LEARNING, DECISION-MAKING, AND INFERENCE WITH LIMITED EXPERIMENTATION

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF ELECTRICAL ENGINEERING
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Khashayar Khosravi

August 2019

Abstract

Over the last decade, machine learning techniques have revolutionized a variety of disciplines including medicine, natural language processing, computer vision, and finance. With this central role that machine learning has in our lives, there exists a vital need to fully investigate the applicability of these techniques, especially in human-centered domains. Indeed, many assumptions behind the theoretical analysis of machine learning algorithms are not satisfied in practice, which can lead to harmful consequences in sensitive applications. One example is in medical decision-making, where the impact of wrong decisions may be severe and irreversible. Furthermore, despite these huge advances in machine learning, the traditional randomized controlled trials are still the gold standard for testing the efficacy of various treatment and policy interventions. Although these trials are very robust and valid with minimal (or no) assumptions, they are very expensive and difficult to run, especially in medical applications. Thus, it is also important to carefully design and customize machine learning techniques that under a mild set of assumptions can provide meaningful and valid guarantees about the treatment effects using observational data. Observational data are usually more accessible and have the advantage of bypassing the burden of running experiments. It is worth noting that, however, the results from observational data should be interpreted with caution and can potentially be used as a preliminary step for conducting targeted trials. In particular, the decision-maker can decide to remove individuals who may adversely be affected by the treatment from the trial or only consider those who benefit from it.

This doctoral dissertation focuses on designing and analyzing data-driven methods with limited experimentation, focusing on three different settings. First, we study the online setting, where the decision-maker needs to personalize treatment decisions sequentially and wishes to reduce the amount of experimentation (randomization). This part contributes to the field of contextual bandit. In the rest of this dissertation, we consider the offline setting where the decision-maker has access to some observational data and wishes to estimate and draw inference about treatment effects. Specifically, we consider panel data models and discuss the treatment effect estimation using matrix completion methods. Moreover, we analyze personalized (non-parametric) inference from observational data with high dimensional covariates and combine non-parametric estimators with sub-sampling techniques to provide valid confidence intervals that are able to adapt to a priori unknown lower-dimensional structure of data. These two parts use techniques from probability theory and statistics, and contribute to the field of learning theory and causal inference.

To my loving family - Arghavan, Alireza, Fariba, and Arash.

Acknowledgments

First, I would like to express my deepest appreciation to my advisor Mohsen Bayati for his tremendous guidance and support throughout my PhD. Not only he taught me the principles of scientific research, but also how to concisely and effectively convey such scientific ideas to an audience of researchers. Regardless of his busy schedule, he always made sure to allocate time so that we could discuss research ideas. I found these discussions extremely valuable, especially when the research was going down a difficult path. In addition to him being a great scientific advisor, Mohsen is a great listener and deeply cares about the well-being of his students. I strive to use the valuable lessons that I learned from Mohsen during my PhD, throughout my life.

I am thankful to Andrea Montanari who helped me greatly during my PhD. I learned a lot from the broad set of probability and statistics courses that Andrea teaches at Stanford. I thank Ramesh Johari for giving me valuable research advice and providing insights on how to accurately formulate and tackle research questions. I am grateful to Tsachy Weissman who, despite his busy schedule, kindly agreed to be in my oral exam committee and provided me with insightful suggestions, and to Kostas Bimpikis for chairing my oral exam. I thank Hamsa Bastani, who I was fortunate to collaborate with early in my PhD. I had many enjoyable conversations with Hamsa about research and career.

I am also thankful to my amazing collaborators at Stanford who made my PhD more fruitful. I learned plenty from them and enjoyed these collaborations: Susan Athey, Nikolay Doudchenko, John Duchi, Nima Hamidi, Guido Imbens, Hakan Inan, Feng Ruan, and Richard Socher. I also spent three amazing months at Microsoft Research New England as an intern. I am thankful to Greg Lewis and Vasilis Syrgkanis for mentoring me throughout this internship.

I thank Stanford Electrical Engineering department for supporting me with a departmental fellowship for one year. I am also grateful to the National Science Foundation, Stanford Cyber Security Initiative, and Stanford Human-Centered Artificial Intelligence for supporting the rest of my PhD.

I would like to thank my amazing friends who made my PhD journey more joyful and fun: Omid Aryan, Javad Azimi, Kamyar Azizzadenesheli, Payam Banazadeh, Amirali Boroumand, Behrad Habib Afshar, Kian Kevin Katanforoosh, Ramtin Keramati, Negar Kouchaki, Mahdis Mahdieh, Shahab Mirjalili, Babak Moatamed, Rad Niazadeh, Koosha Nassiri Nazif, Negar Rahmati, Amirabbas Sadeghian, Alireza Sharafat, Milad Sharif, Alireza Seif, Yasaman Shirian, Reza Takapoui, Saman Tehrani, and many others. They brought me so many amazing memories that I will never forget.

I am very grateful to have enormous encouragement and support from my family, particularly, my parents Alireza and Fariba. Despite living a thousand miles apart, they have never stopped giving me their unconditional love and support. I would also like to thank my caring brother Arash, my compassionate mother-in-law Felora, and my kind sisters-in-law, Delaram and Taraneh. I have been blessed to have their spiritual love and support in my life.

Above all, I would like to thank my beloved wife Arghavan for her endless love and compassion, for creating the most beautiful moments of my life, for rejuvenating my life with new motivations every day, for helping me to get through the tough times, for teaching me how to cook, and most importantly, for being my best friend. I owe you everything.

Contents

List of Figures

Chapter 1

Introduction

Data-driven decision-making has recently received a substantial interest from service providers across a variety of domains, and has greatly improved operational efficiency for numerous firms. A powerful and (increasingly) popular means for data-driven decision-making is running controlled experiments. In fact, companies such as Amazon, Facebook, Google, and Microsoft conduct tens of thousands of experiments every year to improve their operations. For instance, a randomized controlled experiment on Bing ad headlines in 2012 led to a twelve percent increase in revenue (Kohavi and Thomke 2017).

However, a common issue in controlled experiments is that assignment of decisions (or treatments) via randomization generates business opportunity costs and can even damage customers' trust (e.g., Facebook's experiment on emotions, Luca 2014). These challenges are inherent in domains such as healthcare, where experimentation can be unethical and is only allowed in very expensive and highly controlled settings (Sibbald and Roland 1998).

In this dissertation, we tackle these challenges and design new algorithms, provide performance guarantees, and also confirm these theoretical guarantees via simulations. We focus specifically on two types of solutions in this dissertation: (1) reducing unnecessary randomization in experiments by leveraging natural variations in the data or incentive schemes (e.g., combining greedy algorithms and hypodissertation testing to develop an efficient decision-making scheme that utilizes heterogeneity of the users to successfully reduce the experimentation), and (2) bypassing experimentation in situations where experimentation is not feasible and observational data is available (e.g., designing a new class of statistical methods for treatment effect estimation, motivated by matrix completion literature, and also building techniques that allow inference in the non-parametric setting with high-dimensional covariates).

In this dissertation we make three main contributions. In Chapter 2, we consider the online setting where the decision-maker has access to no past data about the treatments and wishes to sequentially personalize the treatment decisions so as to maximize some reward function of interest. Current literature on this topic focuses on algorithms that balance an exploration-exploitation tradeoff, to ensure a sufficient rate of learning while optimizing for reward. In particular, greedy algorithms that exploit current estimates without any exploration may be sub-optimal in general. As mentioned above, however, exploration-free greedy algorithms are desirable in practical settings where exploration may be costly or unethical (e.g., clinical trials). Surprisingly, we find that a simple greedy algorithm can be rate-optimal (achieves asymptotically optimal regret) if there is sufficient randomness in the observed contexts (covariates). We prove that this is always the case for a two-armed bandit under a general class of context distributions that satisfy a condition we term *covariate diversity*. Furthermore, even absent this condition, we show that a greedy algorithm can be rate optimal with positive probability. Thus, standard bandit algorithms may unnecessarily explore. Motivated by these results, we introduce Greedy-First, a new algorithm that uses only observed contexts and rewards to determine whether to follow a greedy algorithm or to explore. We prove that this algorithm is rate-optimal without any additional assumptions on the context distribution or the number of arms. Extensive simulations demonstrate that Greedy-First successfully reduces exploration and outperforms existing (exploration-based) contextual bandit algorithms such as Thompson sampling or upper confidence bound (UCB). This chapter is based on Bastani et al. (2017).

