Incentive-Aware Machine Learning for Social Welfare Maximization under Uncertainty

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Dung (Daniel) Ngo

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Zhiwei Steven Wu, Maria Gini

November, 2024

Acknowledgements

I want to convey my heartfelt thanks to everyone who has supported me throughout the process of completing this dissertation. I am incredibly grateful to my advisors, Zhiwei Steven Wu and Maria Gini, for their unwavering support and guidance. Their mentorship extended beyond academic advice, equipping me with essential skills in and out of research. Under Steven Wu's direction, I had the opportunity to engage in intriguing projects and collaborate with exceptional individuals. He has also been an outstanding role model for my work ethic and a constant inspiration for my research.

Moreover, the unwavering support from my family and peers at the University of Minnesota and Carnegie Mellon University has been a guiding light during this challenging journey.

Dedicated to my family and friends.

Abstract

Machine Learning algorithms are being deployed to aid the decision-making process of high-stakes domains such as hiring, education, and medical trials. Under the influence of these substantial algorithmic decisions, individuals (*i.e.,* the decision-subjects) are more aware of their role and capabilities in dictating the outcome of Machine Learning algorithms. On the other hand, the social planner (*i.e.,* the decision-maker) deploying these machine learning algorithms has to consider the incentives of individuals who are unwilling to follow the algorithmic decisions given to them unquestioningly. This dissertation focuses on incentive-aware machine learning under uncertainty through the lens of a centralized social planner deploying the algorithms for decision-making.

Naturally, some tension exists between the goals of these different stakeholders: an individual's action often springs from self-interested motives. At the same time, the social planner wants to optimize for better social welfare across all individuals in the population. Therefore, the primary focus of this dissertation is to design incentive-aware Machine Learning algorithms that align with the incentives of the decision-subjects while obtaining the desired total outcomes. Concretely, I study the problem of *incentivized exploration* in the context of online recommender systems, where the social planner can only provide signals to facilitate the individual's action choices. Mainly, I extend the existing literature of *incentivized exploration* in three trajectories: structured action sets, heterogeneous populations, and practical applications.

Beyond incentivized exploration, this dissertation also examines other settings where incentives play an essential role in the agent's decision-making process. First, I study the problem of *strategic learning*, where the decision-subjects are strategic and may modify their input to obtain a more favorable prediction outcome. Secondly, I examine how a centralized server can incentivize participation in a collaborative learning setting. Finally, I investigate the setting where the downstream decision-makers face uncertainty in choosing an accurate and loss-minimizing predictor suitable for their downstream task. These works bring a greater understanding of how to create incentive-aware machine learning algorithms under uncertainty that both (1) align with each decision-subject's belief and (2) maximize cumulative social welfare.

Contents

List of Figures

Chapter 1

Introduction

In modern machine learning (ML) algorithms, learning the latent model under uncertainty is at the heart of the decision-making process. For example, a social planner might want to estimate the effectiveness of a new vaccine. Many studies on causal effect estimation rely on randomized trials, in which the social planner randomly assigns each decision-subject to a control group or a treatment group and then estimates the causal effect by comparing the outcomes across different groups. However, in many real-world domains, participation is voluntary and is often susceptible to non-compliance. For instance, patients may hesitate to try a new vaccine when assigned to receive the treatment. Furthermore, voluntary participation may stem from inherent selection bias, *e.g.,* healthier participants are more willing to test out a new vaccine in a medical trial. A naive assignment of interventions based on these brittle stereotypes is often problematic, as it may prompt individuals to modify their attributes strategically to receive more favorable decisions. As a result, their strategic responses may cause a distributional shift and lead to degraded performance of ML models in social domains.

These issues stem from the need to consider individual incentives in machine learning models more. In this dissertation, I provide an overview of various results in the incentive-aware machine learning literature and show how to resolve the tension of conflicting goals between the agents and the social planner deploying these machine learning models. Specifically, this dissertation addresses the following research question: *how to design incentive-aware machine learning algorithms under uncertainty to maximize social welfare.* I investigate this question through two main principle approaches: (1) resolving the uncertainty in decision-subjects' belief about the current state of the system, and (2) resolving the uncertainty in the social planner's intervention choice given the decision-subjects' incentives to best-respond.

Concretely, this dissertation first focuses on the problem of *incentivized exploration*: how to incentivize self-interested individuals (henceforth, *agents*) to explore in an online learning setting (Chapter 3, 4, 5, and 6). In this problem, a social planner interacts with *myopic* agents by providing recommendations but cannot force the agents to always comply with these recommendations. The incentive constraint comes from the agents' need for more certainty about the available alternatives. Specifically, the social planner wants the agents to trade off *exploration* to acquire new information and *exploitation*, making optimal near-term decisions based on the current information. Conversely, the agents prefer to *exploit* and maximize their expected benefits. The critical insight in this setting is that the social planner can incentivize the agents to *explore* by leveraging the essential notion of *information asymmetry*: the social planner observes the entire (long-term) history of action choices and realized reward from the previous agents. In contrast, each agent only knows about their own (short-lived) interaction with the social planner. The literature on incentivized exploration was initialized by Kremer et al. (2014b) and Slivkins (2019a, Ch. 11) offers a comprehensive overview of the problem in various settings.

This dissertation expands on the base model of incentivized exploration in multiple trajectories. In Chapter 3, I study how the social planner can incentivize exploration when the action set is large and structured in the combinatorial semi-bandit setting. Here, the central theme is to take advantage of such a known structure to enable the incentive-compatible algorithm to work with many feasible actions. In Chapter 4, I investigate how to incentivize agents from different subpopulations with different beliefs over the rewards of each action to explore less-known options. We later leverage the idea and techniques in this work to study synthetic control for the panel data setting, where the goal is to estimate counterfactual outcomes for a group of units under different interventions (Chapter 5). Previous work on synthetic control assumes that the test unit's potential outcomes may be expressed 'reasonably well' by the observed outcomes of the donor units. However, despite their ubiquity, such an assumption may only sometimes hold when the test and donor units come from different subpopulations with different beliefs about the rewards. My contribution here is an algorithm that incentivizes units who initially prefer treatment to take control in the post-treatment period instead, thus alleviating the need for the 'overlap' assumption described above. Finally, I leverage the key ideas and techniques from the literature of incentivized exploration to solve incentive issue in determining the optimal matching of two-sided matching markets (Chapter 6, where the agents and the arms are self-interested individuals with personal preferences over their match. In this two-sided matching setting, it is generally insufficient to consider only the incentives from the agents, as is the case in prior work in incentivized exploration. Hence, the main goal is to provide an algorithm that incentivizes both sides of the matching markets to comply with the social planner's (in this case, the matching platform) recommendation and explore alternative matches.

Beyond incentivized exploration for myopic agents (and arms), this dissertation also considers other online learning settings where incentives play a crucial role in decision-making. One such notable setting is that of *strategic learning*: how to recover the causal relationship between observable attributes and outcome of interest in evaluating strategic individuals (Chapter 7). A social planner interacts with strategic agents, who, upon receiving an intervention, have incentives to modify their observable features to obtain a more favorable predicted outcome. A challenge in this setting lies in the unobserved confounding variables that simultaneously influence the agents' observable attributes and outcomes. One key observation is that we can adopt the technical tools from the literature of instrumental variable regression in Chapter 4 to account for the effect of these confounders by viewing the sequence of interventions (*e.g.,* assessment rules in this setting) as valid *instruments* which only affect the observable features but do not directly influence their outcomes. This problem has been studied in the context of performative prediction Perdomo et al. (2020) for a broad setting in which predicting influences the prediction target and generally lies within the intersection of machine learning and causality.

Another practical setting where a decision-maker wants to learn the latent model under uncertainty is centralized *collaborative learning*, where a set of clients (*e.g.,* mobile devices, hospitals) want to jointly learn a model without directly sharing their data (Chapter 8). However, the costs of joining collaborative training (*e.g.,* latency in the network, privacy cost from sharing private information) may dissuade clients from participating in the learning process. Here, the goal is to study the dynamic of the client's participation incentive in collaborative learning to help the service provider design better solutions and clients make more well-informed decisions.

Finally, in Chapter 9, I study the uncertainty in the social planner's intervention choice when the downstream decision-subjects have incentives to best-respond to this intervention through the problem of *model multiplicity*. In many modern domains, a ubiquitous task is to learn an underlying model to make individual probability predictions. For example, a hospital may want to predict the probability that a patient will contract a disease using an image dataset of skin lesions. In this setting, it is common for the hospital to have access to multiple predictive models with nearly equivalent accuracy but disagree on their predictions on many individual samples. This disagreement in outcome prediction may disparate impact the subpopulation if the decision-maker has to choose one predictor over another to make critical downstream decisions. For example, the hospital may have to rely on a single predictor to suggest treatment for each patient. Prior work in model multiplicity has mainly focused on resolving prediction disagreements between models. Here, I aim to leverage tools from the literature of multi-calibration to update predictors such that they approximately agree on (1) individual predictions and (2) best-response actions for each individual in the downstream decision-making tasks.

Chapter 2

Background and Related Work

2.1 Background: Incentivized Exploration

In this section, I provide the general framework of incentivizing exploration in the multi-armed bandits setting as a gentle introduction to the central theme of my research. In multi-armed bandits, we consider the interaction between a *principal* (*i.e.,* the social planner) and a sequence of *agents* over a time horizon T. The set of feasible arms (actions) $\mathcal{A}$ is known and fixed. In each round, a new agent with index t arrives and chooses an action a_t from $\mathcal{A}$. After selecting this action, agent t receives a reward r_t for this action (and nothing else). We are interested in a setting where agents are myopic self-interested individuals and prefer to 'exploit', *i.e.,* choose an action that maximizes their expected reward. At the same time, the principal wants to maximize the *social welfare*, *i.e.,* the cumulative reward from all agents across the time horizon. Hence, the role of the principal is to provide each agent with a signal containing information about interactions with past agents in the form of a recommendation to encourage them to explore alternative options. This tension of explore-exploit tradeoff between the principals and the agents has been studied in Mansour et al. (2015a) for the classical setting of multi-armed bandits.

In this dissertation, I discuss multiple extensions of the standard incentivized exploration problem in multi-armed bandits to capture the population heterogeneity and a more complex structure of the action-reward formulation. Specifically, my contributions to the incentivized exploration problem are:

1. In Chapter 3, I study the incentivized exploration problem in combinatorial semi-bandits, a pragmatic, well-studied setting with structured actions. Each arm is a subset of some ground set, whose elements are called *atoms*. In each round, the algorithm chooses an arm and observes/collects a reward for each atom in this arm. The reward for each atom is drawn independently from some fixed (but unknown) distribution specific to this atom. The set of feasible arms reflects the structure of the problem, *e.g.,* it can comprise all subsets of atoms of a given cardinality or all edge paths in a given graph. Since the number of arms (K) can be exponential in the number of atoms (d), the central theme is replacing the dependence on K in regret bounds for “unstructured" K-armed bandits with a similar dependence on d.

2. In Chapter 4, I focus on the incentivized exploration problem in the multi-armed bandit setting when the population is heterogeneous, *i.e.,* agents have different beliefs about the reward of choosing each action. First, I study a stylized setting with binary actions, *i.e.,* control (action 0) and treatment (action 1) when there are two *types* of agents, where each type initially prefers a different action. That is, agents of type 0 prefer control, and agents of type 1 prefer treatment initially. For example, in COVID-19 vaccine trials, healthier patients in the population may prefer to take the vaccine (treatment). In contrast, some less healthy patients might be more skeptical about the vaccine's effectiveness and would initially prefer not to take it (and prefer control instead). Furthermore, I would extend this framework to allow the agents to choose between finitely many treatments with unknown effects (and no control intervention). The results in this direction are a set of incentive-aware algorithms that can convince agents who initially are non-compliant to explore different actions and achieve sub-linear regret over the time horizon T.

3. In Chapter 5, I study the incentivized exploration problem in the panel data setting where there are non-overlapping sub-populations with different prior beliefs and preferences. Prior work in synthetic control methods assumes that the outcome of a test unit (agent) can be expressed 'reasonably well' in terms of the outcome of donor units. However, this assumption may not hold in settings where units belong to non-overlapping sub-populations. In this work, my goal is to provide an algorithm that can incentivize units that initially prefer treatment to instead take

control in the post-treatment period, which allows valid counterfactual estimation of the outcome for these units.

4. Finally, in Chapter 6, I focus on a practical setting of two-sided matching markets, where both sides of the market have preferences over their potential match. Prior work in incentivized exploration for matching markets mainly focuses on resolving the incentive constraint on the *agent's side*, as the arms are often inanimate objects (*e.g.,* restaurants, movies, treatment choices). However, we observe that it is generally insufficient only to incentivize one side of the matching market to comply with the platform's recommendation, thus requiring a more careful study of *two-sided incentive constraints* in this setting. The main contribution of this work is a novel notion of incentive compatibility and an oracle-efficient recommendation policy that can guarantee full compliance for two-sided matching markets.

2.2 Related Work

Incentivized Exploration Our work draws on techniques from the growing literature of incentivizing exploration (IE) Kremer et al. (2013); Mansour et al. (2015b, 2016); Immorlica et al. (2019); Sellke and Slivkins (2020), where the goal is also to incentivize myopic agents to explore arms in a multi-armed bandit setting (Auer et al., 2002) using information asymmetry techniques from Bayesian persuasion Kamenica and Gentzkow (2011). While our mechanisms are technically similar to those in Mansour et al. (2015b), our work differs in several key aspects. First, prior work in IE —including Mansour et al. (2015b)— does not capture selection bias and cannot be directly applied in our setting to recover causal effects. The mechanism in Mansour et al. (2015b) aims to enforce full compliance (also called *Bayesian incentive-compatibility*) that requires all agents to follow the planner's recommendations: as a result, the mechanism needs to cater to the type of agents that are most difficult to convince. By contrast, our mechanism relies only on the compliance of a partial subset of agents in order to obtain accurate estimates.

There has also been a line of work on mechanisms incentivizing exploration via payments Frazier et al. (2014b); Chen et al. (2018a); Kannan et al. (2017). Several known disadvantages of such payment mechanisms include potential high costs and ethical concerns Groth (2010). See Slivkins (2017) for a detailed discussion.

Most related to our combinatorial semi-bandit result in Chapter 3 is Sellke and Slivkins (2021b), which obtains similar BIC results for the special case of multi-armed bandits with independent priors, both for Thompson Sampling and for initial exploration. A yet unpublished working paper of Simchowitz and Slivkins (2021b) provides a BIC algorithm for initial exploration in reinforcement learning; we build on this result in one of ours. Similar but technically incomparable versions have been studied, *e.g.,* with time-discounted rewards and creating incentives via money (Frazier et al., 2014a).

On the machine learning side, this result is related to the work on combinatorial semi-bandits, starting from György et al. (2007), *e.g.,* (Chen et al., 2013; Kveton et al., 2015, 2014), and the work on Thompson Sampling, see Russo et al. (2018) for a survey. In particular, near-optimal Bayesian regret bounds have been derived in Russo and Van Roy (2014, 2016), and frequentist ones in (Agrawal and Goyal, 2017; Kaufmann et al., 2012). Thompson Sampling has been applied to combinatorial semi-bandits, (*e.g.,* Gopalan et al., 2014; Wen et al., 2015; Degenne and Perchet, 2016; Wang and Chen, 2018), with Bayesian regret bounds derived in Russo and Van Roy (2016).

Instrumental Variable In Chapter 4, we design mechanisms that strategically select instruments to incentivize compliance over time so that we can apply tools from IV regression (Angrist and Krueger, 2001; Angrist and Imbens, 1995; Imbens et al., 1996) to estimate causal effects. Although IV regression is an established tool to estimate causal effects where there is non-compliance in observational studies (see e.g. Bloom et al. (1997a); Angrist (2005)), our results deviate significantly from previous works, due to the dynamic nature of our model. In particular, even if all agents are initially non-compliant, our mechanism can still incentivize compliance over time and estimate treatment effects —whereas directly applying standard IV regression at the onset cannot.

Thematically, our work relates to work on "instrument-armed bandits" by Kallus (2018), which also views arm recommendations as instruments. However, the compliance behavior (modeled as a fixed stochastic mapping from instrument to treatments) is static in Kallus (2018): it does not change over time —even if the planner has obtained accurate estimate(s) of the treatment effect(s). By comparison, since all agents eventually become compliant in our setting, we can achieve sublinear regret w.r.t. the best treatment, which is not achievable in a static compliance model.

Causal Inference and Synthetic Control Methods In Chapter 5, we study incentivized exploration in panel data setting. Popular methods for counterfactual inference using panel data include synthetic control methods Abadie and Gardeazabal (2003); Abadie et al. (2010), difference-in-differences Angrist and Pischke (2009); Ashenfelter and Card (1984); Bertrand et al. (2004), and clustering-based methods Zhang et al. (2019); Dwivedi et al. (2022). Within the literature on synthetic control, our work builds off of the line of work on *robust* synthetic control Amjad et al. (2018, 2019); Agarwal et al. (2020a,b, 2023), which assumes outcomes are generated via a *latent factor model* (e.g. Chamberlain (1984); Liang and Zeger (1986); Arellano and Honore (2000)) and leverages *principal component regression* (PCR) Jolliffe (1982); Massy (1965) to estimate unit counterfactual outcomes. Our work falls in the small but growing line of work at the intersection of synthetic control methods and online learning Chen (2023); Farias et al. (2022); Agarwal et al. (2023), although we are the first to consider unit incentives in this setting. The model used by Agarwal et al. (2023) is particularly relevant to our work, which extends the finite sample guarantees from PCR in the panel data settings to online settings. While Harris et al. (2022) also consider incentives in synthetic control methods (albeit in an offline setting), they consider a principal who can *assign* interventions to units (e.g. can force compliance). As a result, the strategizing they consider is that of units that modify their pre-intervention outcomes to be assigned a more desirable intervention. In contrast, we consider a principal who cannot assign interventions to units, but instead must *persuade* units to take different interventions by providing them with incentive-compatible recommendations.

More broadly, there has been recent interest in causal inference in characterizing optimal treatment policies, which must satisfy some additional constraint(s). For example, Luedtke and van der Laan (2016) study a setting in which the treatment supply is limited. Much like us, Qiu et al. (2021) consider a setting where intervening in the treatment is impossible, but encouraging treatment is feasible. However, their focus is on the setting where the treatment is a limited resource, so their goal is to encourage treatment to those who would benefit most. Since they do not consider panel data and place different behavioral assumptions on the individuals under intervention, the tools and techniques we use to persuade units in our setting differ significantly from theirs. Finally, Qiu et al. (2022); Sun et al. (2021) consider settings with some uncertain *cost* associated with treatment.

Two-sided matching market. The literature on the two-sided matching market is first studied in the seminal work by (Gale and Shapley, 1962). The two-sided matching market has many applications, ranging from streaming platforms to payment systems (Rysman and Wright, 2014) and loan market (Chen and Song, 2013). For a broad overview of these applications, see Rochet and Tirole (2003). The formulation of the two-sided matching problem as a combinatorial semi-bandits problem has been studied by Kasy and Teytelboym (2022). There is a line of work on incentivized exploration in two-sided markets (Li et al., 2024b; Dai et al., 2022; Li et al., 2024a). However, similar to other prior work in incentivized exploration, they only consider the agents' incentives in their algorithms. However, many real-world applications of two-sided matching markets have human decision-subjects on both sides whose incentives need to be considered when the platform designs a matching algorithm. In Section 6.3, we describe a counterexample to illustrate the necessity of a novel incentive mechanism design for two-sided matching markets.

Strategic Learning An active area of research on *strategic learning* aims to develop machine learning algorithms that are capable of making accurate predictions about decision subjects even if they respond *strategically* and potentially *untruthfully* to the choice of the predictive model (Dong et al., 2018; Hardt et al., 2016; Mendler-Dünner et al., 2020; Shavit et al., 2020; Chen et al., 2020; Cummings et al., 2015; Cai et al., 2015; Meir et al., 2012; Hu et al., 2019). Generalizing strategic learning, Perdomo et al. (2020) propose a framework called *performative predictions*, which broadly studies settings where the act of predicting influences the prediction target. Several recent papers have investigated the relationship between strategic learning and causality (Shavit et al., 2020; Bechavod et al., 2021; Miller et al., 2020).

The setting most similar to our result in Chapter 7 is that of Shavit et al. (2020). They consider a strategic classification setting in which an agent's outcome is a linear function of features –some observable and some not. While they assume that an agent's latent attributes can be modified strategically, we model the agent as having an unmodifiable private *type*. Both of these assumptions are reasonable, and some domains may be better described by one model than the other. For example, the model of Shavit et al. may be useful in a setting such as car insurance pricing, where some unobservable factors related to safe driving are modifiable. On the other hand, our model captures settings like university admissions, where confounding factors (e.g., socioeconomic background) are not easily modifiable. Both models are special cases of a broader causal graph. Note that in the model of Shavit et al., $\boldsymbol{\theta}$ violates the backdoor criterion and therefore cannot serve as a valid instrument. Bechavod et al. (2021) consider a setting simpler than ours in which there are no confounding effects from agents' unobserved types on their observable features and outcomes. As a result, the authors can apply standard least squares regression techniques to recover causal parameters.

This work is also related to Miller et al. (2020) shows that designing good incentives for agent improvement in strategic classification is at least as hard as orienting edges in the corresponding causal graph. In contrast to their work, we observe that the assessment rule deployed by the principal can be actively used as a valid *instrument*, which allows us to circumvent this hardness result by performing an *intervention* on the causal graph of the model.

Incentives in collaborative learning. Prior works that study incentive mechanisms in the context of collaborative learning typically aim to encourage clients to join in training in a one-time setting. Here, we list the most relevant works and defer readers to Tu et al. (2022) for a comprehensive discussion of work in the area. Works on collaborative learning such as Blum et al. (2017); Haghtalab et al. (2022); Sim et al. (2020) describe the approach of handling heterogeneous data to learn a common concept by iteratively gathering more data. For the application of federated learning, Donahue and Kleinberg (2021c,a,b) analyzes the coalition structure for federated mean estimation and linear regression problems using hedonic game theory, and Cho et al. (2022) proposes a new FL objective that aims at increasing the number of incentivized clients. Our method differs from the above works in that apart from utility gain; our work considers the cost of joining the federated network as a factor that can dynamically affect client behavior.

More related to our result in Chapter 8 are the works of Blum et al. (2021), which analyzes a framework for incentive-aware data sharing in FL under the notion of envy, and Karimireddy et al. (2022), which introduces an accuracy-based algorithm to maximize the amount of data contributed by each client to learn their local mean parameter. While these two works share some similarities with ours in that they encode some form of joining cost, our method focuses on studying the population dynamics and analyzing the equilibrium when clients best respond to public information.

Finally, we note that the use of financial incentives as a means of motivating clients has been a topic of extensive research. Various economic approaches have been proposed to incentivize participation monetarily, e.g., via auctions/reversed auctions (Kim, 2020; Thi Le et al., 2021), contract theory approaches Saputra et al. (2020), public good approaches Tang and Wong (2021). Our work adds a new perspective to this area by studying how participation changes over time and considering payments only at tipping points of the dynamic.

Network effects. Our work draws inspiration from literature on network effects games (e.g., Katz and Shapiro, 1985; Shapiro and Varian, 1999; Easley and Kleinberg, 2010), where the utility of a client using a service depends on the number of other clients participating in the same service. This effect is often positive, as clients gain more utility from others joining the network. Our model considers accuracy improvement from having more data in a coalition as a natural incentive mechanism. As clients join the FL process, they expect to obtain a more accurate estimator from the combined data of all clients in the coalition. To the best of our knowledge, we are the first to model network effects in FL in this manner, and our work is the first to study the behavior of rational clients in FL under the network effects model. In formulating and analyzing this dynamic perspective on client incentives, we hope our work can lead to future study and development of FL incentive mechanisms that more closely mirror practical FL applications.

Model multiplicity. Within the literature on predictive diversity, our work builds off the line of work focusing on predicting individual probabilities (Marx et al., 2020; D'Amour et al., 2022; Black et al., 2022; Breiman, 2001), where solving an error minimization problem for some prediction tasks can lead to multiple solutions with roughly similar performance in terms of accuracy. Sandroni (2003) showed that one cannot empirically distinguish the outcomes from a predictor encoding the true individual probabilities from one without in isolation, while Al-Najjar and Weinstein (2008); Feinberg and Stewart

(2008) provided comparative tests to differentiate between the true probability predictor and one that is not. Mainly, Feinberg and Stewart (2008) relied on *cross-calibration*, *i.e.,* calibration conditional on the predictions of both models to falsify one of them empirically. For downstream decision-making, Garg et al. (2019) worked on refining predictors and provided an algorithm that produces a predictor f_3 that is cross-calibrated with respect to both f_1 and f_2. An alternative framework studied by Globus-Harris et al. (2022b) seeks to update sub-optimal models for different subsets of the population, following the 'bug bounties' approach used by the software and security communities. Towards the same goal, Roth et al. (2023) proposed an algorithm ("Reconcile") that aims to reconcile different predictors with equivalent errors such that the updated predictors both have lower errors compared to the initial models and approximately agree on their prediction on almost all units. Our work follows the reconciliation idea from Roth et al. (2023) but focuses on predictive accuracy and downstream decision-making. Our model considers reconciling predictors for regression and multi-class classification problems and their impact on the downstream decision-making tasks. In Section 9.2.4, we provide a numerical example where simply reconciling the probability predictions according to Roth et al. (2023) can lead to additional losses in downstream decision-making tasks.

Multi-calibration. Our work draws on techniques from the growing literature on multi-calibration (Hebert-Johnson et al., 2018; Kim et al., 2019; Dwork et al., 2019; Shabat et al., 2020; Jung et al., 2020; Dwork et al., 2021; Jung et al., 2022; Haghtalab et al., 2023; Deng et al., 2023; Noarov et al., 2023; Globus-Harris et al., 2022a, 2023). Multi-calibration has been used as a notion of fairness as it guarantees calibration for any identifiable group. Notably, we leverage the connection between multi-calibration and loss minimization from the omniprediction literature (Gopalan et al., 2022b,a, 2021) in our analysis. Within the framework of multi-calibration, the work most related to ours is that of Zhao et al. (2021), who considered decision calibration with respect to all classes of loss functions. Similar to us, Zhao et al. (2021) takes the perspective of a decision-maker who wants to ensure the predictive models are *indistinguishable* from the true probability when they are used to make downstream decisions. However, two decision-calibrated models can still disagree on their individual predictions and best-response actions for many units. In our motivating example, a hospital with two decision-calibrated predictors may still want to make their predictive models approximately agree on their predictions for almost all individuals in the population and lead to the same downstream decision, *i.e.,* whether to recommend treatment or not based on the predicted probability that a patient has contracted a disease. We formalize this example in Section F.I and provide empirical experiments to show our improvement over their result in Section 4.7.

An independent and concurrent work by Globus-Harris et al. (2024) considers ensembling multiple predictive models for high-dimensional downstream decision-making tasks. While their work also leverages techniques from multi-calibration, their goal is to output an ensemble predictor whose self-estimated expected payoff is accurate and whose induced policy has a payoff at least as high as the maximum self-assessed payoff of individual models. In contrast, our interest is in reducing the downstream decision losses by resolving the differences between equivalent predictors and mitigating model multiplicity for decision-making.

Chapter 3

Incentivizing Combinatorial Bandit Exploration

In this chapter, we focus on the first extension of *incentivized exploration*: **How to incentivize exploration in combinatorial semi-bandits with a large, structured action set**.

3.1 Introduction

We consider *incentivized exploration*: how to incentivize self-interested users to explore. A social planner interacts with self-interested users (henceforth, *agents*) and can make recommendations, but cannot enforce the agents to comply with these recommendations. The agents face uncertainty about the available alternatives. The social planner would like the agents to trade off *exploration* for the sake of acquiring new information and *exploitation*, making optimal near-term decisions based on the current information. The agents, on the other hand, prefer to *exploit*. However, the algorithm can incentivize them to *explore* by leveraging the information collected from the previous users. This problem has been studied since Kremer et al. (2014a), see Slivkins (2019b, Ch. 11) for an overview.

The basic model of incentivized exploration is very stylized and can be extended along two "dimension": more sophisticated economic models for agents' behavior and incentives, and more complex machine-learning models for actions' structures and rewards.

All published work has only dealt with the former, whereas we pursue the latter. In particular, all published work focuses on small, unstructured action sets. Moreover, the case of *independent priors* – when the users' beliefs are independent across actions – is emphasized as the main, paradigmatic special case when specific performance guarantees are derived. However, realistic exploration problems often feature large sets with some known structure that connects actions to one another. A major recurring theme in the vast literature on multi-armed bandits is taking advantage of the available structure so as to enable the algorithm to cope with a large number of actions.

We focus on a paradigmatic, well-studied exploration problem with structured actions: *combinatorial semi-bandits*. Here, each arm is a subset of some ground set, whose elements are called *atoms*. In each round, the algorithm chooses an arm and observes/collects a reward for each atom in this arm. The reward for each atom is drawn independently from some fixed (but unknown) distribution specific to this atom. The set of feasible arms reflects the structure of the problem, *e.g.,* it can comprise all subsets of atoms of a given cardinality or all edge paths in a given graph. Since the number of arms (K) can be exponential in the number of atoms (d), the main theme is replacing the dependence on K in regret bounds for "unstructured" K-armed bandits with a similar dependence on d.

We adopt a standard model for incentivized exploration from Kremer et al. (2014a). The social planner is implemented as a bandit algorithm. Each round corresponds to a new agent which arrives and receives the arm chosen by the algorithm as a recommendation. Agents have Bayesian beliefs, independent across the atoms (but highly correlated across the arms). The algorithm must ensure that following its recommendation is in each agent's best interest, a condition called *Bayesian incentive-compatibility* (*BIC*). Each agent does not observe what happened with the previous agents, but the algorithm does. This information asymmetry is crucial for creating incentives.

Our contributions. We prove that Thompson Sampling is BIC when initialized with at least $n_{\texttt{TS}}$ samples of each atom, where $n_{\texttt{TS}}$ is determined by the prior and scales polynomially in the number of atoms (d). Thompson Sampling (Thompson, 1933) is a well-known bandit algorithm with near-optimal regret bounds and good empirical performance. The initial samples can be provided by another BIC algorithm (more on this below), or procured exogenously, *e.g.,* bought with money.

Next, we consider the problem of *initial exploration*: essentially, design a BIC algorithm that samples each atom at least once. Such algorithms are interesting in their own right and can be used to bootstrap Thompson Sampling, as per above. We present two such algorithms, which build on prior work (Mansour et al., 2020a; Simchowitz and Slivkins, 2021a) and extend it in non-trivial ways. The objective to be optimized is the sufficient number of rounds T_0, and particularly its dependence on d. To calibrate, prior work on incentivized exploration in multi-armed bandits with correlated priors does not provide any guarantees for a super-constant number of arms (K), and is known to have $T_0 > \exp(\Omega(K))$ in some natural examples (Mansour et al., 2020a). In contrast, our algorithms satisfy $T_0 \leq \exp(O(d))$ for a paradigmatic special case, and $T_0 \leq \exp(O(d^3))$ in general.

Finally, what if the prior is *not* independent across atoms? We focus on two arms with arbitrary correlation, a fundamental special case of incentivized exploration, and prove that our analysis of Thompson Sampling extends to the case. This result may be of independent interest.

Discussion. Like all prior work on incentivized exploration, we consider standard, yet idealized models for agents' economic behavior and the machine-learning problem being solved by the social planner. The modeling captures something essential about exploration and incentives in recommendation systems, but is not supposed to capture all the particularities of any specific application scenario. The goal of this paper is to bring more complexity into the machine-learning problem; advancing the economic model is beyond our scope.

We focus on establishing the BIC property and asymptotic guarantees in terms of the number of atoms, without attempting to optimize the dependence on the per-atom Bayesian priors. Our Thompson Sampling result has an encouraging practical implication: a standard, well-performing bandit algorithm plays well with users' incentives, provided a small (in theory) amount of initial data.

(Stylized) motivating examples of incentivized exploration in combinatorial semi-bandits include: recommending online content, *e.g.,* for news or entertainment (with atoms as *e.g.,* specific news articles); recommending complementary products, *e.g.,* a suit that consists of multiple items of clothing; recommending driving directions. In all examples, the social planner corresponds to the online platform issuing the respective recommendations. Such online platforms are often interested in maximizing users' happiness, rather than (or in addition to) the immediate revenue, as a way to ensure user engagement and long-term success.

3.2 Problem Formulation and Preliminaries

Our algorithm operates according to the standard protocol for *combinatorial semi-bandits*, with an ancillary incentive-compatibility constraint, standard in the literature on *incentivized exploration*.

Combinatorial semi-bandits. There are T rounds, d atoms and K arms, where each arm is a subset of atoms. The set $\mathcal{A}$ of feasible arms is fixed and known. In each round t, each atom ℓ generates reward $r_\ell^{(t)} \in [0,1]$. The algorithm chooses an arm $A^{(t)} \in \mathcal{A}$ and observes the reward of each atom in this arm (and nothing else). Algorithm's reward in this round is the total reward of these atoms.

Formally, we write $[T] := \{1\,,\; \dots\;, T\}$ for the set of all rounds and $[d]$ for the set of all atoms, so that arms are subsets $A \subset [d]$. Let θ_ℓ be the expected reward of atom $\ell \in [d]$, and let $\mu(A) = \sum_{\ell \in A} \theta_\ell$ be the expected reward of a given arm $A \subset [d]$. Note that d-armed bandits are a special case when the feasible arms are singleton sets $\{\ell\}$, $\ell \in [d]$.

Stochastic rewards and Bayesian priors. The reward of each atom $\ell \in [0,1]$ is drawn independently in each round from a fixed distribution $\mathcal{D}_\ell$ specific to this atom. This distribution comes from a parametric family, parameterized by the expected reward θ_ℓ. The (realized) problem instance is therefore specified by the *mean reward vector* $\theta = (\theta_1\,,\; \dots\;, \theta_d)$. Initially, each θ_ℓ is drawn independently from a Bayesian prior $\mathcal{P}_\ell$ with support $\Theta \subset [0,1]$. Put differently, the mean reward vector θ is drawn from the product prior $\mathcal{P} = \mathcal{P}_1 \times \cdots \times \mathcal{P}_d$.

Incentive-compatibility. The algorithm must ensure that in each round t, conditional on a particular arm $A = A^{(t)}$ being chosen, the expected reward of this arm is at least as good as that of any other arm. Formally, the algorithm is called *Bayesian incentive-compatible* (BIC) if for each round $t \in [T]$,

$$\mathbb{E}[\, \mu(A) - \mu(A') \mid A^{(t)} = A \,] \geq 0 \qquad \forall \text{ arms } A, A' \in \mathcal{A} \text{ with } \Pr[A^{(t)} = A] > 0. \quad (3.1)$$

This definition is based on the following stylized story . In each round t, a new user arrives to a recommendation system, observes the arm $A^{(t)}$ chosen by our algorithm, and interprets it as a recommendation. Then the user decides which arm to choose (not necessarily the arm recommended), and receives the corresponding reward. Accordingly, the user needs to be incentivized to follow the recommendation. We adopt a standard setup from economic theory (and the prior work on incentivized exploration): each user has the same prior $\mathcal{P}$, knows the algorithm, and wishes to maximize her expected reward. We posit that the user does not observe anything else before making her decision, other than the recommended arm. In particular, she does not observe anything about the previous rounds. Then, (3.1) ensures that she is (weakly) incentivized to follow her recommendation, assuming that the previous users followed theirs. We posit that under (3.1), the user does follow recommendations, and then reports the rewards of all atoms to the algorithm.

We emphasize that this story is not a part of our formal model (although it can be expressed as such if needed). In fact, the story can be extended to allow the algorithm to reveal an arbitrary “message" to each user, but this additional power is useless: essentially, anything that can be achieved with arbitrary messages can also be achieved with arm recommendations. This can easily be proved as a version of Myerson's *direct revelation principle* from theoretical economics.

Conventions. Each atom ℓ satisfies $\Pr[\theta_\ell > 0] > 0$: else, its rewards are all 0, so it can be ignored.

No arm is contained in any the other arm. This is w.l.o.g. for Bernoulli rewards, and more generally if $\Pr[r_\ell^{(t)} = 0 \mid \theta_\ell > 0] > 0$: if $A \subset A'$ for arms A, A' then A cannot be chosen by any BIC algorithm.

W.l.o.g., order the atoms ℓ by their prior mean rewards $\theta_\ell^0 := \mathbb{E}[\theta_\ell]$, so that $\mathbb{E}[\theta_1^0] \geq \cdots \geq \mathbb{E}[\theta_d^0]$.

Let $A^* = \operatorname{argmax}_{A \in \mathcal{A}} \mu(A)$ denote the best arm overall, with some fixed tie-breaking rule.

By a slight abuse of notation, each arm $A \subset [d]$ is sometimes identified with a binary vector $v \in \{0,1\}^d$ such that $v_\ell = 1 \Leftrightarrow \ell \in A$, for each atom $\ell \in [d]$. In particular, we write $A_\ell = v_\ell$.

Thompson Sampling has a very simple definition, generic to many versions of multi-armed bandits. Let $\mathcal{F}_t$ denote the realized history (tuples of chosen actions and realized rewards of all atoms) up to and not including round t. Write $\mathbb{E}^{(t)}[\,\cdot\,] = \mathbb{E}[\,\cdot \mid \mathcal{F}_t\,]$ and $\Pr^{(t)}[\,\cdot\,] = \Pr[\,\cdot \mid \mathcal{F}_t\,]$ as a shorthand for posterior updates. Thompson Sampling in a given round t draws an arm independently at random from distribution $p^{(t)}(A) = \Pr^{(t)}[A^* = A]$, $A \in \mathcal{A}$. If Thompson Sampling is started from some fixed round $t_0 > 1$, this is tantamount to starting the algorithm from round 1, but with prior $\mathcal{P}(\cdot \mid \mathcal{F}_{t_0})$ rather than $\mathcal{P}$. The algorithm is well-defined for an arbitrary prior $\mathcal{P}$.

While this paper is not concerned with computational issues, they are as follows. The posterior update $\mathcal{P}(\cdot \mid \mathcal{F}_t)$ can be performed for each atom ℓ separately: $\mathcal{P}_\ell(\cdot \mid \mathcal{F}_{t,\ell})$, where $\mathcal{F}_{t,\ell}$ is the corresponding history of samples from this atom. A standard implementation draws $\theta'_\ell \in [0,1]$ independently from $\mathcal{P}_\ell(\cdot \mid \mathcal{F}_{t,\ell})$, for each atom ℓ, then chooses the best arm according to these draws: $\operatorname{argmax}_{A\in\mathcal{A}} \sum_{\ell\in A} \theta'_\ell$. The posterior updates $\mathcal{P}_\ell(\cdot \mid \mathcal{F}_{t,\ell})$ and the argmax choice are not computationally efficient in general, and may require heuristics (this is a common situation for all variants of Thompson Sampling). A paradigmatic special case is Beta priors $\mathcal{P}_\ell$ and Bernoulli reward distributions $\mathcal{D}_\ell$, so that the posterior update $\mathcal{P}_\ell(\cdot \mid \mathcal{F}_{t,\ell})$ is another Beta prior.

Composition of BIC algorithms. We rely on a generic observation from Mansour et al. (2020a, 2022) that the composition of two BIC algorithms is also BIC.

Lemma 3.2.1. *Let* `ALG` *be a BIC algorithm which stops after some round T_0. Let* `ALG'`(H) *be another algorithm that initially inputs the history H collected by* `ALG`*, and suppose it is BIC. Consider the composite algorithm:* `ALG` *followed by* `ALG'`(H)*, which stops at the time horizon T. If T_0 is determined before the composite algorithm starts, then this algorithm is BIC.*

3.3 Thompson Sampling is BIC

Our main result is that Thompson Sampling is BIC when initialized with at least $n_{\texttt{TS}}$ samples of each atom, where $n_{\texttt{TS}}$ is determined by the prior and scales polynomially in d, the number of atoms.

Theorem 3.3.1. *Let* `ALG` *be any BIC algorithm such that by some time $T_0 \le T$ (determined by the prior) it almost surely collects at least $n_{\texttt{TS}} = C_{\texttt{TS}} \cdot d^2 \cdot \varepsilon_{\texttt{TS}}^{-2} \cdot \log(\delta_{\texttt{TS}}^{-1})$ samples* *from each atom, where*

$$\varepsilon_{\texttt{TS}} = \min_{A,A' \in \mathcal{A}} \mathbb{E}[(\mu(A) - \mu(A'))_+] \quad \textit{and} \quad \delta_{\texttt{TS}} = \min_{A \in \mathcal{A}} \Pr[A^* = A], \tag{3.2}$$

and $C_{\texttt{TS}}$ is a large enough absolute constant. Consider the composite algorithm which runs ALG *for the first T_0 rounds, followed by Thompson sampling. This algorithm is BIC.*

Note that T_0 and $n_{\texttt{TS}}$ are "constants" once the prior is fixed, in the sense that they do not depend on the time horizon T, the mean reward vector θ, or the rewards in the data collected by ALG.

Remark 3.3.2. For statistical guarantees, consider Bayesian regret, *i.e.,* regret in expectation over the Bayesian prior. Bayesian regret of the composite algorithm in Theorem 3.3.1 is at most T_0 plus Bayesian regret of Thompson Sampling. The latter is $O(\sqrt{dT \log d})$ for any prior (Russo and Van Roy, 2016).

Remark 3.3.3. We can invoke Lemma 3.2.1 since T_0 is determined in advance. So, it suffices to show that each round t of Thompson Sampling satisfies the BIC condition (3.1).

Let us clarify the dependence on d. Note that parameters $\varepsilon_{\texttt{TS}}$ and $\delta_{\texttt{TS}}$ may depend on d through the prior. To separate the dependence on d from that on the prior, we posit that each per-atom prior $\mathcal{P}_\ell$ belongs to a fixed collection $\mathcal{C}$. We make mild non-degeneracy assumptions:[1]

$$\Pr\left[\,\mu(A') < \mathbb{E}[\mu(A)]\,\right] > 0 \qquad \text{for all arms } A \neq A'. \tag{3.3}$$

$$\Pr[\theta_\ell > \tau] > 0 \qquad \text{for all atoms } \ell \in [d] \text{ and some } \tau \in (0,1). \tag{3.4}$$

$$\Pr[\theta_\ell < x] > \text{poly}(1/x) \cdot \exp(-x^{-\alpha}) \qquad \text{for all atoms } \ell \in [d],\ x \in (0, 1/2), \text{ some } \alpha \geq 0. \tag{3.5}$$

Corollary 3.3.4. *Suppose all priors $\mathcal{P}_\ell$ of atoms $\ell \in [d]$ belong to some fixed, finite collection $\mathcal{C}$ of priors and assumptions (3.3-3.5) are satisfied with some absolute constants α, τ. Then $n_{\texttt{TS}} = O_{\mathcal{C}}(d^{3+\alpha} \log d)$, where $O_{\mathcal{C}}$ hides the absolute constants and the dependency on $\mathcal{C}$.*

[1] For the special case of d-armed bandits, assumption (3.3) is necessary and sufficient for the respective arm A to be *explorable*: chosen in some round by some BIC algorithm (Sellke and Slivkins, 2021a).

Remark 3.3.5. The initial data can also be provided to Thompson Sampling exogenously (rather than via a BIC algorithm ALG), *e.g.,* purchased with money. More formally, one would need to provide a collection of (arm, atoms' rewards) datapoints such that each atom is sampled at least $n_{\texttt{TS}}$ times.[2]

Proof Sketch for Theorem 3.3.1. In order to establish the BIC condition in (3.1) for Thompson Sampling, we first observe that $\Pr[A^* = A]$ is a positive prior-dependent constant for all arms A, so it suffices to prove $\mathbb{E}\left[\mathbb{E}^{(t)}[\mu(A) - \mu(A')] \cdot \mathbf{1}_{\{A^*=A\}}\right] \geq 0$ for all A, A'.

Next, to show a lower bound on $\mathbb{E}\left[\,(\mu(A) - \mu(A')) \cdot \mathbf{1}_{\{A^*=A\}}\,\right]$, we will leverage the Harris inequality, which says increasing functions of independent random variables are non-negatively correlated. Observe that the functions $(\mu(A) - \mu(A'))_+$ and $\mathbf{1}_{\{A^*=A\}}$ are co-monotone in each coordinate of θ (*i.e.,* either both increasing or both decreasing in a coordinate). Then, the mixed-monotonicity Harris inequality implies that:

$$\mathbb{E}\left[\,(\mu(A) - \mu(A')) \cdot \mathbf{1}_{\{A^*=A\}}\,\right] = \mathbb{E}\left[\,(\mu(A) - \mu(A'))_+ \cdot \mathbf{1}_{\{A^*=A\}}\,\right] \geq \varepsilon_{\texttt{TS}} \cdot \delta_A \tag{3.6}$$

where $\delta_A = \Pr[A^* = A] \geq \delta_{\texttt{TS}}$.

To finish the proof, we show the expected absolute difference between $\mathbb{E}^{(t)}[\mu(A) - \mu(A')] \cdot \mathbf{1}_{\{A^*=A\}}$ and $(\mu(A) - \mu(A')) \cdot \mathbf{1}_{\{A^*=A\}}$ is upper bounded by $\varepsilon_{\texttt{TS}} \cdot \delta_A$. By regrouping and using triangle inequality as well as $|x \cdot \mathbf{1}_{\{A^*=A\}}| = |x| \cdot \mathbf{1}_{\{A^*=A\}}$, we can upper bound this estimation error by sum of $\mathbb{E}\left[\,|\mathbb{E}^{(t)}\left[\,\mu(A)\,\right] - \mu(A)| \cdot \mathbf{1}_{\{A^*=A\}}\,\right]$ and $\mathbb{E}\left[\,|\mathbb{E}^{(t)}\left[\,\mu(A')\,\right] - \mu(A')| \cdot \mathbf{1}_{\{A^*=A\}}\,\right]$. Since the mean reward of each atom can be estimated by their empirical average, we can apply Bayesian Chernoff and observe that these two terms are $n_{\texttt{TS}}^{-1/2}$ times $O(1)$-sub-Gaussian random variables. By the sub-Gaussian tail bound, we upper bound both terms by $O(n_{\texttt{TS}}^{-1/2}\delta_A\sqrt{\log(1/\delta_A)})$. We conclude by using our choices of $n_{\texttt{TS}}$ and observing that $\delta_{\texttt{TS}} \leq \delta_A$. $\square$

Proof Sketch for Corollary 3.3.4. To derive the dependence of $n_{\texttt{TS}}$ on d, we investigate how the prior-dependent constants $\varepsilon_{\texttt{TS}}$ and $\delta_{\texttt{TS}}$ depends on d. First, we can let $\varepsilon_{\mathcal{C}}$ be a version of $\varepsilon_{\texttt{TS}}$ where the min is taken over all ordered pairs of priors in $\mathcal{C}$. Since $\mathcal{C}$ is finite and satisfies the pairwise non-dominance assumption (3.3), $\varepsilon_{\texttt{TS}} \geq \varepsilon_{\mathcal{C}} > 0$.

[2] A subtlety: the number of samples of each arm should be known in advance. This is because otherwise Bayesian update on this data may become dependent on the data-collection algorithm.

By definition, $\delta_{\texttt{TS}} = \min_{A\in\mathcal{A}} \Pr[A^* = A]$ is the minimum probability that arm A is the best arm overall. Fix an arm A. We observe that the event where arm A is the best arm is more likely than the event where each atom in A is larger than τ, and all other atoms not in A is smaller than τ/d. Hence, we can lower bound $\Pr[A^* = A]$ by $\mathbb{E}\left[\mathbf{1}_{\{\forall \ell\in A, \theta_\ell \geq \tau\}} \cdot \mathbf{1}_{\{\forall x\notin A, \theta_x \leq \tau/d\}}\right]$. Since the prior $\mathcal{P}$ is independent across atoms, we can write the expression above as product of $\mathbb{E}\left[\mathbf{1}_{\{\forall \ell\in A, \theta_\ell \geq \tau\}}\right]$ and $\mathbb{E}\left[\mathbf{1}_{\{\forall x\notin A, \theta_x \leq \tau/d\}}\right]$. As the values $\{\theta_\ell\}_{\ell\in[d]}$ are independent and co-monotone in each in coordinate of θ, repeated application of mixed-monotonicity Harris inequality implies that:

$$\begin{aligned}
\Pr[A^* = A] &\geq \prod_{\ell\in A} \mathbb{E}\left[\mathbf{1}_{\{\theta_\ell \geq \tau\}}\right] \cdot \prod_{x\notin A} \mathbb{E}\left[\mathbf{1}_{\{\theta_x \leq \tau/d\}}\right] = \prod_{\ell\in A} \Pr[\theta_\ell \geq \tau] \cdot \prod_{x\notin A} \Pr[\theta_x \leq \tau/d] \\
&\geq \prod_{\ell=1}^{d} \Pr[\theta_\ell \geq \tau] \Pr[\theta_\ell \leq \tau/d]
\end{aligned}$$

By full support assumption (3.4), we define a prior-dependent constant

$$\rho_\tau = \min_{A\in\mathcal{A}} \Pr[\theta_\ell \geq \tau] > 0$$

Then, by definition of ρ_τ and the non-degeneracy assumption (3.5), the expression above is lower bounded by $\rho_\tau^d \cdot \mathrm{poly}(d^d/(\tau)^d) \cdot \exp(-d(\tau/d)^{-\alpha})$. Plugging this bound into n_{TS}, we obtain $n_{TS} = O_{\mathcal{C}}(d^{3+\alpha}\log d)$. $\square$

3.3.1 The two-arm case with arbitrary correlation

What if the prior is *not* independent across atoms? Our analysis extends to the case of $K = 2$ arms A, A' with arbitrary correlation between the atoms. In fact, we do not assume combinatorial semi-bandit structure, and instead focus on the fundamental special case of incentivized exploration: when one has two arms A, A' with arbitrary joint prior on $(\mu(A),\, \mu(A'))$. [3]

Theorem 3.3.6. *The assertion in Theorem 3.3.1 holds for the case when one has two arms A, A' and an arbitrary joint prior on $(\mu(A),\, \mu(A'))$.*

[3] Equivalently, we have $d = 2$ atoms with an arbitrary joint prior on (θ_1, θ_2), and the feasible arms are the two singleton arms $\{1\}$ and $\{2\}$.

This result completes our understanding of incentivized exploration with two correlated arms: indeed, a necessary and sufficient condition (and the algorithm) are known for collecting the initial data (Mansour et al., 2020a). A similar result for two *independent* arms is in (Sellke and Slivkins, 2021a).

The analysis is very similar to that of Theorem 3.3.1, and omitted. Unfortunately, our technique does not appear to extend to a larger number of arbitrarily correlated arms.

3.4 BIC algorithms for initial exploration

We present two BIC algorithms for *initial exploration*, where the objective is to sample each atom at least once (*i.e.,* choose arms whose union is $[d]$) and complete in N_0 rounds for some N_0 determined by the prior. Such algorithms are interesting in their own right and can be used to bootstrap Thompson Sampling as per Theorem 3.3.1. (To collect n samples of each arm, repeat the algorithm n times.) Both algorithms complete in the number of rounds that is exponential in $\text{poly}(d)$. The first algorithm completes in $\exp(O_{\mathcal{P}}(d))$ rounds but is restricted to arms of the same size and Beta-Bernoulli priors. We obtain $\exp(O_{\mathcal{P}}(d^2))$ for arbitrary sets of arms. The second algorithm sidesteps the Beta-Bernoulli restriction but completes in $\exp(O_{\mathcal{P}}(d^3))$ rounds.

3.4.1 Reduction to K-armed bandits

The first algorithm builds a substantial "super-structure" on top of the "hidden exploration" paradigm from Mansour et al. (2020a). The latter paradigm is defined for K-armed bandits and is proved to work for arms with independent priors, ordered by their prior mean rewards. However, for combinatorial semi-bandits, the arms' priors are highly correlated. We provide a new interpretation of their analysis in terms of exploring a generic sequence of arms that satisfies a certain property (P). Our technical contribution is to construct a sequence of arms and prove that it satisfies Property (P). Note that it suffices to explore a sequence of arms which collectively cover all the atoms.

Throughout this subsection, we make the following assumptions:

the prior $\mathcal{P}_\ell$ for each atom ℓ is a Beta distribution with parameters $(\alpha_\ell, \beta_\ell)$; (3.7)

the reward distributions $\mathcal{D}_\ell$ are Bernoulli distributions. (3.8)

This is a paradigmatic special case for Thompson Sampling (and Bayesian inference in general).

Let $\nu_\ell(n) = \alpha_\ell \,/\, (\alpha_\ell + \beta_\ell + n)$, $n \geq 0$ be the posterior mean reward of atom ℓ when conditioned on n samples of this atom such that each of these samples returns reward 0.

Given any number $n \in \mathbb{N}$, let us define a sequence of $\kappa(n) \leq \infty$ arms $V_1^n\,,\ \ldots\ ,V_{\kappa(n)}^n \in \mathcal{A}$. Let V_1 be a prior-best arm: any arm with the largest prior mean reward. The subsequent arms are defined inductively. Essentially, we pretend that each atom in each arm in the sequence so far has been sampled exactly n times and received 0 each time it has been sampled. The next arm is defined as the posterior-best arm: an arm with a largest posterior reward after seing these samples. Formally, for each $i \geq 2$, we define arm V_i^n given the previous arms $V_1^n\,,\ \ldots\ ,V_{i-1}^n$. For each atom $\ell \in [d]$ define $Z_i^n(\ell) = n$ if this atom is contained in one of the previous arms in the sequence, and set $Z_i^n(\ell) = 0$ otherwise. Then, define V_i^n as a the posterior-best arm if the posterior mean rewards for atoms ℓ are given by $\nu_\ell\,(\,Z_i^n(\ell)\,)$. That is:

$$V_i^n \in \operatorname*{argmax}_{A \in \mathcal{A}} \sum_{\ell \in A} \nu_\ell\,(\,Z_i^n(\ell)\,)\,. \tag{3.9}$$

The sequence stops when the arms therein cover all atoms at least once, and continues infinitely otherwise; this defines $\kappa(n)$. [4]

To state Property (P), we focus on this sequence for a particular, prior-dependent choice of n.

(P) There exist numbers $n_\mathcal{P} \in \mathbb{N}$ and $\tau_\mathcal{P}, \rho_\mathcal{P} \in (0,1)$, determined by the prior $\mathcal{P}$, which satisfy the following. Focus on the sequence of arms $V_1\,,\ \ldots\ ,V_\kappa$, where $\kappa = \kappa(n_\mathcal{P})$ and $V_i = V_i^{n_\mathcal{P}}$ for each $i \in [\kappa]$. Let H_i^N, $i \in [\kappa]$ be a dataset that consists of exactly $N \in \mathbb{N}$ samples of each arm $V_1\,,\ \ldots\ ,V_i$, where each sample contains the reward

[4] In Theorem 3.4.2 and Theorem 3.4.4, we upper-bound $\kappa(n)$ for some prior-dependent $n = n_\mathcal{P}$.

for each atom in the respective arm, and H_0^N is an empty dataset. Then

$$\Pr\left[\, X_i^N \geq \tau_{\mathcal{P}} \,\right] \geq \rho_{\mathcal{P}} \quad \forall i \in [\kappa] \text{ and } N \geq n_{\mathcal{P}}, \tag{3.10}$$

where the random variable X_i^N is defined as

$$X_i^N = \min_{\text{arms } A \neq V_i} \mathbb{E}\left[\, \mu(V_i) - \mu(A) \mid H_{i-1}^N \,\right].$$

In intuition, any given arm V_i can be the posterior best arm with a margin $\tau_{\mathcal{P}}$ and probability at least $\rho_{\mathcal{P}}$ after seeing at least $n_{\mathcal{P}}$ samples of the previous arms $V_1, \ldots, V_{i-1}$.

Given Property (P), prior work guarantees the following (without relying on assumptions (3.7-3.8)).

Theorem 3.4.1 (Mansour et al. (2020a))**.** *Assume Property (P) holds with constants $n_{\mathcal{P}}, \tau_{\mathcal{P}}, \rho_{\mathcal{P}}$ and $\kappa = \kappa(n_{\mathcal{P}})$. Then there exists a BIC algorithm which explores each arm $V_1, \ldots, V_\kappa$ at least $n_{\mathcal{P}}$ times and completes in T_0 rounds, where $T_0 = \kappa \cdot n_{\mathcal{P}} \cdot (1 + d) / (\tau_{\mathcal{P}} \cdot \rho_{\mathcal{P}})$.*

Next, we establish (P). First we state a result for a paradigmatic case when all arms have the same cardinality, then relax it in what follows (with a somewhat weaker guarantee).

Theorem 3.4.2. *Assume Beta-Bernoulli priors (3.7-3.8). Further, assume that*

$$\textit{The arms are all subsets of } [d] \textit{ of a given size } m; \tag{3.11}$$

Then Property (P) holds with $\kappa = \kappa(n_{\mathcal{P}}) = \lceil d/m \rceil$ and

$$n_{\mathcal{P}} = \lceil \beta_d / \alpha_d \rceil \cdot \max_{\ell \in [d]} \lceil \alpha_\ell \rceil \tag{3.12}$$

$$\tau_{\mathcal{P}} = \min_{\textit{atoms } \ell \neq \ell' \in [d],\; n, n' \in \{0,\, n_{\mathcal{P}}\}} |\nu_\ell(n) - \nu_{\ell'}(n')|. \tag{3.13}$$

$$\rho_{\mathcal{P}} = (1 - \theta_1^0)^{d \cdot n_{\mathcal{P}}}, \tag{3.14}$$

as long as $\tau_{\mathcal{P}}$ and $\rho_{\mathcal{P}}$ are strictly positive.

Proof Sketch. For each arm V_i, $i \in [\kappa]$ we consider the event that the dataset H_i^n from

Property (P) contains the reward of 0 for each samples of each atom. We take n to be large enough so that this event makes all arms $V_1, \ldots, V_{i-1}$ look inferior to V_i, in terms of the posterior mean reward. The key is to lower-bound the probability of this event; a non-trivial step here requires Harris inequality. □

We show that T_0, the requisite number of rounds, depends exponentially on the number of atoms d. To this end, we define a suitable parameterization of the priors. To handle $\tau_{\mathcal{P}}$ in Theorem 3.4.2, we posit a lower bound that depends on d, but this dependence is very mild.

Corollary 3.4.3. *Assume Beta-Bernoulli priors (3.7-3.8) and that (3.11) holds. Fix some absolute constants $c_0 \in \mathbb{N}$ and $c, c' \in (0,1)$. Suppose $\mathbb{E}[\theta_\ell] \leq c'$ for all atoms, and the priors satisfy the following non-degeneracy conditions:*

$$\max_{\ell,\ell' \in [d]} \lceil \beta_\ell / \alpha_\ell \rceil \cdot \lceil \alpha_{\ell'} \rceil \leq c_0,$$
$$\min_{\ell,\ell' \in [d],\, n,n' \in \{0,\, c_0\}} |\nu_\ell(n) - \nu_{\ell'}(n')| \geq \Omega(c^{-d})$$

Then there exists a BIC algorithm which samples each atom at least once and completes in

$$N_0 = O\left(c_0\, d \cdot \Phi^d \right)$$

rounds, where $\Phi = c \cdot (1-c')^{-c_0}$ is a constant.

Finally, we handle general feasible sets, *i.e.,* without assumption (3.11). The guarantee becomes slightly weaker, in that we have d^2 in the exponent rather than d.

Theorem 3.4.4. *Assume Beta-Bernoulli priors (3.7-3.8). Then Property (P) holds with $\kappa(n_{\mathcal{P}}) \leq d$ and*

$$n_{\mathcal{P}} = \lceil (\alpha_d + \beta_d)/\alpha_d \rceil \cdot \max_{\ell \in [d]} \lceil \alpha_\ell \rceil \cdot d \tag{3.15}$$

$$\tau_{\mathcal{P}} = \min_{A \neq A' \in \mathcal{A}, n \in \{0, n_{\mathcal{P}}\}^d} |\sum_{\ell \in A} \nu_\ell(n_\ell) - \sum_{\ell' \in A'} \nu_{\ell'}(n_{\ell'})|. \tag{3.16}$$

$$\rho_{\mathcal{P}} = (1 - \theta_1^0)^{d \cdot n_{\mathcal{P}}}, \tag{3.17}$$

as long as $\tau_{\mathcal{P}}$ and $\rho_{\mathcal{P}}$ are strictly positive.

Corollary 3.4.5. *Assume Beta-Bernoulli priors (3.7-3.8). Fix some absolute constants $c_1 \in \mathbb{N}$ and $c_2, c_3 \in (0,1)$. Suppose $\mathbb{E}[\theta_\ell] \leq c_3$ for all atoms, and the priors satisfy the following non-degeneracy conditions:*

$$\max_{\text{atoms } \ell,\ell' \in [d]} \lceil (\alpha_\ell + \beta_\ell)/\alpha_\ell \rceil \cdot \lceil \alpha_{\ell'} \rceil \leq c_1$$
$$\min_{A \neq A' \in \mathcal{A}, n \in \{0, n_{\mathcal{P}}\}^d} \Big|\sum_{\ell \in A} \nu_\ell(n_\ell) - \sum_{\ell' \in A'} \nu_{\ell'}(n_{\ell'})\Big| \geq \Omega(c_2^{-d^2}). \tag{3.18}$$

Then there exists a BIC algorithm which samples each atom at least once and completes in

$$N_0 = O\left(c_1\, d^3 \cdot \Phi^{d^2}\right)$$

rounds, where $\Phi = c_2 \cdot (1 - c_3)^{-c_1}$ is a constant.

We provide some motivation for why (3.18) is a mild assumption. Our intuition is that “typically" $\tau_{\mathcal{P}}$ should be on the order of $e^{-O(d)}$, whereas (3.18) only requires it to be $\geq e^{-\Omega(d^2)}$.

3.4.2 Reduction to incentivized reinforcement learning

Our second algorithm builds on the Hidden Hallucination approach from Simchowitz and Slivkins (2021a), which targets incentivized exploration for episodic reinforcement learning. We use this approach by “encoding" a problem instance of combinatorial semi-bandits as a tabular MDP, so that actions in the MDP correspond to atoms, and feasible trajectories correspond to feasible arms. Then we invoke a theorem in Simchowitz and Slivkins (2021a) and “translate" this theorem back to combinatorial semi-bandits.

More specifically, consider a tabular MDP with deterministic transitions and unique initial state. Each action in this MDP correspond to some atom ℓ; then the action's reward is drawn from the corresponding reward distribution $\mathcal{D}_\ell$. In general, only a subset of actions is feasible at a given state-stage pair of the MDP. Let G be the transition graph in such MDP: it is a rooted directed graph such that the nodes of G correspond to state-stage pairs in the MDP (the root node corresponding to the initial state and stage 0). Each edge (u,v) in G corresponds to some MDP action feasible at u, *i.e.,* to some atom. While different edges in G can correspond to the same atom, we require that any rooted directed path in G cannot contain two edges that correspond to the same atom. Let A_P be the subset of atoms that corresponds to a given rooted directed path P, and let $\mathcal{A}_G = \{\, A_P : \text{rooted directed paths in } G \,\}$ be the family of arms "encoded" by G. A family of arms $\mathcal{A}$ is called *MDP-encodable* if $\mathcal{A} = \mathcal{A}_G$ for some transition graph G as defined above with $O(d^2)$ nodes.

Our result applies to all MDP-encodable feasible sets. In particular, the set of all subsets of exactly m atoms, for some fixed $m \leq d$, is MDP-encodable. To see this, consider an MDP with m stages and d states, where each state $\ell \in [d]$ corresponds to the largest atom already included in the arm, and actions feasible at a given stage i and state ℓ correspond to all atoms larger than ℓ.

Our result allows for arbitrary per-atom priors $\mathcal{P}_\ell$, subject to a minor non-degeneracy condition, and reward distributions $\mathcal{D}_\ell$ that are supported on the same countable set.

Theorem 3.4.6. *Consider a feasible arm set $\mathcal{A}$ that is MDP-encodable, as defined above. Suppose the per-atom priors $\mathcal{P}_\ell$ lie in some fixed, finite collection $\mathcal{C}$ such that any $\mathcal{P}_\ell \in \mathcal{C}$ satisfies $\Pr[\theta_\ell \leq \varepsilon] > 0$ for all $\varepsilon > 0$ and $\mathbb{E}[\theta_\ell] > 0$. Further, suppose all reward distributions $\mathcal{D}_\ell$ that are supported on the same countable set. Fix parameter $\delta \in (0,1)$. There is a BIC algorithm such that with probability at least $1-\delta$ each atom is sampled at least once. This algorithm completes in N_0 rounds, where $N_0 = \Phi_{\mathcal{C}}^{-d^3} \cdot O_{\mathcal{C}}\left(\mathrm{poly}(d) \cdot \log(\delta^{-1})\right)$ for some constant $\Phi_{\mathcal{C}} \in (0,1)$ determined by collection $\mathcal{C}$.*

Remark 3.4.7. While the guarantee in Theorem 3.4.6 holds with probability $1-\delta$, rather than almost surely, it suffices to bootstrap Thompson Sampling in Theorem 3.3.1. To see this, let $\texttt{ALG}$ be an algorithm that runs for $T_0 = N_0 \cdot n_{\texttt{TS}} + d \cdot n_{\texttt{TS}}$ rounds (with $n_{\texttt{TS}}$ from Theorem 3.3.1), and proceeds as follows: first it repeats the algorithm from Theorem 3.4.6 $n_{\texttt{TS}}$ times, and in the remaining rounds it deterministically plays an arm with the largest prior mean reward (this algorithm is BIC). Define the "success event" as one in which $\texttt{ALG}$ samples each atom $\geq n_{\texttt{TS}}$ times in the first $N_0 \cdot N_{\texttt{TS}}$ rounds. Consider *another* algorithm, $\texttt{ALG}^*$, which runs for T_0 rounds, coincides with $\texttt{ALG}$ on the first $N_0 \cdot N_{\texttt{TS}}$ rounds, and in the remaining rounds coincides with $\texttt{ALG}$ on the success event, and otherwise plays some arms so as to sample each atom at least once (this algorithm is not necessarily BIC). Now, if Thompson Sampling is preceded by $\texttt{ALG}^*$, then the analysis in Theorem 3.3.1 guarantees that each round t of Thompson Sampling satisfies the BIC property (3.1), and does so with a strictly positive prior-dependent constant on the right hand side of (3.1). Therefore, the same holds for `ALG`, since it coincides with `ALG`* w.h.p., if the failure probability δ in Theorem 3.4.6 is chosen small enough.

Chapter 4

Incentivizing Compliance with Algorithmic Instruments

In this work, we study the second extension of *incentivized exploration*: **How to incentivize exploration in multi-armed bandits when the population is heterogeneous.**

4.1 Introduction

In many applications, estimating the causal effect of a treatment or intervention is at the heart of a decision-making process. Examples include a study on the effect of a vaccine on immunity, an assessment of the effect of a training program on workers' efficiency, and an evaluation of the effect of a sales campaign on a company's profit. Many studies on causal effects rely on randomized experiments, which randomly assign each individual in a population to a treatment group or a control group and then estimate the causal effects by comparing the outcomes across groups. However, in many real-world domains, participation is voluntary, which can be susceptible to *non-compliance*. For example, people may turn down a vaccine or a drug when they are assigned to receive the treatment Wright (1993). Another example is a randomized evaluation of the Job Training Partnership Act (JTPA) training program Bloom et al. (1997a), where only 60 percent of the workers assigned to be trained chose to receive training, while roughly 2 percent of those assigned to the control group chose to receive training. In many cases, non-compliance can cause selection bias: for example, those who choose to receive the drug or vaccine in a randomized trial tend to be healthier, and those who join the training program might be more productive to begin with.

Although non-compliance in randomized experiments has been well studied in many observational studies (Angrist and Pischke, 2008), there has been little work that studies and models how compliance varies over time. In reality, however, participants' compliance behaviors may not be static: they may change according to their time-varying beliefs about the treatments. If the outcomes from the previous trials suggest that the treatments are effective, then the participants may become more willing to accept the recommendation. For example, those initially weary about a new vaccine may change their mind once they see others take it without experiencing negative symptoms.[1] Motivated by this observation, this paper studies the design of dynamic trial mechanisms that map history– the observations from previous trials–to a treatment recommendation and gradually incentivize compliance over time.

In this work, we introduce a game theoretic model to study the dynamic (non)-compliance behavior due to changing beliefs. In our model, there is a collection of treatments such that each treatment j is associated with an unknown treatment effect θ_j. We study an online learning game, in which a set of T myopic agents arrive sequentially over T rounds. Each agent t has a private unobserved *type* u_t, which determines their heterogeneous prior beliefs about the treatment effects. Each agent's goal is to select a treatment j that maximizes the reward: $\theta_j + g_t^{(u_t)}$, where $g_t^{(u_t)}$ denotes the type-dependent baseline reward (without taking any treatment). We introduce a social planner whose goal is to estimate the effects of underlying treatments and incentive the agents to select the treatment that maximizes long-term cumulative reward. Upon the arrival of each agent t, the planner provides the agent with a random treatment recommendation, which is computed by a policy that maps the history of interactions with the previous $(t-1)$ agents. While agent t does not observe the previous history, they form a posterior belief over the treatment effects based on the recommendation and then select the action that maximizes their expected utility.

Under this model, we provide dynamic trial mechanisms that incentivize compliance

[1] A recent survey shows that many Americans choose to wait before deciding to receive a COVID-19 vaccine Hamel et al. (2021).

over time and accurately estimate the treatment effects. The key technical insight is that the planner's random recommendation at each round can be viewed as an *instrument* that is independent of the agent's private type and only influences the observed rewards through the agent's choice of action. By leveraging this observation, we can perform instrumental variable (IV) regression to recover the treatment effects, as long as some of the agents are compliant with the recommendations. To create compliance incentives, our mechanisms leverage techniques from the literature of *incentivizing exploration* Mansour et al. (2015b); Slivkins (2019a). The key idea is *information asymmetry*: since each agent does not directly observe the previous history, the planner has more information. By strategically mapping previous history to instruments, the planner can incentivize agents to explore treatments that are less preferred a-priori.

We focus on the binary action setting, where each agent can select treatment or control to illustrate our techniques. In the binary setting, we first provide two mechanisms that works with two initial non-compliance situations.

Complete non-compliance. In Section 4.3, we consider a setting where the planner initially has no information about the treatment effect θ, so all agents are initially non-compliant with the planner's recommendations. We provide Algorithm 1 which first lets initial agents choose their preferred arms, then constructs recommendations that incentivize compliance for some later agents. This enables treatment effect estimation through IV regression.

Partial compliance. In Section 4.4, we consider a setting where the planner has an initial estimate of the treatment effect θ (that may be obtained by running Algorithm 1), so they can incentivize some agents to comply. We provide Algorithm 2, which can be viewed as the bandit algorithm *active arm elimination* Even-Dar et al. (2006) which uses IV estimates to compare treatments. Samples collected by Algorithm 1 provide an increasingly accurate estimate $\hat{\theta}$ and incentivize more agents to comply over time.

Regret minimization. In Section 4.5, we show that if the planner first runs Algorithm 1 to obtain an initial treatment effect estimate $\hat{\theta}$ and then runs Algorithm 2 to amplify compliance, then he can achieve $\tilde{O}(\sqrt{T})$ regret w.r.t. the cumulative reward given by enforcing the optimal action for all agents.

Experiments. Lastly, in Section 4.7, we complement our theoretical results with numerical simulations, which allow us to examine how parameters in agents' prior beliefs influence the convergence rate of our recommendation algorithm.

4.2 Treatment-Control Model

We study a sequential game between a *social planner* and a sequence of *agents* over T rounds, where T is known to the social planner. In this manuscript, we will focus on the binary setting with a single treatment. In the binary setting, the treatment of interest has an unknown effect $\theta \in [-1, 1]$. In each round t, a new agent indexed by t arrives with their *private type* u_t drawn independently from a distribution $\mathcal{U}$ over the set of all private types U. Each agent t has two actions to choose from: taking the treatment (denoted as $x_t = 1$) and not taking the treatment, i.e. the control (denoted as $x_t = 0$). Upon arrival, agent t also receives an action recommendation $z_t \in \{0, 1\}$ from the planner. After selecting an action $x_t \in \{0, 1\}$, agent t receives a reward $y_t \in \mathbb{R}$, given by

$$y_t = \theta x_t + g_t^{(u_t)} \tag{4.1}$$

where $g_t^{(u_t)}$ denotes the confounding *baseline reward* which depends on the agent's private type u_t; each is drawn from a sub-Gaussian distribution with a sub-Gaussian norm of σ_g. The social planner's goal is to estimate the treatment effect θ and maximize the total expected reward of all T agents.

History and recommendation policy. The interaction between the planner and the agent t is given by the tuple (z_t, x_t, y_t). For each t, let H_t denote the *history* from round 1 to t, i.e. the sequence of interactions between the social planner and the first t agents, such that $H_t := ((z_1, x_1, y_1), \ldots, (z_t, x_t, y_t))$. Before the game starts, the social planner commits to a recommendation policy $\pi = (\pi_t)_{t=1}^T$ where each $\pi_t \colon (\{0,1\} \times \{0,1\} \times \mathbb{R})^{t-1} \to \Delta(\{0, 1\})$ is a randomized mapping from the history H_{t-1} to recommendation z_t. Policy π is fully known to all agents.

Beliefs, incentives, and action choices. Each agent t knows their place t in the sequential game, and their private type u_t determines a *prior belief* $\mathcal{P}^{(u_t)}$, which is a joint distribution over the treatment effect θ and noisy error term $g^{(u)}$. Agent t selects action x_t as such:

$$x_t := \mathbb{1}\left[\mathop{\mathbb{E}}_{\mathcal{P}^{(u_t)}, \pi_t}[\theta \mid z_t, t] > 0\right]. \tag{4.2}$$

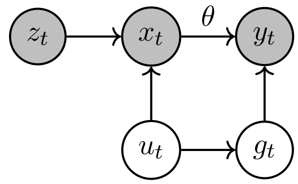


Figure 4.1: Causal diagram of our setting at any time t. Grey nodes are observed, white unobserved. Recommendation z_t influences treatment x_t, which influences outcome y_t. Unobserved type u_t influences both treatment x_t and outcome y_t —the latter via error term $g_t^{(u_t)}$.

An agent t is *compliant* with a recommendation z_t if the agent chooses the recommended action, i.e. $x_t = z_t$. We'll also say that a recommendation is *compliant* if $x_t = z_t$.

Figure 4.1 shows the causal diagram for this setting.

Recommendations as Instruments Unlike the standard multi-armed bandit and previous models on incentivizing exploration Mansour et al. (2015b, 2016), the heterogeneous beliefs in our setting can lead to selection bias. For example, agents who are willing to take the treatment may also have higher baseline rewards. Thus, simply comparing rewards across the treatment group ($x = 1$) and the control group ($x = 0$) will lead to a biased estimate of θ. To overcome this selection bias, we will view the planner's recommendations as instruments and perform *instrumental variable (IV) regression* to estimate θ. There are two criteria for recommendation z_t to be a valid instrument: (1) z_t influences the selection x_t, and (2) z_t is independent from the noisy baseline reward $g^{(u)}$. See Figure 4.1 for a graphical explanation of how these criteria will be satisfied in our setting. Criterion (2) follows because planner chooses z_t randomly, independent of the type u_t. Our goal is to design a recommendation policy to meet criterion (1).

Wald Estimator. Our mechanism periodically solves the following IV regression problem: given a set S of n observations $(x_i, y_i, z_i)_{i=1}^n$, compute an estimate $\hat{\theta}_S$ of θ. We consider the following two-stage least square (2SLS) or Wald estimator (which are equivalent for binary treatments):

$$\hat{\theta}_S = \frac{\sum_{i=1}^n (y_i - \bar{y})(z_i - \bar{z})}{\sum_{i=1}^n (x_i - \bar{x})(z_i - \bar{z})}, \tag{4.3}$$

where $\bar{x}, \bar{y}, \bar{z}$ denote the empirical means of variables x_i, y_i, and z_i respectively.

While existing work on IV regression mostly focuses on asymptotic analyses, we provide a high-probability finite-sample error bound for $\hat{\theta}_S$, which is required by our regret analysis and may be of independent interest.

Theorem 4.2.1 (Finite-sample error bound for Wald estimator)**.** *Let* $z_1, z_2, \ldots, z_n \in \{0,1\}$ *be a sequence of instruments. Suppose there is a sequence of* n *agents such that each agent* i *has their private type* u_i *drawn independently from* $\mathcal{U}$*, selects action* x_i *under instrument* z_i*, and receives reward* y_i*. Let sample set* $S = (x_i, y_i, z_i)_{i=1}^n$*. Let* $A : (\{0,1\}^n \times \{0,1\}^n \times \mathbb{R}^n) \to \mathbb{R}$ *denote the approximation bound for set* S*, such that*

$$A(S, \delta) := \frac{2\sigma_g \sqrt{2n \log(2/\delta)}}{\left|\sum_{i=1}^n (x_i - \bar{x})(z_i - \bar{z})\right|}$$

and the Wald estimator given by (4.3) *satisfies*

$$\left|\hat{\theta}_S - \theta\right| \leq A(S, \delta)$$

with probability at least $1 - \delta$*, for any* $\delta \in (0, 1)$*.*

Proof Sketch. The bound follows by substituting our expressions for y_t, x_t into the IV regression estimator, applying the Cauchy-Schwarz inequality to split the bound into two terms (one dependent on $\{(g_t, z_t)\}_{t=1}^{|S|}$ and one dependent on $\{(x_t, z_t)\}_{t=1}^{|S|}$), and bound the second term with high probability.

Note that the error rate above depends on the covariance between the instruments z and action choices x. In particular, when $\sum_{i=1}^n (x_i - \bar{x})(z_i - \bar{z})$ is linear in n, the error rate becomes $\tilde{O}(1/\sqrt{n})$. In the following sections, we will provide mechanisms that incentivize compliance so that the instruments z are correlated with actions x, enabling us to achieve such an error rate.

4.3 Overcoming Complete Non-Compliance

In this section, we present a recommendation policy that incentivizes compliance to enable IV estimation. We focus on a setting where the agents are initially completely non-compliant: since the planner has no information about the treatment effect in the initial rounds, the recommendations have no influence on agents' action selections. For simplicity of exposition, we will present our policy in a setting where there are two types of agents who are initially "always-takers" and "never-takers." This assumption can be relaxed to have arbitrarily many types and also allow all types to be "always-takers."

Formally, consider two types of agents $i \in \{0,1\}$. For type i, let p_i be the fraction of agents in the population, $\mathcal{P}^{(i)}$ the prior beliefs, and $g^{(i)}$ the baseline reward random variables. Agents of type 1 initially prefer the treatment and type 0 agents prefer control: their prior means for θ satisfy $\mu^{(1)} = \mathbb{E}_{\mathcal{P}^{(1)}}[\theta] > 0$ and $\mu^{(0)} = \mathbb{E}_{\mathcal{P}^{(0)}}[\theta] < 0$.

Our policy (Algorithm 1) splits into two stages. In the first stage, agents take their preferred action according to their prior beliefs: type 0 agents choose control and type 1 treatment. This allows us to collect ℓ_0 and ℓ_1 observations of rewards for $x = 0$ and $x = 1$, respectively. Let $\bar{y}^0$ and $\bar{y}^1$ denote the empirical average rewards for the two actions, respectively. Note that since the baseline rewards $g^{(u)}$ are correlated with the selections x, the difference $(\bar{y}^1 - \bar{y}^0)$ is a biased estimate for θ.

In the second stage, we use this initial set of reward observations to construct valid instruments which incentivize agents of one of the two types to follow both control and treatment recommendations. Without loss of generality, we focus on incentivizing compliance among type 0 agents. Since they already prefer control, the primary difficulty here is to incentivize type 0 agents to comply with treatment recommendations.[2] We leverage the following observation: according to the prior $\mathcal{P}^{(0)}$ of type 0 agents, there is a non-zero probability that the biased estimate $(\bar{y}^1 - \bar{y}^0)$ is so large that θ must be positive.

Formally, consider the following event for the average rewards $\bar{y}^0$ and $\bar{y}^1$:

$$\xi = \left\{ \bar{y}^1 > \bar{y}^0 + \sigma_g \left(\sqrt{\frac{2\log(2/\delta)}{\ell_0}} + \sqrt{\frac{2\log(2/\delta)}{\ell_1}} \right) + G^{(0)} + \frac{1}{2} \right\} \tag{4.4}$$

where $G^{(0)}$ is a constant such that $G^{(0)} > \mathbb{E}_{\mathcal{P}^{(0)}}[g^{(1)} - g^{(0)}]$ and σ_g is the variance parameter for $g^{(0)}$ and $g^{(1)}$.

Assumption 4.3.1 (Knowledge Assumption for Algorithm 1)**.** *Within Section 4.3, the following are common knowledge among agents and planner:*[3]

[2] We could instead incentivize type 1 agents to take control. This would require 1) rewriting event ξ so it indicates that the expectation of θ over $\mathcal{P}^{(1)}$ must be negative and 2) rewriting Algorithm 1 so that control is recommended when exploring. We cannot incentivize both types to comply at the same time.

[3] Assumptions do not hold elsewhere, unless explicitly stated.

1. *Type 0 agents prefer control and type 1 agents prefer treatment. The fraction of agents of type 0 in the population is* $p_0 \geq 0$ *and the fraction of type 1 is* $p_1 > 0$.

2. *Type 0's prior treatment effect mean* $\mu^{(0)}$ *and the probability of event* ξ*, denoted* $\mathbb{P}_{\mathcal{P}^{(0)}}[\xi]$*, over the prior* $\mathcal{P}^{(0)}$ *of type 0.*[4]

Algorithm 1: Overcoming complete non-compliance

Input: exploration probability $\rho \in (0, 1)$, $\ell \in \mathbb{N}$ (w.l.o.g. $\rho\ell \in \mathbb{N}$),
min first stage samples $\ell_0, \ell_1 \in \mathbb{N}$, and failure probability $\delta < \mathbb{P}_{\mathcal{P}_0}[\xi]/8$
1st stage: The first $2 \max(\ell_0/p_0, \ell_1/p_1)$ agents are given no recommendation
2nd stage: Using ℓ_0 control and ℓ_1 treatment samples in the first stage:
if $\bar{y}^1 > \bar{y}^0 + \sigma_g \left(\sqrt{\frac{2\log(2/\delta)}{\ell_0}} + \sqrt{\frac{2\log(2/\delta)}{\ell_1}} \right) + G^{(0)} + \frac{1}{2}$ **then**
    $a^* = 1$
else
    $a^* = 0$
1 **end if**
From the next ℓ agents, pick $\rho\ell$ agents uniformly at random to
be the explore set E
for the next ℓ rounds **do**
    if agent t is in explore set E **then**
        $z_t = 1$
    else
        $z_t = a^*$
    end if
end for

We prove that Algorithm 1 is compliant for agents of type 0 as long as the exploration probability ρ is less than some constant that depends on the prior $\mathcal{P}^{(0)}$. When an agent of type 0 is recommended treatment, they do not know whether this is due to exploration or exploitation. However, with small enough ρ, their expected gain from exploiting exceeds the expected loss from exploring. Hence, the agents comply with the recommendation and take treatment.

---

[4] These assumptions (as well as Assumption 4.4.1 and Assumption 4.5.1) require only partial knowledge of the priors for compliant agents only. They are no more restrictive than the least restrictive (detail-free) assumptions of Mansour et al. (2015b).

Lemma 4.3.2 (Type 0 compliance with Algorithm 1). *Under Assumption 4.3.1, any type 0 agent who arrives in the last ℓ rounds of Algorithm 1 is compliant with any recommendation, as long as the exploration probability ρ satisfies*

$$\rho \leq 1 + \frac{4\mu^{(0)}}{\mathbb{P}_{\mathcal{P}^{(0)}}[\xi] - 4\mu^{(0)}} \tag{4.5}$$

where the event ξ is defined above in Equation (4.4).

Proof Sketch. The proof follows by expressing the compliance condition for type 0 agents as different cases, depending on the recommendation. By keeping the exploration probability ρ small with regard to type 0 agent's prior-dependent probability $\mathbb{P}_{\mathcal{P}^{(0)}}[\xi]$ and the conditional expected treatment effect $\mathbb{E}_{\mathcal{P}^{(0)}}[\theta|\xi]$, the expected gain from exploiting is greater than the expected loss from exploring. Hence, type 0 agents would comply with the recommendation. We further simplify the condition on exploration probability ρ by applying high probability bound on the samples collected from the 1st stage (where no recommendations were given). □

We also provide a separate accuracy guarantee for the treatment effect estimate $\hat{\theta}$ at the end of the Algorithm 1.

Theorem 4.3.3 (Treatment Effect Confidence Interval after Algorithm 1). *With sample set $S_\ell = (x_i, y_i, z_i)_{i=1}^{\ell}$ of ℓ samples collected from the second stage of Algorithm 1 —run with exploration probability ρ small enough so that type 0 agents are compliant (see Lemma 4.3.2),— approximation bound $A(S_\ell, \delta)$ satisfies the following, with probability at least $1-\delta$:*

$$\left|\hat{\theta}_{S_\ell} - \theta\right| \leq A(S_\ell, \delta) \leq \frac{2\sigma_g\sqrt{2\log(5/\delta)}}{\rho(1-\rho)p_0\sqrt{\ell} - (3-\rho)\sqrt{\frac{\rho\log(5/\delta)}{2(1-\rho)}}}$$

for any $\delta \in (0,1)$. Recall σ_g is the variance of $g^{(u_i)}$, p_0 is the fraction of compliant never-takers in the population of agents,[5] *and $A(S_\ell, \delta)$ is defined as in Theorem 4.2.1.*

Proof Sketch. Note that Theorem 4.2.1 applies, so we only have to bound the denominator term which is dependent on $\{(x_t, z_t)\}_{t=1}^{|S_\ell|}$. We assume that Algorithm 1 is initialized with parameters (see Lemma 4.3.2) such that type 0 agent is compliant. We bound the term dependent on $\{(x_t, z_t)\}_{t=1}^{|S_\ell|}$ with high probability. □

[5] We redefine p_0 here to be applicable to more general settings.

4.3.1 Extensions to General Setting

Algorithm 1 can be extended to handle more general settings:

1. There can be arbitrarily many types of agents that do not share the same prior. In this case, let $\mathbb{E}_{\mathcal{P}^{(u)}}[g^0]$ and $\mathbb{E}_{\mathcal{P}^{(u)}}[g^1]$ denote the expected baseline rewards for never-takers and always-takers, respectively, over the prior $\mathcal{P}^{(u)}$ of any type u and $G^{(u)} > \mathbb{E}_{\mathcal{P}^{(u)}}[g^1 - g^0]$. Then, Algorithm 1 can still incentivize any never-taker type u agents to comply as long as the planner has a lower bound on $\mathbb{P}_{\mathcal{P}^{(u)}}[\xi^{(u)}]$, where $\xi^{(u)}$ is defined just as ξ in Equation (4.4), except $G^{(0)}$ is replaced with $G^{(u)}$. Theorem 4.3.3 applies as is.

2. All types can be always-takers (who prefer the treatment). The algorithm can incentivize some of the agents to take control with an event ξ defined without $\bar{y}^0$ and flipped (i.e. the mean treatment reward is much lower than the expected baseline reward).

By Theorem 4.3.3, samples collected from Algorithm 1 produce a confidence interval on the treatment effect θ which decreases proportionally to $1/\sqrt{t}$ by round t. However, it still decreases slowly because the exploration probability ρ is small (see roughly how small in Section 4.7). In Section 4.4, we give an algorithm for which this confidence interval improves quicker and works for arbitrarily many types.

4.4 Overcoming Partial Non-Compliance

In this section, we present a recommendation policy which (1) capitalizes on partial compliance, eventually incentivizing all agents to comply, and (2) determines whether the treatment effect is positive or not (with high probability). Algorithm 2 recommends control and treatment sequentially (one after the other). Lemma 4.4.2 gives conditions for partial compliance from the beginning of Algorithm 2, given access to initial samples which form a crude estimate of the treatment effect. Theorem 4.4.3 demonstrates how rapidly this estimate improves throughout Algorithm 2, which solely depends on the fraction of compliant agents (and not on some fraction like ρ with Algorithm 1). More (and eventually all) types of agents progressively become compliant throughout Algorithm 2.

Assumption 4.4.1 (Knowledge Assumption for Algorithm 2)**.** *Within Section 4.4, the following are common knowledge among agents and the principal:*

1. *The fraction of agents in the population who prefer control is $p_0 \geq 0$; that who prefer treatment is $p_1 \geq 0$.*

2. *For each type u and for some τ (which can differ per u), the probability $\tau\,\mathbb{P}_{\mathcal{P}(u)}[\theta > \tau]$ is known if $\mathbb{E}_{\mathcal{P}(u)}[\theta] < 0$; or $\tau\,\mathbb{P}_{\mathcal{P}(u)}[\theta < -\tau]$ is known if $\mathbb{E}_{\mathcal{P}(u)}[\theta] \geq 0$.*

Algorithm 2: Overcoming partial compliance

Input: samples $S_0 := (x_i, z_i, y_i)_{i=1}^{|S_0|}$ which meet Theorem 4.2.1 conditions, corresponding IV estimate $\hat{\theta}_{S_0}$;[6] time horizon T; number of recommendations for each action per phase h; approximation bound failure probability δ;
Split the remaining rounds (up to T) into consecutive phases of h rounds each, starting with $q = 1$;
Let $\hat{\theta}_0 = \hat{\theta}_{S_0}$ and $A_0 = A(S_0, \delta)$;
while $|\hat{\theta}_{q-1}| \leq A_{q-1}$ **do**
  The next $2h$ agents are recommended control and treatment sequentially;
  Let S_q be samples up to and including phase q, i.e. $S_q := (x_i, z_i, y_i)_{i=1}^{|S_0|+hq} = S_{q-1} + \{\text{round } q \text{ samples}\}$
  Let S_q^{BEST} be the samples with smallest approximation bound so far, i.e. $S_q^{\text{BEST}} = \operatorname{argmin}_{S_r, 0 \leq r \leq q} A(S_r, \delta)$;
  Define $\hat{\theta}_q = \hat{\theta}_{S_q^{\text{BEST}}}$ and $A_q = A(S_q^{\text{BEST}}, \delta)$;
  $q = q + 1$;
end while
For all remaining agents recommend $a^* = \mathbb{1}\left[\hat{\theta}_q > 0\right]$.

We focus on a setting where agents are assumed to have been at least partially compliant in the past, such that we may form an IV estimate from the history. The social planner employs Algorithm 2, which is a modification of the *Active Arms Elimination* algorithm Even-Dar et al. (2006). Treatment and control "race", i.e. are recommended sequentially, until the expected treatment effect is known to be negative or positive (with high probability). Then, the algorithm recommends the "winner" (the action with a higher expected reward) for the remainder of the time horizon T.

[6] Operator $|\cdot|$ denotes the cardinality of a set.

The compliance incentive works as such: when an agent is given a recommendation, they do not know whether it is because the action is still in the "race" or if the action is the "winner". When the algorithm is initialized with samples that form an IV estimate that is sufficiently close to the true treatment effect (according to the agent's prior), then the probability that any recommended action has "won" is high enough such that the agent's expected gain from taking a "winning" action outweighs the expected loss from taking a "racing" one. We formalize this in Lemma 4.4.2.

Lemma 4.4.2 (Algorithm 2 Partial Compliance). *Recall that Algorithm 2 is initialized with input samples $S_0 = (x_i, y_i, z_i)_{i=1}^{|S_0|}$. For any type u with the following prior preference (control or treatment), if S_0 satisfies the following condition, with probability at least $1-\delta$, then all agents of type u will comply with recommendations of Algorithm 2:*

$$A(S_0,\delta) \leq \begin{cases} \tau\, \mathbb{P}_{\mathcal{P}(u)}[\theta > \tau]/4 & \text{if } \mathbb{E}_{\mathcal{P}(u)}[\theta] < 0; \\ \tau\, \mathbb{P}_{\mathcal{P}(u)}[\theta < -\tau]/4 & \text{if } \mathbb{E}_{\mathcal{P}(u)}[\theta] \geq 0, \end{cases}$$

for some $\tau \in (0,1)$, where $A(S_0,\delta)$ is the approximation bound for S_0 and any $\delta \in (0,1)$ (see Theorem 4.2.1).

Proof Sketch. The proof follows by using a "clean event" analysis where the IV estimated treatment effect $\hat{\theta}$ is close to the true treatment effect θ. We split the conditional expected treatment effect $\mathbb{E}_{\mathcal{P}(u)}[\theta]$ into different cases for the value of θ. With an IV estimate that is sufficiently close to the true treatment effect, the expected gain from exploiting (taking the "winning" action) is greater than the expected loss from exploring (taking a recommended action when the "race" is not over) and any agent of type u will comply with the principal's recommendation. □

When a nonzero fraction of agents comply from the beginning, the samples gathered in Algorithm 2 provide treatment effect estimates $\hat{\theta}$ which become increasingly accurate over rounds. In the following Theorem 4.4.3, we provide a high probability guarantee of this accuracy.

Theorem 4.4.3 (Treatment Effect Confidence Interval from Algorithm 2 with Partial Compliance). *With set $S = (x_i, y_i, z_i)_{i=1}^{|S|}$ of $|S|$ samples collected from Algorithm 2 where*

p_c is the fraction of compliant agents in the population, we form an estimate $\hat{\theta}_S$ of the treatment effect θ. With probability at least $1-\delta$,

$$\left|\hat{\theta}_S - \theta\right| \leq A(S,\delta) \leq \frac{8\sigma_g\sqrt{2\log(5/\delta)}}{p_c\sqrt{|S|} - \sqrt{50\log(5/\delta)}}$$

for any $\delta \in (0,1)$, where σ_g is the variance of $g^{(u_i)}$.

Proof Sketch. Note that Theorem 4.2.1 applies, so we only have to bound the denominator term which is dependent on $\{(x_t, z_t)\}_{t=1}^{|S|}$. We assume that Algorithm 2 is initialized with parameters such that $p_c > 0$ fraction of the population complies with all recommendations. We bound the term dependent on $\{(x_t, z_t)\}_{t=1}^{|S|}$ with high probability. □

Agents become compliant during Algorithm 2 for the same reason others become compliant from the beginning: they expect that the estimate $\hat{\theta}$ is sufficiently accurate and it's likely they're getting recommended an action because it won the race. For large enough T, all agents will become compliant. Note that the accuracy improvement in Theorem 4.4.3 relies solely on the proportion of agents p_c who comply from the beginning of Algorithm 2, which relies on the accuracy of the approximation bound given by initial samples S_0. Thus, if the social planner can choose more accurate S_0, then the treatment effect estimate $\hat{\theta}$ given by samples from Algorithm 2 becomes more accurate quicker. In Section 4.5, we present a recommendation policy in which S_0 can be chosen by running Algorithm 1.

4.5 Combined Recommendation Policy

In this section, we present a recommendation policy π_c, which spans T rounds and runs Algorithms 1 and 2 in sequence. This policy achieves $\tilde{O}(\sqrt{T})$ regret for sufficiently large T and produces an estimate $\hat{\theta}$ which deviates from the true treatment effect θ by $O(1/\sqrt{T})$.[7]

Assumption 4.5.1 (Knowledge Assumption for Policy π_c). *In Sections 4.5.1 and 4.5.2, the following are common knowledge among agents and planner:*

[7] We spare the reader the details of the exact bound.

1. All prior-dependent constants given in Assumption 4.4.1

2. For each type u which prefers control, prior mean $\mu^{(u)}$ and a lower bound on the probability $\mathbb{P}_{\mathcal{P}^{(u)}}[\xi^{(u)}]$ (defined in Extension 1 of Algorithm 1 from Section 4.3.1)

4.5.1 Recommendation Policy Formulation

Recommendation policy π_c over T rounds is given as such:

1. Run Algorithm 1 with exploration probability ρ set to incentivize at least $p_{c_1} > 0$ fraction of agents of the population who initially prefer control to comply in Algorithm 1 and ℓ to make at least $p_{c_2} > 0$ fraction of agents comply in Algorithm 2 (see Lemma 4.5.2).

2. Initialize Algorithm 2 with samples from Algorithm 1. At least p_{c_2} fraction of agents comply in Algorithm 2.

We first provide conditions on the phase length ℓ to define policy π_c.

Lemma 4.5.2 (Lower bound on ℓ for Type u Compliance in Algorithm 2)**.** *Recall that S_ℓ denotes the samples collected from the second stage of Algorithm 1. Let S_ℓ be the input samples S_0 in Algorithm 2. Assume that p_{c_1} proportion of agents in the population are compliant with recommendations of Algorithm 1 and length ℓ satisfies:*

$$\ell \geq \begin{cases} \left(\frac{\kappa_1}{\tau\,\mathbb{P}_{\mathcal{P}^{(u)}}[\theta>\tau]} + \kappa_2\right)^2 & \text{if } \underset{\mathcal{P}^{(u)}}{\mathbb{E}}[\theta] < 0 \\ \left(\frac{\kappa_1}{\tau\,\mathbb{P}_{\mathcal{P}^{(u)}}[\theta<-\tau]} + \kappa_2\right)^2 & \text{if } \underset{\mathcal{P}^{(u)}}{\mathbb{E}}[\theta] \geq 0 \end{cases} \tag{4.6}$$

for some $\tau \in (0,1)$ and where $\kappa_1 := \frac{8\sigma_g\sqrt{2\log(5/\delta)}}{p_{c_1}\rho(1-\rho)}$ and $\kappa_2 := (3-\rho)\sqrt{\frac{\rho\log(5/\delta)}{2(1-\rho)}}$ for any $\delta \in (0,1)$. Then any agent of type u will comply with recommendations of Algorithm 2.

Proof Sketch. The proof follows by substituting the value of ℓ into the approximation bound Theorem 4.3.3 and simplifying. The compliance condition follows from Lemma 4.4.2. □

Policy π_c shifts from Algorithm 1 to Algorithm 2 as soon as the condition on ℓ above is satisfied. This is because 1) the treatment effect estimate $\hat{\theta}$ gets more accurate quicker and 2) less regret is accumulated in Algorithm 2 than Algorithm 1.

4.5.2 Regret Analysis

The goal of recommendation policy π_c is to maximize the cumulative reward of all agents. We measure the policy's performance through *regret*. We are interested in minimizing regret, which is specific to the treatment effect θ. Since agents' priors are not exactly known to the social planner, this pseudo-regret is correct for any realization of these priors and treatment effect θ.

Definition 4.5.3. *[Pseudo-regret] The pseudo-regret of a recommendation policy is given as such:*

$$R_\theta(T) = T\max(\theta, 0) - \sum_{t=1}^{T} \theta x_t \tag{4.7}$$

We present regret guarantees for recommendation policy π_c. First, policy π_c achieves sub-linear pseudo-regret.

Lemma 4.5.4 (Pseudo-regret)**.** *The pseudo-regret accumulated from policy π_c is bounded for any $\theta \in [-1, 1]$ as follows, with probability at least $1-\delta$ for any $\delta \in (0, 1)$:*

$$R_\theta(T) \le L_1 + O(\sqrt{T\log(T/\delta)}) \tag{4.8}$$

for sufficiently large time horizon T, where the length of Algorithm 1 is $L_1 = \ell + 2\max\left(\frac{\ell_0}{p_0}, \frac{\ell_1}{p_1}\right)$.

Proof Sketch. The proof follows by observing that Algorithm 2 must end after some $\log(T)$ phases. We can bound the regret of the policy π_c by at most that of Algorithm 2 plus θ per each round of Algorithm 1, or alternatively, we can upper bound it by θ per each round of the policy π_c. □

Policy π_c also achieves sub-linear regret, where the expectation is over the randomness in the priors of the agents. Lemma 4.5.5 provides a basic performance guarantee of our recommendation policy.

Lemma 4.5.5 (Regret)**.** *Policy π_c achieves regret as follows:*

$$\mathbb{E}[R(T)] = O(\sqrt{T\log(T)}) \tag{4.9}$$

for sufficiently large time horizon T.

Proof Sketch. The proof follows by observing that we can set the parameters in Algorithm 1 and Algorithm 2 in terms of the time horizon T while maintaining compliance throughout policy π_c. $\square$

These results are comparable to the pseudo-regret of the classic multi-armed bandit problem, with some added constants factors for the compliance constraints Even-Dar et al. (2006). The pseudo-regret of our policy π_c is asymptotically equivalent to an extension of the detail-free recommendation algorithm of Mansour et al. (2015b), which incentivizes full compliance for all types. However, our policy can finish in a more timely manner and has smaller prior-dependent constants in the asymptotic bound.

4.6 Many Treatments with Unknown Effects

In this section, we introduce a setting that extends the previous binary treatment setting by considering k treatments (and no control). We now consider a treatment effect vector $\theta \in \mathbb{R}^k$; and $x, z \in \{0,1\}^k$ are one-hot encodings of the treatment choice and recommendation, respectively. We assume that $\mathbb{E}[g^{(u_i)}] = 0$.[8] All other terms are defined similarly to those in Section 4.2. Here, the reward $y_i \in \mathbb{R}$ and action choice $x_i \in \{0,1\}^k$ at round i are given as such:[9]

$$\begin{cases} y_i = \langle \theta, x_i \rangle + g^{(u_i)} \\ x_i = \underset{1 \le j \le k}{\operatorname{argmax}} \big(\mathbb{E}_{\mathcal{P}^{(u_i)}, \pi_i}[\theta^j | z_i, i] \big) \end{cases} \tag{4.10}$$

Given sample set $S = (z_i, x_i, y_i)_{i=1}^n$, we compute IV estimate $\hat{\theta}_S$ of θ as such:

$$\hat{\theta}_S = \left(\sum_{i=1}^n z_i x_i^\top \right)^{-1} \sum_{i=1}^n z_i y_i \tag{4.11}$$

Next, we state finite sample approximation results which extend Theorem 4.2.1 to this general setting.

[8]Without this assumption, we run into identifiability issues: we cannot reconstruct the individual treatment effects $\theta^1, \dots, \theta^k$ without fixing some mean $\mathbb{E}[g^{(u)}]$. Yet, for purposes of regret minimization, assuming $\mathbb{E}[g^{(u)}] = 0$ does not change our results.

[9]We bastardize notation by writing $x_i = j$ instead of $x_i = \mathbf{e}_j$ (the k-dimensional unit vector along the jth dimension).

Theorem 4.6.1 (Many Treatments Effect Approximation Bound)**.** *Let* $z_1, \ldots, z_n \in \{0,1\}^k$ *be a sequence of instruments. Suppose there is a sequence of* n *agents such that each agent* i *has private type* u_i *drawn independently from* $\mathcal{U}$*, selects* x_i *under instrument* z_i *and receives reward* y_i*. Let sample set* $S = (x_i, y_i, z_i)_{i=1}^n$*. The approximation bound* $A(S, \delta)$ *is given as such:*[10]

$$A(S,\delta) = \frac{\sigma_g \sqrt{2nk \log(k/\delta)}}{\sigma_{\min}\left(\sum_{i=1}^n z_i x_i^\top\right)},$$

and the IV estimator given by Equation (4.11) *satisfies*

$$\left\|\hat{\theta}_S - \theta\right\|_2 \le A(S,\delta)$$

with probability at least $1-\delta$ *for any* $\delta \in (0,1)$*.*

Proof Sketch. See Section B.VI.3 for the full proof. The bound follows by substituting our expressions for y_t, x_t into the IV regression estimator, applying the Cauchy-Schwarz inequality to split the bound into two terms (one dependent on $\{(g_t^{(u_t)}, z_t)\}_{t=1}^{|S|}$ and one dependent on $\{(x_t, z_t)\}_{t=1}^{|S|}$). We bound the second term with high probability. □

Next, we extend recommendation policy π_c to k treatments (see Definition 4.6.3 in Section B.VI for details).[11]

Assumption 4.6.2 (Knowledge Assumption for General Policy π_c)**.** *Within Section 4.6, the following are common knowledge among agents and planner:*

1. *All agents share a preference ordering over all* k *treatments, i.e. for any type* u*, the prior expected reward* $\mathbb{E}_{\mathcal{P}^{(u)}}[\theta^1] > \mathbb{E}_{\mathcal{P}^{(u)}}[\theta^2] > \cdots > \mathbb{E}_{\mathcal{P}^{(u)}}[\theta^k]$*.*[12]
2. *Prior-dependent constants* $\mathbb{P}_{\mathcal{P}^{(u)}}[\xi^{(u)}]$ *and* $\mathbb{P}_{\mathcal{P}^{(u)}}[G^v > \tau]$ *for some* $\tau \in (0,1)$ *(see Section B.VI.4).*

---

[10] The operator $\sigma_{\min}(\cdot)$ denotes the smallest singular value.