In certain practical settings the decision-maker has access to some past observational data on the treatment and outcomes. Ideally, in these settings, the decision-maker wishes to bypass experimentation by estimating treatment effects from the existing data. One of the fundamental challenges in analyzing observational data is the presence of confounders that can affect both the treatment and outcome, leading to biased estimates for the treatment effect (also known as the "omitted variable bias" or OVB). In Chapter 3, we investigate settings where confounding factors might be unobserved, but we have access to treatment decisions over time. In particular, we consider the panel data model (matrix), where a subset of units (rows) are exposed to a single treatment during a subset of time periods (columns), and the goal is estimating counterfactual (untreated) outcomes for the treated unit/period combinations. Motivated by the literature on matrix completion (Candès and

Recht 2009, Keshavan et al. 2010a,b), we develop a class of matrix completion estimators that uses the observed elements of the matrix of control outcomes corresponding to untreated unit/periods to predict the "missing" elements of the matrix, corresponding to treated units/periods. This approach estimates a matrix that well-approximates the original (incomplete) matrix, but has lower complexity according to the nuclear norm for matrices. From a technical perspective, we generalize results from the matrix completion literature by allowing the patterns of missing data to have a time series dependency structure. We also present new insights concerning the connections between the interactive fixed effects models and the literatures on program evaluation under unconfoundedness as well as on synthetic control methods. If there are few time periods and many units, our method approximates a regression approach where counterfactual outcomes are estimated through a regression of current outcomes on lagged outcomes for the same unit. In contrast, if there are few units and many periods, our proposed method approximates a synthetic control estimator where counterfactual outcomes are estimated through a regression of the lagged outcomes for the treated unit on lagged outcomes for the control units. The advantage of our proposed method is that it moves seamlessly between these two different approaches, utilizing both cross-sectional and within-unit patterns in the data. Simulations illustrate that our proposed matrix completion method outperforms the existing benchmarks and successfully reduces OVB in presence of confounding factors. This chapter is based on Athey et al. (2018).

Finally, in many practical applications in social sciences, the decision-maker wishes to develop methods that not only enable the treatment effect estimation using observational data, but also allow for inference from such data. In these applications, it is plausible that the effect of treatment varies among different individuals depending on potentially high-dimensional attributes (covariates). A central task in such applications is to design adequate estimators for the personalized treatment effect that enjoy strong asymptotic properties that enable valid constructions of confidence intervals. The decision-maker can use these personalized confidence intervals to identify and target a subset of units (population) that benefit the most from the treatment, leading to improved decisions for all individuals. With this motivation, in Chapter 4, we consider non-parametric estimation and inference of conditional moment models in high dimensions. Unfortunately, without any further structural assumptions on the problem, the exponential in dimension rates of approximately $n^{1/D}$ (see, e.g., Stone 1982) cannot be avoided (also known as the "Curse of

Dimensionality"), where D is the dimension of conditioning variable. We show that even when D is larger than the sample size n, estimation and inference is feasible as long as the distribution of the conditioning variable has small intrinsic dimension d, as measured by locally low doubling measures. Our work follows a long line of work in machine learning (Dasgupta and Freund 2008, Kpotufe 2011, Kpotufe and Garg 2013), which is founded on the observation that in many practical applications, the coordinates of X are highly correlated (e.g., an image), despite X being high-dimensional. The latter intuition is formally captured by assuming that the distribution of X has a small doubling measure around the target point x. These works, however, solely establish estimation guarantees and do not characterize the asymptotic distribution of the estimates, so as to enable inference, hypothesis testing and confidence interval construction. Moreover, they only address the regression setting and not the general conditional moment problem, and consequently do not extend to quantile regression, instrumental variable regression, or treatment effect estimation. We generalize these results by providing estimation and asymptotic normality results for the general conditional moment problem, where the estimation and the asymptotic normality rates depend only on the intrinsic dimension, and are independent of the explicit dimension of the conditioning variable. In particular, we show that if the intrinsic dimension of the covariate distribution is equal to d, then the finite sample estimation error of our estimator is of order $n^{-1/(d+2)}$ and our estimate is $n^{1/(d+2)}$-asymptotically normal, irrespective of D. Our estimation is based on a sub-sampled ensemble of the k-nearest neighbors (k-NN) Z-estimator. The sub-sampling size required for achieving these results depends on the unknown intrinsic dimension d. We propose an adaptive data-driven approach for choosing this parameter and prove that it achieves the desired rates. Simulations confirm our theoretical findings and demonstrate that our adaptive sub-sampled k-NN estimator performs well in practice. This chapter is based on Khosravi et al. (2019).

Chapter 2

Reducing Experimentation in Contextual Bandits

2.1 Motivation

Service providers across a variety of domains are increasingly interested in personalizing decisions based on customer characteristics. For instance, a website may wish to tailor content based on an Internet user's web history (Li et al. 2010), or a medical decision-maker may wish to choose treatments for patients based on their medical records (Kim et al. 2011). In these examples, the costs and benefits of each decision depend on the individual customer or patient, as well as their specific context (web history or medical records respectively). Thus, in order to make optimal decisions, the decision-maker must learn a model predicting individual-specific rewards for each decision based on the individual's observed contextual information. This problem is often formulated as a contextual bandit (Auer 2003, Langford and Zhang 2008, Li et al. 2010), which generalizes the classical multi-armed bandit problem (Thompson 1933, Lai and Robbins 1985).

In this setting, the decision-maker has access to K possible decisions (arms) with uncertain rewards. Each arm i is associated with an unknown parameter $\beta_i \in \mathbb{R}^d$ that is predictive of its individual-specific rewards. At each time t, the decision-maker observes an individual with an associated context vector $X_t \in \mathbb{R}^d$. Upon choosing arm i, she realizes a (linear) reward of

$$X_t^\top \beta_i + \varepsilon_{i,t}\,, \tag{2.1.1}$$

where $\varepsilon_{i,t}$ are idiosyncratic shocks. One can also consider nonlinear rewards given by generalized linear models (e.g., logistic, probit, and Poisson regression); in this case, (2.1.1) is replaced with

$$\mu(X_t^\top \beta_i) + \varepsilon_{i,t}\,, \tag{2.1.2}$$

where μ is a suitable *inverse link function* (Filippi et al. 2010, Li et al. 2017). The decision-maker's goal is to maximize the cumulative reward over T different individuals by gradually learning the arm parameters. Devising an optimal policy for this setting is often computationally intractable, and thus, the literature has focused on effective heuristics that are asymptotically optimal, including UCB (Dani et al. 2008, Abbasi-Yadkori et al. 2011), Thompson sampling (Agrawal and Goyal 2013, Russo and Van Roy 2014b), information-directed sampling (Russo and Van Roy 2014a), and algorithms inspired by ϵ-greedy methods (Goldenshluger and Zeevi 2013, Bastani and Bayati 2015).

The key ingredient in designing these algorithms is addressing the *exploration-exploitation tradeoff*. On one hand, the decision-maker must explore or sample each decision for random individuals to improve her estimate of the unknown arm parameters $\{\beta_i\}_{i=1}^K$; this information can be used to improve decisions for future individuals. Yet, on the other hand, the decision-maker also wishes to exploit her current estimates $\{\hat{\beta}_i\}_{i=1}^K$ to make the estimated best decision for the current individual in order to maximize cumulative reward. The decision-maker must therefore carefully balance both exploration and exploitation to achieve good performance. In general, algorithms that fail to explore sufficiently may fail to learn the true arm parameters, yielding poor performance.

However, exploration may be prohibitively costly or infeasible in a variety of practical environments (Bird et al. 2016). In medical decision-making, choosing a treatment that is not the estimated-best choice for a specific patient may be unethical; in marketing applications, testing out an inappropriate ad on a potential customer may result in the costly, permanent loss of the customer. Such concerns may deter decision-makers from deploying bandit algorithms in practice.

In this chapter, we analyze the performance of *exploration-free* greedy algorithms. Surprisingly, we find that a simple greedy algorithm can achieve the same state-of-the-art asymptotic performance guarantees as standard bandit algorithms *if* there is sufficient randomness in the observed contexts (thereby creating natural exploration). In particular, we prove that the greedy algorithm is near-optimal for a two-armed bandit when the context distribution satisfies a condition we term *covariate diversity*; this property requires that the covariance matrix of the observed contexts conditioned on any half space is positive definite. We show that covariate diversity is satisfied by a natural class of continuous and discrete context distributions. Furthermore, even absent covariate diversity, we show that a greedy approach provably converges to the optimal policy with some probability that depends on the problem parameters. Our results hold for arm rewards given by both linear and generalized linear models. Thus, exploration may not be necessary at all in a general class of problem instances, and is only sometimes be necessary in other problem instances.

Unfortunately, one may not know a priori when a greedy algorithm will converge, since its convergence depends on unknown problem parameters. For instance, the decision-maker may not know if the context distribution satisfies covariate diversity; if covariate diversity is not satisfied, the greedy algorithm may be undesirable since it may achieve linear regret some fraction of the time (i.e., it fails to converge to the optimal policy with positive probability). To address this concern, we present Greedy-First, a new algorithm that seeks to reduce exploration when possible by starting with a greedy approach, and incorporating exploration only when it is confident that the greedy algorithm is failing with high probability. In particular, we formulate a simple hypothesis test using observed contexts and rewards to verify (with high probability) if the greedy arm parameter estimates are converging at the asymptotically optimal rate. If not, our algorithm transitions to a standard exploration-based contextual bandit algorithm.

Greedy-First satisfies the same asymptotic guarantees as standard contextual bandit algorithms without our additional assumptions on covariate diversity or any restriction on the number of arms. More importantly, Greedy-First does not perform any exploration (i.e., remains greedy) with high probability if the covariate diversity condition is met. Furthermore, even when covariate diversity is not met, Greedy-First provably reduces the expected amount of forced exploration compared to standard bandit algorithms. This occurs because the vanilla greedy algorithm provably converges to the optimal policy with some probability even for problem instances without covariate diversity; however, it achieves linear regret on average since it may fail a positive fraction of the time. Greedy-First leverages this observation by following a purely greedy algorithm until it detects that this approach has failed. Thus, in any bandit problem, the Greedy-First policy explores less on average than standard algorithms that always explore. Simulations confirm our theoretical results, and demonstrate that Greedy-First outperforms existing contextual bandit algorithms even when covariate diversity is not met.

Finally, Greedy-First provides decision-makers with a natural interpretation for exploration. The hypothesis test for adopting exploration only triggers when an arm has not received sufficiently diverse samples; at this point, the decision-maker can choose to explore that arm by assigning it random individuals, or to discard it based on current estimates and continue with a greedy approach. In this way, Greedy-First reduces the opaque nature of experimentation, which we believe can be valuable for aiding the adoption of bandit algorithms in practice.

2.1.1 Related Literature

There has been significant interest in operational methods for personalizing service decisions as a function of observed user covariates (see, e.g., Ban and Rudin 2014, Bertsimas and Kallus 2014a, Chen et al. 2015, Kallus 2016). We take a sequential decision-making approach with *bandit feedback*, i.e., the decision-maker only observes feedback for her chosen decision and does not observe counterfactual feedback from other decisions she could have made. This obstacle inspires the exploration-exploitation tradeoff in multi-armed bandit problems.

Our work falls within the framework of contextual bandits (or a linear bandit with changing action space), which has been extensively studied in the computer science, operations, and statistics literature (we refer the reader to Chapter 4 of Bubeck and Cesa-Bianchi (2012) for an informative review). This setting was first introduced by Auer (2003) through the LinRel algorithm and was subsequently improved through the OFUL algorithm by Dani et al. (2008) and the LinUCB algorithm by Chu et al. (2011). More recently, Abbasi-Yadkori et al. (2011) proved an upper bound of $\mathcal{O}(d\sqrt{T})$ regret after T time periods when contexts are d-dimensional. (We note that they also prove a "problem-dependent" bound of $\mathcal{O}(d \log T/\Delta)$ if one assumes a constant gap Δ between arm rewards; this bound does not apply to the contextual bandit since there is no such gap between arm rewards.)

As mentioned earlier, this literature typically allows for arbitrary (adversarial) covariate sequences. We consider the case where contexts are generated i.i.d., which is more suited for certain applications (e.g., clinical trials on treatments for a non-infectious disease). In this setting one can achieve exponentially better regret bounds in T. In particular, Goldenshluger and Zeevi (2013) present the OLS Bandit algorithm and prove a corresponding upper bound of $\mathcal{O}(d^3 \log T)$ on its cumulative regret. They also prove a lower bound of $\mathcal{O}(\log T)$ regret for this problem (i.e., the contextual bandit with i.i.d. contexts and linear payoffs).

Greedy Algorithm. However, this substantial literature requires exploration. Greedy policies are desirable in practical settings where exploration may be costly or unethical.

A related but distinct literature on greedy policies exists for discounted Bayesian multi-armed bandit problems. The seminal paper by Gittins (1979) showed that greedily applying an index policy is optimal for a classical multi-armed bandit in Bayesian regret (with a known prior over the unknown parameters). Woodroofe (1979) and Sarkar (1991) extend this result to a Bayesian one armed bandit with a single i.i.d. covariate when the discount factor approaches 1, and Wang et al. (2005b,a) generalize this result with a single covariate and two arms. Mersereau et al. (2009) further model known structure between arm rewards. However, these policies are not greedy in the same sense as ours; in particular, the Gittins index of an arm is not simply the arm parameter estimate, but includes an additional factor that implicitly captures the value of exploration for under-sampled arms (i.e., the variance of the estimate). In fact, recent work has shown a sharp equivalence between the UCB policy (which notably incorporates exploration) and the Gittins index policy as the discount factor approaches one (Russo 2019). In contrast, we consider a greedy policy with respect to *unbiased* arm parameter estimates, i.e., without accounting for the value of exploration (or the variance of our parameter estimates). It is surprising that such a policy can be effective; in fact, we show that it is not rate optimal in general, but is rate optimal for the linear contextual bandit if there is sufficient randomness in the context distribution.

It is also worth noting that, unlike the literature above, we consider undiscounted minimax regret with unknown and deterministic arm parameters. Gutin and Farias (2016) show that the Gittins analysis does not succeed in minimizing Bayesian regret over all sufficiently large horizons, and propose "optimistic" Gittins indices (which incorporate additional exploration) to solve the undiscounted Bayesian multi-armed bandit.

Since the first draft of our work appeared online, there have been two follow-up papers that cite our work and provide additional theoretical and empirical validation for our results. Kannan et al. (2018) consider the case where an adversary selects the observed contexts, but these contexts are then perturbed by white noise; they find that the greedy algorithm can be rate optimal in this setting even for small perturbations. Bietti et al. (2018) perform an extensive empirical study of contextual bandit algorithms on 524 datasets that are publicly available on the OpenML platform. These datasets arise from a variety of applications including medicine, natural language, and sensors. Bietti et al. (2018) find that the greedy algorithm outperforms a wide range of bandit algorithms in cumulative regret on more that

400 datasets. This study provides strong empirical validation of our theoretical findings.