[11] Algorithm 3 extends Algorithm 1 and Algorithm 4 extends Algorithm 2.

[12] This ordering assumption is shared by Mansour et al. (2015b).

Extensions of Algorithms 1 and 2 and Recommendation Policy π_c to k Treatments

We assume that every agent —regardless of type— shares the same prior ordering of the treatments, such that all agents prior expected value for treatment 1 is greater than their prior expected value for treatment 2 and so on. First, Algorithm 3 is a generalization of Algorithm 1 which serves the same purpose: to overcome complete non-compliance and incentivize some agents to comply eventually. The incentivization mechanism works the same as in Algorithm 1, where we begin by allowing all agents to choose their preferred treatment —treatment 1— for the first ℓ rounds. Based on the ℓ samples collected from the first stage, we then define a number of events $\xi_j^{(u)}$ —which are similar to event ξ from Algorithm 1— that each treatment $j \geq 2$ has the largest expected reward of any treatment and treatment 1 has the smallest, according to the prior of type u:

$$\xi_i^{(u)} := \left(\bar{y}_\ell^1 + C \leq \min_{1<j<i} \bar{y}_\ell^j - C \quad \text{and} \quad \max_{1<j<i} \bar{y}_\ell^j + C \leq \mu_i^{(u)} \right), \tag{4.12}$$

where $C = \sigma_g \sqrt{\frac{2\log(3/\delta)}{\ell}} + \frac{1}{4}$ for any $\delta \in (0,1)$ and where $\bar{y}_\ell^1$ denotes the mean reward for treatment 1 over the ℓ samples of the first stage of Algorithm 3. Thus, if we set the exploration probability ρ small enough, then some subset of agents will comply with all recommendations in the second stage of Algorithm 3.

Second, Algorithm 4 is a generalization of Algorithm 2, which is required to start with at least partial compliance and more rapidly and incentivizes more agents to comply eventually. The incentivization mechanism works the same as in Algorithm 1, where we begin by allowing all agents to choose their preferred treatment —treatment 1— for the first ℓ rounds. Based on the ℓ samples collected from the first stage, we then define a number of events —which are similar to event ξ from Algorithm 1— that each treatment $j \geq 2$ has the largest expected reward of any treatment and treatment 1 has the smallest. Thus, if we set the exploration probability ρ small enough, then some subset of agents will comply with all recommendations in the second stage of Algorithm 3.

Definition 4.6.3 (General recommendation policy π_c for k treatments)**.** *Recommendation*

[13]We set the length of each prhase i of the second stage to be ℓ/ρ so that we get ℓ samples of each treatment i and the exploration probability is ρ.

Algorithm 3: Overcoming complete non-compliance for k treatments

Input: exploration probability $\rho \in (0,1)$, minimum number of samples of any treatment $\ell \in \mathbb{N}$ (assume w.l.o.g. $(\ell/\rho) \in \mathbb{N}$), failure probability $\delta \in (0,1)$, compliant type u

1st stage: The first ℓ agents are given no recommendation (they choose treatment 1)

for each treatment $i > 1$ in increasing lexicographic order **do**

  if $\xi_i^{(u)}$ holds, based on the ℓ samples from the first phase and any samples of treatment $2 \le j < i$ collected thus far **then**

    $a_i^* = i$

  else

    $a_i^* = 1$

  end if

  From the next ℓ/ρ agents, pick ℓ agents uniformly at random to be in the explore set E[13]

  for the next ℓ rounds **do**

    if agent t is in explore set E **then**

      $z_t = 1$

    else

      $z_t = a^*$

    end if

  end for

end for

policy π_c over T rounds is given as such:

1. *Run Algorithm 3 with exploration probability ρ set to incentivize at least $p_{c_1} > 1/k$ fraction of agents of the population to comply in Algorithm 3. Let S_0 be the sample set given from Algorithm 3. By Corollary B.VI.5 and Theorem B.VI.6, we can use S_0 as the initial samples in Algorithm 4 to incentivize compliance for any arm a if the approximation bound $A(S_0, \delta)$ given by S_0 is small enough (see Theorem B.VI.6). Thus, we run Algorithm 3 long enough (i.e. we set ℓ large enough) so that the approximation bound is small enough and at least $p_{c_2} > 1/k$ fraction of agents comply with recommendations of every treatment in Algorithm 4.*

2. *Initialize Algorithm 4 with samples S_0 from Algorithm 3. At least $p_{c_2} > 1/k$ fraction of agents comply with recommendations from Algorithm 4 from the beginning and until time horizon T.*

Algorithm 4: Overcoming partial compliance for k treatments

Input: samples $S_0 := (x_i, z_i, y_i)_{i=1}^{|S_0|}$ which meet Theorem 4.6.1 conditions, corresponding IV estimate $\hat{\theta}_{S_0}$, time horizon T, number of recommendations of
each action per phase h, failure probability $\delta \in (0,1)$
Split the remaining rounds (up to T) into consecutive phases of h rounds each, starting with $q = 1$;
Let $\hat{\theta}_0 = \hat{\theta}_{S_0}$ and $A_0 = A(S_0, \delta)$
Initialize set of active treatments: $B = \{\text{all treatments}\}$.
while $|B| > 1$ **do**
  Let $\hat{\theta}^*_{q-1} = \max_{i \in B} \hat{\theta}^i_{q-1}$ be the largest entry i in $\hat{\theta}_{q-1}$
  Recompute $B = \left\{\text{treatments } i : \hat{\theta}^*_{q-1} - \hat{\theta}^i_{q-1} \leq A_{q-1}\right\}$;
  The next $|B|$ agents are recommended each treatment $i \in B$ sequentially in lexicographic order;
  Let S_q^{BEST} be the sample set with the smallest approximation bound so far, i.e. $S_q^{\text{BEST}} = \text{argmin}_{S_r, 0 \leq r \leq q} A(S_r, \delta)$;
  Define $\hat{\theta}_q = \hat{\theta}_{S_q^{\text{BEST}}}$ and $A_q = A(S_q^{\text{BEST}}, \delta)$;
  $q = q + 1$
end while
For all remaining agents, recommend a^* that remains in B.

Similar to the control-treatment setting, we provide the compliance lemmas and proof sketches for Algorithm 3 and Algorithm 2.

Lemma 4.6.4 (Algorithm 3 compliance). *Let event $\xi^{(u)}$ be defined such that $\mathbb{P}[\xi^{(u)}] = \min_i \mathbb{P}[\xi_i^{(u)}]$. In Algorithm 3, any type u agent who arrives in the last ℓ/ρ rounds of Algorithm 3 is compliant with any recommendation if $\mu_i^{(u)} > 0$ for all $1 < i \leq k$, and the exploration probability ρ satisfies:*

$$\rho \leq 1 + \frac{8\left(\mu_j^{(u)} - \mu_i^{(u)}\right)}{\mathbb{P}_{\pi_c, \mathcal{P}^{(u)}}[\xi^{(u)}]} \tag{4.13}$$

Proof Sketch. Let the recommendation policy π here be Algorithm 3.

Part I (Compliance with recommendation for treatment $i > 1$): We first argue that an agent t of type u who is recommended treatment i will not switch to any other treatment j. For treatments $j > i$, there is no information about treatments i or j collected by the algorithm and by assumption, we have $\mu_i^{(u)} \geq \mu_j^{(u)}$. Hence, it suffices to consider when $j < i$. We want to show that

$$\mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u)}}[\theta^j - \theta^1 | z_t = \mathbf{e}_i] \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u)}}[z_t = \mathbf{e}_i] \geq 0. \tag{4.14}$$

Part II (Recommendation for treatment 1): When agent t is recommended treatment 1, they know that they are not in the explore group E. Therefore, they know that the event $\neg\xi_i^{(u)}$ occurred. Thus, in order to prove that Algorithm 3 is BIC for an agent of type u, we need to show the following for any treatment $j > 1$:

$$\mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u)}}[\theta^1 - \theta^j | z_t = \mathbf{e}_1] \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u)}}[z_t = \mathbf{e}_1] = \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u)}}[\theta^1 - \theta^j | \neg\xi_i^{(u)}] \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u)}}[\neg\xi_i^{(u)}] \geq 0 \tag{4.15}$$

We omit the remainder of this proof due to its similarity with the proof of Lemma 4.3.2

□

Lemma 4.6.5 (Algorithm 4 Partial Compliance)**.** *Recall that Algorithm 4 is initialized with input samples* $S_0 = (x_i, y_i, z_i)_{i=1}^{|S_0|}$*. For any type* u*, if* S_0 *satisfies the following condition, then with probability at least* $1-\delta$ *all agents of type* u *will comply with recommendations of Algorithm 4:*

$$A(S_0,\delta) \leq \tau \mathop{\mathbb{P}}_{\mathcal{P}^{(u)}}[\min_{a,b}\left(\left|\theta^a - \theta^b\right|\right) > \tau]/4$$

for some $\tau \in (0,1)$*, where* $A(S_0,\delta)$ *is the approximation bound for* S_0 *and any* $\delta \in (0,1)$ *(see Theorem 4.6.1).*

Proof Sketch. Let the recommendation policy π here be Algorithm 4. We want to show that for any agent at time t with a type $i < u$ in the racing stage and for any two treatments $a, b \in B$:

$$\mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u)}}[\theta^a - \theta^b | z_t = \mathbf{e}_a] \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u)}}[z_t = \mathbf{e}_a] \geq 0$$

We omit the remainder of this proof due to its similarity with the proof of Lemma 4.4.2.

□

In order to incentivize agents of any type u to comply with general extensions

Algorithms 3 and 4, we (again) set exploration probability ρ and length ℓ to satisfy some compliance conditions relative to $\mathbb{P}_{\mathcal{P}(u)}[\xi^{(u)}]$ and $\tau\,\mathbb{P}_{\mathcal{P}(u)}[G^v > \tau]$, respectively (see Section B.VI.4). We present the (expected) regret from the k treatment extension of policy π_c next.

Lemma 4.6.6 (Regret of Policy π_c for k Treatments)**.** *An extension of policy π_c achieves (expected) regret as follows:*

$$\mathbb{E}[R(T)] = O\left(k\sqrt{kT\log(kT)}\right) \tag{4.16}$$

for sufficiently large time horizon T.

Proof Sketch. See Section B.VI.4 for the full proof. The proof follows the same structure as that of Lemma 4.5.5. □

Though our analysis covers a more general k treatment setting than Mansour et al. (2015b) (capturing non-compliance and selection bias), our policy π_c accumulates asymptotically comparable regret in terms of T. See Section B.VI for all other results. Next, in Section 4.7, we implement Algorithm 1 experimentally.

4.7 Numerical Experiments

In this section, we present experiments to evaluate Algorithm 1. We mention previously in the paper that this approximation bound decreases slowly throughout Algorithm 1, because the exploration probability ρ in Algorithm 1 is small. Here, we are interested in (1) *how small* the exploration probability ρ is and (2) *how slowly* the approximation bound on the absolute difference $|\theta - \hat{\theta}|$ decreases as Algorithm 1 progresses (where $\hat{\theta}$ is based on samples from Algorithm 1). These are important to study, because this slow improvement in accuracy is the primary source of inefficiency (in terms of sample size) for policy π_c, which accumulates linear regret during Algorithm 1 (see Lemma 4.5.4) for marginal improvements in estimation accuracy. This motivates the social planner to move to Algorithm 2 —where the estimation accuracy increases much quicker— as soon as possible in policy π_c. Yet, there is also a tradeoff for moving to Algorithm 2 too quickly: if Algorithm 1 is not run for long enough, then only a small portion of agents may comply in Algorithm 2. In order to better inform the choice of hyperparameters in policy π_c (specifically, the compliance paramters p_{c_1} and p_{c_2}), we empirically estimate these quantities experimentally. We defer experiments on Algorithm 2 to the appendix.[14]

Experimental Description. We consider a setting with two types of agents: type 0 who are initially never-takers and type 1 who are initially always-takers. We let each agent's prior on the treatment effect be a truncated Gaussian distribution between -1 and 1. The noisy baseline reward $g_t^{(u_t)}$ for each type u of agents is drawn from a Gaussian distribution $\mathcal{N}(\mu_{g^{(u)}}, 1)$, with its mean $\mu_{g^{(u)}}$ also drawn from a Gaussian prior. We let each type of agent have equal proportion in the population, i.e. $p_0 = p_1 = 0.5$. We are interested in finding the probability of event ξ (as defined in Equation (4.4)) and the exploration probability ρ (as defined in Equation (4.5)). Instead of deriving an explicit formula for $\mathbb{P}_{\mathcal{P}_0}[\xi]$ to calculate the exploration probability ρ, we estimate it using Monte Carlo simulation by running the first stage of Algorithm 1 for 1000 iterations and aggregating the results. After this, Algorithm 1 is run with the previously-found exploration probability ρ over an increasing number of rounds. We repeatedly calculate the IV estimate of the treatment effect and compare it to a naive OLS estimate (that regresses the treatment onto the reward) over the same samples as a benchmark.

Results. In Figure 4.2, we compare the approximation bound on $|\theta - \hat{\theta}|$ between IV estimate $\hat{\theta}$ versus via a naive estimate for a specific, chosen $\rho = 0.001$. In our experiments, the exploration probability ρ generally lies within $[0.001, 0.008]$. In Figure 4.2, we let hidden treatment effect $\theta = 0.5$, type 0 and type 1 agents' priors on the treatment effect be $\mathcal{N}(-0.5, 1)$ and $\mathcal{N}(0.9, 1)$ —each truncated onto $[-1, 1]$,— respectively. We also let the mean baseline reward for type 0 and type 1 agents be $\mu_{g^{(0)}} \sim \mathcal{N}(0, 1)$ and $\mu_{g^{(1)}} \sim \mathcal{N}(0.1, 1)$, respectively. These priors allow us to set the exploration probability $\rho = 0.001$ for Figure 4.2, where IV regression consistently outperforms OLS for any reasonably long run of Algorithm 1.

To our knowledge, these experiments are the first empirical evaluation of an Incentivizing Exploration algorithm. Figure 4.2 shows the effect of small exploration probability $\rho = 0.001$: we need to run Algorithm 1 for a while to produce a decently accurate causal

[14]The code is available here.

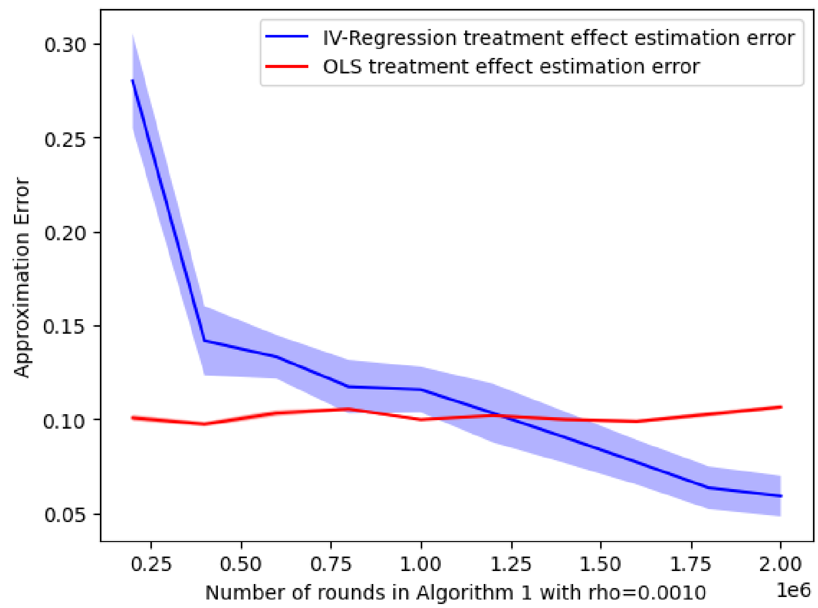


Figure 4.2: Approximation bound using IV regression and OLS during Algorithm 1 with $\rho = 0.001$. Results are averaged over 5 runs; light blue error bars represent one standard error. The OLS estimate converges to a value with an approximation error of around 0.1; whereas the approximation error of our IV estimate steadily decreases over time.

effect estimate.[15]

4.8 Conclusion

In this work, we present a model for how noncompliance changes over time based on beliefs and new information. We observe that recommendations that incentivize (at least partial) compliance can be treated as instrumental variables (IVs), enabling consistent treatment effect estimation via IV regression, even in the presence of non-compliance and confounding. Finally, we provide recommendation mechanisms that provably incentivize compliance and achieve sublinear regret.

Ethical Considerations around Incentivizing Exploration. In this work, we provide recommendation algorithms for incentivizing exploration (IE), each of which consists of two parts with their own ethical considerations: 1) recommending for exploration

[15]We suspect this weakness is likely endemic to previous works in Incentivizing Exploration, as well.

and 2) selectively disclosing information. First, in certain contexts, recommending for exploration of a treatment with unknown effects could place undue risk or harm onto an individual in order to benefit society or maximize well-being overall.[16] Because of this, in certain medical contexts, incentivizing exploration may violate the principle of non-maleficence. Second, incentivizing exploration requires selective information disclosure. While selective disclosure may be required in studies where information is protected or proprietary, it may be (morally) wrong in other settings. Restricting full information about the history of treatment for the purpose of changing an individual's behavior may be manipulative or deceptive and may restrict their autonomy. We do not provide ethical considerations to discourage incentivizing exploration outright: there may be settings where the benefits outweigh the harms. We leave it to study designers and policymakers to weigh these in specific settings.

[16] In moral terms, this may cause a fissure between a utilitarian and a Kantian or purveyor of individual rights.

Chapter 5

Incentive-Aware Synthetic Control

In this chapter, we examine how to leverage the technical tools from Chapter 4 and **incentive exploration in panel data setting to facilitate accurate counterfactual estimation**.

5.1 Introduction

A ubiquitous task in the internet economy is to estimate counterfactual outcomes for a group of *units* (e.g. people, products, geographic regions) under different *interventions* (e.g. marketing campaigns, discount levels, legal regulations) over *time*. Such multi-dimensional data are often referred to as *panel data*, where the different units may be thought of as rows of a matrix and the time-steps as columns. A prominent framework for counterfactual inference using panel data is that of *synthetic control methods* (SCMs) Abadie and Gardeazabal (2003); Abadie et al. (2010), which aim to estimate the counterfactual outcome under *control*, i.e. no treatment, for units that were treated.

In the SCM framework, there is a *pre-intervention* time period, during which all units are under *control*. After which, each unit is given exactly one intervention from a set of possible interventions, which includes control, and they remain under that intervention for the remaining time steps, i.e. the *post-intervention* time period. To estimate the counterfactual under control for a treated unit during the post-intervention period, SCMs broadly follow two steps: (i) during the pre-intervention period, linearly regress the outcomes of a treated unit against the outcomes of the units that remain under control, i.e. the *donor* units; (ii) use the learned linear model and the outcomes of the donor units during the post-intervention period to produce a counterfactual trajectory of the treated unit under control. The way the linear model is learned can vary across different SCM estimators. While SCMs were originally applied to public policy domains Mourtgos et al. (2022); Freire (2018), they are now often used to estimate counterfactuals in various internet domains, ranging from surge pricing for ride-sharing applications Uber (2019) to the effectiveness of advertisements across different geographic regions EBay (2022). From the description of SCMs laid out above, it is clear that for the counterfactual estimate to be valid, the treated unit's potential outcomes under control should be linearly expressible by the observed outcomes of the donor units. To provide statistical guarantees on the performance of SCMs, such intuition has traditionally been made formal by making an *overlap* assumption on the relationship between the outcomes of the donor units and the test unit; for instance of the following form:

Unit Overlap Assumption. *Denote the potential outcome for unit i under intervention d at time t by $y_{i,t}^{(d)} \in \mathbb{R}$. For a given unit i and intervention d, there exists a set of weights $\omega^{(i,d)} \in \mathbb{R}^{N_d}$ such that*

$$\mathbb{E}[y_{i,t}^{(d)}] = \sum_{j \in [\![N_d]\!]} \omega^{(i,d)}[j] \cdot \mathbb{E}[y_{j,t}^{(d)}]$$

for all $t \in [\![T]\!]$, where T is the number of time-steps, N_d is the number of donor units who have received intervention d. The expectation is taken with respect to the randomness in the unit outcomes.

More broadly, previous works that provide theoretical guarantees for SCMs assume there exists some underlying mapping $\omega^{(i,d)}$ (e.g. linear or convex) through which the outcomes of a treated unit (unit i) may be expressed by the outcomes of the N_d donor units. Since such a condition appears to be necessary in order to make valid counterfactual inference, assumptions of this nature are ubiquitous when proving statistical guarantees about SCMs (e.g. Abadie et al. (2010); Amjad et al. (2018); Agarwal et al. (2020b, 2023)).[1]

However, despite their ubiquity, such overlap assumptions may not hold in all domains in which one would like to apply SCMs. For example, consider a streaming service with two service plans: a yearly subscription (treatment) and a pay-as-you-go model (control),

[1]Similar assumptions are also prevalent in the literature on other *matching*-based estimators such as *difference-in-differences* Donald and Lang (2007) or *clustering*-based methods Zhang et al. (2019).

and suppose that the streaming service wants to determine the effectiveness of its subscription program on user engagement (the unit outcome of interest). Under this setting, the subpopulation of streamers who self-select the subscription plan is most likely those who believe they will consume large amounts of content on the platform. In contrast, those who pay-as-they-go most likely believe they will consume less content. The two subpopulations may have very different experiences under the two business plans (due to their differing viewing habits), leading to two different sets of potential outcomes that may not overlap, i.e., are not linearly expressible in terms of each other. This could make it difficult to draw conclusions about the counterfactual user engagement levels of one subpopulation using the realized engagement levels of the other.In this work, our goal is to leverage tools from information design to incentivize the exploration of different treatments by non-overlapping unit subpopulations in order to obtain valid counterfactual estimates using SCMs. Specifically, we adopt tools and techniques from the literature on incentivizing exploration in multi-armed bandits (e.g. Mansour et al. (2015a, 2020b); Slivkins (2021)) to show how the principal—the platform running the SCM—can leverage knowledge gained from previous interactions with similar units to persuade the current unit to take an intervention they would not normally take. The principal achieves this by sending a *signal* or *recommendation* to the unit with information about which intervention is best for them. The principal's *recommendation policy* is designed to be *incentive-compatible*, i.e. it is in the unit's best interests to follow the intervention recommended to them by the policy. In our streaming service example, incentive-compatible signalling may correspond to recommending a service plan for each user based on their usage of the platform in a way that guarantees that units are better off in expectation when purchasing the recommended service plan. Our procedure ensures that the Unit Overlap Assumption becomes satisfied over time, which enables the principal to do valid counterfactual inference using off-the-shelf SCMs after they have interacted with sufficiently many units.

Overview of Results. We introduce an online learning model where the principal observes panel data from a population of units with differing preferences. We show that the Unit Overlap Assumption is indeed a necessary condition to obtain valid counterfactual estimates in our setting. To achieve overlap, we introduce a recommender system for incentivizing exploration when there are two interventions: treatment and control. At a high level, we adapt the "hidden exploration" paradigm from incentivized exploration in bandits to the panel data setting: First, randomly divide units into "exploit" units and "explore" units. For all exploit units, recommend the intervention that maximizes their estimated expected utility, given the data seen so far. For every explore unit, recommend the intervention for which the principal would like the Unit Overlap Assumption to be satisfied. Units are not aware of their exploit/explore designation by the algorithm, and the explore probability is chosen to be low enough such that the units have an incentive to follow the principal's recommendations. After recommendations have been made to sufficiently-many units, we show that with high probability, the Unit Overlap Assumption is satisfied for all units and either intervention of our choosing. This enables the use of existing SCMs to obtain finite sample guarantees for counterfactual estimation for all units under control after running Algorithm 5.

We extend our methods to the setting of synthetic interventions Agarwal et al. (2020b), in which the principal may wish to estimate counterfactual outcomes under different treatments in addition to control. We also provide two hypothesis tests for checking whether the Unit Overlap Assumption holds for a given treated unit and set of donor units. Finally, we empirically evaluate the performance of our recommender system. We find that it enables consistent counterfactual outcome estimation when the Unit Overlap Assumption does not hold *a priori* while existing SCMs alone do not.

5.2 Model

Notation. Subscripts are used to index the unit and time-step, while superscripts are reserved for interventions. We use i to index units, t to index time-steps, and d to index interventions. For $x \in \mathbb{N}$, we use the shorthand $[\![x]\!] := \{1, 2, \ldots, x\}$ and $[\![x]\!]_0 := \{0, 1, \ldots, x-1\}$. $\mathbf{y}[i]$ denotes the i-th component of vector $\mathbf{y}$, where indexing starts at 1. We sometimes use the shorthand $T_1 := T - T_0$ for $T, T_0 \in \mathbb{N}_{>0}$ such that $T > T_0$. $\Delta(\mathcal{X})$ denotes the set of possible probability distributions over $\mathcal{X}$. $\mathbf{0}_d$ is shorthand for the vector $[0, 0, \ldots, 0] \in \mathbb{R}^d$. Finally, we use the notation $a \wedge b := \min\{a, b\}$ and $a \vee b := \max\{a, b\}$.

Our Panel Data Setting. We consider an online setting in which the principal interacts with a sequence of n units one-by-one for T time steps each. We assume

Protocol: Incentivizing Exploration for Synthetic Control
For each unit $i \in [\![n]\!]$:

1. Principal observes pre-intervention outcomes $\mathbf{y}_{i,pre}$
2. Principal recommends an intervention $\widehat{d}_i \in \{0, 1\}$
3. Unit i chooses intervention $d_i \in \{0, 1\}$
4. Principal observes post-intervention outcomes $\mathbf{y}_{i,post}^{(d_i)}$

Figure 5.1: Summary of our setting.

that there is a pre-intervention period of T_0 time-steps, for which each unit $i \in [\![n]\!]$ is under the same intervention, i.e., under *control*. After the pre-intervention period, the principal either recommends the *treatment* to the unit or suggests that they remain under control (possibly using knowledge of the outcomes for units $j < i$). We denote the principal's *recommended* intervention to unit i by $\widehat{d}_i \in \{0, 1\}$, where 1 denotes the treatment and 0 the control. After receiving the recommendation, unit i chooses an intervention $d_i \in \{0, 1\}$ and remains under intervention d_i for the remaining $T - T_0$ time-steps.[2] We use

$$\mathbf{y}_{i,pre} := [y_{i,1}^{(0)}, \ldots, y_{i,T_0}^{(0)}]^\top \in \mathbb{R}^{T_0}$$

to denote unit i's pre-treatment outcomes under control, and

$$\mathbf{y}_{i,post}^{(d)} := [y_{i,T_0+1}^{(d)}, \ldots, y_{i,T}^{(d)}]^\top \in \mathbb{R}^{T-T_0}$$

to refer to unit i's post-intervention potential outcomes under intervention d. We denote the set of possible pre-treatment outcomes by $\mathcal{Y}_{pre}$. See Figure 5.1 for a summary of the setting we consider.

In order to impose structure on how unit outcomes are related for different units, time-steps, and interventions, we assume that outcomes are generated via the following *latent factor model*, a popular assumption in the literature.

[2] We envision a scenario in which both interventions are available to the units and they are free to choose between them.

Assumption 5.2.1 (Latent Factor Model)**.** *Suppose the outcome for unit i at time t under treatment $d \in \{0, 1\}$ takes the following factorized form*

$$\begin{aligned}\mathbb{E}[y_{i,t}^{(d)}] &= \langle \mathbf{u}_t^{(d)}, \mathbf{v}_i \rangle \\ y_{i,t}^{(d)} &= \mathbb{E}[y_{i,t}^{(d)}] + \varepsilon_{i,t}^{(d)}\end{aligned}$$

where $\mathbf{u}_t^{(d)} \in \mathbb{R}^r$ is a latent vector which depends only on the time-step t and intervention d, $\mathbf{v}_i \in \mathbb{R}^r$ is a latent vector which only depends on unit i, and $\varepsilon_{i,t}^{(d)}$ is zero-mean sub-Gaussian random noise with variance at most σ^2. For simplicity, we assume that $|\,\mathbb{E}[y_{i,t}^{(d)}]| \leq 1,\ \forall\, i \in [\![n]\!], t \in [\![T]\!], d \in \{0, 1\}$.

While we assume the existence of such structure in the data, we do *not* assume that the principal knows or gets to observe $\mathbf{u}_t^{(d)}$, $\mathbf{v}_i$, or $\varepsilon_{i,t}^{(d)}$. In the literature on synthetic control, the goal of the principal is to estimate unit-specific counterfactual outcomes under different interventions. Specifically, our target causal parameter is the (counterfactual) average expected post-intervention outcome.[3]

Definition 5.2.2. *(Average expected post-intervention outcome) The average expected post-intervention outcome of unit i under intervention d is*

$$\mathbb{E}[\bar{y}_{i,post}^{(d)}] \coloneqq \frac{1}{T_1} \sum_{t=T_0+1}^{T} \mathbb{E}[y_{i,t}^{(d)}]$$

where the expectation is taken with respect to $(\varepsilon_{i,t}^{(d)})_{T_0 < t \leq T}$.

5.2.1 Recommendations and Beliefs

The data generated by the interaction between the principal and a unit i may be characterized by the tuple $(\mathbf{y}_{i,pre}, \widehat{d}_i, d_i, \mathbf{y}_{i,post}^{(d_i)})$, where $\mathbf{y}_{i,pre}$ are unit i's pre-treatment outcomes, $\widehat{d}_i$ is the intervention recommended to unit i, d_i is the intervention taken by unit i, and $\mathbf{y}_{i,post}^{(d_i)}$ are unit i's post-intervention outcomes under intervention d_i.

[3]Our results hold in the more general setting where the target causal parameter is *any linear* function of the vector of expected post-intervention outcomes. We focus on the average as it is a common quantity of interest in the literature.

Definition 5.2.3 (Interaction History). *The interaction history at unit i is the sequence of outcomes, recommendations, and interventions for all units $j \in [\![i-1]\!]$. Formally,*

$$H_i := \{(\mathbf{y}_{j,pre}, \widehat{d}_j, d_j, \mathbf{y}_{j,post}^{(d_j)})\}_{j=1}^{i-1}.$$

We denote the set of all possible histories at unit i as $\mathcal{H}_i$.

We refer to the way the principal assigns interventions to units as a *recommendation policy.*

Definition 5.2.4 (Recommendation Policy). *A recommendation policy $\pi_i : \mathcal{H}_i \times \mathcal{Y}_{pre} \to \Delta(\{0,1\})$ is a (possibly stochastic) mapping from histories and pre-treatment outcomes to interventions.*

We assume that before the first unit arrives, the principal commits to a sequence of recommendation policies $\{\pi_i\}_{i=1}^n$ which are fully known to all units. Whenever π_i is clear from the context, we use the shorthand $\widehat{d}_i = \pi_i(H_i, \mathbf{y}_{i,pre})$ to denote the recommendation of policy π_i to unit i.

In addition to having a corresponding latent factor, each unit has an associated *belief* over the effectiveness of each intervention. This is made formal through the following definitions.

Definition 5.2.5 (Unit Belief). *Unit i has* prior belief $\mathcal{P}_{\mathbf{v}_i}$, *which is a joint distribution over potential post-intervention outcomes*
$\{\mathbb{E}[\mathbf{y}_{i,post}^{(0)}], \mathbb{E}[\mathbf{y}_{i,post}^{(1)}]\}$. *We use the shorthand $\mu_{v_i}^{(d)} := \mathbb{E}_{\mathcal{P}_{\mathbf{v}_i}}[\bar{y}_{i,post}^{(d)}]$ to refer to a unit's expected average post-intervention outcome, with respect to their prior $\mathcal{P}_{\mathbf{v}_i}$.*

Definition 5.2.6 (Unit Type). *Unit i is of type $\tau \in \{0,1\}$ if $\tau = \arg\max_{d \in \{0,1\}} \mu_{v_i}^{(d)}$. We denote the set of all units of type τ as $\mathcal{T}^{(\tau)}$ and the dimension of the latent subspace spanned by type τ units as $r_\tau := \mathrm{span}(\mathbf{v}_j : j \in \mathcal{T}^{(\tau)})$.*

In other words, a unit's type is the intervention they would take according to their prior without any additional information about the effectiveness of each intervention. We consider the setting in which the possible latent factors associated with each type lie in *mutually orthogonal* subspaces (i.e. the Unit Overlap Assumption is *not* satisfied).[4]

[4]Note that $\sum_\tau r_\tau \leq r$.

See Section 5.3 for a simple theoretical example of such a data-generating process and Section 5.7 for an empirical example. If the principal could get the Unit Overlap Assumption to be satisfied, they could use existing synthetic control methods off-the-shelf to obtain valid finite sample guarantees.

As is standard in the literature on IE, we assume that units are Bayesian-rational and each unit knows its place in the sequence of n units (i.e., their index $i \in [\![n]\!]$). Thus given recommendation $\widehat{d}_i$, unit i selects their intervention d_i such that $d_i \in \arg\max_{d \in \{0,1\}} \mathbb{E}_{\mathcal{P}_{\mathbf{v}_i}}[\bar{y}_{i,post}^{(d)} | \widehat{d}_i]$, i.e. they select the intervention d_i which maximizes their utility in expectation over their prior $\mathcal{P}_{\mathbf{v}_i}$, conditioned on receiving recommendation $\widehat{d}_i$ from the principal.

Definition 5.2.7 (Bayesian Incentive-Compatibility)**.** *We say that a recommendation d is* Bayesian incentive-compatible *(BIC) for unit i if, conditional on receiving intervention recommendation $\widehat{d}_i = d$, unit i's average expected post-intervention outcome under intervention d is at least as large as their average expected post-intervention outcome under any other intervention:*

$$\mathop{\mathbb{E}}_{\mathcal{P}_{\mathbf{v}_i}}[\bar{y}_{i,post}^{(d)} - \bar{y}_{i,post}^{(d')} | \widehat{d}_i = d] \geq 0 \quad \textit{for every } d' \in \{0,1\}.$$

A recommendation policy is BIC if the above condition holds for every intervention which is recommended with non-zero probability.

5.3 Preliminaries

Suppose that the principal would like to estimate counterfactuals for some unit under intervention d. We begin by showing that the Unit Overlap Assumption must be satisfied for this unit under intervention d in order for any algorithm to obtain consistent counterfactual estimates under the setting described in Section 5.2.

Theorem 5.3.1. *Fix a time horizon T, pre-intervention time period T_0, and number of donor units $n^{(0)}$. For any algorithm used to estimate the average post-intervention outcome under control for a test unit, there exists a problem instance such that the produced estimate has constant error whenever the Unit Overlap Assumption is not satisfied for the test unit under control, even as T, T_0, $n^{(0)} \to \infty$.*

The proof of Theorem 5.3.1 proceeds by constructing a family of problem instances under which the Unit Overlap Assumption is not satisfied, before showing that no algorithm can consistently estimate $\mathbb{E}[\bar{y}^{(0)}_{n^{(0)}+1,post}]$ on all instances in the family.

5.3.1 On Estimating Unit-Specific Counterfactual Outcomes

To infer the unit outcomes in the post-intervention period from outcomes in the pre-intervention period, we require the following *linear span inclusion* condition to hold on the latent factors associated with the pre- and post-intervention time-steps:

Assumption 5.3.2 (Linear Span Inclusion)**.** *For each intervention $d \in \{0, 1\}$ and time $t > T_0$, we assume that*

$$\mathbf{u}^{(d)}_t \in \text{span}\{\mathbf{u}^{(0)}_1, \mathbf{u}^{(0)}_2, \ldots, \mathbf{u}^{(0)}_{T_0}\}.$$

Much like the Unit Overlap Assumption, Assumption 5.3.2 is ubiquitous in the literature on robust synthetic control (see, e.g. Agarwal et al. (2020b, 2023); Harris et al. (2022)). In essence, Assumption 5.3.2 requires that the pre-intervention period be "sufficiently informative" about the post-intervention period. Indeed, Assumption 5.3.2 may be thought of as the "transpose" of the Unit Overlap Assumption, where the "overlap" condition is required to hold on the time-steps instead of on the units.[5] To gain intuition for why such an assumption is necessary, consider the limiting case in which $\mathbf{u}^{(0)}_t = \mathbf{0}_r$ for all $t \leq T_0$. Under such a setting, all expected unit outcomes in the pre-intervention time period will be 0, regardless of the underlying unit latent factors, which makes it impossible for the principal to infer anything non-trivial about a unit's post-intervention outcomes from looking at their pre-intervention outcomes alone.

Vertical vs. Horizontal Regression. The following proposition is an important structural result for our main algorithm, as it will allow us to transform the problem of estimating unit counterfactual outcomes from a regression over donor units (*vertical regression*) to a regression over pre-intervention outcomes (*horizontal regression*). While Proposition 5.3.3 is known in the literature on learning from panel data (*e.g.,* (Agarwal et al., 2019, 2020b)), we state it here for completeness.

[5] Note that Assumption 5.3.2 is not a substitute for the Unit Overlap Assumption by Theorem 5.3.1.

Proposition 5.3.3. *Under Assumption 5.2.1 and 5.3.2, there exists a slope vector* $\boldsymbol{\theta}^{(d)} \in \mathbb{R}^{T_0}$ *such that the average expected post-intervention outcome of unit* i *under intervention* d *is given by*

$$\mathbb{E}[\bar{y}_{i,post}^{(d)}] = \frac{1}{T_1}\langle \boldsymbol{\theta}^{(d)}, \mathbb{E}[y_{i,pre}]\rangle.$$

We assume that the principal has knowledge of a valid upper-bound Γ *on the* ℓ_2*-norm of* $\boldsymbol{\theta}^{(d)}$*, i.e.* $\|\boldsymbol{\theta}^{(d)}\|_2 \leq \Gamma$ *for* $d \in \{0, 1\}$*.*

Given historical data of the form $\{\mathbf{y}_{j,pre}, \mathbf{y}_{j,post}^{(d_j)}\}_{j=1}^{i-1}$, we can estimate $\boldsymbol{\theta}^{(d)}$ as $\widehat{\boldsymbol{\theta}}_i^{(d)}$ using some estimation procedure, which in turn allows us to estimate $\mathbb{E}[\bar{y}_{i,post}^{(d)}]$ as $\widehat{\mathbb{E}}[\bar{y}_{i,post}^{(d)}] := \frac{1}{T_1}\langle \widehat{\boldsymbol{\theta}}_i^{(d)}, y_{i,pre}\rangle$ for $d \in \{0, 1\}$. While this is not a synthetic control method in the traditional sense, as it does not construct a "synthetic control" trajectory from the trajectories of the donor units, it turns out that the point estimate it produces (and assumptions it requires) are equivalent to that of a vertical regression-based SCM in many settings (including ours). See Shen et al. (2022) for more background on the similarities and differences between using horizontal regression-based approaches (like ours) and vertical regression-based approaches (like traditional SCMs) in panel data settings. We choose to use a horizontal regression-based approach because it is easier to work with in our setting, where units arrive sequentially (recall Figure 5.1). Importantly, we would still need the Unit Overlap Assumption to be satisfied for intervention d if we used a vertical regression-based method.

For the reader familiar with bandit algorithms, it may be useful to draw the following connection between our setting and that of linear contextual bandits.[6] In particular, it is sometimes helpful to think of $\mathbb{E}[y_{i,pre}]$ as the "context" of unit i, and $\mathbb{E}[\bar{y}_{i,post}^{(d)}]$ as the principal's expected reward of assigning intervention d given $\mathbb{E}[y_{i,pre}]$. Since we observe $y_{i,pre}$ instead of $\mathbb{E}[y_{i,pre}]$ we cannot apply linear contextual bandit algorithms to our panel data setting out-of-the-box. However as we will show in Section 5.4, one can combine ideas from incentivizing exploration in contextual bandits with principled ways of handling the noise in the pre-intervention outcomes to incentivize exploration in panel data settings.

[6]For the reader unfamiliar with bandit algorithms, this paragraph may be skipped without any loss in continuity.

Algorithm 5: Incentivizing Exploration for Synthetic Control: Type 1 units

Input: First stage length N_0, batch size L, number of batches B, failure probability δ, gap $C \in (0, 1)$

Provide no recommendation to first N_0 units.
for *batch* $b = 1, 2, \ldots, B$ **do**
  Select an explore index $i_b \in [\![L]\!]$ uniformly at random.
  for $j = 1, 2, \ldots, L$ **do**
    if $j = i_b$ **then**
      Recommend intervention $\widehat{d}_{N_0+(b-1)\cdot L+j} = 0$ to unit $N_0 + (b-1) \cdot L + j$.
    else
      if $\mu^{(0,1,l)} - \widehat{\mathbb{E}}[\bar{y}^{(1)}_{j,post}] \geq C$ **then**
        Recommend intervention $\widehat{d}_{N_0+(b-1)\cdot L+j} = 0$ to unit $N_0 + (b-1) \cdot L + j$.
      else
        Recommend intervention $\widehat{d}_{N_0+(b-1)\cdot L+j} = 1$ to unit $N_0 + (b-1) \cdot L + j$.
      end
    end
  end
end

Principal Component Regression. The specific algorithm we use to estimate $\boldsymbol{\theta}^{(d)}$ is called *principal component regression* (PCR). At a high level, PCR "de-noises" the matrix of covariates (in our case, the set of all pre-treatment outcomes for all donor units seen so far who underwent intervention d) using hard singular value thresholding, before regressing the (now-denoised) covariates on the outcomes of interest (in our case, the average post-intervention outcome under intervention d). See Section C.I for the mathematical details of how PCR is applied in our setting.

5.4 Incentivized Exploration for Synthetic Control

In this section, we turn our focus to incentivizing type 1 units (i.e. units who *a priori* prefer the treatment) to remain under control in the post-intervention period. The methods we present may also be applied to incentivize type 0 units (i.e. units who *a priori* prefer the control) to take the treatment. We focus on incentivizing control amongst type 1 units in order to be in line with the literature on synthetic control, which aims to estimate counterfactual unit outcomes under control.

The goal of the principal is to design a sequence of recommendation policies that convinces enough units of type 1 to select the control in the post-treatment period, such that the Unit Overlap Assumption is satisfied for both interventions for all units of type 1. At a high level, the principal is able to incentivize units to explore because they have more information than any one particular unit (as they view the realized history H_i before making a recommendation), and so they can selectively reveal information in order to persuade units to take interventions which they would not normally take. Our algorithm (Algorithm 5) is inspired by the "detail free" algorithm for incentivizing exploration in multi-armed bandits in Mansour et al. (2015a). Furthermore, through our discussion in Section 5.3, one may view Algorithm 5 as an algorithm for incentivizing exploration in linear contextual bandit settings with measurement error in the context.

The recommendation policy in Algorithm 5 is split into two stages. In the first stage, the principal provides no recommendations to the first N_0 units. Left to their own devices, these units take their preferred intervention according to their prior belief: type 0 units take control, and type 1 units take the treatment. We choose N_0 such that it is large enough for the Unit Overlap Assumption to be satisfied for all units of type 1 under the treatment with high probability.[7]

In the second stage, we use the initial data collected during the first stage to construct an estimator of the average expected post-intervention outcome for type 1 units under treatment using PCR, as described in Section 5.3.1. By dividing the total number of units in the second stage into phases of L rounds each, the principal can randomly "hide" one *explore* recommendation amongst $L-1$ *exploit* recommendations. When the principal sends an exploit recommendation to unit i, they recommend the intervention which would result in the highest estimated average post-intervention outcome for unit i, where unit i's post-intervention outcomes under treatment are estimated using the data collected during the first stage. On the other hand, the principal recommends that unit i take the control whenever they send an explore recommendation.

Thus under Algorithm 5, if a type 1 unit receives a recommendation to take the

[7]The Unit Overlap Assumption will always be satisfied w.h.p. for type 1 units under the treatment after seeing sufficiently-many type 1 units. The hard part is convincing enough type 1 units to remain under control so that the Unit Overlap Assumption is also satisfied for control.

treatment, they will always follow the recommendation since they can infer they must have received an exploit recommendation. However if a type 1 unit receives a recommendation to take the control, they will be unsure if they have received an explore or an exploit recommendation, and can therefore be incentivized to follow the recommendation as long as L is set to be large enough (i.e. the probability of an explore recommendation is sufficiently low). This works because the sender's estimate of $\mathbb{E}[\bar{y}_{i,post}^{(1)}]$ is more informed than unit i's prior belief, due to the history of outcomes H_i that the principal has observed.

Our algorithm requires the following assumptions on the principal and units' knowledge about the interaction:

Assumption 5.4.1. *[Unit Knowledge for Synthetic Control] We assume that the following are common knowledge among all units and the principal:*

1. *Valid upper- and lower-bounds on the fraction of type* 1 *units in the population, i.e. if the true proportion of type* 1 *units is* $p_1 \in (0,1)$*, the principal and units know* $p_L, p_H \in (0,1)$ *such that* $p_L \leq p_1 \leq p_H$*.*

2. *Valid upper- and lower-bounds on the prior mean for each receiver type and intervention, i.e. values* $\mu^{(d,\tau,h)}, \mu^{(d,\tau,l)} \in \mathbb{R}$ *such that* $\mu^{(d,\tau,l)} \leq \mu_{v_i}^{(d)} \leq \mu^{(d,\tau,h)}$ *for all units* $i \in [\![N]\!]$*, receiver types* $\tau \in \{0,1\}$*, and interventions* $d \in \{0,1\}$*.*

3. *For some* $C \in (0,1)$*, a lower bound on the smallest probability of the event* $\xi_{C,i}$ *over the priors and latent factors of type* 1 *units, denoted by* $\min_{i \in \mathcal{T}^{(1)}} \Pr_{\mathcal{P}_i}[\xi_{C,i}] \geq \zeta_C > 0$*, where*
$$\xi_{C,i} := \left\{ \mu_i^{(0)} \geq \bar{y}_{i,post}^{(1)} + C \right\}$$
and $\bar{y}_{i,post}^{(1)}$ *is the average post-intervention outcome for unit* i *under intervention* 1*.*

4. *A sufficient number of observations* N_0 *needed such that the Unit Overlap Assumption is satisfied for type* 1 *units <u>under the treatment</u> with probability at least* $1-\delta$*, for some* $\delta \in (0,1)$*.*

Parts (1), (2), and (4) posit that bounds on various aggregate statistics about the underlying unit population are common knowledge amongst the principal and all units.

At a high level, part (3) requires that (i) there is a non-zero probability under each unit i's prior that $\bar{y}_{i,post}^{(0)} \geq \bar{y}_{i,post}^{(1)}$ and (ii) the principal will be able to infer a lower bound on this probability given the set of observed outcomes for units in the first phase.[8]

We are now ready to present our main result: a way to initialize Algorithm 5 to ensure that the Unit Overlap Assumption is satisfied with high probability.

Theorem 5.4.2 (Informal; detailed version in Theorem C.III.4)**.** *Under Assumption 5.4.1, if the number of initial units N_0 and the batch size L are chosen to be sufficiently large, then Algorithm 5 satisfies the BIC property (Definition 5.2.7).*

Moreover, if the number of batches B is chosen to be sufficiently large, then the conditions of the Unit Overlap Assumption will be satisfied simultaneously for all type 1 units under control w.h.p.

Proof Sketch. See Section C.III for complete proof details. At a high level, the analysis follows by expressing the compliance condition for type 1 units as different cases depending on the principal's recommendation. In particular, a type 1 unit could receive recommendation $\widehat{d}_i = 0$ for two reasons: (1) Under event $\xi_{C,i}$ when control is indeed the better intervention according to the prior and the observed outcomes of previous units, or (2) when the unit is randomly selected as an explore unit. Using bounds on the probabilities of these two events occurring, we can derive a condition on the minimum phase length L such that a unit's expected gain from exploiting (when the event $\xi_{C,i}$ happens) exceeds the expected (unit) loss from exploring. We then further simplify the condition on the phase length L so that it is computable by the principal by leveraging existing finite sample guarantees for PCR using the samples collected in the first stage when no recommendations are given. $\square$

Theorem 5.4.2 says that after running Algorithm 5 with optimally-chosen parameters, the Unit Overlap Assumption will be satisfied for all type 1 units *under both interventions* with high probability. Therefore after running Algorithm 5 on sufficiently-many units, the principal can use off-the-shelf SCMs (e.g. Agarwal et al. (2020b, 2023)) to obtain counterfactual estimates with valid confidence intervals for all type 1 units under control (w.h.p.). Note that all parameters required to run Algorithm 5 are computable by the principal in polynomial time. See Example C.III.5 for a concrete instantiation of Algorithm 5.

[8] Such "fighting chance" assumptions are very common (and are usually necessary) in the literature on incentivized exploration in bandits.

Algorithm 6: Incentivizing Exploration for Synthetic Interventions: sub-type $\boldsymbol{\tau}$ units

Input: Group size L, number of batches B, failure probability δ, fixed gap $C \in (0, 1)$

Provide no recommendation to first N_0 units.
// Loop through explore interventions
for $\ell = \boldsymbol{\tau}[k], \boldsymbol{\tau}[k-1], \ldots, \boldsymbol{\tau}[2]$ **do**
    for *batch* $b = 1, 2, \cdots, B$ **do**
        Select an explore index $i_b \in [L]$ uniformly at random.
        for $j = 1, 2, \cdots, L$ **do**
            if $j = i_b$ **then**
                Recommend intervention $\widehat{d}_{N_0+(b-1)\cdot L+j} = \ell$
            else
                if $\widehat{\mathbb{E}}[\bar{y}_{i,post}^{(\boldsymbol{\tau}[1])}] + C < \widehat{\mathbb{E}}[\bar{y}_{i,post}^{(d)}] < \mu^{(\ell,\boldsymbol{\tau},l)} - C$ *for every intervention* $d \in \{\ell - 1, \ldots, \boldsymbol{\tau}[k]\}$ **then**
                    Recommend intervention $\widehat{d}_{N_0+(b-1)\cdot L+j} = \ell$
                else
                    Recommend intervention $\widehat{d}_{N_0+(b-1)\cdot L+j} = \boldsymbol{\tau}[1]$
                end
            end
        end
    end
end

5.5 Extension to Synthetic Interventions

Our focus so far has been to incentivize units who *a priori* prefer the (single) treatment to take the control in order to obtain accurate counterfactual estimates for all units under control. In this section, we extend Algorithm 5 to the setting where there are *multiple* treatments. Now our goal is to incentivize enough exploration across all interventions such that the Unit Overlap Assumption is satisfied *for every intervention* and thus our counterfactual estimates are valid simultaneously for all units in the population under every intervention. In order to do so, we use tools from the literature on *synthetic interventions*, which is a generalization of the synthetic control framework to allows for counterfactual estimation under different treatments, in addition to control Agarwal et al. (2020b, 2023).

Our setup is the same as in Section 5.4, with the only difference being that each unit may choose one of $k \geq 2$ interventions after the pre-treatment period. We assume that Assumption 5.2.1, Definition 5.2.4, Definition 5.2.5, and Assumption 5.3.2 are all extended to hold under k interventions. Under this setting, a unit's beliefs induce a preference ordering over different interventions. We capture this through the notion of a unit *sub-type*, which is a generalization of our definition of unit type to the setting with k interventions.

Definition 5.5.1 (Unit Sub-type)**.** *A unit sub-type $\boldsymbol{\tau} \in [k]_0^k$ is a preference ordering over interventions. Unit i is of sub-type $\boldsymbol{\tau}$ if for every $\kappa \leq k$,*

$$\boldsymbol{\tau}[\kappa] = \arg \max_{d \in [\![k]\!]_0 \setminus \{\boldsymbol{\tau}[\kappa']\}_{\kappa' < \kappa}} \mu_{v_i}^{(d)}.$$

Unit i's type is $\boldsymbol{\tau}[1]$. We denote the set of all units of sub-type $\boldsymbol{\tau}$ as $\mathcal{T}^{(\boldsymbol{\tau})}$ and the dimension of the latent subspace spanned by sub-type $\boldsymbol{\tau}$ units as $r_{\boldsymbol{\tau}} := \mathrm{span}(\mathbf{v}_j : j \in \mathcal{T}^{(\boldsymbol{\tau})})$.

Intuitively, a unit sub-type $\boldsymbol{\tau}$ is just the preference ordering over interventions such that intervention $\boldsymbol{\tau}[a]$ is preferred to intervention $\boldsymbol{\tau}[b]$ for $a < b$. Since different unit sub-types have different preference orderings over interventions, we design our algorithm (Algorithm 6) to be BIC for all units of a given sub-type (in contrast to Algorithm 5, which is BIC for all units of a given type). Note that can be up to $k!$ possible sub-types, but the principal will only have to run Algorithm 6 on at most r different sub-types in order to ensure that the Unit Overlap Assumption is satisfied for all units under all interventions with high probability, as long as they know which unit sub-populations overlap and which do not. This is because if the Unit Overlap Assumption is satisfied for a specific unit sub-type whose latent factors lie within a particular subspace, it is also satisfied for all other units whose latent factors lie within the same subspace, and there are of course at most r different subspaces by Assumption 5.2.1.

For a given sub-type $\boldsymbol{\tau}$, the goal of Algorithm 6 is to incentivize sufficient exploration for the Unit Overlap Assumption to be satisfied for all interventions with high probability. As was the case in Section 5.4, the algorithm will still be split into two phases, and units will not be given any recommendation in the first phase. The main difference compared to Algorithm 5 is how the principal chooses the exploit intervention in the second phase.

When there are only two interventions, the principal can compare their estimate of the average counterfactual outcome under treatment to the lower bound on the prior-mean average counterfactual outcome for control for type 1 units (and vice versa for type 0 units). However, a more complicated procedure is required to maintain Bayesian incentive-compatibility when there are more than two interventions.

Instead, in Algorithm 6, the principal will set the exploit intervention as follows: For receiver sub-type $\boldsymbol{\tau}$, Algorithm 6 sets intervention ℓ to be the exploit intervention for unit i only if the sample average of *all* interventions $d \in \{\ell - 1, \ldots, \boldsymbol{\tau}[k]\}$ are both (1) larger than the sample average of intervention $\boldsymbol{\tau}[1]$ by some constant gap C and (2) less than the lower bound on the prior-mean average counterfactual outcome $\mu_i^{(\ell)}$ by C. If no such intervention satisfies both conditions, intervention $\boldsymbol{\tau}[1]$ is chosen to be the exploit intervention.

We require the following "common knowledge" assumption for Algorithm 6, which is analogous to Assumption 5.4.1:

Assumption 5.5.2. *We assume that the following are common knowledge among all units and the principal:*

1. *Valid upper- and lower-bounds on the fraction of sub-type $\boldsymbol{\tau}$ units in the population, i.e. if the true proportion of sub-type $\boldsymbol{\tau}$ units is $p_{\boldsymbol{\tau}} \in (0, 1)$, the principal knows $p_{\boldsymbol{\tau},L}, p_{\boldsymbol{\tau},H} \in (0, 1)$ such that $p_{\boldsymbol{\tau},L} \leq p_{\boldsymbol{\tau}} \leq p_{\boldsymbol{\tau},H}$.*

2. *Valid upper- and lower-bounds on the prior mean for each receiver sub-type and intervention, i.e. values $\mu^{(d,\boldsymbol{\tau},h)}, \mu^{(d,\boldsymbol{\tau},l)} \in \mathbb{R}$ such that $\mu^{(d,\boldsymbol{\tau},l)} \leq \mu_i^{(d)} \leq \mu^{(d,\boldsymbol{\tau},h)}$ for all receiver sub-types $\boldsymbol{\tau}$, interventions $d \in [\![k]\!]_0$ and all units $i \in \mathcal{T}^{(\boldsymbol{\tau})}$.*

3. *For some $C \in (0, 1)$, a lower bound on the smallest probability of the event $\mathcal{E}_{C,i}$ over the prior and latent factors of sub-type $\boldsymbol{\tau}$ units, denoted by $\min_{i \in \mathcal{T}^{(\boldsymbol{\tau})}} \Pr_{\mathcal{P}_i}[\mathcal{E}_{C,i}] \geq \zeta_C > 0$, where*

$$\mathcal{E}_{C,i} := \{\forall j \in [\![k]\!]_0 \backslash \{\boldsymbol{\tau}[1], \boldsymbol{\tau}[k]\} : \bar{y}_{i,post}^{(\boldsymbol{\tau}[1])} + C \leq \bar{y}_{i,post}^{(j)} \leq \mu_i^{(\boldsymbol{\tau}[k])} - C\}$$

   and $\bar{y}_{i,post}^{(\ell)}$ is the average post-intervention outcome for unit i under intervention $\ell \in [\![k]\!]_0$.

4. *A sufficient number of observations N_0 needed such that the Unit Overlap Assumption is satisfied for sub-type $\boldsymbol{\tau}$ units under intervention $\boldsymbol{\tau}[1]$ with probability at least $1 - \delta$, for some $\delta \in (0, 1)$.*

Theorem 5.5.3 (Informal; detailed version in Theorem C.IV.1)**.** *Under Assumption 5.5.2, if the number of initial units N_0 and batch size L are chosen to be sufficiently large, then Algorithm 6 satisfies the BIC property (Definition 5.2.7).*

Moreover, if the number of batches B is chosen to be sufficiently large, then the Unit Overlap Assumption will be satisfied for all units of sub-type $\boldsymbol{\tau}$ under all interventions with high probability.

See Section C.IV for the proof. The main difference compared to the analysis of Algorithm 5 is in the definition of the "exploit" intervention: First, when there are k interventions, an intervention ℓ is chosen as the exploit intervention only when the sample average of each intervention $d \in \{\ell - 1, \ldots, \boldsymbol{\tau}[k]\}$ is (i) larger than the sample average of intervention $\boldsymbol{\tau}[1]$, and (ii) smaller than the prior-mean average outcome of intervention ℓ (with some margin C). Second, instead of choosing between a "previously best" intervention (*i.e.,* an intervention $\boldsymbol{\tau}[j] < \ell$ with the highest sample average) and intervention ℓ to be the 'exploit' intervention, Algorithm 6 always choose between $\boldsymbol{\tau}[1]$ and ℓ. This is because in the former case, conditional on any given intervention being chosen as the exploit intervention, it is an open technical challenge to show that the exploit intervention will have a higher average expected outcome compared to all other interventions.[9]

5.6 Testing Whether the Unit Overlap Assumption Holds

So far, our focus has been on designing algorithms to incentivize subpopulations of units to take interventions for which the Unit Overlap Assumption does not initially hold. In this section we study the complementary problem of designing a procedure for determining whether the Unit Overlap Assumption holds for a given test unit and set of donor units. In particular, we provide two simple hypothesis tests for deciding whether the Unit Overlap Assumption holds, by leveraging existing finite sample guarantees for PCR. Our first hypothesis test is non-asymptotic, and thus is valid under any amount of data. While our second hypothesis test is asymptotic and therefore valid only in the limit, we anticipate that it may be of more practical use.

[9]This observation is in line with work on frequentist-based BIC algorithms for incentivizing exploration in bandits. See Section 6.2 of Mansour et al. (2015a) for more details.

5.6.1 A Non-Asymptotic Hypothesis Test

Our first hypothesis test proceeds as follows: Using the PCR techniques described in Section C.I, we learn a linear relationship between the first $T_0/2$ time-steps in the pre-intervention time period and the average outcome in the second half of the pre-intervention time period using data from the donor units. Using this learned relationship, we then compare our estimate for the test unit with their *true* average outcome in the second half of the pre-intervention time period. If the difference between the two is larger than the confidence interval given by PCR (plus an additional term to account for the noise), we can conclude that the Unit Overlap Assumption is not satisfied with high probability.

In order for our hypothesis test to be valid we require the following assumption, which is analogous to Assumption 5.3.2.

Assumption 5.6.1 (Linear Span Inclusion, revisited)**.** *For every time-step t such that $T_0/2 < t \leq T_0$, we assume that*

$$\mathbf{u}_t^{(0)} \in \text{span}\{\mathbf{u}_1^{(0)}, \ldots, \mathbf{u}_{T_0/2}^{(0)}\}.$$

While Assumption 5.3.2 requires that the post-intervention latent factors for any intervention fall within the linear span of the pre-intervention latent factors, Assumption 5.6.1 requires that the second half of the pre-intervention latent factors fall within the linear span of the first half. This assumption allows us to obtain valid confidence bounds when regressing the second half of the pre-intervention outcomes on the first half. Let $y_{i,pre'} := [y_{i,1}^{(0)}, \ldots, y_{i,T_0/2}^{(0)}]$, $y_{i,pre''} := [y_{i,T_0/2+1}^{(0)}, \ldots, y_{i,T_0}^{(0)}]$, and $\bar{y}_{i,pre''} := \frac{2}{T_0} \sum_{t=T_0/2+1}^{T_0} y_{i,t}^{(0)}$. Under Assumption 5.6.1 we can estimate $\mathbb{E}[\bar{y}_{i,pre''}]$ as $\widehat{\mathbb{E}}[\bar{y}_{i,pre''}] := \langle \widehat{\boldsymbol{\theta}}_i, y_{i,pre'} \rangle$, where $\widehat{\boldsymbol{\theta}}_i$ is computed using the PCR techniques in Section C.I.

We are now ready to introduce our hypothesis test. In what follows, we will leverage the high-probability confidence interval of Agarwal et al. (2023) (Theorem C.III.1). Under Assumption 5.6.1, we can apply Theorem C.III.1 out-of-the-box to estimate $\mathbb{E}[\bar{y}_{i,pre''}]$ using the first $T_0/2$ time-steps as the pre-intervention period and the next $T_0/2$ time-steps as the post-intervention period.

Non-Asymptotic Hypothesis Test. Consider the hypothesis that the Unit Overlap Assumption is not satisfied for a test unit n.

1. Using donor units $\mathcal{I}$ and test unit n, compute $\widehat{\mathbb{E}}[\bar{y}_{n,pre''}]$ via the PCR method of Section C.I.

2. Let $\alpha(\delta)$ be the confidence bound given in Theorem C.III.1 s.t.

$$|\mathbb{E}[\bar{y}_{i,pre''}] - \widehat{\mathbb{E}}[\bar{y}_{n,pre''}]| \leq \alpha(\delta)$$

   with probability at least $1 - \delta$. Specifically, compute $\alpha(\delta) + 2\sigma\sqrt{\frac{\log(1/\delta)}{T_0}}$ for confidence level δ by applying Theorem C.III.1 with the first $T_0/2$ time-steps as the pre-intervention period and the next $T_0/2$ time-steps as the post-intervention period.

3. If $|\widehat{\mathbb{E}}[\bar{y}_{n,pre''}] - \bar{y}_{n,pre''}| > \alpha(\delta) + 2\sigma\sqrt{\frac{\log(1/\delta)}{T_0}}$ then accept the hypothesis. Otherwise, reject.

Theorem 5.6.2. *Under Assumption 5.6.1, if the Unit Overlap Assumption is satisfied for unit n, then*

$$|\widehat{\mathbb{E}}[\bar{y}_{n,pre''}] - \bar{y}_{n,pre''}| \leq \alpha(\delta) + 2\sigma\sqrt{\frac{\log(1/\delta)}{T_0}}$$

with probability $1 - \mathcal{O}(\delta)$, where $\alpha(\delta)$ is the high-probability confidence interval which is defined in Theorem C.III.1 when using the first $T_0/2$ time-steps as the pre-intervention period and the next $T_0/2$ time-steps as the post-intervention period.

Therefore we can conclude that if $|\widehat{\mathbb{E}}[\bar{y}_{i,pre''}] - \bar{y}_{i,pre''}| > \alpha(\delta) + 2\sigma\sqrt{\frac{\log(1/\delta)}{T_0}}$, then the Unit Overlap Assumption will not be satisfied for unit i w.h.p.

5.6.2 An Asymptotic Hypothesis Test

Next we present a hypothesis test which, while only valid in the limit, may be of more practical use when compared to the non-asymptotic hypothesis test of Section 5.6.1. At a high level, our asymptotic hypothesis test leverages Assumption 5.6.1 and the guarantees for synthetic interventions (SI) in Agarwal et al. (2020b) to determine whether the Unit Overlap Assumption holds for a given test unit n. In particular, their results imply that under Assumption 5.6.1 and the Unit Overlap Assumption, a rescaled version of $\widehat{\mathbb{E}}[\bar{y}_{n,pre''}] - \mathbb{E}[\bar{y}_{n,pre''}]$ converges in distribution to the standard normal distribution as $|\mathcal{I}|, T_0, T_1 \to \infty$. While the rescaling amount depends on quantities which are not computable, a relatively simple application of the continuous mapping theorem and Slutsky's theorem imply that we can compute a valid test statistic which converges in distribution to the standard normal.

Notation. In what follows, let $\omega^{(n,0)} \in \mathbb{R}^{|\mathcal{I}|}$ be the linear relationship between the test unit n and donor units $\mathcal{I}$ which is known to exist under the Unit Overlap Assumption, i.e. $\mathbb{E}[y_{n,t}^{(0)}] = \sum_{i \in \mathcal{I}} \omega^{(n,0)}[i] \cdot \mathbb{E}[y_{i,t}^{(0)}]$. Let $\widetilde{\omega}^{(n,0)}$ be the projection of $\omega^{(n,0)}$ onto the subspace spanned by the latent factors of the $\mathcal{I}$ test units. Let $\widehat{\omega}^{(n,0)}$ be the estimate of $\widetilde{\omega}^{(n,0)}$ and $\widehat{\sigma}$ be the estimate of the standard deviation of the noise terms $\{\varepsilon_{i,t} \;:\; i \in \mathcal{I}, t \in [\![T]\!]\}$ given by SI.[10] Finally, let z_α be the z score at significance level $\alpha \in (0,1)$, i.e. $\alpha = \mathbb{P}(x \leq z_\alpha)$, where $x \sim \mathcal{N}(0,1)$.

Asymptotic Hypothesis Test. Consider the hypothesis that the Unit Overlap Assumption is not satisfied for a test unit n.

1. Using donor units $\mathcal{I}$ and test unit n, compute $\widehat{\mathbb{E}}[\bar{y}_{n,pre''}]$ and $\widehat{\omega}^{(n,0)}$ using SI.
2. Accept the hypothesis only if $\frac{\sqrt{T_1}}{\widehat{\sigma}\|\widehat{\omega}^{(n,0)}\|_2} |\widehat{\mathbb{E}}[\bar{y}_{n,pre''}] - \bar{y}_{n,pre''}| > z_{0.95}$.

Theorem 5.6.3 (Informal; detailed version in Theorem C.V.1)**.** *Under Assumption 5.6.1, if $\|\widetilde{\omega}^{(n,0)}\|_2$ is sufficiently large and $\widehat{\omega}^{(n,0)} \to \widetilde{\omega}^{(n,0)}$ at a sufficiently fast rate, then as $|\mathcal{I}|, T_0, T_1 \to \infty$ the Asymptotic Hypothesis Test falsely accepts the hypothesis with probability at most* 5%.

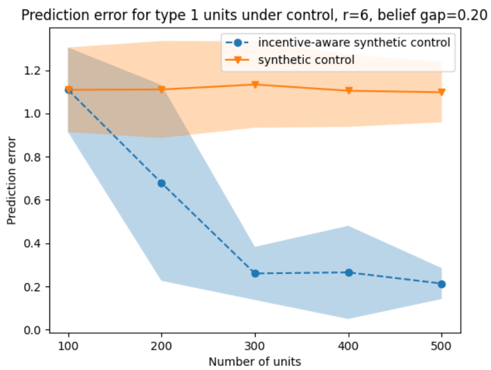
5.7 Numerical Simulations

In this section, we complement our theoretical results with a numerical comparison to standard SCMs which do not take incentives into consideration.

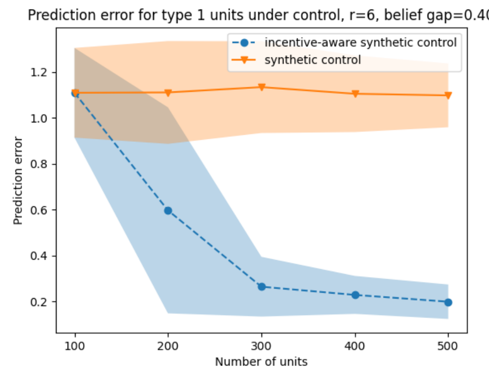
[10] Recall that we only assume knowledge of an upper-bound σ^2 on the variance of $\varepsilon_{i,t}$.

Experiment Setup. We consider the setting of Section 5.4: There are two interventions and two unit types. Type 1 units prefer the treatment and type 0 units prefer the control. Our goal is to incentivize type 1 units to take the control so we can obtain accurate counterfactual estimates of all units under control. See Section C.VI for details about our data-generating process.

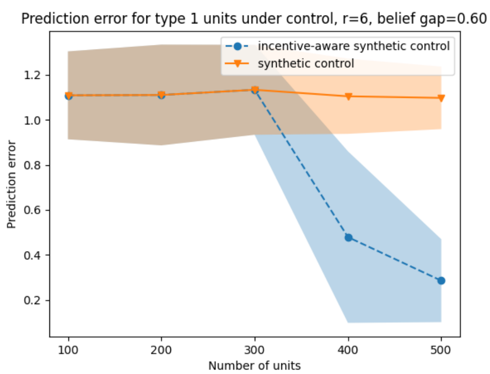
We apply the asymptotic hypothesis test of Section 5.6.2 using all type 0 units as the donor units and a single type 1 unit as the test unit. We accept the null hypothesis, as the test statistic $\frac{\sqrt{T_1}}{\widehat{\sigma}\|\widehat{\omega}^{(n,0)}\|_2}|\widehat{\mathbb{E}}[\bar{y}_{n,pre''}] - \bar{y}_{n,pre''}| \approx 2.24$, which is much larger than the value of $z_{0.95} \approx 1.96$ needed for acceptance. That is, the asymptotic test (correctly) returns that the Unit Overlap Assumption is not satisfied for a test unit of type 1 under control. We consider a simplified setting with a population of 1000 consisting of two unit types: type 0 units that initially prefer intervention 0 and type 1 units that initially prefer intervention 1. We let each type of unit have an equal proportion in the population, *i.e.*, $p_0 = p_1 = 0.5$. We first generate two orthogonal linear subspaces of $\mathbb{R}^2$, from which we sample the latent factors for units of each type. That is, units of type 0 have latent factors $\mathbf{v}_i = [\mathbf{v}_i^{(0)} \quad 0]$, where $\mathbf{v}_i^{(0)} \sim \mathcal{N}(0,1)$ and units of type 1 have latent factors $\mathbf{v}_i = [0 \quad \mathbf{v}_i^{(1)}]$ where $\mathbf{v}_i^{(1)} \sim \mathcal{N}(0,1)$. The time-intervention latent factors $\mathbf{u}_t^{(d)}$ are fixed for the post-treatment period at $[0.1 \quad 1.0]$ for intervention 0 and $[1.0 \quad 0.1]$ for intervention 1. We scale these latent factors by 50 to improve the visibility of the expected outcome. Then, the potential outcome is generated with added independent Gaussian noise $\varepsilon_{i,t}^{(d)} \sim \mathcal{N}(0, 0.1)$. Using Theorem C.III.4 (the formal version of Theorem 5.4.2), we can calculate a lower bound on the phase length L of Algorithm 5 such that the BIC condition (Equation (3.1)) is satisfied for units of type 1 who are recommended the control. We run three different sets of simulations and compare the estimation error $|\widehat{\mathbb{E}}\bar{y}_{n,post}^{(0)} - \mathbb{E}\bar{y}_{n,post}^{(0)}|$ for a new type 1 unit n under control when incentivizing exploration using Algorithm 5 (blue) to the estimation error without incentivizing exploration (orange). For both our method and the incentive-unaware ablation, we use the adaptive PCR method of Agarwal et al. (2023) to estimate counterfactuals. All experiments are repeated 50 times and we report both the average prediction error and the standard deviation for estimating the post-treatment outcome of type 1 units under control.



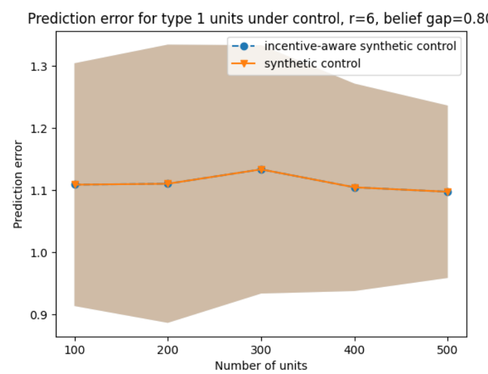
(a) Prior mean gap 0.2



(b) Prior mean gap 0.4



(c) Prior mean gap 0.6

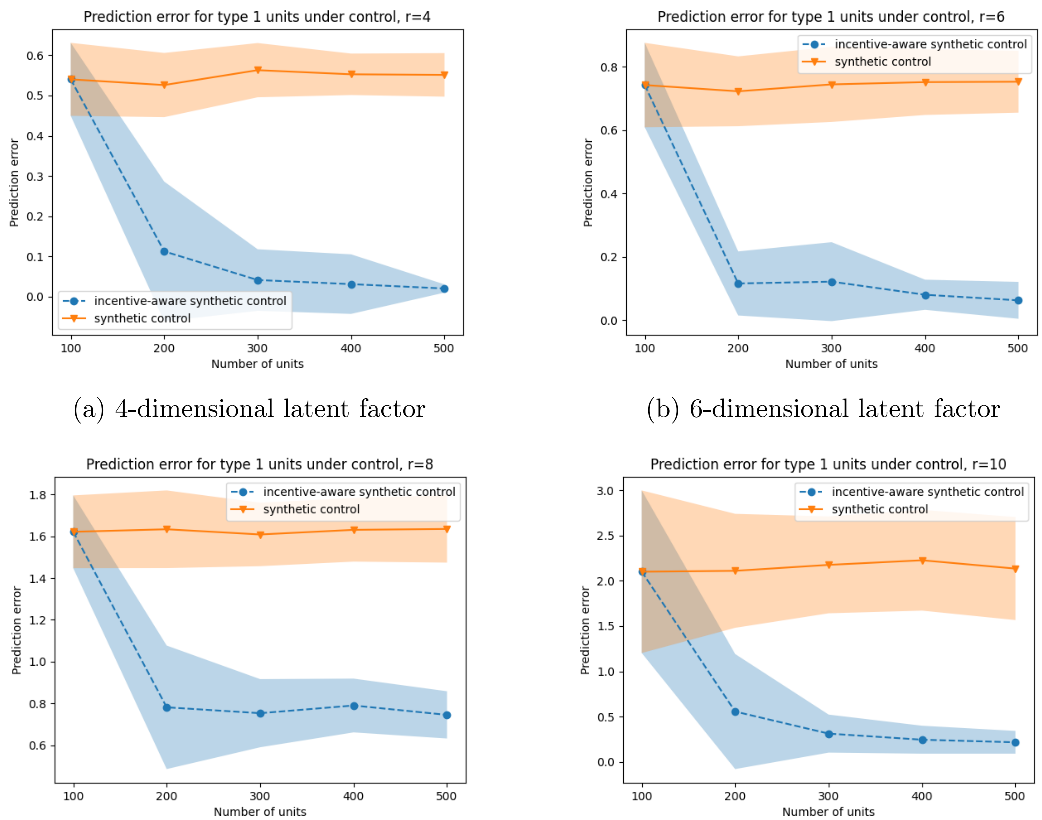


(d) Prior mean gap 0.8

Figure 5.2: Counterfactual estimation error for units of type 1 under control using Algorithm 5 (blue) and synthetic control without incentives (orange) with different gaps between the prior mean reward of control and treatment. Results are averaged over 50 runs, with the shaded regions representing one standard deviation.

Varying Strength of Belief. In Figure 5.2, we explore the effects of changing the units' "strength of beliefs". Specifically, we vary the gap between the prior mean outcome under control and treatment for type 1 units by setting $\mu_i^{(1)} - \mu_i^{(0)} \in \{0.2, 0.4, 0.6, 0.8\}$. As the gap decreases it becomes easier to persuade units to explore, and therefore our counterfactual estimates improve.

Varying Latent Factor Size. In Figure 5.3, we vary $r \in \{2, 4, 6, 8\}$ while keeping all other parameters of the problem instance constant. Increasing r increases the batch size



(a) 4-dimensional latent factor

(b) 6-dimensional latent factor

(c) 8-dimensional latent factor

(d) 10-dimensional latent factor

Figure 5.3: Counterfactual estimation error for units of type 1 under control using Algorithm 5 (blue) and synthetic control without incentives (orange) with increasing number of units and different lengths of the latent factors. Results are averaged over 50 runs, with the shaded regions representing one standard deviation.

L, so we see a modest decrease in performance as r increases.

5.8 Conclusion

We study the problem of non-compliance when estimating counterfactual outcomes using panel data. Our focus is on synthetic control methods, which canonically require an overlap assumption on the unit outcomes in order to provide valid finite sample guarantees. We shed light on this often overlooked assumption, and provide an algorithm for incentivizing units to explore different interventions using tools from information design and online learning. After running our algorithm, the Unit Overlap Assumption will be satisfied with high probability, which allows for the principal to obtain valid finite sample guarantees for all units when using off-the-shelf SCMs to estimate counterfactual outcomes under control. We also extend our algorithm to satisfy the Unit Overlap Assumption for all units and all interventions when there are more than two interventions, and we provide two hypothesis tests for determining if the Unit Overlap Assumption hold for a given test unit and set of donor units. Finally, we complement our theoretical findings with numerical simulations, and observe that our procedure for estimating counterfactual outcomes produces significantly more accurate estimates when compared to existing methods which do not take unit incentives into consideration.

There are several exciting directions for future research. Recent work (Sellke and Slivkins (2022); Sellke (2023)) provides lower bounds on the sample complexity of incentivizing exploration in various multi-armed bandit settings. It would be interesting to prove analogous bounds on the number of units required to incentivize sufficient exploration for the Unit Overlap Assumption to be satisfied for all units and all interventions in our setting. Another avenue for future research is to design difference-in-differences or clustering-based algorithms for incentivizing exploration in other causal inference settings. Finally, it would be interesting to relax other assumptions typically needed for synthetic control, such as the linear span inclusion assumption on the time-step latent factors.

Chapter 6

Incentivizing Exploration in Two-sided Matching Markets

In this chapter, we study the final extension of *incentivized exploration*: **how to incentivize exploration in the practical setting of two-sided matching markets where both sides are human decision-subjects**.

6.1 Introduction

We consider *incentivized exploration*: how to incentivize self-interested users to explore lesser-known options in a bandits setting. In the canonical incentivized exploration problem (Kremer et al., 2014c), a principal interacts with self-interested users and can provide recommendations but cannot force the agents to comply with these recommendations. The users face uncertainty about the available alternatives and can only rely on the principal's recommendation as the primary source of information to drive their decision. The principal would want the users to trade-off between *exploration* to obtain more information and *exploitation* to make optimal near-term decisions given the current information. On the other hand, the users prefer to purely *exploit*. The goal of the mechanism design question is to incentivize users to *explore* by leveraging the information collected by previous users. For a detailed overview of the problem, see Slivkins (2019c).

To the best of our knowledge, prior work in the incentivized exploration literature focuses on a one-sided setting, where the agents are human decision-subjects (*e.g.,* moviegoers, customers, *etc.*) and the arms are inanimate objects (*e.g.,* movies, restaurants, *etc.*). However, many realistic exploration problems often feature interactions between individuals with preferences over their potential matches. One such example is an online job market where job applicants seek to get matched with employers in a one-to-one format, *i.e.,* each job opening only accepts a single candidate. Each job applicant has their preference over which position they want to work in to utilize their skill set best. Similarly, employers want to match with candidates with well-documented track records who they can trust to perform well in the new job. This setting is a canonical example of the one-to-one matching problem studied by Gale and Shapley (1962).

Another prominent example of two-sided matching markets with human decision-subjects on both sides is the financial advising setting (Rose, 2023). In this setting, new clients visit an online platform to match with financial advisors from a financial institute every day. Suppose there are two types of clients: (small) independent businesses and (big) established businesses. Each client's goal is to match with a well-established financial advisor they can trust and provide them with more well-informed investment advice. On the other hand, the financial institute only has a fixed number of financial advisors who all prefer to match with more established clients to improve their professional portfolios. A standard non-incentive-aware platform's matching protocol is as follows: each client and financial advisor submits their top choice. Then, the platform would only match clients and financial advisors who prefer each other while the remaining clients and financial advisors stay unmatched. Only (big) established businesses can match with well-known financial advisors without any intervention from the online platform. In this case, the smaller businesses and new, less established financial advisors would not see any match. This lack of matches for a significant proportion of the matching platform's users is detrimental to both the users and the platform itself.

While preference matching is ubiquitous, it may lead to self-imposed bias where job applicants (clients) only seek out employers (financial advisors) they know beforehand, ignoring other options on the market. At the same time, employers (financial advisors) also need to explore more as they are more favorable to prominent job applicants (clients) than expanding their search for the most suitable candidates. Moreover, in a large market, it is improbable that an employer (financial advisor) can form an accurate preference ordering over job applicants (clients) without interacting with them first. Our goal is to *incentivize exploration* in a centralized matching market, where the platform provides recommendations to both sides of the users to explore alternative options. Such exploration is crucial to any learning algorithm that seeks to find the optimal matching in two-sided markets.

Overview of results. Our main contributions are as follows:

1. To the best of our knowledge, prior work in incentivized exploration only considers the agents' incentives. Instead, in this work, we consider the incentive-aware exploration problem in an online matching market from the perspectives of both agents and arms. In Section 6.3, we provide a counterexample to motivate the necessity of these dual incentive constraints in two-sided matching markets.

2. We provide an end-to-end BIC algorithm with two components: 'warm-start' and accelerated exploration. We developed a novel recommendation policy based on the inverse-gap weighting technique to accelerate exploration with near-optimal regret guarantees. In Section 6.4, we focus on the fundamental special case of incentivized exploration with two types of agents and arms. In Section 6.5, we extend our theoretical results to the setting where multiple types of agents and arms visit the platform at each time-step.

3. In Section 6.6, we provide a set of numerical simulations on synthetic data and show that our end-to-end algorithm is both a) incentive-compatible and b) efficient in terms of regret minimization.

6.2 Preliminary

Notation. We write $[N] = \{1, 2, \cdots, K\}$ for $N \in \mathbb{N}^+$. We use subscripts i, j to denote different agents or arms and superscript $t \in [T]$ to denote different time-steps.

We focus on an online two-sided matching market with a time horizon of T. At time-step $t \in [T]$, the principal interacts with a fresh batch of N agents and N arms and forms N one-to-one matches between agents and arms. If an agent successfully forms a match with some arms, both the agent and arm in this match report their shared utility to the platform and leave. We first focus on the binary type setting, where there are two types of agents and arms, *i.e.,* $i \in [2]$ and $j \in [2]$. We study the more general setting with k types of agents and arms in Section 6.5. At a high level, the type of agents (or arms) determines both their preferences over which arms (or agents) they want to match with initially and the realized reward of their match.

Reward formulation and Bayesian priors. To impose structure on how the shared utilities of matches are related for different agents and arms across different time-steps, we assume that the utilities are generated via the following bilinear model. Concretely, at time-step t, each agent of type i has their user profile $x_i^{(t)} \in \mathbb{R}^d$. Similarly, each arm of type j has profile vector $a_j^{(t)} \in \mathbb{R}^d$. Let $\Sigma \in \mathbb{R}^{d\times d}$ be a latent matrix with rank $r \ll d$. Then, the realized reward of a match between (type i agent, type j arm) is:

$$r_{i,j}^{(t)} = r^{(t)}(x_i^{(t)}, a_j^{(t)}) \coloneqq (x_i^{(t)})^\top \Sigma a_j^{(t)} + \eta_{i,j}^{(t)} \tag{6.1}$$

where $\eta_{i,j}^{(t)} \sim \mathrm{subG}(\sigma)$ is independent zero-mean sub-Gaussian noise with variance at most σ. We write $\mu_{i.j} = x_i^\top \Sigma a_j$ to denote the expected reward of a match between agents of type i and arms of type j, and $\mu_{i,j}^{(0)}$ to denote the prior-mean reward. For simplicity, we assume that $\forall i, j \in [N] : \mu_{i,j} \in [-1, 1]$. Henceforth, we write x_i and a_j to refer to agents of type i and arms of type j.

Preferences. We first focus on the stylized setting with two types of agents and arms. Let $i, j \in [2]$ denote the type of agents and arms, respectively. We are interested in two sets of preferences: agent-to-arm and arm-to-agent. In our motivating example, job applicants want to be matched with compatible employers and employers prefer to be matched with applicants who can perform well. Wlog, we assume that the initial preference ordering is $\mu_{1,1}^{(0)} \geq \mu_{1,2}^{(0)} \geq \mu_{2,2}^{(0)}$ and $\mu_{1,1}^{(0)} \geq \mu_{2,1}^{(0)} \geq \mu_{2,2}^{(0)}$. That is, all agents prefer to match with type 1 arms over type 2 arms, and all arms prefer to match with type 1 agents over type 2 agents.

Incentive-compatibility. Absent incentives and coordination from the platform, the agents and arms would match each other using their initial preferences. However, the platform wants to incentivize both the agents and the arms to explore different options to find the optimal matching and maximize the cumulative reward. In particular, at each time step t, the platform can broadcast a signal $\mathbf{s}^{(t)}$ as a recommendation to all agents and arms. By *direct revelation principle* (Myerson, 2018), this signal is equivalent to directly telling the agents which arms to match with, and vice versa.

Definition 6.2.1 (Two-sided Bayesian Incentive-Compatible Condition)**.** *For time-step $t \in [T]$, the platform's recommendation is $\varepsilon-$*two-sided Bayesian Incentive-Compatible *(ε-BIC) for some $\varepsilon > 0$ if for all agents and arms in that time-step, it satisfies:*

$$\mathbb{E}[r_{i,j}^{(t)} - r_{i,\ell}^{(t)} | \text{rec} = (x_i^{(t)}, a_j^{(t)})] \geq \varepsilon \tag{6.2}$$
$$\mathbb{E}[r_{i,j}^{(t)} - r_{\ell,j}^{(t)} | \text{rec} = (x_i^{(t)}, a_j^{(t)})] \geq \varepsilon \tag{6.3}$$

We assume the following behaviour model of the agents and arms. To ensure that the platform considers the preferences of both sides, we posit that a pair of (agent, arm) only successfully forms a match when both the agent and the arm choose each other.

Assumption 6.2.2 (Behavioral Assumption)**.** *Agents and arms follow recommendations for any ε_0-BIC policy for some fixed $\varepsilon_0 > 0$. If one side rejects the recommendation, then both sides of the recommended (agent, arm) pair do not have a match for that time-step, and the platform receives a reward of* 0 *for that recommended pair. Both the agents and the arms are assumed to be* myopic, i.e., *they will choose the posterior best arms (agents) at the current time-step to match with.*

Reduction to combinatorial semi-bandits. Our first technical insight is to reduce the online two-sided matching problem to a combinatorial semi-bandits problem. Consider the following mapping: at each time-step $t \in [T]$, the set of all feasible matches between agents and arms constitutes the action space $\mathcal{A} \subset \mathbb{R}^{N \times N}$. An *atom* $(x_i^{(t)}, a_j^{(t)})$ is a match between $x_i^{(t)}$ and $a_j^{(t)}$, and there are N^2 total atoms. An *action* $A^{(t)} \in \mathcal{A}$ at time-step t is the combination of matches at that round, where $\left\|A^{(t)}\right\|_1 \leq N$. The interaction protocol is as follows. At each time-step t, a learner arrives at the platform, receives a recommendation for an action $A \in \mathcal{A}$, and chooses an action $A^{(t)} \in \mathcal{A}$. The platform and the learner both observe the reward of each atom in this arm (and nothing else). The algorithm's reward in this time-step is the total reward of these atoms.

Under this reduction, a few technical challenges differentiate our result from that of combinatorial semi-bandits. Particularly, it is unclear how to collect the 'warm-start' samples, which are input to any efficient incentivized exploration algorithm. For a detailed explanation, see Section 6.3.

Regret The objective of the platform is to find the optimal matching between the agents and arms that maximizes the cumulative rewards over the time horizon T using a recommendation policy that satisfies the incentive-compatibility constraint (Definition 6.2.1). We measure the performance of the platform's recommendation policy through the standard notion of regret. In essence, the regret incurred at each time-step is the difference between the reward of choosing the optimal matching and the reward of the chosen matching at that time step.

Definition 6.2.3 (Regret)**.** *Given a time horizon T and a recommendation policy $\{\mathbf{s}^{(t)} : t \in [T]\}$, the regret of this recommendation policy is*

$$\mathrm{Regret}(\{\mathbf{s}^{(t)}_{t\in[T]}\}, T) = \sum_{t=1}^{T} r^{(t)}_{A^*} - r^{(t)}_{\mathbf{s}^{(t)}} \tag{6.4}$$

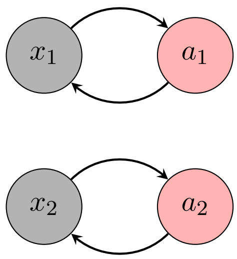
where A^* is the optimal matching.

Moreover, we also look at the Bayesian regret $\mathbb{E}[\mathrm{Regret}(\{\mathbf{s}^{(t)}\}_{t\in[T]}, T)]$, where the expectation is taken over the randomness in the priors of agents and arms.

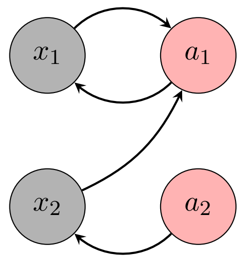
6.3 Necessity of two-sided incentive in matching market

In this section, we provide an example to show the need for two-sided Bayesian incentive-compatible constraints in a two-sided matching market. Consider the stylized setting with two types of agents and arms and a time horizon of T. In the first T_0 time-steps, we assume that the platform runs a black-box recommendation algorithm such that, at the end of T_0 time-steps, the agents always follow the platform's recommendation and take the recommended arm. We show that there exists a problem instance where in the remaining $T - T_0$ time-steps, the platform would incur regret $\Omega(T - T_0)$.

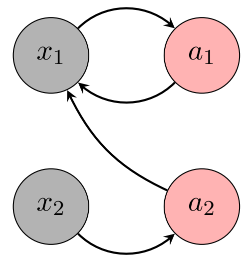
We examine when a stable matching can happen without external incentives from the platform. The number of possible matchings between 2 types of agents and 2 types



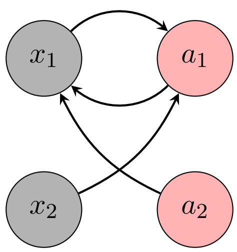
(a) Agent x_1 prefers arm a_1 (and vice versa), and agent x_2 prefers arm a_2 (and vice versa)



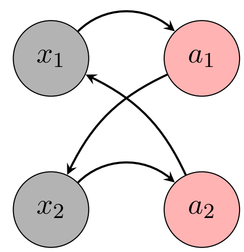
(b) Agent x_1 prefers arm a_1 (and vice versa). Agent x_2 prefers arm a_1. Arm a_2 prefers agent x_2.



(c) Agent x_1 prefers arm a_1 (and vice versa). Agent x_2 prefers arm a_2. Arm a_2 prefers agent x_1.



(d) Agent x_1 prefers arm a_1 (and vice versa). Agent x_2 prefers arm a_1. Arm a_2 prefers agent x_1.



(e) Agent x_1 prefers arm a_1, which prefers agent x_2. Agent x_2 prefers arm a_2, which prefers agent x_1.

Figure 6.1: Possible unique matchings between 2 agents and 2 arms. Blue nodes represent agents, and red nodes represent arms. The directed arrow indicates that the start node prefers the end node. Among all possible matchings, only the first case with two disjoint cyclic subgraphs does not need the platform's interventions to have successful matches for all agents and arms. In any other case, we can always find a blocking pair of nodes that prevent all agents and arms from finding a match.

of arms is $2^4 = 16$ (each agent has 2 choices for which arm they prefer, and vice versa). Due to symmetry among the agents and the arms (*e.g.,* a matching $\{(x_1, a_1), (x_2, a_2)\}$ is equivalent to the matching $\{(x_1, a_2), (x_2, a_2)\}$ by renaming the variables a_1 to a_2), there are 5 possible unique matchings between agents x_1, x_2 to arms a_1, a_2.

Among these unique matchings, only one is stable according to the initial preferences: the matching in Figure 6.1(a) where agents of type 1 prefer arms of type 1 (and vice versa), and agents of type 2 prefer arms of type 2 (and vice versa). If the optimal solution falls into this case (or its isomorphic forms), then the platform does not need to run an incentivized exploration algorithm to achieve optimal matching. However, for the remaining 4 possibilities, there always exists a possible realization of the rewards such that the initial preferences of either the agents or the arms will block an optimal matching (due to incompatible preference from either side) and any non-incentive-aware learning algorithm would incur linear regret.

6.4 Incentivized exploration for two types of agents and arms

In this section, we focus on the fundamental special case of incentivized exploration with two types of agents and arms to show the salient points of our analysis. In essence, the platform first incentivizes all types of agents and arms to match each other and collect samples from these matches. These 'warm-start' samples are then used to compute an accurate estimate of the true reward from each match using an efficient regression oracle. Then, the platform leverages these 'warm-start' samples to accelerate exploration in the second phase and quickly converge to optimal matching.

6.4.1 Initial exploration with Hidden Exploration

We present our first contribution, a BIC algorithm to collect the 'warm-start' samples, where the objective is to sample each atom, *i.e.,* match between an agent and an arm, at least once and completes in T_0 time-steps for some T_0 determined by the prior.

To ensure that the initial exploration can happen, we assume that there exists a pair of (agent, arm) that initially prefers each other, *i.e.,* even without the principal's recommendation, we can always collect samples from this match. We posit that this assumption is necessary and is a weaker assumption than the one made in incentivized exploration for multi-armed bandits, *e.g.,* all agents prefer an arm initially so the principal can always collect samples of this arm at the beginning of the algorithm[1]. Here, we only

[1] Intuitively, if all atoms are not initially 'explorable', then the initial preference graph is Figure 6.1(e) (and its isomorphic forms). Then, no samples can be collected initially and the agents/arms cannot update their posteriors. Hence, no incentivized exploration algorithm can start giving out meaningful recommendations.

assume that one atom is initially 'explorable' by the learner, and all other atoms are not. Without loss of generality, this 'explorable' atom is the match between x_1 and a_1.

In the following algorithm, we show that in the 'worst case' with one 'explorable' atom initially, we can incentivize both the agents and the arms to explore different matches. Intuitively, given enough samples of the 'explorable' atom, we can split the remaining time-steps into phases such that in each phase, a new atom, *i.e.,* a match between an agent and an arm that was previously not 'explorable', can be chosen by the learner upon receiving the principal's recommendation. The incentivized exploration technique within each phase builds on the approach from Mansour et al. (2015b), which is defined for multi-armed bandits. However, the reward priors are highly correlated in two-sided matching markets, and the set of 'explorable' atoms can initially be of size 1. Furthermore, the intricate incentive interplay between agents and arms requires a more careful notion of which action to explore. Our technical contribution here is to provide a sequence of actions and prove that it is possible to incentivize both the agents and the arms to explore, given a standard non-degeneracy condition on the posterior distribution of the reward for each atom.

We make the following non-degeneracy assumption: any action A_{cand} can be the posterior best action with a margin $\tau_{\mathcal{P}}$ and probability at least $\rho_{\mathcal{P}}$ after seeing at least $n_{\mathcal{P}}$ samples of the previous actions.

Assumption 6.4.1 (Fighting chance assumption)**.** *There exists number $n_{\mathcal{P}} \in \mathbb{N}$ and $\tau_{\mathcal{P}}, \rho_{\mathcal{P}} \in (0,1)$ determined by the prior $\mathcal{P}$ such that: for a sequence of actions $A^1_{\text{cand}}, \cdots, A^{N^2}_{\text{cand}}$ defined by* $\text{NextCandidate}(\mathcal{A}, \mathcal{S}, \mathcal{P})$*, Let $\mathcal{S}$ be the dataset containing exactly $k \in \mathbb{N}$ samples of each atom, then*

$$\Pr[\Delta_i^k \geq \tau_{\mathcal{P}}] \geq \rho_{\mathcal{P}} \quad \forall i \in \mathcal{A} \text{ and } k \geq n_{\mathcal{P}}, \tag{6.5}$$

where $\Delta_i^k = \min_{A \neq A_{\text{cand}}} \mathbb{E}[\mu_{A_{\text{cand}}} - \mu_A | \mathcal{S}]$

This non-degeneracy assumption is common in the literature on incentivized exploration to ensure compliance from the first sets of agents and arms at time-step $t = 0$. At a high level, if the non-degeneracy assumption does not hold, then the algorithm cannot incentivize any agent or arm to take a new (unexplored) matching even if the previously taken actions in prior time-steps yield low rewards. In our motivating example, the platform can only observe the matching between big businesses (agents of type 1) and established financial advisors (arms of type 1) in the first few time-steps due to prior preferences. In the worst case, when this matching yields a realized reward of -1 at these time-steps, the agents and arms in all future time-steps will still not follow the platform's recommendation for any other matching without Assumption 6.4.1.

We state our initial sampling algorithm in Line 1 and its theoretical guarantees in Theorem 6.4.2. At a high level, Theorem 6.4.2 ensures that all agents and arms will comply with the recommendation policy by hiding one *explore* recommendation among multiple *exploit* recommendations. When the probability of being given a *exploration* recommendation is low enough, the learner can infer that the expected gain from *exploiting* outweighs the expected loss from *exploring*. Hence, the platform will give recommendations in batches, where only one learner in each batch receives an exploration recommendation. As long as the batch size is big enough, *i.e.,* the exploration probability is small enough, all learners will comply with the given recommendations.

Algorithm 7: Initial sampling: Hidden Exploration

Input: Batch size $L \in \mathbb{N}$, target number of samples $k \in \mathbb{N}$, gap $C \in (0, 1)$.

1: Initialize dataset $\mathcal{S} = \emptyset$;
2: The first k learners choose $A = \{(x_1, a_1)\}$ without recommendations.
3: Let $\widehat{r}^{k}_{1,1}$ be the sample average of these rewards. Add these k samples to $\mathcal{S}$;
4: **for** each phase $\psi = 1$ to N^2 **do**
5: $\quad A^{(\psi)}_{\text{cand}} = \text{NextCandidate}(\mathcal{A}, \mathcal{S}, \mathcal{P})$;
6: $\quad$ **if** $\widehat{r}^{k}_{1,1} \leq \mu^{(0)}_{A^{\psi}_{\text{cand}}} - C$ **then**
1 7: $\quad\quad$ 'Exploit' action $A^* = A^{\psi}_{\text{cand}}$.
8: $\quad$ **else**
9: $\quad\quad$ 'Exploit' action $A^* = \{(x_1, a_1)\}$.
10: $\quad$ **end if**
11: $\quad$ From a set P of the next $L \cdot k$ learners:
12: $\quad$ Pick a set Q of k learners uniformly at random;
13: $\quad$ Every learner $p \in P - Q$ is recommended the 'exploit' action A^*;
14: $\quad$ Every learner $p \in Q$ is recommended action A_{cand}.
15: $\quad$ Add the reward from all $p \in Q$ to $\mathcal{S}$.
16: **end for**

Given Assumption 6.4.1, we have the following guarantees on the 'warm-start' algorithm:

Theorem 6.4.2. *Assuming Assumption 6.4.1 holds with constants $n_{\mathcal{P}}, \tau_{\mathcal{P}}, \rho_{\mathcal{P}}$. Then, Line 1 is two-sided ε-BIC as long as the batch size L is at least*

$$L \geq 1 + \max\left\{\frac{2+2\varepsilon}{\tau_{\mathcal{P}}\cdot\rho_{\mathcal{P}} - 2\varepsilon}, \frac{2\varepsilon}{\mu_{1,2}^{(0)} + \mu_{2,1}^{(0)} - 2\mu_{2,2}^{(0)} + \tau_{\mathcal{P}}\cdot\rho_{\mathcal{P}} - 2\varepsilon}\right\} \tag{6.6}$$

*where $\tau_{\mathcal{P}}, \rho_{\mathcal{P}}$ are defined in Assumption 6.4.1. This algorithm completes in $T_0 = N^2 \cdot n_{\mathcal{P}} \cdot \frac{1+N^2}{\tau_{\mathcal{P}}\cdot\rho_{\mathcal{P}}}$ time-steps. At the end of Line 1, all atoms (*i.e., *matchings) are sampled at least $n_{\mathcal{P}}$ times.*

6.4.2 Accelerated Exploration with Inverse Gap Weighting

Given the data collected by Line 1, the platform wants to accelerate exploration and converge to optimal matching. The platform has to balance *exploitation*, *i.e.,* recommending the empirical best match to minimize regret, and *exploration*, *i.e.,* ensuring that the two-sided BIC condition holds. The theoretical underpinning of our recommendation policy at this stage is *inverse gap weighting*, *i.e.,* recommending a match with probability inversely proportional to the reward gap between that match and the empirical best match. Formally, we let $b^{(t)} = \text{argmax}_{A\in\mathcal{A}}\, \hat{r}_A^{(t)}$ denote the empirical best action at time-step t. Then, the probability of an action A being recommended at time-step t is:

$$p_A^{(t)} = \begin{cases} \frac{1}{N^2+\gamma(\hat{r}_{b^{(t)}}^{(t)} - \hat{r}_A^{(t)})} & \text{if } A \neq b^{(t)} \\ 1 - \sum_{A\neq b^{(t)}} p_A^{(t)} & \text{otherwise} \end{cases} \tag{6.7}$$

where the hyperparameter $\gamma > 0$ shows the tradeoff between exploration and exploitation. Intuitively, a smaller γ leads to more exploration, while a larger γ induces more exploitation. Formally, the algorithm to accelerate exploration is given as follows.

To ensure that γ is adaptive to the samples collected in previous time-steps, we set $\gamma = C_0 \cdot N\sqrt{1/\phi^{(t)}}$, where $\phi^{(t)}$ is the mean squared error of the empirical estimate at time-step t. Similar to Foster and Rakhlin (2020), we assume there exists an efficient regression-oracle that accurately computes $\phi^{(t)}$ at time-step t. With this recommendation policy, we state the theoretical guarantee for accelerated exploration:

Theorem 6.4.3 (Informal)**.** *Given sufficiently many 'warm-start' samples of all atoms,*

Algorithm 8: Accelerated exploration: Inverse-gap weighting

Input: Dataset $\mathcal{S}$, efficient regression-oracle

1: **for** time-step t **do**

2: Initialize $\gamma = C_0 \cdot N\sqrt{1/\phi^{(t)}}$;

3: Compute the estimated reward for each action $A \in \mathcal{A}$ given the samples

1 collected in $\mathcal{S}$ and previous time-steps: $\hat{r}_A^{(t)}$.

4: Compute the 'best' empirical action $b^{(t)} = \operatorname{argmax}_{A\in\mathcal{A}} \hat{r}_A^{(t)}$.

5: Recommend action A to learner according to Equation (6.7), add the reward of the chosen action $r_A^{(t)}$ to the dataset $\mathcal{S}$.

6: **end for**

the inverse gap weighting recommendation policy is two-sided ε-BIC. The total regret during this stage is $O(N\sqrt{dT\log(T)})$, which asymptotically matches the optimal regret of combinatorial semi-bandits.

The detailed theorem statement and analysis of Line 1 are given in Section D.I.2.