Covariate Diversity. The adaptive control theory literature has studied “persistent excitation”: for linear models, this condition ensures that the minimum eigenvalue of the covariance matrix grows at a suitable rate, so that the parameter estimates converge over time (Narendra and Annaswamy 1987, Nguyen 2018). Thus, if persistent excitation holds for each arm, we will eventually recover the true arm rewards. However, the problem remains to derive policies that ensure that such a condition holds for each (optimal) arm; classical bandit algorithms achieve this goal with high probability by incorporating exploration for under-sampled arms. Importantly, a greedy policy that does not incorporate exploration may not satisfy this condition, e.g., the greedy policy may “drop” an arm. The covariate diversity assumption ensures that there is sufficient randomness in the observed contexts, thereby exogenously ensuring that persistent excitation holds for each arm regardless of the sample path taken by the bandit algorithm.

Conservative Bandits. Our approach is also related to recent literature on designing conservative bandit algorithms (Wu et al. 2016, Kazerouni et al. 2016) that operate within a safety margin, i.e., the regret is constrained to stay below a certain threshold that is determined by a baseline policy. This literature proposes algorithms that restrict the amount of exploration (similar to the present work) in order to satisfy a safety constraint. Wu et al. (2016) studies the classical multi-armed bandit, and Kazerouni et al. (2016) generalizes these results to the contextual linear bandit.

Dynamic Pricing. Finally, we note that there are technical parallels between our work and the analysis of the greedy policy and its variants in the dynamic pricing literature (Lattimore and Munos 2014, Broder and Rusmevichientong 2012). In particular, the most commonly-studied dynamic pricing problem (without covariates) can be viewed as a linear bandit problem without changing action space and with a modified reward function (den Boer and Zwart 2013, Keskin and Zeevi 2014b). When there are no covariates, the greedy algorithm has been shown to be undesirable since it provably converges to a suboptimal price (a fixed point known as the “uninformative price”) with nonzero probability (den Boer and Zwart 2013, Keskin and Zeevi 2014b, 2015). Thus, bandit-like algorithms have been proposed, which always explore in order to guarantee convergence to the optimal price (den Boer and Zwart 2013, Keskin and Zeevi 2014a,b, den Boer and Zwart 2015); these approaches have similarities to Greedy-First in that they only explore (i.e., deviate from a greedy strategy) when it is necessary, to ensure that the information envelope or variance grows at the optimal rate.

More recently, some have studied dynamic pricing with changing demand covariates (Cohen et al. 2016, Qiang and Bayati 2016, Javanmard and Nazerzadeh 2019, Ban and Keskin 2018) or a changing demand function (den Boer 2015, Keskin and Zeevi 2015). These changes in the demand environment can help the greedy algorithm explore naturally and achieve asymptotically optimal performance. Our work significantly differs from this line of analysis since we need to learn multiple reward functions (for each arm) simultaneously. Specifically, in dynamic pricing, the decision-maker always receives feedback from the true demand function; in contrast, in the contextual bandit, we only receive feedback from a decision if we choose it, thereby complicating the analysis.

2.1.2 Main Contributions and Organization of this Chapter

We begin by studying conditions under which the greedy algorithm performs well. In §2.2, we introduce the *covariate diversity* condition (Assumption 3), and show that it holds for a general class of continuous and discrete context distributions. In §2.3, we show that when covariate diversity holds, the greedy policy is asymptotically optimal for a two-armed contextual bandit with linear rewards (Theorem 1); this result is extended to rewards given by generalized linear models in Proposition 1. For problem instances with more than two arms or where covariate diversity does not hold, we prove that the greedy algorithm is asymptotically optimal with some probability, and we provide a lower bound on this probability (Theorem 2).

Building on these results, in §2.4, we introduce the Greedy-First algorithm that uses observed contexts and rewards to determine whether the greedy algorithm is failing or not via a hypothesis test. If the test detects that the greedy steps are not receiving sufficient exploration, the algorithm switches to a standard exploration-based algorithm. We show that Greedy-First achieves rate optimal regret bounds without our additional assumptions on covariate diversity or number of arms. More importantly, we prove that Greedy-First remains purely greedy (while achieving asymptotically optimal regret) for almost all problem instances for which a pure greedy algorithm is sufficient (Theorem 3). Finally, for problem instances with more than two arms or where covariate diversity does not hold, we prove that Greedy-First remains exploration-free and rate optimal with some probability, and we provide a lower bound on this probability (Theorem 4). This result implies that Greedy-First reduces exploration on average compared to standard bandit algorithms.

Finally, in §3.7, we run several simulations on synthetic and real datasets to verify our theoretical results. We find that the greedy algorithm outperforms standard bandit algorithms when covariate diversity holds, but can perform poorly when this assumption does not hold. However, Greedy-First outperforms standard bandit algorithms even in the absence of covariate diversity, while remaining competitive with the greedy algorithm in the presence of covariate diversity. Thus, Greedy-First provides a desirable compromise between avoiding exploration and learning the true policy.

2.2 Problem Formulation

We consider a K-armed contextual bandit for T time steps, where T is unknown. Each arm i is associated with an unknown parameter $\beta_i \in \mathbb{R}^d$. For any integer n, let $[n]$ denote the set $\{1, ..., n\}$. At each time t, we observe a new individual with context vector $X_t \in \mathbb{R}^d$. We assume that $\{X_t\}_{t \geq 0}$ is a sequence of i.i.d. samples from some unknown distribution that admits probability density $p_X(\mathbf{x})$ with respect to the Lebesgue measure. If we pull arm $i \in [K]$, we observe a stochastic linear reward (in §2.3.4, we discuss how our results can be extended to generalized linear models)

$$Y_{i,t} = X_t^\top \beta_i + \varepsilon_{i,t} \, ,$$

where $\varepsilon_{i,t}$ are independent σ-subgaussian random variables (see Definition 1 below).

Definition 1. *A random variable Z is σ-subgaussian if for all $\tau > 0$ we have $\mathbb{E}[e^{\tau Z}] \leq e^{\tau^2 \sigma^2 / 2}$.*

We seek to construct a sequential decision-making policy π that learns the arm parameters $\{\beta_i\}_{i=1}^K$ over time in order to maximize expected reward for each individual.

We measure the performance of π by its *cumulative expected regret*, which is the standard metric in the analysis of bandit algorithms (Lai and Robbins 1985, Auer 2003). In particular, we compare ourselves to an oracle policy π^*, which knows the arm parameters $\{\beta_i\}_{i=1}^K$ in advance. Upon observing context X_t, the oracle will always choose the best expected arm $\pi_t^* = \max_{j \in [K]}(X_t^\top \beta_j)$. Thus, if we choose an arm $i \in [K]$ at time t, we incur *instantaneous* *expected regret*

$$r_t \equiv \mathbb{E}_{X \sim p_X}\left[\max_{j \in [K]}(X_t^\top \beta_j) - X_t^\top \beta_i\right],$$

which is simply the expected difference in reward between the oracle's choice and our choice. We seek to minimize the cumulative expected regret $R_T := \sum_{t=1}^{T} r_t$. In other words, we seek to mimic the oracle's performance by gradually learning the arm parameters.

Additional Notation: Let B_R^d be the closed ℓ_2 ball of radius R around the origin in $\mathbb{R}^d$ defined as $B_R^d = \left\{x \in \mathbb{R}^d : \|x\|_2 \leq R\right\}$, and let the volume of a set $S \subset \mathbb{R}^d$ be $\mathrm{vol}(S) \equiv \int_S \mathrm{d}\mathbf{x}$.

2.2.1 Assumptions

We now describe the assumptions required for our regret analysis. Some assumptions will be relaxed in later sections of this chapter as noted below.

Our first assumption is that the contexts as well as the arm parameters $\{\beta_i\}_{i=1}^{K}$ are bounded. This ensures that the maximum regret at any time step t is bounded. This is a standard assumption made in the bandit literature (see e.g., Dani et al. 2008).