6.5 Incentivized exploration for many types of agents and arms

Our focus thus far has been on the stylized setting where there are two types of agents and arms. In this section, we provide a generalization of the previous theoretical results by considering K types of agents and arms. For simplicity, we assume that the initial preference ordering of type i agents is $\mu_{i,1}^0 \geq \mu_{i,2}^0 \geq \cdots \geq \mu_{i,K}^0$. That is, all agents of type i prefer to match with arms of type 1 over arms of type 2, and so on. Similarly, we assume the initial preference ordering of type j arms is $\mu_{1,j}^0 \geq \mu_{2,j}^0 \geq \cdots \mu_{K,j}^0$.

Our combined recommendation policy is still a combination of two stages: collecting 'warm-start' samples, where the goal is to sample each match between an agent of type i and arm of type j at least once and 'accelerated exploration', where the platform quickly converge to the optimal matching. The main difference compared to Section 6.4 is that in the 'warm-start' stage, the platform needs to choose the 'exploit' matching more carefully instead of simply using the prior-best match $\{(x_1, a_1)\}$. We assume the same non-degeneracy assumption to ensure that our algorithm can incentivize all agents and arms to comply with the recommendation for any possible matching. Details of algorithm 1 and its analysis for the 'warm-start' stage with k-types of agents and arms is deferred to Section D.II.

After running Line 1, the platform has access to a dataset containing $n_{\mathcal{P}}$ samples of each matching. Since the recommendation policy in Line 1 and its theoretical guarantee do not change when we have more actions at each time-step, the platform can still employ the same recommendation policy in Equation (6.7) to ensure that the two-sided BIC condition holds for all future learners.

6.6 Numerical Simulations

In this section, we complement our theoretical results with an experiment (Figure 6.2) to compare the incentive compatibility and regret minimization of our combined algorithm with a baseline learning algorithm that does not consider two-sided preferences from both the agents and the arms. For additional experiment details and results, see Section D.III.

Experimental Details We consider a stylized setting with two types of agents and two types of arms as described in Section 6.2. Initially, all agents prefer to match with arms of type 1, and all arms prefer to match with agents of type 1. We specify the prior-mean rewards of these matches as $\mu^0_{1,1} = 0.8, \mu^0_{1,2} = 0.4, \mu^0_{2,1} = 0.4$, and $\mu^0_{2,2} = 0.2$. Our goal is to incentivize all agents and arms to explore all possible alternative matches and minimize regret.

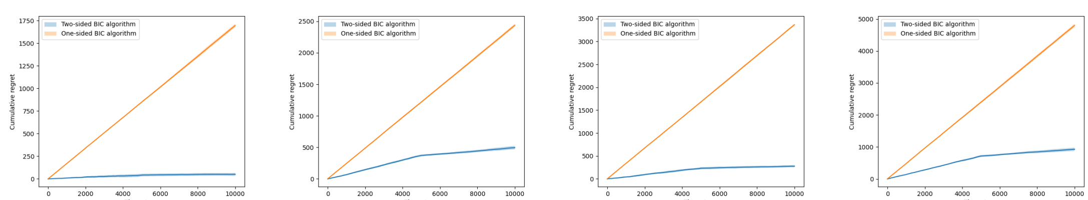
We consider an online setting with a time horizon of $T = 10000$. At each time-step $t \in [T]$, a batch of N units arrives at the platform. In Figure 6.2, we look at the stylized setting of $N = 2$, where there is one agent of each type and one arm of each type at each time-step. The user profile for each agent of type 1 (resp. type 2) is $x^{(t)}_1 = [v^{(t)} \quad 0]$ (resp. $x^{(t)}_2 = [0 \quad v^{(t)}]$) where $v^{(t)} \sim \mathsf{Unif}[0, 1]$. Similarly, the user profile for each arm of type 1 (resp. type 2) is $a^{(t)}_1 = [u^{(t)} \quad 0]$ (resp. $a^{(t)}_2 = [0 \quad u^{(t)}]$) where $u^{(t)} \sim \mathsf{Unif}[0, 1]$. The latent matrix Σ is randomly generated to ensure that, in expectation, the optimal match is $\{(x_1, a_2), (x_2, a_1)\}$, *i.e.,* it is optimal for agents of type 1 to match with arms of type 2, and agents of type 2 to match with arms of type 1. Finally, the realized reward is generated by adding independent Gaussian noise $\eta^{(t)}_{i,j} \sim \mathcal{N}(0, 0.001)$ to each inner product of the user profiles and the latent matrix Σ.

Using Theorem 6.4.2, we calculate a lower bound on the phase length L of Line 1 such that the ε-BIC condition (Definition 6.2.1) is satisfied for all agents and arms. Then, we calculate the number of samples needed to ensure that the efficient oracle in Section 6.4.2 is well-defined. We calculate the regret incurred by the combined algorithm by summing over the gap between the realized reward of the chosen action and the optimal matching $\{(x_1, a_2), (x_2, a_1)\}$ at each time-step.

We test the validity of our theoretical algorithm when using different choices of hyperparameters d and N.

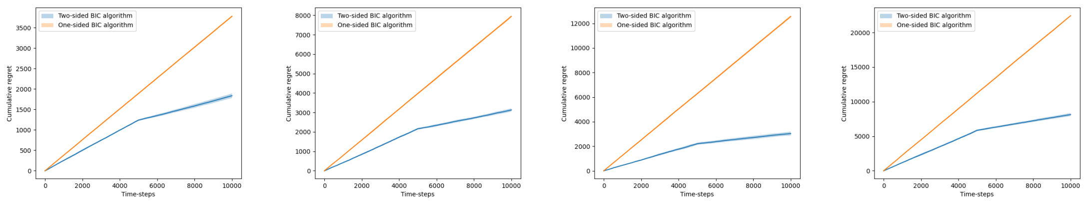
- Varying dimension of user profile vector: In Figure 6.2, we vary the dimension of the user profile vectors, *i.e.,* $d \in \{8, 16, 32, 64\}$, while keeping all other parameters of the problem instance constant. Each experiment is repeated 10 times, and we report the regret and the standard deviation at each time step.

- Varying number of agents and arms at each time-step: In Figure 6.3, we vary the number of agents and arms visiting the platform at each time-step, *i.e.,* $N \in \{4, 8, 16, 32\}$, while keeping all other parameters of the problem instance constant. Each experiment is repeated 10 times, and we report the regret and the standard deviation at each time step.



(a) $d = 8$ (b) $d = 16$ (c) $d = 32$ (d) $d = 64$

Figure 6.2: Cumulative regret using our proposed two-sided BIC algorithm (Line 1 and 1) in blue and the standard one-sided BIC algorithm in orange. We provide the results with increasing dimension d of the user profile vectors: $d \in \{8, 16, 32, 64\}$. Results are averaged over 10 runs, with the shaded region representing one standard deviation. In all experiments, we observe the cumulative regret of our proposed algorithm is significantly lower than that of a one-sided BIC algorithm that does not consider incentives from the arms. As the dimension of the user profile vectors increases, our algorithm incurs slightly more regret yet still outperforms the baseline naive method.



(a) $N = 4$ (b) $N = 8$ (c) $N = 16$ (d) $N = 32$

Figure 6.3: Cumulative regret using our proposed two-sided BIC algorithm (Line 1 and Line 1) in blue and the standard one-sided BIC algorithm in orange. We provide the results with an increasing number of agents and arms N at each time-step: $N \in \{4, 8, 16, 32\}$. Results are averaged over 10 runs, with the shaded region representing one standard deviation. In all experiments, we observe the cumulative regret of our proposed algorithm is significantly lower than that of a one-sided BIC algorithm that does not consider incentives from the arms. As the number of agents and arms at each time-step increases, our algorithm incurs more regret.

Results Our numerical results are consistent with that of prior work in incentivized exploration. In the first stage of collecting 'warm-start' samples (Line 1), we observe linear regret due to the construction of the recommendation policy. In the worst case, Line 1 only collects a new sample of the optimal matching once every batch, resulting in linear regret in this stage. Note that linear regret is also the state-of-the-art regret for the initial sample collection (Mansour et al., 2015b). When the second stage, 'accelerated exploration,' begins, we run the inverse gap weighting algorithm (Line 1), and the regret growth immediately decreases as the platform can explore more efficiently.

In a real-life two-sided matching market, the platform can collect the initial samples by buying them or using existing historical data, thus incurring no regret for the first stage. Then, the platform only has to run Line 1 to incentivize agents and arms in the remaining time-steps and observe sub-linear regret during its running time.

6.7 Conclusion and Future Work

In this work, we present the first results for incentivized exploration in two-sided matching markets, where the agents and arms are individuals with preferences over their matches. Prior work in incentivized exploration focused on settings where only the agents are human decision-subjects (movie-goers, customers, *etc.*) and the arms are inanimate objects (movies, restaurants, *etc.*) without preferences. We characterize the incentive-compatibility constraints in two-sided matching markets and provide a reduction to combinatorial semi-bandits. With this reduction, we present a BIC algorithm that collects 'warm-start' samples and then accelerates exploration to minimize regret. Our end-to-end BIC algorithm asymptotically matches the regret guarantee of the combinatorial semi-bandits without incentive constraints. Finally, we provide numerical simulations to support our theoretical findings.

Chapter 7

Strategic Instrumental Variable Regression

7.1 Introduction

Machine learning (ML) predictions increasingly inform high-stakes decisions for people in areas such as college admissions (Pangburn, 2019; Somvichian-Clausen, 2021), credit scoring (Petrasic et al., 2017; Rice and Swesnik, 2013), employment (Sánchez-Monedero et al., 2020), and beyond. One of the major criticisms against the use of ML in socially consequential domains is the failure of these technologies to identify *causal* relationships among relevant attributes and the outcome of interest (Kusner et al., 2017). The single-minded focus of ML on predictive accuracy has given rise to brittle predictive models that learn to rely on spurious correlations—and at times, harmful stereotypes—to achieve seemingly accurate predictions on held-out test data (Sweeney, 2013; Kusner and Loftus, 2020). The resulting models frequently underperform in deployment, and their predictions can negatively impact decision subjects. As an example of the long-term negative consequences of ML-based decision-making systems, they often prompt individuals to modify their observable attributes *strategically* to receive more favorable predictions—and subsequently, decisions (Hardt et al., 2016). These strategic responses are among the primary causes of distribution shifts (and subsequently, the unsatisfactory performance) of ML in high-stakes decision-making domains. Moreover, recent work has established the potential of these tools to amplify existing social disparities by *incentivizing different effort investments* across distinct groups of subjects (Liu et al., 2020; Heidari et al., 2019; Mouzannar et al., 2019).

The above challenges have led to renewed calls on the ML community to strengthen their understanding of the connections between ML and causality (Pearl, 2019; Schölkopf, 2019). Knowledge of causal relationships among predictive attributes and outcomes of interest promotes several desirable aims: First, ML practitioners can use this knowledge to debug their models and ensure robustness even if the underlying population shifts over time. Second, policymakers can utilize the causal understanding of a domain in their policy choices and examine a decision-making system's compliance with policy goals and societal values (e.g., they can audit the system for unfairness against particular populations (Loftus et al., 2018)). Finally, predictions rooted in causal associations block undesirable pathways of gaming and manipulation and, instead, encourage decision subjects to make meaningful interventions that improve their actual outcomes (as opposed to their assessments alone).

Our work responds to the above calls by offering a new approach to recover causal relationships between observable features and the outcome of interest in the presence of strategic responses—without substantially hampering predictive accuracy. We consider settings where a decision-maker deploys a sequence of models to predict the outcome for a sequence of strategic decision subjects. Often in high-stakes decision-making settings such as the ones mentioned earlier, there are unobserved confounding variables that influence subjects' attributes and outcomes simultaneously. Our key observation is that we can correct for the effect of such confounders by viewing the sequence of **assessment rules as valid *instruments*** which affect subjects' observable features but do not *directly* influence their outcomes. Our main contribution is a general framework that recovers the causal relationships between observed attributes and the outcome of interest by treating assessment rules as instruments.

7.1.1 Our Setting

Next, we describe our theoretical setup in further detail, then proceed to an overview of our findings. For concreteness, we utilize a stylized university admissions scenario as our running example for the remainder of this section. However, the reader should note that our model is applicable to other real-world applications in which confounders taint the causal interpretation of predictive models. For example, in credit lending, lack of access to affordable credit affects not only the applicant's debt, but also their likelihood of default (Collard and Kempson, 2005). In university admissions (which will be our running example), research has shown that the socioeconomic background of a student can impact both their SAT scores and success in college (Sackett et al., 2009).

With the running example in mind, consider a stylized setting in which a university decides whether to admit or reject applicants on a rolling basis[1] based (in part) on how well they are predicted to perform if admitted to the university (See Figure 7.1). We model such interactions as a game between a *principal* (here, the university) and a population of *agents* (here, university applicants) who arrive sequentially over T rounds, indexed by $t = 1, 2, \cdots, T$. In each round t, the principal deploys an assessment rule $\boldsymbol{\theta}_t \in \mathbb{R}^m$, which is used to assign agent t a predicted outcome $\widehat{y}_t \in \mathbb{R}$. In our running example, $\widehat{y}$ could correspond to the applicant's predicted college GPA if admitted. The predicted outcome is calculated based on certain observable/measured attributes of the agent, denoted by $\mathbf{x}_t \in \mathbb{R}^m$. For example, in the case of a university applicant, these attributes may include the applicants' standardized test scores, high school math GPA, science GPA, humanities GPA, and their extracurricular activities. For simplicity, we assume all assessment rules are *linear*, that is, $\hat{y}_t = \mathbf{x}_t^\top \boldsymbol{\theta}_t + \hat{o}_t$ for all t. (Where $\hat{o}_t$ is the current estimate of the expected offset term $\mathbb{E}[o_t]$.) This linear setup corresponds to an instance of the *partially linear regression model* (originally due to Robinson (1988)), a commonly studied setting in both the causal inference and strategic learning literature (e.g., Shavit et al. (2020); Kleinberg and Raghavan (2020); Bechavod et al. (2021)).

Measured vs. latent variables. We assume that the agent best-responds to the assessment rule $\boldsymbol{\theta}_t$ by strategically modifying their observable attributes $\mathbf{x}_t$ to receive a more favorable predicted outcome. Often agents cannot modify the value of their measured attributes (e.g., SAT score) directly, but only through investing effort in certain activities that are difficult to measure. For example, a student might take standardized test preparation courses to improve their SAT scores, or they may spend time studying the respective subjects to improve their math and humanities GPA.

Latent variables: effort investments. We formalize the above hidden investments with a vector $\mathbf{a}_t \in \mathbb{R}^d$, capturing the unobservable efforts agent t invests in d *activities* in

[1] See Safier (2022) for a list of such universities in the United States.

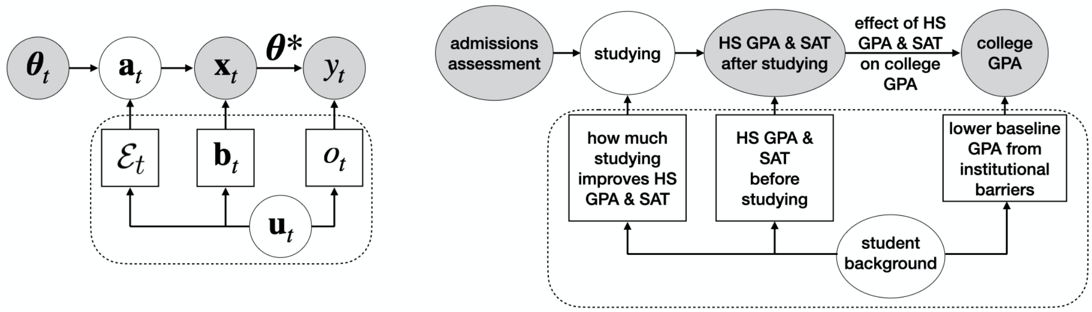


Figure 7.1: Graphical model for our setting (left) along with the way it corresponds to the admissions running example (right). Grey nodes are observed, white unobserved. Observable features $\mathbf{x}_t$ (e.g. high school GPA, SAT scores, etc.) depend on both the agent's private type $\mathbf{u}_t$ (e.g. a student's background) via initial features $\mathbf{b}_t$ (e.g. the SAT score or HS GPA student t would get without studying) and effort conversion matrix $\mathcal{E}_t$ (e.g. how much studying translates to an increase in SAT score for student t) and assessment rule $\boldsymbol{\theta}_t$ via action $\mathbf{a}_t$, which could correspond to studying, taking an SAT prep course, etc). An agent's outcome y_t (e.g. college GPA) is determined by their observable features $\mathbf{x}_t$ (via causal relationship $\boldsymbol{\theta}^*$) and type $\mathbf{u}_t$ (via baseline outcome error term o_t, which could be lower for students from underserved groups due to institutional barriers, discrimination, etc).

response to the assessment rule $\boldsymbol{\theta}_t$. We assume there exists a linear mapping $\mathcal{E}_t$ which translates efforts to changes in observable attributes for agent t. The (k, j)-th entry of this effort conversion matrix defines the change in the k-th observable attribute of agent t, $\mathbf{x}_t$, for one unit increase in the jth coordinate of their effort vector $\mathbf{a}_t$.

Latent variables: agent types. Each agent t has an unobserved private type $\mathbf{u}_t$ that can impact both their observed attributes $\mathbf{x}_t$ and true outcomes y_t. (The type is the confounder we would like to correct for.) In our running example, the type may broadly refer to the student's relevant background factors that cannot be directly observed or measured. For example, the student's type can specify their socioeconomic background factors (including the level of educational support they receive within their immediate family), as well as their interest and skills in specific subjects such as English or Mathematics.[2] Formally, we assume the type $\mathbf{u}_t$ characterizes several relevant latent

[2]Note that later in Section 4.7, we use the terminology of agent *subpopulations*. Subpopulations are distinct from types in that subpopulations determine the *distribution* of types, but individual agents belonging to the same subpopulation may have different types. We will elaborate on this in Section 4.7.

attributes of the agent, which we refer to using the tuple $\mathbf{u}_t := (\mathbf{b}_t, \mathcal{E}_t, o_t)$:

- $\mathbf{b}_t \in \mathbb{R}^m$ specifies agent t's baseline observable attribute values. For example, it can specify the baseline values of high school grades and SAT score the student would have received without any effort spent studying or preparing for standardized tests.
- $\mathcal{E}_t$ specifies agent t's *effort conversion matrix*—that is, how various effort investments in unobservable activities translate to changes in observable features.
- o_t summarizes all other environmental factors that can impact the agent's true outcome when we control for observable attributes. For example, it may reflect the effect of the institutional barriers the student faces on their actual college GPA.

We assume agent t's observable features are affected by their type and effort investments. In particular, we assume they take the form $\mathbf{x}_t = \mathbf{b}_t + \mathcal{E}_t \mathbf{a}_t$.

Agent best responses. We assume the agent selects their effort profile $\mathbf{a}_t$ in order to maximize their predicted outcome $\hat{y}_t$, subject to some *effort* cost $c(\cdot)$ associated with modifying their observable attributes. In particular, we assume the cost function is quadratic, that $c(\mathbf{a}_t) = \frac{1}{2}\|\mathbf{a}_t\|_2^2$. (Note that this assumption is common in the strategic learning literature; see, e.g., Shavit et al. (2020); Mendler-Dünner et al. (2020); Dong et al. (2018)). Formally, we assume agent t selects their effort $\mathbf{a}_t$ by solving the following optimization problem: $\max_{\mathbf{a}} \left\{\hat{y}_t - \frac{1}{2}\|\mathbf{a}\|_2^2\right\}$. It is easy to see that for a given deployed assessment rule $\boldsymbol{\theta}_t$, the agent's best-response effort investment is $\mathbf{a}_t = \mathcal{E}_t^\top \boldsymbol{\theta}_t$.

True causal outcome model. *After* each round, the principal gets to observe the agent's true outcome $y_t \in \mathbb{R}$, which takes the form $y_t = \mathbf{x}_t^\top \boldsymbol{\theta}^* + o_t$. Here $\boldsymbol{\theta}^*$ is the *true causal* relationship between an agent's observable features and outcome. (Recall that $o_t \in \mathbb{R}$ captures the dependence of agent t's outcome y_t on unobservable or unmeasured factors.) We are interested in learning $\boldsymbol{\theta}^* \in \mathbb{R}^m$, which can be interpreted as specifying how interventions impacting the value of $\mathbf{x}$ lead to changes in y. Therefore, we say that an observable feature x_i is *causally relevant* if $\theta_i^* \neq 0$. For convenience, throughout we denote the subset of causally-relevant features by $\mathbf{x}_{\mathcal{C}}$, where $\mathcal{C} \subseteq [n], \forall i \in \mathcal{C}$ if $\theta_i^* \neq 0$.

7.1.2 Overview of Results

Strategic regression as instruments. Since $\mathbf{b}_t$, $\mathcal{E}_t$, and o_t may be correlated with one another, ordinary least squares generally will *not* produce a consistent estimator for $\boldsymbol{\theta}^*$. We make the novel observation that the principal's assessment rule $\boldsymbol{\theta}_t$ is a valid *instrument*, and leverage this observation to recover $\boldsymbol{\theta}^*$ via Two-Stage Least Squares regression (2SLS). Our method applies to both *off-policy* and *on-policy* settings: one can directly apply 2SLS on historical data $\{(\boldsymbol{\theta}_t, \mathbf{x}_t, y_t)\}_{t=1}^T$, or the principal can intentionally deploy a sequence of varying assessment rules (e.g., by making small perturbations on a fixed rule) and then apply 2SLS on the collected data.

Additionally, we show that our recovery of $\boldsymbol{\theta}^*$ can be utilized to improve decision-making across several desired criteria, namely, individual fairness, agent outcomes, and predictive risk.

(Non-)causal assessment rules and fairness. In Section 7.3, we analyze the individual-level disparities that may result if the assessment rule deviates from $\boldsymbol{\theta}^*$. Unlike most existing definitions of individual fairness, which rely on the observed characteristics of individuals, our definition measures the similarity between two individuals solely by comparing their $\mathbf{b}$'s and $\mathcal{E}$'s—that is, we consider two individuals to be similar if they have the same baseline values for causally relevant observable features and similar potentials for improving these observable attributes through effort investments. Individual fairness then requires similar individuals to receive similar decisions. (We note that while our notion of individual fairness may not be easy to estimate using observational data, it is a more fine-grained—and arguably better justified—notion of individual fairness, as it distinguishes between the causally relevant and causally irrelevant facets of observable features.) We show that when making predictions using $\boldsymbol{\theta} = \boldsymbol{\theta}^*$, our notion of individual fairness is satisfied, but when the assessment rule deviates from $\boldsymbol{\theta}^*$, i.e., $\boldsymbol{\theta} \neq \boldsymbol{\theta}^*$, individual fairness may be violated by an arbitrarily large amount.

Agent outcome maximization. Note that a decision-maker can use the assessment rule $\boldsymbol{\theta}$ as a form of intervention to incentivize agents to invest their efforts optimally toward maximizing their outcomes (y). In Section 7.4, we show that utilizing the causal parameters recovered during our 2SLS procedure, one can find the assessment rule maximizing expected agent outcomes.

Predictive risk minimization. Another commonly-studied goal for decision-makers is *predictive risk minimization*, which aims to minimize $\mathbb{E}[(\hat{y}_t - y_t)^2]$, the expected squared difference between an agent's *true* outcome and the outcome predicted by the assessment. Compared to standard regression, this is a more challenging objective since both the prediction $\hat{y}_t$ and outcome y_t depend on the deployed rule $\boldsymbol{\theta}_t$. This leads to a non-convex risk function. In Section 7.5.1, we show that the knowledge of $\boldsymbol{\theta}^*$ enables us to compute an unbiased estimate of the gradient of the predictive risk. As a result, we can apply stochastic gradient descent to find a local minimum of predictive risk function.

Empirical observations. In Section 7.5, we empirically confirm and illustrate the performance of our algorithm. In particular, for a semi-synthetic dataset inspired by our university admissions example, we observe that our methods consistently estimate the true causal relationship between observable features and outcomes (at a rate of $\mathcal{O}(1/\sqrt{T})$), whereas OLS does not. Notably, OLS mistakenly estimates that SAT is causally related to college GPA, even though our experimental setup assumes it is not. On the other hand, our 2SLS-based method avoids this erroneous estimation. We also show that our methods outperform standard SGD methods in the predictive risk minimization setting.

7.2 IV Regression through Strategic Learning

Instrumental variable (IV) regression allows for consistent estimation of the relationship between an outcome and observable features in the presence of confounding terms. In this setting, we view the assessment rules $\{\boldsymbol{\theta}_t\}_{t=1}^T$ as algorithmic instruments and perform IV regression to estimate the true causal relationship $\boldsymbol{\theta}^*$. There are three criteria for $\boldsymbol{\theta}_t$ to be a valid instrument: (1) $\boldsymbol{\theta}_t$ influences the observable features $\mathbf{x}_t$, (2) $\boldsymbol{\theta}_t$ only influences the outcome y_t through $\mathbf{x}_t$, and (3) $\boldsymbol{\theta}_t$ is independent from the private type $\mathbf{u}_t$. By design, criterion (1) and (2) are satisfied. We aim to design a mechanism that satisfies criterion (3) by choosing the assessment rule $\boldsymbol{\theta}_t$ independently of the private type $\mathbf{u}_t$. As can be seen by Figure 7.1, the principal's assessment rule $\boldsymbol{\theta}_t$ satisfies these criteria.

We focus on two-stage least-squares regression (2SLS), a family of techniques for IV estimation. Intuitively, 2SLS can be thought of as estimating the causal relationship $\boldsymbol{\theta}^*$ between $\mathbf{x}_t$ and y_t by perturbing the instrument $\boldsymbol{\theta}$ and measuring the change in $\mathbf{x}_t$ and y_t. This enables us to account for the change in y_t *as a result of* the change in $\mathbf{x}_t$. 2SLS does this by independently estimating the relationship between an *instrumental variable*

$\boldsymbol{\theta}_t$ and the observable features $\mathbf{x}_t$, as well as the relationship between $\boldsymbol{\theta}_t$ and the outcome y_t via simple least squares regression. For more background on the specific version of 2SLS we use, see Section 4.8 of Cameron and Trivedi (2005).

Formally, given a set of observations $\{\boldsymbol{\theta}_t, \mathbf{x}_t, y_t\}_{t=1}^T$, we compute the estimate $\widehat{\boldsymbol{\theta}}$ of the true casual parameters $\boldsymbol{\theta}^*$ from the following process of two-stage least squares regression (2SLS). We use $\widetilde{\boldsymbol{\theta}}_t$ to denote the vector $\begin{bmatrix}\boldsymbol{\theta}_t & 1\end{bmatrix}^\top$.

1. Estimate $\Omega = \mathbb{E}[\mathcal{E}_t\mathcal{E}_t^\top]$, $\mathbb{E}[\mathbf{b}_t^\top]$ using

$$\begin{bmatrix}\widehat{\Omega}\\ \bar{\mathbf{b}}^\top\end{bmatrix} = \left(\sum_{t=1}^T \widetilde{\boldsymbol{\theta}}_t\widetilde{\boldsymbol{\theta}}_t^\top\right)^{-1}\sum_{t=1}^T \widetilde{\boldsymbol{\theta}}_t\mathbf{x}_t^\top$$

2. Estimate $\boldsymbol{\lambda} = \Omega\boldsymbol{\theta}^*$, $(\mathbb{E}[o_t] + \mathbb{E}[\mathbf{b}_t^\top]\boldsymbol{\theta}^*)$ using

$$\begin{bmatrix}\widehat{\boldsymbol{\lambda}}\\ \bar{o} + \bar{\mathbf{b}}^\top\boldsymbol{\theta}^*\end{bmatrix} = \left(\sum_{t=1}^T \widetilde{\boldsymbol{\theta}}_t\widetilde{\boldsymbol{\theta}}_t^\top\right)^{-1}\sum_{t=1}^T \widetilde{\boldsymbol{\theta}}_t y_t$$

3. Estimate $\boldsymbol{\theta}^*$ as $\widehat{\boldsymbol{\theta}} = \widehat{\Omega}^{-1}\widehat{\boldsymbol{\lambda}}$

We assume that $\sum_{t=1}^T \widetilde{\boldsymbol{\theta}}_t\widetilde{\boldsymbol{\theta}}_t^\top$ is invertible, as is standard in the 2SLS literature.

Theorem 7.2.1. *Given a sequence of bounded assessment rules $\{\boldsymbol{\theta}_t\}_{t=1}^T$ and the (observable feature, outcome) pairs $\{(\mathbf{x}_t, y_t)\}_{t=1}^T$ they induce, the distance between the true causal relationship $\boldsymbol{\theta}^*$ and the estimate $\widehat{\boldsymbol{\theta}}$ obtained via IV regression is bounded as*

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 = \widetilde{\mathcal{O}}\left(\frac{\sqrt{mT\log(1/\delta)}}{\sigma_{min}\left(\sum_{t=1}^T \boldsymbol{\theta}_t(\mathbf{x}_t - \bar{\mathbf{b}})^\top\right)}\right)$$

with probability $1-\delta$, if o_t is a bounded random variable.

Proof Sketch. While similar bounds exist for traditional IV regression problems, they do not apply to the strategic learning setting we consider. The bound follows by substituting our expressions for $\mathbf{x}_t$, y_t into the IV regression estimator, applying the Cauchy-Schwarz inequality to split the bound into two terms (one dependent on $\{(\boldsymbol{\theta}_t, \mathbf{x}_t)\}_{t=1}^T$ and one dependent on $\{(\boldsymbol{\theta}_t, o_t)\}_{t=1}^T$), and using a Chernoff bound to bound the term dependent on $\{(\boldsymbol{\theta}_t, o_t)\}_{t=1}^T$ with high probability.

While in some settings, the principal may only have access to *observational* (e.g., batch) data, in other settings, the principal may be able to actively deploy assessment rules on the agent population. We show that in scenarios in which this is possible, the principal can play random assessment rules centered around some "reasonable" assessment rule to achieve an $\mathcal{O}\left(\frac{1}{\sigma_\theta^2\sqrt{T}}\right)$ error bound on the estimated causal relationship $\widehat{\boldsymbol{\theta}}$, where σ_θ^2 is the variance in each coordinate of $\boldsymbol{\theta}_t$. Note that while playing random assessment rules may be seen as unfair in some settings, the principal is free to set the variance parameter σ_θ^2 to an "acceptable" amount for the domain they are working in. We formalize this notion in the following corollary.

Corollary 7.2.2. *If each $\theta_{t,j}$, $j \in 1, \ldots, m$, is drawn independently from some distribution $\mathcal{P}_j$ with variance σ_θ^2, $\mathbf{b}_t$ and $\mathcal{E}_t$ are bounded random variables, $\mathcal{E}_t\mathcal{E}_t^\top$ is full-rank, and $\sigma_{min}(\mathbb{E}[\mathcal{E}_t\mathcal{E}_t^\top]) > 0$, then*

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 = \widetilde{\mathcal{O}}\left(\frac{\sqrt{m\log(1/\delta)}}{\sigma_\theta^2\sqrt{T}}\right)$$

with probability $1-\delta$.

Proof Sketch. We begin by breaking up $\sigma_{min}\left(\sum_{t=1}^T \boldsymbol{\theta}_t(\mathbf{x}_t - \bar{\mathbf{b}})^\top\right)$ into two terms, $\|A\|_2$ and $\sigma_{min}(B)$, where A and B are functions of $\sum_{t=1}^T \boldsymbol{\theta}_t(\mathbf{x}_t - \bar{\mathbf{b}})^\top$. We use the Chernoff and matrix Chernoff inequalities to bound $\|A\|_2$ and $\sigma_{min}(B)$ with high probability respectively.

7.3 (Un)fairness of (Non-)causal Assessments

While making predictions based on causal relationships is important from an ML perspective for reasons of generalization and robustness, the societal implications of using non-causal relationships to make decisions are perhaps an even more persuasive reason to use causally-relevant assessments. In particular, it could be the case that a certain individual is worse at strategically manipulating features which are not causally relevant when compared to their peers. If these attributes are used in the decision-making process, this agent may be unfairly seen by the decision-maker as less qualified than their peers, even if their initial features and ability for improvement is similar to others.

One important criterion for assessing the fairness of a machine learning model at the individual level is that two individuals who have similar merit should receive similar predictions. Dwork et al. (2012) formalize this intuition through the notion of *individual fairness*, which is formally defined as follows.

Definition 7.3.1 (Individual Fairness Dwork et al. (2012))**.** *A mapping* $M : \mathcal{U} \to \Delta(Y)$ *is individually fair if for every* $\mathbf{u}, \mathbf{u'} \in \mathcal{U}$*, we have*

$$D(M(\mathbf{u}), M(\mathbf{u'})) \leq d(\mathbf{u}, \mathbf{u'}),$$

where $\mathbf{u}, \mathbf{u'} \in \mathcal{U}$ *are individuals in population* $\mathcal{U}$*,* $\Delta(Y)$ *is the probability distribution over predictions* Y*,* $D(M(\mathbf{u}), M(\mathbf{u'}))$ *is a distance function which measures the similarity of the predictions received by* $\mathbf{u}$ *and* $\mathbf{u'}$*, and* $d(\mathbf{u}, \mathbf{u'})$ *is a distance function which measures the similarity of the two individuals.*

Recall that in the setting we consider, the mapping between individuals and predictions is defined to be $M(\mathbf{u}) := \mathbf{x}^\top \boldsymbol{\theta} + \hat{o} = (\mathbf{b} + \mathcal{E}\mathcal{E}^\top)^\top \boldsymbol{\theta} + \hat{o}$. The prediction an individual receives is deterministic, so a natural choice for $D(M(\mathbf{u}), M(\mathbf{u'}))$ is $|\hat{y} - \hat{y}'|$. We take a causal perspective when defining a metric $d(\mathbf{u}, \mathbf{u'})$ to measure the similarity of two individuals $\mathbf{u}$ and $\mathbf{u'}$. Intuitively, individuals that have similar initial causally-relevant features and ability to modify causally-relevant features should be treated similarly. Therefore, we define $d(\mathbf{u}, \mathbf{u'})$ to reflect the difference in causally-relevant components of $\mathbf{b}$ & $\mathbf{b'}$ (initial feature values) and $\mathcal{E}\mathcal{E}^\top$ & $\mathcal{E}'\mathcal{E}'^\top$ (ability to manipulate features). With this in mind, we are now ready to define the criterion for individual fairness to be satisfied in the strategic learning setting.

Definition 7.3.2. *In the strategic learning setting, individual fairness is satisfied if*

$$\begin{aligned}
|\hat{y} - \hat{y}'| &\leq d(\mathbf{u}, \mathbf{u'}) \\
&= \|\mathbf{b}_\mathcal{C} - \mathbf{b}'_\mathcal{C}\|_2 + \|(\mathcal{E}\mathcal{E}^\top)_\mathcal{C} - (\mathcal{E}'\mathcal{E}'^\top)_\mathcal{C}\|_2,
\end{aligned}$$

where

$$b_{\mathcal{C},i} = \begin{cases} b_i & \textit{if } i \in \mathcal{C} \\ 0 & \textit{otherwise} \end{cases},$$

$$(\mathcal{E}\mathcal{E}^\top)_{\mathcal{C},ij} = \begin{cases} (\mathcal{E}\mathcal{E}^\top)_{ij} & \textit{if } i \in \mathcal{C} \textit{ or } j \in \mathcal{C} \\ 0 & \textit{otherwise.} \end{cases}$$

Recall that $\mathcal{C} \subseteq \{1, \ldots, n\}$ denotes the set of indices of observable features $\mathbf{x}$ which are causally relevant to y (i.e., $\theta_i^* \neq 0$ for $i \in \mathcal{C}$).

Theorem 7.3.3. *Assessment $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ satisfies individual fairness for any two agents $\mathbf{u}$ and $\mathbf{u}'$.*

Proof Sketch. The full proof follows straightforwardly from the Cauchy-Schwarz inequality and the definition of the matrix operator norm. (Our results are not dependent upon the specific matrix or vector norms used, analogous results will hold for other popular choices of norm.) Throughout the proof we assume that $\|\boldsymbol{\theta}^*\|_2 = 1$ by definition, although our results hold up to constant multiplicative factors if this is not the case.

While $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ satisfies the criterion for individual fairness, this will generally not be the case for an arbitrary assessment $\boldsymbol{\theta} \neq \boldsymbol{\theta}^*$. For instance, consider the case where $d(\mathbf{u}, \mathbf{u}') = 0$ for two agents $\mathbf{u}$ and $\mathbf{u}'$. Under this setting, it is possible to express $|\hat{y} - \hat{y}'|$ using quantities which do not depend on $d(\mathbf{u}, \mathbf{u}')$. As these quantities increase, $|\hat{y} - \hat{y}'|$ increases as well, despite the fact that $d(\mathbf{u}, \mathbf{u}')$ remains constant.

Theorem 7.3.4. *For any deployed assessment rule $\boldsymbol{\theta}$, the gap in predictions between two agents $\mathbf{u}$ and $\mathbf{u}'$ such that $d(\mathbf{u}, \mathbf{u}') = 0$ is*

$$\begin{aligned} |\hat{y} - \hat{y}'| = \Bigg| & \sum_{i \notin \mathcal{C}} (b_i - b_i')\theta_i \\ & + \sum_{i \notin \mathcal{C}} \sum_{j \notin \mathcal{C}} ((\mathcal{E}\mathcal{E}^\top)_{ij} - (\mathcal{E}'\mathcal{E}'^\top)_{ij})\theta_i\theta_j \Bigg| \end{aligned}$$

Note that all components of $\boldsymbol{\theta}$ which appear in Theorem 7.3.4 are outside of the support of $\boldsymbol{\theta}^*$.

In order to illustrate how $|\hat{y} - \hat{y}'|$ can grow while $d(\mathbf{u}, \mathbf{u}')$ remains constant, consider the following example.

Example 7.3.5. Consider a setting in which the distance $d(\mathbf{u}, \mathbf{u}') = 0$ between agents $\mathbf{u}$ and $\mathbf{u}'$, and there is a one-to-one mapping between actions and observable features for each agent, with one agent having an advantage when it comes to manipulating features which are not causally relevant. Formally, let $\boldsymbol{\theta}^* = [\mathbf{0}_{n/2}^\top \ \sqrt{\frac{2}{n}}\mathbf{1}_{n/2}^\top]^\top$, $\mathbf{b} = \mathbf{b}'$, $\mathcal{E} = \boldsymbol{\delta} I_{n \times n}$, and $\mathcal{E}' = \boldsymbol{\delta}' I_{n \times n}$, where $\boldsymbol{\delta} = [\sqrt{n}\mathbf{1}_{n/2}^\top \ \mathbf{1}_{n/2}^\top]^\top$ and $\boldsymbol{\delta}' = [\mathbf{0}_{n/2}^\top \ \mathbf{1}_{n/2}^\top]^\top$.

Under such a setting, the equation in Theorem 7.3.4 simplifies to

$$|\hat{y} - \hat{y}'| = n \sum_{i \notin \mathcal{C}} \theta_i^2 = n \sum_{i=1}^{n/2} \theta_i^2.$$

Suppose now that the assessment $\boldsymbol{\theta}$ puts weight at least $1/\sqrt{n}$ on each observable feature which is not causally relevant. Under such a setting, $|\hat{y} - \hat{y}'| \geq n/2$, meaning that the difference in predictions tends towards infinity as n grows large, despite the fact that $d(\mathbf{u}, \mathbf{u}') = 0$ and $y = y'$!

7.4 Agent Outcome Improvements

In the strategic learning setting, the goal of each agent is clear: they aim to achieve the highest prediction $\hat{y}$ possible, regardless of their true label y. On the other hand, what the goal should be for the principal is less clear, and depends on the specific setting being considered. For example, in some settings it may be enough to discover the causal relationships between observable features and outcomes. However in other settings, the principal may wish to take a more active role. In particular, when making decisions which have real-world consequences, it may be in the principal's best interest to use a decision rule which promotes desirable behavior Kleinberg and Raghavan (2020); Shavit et al. (2020); Harris et al. (2021), i.e., behavior which has the potential to improve the *actual outcome* of an agent.

In the agent outcome improvement setting, the goal of the principal is to maximize the expected outcome $\mathbb{E}[y]$ of an agent drawn from the agent population. In our college admissions example, this would correspond to deploying an assessment rule with the goal of *maximizing* expected student college GPA. Formally, we aim to find $\boldsymbol{\theta}^{AO}$ in a convex

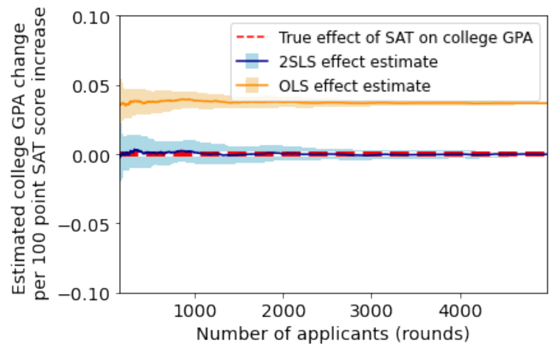


Figure 7.2: OLS versus 2SLS estimates for SAT effect on college GPA over 5000 rounds. Results are averaged over 10 runs, with the error bars (in lighter colors) representing one standard deviation. The red dashed line is the true causal relationship between SAT score and college GPA.

set $\mathcal{S}$ of feasible assessment rules such that the induced expected agent outcome $\mathbb{E}[y]$ is maximized.

After some algebraic manipulation, the optimization becomes $\boldsymbol{\theta}^{AO} = \arg\max_{\boldsymbol{\theta}\in\mathcal{S}} \boldsymbol{\theta}^\top \boldsymbol{\lambda}$, where $\boldsymbol{\lambda} = \mathbb{E}[\mathcal{E}_t\mathcal{E}_t^\top]\boldsymbol{\theta}^*$.

Note that while the principal never directly observes $\mathbb{E}[\mathcal{E}_t\mathcal{E}_t^\top]$ nor $\boldsymbol{\theta}^*$, they estimate $\boldsymbol{\lambda} = \mathbb{E}[\mathcal{E}_t\mathcal{E}_t^\top]\boldsymbol{\theta}^*$ during the second stage of the 2SLS procedure. Therefore, if the principal has already run 2SLS to recover a sufficiently accurate estimate of the causal parameters $\boldsymbol{\theta}^*$, they can estimate the agent outcome-maximizing decision rule by solving the above optimization.

7.5 Experiments

We empirically evaluate our model on a semi-synthetic dataset inspired by our running university admissions example. We compare our 2SLS-based method against ordinary least squares (OLS), which directly regresses observed outcomes y on observable features $\mathbf{x}$. We show that even in our stylized setting with just two observable features, OLS does not recover $\boldsymbol{\theta}^*$, whereas our method does.

University admissions experimental description We constructed a semi-synthetic dataset based on the SATGPA dataset, a collection of real university admissions data.[3]

[3]Originally collected by the Educational Testing Service, the SATGPA dataset is publicly available

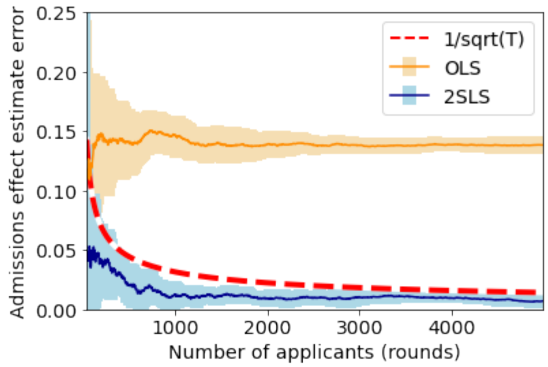


Figure 7.3: OLS effect estimate error $\|\widehat{\boldsymbol{\theta}}_{\text{OLS}} - \boldsymbol{\theta}^*\|_2$ (in orange) and 2SLS estimate error $\|\widehat{\boldsymbol{\theta}}_{\text{2SLS}} - \boldsymbol{\theta}^*\|_2$ (in blue) over 5000 rounds. Results are averaged over 10 runs. Error bars (in lighter colors) represent one standard deviation. 2SLS estimate error decreases at a rate of about $\frac{1}{\sqrt{T}}$ (red dashed line).

The SATGPA dataset contains 6 variables on 1000 students. We use the following: two features (high school (HS) GPA and SAT score) and an outcome (college GPA). Using OLS (which is assumed to be consistent since we have yet to modify the data to include confounding), we find that the effect of [SAT, HS GPA] on college GPA in this dataset is $\boldsymbol{\theta}^* = [0.0015, 0.5895]^\top$. We then construct synthetic data that is based on this original data, yet incorporates confounding factors. For simplicity, we let the true effect $\boldsymbol{\theta}^* = [0, 0.5]^\top$. That is, we assume HS GPA is causally related to college GPA, but SAT score is not.[4] We consider two private types of applicant backgrounds: *disadvantaged* and *advantaged*. Disadvantaged applicants have lower initial HS GPA and SAT ($\mathbf{b}$), lower baseline college GPA (o), and need more effort to improve observable features ($\mathcal{E}$).[5] Each applicants' initial features are randomly drawn from one of two Gaussian distributions, depending on background. Applicants may manipulate both of their features.

Results. In Figure 7.2, we compare the true effect of SAT score on college GPA ($\boldsymbol{\theta}^*$) with the estimates of these quantities given by our method of 2SLS from Section 7.2 ($\hat{\boldsymbol{\theta}}_{\text{2SLS}}$) and with the estimates given by OLS ($\hat{\boldsymbol{\theta}}_{\text{OLS}}$). (An analogous figure for the effects

and can be found here: `https://www.openintro.org/data/index.php?data=satgpa`.

[4]Though this assumption may be contentious, it is based on existing research (e.g., Allensworth and Clark (2020)).

[5]For example, this could be due to the disadvantaged group being systemically underserved or marginalized (and the converse for advantaged group).

of HS GPA is included in the appendix.) In Figure 7.3, we compare the estimation errors of OLS and 2SLS, i.e. $\|\widehat{\boldsymbol{\theta}}_{\text{OLS}} - \boldsymbol{\theta}^*\|_2$ and $\|\widehat{\boldsymbol{\theta}}_{\text{2SLS}} - \boldsymbol{\theta}^*\|_2$.

We find that our 2SLS method converges to the true causal relationship (at a rate of about $\frac{1}{\sqrt{T}}$), whereas OLS has a constant bias. Although our setting assumes that SAT score has no causal relationship with college GPA, OLS mistakenly predicts that, on average, a 100 point increase in SAT score leads to about a 0.05 point increase in college GPA. If SAT were not causally related to collegiate performance in real life, these biased estimates could lead universities to erroneously use SAT scores in admissions decisions. This highlights the advantage of our method, since using a naive parameter estimation method like OLS in the presence of confounding could cause decision-making institutions to deploy assessments which don't accurately reflect the characteristics they are trying to measure.

7.5.1 Predictive Risk Minimization

Analogous to recovering causal relationships and improving agent outcomes, another common goal of the principal in the strategic learning setting is to *minimize predictive risk.* Formally, the goal of the principal in the predictive risk minimization setting is to learn the assessment rule that minimizes the expected squared difference between an agent's true outcome and the outcome predicted by the principal, i.e., $f(\boldsymbol{\theta}_t) = \mathbb{E}[(\widehat{y}_t - y_t)^2]$.

Due to the dependence of $\mathbf{x}_t$ and y_t on $\boldsymbol{\theta}_t$, $f(\boldsymbol{\theta}_t)$ will be non-convex in general, and can have several extrema which are not global minima, even in the case of just one observable feature. When faced with such non-convex optimization problems, gradient descent is often a popular approach due to its simplicity and convergence to local minima in practice.

If the effort conversion matrix $\mathcal{E}$ is the same for all agents, the gradient of population risk function can be written as

$$\nabla_{\boldsymbol{\theta}_t} f(\boldsymbol{\theta}_t) = 2(\mathbb{E}[(\widehat{y}_t - y_t)\mathbf{x}_t] + \mathbb{E}[\widehat{y}_t - y_t]\mathcal{E}\mathcal{E}^\top(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*).$$

In our college admissions example, this would correspond to the setting in which all students' math GPA, SAT scores, etc. improve the same amount given the same effort: this may be a reasonable assumption if the students being considered have the same

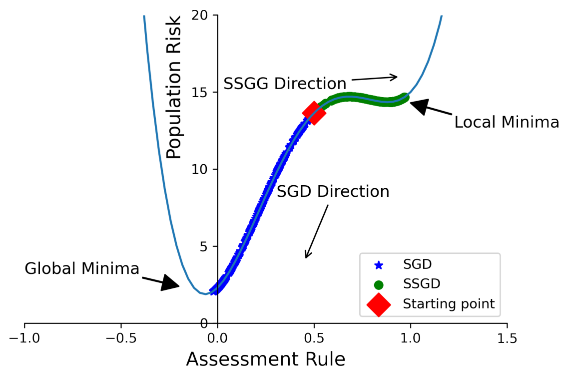


Figure 7.4: Stochastic Gradient Descent (SGD, takes into account x_t, y_t, $\widehat{y}_t$'s dependence on θ_t) vs Simple Stochastic Gradient Descent (SSGD, does not). In the 1D setting, it is possible for the gradient of SSGD to have the wrong sign. When both are initialized at $\theta_0 = 0.5$, SGD is able to follow the gradient and converge to the global minima, while SSGD is not. We ran each method for 1000 time-steps with a decaying learning rate of $\frac{0.001}{\sqrt{T}}$.

ability to learn, despite other differences in background they may have. If $\mathcal{E}\mathcal{E}^\top$ is known to the principal (e.g. through the 2SLS procedure in Section 7.2), then each $(\boldsymbol{\theta}_t, \mathbf{x}_t, y_t)$ tuple can be used to compute an unbiased estimate of $\nabla_{\boldsymbol{\theta}_t} f(\boldsymbol{\theta}_t)$ for use in online gradient descent.

Recent work on *performative prediction* Perdomo et al. (2020); Mendler-Dünner et al. (2020); Miller et al. (2021) examines the use of repeated gradient descent in the strategic learning setting and finds that repeated gradient descent generally converges to *performatively stable* points. There is no direct comparison between performatively stable points and local minima in our setting. In fact, performatively stable points can actually *maximize* predictive risk under some settings. (See Miller et al. (2021) for such an example.) Our methods differ from this line of work because we take $\mathbf{x}_t$, y_t, and $\widehat{y}_t$'s *direct dependence* on the assessment rule $\boldsymbol{\theta}_t$ into account when calculating the gradient of the risk function, whereas these performative prediction models (henceforth *simple stochastic gradient descent* or *SSGD*) do not. While SSGD may be satisfactory for some settings, it produces a biased estimate of the gradient in general, which can lead to unexpected behavior under our setting; by contrast, our gradient estimate is unbiased (see Figure 7.4). Even in situations which SSGD *does* get the sign of the gradient correct, it may converge at a much slower rate, due to its incomplete estimate of the gradient.

7.6 Conclusion

In this work, we establish the possibility of recovering the causal relationship between observable attributes and the outcome of interest in settings where a decision-maker utilizes a series of linear assessment rules to evaluate strategic individuals. Our key observation is that in strategic settings, assessment rules serve as valid instruments (because they causally impact observable attributes but do not directly affect the outcome). This observation enables us to present a 2SLS method to correct for confounding bias in causal estimates. We then demonstrate the potential of the recovered causal coefficients to be utilized for preventing individual-level disparities, improving agent outcomes, and reducing predictive risk minimization.

While our work offers an initial step toward extracting causal knowledge from automated assessment rules, we rely on several simplifying assumptions—all of which mark critical directions for future work. In particular, we assume all assessment rules and the underlying causal model are linear. This assumption allows us to utilize linear IV methods. Extending our work to *non-linear* assessment rules and IV methods is necessary for the applicability of our method to real-world settings. Another critical assumption is the agent's *full knowledge* of the assessment rule and their *rational* response to it, subject to a *quadratic effort cost*. While these are standard assumptions in economic modeling, they need to be empirically verified in the particular decision-making context at hand before our method's outputs can be viewed as reliable estimates of causal relationships.

Chapter 8

Incentivizing Collaborative Learning with Network Effects

8.1 Introduction

Collaborative machine learning schemes such as federated learning (FL) enable a set of clients (*e.g.,* mobile devices, hospitals) to jointly learn a model without directly sharing their private local data McMahan et al. (2017); Kairouz et al. (2021). In real-world applications, the performance of the model on a particular client may not be improved through model sharing due to heterogeneity in the data generated by others in the networks (Li et al., 2021). Moreover, even if a particular client has improved model utility on their local data by joining collaborative training, they may suffer from costs induced by participating. For example, the communication or latency in FL may affect the quality of clients' user experience, and a hospital may suffer some privacy costs for sharing sensitive information about its patients. Understanding client incentives in collaborative learning (CL) is crucial for the service provider to design better solutions and for each client to make better decisions.

There is growing literature on incentive mechanisms in collaborative learning. Most of these works aim to study how to combine the data from all the clients to achieve a low error rate and ensure fairness for participating clients. However, these works are often based on a common assumption that all clients are willing participants who will join the network in order to achieve different learning goals (*e.g.,* minimize the sample complexity

(Blum et al., 2021), learn coalition structures (Donahue and Kleinberg, 2021c), *etc.*) In practice, rational clients can decide to either opt-in or opt-out of collaborative training based on their utility gain and cost as the network grows over time. The formal study of such a dynamic has not been captured in prior works studying incentives in collaborative learning.

In this work, we propose to model clients' behaviors in collaborative learning through the lens of network effects games (Katz and Shapiro, 1985; Shapiro and Varian, 1999). A network effects game models the benefit of each individual after aligning their behavior with the behaviors of others in a coalition. This game closely mirrors expected behavior in collaborative learning, where model sharing can improve the utility compared to only training a model on local data. Traditionally, network effects games have not been looked at in the context of CL or more broadly, data sharing. Existing work in CL typically models client participation as a one-time game where clients arrive and make decisions once before leaving. We instead focus on modeling the dynamics of client participation in a setting where (1) each client incurs a cost when joining the network, and (2) each client joins the network *if and only if* their utility gain outweighs the cost. We then aim to find and analyze an equilibrium, such that no client has an incentive to alter their decision.

Based on such a formulation, we consider two settings: (1) a warm-up example of a stylized global mean estimation problem (Section 8.3) and (2) a more general setting where the gain from joining the collaborative learning process depends on the composition of the coalition (Section 8.4). For each setting, we characterize the dynamics of the client participation game under the network effects model. A key feature of CL is that clients' utility of joining depends on the number of participants/data points. Since different clients may have different utilities and costs, it may be easier to gradually incentivize participation over time. For example, we may target the clients who have lower costs and are more willing to join, and their participation will increase other participants' willingness to join, and so on. Based on this characterization, we design a cost-effective payment scheme that targets a subset of clients over time to reach a desirable equilibrium in Section 8.5. We summarize **our contributions** below:

- We propose to model the clients' behaviors in a dynamic collaborative learning setting as a network effects game, in which clients' utilities depend on the number of local samples of other participants. We want to find the self-fulfilling expectation equilibrium in CL, when the outcome of the clients' best-response strategy matches the shared expectation.

- We characterize all self-fulfilling expectation equilibria of the collaborative network effects game in a global mean estimation problem and in a more general learning problem where the client's participating decision requires knowledge about the composition of the coalition.
- We show that as clients continually best respond to public information, their dynamics naturally converge to an equilibrium where no clients have incentives to alter their decisions.
- Finally, we provide a cost-efficient payment scheme that incentivizes clients to reach a desired equilibrium. We show that for the dynamic to converge to a higher point, the server only needs to give payment until the coalition has reached a *tipping point* determined by the dynamics.

8.2 Problem Formulation and Preliminaries

In this section, we first introduce the setup of our collaborative learning problem. We then characterize the incentive and utility gain for each client from joining the data-sharing process and their interaction with the server. Finally, we propose a general framework of collaborative learning as a dynamic network effects game.

8.2.1 Learning Problem Formulation

We consider a collaborative learning setup where there are M clients in the population who want to solve a common learning problem. For each client $i \in [M]$, we assume that they want to minimize their loss function given by $f_i(\mathbf{w}) = \mathbb{E}_{\xi \sim \mathcal{D}_i}[\ell(\mathbf{w}, \xi)]$, where $\mathcal{D}_i$ is the target data distribution of client i, and $\ell(\mathbf{w}, \xi)$ is the loss function for a model $\mathbf{w}$ for data sample ξ. There are two possibilities: all clients share a target distribution $\mathcal{D}_i = \mathcal{D}$, *e.g.,* global mean estimation, and when clients have different target distribution $\mathcal{D}_i$, *e.g.,* personalized local mean estimation. In practice, each client only has access to its local training dataset with n_i data point sampled i.i.d from data distribution $\mathcal{D}_i$.

Personal cost. If a client i in the population decides to join the coalition, they will incur some cost $c_i > 0$ that may depend on the number of local data samples and the size of the coalition. This cost c_i contains the enrollment and communication cost of joining the coalition. The cost may also represent the privacy loss for the client i from revealing information about their local data to the server. For example, a server can charge each participating hospital a fixed fee for using the shared model. At each time step, each hospital needs to communicate with the server to update its knowledge base and decide to either opt-in or opt-out. Finally, due to the sensitive nature of medical data and how each individual model is trained, the hospitals will also incur some privacy loss for sharing their model with the server.

Client Incentive and Utility Gain. The goal of each client i is to find a model $\mathbf{w}_i^*$ that minimize its true loss function, *i.e.,* $\mathbf{w}_i^* := \arg\min_{\mathbf{w}} f_i(\mathbf{w})$. We assume that each client will perform local training on their own dataset to obtain a local model $\mathbf{w}_i$. Since the local dataset has a small size, the local model $\mathbf{w}_i$ might not generalize well to the target distribution $\mathcal{D}_i$. Hence, a client would consider participating in collaborative learning if the federated model $\mathbf{w}_i^S$ gives a better performance. We define the expected utility gain of joining a coalition S for a client i as the expected improvement from using $\mathbf{w}_S$ minus the cost of joining:

$$U_i(S, c_i) = \mathbb{E}[f(\mathbf{w}_i) - f(\mathbf{w}_i^S)] - c_i$$

where the expectation is taken over randomness in $\mathcal{D}_{j:j \in S}$.

Interaction between clients and server. We assume that at each time step t, the server will post information about the clients who are currently willing to join the coalition, denoted by $S^{(t)}$. Then based on $S^{(t)}$, a client has to decide whether to join the coalition or use only their local data. We assume that if the expected utility gain from joining the coalition $U_i(S^{(t)}, c_i)$ is positive for a client $i \in [M]$, then client i would join the coalition in the next step. That is, given a coalition $S^{(t)}$ and a client $i \in [M]$ with $U_i(S^{(t)}, c_i) \geq 0$, then client i would opt in the data-sharing process. Otherwise, if the expected utility gain $U_i(S^{(t)}, c_i) < 0$, then client i would opt out of the data-sharing process. After T rounds, the coalition will be finalized, and the clients send to the server some aggregation statistics from its local dataset, *e.g.,* for the mean estimation problem, the message would be the empirical average of local data samples. The server then computes and sends back the federated models $\mathbf{w}_i^S$ to all clients in the final coalition.

8.2.2 Collaborative Learning as Dynamic Network Effects Game

We model the dynamics of the population in collaborative learning as a *network effects* game. When all clients share an expectation on the coalition S, each client $i \in [M]$ can calculate their expected utility gain as a function of S and their personal cost, *i.e.,* $U_i(S, c_i)$. After a coalition is formed, all clients can observe the actual coalition composition $\{n_j\}_{j \in S}$, denoted $h(S)$, and update their shared expectations for the next time step. We define the *realization mapping* between the shared expectation S and the actual outcome as $h : P([M]) \rightarrow P([M])$ where $P([M])$ is the power set of all clients, *i.e.,* if the shared expectation is S, then the actual coalition is $h(S)$.

As a result, clients can compute their utility gain with the realized outcome and make their decision. For example, if the actual coalition is larger than expected, *i.e.,* $|h(S^{(t)})| > |S^{(t)}|$ then the utility gain from joining would increases leading to some new clients joining and vice versa. We aim to study the dynamic of the population and characterize a suitable equilibrium under these behaviors.

Definition 8.2.1 (Nash equilibrium)**.** *A coalition S is a Nash equilibrium if the coalition does not change anymore when S is reached at time step t, that is, suppose $S^{(t)} = S$, then $S^{(t+1)} = S$.*

If the clients' shared expectation is perfect, we can define an equilibrium notion. If all clients form a shared expectation on the composition of the coalition S, and if each of them decides to either join or leave the coalition based on this expectation, then the actual coalition matches their shared expectation S. We call this a *self-fulfilling expectation* given a realization mapping $h(S)$:

Definition 8.2.2 (Self-Fulfilling Expectation Equilibrium)**.** *Consider a population where all clients share an expectation on the composition of coalition S. Given a realization mapping $h(S)$, if the actual coalition S exactly matches the shared expectation,* i.e., $h(S) = S$*, then S is called a self-fulfilling expectation equilibrium.*

Characterizing self-fulfilling expectation equilibria is crucial to the server and policymakers who wish to understand and influence where the dynamic converges. In Section 8.3, we study the stylized setting of global mean estimation to introduce our dynamics analysis techniques. Then, in Section 8.4, we look at a more general class of learning problems whose utility function requires knowledge of the entire coalition composition, which requires a more careful analysis of the participation dynamics. Finally, we show that the server can leverage some properties of the equilibria to guarantee that the final coalition reaches the maximum size equilibrium by strategically paying a subset of clients.

8.3 Warm-up: Network Effects game in a Global Mean Estimation Problem

In this section, we investigate how the dynamic behaves in a stylized global mean estimation problem when the utility gain for joining only depends on the number of samples in the coalition. First, we formally describe the setup when there all clients in the population have the same number of local samples. Then, we show that we can always find a set of self-fulfilling expectation equilibrium in this setting with some simple cost functions. Finally, we show that the dynamic naturally converges to an equilibrium where no clients have further incentive to change their decisions.

Mean Estimation Problem Setup. There are M clients in the population who want to solve a global mean estimation problem. Each client i has a fixed number of samples n_i and a personal cost c_i[1]. First, each client i draws their mean parameter μ_i i.i.d from a common distribution with mean θ: $\mu_i \sim \mathcal{N}(\theta, \sigma_\theta^2)$, then draws their samples i.i.d from a unit variant Gaussian with local mean μ_i: $x_{i,j} \sim \mathcal{N}(\mu_i, 1)$. Formally, the clients want to find an estimator $\mathbf{w}$ such that the mean squared error (MSE) $\mathbb{E}[(\theta - \mathbf{w})^2]$ is minimized. We focus our analysis on the case where all clients have the same number of local data, *i.e.,* $\forall i \in [M], n_i = n$.

For a client i, without joining the coalition, their estimator is the empirical average over local data samples $\mathbf{w}_i = {}^1\!/\!_{n_i} \sum_{j=1}^{n_i} x_{i,j}$. If the client decides to join a coalition S,

[1]We only assume the personal cost is fixed in this warm-up example for the sake of simplicity

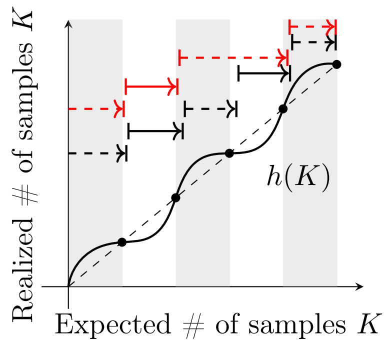


Figure 8.1: Illustration of a realization mapping $h(K)$ and two possible trajectories of our payment mechanism Algorithm 9. The black curve represents the realization mapping $h(K)$ and the gray area represents the region where the cascading behaviors may occur $\{K : h(K) > K\}$. The blue line and the red line show two possible trajectories of Algorithm 9, where the dashed arrows represent the growth of the coalition due to cascading behaviors and the solid arrows represent the growth of the coalition due to payments.

they will send their local estimator w_i to the server and receive a weighted average over the local estimators of participating clients: $\mathbf{w}_i^S = \mathbf{w}_S = {}^1/_{N_S} \sum_{j \in S} w_j \cdot n_j$ where N_S is the total number of samples in the coalition.

If the expected number of clients in the coalition is K, the utility gain function for a client i is:

$$U_i = U_i(K, n, c_i) = \frac{K-1}{Kn} + \frac{3K^2 - 5K + 2}{K^2}\sigma_\theta^2 - c_i \tag{8.1}$$

We abuse the notation and write the expected number of clients in the coalition K as the shared expectation of all clients. Similarly, we define the realization mapping $h(K)$ as the number of clients in the coalition in this setting.

8.3.1 Existence of a Self-fulfilling Expectation Equilibrium

Realization mapping $h(\cdot)$. Since all clients have the same number of local samples, W.L.O.G, we can order them according to their personal cost in ascending order. Moreover, we assume that the cost of joining for a client i is a positive monotonically non-decreasing function of their integer value index i, *i.e.*, $c_i = c(i)$. If all clients share a common belief that K clients will join the coalition, then client i would join only if $U_i(K, n, c_i) \geq 0$. Hence, if any client would join at all, then the set of clients joining would be between

1 and $h(K)$, where $h(K) = \inf\{x \in \mathbb{R} : U_i(x, n, c_i) \geq 0\}$. Figure 8.1 shows a possible realization mapping $h(K)$ as a function of the expectation K.

Moreover, we prove that there always exists an equilibrium under some mild assumption on shared expectation K and the actual number of clients in the coalition $h(K)$. Informally, if the shared expectation K is pessimistic compared to the actual outcome $h(K)$, then we can always find an equilibrium. The clients would update their expectations and join the coalition due to a new higher utility gain. However, as the $h(\cdot)$ function is non-decreasing, there exist some clients whose utility gain is less than the cost of joining and they do not have incentives to join the coalition.

Lemma 8.3.1. *In the mean estimation problem where all clients have the same number of local samples, for any positive number of clients in the coalition K with $h(K) \geq K$, we can find a self-fulfilling expectation equilibrium $h(K^*) = K^*$ to the right of K with $K^* \geq K$; and for any point in between $K' \in (K, K^*)$, we have $h(K') > K'$.*

8.4 Collaborative Learning with Coalition-dependent Utilities

In the preceding section, our analysis was predicated on the assumption that the clients' utility gain is solely determined by the number of samples. Although the number of samples undoubtedly plays a crucial role, it may not be the sole determinant of the clients' utility gain. The composition of the coalition, for instance, could also be of consideration. For instance, having 10 datasets with 1000 samples each may not yield the same utility gain as having 1000 datasets with only 10 samples each. To further illustrate this point, we present two specific examples of collaborative learning problems where the utility gain cannot be adequately determined by the number of data points alone.

Example 8.4.1 (Global Mean Estimation with Heterogeneous clients)**.** Consider the mean estimation problem defined in Section 8.3 with different n_i's. Then each client's utility gain will depend on the composition of the coalition $\{n_j\}_{j\in S}$ as

$$U_i(S,c_i) = \frac{1}{n_i} - \frac{1}{N_S} \\ - \left(\frac{\sum_{i\neq j} n_j^2 - (N_S - n_i)^2}{N_S^2} - \frac{2(N_S - n_i)}{N_S}\right)\sigma_\theta^2 - c_i$$

Example 8.4.2 (Learning with shared representation)**.** Suppose each client i has a mean vector $\mu_i \in \mathbb{R}^d$ and all the vectors share common support of size k, *i.e.,* , $\mu_1, \ldots, \mu_M$ lie in the same k-dimensional subspace. The clients do not know the low-rank support and they want to estimate their own mean vector μ_i by drawing noisy samples of μ_i. Each client chooses to join the coalition or not. The server collects some aggregated statistics from each participating client, estimates the common support, and sends the estimated support back to the clients. In this problem, the probability of identifying the correct support depends on the scale of the noises in each participating client's samples.

In this section, we venture into a wider range of utility functions that incorporate the coalition composition, offering a more comprehensive approach to accommodate more general collaborative learning problems. Despite the added complexity introduced by these utility functions, we still characterize the dynamic by defining a more sophisticated realization mapping $h(\cdot)$. In particular, we will continue to observe "cascading behaviors" when $h(K) \geq K$, where the participation of some clients triggers a chain effect, leading to an increasing number of clients joining the coalition.

8.4.1 Utility function and agent model

The dynamics of this network effect game are intrinsically tied to the utility functions of the clients involved. Without careful modeling of possible utility functions, deriving meaningful insights for the dynamic becomes an exceedingly challenging task. First of all, the clients certainly care about the total number of data points and their costs for participation. In addition, as we have seen in Example 8.4.1 and Example 8.4.2, the clients may also place significance on the coalition composition. We thus extend our modeling of client utility to include this dependency on the coalition composition. We formulate the dependency in a general form to accommodate a wide range of functions, while keeping the problem tractable. More specifically, we examine a scenario where a client i's utility gain, given their decision to join and the set of clients who also choose to join (S), depends on two key factors:

(1) The composition of the coalition $\{n_j\}_{j\in S}$, represented by a function $f(\{n_j\}_{j\in S})$. We write $f(\{n_j\}_{j\in S}) = f(S)$ as shorthand and assume that $f(\cdot)$ is monotone: if $S \subseteq S'$, then $f(S) \leq f(S')$.

(2) A personalized utility change $V(N_S, n_i, c_i)$ that depends on the total number of data points N_S, client i's dataset size n_i, and client i's cost of joining $c_i = c(N_S, n_i)$, *i.e.,* the personal cost c_i is a non-negative function of N_S and n_i. We assume that $V(\cdot)$ is increasing in N_S, that is, with fixed n_i, if $N_S \leq N'_S$, then $V(N_S, n_i, c_i) \leq V(N'_S, n_i, c_i)$. Moreover, this assumption implies that once a client i has joined the coalition at a previous time step, the increase in their cost of staying in a coalition c_i will not exceed the personal utility gained from being in the coalition when more clients join the coalition.

More specifically, we assume that when client i chooses to join and the set of clients who choose to join is S, client i's utility gain can be represented as

$$U_i(S, c_i) = g(f(S), V(N_S, n_i, c_i)),$$

where $g(x, y)$ is a function that captures how each client's utility gain depends on the coalition structure $f(S)$ and her personalized utility change $V(N_S, n_i, c_i)$. We assume that $g(x, y)$ is monotonically increasing in both parameters, that is, $g(x, y) \geq g(x', y)$ for $x \geq x'$ fixing any y and $g(x, y) \geq g(x, y')$ for $y \geq y'$ fixing any x. Examples of such $g(\cdot)$ functions include:

- $U_i(S, c_i) = f(S) + V(N_S, n_i, c_i)$;
- $U_i(S, c_i) = f(S) \cdot V(N_S, n_i, c_i)$;
- Every client incur the same cost c of joining, and $U_i(S, c) = g(f(S), V(N_S, n_i)) - c$, where $g(x, y)$ is a function that is increasing in both x and y.

We give several specific collaborative learning problems that align with this framework:

Mean estimation: For Example 8.4.1, we have $U_i(S, c_i) = f(S) + V(N_S, n_i, c_i)$ with $f(S) = \frac{1}{-N_S} - \left(\frac{\sum_{j\in S} n_j^2}{N_S^2} - 3\right)\sigma_\theta^2$ and $V(N_S, n_i, c_i) = \frac{1}{n_i} + \left(\frac{2n_i^2}{N_S^2} - \frac{4n_i}{N_S}\right)\sigma_\theta^2 - c(N_S, n_i)$.

Learning with shared representation: For Example 8.4.2, suppose each client incurs a fixed enrollment cost c if they choose to join (e.g. they pay an enrollment fee to the server to get the shared support), then we have $U_i(S, c) = g(f(S), V(n_i)) - c = g\left(\delta, \frac{1}{\sqrt{n_i}}\right) - c$ where the probability that the server fails to learn the shared support δ is a function of $\{n_j\}_{j\in S}$. We can show that $g\left(\delta, \frac{1}{\sqrt{n_i}}\right)$ is increasing in $1/\sqrt{n_i}$ for any $\delta \in [0, 1]$.

A client decides whether to join or not each time step as described in Section 3.2. Let $S^{(t)}$ be the coalition (the set of clients who decide to join) on time step t. Then on time step $t+1$, a client i will join if and only if

$$U_i(S^{(t)}, c_i) = g(f(S^{(t)}), V(N_{S^{(t)}}, n_i, c_i)) \geq 0.$$

8.4.2 Realization mapping based on inferred coalitions

When the clients' utilities depend not only on N_S but also on the coalition S, we cannot use the characterization in the previous section. The straightforward realization mapping $h(\cdot)$ would be a mapping from a subset of clients to another subset, and there are exponentially many possible subsets. In this section, we show that, surprisingly, we can still predict the equilibria and the dynamic by a simple realization mapping $\tilde{h} : \mathbb{Z}^+ \to \mathbb{Z}^+$ that takes in an expected total number of data points and outputs a "guess" about the outcome.

The function $\tilde{h}(\cdot)$ is defined as follows. For an integer K, we define an "inferred coalition" $IC(K)$ as our guess about the coalition S when we only know that the total number of data points N_S is equal to K. When $N_S = K$, client i's utility is equal to

$$U_i(S, c_i) = g(f(S), V(K, n_i, c_i)).$$

Since $f(S)$ is shared among all the clients, the clients with the highest $V(K, n_i, c_i)$ are most willing to join. Therefore we can guess S based on K as follows: we first order the clients in decreasing order of $V(K, n_i, c_i)$, and we define $IC(K) = \{1, \ldots, z\}$ as the first z clients in the ordering who hold at least K data points in total, $z = \arg\min_x \sum_{j=1}^{x} n_j \geq K$. Now assume that $IC(K)$ is the actual coalition at a particular time step. Then we can further infer the set of clients who will join the next time step: we define "inferred outcome" $IO(K)$ to be the set of clients who will join the next time step assuming that $S = IC(K)$ and $N_S = K$,

$$IO(K) = \{i : g(f(IC(K)), V(K, n_i, c_i)) \geq 0\},$$

and we define $\tilde{h}(K)$ as the total number of data points provided by $IO(K)$,

$$\tilde{h}(K) = \sum_i n_i \cdot \mathbf{1}\left[g(f(IC(K)), V(K, n_i, c_i)) \geq 0\right].$$

8.4.3 Characterizing equilibria and dynamic based on the realization mapping

In this section, we characterize the equilibria and the dynamic by the realization mapping based on inferred coalition mapping function $\tilde{h} : \mathbb{Z}^+ \to \mathbb{Z}^+$. We show that (1) the equilibria (Definition 8.2.1) of the dynamic correspond to the self-fulfilling equilibria defined by $\tilde{h}(\cdot)$, *i.e.,* , crossing points K with $\tilde{h}(K) = K$; and (2) we will still observe cascading behaviors in the region $\{K : \tilde{h}(K) > K\}$.

Our first observation is that there is a mapping from the Nash equilibria (Definition 8.2.1) to the self-fulfilling equilibria defined by $\tilde{h}(\cdot)$.

Theorem 8.4.3. *For any equilibrium S, we have $\tilde{h}(N_S) = N_S$; for any K with $\tilde{h}(K) = K$, there exists an S with $N_S = K$ that is an equilibrium.*

Characterizing the dynamics becomes more intricate in this context. Firstly, unlike in the previous section where convergence to equilibrium is guaranteed (as demonstrated in Lemma 8.3.1), the dynamic $S^{(1)}, S^{(2)}, \ldots$ may not always converge and can potentially become trapped in a loop. Secondly, our realization mapping $\tilde{h}(K)$ does not provide a perfect prediction for the next time step's outcome after receiving K data points. It is possible for the inferred coalition to deviate from the actual coalition, leading to a realized outcome that differs from $\tilde{h}(K)$. However, we demonstrate that we can still observe "cascading behaviors" within the region $K : \tilde{h}(K) > K$.

We first prove that the realization mapping function $\tilde{h}(\cdot)$ can serve as a lower bound of future dataset size in the region of $\tilde{h}(K) \geq K$. For simplicity, in the rest of the section, we write $N_t = N_{S^{(t)}}$.

Lemma 8.4.4. *For any $t, S^{(t)}, S^{(t+1)}$, if the number of data points N_t falls in the region $\{K : \tilde{h}(K) \geq K\}$ and on time step $t+1$, we have $N_{t+1} \geq N_t$, then we must have $N_{t+2} \geq \tilde{h}(N_t)$ and the number of data points will never drop below $\tilde{h}(N_t)$ after time step $t+2$, that is, $N_\tau \geq \tilde{h}(N_t)$ for all $\tau \geq t+2$.*

Based on the lower bound, we prove the cascading behaviors as follows.

Theorem 8.4.5 (Cascading behaviors)**.** *Suppose on a time step t, the number of data points N_t is in the region $\{K : \tilde{h}(K) > K\}$ and on time step $t+1$, we have $N_{t+1} \geq N_t$, then within $(\sum_{j=1}^{M} n_j - \tilde{h}(N_t))^2$ time steps, we will have the number of data points in the region $\{K : \tilde{h}(K) \leq K\}$ and strictly higher than N_t.*

8.5 Payment Schemes Aided by Network Effects

In this section, we use our previous characterization in Section 8.4 to design cost-effective payment schemes that achieve and maintain a desired coalition size. Our approach leverages network effects by targeting clients who are most willing to join and providing them with the necessary payments to trigger a chain reaction of more clients joining. This helps us save costs while reaching and sustaining the desired coalition size.

Suppose the designer can give payments to the clients to encourage participation. If the designer promises a payment $P^{(t+1)}(i)$ to client i if she joins on time step $t+1$, then client i will join on time step $t+1$ if and only if $U_i(S^{(t)}, c_i) + P^{(t+1)}(i) \geq 0$.

The objective of the designer is to establish a stable coalition with a minimum of N data points, while minimizing the overall payment required. A straightforward approach would involve selecting clients with the lowest costs and fully covering their expenses until the coalition accumulates at least N data points. In this case, the total payment required can be expressed as $\sum_{i:c_i \leq W} c_i$, where W is the minimum value that guarantees $\sum_{i:c_i \leq W} n_i \geq N$. However, not all these payments are necessary, as clients also benefit from the data contributed by others, and the utility gain increases with a larger coalition.

We thus propose a payment scheme that leverages network effects and gradually acquires data by recruiting a small number of clients in each round. The participation of these clients increases the willingness of other clients to join, enabling us to make lower payments, and so forth. Specifically, our payment scheme follows an iterative process, targeting the clients most inclined to join (those within the inferred coalitions) and providing the necessary payment to initiate cascading behaviors.

Algorithm 9: Network effects aided payment scheme

Input: Clients $1, \cdots, M$ with personal cost $c_1, \cdots, c_M$ and number of local samples $n_1, \cdots, n_M$, respectively; target number of data points N; function $f(\cdot), V(\cdot)$.

$S = \emptyset,\ LB = 0$

while $LB < N$ **do**

(1) Let N_S be the current number of samples in the coalition.

(2) Let K be the smallest number with $K > N_S$ and $\tilde{h}(K) \geq K$, which represents the closest "tipping point" that will trigger the cascading behaviors.

(3) Target clients in the inferred coalition $IC(K)$ who have not joined yet and promise them payments so that they will join on the next time step; that is, for a client $i \in IC(K)$ but $i \notin S$, promise a payment of $-U_i(S, c_i)$ if i joins the next time step.

(4) Wait one time step and we have $S \leftarrow S \cup IC(K)$; give the promised payments.

(5) While $\tilde{h}(N_S) \leq N_S$, let more clients join from cascading effect and update $LB \leftarrow \max\{LB, \tilde{h}(N_S)\}$ and $S \leftarrow \{i : U_i(S, c_i) \geq 0\}$.

end while

Theorem 8.5.1. *Algorithm 9 will end. Suppose Algorithm 9 ends on time step t, then the number of data points will never drop below N afterward, i.e., we have $N_{S(\tau)} \geq N$ for all $\tau \geq t$.*

Figure 8.1 shows two possible trajectories of Algorithm 9.

8.6 Experiment

In this section, we present numerical experiments to evaluate the payment scheme aided by network effects described in Section 8.5. When the initial coalition is empty, the server can give payment to some clients to offset their costs and encourage participation. We focus on showing that on average, the payment scheme in Algorithm 9 leads to a lower total payment amount compared to naively paying the full cost of all clients for the global mean estimation problem.

Experimental Description We consider the setting of 20 clients learning a shared global mean parameter. For each client i, we let their local sample size n_i be randomly drawn from a uniform distribution between $[1, 5]$. We consider two cost distributions: (1) each client i has their personal cost c_i drawn from a uniform distribution between $[0, 0.5]$ and (2) the personal cost c_i is drawn from a truncated Gaussian distribution between $[0, 0.5]$ with mean 0.3 and standard deviation 0.1. We assume that at the beginning, all clients share an expectation that there are no clients in the coalition. The goal of the server is to have all clients join the coalition. The server can employ a naive payment method where it pays the entire cost of all clients in the population. Alternatively, the server can follow the payment scheme aided by network effects in Algorithm 9 until all clients have joined the coalition. For each cost distribution, we run the simulation for 20 rounds and calculate the total payment by using the naive method and using Algorithm 9. The results in Figure 8.2 are averaged over 10 random runs and the error bars represent one standard deviation. Our experiments are run on a 2020 MacBook Air laptop with 16GB of RAM.

Results. In Figure 8.2, we plot the total payment by the naive method and Algorithm 9 on two different cost distributions over time. We observe that when the personal costs c_i are drawn from a uniform distribution, the total payment using Algorithm 9 is much smaller than that of the naive method. On the other hand, when the personal costs c_i are drawn from the truncated Gaussian distribution, there is little difference between the total payments of the two methods. This result is potentially due to the lack of cascading effect in line (5) of Algorithm 9: given a realization mapping $\tilde{h}$, if only a few clients join "for free" at each time step, then the savings on total payment with network effects is negligible.

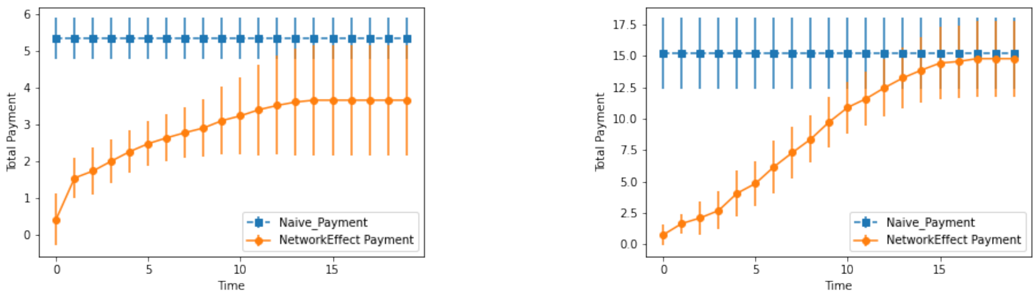


Figure 8.2: Global mean estimation for 20 clients with different local sample sizes and uniform cost between $[0, 0.5]$ (left) and truncated Gaussian cost with mean 0.3 and standard deviation 0.1 (right). The payments are averaged over 10 random runs and the error bars represent one standard deviation.

8.7 Conclusion, Limitations, and Future Work

We have initiated the study of clients' behaviors in collaborative learning under a network effects game model, where each client's participation decision changes as the network grows over time. We showed that whenever the entire population shares an expectation on the coalition size, we can characterize the dynamics of the client's participation game for some classes of learning problems. We identified self-fulfilling expectation equilibrium points, where the coalition after all clients have computed the best-response strategy matches the shared expectation. Moreover, we showed that without explicit intervention from the server, the client's participation dynamic naturally converges to self-fulfilling expectation equilibria. Based on this characterization, we proposed a cost-efficient payment scheme that incentivizes clients so that the final coalition reaches a desirable equilibrium. The goal of our payment mechanism is to kick-start the participation dynamic while minimizing the total amount of payment sent by the server. We are making strict positive improvements on the total population welfare over the setting where there is no payment and there is no client who wants to voluntarily join the coalition.

Our framework offers insights into clients' participation dynamics in FL. A limitation of our work is that we did not consider more realistic learning problems. However, we believe that our work provides a meaningful contribution to the study of incentives in FL, and hope that continuing to understand client participation dynamics can lead to the design of explicit incentive mechanisms that induce desirable final outcomes in practical FL applications.

Chapter 9

Reconciling Model Multiplicity for Downstream Decision Making

9.1 Introduction

In many applications, individual probability prediction is at the heart of a decision-making process. For example, in the Job Training Partnership Act (JTPA) training program (Bloom et al., 1997b), a decision-maker may want to predict whether an individual is employed or not before assigning them to training; or in medical trials, a doctor wants to predict the probability that the patient has contracted a disease before recommending them a treatment. Since the hospital does not know the true individual probability that a particular patient is ill, they can only evaluate the individual probability predictions through its average outcome over a sufficiently large sample set. For a predictive task, the standard convention is to choose the model that maximizes *accuracy*. However, previous work has shown that it is common to have multiple predictive models with similar accuracy but substantially different properties (Chen et al., 2018b; Rodolfa et al., 2020; D'Amour et al., 2022). This phenomenon is called *predictive* (or model) multiplicity, a line of work studied by Breiman (2001); Marx et al. (2020); Black et al. (2022).

In a predictive multiplicity scenario, the decision-maker may have two or more predictive models that are nearly equivalent in terms of accuracy but disagree on their predictions on many individual samples. In our motivating example (Figure 9.1), the hospital has access to two models f_1 and f_2 predicting the probability of disease which are equally accurate on average over the entire population in terms of squared loss, but their predictions on a subpopulation may vastly differ. This disagreement in outcome prediction may have a disparate impact on the subpopulation if the hospital has to choose one predictor over the other to make important downstream decisions. For example, they might select a treatment based on the predicted probability that a patient has contracted a disease. Formally, given a predictive model f and a decision-making loss function $\ell(y, \cdot)$, the decision-maker wants to choose a best-response action, *i.e.,* the action a that minimizes $\mathbb{E}_{y \sim f}[\ell(', a)]$. When two models f_1 and f_2 have nearly equivalent accuracy but lead to different best-response actions, the decision-maker would not be able to identify which best-response action to take for individual patients. While predictive multiplicity offers great flexibility for the decision-maker in the model selection process, it also places an additional burden on the decision-maker to correctly navigate such freedom and justify how they use a predictive model to make downstream decisions.

Roth et al. (2023) attempts to address the model multiplicity issue by resolving prediction disagreement between models. Adapting techniques from the literature of multi-calibration (Hebert-Johnson et al., 2018), they provide a procedure called "Reconcile" that updates the predictive models to minimize their disagreements and improve the accuracy of each model. However, we show simple settings where the reconciled predictions from Roth et al. (2023) can lead the downstream decision-makers to take actions with substantially higher losses. We visually demonstrate this scenario in Figure 9.1. For a more detailed discussion on the limitation of prior work, see Section 9.2.4 and Section F.I. This motivates the study of how to reconcile predictive multiplicity with an explicit focus on its impact on downstream decisions.

In this work, our goal is to leverage tools from multi-calibration (Hebert-Johnson et al., 2018) to alleviate the model multiplicity issue in high-dimensional decision-making tasks for multiple decision-makers with multiple decision-making loss functions. Specifically, we show how the decision-maker can update the input predictors so they approximately agree on (1) individual predictions and (2) best-response actions for each individual in the downstream decision-making task. Our procedure ensures that the number of disagreements in best-response actions decreases over time, which enables the decision-makers to confidently use either of the updated predictors to justify their decisions.

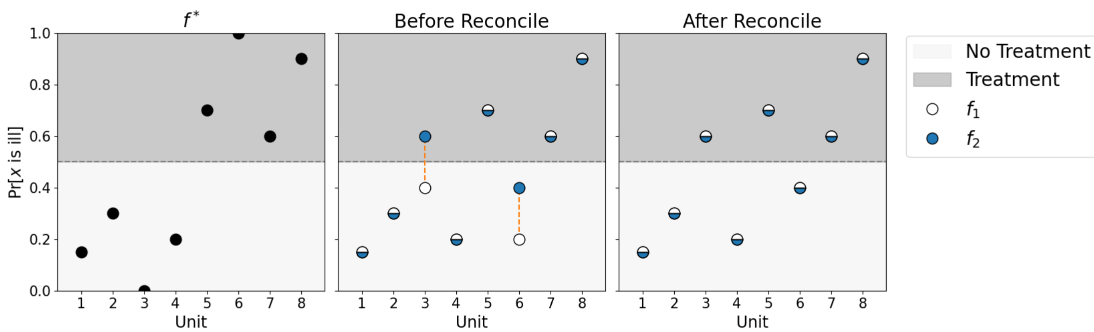


Figure 9.1: An illustrative example of the drawback in a prior work's attempt at addressing model multiplicity. Consider a stylized binary classification problem on a dataset with 8 units (patients) and the hospital deciding between two actions (treatment vs. no treatment). Treatment is assigned if the predicted probability is above 1/2. **Left**: The true probability that each patient is labeled 'ill'. **Middle**: The predicted probability that each patient is ill according to f_1 (white) and f_2 (blue). While these two predictors have almost the same accuracy (in squared loss), their individual probability predictions for patients 3 and 6 vastly differ. **Right**: After running the Reconcile procedure of Roth et al. (2023), the individual probability predictions agree everywhere. However, the best-response action of unit 3 changed from correct (no treatment) to incorrect (treatment). If the hospital uses the updated f_1 to make their treatment recommendation, they would incur more loss than before had they not updated the predictor using Reconcile. This example is formalized in Theorem 9.2.7.

Overview of Paper. We study the problem of reconciling model multiplicity for multiple downstream decision-making tasks, where the decision-makers have multiple predictive models with nearly equivalent squared loss but may lead to vastly different best-response actions for a significant number of individuals in the population. Our key contributions are summarized as follows.

- In Section 9.2, we formulate the problem of model multiplicity from the perspective of the decision-makers. In Section 9.2.4, we formalize our motivating example in Figure 9.1 and show that it is insufficient to only update two predictive models so that they have improved squared loss and nearly agree on their individual predictions almost everywhere.

- In Section 9.3, we introduce an algorithm, ReDCal, that outputs predictive models that are (1) calibrated to a finite set of downstream decision-making tasks and (2) approximately agree on their predictions and best-response actions almost everywhere for each downstream task.

- In Section 9.3.2, we extend our analysis to the setting where one does not have direct access to the true distribution and instead only has a validation dataset with samples drawn i.i.d from the underlying distribution. We show that the guarantees obtained using the empirical distribution can be translated to the unknown underlying distribution.

- In Section 9.3.3, we extend our analysis to the setting where one has $k > 2$ predictors that need to be reconciled. We provide an efficient "contestation" algorithm that only requires approximately k times more samples than the base model that reconciles 2 predictors. In Section F.V, we further generalize our result to the setting where the action space of the downstream decision-making tasks is infinitely large.

- Finally, in Section 9.4, we empirically evaluate the performance of the proposed algorithm on real-world datasets and show our improvement over the benchmark prior work in resolving disagreement in downstream decision-making tasks.

9.2 Problem Formulation

Notation. Throughout this paper, we use subscripts i to index different predictions, superscripts t to index different time-steps, and a to index actions. For $K \in \mathbb{N}$, we use the shorthand $[K] \coloneqq \{1, 2, \cdots, K\}$. $\Delta([k])$ denote the set of possible distributions over $[k]$.

We consider the prediction problem with random variables x and y, where $x \in [k]$ represents the features and $y \in \mathcal{Y}$ represents the labels. We focus on the regression problem in which the label domain is real-valued and bounded: $\mathcal{Y} \subset [0, 1]^d$. Our formulation also permits the multi-class classification problem by writing the label y's as one-hot vectors.

We denote $\mathcal{D} \in \Delta([k] \times \mathcal{Y})$ as the true distribution over the pairs of features-label (x, y). In practice, we will not have access to $\mathcal{D}$, and instead only know a set of n data points D sampled i.i.d from $\mathcal{D}$. In such case, we consider the dataset $D = \{(x_1, y_1), \cdots, (x_n, y_n)\}$

to be the *empirical distribution* over D, which is a discrete distribution that place uniform weight $1/n$ on each sample $(x, y) \in D$.

A predictor is a map $f : [k] \to [0,1]^d$. Our goal is to find the Bayes optimal predictor $f^* : [k] \to [0,1]^d$ such that for all $x \in [k]$, $f^*(x) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[y|x]$ is the *conditional label expectation given* x.

9.2.1 Model Evaluation

Given a predictor $f \in [0,1]^d$, we evaluate f via its squared error, *i.e.,* its expected deviation from the true label. We formalize this objective in the following definition.

Definition 9.2.1 (Brier Score)**.** *The squared error (also known as Brier score) of a predictor f evaluated on distribution $\mathcal{D}$ is given as:*

$$B(f, \mathcal{D}) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\|f(x) - y\|_2^2]$$

When we only have a dataset $D = \{(x_1, y_1), \cdots, (x_n, y_n)\}$, the empirical Brier score is given as:

$$B(f, D) = \frac{1}{n}\sum_{i=1}^{n} \|f(x_i) - y_i\|_2^2$$

Note that we use the Brier score as our metric because it can be accurately estimated given access to only the samples from the distribution. Moreover, among all possible predictors, the Brier score is minimized by the Bayes optimal predictor f^*.

Lemma 9.2.2. *Fix any distribution $\mathcal{D}$ and let $f^*(x) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[y|x]$ represent the true conditional label encoded by $\mathcal{D}$. Let $f : [k] \to [0,1]^d$ be any other model. Then we have $B(f^*, \mathcal{D}) \le B(f, \mathcal{D})$.*

Hence, given two predictors f_1 and f_2, if we can verify empirically from the observable data that $B(f_1, \mathcal{D}) \le B(f_2, \mathcal{D})$, then we can empirically falsify that f_2 encodes the true conditional label.

9.2.2 Downstream Decision-Making Tasks and Loss Functions

Beyond our initial goal of finding a good estimate for the true conditional predictor f^*, we are also interested in using our predictors for downstream decision-making problems.

Formally, we consider a loss minimization problem, where the decision-maker has a set of possible actions $\mathcal{A}$ and a loss function $\ell : \mathcal{Y} \times \mathcal{A} \to [0, d]$. Wlog, we only consider finite action set $\mathcal{A} = [A]$, *i.e.,* there are A possible actions[1]. In this paper, we assume that the loss function does not directly depend on the features x and is linear in y. That is, for each action $a \in \mathcal{A}$, there exists some $\ell_a \in [0, 1]^d$ such that

$$\ell(y, a) = \langle y, \ell_a \rangle \tag{9.1}$$

We write $\mathcal{L} = \{\ell : \mathcal{Y} \times \mathcal{A} \to [0, d]\}$ to denote a finite family of loss functions. We consider the setting with multiple different decision-makers, each using a different linear loss function in $\mathcal{L}$. For any loss function $\ell : \mathcal{Y} \times \mathcal{A} \to \mathbb{R}$, we can rescale each coordinate of ℓ_a to be between $[0, 1]$.

Given a predictor f and a loss function ℓ, the decision-maker selects an action $a \in \mathcal{A}$ that minimizes the expected loss. We define the best-response policy taken by the decision-maker as follows.

Definition 9.2.3 (Best-response policy)**.** *Given a loss function ℓ, a predictor f and the action set $\mathcal{A}$, the best-response policy for ℓ is given as*

$$\pi_\ell^{\mathsf{BR}}(f(x)) = \operatorname*{argmin}_{a \in \mathcal{A}} \langle f(x), \ell_a \rangle.$$

9.2.3 Calibration

In our setting, we consider the decision-maker only having access to some pre-trained predictors f given by a third-party. For instance, a data scientist trained a pair of models on an image dataset without exact knowledge of how the downstream decision-maker will use such predictors. We may imagine the decision-maker as a hospital considering whether to recommend treatment to certain patients based on the predicted probability that the patient has contracted a skin disease. Since the hospital's treatment-recommendation algorithm is not known to the public (and the data scientist), we assume that the data scientists initially aim to minimize the squared error in their predictions.

Since the hospital believes that the input predictors may not perform well according

[1]Our technical results can be generalized to a setting with infinitely many actions, *i.e.,* $|\mathcal{A}| = \infty$. For details, see Section F.V.

to their own loss function, they want the data scientist to convey trust through other performance guarantees of the predictors. One such guarantee is multi-calibration with respect to a finite set of loss functions $\mathcal{L} \ni \ell$ and a set of events $\mathcal{E}$ on the best-response policy, *i.e.,* if the loss function ℓ belongs to $\mathcal{L}$, the decision-maker should be able to accurately compute the expected loss of choosing an action using the best-response policy π_ℓ^{BR}. Formally, we let $E_{\ell,a}(f(x), x)$ denote the action selection events:

Definition 9.2.4 (Best-response Events)**.** *Given a predictor f, let $\mathcal{E}$ be the set of best-response events, where each event $E_{\ell,a} \in \mathcal{E}$ for some loss function ℓ and action $a \in \mathcal{A}$ is defined as*

$$E_{\ell,a}(f(x), x) = \mathbf{1}\{x : \pi_\ell^{\mathsf{BR}}(f(x)) = a\}$$

Given a set of events $\mathcal{E}$, we can define an approximate notion of multi-calibration with respect to $\mathcal{E}$.

Definition 9.2.5 (β-approximate decision calibration)**.** *A predictor f is $\beta-$decision calibrated with respect to the set of best-response events $\mathcal{E}$ if for all $E_{\ell,a} \in \mathcal{E}$, we have:*

$$\left\| \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(y - f(x)) \cdot E_{\ell,a}(f(x), x)] \right\|_2 \leq \beta.$$

This definition follows from an equivalent definition of decision calibration in Zhao et al. (2021). The main difference is we define calibration with respect to a set of events on the best-response actions following the formulation of multi-calibration for online learning in Noarov et al. (2023) and a generalization of multi-calibration in Deng et al. (2023). This definition implies that if a predictor f is $\beta-$decision calibrated with respect to the best-response events $\mathcal{E}$, then the decision-maker can accurately estimate the expected loss from using f to make decisions.

Lemma 9.2.6 ((Zhao et al., 2021))**.** *For all $a, a' \in \mathcal{A}, \ell \in \mathcal{L}$, if f is β-decision-calibrated with respect to the best-response events $\mathcal{E}$, then the loss estimation satisfies*

$$\left| \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell(y, a') \cdot E_{\ell,a}(f(x), x)] - \mathop{\mathbb{E}}_{x\sim\mathcal{D}_{[k]}}[\langle f(x), \ell_{a'} \rangle \cdot E_{\ell,a}(f(x), x)] \right| \leq \beta\sqrt{d}$$

9.2.4 Limitations of Prior Works

In this section, we show that, for the predictive model multiplicity problem under decision-making, improving the accuracy until the two predictors agree on their predictions almost everywhere is not a sufficient solution to our problem. In our analysis below, we consider a stylized problem with $\mathcal{Y} = \{0, 1\}$ and $\mathcal{A} = \{0, 1\}$, *i.e.,* binary class and binary action space. A predictor here is $f : [k] \to [0, 1]$, and the optimal predictor is $f^*(x) = \Pr_{(x',y')\sim\mathcal{D}}[y' = 1|x' = x]$. As a shorthand, we denote $f_1(1)$ as the probability of unit 1 being labelled 1. The loss is defined as

$$\ell(0,0) = \ell(1,1) = 0, \quad \ell(1,0) = \ell(0,1) = 1, \tag{9.2}$$

That is, for any x, the best-response policy is to take action 0 if $f(x) \leq 1/2$ and action 1 otherwise.

Reconcile individual predictions. Prior work by Roth et al. (2023) considers the model multiplicity problem for individual probability predictions. Their proposed algorithm, Reconcile (Line 1), returns a pair of predictors that has a smaller Brier score than the input predictors and approximately agree on their predictions on almost all units. In the following theorem, we show that the best-response policy induced by the predictors updated by Reconcile might lead to a higher expected loss than the ones they started with.

Theorem 9.2.7. *For any $\alpha \in (0, 1/3)$, $\eta \in (0, 1)$, there exists a pair of predictors f_1, f_2 such that after running Line 1 (Reconcile), the output models f_1^T, f_2^T satisfy $f_1^T(x) = f_2^T(x), \forall x \in [k]$, but there exists a loss function ℓ such that at least one of f_1^T, f_2^T induce worse losses compared to the original models f_1, f_2. That is, for some $i \in [2]$,*

$$E_{(x,y)\sim\mathcal{D}}[\ell(y, \pi^{\mathsf{BR}}(f_i^T(x))) - \ell(y, \pi^{\mathsf{BR}}(f_i(x)))] = 1/2 > 0 \tag{9.3}$$

Proof. Consider a setting with $[k] = [2]$ and $\Pr[x = 1] = \Pr[x = 2] = 0.5$. For any

$0 < \alpha < 1/3$, let $\phi \geq \alpha$, we consider the two predictors f_1, f_2 defined as follows:

$$f_1(1) = 1/2 - \phi/2, \quad f_1(2) = 1/2 - 3\phi/2, \quad f_2(1) = 1/2 + \phi/2, \quad f_2(2) = 1/2 - \phi/2 \tag{9.4}$$

and the true probability of each unit being labeled 1 are $f^*(1) = 0$ and $f^*(2) = 1$.

The Brier scores of f_1 and f_2 differ only by ϕ^2, but their individual predictions differ for both features $x = 1$ and $x = 2$. We can run Line 1 and patch f_1 to get the updated model f_1^T with

$$f_1^T(1) = 1/2 + \phi/2 = f_2(1), \quad f_1^T(2) = 1/2 - \phi/2 = f_2(2).$$

Consider the loss function ℓ defined in equation 9.2. The change in expected loss after patching f_1 is

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell(y, \pi_\ell^{\mathsf{BR}}(f_1^T(x))) - \ell(y, \pi_\ell^{\mathsf{BR}}(f_1(x)))] = 1/2 > 0.$$

Therefore, no matter how small we let α and η be, the loss of predictor f_1 increases by a constant amount after running Line 1. $\square$

Moreover, we provide a counterexample to show that it is insufficient to only ensure each individual predictor is approximately decision-calibrated using Line 1. For a detailed description of this counterexample and the theoretical guarantees of Line 1, see Section F.I.

9.3 Reconcile for Decision Making

Suppose we are given two predictors $f_1, f_2 : [k] \to [0,1]^d$. We consider the model multiplicity problem with respect to downstream decision-making – where f_1, f_2 with nearly equivalent accuracy in terms of Brier score differ in their induced decision-making policies, but we cannot falsify either of the two from the data. Informally, our goal is to return a pair of models f_1', f_2' such that: (1) for both $i \in \{1, 2\}$, f_i' is more accurate than f_i in terms of Brier score; (2) for both $i \in \{1, 2\}$, the best-response policy induced by f_i' has no larger expected loss than that of f_i; (3) f_1' and f_2' approximately agree almost everywhere, indicating limited room for additional improvement.

To this end, we are interested in the region where the two predictors disagree substantially with respect to the downstream decision-making task. We define the disagreement region as follows:

Definition 9.3.1 (Disagreement Event)**.** *For f_1, f_2, margin $\alpha > 0$, and a loss function ℓ, the disagreement event is defined for a pair of best-response actions $a_1, a_2 \in \mathcal{A}$ where $a_1 \neq a_2$ as*

$$E^{\alpha}_{\ell,a_1,a_2}(f_1(x), f_2(x), x) = \mathbb{I}\Big[x \in \big\{x : \pi^{\mathsf{BR}}_{\ell}(f_1(x)) = a_1, \pi^{\mathsf{BR}}_{\ell}(f_2(x)) = a_2, \\ \langle f_1(x), \ell_{a_2} - \ell_{a_1}\rangle > \alpha \text{ or } \langle f_2(x), \ell_{a_1} - \ell_{a_2}\rangle > \alpha\big\}\Big],$$

As shorthand, we denote $E_{\ell,a_1,a_2}(x) = E^{\alpha}_{\ell,a_1,a_2}(f_1(x), f_2(x), x)$ when the predictors f_1, f_2 and the margin α are clear from context. For a finite family of loss functions $\mathcal{L}$, we can always iterate through $\mathcal{L}$ to identify the tuple (ℓ, a_1, a_2) that defines a disagreement region between f_1 and f_2.

We say the two models approximately agree with each other when the size of the disagreement event is small enough, *i.e.,* its probability mass $\mu(E_{\ell,a_1,a_2})$ on the underlying distribution $\mathcal{D}$ is small.

9.3.1 The Reconcile Procedure

In this section, we propose our main algorithm ReDCal (Line 1). Whenever the decision-maker observes a large disagreement event E_{ℓ,a_1,a_2}, the best-response action and its corresponding expected loss given by at least one of the predictors must be incorrect. For example, at time step t and unit x, if the gap between the losses of taking a_1 and a_2 according to f_2 is substantially different from the loss gap observed on the data, then the decision-maker can induce that f_2 must have been wrong in its prediction for x. Then, the decision-maker would want to 'patch' f_2 in this time-step.

The calibration procedure within each time-step is divided into two stages. In the first stage, we update model f_2 to f_2' by minimizing the mean prediction error on the disagreement event, *i.e.,* minimizing $\mathbb{E}[\|f_2'(x) - y | E_{\ell,a_1,a_2}(x) = 1\|]$. Following the intuition from multi-calibration, this update operation would improve the Brier score and produce a more accurate predictor. However, the updated model f_2' is not guaranteed to induce the correct best-response action and could instead induce some other actions that might lead to a larger expected loss. To cope with this, in the second stage, we further update f_2' to a model f_2'' that is approximately decision-calibrated within event E_{ℓ,a_1,a_2} using Line 1. Since the loss estimation given by f_2'' is accurate for all best-response events within E_{ℓ,a_1,a_2} and we are taking actions to minimize estimated loss, we can now safely take the best-response action induced by f_2''. The formal description of the algorithm is given by Algorithm 1.

Algorithm 10: Decision Calibration

Input: Predictor f, loss family $\mathcal{L}$, $\beta > 0$, event E

1: Let $f^0 = f$.
2: **while** f^t is not β-MC w.r.t events $E_{\ell,a} \cap E$ for some $\ell \in \mathcal{L}$ **do**
3: Let $\ell^t, a^t = \text{argmax}_{\ell,a} \| \mathbb{E}_{(x,y)\sim\mathcal{D}}[(y - f^t(x))E_{\ell,a}(f^t(x), x)]\|_2$
1 4: Let $\phi^t = \mathbb{E}_{(x,y)\sim\mathcal{D}}[y - f^t(x)|E_{\ell^t,a^t}(f^t(x), x) = 1]$
5: Patch $f^{t+1}(x) = \text{proj}_{[0,1]^d}(f^t(x) + \phi^t E_{\ell^t,a^t}(f^t(x), x))$
6: $t = t + 1$.
7: **end while**

Output: f^t

Algorithm 11: Reconcile Decision Calibration (ReDCal)

Input: $f_1, f_2, \mathcal{L}, \eta > 0, \alpha > 0, \beta > 0$

1: Let $f_1^0 = f_1, f_2^0 = f_2$ and $t = 0$.
2: **while** $\mu(E_{\ell,a_1,a_2}) \geq \eta$ for some $a_1, a_2 \in \mathcal{A}$ and $\ell \in \mathcal{L}$ **do**
3: Let $\ell^t, a_1^t, a_2^t = \text{argmax}_{\ell,a,a'} \mu(E_{\ell,a,a'})$, $E^t = E_{\ell^t,a_1^t,a_2^t}$.
4: Pick

$$i^t = \underset{i\in\{1,2\}}{\text{argmax}} \Big| \underset{(x,y)\sim\mathcal{D}}{\mathbb{E}}[\ell^t(y, a_1^t) - \ell^t(y, a_2^t)|E^t(x) = 1] - \underset{x\sim[k]}{\mathbb{E}}[\ell^t(f_i(x), a_1^t) - \ell^t(f_i(x), a_2^t)|E^t(x) = 1]\Big|.$$

1

5: Denote $f_{i^t}^t$ as f_i^t. Let $\phi^t = \mathbb{E}_{(x,y)\sim\mathcal{D}}[y|E^t(x) = 1] - \mathbb{E}_{x\sim\mathcal{D}_{[k]}}[f_i^t(x)|E^t(x) = 1]$.
6: Patch $f^t(x) = \text{proj}_{[0,1]^d}(f_i^t(x) + \phi^t E^t(x))$.
7: Let $f_i^{t+1} = \text{Decision-Calibration}(f^t, \mathcal{L}, \beta, E^t)$. $t = t + 1$.
8: **end while**

Output: f_1^t, f_2^t

We provide the theoretical guarantees of our proposed algorithm below. At a high level, Line 1 produces a pair of models with improved accuracy and approximately agrees on the best-response action almost everywhere. Unlike prior work in model multiplicity, our theoretical guarantees in this section (and their extensions in later sections) do not require the input predictors f_1, f_2 to have similar accuracy. Using techniques from multi-calibration, our proposed algorithms can reconcile predictors with distinct accuracy, which could be beneficial to downstream decision-makers depending on the context. For the formal proofs of Theorem 9.3.2, see Section F.II.

Theorem 9.3.2. *For any pair of models $f_1, f_2 : [k] \to [0,1]^d$, any distribution $\mathcal{D}$, family of loss functions $\mathcal{L}$, any loss margin $\alpha > 0$, disagreement region mass $\eta > 0$, and decision-calibration tolerance $\beta > 0$, Algorithm 1 updates f_1 and f_2 for T_1 and T_2 time-steps, respectively, and outputs a pair of models (f_1^T, f_2^T), such that:*

1. *Line 1 terminates within $T = T_1 + T_2 \leq \frac{4 \cdot d \cdot (B(f_1,\mathcal{D}) + B(f_2,\mathcal{D}))}{\alpha^2 \eta}$ time-steps.*
2. *The Brier scores of the final models are lower than that of the input models (f_1, f_2):*

$$B(f_1^T, \mathcal{D}) \leq B(f_1, \mathcal{D}) - T_1 \cdot \frac{\alpha^2 \eta}{4d} \quad \text{and} \quad B(f_2^T, \mathcal{D}) \leq B(f_2, \mathcal{D}) - T_2 \cdot \frac{\alpha^2 \eta}{4d}$$

3. *All the downstream decision-making losses of the final models do not increase by much compared to that of the input models (f_1, f_2): for each $i \in \{1, 2\}$ and for all $\ell \in \mathcal{L}$,*

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell(y, \pi_\ell^{\mathsf{BR}}(f_i^T(x))] - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell(y, \pi_\ell^{\mathsf{BR}}(f_i(x))] \leq T_i \beta \sqrt{d} A$$

4. *The final models approximately agree on their best-response actions almost everywhere. That is, the disagreement region E_{ℓ,a_1,a_2} calculated using f_1^T, f_2^T has small mass. $\forall \ell \in \mathcal{L}$,*

$$\mu(E_{\ell,a_1,a_2}) < \eta \quad \text{for all } a_1, a_2 \in \mathcal{A} \quad s.t\ a_1 \neq a_2$$

Remark 9.3.3. In the third result of Theorem 9.3.2, the increase in downstream decision-making loss at each time-step only depends on the decision-calibrate tolerance

β, dimension d, and number of actions A. Since the total number of time-steps does not depend on β, we can set $\beta = \alpha/T\sqrt{d}A$ to ensure the loss of taking the best-response action does not degrade by more than α. Moreover, in Section 4.7, we empirically observe that the loss only increases minimally.

9.3.2 Finite Sample Analysis

In Section 9.3.1, we have presented an algorithm, ReDCal, to reconcile two predictors with similar Brier scores, assuming the decision-makers have direct access to the probability distribution $\mathcal{D}$. In practice, the decision-makers will only have a dataset $D = \{(x_1, y_1), \cdots, (x_n, y_n)\}$ containing n i.i.d samples drawn from $\mathcal{D}$. In this section, we will instead run Line 1 on the empirical distribution over D and show that its guarantees can translate to the underlying distribution $\mathcal{D}$ with high probability. To prevent data leakage, we assume that the dataset D is drawn independently of the predictors f_1 and f_2, *i.e.,* the dataset contains freshly drawn data that was not used to train either of the predictors that we want to reconcile. For the formal proofs, see Section F.III.

At a high level, since the samples in D are independently and identically distributed, we can apply Chernoff-Hoeffding inequality to show that, with high probability, the in-sample quantities are approximately equal to out-sample quantities. We summarize the results in the theorem below.

Theorem 9.3.4. *Fix any distribution $\mathcal{D}$ and dataset D containing n samples drawn i.i.d from $\mathcal{D}$. For any pair of models $f_1, f_2 : [k] \to [0,1]^d$, family of loss functions $\mathcal{L}$, loss margin $\alpha > 0$, disagreement region mass $\eta > 0$, and decision-calibration tolerance $\beta > 0$, Line 1 run over the empirical distribution D updates f_1 and f_2 for T_1 and T_2 time-steps, respectively, and outputs a pair of predictors (f_1^T, f_2^T) such that, with probability at least $1-\delta$ over the randomness of $D \sim \mathcal{D}^n$,*

1. *The total number of time-steps for Line 1 and Line 1 is*

$$T = T_1 + T_2 \leq 2d/\min\{\beta^2, \eta\alpha^2/4d\}\frac{2d}{\min\{\beta^2, \eta\alpha^2/4d\}}$$

2. *For $i \in \{1, 2\}$, the Brier scores of the final models are lower than that of the input*

models:

$$B(f_i^{T_i}, \mathcal{D}) \leq B(f_i, \mathcal{D}) - (T_i - \eta) \cdot \min\left\{\beta^2, \eta\alpha^2/(4d)\right\}$$

3. *For $i \in \{1, 2\}$ and for all $\ell \in \mathcal{L}$, the downstream decision-making losses of the final models do not increase by much compared to that of the input models:*

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell(y, \pi_\ell^{\mathsf{BR}}(f_i^T(x)) - \ell(y, \pi_\ell^{\mathsf{BR}}(f_i(x))] \leq 2T_i\beta\sqrt{d}A$$

4. *The final models approximately agree on their best-response actions almost everywhere. That is, the disagreement region E_{ℓ,a_1,a_2} calculated using f_1^T, f_2^T has small mass: $\forall \ell \in \mathcal{L}$,*

$$\mu(E_{\ell,a_1,a_2}) \leq 2\eta \quad \textit{for all } a_1, a_2 \in \mathcal{A} \quad \textit{s.t } a_1 \neq a_2$$

if $n \geq \Omega\left(d^2/(\eta^2 \min\{\beta, \eta\alpha^2/d\}) \cdot \left(\ln(dA|\mathcal{L}|) - \ln(\delta \min\{\beta, \eta\alpha^2\})\right)\right).$

9.3.3 Extension to Reconciling Multiple Predictors

In the previous sections, we addressed reconciling two predictors for downstream decision-making tasks. When there are $k > 2$ predictors that need to be reconciled, a straightforward approach to generalize our previous analysis is applying ReDCal iteratively until each pair among the k predictors "agrees" on their individual prediction and best-response action. However, since there are $O(k^2)$ pairs of predictors and this approach requires fresh calibration data for each new reconciliation procedure, the data requirement significantly increases. Another approach is to simply apply ReDCal $(k-1)$ times for each new predictor. However, running ReDCal sequentially does not minimize disagreement among all predictors. For instance, initially reconciling f_1 with f_2 minimizes their disagreement, but after f_1 undergoes further reconciliations with other predictors, its updated version f_1^t may still end up disagreeing with f_2 for sufficiently many units. This disagreement indicates a falsification potential and suggests room for further improvements in f_1.

To address this problem, we introduce a "contestation" method. We select a base model, f, and focus updates on events where f disagrees with the other models, f_j. Using

Definition 9.3.1, a disagreement event $E^{\alpha}_{\ell,a_1,a_2}(f(x), f_j(x), x)$ in this setting is identified by a pair of actions (a_1, a_2), a loss function ℓ, and a contestant predictor f_j. At each iteration, we select the disagreement event with the largest size and "patch" the model whose estimated loss on the event has a larger error. The detailed algorithm is written in Line 1 and theoretical guarantees are in Theorem F.IV.1.

To maintain the same guarantee as in Theorem 9.3.4, we need roughly k times more samples for Algorithm 1 compared to ReDCal. Since each iteration decreases the Brier score of one predictor, the cumulative number of necessary updates across k predictors, *i.e.,* the total number of time-steps, is approximately multiplied by $O(k)$. As each set of potential output predictor is identified by a sequence of patching operations, the number of output predictors grows exponentially with k. Using union bounds and Chernoff-Hoeffding inequality suggests that the sample complexity scales up linearly with k. The detailed algorithm, formal theorem, and its proof are in Appendix F.IV.

9.4 Experiments

In this section, we complement our theoretical results with a set of experiments on real-world datasets to show our improvement in decreasing decision-making loss compared to prior work.

9.4.1 Imagenet Multi-class Classification

Experiment Setup. We use the ImageNet dataset (Deng et al., 2009) and two pre-trained models provided by pyTorch (inception-v3 (Szegedy et al., 2015) and resnet50 (He et al., 2015)). Among the 50000 validation samples, we use 40000 samples for calibration and 10000 samples for testing.

We investigate how the downstream decision loss changes with the four calibration algorithms: Reconcile (Line 1), Decision-Calibration (Line 1), ReDCal (Line 1), and the combination of running ReDCal after Decision Calibration as post-process. We run each calibration algorithm 500 times. For each run, we first randomly draw 100 classes from the 1000 classes of ImageNet. Then, we randomly generate a loss function such that, for each $y \in \mathcal{Y}$, $a \in \mathcal{A}$, $\ell(y, a) \sim \mathsf{Normal}(0, 1)$. For each randomly generated loss function ℓ, we compare the expected losses derived from the best-response policies based on predictors f_1 and f_2 against those based on the optimal predictor f^*. Formally, the loss gap at timestep t using predictor f_i is defined as

$$\text{LossGap}(f_i^t) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell(y, \pi_\ell^{\mathsf{BR}}(f_i^t(x))) - \ell(y, \pi_\ell^{\mathsf{BR}}(f^*(x)))].$$

The hyperparameters are chosen as follows: loss margin $\alpha = 0.001$, disagreement region mass $\eta = 0.01$, decision-calibration tolerance $\beta = 0.00001$, and the number of actions $K = 10$.

Results. The results are shown in Figure 9.2. Compared to Reconcile, ReDCal converges within a similar number of time-step and decreases the loss by a larger amount on the test dataset. Moreover, ReDCal further decrease the loss when used as a post-process after Decision-Calibration terminates.

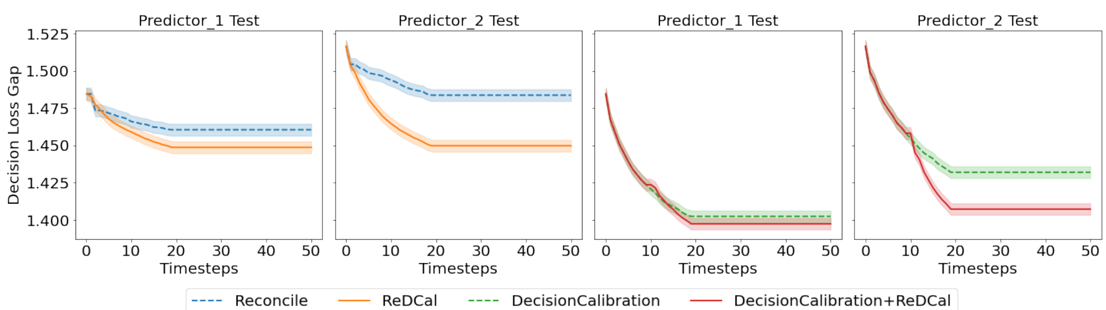


Figure 9.2: ReDCal decreases decision loss on Imagenet. We plot the gap between the average loss of each predictor and the optimal loss had we known the true predictor f^* on the test set. **Left two figures:** We compare the LossGap of ReDCal (orange) with Reconcile (blue). Compared to Reconcile, our algorithm converges at a similar rate and decreases the loss by a larger amount on the test dataset. **Right two figures:** We compare the LossGap of Decision-Calibration (green) to that of Decision-Calibration with a run of ReDCal as post-process (red). Our proposed algorithm can improve upon Decision-Calibration output predictors and further reduce the loss on the test dataset. Results are averaged over 500 runs and the shaded region indicates ± 1 standard errors.

9.4.2 HAM10000 Multi-class Classification

We use the HAM10000 dataset (Tschandl et al., 2018) on pigmented skin lesions to predict the probability that a patient has contracted one of 7 possible skin diseases. The results are shown in Figure 9.3. For a detailed description and discussion of the results of this experiment, see Section F.VI.

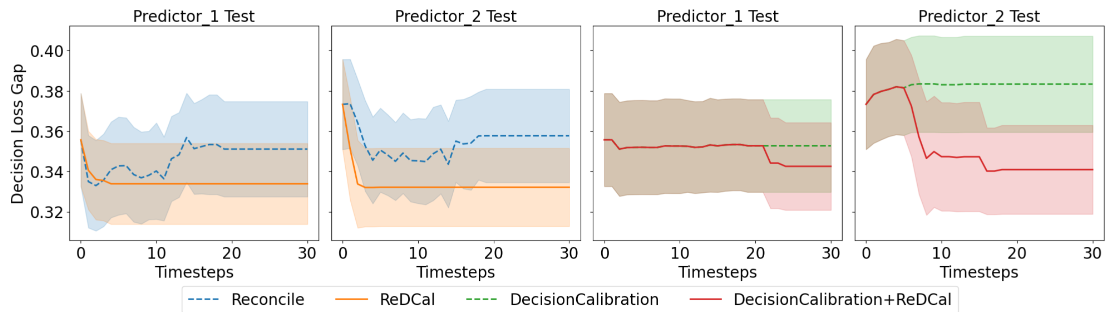


Figure 9.3: ReDCal decreases decision loss on the HAM10000 dataset. The comparisons and results are similar to that of Figure 9.2. Results are averaged over 10 runs and the shaded region indicates ±1 standard errors.

9.5 Conclusion

Predictive multiplicity is a phenomenon in machine learning where the decision-makers have multiple predictors with nearly equivalent squared loss but vastly different individual predictions. Leveraging technique from the multi-calibration literature, we propose an algorithm, ReDCal, that updates a pair of predictors until they approximately agree almost everywhere on (1) individual predictions, (2) best-response actions in the downstream decision-making task, and (3) following the best-response actions incur losses that are close to the optimal loss. We further generalize this result to the settings where one has more than two predictive models and infinitely many possible downstream actions. Our theoretical results help alleviate the problem of predictive multiplicity in model selection. Finally, we provide experiments using real-world datasets to show that our proposed algorithm achieves lower decision loss compared to existing work. While we do not provide examples of domain-specific loss functions as part of our analysis and experiments, we hope that our findings can aid future studies on the impact of model multiplicity in decision-making.

References

Alberto Abadie and Javier Gardeazabal. The economic costs of conflict: A case study of the basque country. *American economic review*, 93(1):113–132, 2003.

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American statistical Association*, 105(490):493–505, 2010.

Anish Agarwal, Devavrat Shah, Dennis Shen, and Dogyoon Song. Model agnostic high-dimensional error-in-variable regression. *CoRR*, abs/1902.10920, 2019. URL `http://arxiv.org/abs/1902.10920`.

Anish Agarwal, Devavrat Shah, and Dennis Shen. On model identification and out-of-sample prediction of principal component regression: Applications to synthetic controls. *arXiv preprint arXiv:2010.14449*, 2020a.

Anish Agarwal, Devavrat Shah, and Dennis Shen. Synthetic interventions. *arXiv preprint arXiv:2006.07691*, 2020b.

Anish Agarwal, Keegan Harris, Justin Whitehouse, and Zhiwei Steven Wu. Adaptive Principal Component Regression with Applications to Panel Data. Papers 2307.01357, arXiv.org, July 2023. URL `https://ideas.repec.org/p/arx/papers/2307.01357.html`.

Shipra Agrawal and Navin Goyal. Near-optimal regret bounds for thompson sampling. *J. of the ACM*, 64(5):30:1–30:24, 2017. Preliminary version in *AISTATS 2013*.

Nabil I. Al-Najjar and Jonathan Weinstein. Comparative testing of experts. *Econometrica*, 76(3):541–559, 2008. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/40056456.

Elaine M. Allensworth and Kallie Clark. High school gpas and act scores as predictors of college completion: Examining assumptions about consistency across high schools. *Educational Researcher*, 49(3):198–211, 2020. doi: 10.3102/0013189X20902110. URL https://doi.org/10.3102/0013189X20902110.

Muhammad Amjad, Devavrat Shah, and Dennis Shen. Robust synthetic control. *The Journal of Machine Learning Research*, 19(1):802–852, 2018.

Muhammad Amjad, Vishal Misra, Devavrat Shah, and Dennis Shen. mrsc: Multi-dimensional robust synthetic control. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 3(2):1–27, 2019.

Joshua Angrist. Instrumental variables methods in experimental criminological research: What, why, and how? Working Paper 314, National Bureau of Economic Research, September 2005. URL http://www.nber.org/papers/t0314.

Joshua Angrist and Alan Krueger. Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives*, 15(4):69–85, 2001.

Joshua D. Angrist and Guido W. Imbens. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American Statistical Association*, 90(430):431–442, 1995. doi: 10.1080/01621459.1995.10476535. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1995.10476535.

Joshua D. Angrist and Jörn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, December 2008. ISBN 0691120358.

Joshua D. Angrist and Jörn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, 2009. ISBN 9780691120348.

Manuel Arellano and Bo Honore. Panel data models: Some recent developments. *Handbook of Econometrics*, 02 2000.

Orley C Ashenfelter and David Card. Using the longitudinal structure of earnings to estimate the effect of training programs, 1984.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 2002. doi: 10.1023/A:1013689704352.

Yahav Bechavod, Katrina Ligett, Zhiwei Steven Wu, and Juba Ziani. Gaming helps! learning from strategic interactions in natural dynamics. In Arindam Banerjee and Kenji Fukumizu, editors, *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pages 1234–1242. PMLR, 2021. URL http://proceedings.mlr.press/v130/bechavod21a.html.

Dennis S Bernstein. *Matrix mathematics: theory, facts, and formulas*. Princeton University Press, 2009.

Marianne Bertrand, Esther Duflo, and Sendhil Mullainathan. How much should we trust differences-in-differences estimates? *The Quarterly journal of economics*, 119(1): 249–275, 2004.

Emily Black, Manish Raghavan, and Solon Barocas. Model multiplicity: Opportunities, concerns, and solutions. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 850–863, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533149. URL https://doi.org/10.1145/3531146.3533149.

Howard S. Bloom, Larry L. Orr, Stephen H. Bell, George Cave, Fred Doolittle, Winston Lin, and Johannes M. Bos. The benefits and costs of jtpa title ii-a programs: Key findings from the national job training partnership act study. *The Journal of Human Resources*, 32(3):549–576, 1997a. ISSN 0022166X. URL http://www.jstor.org/stable/146183.

Howard S. Bloom, Larry L. Orr, Stephen H. Bell, George Cave, Fred Doolittle, Winston Lin, and Johannes M. Bos. The benefits and costs of jtpa title ii-a programs: Key findings from the national job training partnership act study. *The Journal of Human*

Resources, 32(3):549–576, 1997b. ISSN 0022166X. URL http://www.jstor.org/stable/146183.

Avrim Blum, Nika Haghtalab, Ariel D Procaccia, and Mingda Qiao. Collaborative pac learning. In *Advances in Neural Information Processing Systems*, 2017.

Avrim Blum, Nika Haghtalab, Richard L. Phillips, and Han Shao. One for one, or all for all: Equilibria and optimality of collaboration in federated learning. In *International Conference on Machine Learning*, 2021.

Leo Breiman. Statistical modeling: the two cultures. *Statist. Sci.*, 16(3):199–231, 2001. ISSN 0883-4237. doi: 10.1214/ss/1009213726. URL http://dx.doi.org/10.1214/ss/1009213726. With comments and a rejoinder by the author.

Yang Cai, Constantinos Daskalakis, and Christos H. Papadimitriou. Optimum statistical estimation with strategic data sources. In Peter Grünwald, Elad Hazan, and Satyen Kale, editors, *Proceedings of The 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3-6, 2015*, volume 40 of *JMLR Workshop and Conference Proceedings*, pages 280–296. JMLR.org, 2015. URL http://proceedings.mlr.press/v40/Cai15.html.

A Colin Cameron and Pravin K Trivedi. *Microeconometrics: methods and applications*. Cambridge University Press, 2005.

Gary Chamberlain. Panel data. In Z. Griliches† and M. D. Intriligator, editors, *Handbook of Econometrics*, volume 2, chapter 22, pages 1247–1318. Elsevier, 1 edition, 1984. URL https://EconPapers.repec.org/RePEc:eee:ecochp:2-22.

Bangrui Chen, Peter Frazier, and David Kempe. Incentivizing exploration by heterogeneous users. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 798–818. PMLR, 06–09 Jul 2018a. URL http://proceedings.mlr.press/v75/chen18a.html.

Chaofan Chen, Kangcheng Lin, Cynthia Rudin, Yaron Shaposhnik, Sijia Wang, and Tong Wang. An interpretable model with globally consistent explanations for credit risk, 2018b.

Jiafeng Chen. Synthetic control as online linear regression. *Econometrica*, 91(2):465–491, 2023.

Jiawei Chen and Kejun Song. Two-sided matching in the loan market. *International Journal of Industrial Organization*, 31(2):145–152, 2013. ISSN 0167-7187. doi: https://doi.org/10.1016/j.ijindorg.2012.12.002. URL https://www.sciencedirect.com/science/article/pii/S0167718712001245.

Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit: General framework and applications. In *20th Intl. Conf. on Machine Learning (ICML)*, pages 151–159, 2013.

Yiling Chen, Yang Liu, and Chara Podimata. Learning strategy-aware linear classifiers. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15265–15276. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/ae87a54e183c075c494c4d397d126a66-Paper.pdf.

Yae Jee Cho, Divyansh Jhunjhunwala, Tian Li, Virginia Smith, and Gauri Joshi. To federate or not to federate: Incentivizing client participation in federated learning. *arXiv preprint arXiv:2205.14840*, 2022.

Sharon Collard and Elaine Kempson. *Affordable credit: The way forward.* Policy Press, 2005.

Rachel Cummings, Stratis Ioannidis, and Katrina Ligett. Truthful linear regression. In Peter Grünwald, Elad Hazan, and Satyen Kale, editors, *Proceedings of The 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3-6, 2015*, volume 40 of *JMLR Workshop and Conference Proceedings*, pages 448–483. JMLR.org, 2015. URL http://proceedings.mlr.press/v40/Cummings15.html.

Xiaowu Dai, Yuan, Qi, and Michael I. Jordan. Incentive-aware recommender systems in two-sided markets, 2022.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022.

Rémy Degenne and Vianney Perchet. Combinatorial semi-bandit with known covariance. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Zhun Deng, Cynthia Dwork, and Linjun Zhang. Happymap: A generalized multi-calibration method, 2023.

Kate Donahue and Jon Kleinberg. Optimality and stability in federated learning: A game-theoretic approach. In *Advances in Neural Information Processing Systems*, 2021a.

Kate Donahue and Jon Kleinberg. Models of fairness in federated learning, 2021b. URL `https://arxiv.org/abs/2112.00818`.

Kate Donahue and Jon Kleinberg. Model-sharing games: Analyzing federated learning under voluntary participation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 5303–5311, 2021c.

Stephen G Donald and Kevin Lang. Inference with difference-in-differences and other panel data. *The review of Economics and Statistics*, 89(2):221–233, 2007.

Jinshuo Dong, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu. Strategic classification from revealed preferences. In Éva Tardos, Edith Elkind, and Rakesh Vohra, editors, *Proceedings of the 2018 ACM Conference on Economics and Computation, Ithaca, NY, USA, June 18-22, 2018*, pages 55–70. ACM, 2018. doi: 10.1145/3219166.3219193. URL `https://doi.org/10.1145/3219166.3219193`.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2019. URL `http://archive.ics.uci.edu/ml`.

Raaz Dwivedi, Katherine Tian, Sabina Tomkins, Predrag Klasnja, Susan Murphy, and Devavrat Shah. Counterfactual inference for sequential experiments. *arXiv preprint arXiv:2202.06891*, 2022.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. Fairness through awareness. In Shafi Goldwasser, editor, *Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, January 8-10, 2012*, pages 214–226. ACM, 2012. doi: 10.1145/2090236.2090255. URL https://doi.org/10.1145/2090236.2090255.

Cynthia Dwork, Michael P. Kim, Omer Reingold, Guy N. Rothblum, and Gal Yona. Learning from outcomes: Evidence-based rankings. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 106–125, 2019. doi: 10.1109/FOCS.2019.00016.

Cynthia Dwork, Michael P. Kim, Omer Reingold, Guy N. Rothblum, and Gal Yona. Outcome indistinguishability. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2021, page 1095–1108, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380539. doi: 10.1145/3406325.3451064. URL https://doi.org/10.1145/3406325.3451064.

David Easley and Jon Kleinberg. *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge university press, 2010.

EBay, Jun 2022. URL https://partnerhelp.ebay.com/helpcenter/s/article/Incrementality-Testing-by-Location-IP?language=en_US.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research (JMLR)*, 7:1079–1105, 2006.

Vivek Farias, Ciamac Moallemi, Tianyi Peng, and Andrew Zheng. Synthetically controlled bandits. *arXiv preprint arXiv:2202.07079*, 2022.

Yossi Feinberg and Colin Stewart. Testing multiple forecasters. *Econometrica*, 76(3):561–582, 2008. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/40056457.

Dylan J. Foster and Alexander Rakhlin. Beyond ucb: Optimal and efficient contextual bandits with regression oracles, 2020. URL https://arxiv.org/abs/2002.04926.

Peter Frazier, David Kempe, Jon Kleinberg, and Robert Kleinberg. Incentivizing exploration. In *Proceedings of the Fifteenth ACM Conference on Economics and Computation*, EC '14, page 5–22, New York, NY, USA, 2014a. Association for Computing Machinery. ISBN 9781450325653. doi: 10.1145/2600057.2602897. URL https://doi.org/10.1145/2600057.2602897.

Peter Frazier, David Kempe, Jon Kleinberg, and Robert Kleinberg. Incentivizing exploration. In *Proceedings of the Fifteenth ACM Conference on Economics and Computation*, EC '14, page 5–22, New York, NY, USA, 2014b. Association for Computing Machinery. ISBN 9781450325653. doi: 10.1145/2600057.2602897. URL https://doi.org/10.1145/2600057.2602897.

Danilo Freire. Evaluating the effect of homicide prevention strategies in são paulo, brazil: A synthetic control approach. *Latin American Research Review*, 53(2):231–249, 2018.

D. Gale and L. S. Shapley. College admissions and the stability of marriage. *The American Mathematical Monthly*, 69(1):9–15, 1962. ISSN 00029890, 19300972. URL http://www.jstor.org/stable/2312726.

Sumegha Garg, Michael P. Kim, and Omer Reingold. Tracking and improving information in the service of fairness. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, EC '19, page 809–824, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450367929. doi: 10.1145/3328526.3329624. URL https://doi.org/10.1145/3328526.3329624.

Ira Globus-Harris, Varun Gupta, Christopher Jung, Michael Kearns, Jamie Morgenstern, and Aaron Roth. Multicalibrated regression for downstream fairness, 2022a. URL https://arxiv.org/abs/2209.07312.

Ira Globus-Harris, Michael Kearns, and Aaron Roth. An algorithmic framework for bias bounties. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 1106–1124, New York, NY, USA, 2022b. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533172. URL https://doi.org/10.1145/3531146.3533172.

Ira Globus-Harris, Declan Harrison, Michael Kearns, Aaron Roth, and Jessica Sorrell. Multicalibration as boosting for regression, 2023. URL https://arxiv.org/abs/2301.13767.

Ira Globus-Harris, Varun Gupta, Michael Kearns, and Aaron Roth. Model ensembling for constrained optimization, 2024.

Joshua Goodman, Oded Gurantz, and Jonathan Smith. Take two! sat retaking and college enrollment gaps. *American Economic Journal: Economic Policy*, 12(2):115–58, May 2020. doi: 10.1257/pol.20170503. URL https://www.aeaweb.org/articles?id=10.1257/pol.20170503.

Aditya Gopalan, Shie Mannor, and Y. Mansour. Thompson sampling for complex online problems. In *ICML*, 2014.

Parikshit Gopalan, Adam Tauman Kalai, Omer Reingold, Vatsal Sharan, and Udi Wieder. Omnipredictors, 2021. URL https://arxiv.org/abs/2109.05389.

Parikshit Gopalan, Lunjia Hu, Michael P. Kim, Omer Reingold, and Udi Wieder. Loss minimization through the lens of outcome indistinguishability, 2022a. URL https://arxiv.org/abs/2210.08649.

Parikshit Gopalan, Michael P. Kim, Mihir Singhal, and Shengjia Zhao. Low-degree multicalibration, 2022b. URL https://arxiv.org/abs/2203.01255.

Susan W Groth. Honorarium or coercion: use of incentives for participants in clinical research. *The Journal of the New York State Nurses' Association*, 41(1):11, 2010.

András György, Tamás Linder, Gábor Lugosi, and György Ottucsák. The on-line shortest path problem under partial monitoring. *J. of Machine Learning Research (JMLR)*, 8:2369–2403, 2007.

Nika Haghtalab, Michael I Jordan, and Eric Zhao. On-demand sampling: Learning optimally from multiple distributions. In *Advances in Neural Information Processing Systems*, 2022.

Nika Haghtalab, Michael I. Jordan, and Eric Zhao. A unifying perspective on multicalibration: Game dynamics for multi-objective learning, 2023.

Liz Hamel, Ashley Kirzinger, Lunna Lopes, Grace Sparks, Audrey Kearney, Mellisha Stokes, and Mollyann Brodie. Kff covid-19 vaccine monitor: May 2021, 2021. URL https://www.kff.org/coronavirus-covid-19/poll-finding/kff-covid-19-vaccine-monitor-may-2021/.

Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, ITCS '16, page 111–122, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450340571. doi: 10.1145/2840728.2840730. URL https://doi.org/10.1145/2840728.2840730.

Keegan Harris, Hoda Heidari, and Steven Z. Wu. Stateful strategic regression. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 28728–28741, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/f1404c2624fa7f2507ba04fd9dfc5fb1-Abstract.html.

Keegan Harris, Anish Agarwal, Chara Podimata, and Zhiwei Steven Wu. Strategyproof decision-making in panel data settings and beyond. *arXiv preprint arXiv:2211.14236*, 2022.

T. E. Harris. A lower bound for the critical probability in a certain percolation process. *Mathematical Proceedings of the Cambridge Philosophical Society*, 56(1):13–20, 1960.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

Ursula Hebert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (Computationally-identifiable) masses. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages

1939–1948. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/hebert-johnson18a.html.

Hoda Heidari, Vedant Nanda, and Krishna P. Gummadi. On the long-term impact of algorithmic decision policies: Effort unfairness and feature segregation through social learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 2692–2701. PMLR, 2019. URL http://proceedings.mlr.press/v97/heidari19a.html.

Lily Hu, Nicole Immorlica, and Jennifer Wortman Vaughan. The disparate effects of strategic manipulation. In danah boyd and Jamie H. Morgenstern, editors, *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, pages 259–268. ACM, 2019. doi: 10.1145/3287560.3287597. URL https://doi.org/10.1145/3287560.3287597.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks, 2018.

Guido Imbens, Joshua Angrist, and Donald Rubin. Identification of causal effects using instrumental variables. *Journal of Econometrics*, 71(1-2):145–160, 1996.

Nicole Immorlica, Jieming Mao, Aleksandrs Slivkins, and Zhiwei Steven Wu. Bayesian exploration with heterogeneous agents. In *The World Wide Web Conference*, WWW '19, page 751–761, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366748. doi: 10.1145/3308558.3313649. URL https://doi.org/10.1145/3308558.3313649.

Ian T Jolliffe. A note on the use of principal components in regression. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 31(3):300–303, 1982.

Kwang-Sung Jun, Rebecca Willett, Stephen Wright, and Robert Nowak. Bilinear bandits with low-rank structure, 2019.

Christopher Jung, Changhwa Lee, Mallesh M. Pai, Aaron Roth, and Rakesh Vohra. Moment multicalibration for uncertainty estimation, 2020.

Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Batch multivalid conformal prediction, 2022.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1–2):1–210, 2021.

Nathan Kallus. Instrument-armed bandits. *ArXiv*, abs/1705.07377, 2018.

Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6):2590–2615, October 2011. doi: 10.1257/aer.101.6.2590. URL `https://www.aeaweb.org/articles?id=10.1257/aer.101.6.2590`.

Sampath Kannan, Michael Kearns, Jamie Morgenstern, Mallesh Pai, Aaron Roth, Rakesh Vohra, and Zhiwei Steven Wu. Fairness incentives for myopic agents. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, EC '17, page 369–386, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450345279. doi: 10.1145/3033274.3085154. URL `https://doi.org/10.1145/3033274.3085154`.

Sai Praneeth Karimireddy, Wenshuo Guo, and Michael I. Jordan. Mechanisms that incentivize data sharing in federated learning, 2022. URL `https://arxiv.org/abs/2207.04557`.

Maximilian Kasy and Alexander Teytelboym. Matching with semi-bandits. *The Econometrics Journal*, 26(1):45–66, 09 2022. ISSN 1368-4221. doi: 10.1093/ectj/utac021. URL `https://doi.org/10.1093/ectj/utac021`.

Michael L. Katz and Carl Shapiro. Network externalities, competition, and compatibility. *The American Economic Review*, 75(3):424–440, 1985. ISSN 00028282. URL `http://www.jstor.org/stable/1814809`.

Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *23rd Intl. Conf. on Algorithmic Learning Theory (ALT)*, pages 199–213, 2012.

Michael P. Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 247–254, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450363242. doi: 10.1145/3306618.3314287. URL https://doi.org/10.1145/3306618.3314287.

Sungwook Kim. Incentive design and differential privacy based federated learning: A mechanism design perspective. *IEEE Access*, 8:187317–187325, 2020. doi: 10.1109/ACCESS.2020.3030888.

Jon M. Kleinberg and Manish Raghavan. How do classifiers induce agents to invest effort strategically? *ACM Trans. Economics and Comput.*, 8(4):19:1–19:23, 2020. doi: 10.1145/3417742. URL https://doi.org/10.1145/3417742.

Ilan Kremer, Yishay Mansour, and Motty Perry. Implementing the "wisdom of the crowd". In Michael J. Kearns, R. Preston McAfee, and Éva Tardos, editors, *Proceedings of the fourteenth ACM Conference on Electronic Commerce, EC 2013, Philadelphia, PA, USA, June 16-20, 2013*, pages 605–606. ACM, 2013. doi: 10.1145/2492002.2482542. URL https://doi.org/10.1145/2492002.2482542.

Ilan Kremer, Yishay Mansour, and Motty Perry. Implementing the “wisdom of the crowd". *J. of Political Economy*, 122(5):988–1012, 2014a. Preliminary version in *ACM EC 2013*.

Ilan Kremer, Yishay Mansour, and Motty Perry. Implementing the “wisdom of the crowd”. *Journal of Political Economy*, 122(5):988–1012, 2014b. ISSN 00223808, 1537534X. URL http://www.jstor.org/stable/10.1086/676597.

Ilan Kremer, Yishay Mansour, and Motty Perry. Implementing the “wisdom of the crowd”. *Journal of Political Economy*, 122(5):988–1012, 2014c. ISSN 00223808, 1537534X. URL http://www.jstor.org/stable/10.1086/676597.

Matt Kusner and Joshua Loftus. The long road to fairer algorithms. *Nature*, 578:34–36, 02 2020. doi: 10.1038/d41586-020-00274-3.

Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4066–4076, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/a486cd07e4ac3d270571622f4f316ec5\-Abstract.html.

Branislav Kveton, Zheng Wen, Azin Ashkan, Hoda Eydgahi, and Brian Eriksson. Matroid bandits: Fast combinatorial optimization with learning. In *13th Conf. on Uncertainty in Artificial Intelligence (UAI)*, pages 420–429, 2014.

Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvári. Tight regret bounds for stochastic combinatorial semi-bandits. In *18th Intl. Conf. on Artificial Intelligence and Statistics (AISTATS)*, 2015.

Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *International Conference on Machine Learning*, pages 6357–6368. PMLR, 2021.

Yuantong Li, Guang Cheng, and Xiaowu Dai. Dynamic online recommendation for two-sided market with bayesian incentive compatibility, 2024a.

Yuantong Li, Guang Cheng, and Xiaowu Dai. Two-sided competing matching recommendation markets with quota and complementary preferences constraints, 2024b.

Kung-Yee Liang and Scott L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 04 1986. ISSN 0006-3444. doi: 10.1093/biomet/73.1.13. URL https://doi.org/10.1093/biomet/73.1.13.

Lydia T. Liu, Ashia Wilson, Nika Haghtalab, Adam Tauman Kalai, Christian Borgs, and Jennifer T. Chayes. The disparate equilibria of algorithmic decision making when individuals invest rationally. In Mireille Hildebrandt, Carlos Castillo, L. Elisa Celis, Salvatore Ruggieri, Linnet Taylor, and Gabriela Zanfir-Fortuna, editors, *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain,* *January 27-30, 2020*, pages 381–391. ACM, 2020. doi: 10.1145/3351095.3372861. URL https://doi.org/10.1145/3351095.3372861.

Joshua R. Loftus, Chris Russell, Matt J. Kusner, and Ricardo Silva. Causal reasoning for algorithmic fairness. *CoRR*, abs/1805.05859, 2018. URL http://arxiv.org/abs/1805.05859.

Alexander R Luedtke and Mark J van der Laan. Optimal individualized treatments in resource-limited settings. *The international journal of biostatistics*, 12(1):283–303, 2016.

Yishay Mansour, Aleksandrs Slivkins, and Vasilis Syrgkanis. Bayesian incentive-compatible bandit exploration. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, pages 565–582, 2015a.

Yishay Mansour, Aleksandrs Slivkins, and Vasilis Syrgkanis. Bayesian incentive-compatible bandit exploration. *In 15th ACM Conf. on Economics and Computation (ACM EC)*, 2015b.

Yishay Mansour, Aleksandrs Slivkins, Vasilis Syrgkanis, and Zhiwei Steven Wu. Bayesian exploration: Incentivizing exploration in bayesian games. In Vincent Conitzer, Dirk Bergemann, and Yiling Chen, editors, *Proceedings of the 2016 ACM Conference on Economics and Computation, EC '16, Maastricht, The Netherlands, July 24-28, 2016*, page 661. ACM, 2016. doi: 10.1145/2940716.2940755. URL https://doi.org/10.1145/2940716.2940755.

Yishay Mansour, Aleksandrs Slivkins, and Vasilis Syrgkanis. Bayesian incentive-compatible bandit exploration. *Operations Research*, 68(4):1132–1161, 2020a. Preliminary version in *ACM EC 2015*.

Yishay Mansour, Aleksandrs Slivkins, and Vasilis Syrgkanis. Bayesian incentive-compatible bandit exploration. *Operations Research*, 68(4):1132–1161, 2020b.

Yishay Mansour, Aleksandrs Slivkins, Vasilis Syrgkanis, and Steven Wu. Bayesian exploration: Incentivizing exploration in Bayesian games. *Operations Research*, 70(2), 2022. Preliminary version in *ACM EC 2016*.

Charles Marx, Flavio Calmon, and Berk Ustun. Predictive multiplicity in classification. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6765–6774. PMLR, 13–18 Jul 2020. URL `https://proceedings.mlr.press/v119/marx20a.html`.

William F Massy. Principal components regression in exploratory statistical research. *Journal of the American Statistical Association*, 60(309):234–256, 1965.

H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agøura y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. *International Conference on Artificial Intelligenece and Statistics (AISTATS)*, 2017.

Reshef Meir, Ariel D. Procaccia, and Jeffrey S. Rosenschein. Algorithms for strategyproof classification. *Artificial Intelligence*, 186:123–156, 2012. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2012.03.008. URL `https://www.sciencedirect.com/science/article/pii/S000437021200029X`.

Celestine Mendler-Dünner, Juan C. Perdomo, Tijana Zrnic, and Moritz Hardt. Stochastic optimization for performative prediction. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL `https://proceedings.neurips.cc/paper/2020/hash/33e75ff09dd601bbe69f351039152189-Abstract.html`.

John Miller, Smitha Milli, and Moritz Hardt. Strategic classification is causal modeling in disguise. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 6917–6926. PMLR, 2020. URL `http://proceedings.mlr.press/v119/miller20b.html`.

John Miller, Juan C. Perdomo, and Tijana Zrnic. Outside the echo chamber: Optimizing the performative risk. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual*

Event, volume 139 of *Proceedings of Machine Learning Research*, pages 7710–7720. PMLR, 2021. URL http://proceedings.mlr.press/v139/miller21a.html.

Scott M Mourtgos, Ian T Adams, and Justin Nix. Elevated police turnover following the summer of george floyd protests: A synthetic control study. *Criminology & Public Policy*, 21(1):9–33, 2022.

Hussein Mouzannar, Mesrob I. Ohannessian, and Nathan Srebro. From fair decision making to social equality. In danah boyd and Jamie H. Morgenstern, editors, *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, pages 359–368. ACM, 2019. doi: 10.1145/3287560.3287599. URL https://doi.org/10.1145/3287560.3287599.

Roger B. Myerson. *Revelation Principle*, pages 11646–11652. Palgrave Macmillan UK, London, 2018. ISBN 978-1-349-95189-5. doi: 10.1057/978-1-349-95189-5_2362. URL https://doi.org/10.1057/978-1-349-95189-5_2362.

Georgy Noarov, Ramya Ramalingam, Aaron Roth, and Stephan Xie. High-dimensional prediction for sequential decision making, 2023.

DJ Pangburn. Schools are using software to help pick who gets in. what could go wrong? https://www.fastcompany.com/90342596/schools-are-quietly-turning-to-ai-to-\help-pick-who-gets-in-what-could-go-\wrong, 2019.

Judea Pearl. The seven tools of causal inference, with reflections on machine learning. *Commun. ACM*, 62(3):54–60, 2019. doi: 10.1145/3241036. URL https://doi.org/10.1145/3241036.

Juan C. Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 7599–7609. PMLR, 2020. URL http://proceedings.mlr.press/v119/perdomo20a.html.

Kevin Petrasic, Benjamin Saul, James Greig, and Matthew Bornfreund. Algorithms and bias: What lenders need to know. *White & Case*, 2017.

Hongxiang Qiu, Marco Carone, Ekaterina Sadikova, Maria Petukhova, Ronald C Kessler, and Alex Luedtke. Optimal individualized decision rules using instrumental variable methods. *Journal of the American Statistical Association*, 116(533):174–191, 2021.

Hongxiang Qiu, Marco Carone, and Alex Luedtke. Individualized treatment rules under stochastic treatment cost constraints. *Journal of Causal Inference*, 10(1):480–493, 2022.

Lisa Rice and Deidre Swesnik. Discriminatory effects of credit scoring on communities of color. *Suffolk UL Rev.*, 46:935, 2013.

P. M. Robinson. Root-n-consistent semiparametric regression. *Econometrica*, 56(4):931–954, 1988. ISSN 00129682, 14680262. URL `http://www.jstor.org/stable/1912705`.

Jean-Charles Rochet and Jean Tirole. Platform Competition in Two-Sided Markets. *Journal of the European Economic Association*, 1(4):990–1029, 06 2003. ISSN 1542-4766. doi: 10.1162/154247603322493212. URL `https://doi.org/10.1162/154247603322493212`.

Kit T. Rodolfa, Erika Salomon, Lauren Haynes, Iván Higuera Mendieta, Jamie Larson, and Rayid Ghani. Case study: predictive fairness to reduce misdemeanor recidivism through social service interventions. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 142–153, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3372863. URL `https://doi.org/10.1145/3351095.3372863`.

Julia Elisabeth Rose. Same same, not different: client-advisor matching in the finance industry. *Not Different: Client-Advisor Matching in the Finance Industry (April 17, 2023)*, 2023.

Aaron Roth, Alexander Tolbert, and Scott Weinstein. Reconciling individual probability forecasts. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 101–110, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701924. doi: 10.1145/3593013.3593980. URL `https://doi.org/10.1145/3593013.3593980`.

Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

Daniel Russo and Benjamin Van Roy. An information-theoretic analysis of thompson sampling. *J. of Machine Learning Research (JMLR)*, 17:68:1–68:30, 2016.

Daniel Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, and Zheng Wen. A tutorial on thompson sampling. *Foundations and Trends in Machine Learning*, 11(1): 1–96, 2018. Published with *Now Publishers* (Boston, MA, USA). Also available at `https://arxiv.org/abs/1707.02038`.

Marc Rysman and Julian Wright. The economics of payment cards. *Review of Network Economics*, 13(3):303–353, 2014. URL `https://EconPapers.repec.org/RePEc:bpj:rneart:v:13:y:2014:i:3:p:303-353:n:4`.

Paul R Sackett, Nathan R Kuncel, Justin J Arneson, Sara R Cooper, and Shonna D Waters. Socioeconomic status and the relationship between the sat® and freshman gpa: An analysis of data from 41 colleges and universities. research report no. 2009-1. *College Board*, 2009.

Rebecca Safier. Complete list: Colleges with rolling admissions, 2022. URL `https://blog.prepscholar.com/colleges-with-rolling-admissions`.

Javier Sánchez-Monedero, Lina Dencik, and Lilian Edwards. What does it mean to 'solve' the problem of discrimination in hiring?: social, technical and legal perspectives from the UK on automated hiring systems. In Mireille Hildebrandt, Carlos Castillo, L. Elisa Celis, Salvatore Ruggieri, Linnet Taylor, and Gabriela Zanfir-Fortuna, editors, *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, pages 458–468. ACM, 2020. doi: 10.1145/3351095.3372849. URL `https://doi.org/10.1145/3351095.3372849`.

Alvaro Sandroni. The reproducible properties of correct forecasts. *International Journal of Game Theory*, 32(1):151–159, 2003. URL `https://EconPapers.repec.org/RePEc:spr:jogath:v:32:y:2003:i:1:p:151-159`.

Yuris Mulya Saputra, Diep N Nguyen, Dinh Thai Hoang, Thang X Vu, Eryk Dutkiewicz, and Symeon Chatzinotas. Federated learning meets contract theory: Economic-efficiency framework for electric vehicle networks. *IEEE Transactions on Mobile Computing*, 21(8):2803–2817, 2020.

Bernhard Schölkopf. Causality for machine learning. *CoRR*, abs/1911.10500, 2019. URL http://arxiv.org/abs/1911.10500.

Mark Sellke. Incentivizing exploration with linear contexts and combinatorial actions, 2023.

Mark Sellke and Aleksandrs Slivkins. Sample complexity of incentivized exploration. *CoRR*, abs/2002.00558, 2020. URL https://arxiv.org/abs/2002.00558.

Mark Sellke and Aleksandrs Slivkins. The price of incentivizing exploration: A characterization via thompson sampling and sample complexity. In *22th ACM Conf. on Economics and Computation (ACM-EC)*, 2021a.

Mark Sellke and Aleksandrs Slivkins. The price of incentivizing exploration: A characterization via thompson sampling and sample complexity, 2021b.

Mark Sellke and Aleksandrs Slivkins. The price of incentivizing exploration: A characterization via thompson sampling and sample complexity, 2022.

Eliran Shabat, Lee Cohen, and Yishay Mansour. Sample complexity of uniform convergence for multicalibration. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 13331–13340. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/9a96876e2f8f3dc4f3cf45f02c61c0c1-Paper.pdf.

C. Shapiro and H.R. Varian. *Information Rules: A Strategic Guide to the Network Economy*. Strategy/Technology / Harvard Business School Press. Harvard Business School Press, 1999. ISBN 9780875848631. URL https://books.google.com/books?id=aE_J4Iv_PVEC.

Yonadav Shavit, Benjamin L. Edelman, and Brian Axelrod. Causal strategic linear regression. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 8676–8686. PMLR, 2020. URL http://proceedings.mlr.press/v119/shavit20a.html.

Dennis Shen, Peng Ding, Jasjeet Sekhon, and Bin Yu. Same root different leaves: Time series and cross-sectional methods in panel data. *arXivorg*, 2022.

Rachael Hwee Ling Sim, Yehong Zhang, Mun Choon Chan, and Bryan Kian Hsiang Low. Collaborative machine learning with incentive-aware model rewards. *CoRR*, abs/2010.12797, 2020. URL `https://arxiv.org/abs/2010.12797`.

Max Simchowitz and Aleksandrs Slivkins. Incentives and exploration in reinforcement learning, 2021a. Working paper, available at `https://arxiv.org/abs/2103.00360`.

Max Simchowitz and Aleksandrs Slivkins. Exploration and incentives in reinforcement learning. *CoRR*, abs/2103.00360, 2021b. URL `https://arxiv.org/abs/2103.00360`.

Aleksandrs Slivkins. Incentivizing exploration via information asymmetry. *XRDS*, 24 (1):38–41, September 2017. ISSN 1528-4972. doi: 10.1145/3123744. URL `https://doi.org/10.1145/3123744`.

Aleksandrs Slivkins. Introduction to multi-armed bandits. *Found. Trends Mach. Learn.*, 12(1-2):1–286, 2019a. doi: 10.1561/2200000068. URL `https://doi.org/10.1561/2200000068`.

Aleksandrs Slivkins. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, November 2019b. Published with *Now Publishers* (Boston, MA, USA). Also available at `https://arxiv.org/abs/1904.07272`. Latest online revision: Jan 2022.

Aleksandrs Slivkins. Introduction to multi-armed bandits, 2019c. URL `https://arxiv.org/abs/1904.07272`.

Aleksandrs Slivkins. Exploration and persuasion, 2021.

Austa Somvichian-Clausen. Experts see new roles for artificial intelligence in college admissions process. *The Hill*, 2021.

Hao Sun, Evan Munro, Georgy Kalashnov, Shuyang Du, and Stefan Wager. Treatment allocation under uncertain costs. *arXiv preprint arXiv:2103.11066*, 2021.

Latanya Sweeney. Discrimination in online ad delivery. *ACM Queue*, 11(3):10, 2013. doi: 10.1145/2460276.2460278. URL https://doi.org/10.1145/2460276.2460278.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision, 2015.

Ming Tang and Vincent W.S. Wong. An incentive mechanism for cross-silo federated learning: A public goods perspective. In *IEEE INFOCOM 2021 - IEEE Conference on Computer Communications*, pages 1–10, 2021. doi: 10.1109/INFOCOM42981.2021.9488705.

Tra Huong Thi Le, Nguyen H. Tran, Yan Kyaw Tun, Minh N. H. Nguyen, Shashi Raj Pandey, Zhu Han, and Choong Seon Hong. An incentive mechanism for federated learning in wireless cellular networks: An auction approach. *IEEE Transactions on Wireless Communications*, 20(8):4874–4887, 2021. doi: 10.1109/TWC.2021.3062708.

William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5(1), August 2018. ISSN 2052-4463. doi: 10.1038/sdata.2018.161. URL http://dx.doi.org/10.1038/sdata.2018.161.

Xuezhen Tu, Kun Zhu, Nguyen Cong Luong, Dusit Niyato, Yang Zhang, and Juan Li. Incentive mechanisms for federated learning: From economic and game theoretic perspective. *IEEE Transactions on Cognitive Communications and Networking*, 2022.

Uber, Jun 2019. URL https://www.uber.com/blog/causal-inference-at-uber/.

R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. ISBN 9781108415194. URL https://books.google.com/books?id=J-VjswEACAAJ.

Siwei Wang and Wei Chen. Thompson sampling for combinatorial semi-bandits. In

Proceedings of the 35th International Conference on Machine Learning, volume 80 of *Proceedings of Machine Learning Research*, pages 5114–5122. PMLR, 10–15 Jul 2018.

Zheng Wen, Branislav Kveton, and Azin Ashkan. Efficient learning in large-scale combinatorial semi-bandits. In *32nd Intl. Conf. on Machine Learning (ICML)*, pages 1113–1122, 2015.

EC Wright. Non-compliance–or how many aunts has matilda? *Lancet (London, England)*, 342(8876):909—913, October 1993. ISSN 0140-6736. doi: 10.1016/0140-6736(93)91951-h. URL https://doi.org/10.1016/0140-6736(93)91951-h.

Yingying Zhang, Huixia Judy Wang, and Zhongyi Zhu. Quantile-regression-based clustering for panel data. *Journal of Econometrics*, 213(1):54–67, 2019.

Shengjia Zhao, Michael Kim, Roshni Sahoo, Tengyu Ma, and Stefano Ermon. Calibrating predictions to decisions: A novel approach to multi-class calibration. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 22313–22324. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/bbc92a647199b832ec90d7cf57074e9e-Paper.pdf.

Appendix A

Appendix for Incentivizing Combinatorial Bandit Exploration

A.I Motivating examples

We spell out three examples of incentivized exploration that are specific to combinatorial semi-bandits. We frame each example as a recommendation system that recommends "actions" with a combinatorial structure. A user then can choose to follow such recommendation (i.e., pursue the action), or deviate from it. The BIC condition ensures the former, in which case the user reports a "reward" for each atom in the action. Thus:

- Recommending online content such as news/entertainment articles. Here, an "action" is a slate of articles. The platform recommends each user a slate of articles, and the user can choose to "follow the recommendation": look at each article in the slate. The feedback for each article is either generated automatically from the interaction (*e.g.,* dwell time or how far the user has scrolled) or entered explicitly (*e.g.,* "thumbs up" or "thumbs down").

- Recommending complementary products: *e.g.,* a three-piece suit or the contents of a child's pencil box. Thus, an "action" is a slate of products. A user who follows such a recommendation buys each item in the slate, and eventually reports its quality (*e.g.,* "thumbs up" or "thumbs down").

- Recommending driving directions. Here, an "action" consists of a route, which is a sequence of segments. A driver can choose to follow this route, in which case the travel time for each segment is reported (*e.g.,* automatically by an app).

Our framing differs from the "usual" framing for combinatorial semi-bandits, much like the framing for incentivized exploration differs from that for multi-armed bandits.

A.II BIC analysis for Thompson Sampling

This appendix provides the proofs for Section 3.3, the BIC analysis of Thompson Sampling. Specifically, we prove Theorem 3.3.1 (that Thompson Sampling is BIC when initialized with sufficiently many samples) and Corollary 3.3.4 (that the sufficient number of samples is polynomial in d).

Proof of Theorem 3.3.1 By definition, Thompson sampling is BIC at a particular round $t > T_0$ if and only if we have $\mathbb{E}[\mu(A) - \mu(A')|A^{(t)} = A] \geq 0$ for all (i, j) such that $i \neq j$. This condition can be written as:

$$\begin{aligned}
\mathbb{E}[\mu(A) - \mu(A')|A^{(t)} = A] &= \frac{\mathbb{E}\left[\mathbb{E}^{(t)}[\mu(A) - \mu(A')]\Pr^{(t)}[A^{(t)} = A]\right]}{\Pr[A^{(t)} = A]} \\
&= \frac{\mathbb{E}\left[\mathbb{E}^{(t)}[\mu(A) - \mu(A')]\Pr^{(t)}[A^{*} = A]\right]}{\Pr[A^{*} = A]} \\
&\qquad\text{(by definition of Thompson Sampling)}
\end{aligned}$$

Observe that the denominator $\Pr[A^* = A]$ is a positive prior-dependent constant. Hence, we only need to bound the numerator to satisfy the BIC condition.

Fixing arms A, A', we can rewrite the numerator as:

$$\begin{aligned}
\mathbb{E}\left[\overset{(t)}{\mathbb{E}}[\mu(A) - \mu(A')]\overset{(t)}{\Pr}[A^* = A]\right] &= \mathbb{E}\left[\overset{(t)}{\mathbb{E}}\left[\overset{(t)}{\mathbb{E}}[\mu(A) - \mu(A')] \cdot \mathbf{1}_{\{A^*=A\}}\right]\right] \\
&= \mathbb{E}\left[\overset{(t)}{\mathbb{E}}[\mu(A) - \mu(A')] \cdot \mathbf{1}_{\{A^*=A\}}\right]
\end{aligned}$$

For Thompson sampling to be BIC, it suffices to show that

$$\mathbb{E}\left[\overset{(t)}{\mathbb{E}}[\mu(A)-\mu(A')]\mathbf{1}_{\{A^*=A\}}\right]\geq 0$$

We first prove our observation that the functions $(\mu(A)-\mu(A'))_+$ and $\mathbf{1}_{\{A^*=A\}}$ are co-monotone in each coordinate of θ, which means they are both increasing in some coordinates while both decreasing in the other coordinates. Specifically for any ℓ-th coordinate of θ, they are both increasing in θ_ℓ (given all other coordinates in θ stay the same) if A_ℓ (the ℓ-th coordinate of A) equals to 1. Otherwise, if $A_\ell = 0$, they are both decreasing in θ_ℓ.

Given any θ and θ' having the same coordinates $\theta_x = \theta'_x$ $(x \in [d])$ except for the ℓ-th coordinate, $\theta_\ell > \theta'_\ell$, and for any arm $A' \neq A$,

$$\begin{aligned}
&\langle\theta, A-A'\rangle - \langle\theta', A-A'\rangle && \text{(A.1)}\\
&= \sum_{x=1}^{d}\theta_x(A_x - A'_x) - \sum_{x=1}^{d}\theta'_x(A_x-A'_x) && \text{(A.2)}\\
&= (\theta_\ell-\theta'_\ell)(A_\ell-A'_\ell) && \text{(other coordinates are the same except } \ell)\\
&\begin{cases}\geq 0, & \text{if } A_\ell = 1\\ \leq 0, & \text{if } A_\ell = 0\end{cases} && (\theta_\ell > \theta'_\ell \text{ and } A_\ell, A'_\ell \in \{0,1\})
\end{aligned}$$

Note that $\mathbf{1}_{\{A^*=A\}} = \mathbf{1}_{\{\mu(A)-\mu(A')\geq 0, \forall A'\neq A\}}$. So if $A_\ell = 1$, then $\mu(A)-\mu(A')$ are increasing in θ_ℓ for all $A' \neq A$, especially for $A_k = A'$. Hence, $\mathbf{1}_{\{A^*=A\}}$ and $\mu(A)-\mu(A')$ are both increasing in θ_ℓ. Otherwise, if $A_\ell = 0$, they are both decreasing in θ_ℓ.

Hence, we can apply Remark A.IV.3 to lower bound the expression above as follows:

$$\begin{aligned}
\mathbb{E}[\mu(A)-\mu(A')\cdot\mathbf{1}_{\{A^*=A\}}] &= \mathbb{E}[(\mu(A)-\mu(A'))_+\cdot\mathbf{1}_{\{A^*=A\}}]\\
&\geq \min_{A,A'\in\mathcal{A}}\mathbb{E}[(\mu(A)-\mu(A'))_+]\Pr[A^*=A]\\
&= \varepsilon_{\mathtt{TS}}\delta_A
\end{aligned}$$

where $\delta_A = \Pr[A^* = A] \geq \delta_{\mathtt{TS}}$.

To finish the proof, we need the following inequality to hold:

$$\mathbb{E}[|\overset{T_0}{\mathbb{E}}[\langle\theta, A - A'\rangle] \cdot \mathbf{1}_{\{A^*=A\}} - \langle\theta, A - A'\rangle \cdot \mathbf{1}_{\{A^*=A\}}|] \leq \varepsilon_{\mathtt{TS}}\delta_A \tag{A.3}$$

where $\delta_A = \Pr[A^* = A]$. Regrouping and using triangle inequality on the LHS of Equation (A.3), we have:

$$\begin{aligned}
&\mathbb{E}\left[|\overset{T_0}{\mathbb{E}}[\langle\theta, A - A'\rangle] \cdot \mathbf{1}_{\{A^*=A\}} - \langle\theta, A - A'\rangle \cdot \mathbf{1}_{\{A^*=A\}}|\right] \\
&\leq \mathbb{E}\left[|\overset{T_0}{\mathbb{E}}[\langle\theta, A\rangle] - \langle\theta, A\rangle| \cdot \mathbf{1}_{\{A^*=A\}}\right] + \mathbb{E}\left[|\overset{T_0}{\mathbb{E}}[\langle\theta, A'\rangle] - \langle\theta, A'\rangle| \cdot \mathbf{1}_{\{A^*=A\}}\right]
\end{aligned} \tag{A.4}$$

The final step is to bound each individual summand in the inequality above. By the Bayesian Chernoff Bound (Lemma A.IV.1), we have $\left\|\mathbb{E}^{T_0}[\theta] - \theta\right\|$ is a $(n_{\mathtt{TS}}^{-1/2}\sqrt{d})$ times $O(1)$-sub-Gaussian random variable. Then, by Cauchy-Schwarz inequality, we have

$$\begin{aligned}
|\overset{T_0}{\mathbb{E}}[\langle\theta, A\rangle] - \langle\theta, A\rangle| &\leq \left\|\overset{T_0}{\mathbb{E}}[\theta] - \theta\right\| \cdot \|A\| \\
&\leq \sqrt{d}\left\|\overset{T_0}{\mathbb{E}}[\theta] - \theta\right\|
\end{aligned}$$

Hence, $|\mathbb{E}^{T_0}[\langle\theta, A\rangle] - \langle\theta, A\rangle|$ and $|\mathbb{E}^{T_0}[\langle\theta, A'\rangle] - \langle\theta, A'\rangle|$ have magnitude at most as large as a $(n_{\mathtt{TS}}^{-1/2} \cdot d)$ times $O(1)$-sub-Gaussian random variable. Then, we can apply Lemma A.IV.4 to both terms in the inequality (A.4) above and upper bound it by at most $O\left(\delta_A \cdot n_{\mathtt{TS}}^{-1/2} \cdot d\sqrt{\log\left(\delta_A^{-1}\right)}\right)$. Then, using our choice of $n_{\mathtt{TS}}$ and $\delta_{\mathtt{TS}} \leq \delta_A$, we arrive at the conclusion.

Proof of Corollary 3.3.4 Recall that by Theorem 3.3.1, we have $n_{\mathtt{TS}} = C_{\mathtt{TS}} \cdot d^2 \cdot \varepsilon_{\mathtt{TS}}^{-2} \cdot \log(\delta_{\mathtt{TS}}^{-1})$. Let $\varepsilon_{\mathcal{C}}$ be the version of $\varepsilon_{\mathtt{TS}}$ where the min is taken over all ordered pairs of priors in $\mathcal{C}$. Then we have $\varepsilon_{\mathtt{TS}} \geq \varepsilon_{\mathcal{C}}$. Since $\mathcal{C}$ is finite and satisfy the pairwise non-dominance assumption, $\varepsilon_{\mathcal{C}}$ is strictly positive.

By definition, $\delta_{\mathtt{TS}} = \min_{A\in\mathcal{A}} \Pr[A^* = A]$. Fix an arm A. Then, we can decompose the probability of arm A being the best arm as:

$$\begin{aligned}\Pr[A^* = A] &= \Pr[\langle \theta, A - A' \rangle \geq 0, \forall A' \neq A] \\ &= \Pr\left[\sum_{\ell \in A} \theta_\ell - \sum_{x \in A'} \theta_x \geq 0, \forall A' \neq A\right] \qquad \text{(A.5)}\end{aligned}$$

We observe that the event when A is the best arm is more likely than the event when each atom in A is larger than τ, and all other atoms not in A is smaller than τ/d. Hence, we can lower bound the probability above as:

$$\begin{aligned}&\Pr\left[\sum_{\ell \in A} \theta_\ell - \sum_{x \in A'} \theta_x \geq 0, \forall A' \neq A\right] \\ &\geq \Pr\left[\forall \ell \in A, \theta_\ell \geq \tau \quad \text{and} \quad \forall x \notin A, \theta_x \leq \tau/d\right] \\ &= \Pr\left[\forall \ell \in A, \theta_\ell \geq \tau\right] \cdot \Pr\left[\forall x \notin A, \theta_x \leq \tau/d\right] \quad \text{(the prior is independent across atoms)} \\ &= \mathbb{E}\left[\prod_{\ell \in A} \mathbf{1}_{\{\theta_\ell \geq \tau\}}\right] \cdot \mathbb{E}\left[\prod_{x \notin A} \mathbf{1}_{\{\theta_x \leq \tau/d\}}\right]\end{aligned}$$

Observe that the values $\{\theta_\ell\}_{\ell \in [d]}$ are independent, and each function $\mathbf{1}_{\{\theta_\ell \geq \tau\}}$ (and $\mathbf{1}_{\{\theta_\ell \leq \tau/d\}}$) are co-monotone in each coordinate of θ. Then, repeated application of mixed-monotone Harris inequality (see Remark A.IV.3) implies that

$$\begin{aligned}\Pr[A^* = A] &\geq \prod_{\ell \in A} \mathbb{E}[\mathbf{1}_{\{\theta_\ell \geq \tau\}}] \cdot \prod_{x \notin A} \mathbb{E}[\mathbf{1}_{\{\theta_x \leq \tau/d\}}] \qquad \text{(mixed-monotonicity Harris)} \\ &= \prod_{\ell \in A} \Pr[\theta_\ell \geq \tau] \cdot \prod_{x \notin A} \Pr[\theta_x \leq \tau/d] \\ &\geq \prod_{\ell=1}^{d} \Pr[\theta_\ell \geq \tau] \Pr[\theta_\ell \leq \tau/d]\end{aligned}$$

By the full support assumption Equation (3.4), we define a prior-dependent constant $\rho_\tau = \min_{\ell \in [d]} \Pr[\theta_\ell \geq \tau] > 0$. Then, by definition of ρ_τ and the non-degeneracy assumption Equation (3.5), we have:

$$\prod_{\ell=1}^{d} \Pr[\theta_\ell \geq \tau] \Pr[\theta_\ell \leq {}^{\tau}/_{d}] \geq \prod_{\ell=1}^{d} \rho_\tau^d \cdot \mathrm{poly}({}^{d}/_{\tau}) \cdot \exp(-({}^{\tau}/_{d})^{-\alpha})$$
$$\geq \rho_\tau^d \cdot \mathrm{poly}({}^{d^d}/_{(\tau)^d}) \cdot \exp(-d({}^{\tau}/_{d})^{-\alpha})$$

Substituting this expressions and $\varepsilon_{\mathtt{TS}}$ into $n_{\mathtt{TS}}$, we have $n_{\mathtt{TS}} = O_{\mathcal{C}}(d^{3+\alpha} \log d)$.

A.III Initial exploration: reduction to K-armed bandits

A.III.1 Proof of Theorem 3.4.1

Recall that we build on an approach from Mansour et al. (2020a), encapsulated in Theorem 3.4.1. Let us clarify how this theorem follows from the material in Mansour et al. (2020a).

The algorithm from Mansour et al. (2020a) is modified in two ways: it explores the arms in the order given by the sequence $V_1 , \ \ldots \ , V_\kappa$, and the observed outcome from playing a given arm now includes the rewards for all atoms in this arm. Let us spell out the resulting algorithm, for completeness.

Algorithm 12: Hidden Exploration (modification of Algorithm 3 in Mansour et al. (2020a))

Parameters: $L, N \in \mathbb{N}$

1 For the first N rounds, recommend arm V_1.

2 Let $s_1 = \left(r_\ell^{(t)} :\ \ell \in V_1,\, t \in [N] \right)$ be the tuple of all observed per-atom rewards from arm V_1;

3 **for** *each arm* V_i *in increasing order of* i **do**

4 Let $A^* = \mathrm{argmax}_{A \in \mathcal{A}} \mathbb{E}\left[\, \mu(A) \mid s_1\,,\ \ldots\ , s_{i-1} \right]$, breaking ties favoring smaller index;

5 From the set P of the next $L \cdot N$ rounds, pick a set Q of N rounds uniformly at random;

6 Every agent $p \in P - Q$ is recommended arm A^*;

7 Every agent $p \in Q$ is recommended arm V_i;

8 Let s_i^N be the tuple of all per-atom rewards from arm V_i observed in rounds $t \in Q$;

9 **end**

The analysis in Section 5.2 of Mansour et al. (2020a) carries over seamlessly to combinatorial semi-bandits, and yields the following guarantee:

Lemma A.III.1 (Mansour et al. (2020a))**.** *Assume Property (P) holds with constants $n_\mathcal{P}, \tau_\mathcal{P}, \rho_\mathcal{P}$ and $\kappa = \kappa(n_\mathcal{P}) < \infty$. Then Algorithm 12 with parameters $N \geq n_\mathcal{P}$ and L satisfying* (A.6) *is BIC:*

$$L \geq 1 + \frac{\mu^0_{\max} - \mu^0_{\min}}{\tau_\mathcal{P} \cdot \rho_\mathcal{P}}, \tag{A.6}$$

where $\mu^0_{\max} = \max_{A \in \mathcal{A}} \mathbb{E}[\mu(A)]$ and $\mu^0_{\min} = \min_{A \in \mathcal{A}} \mathbb{E}[\mu(A)]$.

Proof of Theorem 3.4.1. It remains to interpret the quantities in Lemma A.III.1. According to Lemma A.III.1, Algorithm 12 is BIC with parameters $N \geq n_\mathcal{P}$ and L satisfying Equation (A.6). It suffices to take $N = n_\mathcal{P}$. Since $\theta^0_\ell \in [0,1]$ for any $\ell \in [d]$, we have $0 \leq \mu^0_{\min} \leq \mu^0_{\max} \leq d$ and $0 \leq \mu^0_{\max} - \mu^0_{\min} \leq d$. Additionally, $\tau_\mathcal{P}, \rho_\mathcal{P} \in (0,1)$. So

$$1 + \frac{\mu^0_{\max} - \mu^0_{\min}}{\tau_\mathcal{P} \cdot \rho_\mathcal{P}} \leq 1 + \frac{d}{\tau_\mathcal{P} \cdot \rho_\mathcal{P}} \leq \frac{1+d}{\tau_\mathcal{P} \cdot \rho_\mathcal{P}}.$$

And thus it suffices to take $L = \frac{1+d}{\tau_\mathcal{P} \cdot \rho_\mathcal{P}}$. Then we have the total number of rounds

$T_0 = \kappa \cdot N \cdot L = \kappa \cdot n_{\mathcal{P}} \cdot (1+d) \,/\, (\tau_{\mathcal{P}} \cdot \rho_{\mathcal{P}})$. □

A.III.2 Restricted family of arms

Firstly, according to Assumption (3.11) and the reward support $\Theta \subset [0,1]$, we observe that:

$$\text{the prior/posterior-best arm contains the } m \text{ prior/posterior-best atoms;} \tag{A.7}$$

$$\text{the second prior/posterior-best arm contains the } m-1 \text{ prior/posterior-best atoms.} \tag{A.8}$$

Then, according to our choice of $n_{\mathcal{P}}$ (3.12), we will prove that $\kappa = \kappa(n_{\mathcal{P}})$ is finite (i.e. our arm sequence will contain all atoms at least once). Actually, we will prove $\kappa = \lceil d/m \rceil$ by proving the following Claim A.III.2.

Claim A.III.2. *Assume Beta-Bernoulli priors (3.7-3.8), all arms have a fixed size (3.11), and $n_{\mathcal{P}}$ satisfies (3.12). Then the arm sequence $V_1, V_2, \cdots$, where $V_i = V_i^{n_{\mathcal{P}}}$, have the following properties:*

$$V_i = \{\, (i-1)m + \ell : \ell \in [m] \,\}, \quad i \in [\lceil d/m \rceil - 1]; \tag{A.9}$$

$$V_i \supset \{(i-1)m + \ell : \ell \in [m], (i-1)m + \ell \le d\}, \quad i = \lceil d/m \rceil. \tag{A.10}$$

And thus $\kappa(n_{\mathcal{P}}) = \lceil d/m \rceil$.

Proof. We will prove by induction on phase i. For phase $i = 1$, V_1 is the prior-best arm. According to the observation (A.7), V_1 contains the largest m prior-best atoms, which is $[m]$.

Suppose the induction hypothesis is true for all phases up to some phase $i \in [\lceil d/m \rceil - 1]$. Denote B_i as a subset of atoms having been contained at least once in the first i arms and $\bar{B}_i$ as the complement subset of atoms. Then

$$B_i = \bigcup_{j \in [i]} V_j = [im] \text{ and } \bar{B}_i = [d] - [im].$$

Recall the definition of $Z_\ell^{n_\mathcal{P}}$ and $\nu_\ell(n)$, we have for each atom $\ell \in B_i$ and $\ell' \in \bar{B}_i$:

$$Z_\ell^{n_\mathcal{P}} = n_\mathcal{P} \text{ and } \nu_\ell(Z_\ell^{n_\mathcal{P}}) = \nu_\ell(n_\mathcal{P}) = \alpha_\ell/(\alpha_\ell + \beta_\ell + n_\mathcal{P});$$

$$Z_{\ell'}^{n_\mathcal{P}} = 0 \text{ and } \nu_{\ell'}(Z_{\ell'}^{n_\mathcal{P}}) = \nu_{\ell'}(0) = \alpha_{\ell'}/(\alpha_{\ell'} + \beta_{\ell'}) = \theta_{\ell'}^0.$$

Since $\theta_1^0 \geq, \dots, \geq \theta_d^0$, we have:

$$\nu_{\ell'}(Z_{\ell'}^{n_\mathcal{P}}) \text{ decreases in } \ell' \in \bar{B}_i. \tag{A.11}$$

By definition of $n_\mathcal{P}$ and $\theta_{\ell'}^0 \geq \theta_d^0$, we have:

$$\alpha_\ell/(\alpha_\ell + \beta_\ell + n_\mathcal{P}) < \alpha_d/(\alpha_d + \beta_d) \leq \alpha_{\ell'}/(\alpha_{\ell'} + \beta_{\ell'}) = \theta_{\ell'}^0.$$

Thus:

$$\nu_\ell(Z_\ell^{n_\mathcal{P}}) < \nu_{\ell'}(Z_{\ell'}^{n_\mathcal{P}}), \forall \ell \in B_i, \ell' \in \bar{B}_i. \tag{A.12}$$

Combining (A.11)-(A.12) and according to the observation (A.7), we have V_{i+1} for phase $i+1$. If $i \in [\lceil d/m \rceil - 2]$, we have $|\bar{B}_i| = d - im \geq m+1$. Thus $im+1 < \cdots < (i+1)m \leq d-1$ and $V_{i+1} = \{im + \ell : \ell \in [m]\}$. Otherwise for $i = \lceil d/m \rceil - 1$, we have $1 \leq |\bar{B}_i| \leq m$. Thus $V_{i+1} \supset \bar{B}_i = \{im + \ell : \ell \in [m] \text{ and } im + \ell \leq d\}$. Thus, the induction hypothesis is true for phase $i+1$ and we complete the induction proof. And since $V_1, \dots, V_{\lceil d/m \rceil}$ contain all atoms, we have $\kappa(n_\mathcal{P}) = \lceil d/m \rceil$. $\square$

Secondly, we define an event and give a lower bound of the probability of this event. Given any $n_\mathcal{P}, N \in \mathbb{N}$ ($n_\mathcal{P} \leq N$) and $H_i^N, \forall i \in [\kappa]$, define an event $\mathcal{E}_i$ for each $i \in [\kappa]$ saying that the first $n_\mathcal{P}$ reward samples of each atom in $\bigcup_{j \in [i]} V_j$ are 0. Formally,

$$\mathcal{E}_i = \left\{ r_\ell^{(t)} = 0, \forall \ell \in \bigcup_{j \in [i]} V_j, t \in [n_\mathcal{P}] \right\}, \forall i \in [\kappa]. \tag{A.13}$$

where we abuse the notation of $r_\ell^{(t)}$ as the t-th round that atom ℓ is being contained. Since H_0^N is an empty data set, we define $\mathcal{E}_0$ is a full event, which gives no information wherever it applies.

Then, according to our choice of $\rho_{\mathcal{P}}$(3.14), we will lower bound the probability of the event defined above in the following claim.

Claim A.III.3. *Assume Beta-Bernoulli priors (3.7-3.8) and* $\rho_{\mathcal{P}}$ *satisfies (3.14). Then for any given* $n_{\mathcal{P}} \leq N$*, with the definition of* $\mathcal{E}_i, \forall i \in [\kappa]$ (A.13)*, we have:*

$$\Pr[\,\mathcal{E}_i\,] \geq \rho_{\mathcal{P}}, \forall i \in [\kappa]. \tag{A.14}$$

Proof. Firstly, by the prior and reward independence among each atoms:

$$\begin{aligned}
\Pr[\,\mathcal{E}_i\,] &= \Pr\left[r_\ell^{(t)} = 0, \forall \ell \in \bigcup_{j\in[i]} V_j, t \in [n_{\mathcal{P}}] \right] \\
&= \prod_{\ell \in \bigcup_{j\in[i]} V_j} \Pr\left[r_\ell^{(t)} = 0, \forall t \in [n_{\mathcal{P}}] \right]
\end{aligned}$$

Secondly, for each given atom ℓ, by the independence among realized rewards conditioned on the mean reward drawn from the prior and iteratively using Harris Inequality:

$$\begin{aligned}
\Pr\left[r_\ell^{(t)} = 0, \forall t \in [n_{\mathcal{P}}] \right] &= \mathop{\mathbb{E}}_{\theta_\ell}\left[\Pr_{r_\ell^{(t)}}\left[r_\ell^{(t)} = 0, \forall t \in [n_{\mathcal{P}}] \right] \mid \theta_\ell \right] \\
&= \mathop{\mathbb{E}}_{\theta_\ell}\left[\prod_{t\in[n_{\mathcal{P}}]} \Pr_{r_\ell^{(t)}}\left[r_\ell^{(t)} = 0 \mid \theta_\ell \right] \right] && \text{(conditional independence)} \\
&= \mathop{\mathbb{E}}_{\theta_\ell}\left[\prod_{t\in[n_{\mathcal{P}}]} (1-\theta_\ell) \right] && \text{(Bernoulli rewards (3.8))} \\
&\geq \prod_{t\in[n_{\mathcal{P}}]} \mathop{\mathbb{E}}_{\theta_\ell}[\,(1-\theta_\ell)\,] && \text{(Harris inequality)} \\
&= (1-\theta_\ell^0)^{n_{\mathcal{P}}}
\end{aligned}$$

Combining both and recall that $\theta_1^0 \geq \cdots \geq \theta_d^0$:

$$\Pr[\,\mathcal{E}_i\,] \geq \prod_{\ell \in \bigcup_{j\in[i]} V_j} (1-\theta_\ell^0)^{n_{\mathcal{P}}} \geq \prod_{\ell\in[d]} (1-\theta_\ell^0)^{n_{\mathcal{P}}} \geq (1-\theta_1^0)^{d n_{\mathcal{P}}} = \rho_{\mathcal{P}}.$$

□

Note that this second part analysis does not rely on (3.11) and we will reuse that part for the proof of general feasible arm set case in Section A.III.3.

Thirdly, according to our choice of $\tau_{\mathcal{P}}$ (3.13), we will prove the following claim, which says the expectation X_i^N conditioned on the event $\mathcal{E}_i$ almost surely $\geq \tau_{\mathcal{P}}$ for any phase $i \in [\kappa]$ and any $N \geq n_{\mathcal{P}}$.

Claim A.III.4. *Assume Beta-Bernoulli priors (3.7-3.8), all arms have a fixed size (3.11) and $\tau_{\mathcal{P}}$ satisfies (3.13). Then for any given $n_{\mathcal{P}}$, we have:*

$$\Pr\left[\, X_i^N \geq \tau_{\mathcal{P}} \mid \mathcal{E}_{i-1} \,\right] = 1, \forall i \in [\kappa], N \geq n_{\mathcal{P}}. \tag{A.15}$$

Proof. For each phase $i \in [\kappa]$, let A_i is the second prior/posterior-best arm conditioned on H_{i-1} and $\mathcal{E}_{i-1}$. According to observation (A.7)-(A.8) and the definition of $\tau_{\mathcal{P}}$:

$$\begin{aligned}
\min_{\text{arm } A \neq V_i} \mathbb{E}\left[\, \mu(V_i) - \mu(A) \mid H_{i-1}^N, \mathcal{E}_{i-1} \,\right] &= \mathbb{E}\left[\, \mu(V_i) - \mu(A_i) \mid H_{i-1}^N, \mathcal{E}_{i-1} \,\right] \\
&= \sum_{\ell \in V_i} \nu_\ell(Z_\ell^{n_{\mathcal{P}}}) - \sum_{\ell' \in A_i} \nu_{\ell'}(Z_{\ell'}^{n_{\mathcal{P}}}) \\
&= \min_{\ell \in V_i} \nu_\ell(Z_\ell^{n_{\mathcal{P}}}) - \min_{\ell' \in A_i} \nu_\ell(Z_{\ell'}^{n_{\mathcal{P}}}) \\
&\geq \min_{\ell, \ell' \in [d], n, n' \in \{0, n_{\mathcal{P}}\}} |\nu_\ell(n) - \nu_{\ell'}(n')| \\
&= \tau_{\mathcal{P}}.
\end{aligned}$$

Thus we have Equation (A.15). $\square$

At last, combining the claims above, we have for each $i \in [\kappa]$:

$$\begin{aligned}
\Pr\left[\, X_i^N \geq \tau_{\mathcal{P}} \,\right] &\geq \Pr\left[\, \mathcal{E}_{i-1} \,\right] \cdot \Pr\left[\, X_i^N \geq \tau_{\mathcal{P}} \mid \mathcal{E}_{i-1} \,\right] \\
&\geq \rho_{\mathcal{P}} \cdot \Pr\left[\, X_i^N \geq \tau_{\mathcal{P}} \mid \mathcal{E}_{i-1} \,\right] && \text{(Claim A.III.4)} \\
&= \rho_{\mathcal{P}} \cdot 1 && \text{(Claim A.III.3)} \\
&= \rho_{\mathcal{P}},
\end{aligned}$$

which implies (P).

A.III.3 Arbitrary family of arms

We prove Theorem 3.4.4, reusing much of the proof of Theorem 3.4.2. While the parameters in Theorem 3.4.4 give a weaker bound on the number of rounds, the proof becomes more intuitive.

Firstly, we prove κ is finite in this following claim.

Claim A.III.5. *Assume Beta-Bernoulli priors (3.7-3.8) and* $n_{\mathcal{P}}$ *satisfies (3.15). Then* $\kappa(n_{\mathcal{P}}) \le d$.

Proof. Denote the explored atom set up to phase i as $B_i = \bigcup_{j\in[i]} V_j$ and the unexplored atom set as $\bar{B}_i = [d] - B_i$. Denote $B_0 = \emptyset$ and $\bar{B}_0 = [d]$. Fixed $i \ge 0$. Recall the definition of $Z_\ell^{n_{\mathcal{P}}}$ and $\nu_\ell(n)$, we have for each atom $\ell \in B_i$ and $\ell' \in \bar{B}_i$:

$$Z_\ell^{n_{\mathcal{P}}} = n_{\mathcal{P}} \text{ and } \nu_\ell(Z_\ell^{n_{\mathcal{P}}}) = \nu_\ell(n_{\mathcal{P}}) = \alpha_\ell/(\alpha_\ell + \beta_\ell + n_{\mathcal{P}});$$

$$Z_{\ell'}^{n_{\mathcal{P}}} = 0 \text{ and } \nu_{\ell'}(Z_{\ell'}^{n_{\mathcal{P}}}) = \nu_{\ell'}(0) = \alpha_{\ell'}/(\alpha_{\ell'} + \beta_{\ell'}).$$

By definition of $n_{\mathcal{P}}$ and $\theta_{\ell'}^0 \ge \theta_d^0$, we have:

$$\nu_{\ell'}(Z_{\ell'}^{n_{\mathcal{P}}}) = \alpha_\ell/(\alpha_\ell + \beta_\ell + n_{\mathcal{P}}) < \frac{1}{d}\alpha_d/(\alpha_d + \beta_d) \le \frac{1}{d}\alpha_{\ell'}/(\alpha_{\ell'} + \beta_{\ell'}) = \frac{1}{d}\nu_{\ell'}(Z_{\ell'}^{n_{\mathcal{P}}}).$$

Thus:

$$\sum_{\ell\in B_i} \nu_\ell(Z_\ell^{n_{\mathcal{P}}}) < d \cdot \frac{1}{d}\nu_{\ell'}(Z_{\ell'}^{n_{\mathcal{P}}}) = \nu_{\ell'}(Z_{\ell'}^{n_{\mathcal{P}}}), \forall \ell' \in \bar{B}_i. \tag{A.16}$$

Then according to the definition of V_{i+1}, we know V_{i+1} contains at least one atom $\ell' \in \bar{B}_i$. So the number of uncovered atoms, i.e. $|\bar{B}_i|$, decreases at least 1 after each phase. Thus it takes at most d phases to cover all atoms, which implies $\kappa \le d$. □

Secondly, we reuse the definition of $\mathcal{E}_i$ (A.13) and Claim A.III.3 to give a lower bound of $\Pr[\mathcal{E}_i]$, since this part in Section A.III.2 don't rely on Assumption (3.11).

Thirdly, according to our definition of $\tau_{\mathcal{P}}$ (3.16), we have the following claim similar to Claim A.III.4.

Claim A.III.6. *Assume Beta-Bernoulli priors (3.7-3.8) and $\tau_{\mathcal{P}}$ satisfies (3.16). Then for any given $n_{\mathcal{P}}$, we have:*

$$\Pr\left[\, X_i^N \geq \tau_{\mathcal{P}} \mid \mathcal{E}_{i-1} \,\right] = 1, \forall i \in [\kappa], N \geq n_{\mathcal{P}}. \tag{A.17}$$

Proof. For each phase $i \in [\kappa]$, we have:

$$\begin{aligned}
\min_{\text{arm } A \neq V_i} \mathbb{E}\left[\, \mu(V_i) - \mu(A) \mid H_{i-1}^N, \mathcal{E}_{i-1} \,\right] &\geq \min_{A \neq A' \in \mathcal{A}} \mathbb{E}\left[\, \mu(A) - \mu(A') \mid H_{i-1}^N, \mathcal{E}_{i-1} \,\right] \\
&= \min_{A \neq A' \in \mathcal{A}, n, n' \in \{0, n_{\mathcal{P}}\}^d} |\sum_{\ell \in A} \nu_\ell(n) - \sum_{\ell' \in A'} \nu_\ell(n')| \\
&= \tau_{\mathcal{P}}.
\end{aligned}$$

Then we have Equation (A.17). $\square$

Similar to the last step in Section A.III.2, combining Claim A.III.5, Claim A.III.3 and Claim A.III.6, we have for each $i \in [\kappa]$:

$$\begin{aligned}
\Pr\left[\, X_i^N \geq \tau_{\mathcal{P}} \,\right] &\geq \Pr\left[\, \mathcal{E}_{i-1} \,\right] \cdot \Pr\left[\, X_i^N \geq \tau_{\mathcal{P}} \mid \mathcal{E}_{i-1} \,\right] \\
&\geq \rho_{\mathcal{P}} \cdot \Pr\left[\, X_i^N \geq \tau_{\mathcal{P}} \mid \mathcal{E}_{i-1} \,\right] && \text{(Claim A.III.6)} \\
&= \rho_{\mathcal{P}} \cdot 1 && \text{(Claim A.III.3)} \\
&= \rho_{\mathcal{P}},
\end{aligned}$$

which implies (P).

A.III.4 Motivation for non-degeneracy assumption

Let us provide some motivation for why (3.18) is a mild assumption.

Fix a vector $n \in \mathbb{N}^d$ and define

$$\tau_{\mathcal{P}}(n) = \min_{A \neq A' \in \mathcal{A}} |\sum_{\ell \in A} \nu_\ell(n_\ell) - \sum_{\ell' \in A'} \nu_{\ell'}(n_{\ell'})|. \tag{A.18}$$

Our intuition is as follows: $\tau_{\mathcal{P}}(n)$ is defined as the smallest difference between $e^{O(d)}$ numbers in $[-d, d]$, so typical situation should be that $\tau_{\mathcal{P}}$ is on the order of $e^{-O(d)}$, whereas our assumption only requires it to be larger than $e^{-\Omega(d^2)}$.

We make this intuition precise, in a sense defined below. We argue that $\tau_{\mathcal{P}}(n)$ is "not too small" for "all but a few" problem instances. More formally, we define a distribution over problem instances such that $\tau_{\mathcal{P}}(n) \geq \Omega(c_2^{-d^2})$ with very high probability. For instance, we can make it hold with probability at least $1-\delta/2^d$ for some small $\delta>0$.

(However, we do not construct one distribution that works for all relevant vectors n at once, although we suspect that our technique, based on Esseen inequality, might be extended there.)

So, let us construct the desired distribution over problem instances. We fix d and the set of feasible arms, and we only vary the per-atom priors. Recall that the prior $\mathcal{P}_\ell$ for a given atom $\ell \in [d]$ is specified by a pair of numbers, $(\alpha_\ell,\, \beta_\ell)$. Further, recall that $\nu_\ell(n_\ell) = \alpha_\ell \,/\, (\alpha_\ell + \beta_\ell + n_\ell)$, $n_\ell \in \mathbb{N}$. We require that $\nu_\ell(n_\ell)$ is distributed uniformly on some interval.

Lemma A.III.7. *Fix vector $n \in \mathbb{N}^d$. Suppose for each atom $\ell \in [d]$, the pair $(\alpha_\ell, \beta_\ell)$ is drawn independently from some distribution such that $\nu_\ell(n_\ell)$ is uniformly distributed in some interval $[a_\ell, b_\ell]$. Fix $\delta \in (0,1)$. Then it holds that*

$$\Pr\left[\tau_{\mathcal{P}}(n) < \frac{\delta}{2\cdot 8^d}\right] \leq \frac{\delta}{2^d}. \tag{A.19}$$

Remark A.III.8. One way to ensure that $\nu_\ell(n_\ell)$ is uniformly distributed is as follows. Fix atom ℓ, and parameters β_ℓ and n_ℓ. Let $\nu_\ell = \nu_\ell(n_\ell)$. Note that

$$\alpha_\ell = \Phi(\nu_\ell) := \frac{\nu_\ell \cdot (\beta_\ell + n_\ell)}{1-\nu_\ell}. \tag{A.20}$$

Now, just let α_ℓ be distributed as $\Phi(Y)$, where Y is uniform on $[a_\ell, b_\ell]$ interval, where $0 \leq a_\ell \leq \alpha_\ell \leq b_\ell \leq 1$. Observe that by change-of-variable, when α_ℓ is distributed with $\Phi(Y)$, then ν_ℓ is distributed uniformly on $[a_\ell, b_\ell]$.

To prove Lemma A.III.7, we invoke the following tool from *anti-concentration.*

Theorem A.III.9 (Esseen inequality)**.** *Let Y be a random variable. Consider its* characteristic function,

$$\psi_Y(\lambda) = \mathbb{E}\left[e^{i\lambda Y}\right], \quad \lambda \in \mathbb{R}.$$

Then for any $x > 0$ it holds that

$$Q_Y(x) \coloneqq \sup_{y \in \mathbb{R}} \Pr\left[\, |Y - y| \leq x \,\right] \leq x \int_{-2\pi/x}^{2\pi/x} |\psi_Y(\lambda)| \, d\lambda.$$

Proof of Lemma A.III.7. Let $\nu_\ell = \nu_\ell(n_\ell)$ for each atom $\ell \in [d]$. Focus on

$$X := \sum_{\ell \in A} \nu_\ell - \sum_{\ell' \in A'} \nu_{\ell'}.$$

We treat X as a random variable, under the distribution over the $(\alpha_\ell, \beta_\ell)$ pairs. Without loss of generality, from here on assume that arms A and A' are disjoint subsets of atoms.

We will use Theorem A.III.9 to prove that $Q_X(x) \leq 2x$ for any $x > 0$.

Since X is a sum of independent random variables $\pm\nu_\ell$, $\ell \in A \cup A'$, the characteristic function of X is the product of the respective characteristic functions

$$\psi_X(\lambda) = \prod_{\ell \in A \cup A'} \psi_\ell(\lambda),$$

where $\psi_\ell(\lambda) = \psi_{\nu_\ell}(\lambda)$ for $\ell \in A$, and $\psi_\ell(\lambda) = \psi_{-\nu_\ell}(\lambda)$ for $\ell \in A'$.

From here on, fix some atom $\ell \in A$. Since $|\psi_Y(\lambda)| \leq 1$ for any random variable Y, it follows that

$$|\psi_X(\lambda)| \leq |\psi_{\nu_\ell}(\lambda)|.$$

For the rest of the proof we focus on the characteristic function for ν_ℓ, $\psi(\cdot) := \psi_{\nu_\ell}(\cdot)$.

Recall that ν_ℓ is distributed uniformly on some interval $[a, b] = [a_\ell, b_\ell]$. A known fact about characteristic function of uniform distribution is that

$$\psi(\lambda) = \frac{e^{i\lambda b} - e^{i\lambda a}}{i\lambda(b-a)}.$$

The rest of the proof is a simple but somewhat tedious integration. By Esseen inequality,

$$Q_X(x) \leq x \int_{-2\pi/x}^{2\pi/x} |\psi(\lambda)| d\lambda \tag{A.21}$$

$$= x \int_{-2\pi/x}^{2\pi/x} |\frac{e^{i\lambda b} - e^{i\lambda a}}{i\lambda(b-a)}| d\lambda \tag{A.22}$$

$$= x \int_{-2\pi/x}^{2\pi/x} \frac{\sqrt{2 - 2\cos(\lambda b - \lambda a)}}{|\lambda(b-a)|} d\lambda \tag{A.23}$$

Let $u = \lambda(b-a)$. Then by substitution we have:

$$Q_X(x) \leq x \int_{-2\pi(b-a)/x}^{2\pi(b-a)/x} \sqrt{\frac{2 - 2\cos(u)}{u^2}} du \tag{A.24}$$

$$= x \int_{-2\pi(b-a)/x}^{2\pi(b-a)/x} \sqrt{\frac{2 - 2\sum_{n=0}^{\infty} \frac{(-1)^n u^{2n}}{(2n)!}}{u^2}} du \qquad \text{(by Maclaurin series of } \cos(u))$$

$$= x \int_{-2\pi(b-a)/x}^{2\pi(b-a)/x} \sqrt{2}\sqrt{\frac{1}{2} - \frac{u^2}{4!} + \cdots} du \tag{A.25}$$

When we have $|u| \leq 1$, the terms in the integrand are decreasing. Hence, the entire integrand can be upper bounded by 1. Otherwise, when $|u| \geq 1$, we can upper bound the integrand by $2/|u|$. Hence, the concentration function $Q_X(t)$ is upper bounded by:

$$\begin{aligned}
Q_X(x) &\leq x \int_{-2\pi(b-a)/x}^{-1} \frac{2}{|u|} du + \int_{-1}^{1} 1 du + \int_{1}^{2\pi(b-a)/x} \frac{2}{|u|} du \\
&= x \int_{1}^{2\pi(b-a)/x} \frac{-2}{|v|} dv + \int_{-1}^{1} 1 du + \int_{1}^{2\pi(b-a)/x} \frac{2}{|u|} du \qquad \text{(substitution } v = -u) \\
&= x \left(-2\log(v) \Big|_{1}^{2\pi(b-a)/x} + 2 + 2\log(u) \Big|_{1}^{2\pi(b-a)/x} \right) \\
&= 2x
\end{aligned}$$

For $x = \delta/2 \cdot 8^d$, we have $Q_X(\delta/2 \cdot 8^d) \leq \delta/8^d$. Observe that since A and A' are subsets of atoms, there are at most 2^d possible choices for each arm A and A'. Hence, for a fixed vector n, there are 4^d possible values of X. By union bound, we have $\Pr[\tau_{\mathcal{P}}(n) \leq \delta/2 \cdot 8^d] \leq \delta/2^d$. $\square$

A.IV Probabilistic tools from prior work

In this appendix, we spell out some probabilistic tools from prior work that we rely on.

Bayesian Chernoff Bound

We use an easy corollary of the Bayesian Chernoff Bound from Sellke and Slivkins (2021a).

Lemma A.IV.1 (Sellke and Slivkins (2021a)). *Fix round t and parameters $\varepsilon, \tau > 0$. Suppose algorithm's history $\mathcal{F}_t$ almost surely contains at least ε^{-2} samples of each atom. Let $\tilde{\theta}$ be a posterior sample for the mean reward θ,* i.e., *$\tilde{\theta}$ is an independent sample from the posterior distribution on θ given $\mathcal{F}_t$. Then for some universal absolute constant C, we have*

$$\Pr\left[\left\|\tilde{\theta} - \theta\right\| \geq \tau\varepsilon \right] \leq C \cdot e^{-\tau^2/C}, \tag{A.26}$$

$$\Pr\left[\|\mathbb{E}[\theta|\mathcal{F}_t] - \theta\| \geq \tau\varepsilon \right] \leq C \cdot e^{-\tau^2/C}. \tag{A.27}$$

Proof. Sellke and Slivkins (2021a) contains this result for $d = 1$ atoms. Here, we apply the result from Sellke and Slivkins (2021a) to each atom separately, using the fact that the Bayesian update is independent across atoms. □

Harris Inequality

We invoke Harris Inequality about correlated random variables.

Theorem A.IV.2 (Harris (1960)). *Let $f, g : \mathbb{R}^n \to \mathbb{R}$ be nondecreasing functions. Let $X_1, \cdots, X_n$ be independent real-valued random variables and define the random vector $X = (X_1, \cdots, X_n)$ taking values in $\mathbb{R}^n$. Then*

$$\mathbb{E}\left[f(X)\ g(X) \right] \geq \mathbb{E}[f(X)]\ \mathbb{E}[g(X)]$$

Similarly, if f is nonincreasing and g is nondecreasing then

$$\mathbb{E}\left[f(X)\ g(X) \right] \leq \mathbb{E}[f(X)]\ \mathbb{E}[g(X)]$$

Remark A.IV.3 (Mixed-monotonicity Harris inequality)**.** If f and g are both increasing or both decreasing in each coordinate, then the results of Theorem A.IV.2 still hold since we can simply negate some coordinates in the parameterization, i.e. we view f and g as increasing function of $-x_i$ instead of decreasing function of x_i. We refer to this in the proof as the mixed-monotonicity Harris inequality to highlight this subtle modification.

Tails of sub-Gaussian distribution

Lemma A.IV.4. *If random variable X is $O(1)$-sub-Gaussian and event E has probability* $\Pr[E] \leq p$*, then* $\mathbb{E}[|X \cdot 1_E|] \leq O(p\sqrt{\log(1/p)})$

A.V Initial exploration: reduction to Incentivized RL

This appendix spells out the analysis for Section 3.4.2: the approach for initial exploration by reduction to incentivized reinforcement learning (RL) (Simchowitz and Slivkins, 2021a).

In this section, we build on an algorithm from Simchowitz and Slivkins (2021a), called `HiddenHallucination`, and their guarantee for this algorithm. We state their setup and guarantee below. (The specification of their algorithm is unimportant for our presentation.) Then we use it to prove Theorem 3.4.6.

Incentivized RL: the setup and the guarantee

The setting is as follows. Consider an MDP with S states, A actions and H stages, where H is the time horizon. We write $x \in [S]$, $a \in [A]$, and $h \in [H]$ to represent states, actions, and stages, respectively. In the following analysis, we often refer to (x, a, h) triples. We consider a set of feasible (x, a, h) triples called $\texttt{FEASIBLE} \subset [S] \times [A] \times [H]$. (In Simchowitz and Slivkins (2021a), $\texttt{FEASIBLE} = [S] \times [A] \times [H]$, but we will extend their result to an arbitrary `FEASIBLE`.)

The "true" MDP is denoted by ϕ. Let $r_\phi(x, a, h)$ be the expected reward if action a is chosen at state x and stage h. We posit a Bayesian model: ϕ is chosen from a Bayesian prior $\mathcal{P}$.

Then, we consider the setting of *episodic RL*, where in each episode t an algorithm chooses a policy $\pi^{(t)}$ in this MDP.[1] The chosen policies must satisfy a similar BIC condition: for each round $t \in [T]$,

$$\mathbb{E}[\, V(\pi) - \mu(\pi') \mid \pi^{(t)} = \pi \,] \geq 0 \qquad \forall \text{ policies } \pi, \pi' \in \mathcal{A} \text{ with } \Pr[\pi^{(t)} = \pi] > 0,$$

where $V(\pi)$ is the value (expected reward) of policy π. Essentially, this is the same condition as (3.1), where arms are replaced with policies.