Assumption 1 (Parameter Set)**.** *There exists a positive constant $x_{\max}$ such that the context probability density p_X has no support outside the ball of radius $x_{\max}$, i.e., $\|X_t\|_2 \leq x_{\max}$ for all t. There also exists a constant $b_{\max}$ such that $\|\beta_i\|_2 \leq b_{\max}$ for all $i \in [K]$.*

Second, we assume that the context probability density p_X satisfies a margin condition, which comes from the classification literature (Tsybakov 2004). We do not require this assumption to prove convergence of the greedy algorithm, but the rate of convergence differs depending on whether it holds. In particular, Goldenshluger and Zeevi (2009) prove matching upper and lower bounds demonstrating that all bandit algorithms achieve $\mathcal{O}(\log T)$ regret when the margin condition holds, but they can achieve up to $\mathcal{O}(\sqrt{T})$ regret when this condition is violated. We can obtain analogous results for the simple greedy algorithm as well (see Appendix A.5.2 for details). This is because the margin condition rules out unusual context distributions that become unbounded near the decision boundary (which has zero measure), thereby making learning difficult.

Assumption 2 (Margin Condition)**.** *There exists a constant $C_0 > 0$ such that for each $\kappa > 0$:*

$$\forall\ i \neq j : \quad \mathbb{P}_X\Big[0 < |X^\top(\beta_i - \beta_j)| \leq \kappa\Big] \leq C_0 \kappa\,.$$

Thus far, we have made generic assumptions that are standard in the bandit literature. Our third assumption introduces the covariate diversity condition, which is essential for proving that the greedy algorithm always converges to the optimal policy. This condition guarantees that no matter what our arm parameter estimates are at time t, there is a diverse set of possible contexts (supported by the context probability density p_X) under which each arm may be chosen.

Assumption 3 (Covariate Diversity)**.** *There exists a positive constant λ_0 such that for each vector $\mathbf{u} \in \mathbb{R}^d$ the minimum eigenvalue of $\mathbb{E}_X\left[XX^\top \mathbb{I}\{X^\top \mathbf{u} \geq 0\}\right]$ is at least λ_0, i.e.,*

$$\lambda_{\min}\Big(\mathbb{E}_X\left[XX^\top \mathbb{I}\{X^\top \mathbf{u} \geq 0\}\right]\Big) \geq \lambda_0 \,.$$

Assumption 3 holds for a general class of distributions. For instance, if the context probability density p_X is bounded below by a nonzero constant in an open set around the origin, then it would satisfy covariate diversity. This includes common distributions such as the uniform or truncated gaussian distributions. Furthermore, discrete distributions such as the classic Rademacher distribution on binary random variables also satisfy covariate diversity.

Remark 2.2.1. *As discussed in the related literature, the adaptive control theory literature has studied "persistent excitation," which is reminiscent of the covariate diversity condition without the indicator function $\mathbb{I}\{X^\top \mathbf{u} \geq 0\}$. If persistent excitation holds for each arm, then the minimum eigenvalue of the corresponding covariance matrix grows at a suitable rate, and the arm parameter estimate converges over time. However, a greedy policy that does not incorporate exploration may not satisfy this condition, e.g., the greedy policy may drop an arm. Assumption 3 ensures that there is sufficient randomness in the observed contexts, thereby exogenously ensuring that persistent excitation holds for each arm (see Lemma 4), regardless of the sample path taken by the bandit algorithm.*

2.2.2 Examples of Distributions Satisfying Assumptions 1-3

While Assumptions 1-2 are generic, it is not straightforward to verify Assumption 3. The following lemma provides sufficient conditions (that are easier to check) that guarantee Assumption 3.

Lemma 1. *If there exists a set $W \subset \mathbb{R}^d$ that satisfies conditions (a), (b), and (c) given below, then p_X satisfies Assumption 3.*

(a) W is symmetric around the origin; i.e., if $\mathbf{x} \in W$ then $-\mathbf{x} \in W$.

(b) There exist positive constants $a, b \in \mathbb{R}$ such that for all $\mathbf{x} \in W$, $a \cdot p_X(-\mathbf{x}) \leq b \cdot p_X(\mathbf{x})$.

(c) There exists a positive constant λ such that $\int_W \mathbf{x}\mathbf{x}^\top p_X(\mathbf{x})\mathrm{d}\mathbf{x} \succeq \lambda I_d$. For discrete distributions, the integral is replaced with a sum.

We now use Lemma 1 to demonstrate that covariate diversity holds for a wide range of continuous and discrete context distributions, and we explicitly provide the corresponding constants. It is straightforward to verify that these examples also satisfy Assumptions 1 and 2.

1. **Uniform Distribution.** Consider the uniform distribution over an arbitrary bounded set V that contains the origin. Then, there exists some $R > 0$ such that $B_R^d \subset V$. Taking $W = B_R^d$, we note that conditions (a) and (b) of Lemma 1 follow immediately. We now check condition (c) by first stating the following lemma (see Appendix A.1 for proof):

   Lemma 2. $\int_{B_R^d} \mathbf{x}\mathbf{x}^\top \mathrm{d}\mathbf{x} = \left[\frac{R^2}{d+2}\mathrm{vol}(B_R^d)\right] I_d$ *for any $R > 0$.*

   By definition, $p_X(\mathbf{x}) = 1/\mathrm{vol}(V)$ for all $\mathbf{x} \in V$, and $\mathrm{vol}(B_R^d) = R^d \mathrm{vol}(B_{x_{\max}}^d)/x_{\max}^d$. Applying Lemma 2, we see that condition (c) of Lemma 1 holds with constant $\lambda = R^{d+2}/[(d+2)x_{\max}^d]$.

2. **Truncated Multivariate Gaussian Distribution.** Let p_X be a multivariate Gaussian distribution $\mathsf{N}(\mathbf{0}_d, \Sigma)$, truncated to 0 for all $\|\mathbf{x}\|_2 \geq x_{\max}$. The density after renormalization is

$$p_X(\mathbf{x}) = \frac{\exp\left(-\frac{1}{2}\mathbf{x}^\top \Sigma^{-1}\mathbf{x}\right)}{\int_{B_{x_{\max}}^d} \exp\left(-\frac{1}{2}\mathbf{z}^\top \Sigma^{-1}\mathbf{z}\right)\mathrm{d}\mathbf{z}} \mathbb{I}(\mathbf{x} \in B_{x_{\max}}^d)\,.$$

   Taking $W = B_{x_{\max}}^d$, conditions (a) and (b) of Lemma 1 follow immediately. Condition (c) of Lemma 1 holds with constant

$$\lambda = \frac{1}{(2\pi)^{d/2}|\Sigma|^{d/2}} \exp\left(-\frac{x_{\max}^2}{2\lambda_{\min}(\Sigma)}\right) \frac{x_{\max}^2}{d+2} \mathrm{vol}(B_{x_{\max}}^d)\,,$$

as shown in Lemma 12 in Appendix A.1.

3. **Gibbs Distributions with Positive Covariance.** Consider the set $\{\pm 1\}^d \subset \mathbb{R}^d$ equipped with a discrete probability density p_X, which satisfies

$$p_X(\mathbf{x}) = \frac{1}{Z} \exp\left(\sum_{1 \le i,j \le d} J_{ij} x_i x_j \right) ,$$

for any $\mathbf{x} = (x_1, x_2, \ldots, x_d) \in \{\pm 1\}^d$. Here, $J_{ij} \in \mathbb{R}$ are (deterministic) parameters, and Z is a normalization term known as the *partition function* in the statistical physics literature. We define $W = \{\pm 1\}^d$, satisfying conditions (a) and (b) of Lemma 1. Furthermore, condition (c) follows by definition since the covariance of the distribution is positive-definite. This class of distributions includes the well-known Rademacher distribution (by setting all $J_{ij} = 0$).

Finally, note that any product of these distributions would also satisfy our assumptions.

Remark 2.2.2. *A special case under which the conditions in Lemma 1 hold is when W is the entire support of the distribution P_X (this is the case in the Gaussian and Gibbs distributions, where $W = B^d_{x_{\max}}$ and $W = \{\pm 1\}^d$ respectively). Now, let $X^{(1)}$ be a random vector that satisfies this special case and has mean 0. Let $X^{(2)}$ be another vector that is independent of $X^{(1)}$ and satisfies the general form of Lemma 1. Then it is easy to see that $X = (X^{(1)}, X^{(2)})$ also satisfies the conditions in Lemma 1. (Parts (a) and (b) clearly hold; to see why (c) holds, note that the cross diagonal entries in $XX^\top$ are zero since $X^{(1)}$ has mean 0.) This construction illustrates how covariate diversity works for distributions that contain a mixture of discrete and continuous components.*

2.3 Greedy Bandit

Notation. Let the *design matrix* $\mathbf{X}$ be the $T \times d$ matrix whose rows are X_t. Similarly, for $i \in [K]$, let Y_i be the length T vector of potential outcomes $X_t^\top \beta_i + \varepsilon_{i,t}$. Since we only obtain feedback when arm i is played, entries of Y_i may be missing. For any $t \in [T]$, let $\mathcal{S}_{i,t} = \{j \mid \pi_j = i\} \cap [t]$ be the set of times when arm i was played within the first t time steps. We use notation $\mathbf{X}(\mathcal{S}_{i,t}), Y(\mathcal{S}_{i,t})$, and $\varepsilon(\mathcal{S}_{i,t})$ to refer to the design matrix, the outcome vector, and vector of idiosyncratic shocks respectively, for the observations in time periods in $\mathcal{S}_{i,t}$. We estimate β_i at time t based on $\mathbf{X}(\mathcal{S}_{i,t})$ and $Y(\mathcal{S}_{i,t})$, using ordinary least squares (OLS) regression that is defined below. We denote this estimator $\hat{\beta}_{\mathbf{X}(\mathcal{S}_{i,t}),Y(\mathcal{S}_{i,t})}$, or $\hat{\beta}(\mathcal{S}_{i,t})$ for short.

Definition 2 (OLS Estimator)**.** *For any $\mathbf{X}_0 \in \mathbb{R}^{n\times d}$ and $Y_0 \in \mathbb{R}^{n\times 1}$, the OLS estimator is $\hat{\beta}_{\mathbf{X}_0,Y_0} \equiv \arg\min_\beta \|Y_0 - \mathbf{X}_0\beta\|_2^2$, which is equal to $(\mathbf{X}_0^\top \mathbf{X}_0)^{-1}\mathbf{X}_0^\top Y_0$ when $\mathbf{X}_0^\top \mathbf{X}_0$ is invertible.*

We now describe the greedy algorithm and provide performance guarantees when covariate diversity holds.

2.3.1 Algorithm

At each time step, we observe a new context X_t and use the current arm estimates $\hat{\beta}(\mathcal{S}_{i,t-1})$ to play the arm with the highest estimated reward, i.e., $\pi_t = \arg\max_{i\in[K]} X_t^\top \hat{\beta}(\mathcal{S}_{i,t-1})$. Upon playing arm π_t, a reward $Y_{\pi_t,t} = X_t^\top \beta_{\pi_t} + \varepsilon_{\pi_t,t}$ is observed. We then update our estimate for arm π_t but we need not update the arm parameter estimates for other arms as $\hat{\beta}(\mathcal{S}_{i,t-1}) = \hat{\beta}(\mathcal{S}_{i,t})$ for $i \neq \pi_t$. The update formula is given by

$$\hat{\beta}(\mathcal{S}_{\pi_t,t}) = \left[\mathbf{X}(\mathcal{S}_{\pi_t,t})^\top \mathbf{X}(\mathcal{S}_{\pi_t,t})\right]^{-1} \mathbf{X}(\mathcal{S}_{\pi_t,t})^\top \mathbf{Y}(\mathcal{S}_{\pi_t,t})\,.$$

We do not update the parameter of arm π_t if $\mathbf{X}(\mathcal{S}_{\pi_t,t})^\top \mathbf{X}(\mathcal{S}_{\pi_t,t})$ is not invertible. The pseudo-code for the algorithm is given in Algorithm 1.

Algorithm 1 Greedy Bandit

```
Initialize $\hat{\beta}(\mathcal{S}_{i,0}) = 0 \in \mathbb{R}^d$ for $i \in [K]$
for $t \in [T]$ do
    Observe $X_t \sim p_X$
    $\pi_t \leftarrow \arg\max_i X_t^\top \hat{\beta}(\mathcal{S}_{i,t-1})$ (break ties randomly)
    $\mathcal{S}_{\pi_t,t} \leftarrow \mathcal{S}_{\pi_t,t-1} \cup \{t\}$
    Play arm $\pi_t$, observe $Y_{\pi_t,t} = X_t^\top \beta_{\pi_t} + \varepsilon_{\pi_t,t}$
    If $\mathbf{X}(\mathcal{S}_{\pi_t,t})^\top \mathbf{X}(\mathcal{S}_{\pi_t,t})$ is invertible*, update the arm parameter $\hat{\beta}(\mathcal{S}_{\pi_t,t})$ via
        $$\hat{\beta}(\mathcal{S}_{\pi_t,t}) \leftarrow \left[\mathbf{X}(\mathcal{S}_{\pi_t,t})^\top \mathbf{X}(\mathcal{S}_{\pi_t,t})\right]^{-1} \mathbf{X}(\mathcal{S}_{\pi_t,t})^\top \mathbf{Y}(\mathcal{S}_{\pi_t,t})$$
end for
```

*In practice, until $\mathbf{X}(\mathcal{S}_{\pi_t,t})^\top \mathbf{X}(\mathcal{S}_{\pi_t,t})$ becomes non-singular, using ridge regression or pseudo inverse for estimating $\hat{\beta}(\mathcal{S}_{\pi_t,t})$ may reduce the regret. See also Remark 2.3.2.

2.3.2 Regret Analysis of Greedy Bandit with Covariate Diversity

We establish an upper bound of $\mathcal{O}(\log T)$ on the cumulative expected regret of the Greedy Bandit for the two-armed contextual bandit when covariate diversity is satisfied.

Theorem 1. *If $K = 2$ and Assumptions 1-3 are satisfied, the cumulative expected regret of the Greedy Bandit at time $T \geq 3$ is at most*

$$\begin{aligned} R_T(\pi) &\leq \frac{128 C_0 \bar{C} x_{\max}^4 \sigma^2 d (\log d)^{3/2}}{\lambda_0^2} \log T \\ &\quad + \bar{C} \left(\frac{128 C_0 x_{\max}^4 \sigma^2 d (\log d)^{3/2}}{\lambda_0^2} + \frac{160 b_{\max} x_{\max}^3 d}{\lambda_0} + 2 x_{\max} b_{\max} \right) \qquad (2.3.1) \\ &\leq C_{GB} \log T = \mathcal{O}\left(\log T\right) \ , \end{aligned}$$

where the constant C_0 is defined in Assumption 2 and

$$\bar{C} = \left(\frac{1}{3} + \frac{7}{2} (\log d)^{-0.5} + \frac{38}{3} (\log d)^{-1} + \frac{67}{4} (\log d)^{-1.5} \right) \in (1/3, 52) \, . \qquad (2.3.2)$$

We prove an analogous result for the greedy algorithm in the case where arm rewards are given by generalized linear models (see Appendix 2.3.4 and Proposition 1 for details).