We only need the guarantee for `HiddenHallucination` for an MDP with deterministic transitions (but randomized rewards). This guarantee depends on the following prior-dependent quantities:

$$\mathrm{q_{pun}}(\varepsilon) \coloneqq \Pr[r_\phi(x,a,h) \leq \varepsilon, \forall (x,a,h) \in \texttt{FEASIBLE}], \tag{A.28}$$

$$\mathrm{r_{alt}} \coloneqq \min_{(x,a,h)\in\texttt{FEASIBLE}} \mathbb{E}[r_\phi(x,a,h)]. \tag{A.29}$$

The guarantee is stated as follows.

Theorem A.V.1. *Consider an arbitrary prior $\mathcal{P}$. Fix parameters $\delta \in (0,1]$. Assume that $\mathrm{r_{alt}} > 0$ and $\mathrm{q_{pun}} = \mathrm{q_{pun}}(\varepsilon_{\mathrm{pun}}) > 0$, where $\varepsilon_{\mathrm{pun}} = \mathrm{r_{alt}}/18H$. Consider* `HiddenHallucination` *with punishment parameter $\varepsilon_{\mathrm{pun}}$, appropriately chosen phase length $\mathrm{n_{ph}}$, and large enough target $n = \mathrm{n_{lrn}}$. This algorithm is guaranteed to explore all $(x,a,h) \in$* `FEASIBLE` *with probability at least $1-\delta$ by round N_0, where n and N_0 are specified below.*

For some absolute constants c_1, c_2, it suffices to take

$$n = n_{\mathrm{lrn}} \geq c_1 \cdot \mathrm{r_{alt}^{-2}} H^4 \left(S + \log \frac{SAH}{\delta \cdot \mathrm{r_{alt}} \cdot \mathrm{q_{pun}}} \right),$$
$$N_0 = c_2 \cdot n \cdot \mathrm{q_{pun}} \cdot \mathrm{r_{alt}^{-3}} \cdot SAH^4$$

In particular, for any $n \geq 1$, one can obtain $N_0 \leq n \cdot \mathrm{q_{pun}} \cdot \mathrm{poly}\left(\mathrm{r_{alt}^{-1}} SAH\right) \cdot \log\left(\delta^{-1} \mathrm{q_{pun}^{-1}}\right)$.

As we mentioned above, Simchowitz and Slivkins (2021a) guarantees Theorem A.V.1

[1] By default, MDP policies are Markovian and deterministic.

for `FEASIBLE` $= [S] \times [A] \times [H]$. Below, we show how to extend it to an arbitrary `FEASIBLE` $= [S] \times [A] \times [H]$.

Theorem A.V.2 (Simchowitz and Slivkins (2021a))**.** *The guarantee in Theorem A.V.1 holds for* `FEASIBLE` $= [S] \times [A] \times [H]$.

Remark A.V.3. The relevant result, Theorem 5.5 in Simchowitz and Slivkins (2021a) is stated for MDPs with randomized transitions. In this more general formulation, (A.28) and (A.29) are conditioned on an object called *censored ledger*, and then a minimum is taken over all such objects. However, this conditioning vanishes when the MDP transitions are deterministic. (This follows easily from Lemma 6.2 in Simchowitz and Slivkins (2021a), essentially because censored ledgers do not carry any useful information.) We present a version without censored ledgers, because defining them is quite tedious.

Proof Sketch for Theorem A.V.1. Start with an arbitrary `FEASIBLE`. We modify the MDP as follows. Add two terminal state, `GOOD` and `BAD`, such that where we deterministically transition into `BAD` if $(x, a, h) \notin$ `FEASIBLE`. Otherwise, at the end of the MDP, we go into `GOOD`. We let `BAD` yield reward 0, and `GOOD` yield reward $H + 1$. With this modified MDP, even if all (x, a, h) triples are allowed, any BIC algorithm would only choose feasible policies, *i.e.,* policies that only choose $(x, a, h) \in$ `FEASIBLE`. So, we conclude by invoking Theorem A.V.2. □

Proof of Theorem 3.4.6 for combinatorial semi-bandits

Now, let us go back to combinatorial semi-bandits and prove Theorem 3.4.6. We start with an instance of combinatorial semi-bandits and construct an MDP as specified in Section 3.4.2. Then we invoke Theorem A.V.1. To state the final guarantee, it remains to interpret (and simplify) the quantities in Theorem A.V.1 for a particular MDP obtained with our construction.

First, recall that $H = A = d$ and $S \leq d$, where d is the number of atoms.

Second, $r_\phi(x, a, h)$ is simply θ_ℓ, the expected reward of the corresponding atom ℓ. Accordingly,

$$\mathrm{r}_{\text{alt}} = \min_{\text{atoms } \ell \in [d]} \mathbb{E}[\theta_\ell] \geq \min_{\text{priors } \mathcal{P}_\ell \in \mathcal{C}} \mathop{\mathbb{E}}_{\theta_\ell \sim \mathcal{P}_\ell} [\theta_\ell] := \varepsilon_0.$$

Note that ε_0 is determined by the collection $\mathcal{C}$ of feasible per-atom priors.

Finally, observe that $\mathrm{q}_{\mathrm{pun}}$ is the probability of all (x, a, h) triples have low reward:

$$\mathrm{q}_{\mathrm{pun}}(\varepsilon) = \Pr[\theta_\ell \leq \varepsilon : \; \forall \text{ atoms } \ell \in [d]]$$

We can divide all (x, h, a) triples into classes, where each class represents an atom. Since our prior is independent across the atoms,

$$\begin{aligned}
\mathrm{q}_{\mathrm{pun}} &= \mathrm{q}_{\mathrm{pun}}(\mathrm{r}_{\mathrm{alt}}) \geq \mathrm{q}_{\mathrm{pun}}(\varepsilon_0) \\
&= \Pr\left[\, \theta_\ell \leq \varepsilon_0 : \; \forall \text{ atoms } \ell \in [d] \,\right] \\
&= \prod_{\ell \in [d]} \Pr\left[\, \theta_\ell \leq \varepsilon_0 \,\right] \\
&\geq \left(\min_{\text{priors } \mathcal{P}_\ell \in \mathcal{C}} \; \Pr_{\theta_\ell \sim \mathcal{P}_\ell} \left[\, \theta_\ell \leq \varepsilon_0 \,\right] \right)^d .
\end{aligned}$$

Again, the expression in $(\cdot)$ is determined by the collection $\mathcal{C}$.

Appendix B

Appendix for Incentivizing Compliance with Algorithmic Instruments

B.I Theorems and Lemmas

Theorem B.I.1. *(Chernoff Bound for unbounded sub-Gaussian random variables) Let $X_1, \ldots, X_n$ be independent sub-Gaussian random variables with parameter σ. Let $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. For all $\varepsilon > 0$,*

$$\mathbb{P}\left[\left|\overline{X}\right| \geq \varepsilon\right] \leq \exp\left\{\frac{-n\varepsilon^2}{2\sigma^2}\right\}.$$

Corollary B.I.2. *(High probability bound on the sum of unbounded sub-Gaussian random variables) For any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\left|\overline{X}\right| < \sigma\sqrt{\frac{2\log(1/\delta)}{n}}$$

Theorem B.I.3. *(Chernoff/Hoeffding's inequality) Let $X_1, \ldots, X_n$ be independent and bounded random variables such that $a \leq X_i \leq b$ for all i. Then*

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X_i] \geq \varepsilon\right] \leq \exp\left(\frac{-2n\varepsilon^2}{(b-a)^2}\right)$$

Corollary B.I.4. *(High probability upper bound on the sum of bounded random variables) For any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\mathbb{E}[X] - \frac{1}{n}\sum_{i=1}^{n} X_i \leq (b-a)\sqrt{\frac{\log(1/\delta)}{2n}},$$

where $X_i \in [a,b]$ for all i from 1 to n.

Lemma B.I.5. *(Cauchy-Schwarz Inequality) For any n-dimensional vectors $u, v \in \mathbb{R}^n$, the L^2-norm of the inner product of u and v is less than or equal to the L^2-norm of u times the L^2-norm of v, i.e.*

$$\|\langle u, v\rangle\|_2 \leq \|u\|_2 \cdot \|v\|_2 \, .$$

Alternatively, for any $m \times n$-dimensional matrices $A \in \mathbb{R}^{m\times n}$ and n-dimensional vector $v \in \mathbb{R}^n$, the L^2-norm of the dot product of A and v is less than or equal to the spectral norm of A times the L^2-norm of v, i.e.

$$\|Av\|_2 \leq \|A\|_2 \cdot \|v\|_2 \, .$$

Theorem B.I.6. *(Matrix Chernoff) Consider a finite sequence X_k of independent, random, self-adjoint matrices with common dimension d. Assume that:*

$$0 \leq \lambda_{\min}(X_k) \quad \textit{and} \quad \lambda_{\max}(X_k) \leq \omega \quad \textit{for each index } k.$$

Introduce the random matrix $Y = \sum_k X_k$. Define the minimum eigenvalue μ_{min} and maximum eigenvalue μ_{max} of the expectation $\mathbb{E}[Y]$.

$$\mu_{min} = \lambda_{\min}\{\mathbb{E}[Y]\} = \lambda_{\min}\left\{\sum_k \mathbb{E}[X_k]\right\}, \quad \textit{and}$$
$$\mu_{\max} = \lambda_{\max}\{\mathbb{E}[Y]\} = \lambda_{\max}\left\{\sum_k \mathbb{E}[X_k]\right\}$$

Then, for $\theta > 0$,

$$\mathbb{E}[\lambda_{\min}(Y)] \geq \frac{1-e^{-\theta}}{\theta}\mu_{\min} - \frac{1}{\theta}L\log d, \quad \textit{and}$$
$$\mathbb{E}[\lambda_{\max}(Y)] \leq \frac{e^{\theta}-1}{\theta}\mu_{\max} + \frac{1}{\theta}L\log d$$

Furthermore,

$$\mathbb{P}[\lambda_{\min}(Y) \leq (1-\varepsilon)\mu_{\min}] \leq d\left[\frac{e^{-\varepsilon}}{(1-\varepsilon)^{1-\varepsilon}}\right]^{\mu_{\min}/\omega} \quad \textit{for } \varepsilon \in [0,1)$$
$$\mathbb{P}[\lambda_{\max}(Y) \leq (1+\varepsilon)\mu_{\max}] \leq d\left[\frac{e^{\varepsilon}}{(1+\varepsilon)^{1+\varepsilon}}\right]^{\mu_{\max}/\omega} \quad \textit{for } \varepsilon \geq 0$$

Theorem B.I.7. *(Union bound): For a countable set of events $A_1, A_2, \ldots$, we have*

$$\mathbb{P}\left[\bigcup_i A_i\right] \leq \sum_i \mathbb{P}(A_i)$$

B.II IV Estimator Proof for Control-Treatment Setting

Recall that our reward model can be stated as the following equation:

$$y_i = \theta x_i + g_i^{(u_i)} \tag{B.1}$$

To analyze the Wald estimator, we introduce two conditional probabilities that an agent chooses the treatment given a recommendation $\hat{\gamma}_0$ and $\hat{\gamma}_1$, given as proportions over a set of n samples $(x_i, z_i)_{i=1}^n$ and formally defined as

$$\hat{\gamma}_0 = \underset{(x_i,z_i)_{i=1}^n}{\hat{\mathbb{P}}}[x_i = 1 | z_i = 0] = \frac{\sum_{i=1}^n x_i(1-z_i)}{\sum_{i=1}^n (1-z_i)^2} \quad \text{and}$$
$$\hat{\gamma}_1 = \underset{(x_i,z_i)_{i=1}^n}{\hat{\mathbb{P}}}[x_i = 1 | z_i = 1] = \frac{\sum_{i=1}^n x_i z_i}{\sum_{i=1}^n z_i^2}$$

Then, we can write the action choice x_i as such:

$$\begin{aligned} x_i &= \hat{\gamma}_1 z_i + \hat{\gamma}_0(1 - z_i) + \eta_i \\ &= \hat{\gamma} z_i + \hat{\gamma}_0 + \eta_i \end{aligned}$$

where $\eta_i = x_i - \hat{\gamma}_1 z_i - \hat{\gamma}_0(1 - z_i)$ and $\hat{\gamma} = \hat{\gamma}_1 - \hat{\gamma}_0$ is the in-sample *compliance coefficient*. Now, we can rewrite the reward y_i as

$$\begin{aligned} y_i &= \theta\left(\hat{\gamma} z_i + \hat{\gamma}_0 + \eta_i\right) + g_i^{(u_i)} \\ &= \underbrace{\theta\,\hat{\gamma}}_{\beta} z_i + \theta\hat{\gamma}_0 + \theta\eta_i + g_i^{(u_i)} \end{aligned}$$

Let operator $\bar{\cdot}$ denote the sample mean, e.g. $\bar{y} := \frac{1}{n}\sum_{i=1}^n y_i$ and $\bar{g} := \frac{1}{n}\sum_{i=1}^n g_i^{(u_i)}$. $\bar{\eta} = \frac{1}{n}\sum_{i=1}^n \eta_i = 0$, by definition.

Then,

$$\bar{y} = \beta\bar{z} + \theta\hat{\gamma}_0 + \theta\bar{\eta} + \bar{g} + \bar{\varepsilon}$$

Thus, the centered reward and treatment choice at round i are given as:

$$\begin{cases} y_i - \bar{y} = \theta(x_i - \bar{x}) + g_i^{(u_i)} - \bar{g} \\ y_i - \bar{y} = \beta(z_i - \bar{z}) + \theta(\eta_i - \bar{\eta}) + g_i^{(u_i)} - \bar{g} \\ x_i - \bar{x}_i = \hat{\gamma}(z_i - \bar{z}) + \eta_i - \bar{\eta} \end{cases} \tag{B.2}$$

This formulation of the centered reward $y_i - \bar{y}$ allows us to express and bound the error between the treatment effect θ and its instrumental variable estimate $\hat{\theta}_S$, which we show in the following Theorem 4.2.1.

Theorem 4.2.1 (Finite-sample error bound for Wald estimator)**.** *Let $z_1, z_2, \ldots, z_n \in \{0, 1\}$ be a sequence of instruments. Suppose there is a sequence of n agents such that each agent i has their private type u_i drawn independently from $\mathcal{U}$, selects action x_i under instrument z_i, and receives reward y_i. Let sample set $S = (x_i, y_i, z_i)_{i=1}^n$. Let*

$A : (\{0,1\}^n \times \{0,1\}^n \times \mathbb{R}^n) \to \mathbb{R}$ denote the approximation bound for set S, such that

$$A(S, \delta) := \frac{2\sigma_g \sqrt{2n \log(2/\delta)}}{|\sum_{i=1}^n (x_i - \bar{x})(z_i - \bar{z})|}$$

and the Wald estimator given by (4.3) satisfies

$$\left| \hat{\theta}_S - \theta \right| \leq A(S, \delta)$$

with probability at least $1 - \delta$, for any $\delta \in (0, 1)$.

Proof. Given a sample set $S = (x_i, y_i, z_i)_{i=1}^n$ of size n, we form an estimate of the treatment effect $\hat{\theta}_S$ via a Two-Stage Least Squares (2SLS). In the first stage, we regress $y_i - \bar{y}$ onto $z_i - \bar{z}$ to get the empirical estimate $\hat{\beta}_S$ and $x_i - \bar{x}$ onto $z_i - \bar{z}$ to get $\hat{\gamma}_S$ as such:

$$\hat{\beta}_S := \frac{\sum_{i=1}^n (y_i - \bar{y})(z_i - \bar{z})}{\sum_{i=1}^n (z_i - \bar{z})^2} \qquad \text{and} \qquad \hat{\gamma}_S := \frac{\sum_{i=1}^n (x_i - \bar{x})(z_i - \bar{z})}{\sum_{i=1}^n (z_i - \bar{z})^2} \tag{B.3}$$

In the second stage, we take the quotient of these two empirical estimates as the predicted treatment effect $\hat{\theta}_S$, i.e.

$$\begin{aligned} \hat{\theta}_S = \frac{\hat{\beta}_S}{\hat{\gamma}_S} &= \left(\frac{\sum_{i=1}^n (y_i - \bar{y})(z_i - \bar{z})}{\sum_{i=1}^n (z_i - \bar{z})^2} \right) \left(\frac{\sum_{i=1}^n (z_i - \bar{z})^2}{\sum_{i=1}^n (x_i - \bar{x})(z_i - \bar{z})} \right) \\ &= \frac{\sum_{i=1}^n (y_i - \bar{y})(z_i - \bar{z})}{\sum_{i=1}^n (x_i - \bar{x})(z_i - \bar{z})} \end{aligned} \tag{B.4}$$

Next, we can express the absolute value of the difference between the true treatment effect θ and the IV estimate of the treatment effect $\hat{\theta}_S$ given a sample set S of size n as such:

$$
\begin{aligned}
\left|\hat{\theta}_S - \theta\right| &= \left|\frac{\sum_{i=1}^n (y_i - \bar{y})(z_i - \bar{z})}{\sum_{i=1}^n (x_i - \bar{x})(z_i - \bar{z})} - \theta\right| \\
&= \left|\frac{\sum_{i=1}^n \left(\theta(x_i - \bar{x}) + g_i^{(u_i)} - \bar{g}\right)(z_i - \bar{z})}{\sum_{i=1}^n (x_i - \bar{x})(z_i - \bar{z})} - \theta\right| && \text{(by Equation (B.2))} \\
&= \left|\theta + \frac{\sum_{i=1}^n \left(g_i^{(u_i)} - \bar{g}\right)(z_i - \bar{z})}{\sum_{i=1}^n (x_i - \bar{x})(z_i - \bar{z})} - \theta\right| \\
&= \frac{\left|\sum_{i=1}^n \left(g_i^{(u_i)} - \bar{g}\right)(z_i - \bar{z})\right|}{\left|\sum_{i=1}^n (x_i - \bar{x})(z_i - \bar{z})\right|} && \text{(B.5)}
\end{aligned}
$$

In order to complete our proof, we demonstrate an upper bound on the numerator $\left|\sum_{i=1}^n \left(g_i^{(u_i)} - \bar{g}\right)(z_i - \bar{z})\right|$ of Equation (B.5) in the last line above in Lemma B.II.1. □

Lemma B.II.1. *For all $\delta \in (0, 1)$, with probability at least $1 - \delta$, we have*

$$\left|\sum_{i=1}^n \left(g_i^{(u_i)} - \bar{g}\right)(z_i - \bar{z})\right| \le 2\sigma_g \sqrt{2n \log(2/\delta)} \tag{B.6}$$

if the set of $g_i^{(u_i)}$ are i.i.d. sub-Gaussian random variables with sub-Gaussian norm σ_g.

Proof. We can rewrite the left hand side as follows

$$
\begin{aligned}
&\left|\sum_{i=1}^{n}\left(g_i^{(u_i)}-\bar{g}\right)(z_i-\bar{z})\right| \\
&=\left|\sum_{i=1}^{n}\left(g_i^{(u_i)}-\mathbb{E}[g^{(u)}]+\mathbb{E}[g^{(u)}]-\bar{g}\right)(z_i-\bar{z})\right| \\
&=\left|\sum_{i=1}^{n}\left(g_i^{(u_i)}-\mathbb{E}[g^{(u)}]\right)z_i-\sum_{i=1}^{n}\left(g_i^{(u_i)}-\mathbb{E}[g^{(u)}]\right)\bar{z}+\sum_{i=1}^{n}\left(\mathbb{E}[g^{(u)}]-\bar{g}\right)(z_i-\bar{z})\right| \\
&=\left|\sum_{i=1}^{n}\left(g_i^{(u_i)}-\mathbb{E}[g^{(u)}]\right)z_i-\sum_{i=1}^{n}\left(g_i^{(u_i)}-\mathbb{E}[g^{(u)}]\right)\bar{z}\right| && \left(\textstyle\sum_{i=1}^{n}(z_i-\bar{z})=0\right) \\
&\le\left|\sum_{i=1}^{n}\left(g_i^{(u_i)}-\mathbb{E}[g^{(u)}]\right)z_i\right|+\left|\sum_{i=1}^{n}\left(g_i^{(u_i)}-\mathbb{E}[g^{(u)}]\right)\right| \\
& && \text{(by the triangle inequality and } |\bar{z}|\le 1)
\end{aligned}
$$

Now, if $g_i^{(u_i)}$ is sub-Gaussian, then the last line in the system of inequalities above is given as:

$$
\begin{aligned}
&\left|\sum_{i=1}^{n}\left(g_i^{(u_i)}-\mathbb{E}[g^{(u)}]\right)z_i\right|+\left|\sum_{i=1}^{n}\left(g_i^{(u_i)}-\mathbb{E}[g^{(u)}]\right)\right| \\
&\le\left|\sigma_g\sqrt{2n_1\log(1/\delta_1)}\right|+\left|\sigma_g\sqrt{2n\log(1/\delta_2)}\right| \\
& && \text{(by Corollary B.I.2, where } n_1:=\textstyle\sum_{i=1}^{n}z_i) \\
&\le 2\sigma_g\sqrt{2n\log(2/\delta)} && \text{(since } n_1\le n \text{ and by Theorem B.I.7, where } \delta_1=\delta_2=\delta/2)
\end{aligned}
$$

This recovers the stated bound and finishes the proof for Theorem 4.2.1. □

Next, we demonstrate a lower bound on the denominator of Theorem 4.2.1, in terms of the level of compliance at each phase of Algorithms 1 and 2.

Theorem B.II.2 (Lower bound on $|\sum_{i=1}^{n}(x_i-\bar{x})(z_i-\bar{z})|$ for a type 0 compliant sample set)**.** *Let $S=(x_i,y_i,z_i)_{i=1}^{n}$ denote a sample set which satisfies the conditions of Theorem 4.2.1. Furthermore, assume that there are p_c fraction of agents in the population who would be compliant. Recall that $\bar{z}=\frac{1}{n}\sum_{i=1}^{n}z_i$ and $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$. Then, the denominator of the approximation bound $A(S,\delta)$ (from Theorem 4.2.1) is lower bounded* *as such:*

$$\left|\sum_{i=1}^{n}(x_i-\bar{x})(z_i-\bar{z})\right| \geq \begin{cases} n\bar{z}(1-\bar{z}) & \textit{if } p_c = 1; \\ n\bar{z}(1-\bar{z})p_c - (3-\bar{z})\sqrt{\frac{n\bar{z}\log(3/\delta)}{2(1-\bar{z})}} & \textit{otherwise.} \end{cases}$$

The second case above occurs with probability at least $1-\delta$ *for any* $\delta \in (0,1)$.

Proof. In this theorem, we formulate the denominator of the approximation bound in Theorem 4.2.1 in terms of $\bar{z}$, since $\bar{z}$ is determined by the social planner. For any type u, let $u \in U_c$ denote that agents of type u comply; let $u \in U_0$ denote that agents of type u are never-takers (agents which prefer control, according to their prior); and let $u \in U_1$ denote that agents of type u are always-takers (agents which prefer treatment, according to their prior). Let p_0 and p_1 be the fractions of never-takers and always-takers, respectively.

Next, we expand the binomial in the denominator and arrive at the following simplified form:

$$\left|\sum_{i=1}^{n}(x_i-\bar{x})(z_i-\bar{z})\right| = \left|\sum_{i=1}^{n} x_i z_i - \bar{z}\sum_{i=1}^{n} x_i - \bar{x}\sum_{i=1}^{n} z_i + \bar{z}\bar{x}\right| = \left|\sum_{i=1}^{n} x_i z_i - n\bar{z}\bar{x}\right| \tag{B.7}$$

First, observe that at any round i, the product $x_i = 1$ only when agent i is a non-compliant always-taker or when $z_i = 1$ and agent i is compliant. Formally, for any agent i, action choice $x_i = 1$ is equivalent to the following:

$$x_i = 1 \equiv (z_i = 1 \wedge u_i \in U_c) \vee (u_i \in U_1 \wedge u_i \notin U_c) \tag{B.8}$$

Then, the sum $\sum_{i=1}^n x_i$ can be expressed as follows:

$$\begin{aligned}
\sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} \mathbb{1}\left[(z_i = 1 \wedge u_i \in U_c) \vee (u_i \in U_1 \wedge u_i \notin U_c)\right] \\
&= \sum_{i=1}^{n} \mathbb{1}\left[(z_i = 1 \wedge u_i \in U_c\right] + \sum_{i=1}^{n} \mathbb{1}\left[u_i \in U_1 \wedge u_i \notin U_c\right] \\
&= \left(\sum_{i=1}^{n} z_i\right)\left(\hat{p}_{c:z_i=1} + n\hat{p}_{nc1}\right) \\
&= n\left(\bar{z}\hat{p}_{c:z_i=1} + \hat{p}_{nc1}\right) \qquad \text{(B.9)}
\end{aligned}$$

where we define $\hat{p}_{c:z_i=1}$ as the empirical proportion of agents with types in U_c when the recommendation $z = 1$ and $\hat{p}_{nc1}$ as the empirical proportion of non-compliant always-takers. Formally, $\hat{p}_{c:z_i=1} = \frac{1}{n}\sum_{i=1}^n \mathbb{1}[u_i \in U_c, z_i = 1]$ and $\hat{p}_{nc1} = \frac{1}{n}\sum_{i=1}^n \mathbb{1}[u_i \in U_1 \wedge u_i \notin U_c]$. Define p_{nc1} to be the proportion of non-compliant always-takers in the population of agents. Then, in expectation over the randomness of how agents arrive, $\mathbb{E}[\hat{p}_{c:z_i=1}] = p_c$ and $\mathbb{E}[\hat{p}_{nc1}] = p_{nc1}$.

Next, we rewrite the sum $\sum_{i=1}^n x_i z_i$ in terms of $\bar{z}$ and some population constants. Observe that at any round i, the product $x_i z_i = 1$ only when both $x_i = 1$ and $z_i = 1$. Thus, by Equation (B.8), for any agent i, the event $x_i z_i = 1$ is equivalent to the following:

$$\begin{aligned}
x_i z_i = 1 &\equiv z_i = 1 \wedge ((z_i = 1 \wedge u_i \in U_c) \vee (u_i \in U_1 \wedge u_i \notin U_c)) \\
&\equiv z_i = 1 \wedge (u_i \in U_c \vee (u_i \in U_1 \wedge u_i \notin U_c))
\end{aligned}$$

Then, the sum $\sum_{i=1}^n x_i z_i$ can be expressed as follows:

$$\begin{aligned}
\sum_{i=1}^{n} x_i z_i &= \sum_{i=1}^{n} \mathbb{1}\left[z_i = 1 \wedge (u_i \in U_c \vee (u_i \in U_1 \wedge u_i \notin U_c))\right] \\
&= \left(\sum_{i=1}^{n} \mathbb{1}[z_i = 1]\right)\left(\hat{p}_c|_{z_i=1} + \hat{p}_{nc1}|_{z_i=1}\right) \\
&= n\bar{z}\left(\hat{p}_{c:z_i=1} + \hat{p}_{nc1:z_i=1}\right) \qquad \text{(B.10)}
\end{aligned}$$

where we define $\hat{p}_{nc1:z_i=1}$ as the empirical proportions of non-compliant always-takers who arrive when $z_i = 1$ —i.e. $\hat{p}_{nc1:z_i=1} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[u_i \in U_1 \wedge u_i \notin U_c, z_i = 1]$. In expectation over the randomness of how agents arrive, $\mathbb{E}[\hat{p}_{nc1:z_i=1}] = p_{nc1}$.

Finally, by Equations (B.7), (B.9), and (B.10), we can provide a high probability lower bound on the denominator as such:

$$
\begin{aligned}
&\left|\sum_{i=1}^{n}(x_i - \bar{x})(z_i - \bar{z})\right| \\
&= \left|\sum_{i=1}^{n} x_i z_i - n\bar{z}\bar{x}\right| && \text{(by Equation (B.7))} \\
&= \left|\sum_{i=1}^{n} x_i z_i - n\bar{z}\bar{x}\right| && \text{(by Equation (B.7))} \\
&= \left|n\bar{z}\left(\hat{p}_{c:z_i=1} + \hat{p}_{nc1:z_i=1}\right) - n\bar{z}\left(\bar{z}\hat{p}_{c:z_i=1} + \hat{p}_{nc1}\right)\right| && \text{(by Equations (B.9) and (B.10))} \\
&= \left|n\bar{z}\left((1-\bar{z})\hat{p}_{c:z_i=1} + \hat{p}_{nc1:z_i=1} - \hat{p}_{nc1}\right)\right| \\
&\geq \left|n\bar{z}\left((1-\bar{z})\left(p_c - \sqrt{\frac{\log(1/\delta_1)}{2n\bar{z}}}\right) + p_{nc1} - \sqrt{\frac{\log(1/\delta_2)}{2n\bar{z}}} - \left(p_{nc1} + \sqrt{\frac{\log(1/\delta_3)}{2n}}\right)\right)\right| \\
& && \text{(by Theorem B.I.3)} \\
&\geq \left|n\bar{z}\left((1-\bar{z})p_c - (1-\bar{z})\sqrt{\frac{\log(3/\delta)}{2n\bar{z}}} - \sqrt{\frac{\log(3/\delta)}{2n\bar{z}}} - \sqrt{\frac{\log(3/\delta)}{2n}}\right)\right| \\
& && \text{(by Theorem B.I.7 where } \delta_1 = \delta_2 = \delta_3 = \delta/3) \\
&\geq \left|n\bar{z}\left((1-\bar{z})p_c - (3-\bar{z})\sqrt{\frac{\log(3/\delta)}{2n\bar{z}}}\right)\right| \\
&= n\bar{z}(1-\bar{z})p_c - (3-\bar{z})\sqrt{\frac{n\bar{z}\log(3/\delta)}{2(1-\bar{z})}}
\end{aligned}
$$

with probability at least $1 - \delta$ for any $\delta \in (0, 1)$. □

B.III Missing Proofs for Section 3.8

Claim B.III.1. *For any agent t at round t with recommendation policy π_t with a positive probability of recommending either control or treatment, according to the prior $\mathcal{P}^{(u_t)}$, i.e. $\mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t = 0] > 0$ and $\mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t = 1] > 0$. Furthermore, $a^{(u)}$ and $b^{(u)}$ denote the initially preferred and unpreferred actions for any type u, i.e. $a^{(u)} := \mathbb{1}[\mathbb{E}_{\mathcal{P}^{(u)}}[\theta] \geq 0]$ and*

$b^{(u)} := \mathbb{1}[\mathbb{E}_{\mathcal{P}^{(u)}}[\theta] < 0]$. *Formally, the following holds:*

$$\left\{(-1)^{a^{(u_t)}} \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|z_t = b^{(u_t)}] \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t = b^{(u_t)}] \geq 0\right\}$$
$$\Rightarrow \left\{(-1)^{b^{(u_t)}} \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|z_t = a^{(u_t)}] \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t = a^{(u_t)}] < 0\right\}$$

Proof. Note that $a^{(u)}$ is defined in such a way that $(-1)^{a^{(u)}} \mathbb{E}_{\mathcal{P}^{(u)}}[\theta] < 0$ always: if agents of type u prefer initially control, then $a^{(u)} = 0$ and $(-1)^{a^{(u)}} \mathbb{E}_{\mathcal{P}^{(u)}}[\theta] = \mathbb{E}_{\mathcal{P}^{(u)}}[\theta] < 0$; if agents of type u initially prefer treatment, then $a^{(u)} = 1$ and $(-1)^{a^{(u)}} \mathbb{E}_{\mathcal{P}^{(u)}}[\theta] = -\mathbb{E}_{\mathcal{P}^{(u)}}[\theta] < 0$. Then,

Recall that we assume that type 0 agents prefer the control, i.e. the expected treatment effect $\mathop{\mathbb{E}}_{\mathcal{P}^{(0)}}[\theta] < 0$. Then:

$$\begin{aligned}
&(-1)^{a^{(u_t)}} \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|z_t = b^{(u_t)}] \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t = b^{(u_t)}] \\
&\quad - (-1)^{b^{(u_t)}} \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|z_t = a^{(u_t)}] \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t = a^{(u_t)}] \\
&= (-1)^{a^{(u_t)}} \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|z_t = b^{(u_t)}] \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t = b^{(u_t)}] \\
&\quad + (-1)^{a^{(u_t)}} \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|z_t = a^{(u_t)}] \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t = a^{(u_t)}] \\
&= (-1)^{a^{(u_t)}} \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta] < 0.
\end{aligned}$$

Therefore, given that both $\mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t = 0] > 0$ and $\mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t = 1] > 0$ and, by assumption, $(-1)^{a^{(u_t)}} \mathbb{E}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|z_t = b^{(u_t)}] \mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t = b^{(u_t)}] \geq 0$, then it must be that $(-1)^{b^{(u_t)}} \mathbb{E}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|z_t = a^{(u_t)}] \mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t = a^{(u_t)}] < 0$. □

B.III.1 Algorithm 1 Proofs and Extension 1

Lemma 4.3.2 (Type 0 compliance with Algorithm 1)**.** *Under Assumption 4.3.1, any type 0 agent who arrives in the last ℓ rounds of Algorithm 1 is compliant with any recommendation, as long as the exploration probability ρ satisfies*

$$\rho \leq 1 + \frac{4\mu^{(0)}}{\mathbb{P}_{\mathcal{P}^{(0)}}[\xi] - 4\mu^{(0)}} \tag{4.5}$$

where the event ξ is defined above in Equation (4.4).

Proof. Let the event $\xi = \xi^{(0)}$ (as given by Definition B.III.2). By Lemma B.III.3, if ρ satisfies the following condition, then any type 0 agent will comply with any recommendation of the last ℓ rounds of Algorithm 1:

$$\rho \leq 1 + \frac{4\mu^{(0)}}{\mathbb{P}_{\mathcal{P}^{(0)}}[\xi^{(0)}] - 4\mu^{(0)}} \tag{B.11}$$

□

Definition B.III.2 (Extension 1 of Algorithm 1)**.** *Here, we formalize the recommendation policy of Extension 1 in Section 4.3.1, which modifies Algorithm 1 in two ways:*

1. *We redefine event ξ as $\xi^{(u)}$ such that it is relative to any type u, defined as follows:*

$$\xi^{(u)} = \left\{ \bar{y}^1 > \bar{y}^0 + \sigma_g \left(\sqrt{\frac{2\log(2/\delta)}{\ell_0}} + \sqrt{\frac{2\log(2/\delta)}{\ell_1}} \right) + G^{(u_t)} + \frac{1}{2} \right\}, \tag{B.12}$$

   where $G^{(u_t)}$ is an upper bound on the difference between the prior mean of the treatment versus the control according to type u, i.e. $G^{(u_t)} > \mathbb{E}_{\mathcal{P}^{(u)}}[g^1 - g^0]$, and where $\mathbb{E}_{\mathcal{P}^{(u)}}[g^0]$ and $\mathbb{E}_{\mathcal{P}^{(u)}}[g^1]$ are the expected baseline rewards for initial never-takers and always-takers.

2. *If we are trying to incentivize compliance for always-takers, then those agents in the exploration set E are recommended control (rather than treatment, as described in the pseudocode for Algorithm 1).*

Lemma B.III.3 (Arbitrary Type Compliance with Extension 1 of Algorithm 1)**.** *Under Assumption 4.3.1, any type u_t agent who arrives at round t in the last ℓ rounds of Extension 1 of Algorithm 1 (given in Definition B.III.2) is compliant with any recommendation z_t, as long as the exploration probability ρ satisfies:*

$$\rho \leq 1 + \frac{4\mu^{(u_t)}}{\mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[\xi^{(u_t)}] - 4\mu^{(u_t)}} \tag{B.13}$$

where the event $\xi^{(u_t)}$ is defined in Definition B.III.2.

Proof. This proof follows a similar structure to the Sampling Stage BIC proof in Mansour et al. (2015b).

We will prove compliance for any type u in the more general Extension 1 of Algorithm 1, as given in Definition B.III.2, which admits arbitrarily many types and the option to incentivize initial always-takers, instead of initial never-takers, to comply.

Let recommendation policy π be that described in Definition B.III.2, i.e. Extension 1 of Algorithm 1 which admits arbitrarily many types and allows for the exploration recommendations to be given in order to incentivize initial always-takers, instead of initial never-takers, to comply. Throughout this proof, we will assume that the exploration set E is defined relative to the initial preference of any agent of type u_t, who we are proving compliance for.

According to the selection function in Equation (4.2), if any agent t expects the treatment effect θ to be positive, they will select the treatment $x_t = 1$. Conversely, if they expect the treatment effect θ to be negative, they will select control $x_t = 0$. Thus, for any agent of type u_t at round t, proving compliance entails the expected treatment effect θ over the prior of type u_t and policy π_t is positive given that the recommendation $z_t = 1$ and negative given that the recommendation $z_t = 0$, i.e.

$$\mathop{\mathbb{E}}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta | z_t = 1] \geq 0 \qquad \text{and} \qquad \mathop{\mathbb{E}}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta | z_t = 0] < 0.$$

Next, we show that we can reduce our proof to demonstrating only one of the above statements, depending on the prior preference of type u. Let $a^{(u)}$ and $b^{(u)}$ denote the prior preferred and unpreferred actions for any type u, i.e. $a^{(u)} := \mathbb{1}[\mathbb{E}_{\mathcal{P}^{(u)}}[\theta] \geq 0]$ and $b^{(u)} := \mathbb{1}[\mathbb{E}_{\mathcal{P}^{(u)}}[\theta] < 0]$. Because policy π (Algorithm 1 extension) is designed in a such way that at any round t in the last ℓ rounds, treatment or control is recommended each with positive probability —i.e. $\mathbb{P}_{\pi_t, \mathcal{P}^{(u_t)}}[z_t = 1] > 0$ and $\mathbb{P}_{\pi_t, \mathcal{P}^{(u_t)}}[z_t = 0] > 0$,— Claim B.III.1 applies and the following holds:

$$\left\{ (-1)^{a^{(u_t)}} \mathop{\mathbb{E}}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta | z_t = b^{(u_t)}] \mathop{\mathbb{P}}_{\pi_t, \mathcal{P}^{(u_t)}}[z_t = b^{(u_t)}] \geq 0 \right\}$$
$$\Rightarrow \left\{ (-1)^{b^{(u_t)}} \mathop{\mathbb{E}}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta | z_t = a^{(u_t)}] \mathop{\mathbb{P}}_{\pi_t, \mathcal{P}^{(u_t)}}[z_t = a^{(u_t)}] < 0 \right\}.$$

Thus, at round t, in order to prove compliance for agents of type u_t with prior preferred and unpreferred actions $a^{(u_t)}$ and $b^{(u_t)}$, respectively, it suffices to demonstrate that $(-1)^{a^{(u_t)}}\,\mathbb{E}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|z_t=b^{(u_t)}]\,\mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t=b^{(u_t)}]\geq 0$. The remainder of the proof is devoted to demonstrating this.

We first rewrite $\underset{\pi_t,\mathcal{P}^{(u_t)}}{\mathbb{E}}[\theta|z_t=b^{(u_t)}]\underset{\pi_t,\mathcal{P}^{(u_t)}}{\mathbb{P}}[z_t=b^{(u_t)}]$ in terms of the event $\xi^{(u_t)}$:

$$
\begin{aligned}
&\underset{\pi_t,\mathcal{P}^{(u_t)}}{\mathbb{E}}[\theta|z_t=b^{(u_t)}]\underset{\pi_t,\mathcal{P}^{(u_t)}}{\mathbb{P}}[z_t=b^{(u_t)}]\\
&=\underset{\pi_t,\mathcal{P}^{(u_t)}}{\mathbb{E}}[\theta|z_t=b^{(u_t)}\ \&\ t\notin E]\underset{\pi_t,\mathcal{P}^{(u_t)}}{\mathbb{P}}[z_t=b^{(u_t)}\ \&\ t\notin E]\\
&\quad+\underset{\pi_t,\mathcal{P}^{(u_t)}}{\mathbb{E}}[\theta|z_t=b^{(u_t)}\ \&\ t\in E]\underset{\pi_t,\mathcal{P}^{(u_t)}}{\mathbb{P}}[z_t=b^{(u_t)}\ \&\ t\in E]\\
&\hspace{12em}\text{(for explore set E defined relative to recommend action $b^{(u_t)}$)}\\
&=\underset{\pi_t,\mathcal{P}^{(u_t)}}{\mathbb{E}}[\theta|\xi\ \&\ t\notin E]\underset{\pi_t,\mathcal{P}^{(u_t)}}{\mathbb{P}}[\xi\ \&\ t\notin E]\\
&+\underset{\pi_t,\mathcal{P}^{(u_t)}}{\mathbb{E}}[\theta|z_t=b^{(u_t)}\ \&\ t\in E]\underset{\pi_t,\mathcal{P}^{(u_t)}}{\mathbb{P}}[z_t=b^{(u_t)}\ \&\ t\in E]\\
&\hspace{3em}\text{(since the only way $z_t=b^{(u_t)}$ when exploiting (i.e. when $t\notin E$) is when event ξ occurs)}\\
&=\underset{\pi_t,\mathcal{P}^{(u_t)}}{\mathbb{E}}[\theta|\xi\ \&\ t\notin E]\underset{\pi_t,\mathcal{P}^{(u_t)}}{\mathbb{P}}[\xi\ \&\ t\notin E]+\underset{\pi_t,\mathcal{P}^{(u_t)}}{\mathbb{E}}[\theta|t\in E]\underset{\pi_t,\mathcal{P}^{(u_t)}}{\mathbb{P}}[t\in E],\\
&\hspace{22em}\text{($t\in E\Rightarrow z_t=1$ by definition of E)}\\
&=\underset{\pi_t,\mathcal{P}^{(u_t)}}{\mathbb{E}}[\theta|\xi]\underset{\pi_t,\mathcal{P}^{(u_t)}}{\mathbb{P}}[\xi]\underset{\pi_t,\mathcal{P}^{(u_t)}}{\mathbb{P}}[t\notin E]+\underset{\pi_t,\mathcal{P}^{(u_t)}}{\mathbb{E}}[\theta]\underset{\pi_t,\mathcal{P}^{(u_t)}}{\mathbb{P}}[t\in E]\\
&\hspace{24em}\text{(since $\theta\perp t\in E$ and $\xi\perp t\notin E$)}\\
&=(1-\rho)\underset{\pi_t,\mathcal{P}^{(u_t)}}{\mathbb{E}}[\theta|\xi]\underset{\pi_t,\mathcal{P}^{(u_t)}}{\mathbb{P}}[\xi]+\rho\underset{\pi_t,\mathcal{P}^{(u_t)}}{\mathbb{E}}[\theta]\qquad\text{(since agent $t\in E$ with probability ρ)}\\
&=(1-\rho)\underset{\pi_t,\mathcal{P}^{(u_t)}}{\mathbb{E}}[\theta|\xi]\underset{\pi_t,\mathcal{P}^{(u_t)}}{\mathbb{P}}[\xi]+\rho\mu^{(u)}\qquad\text{(by definition, $\underset{\pi_t,\mathcal{P}^{(u_t)}}{\mathbb{E}}[\theta]=\mu^{(u)}$)}
\end{aligned}
\tag{B.14}
$$

Now, we can rewrite our compliance condition as such:

$$
\begin{aligned}
&(-1)^{a^{(u_t)}}\underset{\pi_t,\mathcal{P}^{(u_t)}}{\mathbb{E}}[\theta|z_t=b^{(u_t)}]\underset{\pi_t,\mathcal{P}^{(u_t)}}{\mathbb{P}}[z_t=b^{(u_t)}]\geq 0\\
&\equiv(-1)^{a^{(u_t)}}\left((1-\rho)\underset{\pi_t,\mathcal{P}^{(u_t)}}{\mathbb{E}}[\theta|\xi]\underset{\pi_t,\mathcal{P}^{(u_t)}}{\mathbb{P}}[\xi]+\rho\mu^{(u)}\right)\geq 0.
\end{aligned}
$$

Now, we rewrite this compliance condition strictly in terms of the exploration probability ρ and relative to a number of constants which depend on the prior $\mathcal{P}^{(u_t)}$. Thus, if we set ρ to satisfy the following condition (in Equation (B.15)), then all agents of type u will comply with recommendations from policy π (Algorithm 1 extension):

$$(-1)^{a^{(u_t)}} \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|z_t = b^{(u_t)}] \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t = b^{(u_t)}] \geq 0$$

$$(-1)^{a^{(u_t)}}\left((1-\rho) \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|\xi^{(u_t)}] \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[\xi^{(u_t)}] + \rho\mu^{(u_t)}\right) \geq 0 \qquad \text{(by Equation (B.14))}$$

$$(-1)^{a^{(u_t)}}\left(\mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|\xi^{(u_t)}] \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[\xi^{(u_t)}] - \rho \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|\xi^{(u_t)}] \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[\xi^{(u_t)}] + \rho\mu^{(u_t)}\right) \geq 0$$

$$(-1)^{a^{(u_t)}+1}\rho\left(\mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|\xi^{(u_t)}] \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[\xi^{(u_t)}] - \mu^{(u_t)}\right) \geq (-1)^{a^{(u_t)}+1} \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|\xi^{(u_t)}]\, \mathbb{P}[\xi^{(u_t)}]$$

$$\rho \leq \frac{\mathbb{E}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|\xi^{(u_t)}]\, \mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[\xi^{(u_t)}]}{\mathbb{E}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|\xi^{(u_t)}]\, \mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[\xi^{(u_t)}] - \mu^{(0)}}$$

$$\text{(since } (-1)^{a^{(u_t)}+1}\rho\left(\mathbb{E}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|\xi^{(u_t)}]\, \mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[\xi^{(u_t)}] - \mu^{(u_t)}\right) < 0 \text{ for any } u_t{}^{1})$$

$$\rho \leq 1 + \frac{\mu^{(u_t)}}{\mathbb{E}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|\xi^{(u_t)}]\, \mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[\xi^{(u_t)}] - \mu^{(u_t)}} \tag{B.15}$$

Finally, we can further simplify the upper bound on ρ given in Equation (B.15) above by showing that $\mathbb{E}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|\xi^{(u_t)}]$ satisfies some constant lower bound. This will complete our proof.

For any type u, the baseline reward $g^{(u)}$ is a random variable independently distributed according to a sub-Gaussian distribution with variance $\sigma^{(u)}$ which is bounded above by σ_g, i.e. $\sigma^{(u)} < \sigma_g$ for any u. Furthermore, recall that $G^{(u_t)} > \mathbb{E}_{\mathcal{P}^{u_t}}[g^1 - g^0]$, where $\mathbb{E}_{\mathcal{P}^{u_t}}[g^1]$ and $\mathbb{E}_{\mathcal{P}^{u_t}}[g^0]$ are the expected value of the baseline rewards of always-takers and never-takers over the prior of type u_t, respectively.

Now, we define 3 clean events: $\mathcal{C}_0$ and $\mathcal{C}_1$ pertain to these baseline reward random variables, and $\mathcal{C}_2$ occurs when the first stage of Algorithm 1 generates at least ℓ_0 control

[1]This point is not entirely obvious: If $a^{(u_t)} = 0$, then $(-1)^{a^{(u_t)}+1} < 0$ and $\mathbb{E}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|\xi^{(u_t)}] - \mu^{(u_t)} > 0$, since $\mathbb{E}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|\xi^{(u_t)}] > 0$ and $\mu^{(u_t)} < 0$. If $a^{(u_t)} = 1$, then $(-1)^{a^{(u_t)}+1} > 0$ and $\mathbb{E}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|\xi^{(u_t)}] - \mu^{(u_t)} < 0$, since $\mathbb{E}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|\xi^{(u_t)}] < 0$ and $\mu^{(u_t)} > 0$.

samples and at least ℓ_1 treatment samples:

$$\mathcal{C}_0 := \left\{ \bar{y}^0 = \frac{1}{\sum_{t=1}^{\ell} \mathbb{1}[\mu^{(u_t)} < 0]} \sum_{t=1}^{\ell} g^{(u_t)} \mathbb{1}[\mu^{(u_t)} < 0] \leq \sigma_g \sqrt{\frac{2\log(1/\delta_0)}{\ell_0}} - \mathop{\mathbb{E}}_{\pi_t, \mathcal{P}^{(u_t)}}[g^0] \right\} \tag{B.16}$$

$$\mathcal{C}_1 := \left\{ \bar{y}^1 = \frac{1}{\sum_{t=1}^{\ell} \mathbb{1}[\mu^{(u_t)} > 0]} \sum_{t=1}^{\ell} g^{(u_t)} \mathbb{1}[\mu^{(u_t)} > 0] \geq -\sigma_g \sqrt{\frac{2\log(1/\delta_1)}{\ell_1}} - \mathop{\mathbb{E}}_{\pi_t, \mathcal{P}^{(u_t)}}[g^1] \right\} \tag{B.17}$$

$$\mathcal{C}_2 := \left\{ \ell_1 \leq \sum_{i=1}^{\ell'} x_i \leq \ell' - \ell_0 \right\} \tag{B.18}$$

where $\ell' = 2\max(\ell_0/p_0, \ell_1/p_1)$ is the number of rounds in the first stage of Algorithm 1. Let $\delta_0 = \delta_1 = \mathbb{P}_{\pi_t, \mathcal{P}^{(u_t)}}[\xi^{(u_t)}]/24$. Furthermore, event $\mathcal{C}_2$ occurs when the binomial random variable with success $u_t = x_t = 1$ (since $x_t = u_t$ in the first stage of Algorithm 1) and success probability p_1 is lower bounded by ℓ_1 and upper bounded by $\ell' - \ell_0$. For $\ell' = 2\max(\ell_0/p_0, \ell_1/p_1)$ total trials, the probability of this event is less than $\mathbb{P}_{\pi_t, \mathcal{P}^{(u_t)}}[\xi^{(u_t)}]/24$.

Now, define another clean event $\mathcal{C}$ where all $\mathcal{C}_0$, $\mathcal{C}_1$, and $\mathcal{C}_2$ happen simultaneously. Letting $\delta = \delta_0 + \delta_1 + \delta_2$, the event $\mathcal{C}$ occurs with probability at least $1 - \delta$ where $\delta < \mathop{\mathbb{P}}_{\pi_t, \mathcal{P}^{(u_t)}}[\xi^{(u_t)}]/8$. We can now rewrite $\mathbb{E}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta | \xi^{(u_t)}]\, \mathbb{P}_{\pi_t, \mathcal{P}^{(u_t)}}[\xi^{(u_t)}]$ in terms of event

$\mathcal{C}$:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|\xi^{(u_t)}] \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[\xi^{(u_t)}] && \text{(B.19)}\\
&= \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|\xi^{(u_t)},\mathcal{C}] \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[\xi^{(u_t)},\mathcal{C}] + \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|\xi^{(u_t)},\neg\mathcal{C}] \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[\xi^{(u_t)},\neg\mathcal{C}]\\
&\geq \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|\xi^{(u_t)},\mathcal{C}] \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[\xi^{(u_t)},\mathcal{C}] - \delta\\
&\qquad\qquad\qquad\qquad \text{(since } \mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[\neg\mathcal{C}] < \delta \text{ and } \theta \geq -1 \text{ by definition)}\\
&\geq \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|\xi^{(u_t)},\mathcal{C}] \left(\mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[\xi^{(u_t)}] - \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[\neg\mathcal{C}] \right) - \delta\\
&\geq \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|\xi^{(u_t)},\mathcal{C}] \left(\mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[\xi^{(u_t)}] - \delta \right) - \delta\\
&= \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|\xi^{(u_t)},\mathcal{C}] \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[\xi^{(u_t)}] - \delta \left(1 + \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|\xi^{(u_t)},\mathcal{C}] \right)\\
&\geq \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|\xi^{(u_t)},\mathcal{C}] \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[\xi^{(u_t)}] - 2\delta \qquad \text{(since } \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|\xi^{(u_t)},\mathcal{C}] \leq 1) && \text{(B.20)}
\end{aligned}$$

This comes down to finding a lower bound on the denominator of the expression above. We can reduce the dependency of the denominator to a single prior-dependent constant $\mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[\xi^{(u_t)}]$ if we lower bound the prior-dependent expected value $\mathbb{E}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|\xi^{(u_t)}]$. That way, assuming we know the prior and can calculate the probability of event $\xi^{(u_t)}$, we can pick an appropriate exploration probability ρ to satisfy the compliance condition for all agents of type 0. Then:

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|\xi^{(u_t)},\mathcal{C}] \\
&= \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}\left[\theta\middle|\bar{y}^1 > \bar{y}^0 + \sigma_g\left(\sqrt{\frac{2\log(1/\delta)}{\ell_0}} + \sqrt{\frac{2\log(1/\delta)}{\ell_1}}\right) + G^{(u_t)} + \frac{1}{2},\mathcal{C}\right] \\
&\geq \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}\left[\theta\middle|\theta > \frac{1}{\ell_0}\sum_{t=1}^{\ell_0} g^0 - \frac{1}{\ell_1}\sum_{t=1}^{\ell_1}\theta + \sigma_g\left(\sqrt{\frac{2\log(1/\delta)}{\ell_0}} + \sqrt{\frac{2\log(1/\delta)}{\ell_1}}\right) + G^{(u_t)} + \frac{1}{2},\mathcal{C}\right] \\
&= \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}\left[\theta\middle|\theta > -\left(\mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[g^0] - \sigma_g\sqrt{\frac{2\log(1/\delta_1)}{\ell_0}}\right) + \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[g^1] - \sigma_g\sqrt{\frac{2\log(1/\delta_2)}{\ell_1}}\right. \\
&\qquad\qquad \left. + \sigma_g\left(\sqrt{\frac{2\log(1/\delta)}{\ell_0}} + \sqrt{\frac{2\log(1/\delta)}{\ell_1}}\right) + G^{(u_t)} + \frac{1}{2},\mathcal{C}\right] \\
&\hspace{25em}\text{(by event } \mathcal{C}\text{)} \\
&= \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}\left[\theta\middle|\theta > -\sigma_g\left(\sqrt{\frac{2\log(2/\delta)}{\ell_0}} - \sqrt{\frac{2\log(2/\delta)}{\ell_1}}\right) + \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[g^1 - g^0]\right. \qquad \text{(B.21)} \\
&\qquad\qquad \left. +\sigma_g\left(\sqrt{\frac{2\log(2/\delta)}{\ell_0}} + \sqrt{\frac{2\log(2/\delta)}{\ell_1}}\right) + G^{(u_t)} + \frac{1}{2},\mathcal{C}\right] \\
&\hspace{15em}\text{(by Theorem B.I.7, where } \delta_1 = \delta_2 = \delta/2\text{)} \\
&> \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}\left[\theta\middle|\theta > \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[g^1 - g^0] - \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[g^1 - g^0] + \frac{1}{2}\right] \qquad \text{(by definition of } G^{(u_t)}\text{)} \\
&> \frac{1}{2} \qquad \text{(B.22)}
\end{aligned}
$$

Hence, the term $\mathbb{E}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|\xi^{(u_t)}]\,\mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[\xi^{(u_t)}]$ satisfies the following lower bound:

$$
\begin{aligned}
\mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|\xi^{(u_t)}]\mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[\xi^{(u_t)}] &\geq \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|\xi^{(u_t)},\mathcal{C}]\mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[\xi^{(u_t)}] - 2\delta \quad \text{(by Equation (B.20))} \\
&> \frac{1}{2}\mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[\xi^{(u_t)}] - 2\delta \qquad \text{(by Equation (B.22))} \\
&= \frac{\mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[\xi^{(u_t)}]}{4} + \frac{\mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[\xi^{(u_t)}]}{4} - 2\delta \\
&> \frac{\mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[\xi^{(u_t)}]}{4} \qquad \text{(since } \delta < \mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[\xi^{(u_t)}]/8\text{)}
\end{aligned}
$$

Substituting this into Equation (B.15), we arrive at a lower bound to set the exploration probability ρ for the agent any round t with type u_t to comply with recommendation policy π_t (extension of Algorithm 1):

$$\rho \leq 1 + \frac{4\mu^{(u_t)}}{\mathbb{P}_{\pi_t, \mathcal{P}^{(u_t)}}[\xi^{(u_t)}] - 4\mu^{(u_t)}}$$

□

Theorem 4.3.3 (Treatment Effect Confidence Interval after Algorithm 1)**.** *With sample set $S_\ell = (x_i, y_i, z_i)_{i=1}^{\ell}$ of ℓ samples collected from the second stage of Algorithm 1 —run with exploration probability ρ small enough so that type 0 agents are compliant (see Lemma 4.3.2),— approximation bound $A(S_\ell, \delta)$ satisfies the following, with probability at least $1-\delta$:*

$$\left|\hat{\theta}_{S_\ell} - \theta\right| \leq A(S_\ell, \delta) \leq \frac{2\sigma_g\sqrt{2\log(5/\delta)}}{\rho(1-\rho)p_0\sqrt{\ell} - (3-\rho)\sqrt{\frac{\rho\log(5/\delta)}{2(1-\rho)}}}$$

for any $\delta \in (0,1)$. Recall σ_g is the variance of $g^{(u_i)}$, p_0 is the fraction of compliant never-takers in the population of agents,[2] *and $A(S_\ell, \delta)$ is defined as in Theorem 4.2.1.*

Proof. First, Theorem 4.2.1 demonstrates, for any $\delta_1 \in (0,1)$, with probability at least $1-\delta_1$ that the approximation bound

$$|\theta - \hat{\theta}_{S_\ell}| \leq A(S_\ell, \delta) = \frac{2\sigma_g\sqrt{2\ell\log(2/\delta_1)}}{\left|\sum_{i=1}^{\ell}(x_i - \bar{x})(z_i - \bar{z})\right|}. \tag{B.23}$$

Next, recall that the mean recommendation $\bar{z} = \rho$ for exploration probability ρ in the second stage of Algorithm 1. We assume Algorithm 1 to be initialized with parameters (see Lemma 4.3.2 for details) such that its recommendations are compliant for agents of type 0. In the worst case, only type 0 agents are compliant. Therefore, Theorem B.II.2 implies that, for any $\delta_2 \in (0,1)$, with probability at least $1-\delta_2$ that

$$\left|\sum_{i=1}^{\ell}(x_i - \bar{x})(z_i - \bar{z})\right| \geq \rho\ell\left(p_0(1-\rho) - \sqrt{\frac{(1-\rho)\log(1/\delta_2)}{2\ell}}\right). \tag{B.24}$$

With a union bound over Equations (B.23) and (B.24) while letting $\delta_1 = \delta_2 = \frac{\delta}{3}$ for any

$\delta \in (0,1)$, we conclude: with probability at least $1-\delta$,

$$A(S_\ell,\delta) \le \frac{2\sigma_g\sqrt{2\ell\log(3/\delta)}}{\rho\ell\left(p_0(1-\rho)-\sqrt{\frac{(1-\rho)\log(3/\delta)}{2\ell}}\right)} = \frac{2\sigma_g\sqrt{2\log(3/\delta)}}{\rho\left(p_0(1-\rho)\sqrt{\ell}-\sqrt{\frac{(1-\rho)\log(3/\delta)}{2}}\right)}$$

□

B.IV Missing Proofs for Section 4.4

B.IV.1 Algorithm 2 Proofs

Lemma 4.4.2 (Algorithm 2 Partial Compliance). *Recall that Algorithm 2 is initialized with input samples $S_0 = (x_i, y_i, z_i)_{i=1}^{|S_0|}$. For any type u with the following prior preference (control or treatment), if S_0 satisfies the following condition, with probability at least $1-\delta$, then all agents of type u will comply with recommendations of Algorithm 2:*

$$A(S_0,\delta) \le \begin{cases} \tau\,\mathbb{P}_{\mathcal{P}^{(u)}}[\theta > \tau]/4 & \text{if } \mathbb{E}_{\mathcal{P}^{(u)}}[\theta] < 0; \\ \tau\,\mathbb{P}_{\mathcal{P}^{(u)}}[\theta < -\tau]/4 & \text{if } \mathbb{E}_{\mathcal{P}^{(u)}}[\theta] \ge 0, \end{cases}$$

for some $\tau \in (0,1)$, where $A(S_0,\delta)$ is the approximation bound for S_0 and any $\delta \in (0,1)$ (see Theorem 4.2.1).

Proof. Just as in the proof for Lemma B.III.3, let $a^{(u)}$ and $b^{(u)}$ denote the prior preferred and unpreferred actions for agents of any type u, i.e. $a^{(u)} := \mathbb{1}[\mathbb{E}_{\mathcal{P}^{(u)}}[\theta] \ge 0]$ and $b^{(u)} := \mathbb{1}[\mathbb{E}_{\mathcal{P}^{(u)}}[\theta] < 0]$. Let π denote the recommendation policy defined by Algorithm 2. At any round t of Algorithm 2, recommendation policy π_t has a positive probability of recommending either control or treatment, according to the prior $\mathcal{P}^{(u_t)}$ for type u_t, i.e. $\mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t = 0] > 0$ and $\mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t = 1] > 0$. Thus, by Claim B.III.1, the following holds:

$$\left\{(-1)^{a^{(u_t)}} \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|z_t = b^{(u_t)}] \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t = b^{(u_t)}] \ge 0\right\}$$
$$\Rightarrow \left\{(-1)^{b^{(u_t)}} \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|z_t = a^{(u_t)}] \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t = a^{(u_t)}] < 0\right\}$$

and it suffices to prove the premise $(-1)^{a^{(u_t)}}\,\mathbb{E}_{\mathcal{P}^{(u_t)},\pi_t}[\theta|z_t = b^{(u_t)}]\,\mathbb{P}_{\mathcal{P}^{(u_t)},\pi_t}[z_t = b^{(u_t)}] \ge 0$

in order to prove that agent t of type u_t complies with recommendation z_t.

Recall that the sample set S_q^{BEST} is made up of the best samples up until phase q of Algorithm 2, i.e. the samples which produce the smallest approximation bound A_q. The treatment effect estimate derived from set S_q^{BEST} is denoted $\hat{\theta}_q$. We define the event $\mathcal{C}$ as the event that the treatment effect estimate $\hat{\theta}_q$ satisfies the approximation bound A_q at every phase q throughout Algorithm 2:

$$\mathcal{C} := \left\{ \forall q \geq 0 : |\theta - \hat{\theta}_q| < A_q \right\}. \tag{B.25}$$

By Theorem 4.2.1, for event $\mathcal{C}$, the failure probability $\mathbb{P}[\neg\mathcal{C}] \leq \delta$. Furthermore, we assume here that

$$\delta \leq \begin{cases} \frac{\tau \, \mathbb{P}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta \geq \tau]}{2(\tau \, \mathbb{P}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta \geq \tau] + 1)} & \text{if } \mu^{(u_t)} < 0; \\ \frac{\tau \, \mathbb{P}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta < -\tau]}{2(\tau \, \mathbb{P}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta < -\tau] + 1)} & \text{if } \mu^{(u_t)} \geq 0. \end{cases}$$

Therefore, since $|\theta| \leq 1$, we have:

$$\begin{aligned}
&(-1)^{a^{(u_t)}} \mathop{\mathbb{E}}_{\mathcal{P}^{(u_t)}, \pi_t}[\theta | z_t = b^{(u_t)}] \mathop{\mathbb{P}}_{\mathcal{P}^{(u_t)}, \pi_t}[z_t = b^{(u_t)}] \\
&= (-1)^{a^{(u_t)}} \left(\mathop{\mathbb{E}}_{\mathcal{P}^{(u_t)}, \pi_t}[\theta | z_t = b^{(u_t)}, \mathcal{C}] \mathop{\mathbb{P}}_{\mathcal{P}^{(u_t)}, \pi_t}[z_t = b^{(u_t)}, \mathcal{C}] \right) \\
&\quad + (-1)^{a^{(u_t)}} \left(\mathop{\mathbb{E}}_{\mathcal{P}^{(u_t)}, \pi_t}[\theta | z_t = b^{(u_t)}, \neg\mathcal{C}] \mathop{\mathbb{P}}_{\mathcal{P}^{(u_t)}, \pi_t}[z_t = b^{(u_t)}, \neg\mathcal{C}] \right) \\
&\geq (-1)^{a^{(u_t)}} \left(\mathop{\mathbb{E}}_{\mathcal{P}^{(u_t)}, \pi_t}[\theta | z_t = b^{(u_t)}, \mathcal{C}] \mathop{\mathbb{P}}_{\mathcal{P}^{(u_t)}, \pi_t}[z_t = b^{(u_t)}, \mathcal{C}] - (-1)^{a^{(u_t)}} \delta \right) \\
&\geq (-1)^{a^{(u_t)}} \left(\mathop{\mathbb{E}}_{\mathcal{P}^{(u_t)}, \pi_t}[\theta | z_t = b^{(u_t)}, \mathcal{C}] \mathop{\mathbb{P}}_{\mathcal{P}^{(u_t)}, \pi_t}[z_t = b^{(u_t)}, \mathcal{C}] \right) - \frac{\tau \, \mathbb{P}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta \geq \tau]}{2\tau \, \mathbb{P}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta \geq \tau] + 2}
\end{aligned}$$

In order to lower bound the last line above, we marginalize $\mathbb{E}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta | z_t = b^{(u_t)}, \mathcal{C}] \, \mathbb{P}_{\mathcal{P}^{(u_t)}, \pi_t}[z_t = b^{(u_t)}, \mathcal{C}]$ based on four possible ranges which θ lies on:

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{\mathcal{P}^{(u_t)},\pi_t}[\theta|z_t = b^{(u_t)}, \mathcal{C}] \mathop{\mathbb{P}}_{\mathcal{P}^{(u_t)},\pi_t}[z_t = b^{(u_t)}, \mathcal{C}] \\
&= \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|z_t = b^{(u_t)}, \mathcal{C}, (-1)^{a^{(u_t)}}\theta \geq \tau] \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t = b^{(u_t)}, \mathcal{C}, (-1)^{a^{(u_t)}}\theta \geq \tau] \\
&+ \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|z_t = b^{(u_t)}, \mathcal{C}, 0 \leq (-1)^{a^{(u_t)}}\theta < \tau] \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t = b^{(u_t)}, \mathcal{C}, 0 \leq (-1)^{a^{(u_t)}}\theta < \tau] \\
&+ \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|z_t = b^{(u_t)}, \mathcal{C}, -2A_q < (-1)^{a^{(u_t)}}\theta < 0] \, \mathbb{P}[z_t = b^{(u_t)}, \mathcal{C}, -2A_q < (-1)^{a^{(u_t)}}\theta < 0] \\
&+ \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|z_t = b^{(u_t)}, \mathcal{C}, (-1)^{a^{(u_t)}}\theta \leq -2A_q] \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t = b^{(u_t)}, \mathcal{C}, (-1)^{a^{(u_t)}}\theta \leq -2A_q]
\end{aligned} \tag{B.26}
$$

Because A_q is the smallest approximation bound derived from samples collected over any phase q of Algorithm 2 (including the initial sample set S_0), the following holds:

$$
\begin{aligned}
2A_q &\leq 2A(S_0, \delta) \\
&\leq \frac{\tau \, \mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[(-1)^{a^{(u_t)}}\theta \geq \tau]}{2} \\
&\quad \text{(by assumption } A(S_0, \delta) \leq \tau \, \mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[(-1)^{a^{(u_t)}}\theta \geq \tau]/4) \\
&\leq \tau
\end{aligned}
$$

Conditional on $\mathcal{C}$, $|\theta - \hat{\theta}_q| < A_q$. Thus, we may reduce two of the terms in Equation (B.26) above, based on the structure of Algorithm 2. First, from the first term in Equation (B.26), note that

$$
(-1)^{a^{(u_t)}}\theta \geq \tau \geq 2A_q \quad \Rightarrow \quad (-1)^{a^{(u_t)}}\hat{\theta}_q \geq \tau - A_q \geq A_q,
$$

which invokes the stopping criterion for the while loop in Algorithm 2. Thus, type u_t's preferred action $a^{(u_t)}$ must have been eliminated from the race before phase $q = 1$ and the unpreferred action $b^{(u_t)}$ is recommended almost surely throughout Algorithm 2, i.e.

$$
\mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t = b^{(u_t)}, \mathcal{C}, (-1)^{a^{(u_t)}}\theta \geq \tau] = \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[\mathcal{C}, (-1)^{a^{(u_t)}}\theta \geq \tau].
$$

Second, from the last term in Equation (B.26), note that

$$(-1)^{a^{(u_t)}}\theta \leq -2A_q \quad \Rightarrow \quad (-1)^{a^{(u_t)}}\hat{\theta}_q \leq -A_q$$

by phase $q = 1$ and the unpreferred action $b^{(u_t)}$ is recommended almost never, i.e.

$$\mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t = b^{(u_t)}, \mathcal{C}, (-1)^{a^{(u_t)}}\theta < -2A_q] = 0.$$

Substituting these probabilities back into Equation (B.26), we proceed:

$$
\begin{aligned}
&(-1)^{a^{(u_t)}} \mathop{\mathbb{E}}_{\mathcal{P}^{(u_t)},\pi_t}[\theta | z_t = b^{(u_t)}, \mathcal{C}] \mathop{\mathbb{P}}_{\mathcal{P}^{(u_t)},\pi_t}[z_t = b^{(u_t)}, \mathcal{C}] \\
=&(-1)^{a^{(u_t)}} \Bigg(\mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta | \mathcal{C}, (-1)^{a^{(u_t)}}\theta \geq \tau] \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[\mathcal{C}, (-1)^{a^{(u_t)}}\theta \geq \tau] \\
&\quad + \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta | z_t = b^{(u_t)}, \mathcal{C}, 0 \leq (-1)^{a^{(u_t)}}\theta < \tau] \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t = b^{(u_t)}, \mathcal{C}, 0 \leq (-1)^{a^{(u_t)}}\theta < \tau] \\
&\quad + \mathbb{E}[\theta | z_t = b^{(u_t)}, \mathcal{C}, -2A_q < (-1)^{a^{(u_t)}}\theta < 0] \, \mathbb{P}[z_t = b^{(u_t)}, \mathcal{C}, -2A_q < (-1)^{a^{(u_t)}}\theta < 0] \Big) \\
\geq& (-1)^{a^{(u_t)}} \Big((-1)^{a^{(u_t)}} \tau \, \mathbb{P}[\mathcal{C}, (-1)^{a^{(u_t)}}\theta \geq \tau] + 0 \cdot \mathbb{P}[z_t = b^{(u_t)}, \mathcal{C}, 0 \leq (-1)^{a^{(u_t)}}\theta < \tau] \\
&\qquad\qquad - (-1)^{a^{(u_t)}} 2A_q \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t = b^{(u_t)}, \mathcal{C}, -2A_q < (-1)^{a^{(u_t)}}\theta < 0] \Bigg) \\
\geq& (-1)^{a^{(u_t)}} \left((-1)^{a^{(u_t)}} \tau \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[\mathcal{C}, (-1)^{a^{(u_t)}}\theta \geq \tau] - (-1)^{a^{(u_t)}} 2A_q \right) \\
\geq& \tau \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[\mathcal{C}, (-1)^{a^{(u_t)}}\theta \geq \tau] - \frac{\tau \, \mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[(-1)^{a^{(u_t)}}\theta \geq \tau]}{2} \\
\geq& \tau \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[\mathcal{C} | (-1)^{a^{(u_t)}}\theta \geq \tau] \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[(-1)^{a^{(u_t)}}\theta \geq \tau] - \frac{\tau \, \mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[(-1)^{a^{(u_t)}}\theta \geq \tau]}{2} \\
\geq& (1-\delta)\tau \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[(-1)^{a^{(u_t)}}\theta \geq \tau] - \frac{\tau \, \mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[(-1)^{a^{(u_t)}}\theta \geq \tau]}{2} \\
=& \left(\frac{1}{2} - \delta \right) \tau \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[(-1)^{a^{(u_t)}}\theta \geq \tau] \\
\geq& \left(\frac{1}{2} - \frac{\tau \, \mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[(-1)^{a^{(u_t)}}\theta \geq \tau]}{2\tau \, \mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[(-1)^{a^{(u_t)}}\theta \geq \tau] + 2} \right) \tau \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[(-1)^{a^{(u_t)}}\theta \geq \tau] \\
=& \frac{\tau \, \mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[(-1)^{a^{(u_t)}}\theta \geq \tau]}{2\tau \, \mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[(-1)^{a^{(u_t)}}\theta \geq \tau] + 2}
\end{aligned}
$$

Putting everything together, we get that

$$\begin{aligned}
&(-1)^{a^{(u_t)}} \mathop{\mathbb{E}}_{\mathcal{P}^{(u_t)},\pi_t}[\theta | z_t = b^{(u_t)}] \mathop{\mathbb{P}}_{\mathcal{P}^{(u_t)},\pi_t}[z_t = b^{(u_t)}] \\
&\geq (-1)^{a^{(u_t)}} \Bigg(\mathop{\mathbb{E}}_{\mathcal{P}^{(u_t)},\pi_t}[\theta | z_t = b^{(u_t)}, \mathcal{C}] \mathop{\mathbb{P}}_{\mathcal{P}^{(u_t)},\pi_t}[z_t = b^{(u_t)}, \mathcal{C}] \\
&\quad + \mathop{\mathbb{E}}_{\mathcal{P}^{(u_t)},\pi_t}[\theta | z_t = b^{(u_t)}, \neg\mathcal{C}] \mathop{\mathbb{P}}_{\mathcal{P}^{(u_t)},\pi_t}[z_t = b^{(u_t)}, \neg\mathcal{C}] \Bigg) \\
&\geq \frac{\tau \, \mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[(-1)^{a^{(u_t)}}\theta \geq \tau]}{2\tau \, \mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[(-1)^{a^{(u_t)}}\theta \geq \tau] + 2} - \frac{\tau \, \mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[(-1)^{a^{(u_t)}}\theta \geq \tau]}{2\tau \, \mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[(-1)^{a^{(u_t)}}\theta \geq \tau] + 2} \\
&= 0
\end{aligned}$$

Therefore, so long as

$$A(S_0, \delta) \leq \tau \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[(-1)^{a^{(u_t)}}\theta \geq \tau]/4 \quad \text{and} \quad \delta < \tau \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[(-1)^{a^{(u_t)}}\theta \geq \tau]/2,$$

any agent of type u_t will comply with recommendations from Algorithm 2. □

B.IV.2 Lemma B.IV.1 and Theorem B.IV.2: Full Compliance and Subsequent Estimation Bound

Lemma B.IV.1 (Algorithm 2 Full Compliance)**.** *Suppose that some fraction $p_c > 0$ of agents is compliant from the beginning of Algorithm 2 and assume that $p_c < 1$. Of all types u which were not compliant from the beginning, let type u^* agents be the most resistant to compliance. Suppose that phase q satisfies one of the following bounds (depending on whether type u^* agents prefer control or treatment):*

$$q \geq \begin{cases} \left(\frac{1}{2hp_c} \left(\frac{(32\sigma_g \sqrt{2\log(5/\delta)}}{\tau \, \mathbb{P}_{\mathcal{P}^{(u^*)}}[\theta > \tau]} + \sqrt{50\log(5/\delta)} \right) \right)^2 & \textit{if } \mathbb{E}_{\mathcal{P}^{(u^*)}}[\theta] < 0 \\ \left(\frac{1}{2hp_c} \left(\frac{(32\sigma_g \sqrt{2\log(5/\delta)}}{\tau \, \mathbb{P}_{\mathcal{P}^{(u^*)}}[\theta < -\tau]} + \sqrt{50\log(5/\delta)} \right) \right)^2 & \textit{if } \mathbb{E}_{\mathcal{P}^{(u^*)}}[\theta] \geq 0, \end{cases}$$

for some $\tau \in (0,1)$ and any $\delta \in (0,1)$. Then, with probability at least $1-\delta$, for any phase q greater or equal to the following lower bound all agents will comply with recommendations from Algorithm 2.

Proof. First, recall that the set S_q is made up of the input samples S_0 plus samples collected following Algorithm 2 over all phases up to q. Let $S_{q-0} = (x_i, y_i, z_i)_{i=1}^{2hq}$ denote S_q sans S_0 (i.e. just samples collected following Algorithm 2 up to phase q). Note that for any $\delta \in (0,1)$, the approximation bound $A_q \leq A(S_{q-0}, \delta)$.

We want to prove that type u^* is compliant by and beyond phase q. By Lemma 4.4.2, it suffices to prove that the approximation bound A_q satisfies the following upper bound with probability at least $1-\delta$:[3]

$$\begin{cases} A_q \leq \tau\,\mathbb{P}_{\mathcal{P}(u^*)}[\theta > \tau]/4 & \text{if } \mathbb{E}_{\mathcal{P}(u^*)}[\theta] < 0; \\ A_q \leq \tau\,\mathbb{P}_{\mathcal{P}(u^*)}[\theta < -\tau]/4 & \text{if } \mathbb{E}_{\mathcal{P}(u^*)}[\theta] \geq 0, \end{cases}$$

for any $\delta \in (0,1)$ and some $\tau \in (0,1)$.

In order to prove this, recall that each phase q of Algorithm 2 is $2hq$ rounds long and the mean recommendation $\bar{z} = \frac{1}{2}$. By assumption, p_c proportion of agents are compliant. Thus, by Theorem B.II.2, with probability $1-\delta$ for any $\delta \in (0,1)$, the set S_{q-0} satisfies:

$$A(S_{q-0}, \delta) \leq \frac{8\sigma_g \sqrt{2\log(3/\delta)}}{p_c\sqrt{|S_{q-0}|} - \sqrt{\log(3/\delta)}}$$

By assumption, q satisfies the following lower bound for some $\tau \in (0,1)$:

$$q \geq \begin{cases} \left(\frac{1}{2hp_c}\left(\frac{(32\sigma_g\sqrt{2\log(5/\delta)}}{\tau\,\mathbb{P}_{\mathcal{P}(u^*)}[\theta>\tau]} + \sqrt{50\log(5/\delta)}\right)\right)^2 & \text{if } \mathbb{E}_{\mathcal{P}(u^*)}[\theta] < 0 \\ \left(\frac{1}{2hp_c}\left(\frac{(32\sigma_g\sqrt{2\log(5/\delta)}}{\tau\,\mathbb{P}_{\mathcal{P}(u^*)}[\theta<-\tau]} + \sqrt{50\log(5/\delta)}\right)\right)^2 & \text{if } \mathbb{E}_{\mathcal{P}(u^*)}[\theta] \geq 0, \end{cases}$$

Substituting these lower bound values for q in $A(S_{q-0}, \delta)$, we get that the approximation bound A_q satisfies the following inequalities (since $A_q \leq A(S_{q-0}, \delta)$):

$$A_q \leq A(S_{q-0}, \delta) \leq \begin{cases} \tau\,\mathbb{P}_{\mathcal{P}(u^*)}[\theta > \tau]/4 & \text{if } \mathbb{E}_{\mathcal{P}(u^*)}[\theta] < 0; \\ \tau\,\mathbb{P}_{\mathcal{P}(u^*)}[\theta < -\tau]/4 & \text{if } \mathbb{E}_{\mathcal{P}(u^*)}[\theta] \geq 0. \end{cases}$$

[3] Lemma 4.4.2 doesn't exactly state this: it states that any type u will be compliant if the input samples S_0 satisfy the above bounds. Yet, we can simply imagine that phase q is . Proving compliance starting from any phase $q > 0$ is just the same as proving compliance from phase 0. Intuitively, you can imagine we simply run Algorithm 2 starting at phase q initialized with the samples collected up until phase q.

□

Finally, after Algorithm 2 has become compliant for both types of agents, we achieve the following accuracy guarantee for the final treatment estimate $\hat{\theta}_S$.

Theorem B.IV.2 (Treatment Effect Confidence Interval from Algorithm 2 with Full Compliance)**.** *Suppose sample set* $S = (x_i, y_i, z_i)_{i=1}^{|S|}$ *is collected from Algorithm 2 during* $|S|$ *rounds when all agents comply. We form estimate* $\hat{\theta}_S$ *of the treatment effect* θ*. For any* $\delta \in (0,1)$*, with probability at least* $1-\delta$*,*

$$\left|\hat{\theta}_S - \theta\right| \leq 8\sigma_g\sqrt{\frac{2\log(2/\delta)}{|S|}}$$

Proof. We assume Algorithm 2 is initialized and allowed to run long enough such that both types 0 and 1 become compliant at some point. From samples $S = (x_i, y_i, z_i)_{i=1}^{|S|}$ collected during these rounds (from i to $|S|$)), we form an estimate $\hat{\theta}_S$ of the treatment effect θ. By Theorem 4.2.1, this estimate satisfies the following bound with probability at least $1-\delta$ for any $\delta \in (0,1)$:

$$|\hat{\theta}_S - \theta| \leq A(S,\delta) = \frac{2\sigma_g\sqrt{2|S|\log(2/\delta)}}{\left|\sum_{i=1}^{|S|}(x_i - \bar{x})(z_i - \bar{z})\right|}. \tag{B.27}$$

Recall that $\bar{z} = \frac{1}{2}$ throughout Algorithm 2. Then, by Theorem B.II.2, the denominator of the bound in Equation (B.27) above satisfies the following bound:

$$\left|\sum_{i=1}^{|S|}(x_i - \bar{x})(z_i - \bar{z})\right| \geq \frac{|S|}{4}. \tag{B.28}$$

Therefore, by Equations (B.27) and (B.28), the confidence interval $|\hat{\theta}_S - \theta|$ satisfies the following upper bound:

$$|\hat{\theta}_S - \theta| \leq 8\sigma_g\sqrt{\frac{2\log(2/\delta)}{|S|}}$$

□

B.IV.3 Proof of Lemma 4.5.2

Lemma 4.5.2 (Lower bound on ℓ for Type u Compliance in Algorithm 2)**.** *Recall that S_ℓ denotes the samples collected from the second stage of Algorithm 1. Let S_ℓ be the input samples S_0 in Algorithm 2. Assume that p_{c_1} proportion of agents in the population are compliant with recommendations of Algorithm 1 and length ℓ satisfies:*

$$\ell \geq \begin{cases} \left(\frac{\kappa_1}{\tau\,\mathbb{P}_{\mathcal{P}^{(u)}}[\theta > \tau]} + \kappa_2\right)^2 & \text{if } \underset{\mathcal{P}^{(u)}}{\mathbb{E}}[\theta] < 0 \\ \left(\frac{\kappa_1}{\tau\,\mathbb{P}_{\mathcal{P}^{(u)}}[\theta < -\tau]} + \kappa_2\right)^2 & \text{if } \underset{\mathcal{P}^{(u)}}{\mathbb{E}}[\theta] \geq 0 \end{cases} \tag{4.6}$$

for some $\tau \in (0,1)$ and where $\kappa_1 := \frac{8\sigma_g\sqrt{2\log(5/\delta)}}{p_{c_1}\rho(1-\rho)}$ and $\kappa_2 := (3-\rho)\sqrt{\frac{\rho\log(5/\delta)}{2(1-\rho)}}$ for any $\delta \in (0,1)$. Then any agent of type u will comply with recommendations of Algorithm 2.

Proof. By assumption, Algorithm 1 is initialized so that agents of type 0 comply and we collect S_ℓ samples from the second stage. Then, for any $\delta \in (0,1)$ and some $\tau \in (0,1)$, approximation bound $A(S_\ell, \delta)$ satisfies:

$$\begin{aligned} A(S_\ell,\delta) &\leq \frac{2\sigma_g\sqrt{2\log(3/\delta)}}{\rho\left(p_0(1-\rho)\sqrt{\ell} - \sqrt{\frac{(1-\rho)\log(3/\delta)}{2}}\right)} && \text{(by Theorem 4.3.3)} \\ &\leq \frac{2\sigma_g\sqrt{2\log(3/\delta)}}{\rho\left(p_0(1-\rho)\left(\frac{8\sigma_g\sqrt{2\log(3/\delta)}}{p_0\rho(1-\rho)\tau\,\mathbb{P}_{\mathcal{P}^{(0)}}[\theta>\tau]} + \frac{\sqrt{(1-\rho)\log(3/\delta)}}{2p_0(1-\rho)}\right) - \sqrt{\frac{(1-\rho)\log(3/\delta)}{2}}\right)} \\ & && \text{(by Equation (4.6))} \\ &\leq \frac{\tau\,\mathbb{P}_{\mathcal{P}^{(0)}}[\theta > \tau]}{4} \end{aligned}$$

Thus, by Lemma 4.4.2, if we let the samples S_ℓ collected from the second stage of Algorithm 1 be the input samples S_0 in Algorithm 2, i.e. $S_0 = S_\ell$, then that agents of type 0 will comply with recommendations of Algorithm 2.

□

Racing Stage First Part Estimation Bound

Theorem 4.4.3 (Treatment Effect Confidence Interval from Algorithm 2 with Partial Compliance)**.** *With set $S = (x_i, y_i, z_i)_{i=1}^{|S|}$ of $|S|$ samples collected from Algorithm 2 where*

p_c is the fraction of compliant agents in the population, we form an estimate $\hat{\theta}_S$ of the treatment effect θ. With probability at least $1-\delta$,

$$\left|\hat{\theta}_S - \theta\right| \leq A(S,\delta) \leq \frac{8\sigma_g\sqrt{2\log(5/\delta)}}{p_c\sqrt{|S|} - \sqrt{50\log(5/\delta)}}$$

for any $\delta \in (0,1)$, where σ_g is the variance of $g^{(u_i)}$.

Proof. First, Theorem 4.2.1 demonstrates, for any $\delta_1 \in (0,1)$, with probability at least $1-\delta_1$ that the approximation bound

$$\left|\hat{\theta}_S - \theta\right| \leq A(S,\delta) = \frac{2\sigma_g\sqrt{2|S|\log(2/\delta_1)}}{\left|\sum_{i=1}^{\ell}(x_i-\bar{x})(z_i-\bar{z})\right|}. \tag{B.29}$$

Next, recall that the mean recommendation $\bar{z} = \frac{1}{2}$ throughout Algorithm 2. We assume Algorithm 2 to be initialized with parameters such that its recommendations are compliant for agents of type 0. In the worst case, only type 0 agents are compliant. Therefore, Theorem B.II.2 implies that, for any $\delta_2 \in (0,1)$, with probability at least $1-\delta_2$ that

$$\left|\sum_{i=1}^{|S|}(x_i-\bar{x})(z_i-\bar{z})\right| \geq \frac{|S|}{4}\left(p_0 - \sqrt{\frac{\log(1/\delta_2)}{|S|}}\right). \tag{B.30}$$

With a union bound over Equations (B.29) and (B.30) while letting $\delta_1 = \delta_2 = \frac{\delta}{3}$ for any $\delta \in (0,1)$, we conclude: with probability at least $1-\delta$,

$$A(S,\delta) \leq \frac{8\sigma_g\sqrt{2\log(3/\delta)}}{p_0\sqrt{|S|} - \sqrt{\log(3/\delta)}}$$

$\square$

B.V Missing Regret Proofs for Section 3.10

B.V.1 Pseudo-regret Proof

Lemma 4.5.4 (Pseudo-regret)**.** *The pseudo-regret accumulated from policy π_c is bounded for any $\theta \in [-1, 1]$ as follows, with probability at least $1 - \delta$ for any $\delta \in (0, 1)$:*

$$R_\theta(T) \leq L_1 + O(\sqrt{T \log(T/\delta)}) \tag{4.8}$$

for sufficiently large time horizon T, where the length of Algorithm 1 is $L_1 = \ell + 2\max\left(\frac{\ell_0}{p_0}, \frac{\ell_1}{p_1}\right)$.

Proof. Recall that the clean event $\mathcal{C}$, as defined in the proof of Lemma 4.4.2, entails that the approximation bound over all rounds. If event $\mathcal{C}$ fails (i.e. $\neg\mathcal{C}$ holds), then the pseudo-regret may only be bounded by the maximum possible value, which is at most $T|\theta|$.

Assume now that $\mathcal{C}$ holds for every round L_2 in Algorithm 2. Then, the absolute value of the treatment $|\theta| \leq |\hat{\theta}_{S_{L_2}}| + A(S_{L_2}, \delta)$ where $A(S_{L_2}, \delta)$ is the approximation bound based on the samples S_{L_2} collected from L_2 rounds of Algorithm 2. Before the stopping criterion of Algorithm 2 is invoked, we also have $|\hat{\theta}_{S_{L_2}}| \leq A(S_{L_2}, \delta)$. Hence, the treatment effect absolute value satisfies the following inequalities:

$$|\theta| \leq 2A(S_{L_2}, \delta) \leq \frac{16\sigma_g\sqrt{2\log(5T/\delta)}}{p_{c_2}\sqrt{L_2} - \sqrt{50\log(5T/\delta)}}$$

Assuming that $L_2 \geq \frac{200\log(5T/\delta)}{p_{c_2}^2}$, then $p_{c_2}\sqrt{L_2} - \sqrt{50\log(5T/\delta)} \geq \frac{p_{c_2}\sqrt{L_2}}{2}$ and, to carry on:

$$\begin{aligned} |\theta| &\leq \frac{32\sigma_g\sqrt{2\log(5T/\delta)}}{p_{c_2}\sqrt{L_2}} \\ \Rightarrow L_2 &\leq \frac{2048\sigma_g^2\log(5T/\delta)}{p_{c_2}^2|\theta|^2} \end{aligned}$$

Therefore, a winner is declared —i.e. the treatment effect is definitively either positive or not— by the following round of Algorithm 2:

$$L_2^* = \frac{2048\sigma_g^2 \log(5T/\delta)}{p_{c_2}^2 |\theta|^2}$$

After this, the winner is recommended for the remainder of the rounds and (because we assume event $\mathcal{C}$ holds) no regret is accumulated for the remaining rounds (until time horizon T).

Note that, by the length the assumption that $L_2 \geq \frac{200\log(5T/\delta)}{p_{c_2}^2}$ holds trivially for L_2^* if $|\theta| \leq \frac{16\sigma_g}{5}$. Otherwise, we simply assume that $L_2 \geq \frac{200\log(5T/\delta)}{p_{c_2}^2}$ holds. We may incorporate this bound into a necessary lower bound on the time horizon T, such that we assume $T \geq L_1 + L_2 \geq L_1 + \frac{200\log(5T/\delta)}{p_{c_2}^2}$.

Now, we demonstrate the amount of regret accumulated during these n^* rounds of Algorithm 2 before a winner is declared. During each phase Algorithm 2, each control and treatment each get recommended $n^*/2$ times.

Furthermore, recall that p_{c_2} fraction of agents are compliant throughout all rounds of Algorithm 2. Without loss of generality, assume that these agents initially prefer control and the rest of the $1 - p_{c_2}$ fraction of agents prefer treatment (and do not comply). Then, if the treatment $\theta < 0$, then in expectation over the randomness of the arrival of agents, the regret is on average $(1 - p_{c_2}/2)|\theta|$. Then, the total accumulated regret throughout Algorithm 2 in policy π_c is given as such:

$$R_2(T) \leq \frac{2048(1 - p_{c_2}/2)\sigma_g^2 \log(5T/\delta)}{p_{c_2}^2 |\theta|} \tag{B.31}$$

On the other hand, if treatment $\theta \geq 0$, then on average, the regret for each phase is $p_{c_2}|\theta|/2$, then the total accumulated regret for Algorithm 2 is given:

$$R_2(T) \leq \frac{1024\sigma_g^2 \log(5T/\delta)}{p_{c_2} |\theta|} \tag{B.32}$$

Observe that the pseudo-regret for each round t of the policy π_c over the entire T rounds is at most that of Algorithm 2 plus $|\theta|$ per round of Algorithm 1. Recall that, by policy π_c, there are L_1 total rounds of Algorithm 1. Alternatively, we can also upper bound the total pseudo-regret by $|\theta|$ per each round. Therefore, the total accumulated pseudo-regret for policy π_c once we reach the time horizon T is bounded as follows.

If the treatment effect $\theta \geq 0$, then the total regret of policy π_c satisfies the following bound with probability at least $1-\delta$, for any $\delta \in (0,1)$ which satisfies the condition that $L_1 \geq \frac{2\sqrt{\log(5T/\delta)}}{p_{c_2}}$:

$$R(T) \leq \min\left\{L_1\rho|\theta| + \frac{1024\sigma_g^2 \log(5T/\delta)}{p_{c_2}|\theta|}, T|\theta|\right\}$$

We can solve for $|\theta|$ in terms of T at the point when $T|\theta|$ is the better regret:

$$\begin{aligned}
T|\theta| &= \frac{1024\sigma_g^2 \log(5T/\delta)}{p_{c_2}|\theta|} \\
\Rightarrow |\theta|^2 &= \frac{1024\sigma_g^2 \log(5T/\delta)}{p_{c_2}T} \\
\Rightarrow |\theta| &= 32\sigma_g\sqrt{\frac{\log(5T/\delta)}{p_{c_2}T}}
\end{aligned}$$

Substituting this expression for $|\theta|$ back into our expression for the total pseudo-regret, we get the following:

$$\begin{aligned}
R(T) &\leq \min\left\{L_1\rho|\theta| + \frac{1024\sigma_g^2 \log(5T/\delta)}{32\sigma_g p_{c_2}\sqrt{\frac{\log(5T/\delta)}{p_{c_2}T}}}, 32\sigma_g T\sqrt{\frac{\log(5T/\delta)}{p_{c_2}T}}\right\} \\
&= \min\left\{L_1\rho|\theta| + 32\sigma_g\sqrt{\frac{T\log(5T/\delta)}{p_{c_2}}}, 32\sigma_g\sqrt{\frac{T\log(5T/\delta)}{p_{c_2}}}\right\} \\
&\leq L_1\rho + O(\sqrt{T\log(T/\delta)})
\end{aligned}$$

The $L_1\rho$ regret for Algorithm 1 in the last line above is given because $|\theta| \leq 1$.

Following a similar analysis, if the treatment effect $\theta < 0$, then the total pseudo-regret accumulated following policy π_c satisfies the following bound with probability at least

$1-\delta$ for any $\delta \in (0,1)$ which satisfies the condition that $L_1 \geq \frac{2\sqrt{\log(5T/\delta)}}{p_{c_2}}$:

$$R(T) \leq \min\left\{L_1|\theta| + \frac{2048(1-p_{c_2}/2)\sigma_g^2 \log(5T/\delta)}{p_{c_2}^2|\theta|}, T|\theta|\right\} \tag{B.33}$$

$$\leq L_1 + O(\sqrt{T\log(T/\delta)}) \tag{B.34}$$

Note that (as stated above) this regret holds only if the time horizon T is sufficiently large such that $T \geq L_1 + \frac{200\log(5T/\delta)}{p_{c_2}^2}$. $\square$

B.V.2 Regret Proof

Lemma 4.5.5 (Regret)**.** *Policy π_c achieves regret as follows:*

$$\mathbb{E}[R(T)] = O(\sqrt{T\log(T)}) \tag{4.9}$$

for sufficiently large time horizon T.

Proof. We can set parameters $\delta, \ell_0, \ell_1, \ell$, and ρ in terms of the time horizon T, in order to both guarantee compliance throughout policy π_c and to obtain sublinear (expected) regret bound relative to T.

First, to guarantee sublinear expected regret, we must guarantee that $\delta = 1/T^2$. To meet our compliance conditions for Algorithm 2, we must set

$$\delta \leq \frac{\tau\, \mathbb{P}_{\mathcal{P}(u)}[|\theta| \geq \tau]}{2\left(\tau\, \mathbb{P}_{\mathcal{P}(u)}[|\theta| \geq \tau] + 1\right)},$$

for some τ. These may be expressed as conditions on the time horizon T: for any $\delta \in (0,1)$ which satisfies the above compliance conditions, we set T sufficiently large to satisfy the following condition:

$$T \geq \frac{1}{\sqrt{\delta}} \geq \sqrt{\frac{2\left(\tau\, \mathbb{P}_{\mathcal{P}(u)}[|\theta| \geq \tau] + 1\right)}{\tau\, \mathbb{P}_{\mathcal{P}(u)}[|\theta| \geq \tau]}} \tag{B.35}$$

Second, recall that p_0 and p_1 denote the fractions in the population of agents who are never-takers and always-takers, respectively. Furthermore, recall that p_{c_1} and p_{c_2} denote the fractions of agents who comply with Algorithm 1 and Algorithm 2, respectively.

Assume that the length of the first stage of Algorithm 1 is non-zero and the exploration probability ρ is set to be small enough in order to guarantee compliance throughout Algorithm 1. The length L_1 of Algorithm 1 must be sufficiently large so that p_c fraction of agents comply in Algorithm 2, as well. However, in order to guarantee sublinear regret, we also need that

$$T \geq L_1^2 = (2\max(\ell_0/p_0, \ell_1/p_1) + \ell)^2 \tag{B.36}$$

Recall that the clean event $\mathcal{C}$, as defined in the proof of Lemma 4.4.2, entails that the approximation bound over all rounds. This event $\mathcal{C}$ holds with probability at least $1-\delta$ for any $\delta \in (0,1)$. Conditional on the failure event $\neg\mathcal{C}$, policy π_c accumulates at most linear pseudo-regret in terms of T, i.e. $T|\theta|$. Thus, in expectation it accumulates at most $T|\theta|\delta$ regret.

Then, with the above assumptions on T in mind, the expected regret of policy π_c is:

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathcal{P}^{(u)}}[R(T)] &= \mathbb{E}[R(T)|\neg\mathcal{C}] \mathop{\mathbb{P}}_{\mathcal{P}^{(u)}}[\neg\mathcal{C}] + \mathbb{E}[R(T)|\mathcal{C}] \mathop{\mathbb{P}}_{\mathcal{P}^{(u)}}[\mathcal{C}] \\
&\leq T\delta + \left(L_1 + O\left(\sqrt{T\log(T/\delta)}\right)\right) \\
&= \frac{1}{T} + \left(\sqrt{T} + O\left(\sqrt{T\log(T^3)}\right)\right) \\
&= \frac{1}{T} + O\left(\sqrt{T\log(T)}\right) \\
&= O\left(\sqrt{T\log(T)}\right)
\end{aligned}$$

Therefore, assuming that all hyperparameters $\delta, \ell_0, \ell_1, \ell$, and ρ are set to incentivize compliance of some nonzero proportion of agents throughout π_c and assuming that T is sufficiently large so as to satisfy both Equations (B.35) and (B.36) above, policy π_c achieves sublinear regret. □

B.VI Missing Proofs and Materials for Section 3.11

B.VI.1 Model

We now consider a general setting for the sequential game between a social planner and a sequence of agents over T rounds, as first mentioned in Section 4.2. In this setting, there are k treatments of interest, each with unknown treatment effect. In each round t, a new agent indexed by t arrives with their private type u_t drawn independently from a distribution $\mathcal{U}$ over the set of all private types U. Each agent t has k actions to choose from, numbered 1 to k. Let $x_t \in \mathbb{R}^k$ be a one-hot encoding of the action choice at round t, i.e. a k-dimensional unit vector in the direction of the action. For example, if the agent at round t chooses action 2, then $x_t = \mathbf{e}_2 = (0, 1, 0, \cdots, 0) \in \mathbb{R}^k$. Additionally, agent t receives an action recommendation $z_t \in \mathbb{R}^k$ from the planner upon arrival. After selecting action $x_t \in \mathbb{R}^k$, agent t receives a reward $y_t \in \mathbb{R}$, given by

$$y_t = \langle \theta, x_t \rangle + g_t^{(u_t)} \tag{B.37}$$

where $g_t^{(u_t)}$ denotes the confounding *baseline reward* which depends on the agent's private type u_t. Each $g_t^{(u_t)}$ is drawn from a sub-Gaussian distribution with a sub-Gaussian norm of σ_g. The social planner's goal is to estimate the treatment effect vector $\theta \in \mathbb{R}^k$ and maximize the total expected reward of all T agents.

History, beliefs, and action choice. As in the body of the paper, the history H_t is made up of all tuples (z_i, x_i, y_i) over all rounds from $i = 1$ to t. Additionally, before the game starts, the social planner commits to recommendation policy π, which is known to all agents. Each agent also knows the number of the round t when they arrive. Their private type u_t maps to their prior belief $\mathcal{P}^{(u_t)}$, which is a joint distribution over the treatment effect θ and noisy error term $g^{(u)}$. With all this information, the agent t selects the action x_t which they expect to produce the most reward:

$$x_t := \mathbf{e}_{a_t} \qquad \text{where} \qquad a_t := \operatorname*{argmax}_{1 \le j \le k} \mathop{\mathbb{E}}_{\mathcal{P}^{(u_t)}, \pi_t} \left[\theta^j \mid z_t, t \right]. \tag{B.38}$$

B.VI.2 Instrumental Variable Estimate and Finite Sample Approximation Bound

As in the body of the paper, we view the planner's recommendations as instruments and perform *instrumental variable (IV) regression* to estimate θ.

IV Estimator for $k > 1$ Treatments Our mechanism periodically solves the following IV regression problem: given a set S of n observations $(x_i, y_i, z_i)_{i=1}^n$, compute an estimate $\hat{\theta}_S$ of θ. We consider the following two-stage least square (2SLS) estimator:

$$\hat{\theta}_S = \left(\sum_{i=1}^n z_i x_i^\top\right)^{-1} \sum_{i=1}^n z_i y_i, \tag{B.39}$$

where $(\cdot)^{-1}$ denotes the pseudoinverse.

To analyze the 2SLS estimator, we introduce a *compliance matrix* of conditional probabilities that an agent chooses some treatment given a recommendation $\hat{\Gamma}$, given as proportions over a set of n samples $S = (x_i, z_i)_{i=1}^n$, where any entry in $\hat{\Gamma}$ is given as such:

$$\hat{\Gamma}_{ab}(S) = \hat{\mathbb{P}}_S[x = \mathbf{e}_a | z = \mathbf{e}_b] = \frac{\sum_{i=1}^n \mathbb{1}[x = \mathbf{e}_a, z = \mathbf{e}_b]}{\sum_{i=1}^n \mathbb{1}[z = \mathbf{e}_b]} \tag{B.40}$$

Then, we can write the action choice x_i as such:

$$x_i = \hat{\Gamma} z_i + \eta_i, \tag{B.41}$$

where $\eta_i = x_i - \hat{\Gamma} z_i$. Now, we can rewrite the reward y_i at round i as such:

$$\begin{aligned} y_i &= \big\langle \theta, (\hat{\Gamma} z_i + \eta_i) \big\rangle + g_i^{(u_i)} \\ &= \langle \underbrace{\theta \hat{\Gamma}}_{\beta}, z_i \rangle + \langle \theta, \eta_i \rangle + g_i^{(u_i)} \\ &= \langle \beta, z_i \rangle + \langle \theta, \eta_i \rangle + g_i^{(u_i)}. \end{aligned}$$

This formulation allows us to express and bound the error between the treatment effect θ and its IV estimate $\hat{\theta}$ in Theorem 4.6.1.

B.VI.3 Proof of Theorem 3.11.1

Theorem 4.6.1 (Many Treatments Effect Approximation Bound)**.** *Let $z_1, \ldots, z_n \in \{0,1\}^k$ be a sequence of instruments. Suppose there is a sequence of n agents such that each agent i has private type u_i drawn independently from $\mathcal{U}$, selects x_i under instrument z_i and receives reward y_i. Let sample set $S = (x_i, y_i, z_i)_{i=1}^n$. The approximation bound $A(S, \delta)$ is given as such:*[4]

$$A(S,\delta) = \frac{\sigma_g \sqrt{2nk\log(k/\delta)}}{\sigma_{\min}\left(\sum_{i=1}^n z_i x_i^\top\right)},$$

and the IV estimator given by Equation (4.11) satisfies

$$\left\|\hat{\theta}_S - \theta\right\|_2 \leq A(S,\delta)$$

with probability at least $1-\delta$ for any $\delta \in (0,1)$.