Remark 2.3.1. *Goldenshluger and Zeevi (2013) established a lower bound of $\mathcal{O}(\log T)$ for any algorithm in a two-armed contextual bandit. While they do not make Assumption 3, the distribution used in their proof satisfies Assumption 3; thus their result applies to our setting. Combined with our upper bound (Theorem 1), we conclude that the Greedy Bandit is rate optimal.*

2.3.3 Proof Strategy

Notation. Let $\mathcal{R}_i = \left\{ \mathbf{x} \in \mathcal{X} : \mathbf{x}^\top \beta_i \geq \max_{j \neq i} \mathbf{x}^\top \beta_j \right\}$ denote the true set of contexts where arm i is optimal. Then, let $\hat{\mathcal{R}}_{i,t}^{\pi} = \left\{ \mathbf{x} \in \mathcal{X} : \mathbf{x}^\top \hat{\beta}(\mathcal{S}_{i,t-1}) \geq \max_{j \neq i} \mathbf{x}^\top \hat{\beta}(\mathcal{S}_{j,t-1}) \right\}$ denote the estimated set of contexts at time t where arm i appears optimal; in other words, if the context $X_t \in \hat{\mathcal{R}}_{i,t}^{\pi}$, then the greedy policy will choose arm i at time t. (since we assume without loss of generality that ties are broken randomly as selected by π and thus, $\{\mathcal{R}_i\}_{i=1}^K$ and $\{\hat{\mathcal{R}}_{i,t}^{\pi}\}_{i=1}^K$ partition the context space $\mathcal{X}$.)

For any $t \in [T]$, let $\mathcal{H}_{t-1} = \sigma\left(\mathbf{X}_{1:t}, \pi_{1:t-1}, Y_1(\mathcal{S}_{1,t-1}), Y_2(\mathcal{S}_{2,t-1}), \ldots, Y_K(\mathcal{S}_{K,t-1})\right)$ denote the σ-algebra containing all observed information up to time t before taking an action; thus, our policy π_t is $\mathcal{H}_{t-1}$-measurable. Furthermore, let $\mathcal{H}_{t-1}^{-}$ be the σ-algebra containing all observations *before* time t, i.e., $\mathcal{H}_{t-1}^{-} = \sigma\left(\mathbf{X}_{1:t-1}, \pi_{1:t-1}, Y_1(\mathcal{S}_{1,t-1}), Y_2(\mathcal{S}_{2,t-1}), \ldots, Y_K(\mathcal{S}_{K,t-1})\right)$.

Define $\hat{\Sigma}(\mathcal{S}_{i,t}) = \mathbf{X}(\mathcal{S}_{i,t})^\top \mathbf{X}(\mathcal{S}_{i,t})$ as the sample covariance matrix for observations from arm i up to time t. We may compare this to the expected covariance matrix for arm i under the greedy policy, defined as $\tilde{\Sigma}_{i,t} = \sum_{k=1}^{t} \mathbb{E}\left[X_k X_k^\top \mathbb{I}[X_k \in \hat{\mathcal{R}}_{i,k}^{\pi}] \mid \mathcal{H}_{k-1}^{-}\right]$.

Proof Strategy. Intuitively, covariate diversity (Assumption 3) guarantees that there is sufficient randomness in the observed contexts, which creates natural "exploration." In particular, no matter what our current arm parameter estimates $\{\hat{\beta}\left(\mathcal{S}_{1,t}\right), \hat{\beta}\left(\mathcal{S}_{2,t}\right)\}$ are at time t, each arm will be chosen by the greedy policy with at least some constant probability (with respect to p_X) depending on the observed context. We formalize this intuition in the following lemma.

Lemma 3. *Given Assumptions 1 and 3, the following holds for any* $\mathbf{u} \in \mathbb{R}^d$*:*

$$\mathbb{P}_X[\mathbf{x}^\top \mathbf{u} \geq 0] \geq \frac{\lambda_0}{x_{\max}^2}\,.$$

Proof. For any observed context $\mathbf{x}$, note that $\mathbf{x}\mathbf{x}^\top \preceq x_{\max}^2 I_d$ by Assumption 1. Re-stating Assumption 3 for each $\mathbf{u} \in \mathbb{R}^d$, we can write

$$\lambda_0 I_d \preceq \int \mathbf{x}\mathbf{x}^T \mathbb{I}(\mathbf{x}^\top \mathbf{u} \geq 0) p_X(\mathbf{x}) \mathrm{d}\mathbf{x} \preceq x_{\max}^2 I_d \int \mathbb{I}(\mathbf{x}^\top u \geq 0) p_X(\mathbf{x}) \mathrm{d}\mathbf{x} = x_{\max}^2 \mathbb{P}_X[\mathbf{x}^\top \mathbf{u} \geq 0] I_d,$$

since the indicator function and p_X are both nonnegative. □

Taking $\mathbf{u} = \hat{\beta}\left(\mathcal{S}_{1,t}\right) - \hat{\beta}\left(\mathcal{S}_{2,t}\right)$, Lemma 3 implies that arm 1 will be pulled with probability at least $\lambda_0 / x_{\max}^2$ at each time t; the claim holds analogously for arm 2. Thus, each arm will be played at least $\lambda_0 T / x_{\max}^2 = \mathcal{O}(T)$ times in expectation. However, this is not sufficient to guarantee that each arm parameter estimate $\hat{\beta}_i$ converges to the true parameter β_i. In Lemma 4, we establish a sufficient condition for convergence.

First, we show that covariate diversity guarantees that the minimum eigenvalue of each arm's expected covariance matrix $\tilde{\Sigma}_{i,t}$ under the greedy policy grows linearly with t. This result implies that not only does each arm receive a sufficient number of observations under the greedy policy, but also that these observations are sufficiently diverse (in expectation). Next, we apply a standard matrix concentration inequality (see Lemma 14 in Appendix A.2) to show that the minimum eigenvalue of each arm's sample covariance matrix $\hat{\Sigma}(\mathcal{S}_{i,t})$ also grows linearly with t. This will guarantee the convergence of our regression estimates for each arm parameter.

Lemma 4. *Take* $C_1 = \lambda_0/(40x_{\max}^2)$*. Given Assumptions 1 and 3, the following holds for the minimum eigenvalue of the empirical covariance matrix of each arm* $i \in [2]$*:*

$$\mathbb{P}\left[\lambda_{\min}\left(\hat{\Sigma}(\mathcal{S}_{i,t})\right) \geq \lambda_0 t/4\right] \geq 1 - \exp(\log d - C_1 t)\,.$$

Remark 2.3.2. *Note that the above lemma in particular implies that, the probability that the matrix* $\hat{\Sigma}(\mathcal{S}_{i,t}) = \mathbf{X}(\mathcal{S}_{i,t})^\top \mathbf{X}(\mathcal{S}_{i,t})$ *is singular, is upper bounded by* $\exp(\log d - C_1 t)$*. Hence, as* t *grows, this probability vanishes at an exponential rate, ensuring that its contribution to the final regret is at most a constant. In fact, The second term in the regret upper bound stated in Lemma 6 captures this term. Therefore, in practice, using ridge regression or pseudo inverse until* $\mathbf{X}(\mathcal{S}_{i,t})^\top \mathbf{X}(\mathcal{S}_{i,t})$ *becomes non-singular may reduce the final regret by a constant (which may depend on* d*). However, this would not alter the final regret of* $\mathcal{O}(\log T)$*.*

Proof. Without loss of generality, take $i = 1$. For any $k \leq t$, let $\mathbf{u}_k = \hat{\beta}(\mathcal{S}_{1,k}) - \hat{\beta}(\mathcal{S}_{2,k})$; by the greedy policy, we pull arm 1 if $X_k^\top \mathbf{u}_{k-1} > 0$ and arm 2 if $X_k^\top \mathbf{u}_{k-1} < 0$ (ties are broken randomly using a fair coin flip W_k). Thus, the estimated set of optimal contexts for arm 1 is

$$\hat{\mathcal{R}}_{1,k} = \left\{\mathbf{x} \in \mathcal{X} : \mathbf{x}^\top \mathbf{u}_{k-1} > 0\right\} \cup \left\{\mathbf{x} \in \mathcal{X} : \mathbf{x}^\top \mathbf{u}_{k-1} = 0, W_k = 0\right\}.$$