Proof. Given a sample set $S = (x_i, y_i, z_i)_{i=1}^n$ of size n, we form an estimate of the treatment effect $\hat{\theta}_S$ via Two-Stage Least Squares regression (2SLS). In the first stage, we regress y_i onto z_i to get the empirical estimate $\hat{\beta}_S$ and x_i onto z_i to get $\hat{\Gamma}_S$ as such:

$$\hat{\beta}_S := \left(\sum_{i=1}^n z_i z_i^\top\right)^{-1}\left(\sum_{i=1}^n z_i y_i\right) \qquad \text{and} \qquad \hat{\Gamma}_S := \left(\sum_{i=1}^n z_i z_i^\top\right)^{-1}\left(\sum_{i=1}^n z_i x_i^\top\right) \tag{B.42}$$

Now, note that by definition $\theta = (\hat{\Gamma})^{-1}\beta$. In the second stage, we take the inverse of $\hat{\Gamma}$ times $\hat{\beta}$ as the predicted causal effect vector $\hat{\theta}_S$, i.e.

$$\begin{aligned}
\hat{\theta}_S &= \hat{\Gamma}_S^{-1}\hat{\beta}_S \\
&= \left(\sum_{i=1}^n z_i x_i^\top\right)^{-1}\left(\sum_{i=1}^n z_i z_i^\top\right)\left(\sum_{i=1}^n z_i z_i^\top\right)^{-1}\sum_{i=1}^n z_i y_i \\
&= \left(\sum_{i=1}^n z_i x_i^\top\right)^{-1}\sum_{i=1}^n z_i y_i
\end{aligned}$$

[4]The operator $\sigma_{\min}(\cdot)$ denotes the smallest singular value.

Hence, the L2-norm of the difference between θ and $\hat{\theta}_S$ is given as:

$$\begin{aligned}
\left\|\hat{\theta}_S - \theta\right\|_2 &= \left\|\left(\sum_{i=1}^{n} z_i x_i^\top\right)^{-1} \sum_{i=1}^{n} z_i y_i - \theta\right\|_2 \\
&= \left\|\left(\sum_{i=1}^{n} z_i x_i^\top\right)^{-1} \sum_{i=1}^{n} z_i \left(\langle\theta, x_i\rangle + g_i^{(u_i)}\right)^\top - \theta\right\|_2 \\
&= \left\|\left(\sum_{i=1}^{n} z_i x_i^\top\right)^{-1} \left(\sum_{i=1}^{n} z_i x_i^\top \theta + \sum_{i=1}^{n} z_i g_i^{(u_i)}\right) - \theta\right\|_2 \\
&= \left\|\theta + \left(\sum_{i=1}^{n} z_i x_i^\top\right)^{-1} \sum_{i=1}^{n} z_i g_i^{(u_i)} - \theta\right\|_2 \\
&= \left\|\left(\sum_{i=1}^{n} z_i x_i^\top\right)^{-1} \sum_{i=1}^{n} z_i g_i^{(u_i)}\right\|_2 \\
&\leq \left\|\left(\sum_{i=1}^{n} z_i x_i^\top\right)^{-1}\right\|_2 \left\|\sum_{i=1}^{n} z_i g_i^{(u_i)}\right\|_2 \qquad \text{(by Lemma B.I.5)} \\
&= \frac{\left\|\sum_{i=1}^{n} z_i g_i^{(u_i)}\right\|_2}{\sigma_{\min}\left(\sum_{i=1}^{n} z_i x_i^\top\right)}
\end{aligned}$$

Finally, we may bound $\left\|\hat{\theta}_S - \theta\right\|_2$ by upper bounding $\left\|\sum_{i=1}^{n} z_i g_i^{(u_i)}\right\|_2$ in the following lemma B.VI.1.

Lemma B.VI.1. *For any $\delta \in (0, 1)$, with probability at least $1 - \delta$, we have*

$$\left\|\sum_{i=1}^{n} z_i g_i^{(u_i)}\right\|_2 \leq \sigma_g \sqrt{2nk \log(k/\delta)} \tag{B.43}$$

Proof. Recall that the baseline reward $g^{(u)}$ is an independently distributed random variable which, by assumption, has a mean of zero, i.e. $\mathbb{E}[g^{(u)}] = 0$. Because of these properties of $g^{(u)}$, with probability at least $1 - \delta$ for any $\delta \in (0, 1)$, the numerator above satisfies the following upper bound:

$$
\begin{aligned}
\left\|\sum_{i=1}^{n} g_i^{(u_i)} z_i\right\|_2 &= \sqrt{\sum_{j=1}^{k}\left(\sum_{i=1}^{n} g_i^{(u_i)} \mathbb{1}[z_i = \mathbf{e}_j]\right)^2} \\
&= \sqrt{\sum_{j=1}^{k}\left(\sum_{i=1}^{n_j} g_i^{(u_i)}\right)^2} && \text{(where } n_j = \textstyle\sum_{i=1}^{n} \mathbb{1}[z_i = \mathbf{e}_j]\text{)} \\
&\leq \sqrt{\sum_{j=1}^{k}\left(\sigma_g \sqrt{2n_j \log(1/\delta_j)}\right)^2} \\
& && \text{(by Corollary B.I.2 and, by assumption, } \mathbb{E}[g^{(u)}] = 0\text{)} \\
&\leq \sqrt{\sum_{j=1}^{k}\left(\sigma_g \sqrt{2n_j \log(k/\delta)}\right)^2} \\
& && \text{(by a Theorem B.I.7 where } \delta_j = \tfrac{\delta}{k} \text{ for all } j\text{)} \\
&\leq \sqrt{k\left(\sigma_g \sqrt{2n \log(k/\delta)}\right)^2} && \text{(since } n_j \leq n \text{ for all } j\text{)} \\
&= \sigma_g \sqrt{2nk \log(k/\delta)}
\end{aligned}
$$

□

This recovers the stated bound and finishes the proof for Theorem 4.6.1. □

Next, we demonstrate a lower bound which the denominator of the approximation bound $A(S, \delta)$ in Theorem 4.6.1 equals $O(1/\sqrt{|S|})$, where $|S|$ is the size of sample set S.

Theorem B.VI.2 (Treatment Effect Confidence Interval for General k Treatments)**.** *Let $z_1, \ldots, z_n \in \{0,1\}^k$ be a sequence of instruments. Suppose there is a sequence of n agents such that each agent i has private type u_i drawn independently from $\mathcal{U}$, selects x_i under instrument z_i and receives reward y_i. Assume that each agent initially prefers treatment 1, i.e. $x = \mathbf{e}_1$. Let sample set $S = (x_i, y_i, z_i)_{i=1}^n$. Let r be the proportion of recommendations for each treatment $j > 1$ and let $1 - (k-1)r$ be the proportion of recommendations for treatment 1. Let p_c fraction of agents in the population of agents be compliant over the rounds from which S is collected. For any $\delta \in (0,1)$, if $n \geq \frac{rp_c^2}{\log(k/\delta)}$,* *then the approximation bound $A(S,\delta)$ is given as such:*

$$A(S,\delta) \leq \frac{\sigma_g\sqrt{2k\log(k/\delta)}}{\alpha\sqrt{n}} = O\left(\sqrt{\frac{\log(1/\delta)}{n}}\right),$$

and the IV estimator given by Equation (4.11) satisfies

$$\left\|\hat{\theta}_S - \theta\right\|_2 \leq A(S,\delta)$$

with probability at least $1-\delta$, where $\alpha > 0$ is a constant of proportionality given in Claim B.VI.3 below.

Proof. Note that Theorem 4.6.1 holds in this case and it suffices to demonstrate that the denominator is bounded by $\frac{1}{\alpha n}$.

Claim B.VI.3 (Proportionality of the Denominator of the Approximation Bound for k Treatments)**.** *Given all assumptions in Theorem B.VI.2 above, the denominator $\sigma_{\min}\left(\sum_{i=1}^n z_i x_i^\top\right)$ of the approximation bound $A(S,\delta)$ is positive and increases proportionally to n. Formally, $\sigma_{\min}\left(\sum_{i=1}^n z_i x_i^\top\right) = \Omega(n)$.*

Proof. Recall that we assume that every agent initially prefers treatment 1. Thus, whenever agent is recommended any treatment greater than 1 and does not comply, the agent takes treatment 1. At any round i, if agent i is always compliant, then $x_i = z_i$; if not, then $x_i = \mathbf{e}_1$. (If $z_i = \mathbf{e}_1$, then $x_i = z_i = \mathbf{e}_1$ always.) Furthermore, at any round i when $x_i = z_i$, the outer product $z_i x_i^\top = diag(z_i)$, i.e. a diagonal matrix where the diagonal equals z_i. If $x_i = \mathbf{e}_1$, then the outer product

$$z_i x_i^\top = \begin{pmatrix} \uparrow & \uparrow & & \uparrow \\ z_i & \mathbf{0} & \cdots & \mathbf{0} \\ \downarrow & \downarrow & & \downarrow \end{pmatrix},$$

which is a $k \times k$ matrix where the first column is z_i and all other entries are 0. Thus, as long as we have at least one sample of each treatment, i.e. at least one round i where $x_i = z_i = \mathbf{e}_j$ for all $1 \leq j \leq k$, then the sum $\sum_{i=1}^n z_i x_i^\top$ is a lower triangular matrix with all positive entries in the diagonal. To illustrate this, let $\mathbf{A}$ denote the expected mean values of the sum $\sum_{i=1}^{n} z_i x_i^\top$, such that

$$\mathbb{E}\left[\sum_{i=1}^{n} z_i x_i^\top\right] = n \begin{pmatrix} 1-rk & 0 & \cdots & & \cdots & 0 \\ r(1-p_c) & rp_c & 0 & \cdots & \cdots & \vdots \\ r(1-p_c) & 0 & rp_c & 0 & \cdots & \\ \vdots & \vdots & 0 & \ddots & & \vdots \\ \vdots & \vdots & \vdots & & \ddots & 0 \\ r(1-p_c) & 0 & \cdots & \cdots & 0 & rp_c \end{pmatrix} = n\mathbf{A}.$$

Note that

$$\mathbb{E}\left[\left(\sum_{i=1}^{n} z_i x_i^\top\right)^\top \left(\sum_{i=1}^{n} z_i x_i^\top\right)\right] = \mathbb{E}\left[\left(\sum_{i=1}^{n} z_i x_i^\top\right)\right]^\top \mathbb{E}\left[\sum_{i=1}^{n} z_i x_i^\top\right]$$

$$= n^2 \begin{pmatrix} (1-rk)^2 + (k-1)r^2(1-p_c)^2 & r^2 p_c(1-p_c) & \cdots & & \cdots & r^2 p_c(1-p_c) \\ r^2 p_c(1-p_c) & r^2 p_c^2 & 0 & \cdots & \cdots & 0 \\ r^2 p_c(1-p_c) & 0 & r^2 p_c^2 & 0 & \cdots & \vdots \\ \vdots & \vdots & 0 & \ddots & & \vdots \\ \vdots & \vdots & \vdots & & \ddots & 0 \\ r^2 p_c(1-p_c) & 0 & \cdots & \cdots & 0 & r^2 p_c^2 \end{pmatrix}.$$

Furthermore, let $\hat{\mathbf{A}}$ denote the empirical approximation of $\mathbf{A}$ over our n samples, given as such:

$$\sum_{i=1}^{n} z_i x_i^\top = n \begin{pmatrix} 1-rk & 0 & \cdots & & \cdots & 0 \\ r(1-\hat{p}_{c,2}) & r\hat{p}_{c,2} & 0 & \cdots & \cdots & \vdots \\ r(1-\hat{p}_{c,3}) & 0 & r\hat{p}_{c,3} & 0 & \cdots & \vdots \\ \vdots & \vdots & 0 & \ddots & & \vdots \\ \vdots & \vdots & \vdots & & \ddots & 0 \\ r(1-\hat{p}_{c,k}) & 0 & \cdots & \cdots & 0 & r\hat{p}_{c,k} \end{pmatrix} = n\hat{\mathbf{A}},$$

where for any $j \geq 2$, the empirical proportion of agents who comply with the recommended treatment j is denoted as $\hat{p}_{c,j}$. Note that, since $\mathbb{E}[\hat{p}_{c,j}] = p_c$ for all $j \geq 2$, the expected value $\mathbb{E}\left[\hat{\mathbf{A}}\right] = \mathbf{A}$. We may bound the difference between $\hat{p}_{c,j}$ and p_c with high probability, based on the number of times each treatment j is recommended, which is rn. Over n samples, with probability at least $1-\delta_j$ for any $\delta_j \in (0,1)$ for any treatment j, the proportion $\hat{p}_{c,j}$ satisfies the following: $\hat{p}_{c,j} \geq p_c - \sqrt{\frac{\log(1/\delta_j)}{2rn}}$. In order for this bound to hold for all $j \geq 2$, let $\delta_2 = \delta_3 = \cdots = \delta_k = \delta/k$. Then, by a union bound, with probability $1-\delta$ for any $\delta \in (0,1)$, the bound $\hat{p}_{c,j} \geq p_c - \sqrt{\frac{\log(k/\delta)}{2rn}}$ holds simultaneously for all $2 \leq j \leq k$. Thus, for any $\delta \in (0,1)$ and $n \geq \frac{rp_c^2}{\log(k/\delta)}$, each entry in the diagonal of $\hat{\mathbf{A}}$ is positive. Thus, (since it is a triangular matrix) the eigenvalues of $\hat{\mathbf{A}}$ equal the entries in the diagonal and are all positive. Furthermore, because $\text{rank}(\hat{\mathbf{A}}) = \text{rank}(\hat{\mathbf{A}}^\top \hat{\mathbf{A}})$, the singular values of $\hat{\mathbf{A}}$ are all positive, as well.

Thus, for $n \geq \frac{rp_c^2}{\log(k/\delta)}$, the minimum singular value

$$\sigma_{\min}\left(\sum_{i=1}^{n} z_i x_i^\top\right) = n\sigma_{\min}\{\hat{\mathbf{A}}\} = n\alpha = \Omega(n),$$

where $\alpha = \sigma_{\min}(\hat{\mathbf{A}}) > 0$ is some (possibly small) constant of proportionality. ◻

Thus, by Claim B.VI.3 and Theorem 4.6.1, the approximation bound

$$A(S,\delta) \leq \frac{\sigma_g\sqrt{2nk\log(k/\delta)}}{n\alpha} = O\left(\sqrt{\frac{\log(1/\delta)}{n}}\right).$$

◻

Corollary B.VI.4 (Treatment Effect Confidence Interval for General k Treatments)**.** *Given all assumptions in Theorem B.VI.2, plus the assumptions that the minimum compliance rate for any arm is at least $1/k$ and the minimum proportion of treatment 1 recommendations is at least $1/k$, for any $\delta \in (0,1)$, with a large enough sample size n, the approximation bound $A(S,\delta)$ is given as such:*

$$A(S,\delta) = O\left(k\sqrt{\frac{k\log(1/\delta)}{n}}\right) \tag{B.44}$$

Proof. Note that Claim B.VI.3 holds in this case and it suffices to demonstrate that the α is bounded by $\frac{1}{k}$. We focus on the denominator $\sigma_{\min}\left(\sum_{i=1}^{n} z_i x_i^\top\right)$ of the approximation bound $A(S,\delta)$. Note that since z_i and x_i are one-hot encoded vectors, we have:

$$
\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{n} z_i x_i^\top\right] &= \sum_{j=1}^{k}\sum_{i\in S_j} x_i \left(\sum_{i\in S_j} x_i\right)^\top && \text{(where } S_j = \{i : z_i = \mathbf{e}_j\}\text{)}\\
&= n\sum_{j=1}^{k} v_j v_j^\top && \text{(where vector } v_j = (v_{j1}, 0, \cdots 0, v_{jj}, 0, \cdots 0) \in \mathbb{R}^k\text{)}
\end{aligned}
$$

where $\forall j : v_{j1} = \frac{r}{k}(1-p_j)$ is the probability of getting a treatment 1 sample when the recommendation is $j > 1$, the term $v_{11} = 1 - r$ is the probability of recommending treatment 1 (since we assume agents always comply with treatment 1 recommendations), and the term $v_{jj} = \frac{r}{k}p_j$ is the probability of getting a treatment j sample when the recommendation is $j > 1$. By definition, we can write the denominator term squared as:

$$
\begin{aligned}
&\sigma_{\min}\left(\mathbb{E}\left[\sum_{i=1}^{n} z_i x_i^\top\right]\right)^2\\
&= \min_{a:\|a\|=1} a^\top \left(n\sum_{j=1}^{k} v_j v_j^\top\right) a\\
&= n\left[\min_{a:\|a\|=1} (a_1 v_{11})^2 + (a_1 v_{21} + a_2 v_{22})^2 + \cdots + (a_1 v_{k1} + a_k v_{kk})^2\right].
\end{aligned}
$$

Also, without loss of generality, assume that $a_1 > 0$ and $\forall j > 1 : a_j \le 0$.
Substituting the expression above with algorithm-specific variables, we have:

$$
\begin{aligned}
&(a_1 v_{11})^2 + (a_1 v_{21} + a_2 v_{22})^2 + \cdots + (a_1 v_{k1} + a_k v_{kk})^2\\
&= a_1^2(1-r)^2 + \frac{r^2}{k^2}a_1^2\sum_{j=2}^{k}(1-p_j)^2 + \sum_{j=2}^{k} a_j^2 p_j^2 + \frac{2r^2}{k^2}a_1\sum_{j=2}^{k} a_j p_j(1-p_j)\\
&\ge a_1^2(1-r)^2 + \frac{r^2}{k^2}a_1^2\sum_{j=2}^{k}(1-p_j)^2 + p_{\min}^2\sum_{j=2}^{k} a_j^2 + \frac{2r^2}{k^2}a_1\sum_{j=2}^{k} a_j\frac{1}{4}\\
&\ge a_1^2(1-r)^2 + \frac{r^2}{k^2}a_1^2\sum_{j=2}^{k}(1-p_j)^2 + p_{\min}^2(1-a_1^2) - \frac{r^2 a_1}{2k^2}\sqrt{(k-1)(1-a_1^2)}
\end{aligned}
$$

where the second line is direct substitution, the third line comes from lower bounding all p_j^2 terms with the minimum compliance rate $p_{\min}^2$ and lower bounding $p_j(1-p_j)$ by 1/4. The last line comes from the fact that $\|a\|=1$ and from applying Lemma B.I.5 on the last term. Since we assume that the probability of recommending treatment 1 is $1-r \geq \frac{1}{k}$, we have:

$$\begin{aligned}
&\sigma_{\min}\left(\mathbb{E}\left[\sum_{i=1}^{n} z_i x_i^\top\right]\right)^2 \\
&\geq n\Bigg[\min_{a:\|a\|=1} \frac{a^2}{k^2} + \frac{\left(1-\frac{1}{k}\right)^2}{k^2} a_1^2 \sum_{j=2}^{k}(1-p_j)^2 + p_{\min}^2(1-a_1)^2 \\
&\quad - \frac{\left(1-\frac{1}{k}\right)^2}{2k^2} a_1\sqrt{(k-1)(1-a_1^2)}\Bigg] \\
&\geq n\left[\min_{a:\|a\|=1} \frac{a_1^2}{k^2} + p_{\min}^2(1-a_1)^2 - \frac{(k-1)^2}{2k^4} a_1\sqrt{(k-1)(1-a_1^2)}\right] \\
&\geq n\left[\min_{a:\|a\|=1} \frac{a_1^2}{k^2} + p_{\min}^2(1-a_1)^2 - \frac{a_1\sqrt{(k-1)(1-a_1^2)}}{2k^2}\right] \\
&= n\left[\left(a_1\sqrt{\frac{1}{k^2}-p_{\min}^2} - \frac{1}{4k^2}\sqrt{\frac{(k-1)(1-a_1^2)}{\frac{1}{k^2}-p_{\min}^2}}\right)^2 + p_{\min}^2 - \frac{(k-1)(1-a_1^2)}{16k^4\left(\frac{1}{k^2}-p_{\min}^2\right)}\right] \\
&\geq n\left[p_{\min}^2 - \frac{(k-1)(1-a_1^2)}{16k^4\left(\frac{1}{k^2}-p_{\min}^2\right)}\right] \\
&\geq n\left[p_{\min}^2 - \frac{(1-a_1^2)}{16k-16k^3 p_{\min}^2}\right]
\end{aligned}$$

Since we assume that the minimum compliance rate $p_{\min} \geq \frac{1}{k}$, we have:

$$p_{\min}^2 \geq \frac{1}{k^2} \Rightarrow 16k - 16k^3 p_{\min}^2 \leq 0$$

Therefore, we have $\alpha = \frac{1}{k}$ and

$$\sigma_{\min}\left(\mathbb{E}\left[\sum_{i=1}^{n} z_i x_i^\top\right]\right)^2 \geq \frac{n}{k^2}$$

We apply Theorem B.I.6 to this matrix to get that, with probability at least $1-\delta$, for $\delta \in (0,1)$:

$$\sigma_{\min}\left(\sum_{i=1}^{n} z_i x_i^\top\right) \geq \sqrt{\frac{n}{k^2} - \log(k/\delta)}$$

Hence, we have the approximation bound $A(S,\delta)$ for Algorithm 4 is given as

$$A(S,\delta) \leq \frac{\sigma_g\sqrt{2nk\log(k/\delta)}}{\sqrt{\frac{n}{k^2} - \log(k/\delta)}\sqrt{n}} = O\left(k\sqrt{\frac{k\log(1/\delta)}{n}}\right)$$

B.VI.4 Extensions of Algorithms 1 and 2 and Recommendation Policy to k Treatments

We assume that every agent —regardless of type— shares the same prior ordering of the treatments, such that all agents prior expected value for treatment 1 is greater than their prior expected value for treatment 2 and so on. First, Algorithm 3 is a generalization of Algorithm 1 which serves the same purpose: to overcome complete non-compliance and incentivize some agents to comply eventually. The incentivization mechanism works the same as in Algorithm 1, where we begin by allowing all agents to choose their preferred treatment —treatment 1— for the first ℓ rounds. Based on the ℓ samples collected from the first stage, we then define a number of events $\xi_j^{(u)}$ —which are similar to event ξ from Algorithm 1— that each treatment $j \geq 2$ has the largest expected reward of any treatment and treatment 1 has the smallest, according to the prior of type u:

$$\xi_i^{(u)} := \left(\bar{y}_\ell^1 + C \leq \min_{1<j<i} \bar{y}_\ell^j - C \quad \text{and} \quad \max_{1<j<i} \bar{y}_\ell^j + C \leq \mu_i^{(u)}\right), \tag{B.45}$$

where $C = \sigma_g\sqrt{\frac{2\log(3/\delta)}{\ell}} + \frac{1}{4}$ for any $\delta \in (0,1)$ and where $\bar{y}_\ell^1$ denotes the mean reward for treatment 1 over the ℓ samples of the first stage of Algorithm 3. Thus, if we set the exploration probability ρ small enough, then some subset of agents will comply with all recommendations in the second stage of Algorithm 3.

Second, Algorithm 4 is a generalization of Algorithm 2, which is required to start with at least partial compliance and more rapidly and incentivizes more agents to comply eventually. The incentivization mechanism works the same as in Algorithm 1, where we begin by allowing all agents to choose their preferred treatment —treatment 1— for the first ℓ rounds. Based on the ℓ samples collected from the first stage, we then define a number of events —which are similar to event ξ from Algorithm 1— that each treatment $j \geq 2$ has the largest expected reward of any treatment and treatment 1 has the smallest. Thus, if we set the exploration probability ρ small enough, then some subset of agents will comply with all recommendations in the second stage of Algorithm 3.

□

Corollary B.VI.5. *(Pairwise Treatment Effect Confidence Interval for General k Treatments) Given all assumptions in Corollary B.VI.4, the pairwise approximation bound between any two particular arms a, b is given as*

$$|(\theta^a - \theta^b) - (\hat{\theta}^a - \hat{\theta}^b)| = A^{ab}(S, \delta) \leq \sqrt{2}A(S, \delta)$$

where $\hat{\theta}^a$ and $\hat{\theta}^b$ are the IV estimate for the treatment effect of arm a and arm b, respectively.

Proof. We have:

$$\begin{aligned}
|(\theta^a - \theta^b) - (\hat{\theta}^a - \hat{\theta}^b)| &= |(\hat{\theta}^a - \theta^a) + (\hat{\theta}^b - \theta^b)| \\
&\leq |\hat{\theta}^a - \theta^a| + |\hat{\theta}^b - \theta^b| && \text{(by Triangle Inequality)} \\
&\leq \sqrt{2}\sqrt{(\hat{\theta}^a - \theta^a)^2 + (\hat{\theta}^b - \theta^b)^2} && \text{(by Lemma B.I.5)} \\
&\leq \sqrt{2}\sqrt{\sum_{i=1}^{k}(\hat{\theta}^i - \theta^i)^2} \\
&\leq \sqrt{2}A(S, \delta)
\end{aligned}$$

This recovers the stated bound and we only pay a small constant ($\sqrt{2}$) to obtain a pairwise approximation bound from our IV estimator. □

Theorem B.VI.6. *Let G^{vw} denote the gap between the causal effects of any arms v and w, i.e. $G^{vw} := \theta^v - \theta^w$ and let G^v denote the smallest gap for arm v, i.e. $G^v := \theta^v - \max_{w \neq v} \theta^w = \min_{w \neq v} \theta^v - \theta^w$.*

Let $A^{vw}(S, \delta)$ denote a high-probability upper bound (with probability at least $1 - \delta$) on the difference between the true gap G^{vw} (for causal effects θ^v and θ^w for arms v and *w) and its estimate $\widehat{G}^{vw}$ based on the sample set S, i.e.*

$$\left|G^{vw} - \widehat{G}^{vw}\right| = \left|\theta^v - \theta^w - \left(\hat{\theta}^v - \hat{\theta}^w\right)\right| < A^{vw}(S, \delta).$$

Furthermore, let $A^v(S, \delta)$ denote a high-probability upper bound on the difference between the true minimum gap G^v for arm v and its empirical estimate $\widehat{G}^v$ based on sample set S, i.e.

$$\left|G^{v} - \widehat{G}^{v}\right| = \left|\theta^v - \theta^{w_{\min}} - \left(\hat{\theta}^v - \hat{\theta}^{w_{\min}}\right)\right| < A^{v}(S, \delta),$$

where $w_{\min} = \mathrm{argmin}_{w \neq v}\, \theta^v - \theta^w$. For shorthand, let A^v_q denote the best (i.e. smallest) approximation bound $A^v(S_q^{BEST}, \delta)$ by phase q.

Recall that Algorithm 4 is initialized with samples $S_0 = (x_i, y_i, z_i)_{i=1}^{|S_0|}$. Any agent at time t with type u_t will comply with recommendation $z_t = \mathbf{e}_v$ for arm v from policy π_t according to Algorithm 4, if the following holds for some $\tau \in (0,1)$:

$$A^v(S_0, \delta) \leq \tau \underset{\pi_t, \mathcal{P}^{(u_t)}}{\mathbb{P}}[G^v \geq \tau]/4.$$

Proof. Any agent at time t with type u_t will comply with an arm v recommendation $z_t = \mathbf{e}_v$ from policy π_t following Algorithm 4, if the following holds: For any two treatments $v, w \in B$,

$$\underset{\pi_t, \mathcal{P}^{(u_t)}}{\mathbb{E}}[\theta^v - \theta^w | z_t = \mathbf{e}_v] \underset{\pi_t, \mathcal{P}^{(u_t)}}{\mathbb{P}}[z_t = \mathbf{e}_v] \geq 0.$$

We will prove a stronger statement:

$$\underset{\pi_t, \mathcal{P}^{(u_t)}}{\mathbb{E}}[\theta^v - \max_{w \neq v} \theta^w | z_t = \mathbf{e}_v] \underset{\pi_t, \mathcal{P}^{(u_t)}}{\mathbb{P}}[z_t = \mathbf{e}_v] \geq 0.$$

We can prove this in largely the same way as we proved Lemma 4.4.2 in Section B.IV.1: we simply replace θ and A^v_q in the proof for Lemma 4.4.2 with G^v and A^v_q, respectively.

The clean event C is given as:

$$\mathcal{C} := \left(\forall q \geq 0 : |G^v - \widehat{G}^v| < A^v_q\right).$$

By Corollary B.VI.4, for event $\mathcal{C}$, the failure probability $\mathbb{P}[\neg\mathcal{C}] \leq \delta$. We assume that

$$\delta \leq \frac{\tau\,\mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[G^v \geq \tau]}{2\tau\,\mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[G^v \geq \tau] + 2}.$$

First, we marginalize $\mathbb{E}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta^v - \max_{w\neq v}\theta^w | z_t = \mathbf{e}_v]\,\mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t = \mathbf{e}_v]$ based on the clean event $\mathcal{C}$, such that

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[G^v | z_t = \mathbf{e}_v] \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t = \mathbf{e}_v] \\
&= \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[G^v | z_t = \mathbf{e}_v, \mathcal{C}] \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t = \mathbf{e}_v, \mathcal{C}] + \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[G^v | z_t = \mathbf{e}_v, \neg\mathcal{C}] \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t = \mathbf{e}_v, \neg\mathcal{C}] \\
&\geq \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[G^v | z_t = \mathbf{e}_v, \mathcal{C}] \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t = \mathbf{e}_v, \mathcal{C}] - \delta \\
&\geq \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[G^v | z_t = \mathbf{e}_v, \mathcal{C}] \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t = \mathbf{e}_v, \mathcal{C}] - \delta \\
&\geq \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[G^v | z_t = \mathbf{e}_v, \mathcal{C}] \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t = \mathbf{e}_v, \mathcal{C}] - \frac{\tau\,\mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[G^v \geq \tau]}{2\tau\,\mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[G^v \geq \tau] + 2}.
\end{aligned}$$

Next, we marginalize $\mathbb{E}_{\mathcal{P}^{(u_t)},\pi_t}[G^v | z_t = \mathbf{e}_v, \mathcal{C}]\,\mathbb{P}_{\mathcal{P}^{(u_t)},\pi_t}[z_t = \mathbf{e}_v, \mathcal{C}]$ based on four possible ranges which G^v lies on:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathcal{P}^{(u_t)},\pi_t}[G^v | z_t = \mathbf{e}_v, \mathcal{C}] \mathop{\mathbb{P}}_{\mathcal{P}^{(u_t)},\pi_t}[z_t = \mathbf{e}_v, \mathcal{C} \\
&= \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[G^v | z_t = \mathbf{e}_v, \mathcal{C}, G^v \geq \tau] \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t = \mathbf{e}_v, \mathcal{C}, G^v \geq \tau] \\
&\quad + \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[G^v | z_t = \mathbf{e}_v, \mathcal{C}, 0 \leq G^v < \tau] \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t = \mathbf{e}_v, \mathcal{C}, 0 \leq G^v < \tau] \\
&\quad + \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[G^v | z_t = \mathbf{e}_v, \mathcal{C}, -2A_q^v < G^v < 0] \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t = \mathbf{e}_v, \mathcal{C}, -2A_q^v < G^v < 0] \\
&\quad + \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[G^v | z_t = \mathbf{e}_v, \mathcal{C}, G^v \leq -2A_q^v] \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t = \mathbf{e}_v, \mathcal{C}, G^v \leq -2A_q^v]
\end{aligned} \tag{B.46}$$

Because A_q^v is the smallest approximation bound derived from samples collected over any phase q of Algorithm 4 (including the initial sample set S_0), the following holds:

$$\begin{aligned}
2A_q^v &\le 2A^v(S_0,\delta) \\
&\le \frac{\tau\,\mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[G^v \ge \tau]}{2} && \text{(by assumption } A^v(S_0,\delta) \le \tau\,\mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[G^v \ge \tau]/4) \\
&\le \tau
\end{aligned}$$

Conditional on $\mathcal{C}$, $|G^v - \widehat{G}_q^v| < A_q^v$. Thus, if $G^v \ge \tau \ge 2A_q^v$, then $\widehat{G}_q^v \ge \tau - A_q^v \ge A_q^v$, which invokes the stopping criterion for the while loop in Algorithm 4. Thus, all other arms must have been eliminated from the race before phase $q = 1$ and arm v is recommended almost surely throughout Algorithm 4, i.e. $\mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t = \mathbf{e}_v, \mathcal{C}, G^v \ge \tau] = \mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[\mathcal{C}, G^v \ge \tau]$. Similarly, if $G^v \le -2A_q^v$, then $\widehat{G}_q^v \le -A_q^v$ by phase $q = 1$ and arm v is recommended almost never, i.e. $\mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t = \mathbf{e}_v, \mathcal{C}, G^v < -2A_q^v] = 0$. Substituting in these probabilities (and substituting minimum possible expected values), we proceed:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathcal{P}^{(u_t)},\pi_t}[G^v | z_t = \mathbf{e}_v, \mathcal{C}] \mathop{\mathbb{P}}_{\mathcal{P}^{(u_t)},\pi_t}[z_t = \mathbf{e}_v, \mathcal{C}] \\
&\ge \tau \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[\mathcal{C}, G^v \ge \tau] - 2A_q^v \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t = \mathbf{e}_v, \mathcal{C}, -2A_q^v < G^v < 0] \\
&\ge \tau \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[\mathcal{C}, G^v \ge \tau] - \frac{\tau\,\mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[G^v \ge \tau]}{2} \\
&\ge \tau \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[\mathcal{C} | G^v \ge \tau] \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[G^v \ge \tau] - \frac{\tau\,\mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[G^v \ge \tau]}{2} \\
&\ge (1-\delta)\tau \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[G^v \ge \tau] - \frac{\tau\,\mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[G^v \ge \tau]}{2} \\
&\ge \left(\frac{1}{2} - \frac{\tau\,\mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[G^v \ge \tau]}{2\tau\,\mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[G^v \ge \tau] + 2}\right) \tau \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[G^v \ge \tau] \\
&= \frac{\tau\,\mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[G^v \ge \tau]}{2\tau\,\mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[G^v \ge \tau] + 2}.
\end{aligned}$$

Putting everything together, we get that

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathcal{P}^{(u_t)},\pi_t}[G^v|z_t=\mathbf{e}_v]\mathop{\mathbb{P}}_{\mathcal{P}^{(u_t)},\pi_t}[z_t=\mathbf{e}_v] \\
&= \mathop{\mathbb{E}}_{\mathcal{P}^{(u_t)},\pi_t}[G^v|z_t=\mathbf{e}_v,\mathcal{C}]\mathop{\mathbb{P}}_{\mathcal{P}^{(u_t)},\pi_t}[z_t=\mathbf{e}_v,\mathcal{C}] \\
&\quad+ \mathop{\mathbb{E}}_{\mathcal{P}^{(u_t)},\pi_t}[G^v|z_t=\mathbf{e}_v,\neg\mathcal{C}]\mathop{\mathbb{P}}_{\mathcal{P}^{(u_t)},\pi_t}[z_t=\mathbf{e}_v,\neg\mathcal{C}] \\
&\geq \frac{\tau\,\mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[G^v\geq\tau]}{2\tau\,\mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[G^v\geq\tau]+2} - \frac{\tau\,\mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[G^v\geq\tau]}{2\tau\,\mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[G^v\geq\tau]+2} \\
&= 0.
\end{aligned}$$

Thus, as long as $A^v(S_0,\delta)\leq\tau\,\mathbb{P}_{\pi,\mathcal{P}^{(u)}}[G^v\geq\tau]/4$, any agent of type u will comply with a recommendation of arm v from recommendation policy π according to Algorithm 4. □

Finally, we present the (expected) regret from the k treatment extension of policy π_c given in Definition 4.6.3.

Lemma 4.6.6 (Regret of Policy π_c for k Treatments)**.** *An extension of policy π_c achieves (expected) regret as follows:*

$$\mathbb{E}[R(T)] = O\left(k\sqrt{kT\log(kT)}\right) \tag{4.16}$$

for sufficiently large time horizon T.

Proof. Let θ^* be the best treatment effect overall and the gap between θ^* and any treatment effect θ^i be $\Delta_i=|\theta^*-\theta_i|$. Recall that the clean event $\mathcal{C}$ entails that the approximation bound holds for all rounds. If event $\mathcal{C}$ fails, then we can only bound the pseudo-regret by the maximum value, which is at most $T\min_i\Delta_i$.

For the rest of this proof, assume that the event $\mathcal{C}$ holds for every round of Algorithm 4. This proof follows the standard technique from Even-Dar et al. (2006). Since $\mathcal{C}$ holds, we have $\Delta_i\leq A(S_{L_2},\delta)+|\hat{\theta}^*-\hat{\theta}^i|$ for any treatment i, where $A(S_{L_2},\delta)$ is the approximation bound based on S_{L_2} samples of Algorithm 4. Before the stopping criteria is invoked, we also have $|\hat{\theta}^*-\hat{\theta}^i|\leq A(S_{L_2,\delta})$. Hence, the gap between the best treatment effect and any other treatment effect is:

$$\Delta_i \leq 2A(S_{L_2}, \delta) \leq \frac{2\sigma_g\sqrt{2k\log(2kT/\delta)}}{\sqrt{\frac{L_2}{k^2} - \log(k/\delta)}},$$

where σ_g is the variance parameter for the baseline reward $g^{(u)}$ and α_2 is defined as in Claim B.VI.3 relative to the proportion $r_2 = 1/|B|$ of recommendations for each treatment during Algorithm 4 and the proportion of compliant agents p_{c_2}. Hence, we must have eliminated treatment i by round

$$L_2 = \frac{8k\sigma_g^2\log(2kT/\delta)}{\Delta_i^2\left(\frac{1}{k^2} - \log(k/\delta)\right)},$$

assuming that $L_2 \geq \frac{p_c^2}{k\log(k/\delta)}$ (in order to satisfy the criterion for Theorem B.VI.2). During Algorithm 4, the social planner gives out $|B|$ recommendations for each treatment $i \in B$ sequentially. Hence, the contribution of each treatment i for each phase is Δ. Conditioned on event $\mathcal{C}$, the treatment a^* at the end of Algorithm 4 is the best treatment overall; so, no more regret is collected after Algorithm 4 is finished.

If treatment 1 is not the winner, then we accumulate $R_1(T) = \Delta_i\left((1-k\rho)L_1 + L_2/k\right)$ regret for treatment 1. If some other treatment $i > 1$ is not the winner, then we accumulate $R_i(T) = \Delta_i\left(\rho L_1 + L_2/k\right)$ regret for treatment i. Hence, the total regret accumulated in Algorithm 4 is:

$$R(T) \leq \Delta_i\left((1-\rho)L_1 + \left(\frac{k-1}{k}\right)L_2\right) \leq (1-\rho)L_1\Delta_i + \frac{8(k-1)\sigma_g^2\log(2kT/\delta)}{\Delta_i\left(\frac{1}{k^2} - \log(k/\delta)\right)}$$

Observe that the pseudo-regret of the combined recommendation policy is at most that of Algorithm 4 plus $\Delta = \min_i \Delta_i$ per each round of Algorithm 3. Alternatively, we can also upper bound the regret by Δ per each round of the combined recommendation policy. Following the same argument as Lemma 4.5.4, we can derive the pseudo-regret of the policy π_c for k treatments:

$$R(T) \leq \min\left(L_1(1-\rho)\Delta_i + \frac{8(k-1)\sigma_g^2 \log(2kT/\delta)}{\Delta_i\left(\frac{1}{k^2} - \log\left(k/\delta\right)\right)}, T\Delta\right)$$
$$\leq L_1 + O(k\sqrt{kT\log(kT/\delta)}).$$

For the expected regret, we can set the parameters δ and L_1 in terms of the time horizon T, in order to both guarantee compliance throughout policy π_c and to obtain sublinear expected regret bound relative to T.

First, we must guarantee that the failure probability δ in Algorithm 4 is small, i.e. $\delta = 1/T^2$. To meet our compliance condition for Algorithm 4, we must set

$$\delta \leq \frac{\tau\,\mathbb{P}_{\mathcal{P}^{(u)}}[\theta \geq \tau]}{2(\tau\,\mathbb{P}_{\mathcal{P}^{(u)}}[\theta \geq \tau] + 1)}$$

for some constant $\tau \in (0,1)$. Hence, we can set T sufficiently large such that, for any $\delta in (0,1)$, we have

$$T \geq \frac{1}{\sqrt{\delta}} \geq \sqrt{\frac{2(\tau\,\mathbb{P}_{\mathcal{P}^{(u)}}[\theta \geq \tau] + 1)}{\tau\,\mathbb{P}_{\mathcal{P}^{(u)}}[\theta \geq \tau]}}$$

We also recall that the length L_1 of Algorithm 3 needs to be sufficiently large so that p_{c_2} fraction of agents comply in Algorithm 4. Moreover, we accumulate linear regret in each round of Algorithm 3. Hence, in order to guarantee sublinear regret, we also require that T satisfies the following:

$$T \geq L_1^2 = (\ell + \ell/\rho)^2$$

Finally, recall that the clean event $\mathcal{C}$ in Algorithm 4 holds with probability at least $1-\delta$ for any $\delta \in (0,1)$. Conditioned on the failure event $\neg\mathcal{C}$, policy π_c accumulates at most linear pseudo-regret in terms of T. Thus, in expectation, it accumulates at most $T\max_{i,j}|\theta^i - \theta^j|\delta$ regret

Therefore, we can derive the expected regret of k treatment recommendation policy

π_c as:

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathcal{P}^{(u)}}[R(T)] &= \mathbb{E}[R(T)|\neg\mathcal{C}] \mathop{\mathbb{P}}_{\mathcal{P}^{(u)}}[\neg\mathcal{C}] + \mathbb{E}[R(T)|\mathcal{C}] \mathop{\mathbb{P}}_{\mathcal{P}^{(u)}}[\mathcal{C}] \\
&\leq T\delta + (L_1 + O(\sqrt{kT\log(kT/\delta)})) \\
&= \frac{1}{T} + (\sqrt{T} + O(k\sqrt{kT\log(kT)})) \\
&= \frac{1}{T} + O(k\sqrt{kT\log(kT)}) \\
&= O(k\sqrt{kT\log(kT)})
\end{aligned}$$

Therefore, assuming that all hyperparameters δ, L_1 are set to incentivize compliance for some nonzero proportion of agents throughout π_c and assuming that T is sufficiently large so as to satisfy the conditions above, policy π_c (for k treatments) achieves sublinear regret. □

B.VII Additional Experiment Details from Section 3.12

In this section, we present additional experiments to evaluate Algorithm 1 and Algorithm 2, which were previously omitted from Section 4.7. We are interested in (1) the effect of different prior choices on the exploration probability ρ, (2) comparing the approximation bound in Algorithm 1 to that of Algorithm 2 and (3) the total regret accumulated by the combined recommendation policy. Firstly, we observed that the exploration probability ρ in Figure 4.2 is small, leading to slow improvement in accuracy of Algorithm 1. Since ρ depends on the event ξ (as defined in Equation (4.5)), we want to investigate whether changes in the agents' priors would increase the exploration probability. Secondly, we claimed earlier in the paper that the estimation accuracy increases much quicker in Algorithm 2 compared to Algorithm 1. This improvement motivates the social planner to move to Algorithm 2, granted there is a large enough portion of agents that comply with the recommendations. Finally, while we provide a regret guarantee in Lemma 4.5.4, it is not immediately clear how the magnitude of Algorithm 1 length L_1 would affect the overall regret. There is a tradeoff: if we run Algorithm 1 for a small number of rounds, then it would not affect the regret by a significant amount, but a portion of the agents in Algorithm 2 may not comply. For our combined recommendation policy, we run Algorithm 1 until it is guaranteed that type 0 agents will comply in Algorithm 2.

Experimental Description For Algorithm 1, we consider a setting with two types of agents: type 0 who are initially never-takers, and type 1 who are initially always takers. For Algorithm 2, we consider a setting with two types of agents: type 0 who are compliant, and type 1 who are initially always-takers. We let each agent's prior on the treatment be a truncated Gaussian distribution between -1 and 1. The noisy baseline reward $g_t^{(u_t)}$ for each type u of agents is drawn from a Gaussian distribution $\mathcal{N}(\mu_{g^{(u)}}, 1)$, with its mean $\mu_{g^{(u)}}$ also drawn from a Gaussian prior. We let each type of agents have equal proportion in the population, i.e. $p_0 = p_1 = 0.5$.

For the first experiment, we are interested in finding the correlation between the exploration probability ρ and different prior parameters, namely the difference between mean baseline rewards $\mu_{g^{(1)}} - \mu_{g^{(0)}}$ and the variance of Gaussian prior on the treatment effect θ. Similar to the experiment in Section 4.7, we use Monte Carlo simulation by running the first stage of Algorithm 1 with varying choices of the two prior parameters above. From these initial samples, we calculate the probability of event ξ, and subsequently the exploration probability ρ. For the second experiment, we are interested in finding when agents of type 1 also comply with the recommendations. This shift in compliance depends on a constant τ (as defined in Lemma 4.4.2). We find two values of the constant τ that minimizes the number of samples needed to guarantee that agents of type 0 and type 1 are compliant in Algorithm 2 (as defined in Lemma 4.5.2). After this, Algorithm 2 is run for increasing number of rounds. Similar to the Algorithm 1 experiment, we repeated calculate the IV estimate of the treatment effect and compare it to the naive OLS estimate over the same samples as a benchmark. On a separate attempt, we evaluate the combined recommendation policy by running Algorithm 1 and Algorithm 2 successively using the priors above. We calculate the accumulated regret of this combined policy using the pseudo-regret notion (as defined in Definition 4.5.3).

Results In Table B.1 and Table B.2, we calculate the exploration probability ρ with different initialization of the agents' priors. In Table B.1, we let the mean baseline reward of type 1 $\mu_{g^{(1)}}$ be drawn from $\mathcal{N}(0.5, 1)$ and the mean baseline reward of type 0 $\mu_{g^{(0)}}$ be drawn from $\mathcal{N}(c, 1)$ with $c \in [0, 1]$. The gap between these priors is defined

Expected gap $\mathbb{E}[\mu_{g^{(1)}} - \mu_{g^{(0)}}]$	Exploration probability ρ
-0.5	0.004480
-0.4	0.004975
-0.3	0.007936
-0.2	0.003984
-0.1	0.003488
0.1	0.003984
0.2	0.004480
0.3	0.003488
0.4	0.004480
0.5	0.003488

Table B.1: Upper bounds on exploration probability ρ to incentivize partial compliance with respect to different gaps $\mathbb{E}[\mu_{g^{(1)}} - \mu_{g^{(0)}}]$

Variance in prior over treatment effect	Exploration probability ρ
0.1	0.002561
0.2	0.003112
0.3	0.002561
0.4	0.002561
0.5	0.003982
0.6	0.004643
0.7	0.005422
0.8	0.002790
0.9	0.001389
1	0.003488

Table B.2: Upper bounds on exploration probability ρ to incentivize partial compliance with respect to different variances in the prior over treatment effect θ

as $\mathbb{E}[\mu_{g^{(1)}}] - \mathbb{E}[\mu_{g^{(0)}}]$. We observe that the exploration probability does not change monotonically with increasing gap between mean baseline reward. In Table B.2, we calculate the exploration probability ρ with different variance in prior over treatment effect θ. Similarly, in Table B.1, we observe that the exploration probability ρ does not change monotonically with increasing variance in prior over θ. In both tables, ρ value lies between $[0.001, 0.008]$, which implies infrequent exploration by Algorithm 1. This slow rate of exploration is also reflected in Figure 4.2, which motivates the social planner to transition to Algorithm 2.

In Figure B.1, we compare the approximation bound on $|\theta - \hat{\theta}|$ between the IV estimate $\hat{\theta}$ and the naive estimate for Algorithm 2. In our experiments, the constant τ generally lies within $[0.4, 0.6]$. Similar to the experiment in Section 4.7, we let the hidden treatment effect $\theta = 0.5$, type 0 and type 1 agents' priors on the treatment effect be $\mathcal{N}(-0.5, 1)$ and $\mathcal{N}(0.9, 1)$ — each truncated into $[-1, 1]$ — respectively. We also let the mean baseline reward for type 0 and type 1 agents be $\mu_{g^{(0)}} \sim \mathcal{N}(0, 1)$ and $\mu_{g^{(1)}} \sim \mathcal{N}(0.1, 1)$, respectively. With these priors, we have found a suitable value of $\tau = 0.43$ for Algorithm 2. Instead of using the theoretical bound on ℓ in Lemma 4.5.2, we compare the approximation bound $|\theta - \hat{\theta}|$ with the conditions in Lemma 4.4.2. In Figure B.1, the IV estimate consistently outperform the naive estimate for any number

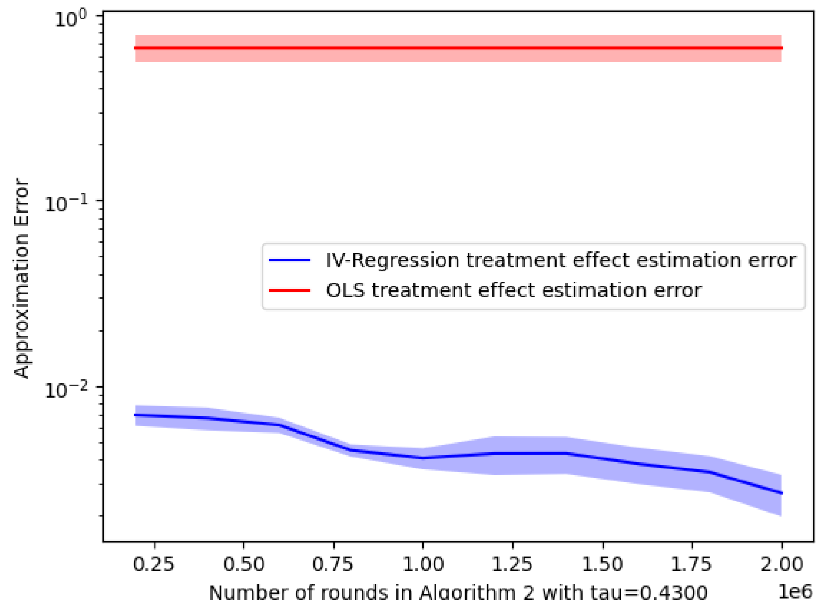


Figure B.1: Approximation bound using IV regression and OLS during Algorithm 2 with $\tau = 0.43$. The y-axis uses a log scale. Results are averaged over 5 runs; error bars represent one standard error.

of rounds. Furthermore, we observe that the scale of the IV estimate approximation bound in Figure B.1 is much smaller than that of Figure 4.2. This difference shows the improvement of Algorithm 2 over Algorithm 1 on estimating the treatment effect θ. It takes Algorithm 2 a small number of rounds to get a better estimate than Algorithm 1 due to the small exploration probability ρ.

Appendix C

Appendix for Incentive-Aware Synthetic Control

C.I Further Background on Principal Component Regression

Let $\mathcal{I}_i^{(d)}$ be the set of units who have received intervention d before unit i arrives, and let $n_i^{(d)}$ be the number of such units. We use

$$Y_{pre,i}^{(d)} := [\mathbf{y}_{j,pre}^\top \ : \ j \in \mathcal{I}_i^{(d)}] \in \mathbb{R}^{n_i^{(d)} \times T_0}$$

to denote the matrix of pre-treatment outcomes corresponding to the subset of units who have undergone intervention d before unit i arrives, and

$$Y_{post,i}^{(d)} := \left[\sum_{t=T_0+1}^{T} y_{j,t}^{(d)} : j \in \mathcal{I}_i^{(d)} \right] \in \mathbb{R}^{n_i^{(d)} \times 1}$$

be the vector of the sum of post-intervention outcomes for the subset of units who have undergone intervention d before unit i arrives. We denote the singular value decomposition of $Y_{pre,i}^{(d)}$ as

$$Y_{pre,i}^{(d)} = \sum_{\ell=1}^{n_i^{(d)} \wedge T_0} s_\ell^{(d)} \widehat{\mathbf{u}}_\ell^{(d)} (\widehat{\mathbf{v}}_\ell^{(d)})^\top,$$

where $\{s_\ell^{(d)}\}_{\ell=1}^{n_i^{(d)} \wedge T_0}$ are the singular values of $Y_{pre,i}^{(d)}$, and $\widehat{\mathbf{u}}_\ell^{(d)}$ and $\widehat{\mathbf{v}}_\ell^{(d)}$ are orthonormal column vectors. We assume that the singular values are ordered such that $s_1(Y_{pre,i}^{(d)}) \geq \cdots \geq s_{n_i^{(d)} \wedge T_0}(Y_{pre,i}^{(d)}) \geq 0$. For some threshold value r, we use

$$\widehat{Y}_{pre,i}^{(d)} := \sum_{\ell=1}^{r} s_\ell^{(d)} \widehat{\mathbf{u}}_\ell^{(d)} (\widehat{\mathbf{v}}_\ell^{(d)})^\top$$

to refer to the truncation of $Y_{pre,i}^{(d)}$ to its top r singular values. We define the projection matrix onto the subspace spanned by the top r right singular vectors as $\widehat{\mathbf{P}}_{i,r}^{(d)} \in \mathbb{R}^{r \times r}$, given by $\widehat{\mathbf{P}}_{i,r}^{(d)} := \sum_{\ell=1}^{r} \widehat{\mathbf{v}}_\ell^{(d)} \left(\widehat{\mathbf{v}}_\ell^{(d)}\right)^\top$. Equipped with this notation, we are now ready to define the procedure for estimating $\boldsymbol{\theta}^{(d)}$ using (regularized) principal component regression.

Definition C.I.1 (Regularized Principal Component Regression)**.** *Given regularization parameter $\rho \geq 0$ and truncation level $r \in \mathbb{N}$, for $d \in \{0, 1\}$ and $i \geq 1$, let $\mathcal{V}_i^{(d)} := \left(\widehat{Y}_{pre,i}^{(d)}\right)^\top \widehat{Y}_{pre,i}^{(d)} + \rho \widehat{\mathbf{P}}_{i,r}^{(d)}$. Then, regularized PCR estimates $\boldsymbol{\theta}^{(d)}$ as*

$$\widehat{\boldsymbol{\theta}}_i^{(d)} := \left(\mathcal{V}_i^{(d)}\right)^{-1} \widehat{Y}_{pre,i}^{(d)} Y_{i,post}^{(d)},$$

where $\boldsymbol{\theta}^{(d)}$ is defined as in Proposition 5.3.3. The average post-intervention outcome for unit i under intervention d may then be estimated as $\widehat{\mathbb{E}}[\bar{y}_{i,post}^{(d)}] := \frac{1}{T_1} \langle \widehat{\boldsymbol{\theta}}_i^{(d)}, y_{i,pre} \rangle$.

Observe that if $\rho = 0$, then $\mathcal{V}_i^{(d)} := \left(\widehat{Y}_{pre,i}^{(d)}\right)^\top \widehat{Y}_{pre,i}^{(d)}$ and we recover the non-regularized version of PCR. Under the Unit Overlap Assumption and Assumption 5.3.2, work on robust synthetic control uses (regularized) PCR to obtain consistent estimates for counterfactual post-intervention outcomes under different interventions. Intuitively, the Unit Overlap Assumption is required for consistent estimation because there needs to be sufficient information about the test unit contained in the data from donor units in order to accurately make predictions about the test unit using a model learned from the donor units' data.

C.II Appendix for Section 5.3: Preliminaries

Theorem 5.3.1. *Fix a time horizon T, pre-intervention time period T_0, and number of donor units $n^{(0)}$. For any algorithm used to estimate the average post-intervention outcome under control for a test unit, there exists a problem instance such that the produced estimate has constant error whenever the Unit Overlap Assumption is not satisfied for the test unit under control, even as T, T_0, $n^{(0)} \to \infty$.*

Proof. Consider the setting in which all type 0 units have latent factor $\mathbf{v}_0 = [0\ 1]^\top$ and all type 1 units have latent factor $\mathbf{v}_1 = [1\ 0]^\top$. Suppose that $\mathbf{u}_t^{(0)} = [1\ 0]^\top$ if $t \mod 2 = 0$ and $\mathbf{u}_t^{(0)} = [0\ 1]^\top$ if $t \mod 2 = 1$ if $t \leq T_0$. For all $t > T_0$, let $\mathbf{u}_t^{(0)} = [H\ 1]^\top$ for some H in range $[-c, c]$, where $c > 0$. Furthermore, suppose that there is no noise in the unit outcomes. Under this setting, the (expected) outcomes for type 0 units under control are

$$\mathbb{E}[y_{0,t}^{(0)}] = \begin{cases} 0 \text{ if } t \mod 2 = 0, t \leq T_0 \\ 1 \text{ if } t \mod 2 = 1, t \leq T_0 \\ 1 \text{ if } t > T_0 \end{cases}$$

and the outcomes for type 1 units under control are

$$\mathbb{E}[y_{1,t}^{(0)}] = \begin{cases} 1 \text{ if } t \mod 2 = 0, t \leq T_0 \\ 0 \text{ if } t \mod 2 = 1, t \leq T_0 \\ H \text{ if } t > T_0. \end{cases}$$

Suppose that $H \sim \mathsf{Unif}[-c, c]$ and consider a principal who wants to estimate $\mathbb{E}[\bar{y}_{1,post}^{(0)}]$ using just the set of outcomes $\mathbb{E}[\mathbf{y}_{0,pre}]$, $\mathbb{E}[\mathbf{y}_{0,post}^{(0)}]$, and $\mathbb{E}[\mathbf{y}_{1,pre}]$. Since these outcomes do not contain any information about H, any estimator $\widehat{\mathbb{E}}[\bar{y}_{1,post}^{(0)}]$ cannot be a function of H and thus will be at least a constant distance away from the true average post-intervention outcome $\mathbb{E}[\bar{y}^{(0)}_{1,post}]$ in expectation over H. That is,

$$\mathop{\mathbb{E}}_{H}|\mathbb{E}[\bar{y}^{(0)}_{1,post}] - \widehat{\mathbb{E}}[\bar{y}^{(0)}_{1,post}]| = \mathop{\mathbb{E}}_{H}|H - \widehat{\mathbb{E}}[\bar{y}^{(0)}_{1,post}]|$$
$$= \frac{1}{2c}(c^2 + (\widehat{\mathbb{E}}[\bar{y}^{(0)}_{1,post}])^2) \geq \frac{c}{2}.$$

Note that our choice of T and T_0 was arbitrary, and that estimation does not improve as the number of donor units increases since (1) there is no noise in the outcomes and (2) all type 0 units have the same latent factor. Therefore any estimator for $\mathbb{E}[\bar{y}^{(0)}_{1,post}]$ is inconsistent in expectation over H as $n^{(0)}, T, T_0 \to \infty$. This implies our desired result because if estimation error is constant in expectation over problem instances, there must exist at least one problem instance with constant estimation error. □

C.III Appendix for Section 7.2: Incentivized Exploration for Synthetic Control

C.III.1 Causal Parameter Recovery

Theorem C.III.1 (Theorem G.3 of Agarwal et al. (2023))**.** *Let $\delta \in (0,1)$ be an arbitrary confidence parameter and $\rho > 0$ be chosen to be sufficiently small. Further, assume that Assumption 5.2.1 and Unit Overlap Assumption are satisfied, there is some $i_0 \geq 1$ such that $\text{rank}(\mathbf{X}_{i_0}](d)) = r$, and $\text{snr}_i(d) \geq 2$ for all $i \geq i_0$. Then, with probability at least $1 - \mathcal{O}(k\delta)$, simultaneously for all interventions $d \in [k]_0$,*

$$|\widehat{\mathbb{E}}[\bar{y}^{(d)}_{i,post}] - \bar{\mathbb{E}}[y^{(d)}_{i,post}]| \leq \alpha(\delta),$$
$$\textit{where } \alpha(\delta) := \frac{3\sqrt{T_0}}{\widehat{\text{snr}}_i(d)}\left(\frac{L(\sqrt{74} + 12\sqrt{6}\kappa(\mathbf{Z}_i(d)))}{(T-T_0)\cdot\widehat{\text{snr}}_i(d)} + \frac{\sqrt{2\text{err}_i(d)}}{\sqrt{T-T_0}\cdot\sigma_r(\mathbf{Z}_i(d))}\right)$$
$$+ \frac{2L\sqrt{24T_0}}{(T-T_0)\cdot\widehat{\text{snr}}_i(d)} + \frac{12L\kappa(\mathbf{Z}_i(d))\sqrt{3T_0}}{(T-T_0)\cdot\widehat{\text{snr}}_i(d)} + \frac{2\sqrt{\text{err}_i(d)}}{\sqrt{T-T_0}\cdot\sigma_r(\mathbf{Z}_i(d))}$$
$$+ \frac{L\sigma\sqrt{2\log(k/\delta)}}{T-T_0} + \frac{L\sigma\sqrt{148\log(k/\delta)}}{\widehat{\text{snr}}_i(d)(T-T_0)}$$
$$+ \frac{24\sigma\kappa(\mathbf{Z}_i(d))\sqrt{6\log(k/\delta)}}{\widehat{\text{snr}}_i(d)(T-T_0)} + \frac{2\sigma\sqrt{\text{err}_i(d)\log(k/\delta)}}{\sigma_r(\mathbf{Z}_i(d))\sqrt{T-T_0}}$$

where $\widehat{\mathbb{E}}[\bar{y}_{i,post}^{(d)}] \coloneqq \frac{1}{T-T_0} \cdot \langle \widehat{\theta}_i(d), y_{i,pre} \rangle$ *is the estimated average post-intervention outcome for unit* i *under intervention* d *and* $\left\| \theta_i^{(d)} \right\| \leq L$*.*

Corollary C.III.2. *Given a gap* ε *and the same assumptions as in Theorem C.III.1, the probability that* $|\widehat{\mathbb{E}}[\bar{y}_{i,post}^{(d)}] - \mathbb{E}[\bar{y}_{i,post}^{(d)}]| \leq \varepsilon$ *is at least* $1-\delta$*, where*

$$\delta \leq \frac{\log(n^{(d)}) \vee k}{\exp\left(\left(\sqrt{\frac{\sigma_{r_d}(\mathbf{Z}_i(d))\varepsilon - (A+F)(\sqrt{n^{(d)}}+\sqrt{T_0})}{D} + \frac{\alpha^2}{4D^2}} - \frac{\alpha}{2D}\right)^2\right)}$$

with

- $A = 3\sqrt{T_0}\left(\frac{\left\|\theta_i^{(d)}\right\|(\sqrt{74}+12\sqrt{6}\kappa(\mathbf{Z}_i(d)))}{T-T_0} + \frac{\sqrt{2}}{\sqrt{T-T_0}}\right),$
- $F = \frac{2\left\|\theta_i^{(d)}\right\|\sqrt{24T_0}}{T-T_0} + \frac{12\left\|\theta_i^{(d)}\right\|\kappa(\mathbf{Z}_i(d))\sqrt{3T_0}}{T-T_0} + \frac{2}{\sqrt{T-T_0}},$
- $D = \frac{\left\|\theta_i^{(d)}\right\|\sigma\sqrt{148}}{T-T_0} + \frac{24\sigma\kappa(\mathbf{Z}_i(d))\sqrt{6}}{T-T_0} + \frac{2\sigma}{\sqrt{T-T_0}},$
- $E = \frac{\sqrt{2}\left\|\theta_i^{(d)}\right\|\sigma}{T-T_0},$
- $\alpha = A + F + D(\sqrt{n^{(d)}} + \sqrt{T_0}) + \sigma_{r_d}(\mathbf{Z}_i(d))E$

Proof. We begin by setting the right-hand side of Theorem C.III.1 to be ε. The goal is to write the failure probability δ as a function of ε. Then, using the notations above, we can write

$$\varepsilon = \frac{A}{(\widehat{\mathrm{snr}}_i(d))^2} + \frac{F}{\widehat{\mathrm{snr}}_i(d)} + \frac{D\sqrt{\log(k/\delta)}}{\widehat{\mathrm{snr}}_i(d)} + E\sqrt{\log(k/\delta)}$$

First, we take a look at the signal-to-noise ratio $\widehat{\mathrm{snr}}_i(d)$. By definition, we have:

$$\begin{aligned}\widehat{\mathrm{snr}}_i(d) &= \frac{\sigma_r(\mathbf{Z}_i(d))}{U_i} \\ &= \frac{\sigma_r(\mathbf{Z}_i(d))}{\sqrt{n^{(d)}} + \sqrt{T_0} + \sqrt{\log(\log(n^{(d)})/\delta)}}\end{aligned}$$

Hence,

$$\frac{1}{\widehat{\text{snr}}_i(d)} = \frac{\sqrt{n^{(d)}} + \sqrt{T_0} + \sqrt{\log(\log(n^{(d)})/\delta)}}{\sigma_r(\mathbf{Z}_i(d))}$$

Observe that since $\widehat{\text{snr}}_i(d) \geq 2$, we have $\frac{1}{(\widehat{\text{snr}}_i(d))^2} \leq \frac{1}{\widehat{\text{snr}}_i(d)}$. Hence, we can write an upper bound on ε as:

$$\begin{aligned}
\varepsilon &\leq \frac{A+F}{\widehat{\text{snr}}_i(d)} + \frac{D\sqrt{\log(k/\delta)}}{\widehat{\text{snr}}_i(d)} + E\sqrt{\log(k/\delta)} \\
&\leq \frac{(A+F)(\sqrt{n^{(d)}} + \sqrt{T_0})}{\sigma_{r_d}(\mathbf{Z}_i(d))} + \frac{(A+F)\sqrt{\log(\log(n^{(d)})/\delta)}}{\sigma_{r_d}(\mathbf{Z}_i(d))} \\
&\quad + \frac{D\left(\sqrt{n} + \sqrt{d} + \sqrt{\log(\log(n^{(d)})/\delta)}\right)\sqrt{\log(k/\delta)}}{\sigma_r(\mathbf{Z}_i(d))} + E\sqrt{\log(k/\delta)}
\end{aligned}$$

Since $\log(x)$ is a strictly increasing function for $x > 0$, we can simplify the above expression as:

$$\begin{aligned}
\varepsilon \leq{}& \frac{(A+F)(\sqrt{n^{(d)}} + \sqrt{T_0})}{\sigma_{r_d}(\mathbf{Z}_i(d))} + \frac{(A+F)\sqrt{\log(\log(n^{(d)}) \vee k/\delta)}}{\sigma_{r_d}(\mathbf{Z}_i(d))} \\
&+ \frac{D(\sqrt{n} + \sqrt{d})\sqrt{\log(\log(n^{(d)}) \vee k/\delta)}}{\sigma_{r_d}(\mathbf{Z}_i(d))} + \frac{D\log(\log(n^{(d)}) \vee k/\delta)}{\sigma_{r_d}(\mathbf{Z}_i(d))} + E\sqrt{\log(\log(n^{(d)}) \vee k/\delta)}
\end{aligned}$$

Subtracting the first term from both sides and multiplying by $\sigma_r(\mathbf{Z}_i(d))$, we have:

$$\begin{aligned}
&\sigma_{r_d}(\mathbf{Z}_i(d))\varepsilon - (A+F)(\sqrt{n^{(d)}} + \sqrt{T_0}) \\
&\leq (A+F)\sqrt{\log(\log(n^{(d)}) \vee k/\delta)} + D(\sqrt{n} + \sqrt{d})\sqrt{\log(\log(n^{(d)}) \vee k/\delta)} \\
&+ D\log(\log(n^{(d)}) \vee k/\delta) + E\sigma_{r_d}(\mathbf{Z}_i(d))\sqrt{\log(\log(n^{(d)}) \vee k/\delta)} \\
&= (A + F + D(\sqrt{n} + \sqrt{d}) + E\sigma_{r_d}(\mathbf{Z}_i(d)))\sqrt{\log(\log(n^{(d)}) \vee k/\delta)} + D\log(\log(n^{(d)}) \vee k/\delta)
\end{aligned}$$

Let $\alpha = A + F + D(\sqrt{n} + \sqrt{d}) + E\sigma_{r_d}(\mathbf{Z}_i(d))$, we can rewrite the inequality above as:

$$\sigma_r(\mathbf{Z}_i(d))\varepsilon - (A+F)(\sqrt{n^{(d)}} + \sqrt{T_0}) \leq \alpha\sqrt{\log(\log(n^{(d)}) \vee k/\delta)} + D\log(\log(n^{(d)}) \vee k/\delta)$$

Then, we can complete the square and obtain:

$$
\begin{aligned}
&\log\left({}^{\log(n^{(d)})\vee k}/_{\delta}\right) + \frac{\alpha}{D}\sqrt{\log\left({}^{\log(n^{(d)})\vee k}/_{\delta}\right)} + \frac{\alpha^2}{4D^2} \\
&\quad \geq \frac{\sigma_{r_d}(\mathbf{Z}_i(d))\varepsilon - (A+F)(\sqrt{n^{(d)}} + \sqrt{T_0})}{D} + \frac{\alpha^2}{4D^2} \\
&\iff \left(\sqrt{\log\left({}^{\log(n^{(d)})\vee k}/_{\delta}\right)} + \frac{\alpha}{2D}\right)^2 \geq \frac{\sigma_{r_d}(\mathbf{Z}_i(d))\varepsilon - (A+F)(\sqrt{n^{(d)}} + \sqrt{T_0})}{D} + \frac{\alpha^2}{4D^2} \\
&\iff \sqrt{\log\left({}^{\log(n^{(d)})\vee k}/_{\delta}\right)} + \frac{\alpha}{2D} \geq \sqrt{\frac{\sigma_{r_d}(\mathbf{Z}_i(d))\varepsilon - (A+F)(\sqrt{n^{(d)}} + \sqrt{T_0})}{D} + \frac{\alpha^2}{4D^2}} \\
&\iff \sqrt{\log\left({}^{\log(n^{(d)})\vee k}/_{\delta}\right)} \geq \sqrt{\frac{\sigma_{r_d}(\mathbf{Z}_i(d))\varepsilon - (A+F)(\sqrt{n^{(d)}} + \sqrt{T_0})}{D} + \frac{\alpha^2}{4C^2}} - \frac{\alpha}{2D} \\
&\iff \log\left({}^{\log(n^{(d)})\vee k}/_{\delta}\right) \geq \left(\sqrt{\frac{\sigma_{r_d}(\mathbf{Z}_i(d))\varepsilon - (A+F)(\sqrt{n^{(d)}} + \sqrt{T_0})}{D} + \frac{\alpha^2}{4D^2}} - \frac{\alpha}{2D}\right)^2 \\
&\iff \frac{\log(n^{(d)}) \vee k}{\delta} \\
&\quad \geq \exp\left(\left(\sqrt{\frac{\sigma_{r_d}(\mathbf{Z}_i(d))\varepsilon - (A+F)(\sqrt{n^{(d)}} + \sqrt{T_0})}{D} + \frac{\alpha^2}{4D^2}} - \frac{\alpha}{2D}\right)^2\right) \\
&\iff \delta \leq \frac{\log(n^{(d)}) \vee k}{\exp\left(\left(\sqrt{\frac{\sigma_{r_d}(\mathbf{Z}_i(d))\varepsilon - (A+F)(\sqrt{n^{(d)}} + \sqrt{T_0})}{D} + \frac{\alpha^2}{4D^2}} - \frac{\alpha}{2D}\right)^2\right)}
\end{aligned}
$$

□

C.III.2 Proof of Theorem 4.4.2

Validity of Assumption 5.4.1.3 First, we note that in the first stage of Algorithm 5, the principal does not provide any recommendation to the units and instead lets them pick their preferred intervention. The goal of this first stage is to ensure the linear span inclusion assumption (Unit Overlap Assumption) is satisfied for type 1 units and intervention 1. This condition is equivalent to having enough samples of type 1 units such that the set of latent vectors $\{v_i\}_{i\in\mathcal{I}^{(1)}}$ spans the latent vector space S_1. We invoke the following theorem from Vershynin (2018) that shows $\text{span}(\{v_i\}_{i\in\mathcal{I}^{(1)}}) = S_1$ with high probability:

Theorem C.III.3 (Theorem 4.6.1 of Vershynin (2018)). *Let A be an $m \times n$ matrix whose rows A_i are independent, mean zero, sub-gaussian isotropic random vectors in $\mathbb{R}^n$. Then for any $t \geq 0$ we have with probability at least $1 - 2\exp(-t^2)$:*

$$\sqrt{m} - c_{Ver}K^2(\sqrt{n} + t) \leq s_n(A) \leq s_1(A) \leq \sqrt{m} + c_{Ver}K^2(\sqrt{n} + t) \tag{C.1}$$

where $K = \max_i \|A_i\|_{\psi_2}$ and c_{Ver} is an absolute constant.

Hence, after observing $N_0^{(1)}$ samples of type 1 units taking intervention 1, the linear span inclusion assumption is satisfied with probability at least

$$1 - 2\exp\left(-\left(\frac{\sqrt{N_0^{(1)}}}{c_{\text{Ver}}K^2} - \sqrt{r_1}\right)^2\right)$$

Theorem C.III.4. *Suppose the there are two interventions, and assume that Assumption 5.4.1 holds for some constant gap $C \in (0, 1)$. If N_0 is chosen to be large enough such that Unit Overlap Assumption is satisfied for all units of type 1 under treatment with probability at least $1 - \delta_0$, then Algorithm 5 with parameters δ, B, C is BIC for all units of type 1 if*

$$L \geq 1 + \max_{i \in \mathcal{T}^{(1)}} \left\{ \frac{\mu_{v_i}^{(1)} - \mu_{v_i}^{(0)}}{\left(C - \alpha(\delta_{PCR}) - \sigma\sqrt{\frac{2\log(1/\delta_{\varepsilon_i})}{T_1}}\right)} \cdot \frac{1}{\Pr\left[\mu_{v_i}^{(0)} - \bar{y}_{i,post}^{(1)} \geq C - \alpha(\delta_{PCR}) - \sigma\sqrt{\frac{2\log(1/\delta_{\varepsilon_i})}{T_1}}\right] - 2\delta} \right\}$$

where $\alpha(\delta_{PCR})$ is the high-probability confidence bound defined in Theorem C.III.1, $\delta = \delta_{\varepsilon_i} + \delta_0 + \delta_{PCR}$ and $\delta_{\varepsilon_i} \in (0, 1)$ is the failure probability of the Chernoff-Hoeffding *bound on the average of sub-Gaussian random noise* $\{\varepsilon_{i,t}^{(1)}\}_{t=T_0+1}^{T}$, *and*

$$\delta_{PCR} \leq \frac{\log(N_0^{(1)}) \vee k}{\exp\left(\left(\sqrt{\frac{\sigma_r(Y_{pre,N_0}^{(1)})(C/2)-(A+F)(\sqrt{N_0^{(1)}}+\sqrt{T_0})}{D}+\frac{\alpha^2}{4D^2}}-\frac{\alpha}{2D}\right)^2\right)}$$

and $\kappa\left(Y_{pre,N_0}^{(1)}\right) = \sigma_1\left(Y_{pre,N_0}^{(1)}\right)/\sigma_{r_1}\left(Y_{pre,N_0}^{(1)}\right)$ *is the condition number of the matrix of observed pre-intervention outcomes for the subset of the first* N_0 *units who have taken the treatment. The remaining variables in* δ_{PCR} *are defined as*

$$\begin{aligned}
A &= 3\sqrt{T_0}\left(\frac{\Gamma(\sqrt{74}+12\sqrt{6}\kappa(Y_{pre,N_0}^{(1)})}{T-T_0}+\frac{\sqrt{2}}{\sqrt{T-T_0}}\right),\\
F &= \frac{2\Gamma\sqrt{24T_0}}{T-T_0}+\frac{12\Gamma\kappa(Y_{pre,N_0}^{(1)})\sqrt{3T_0}}{T-T_0}+\frac{2}{\sqrt{T-T_0}},\\
D &= \frac{\Gamma\sigma\sqrt{148}}{T-T_0}+\frac{24\sigma\kappa(Y_{pre,N_0}^{(1)})\sqrt{6}}{T-T_0}+\frac{2\sigma}{\sqrt{T-T_0}},\\
E &= \frac{\sqrt{2}\Gamma\sigma}{T-T_0}\\
\alpha &= A+F+D(\sqrt{N_0^{(1)}}+\sqrt{T_0})+\sigma_{r_1}(Y_{pre,N_0}^{(1)})E.
\end{aligned}$$

Moreover if the number of batches B *is chosen to be large enough such that with probability at least* $1-\delta$, $\mathrm{rank}([\mathbb{E}[\mathbf{y}_{i,pre}]^\top : \widehat{d}_i = 0, i \in \mathcal{T}^{(1)}]_{i=1}^{N_0+B\cdot L}) = \mathrm{rank}([\mathbb{E}[\mathbf{y}_{j,pre}]^\top]_{j\in\mathcal{T}^{(1)}})$, *then the Unit Overlap Assumption will be satisfied for all type* 1 *units under control with probability at least* $1-2\delta$.

Proof. At a particular time step t, unit i of type 1 can be convinced to pick control if $\mathbb{E}_{v_i}[\bar{y}_{i,post}^{(0)} - \bar{y}_{i,post}^{(1)}|\widehat{d}=0]\Pr[\widehat{d}=0] \geq 0$. There are two possible disjoint events under which unit i is recommended intervention 0: either intervention 0 is better empirically, i.e. $\mu_{v_i}^{(0)} \geq \widehat{\mathbb{E}}[\bar{y}_{i,post}^{(1)}] + C$, or intervention 1 is better and unit i is chosen for exploration.

Hence, we have

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{v_i}[\bar{y}^{(0)}_{i,post} - \bar{y}^{(1)}_{i,post} | \widehat{d} = 0] \Pr[\widehat{d} = 0] \\
&= \mathbb{E}[\bar{y}^{(0)}_{i,post} - \bar{y}^{(1)}_{i,post} | \mu^{(0)}_{v_i} \geq \widehat{\mathbb{E}}[\bar{y}^{(1)}_{i,post}] + C] \Pr[\mu^{(0)}_{v_i} \geq \widehat{\mathbb{E}}[\bar{y}^{(1)}_{i,post}] + C] \left(1 - \frac{1}{L}\right) \\
&\quad + \frac{1}{L} \mathop{\mathbb{E}}_{v_i}[\bar{y}^{(0)}_{i,post} - \bar{y}^{(1)}_{i,post}] \\
&= \left(\mu^{0}_{v_i} - \mathop{\mathbb{E}}_{v_i}[\bar{y}^{(1)}_{i,post} | \mu^{(0)}_{v_i} \geq \widehat{\mathbb{E}}[\bar{y}^{(1)}_{i,post}] + C]\right) \Pr[\mu^{(0)}_{v_i} \geq \widehat{\mathbb{E}}[\bar{y}^{(1)}_{i,post}] + C] \left(1 - \frac{1}{L}\right) \\
&\quad + \frac{1}{L}(\mu^{(0)}_{v_i} - \mu^{(1)}_{v_i})
\end{aligned}
$$

Rearranging the terms and taking the maximum over all units of type 1 gives the lower bound on phase length L:

$$L \geq 1 + \frac{\mu^{(1)}_{v_i} - \mu^{(0)}_{v_i}}{(\mu^{(0)}_{v_i} - \mathbb{E}_{v_i}[\widehat{\mathbb{E}}[\bar{y}^{(1)}_{i,post}] | \widehat{\xi}_{C,i}]) \Pr_{v_i}[\widehat{\xi}_{C,i}])}$$

To complete the analysis, we want to find a lower bound for the terms in the denominator. That is, we want to lower bound

$$\left(\mu^{(0)}_{v_i} - \mathop{\mathbb{E}}_{v_i}[\bar{y}^{(1)}_{i,post} | \mu^{(0)}_{v_i} \geq \widehat{\mathbb{E}}[\bar{y}^{(1)}_{i,post}] + C]\right) \Pr_{v_i}[\mu^{(0)}_{v_i} \geq \widehat{\mathbb{E}}[\bar{y}^{(1)}_{i,post}] + C]$$

Let ξ_0 denote the event that Unit Overlap Assumption is satisfied for type 1 units under treatment. From Assumption 5.4.1, we know that event ξ_0 occurs with probability at least $1 - \delta_0$ for some $\delta_0 \in (0, 1)$. Let ξ_{PCR} denote the event that the high-probability concentration bound holds for units of type 1 and intervention 1. Finally, let ξ_{ε_i} denote the event where the Chernoff-Hoeffding bound holds on the average of the sub-Gaussian noise $\varepsilon^{(1)}_{i,t}$, that is with probability at least $1 - \delta_{\varepsilon_i}$, we have:

$$|\frac{1}{T_1} \sum_{t=T_0+1}^{T} \varepsilon^{(1)}_{i,t}| \leq \sigma \sqrt{\frac{2\log(1/\delta_{\varepsilon_i})}{T_1}}$$

Define another clean event ξ where the events ξ_0, ξ_{ε_i} and ξ_{PCR} happen simultaneously with probability at least $1-\delta$, where $\delta = \delta_0 + \delta_{PCR} + \delta_{\varepsilon_i}$. Then, we have:

$$\begin{aligned}
&(\mu_{v_i}^{(0)} - \mathop{\mathbb{E}}_{v_i}[\bar{y}_{i,post}^{(1)} | \widehat{\xi}_{C,i}]) \Pr[\widehat{\xi}_{C,i}] \\
&= (\mu_{v_i}^{(0)} - \mathop{\mathbb{E}}_{v_i}[\bar{y}_{i,post}^{(1)} | \widehat{\xi}_{C,i}, \xi]) \Pr[\widehat{\xi}_{C,i}, \xi] + (\mu_{v_i}^{(0)} - \mathop{\mathbb{E}}_{v_i}[\bar{y}_{i,post}^{(1)} | \widehat{\xi}_{C,i}, \neg\xi]) \Pr[\widehat{\xi}_{C,i}, \neg\xi] \\
&\geq (\mu_{v_i}^{(0)} - \mathop{\mathbb{E}}_{v_i}[\bar{y}_{i,post}^{(1)} | \widehat{\xi}_{C,i}, \xi]) \Pr[\widehat{\xi}_{C,i}, \xi] - 2\delta \quad (\text{since } \Pr[\neg\xi] < \delta,\ \mu_{v_i}^{(0)} - \mathbb{E}_{v_i}[\bar{y}_{i,post}^{(1)}] \geq -2)
\end{aligned}$$

We proceed to lower bound the expression above as follows:

$$\begin{aligned}
&\mu_{v_i}^{(0)} - \mathop{\mathbb{E}}_{v_i}[\bar{y}_{i,post}^{(1)} | \widehat{\xi}_{C,i}, \xi] \\
&= \mu_{v_i}^{(0)} - \mathop{\mathbb{E}}_{v_i}\left[\bar{y}_{i,post}^{(1)} \middle| \mu_{v_i}^{(0)} \geq \widehat{\mathbb{E}}[\bar{y}_{i,post}^{(1)} + C], |\widehat{\mathbb{E}}[\bar{y}_{i,post}^{(1)}] - \mathbb{E}[\bar{y}_{i,post}^{(1)}]| \leq \alpha(\delta_{PCR}),\right. \\
&\qquad \left. |\frac{1}{T_1}\sum_{t=T_0+1}^{T} \varepsilon_{i,t}^{(1)}| \leq \sigma\sqrt{\frac{2\log(1/\delta_{\varepsilon_i})}{T_1}}\right] \\
&= \mu_{v_i}^{(0)} - \mathop{\mathbb{E}}_{v_i}\left[\bar{y}_{i,post}^{(1)} \middle| \mu_{v_i}^{(0)} - \mathbb{E}[\bar{y}_{i,post}^{(1)}] \geq C - \alpha(\delta_{PCR}), |\frac{1}{T_1}\sum_{t=T_0+1}^{T} \varepsilon_{i,t}^{(1)}| \leq \sigma\sqrt{\frac{2\log(1/\delta_{\varepsilon_i})}{T_1}}\right] \\
&= \mu_{v_i}^{(0)} - \mathbb{E}[v_i]\left[\bar{y}_{i,post}^{(1)} \middle| \mu_{v_i}^{(0)} - \bar{y}_{i,post}^{(1)} \geq C - \alpha(\delta_{PCR}) - \sigma\sqrt{\frac{2\log(1/\delta_{\varepsilon_i})}{T_1}}\right] \\
&\geq C - \alpha(\delta_{PCR}) - \sigma\sqrt{\frac{2\log(1/\delta_{\varepsilon_i})}{T_1}}
\end{aligned}$$

Furthermore, we can write the probability of the joint event $\widehat{\xi}_{C,i}, \xi$ as:

$$\begin{aligned}
&\Pr[\widehat{\xi}_{C,i}, \xi] \\
&= \Pr\left[\mu_{v_i}^{(0)} \geq \widehat{\mathbb{E}}[\bar{y}_{i,post}^{(1)} + C], |\widehat{\mathbb{E}}[\bar{y}_{i,post}^{(1)}] - \mathbb{E}[\bar{y}_{i,post}^{(1)}]| \leq \alpha(\delta_{PCR}),\right. \\
&\qquad \left. |\frac{1}{T_1}\sum_{t=T_0+1}^{T} \varepsilon_{i,t}^{(1)}| \leq \sigma\sqrt{\frac{2\log(1/\delta_{\varepsilon_i})}{T_1}}\right] \\
&= \Pr\left[\mu_{v_i}^{(0)} - \bar{y}_{i,post}^{(1)} \geq C - \alpha(\delta_{PCR}) - \sigma\sqrt{\frac{2\log(1/\delta_{\varepsilon_i})}{T_1}}\right]
\end{aligned}$$

Hence, we can derive the following lower bound on the denominator of L:

$$\begin{aligned}
&\mu_{v_i}^{(0)} - \mathop{\mathbb{E}}_{v_i}[\bar{y}_{i,post}^{(1)}|\widehat{\xi}_{C,i}] \\
&\geq \left(C - \alpha(\delta_{PCR}) - \sigma\sqrt{\frac{2\log(1/\delta_{\varepsilon_i})}{T_1}}\right) \cdot \\
&\quad \Pr\left[\mu_{v_i}^{(0)} - \bar{y}_{i,post}^{(1)} \geq C - \alpha(\delta_{PCR}) - \sigma\sqrt{\frac{2\log(1/\delta_{\varepsilon_i})}{T_1}}\right] - 2\delta
\end{aligned}$$

Applying this lower bound to the expression of L and taking the maximum over type 1 units, we have:

$$\begin{aligned}
L \geq 1 + \max_{i \in \mathcal{I}^{(1)}} \Bigg\{ &\frac{\mu_{v_i}^{(1)} - \mu_{v_i}^{(0)}}{\left(C - \alpha(\delta_{PCR}) - \sigma\sqrt{\frac{2\log(1/\delta_{\varepsilon_i})}{T_1}}\right)} \times \\
&\frac{1}{\Pr\left[\mu_{v_i}^{(0)} - \bar{y}_{i,post}^{(1)} \geq C - \alpha(\delta_{PCR}) - \sigma\sqrt{\frac{2\log(1/\delta_{\varepsilon_i})}{T_1}}\right] - 2\delta} \Bigg\}
\end{aligned}$$

□

Example C.III.5. Consider the following data-generating process: Suppose there are two receiver types, each with equal probability; type 1 units prefer the treatment and type 0 units prefer control. Let $v_i \sim \mathsf{Unif}[0.25, 0.75]$. If unit i is a type 1 unit, they have latent factor $\mathbf{v}_i = [v_i \; 0]$. Otherwise they have latent factor $\mathbf{v}_i = [0 \; v_i]$. Suppose that the pre-intervention period is of length $T_0 = 2$ and the post-intervention period is of length $T_1 = 1$. We leave $\mathbf{u}_1^{(0)}, \mathbf{u}_2^{(0)}, \mathbf{u}_3^{(0)}, \mathbf{u}_3^{(1)}$ unspecified.

Since $r_1 = 1$, we only need to incentivize a single type 1 unit to take the control in order for the Unit Overlap Assumption to be satisfied for type 1 units under control. Therefore, for a given confidence level δ, it suffices to set $N_0 = B = \frac{\log(1/\delta)}{\log(2)}$. After the first N_0 time-steps, the principal will know all of the parameters necessary to compute the batch size L.

Since $T_1 = 1$, unit i's prior is over $\{\mathbb{E}[y_{i,3}^{(0)}], \mathbb{E}[y_{i,3}^{(1)}]\}$. For simplicity, suppose that (1) all type 1 units believe that $\mathbb{E}[y_{i,3}^{(0)}] \sim \mathsf{Unif}[0, 0.5]$ and $\mathbb{E}[y_{i,3}^{(1)}] \sim \mathsf{Unif}[0, 1]$ and (2) there is no noise in the post-intervention outcome. Under this setting, the event $\xi_{C,i}$ simplifies to $\xi_{C,i} = \left\{0.25 \leq \mathbb{E}[y_{i,3}^{(1)}] + C\right\}$ for any unit i and constant $C \in (0,1)$. Therefore

$$\zeta_C = \Pr_{\mathcal{P}_i}[\mathbb{E}[y_{i,3}^{(1)}] \leq 0.25 - C] = \max\{0, 0.25 - C\}.$$

C.IV Appendix for Section 5.5: Extension to Synthetic Interventions

Theorem C.IV.1. *Suppose that Assumption 5.5.2 holds for some constant gap $C \in (0,1)$. If the number of initial units N_0 is chosen to be large enough such that Unit Overlap Assumption is satisfied for all units of subtype $\boldsymbol{\tau}$ under intervention $\boldsymbol{\tau}[1]$ with probability $1-\delta$, then Algorithm 6 with parameters δ, L, B, C is BIC for all units of subtype $\boldsymbol{\tau}$ if:*

$$L \geq 1 + \max_{i \in \mathcal{T}^{(\boldsymbol{\tau})}} \left\{ \frac{\mu_i^{(\boldsymbol{\tau}[1])} - \mu_i^{(\boldsymbol{\tau}[k])}}{\left(C - 2\alpha(\delta_{PCR}) - 2\sigma\sqrt{\frac{2\log(1/\delta_{\varepsilon_i})}{T_1}}\right)(1-\delta)\Pr\left[\mathcal{E}_{C,i}\right] - 2\delta} \right\}$$

Moreover, if the number of batches B is chosen to be large enough such that with probability at least $1-\delta$, we have $\text{rank}([\mathbb{E}[\mathbf{y}_{i,pre}]^\top : \widehat{d}_i = \ell \neq \boldsymbol{\tau}[1], i \in \mathcal{T}^{(\tau)}]_{i=1}^{N_0+B\cdot L}) = \text{rank}([\mathbb{E}[\mathbf{y}_{j,pre}]^\top]_{j\in\mathcal{T}^{(\tau)}})$*, then the Unit Overlap Assumption will be satisfied for all units of subtype $\boldsymbol{\tau}$ under all interventions with probability at least $1 - \mathcal{O}(k\delta)$.*

Proof. According to our recommendation policy, unit i is either recommended intervention $\boldsymbol{\tau}[1]$ or intervention ℓ. We will prove that in either case, unit i will comply with the principal's recommendation.

Let $\widehat{\mathcal{E}}_{C,i}^{(\ell)}$ denote the event that the exploit intervention for unit i is intervention ℓ. Formally, we have

$$\widehat{\mathcal{E}}_{C,i}^{(\ell)} = \left\{ \widehat{\mathbb{E}}[\bar{y}_{i,post}^{(\boldsymbol{\tau}[1])}] \leq \min_{1<j<\ell} \widehat{\mathbb{E}}[\bar{y}_{i,post}^{(\boldsymbol{\tau}[j])}] - C \quad \text{and} \quad \max_{1\leq j<\ell} \widehat{\mathbb{E}}[\bar{y}_{i,post}^{(\boldsymbol{\tau}[j])}] \leq \mu_{v_i}^{(\ell)} - C \right\}$$

When unit i is recommended intervention ℓ: When a unit $i \in \mathcal{T}^{(\boldsymbol{\tau})}$ is recommended intervention ℓ, we argue that this unit will not switch to any other intervention $j \neq \ell$. Because of our ordering of the prior mean reward, any intervention $j > \ell$ has had no sample collected by a type $\boldsymbol{\tau}$ unit and $\mu_{v_i}^{(\boldsymbol{\tau}[j])} \leq \mu_{v_i}^{(\boldsymbol{\tau}[\ell])}$. Hence, we only need to focus on the cases where $j < \ell$. For the recommendation policy to be BIC, we need to show that

$$\mathbb{E}[\mu_{v_i}^{(\ell)} - \bar{y}_{i,post}^{(\boldsymbol{\tau}[j])} | \widehat{d} = \ell] \Pr[\widehat{d} = \ell] \geq 0$$

There are two possible disjoint events under which unit i is recommended intervention ℓ: either ℓ is determined to be the 'exploit' intervention or unit i is chosen as an 'explore' unit. Since being chosen as an explore unit does not imply any information about the rewards, we can derive that conditional on being in an 'explore' unit, the expected gain for unit i to switch to intervention j is simply $\mu_{v_i}^{(\ell)} - \mu_{v_i}^{(\boldsymbol{\tau}[j])}$. On the other hand, if intervention ℓ is the 'exploit' intervention, then event $\widehat{\mathcal{E}}_{C,i}^{(\ell)}$ has happened. Hence, we can rewrite the left-hand side of the BIC condition above as:

$$\mathbb{E}[\mu_{v_i}^{(\ell)} - \bar{y}_{i,post}^{(\boldsymbol{\tau}[j])} | \widehat{d} = \ell] \Pr[\widehat{d} = \ell] = \mathbb{E}[\mu_{v_i}^{(\ell)} - \bar{y}_{i,post}^{(\boldsymbol{\tau}[j])} | \widehat{\mathcal{E}}_{C,i}^{(\ell)}] \Pr[\widehat{\mathcal{E}}_{C,i}^{(\ell)}] \left(1 - \frac{1}{L}\right) + \frac{1}{L}(\mu_{v_i}^{(\ell)} - \mu_{v_i}^{(j)})$$

Rearranging the terms, we have the following lower bound on the phase length L for the algorithm to be BIC for type 1 units:

$$L \geq 1 + \frac{\mu_{v_i}^{(\boldsymbol{\tau}[j])} - \mu_{v_i}^{(\ell)}}{\mathbb{E}[\mu_{v_i}^{(\ell)} - \bar{y}_{i,post}^{(\boldsymbol{\tau}[j])} | \widehat{\mathcal{E}}_{C,i}^{(\ell)}] \Pr[\widehat{\mathcal{E}}_{C,i}^{(\ell)}]}$$

To complete the analysis, we need to lower bound the denominator of the expression above. Consider the following event $\mathcal{C}$:

$$\mathcal{C} = \{\forall j \in [k] : |\widehat{\mathbb{E}}[\bar{y}_i^{(\boldsymbol{\tau}[j])}] - \mathbb{E}[\bar{y}_{i,post}^{(\boldsymbol{\tau}[j])}]| \leq \alpha(\delta_{PCR})\}$$

Let $\mathcal{E}_0$ denote the event that Unit Overlap Assumption is satisfied for units of type $\boldsymbol{\tau}$ under intervention $\boldsymbol{\tau}[1]$. From Assumption 5.5.2, we know that event $\mathcal{E}_0$ occurs with probability at least $1 - \delta_0$ for some $\delta_0 \in (0, 1)$. Furthermore, let $\mathcal{E}_{\varepsilon_i}$ denote the event where the Chernoff-Hoeffding bound on the noise sequences $\{\varepsilon_{i,t}^{(\boldsymbol{\tau}[j])}\}_{t=T_0+1}^{T}$ holds, that is with probability at least $1 - \delta_{\varepsilon_i}$, we have:

$$\left|\frac{1}{T_1} \sum_{t=T_0+1}^{T} \varepsilon_{i,t}^{(\boldsymbol{\tau}[j])}\right| \leq \sigma \sqrt{\frac{2 \log(1/\delta_{\varepsilon_i})}{T_1}}$$

Then, we can define a clean event $\mathcal{E}$ where the events $\mathcal{C}$, $\mathcal{E}_0$ and $\mathcal{E}_{\varepsilon_i}$ happens simultaneously with probability at least $1-\delta$, where $\delta = \delta_{\varepsilon}^{PCR} + \delta_{\varepsilon_i} + \delta_0$.

Note that since $|\bar{y}_{i,post}^{(\boldsymbol{\tau}[j])}| \leq 1$, we can rewrite the denominator as:

$$\begin{aligned}
&\mathbb{E}[\mu_{v_i}^{(\ell)} - \bar{y}_{i,post}^{(\boldsymbol{\tau}[j])} | \widehat{\mathcal{E}}_{C,i}] \Pr[\widehat{\mathcal{E}}_{C,i}^{(\ell)}] \\
&= \mathbb{E}[\mu_{v_i}^{(\ell)} - \bar{y}_{i,post}^{(\boldsymbol{\tau}[j])} | \widehat{\mathcal{E}}_{C,i}^{(\ell)}, \mathcal{E}] \Pr[\widehat{\mathcal{E}}_{C,i}^{(\ell)}, \mathcal{E}] + \mathbb{E}[\mu_{v_i}^{(\ell)} - \bar{y}_{i,post}^{(\boldsymbol{\tau}[j])} | \widehat{\mathcal{E}}_{C,i}^{(\ell)}, \neg\mathcal{E}] \Pr[\widehat{\mathcal{E}}_{C,i}^{(\ell)}, \neg\mathcal{E}] \\
&\geq \mathbb{E}[\mu_{v_i}^{(\ell)} - \bar{y}_{i,post}^{(\boldsymbol{\tau}[j])} | \widehat{\mathcal{E}}_{C,i}^{(\ell)}, \mathcal{E}] \Pr[\widehat{\mathcal{E}}_{C,i}^{(\ell)}, \mathcal{E}] - 2\delta
\end{aligned}$$

Define an event $\mathcal{E}_i^{(\ell)}$ on the true average expected post-intervention outcome of each intervention as follows:

$$\begin{aligned}
\mathcal{E}_i^{(\ell)} := \Bigg\{ &\bar{y}_{i,post}^{(\boldsymbol{\tau}[1])} \leq \min_{1<j<\ell} \bar{y}_{i,post}^{(\boldsymbol{\tau}[j])} - C - 2\alpha(\delta_{PCR}) - 2\sigma\sqrt{\frac{2\log(1/\delta_{\varepsilon_i})}{T_1}} \\
&\text{and} \quad \max_{1\leq j<\ell} \bar{y}_{i,post}^{(\boldsymbol{\tau}[j])} \leq \mu_{v_i}^{(\ell)} - C - 2\alpha(\delta_{PCR}) - 2\sigma\sqrt{\frac{2\log(1/\delta_{\varepsilon_i})}{T_1}} \Bigg\}
\end{aligned}$$

We can observe that under event $\mathcal{E}$, event $\widehat{\mathcal{E}}_{C,i}^{(\ell)}$ is implied by event $\mathcal{E}_i^{(\ell)}$. Hence, we have:

$$\Pr[\mathcal{E}, \mathcal{E}_i^{(\ell)}] \leq \Pr[\mathcal{E}, \widehat{\mathcal{E}}_{C,i}^{(\ell)}]$$

We can rewrite the left-hand side as:

$$\Pr[\mathcal{C}, \mathcal{E}_i] = \Pr[\mathcal{C} | \mathcal{E}_i^{(\ell)}] \Pr[\mathcal{E}_i^{(\ell)}] \geq (1-\delta) \Pr[\mathcal{E}_i^{(\ell)}]$$

Substituting these expressions using $\mathcal{E}_i^{(\ell)}$ for the ones using $\widehat{\mathcal{E}}_{C,i}$ in the denominator gives:

$$\begin{aligned}
&\mathbb{E}[\mu_{v_i}^{(\ell)} - \bar{y}_{i,post}^{(\boldsymbol{\tau}[j])} | \widehat{\mathcal{E}}_{C,i}^{(\ell)}] \Pr[\widehat{\mathcal{E}}_{C,i}^{(\ell)}] \\
&\geq \left(C - 2\alpha(\delta_{PCR}) - 2\sigma\sqrt{\frac{2\log(1/\delta_{\varepsilon_i})}{T_1}} \right) (1-\delta) \Pr[\mathcal{E}_i^{(\ell)}] - 2\delta
\end{aligned}$$

Applying this lower bound to the expression of L and taking the maximum over type 1 units, we have:

$$L \geq 1 + \max_{i \in \mathcal{I}^{(1)}} \left\{ \frac{\mu_{v_i}^{(\boldsymbol{\tau}[1])} - \mu_{v_i}^{(\boldsymbol{\tau}[k])}}{\left(C - 2\alpha(\delta_{PCR}) - 2\sigma\sqrt{\frac{2\log(1/\delta_{\varepsilon_i})}{T_1}}\right)(1-\delta)\Pr\left[\mathcal{E}_{C,i}^{(\ell)}\right] - 2\delta} \right\}$$

When unit i is recommended intervention $\boldsymbol{\tau}[1]$: When unit i gets recommended intervention $\boldsymbol{\tau}[1]$, they know that they are not in the explore group. Hence, the event $\widehat{\mathcal{E}}_{C,i}$ did not happen, and the BIC condition, in this case, can be written as: for any intervention $\boldsymbol{\tau}[j] \neq \boldsymbol{\tau}[1]$,

$$\mathbb{E}[\bar{y}_{i,post}^{(\boldsymbol{\tau}[1])} - \bar{y}_{i,post}^{(\boldsymbol{\tau}[j])} | \neg\widehat{\mathcal{E}}_{C,i}^{(\ell)}] \Pr[\neg\widehat{\mathcal{E}}_{C,i}^{(\ell)}] \geq 0$$

Similar to the previous analysis on the recommendation of intervention ℓ, it suffices to only consider interventions $\boldsymbol{\tau}[j] < \ell$. We have:

$$\begin{aligned}\mathbb{E}[\bar{y}_{i,post}^{(\boldsymbol{\tau}[1])} - \bar{y}_{i,post}^{(\boldsymbol{\tau}[j])} | \neg\widehat{\mathcal{E}}_{C,i}^{(\ell)}] \Pr[\neg\widehat{\mathcal{E}}_{C,i}^{(\ell)}] &= \mathbb{E}[\bar{y}_{i,post}^{(\boldsymbol{\tau}[1])} - \bar{y}_{i,post}^{(\boldsymbol{\tau}[j])}] - \mathbb{E}[\bar{y}_{i,post}^{(\boldsymbol{\tau}[1])} - \bar{y}_{i,post}^{(\boldsymbol{\tau}[j])} | \widehat{\mathcal{E}}_{C,i}^{(\ell)}] \Pr[\widehat{\mathcal{E}}_{C,i}^{(\ell)}] \\ &= \mu_{v_i}^{(\boldsymbol{\tau}[1])} - \mu_{v_i}^{(\boldsymbol{\tau}[j])} + \mathbb{E}[\bar{y}_{i,post}^{(\boldsymbol{\tau}[j])} - \bar{y}_{i,post}^{(\boldsymbol{\tau}[1])} | \widehat{\mathcal{E}}_{C,i}^{(\ell)}] \Pr[\widehat{\mathcal{E}}_{C,i}^{(\ell)}]\end{aligned}$$

By definition, we have $\mu_{v_i}^{(\boldsymbol{\tau}[1])} \geq \mu_{v_i}^{(\boldsymbol{\tau}[j])}$. Hence, it suffices to show that for any intervention $1 < \boldsymbol{\tau}[j] < \ell$, we have:

$$\mathbb{E}[\bar{y}_{i,post}^{(\boldsymbol{\tau}[j])} - \bar{y}_{i,post}^{(\boldsymbol{\tau}[1])} | \widehat{\mathcal{E}}_{C,i}^{(\ell)}] \Pr[\widehat{\mathcal{E}}_{C,i}^{(\ell)}] \geq 0$$

Observe that with the event $\mathcal{E}$ defined above, we can write:

$$\begin{aligned}&\mathbb{E}[\bar{y}_{i,post}^{(\boldsymbol{\tau}[j])} - \bar{y}_{i,post}^{(\boldsymbol{\tau}[1])} | \widehat{\mathcal{E}}_{C,i}^{(\ell)}] \Pr[\widehat{\mathcal{E}}_{C,i}^{(\ell)}] \\ &= \mathbb{E}[\bar{y}_{i,post}^{(\boldsymbol{\tau}[j])} - \bar{y}_{i,post}^{(\boldsymbol{\tau}[1])} | \widehat{\mathcal{E}}_{C,i}^{(\ell)}, \mathcal{E}] \Pr[\widehat{\mathcal{E}}_{C,i}^{(\ell)}, \mathcal{E}] + \mathbb{E}[\bar{y}_{i,post}^{(\boldsymbol{\tau}[j])} - \bar{y}_{i,post}^{(\boldsymbol{\tau}[1])} | \widehat{\mathcal{E}}_{C,i}^{(\ell)}, \neg\mathcal{C}] \Pr[\widehat{\mathcal{E}}_{C,i}^{(\ell)}, \neg\mathcal{E}] \\ &\geq \mathbb{E}[\bar{y}_{i,post}^{(\boldsymbol{\tau}[j])} - \bar{y}_{i,post}^{(\boldsymbol{\tau}[1])} | \widehat{\mathcal{E}}_{C,i}^{(\ell)}, \mathcal{E}] \Pr[\widehat{\mathcal{E}}_{C,i}^{(\ell)}, \mathcal{E}] - 2\delta\end{aligned}$$

When event $\widehat{\mathcal{E}}_{C,i}^{(\ell)}$ happens, we know that $\widehat{\mathbb{E}}[\bar{y}_{i,post}^{(\boldsymbol{\tau}[j])}] \geq \widehat{\mathbb{E}}[\bar{y}_{i,post}^{(\boldsymbol{\tau}[1])}] + C$. Furthermore, when event $\mathcal{E}$ happens, we know that $\bar{y}_{i,post}^{(\boldsymbol{\tau}[j])} \geq \widehat{\mathbb{E}}[\bar{y}_{i,post}^{(\boldsymbol{\tau}[j])}] - \alpha(\delta_{PCR}) - \sigma\sqrt{\frac{2\log(1/\delta_{\varepsilon_i})}{T_1}}$ and $\widehat{\mathbb{E}}_{[}\bar{y}_{i,post}^{(\boldsymbol{\tau}[1])}] \geq \bar{y}_{i,post}^{(\boldsymbol{\tau}[1])} - \alpha(\delta_{PCR}) - \sigma\sqrt{\frac{2\log(1/\delta_{\varepsilon_i})}{T_1}}$. Hence, when these two events $\widehat{\mathcal{E}}_{C,i}^{(\ell)}$ and $\mathcal{E}$ happen simultaneously, we have $\bar{y}_{i,post}^{(\boldsymbol{\tau}[j])} \geq \bar{y}_{i,post}^{(\boldsymbol{\tau}[1])} + C - 2\alpha(\delta_{PCR}) - 2\sigma\sqrt{\frac{2\log(1/\delta_{\varepsilon_i})}{T_1}}$.

Therefore, the lower bound can be written as:

$$\mathbb{E}[\bar{y}_{i,post}^{(\boldsymbol{\tau}[j])} - \bar{y}_{i,post}^{(\boldsymbol{\tau}[1])} | \widehat{\mathcal{E}}_{C,i}^{(\ell)}] \Pr[\widehat{\mathcal{E}}_{C,i}^{(\ell)}] \geq \left(C - 2\alpha(\delta_{PCR}) - 2\sigma\sqrt{\frac{2\log(1/\delta_{\varepsilon_i})}{T_1}} \right) \Pr[\widehat{\mathcal{E}}_{C,i}^{(\ell)}, \mathcal{E}] - 2\delta$$

Similar to the previous analysis when unit i gets recommended intervention ℓ, we have:

$$\begin{aligned}
&\mathbb{E}[\bar{y}_{i,post}^{(\boldsymbol{\tau}[j])} - \bar{y}_{i,post}^{(\boldsymbol{\tau}[1])} | \widehat{\mathcal{E}}_{C,i}] \Pr[\widehat{\mathcal{E}}_{C,i}] \\
&\geq \left(C - 2\alpha(\delta_{PCR}) - 2\sigma\sqrt{\frac{2\log(1/\delta_{\varepsilon_i})}{T_1}} \right) (1-\delta) \Pr[\mathcal{E}_i^{(\ell)}] - 2\delta
\end{aligned}$$

Choosing a large enough C such that the right-hand side is non-negative, we conclude the proof. □

C.V Appendix for Section 5.6: Testing Whether the Unit Overlap Assumption Holds

Theorem 5.6.2. *Under Assumption 5.6.1, if the Unit Overlap Assumption is satisfied for unit n, then*

$$|\widehat{\mathbb{E}}[\bar{y}_{n,pre''}] - \bar{y}_{n,pre''}| \leq \alpha(\delta) + 2\sigma\sqrt{\frac{\log(1/\delta)}{T_0}}$$

with probability $1 - \mathcal{O}(\delta)$, where $\alpha(\delta)$ is the high-probability confidence interval which is defined in Theorem C.III.1 when using the first $T_0/2$ time-steps as the pre-intervention period and the next $T_0/2$ time-steps as the post-intervention period.

Proof.

$$\begin{aligned}
|\widehat{\mathbb{E}}[\bar{y}_{i,pre''}] - \bar{y}_{i,pre''}| &\leq |\widehat{\mathbb{E}}[\bar{y}_{i,pre''}] - \mathbb{E}[\bar{y}_{i,pre''}]| + |\,\mathbb{E}[\bar{y}_{i,pre''}] - \bar{y}_{i,pre''}| \\
&\leq \alpha(\delta) + \left| \frac{2}{T_0} \sum_{t=T_0/2+1}^{T_0} \varepsilon_{i,t}^{(0)} \right|
\end{aligned}$$

with probability at least $1 - \mathcal{O}(\delta)$, where the second inequality follows from Theorem C.III.1. The result follows from a Hoeffding bound. □

Theorem C.V.1. *Under Assumption 5.6.1, if*

$$\frac{r^{3/2}\sqrt{\log(T_0|\mathcal{I}|)}}{\|\widetilde{\omega}^{(n,0)}\|_2 \cdot \min\{T_0, |\mathcal{I}|, T_0^{1/4}|\mathcal{I}|^{1/2}\}} = o(1) \tag{C.2}$$

and

$$\frac{1}{\sqrt{T_1}\|\widetilde{\omega}^{(n,0)}\|_2} \sum_{t=T_0+1}^{T} \sum_{i\in\mathcal{I}} \mathbb{E}[y_{i,t}^{(0)}] \cdot (\widehat{\omega}^{(n,0)}[i] - \widetilde{\omega}^{(n,0)}[i]) = o_p(1) \tag{C.3}$$

then as $|\mathcal{I}|, T_0, T_1 \to \infty$ the Asymptotic Hypothesis Test falsely accepts the hypothesis with probability at most 5%, where $\widehat{\sigma}^2$ is defined as in Equation (4) and $\widehat{\omega}^{(n,0)}$ is defined as in Equation (5) in Agarwal et al. (2020b).

Condition (C.2) requires that the ℓ_2 norm of $\widetilde{\omega}^{(n,0)}$ is sufficiently large, and condition (C.3) requires that the estimation error of $\widehat{\omega}^{(n,0)}$ decreases sufficiently fast. See Section 5.3 of Agarwal et al. (2020b) for more details.

Proof. We use $x \xrightarrow{p} y$ (resp. $x \xrightarrow{d} y$) if x converges in probability (resp. distribution) to y. Let σ' denote the true standard deviation of $\varepsilon_{i,t}$. By Theorem 3 of Agarwal et al. (2020b) we know that as $|\mathcal{I}|, T_0, T_1 \to \infty$,

1. $\frac{\sqrt{T_1}}{\sigma'\|\widetilde{\omega}^{(n,0)}\|_2}(\widehat{\mathbb{E}}[\bar{y}_{n,pre''}] - \mathbb{E}[\bar{y}_{n,pre''}]) \xrightarrow{d} \mathcal{N}(0,1)$
2. $\widehat{\sigma} \xrightarrow{p} \sigma'$ and
3. $\widehat{\omega}^{(n,0)} \xrightarrow{p} \widetilde{\omega}^{(n,0)}$.

By the continuous mapping theorem, we know that $\frac{1}{\widehat{\sigma}} \xrightarrow{p} \frac{1}{\sigma'}$ and $\frac{1}{\|\widehat{\omega}^{(n,0)}\|_2} \xrightarrow{p} \frac{1}{\|\widetilde{\omega}^{(n,0)}\|_2}$. We can rewrite $\widehat{\mathbb{E}}[\bar{y}_{n,pre''}] - \mathbb{E}[\bar{y}_{n,pre''}]$ as $\widehat{\mathbb{E}}[\bar{y}_{n,pre''}] - \bar{y}_{n,pre''} + \bar{\varepsilon}_{n,pre''}$, and we know that $\bar{\varepsilon}_{n,pre''} \xrightarrow{p} 0$. Applying Slutksy's theorem several times obtains the desired result. □

C.VI Appendix for Section 5.7: Numerical Simulations

If unit i is of type 1 (resp. type 0), we generate latent factor $\mathbf{v}_i = [0 \ \cdots \ 0 \ v_i[1] \ \cdots \ v_i[r/2]]$ (resp. $\mathbf{v}_i = [v_i[1] \ \cdots \ v_i[r/2] \ 0 \ \cdots \ 0]$), where $\forall j \in [r/2] : v_i[j] \sim \mathsf{Unif}(0,1)$. We consider an online setting where 500 units of alternating types arrive sequentially.[1] We consider a pre-intervention time period of length $T_0 = 100$ with latent factors as follows:

- If $t \mod 2 = 0$, generate latent factor $\mathbf{u}_t^{(0)} = [0 \cdots 0 \; u_t^{(0)}[1] \cdots u_t^{(0)}[r/2]]$, where $\forall \ell \in [r/2] : u_t^{(0)}[\ell] \sim \mathsf{Unif}[0.25, 0.75]$.

- If $t \mod 2 = 1$, generate latent factor $\mathbf{u}_t^{(0)} = [u_t^{(0)}[1] \cdots u_t^{(0)}[r/2] \; 0 \cdots 0]$, where $\forall \ell \in [r/2] : u_t^{(0)}[\ell] \sim \mathsf{Unif}[0.25, 0.75]$.

We consider a post-intervention period of length $T_1 = 100$ and generate the following post-intervention latent factors:

- If $d = 0$, we set $\mathbf{u}_t^{(d)} = [u_t^{(d)}[1] \; \cdots \; u_t^{(d)}[r]] : t \in [T_0 + 1, T]$, where $\forall \ell \in [r] : u_t^{(d)}[\ell] \sim \mathsf{Unif}[0, 1]$.

- If $d = 1$, we set $\mathbf{u}_t^{(d)} = [u_t^{(d)}[1] \; \cdots \; u_t^{(d)}[r]] : t \in [T_0 + 1, T]$, where $\forall \ell \in [r] : u_t^{(d)}[\ell] \sim \mathsf{Unif}[-1, 0]$.

Finally, outcomes are generated by adding independent Gaussian noise $\varepsilon_{i,t}^{(d)} \sim \mathcal{N}(0, 0.01)$ to each inner product of latent factors.

[1]The alternating arrival of different unit types is done only for convenience. The order of unit arrival is unknown to the algorithms.

Appendix D

Appendix for Incentivized Exploration in Two-sided Matching Markets

D.I Proofs of incentivized exploration for two types of agents and arms

D.I.1 Warm-start proofs

In this proof, we show that all agents and arms of type $i, j \in [2]$ are willing to change their initial preferences and comply with the platform's recommendation. In the beginning, the platform wants to incentivize all agents of type 1 and arms of type 1 to change their initial preference and choose arms of type 2 (respectively, agents of type 2 instead). Then, the platform uses these collected samples to convince agents of type 2 to match with arms of type 2, thus completing the incentive-aware algorithm 1 for all types of agents and arms.

Proof of Theorem 6.4.2.

Proof. First, we show that agents of type 1 and arms of type 1 are willing to change their initial preference and follow the platform's recommendation. Then, we show that agents of type 2 and arms of type 2 will follow the recommendation and match each other.

Recommended action is $A^{(t)} = \{(x_1, a_2), (x_2, a_1)\}$. For both agents of type 1 and arms of type 1 to change their initial preferences, we want to show that:

$$\mathbb{E}[\mu_{1,2} - \mu_{1,1} | A^{(t)} = \{(x_1, a_2), (x_2, a_1)\}] \geq \varepsilon \tag{D.1}$$

and

$$\mathbb{E}[\mu_{2,1} - \mu_{1,1} | A^{(t)} = \{(x_1, a_2), (x_2, a_1)\}] \geq \varepsilon \tag{D.2}$$

Combining these conditions, we instead will prove the following:

$$\mathbb{E}[\mu_{1,2} + \mu_{2,1} - 2\mu_{1,1} | A^{(t)} = \{(x_1, a_2), (x_2, a_1)\}] \geq 2\varepsilon \tag{D.3}$$

Let $A^0 = \{(x_1, a_1), (x_1, a_1)\}$ be a dummy action whose reward is twice the reward from choosing the prior-best atom (x_1, a_1). Then, our goal is to show that $\mathbb{E}[\mu_A - \mu_{A^0} | A^{(t)} = A] \geq 2\varepsilon$.

Define the following two events:

$$\xi_1 = \{\text{exploit: } \mathbb{E}[\mu_A - \mu_{A^0} | S^k_{A^0}] > 0\} \tag{D.4}$$

and

$$\xi_2 = \{\text{explore: } \mathbb{E}[\mu_A - \mu_{A^0} | S^k_{A^0}] \leq 0 \text{ and selected for exploration}\} \tag{D.5}$$

Then, we can write

$$E[\mu_A - \mu_{A^0} | A^{(t)} = A] \geq \mathbb{E}[\mu_A - \mu_{A^0} | \xi_1] \Pr[\xi_1] + \mathbb{E}[\mu_A - \mu_{A^0} | \xi_2] \Pr[\xi_2]$$

Let $\Delta^k_{A,A^0} \coloneqq \mathbb{E}[\mu(A) - \mu(A^0) | S^k_{A^0}]$. Then, we have:

$$\begin{aligned}
\Pr[\xi_2] &= \Pr\left[\mathbb{E}[\mu_A - \mu_{A^0} | S^k_{A^0}] \leq 0 \text{ and selected for exploration}\right] \\
&= \Pr[\Delta^k_{A,A^0} \leq 0] \Pr[\text{selected} | \Delta^k_{A,A^0} \leq 0] \\
&= \frac{1}{L} \cdot \Pr[\Delta^k_{A,A^0} \leq 0]
\end{aligned}$$

where the first equality is by definition and the second equality is due to Δ^k_{A,A^0} being independent of the event that the learner is selected for exploration. Then, we can write

$$\begin{aligned}
&\mathbb{E}[\Delta^k_{A,A^0}|\xi_2]\Pr[\xi_2] \\
&= \mathbb{E}[\Delta^k_{A,A^0}|\Delta^k_{A,A^0} \leq 0 \text{ and selected}]\Pr[\Delta^k_{A,A^0} \leq 0] \cdot \frac{1}{L} \\
&= \mathbb{E}[\Delta^k_{A,A^0}|\Delta^k_{A,A^0} \leq 0]\Pr[\Delta^k_{A,A^0} \leq 0] \cdot \frac{1}{L}
\end{aligned}$$

Hence, the left-hand side of the dual BIC condition is

$$\begin{aligned}
&\mathbb{E}[\Delta^k_{A,A^0}|A^{(t)} = A]\Pr[A^{(t)} = A] \\
&= \mathbb{E}[\Delta^k_{A,A^0}|\Delta^k_{A,A^0} > 0]\Pr[\Delta^k_{A,A^0} > 0] \\
&+ \mathbb{E}[\Delta^k_{A,A^0}|\Delta^k_{A,A^0} \text{ and selected}]\Pr[\Delta^k_{A,A^0} < 0 \text{ and selected}] \\
&= \mathbb{E}[\Delta^k_{A,A^0}|\Delta^k_{A,A^0} > 0] \cdot Pr[\Delta^k_{A,A^0} > 0] \\
&+ \frac{1}{L} \cdot \mathbb{E}[\Delta^k_{A,A^0} \leq 0|\Delta^k_{A,A^0} \leq 0] \\
&= \left(1 - \frac{1}{L}\right) \cdot \mathbb{E}[\Delta^k_{A,A^0}|\Delta^k_{A,A^0} > 0] \cdot \Pr[\Delta^k_{A,A^0} > 0] \\
&+ \frac{1}{L} \cdot \Big(\mathbb{E}[\Delta^k_{A,A^0}|\Delta^k_{A,A^0} > 0] \cdot \Pr[\Delta^k_{A,A^0} > 0] + \mathbb{E}[\Delta^k_{A,A^0}|\Delta^k_{A,A^0} \leq 0] \cdot \Pr[\Delta^k_{A,A^0} \leq 0]\Big) \\
&= \left(1 - \frac{1}{L}\right) \cdot \mathbb{E}[\Delta^k_{A,A^0}|\Delta^k_{A,A^0} > 0] \cdot \Pr[\Delta^k_{A,A^0} > 0] + \frac{1}{L} \cdot \mathbb{E}[\Delta^k_{A,A^0}] \\
&= \frac{L-1}{L} \cdot \mathbb{E}[\Delta^k_{A,A^0}|\Delta^k_{A,A^0} > 0] \cdot \Pr[\Delta^k_{A,A^0} > 0] + \frac{1}{L} \cdot \left(\mu^0_A - \mu^0_{A^0}\right)
\end{aligned}$$

For the dual BIC condition to hold, we can set

$$\begin{aligned}
&\frac{L-1}{L} \cdot \mathbb{E}[\Delta^k_{A,A^0}|\Delta^k_{A,A^0} > 0]\Pr[\Delta^k_{A,A^0} > 0] + \frac{1}{L} \cdot (\mu^{(0)}_A - \mu^{(0)}_{A^0}) \geq 2\varepsilon \\
&\iff L \geq \frac{\mathbb{E}[\Delta^k_{A,A^0}|\Delta^k_{A,A^0} > 0]\Pr[\Delta^k_{A,A^0} > 0] - (\mu^{(0)}_A - \mu^{(0)}_{A^0})}{\mathbb{E}[\Delta^k_{A,A^0}|\Delta^k_{A,A^0} > 0]\Pr[\Delta^k_{A,A^0} > 0] - 2\varepsilon} \\
&\iff L \geq 1 - \frac{\mu^{(0)}_A - \mu^{(0)}_{A^0} - 2\varepsilon}{\mathbb{E}[\Delta^k_{A,A^0}|\Delta^k_{A,A^0} > 0]\Pr[\Delta^k_{A,A^0} > 0] - 2\varepsilon} \\
&\iff L \geq 1 + \frac{\mu^{(0)}_{A^0} - \mu^{(0)}_A + 2\varepsilon}{\mathbb{E}[\Delta^k_{A,A^0}|\Delta^k_{A,A^0} > 0]\Pr[\Delta^k_{A,A^0} > 0] - 2\varepsilon}
\end{aligned}$$

The expression above can be simplified by using definitions of $\tau_\mathcal{P}, \rho_\mathcal{P}$ and observing that $\mu^{(0)}_{A^0} - \mu^{(0)}_{A} \leq 2$ to get

$$L \geq 1 + \frac{2+2\varepsilon}{\tau_\mathcal{P} \cdot \rho_\mathcal{P} - 2\varepsilon} \tag{D.6}$$

Recommended action is $A^{(t)} = \{(x_1, a_1), (x_2, a_2)\}$. We want to show that all agents and arms will comply with this recommendation. That is, we want to show

$$\mathbb{E}[\mu_{2,2} - \mu_{2,1} | A^{(t)} = \{(x_1, a_1), (x_2, a_2)\}] \geq \varepsilon$$
$$\mathbb{E}[\mu_{2,2} - \mu_{1,2} | A^{(t)} = \{(x_1, a_1), (x_2, a_2)\}] \geq \varepsilon$$
$$\mathbb{E}[\mu_{1,1} - \mu_{1,2} | A^{(t)} = \{(x_1, a_1), (x_2, a_2)\}] \geq \varepsilon$$
$$\mathbb{E}[\mu_{1,1} - \mu_{2,1} | A^{(t)} = \{(x_1, a_1), (x_2, a_2)\}] \geq \varepsilon$$

Let $A_{2,2} = \{(x_2, a_2), (x_2, a_2)\}$ and $A_{1,1} = \{(x_1, a_1), (x_1, a_1)\}$ be a pair of dummy actions with reward twice that of atom (x_2, a_2) and (x_1, a_1), respectively. Let $A^0 = \{(x_1, a_2), (x_2, a_1)\}$ denote the prior-best actions for x_2 and a_2. Then, we can combine these conditions and show that:

$$\mathbb{E}[\mu_{A_{2,2}} - \mu_{A^0} | A^{(t)} = \{(x_1, a_1), (x_2, a_2)\}] \geq 2\varepsilon$$
$$\mathbb{E}[\mu_{A_{1,1}} - \mu_{A^0} | A^{(t)} = \{(x_1, a_1), (x_2, a_2)\}] \geq 2\varepsilon$$

First, we consider the incentives of x_1 and a_1. We have:

$$\begin{aligned}
&\mathbb{E}[\mu_{A_{1,1}} - \mu_{A^0} | \neg\text{explore}] \Pr[\neg\text{explore}] \\
&\mathbb{E}[\mu_{A_{1,1}} - \mu_{A^0}] - \mathbb{E}[\mu_{A_{1,1}} - \mu_{A^0} | \text{explore}] \Pr[\text{explore}] \\
&= 2\mu^{(0)}_{1,1} - (\mu^{(0)}_{1,2} + \mu^{(0)}_{2,1}) + \mathbb{E}[\mu_{A^0} - \mu_{A_{1,1}} | \text{explore}] \Pr[\text{explore}]
\end{aligned}$$

Since the first term is non-negative according to the initial preference ordering, it suffices to show that $\mathbb{E}[\mu_{A^0} - \mu_{A_{1,1}} | \text{explore}] \Pr[\text{explore}] \geq 2\varepsilon$. This inequality holds from the previous analysis for recommending $A_{\text{cand}} = \{(x_1, a_2), (x_2, a_1)\}$.

Then, we consider the incentives of x_2 and a_2. By construction, when agent x_2 receives a recommendation for arm a_2, they can infer that they are not in the explore group. Hence, it suffices to show that $\mathbb{E}[\mu_{A_{2,2}} - \mu_{A^0} | \neg\text{explore}] \Pr[\neg\text{explore}] \geq 2\varepsilon$. We have:

$$\begin{aligned}
&\mathbb{E}[\mu_{A_{2,2}} - \mu_{A^0} | \neg\text{explore}] \Pr[\neg\text{explore}] \\
&= \mathbb{E}[\mu_{A_{2,2}} - \mu_{A^0}] - \mathbb{E}[\mu_{A_{2,2}} - \mu_{A^0} | \text{explore}] \Pr[\text{explore}] \\
&= (2\mu_{2,2}^{(0)} - \mu_{2,1}^{(0)} - \mu_{1,2}^{(0)}) + \mathbb{E}[\mu_{A^0} - \mu_{A_{2,2}} | \text{explore}] \Pr[\text{explore}]
\end{aligned}$$

Define the following events:

$$\begin{aligned}
\xi_3 &= \{\mathbb{E}[\mu_{A^0} - \mu_{A_{2,2}} | S_{1,1}^k] > 0\} \\
\xi_4 &= \{\mathbb{E}[\mu_{A^0} - \mu_{A_{2,2}} | S_{1,1}^k] \leq 0\}
\end{aligned}$$

Then, we can write:

$$\begin{aligned}
&\mathbb{E}[\mu_{A^0} - \mu_{A_{2,2}} | \text{explore}] \Pr[\text{explore}] \\
&= \mathbb{E}[\mu_{A^0} - \mu_{A_{2,2}} | \xi_3] \Pr[\xi_3] + \mathbb{E}[\mu_{A^0} - \mu_{A_{2,2}} | \xi_4] \Pr[\xi_4]
\end{aligned}$$

Let $\Delta_{A^0, A_{2,2}}^k = \mathbb{E}[\mu_{A^0} - \mu_{A_{2,2}} | S_{1,1}^k]$. Then, we have:

$$\begin{aligned}
\Pr[\xi_3] &= \Pr[\Delta_{A^0, A_{2,2}}^k \leq 0 | \text{selected for exploration}] \Pr[\text{selected for exploration}] \\
&= \Pr[\Delta_{A^0, A_{2,2}}^k \leq 0] \Pr[\text{selected for exploration}]
\end{aligned}$$

Furthermore, we have

$$\begin{aligned}
&\mathbb{E}[\mu_{A^0} - \mu_{A_{2,2}} | \text{explore}] \Pr[\text{explore}] \\
&= \mathbb{E}[\mathbb{E}[\mu_{A^0} - \mu_{A_{2,2}} | S_{1,1}^k] | \text{explore}] \Pr[\text{explore}] \\
&= \mathbb{E}[\Delta_{A^0, A_{2,2}}^k | \text{explore}] \Pr[\text{explore}]
\end{aligned}$$

where the first equality is by the law of iterated expectation and the second equality is by definition of $\Delta_{A^0, A_{2,2}}^k$.

Therefore, we have:

$$\begin{aligned}
&\mathbb{E}[\mu_{A^0} - \mu_{A_{2,2}} | \text{explore}] \Pr[\text{explore}] \\
&= \mathbb{E}[\Delta^k_{A^0, A_{2,2}} | \xi_3] \Pr[\xi_3] + \mathbb{E}[\Delta^k_{A^0, A_{2,2}} | \xi_4] \Pr[\xi_4] \\
&= \mathbb{E}[\Delta^k_{A^0, A_{2,2}} | \xi_3] \Pr[\xi_3] + \mathbb{E}[\Delta^k_{A^0, A_{2,2}} | \Delta_{A^0, A_{2,2}} < 0] \Pr[\Delta^k_{A^0, A_{2,2}} < 0] \cdot \frac{1}{L} \\
&= \left(1 - \frac{1}{L}\right) \mathbb{E}[\Delta^k_{A^0, A_{2,2}} | \xi_3] \Pr[\xi_3] + \frac{1}{L} \cdot \mathbb{E}[\Delta^k_{A^0, A_{2,2}}] \\
&= \frac{L-1}{L} \mathbb{E}[\Delta^k_{A^0, A_{2,2}} | \xi_3] \Pr[\xi_3] + \frac{1}{L} \cdot (\mu^{(0)}_{1,2} + \mu^{(0)}_{2,1} - 2\mu^{(0)}_{2,2})
\end{aligned}$$

The BIC condition can be written as:

$$\begin{aligned}
&\mathbb{E}[\mu_{A_{2,2}} - \mu_{A^0} | \neg\text{explore}] \Pr[\neg\text{explore}] \\
&= 2\mu^{(0)}_{2,2} - (\mu^{(0)}_{1,2} + \mu^{(0)}_{2,1}) + \frac{L-1}{L} \cdot \mathbb{E}[\Delta^k_{A^0, A_{2,2}} | \xi_3] \Pr[\xi_3] + \frac{1}{L} \cdot ((\mu^{(0)}_{1,2} + \mu^{(0)}_{2,1}) - 2\mu^{(0)}_{2,2}) \\
&= \frac{L-1}{L} \left(\mu^{(0)}_{1,2} + \mu^{(0)}_{2,1} - 2\mu^{(0)}_{2,2} + \mathbb{E}[\Delta^k_{A^0, A_{2,2}} | \xi_3] \Pr[\xi_3]\right)
\end{aligned}$$

Solving for L, we obtain the following condition:

$$L \geq 1 + \frac{2\varepsilon}{\mu^{(0)}_{1,2} + \mu^{(0)}_{2,1} - 2\mu^{(0)}_{2,2} + \mathbb{E}[\Delta^k_{A^0, A_{2,2}} | \xi_3] \Pr[\xi_3] - 2\varepsilon}$$

To ensure that this lower bound is not vacuous, we choose ε small enough such that the denominator is positive. □

D.I.2 Accelerated Exploration Proofs

First, we state the following theorem from Jun et al. (2019) on finite sample error for low-rank bilinear bandits.

Theorem D.I.1 ((Jun et al., 2019)). *There exists a constant C such that for*

$$n_{\mathcal{P}} \geq C \cdot \sigma^2 (g_0^2 + g_1^2) \cdot \frac{\kappa^6}{d\sigma^2_{\min}(K^*)} r(r + \log(d))$$

with probability at least $1 - 2/d^{3/2}$, *we have*

$$\left\|\hat{\Gamma} - \Gamma^*\right\|_F \leq C_1 \kappa^2 \sigma \sqrt{\frac{dr}{n_{\mathcal{P}}}} \tag{D.7}$$

where C_1 *is an absolute constant,* Γ^* *is the mean reward matrix defined by* $\Gamma^*_{i,j} = \mu_{i,j}$ *with rank* r, $\hat{\Gamma}$ *is the noisy estimate of* K^* *using* $n_{\mathcal{P}}$ *samples of each atom,* $\kappa = \sigma_{\max}(\Gamma^*)/\sigma_{\min}(\Gamma^*)$. *Let* $\Gamma^* = URV^\top$ *be the SVD of* Γ^*. *Let* (g_0, g_1) *are the smallest values such that for all* $i, j \in [d]$

$$\sum_{s=1}^{r} U_{is}^2 \leq g_0 r/d \quad \sum_{s=1}^{r} V_{js}^2 \leq g_0 r/d$$
$$|\sum_{s=1}^{r} U_{is}(\sigma_s(\Gamma^*)/\sigma_{\max}(\Gamma^*))V_{js}| \leq g_1 \sqrt{\frac{r}{d^2}}$$

Proof of Theorem 6.4.3. We begin by stating the formal theorem for accelerated exploration:

Theorem D.I.2 (Accelerated Exploration BIC)**.** *Given* $n_{\mathcal{P}}$ *samples of all atoms where*

$$n_{\mathcal{P}} \geq \frac{N^6 C_1^2 \kappa^4 \sigma^2 dr}{4C_0^2(\Delta^{(t)}_{(b^{(t)})} - \varepsilon N^2)^2}$$

the inverse gap weighting recommendation policy is two-sided ε*-BIC. The total regret during this stage is* $O(N\sqrt{dT\log(T)})$, *which asymptotically matches the optimal regret of combinatorial semi-bandits.*

Proof. We want to show that given a recommendation for any action $A \in \mathcal{A}$, the learner would not switch to some other action A'. Formally, we want to ensure the following condition:

$$\mathbb{E}[\mu_A - \mu_{A'}|\mathrm{rec}^{(t)} = A]\Pr[\mathrm{rec}^{(t)} = A] \geq \varepsilon$$

Let $\Delta^{(t)}_{A,A'} = \mathbb{E}[\mu_A - \mu_{A'}|\mathcal{S}]$ denote the posterior gap between action A and A' given the data collected during the warm-start stage. Let $\Delta^{(t)}_A = \min_{A' \neq A} \Delta^{(t)}_{A,A'}$ denote the minimal posterior gap between action A and any other action. Then, when action A is recommended at time-step t, it means either 1) A is indeed the posterior best action at this time-step and $\Delta_A^{(t)} > 0$ or 2) A is not the posterior best action and $\Delta_A^{(t)} \geq 0$. We have

$$
\begin{aligned}
&\mathbb{E}[\Delta_A^{(t)} | \mathrm{rec}^{(t)} = A] \Pr[\mathrm{rec}^{(t)} = A] \\
&= \mathbb{E}[\mathbb{E}[\mu_A - \max_{A' \in \mathcal{A}} \mu_{A'} | \mathcal{S}] | \mathrm{rec}^{(t)} = A] \Pr[A^{(t)} = A] \\
&= \mathbb{E}[\mathbb{E}[\mu_A | \mathcal{S}] - \max_{A' \in \mathcal{A}} \mathbb{E}[\mu_{A'} | \mathcal{S}] | A^{(t)} = b^{(t)}] \cdot \Pr[A^{(t)} = b^{(t)}] \\
&+ \mathbb{E}[\mathbb{E}[\mu_A | \mathcal{S}] - \max_{A' \in \mathcal{A}} \mathbb{E}[\mu_{A'} | \mathcal{S}] | A^{(t)} \neq b^{(t)}] \cdot \Pr[A^{(t)} \neq b^{(t)}]
\end{aligned}
$$

We proceed to analyze the lower bound for each case separately.

Exploitation: Recommended action $A^{(t)} = b^{(t)}$. By construction, the posterior best action is recommended with probability $p_{b^{(t)}}^{(t)} = 1 - \sum_{A \neq b^{(t)}} \frac{1}{N^2 + \gamma(\hat{r}_{b^{(t)}}^{(t)} - \hat{r}_A^{(t)})}$. Since $\gamma > 0$, we observe that the probability of recommending any other action $A \neq b^{(t)}$ is at most $1/N^2$. Hence, we have $p_{b^{(t)}}^{(t)} \geq 1/N^2$. Therefore, we can write the reward gap in this case as:

$$
\mathbb{E}[\mathbb{E}[\mu_A | \mathcal{S}] - \max_{A' \in \mathcal{A}} \mathbb{E}[\mu_{A'} | \mathcal{S}] | A^{(t)} = b^{(t)}] \cdot \Pr[A^{(t)} = b^{(t)}] \geq \frac{1}{N^2} \cdot \Delta_{b^{(t)}}^{(t)}
$$

Exploration: Recommended action $A^{(t)} \neq b^{(t)}$. The reward gap in this case can be written as follows.

$$
\begin{aligned}
&\mathbb{E}[\mathbb{E}[\mu_A|\mathcal{S}] - \max_{A' \in \mathcal{A}} \mathbb{E}[\mu_{A'}|\mathcal{S}] | A^{(t)} \neq b^{(t)}] \cdot \Pr[A^{(t)} \neq b^{(t)}] \\
&= \sum_{A \neq b^{(t)}} p_A^{(t)} \left(\mathbb{E}[\mu_A - \mu_{b^{(t)}}|\mathcal{S}]\right) \\
&= \sum_{A \neq b^{(t)}} \frac{1}{N^2 \gamma(\hat{r}_{b^{(t)}}^{(t)} - \hat{r}_A^{(t)})} \left(\mathbb{E}[\mu_{b^{(t)}} - \mu_A|\mathcal{S}]\right) \\
&= -\,\mathbb{E}\left[\sum_{A \neq b^{(t)}} \frac{1}{\gamma} \cdot \frac{\gamma(\hat{r}_{b^{(t)}}^{(t)} - \hat{r}_A^{(t)})}{N^2 + \gamma(\hat{r}_{b^{(t)}}^{(t)} - \hat{r}_A^{(t)})}\right] \\
&< -\frac{N^2 - 1}{\gamma} \qquad \left(\text{since } \frac{\gamma(\hat{r}_{b^{(t)}}^{(t)} - \hat{r}_A^{(t)})}{N^2 + \gamma(\hat{r}_{b^{(t)}}^{(t)} - \hat{r}_A^{(t)})} < 1\right) \\
&< -\frac{N^2}{\gamma}
\end{aligned}
$$

Hence, for the BIC condition to hold, it suffices to show that

$$
\begin{aligned}
&\mathbb{E}[\Delta_A^{(t)}|\mathrm{rec}^{(t)} = A] \Pr[\mathrm{rec}^{(t)} = A] \geq \varepsilon \\
\iff\ & \frac{\Delta_{b^{(t)}}^{(t)}}{N^2} - \frac{N^2}{\gamma} \geq \varepsilon \\
\iff\ & \gamma \geq \frac{N^4}{\Delta_{b^{(t)}}^{(t)} - \varepsilon N^2}
\end{aligned}
$$

By definition, we have $\gamma = C_0 \cdot N\sqrt{1/\phi^{(t)}}$. Then, combining with the condition above, we derive the requirement for the minimum prediction error at time-step t as:

$$
\begin{aligned}
&\gamma \geq \frac{N^4}{\Delta_{b^{(t)}}^{(t)} - \varepsilon N^2} \\
\iff\ & C_0 \cdot \frac{N}{\sqrt{\phi^{(t)}}} \geq \frac{N^4}{\Delta_{b^{(t)}}^{(t)} - \varepsilon N^2} \\
\iff\ & \phi^{(t)} \leq \frac{C_0^2 \cdot (\Delta_{b^{(t)}}^{(t)} - \varepsilon N^2)^2}{N^6}
\end{aligned}
$$

Then, we use the theoretical guarantee of Theorem 2 in Jun et al. (2019) for bilinear bandits: Hence, it suffices to have

$$\begin{aligned}
\phi^{(t)} &\leq \frac{C_0^2 \cdot (\Delta_{b^{(t)}}^{(t)} - \varepsilon N^2)^2}{N^6} \\
\frac{C_1^2 \kappa^4 \sigma^2 dr}{4 n_{\mathcal{P}}} &\leq \frac{C_0^2 \cdot (\Delta_{b^{(t)}}^{(t)} - \varepsilon N^2)^2}{N^6} \\
n_{\mathcal{P}} &\geq \frac{N^6 C_1^2 \kappa^4 \sigma^2 dr}{4 C_0^2 (\Delta_{(b^{(t)})}^{(t)} - \varepsilon N^2)^2}
\end{aligned}$$

Regret Analysis Following the analysis of Foster and Rakhlin (2020), with probability at least $1-\delta$, the regret upper bound of the inverse gap weighting algorithm is $O(N\sqrt{T \cdot \phi^{(T)} \log(2/\delta)})$. □

D.II General setting proofs

D.II.1 Initial exploration with Hidden Exploration for K types of agents and arms

In this section, we provide the theoretical results when there are K types of agents and arms visiting the platform at each time-step t. For simplicity, we assume that the initial preference ordering of type i agents is $\mu_{i,1}^0 \geq \mu_{i,2}^0 \geq \cdots \geq \mu_{i,K}^0$. Similarly, we assume that the initial preference ordering of type j agents is $\mu_{1,j}^0 \geq \cdots \geq \mu_{K,j}^0$. Our setup is similar to the binary case with two types of agents and arms as described in Section 3.2, with the only difference being that each agent can choose one of K types of arms (and respectively, each arm can choose one of K agents) at each time-step. We assume the same non-degeneracy condition Assumption 6.4.1 and the platform's knowledge of the prior-dependent constants in this condition. Since different types of agents and arms have different preference orderings over their matches, we design a two-sided BIC recommendation policy (Line 1) for each type of agent and arm (in contrast to Line 1, which is two-sided BIC for all agents and arms of both types). Note that, with K different types of agents and arms, the platform can run this two-sided BIC recommendation policy at most K^2 times to ensure that all types of agents and arms will comply with the recommendation and match with less-preferred alternatives. This is because after agents of type 1 and arms of type 1 have successfully matched with all alternative options, *i.e.,* agents of type 1 have matched with arms of type 2, then arms of type 3 and so on, and arms of type 1 have matched with agents of type 2, then agents of type 3, and so on, then with high probability, we can form an accurate estimate for a row/column of the latent matrix Σ.

Given a pair of agent type and arm type (i, j), the goal of Line 1 is to incentivize sufficient exploration such that the platform can collect sufficiently many samples of all matches (i, ℓ) and (ℓ, j), for all type $\ell \in [k]$. As was the case in Line 1, the 'warm-start' algorithm will still be split into two stages, and all agents and arms are not given any recommendation in the first stage. The main difference compared to Line 1 is how the platform chooses an 'exploit' match in the second stage to ensure the two-sided BIC condition.

When there are only two possible matchings, for example, $\{(x_1, a_1), (x_2, a_2)\}$ and $\{(x_1, a_2), (x_2, a_1)\}$, then the platform can compare $\widehat{r}^{(k)}_{1,1}$, which is the estimate of the average reward from matches between agents of type 1 and arms of type 1, to the lower bound on the prior-mean reward of an alternative match, *e.g.,* (x_1, a_2) or (x_2, a_1). However, a more complicated procedure is required to ensure the two-sided BIC condition when there are more than two possible matchings.

At a high level, in Line 1, the platform will set the 'exploit' matching as follows. For agents of type i and arms of type j, Line 1 sets a matching (x_i, a_ℓ) or (x_ℓ, a_j) as the 'exploit' matching if and only if the sample average of *all* matching $\{(x_i, a_\nu)$ for $\nu \in \{1, \cdots, \ell\}\}$ (and vice versa, $\{x_\nu, a_j\}$ for $\nu \in \{1, \cdots, \ell\}$) are both (1) larger than the sample average of the matching (x_1, a_1) by some constant gap C, and (2) less than the prior-mean reward $\mu^{(0)}_{i,\ell}$ (or $\mu^{(0)}_{j,\ell}$) by C. If no matching satisfies both of these conditions, then the prior-best matching (x_1, a_1) is chosen to be the 'exploit' matching. We provide the formal theoretical guarantee in the following theorem.

Theorem D.II.1. *Under Assumption 6.4.1, if the number of initial samples k and the batch size L is chosen to be large enough, such that*

$$L \geq 1 + \frac{2 + 2\varepsilon}{\tau_{\mathcal{P}} \cdot \rho_{\mathcal{P}} - 2\varepsilon} \tag{D.8}$$

then Line 1 is two-sided BIC for all agents of type i and arms of type i.

Algorithm 13: Initial sampling: Hidden Exploration for agents of type i and arms of type j

Input: Batch size $L \in \mathbb{N}$, target number of samples $k \in \mathbb{N}$, gap $C \in (0, 1)$.

Initialize dataset $\mathcal{S} = \emptyset$;
The first k learners choose $A = \{(x_1, a_1)\}$ without recommendations.
Let $\widehat{r}^k_{1,1}$ be the sample average of these rewards. Add these k samples to $\mathcal{S}$;
for each phase $\psi = 1$ to K^2 **do**
  if $\psi \leq K$ **then**
    $A^{(\psi)}_{\text{cand}} \ni \{(x_i, a_\psi)\}$;
  else
    $A^{(\psi)}_{\text{cand}} \ni \{(x_\psi, a_j)\}$
  end if
1 **if** $\widehat{r}^k_{i,1} + C \leq \widehat{r}^k_{i,g} \leq \mu^{(0)}_{A^{\psi}_{\text{cand}}} - C$ for all types $g \in \{2, \cdots, \psi\}$ and $\widehat{r}^k_{1,j} + C \leq \widehat{r}^k_{h,j} \leq \mu^{(0)}_{A^{\psi}_{\text{cand}}} - C$ for all types $h \in \{2, \cdots, \psi\}$ **then**
    'Exploit' action $A^* = A^{\psi}_{\text{cand}}$.
  else
    'Exploit' action $A^* = \{(x_1, a_1)\}$.
  end if
  From set P of $L \cdot k$ learners, pick set Q of k learners uniformly at random;
  Every learner $p \in P - Q$ is recommended the 'exploit' action A^*;
  Every learner $p \in Q$ is recommended action A_{cand}.
  Add the reward from all $p \in Q$ to $\mathcal{S}$.
end for

Proof. According to our recommendation policy, a learner is either recommended the matching $\{(x_1, a_1)\}$ or some matching $\{(x_\ell, a_\nu)\}$. We will prove that, in either case, the learner will comply with the platform's recommendation.

Consider the event $\widehat{\xi}^{(t)}_{(x_\ell, a_\nu)}$ that the action $\{(x_\ell, a_\nu)\}$ is the exploit intervention for learner t. Formally, we have:

$$\widehat{\xi}^{(t)}_{(x_\ell, a_\nu)} = \Big\{ \widehat{r}^k_{1,1} + C \leq \widehat{r}^k_{1,\nu} \leq \mu^{(0)}_{A^{\psi}_{\text{cand}}} - C \\ \text{and } \widehat{r}^k_{1,1} + C \leq \widehat{r}^k_{\nu,1} \leq \mu^{(0)}_{A^{\psi}_{\text{cand}}} - C \text{ for all types } \nu \in \{2, \cdots, \psi\} \Big\} \tag{D.9}$$

Recommended action is $A^{(t)} = \{(x_\ell, a_\nu)\}$. When the learner at time-step t is recommended action $A^{(t)}$, we argue that the learner will not switch to any other action. Due to our simplifying assumption on the prior-mean reward, any action (x_h, a_g) for $h > \psi, g > \psi$ has had no samples collected in previous time-steps and $\mu^{(0)}_{h,g} \leq \mu^{(0)}_{\ell,\nu}$. Hence, we only need to focus on the cases when $h < \ell, g < \nu$. For the recommendation policy to be two-sided BIC, we need to show that:

$$\mathbb{E}[\mu_{\ell,\nu} - \mu_{\ell,g} | (\ell,\nu) \in A^{(t)}] \Pr[(\ell,\nu) \in A^{(t)}] \geq \varepsilon$$
$$\mathbb{E}[\mu_{\ell,\nu} - \mu_{h,\nu} | (\ell,\nu) \in A^{(t)}] \Pr[(\ell,\nu) \in A^{(t)}] \geq \varepsilon$$

We combine these conditions and prove the following:

$$\mathbb{E}[2\mu_{\ell,\nu} - \mu_{\ell,g} - \mu_{h,\nu} | (\ell,\nu) \in A^{(t)}] \geq 2\varepsilon$$

Let $A' = \{(\ell,h),(h,\nu)\}$ and $A = \{(\ell,\nu),(\ell,\nu)\}$ denote the dummy action whose reward is twice the reward of choosing atom (ℓ,ν). Then, our goal is to show that $E[\mu_A - \mu_{A'} | A^{(t)} = A] \Pr[A^{(t)} = A] \geq 2\varepsilon$. There are two possible disjoint events under which learner t is recommended the matching (x_ℓ, a_ν): either (x_ℓ, a_ν) is determined to be part of an 'exploit' action or the learner t is chosen to be in the 'explore' group. Since being chosen to be in the 'explore' group does not imply any information about the reward of this atom (or any other atoms in the atom set), we can derive that, conditional on being chosen as an 'explore' learner, the expected gain for learner t to switch to another atom (h,g) is simply $\mu^{(0)}_A - \mu^{(0)}_{A'}$. On the other hand, if atom (ℓ,ν) is in the 'exploit' action, then we know that event $\xi^{(t)}_{(\ell,\nu)}$ has happened. Hence, we can rewrite the left-hand side of the two-sided BIC condition as:

$$\begin{aligned}
&\mathbb{E}[\mu_A - \mu_{A'} | A^{(t)} = A] \Pr[A^{(t)} = A] \\
&= \left(1 - \frac{1}{L}\right) \mathbb{E}[\mu_A - \mu_{A'} | \widehat{\xi}^{(t)}_{(\ell,\nu)}] \Pr[\widehat{\xi}^{(t)}_{(\ell,\nu)}] + \frac{1}{L}(\mu^{(0)}_A - \mu^{(0)}_{A'})
\end{aligned}$$

Rearranging the terms, we have the following lower bound on the phase length L for the algorithm to be two-sided BIC:

$$L \geq 1 + \frac{\mu_{A'}^{(0)} - \mu_A^{(0)} + 2\varepsilon}{\mathbb{E}[\mu_A - \mu_{A'} | \xi_{(\ell,\nu)}^{(t)}] \Pr[\xi_{(\ell,\nu)}^{(t)}] - 2\varepsilon}$$

The expression above can be simplified using the definitions of $\tau_{\mathcal{P}}, \rho_{\mathcal{P}}$ and observing that $\mu_{A'}^{(0)} - \mu_A^{(0)} \leq 2$ to get

$$L \geq 1 + \frac{1 + 2\varepsilon}{\tau_{\mathcal{P}} \cdot \rho_{\mathcal{P}} - 2\varepsilon}$$

Recommended action is $A^{(t)} = \{(x_1, a_1)\}$. When the learner is recommended action $\{(x_1, a_1)\}$, then they know that they are not in the 'explore' group. Hence, the event $\xi_{(x_\ell,a_\nu)}^{(t)}$ did not happen, and the two-sided BIC condition can be written as: for any action $(g, h) \in [K] \times [K]$,

$$\mathbb{E}[\mu_{1,1} - \mu_{g,h} | \neg\xi_{(x_h,a_g)}^{(t)}] \Pr[\neg\xi_{(x_h,a_g)}^{(t)}] \geq \varepsilon$$

It suffices to only consider atom $(h, g) \in [\psi] \times [\psi]$. We have:

$$\begin{aligned}
\mathbb{E}[\mu_{1,1} - \mu_{h,g} | \neg\widehat{\xi}_{(x_h,a_g)}^{(t)}] \Pr[\neg\widehat{\xi}_{(x_h,a_g)}^{(t)}] &= \mathbb{E}[\mu_{1,1} - \mu_{g,h}] - \mathbb{E}[\mu_{1,1} - \mu_{h,g} | \widehat{\xi}_{(x_h,a_g)}^{(t)}] \Pr[\widehat{\xi}_{(x_h,a_g)}^{(t)}] \\
&= \mu_{1,1}^{(0)} - \mu_{h,g}^{(0)} + \mathbb{E}[\mu_{g,h} - \mu_{1,1} | \widehat{\xi}_{(x_h,a_g)}^{(t)}] \Pr[\widehat{\xi}_{(x_h,a_g)}^{(t)}]
\end{aligned}$$

By definition, we have $\mu_{1,1}^{(0)} - \mu_{g,h}^{(0)}]$. Hence, it suffices to show that for any action $(g, h) \in [\psi] \times [\psi]$,

$$\mathbb{E}[\mu_{g,h} - \mu_{1,1} | \widehat{\xi}_{(x_h,a_g)}^{(t)}] \Pr[\widehat{\xi}_{(x_h,a_g)}^{(t)}] \geq \varepsilon$$

Let $\xi_{g,h}$ be the 'exploit' event defined on the true mean reward. Formally,

$$\xi_{g,h} = \Big\{ \mu_{1,1} + C \leq \mu_{i,\nu} \leq \mu_{\ell,\nu} - C$$

$$\text{and } \mu_{1,1} + C \leq \mu_{\nu,j} \leq \mu_{\ell,\nu} - C \text{ for all types } h \in \{2, \cdots, \ell\}, g \in \{2, \cdots, \nu\} \Big\}$$

We can rewrite the two-sided BIC condition as:

$$\begin{aligned}
&\mathbb{E}[\mu_{g,h} - \mu_{1,1} | \widehat{\xi}^{(t)}_{(x_h,a_g)}] \Pr[\widehat{\xi}^{(t)}_{(x_h,a_g)}] \\
&= \mathbb{E}[\mu_{g,h} - \mu_{1,1} | \widehat{\xi}^{(t)}_{(x_h,a_g)}, \xi_{g,h}] \Pr[\widehat{\xi}^{(t)}_{(x_h,a_g)}, \xi_{g,h}] \\
&\quad + \mathbb{E}[\mu_{g,h} - \mu_{1,1} | \widehat{\xi}^{(t)}_{(x_h,a_g)}, \neg\xi_{g,h}] \Pr[\widehat{\xi}^{(t)}_{(x_h,a_g)}, \neg\xi_{g,h}] \\
&\geq \mathbb{E}[\mu_{g,h} - \mu_{1,1} | \widehat{\xi}^{(t)}_{(x_h,a_g)}, \xi_{g,h}] \Pr[\widehat{\xi}^{(t)}_{(x_h,a_g)}, \xi_{g,h}] - 2\delta
\end{aligned}$$

where $\delta = 2/d^{3/2}$ is the failure probability that the estimated reward is not close to the true mean reward of each atom (Theorem D.I.1).

Observe that when event $\widehat{\xi}^{(t)}_{(x_h,a_g)}$ happens, then we know that $\widehat{r}^{(t)}_{g,h} \geq \widehat{r}^{(t)}_{1,1} + C$. Furthermore, when event $\xi_{h,g}$ happens, we know that $\mu_{g,h} \geq \widehat{r}^{(t)}_{h,g} - C_1\kappa^2\sigma\sqrt{\frac{dr}{n_{\mathcal{P}}}}$. Hence, when these two events happen simultaneously, we have $\mu_{g,h} \geq \mu_{1,1} + C - 2C_1\kappa^2\sigma\sqrt{\frac{dr}{n_{\mathcal{P}}}}$. Choosing C to be large enough such that the bound is not vacuous, we can conclude the proof. □

D.II.2 Accelerated Exploration for K types of agents and arms

From Foster and Rakhlin (2020), we know that the recommendation policy described in Section 6.4.2 can be adapted to an increasing number of actions at each time-step. That is, the platform will still estimate the reward of each atom using the 'warm-start' samples collected by Line 1 and provide recommendations according to Equation (6.7). Hence, when there are k types of agents and arms visiting the platform at each time-step, the platform can still leverage the inverse-gap weighting technique to provide recommendations to each learner. The regret guarantee, in this case, is the same as in the binary setting with only two types of agents and arms (Foster and Rakhlin, 2020).

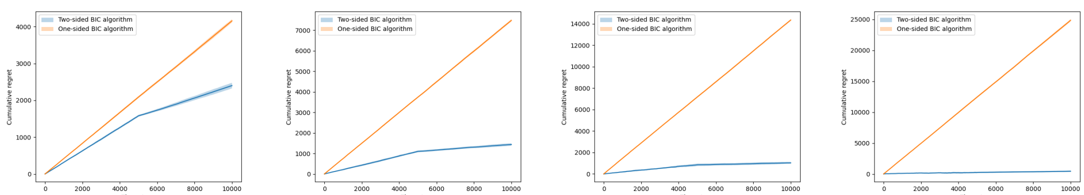
D.III Experiment Detail

In this section, we provide additional experimental details and analysis of Section 4.7 that were previously omitted from the main body.

For the benchmark one-sided BIC algorithm, we consider the incentivizing exploration algorithm for only the agents from Mansour et al. (2015b). Particularly, for Figure 6.2

and 6.3, we consider Algorithm 1 from Mansour et al. (2015b), where the goal is to incentivize the agents to choose the lesser arm (arm 2). For the following experiments with more types of agents and arms, we consider the detailed-free algorithm 4 in Mansour et al. (2015b).

In this experiment, we test the validity of our theoretical algorithms in Section D.II when there are K types of agents and arms visiting the platform at each time-step. In Figure D.1, we vary the number of types of agents and arms $K \in \{4, 8, 16, 32\}$, while keeping all other parameters of the problem instant constant. Each experiment is repeated 10 times, and we report the regret and standard deviation at each time-step.



(a) $K = 4$ (b) $K = 8$ (c) $K = 16$ (d) $K = 32$

Figure D.1: Cumulative regret using our proposed two-sided BIC algorithm (Line 1 and Line 1) in blue and the standard one-sided BIC algorithm in orange. We provide the results with an increasing number of types of agents and arms K at each time-step: $K \in \{4, 8, 16, 32\}$. Results are averaged over 10 runs, with the shaded region representing one standard deviation. In all experiments, we observe the cumulative regret of our proposed algorithm is significantly lower than that of a one-sided BIC algorithm that does not consider incentives from the arms. As the number of types of agents and arms at each time-step increases, our algorithm incurs more regret as the platform needs more exploration before determining the optimal matching.

Appendix E

Appendix for Strategic Instrumental Variable Regression

E.I Parameter estimation in the causal setting

E.I.1 Ordinary least squares is not consistent

The least-squares estimate of $\boldsymbol{\theta}^*$ is given as

$$\widehat{\boldsymbol{\theta}}_{LS} = \left(\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t^\top\right)^{-1} \sum_{t=1}^{T} \mathbf{x}_t y_t.$$

However, $\widehat{\boldsymbol{\theta}}_{LS}$ is not a consistent estimator of $\boldsymbol{\theta}^*$. To see this, let us plug in our expression for y_t into our expression for $\widehat{\boldsymbol{\theta}}_{LS}$. We get

$$\widehat{\boldsymbol{\theta}}_{LS} = \left(\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t^\top\right)^{-1} \sum_{t=1}^{T} \mathbf{x}_t (\mathbf{x}_t^\top \boldsymbol{\theta}^* + o_t)$$

After distributing terms and simplifying, we get

$$\widehat{\boldsymbol{\theta}}_{LS} = \boldsymbol{\theta}^* + \left(\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t^\top\right)^{-1} \sum_{t=1}^{T} \mathbf{x}_t o_t.$$

$\mathbf{x}_t$ and o_t are not independent due to their shared dependence on the agent's private type u_t. Because of this, $\left(\sum_{t=1}^T \mathbf{x}_t\mathbf{x}_t^\top\right)^{-1}\sum_{t=1}^T \mathbf{x}_t o_t$ will generally not equal $\mathbf{0}_m$, even as the number of data points (agents) grows large. To see this, recall that $\mathbf{x}_t = \mathbf{b}_t + \mathcal{E}_t\mathbf{a}_t$, so $\sum_{t=1}^T \mathbf{x}_t o_t = \sum_{t=1}^T (\mathbf{b}_t + \mathcal{E}_t\mathcal{E}_t^\top\boldsymbol{\theta}_t)o_t$. o_t and $\mathbf{b}_t$ are both determined by the agent's private type. Take the example where $\mathbf{b}_t = [o_t, 0, \ldots, 0]^\top$. In this setting, $\sum_{t=1}^T \mathbf{b}_t o_t = [o_t^2, 0, \ldots, 0]^\top$, which will always be greater than 0 unless $o_t = 0$, $\forall t$.

E.I.2 2SLS derivations

Define $\widetilde{\boldsymbol{\theta}}_t = \begin{bmatrix}\boldsymbol{\theta}_t \\ 1\end{bmatrix}$. $\mathbf{x}_t$ can now be written as $\mathbf{x}_t = \begin{bmatrix}\mathcal{E}_t\mathcal{E}_t^\top & \mathbf{b}_t\end{bmatrix}\begin{bmatrix}\boldsymbol{\theta}_t \\ 1\end{bmatrix}$.

Lemma E.I.1. *Using OLS, we can estimate* $\begin{bmatrix}\mathbb{E}[\mathcal{E}_t\mathcal{E}_t^\top] \\ \mathbb{E}[\mathbf{b}_t]^\top\end{bmatrix}$ *as*

$$\begin{aligned}
\begin{bmatrix}\widehat{\Omega} \\ \bar{\mathbf{b}}^\top\end{bmatrix} &= \left(\sum_{t=1}^T \widetilde{\boldsymbol{\theta}}_t\widetilde{\boldsymbol{\theta}}_t^\top\right)^{-1}\sum_{t=1}^T \widetilde{\boldsymbol{\theta}}_t\mathbf{x}_t^\top \\
&= \left(\sum_{t=1}^T \widetilde{\boldsymbol{\theta}}_t\widetilde{\boldsymbol{\theta}}_t^\top\right)^{-1}\begin{bmatrix}\sum_{t=1}^T \boldsymbol{\theta}_t\mathbf{x}_t^\top \\ \sum_{t=1}^T \mathbf{x}_t^\top\end{bmatrix},
\end{aligned}$$

where $\widehat{\Omega} = \left(\sum_{t=1}^T \boldsymbol{\theta}_t\boldsymbol{\theta}_t^\top\right)^{-1}\sum_{t=1}^T \boldsymbol{\theta}_t(\mathbf{x}_t - \bar{\mathbf{b}})^\top$.

Proof. In order to calculate $\widehat{\Omega}$, we will make use of the following fact:

Fact E.I.2 (Block Matrix Inversion (Bernstein (2009)))**.** *If a matrix P is partitioned into four blocks, it can be inverted blockwise as follows:*

$$\begin{aligned}
P &= \begin{bmatrix}A & B \\ C & D\end{bmatrix}^{-1} \\
&= \begin{bmatrix}A^{-1} + A^{-1}BE^{-1}CA^{-1} & -A^{-1}BE^{-1} \\ -E^{-1}CA^{-1} & E^{-1}\end{bmatrix},
\end{aligned}$$

where A and D are square matrices of arbitrary size, and B and C are conformable for partitioning. Furthermore, A and the Schur complement of A in P ($E = D - CA^{-1}B$) must be invertible.

Let $A = \sum_{t=1}^T \boldsymbol{\theta}_t\boldsymbol{\theta}_t^\top$, $B = \sum_{t=1}^T \boldsymbol{\theta}_t$, $C = \sum_{t=1}^T \boldsymbol{\theta}_t^\top$, and $D = \sum_{t=1}^T 1 = T$. Note that A is invertible by assumption and E is a scalar, so is trivially invertible unless $CA^{-1}B = T$.

Using this formulation, observe that

$$\bar{\mathbf{b}}^\top = -E^{-1}CA^{-1}\sum_{t=1}^{T}\boldsymbol{\theta}_t\mathbf{x}_t^\top + E^{-1}\sum_{t=1}^{T}\mathbf{x}_t^\top$$

and

$$\begin{aligned}\widehat{\Omega} = A^{-1}\sum_{t=1}^{T}\boldsymbol{\theta}_t\mathbf{x}_t^\top + A^{-1}BE^{-1}CA^{-1}\sum_{t=1}^{T}\boldsymbol{\theta}_t\mathbf{x}_t^\top \\ - A^{-1}BE^{-1}\sum_{t=1}^{T}\mathbf{x}_t^\top\end{aligned}$$

Rearranging terms, we see that $\widehat{\boldsymbol{\lambda}}$ can be written as

$$\begin{aligned}\widehat{\Omega} &= A^{-1}\sum_{t=1}^{T}\boldsymbol{\theta}_t\mathbf{x}_t^\top \\ &+ A^{-1}B(E^{-1}CA^{-1}\sum_{t=1}^{T}\boldsymbol{\theta}_t\mathbf{x}_t^\top - E^{-1}\sum_{t=1}^{T}\mathbf{x}_t^\top) \\ &= A^{-1}\sum_{t=1}^{T}\boldsymbol{\theta}_t\mathbf{x}_t^\top - A^{-1}B\bar{\mathbf{b}}^\top\end{aligned}$$

Finally, plugging in for A and B, we see that

$$\begin{aligned}\widehat{\Omega} &= \left(\sum_{t=1}^{T}\boldsymbol{\theta}_t\boldsymbol{\theta}_t^\top\right)^{-1}\sum_{t=1}^{T}\boldsymbol{\theta}_t\mathbf{x}_t^\top \\ &- \left(\sum_{t=1}^{T}\boldsymbol{\theta}_t\boldsymbol{\theta}_t^\top\right)^{-1}\sum_{t=1}^{T}\boldsymbol{\theta}_t\bar{\mathbf{b}}^\top \\ &= \left(\sum_{t=1}^{T}\boldsymbol{\theta}_t\boldsymbol{\theta}_t^\top\right)^{-1}\sum_{t=1}^{T}\boldsymbol{\theta}_t(\mathbf{x}_t - \bar{z})^\top\end{aligned}$$

□

Similarly, we can write y_t as $y_t = \begin{bmatrix}\boldsymbol{\theta}_t^\top & 1\end{bmatrix}\begin{bmatrix}\mathcal{E}_t\mathcal{E}_t^\top\boldsymbol{\theta}^* \\ o_t + \mathbf{b}_t^\top\boldsymbol{\theta}^*\end{bmatrix}$.

Lemma E.I.3. *Using OLS, we can estimate* $\begin{bmatrix} \mathbb{E}[\mathcal{E}_t\mathcal{E}_t^\top]\boldsymbol{\theta}^* \\ \mathbb{E}[o_t] + \mathbb{E}[\mathbf{b}_t^\top]\boldsymbol{\theta}^* \end{bmatrix}$ *as*

$$\begin{bmatrix} \widehat{\boldsymbol{\lambda}} \\ \bar{o} + \bar{\mathbf{b}}^\top\boldsymbol{\theta}^* \end{bmatrix} = \left(\sum_{t=1}^{T} \widetilde{\boldsymbol{\theta}}_t\widetilde{\boldsymbol{\theta}}_t^\top\right)^{-1} \sum_{t=1}^{T} \widetilde{\boldsymbol{\theta}}_t y_t$$
$$= \left(\sum_{t=1}^{T} \widetilde{\boldsymbol{\theta}}_t\widetilde{\boldsymbol{\theta}}_t^\top\right)^{-1} \begin{bmatrix} \sum_{t=1}^{T} \boldsymbol{\theta}_t y_t^\top \\ \sum_{t=1}^{T} y_t^\top \end{bmatrix},$$

where $\widehat{\boldsymbol{\lambda}} = \left(\sum_{t=1}^{T} \boldsymbol{\theta}_t\boldsymbol{\theta}_t^\top\right)^{-1} \sum_{t=1}^{T} \boldsymbol{\theta}_t(y_t - \bar{o} - \bar{\mathbf{b}}^\top\boldsymbol{\theta}^*)$.

Proof. The proof follows similarly to the proof of the previous lemma. Let $A = \sum_{t=1}^{T} \boldsymbol{\theta}_t\boldsymbol{\theta}_t^\top$, $B = \sum_{t=1}^{T} \boldsymbol{\theta}_t$, $C = \sum_{t=1}^{T} \boldsymbol{\theta}_t^\top$, and $D = \sum_{t=1}^{T} 1 = T$. Note that A is invertible by assumption and E is a scalar, so is trivially invertible unless $CA^{-1}B = T$.

Using this formulation, observe that

$$\bar{o}^\top + \bar{z}^\top\boldsymbol{\theta}^* = -E^{-1}CA^{-1}\sum_{t=1}^{T} \boldsymbol{\theta}_t y_t + E^{-1}\sum_{t=1}^{T} y_t$$

and

$$\begin{aligned} \widehat{\boldsymbol{\lambda}} &= A^{-1}\sum_{t=1}^{T} \boldsymbol{\theta}_t y_t \\ &\quad + A^{-1}B\left(E^{-1}CA^{-1}\sum_{t=1}^{T} \boldsymbol{\theta}_t y_t - E^{-1}\sum_{t=1}^{T} y_t\right) \\ &= A^{-1}\sum_{t=1}^{T} \boldsymbol{\theta}_t y_t - A^{-1}B\left(\bar{o}^\top + \bar{z}^\top\boldsymbol{\theta}^*\right) \\ &= \left(\sum_{t=1}^{T} \boldsymbol{\theta}_t\boldsymbol{\theta}_t^\top\right)^{-1} \sum_{t=1}^{T} \boldsymbol{\theta}_t\left(y_t - \bar{o}^\top - \bar{z}^\top\boldsymbol{\theta}^*\right) \end{aligned}$$

□

Theorem E.I.4. *We can estimate* $\boldsymbol{\theta}^*$ *as*

$$\widehat{\boldsymbol{\theta}} = \widehat{\Omega}^{-1}\widehat{\boldsymbol{\lambda}} = \left(\sum_{t=1}^{T} \boldsymbol{\theta}_t(\mathbf{x}_t - \bar{\mathbf{b}})^\top\right)^{-1} \sum_{t=1}^{T} \boldsymbol{\theta}_t(y_t - \bar{o} - \bar{z}^\top\boldsymbol{\theta}^*)$$

Proof. This follows immediately from the previous two lemmas. □

E.I.3 2SLS is consistent

Consider the two-stage least squares (2SLS) estimate of $\boldsymbol{\theta}^*$,

$$\widehat{\boldsymbol{\theta}}_{IV} = \left(\sum_{t=1}^{T} \boldsymbol{\theta}_t(\mathbf{x}_t - \bar{\mathbf{b}})^\top\right)^{-1} \sum_{t=1}^{T} \boldsymbol{\theta}_t(y_t - \bar{o} - \bar{z}^\top \boldsymbol{\theta}^*)$$

Plugging in for y_t and simplifying, we get

$$\widehat{\boldsymbol{\theta}}_{IV} = \boldsymbol{\theta}^* + \left(\sum_{t=1}^{T} \boldsymbol{\theta}_t(\mathbf{x}_t - \bar{\mathbf{b}})^\top\right)^{-1} \sum_{t=1}^{T} \boldsymbol{\theta}_t(o_t - \bar{o})$$

To see that $\widehat{\boldsymbol{\theta}}_{IV}$ *is* a consistent estimator of $\boldsymbol{\theta}^*$, we show $\lim_{T\to\infty} \mathbb{E}\|\widehat{\boldsymbol{\theta}}_{IV} - \boldsymbol{\theta}^*\|_2^2 = 0$.

$$\mathbb{E}\|\widehat{\boldsymbol{\theta}}_{IV} - \boldsymbol{\theta}^*\|_2^2 = \mathbb{E}\left\|\left(\sum_{t=1}^{T} \boldsymbol{\theta}_t(\mathbf{x}_t - \bar{\mathbf{b}})^\top\right)^{-1} \sum_{t=1}^{T} \boldsymbol{\theta}_t(o_t - \bar{o})\right\|_2^2$$

$o_t - \bar{o}$ and $\boldsymbol{\theta}_t$ are uncorrelated, so $\sum_{t=1}^{T} \boldsymbol{\theta}_t(o_t - \bar{o})$ will go to zero as $T \to \infty$. On the other hand, $\sum_{t=1}^{T} \boldsymbol{\theta}_t(\mathbf{x}_t - \bar{\mathbf{b}})^\top$ will approach $T\mathbb{E}[\boldsymbol{\theta}_t(\mathbf{x}_t - \bar{\mathbf{b}})^\top]$. $\boldsymbol{\theta}_t$ and $\mathbf{x}_t - \bar{\mathbf{b}}$ *are* correlated, so $\mathbb{E}[\boldsymbol{\theta}_t(\mathbf{x}_t - \bar{\mathbf{b}})^\top] \neq \mathbf{0}$ in general.

E.II Causal parameter recovery derivations

E.II.1 Proof of Theorem 7.2.1

Recall that $\widehat{\boldsymbol{\theta}} = \left(\sum_{t=1}^{T} \boldsymbol{\theta}_t(\mathbf{x}_t - \bar{\mathbf{b}})^\top\right)^{-1} \sum_{t=1}^{T} \boldsymbol{\theta}_t(y_t - \bar{o} - \bar{\mathbf{b}}^\top \boldsymbol{\theta}^*)$ from Appendix E.I.2. Plugging this into $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2$, we get

$$\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right\|_2 =$$
$$\left\|\left(\sum_{t=1}^{T} \boldsymbol{\theta}_t(\mathbf{x}_t - \bar{\mathbf{b}})^\top\right)^{-1} \left(\sum_{t=1}^{T} \boldsymbol{\theta}_t(y_t - \bar{o} - \bar{\mathbf{b}}^\top \boldsymbol{\theta}^*)\right) - \boldsymbol{\theta}^*\right\|_2$$

Next, we substitute in our expression for y_t and simplify, obtaining

$$
\begin{aligned}
&\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 \\
&= \Bigg\| \left(\sum_{t=1}^{T} \boldsymbol{\theta}_t (\mathbf{x}_t - \bar{\mathbf{b}})^\top \right)^{-1} \\
&\left(\sum_{t=1}^{T} \boldsymbol{\theta}_t (\mathbf{x}_t^\top \boldsymbol{\theta}^* + o_t - \bar{o} - \bar{\mathbf{b}}^\top \boldsymbol{\theta}^*) \right) - \boldsymbol{\theta}^* \Bigg\|_2 \\
&= \Bigg\| \left(\sum_{t=1}^{T} \boldsymbol{\theta}_t (\mathbf{x}_t - \bar{\mathbf{b}})^\top \right)^{-1} \left(\sum_{t=1}^{T} \boldsymbol{\theta}_t (\mathbf{x}_t - \bar{\mathbf{b}})^\top \boldsymbol{\theta}^* \right) \\
&+ \left(\sum_{t=1}^{T} \boldsymbol{\theta}_t (\mathbf{x}_t - \bar{\mathbf{b}})^\top \right)^{-1} \left(\sum_{t=1}^{T} \boldsymbol{\theta}_t (o_t - \bar{o}) \right) - \boldsymbol{\theta}^* \Bigg\|_2 \\
&= \left\| \boldsymbol{\theta}^* + \left(\sum_{t=1}^{T} \boldsymbol{\theta}_t (\mathbf{x}_t - \bar{\mathbf{b}})^\top \right)^{-1} \left(\sum_{t=1}^{T} \boldsymbol{\theta}_t (o_t - \bar{o}) \right) - \boldsymbol{\theta}^* \right\|_2 \\
&= \left\| \left(\sum_{t=1}^{T} \boldsymbol{\theta}_t (\mathbf{x}_t - \bar{\mathbf{b}})^\top \right)^{-1} \left(\sum_{t=1}^{T} \boldsymbol{\theta}_t (o_t - \bar{o}) \right) \right\|_2 \\
&\leq \left\| \left(\sum_{t=1}^{T} \boldsymbol{\theta}_t (\mathbf{x}_t - \bar{\mathbf{b}})^\top \right)^{-1} \right\|_2 \left\| \sum_{t=1}^{T} \boldsymbol{\theta}_t (o_t - \bar{o}) \right\|_2 \\
&\leq \frac{\left\| \sum_{t=1}^{T} \boldsymbol{\theta}_t (o_t - \bar{o}) \right\|_2}{\sigma_{min} \left(\sum_{t=1}^{T} \boldsymbol{\theta}_t (\mathbf{x}_t - \bar{\mathbf{b}})^\top \right)}
\end{aligned}
$$

We now bound the numerator and denominator separately with high probability.

E.II.2 Bound on numerator

$$
\begin{aligned}
\left\| \sum_{t=1}^{T} \boldsymbol{\theta}_t (o_t - \bar{o}) \right\|_2 &= \left\| \sum_{t=1}^{T} \boldsymbol{\theta}_t (o_t - \mathbb{E}[o_t] + \mathbb{E}[o_t] - \bar{o}) \right\|_2 \\
&\leq \left\| \sum_{t=1}^{T} \boldsymbol{\theta}_t (o_t - \mathbb{E}[o_t]) \right\|_2 + \left\| \sum_{t=1}^{T} \boldsymbol{\theta}_t (\mathbb{E}[o_t] - \bar{o}) \right\|_2
\end{aligned}
$$

Bound on first term

$$\begin{aligned}&\left\|\sum_{t=1}^{T}\boldsymbol{\theta}_t(o_t-\mathbb{E}[o_t])\right\|_2\\&=\left(\sum_{j=1}^{m}\left(\sum_{t=1}^{T}\theta_{t,j}(o_t-\mathbb{E}[o_t])\right)^2\right)^{1/2}\end{aligned}$$

Since $(o_t-\mathbb{E}[o_t])$ is a zero-mean bounded random variable with variance parameter σ_g^2, the product $\theta_{t,j}(o_t-\mathbb{E}[o_t])$ will also be a zero-mean bounded random variable with variance at most $\beta^2\sigma_g^2$. In order to bound $\left(\sum_{j=1}^{m}\left(\sum_{t=1}^{T}\theta_{t,j}(o_t-\mathbb{E}[o_t])\right)^2\right)^{1/2}$ with high probability, we make use of the following lemma. Note that bounded random variables are sub-Gaussian random variables.

Lemma E.II.1 (High probability bound on the sum of unbounded sub-Gaussian random variables)**.** *Let* $x_t\sim subG(0,\sigma^2)$*. For any* $\delta\in(0,1)$*, with probability at least* $1-\delta$*,*

$$\left|\sum_{t=1}^{T}x_t\right|\le\sigma\sqrt{2T\log(1/\delta)}$$

Applying Lemma E.II.1 to $\left(\sum_{j=1}^{m}\left(\sum_{t=1}^{T}\theta_{t,j}(o_t-\mathbb{E}[o_t])\right)^2\right)^{1/2}$, we get

$$\begin{aligned}&\sqrt{\sum_{j=1}^{m}\left(\sum_{t=1}^{T}\theta_{t,j}(o_t-\mathbb{E}[o_t])\right)^2}\\&\le\sqrt{\sum_{j=1}^{m}\left(\beta\sigma_g\sqrt{2T\log(1/\delta_j)}\right)^2}\\&\le\sqrt{\sum_{j=1}^{m}\beta^2\sigma_g^2 2T\log(m/\delta)}&&\text{(by a union bound, where }\delta_j=\delta/m\text{ for all }j)\\&\le\beta\sigma_g\sqrt{2Tm\log(m/\delta)}\end{aligned}$$

with probability at least $1-\delta$.

Bound on second term

$$
\begin{aligned}
&\left\|\sum_{t=1}^{T}\boldsymbol{\theta}_t(\mathbb{E}[o_t]-\bar{o})\right\|_2\\
&=\left\|\sum_{t=1}^{T}\boldsymbol{\theta}_t\left(\mathbb{E}[o_t]-\frac{1}{T}\sum_{s=1}^{T}o_s\right)\right\|_2\\
&=\left\|\sum_{t=1}^{T}\boldsymbol{\theta}_t\frac{1}{T}\sum_{s=1}^{T}(\mathbb{E}[o_t]-o_s)\right\|_2\\
&=\left(\sum_{j=1}^{m}\left(\sum_{t=1}^{T}\theta_{t,j}\frac{1}{T}\sum_{s=1}^{T}(\mathbb{E}[o_t]-o_s)\right)^2\right)^{1/2}\\
&\leq\left(\sum_{j=1}^{m}\left(\sum_{t=1}^{T}|\theta_{t,j}|\frac{1}{T}\left|\sum_{s=1}^{T}\mathbb{E}[o_t]-o_s\right|\right)^2\right)^{1/2}
\end{aligned}
$$

After applying Lemma E.II.1, we get

$$
\begin{aligned}
&\left\|\sum_{t=1}^{T}\boldsymbol{\theta}_t(\mathbb{E}[o_t]-\bar{o})\right\|_2\\
&\leq\left(\sum_{j=1}^{m}\left(\sum_{t=1}^{T}|\theta_{t,j}|\frac{1}{T}\sigma_g\sqrt{2T\log(1/\delta_j)}\right)^2\right)^{1/2}\\
&\leq\left(\sum_{j=1}^{m}\left(\beta\sigma_g\sqrt{2T\log(1/\delta_j)}\right)^2\right)^{1/2}\\
&\leq\left(\sum_{j=1}^{m}\beta^2\sigma_g^2 2T\log(m/\delta)\right)^{1/2}\\
&\leq\beta\sigma_g\sqrt{2Tm\log(m/\delta)}
\end{aligned}
$$

with probability at least $1-\delta$

E.II.3 Proof of Corollary 7.2.2

Next let's bound the denominator. By plugging in the expression for $\mathbf{x}_t$, we see that

$$\begin{aligned}
&\sigma_{min}\left(\sum_{t=1}^{T}\boldsymbol{\theta}_t(\mathbf{x}_t-\bar{\mathbf{b}})^\top\right)\\
&=\sigma_{min}\left(\sum_{t=1}^{T}\boldsymbol{\theta}_t(\mathbf{b}_t-\bar{\mathbf{b}})^\top+\boldsymbol{\theta}_t\boldsymbol{\theta}_t^\top\mathcal{E}_t\mathcal{E}_t^\top\right)\\
&=\sigma_{min}\left(A+B\right),
\end{aligned}$$

where $A=\sum_{t=1}^{T}\boldsymbol{\theta}_t(\mathbf{b}_t-\bar{\mathbf{b}})^\top$ and $B=\sum_{t=1}^{T}\boldsymbol{\theta}_t\boldsymbol{\theta}_t^\top\mathcal{E}_t\mathcal{E}_t^\top$. By definition,

$$\sigma_{min}(A+B)=\min_{\mathbf{a},\|\mathbf{a}\|_2=1}\|(A+B)\mathbf{a}\|_2.$$

Via the triangle inequality,

$$\begin{aligned}
\sigma_{min}(A+B)&\geq\min_{\mathbf{a},\|\mathbf{a}\|_2=1}\left(\|B\mathbf{a}\|_2-\|A\mathbf{a}\|_2\right)\\
&\geq\min_{\mathbf{a},\|\mathbf{a}\|_2=1}\|B\mathbf{a}\|_2-\|A\|_2\\
&\geq\sigma_{min}(B)-\|A\|_2
\end{aligned}$$

.

Bounding $\|A\|_2$

$$\begin{aligned}
\|A\|_2&=\left\|\sum_{t=1}^{T}\boldsymbol{\theta}_t(\mathbf{b}_t-\mathbb{E}[\mathbf{b}_t]+\mathbb{E}[\mathbf{b}_t]-\bar{\mathbf{b}})^\top\right\|_2\\
&\leq\left\|\sum_{t=1}^{T}\boldsymbol{\theta}_t(\mathbf{b}_t-\mathbb{E}[\mathbf{b}_t])^\top\right\|_2+\left\|\sum_{t=1}^{T}\boldsymbol{\theta}_t(\mathbb{E}[\mathbf{b}_t]-\bar{\mathbf{b}})^\top\right\|_2
\end{aligned}$$

Bound on first term

$$\begin{aligned}
&\left\|\sum_{t=1}^{T}\boldsymbol{\theta}_t(\mathbf{b}_t-\mathbb{E}[\mathbf{b}_t])^\top\right\|_2\leq\left\|\sum_{t=1}^{T}\boldsymbol{\theta}_t(\mathbf{b}_t-\mathbb{E}[\mathbf{b}_t])^\top\right\|_F\\
&\leq\left(\sum_{i=1}^{m}\sum_{j=1}^{m}\left(\sum_{t=1}^{T}\theta_{t,i}(z_{t,j}-\mathbb{E}[z_{t,j}])\right)^2\right)^{1/2}
\end{aligned}$$

Notice that $\theta_{t,i}(z_{t,j} - \mathbb{E}[z_{t,j}])$ is a zero-mean bounded random variable with variance at most $\beta^2\sigma_z^2$. Applying Lemma E.II.1, we can see that

$$\begin{aligned}
&\left\|\sum_{t=1}^{T} \boldsymbol{\theta}_t(\mathbf{b}_t - \mathbb{E}[\mathbf{b}_t])^\top\right\|_2 \\
&\leq \left(\sum_{i=1}^{m}\sum_{j=1}^{m}\left(\beta\sigma_z\sqrt{2T\log(1/\delta_{i,j})}\right)^2\right)^{1/2} \\
&\leq \left(\sum_{i=1}^{m}\sum_{j=1}^{m}\beta^2\sigma_z^2 2T\log(m^2/\delta)\right)^{1/2} \\
&\leq \left(m^2\beta^2\sigma_z^2 2T\log(m^2/\delta)\right)^{1/2} \\
&\leq m\beta\sigma_z\sqrt{2T\log(m^2/\delta)}
\end{aligned}$$

with probability at least $1-\delta$.

Bound on second term

$$\begin{aligned}
&\left\|\sum_{t=1}^{T} \boldsymbol{\theta}_t(\mathbb{E}[\mathbf{b}_t] - \bar{\mathbf{b}})^\top\right\|_2 \\
&= \left\|\sum_{t=1}^{T} \boldsymbol{\theta}_t \frac{1}{T}\sum_{s=1}^{T}(\mathbb{E}[\mathbf{b}_t] - \mathbf{b}_j)^\top\right\|_2 \\
&\leq \left\|\sum_{t=1}^{T} \boldsymbol{\theta}_t \frac{1}{T}\sum_{s=1}^{T}(\mathbb{E}[\mathbf{b}_t] - \mathbf{b}_j)^\top\right\|_F \\
&\leq \left(\sum_{i=1}^{m}\sum_{j=1}^{m}\left(\sum_{t=1}^{T}\theta_{t,i}\frac{1}{T}\sum_{s=1}^{T}(\mathbb{E}[z_{t,j}] - z_j)\right)^2\right)^{1/2} \\
&\leq \left(\sum_{i=1}^{m}\sum_{j=1}^{m}\left(\sum_{t=1}^{T}|\theta_{t,i}|\frac{1}{T}\left|\sum_{s=1}^{T}(\mathbb{E}[z_{t,j}] - z_j)\right|\right)^2\right)^{1/2}
\end{aligned}$$

By applying Lemma E.II.1, we obtain

$$
\begin{aligned}
&\left\|\sum_{t=1}^{T} \boldsymbol{\theta}_t(\mathbb{E}[\mathbf{b}_t] - \bar{\mathbf{b}})^\top\right\|_2 \\
&\leq \left(\sum_{i=1}^{m}\sum_{j=1}^{m}\left(\sum_{t=1}^{T}|\theta_{t,i}|\frac{1}{T}\sigma_z\sqrt{2T\log(1/\delta_{i,j})}\right)^2\right)^{1/2} \\
&\leq \left(\sum_{i=1}^{m}\sum_{j=1}^{m}\left(\beta\sigma_z\sqrt{2T\log(1/\delta_{i,j})}\right)^2\right)^{1/2} \\
&\leq \left(\sum_{i=1}^{m}\sum_{j=1}^{m}\beta^2\sigma_z^2 2T\log(m^2/\delta)\right)^{1/2} \\
&\leq m\beta\sigma_z\sqrt{2T\log(m^2/\delta)}
\end{aligned}
$$

Bounding $\sigma_{min}(B)$

Next we bound $\sigma_{min}(B) = \sigma_{min}(\sum_{t=1}^{T}\boldsymbol{\theta}_t\boldsymbol{\theta}_t^\top\mathcal{E}_t\mathcal{E}_t^\top)$. We can write $\mathcal{E}_t\mathcal{E}_t^\top$ as $\mathbb{E}[\mathcal{E}_t\mathcal{E}_t^\top] + \varepsilon_t$. Note that since each element of $\mathcal{E}_t$ is bounded, each element of $\varepsilon_t \in \mathbb{R}^{m\times m}$ will be bounded as well. Using this formulation,

$$
\begin{aligned}
\sigma_{min}(B) &= \sigma_{min}\left(\sum_{t=1}^{T}\boldsymbol{\theta}_t\boldsymbol{\theta}_t^\top(\mathbb{E}[\mathcal{E}_t\mathcal{E}_t^\top] + \varepsilon_t)\right) \\
&= \sigma_{min}\left(\sum_{t=1}^{T}\boldsymbol{\theta}_t\boldsymbol{\theta}_t^\top\mathbb{E}[\mathcal{E}_t\mathcal{E}_t^\top] + \sum_{t=1}^{T}\boldsymbol{\theta}_t\boldsymbol{\theta}_t^\top\varepsilon_t)\right) \\
&\geq \sigma_{min}\left(\sum_{t=1}^{T}\boldsymbol{\theta}_t\boldsymbol{\theta}_t^\top\mathbb{E}[\mathcal{E}_t\mathcal{E}_t^\top]\right) - \left\|\sum_{t=1}^{T}\boldsymbol{\theta}_t\boldsymbol{\theta}_t^\top\varepsilon_t\right\|_2 \\
&\geq \sigma_{min}\left(\sum_{t=1}^{T}\boldsymbol{\theta}_t\boldsymbol{\theta}_t^\top\mathbb{E}[\mathcal{E}_t\mathcal{E}_t^\top]\right) - \left\|\sum_{t=1}^{T}\boldsymbol{\theta}_t\boldsymbol{\theta}_t^\top\varepsilon_t\right\|_F
\end{aligned}
$$

We proceed by bounding each term separately.

Bound on first term

$$\sigma_{min}\left(\sum_{t=1}^{T}\boldsymbol{\theta}_t\boldsymbol{\theta}_t^\top\mathbb{E}[\mathcal{E}_t\mathcal{E}_t^\top]\right)$$
$$\geq \sigma_{min}(\mathbb{E}[\mathcal{E}_t\mathcal{E}_t^\top])\sigma_{min}(\sum_{t=1}^{T}\boldsymbol{\theta}_t\boldsymbol{\theta}_t^\top)$$

Let $c = \sigma_{min}(\mathbb{E}[\mathcal{E}_t\mathcal{E}_t^\top])$. We assume that $\mathcal{E}_t$ is distributed such that $c > 0$. Therefore,

$$\sigma_{min}\left(\sum_{t=1}^{T}\boldsymbol{\theta}_t\boldsymbol{\theta}_t^\top\mathbb{E}[\mathcal{E}_t\mathcal{E}_t^\top]\right) \geq c\sigma_{min}(\sum_{t=1}^{T}\boldsymbol{\theta}_t\boldsymbol{\theta}_t^\top).$$

Next, we use the matrix Chernoff bound to bound

$$c\sigma_{min}(\sum_{t=1}^{T}\boldsymbol{\theta}_t\boldsymbol{\theta}_t^\top) = c\lambda_{min}(\sum_{t=1}^{T}\boldsymbol{\theta}_t\boldsymbol{\theta}_t^\top)$$

with high probability.

Theorem E.II.2 (Matrix Chernoff). *Consider a finite sequence $\{X_t\}_{t=1}^T$ of independent, random, Hermitian matrices with common dimension d. Assume that*

$$0 \leq \lambda_{min}(X_t) \text{ and } \lambda_{max}(X_t) \leq L \text{ for each index } t$$

Introduce the random matrix

$$Y = \sum_{t=1}^{T} X_t.$$

Define the minimum eigenvalue μ_{min} of the expectation $\mathbb{E}[Y]$:

$$\mu_{min} = \lambda_{min}(\mathbb{E}[Y]) = \lambda_{min}\left(\sum_{t=1}^{T}\mathbb{E}[X_t]\right)$$

Then,

$$P(\lambda_{min}(Y) \leq (1-\varepsilon)\mu_{min}) \leq d\left(\frac{e^{-\varepsilon}}{(1-\varepsilon)^{1-\varepsilon}}\right)^{\mu_{min}/L}$$

for $\varepsilon \in [0, 1)$.

Let $Y = \sum_{t=1}^{T} \boldsymbol{\theta}_t \boldsymbol{\theta}_t^\top$. In our setting,

$$\begin{aligned}\mu_{min} &= \lambda_{min}\left(\sum_{t=1}^{T} \mathbb{E}[\boldsymbol{\theta}_t \boldsymbol{\theta}_t^\top]\right) \\ &= T\lambda_{min}\left(\mathbb{E}[\boldsymbol{\theta}_t \boldsymbol{\theta}_t^\top]\right) \\ &= T\lambda_{min}\left(\sigma_\theta^2 \mathbb{I}_{m\times m} + \mathbb{E}[\boldsymbol{\theta}_t]\mathbb{E}[\boldsymbol{\theta}_t^\top]\right)\end{aligned}$$

$\sigma_\theta^2 \mathbb{I}_{m\times m}$ and $\mathbb{E}[\boldsymbol{\theta}_t]\mathbb{E}[\boldsymbol{\theta}_t^\top]$ commute, so

$$\begin{aligned}\mu_{min} &= T\left(\lambda_{min}\left(\sigma_\theta^2 \mathbb{I}_{m\times m}\right) + \lambda_{min}\left(\mathbb{E}[\boldsymbol{\theta}_t]\mathbb{E}[\boldsymbol{\theta}_t^\top]\right)\right) \\ &= T\lambda_{min}\left(\sigma_\theta^2 \mathbb{I}_{m\times m}\right) \\ &= T\sigma_\theta^2 \lambda_{min}\left(\mathbb{I}_{m\times m}\right) \\ &= T\sigma_\theta^2\end{aligned}$$

$$\lambda_{max}(\boldsymbol{\theta}_t \boldsymbol{\theta}_t^\top) = \beta m,$$

so let $L = \beta m$.

Picking $\varepsilon = 1/2$ and applying the matrix Chernoff bound to $\lambda_{min}(\sum_{t=1}^{T} \boldsymbol{\theta}_t \boldsymbol{\theta}_t^\top)$, we obtain

$$P\left(\lambda_{min}\left(\sum_{t=1}^{T} \boldsymbol{\theta}_t \boldsymbol{\theta}_t^\top\right) \leq \frac{1}{2}T\sigma_\theta^2\right) \leq d\left(\frac{1}{2}e\right)^{-\frac{T\sigma_\theta^2}{2\beta m}}$$

By rearranging terms, we see that if $T \geq \frac{2\beta m}{\sigma_\theta^2 \log \frac{1}{2}e} \log \frac{d}{\delta}$, then

$$\lambda_{min}\left(\sum_{t=1}^{T} \boldsymbol{\theta}_t \boldsymbol{\theta}_t^\top\right) \geq \frac{1}{2}T\sigma_\theta^2$$

with probability at least $1 - \delta$.

Bound on second term

$$\left\|\sum_{t=1}^{T} \boldsymbol{\theta}_t \boldsymbol{\theta}_t^\top \varepsilon_t\right\|_F = \left(\sum_{i=1}^{m}\sum_{j=1}^{m}\left(\sum_{t=1}^{T} \theta_{t,i}\theta_{t,j}\varepsilon_{t,i,j}\right)^2\right)^{1/2}$$

Since each $\varepsilon_{t,i,j}$ is a bounded zero-mean random variable, $\theta_{t,i}\theta_{t,j}\varepsilon_{t,i,j}$ is also a bounded zero-mean random variable, with variance at most $\beta^4\sigma_{\mathcal{E}}^2$ We can now apply Lemma E.II.1:

$$\begin{aligned}
&\left\|\sum_{t=1}^{T} \boldsymbol{\theta}_t \boldsymbol{\theta}_t^\top \varepsilon_t\right\|_F \\
&\leq \left(\sum_{i=1}^{m}\sum_{j=1}^{m}\left(\beta^2\sigma_{\mathcal{E}}\sqrt{2T\log(1/\delta_{i,j})}\right)^2\right)^{1/2} \\
&\leq \left(\sum_{i=1}^{m}\sum_{j=1}^{m}\beta^4\sigma_{\mathcal{E}}^2 2T\log(m^2/\delta)\right)^{1/2} \\
&\leq \left(m^2\beta^4\sigma_{\mathcal{E}}^2 2T\log(m^2/\delta)\right)^{1/2} \\
&\leq m\beta^2\sigma_{\mathcal{E}}\sqrt{2T\log(m^2/\delta)}
\end{aligned}$$

with probability at least $1-\delta$.

Putting everything together

Putting everything together, we have that

$$\begin{aligned}
&\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 \leq \\
&\frac{2\beta\sigma_g\sqrt{2m\log(m/\delta)}}{\frac{1}{2}c\sqrt{T}\sigma_\theta^2 - m\beta^2\sigma_{\mathcal{E}}\sqrt{2\log(m^2/\delta)} - 2m\beta\sigma_z\sqrt{2\log(m^2/\delta)}}
\end{aligned}$$

with probability at least $1-6\delta$.

E.III Individual Fairness Derivations

E.III.1 Proof of Theorem 7.3.3

Proof.

$$
\begin{aligned}
&|\hat{y} - \hat{y}'| = \\
&|(\mathbf{b} - \mathbf{b}')^\top \boldsymbol{\theta}^* + \boldsymbol{\theta}^{*\top}(\mathcal{E}\mathcal{E}^\top - \mathcal{E}'\mathcal{E}'^\top)\boldsymbol{\theta}^*| \\
&= |(\mathbf{b}_\mathcal{C} - \mathbf{b}'_\mathcal{C})^\top \boldsymbol{\theta}^* + \boldsymbol{\theta}^{*\top}((\mathcal{E}\mathcal{E}^\top)_\mathcal{C} - (\mathcal{E}'\mathcal{E}'^\top)_\mathcal{C})\boldsymbol{\theta}^*| \\
&\leq \|\mathbf{b}_\mathcal{C} - \mathbf{b}'_\mathcal{C}\|_2 \|\boldsymbol{\theta}^*\|_2 + \|\boldsymbol{\theta}^*\|_2 \|((\mathcal{E}\mathcal{E}^\top)_\mathcal{C} - (\mathcal{E}'\mathcal{E}'^\top)_\mathcal{C})\boldsymbol{\theta}^*\|_2 \\
&\leq \|\mathbf{b}_\mathcal{C} - \mathbf{b}'_\mathcal{C}\|_2 + \max_{\boldsymbol{\theta}, \|\boldsymbol{\theta}\|_2 = 1} \|((\mathcal{E}\mathcal{E}^\top)_\mathcal{C} - (\mathcal{E}'\mathcal{E}'^\top)_\mathcal{C})\boldsymbol{\theta}^*\|_2 \\
&\leq \|\mathbf{b}_\mathcal{C} - \mathbf{b}'_\mathcal{C}\|_2 + \|((\mathcal{E}\mathcal{E}^\top)_\mathcal{C} - (\mathcal{E}'\mathcal{E}'^\top)_\mathcal{C})\|_2
\end{aligned}
$$

□

E.III.2 Proof of Theorem 7.3.4

Proof. Let

$$
\begin{aligned}
b_{\widetilde{\mathcal{C}},i} &= \begin{cases} b_i & \text{if } i \notin \mathcal{C} \\ 0 & \text{otherwise} \end{cases}, \\
(\mathcal{E}\mathcal{E}^\top)_{\widetilde{\mathcal{C}},ij} &= \begin{cases} (\mathcal{E}\mathcal{E}^\top)_{ij} & \text{if } i, j \notin \mathcal{C} \\ 0 & \text{otherwise.} \end{cases}
\end{aligned}
$$

$$
\begin{aligned}
&|\hat{y} - \hat{y}'| = |(\mathbf{b}_\mathcal{C} - \mathbf{b}'_\mathcal{C})^\top \boldsymbol{\theta} + \boldsymbol{\theta}^\top((\mathcal{E}\mathcal{E}^\top)_\mathcal{C} - (\mathcal{E}'\mathcal{E}'^\top)_\mathcal{C})\boldsymbol{\theta} \\
&+ (\mathbf{b}_{\widetilde{\mathcal{C}}} - \mathbf{b}'_{\widetilde{\mathcal{C}}})^\top \boldsymbol{\theta} + \boldsymbol{\theta}^\top((\mathcal{E}\mathcal{E}^\top)_{\widetilde{\mathcal{C}}} - (\mathcal{E}'\mathcal{E}'^\top)_{\widetilde{\mathcal{C}}})\boldsymbol{\theta}| \\
&= |(\mathbf{b}_{\widetilde{\mathcal{C}}} - \mathbf{b}'_{\widetilde{\mathcal{C}}})^\top \boldsymbol{\theta} + \boldsymbol{\theta}^\top((\mathcal{E}\mathcal{E}^\top)_{\widetilde{\mathcal{C}}} - (\mathcal{E}'\mathcal{E}'^\top)_{\widetilde{\mathcal{C}}})\boldsymbol{\theta}| \\
&= \left| \sum_{i \notin \mathcal{C}} (b_i - b'_i)\theta_i + \sum_{i \notin \mathcal{C}} \sum_{j \notin \mathcal{C}} ((\mathcal{E}\mathcal{E}^\top)_{ij} - (\mathcal{E}'\mathcal{E}'^\top)_{ij})\theta_i \theta_j \right|
\end{aligned}
$$

□

E.III.3 Example 7.3.5 Derivations

$$d(\mathbf{u}, \mathbf{u}') = \|b_{\mathcal{C}} - b'_{\mathcal{C}}\|_2 + \|(\mathcal{E}\mathcal{E}^\top)_{\mathcal{C}} - (\mathcal{E}'\mathcal{E}'^\top)_{\mathcal{C}}\|_2$$
$$= \|(\boldsymbol{\delta} - \boldsymbol{\delta}')_{\mathcal{C}} I_{n\times n}\|_2 = \|\mathbf{0}_{n\times n}\|_2 = 0,$$

where

$$\delta_{\mathcal{C},i} = \begin{cases} \delta_i & \text{if } i \in \mathcal{C} \\ 0 & \text{otherwise.} \end{cases}$$

$$|\hat{y} - \hat{y}'| = \left| 0 + \sum_{i \notin \mathcal{C}} (n-0)\theta_i^2 \right| = n \sum_{i=1}^{n/2} \theta_i^2$$

E.IV Agent outcome maximization derivations

Derivation of $\boldsymbol{\theta}^{AO}$

$$\boldsymbol{\theta}^{AO} = \arg\max_{\boldsymbol{\theta}\in\mathcal{S}} \quad \mathbb{E}[y_t]$$

Substituting in for y_t:

$$\boldsymbol{\theta}^{AO} = \arg\max_{\boldsymbol{\theta}\in\mathcal{S}} \quad \mathbb{E}[\mathbf{x}_t^\top \boldsymbol{\theta}^* + o_t]$$

$$\boldsymbol{\theta}^{AO} = \arg\max_{\boldsymbol{\theta}\in\mathcal{S}} \quad \mathbb{E}[\mathbf{x}_t^\top \boldsymbol{\theta}^*] + \mathbb{E}[o_t]$$

$$\boldsymbol{\theta}^{AO} = \arg\max_{\boldsymbol{\theta}\in\mathcal{S}} \quad \mathbb{E}[\mathbf{x}_t^\top \boldsymbol{\theta}^*]$$

Substitute in for $\mathbf{x}_t$:

$$\boldsymbol{\theta}^{AO} = \arg\max_{\boldsymbol{\theta}\in\mathcal{S}} \quad \mathbb{E}[(\mathbf{b}_t^\top + \boldsymbol{\theta}^\top \mathcal{E}_t \mathcal{E}_t^\top)\boldsymbol{\theta}^*]$$

$$\boldsymbol{\theta}^{AO} = \arg\max_{\boldsymbol{\theta}\in\mathcal{S}} \quad \mathbb{E}[\mathbf{b}_t^\top \boldsymbol{\theta}^*] + \mathbb{E}[\boldsymbol{\theta}^\top \mathcal{E}_t \mathcal{E}_t^\top \boldsymbol{\theta}^*]$$

$$\boldsymbol{\theta}^{AO} = \arg\max_{\boldsymbol{\theta}\in\mathcal{S}} \quad \mathbb{E}[\boldsymbol{\theta}^\top \mathcal{E}_t \mathcal{E}_t^\top \boldsymbol{\theta}^*]$$

$$
\begin{aligned}
\boldsymbol{\theta}^{AO} &= \arg\max_{\boldsymbol{\theta}\in\mathcal{S}} \quad \boldsymbol{\theta}^\top \mathbb{E}[\mathcal{E}_t\mathcal{E}_t^\top]\boldsymbol{\theta}^* \\
&= \arg\max_{\boldsymbol{\theta}\in\mathcal{S}} \quad \sum_{i\in\mathcal{C}}\sum_{j\in\mathcal{C}}\sum_{k=1}^{d}\mathbb{E}[w_{ik}w_{jk}]\theta_i\theta_j^* \\
&+ \sum_{i\notin\mathcal{C}}\sum_{j\in\mathcal{C}}\sum_{k=1}^{d}\mathbb{E}[w_{ik}w_{jk}]\theta_i\theta_j^*
\end{aligned}
$$

E.V Predictive risk minimization derivations

E.V.1 Population gradient derivation

The gradient of the population risk function $f(\boldsymbol{\theta}_t) = \mathbb{E}[(\widehat{y}_t - y_t)^2]$ can be derived as follows

$$
\begin{aligned}
\nabla_{\boldsymbol{\theta}_t} f(\boldsymbol{\theta}_t) &= \mathbb{E}[\nabla_{\boldsymbol{\theta}_t}(\widehat{y}_t - y_t)^2] \\
&= 2\mathbb{E}[(\widehat{y}_t - y_t)\nabla_{\boldsymbol{\theta}_t}(\widehat{y}_t - y_t)] \\
&= 2\mathbb{E}[(\widehat{y}_t - y_t)\nabla_{\boldsymbol{\theta}_t}(\mathbf{x}_t^\top\boldsymbol{\theta}_t - \mathbf{x}_t^\top\boldsymbol{\theta}^* - o_t)] \\
&= 2\mathbb{E}[(\widehat{y}_t - y_t)\nabla_{\boldsymbol{\theta}_t}(\mathbf{x}_t^\top(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*))] \\
&= 2\mathbb{E}[(\widehat{y}_t - y_t)\nabla_{\boldsymbol{\theta}_t}((\mathbf{b}_t^\top + \boldsymbol{\theta}_t\mathcal{E}_t\mathcal{E}_t^\top)(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*))] \\
&= 2\mathbb{E}[(\widehat{y}_t - y_t)(\mathbf{b}_t + \mathcal{E}_t\mathcal{E}_t^\top(2\boldsymbol{\theta}_t - \boldsymbol{\theta}^*))] \\
&= 2\mathbb{E}[(\widehat{y}_t - y_t)(\mathbf{x}_t + \mathcal{E}_t\mathcal{E}_t^\top(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*))]
\end{aligned}
$$

E.VI Omitted experiments

In this section, we present additional details for our experiments in Section 4.7. At the end, we provide more information regarding the dataset and computation resources used.

E.VI.1 University admissions full experimental description

We construct a semi-synthetic dataset based on an example of university admissions with disadvantaged and advantaged students from Hu et al. (2019). From a real dataset of the high school (HS) GPA, SAT score, and college GPA of 1000 college students, we estimate the causal effect of observed features [SAT, HS GPA] on college GPA to be $\boldsymbol{\theta}^* = [0.00085, 0.49262]^\top$ using OLS (which is assumed to be consistent, since we have yet to modify the data to include confounding). We then use this dataset to construct synthetic data which looks similar, yet incorporates confounding factors. For simplicity, we let the true causal effect parameters $\boldsymbol{\theta}^* = [0, 0.5]^\top$. That is, we assume there is a significant causal relationship between college performance and HS GPA, but not SAT score.[1] We consider two types of student backgrounds, those from a *disadvantaged group* and those from an *advantaged group*. We assume disadvantaged applicants have, on average, lower HS GPA and SAT $\mathbf{b}_t$, lower baseline college GPA o_t, and require more effort to improve observable features (reflected in $\mathcal{E}_t$): this could be due to disadvantaged groups being systemically underserved, marginalized, or abjectly discriminated against (and the converse for advantaged groups). Initial features $\mathbf{b}_t$ are constructed as such: For any disadvantaged applicant t, their initial SAT features $z_t^{\text{SAT}} \sim \mathcal{N}(800, 200)$ and initial HS GPA $z_t^{\text{HS GPA}} \sim \mathcal{N}(1.8, 0.5)$. For any advantaged applicant t, $z_t^{\text{SAT}} \sim \mathcal{N}(1000, 200)$ and $z_t^{\text{HS GPA}} \sim \mathcal{N}(2.2, 0.5)$. We truncate SAT scores between 400 to 1600 and HS GPA between 0 to 4. For any applicant t, we randomly deploy assessment rule $\boldsymbol{\theta}_t = [\theta_t^{\text{SAT}}, \theta_t^{\text{HS GPA}}]^\top$ where $\theta_t^{\text{SAT}} \sim \mathcal{N}(1, 10)$ and $\theta_t^{\text{HS GPA}} \sim \mathcal{N}(1, 2)$. $\boldsymbol{\theta}_t$ need not be zero-mean, so universities can play a reasonable assessment rule with slight perturbations while still being able to perform unbiased causal estimation. Components of the average effort conversion matrix $\mathbb{E}[\mathcal{E}_t]$ are smaller for disadvantaged applicants, which makes their mean improvement worse (see Figure E.1). We set the expected effort conversion term $\mathbb{E}[\mathcal{E}_t] = \begin{pmatrix} 10 & 0 \\ 0 & 1 \end{pmatrix}$ for simplicity. Each row of $\mathbb{E}[\mathcal{E}_t]$ corresponds to effort expended to change a specific feature. For example, entries in the first row of $\mathbb{E}[\mathcal{E}_t]$ correspond to effort expended to change one's SAT score. For each applicant t, we perturb $\mathbb{E}[\mathcal{E}_t]$ with random noise drawn from $\mathcal{N}(0.5, 0.25)$ to the top left entry and noise drawn from $\mathcal{N}(0.1, 0.01)$ the bottom right entry to produce $\mathcal{E}_t$. We add this noise to $\mathbb{E}[\mathcal{E}_t]$ to produce $\mathcal{E}_t$ for advantaged applicants and subtract for disadvantaged applicants: thus, it takes more effort, on average, for members of disadvantaged groups to improve their HS GPA and SAT scores than members of advantaged groups. Finally, we construct college GPA (true outcome y_t) by multiplying observed features $\mathbf{x}_t$ by the true effect parameters $\boldsymbol{\theta}^*$. We then add confounding error o_t where $o_t \sim \mathcal{N}(0.5, 0.2)$ for disadvantaged applicants

[1] Though this assumption may be contentious, it is based on existing research Allensworth and Clark (2020).

and $o_t \sim \mathcal{N}(1.5, 0.2)$ for advantaged applicants. Disadvantage applicants could have lower baseline outcomes, e.g. due to institutional barriers or discrimination. While the setting we consider is simplistic, Figures E.1 and E.2 demonstrate that our semi-synthetic admissions data behaves reasonably.[2]

E.VI.2 Experimental Details

We evaluate our model on a semi-synthetic dataset based on our running university admission example Dua and Graff (2019). The dataset we base our experiments off of is publicly available at `www.openintro.org/data/index.php?data=satgpa`. This dataset does not contain personally identifiable information or offensive content. Since this is a publicly available dataset, no consent from the people whose data we are using was required. We ran our experiments on a 2020 MacBook Air laptop with 16GB of RAM.

E.VII Comparison with Shavit et al.

The setting most similar to ours is that of Shavit et al.. They consider a strategic classification setting in which an agent's outcome is a linear function of features –some observable and some not (see Figure E.4 for a graphical representation of their model). While they assume that an agent's hidden attributes can be modified strategically, we choose to model the agent as having an unmodifiable private *type*. Both of these assumptions are reasonable, and some domains may be better described by one model than the other. For example, the model of Shavit et al. may be useful in a setting such as car insurance pricing, where some unobservable factors which lead to safe driving are modifiable. On the other hand, settings like our college admissions example in which the unobservable features which contribute to college success (i.e. socioeconomic status, lack of resources, etc., captured in o_t) are not easily modifiable.

One benefit of our setting is that we are able to use $\boldsymbol{\theta}_t$ as a valid instrument to recover the true relationship $\boldsymbol{\theta}^*$ between observable features and outcomes. This is generally not possible in the model of Shavit et al. (2020), since $\boldsymbol{\theta}_t$ violates the backdoor criterion as long as there exists any hidden features $\mathbf{h}_t$ and is therefore not a valid instrument.

[2] For example, the mean shift in SAT scores from the first to second exam is 46 points Goodman et al. (2020). In our data, the mean shift for disadvantaged and advantaged applicants is about 36 points and 91 points, respectively.

Another difference between our setting and theirs is that we allow for a heterogeneous population of agents, while they do not. Specifically, they assume that each agent's mapping from actions to features is the same, while our model is capable of handling mappings which vary from agent-to-agent.

A natural question is whether or not there exists a general model which captures the setting of both Shavit et al. and ours. We provide such a model in Figure E.5. In this setting, an agent has both observable and unobservable features, both of which are affected by the assessment rule $\boldsymbol{\theta}_t$ deployed and the agent's private type u_t. However, much like the setting of Shavit et al., $\boldsymbol{\theta}_t$ violates the backdoor criterion, so it cannot be used as a valid instrument in order to recover the true relationship between observable features and outcomes. Moreover, the following toy example illustrates that *no* form of true parameter recovery can be performed when an agent's unobservable features are modifiable.

Example E.VII.1. Consider the one-dimensional setting

$$y_t = \theta^* x_t + \beta^* h_t,$$

where x_t is an agent's observable, modifiable feature and h_t is an unobservable, modifiable feature. If the relationship between x_t and h_t is unknown, then it is generally impossible to recover the true relationship between x_t, h_t, and outcome y_t. To see this, consider the setting where h_t and x_t are highly correlated. In the extreme case, take $h_t = x_t$, $\forall t$. (Note we use equality to indicate identical feature values, not a causal relationship.) In this setting, the models $\theta^* = 1$, $\beta^* = 1$ and $\theta^* = 2$, $\beta^* = 0$ produce the same outcome y_t for all $x \in \mathbb{R}$, making it impossible to distinguish between the two models, even in the limit of infinite data.

E.VIII Setting of SGD comparison

In 1D, the derivative of $\mathbb{E}[(\widehat{y}_t - y_t)]^2$ which accounts for x_t and y_t's dependence on $\boldsymbol{\theta}_t$ takes the form

$$\Delta = 2(\mathbb{E}[(\widehat{y}_t - y_t)x_t] + \mathbb{E}_{x_t}[\widehat{y}_t - y_t]\mathcal{E}^2(\theta_t - \theta^*)).$$

By plugging in for x_t, y_t, $\widehat{y}_t$ and simplifying, we can write the derivative as

$$\Delta = 2\left(\mathbb{E}[b_t^2] + \mathcal{E}^4\theta_t^2(\theta_t - \theta^*) - \mathbb{E}[o_t b_t] + \mathcal{E}^4\theta_t(\theta_t - \theta^*)^2\right).$$

The derivative of $\mathbb{E}[(\widehat{y}_t - y_t)]^2$ which does *not* account for x_t and y_t's dependence on $\boldsymbol{\theta}_t$ can be written as

$$\begin{aligned}\Delta' &= 2\left(\mathbb{E}[(\widehat{y}_t - y_t)x_t]\right)\\ &= 2\left(\mathbb{E}[b_t^2] + \mathcal{E}^4\theta_t^2(\theta_t - \theta^*) - \mathbb{E}[o_t b_t]\right).\end{aligned}$$

As can be seen by comparing the two equations, there is an extra $\mathcal{E}^4\theta_t(\theta_t - \theta^*)^2$ term present in Δ that is not in Δ'. This can cause Δ and Δ' to have opposite signs under certain scenarios, e.g. when $\mathbb{E}[b_t^2] + \mathcal{E}^4\theta_t^2(\theta_t - \theta^*) - \mathbb{E}[o_t b_t]$ is negative and $\mathcal{E}^4\theta_t(\theta_t - \theta^*)^2$ is sufficiently large. To generate Figure 7.4, we set $\mathbb{E}[b_t] = 0$, $\mathbb{E}[b_t^2] = 0.3$, $\mathbb{E}[o_t] = 0$, $\mathbb{E}[g_t^2] = 15$, $\mathbb{E}[o_t b_t] = -6.5$, $\mathcal{E} = 3$, $\theta^* = 1$, and $\theta_0 = 0.5$. To generate Figure E.6a and E.6b, we changed θ^* to be 0.7.

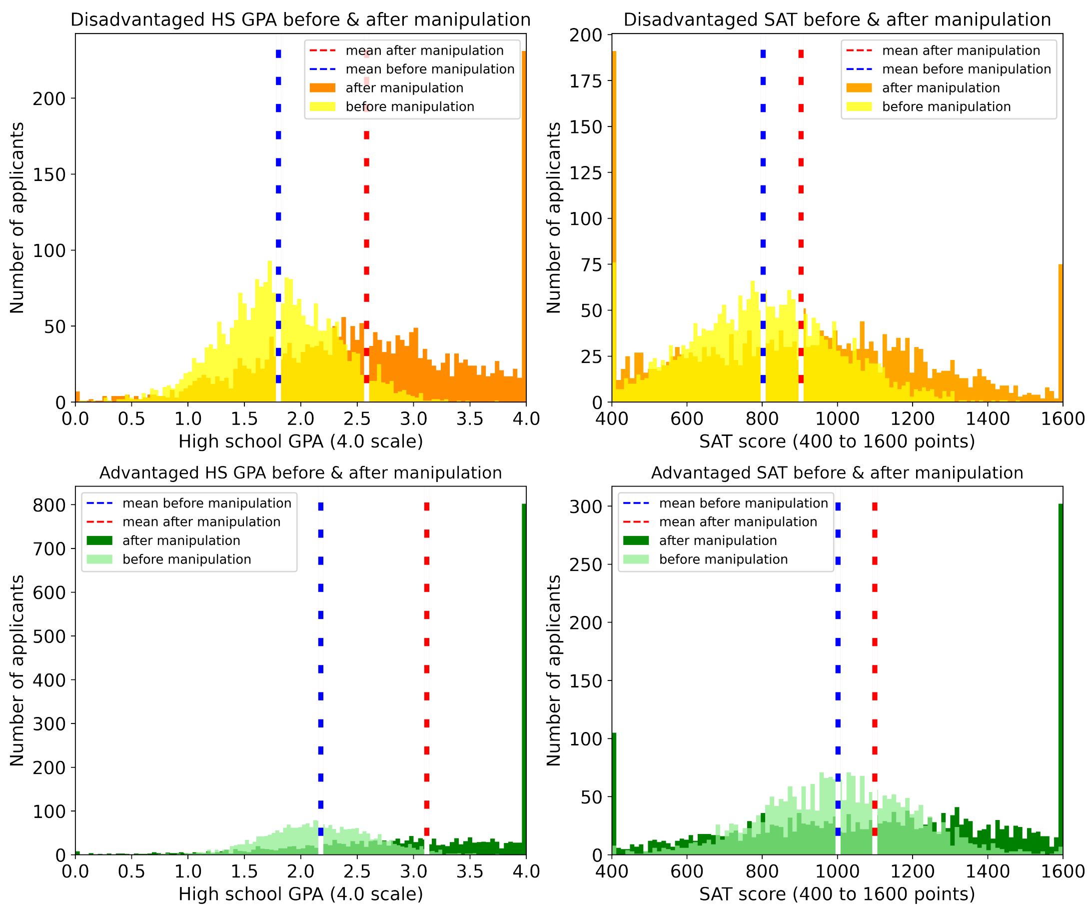


Figure E.1: Distributions of unobserved features **b** (in lighter colors), i.e. initial HS GPA (two left figures) and SAT (two right figures), and observed features **x** (darker colors) for disadvantaged (two top figures in yellow and orange) and advantaged students (two bottom figures in green).

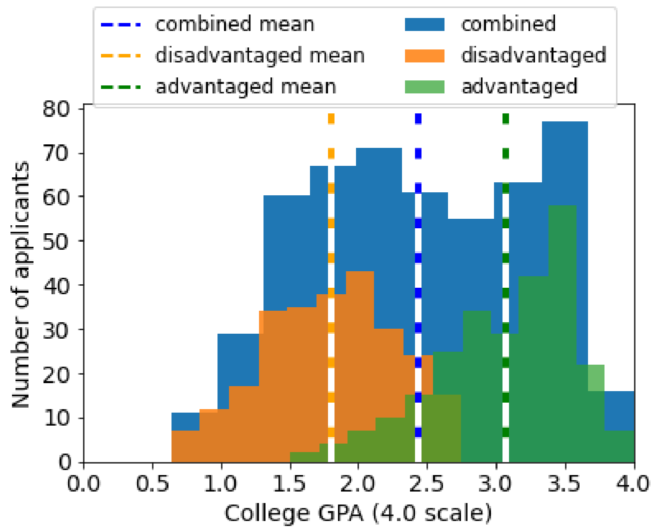


Figure E.2: Distribution of college GPAs (outcomes y) for disadvantaged students (orange), advantaged students (green), and both combined (blue).

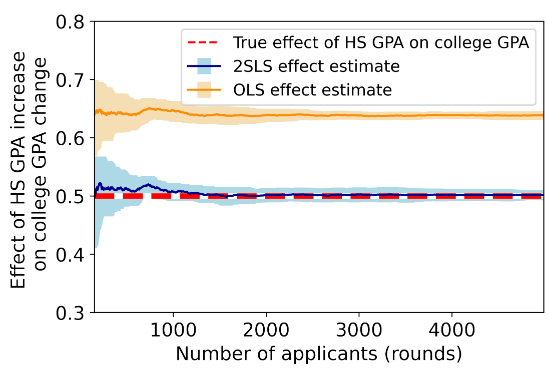


Figure E.3: OLS versus 2SLS estimates for high school GPA effect on college GPA over 5000 rounds. Results are averaged over 10 runs, with the error bars (in lighter colors) representing one standard deviation. The red dashed line is the true causal effect of each high school GPA on college GPA.

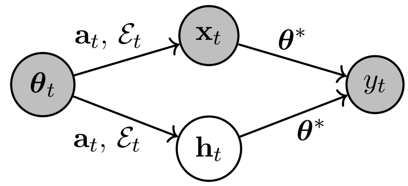


Figure E.4: Graphical model of Shavit et al.. Observable features $\mathbf{x}_t$ (e.g. the type of car a person drives) and unobservable features $\mathbf{h}_t$ (e.g. how defensive of a diver someone is) are affected by $\boldsymbol{\theta}_t$ through action $\mathbf{a}_t$ (e.g. buying a new car) and common action conversion matrix $\mathcal{E}$ (representing, in part, the cost to a person of buying a new car). Outcome y_t (in this example, the person's chance of getting in an accident) is affected by $\mathbf{x}_t$ and $\mathbf{h}_t$ through the true causal relationship $\boldsymbol{\theta}_t$. Note that causal parameter recovery is not possible in this setting unless all features are observable.

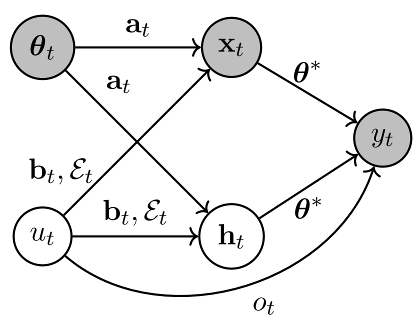
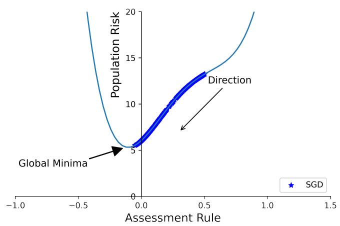
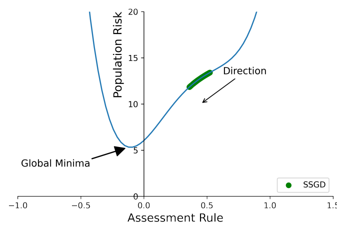


Figure E.5: Graphical model which captures both our setting and that of Shavit et al.. In this setting, observable features $\mathbf{x}_t$ *and* unobservable features $\mathbf{h}_t$ are affected by $\boldsymbol{\theta}_t$ through action $\mathbf{a}_t$. The agent's private type u_t affects $\mathbf{x}_t$ and $\mathbf{h}_t$ through initial feature values $\mathbf{b}_t$ and action conversion matrix $\mathcal{E}_t$. The agent's outcome y_t depends on $\mathbf{x}_t$ and $\mathbf{h}_t$ through the causal relationship $\boldsymbol{\theta}^*$ and u_t through confounding term o_t. Note that much like the setting of Shavit et al. (2020), causal parameter recovery is not possible in this setting unless all features are observable.



(a) SGD with invex function



(b) SSGD with invex function

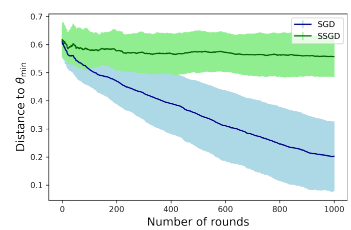


Figure E.6: Convergence rate of Stochastic Gradient Descent vs Simple Stochastic Gradient Descent for simple 1D setting. Even when SSGD converges, it may do so at a much slower rate, due to the inexact measure of the gradient. We ran both methods for 1000 time-steps with a decaying learning rate of $\frac{0.001}{\sqrt{T}}$. Results are averaged over 10 runs, with the error bars (in lighter colors) representing one standard deviation.

Appendix F

Appendix for Reconciling Model Multiplicity for Downstream Decision Making

F.I Limitation of Prior Work (Continue)

We provide the Line 1 from (Roth et al., 2023) and their theoretical guarantees for completion. First, given two predictors f_1 and f_2, define the disagreement region as:

$$U_\varepsilon(f_1, f_2) \coloneqq \{x : |f_1(x) - f_2(x)| > \varepsilon\}$$

which can be further divided into two partitions:

$$U_\varepsilon^{>}(f_1, f_2) = \{x \in U_\varepsilon(f_1, f_2) : f_1(x) > f_2(x)\}$$
$$U_\varepsilon^{<}(f_1, f_2) = \{x \in U_\varepsilon(f_1, f_2) : f_1(x) < f_2(x)\}$$

Theorem F.I.1 (Reconcile (Roth et al., 2023))**.** *For any pair of models $f_1, f_2 : [k] \to [0, 1]$, any distribution $\mathcal{D}$, and any $\alpha, \eta > 0$, Line 1 runs for $T = T_1 + T_2$ many rounds and outputs a pair of models $(f_1^{T_1}, f_2^{T_2})$ such that:*

1. $T \le (B(f_1, \mathcal{D}) + B(f_2, \mathcal{D})) \cdot \frac{16}{\eta\alpha^2}$
2. $B(f_1^{T_1}, \mathcal{D}) \le B(f_1, \mathcal{D}) - T_1 \cdot \frac{\eta\alpha^2}{16}$ *and* $B(f_2^{T_2}, \mathcal{D}) \le B(f_2, \mathcal{D}) - T_2 \cdot \frac{\eta\alpha^2}{16}$
3. $\mu(U_\varepsilon(f_1^{T_1}, f_2^{t_2})) \le \eta$

Furthermore, we provide a counterexample that shows ensuring each individual predictor is decision-calibrated by running Line 1 is not sufficient.

Decision-Calibrated predictions. Another baseline algorithm we consider is to run Algorithm 1 separately for both f_1, f_2. However, in the following theorem, we show that the updated predictors f_1' and f_2' can still disagree with each other on the best-response actions for substantially many units, indicating room for further improvement.

Theorem F.I.2. *For any $\eta \in (0, 1/4)$ and $\beta \in (0, 1/2)$, there exists a pair of predictors f_1, f_2 and a loss function ℓ, such that after running Decision-Calibration (Zhao et al., 2021), the resulting models f_1^T, f_2^T are β-decision-calibrated with respect to the loss function ℓ. There exists a set of units x with probability mass 2η where f_1^T and f_2^T disagree on the individual best-response actions.*

Proof. For any $\eta \in (0, 1/4), \beta \in (0, 1/2)$, let $[k] = [4]$, with $\Pr[1] = \Pr[4] = 1/2 - \eta$ and $\Pr[x = 2] = \Pr[x = 3] = \eta$. Consider the predictors f_1, f_2 as follows:

$$f_1(1) = f_1(2) = f_2(1) = f_2(3) = \frac{\beta}{4} - 2\eta\beta + 2\eta, \tag{F.1}$$

$$f_1(3) = f_1(4) = f_2(2) = f_2(4) = 1 - \frac{\beta}{2}, \tag{F.2}$$

$$f^*(1) = \frac{\beta}{2}, \quad f^*(2) = f^*(3) = f^*(4) = 1 - \frac{\beta}{2}. \tag{F.3}$$

Notice that

$$\frac{\beta}{4} - 2\eta\beta + 2\eta = \frac{\beta}{4} + 2\eta(1 - \beta) < \frac{1}{4} + \frac{1}{4} = \frac{1}{2}.$$

The best-response policy for each predictor is

$$\pi_\ell^{\mathsf{BR}}(f_1(1)) = \pi_\ell^{\mathsf{BR}}(f_1(2)) = 0, \;\; \pi_\ell^{\mathsf{BR}}(f_1(3)) = \pi_\ell^{\mathsf{BR}}(f_1(4)) = 1, \tag{F.4}$$

$$\pi_\ell^{\mathsf{BR}}(f_2(1)) = \pi_\ell^{\mathsf{BR}}(f_2(3)) = 0, \;\; \pi_\ell^{\mathsf{BR}}(f_2(2)) = \pi_\ell^{\mathsf{BR}}(f_1(4)) = 1. \tag{F.5}$$

For each best-response event, we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[f^*(x)E_1(f_1(x),x)] &= \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[f_1(x)E_1(f_1(x),x)],\\
\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[f^*(x)E_2(f_1(x),x)] &= \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[f_1(x)E_2(f_1(x),x)],\\
\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[f^*(x)E_1(f_2(x),x)] &= \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[f_2(x)E_1(f_2(x),x)],\\
\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[f^*(x)E_2(f_2(x),x)] &= \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[f_2(x)E_2(f_2(x),x)].
\end{aligned}$$

That is, f_1, f_2 are already decision-calibrated, so running Decision Calibration will not further improve either of the two predictors. However, based on our definition of disagreement events, we still have

$$E_{1,2} = \{2\}, \quad E_{2,1} = \{1\},$$

each with size

$$\mu(E_{1,2}) = \mu(E_{2,1}) = \eta.$$

However, we observe that f_1 and f_2 still disagree on the best-response action for units $x = 2$ and $x = 3$. We can further reduce the differences in best-response actions using our algorithm Line 1. $\square$

F.II Proofs of Section 9.3.1: Reconcile for Decision Making

First, we show that if a disagreement event has a large probability mass, then at least one of f_1, f_2 has a large prediction error within the region:

Lemma F.II.1. *Fix any two predictors $f_1, f_2 : [k] \to [0,1]^d$ and $\alpha, \eta > 0$. If $\mu(E_{\ell,a_1,a_2}) > \eta$ for some $a_1, a_2 \in \mathcal{A}$, then we have*

$$\| \mathop{\mathbb{E}}_{x\sim[k]} \left[f_i(x) - f^*(x) \middle| E_{\ell,a_1,a_2}(x) \right] \| \geq \frac{\alpha}{2\sqrt{d}} \tag{F.6}$$

for some $i \in \{1, 2\}$.

Proof. By definition of event E_{ℓ,a_1,a_2}, we have

$$\mathop{\mathbb{E}}_{x\sim[k]}[\ell(f_1(x),a_2)-\ell(f_1(x),a_1)+\ell(f_2(x),a_1)-\ell(f_2(x),a_2)]|E_{\ell,a_1,a_2}(x)]\geq\alpha$$

Also, we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{x\sim[k]}[\ell(f_1(x),a_2)-\ell(f_1(x),a_1)+\ell(f_2(x),a_1)-\ell(f_2(x),a_2)]|E_{\ell,a_1,a_2}(x)]\\
=&\mathop{\mathbb{E}}_{x\sim[k]}[\langle f_1(x)-f_2(x),\ell_{a_2}-\ell_{a_1}\rangle|E_{\ell,a_1,a_2}(x)]\\
=&\mathop{\mathbb{E}}_{x\sim[k]}[\langle f_1(x)-f^*(x),\ell_{a_2}-\ell_{a_1}\rangle|E_{\ell,a_1,a_2}(x)]\\
&\quad+\mathop{\mathbb{E}}_{x\sim[k]}[\langle f^*(x)-f_2(x),\ell_{a_2}-\ell_{a_1}\rangle|E_{\ell,a_1,a_2}(x)]\\
\leq&\|\mathop{\mathbb{E}}_{x\sim[k]}\left[f_1(x)-f^*(x)\middle|E_{\ell,a_1,a_2}(x)\right]\|_2\sqrt{d}+\|\mathop{\mathbb{E}}_{x\sim[k]}\left[f_2(x)-f^*(x)\middle|E_{\ell,a_1,a_2}(x)\right]\|_2\sqrt{d}\\
=&\sqrt{d}\cdot\left(\|\mathop{\mathbb{E}}_{x\sim[k]}\left[f_1(x)-f^*(x)\middle|E_{\ell,a_1,a_2}(x)\right]\|_2+\|\mathop{\mathbb{E}}_{x\sim[k]}\left[f_2(x)-f^*(x)\middle|E_{\ell,a_1,a_2}(x)\right]\|_2\right)
\end{aligned}$$

where the last inequality comes from Cauchy-Schwartz and that ℓ is bounded in $[0,1]$.

Combining the above inequalities, we have

$$\sqrt{d}\cdot\left(\left\|\mathop{\mathbb{E}}_{x\sim[k]}\left[f_1(x)-f^*(x)\middle|E_{\ell,a_1,a_2}(x)\right]\right\|_2+\left\|\mathop{\mathbb{E}}_{x\sim[k]}\left[f_2(x)-f^*(x)\middle|E_{\ell,a_1,a_2}(x)\right]\right\|_2\right)\geq\alpha.$$

Therefore, for some $i\in\{1,2\}$, we have

$$\left\|\mathop{\mathbb{E}}_{x\sim[k]}\left[f_i(x)-f^*(x)\middle|E_{\ell,a_1,a_2}(x)\right]\right\|_2\geq\frac{\alpha}{2\sqrt{d}}.$$

□

This lemma indicates that, if we have two predictors f_1, f_2 that create a large disagreement event, we can falsify at least one of the models. We now show that these events also provide a directly actionable way to improve one of the models.

Lemma F.II.2. *For any predictor $f:[k]\to[0,1]^d$, any event $E\in\mathcal{E}$, and distribution*

$\mathcal{D}$. *Let* $\phi = \mathbb{E}_{(x,y)\sim\mathcal{D}}[y - f(x)|E(x) = 1]$. *We patch* f *as*

$$f'(x) = \mathrm{proj}_{[0,1]^d}(f(x) + \phi E(x)), \quad \textit{where} \;\; \mathrm{proj}_{[0,1]^d}(y) = \operatorname*{argmin}_{y' \in [0,1]^d} \|y - y'\|_2.$$

Then,

$$B(f, \mathcal{D}) - B(f', \mathcal{D}) \geq \|\phi\|_2^2 \mu(E).$$

Proof.

$$\begin{aligned} B(f, \mathcal{D}) - B(f', \mathcal{D}) =& \mathbb{E}\left[\|f(x) - y\|_2^2 - \|f'(x) - y\|_2^2\right] \\ \geq& \mathbb{E}\left[\|f(x) - y\|_2^2 - \|f(x) + \phi E(x) - y\|_2^2\right] \\ & \text{(since projection is non-expansive)} \\ =& \mathbb{E}\left[2\langle y - f(x), \phi E(x)\rangle - \|\phi E(x)\|_2^2\right] \\ \geq& \|\phi\|_2^2 \cdot \mu(E) \end{aligned}$$

□

Therefore, whenever we have two predictors that have a large disagreement event, we can always falsify at least one of the predictors and improve it through patching, causing the Brier score to decrease by a large amount. Similarly, for a fixed predictor, if one of its best-response events has a large calibration error, we can patch the predictor within the event to decrease the Brier score. As the Brier score is bounded in $[0, d]$, these two observations imply that the number of time-steps for both Line 1 and its subroutine Line 1 are bounded.

Other than the Brier score, we also care about minimizing the loss of the downstream decision-making task. We now show that, after a further update through the subroutine Line 1, the loss does not increase much at each time-step of Line 1:

Lemma F.II.3. *For any predictors* f_1, f_2*, loss function* $\ell \in \mathcal{L}$ *and any distribution* $\mathcal{D}$*, at any time-step* t *of Line 1, the predictors satisfies*

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(y, \pi_\ell^{\mathsf{BR}}(f_i^{t+1}(x))) - \ell(y, \pi_\ell^{\mathsf{BR}}(f_i^t(x)))] \leq \beta\sqrt{d}A,$$

for all $i \in \{1, 2\}$.

Proof. At each round, we define the set $\Delta_a^t \subseteq E^t$ as

$$\Delta_a^t = \{x \in E^t : \pi_\ell^{\mathsf{BR}}(f_i^{t+1}(x)) = a\}.$$

Then, we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell(y, \pi_\ell^{\mathsf{BR}}(f_i^{t+1}(x))) - \ell(y, \pi_\ell^{\mathsf{BR}}(f_i^t(x)))] \\
=&\sum_{a\in\mathcal{A}} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(\ell(y,a) - \ell(y, a_i^t))\Delta_a^t(x)].
\end{aligned}$$

For each term in the summation, we can upper-bound it as

$$\begin{aligned}
&\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(\ell(y,a) - \ell(y, a_i^t))\Delta_a^t(x)] && \text{(F.7)}\\
=&\langle \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[y\Delta_a^t(x)], \ell_a - \ell_{a_i^t}\rangle && \text{(Linearity of Expectation)}\\
\le&\langle \mathop{\mathbb{E}}_{x\sim\mathcal{D}_{[k]}}[f_i^{t+1}(x)\Delta_a^t(x)], \ell_a - \ell_{a_i^t}\rangle + \beta\sqrt{d} && \text{(Since } f_i^{t+1} \text{ is } \beta\text{-calibrated)}\\
\le&\beta\sqrt{d}. && \text{(Since } a \text{ is the new Best-response action)}
\end{aligned}$$

Summing these actions together, we have

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell(y, \pi_\ell^{\mathsf{BR}}(f_i^{t+1}(x))) - \ell(y, \pi_\ell^{\mathsf{BR}}(f_i^t(x)))] \le \beta\sqrt{d}A. \tag{F.8}$$

□

Instead of setting a fixed β, we can calculate a different β^t at each round, which allows a smaller increase in loss.

Lemma F.II.4. *At each round t, if a_i^t is not the best action on E^t in average, i.e.*

$$\delta^t = \max_{a\in\mathcal{A}} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(\ell(y, a_i^t) - \ell(y,a))E^t(x)] > 0,$$

then we can set $\beta^t \le \delta^t/\sqrt{d}$, such that

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell(y, \pi_\ell^{\mathsf{BR}}(f_i^{t+1}(x))) - \ell(y, \pi_\ell^{\mathsf{BR}}(f_i^t(x)))] \le 0.$$

Proof. We can write the change in loss at each round as

$$\begin{aligned}
&\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell(y,\pi_\ell^{\mathsf{BR}}(f_i^{t+1}(x))) - \ell(y,\pi_\ell^{\mathsf{BR}}(f_i^t(x)))] \\
=&\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell(y,\pi_\ell^{\mathsf{BR}}(f_i^{t+1}(x))) - \ell(y,a_i^t))E^t(x)] \\
=&\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell(y,\pi_\ell^{\mathsf{BR}}(f_i^{t+1}(x))) - \ell(y,a'))E^t(x)] + \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell(y,a') - \ell(y,a_i^t))E^t(x)] \\
=&\sum_{a\in\mathcal{A}}\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(\ell(y,a) - \ell(y,a'))\Delta_a^t(x)] + \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell(y,a') - \ell(y,a_i^t))E^t(x)],
\end{aligned}$$

for any $a' \in \mathcal{A}$.

We can use the same analysis as in Lemma F.II.3 to get

$$\sum_{a\in\mathcal{A}}\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(\ell(y,a) - \ell(y,a'))\Delta_a^t(x)] \le \beta^t\sqrt{d}.$$

For the second term, we would want the loss to be as small as possible, so we can choose $a' = \operatorname*{argmin}_{a\in\mathcal{A}} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell(y,a)\cdot E^t(x)]$ and let

$$\delta^t = -\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(\ell(y,a') - \ell(y,a_i^t))E^t(x)], \tag{F.9}$$

then δ^t is maximized and $\delta^t \ge 0$ by definition.

The total change in loss in this round can be written as

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell(y,\pi_\ell^{\mathsf{BR}}(f_i^{t+1}(x))) - \ell(y,\pi_\ell^{\mathsf{BR}}(f_i^t(x)))] \le \beta^t\sqrt{d} - \delta^t.$$

If $\delta^t > 0$, we can set $\beta^t \le \delta^t/\sqrt{d}$ to ensure the loss does not increase at this round.

□

F.II.1 Proof of Theorem 9.3.2

Proof. By Lemma F.II.1 and F.II.2, for any $i \in \{1, 2\}$, at time-step t, we have the inequality

$$B(f_i^t, \mathcal{D}) - B(f_i^{t+1}, \mathcal{D}) \geq \frac{\alpha^2 \eta}{4d}.$$

Taking the sum over all time-steps, we have for any $i \in \{1, 2\}$,

$$B(f_i, \mathcal{D}) - B(f_i^T, \mathcal{D}) \geq T_i \cdot \frac{\alpha^2 \eta}{4d}.$$

Since the Brier score is always non-negative, we have

$$T_i \leq \frac{4d \cdot B(f_i, \mathcal{D})}{\alpha^2 \eta}.$$

Second, using Lemma F.II.3 and summing over all time-steps, we have

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell(y, \pi_\ell^{\mathsf{BR}}(f_i^T(x))] - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell(y, \pi_\ell^{\mathsf{BR}}(f_i(x))] \tag{F.10}$$

$$= \sum_{t=1}^{T} \mathbb{I}[i_t = i] \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell(y, \pi_\ell^{\mathsf{BR}}(f_i^{t+1}(x))) - \ell(y, \pi_\ell^{\mathsf{BR}}(f_i^t(x)))] \tag{F.11}$$

$$\leq T_i \cdot \beta\sqrt{d}A. \tag{F.12}$$

Finally, the halting condition implies that $\mu(E_{\ell,a_1,a_2}) < \eta$ for all $a_1, a_2 \in \mathcal{A}$. □

F.III Proofs of Section 9.3.2: Finite Sample Analysis

First, to make our argument that in-sample quantities translate to out-sample quantities, it is useful for the patching operations to use values that are rounded to a finite grid, rather than the precise value from the arbitrary sample. We define the finite grid as follows:

Definition F.III.1. *For any integer $m > 0$, let $1/m$ denote the $m + 1$ grid points,*

$$\left[\frac{1}{m}\right] = \left\{0, \frac{1}{m}, \frac{2}{m}, \ldots, \frac{m-1}{m}, 1\right\}.$$

For any value $v \in [0,1]^d$, let $Round(v;m) = \operatorname{argmin}_{v' \in [1/m]^d} \|v - v'\|_2$ denote the closest grid point to v in $[1/m]^d$.

At each time-step in Line 1 and Line 1, denote $\widetilde{\phi}^t = Round(\phi; m)$, and we patch the predictors using $\widetilde{\phi}^t$ instead of ϕ^t, *i.e.,* we update f^t to f^{t+1} as

$$f^{t+1}(x) = \operatorname{proj}_{[0,1]^d}(f^t(x) + \widetilde{\phi}^t E^t(f^t(x), x)).$$

With this new patching operation, we can perform a similar analysis in Section 9.3.1 to show that the Brier score decreases at each iteration, and therefore the algorithm terminates within a finite number of time-steps. We denote the maximum number of time-steps, counting both Line 1 and Line 1, as $T_{\max}$.

Then, we can count the total number of possible predictors outputted by Line 1 by observing that, for a fixed pair of input predictors, each pair of output predictors can be encoded as a sequence of tuple, $\{(i^t, E^t, \Delta^t)\}_{t\in[T]}$. Here, index $i^t \in \{1,2\}$, event $E^t \in \{E_{\ell,a_1,a_2} : \ell \in \mathcal{L}, a_1, a_2 \in \mathcal{A}\} \cup \{E_{\ell,a} \cap E_{\ell',a_1,a_2} : \ell, \ell' \in \mathcal{L}, a_1, a_2, a \in \mathcal{A}\}$, and $\Delta_t \in [1/m]^d$ are all chosen from a finite set, and the length of the sequence, T, is also bounded. Specifically, for a fixed input f_1, f_2, we denote S to be the set of all possible predictors outputted by Line 1. Then, its size satisfies

$$|S| \le (4|\mathcal{L}|^2 A^3 (m+1)^d)^{T_{\max}+1}$$

We show that the number of predictors outputted by Line 1 is bounded:

Lemma F.III.2. *Fix any pair of predictors $f_1, f_2 : [k] \to [0,1]^d$ and any $\eta, \alpha, \beta > 0$. Then the total number of possible predictors outputted by Line 1 is at most $|S|$ such that, for any distribution $\mathcal{D}$ on which Line 1 is run, the output predictors $(f_1^t, f_2^t) \in S$.*

Proof. First, notice that a sequence of quantities $\{(i^t, E^t, \Delta^t)\}_{t\in[T]}$ defines the pair of predictors outputted by Line 1.

Let S denote the pairs of functions induced by all such trajectories defined above. Here, $i^t \in \{1,2\}, E^t \in \{E_{\ell,a_1,a_2} : \ell \in \mathcal{L}, a_1, a_2 \in \mathcal{A}\} \cup \{E_{\ell,a} \cap E_{\ell',a_1,a_2} : \ell, \ell' \in \mathcal{L}, a_1, a_2, a \in \mathcal{A}\}$, and $\Delta_t \in [1/m]^d$. Therefore, there are

$$|S| \leq \sum_{t=1}^{T} \left(2(|\mathcal{L}|A^2 + |\mathcal{L}|^2 A^3)(m+1)^d\right)^t \leq (4|\mathcal{L}|^2 A^3 (m+1)^d)^{T_{\max}+1}$$

output predictors. □

F.III.1 Finite Grid

With this new patching operation, we can show that the Brier score decreases on the empirical distribution D, corresponding to Lemma F.II.2:

Lemma F.III.3. *Fix any event E. Let $\phi = \mathbb{E}_{(x,y)\sim D}[y - f(x)|E(x) = 1]$. For any predictor f, we patch f as $f'(x) = \text{proj}_{\Delta}(f(x) + \tilde{\phi}E(x))$. Then,*

$$B(f, D) - B(f', D) \geq \|\phi\|_2^2 \mu(E) - \frac{d}{4m^2}$$

Proof. Let $\tilde{f}'(x) = f(x) + \phi E(x)$. Then, we have

$$\begin{aligned} B(f, D) - B(f', D) =& B(f, D) - B(\tilde{f}', D) + B(\tilde{f}', D) - B(f', D) \\ \geq& \|\phi\|_2^2 \mu(E) + \mathbb{E}[\|f(x) + \phi E(x) - y\|_2^2 - \left\|f(x) + \tilde{\phi}E(x) - y\right\|_2^2] \\ & \hfill \text{(Lemma F.II.2)} \\ =& \|\phi\|_2^2 \mu(E) - \mathbb{E}\left[\left\|\tilde{\phi} - \phi\right\|_2^2\right] \mu(E) \end{aligned}$$

By definition $\tilde{\phi}$, we know that each index of $|\tilde{\phi} - \phi|$ is in $[0, \frac{1}{2m}]$. Therefore, we have

$$B(f, D) - B(f', D) \geq \|\phi\|_2^2 \cdot \mu(E) - \frac{d}{4m^2}.$$

□

Since the Brier score is within the range $[0, d]$, and it decreases at each iteration, we can show that, if we set m large enough, Line 1 terminates within a finite number of iterations:

Lemma F.III.4. *For any predictor f_1, f_2. Let $m \geq \left\lceil \sqrt{\frac{d}{2\min\{\beta^2, \eta\alpha^2/4d\}}} \right\rceil$. The Brier score at each iteration of Line 1 and Line 1 satisfies*

$$B(f^t, D) - B(f^{t+1}, D) > \frac{\min\{\beta^2, \eta\alpha^2/(4d)\}}{2}.$$

Counting both Line 1 and its subroutine Line 1, the total number of iterations T satisfies

$$T \leq \frac{2d}{\min\{\beta^2, \eta\alpha^2/4d\}}.$$

Proof. By Line 1, we have by definition of β-decision calibration that

$$\mu(E^t \cap E_{\ell,a}) \cdot \left\| \mathop{\mathbb{E}}_{(x,y)\sim D}[(y - f(x)|E^t(x) \cdot E_{\ell,a}(x) = 1] \right\|_2^2 \tag{F.13}$$

$$\geq \mu(E^t \cap E_{\ell,a})^2 \cdot \left\| \mathop{\mathbb{E}}_{(x,y)\sim D}[(y - f(x)|E^t(x) \cdot E_{\ell,a}(x) = 1] \right\|_2^2 \tag{F.14}$$

$$= \left\| \mathop{\mathbb{E}}_{(x,y)\sim D}[(y - f(x) \cdot E^t(x) \cdot E_{\ell,a}(x)] \right\|_2^2 > \beta^2. \tag{F.15}$$

In Line 1, we have by Lemma F.II.1 that

$$\mu(E^t) \left\| \mathop{\mathbb{E}}_{(x,y)\sim D}[(y - f(x)|E^t(x) = 1] \right\|_2^2 > \frac{\eta\alpha^2}{4d}. \tag{F.16}$$

Therefore, for any ϕ and event E that we patch in Line 1 or Line 1, they satisfy

$$\|\phi\|_2^2 \cdot \mu(E) > \min\{\beta^2, \frac{\eta\alpha^2}{4d}\}.$$

Letting $m \geq \left\lceil \sqrt{\frac{d}{2\min\{\beta^2, \eta\alpha^2/4d\}}} \right\rceil$, we can ensure

$$B(f^t, D) - B(f^{t+1}, D) \geq \frac{\|\phi\|_2^2 \cdot \mu(E)}{2} > \frac{\min\{\beta^2, \eta\alpha^2/(4d)\}}{2}.$$

Since the Brier score is in the range $[0, d]$, we can bound the total number of iterations of both Line 1 and Line 1 as

$$T \leq \frac{2d}{\min\{\beta^2, \eta\alpha^2/4d\}}.$$

□

F.III.2 Proof of Theorem 9.3.4

First, we show that, for a fixed predictor f and event $E_{\ell,a}$, the in-sample prediction error is approximately accurate. The deviation bound of the Brier score and calibration error can then be directly implied.

Lemma F.III.5. *Fix any f, $E_{\ell,a}$, with probability at least $1-\delta'$, we have*

$$\left\| \mathop{\mathbb{E}}_{\mathcal{D}}[(y-f(x))E_{\ell,a}(f(x),x)] - \frac{1}{n}\sum_{i=1}^{n}[(y_i - f(x_i))E_{\ell,a}(f(x_i),x_i)] \right\|_2 \leq \sqrt{\frac{d\ln(2d/\delta')}{2n}}.$$

Proof. Fix an index $j \in [d]$, we know $\mathbb{E}_{(x,y)\sim\mathcal{D}}[(y-f(x))_j \cdot E_{\ell,a}(f(x),x)] \in [0,1]$ and

$$\mathop{\mathbb{E}}_{D}\left[\frac{1}{n}\sum_{i=1}^{n}[(y_i - f(x_i))_j \cdot E_{\ell,a}(f(x_i),x_i)]\right] = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(y-f(x))_j \cdot E_{\ell,a}(f(x),x)].$$

Since (x_i, y_i) is drawn i.i.d. from $\mathcal{D}$, we can use Hoeffding's inequality to get, with probability δ'/d,

$$\left| \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(y-f(x))_j \cdot E_{\ell,a}(f(x),x)] - \frac{1}{n}\sum_{i=1}^{n}[(y_i - f(x_i))_j \cdot E_{\ell,a}(f(x_i),x_i)] \right|$$
$$\leq \sqrt{\frac{\ln(2d/\delta')}{2n}}$$

Using union bound, we have that with probability $1-\delta'$, the above inequality holds for all $j \in [d]$. Then, we have

$$\left\| \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(y - f(x)) \cdot E_{\ell,a}(f(x), x)] - \frac{1}{n}\sum_{i=1}^{n}[(y_i - f(x_i)) \cdot E_{\ell,a}(f(x_i), x_i)] \right\|_2$$
$$\leq \sqrt{\sum_{j\in d}\left(\sqrt{\frac{\ln(2d/\delta')}{2n}}\right)^2} = \sqrt{\frac{d\ln(2d/\delta')}{2n}}.$$

□

The deviation bound of the Brier score and calibration error can be directly implied by Lemma F.III.5. We summarize them in the lemmas below:

Lemma F.III.6. *For a fixed f, with probability at least $1 - \delta'$, $|B(f, \mathcal{D}) - B(f, D)| \leq \sqrt{\frac{d\ln(2d/\delta')}{2n}}$.*

Proof. Using triangle inequality, we have

$$\begin{aligned} |B(f,\mathcal{D}) - B(f,D)| &= \left| \left\| \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[y - f(x)] \right\|_2 - \left\| \frac{1}{n}\sum_{i=1}^{n}[y_i - f(x_i)] \right\|_2 \right| \\ &\leq \left\| \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[y - f(x)] - \frac{1}{n}\sum_{i=1}^{n}[y_i - f(x_i)] \right\|_2 \\ &\leq \sqrt{\frac{d\ln(2d/\delta')}{2n}}. \end{aligned}$$

□

Lemma F.III.7. *For a fixed f, any loss function $\ell \in \mathcal{L}$ and $\mathcal{E} = \{E_{\ell,a} \cap E_{\ell',a_1,a_2} : \ell, \ell' \in \mathcal{L}, a, a_1, a_2 \in \mathcal{A}\}$, with probability $1 - \delta'$, we have*

$$\left\| \mathop{\mathbb{E}}_{\mathcal{D}}[(y - f(x))E(x)] - \frac{1}{n}\sum_{i=1}^{n}[(y_i - f(x_i))E(x)] \right\|_2 \leq \sqrt{\frac{3d\ln(2dA|\mathcal{L}|/\delta')}{2n}}$$

for all $E \in \mathcal{E}$.

Proof. The claim follows by using a union bound over the events in $\mathcal{E}^t$, using Lemma F.III.5, and that $|\mathcal{E}^t| = A^3|\mathcal{L}|^2$. □

For a fixed pair of predictors, we can also show that the empirical size of the disagreement events E_{ℓ,a_1,a_2} is approximately correct with high probability:

Lemma F.III.8. *Fix any pair of predictors $(f_1, f_2) \in S$, with probability at least $1-\delta'$ over D, we have*

$$\left|\mu(E_{\ell,a_1,a_2}) - \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[E_{\ell,a_1,a_2}(x_i) = 1]\right| \le \sqrt{\frac{2\ln(2A|\mathcal{L}|/\delta')}{2n}}.$$

for all $a_1, a_2 \in \mathcal{A}$ with $a_1 \neq a_2$ and for all $\ell \in \mathcal{L}$.

Proof. We know $\mathbb{I}[E_{\ell,a_1,a_2}(x_i) = 1] \in [0, 1]$ and

$$\mathop{\mathbb{E}}_{D}\left[\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[E_{\ell,a_1,a_2}(x_i)\right] = \mu(E_{\ell,a_1,a_2}).$$

Since (x_i, y_i) is drawn i.i.d. from $\mathcal{D}$, we can use Hoeffding's inequality to get, with probability $1 - \delta'/(A^2|\mathcal{L}|)$,

$$\left|\mu(E_{\ell,a_1,a_2}) - \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[E_{\ell,a_1,a_2}(x_i)]\right| \le \sqrt{\frac{2\ln(2A|\mathcal{L}|/\delta')}{2n}}.$$

Using union bound over all pairs of $a_1, a_2 \in \mathcal{A}$ and $\ell \in \mathcal{L}$, we know the above inequality holds for all a_1, a_2 and $\ell \in \mathcal{L}$ with probability at least $1 - \delta'$. $\square$

We summarize the above results in the theorem below. Theorem 9.3.4 follows by solving for n in the 2-4th guarantees below.

Theorem F.III.9. *Fix any distribution $\mathcal{D}$ and dataset $D \sim \mathcal{D}$ containing n samples drawn i.i.d from $\mathcal{D}$. For any pair of predictors $f_1, f_2 : [k] \to [0, 1]^d$, loss margin $\alpha > 0$, disagreement region mass $\eta > 0$, and decision-calibration tolerance $\beta > 0$, Line 1 run over the empirical distribution D updates predictors f_1 and f_2 for T_1 and T_2 time-steps, respectively, and outputs a pair of predictors (f_1^T, f_2^T) such that, with probability at least $1 - \delta$ over the randomness of $D \sim \mathcal{D}^n$,*

1. *The total number of time-steps for Line 1 and Line 1 is*

$$T = T_1 + T_2 \le \frac{2d}{\min\{\beta^2, \eta\alpha^2/4d\}}$$

2. *For $i \in \{1, 2\}$, the Brier scores of the final models are lower than that of the input models:*

$$B(f_i^{T_i}, \mathcal{D}) \leq B(f_i, \mathcal{D}) - T_i \cdot \min\left\{\beta^2, \eta\alpha^2/(4d)\right\} + \sqrt{(d \ln(6d|S|/\delta)/(2n)}$$

3. *For $i \in \{1, 2\}$, the downstream decision-making losses of the final models do not increase by much compared to that of the input models:*

$$\begin{aligned}&\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell(y, \pi^{\mathsf{BR}}(f_i^T(x)) - \ell(y, \pi^{\mathsf{BR}}(f_i(x))]\\ &\leq \left(\beta + \sqrt{(3d\ln(6dA|S||\mathcal{L}|/\delta)/(2n)}\right)\sqrt{d}AT_i\end{aligned}$$

4. *The final models approximately agree on their best-response actions almost everywhere. That is, the disagreement region E_{a_1,a_2} calculated using f_1^T, f_2^T has small mass.*

$$\mu(E_{\ell,a_1,a_2}) \leq \eta + \sqrt{(2\ln(6A|S||\mathcal{L}|/\delta)/(2n)} \quad \textit{for all } a_1, a_2 \in \mathcal{A} \quad \textit{s.t } a_1 \neq a_2$$

Here, S is the set of all possible predictors outputted by Line 1 satisfying

$$\ln(|S|) \leq \left(\frac{2d}{\min\{\beta^2, \eta\alpha^2/4d\}} + 1\right) \ln\left(4|\mathcal{L}|^2 A^3 \left(\left\lceil\sqrt{\frac{d}{2\min\{\beta^2, \eta\alpha^2/4d\}}}\right\rceil + 1\right)^d\right)$$

Proof. The upper bound on T holds true with probability 1. For the remaining three guarantees, we show that each of them holds with probability at least $1 - \delta/3$ over the randomness of D.

Brier Score. First, by Lemma F.III.6 and using union bound over all possible output predictors $(f_1, f_2) \in S$, we have with probability at least $1 - \delta/3$ that

$$|B(f_i, \mathcal{D}) - B(f_i, D)| \leq \sqrt{\frac{d\ln(6d|S|/\delta)}{2n}}.$$

By Lemma F.III.4, and summing over all iterations, we have

$$B(f_i^{T_i}, D) \leq B(f_i, D) - T_i \cdot \min\left\{\frac{\beta^2}{2}, \frac{\eta\alpha^2}{8d}.\right\}$$

Therefore,

$$\begin{aligned} B(f_i^{T_i}, \mathcal{D}) \leq & B(f_i^{T_i}, D) + \sqrt{\frac{d\ln(6d|S|/\delta)}{2n}} \\ \leq & B(f_i, \mathcal{D}) - T_i \cdot \min\left\{\frac{\beta^2}{2}, \frac{\eta\alpha^2}{8d}\right\} + \sqrt{\frac{d\ln(6d|S|/\delta)}{2n}} \end{aligned}$$

for $i \in \{1, 2\}$.

Expected Loss. Using union bound over all predictors in S, by Lemma F.III.7, we have, with probability at least $1 - \delta/3$,

$$\begin{aligned} & \left\|\mathop{\mathbb{E}}_{\mathcal{D}}[(y - f(x))E_{\ell,a}(f(x), x)] - \frac{1}{n}\sum_{i=1}^{n}[(y_i - f(x_i))E_{\ell,a}(f(x_i), x_i)]\right\|_2 \\ & \leq \sqrt{\frac{3d\ln(6dA|S||\mathcal{L}|/\delta)}{2n}} \end{aligned}$$

for all predictors f, action $a \in \mathcal{A}$ and loss $\ell \in \mathcal{L}$. Using similar method as in Lemma 9.2.6, we define the set $\Delta_a^t \subseteq E^t$ as

$$\Delta_a^t = \{x \in E^t : \pi_\ell^{\mathsf{BR}}(f_i^{t+1}(x)) = a\}.$$

Then, we have

$$\begin{aligned} & \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell(y, \pi_\ell^{\mathsf{BR}}(f_i^{t+1}(x))) - \ell(y, \pi_\ell^{\mathsf{BR}}(f_i^t(x)))] \\ = & \sum_{a\in\mathcal{A}} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(\ell(y, a) - \ell(y, a_i^t))\Delta_a^t(x)]. \end{aligned}$$

For each term in the summation,

$$\begin{aligned}
&\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(\ell(y,a)-\ell(y,a_i^t))\Delta_a^t(x)] && \text{(F.17)}\\
=&\langle \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[y\Delta_a^t(x)], \ell_a-\ell_{a_i^t}\rangle && \text{(Linearity of Expectation)}\\
\leq&\langle \mathop{\mathbb{E}}_{x\sim\mathcal{D}_{[k]}}[f_i^{t+1}(x)\Delta_a^t(x)], \ell_a-\ell_{a_i^t}\rangle + \left(\beta+\sqrt{\frac{3d\ln(6dA|S||\mathcal{L}|/\delta)}{2n}}\right)\sqrt{d}\\
& && \text{(Lemma F.III.7 and } \beta\text{-calibrated)}\\
\leq&\left(\beta+\sqrt{\frac{3d\ln(6dA|S||\mathcal{L}|/\delta)}{2n}}\right)\sqrt{d}. && \text{(Since } a \text{ is the new Best-response action)}
\end{aligned}$$

Summing these actions together, we have

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell(y,\pi_\ell^{\mathsf{BR}}(f_i^{t+1}(x)))-\ell(y,\pi_\ell^{\mathsf{BR}}(f_i^t(x)))]\leq\beta\sqrt{d}A+\sqrt{\frac{3\ln(6dA|S||\mathcal{L}|/\delta)}{2n}}dA. \tag{F.18}$$

Summing over all iterations, we conclude that, with probability at least $1-\delta/3$,

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell(y,\pi^{\mathsf{BR}}(f_i^{T_i}(x)))-\ell(y,\pi^{\mathsf{BR}}(f_i(x)))]\leq T_i(\beta\sqrt{d}A+\sqrt{\frac{3\ln(6dA|S||\mathcal{L}|/\delta)}{2n}}dA)$$

for all $i\in\{1,2\}$.

Disagreement Event. By Lemma F.III.8, with probability at least $1-\delta/(3|S|)$, we have, for all $\ell\in\mathcal{L}$, $a_1,a_2\in\mathcal{A}$ with $a_1\neq a_2$ and all $(f_1,f_2)\in S$,

$$\left|\mu(E_{\ell,a_1,a_2}(f_1(x),f_2(x),x)-\frac{1}{n}\sum_{i=1}^n\mathbb{I}[E_{\ell,a_1,a_2}(f_1(x_i),f_2(x_i),x_i))]\right|\leq\sqrt{\frac{2\ln(6A|S||\mathcal{L}|/\delta)}{2n}}.$$

From the while loop condition in Line 1, we know that

$$\frac{1}{n}\sum_{i=1}^n\mathbb{I}[E_{\ell,a_1,a_2}(f_1^{T_1}(x_i),f_2^{T_2}(x_i),x_i)]\leq\eta$$

Then, using union bound over all $(f_1,f_2)\in S$ and $\ell\in\mathcal{L}$, with probability at least $1-\delta/3$, we have the guarantee

$$
\begin{aligned}
&\mu(E_{\ell,a_1,a_2}(f_1^{T_1}(x), f_2^{T_2}(x), x)) \\
\leq &\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[E_{\ell,a_1,a_2}(f_1^{T_1}(x_i), f_2^{T_2}(x_i), x_i)] + \sqrt{\frac{2\ln(6A|S||\mathcal{L}|/\delta)}{2n}} \qquad \text{(F.19)} \\
\leq &\eta + \sqrt{\frac{2\ln(6A|S||\mathcal{L}|/\delta)}{2n}} \qquad \text{(F.20)}
\end{aligned}
$$

for all $a_1, a_2 \in \mathcal{A}, a_1 \neq a_2$ and $\ell \in \mathcal{L}$.

Finally, using results from Lemma F.III.2, value of $T_{\max}$, and $m = \left\lceil \sqrt{\frac{d}{2\min\{\beta^2, \eta\alpha^2/4d\}}} \right\rceil$, we conclude by showing

$$
\begin{aligned}
\ln(|S|) \leq &\ln\left((4|\mathcal{L}|^2 A^3 (m+1)^d)^{T_{\max}+1}\right) \\
= &(T_{\max}+1)\ln(4|\mathcal{L}|^2 A^3 (m+1)^d) \\
= &\left(\frac{2d}{\min\{\beta^2, \eta\alpha^2/4d\}} + 1\right)\ln\left(4|\mathcal{L}|^2 A^3 \left(\left\lceil \sqrt{\frac{d}{2\min\{\beta^2, \eta\alpha^2/4d\}}} \right\rceil + 1\right)^d\right)
\end{aligned}
$$

□

F.IV Proofs of Section 9.3.3: Extension to Multiple Predictors

Here, we present the algorithm we described in Section 9.3.3. As a shorthand, we write $E^{\alpha}_{\ell,a_1,a_2}(f(x), f_j(x), x)$ as $E_{\ell,a_1,a_2,j}$.

With Line 1, we can obtain the following guarantee.

Theorem F.IV.1. *Fix any distribution $\mathcal{D}$ and dataset $D \sim \mathcal{D}$ containing n samples drawn i.i.d from $\mathcal{D}$. For any k predictors $f_1, f_2, \ldots, f_k : [k] \to [0,1]^d$, family of loss functions $\mathcal{L}$, loss margin $\alpha > 0$, disagreement region mass $\eta > 0$, and decision-calibration tolerance $\beta > 0$, Line 1 run over the empirical distribution D updates each f_i for T_i time-steps, and outputs k predictors $(f_1^T, \ldots, f_k^T)$ such that, with probability at least $1-\delta$ over the randomness of $D \sim \mathcal{D}^n$,*

1. *The total number of time-steps for Line 1 and Line 1 is*

$$T = \sum_{i\in[k]} T_i \le \frac{2dk}{\min\{\beta^2, \eta\alpha^2/4d\}}$$

2. *For $i \in [k]$, the Brier scores of the final models are lower than that of the input models:*

$$B(f_i^{T_i}, \mathcal{D}) \le B(f_i, \mathcal{D}) - (T_i - \eta) \cdot \min\left\{\beta^2, \eta\alpha^2/(4d)\right\}$$

3. *For $i \in [k]$ and for all $\ell \in \mathcal{L}$, the downstream decision-making losses of the final models do not increase by much compared to that of the input models:*

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell(y, \pi_\ell^{\mathsf{BR}}(f_i^T(x)) - \ell(y, \pi_\ell^{\mathsf{BR}}(f_i(x))] \le 2T_i\beta A\sqrt{d}$$

4. *The final models approximately agree on their best-response actions almost everywhere. That is, the disagreement region $E_{\ell,a_1,a_2,j}$ calculated using f_1^T, f_j^T has small mass: For all $j \in [k]$, $\ell \in \mathcal{L}$, and $a_1, a_2 \in \mathcal{A}$ s.t. $a_1 \ne a_2$,*

$$\mu(E_{\ell,a_1,a_2,j}) \le 2\eta$$

if $n \ge \Omega\left(d^2k/(\eta^2\min\{\beta, \eta\alpha^2/d\}) \cdot (\ln(dkA|\mathcal{L}|) - \ln(\delta\min\{\beta, \eta\alpha/d\}))\right)$.

Proof. The proof follows the same idea and structure as the proof of theorem 9.3.4.

Number of Iterations. By lemma F.III.3, at each iteration of Line 1 and Line 1, the Brier score of some predictor $j \in [k]$ satisfies

$$B(f_j^t, D) - B(f_j^{t+1}, D) > \frac{\min\{\beta^2, \eta\alpha^2/(4d)\}}{2}.$$

Since the Brier score is non-negative and we have k predictors in total, we have that, after $T = \sum_{i=1}^k T_i$ iterations,

$$\sum_{j\in[k]} B(f_j, D) \ge \sum_{j\in[k]} (B(f_j, D) - B(f_j^T, D)) > T \cdot \frac{\min\{\beta^2, \eta\alpha^2/(4d)\}}{2}.$$

Since Brier score is in the range $[0, d]$, we can solve for T in the inequality and have the upper bound on the number of iterations

$$T_{\max} \leq \frac{2dk}{\min\{\beta^2, \eta\alpha^2/(4d)\}}.$$

Brier Score. We first count the number of possible predictors outputted by Line 1. At each iteration in Line 1 or its subroutine Line 1, we patch on a predictor f_i^t for some $i^t \in [k]$, on the event $E \in \{E_{\ell,a_1,a_2,j} : \ell \in \mathcal{L}, a_1, a_2 \in \mathcal{A}, j \in [k]\} \cup \{E_{\ell,a_1,a_2,j} \cap E_{\ell',a} : \ell, \ell' \in \mathcal{L}, a_1, a_2, a \in \mathcal{A}, j \in [k]\}$, with the value $\Delta_t \in [1/m]^d$. Therefore, at each iteration, the number of different patching operations satisfies

$$(|\mathcal{L}|A^2k + |\mathcal{L}|^2A^3k)(m+1)^d \leq 2|\mathcal{L}|^2A^3k(m+1)^d.$$

For fixed k predictors, the output predictors are defined by the sequence of patches. The total number of different sets of output predictors are

$$|S| \leq \sum_{i=1}^{T}(2|\mathcal{L}|^2A^3k(m+1)^d)^t \leq (2|\mathcal{L}|^2A^3k(m+1)^d)^{T_{\max}+1}.$$

Using lemma F.III.6 and using union bound over all possible output predictors, we have with probability at least $1 - \delta/3$ that

$$|B(f_i, D) - B(f_i, \mathcal{D})| \leq \sqrt{\frac{d\ln(6dk|S|/\delta)}{2n}}.$$

Summing over all iterations, we have

$$B(f_i^{T_i}, \mathcal{D}) \leq B(f_i^{T_i}, \mathcal{D}) - T_i \cdot \min\left\{\frac{\beta^2}{2}, \frac{\eta\alpha^2}{8d}\right\} + \sqrt{\frac{d\ln(6dk|S|/\delta)}{2n}}.$$

Expected Loss. Using lemma F.III.6 and a union bound over the events in $\mathcal{E} = \{E_{\ell,a_1,a_2,j} \cap E_{\ell',a} : \ell, \ell' \in \mathcal{L}, a_1, a_2, a \in \mathcal{A}, j \in [k]\}$, we have with probability $1 - \delta/3$ that

$$\|\mathop{\mathbb{E}}_{\mathcal{D}}[(y - f(x)]E(x) - \frac{1}{n}\sum_{i=1}^{n}[(y_i - f(x_i))E(x)]\| \leq \sqrt{\frac{3d\ln(6dkA|\mathcal{L}||S|/\delta)}{2n}}$$

for all predictors f, action $a \in \mathcal{A}$, loss $\ell \in \mathcal{L}$. The rest of the steps are the same as in the proof of theorem 9.3.4. We have, at iteration t,

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell(y, \pi_\ell^{\mathsf{BR}}(f_i^{t+1}(x))) - \ell(y, \pi_\ell^{\mathsf{BR}}(f_i^t(x)))] \le \beta\sqrt{d}A + \sqrt{\frac{3\ln(6dkA|\mathcal{L}||S|/\delta)}{2n}}dA.$$

The result follows by summing over all iterations.

Disagreement event is small. The stopping condition at line 2 of line 1 implies that the final output models satisfies

$$\mu(E_{\ell,a_1,a_2,j}) = \frac{1}{n}\sum_{i=1}^{n} E^{\alpha}_{\ell,a_1,a_2,j}(f_1^T(x_i), f_j^T(x_i), x_i) \le \eta$$

for all $a_1, a_2 \in \mathcal{A}$, $\ell \in \mathcal{L}$, and $j \in [k]\backslash\{1\}$. That is, the disagreement region between our base model f_1^T and any of the other $k-1$ models is small. Using Hoeffding's inequality and union bound, we have that, for any predictors $(f_1, \dots, f_k) \in S$, with probability $1 - \delta/3$ over D,

$$\left|\mu(E_{\ell,a_1,a_2,j} - \frac{1}{n}\sum_{i=1}^{n} E_{\ell,a_1,a_2,j}(x_i)\right| \le \sqrt{\frac{\ln(6kA|S||\mathcal{L}|/\delta)}{n}}$$

for all $a_1, a_2 \in \mathcal{A}$ with $a_1 \ne a_2$, all $\ell \in \mathcal{L}$, and predictor $j \in [k]$. Combining this with the bound on $\frac{1}{n}\sum_{i=1}^{n} E_{\ell,a_1,a_2,j}(x_i)$ given by the stopping condition, we can obtain the fourth guarantee. □

Note that, with the fourth condition in Theorem F.IV.1, we can also obtain that the size of the disagreement event between other predictors, f_i, f_j with $i, j \ne 1$ satisfies

$$\mu(E^{2\alpha}_{a_1,a_2,\ell}(f_i(x), f_j(x), x)) \le 2\eta A.$$

For each $x \in E^{2\alpha}_{a_1,a_2,\ell}(f_i(x), f_j(x), x))$, it must be either $x \in E^{\alpha}_{a_1,a,\ell}(f(x), f_i(x), x)$ or

$x \in E^{\alpha}_{a_1,a,\ell}(f(x), f_j(x), x)$ for some $a \in \mathcal{A}$. Therefore,

$$\begin{aligned}
&\mu(E^{2\alpha}_{a_1,a_2,\ell}(f_i(x), f_j(x), x)) \\
\leq &\sum_{a \in \mathcal{A}\setminus\{a_1,a_2\}} \big(\mu(E^{\alpha}_{a_1,a_2,\ell}(f(x), f_i(x), x)) + \mu(E^{\alpha}_{a_1,a_2,\ell}(f(x), f_j(x), x))\big) \\
\leq &\sum_{a \in \mathcal{A}\setminus\{a_1,a_2\}} 2\eta \leq 2A\eta.
\end{aligned}$$

F.V Extension to Infinitely Large Action Set

We extend our results to a setting where the action space $\mathcal{A}$ of the downstream decision-making tasks is infinitely large. Now that we cannot enumerate all actions in the action space, the disagreement event in this setting can not specify the best-response action. For a loss $\ell \in \mathcal{L}$, one approach is to define the event as

$$\begin{aligned}
E^{\alpha}_{\ell}(f_1(x), f_2(x), x) = \mathbb{I}\big[&\langle f_1(x),\ \ell_{\pi^{\mathsf{BR}}(f_2(x))} - \ell_{\pi^{\mathsf{BR}}(f_1(x))}\rangle > \alpha, or \\
&\langle f_2(x),\ \ell_{\pi^{\mathsf{BR}}(f_1(x))} - \ell_{\pi^{\mathsf{BR}}(f_2(x))}\rangle > \alpha\big].
\end{aligned}$$

We now want to "patch" a predictor on the disagreement event, such that the predictor's Brier score decreases by a large amount after patching. However, using this definition of disagreement event, we cannot guarantee that at least one of f_1 and f_2 has a large prediction error. For example, consider $y_1 = y_2 = [1/2, 1/2]$, predictors $f_1(x_1) = f_2(x_2) = [1, 0]$, $f_1(x_2) = f_2(x_1) = [0, 1]$, and the loss $\ell_a \in \{[x, 1-x] : x \in [0, 1]\}$. According to f_1, the loss vector of the best-response action of x_1 is $\ell_{a_1} = [0, 1]$, and that of x_2 is $\ell_{a_2} = [1, 0]$. According to f_2, the loss vector of the best-response action of x_1 and x_2 is ℓ_{a_2} and ℓ_{a_1}, respectively. Clearly, data x_1 and x_2 are both in the disagreement event E^{α}_{ℓ} for any $\alpha < 1$, but the mean estimation error of f_1 and f_2 on the disagreement event $\{x_1, x_2\}$ are both 0, since $1/2([0, 1] + [1, 0]) = [1/2, 1/2]$.

A solution to this problem is to further separate the E^{α}_{ℓ} into $2d$ sub-events that focus on the disagreement per-label. The union of these $2d$ sub-events is equivalent to E^{α}_{ℓ}, as shown in Lemma F.V.3. Formally, we have the following definition.

Definition F.V.1 (Per-label Disagreement Event)**.** *For each a loss $\ell \in \mathcal{L}$, index $i \in [d]$,* *we define*

$$E^{\alpha}_{\ell,i,>}(f_1(x), f_2(x), x) = \mathbb{I}\big[f_1(x)_i - f_2(x)_i > \alpha/d, \quad E^{\alpha}_{\ell}(f_1(x), f_2(x), x) = 1\big].$$
$$E^{\alpha}_{\ell,i,<}(f_1(x), f_2(x), x) = \mathbb{I}\big[f_1(x)_i - f_2(x)_i < -\alpha/d, \quad E^{\alpha}_{\ell}(f_1(x), f_2(x), x) = 1\big].$$

To ensure the predictor's expected loss is accurate, we define a per-label best-response event using a discretized approach, approximating each index of the loss vector ℓ_a with a finite grid $[1/\mathcal{T}]$ (Definition F.III.1). Then, if a predictor is approximately accurate with respect to each per-label best-response event, then its estimated loss for the best-response actions are approximately accurate.

Definition F.V.2 (Per-label Best-Response Event)**.** *Given a predictor f, a loss ℓ, and a level $\tau \in [1/\mathcal{T}]$, for each index $i \in [d]$, we define the best-response event as*

$$E_{i,\tau}(f(x), x) = \mathbb{I}\{x : \tau - 1/\mathcal{T} < \ell_{\pi^{\mathsf{BR}}(f(x)),i} \leq \tau\}.$$

where $\ell_{\pi^{\mathsf{BR}}(f(x)),i}$ denote the i-th index of $\ell_{\pi^{\mathsf{BR}}(f(x))}$.

With the disagreement and best-response events now clearly defined, we can adapt our existing algorithm (Line 1) to this new context: at each iteration, we find the largest per-label disagreement events and patch on the predictor with the larger estimation error; then, in a subroutine similar to Line 1, we iteratively calibrate the predictor we just patched on, using the per-label best-response events. This algorithm can maintain similar guarantees as Theorem 9.3.4, with a data size $\tilde{\Omega}\left(d^2/(\eta^2\beta\min\{\beta^2, \eta\alpha^2/d^3\})\right)$. The formal algorithm is written in Line 1.

Given our approximation of each action feature ℓ_a using a finite grid $[1/\mathcal{T}]$, it may seem intuitive to approximate the entire infinite action set as $[1/\mathcal{T}]^d$ and simply run Line 1. However, such an approach yields an action set whose size grows exponentially with d, which in turn would require a data sample size that also grows exponentially to achieve guarantees comparable to those in Theorem 9.3.4. In contrast, our algorithm 1 operates with a polynomial data requirement relative to d, making it more feasible for practical application.

We now present the formal proofs related to Line 1.

Lemma F.V.3. *For any loss $\ell \in \mathcal{L}$, we have*

$$E_\ell^\alpha = \bigcup_{i\in[d],\cdot\in\{>,<\}} E_{\ell,i,\cdot}^\alpha.$$

Proof. By definition, we may get

$$\bigcup_{i\in[d],\cdot\in\{>,<\}} E_{\ell,i,\cdot}^\alpha \subseteq E_\ell^\alpha.$$

For the other direction, assume for the sake of contradiction that there is an element $x \in E_\ell^\alpha$ such that $x \notin E_{\ell,i,\cdot}^\alpha$ for all $i \in [d]$ and $\cdot \in \{>,<\}$. Then, this means

$$|f_1(x)_i - f_2(x)_i| < \alpha/d$$

for all $i \in [d]$. We can then get the inequality

$$\begin{aligned}
\langle f_1(x) - f_2(x), \ell_{\pi^{\mathsf{BR}}(f_2(x))} - \ell_{\pi^{\mathsf{BR}}(f_1(x))}\rangle &= \sum_{i\in[d]} (f_1(x)_i - f_2(x)_i)\cdot(\ell_{\pi^{\mathsf{BR}}(f_2(x))} - \ell_{\pi^{\mathsf{BR}}(f_1(x))}) \\
&\le \sum_{i\in[d]} |f_1(x)_i - f_2(x)_i| \cdot \left|\ell_{\pi^{\mathsf{BR}}(f_2(x))} - \ell_{\pi^{\mathsf{BR}}(f_1(x))}\right| \\
&< d\cdot\alpha/d = \alpha.
\end{aligned}$$

On the other hand, by definition of event E_ℓ^α, we have for all $x \in E_\ell^\alpha$,

$$\langle f_1(x) - f_2(x), \ell_{\pi^{\mathsf{BR}}(f_2(x))} - \ell_{\pi^{\mathsf{BR}}(f_1(x))}\rangle > \alpha,$$

which leads to a contradiction. □

The following lemma shows that, if f is β-index-wise decision-calibrated, then the loss estimation is approximately accurate.

Lemma F.V.4. *If the predictor f is β-index-wise decision-calibrated, i.e. for all $\ell \in \mathcal{L}$, $i \in [d]$, $\tau \in [1/\mathcal{T}]$,*

$$\mathop{\mathbb{E}}_{(x,y)\in\mathcal{D}}[\|y - f(x)\|_2 \cdot E_{\ell,i,\tau}(f(x), x)] \le \beta,$$

then, the loss estimation given by f is approximately accurate for all $\ell \in \mathcal{L}$:

$$\mathop{\mathbb{E}}_{(x,y)\in\mathcal{D}}[\langle y, \ell_{\pi^{\mathsf{BR}}(f(x))}\rangle - \mathop{\mathbb{E}}_{(x,y)\in\mathcal{D}}[\langle f(x), \ell_{\pi^{\mathsf{BR}}(f(x))}\rangle] \leq d\mathcal{T}\beta + d/\mathcal{T}.$$

Letting $\mathcal{T} \geq 1/\sqrt{\beta}$, we get

$$\mathop{\mathbb{E}}_{(x,y)\in\mathcal{D}}[\langle y, \ell_{\pi^{\mathsf{BR}}(f(x))}\rangle - \mathop{\mathbb{E}}_{(x,y)\in\mathcal{D}}[\langle f(x), \ell_{\pi^{\mathsf{BR}}(f(x))}\rangle] \leq 2d\sqrt{\beta}.$$

Proof. We can rewrite the expected loss as

$$\begin{aligned}
\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\langle y, \ell_{\pi^{\mathsf{BR}}(f(x))}\rangle] &= \sum_{i\in[d]} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[y_i \cdot \ell_{\pi^{\mathsf{BR}}(f(x)),i}] \\
&\qquad \text{(here } \ell_{\pi^{\mathsf{BR}}(f(x)),i} \text{ represents the } i\text{-th index of } \ell_{\pi^{\mathsf{BR}}(f(x))}) \\
&= \sum_{i\in[d]} \sum_{\tau\in[1/\mathcal{T}]} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[y_i \cdot \ell_{\pi^{\mathsf{BR}}(f(x)),i} \cdot E_{\ell,i,\tau}(f(x),x)] \\
&\leq \sum_{i\in[d]} \sum_{\tau\in[1/\mathcal{T}]} \tau \cdot \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[y_i \cdot E_{\ell,i,\tau}(f(x),x)].
\end{aligned}$$

Similarly, we can rewrite the estimated loss as

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\langle f(x), \ell_{\pi^{\mathsf{BR}}(f(x))}\rangle] > \sum_{i\in[d]} \sum_{\tau\in[1/\mathcal{T}]} (\tau - 1/\mathcal{T}) \cdot \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[f(x)_i \cdot E_{\ell,i,\tau}(f(x),x)].$$

The estimation error is

$$\begin{aligned}
&\mathop{\mathbb{E}}_{(x,y)\in\mathcal{D}}[\langle y, \ell_{\pi^{\mathsf{BR}}(f(x))}\rangle - \mathop{\mathbb{E}}_{(x,y)\in\mathcal{D}}[\langle f(x), \ell_{\pi^{\mathsf{BR}}(f(x))}\rangle] \\
=& \sum_{i\in[d]} \sum_{\tau\in[1/\mathcal{T}]} \tau \cdot \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[y_i \cdot E_{\ell,i,\tau}(f(x),x)] \\
&- \sum_{i\in[d]} \sum_{\tau\in[1/\mathcal{T}]} (\tau - 1/\mathcal{T}) \cdot \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[f(x)_i \cdot E_{\ell,i,\tau}(f(x),x)] \\
=& \sum_{i\in[d]} \sum_{\tau\in[1/\mathcal{T}]} \tau \cdot \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(y_i - f(x)_i) \cdot E_{\ell,i,\tau}(f(x),x)] \\
&+ \sum_{i\in[d]} \sum_{\tau\in[1/\mathcal{T}]} 1/\mathcal{T} \cdot \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[f(x)_i \cdot E_{\ell,i,\tau}(f(x),x)].
\end{aligned}$$

Since f is β-index-wise decision-calibrated, we have

$$\mathop{\mathbb{E}}_{(x,y)\in\mathcal{D}}[|y_i - f(x)_i| \cdot E_{\ell,i,\tau}(f(x),x)] \leq \beta.$$

Therefore,

$$\mathop{\mathbb{E}}_{(x,y)\in\mathcal{D}}[\langle y, \ell_{\pi^{\mathsf{BR}}(f(x))}\rangle - \mathop{\mathbb{E}}_{(x,y)\in\mathcal{D}}[\langle f(x), \ell_{\pi^{\mathsf{BR}}(f(x))}\rangle] \leq d\mathcal{T}\beta + d/\mathcal{T}$$

$\square$

We are now ready to give the formal guarantees of Line 1.

Theorem F.V.5. *Fix any distribution $\mathcal{D}$ and dataset $D \sim \mathcal{D}$ containing n samples drawn i.i.d from $\mathcal{D}$. For any pair of predictors $f_1, f_2 : [k] \to [0,1]^d$, family of loss functions $\mathcal{L}$, loss margin $\alpha > 0$, disagreement region mass $\eta > 0$, and decision-calibration tolerance $\beta > 0$, Line 1 run over the empirical distribution D updates f_1 and f_2 for T_1 and T_2 time-steps, respectively, and outputs a pair of predictors (f_1^T, f_2^T) such that, with probability at least $1-\delta$ over the randomness of $D \sim \mathcal{D}^n$,*

1. *The total number of time-steps for Line 1 and Line 1 is*

$$T = T_1 + T_2 \leq \frac{2d}{\min\{\beta^2, \eta\alpha^2/2d^3\}}$$

2. *For $i \in \{1,2\}$, the Brier scores of the final models are lower than that of the input models:*

$$B(f_i^{T_i}, \mathcal{D}) \leq B(f_i, \mathcal{D}) - (T_i - \eta) \cdot \min\left\{\beta^2, \eta\alpha^2/(2d^3)\right\}$$

3. *For $i \in \{1,2\}$ and for all $\ell \in \mathcal{L}$, the downstream decision-making losses of the final models do not increase by much compared to that of the input models:*

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell(y, \pi_\ell^{\mathsf{BR}}(f_i^T(x)) - \ell(y, \pi_\ell^{\mathsf{BR}}(f_i(x))] \leq 2T_i\sqrt{\beta}d$$

4. *The final models approximately agree on their best-response actions almost everywhere. That is, the disagreement region E_ℓ^α calculated using f_1^T, f_2^T has small mass:*

$\forall \ell \in \mathcal{L}$,

$$\mu(E_\ell^\alpha) \le 2\eta$$

if $n \ge \Omega\left(d^2/(\eta^2\beta\min\{\beta^2,\eta\alpha^2/d^3\}) \cdot \left(\ln(d\mathcal{T}|\mathcal{L}|) - \ln(\delta\min\{\beta^2,\eta\alpha^2/d^3\})\right)\right)$.

Proof. **Number of time-steps**. By Lemma F.II.2, at each iteration t in Line 1 of its subroutine Line 1, the Brier score of f^t decreases by $\min\{\beta^2, \eta\alpha^2/2d^3\}/2$, when $m \ge \lceil\sqrt{d/2\min\{\beta^2,\eta\alpha^2/2d^3\}}\rceil$. Since Brier score is bounded in $[0, d]$, the total number of time-steps satisfies

$$T_{\max} = T_1 + T_2 \le \frac{4d}{\min\{\beta^2,\eta\alpha^2/2d^2\}}.$$

Number of Output Predictors. We count the number of pairs of output predictors. The patching event at each iteration is in the set $\{E_{\ell,i,\cdot} : \ell \in \mathcal{L}, i \in [d], \cdot \in \{<,>\}\} \cup \{E_{\ell,i,\cdot} \cap E_{\ell',i',\tau} : \ell,\ell' \in \mathcal{L}, i,i' \in [d], \tau \in \mathcal{T}, \cdot \in \{<,>\}\}$. The total number of event is at most $2|\mathcal{L}|d + 2|\mathcal{L}|^2d^2|\mathcal{T}|$. Since each pair of output predictors is defined by the sequence of patchings and $\Delta^t \in [1/m]^d$, we have the total number of possible output predictors is

$$|S| \le ((4d^2|\mathcal{L}|^2|\mathcal{T}|)(m+1)^d)^{T_{\max}+1}.$$

Therefore,

$$\ln(|S|) \le \Omega\left(\frac{d^2}{\min\{\beta^2,\eta\alpha^2/2d^2\}}\ln(d|\mathcal{L}|\mathcal{T}/\min\{\beta^2,\eta\alpha^2/2d^2\})\right).$$

Brier Score. The result directly follows using the same argument as in Theorem 9.3.4. We can obtain

$$B(f_i^{T_i},\mathcal{D}) \le B(f_i,\mathcal{D}) - T_i \cdot \min\left\{\beta^2,\eta\alpha^2/(2d^3)\right\} + \sqrt{(d\ln(6d|S|/\delta)/(2n)}.$$

Loss Estimation. Using Lemma F.III.5 and a union bound over $\{E_{\ell,i,\cdot} \cap E_{\ell',i',\tau} : \ell,\ell' \in \mathcal{L}, i,i' \in [d], \tau \in [1/\mathcal{T}], \cdot \in \{<,>\}\}$ and $|S|$, we have with probability at least

$1 - \delta/3$,

$$\left\| \mathop{\mathbb{E}}_{\mathcal{D}}[(y - f(x))E_{\ell,a}(f(x), x)] - \frac{1}{n}\sum_{i=1}^{n}[(y_i - f(x_i))E_{\ell,a}(f(x_i), x_i)] \right\|_2$$
$$\leq \sqrt{\frac{3d\ln(6d\mathcal{T}|S||\mathcal{L}|/\delta)}{2n}}.$$

Using Lemma F.V.4 and summing over all time-steps, we can obtain the result following the same steps as the proof in Theorem 9.3.4:

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell(y, \pi^{\mathsf{BR}}(f_i^T(x)) - \ell(y, \pi^{\mathsf{BR}}(f_i(x))] \leq \left(\sqrt{\beta}d + \sqrt{(3d\ln(6d\mathcal{T}|S||\mathcal{L}|/\delta)/(2n)}\right) T_i$$

Disagreement Event. Using union bound over the disagreement events $\{E_\ell : \ell \in \mathcal{L}, \}$ and the set of possible output predictors, we have, for all $\ell \in \mathcal{L}$, and $(f_1, f_2) \in S$,

$$\left| \mu(E_{\ell,a_1,a_2}(f_1(x), f_2(x), x) - \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}[E_{\ell,a_1,a_2}(f_1(x_i), f_2(x_i), x_i))] \right| \leq \sqrt{\frac{\ln(6|S||\mathcal{L}|/\delta)}{2n}}.$$

From the while loop condition in Line 1, we know

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}[E_\ell(f_1^{T_1}(x_i), f_2^{T_2}(x_i), x_i)] \leq \eta.$$

Therefore,

$$\mu(E_\ell) \leq \eta + \sqrt{(\ln(6|S||\mathcal{L}|/\delta)/(2n)} \quad \text{for all } \ell \in \mathcal{L}$$

The theorem follows from solving n in the conditions 2 through 4. □

F.VI Additional Experiment Detail

Our ImagineNet experiments are run on a Macbook Pro with 32GB of RAM. The experiment on the ImageNet dataset uses pre-trained models provided by pyTorch, which do not require additional training. The experiment on the HAM10000 dataset includes neural network models trained using pyTorch on NVIDIA GA100 GPU (80 GB of RAM) with 2 compute workers loading the data. The total time to train the two neural network models with approximately 88% Top-1 accuracy takes less than 5 minutes.

F.VI.1 Additional experiment on ImageNet dataset

Brier score. In Figure F.1, we compare the Brier score of ReDCal (Line 1) with the two baseline algorithms on both the calibration and the test datasets. Compared to Reconcile, our algorithm decreases the Brier score by a smaller amount on the test dataset. The combined algorithm of Decision-Calibration with ReDCal as post-process achieves the most substantial decrease in the Brier score.

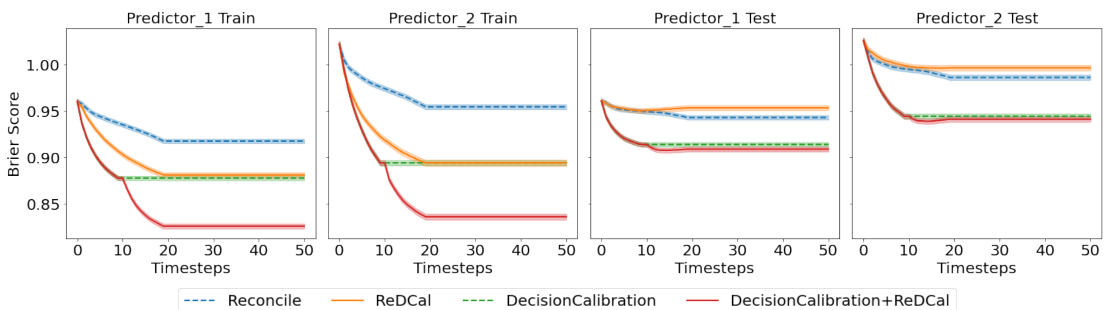


Figure F.1: ReDCal decreases Brier score on Imagenet. Compared to Reconcile, our algorithm decreases the Brier score by a smaller amount on the test dataset. Decision-Calibration with ReDCal as post-process achieves the most substantial decrease in the Brier score.

Decision loss on calibration dataset. In Figure F.2, we compare the decision gap of our proposed algorithm with the two baseline algorithms on the training dataset.

Decision loss comparison for high-dimensional classification problem. In Figure F.3, we compare the decision loss gap of our proposed algorithm with the two baseline algorithms on the testing dataset, using $d = 10, 100$, and 1000 classes. We plot the average loss gap of the two predictors. The hyperparameters are: disagreement margin $\alpha = 0.1/d$, decision-calibration tolerance $\beta = 0.001/d$, disagreement region mass $\eta = 0.01$, number of actions $K = 10$.

F.VI.2 HAM1000 Classification

Experiment Setup. We use the HAM10000 dataset (Tschandl et al., 2018) (licensed CC BY-NC 4.0) on pigmented skin lesions to predict the probability that a patient

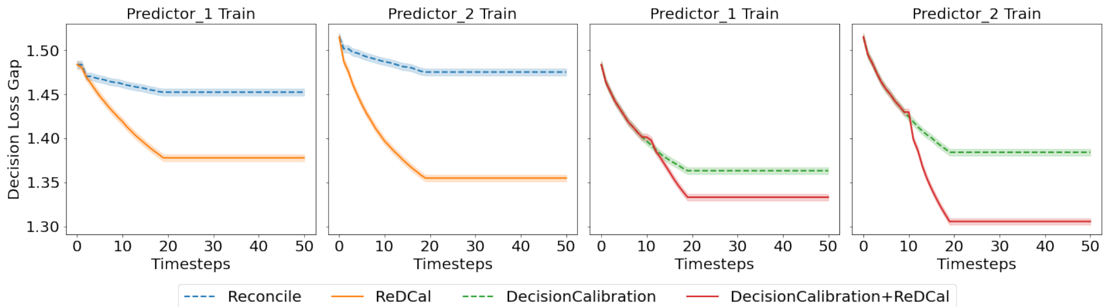


Figure F.2: ReDCal decreases decision loss on Imagenet. We plot the gap between the average loss of each predictor and the optimal loss had we known the true predictor f^* on the calibration set. **Left two figures:** We compare the LossGap of ReDCal (orange) with Reconcile (blue). Compared to Reconcile, our algorithm converges at a similar rate and decreases the loss by a larger amount on the test dataset. **Right two figures:** We compare the LossGap of Decision-Calibration (green) to that of Decision-Calibration with a run of ReDCal as post-process (red). Our proposed algorithm can improve upon Decision-Calibration output predictors and further reduce the loss on the test dataset. Results are averaged over 500 runs and the shaded region indicates ± 1 standard errors.

has contracted one of 7 possible skin diseases: 'akiec', 'bcc', 'bkl', 'df', 'nv', 'vasc', and 'mel'. We split the dataset into train/validation/test sets, with 20% of the data are used for validation and 20% are used for testing. We use the train set to train two neural networks using pyTorch with resnet50 (He et al., 2015) and densenet121 (Huang et al., 2018) architectures and learn two models with around 88% top-1 accuracy. From each model, we output the individual probability prediction for each of the 7 possible labels. We use the validation set to calibrate the predictors using our proposed algorithm and the two baseline algorithms, and the test set to measure the final performance.

We run each calibration algorithm 10 times. At each run, we draw a fresh loss function created based on the loss function motivated by medical domain knowledge in Zhao et al. (2021) and additional random noise drawn from $\mathsf{Normal}(0, 1)$. There are two possible actions for the decision-maker: treatment ($a = [1, 0]$) or no treatment ($a = [0, 1]$). Given a loss function ℓ and a predictor f, the decision-maker will choose an action that minimizes their loss.

For each calibration algorithm, we calculate (1) the Brier score of the updated predictors and (2) the differences between the optimal loss had we known y and the actual loss from taking the best-response actions induced by each predictor.

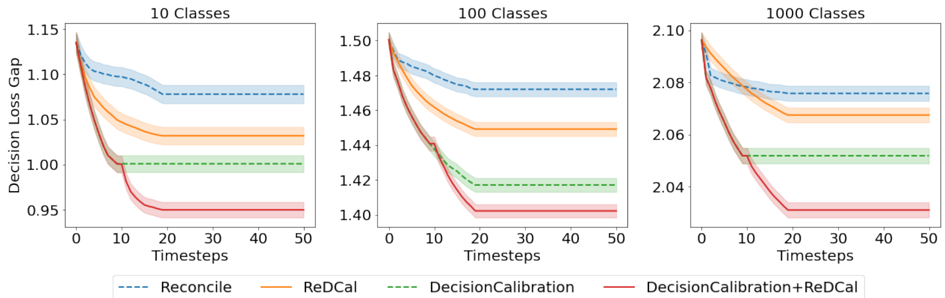


Figure F.3: ReDCal decreases decision loss on Imagenet. The takeaway results are similar to Figure 9.2. As the number of classes in the multi-class classification problem grows from 10 to 1000, ReDCal still outperforms Reconcile in decreasing decision loss on the test dataset. When we have 1000 classes, ReDCal converges slower than Reconcile. Furthermore, ReDCal can further decrease the decision loss when it is used as a post-process after Decision Calibration terminates.

The hyperparameters for Line 1 are chosen as follows: loss margin $\alpha = 0.1$, target disagreement region mass $\eta = 0.01$, and decision-calibration tolerance $\beta = 0.000001$.

Results. The takeaway results are similar to those of the ImageNet experiment.

Brier score. In Figure F.4, we compare the Brier score of ReDCal (Line 1) with the two baseline algorithms on both the calibration and the test datasets. Compared to Reconcile, our proposed algorithm decreases the Brier score by a smaller amount.

Decision loss. In Figure F.5, we compare the decision loss gap of our proposed algorithm with the two baseline calibration algorithms. Compared to Reconcile, our algorithm decreases the decision loss by a larger amount on the test dataset. Furthermore, while Decision-Calibration already decreases the decision loss, our algorithm can further improve upon their result when it is used as a post-process after Decision-Calibration terminates.

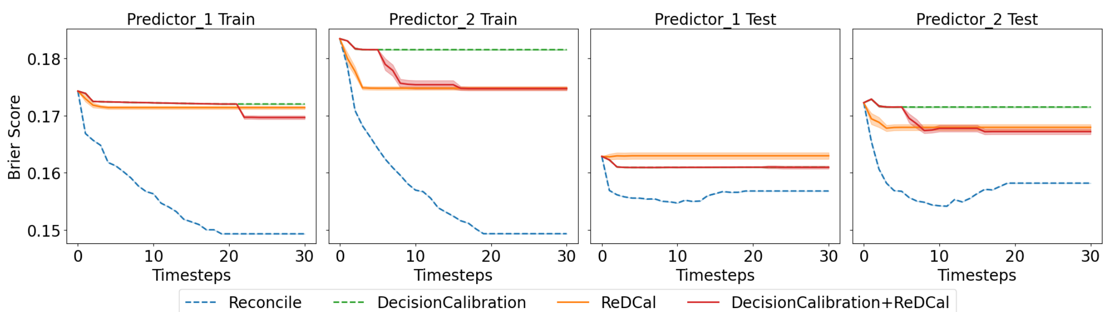


Figure F.4: Brier score of the updated predictors using Line 1 (orange) and two benchmark algorithms: Line 1 (dashed-blue) and Line 1 (dashed-green). Our algorithm reduces the Brier score by a smaller amount compared to Line 1. Results are averaged over 10 runs and the shaded region indicates ± 1 standard error.

Algorithm 14: Reconcile (Roth et al., 2023)

Input: $f_1, f_2, \eta > 0, \alpha > 0$

Let $f_1^0 = f_1, f_2^0 = f_2$.

while $\mu(U_\alpha(f_1^{t_1}, f_2^{t_2})) \geq \eta$ **do**

For each $\bullet \in \{>, <\}$ and $i \in \{1, 2\}$, let:

$$v_*^\bullet = \mathbb{E}[y | x \in U_\varepsilon^\bullet(f_1^{t_1}, f_2)] \quad v_i^\bullet = \mathbb{E}[f_i^{t_i}(x) | x \in U_\varepsilon^\bullet(f_1^{t_1}, f_2)]$$

Let

$$(i_t, \bullet_t) = \underset{i \in \{1,2\}, \bullet \in \{>,<\}}{\operatorname{argmax}} \mu(U_\varepsilon^\bullet(f_1^{t_1}, f_2^{t_2})) \cdot (v_*^\bullet - v_i^\bullet)^2$$

breaking ties arbitrarily.

Let:

$$g_t(x) = \begin{cases} 1 & x \in U_\varepsilon^{\bullet_t}(f_1^{t_1}, f_2^{t_2}) \\ 0 & \text{otherwise} \end{cases}$$

Let

$$\widetilde{\Delta}_t = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[y | g_t(x) = 1] - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[f_{i_t}^{t_{i_t}}(x) | g_t(x) = 1]$$

$$\Delta_t = \text{Round}(\widetilde{\Delta}_t; m)$$

Let $f_i^{t_i+1}(x) = h(x, f_i^{t_i}, g_t, \Delta_t), t_i = t_i + 1, t = t + 1$.

end while

Output: $(f_1^{t_1}, f_2^{t_2})$

Algorithm 15: Reconcile Decision Calibration for Multiple Predictors (ReDCal-Multi)

Input: $f_1, f_2, \ldots, f_k, \mathcal{L}, \eta > 0, \alpha > 0, \beta > 0$

Let $f_i^0 = f_i$ for all $i \in [k]\backslash\{1\}$, and $t = 0$. We use f_1 as the base predictor.

while $\mu(E_{\ell,a_1,a_2,j}) \geq \eta$ for some $a_1, a_2 \in \mathcal{A}$, $\ell \in \mathcal{L}$, and $j \in [k]\backslash\{1\}$ **do**

  Let $\ell^t, a_1^t, a_2^t, j^t = \text{argmax}_{\ell,a,a'}\, \mu(E_{\ell,a,a'})$, $E^t = E_{\ell^t,a_1^t,a_2^t,j^t}$.

  Pick

$$i^t = \underset{i\in\{1,j^t\}}{\text{argmax}} \Big| \underset{(x,y)\sim\mathcal{D}}{\mathbb{E}}[\ell^t(y, a_1^t) - \ell^t(y, a_2^t)|E^t(x) = 1] - \underset{x\sim[k]}{\mathbb{E}}[\ell^t(f_i(x), a_1^t) - \ell^t(f_i(x), a_2^t)|E^t(x) = 1]\Big|.$$

  Denote $f_{i^t}^t$ as f_i^t.
  Let $\phi^t = \mathbb{E}_{(x,y)\sim\mathcal{D}}[y|E^t(x) = 1] - \mathbb{E}_{x\sim\mathcal{D}_{[k]}}[f_i^t(x)|E^t(x) = 1]$.

  Patch $f^t(x) = \text{proj}_{[0,1]^d}(f_i^t(x) + \phi^t E^t(x))$.

  Let $f_i^{t+1} = \text{Decision-Calibration}(f^t,\, \mathcal{L},\, \beta,\, E^t)$. $t = t + 1$.

end while

Output: $f_1^t, \ldots, f_k^t$

Algorithm 16: Decision Calibration for Infinite Action Set

Input: Predictor f, loss family $\mathcal{L}$, $\beta > 0$, event E

Let $f^0 = f$.

while f^t is not index-wise β-MC w.r.t $E_{\ell,i,\tau} \cap E$ for loss $\ell \in \mathcal{L}$, index $i \in [d]$, $\tau \in [1/\mathcal{T}]$ **do**

  Let $\ell^t, i^t, \tau^t = \text{argmax}_{\ell,i,\tau} \parallel \mathbb{E}_{(x,y)\sim\mathcal{D}}[(y - f^t(x))E_{\ell,i,\tau}(f^t(x), x)]\|_2$

  Let $\phi^t = \mathbb{E}_{(x,y)\sim\mathcal{D}}[y - f^t(x)|E_{\ell^t,i^t,\tau^t}(f^t(x), x) = 1]$

  Patch $f^{t+1}(x) = \text{proj}_{[0,1]^d}(f^t(x) + \phi^t E_{\ell^t,i^t,\tau^t}(f^t(x), x))$

  $t = t + 1$.

end while

Output: f^t

Algorithm 17: Reconcile Decision Calibration for Infinite Action Set (ReDCal-Inf)

Input: $f_1, f_2, \mathcal{L}, \eta > 0, \alpha > 0, \beta > 0$

1: Let $f_1^0 = f_1, f_2^0 = f_2$ and $t = 0$.

2: **while** $\mu(E_\ell) \geq \eta$ for some $\ell \in \mathcal{L}$ **do**

3: Let $\ell^t, j^t, \cdot^t = \text{argmax}_{\ell, j, \cdot}\, \mu(E_{\ell,j,\cdot})$, $E^t = E_{\ell^t, j^t, \cdot^t}$.

4: Pick

1
$$i^t = \underset{i \in \{1,2\}}{\text{argmax}} \Bigg| \underset{(x,y)\sim\mathcal{D}}{\mathbb{E}}[\ell^t(y, \pi^{\mathsf{BR}}(f_1^t(x))) - \ell^t(y, \pi^{\mathsf{BR}}(f_2^t(x)))|E^t(x) = 1] - \underset{x\sim[k]}{\mathbb{E}}[\ell^t(f_i(x), \pi^{\mathsf{BR}}(f_1^t(x))) - \ell^t(f_i(x), \pi^{\mathsf{BR}}(f_2^t(x)))|E^t(x) = 1]\Bigg|.$$

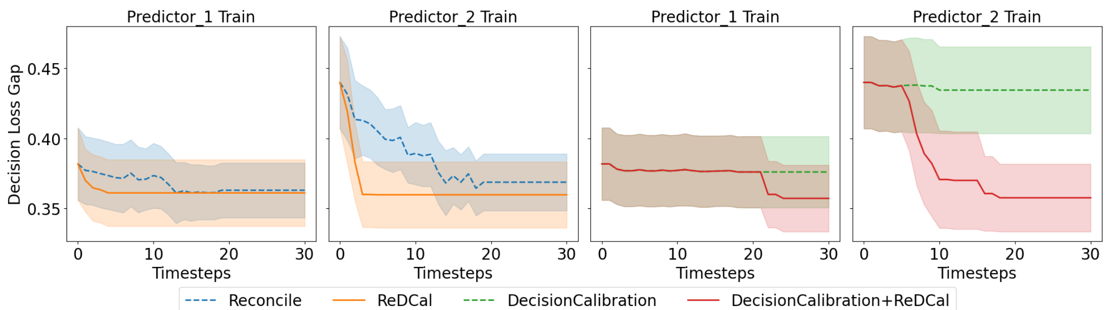
5: Denote $f_{i^t}^t$ as f_i^t.
Let $\phi^t = \mathbb{E}_{(x,y)\sim\mathcal{D}}[y|E^t(x) = 1] - \mathbb{E}_{x\sim\mathcal{D}_{[k]}}[f_i^t(x)|E^t(x) = 1]$.

6: Patch $f^t(x) = \text{proj}_{[0,1]^d}(f_i^t(x) + \phi^t E^t(x))$.

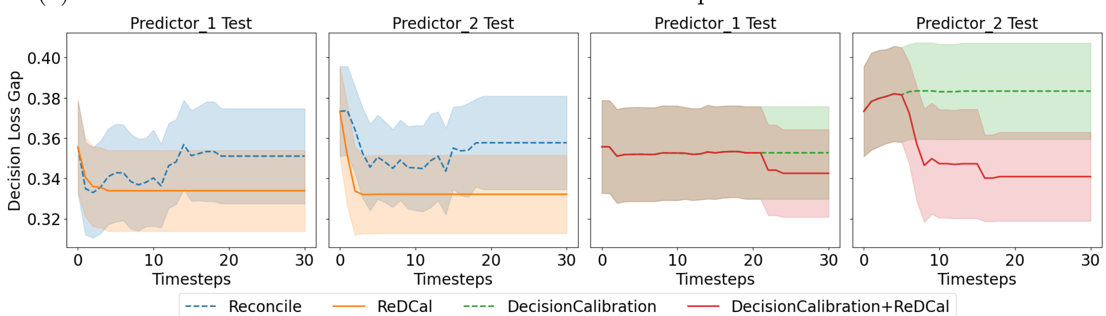
7: Let $f_i^{t+1} = \text{Decision-Calibration-Inf}(f^t,\ \mathcal{L},\ \beta,\ E^t)$. $t = t + 1$.

8: **end while**

Output: f_1^t, f_2^t



(a) ReDCal decreases the decision loss on the validation partition of HAM10000 dataset.



(b) ReDCal decreases the decision loss on the test partition of HAM10000 dataset.

Figure F.5: In Figure F.5a and Figure F.5b, we plot the gap between optimal loss had we know the true label y and the loss from taking best-response actions induced by the calibrated predictors on the validation set and test set, respectively. In the left two figures, we compare Line 1 (orange) with Line 1 (blue). While the average loss of predictors updated using Line 1 may increase on the test set, our algorithm quickly converges and produces predictors with lower decison-making loss. In the right two figures, we compare Line 1 (green) to Line 1 with an additional run of Line 1 (red) as post-process. We observe that running our algorithm as post-process can still further decrease the loss compared to just running Line 1 on its own. Results are averaged over 10 runs and the shaded region indicates ± 1 standard errors.