First, we seek to bound the minimum eigenvalue of the expected covariance matrix $\tilde{\Sigma}_{1,t} = \sum_{k=1}^{t} \mathbb{E}\left[X_k X_k^\top \mathbb{I}[X_k \in \hat{\mathcal{R}}_{1,k}] \mid \mathcal{H}_{k-1}^{-}\right]$. Expanding one term in the sum, we can write

$$\begin{aligned}
\mathbb{E}\left[X_k X_k^\top \mathbb{I}[X_k \in \hat{\mathcal{R}}_{1,k}] \mid \mathcal{H}_{k-1}^{-}\right] &= \mathbb{E}\left[X_k X_k^\top \left(\mathbb{I}[X_k^\top \mathbf{u}_{k-1} > 0] + \mathbb{I}[X_k^\top \mathbf{u}_{k-1} = 0, W_k = 0\right) \mid \mathcal{H}_{k-1}^{-}\right] \\
&= \mathbb{E}_X\left[XX^\top \left(\mathbb{I}[X^\top \mathbf{u}_{k-1} > 0] + \frac{1}{2}\mathbb{I}[X^\top \mathbf{u}_{k-1} = 0]\right)\right] \\
&\geq \lambda_0/2\,,
\end{aligned}$$

where the last line follows from Assumption 3. Since the minimum eigenvalue function

$\lambda_{\min}(\cdot)$ is concave over positive semi-definite matrices, we can write

$$\begin{aligned}\lambda_{\min}\left(\tilde{\Sigma}_{1,t}\right) &= \lambda_{\min}\left(\sum_{k=1}^{t}\mathbb{E}\left[XX^\top\mathbb{I}[X\in\hat{\mathcal{R}}_{1,k}]\mid\mathcal{H}_{k-1}^{-}\right]\right)\\ &\geq \sum_{k=1}^{t}\lambda_{\min}\left(\mathbb{E}\left[XX^\top\mathbb{I}[X\in\hat{\mathcal{R}}_{1,k}]\mid\mathcal{H}_{k-1}^{-}\right]\right)\geq\frac{\lambda_0 t}{2}\,.\end{aligned}$$

Next, we seek to use matrix concentration inequalities (Lemma 14 in Appendix A.2) to bound the minimum eigenvalue of the sample covariance matrix $\hat{\Sigma}(\mathcal{S}_{1,t})$. To apply the concentration inequality, we also need to show an upper bound on the maximum eigenvalue of $X_kX_k^\top$; this follows trivially from Assumption 1 using the Cauchy-Schwarz inequality:

$$\lambda_{\max}(X_kX_k^\top)=\max_{\mathbf{u}}\frac{\|X_kX_k^\top\mathbf{u}\|_2}{\|\mathbf{u}\|_2}\leq\frac{\|X_k\|_2^2\|\mathbf{u}\|_2}{\|\mathbf{u}\|_2}\leq x_{\max}^2.$$

We can apply Lemma 14, taking the finite adapted sequence $\{X_k\}$ to be $\left\{X_kX_k^\top\mathbb{I}[X_k\in\hat{\mathcal{R}}_{1,k}]\right\}$, so that $Y=\hat{\Sigma}(\mathcal{S}_{1,t})$ and $W=\tilde{\Sigma}_{1,t}$. We also take $R=x_{\max}^2$ and $\gamma=1/2$. Thus, we have

$$\begin{aligned}\mathbb{P}_X\left[\lambda_{\min}\left(\hat{\Sigma}(\mathcal{S}_{1,t})\right)\leq\frac{\lambda_0 t}{4}\ \text{ and }\ \lambda_{\min}\left(\tilde{\Sigma}_{1,t}\right)\geq\frac{\lambda_0 t}{2}\right] &\leq d\left(\frac{e^{-0.5}}{0.5^{0.5}}\right)^{\frac{\lambda_0}{4x_{\max}^2}t}\\ &\leq\exp\left(\log d-\frac{0.1\lambda_0}{4x_{\max}^2}t\right),\end{aligned}$$

using the fact $-0.5-0.5\log(0.5)\leq-0.1$. As we showed earlier, $\mathbb{P}_X\left(\lambda_{\min}\left(\tilde{\Sigma}_{1,t}\right)\geq\frac{\lambda_0 t}{2}\right)=1$. This proves the result. □

Next, Lemma 5 guarantees with high probability that each arm's parameter estimate has small ℓ_2 error with respect to the true parameter if the minimum eigenvalue of the sample covariance matrix $\hat{\Sigma}(\mathcal{S}_{i,t})$ has a positive lower bound. Note that we cannot directly use results on the convergence of the OLS estimator since the set of samples $\mathcal{S}_{i,t}$ from arm i at time t are not i.i.d. (we use the arm estimate $\hat{\beta}(\mathcal{S}_{i,t-1})$ to decide whether to play arm i at time t; thus, the samples in $\mathcal{S}_{i,t}$ are correlated.). Instead, we use a Bernstein concentration inequality to guarantee convergence with adaptive observations.

Lemma 5. *Taking $C_2=\lambda^2/(2d\sigma^2x_{\max}^2)$ and $n\geq|\mathcal{S}_{i,t}|$, we have for all $\lambda,\chi>0$,*

$$\mathbb{P}\left[\|\hat{\beta}(\mathcal{S}_{i,t})-\beta_i\|_2\geq\chi\ \ \textit{and}\ \ \lambda_{\min}\left(\hat{\Sigma}(\mathcal{S}_{i,t})\right)\geq\lambda t\right]\leq 2d\exp\left(-C_2t^2\chi^2/n\right).$$

Proof of Lemma 5. We begin by noting that if the event $\lambda_{\min}\left(\hat{\Sigma}(\mathcal{S}_{i,t})\right) \geq \lambda t$ holds, then

$$\begin{aligned}\|\hat{\beta}(\mathcal{S}_{i,t}) - \beta_i\|_2 &= \| \left(\mathbf{X}(\mathcal{S}_{i,t})^\top \mathbf{X}(\mathcal{S}_{i,t})\right)^{-1} \mathbf{X}(\mathcal{S}_{i,t})^\top \varepsilon(\mathcal{S}_{i,t})\|_2 \\ &\leq \| \left(\mathbf{X}(\mathcal{S}_{i,t})^\top \mathbf{X}(\mathcal{S}_{i,t})\right)^{-1} \|_2 \|\mathbf{X}(\mathcal{S}_{i,t})^\top \varepsilon(\mathcal{S}_{i,t})\|_2 \ \leq \ \frac{1}{\lambda t}\|\mathbf{X}(\mathcal{S}_{i,t})^\top \varepsilon(\mathcal{S}_{i,t})\|_2 .\end{aligned}$$

As a result, we can write

$$\begin{aligned}&\mathbb{P}\left[\|\hat{\beta}(\mathcal{S}_{i,t}) - \beta_i\|_2 \geq \chi \ \text{ and } \ \lambda_{\min}\left(\hat{\Sigma}(\mathcal{S}_{i,t})\right) \geq \lambda t\right] \\ &\quad = \mathbb{P}\left[\|\hat{\beta}(\mathcal{S}_{i,t}) - \beta_i\|_2 \geq \chi \ \mid \ \lambda_{\min}\left(\hat{\Sigma}(\mathcal{S}_{i,t})\right) \geq \lambda t\right] \mathbb{P}\left[\lambda_{\min}\left(\hat{\Sigma}(\mathcal{S}_{i,t})\right) \geq \lambda t\right] \\ &\quad \leq \mathbb{P}\left[\|\mathbf{X}(\mathcal{S}_{i,t})^\top \varepsilon(\mathcal{S}_{i,t})\|_2 \geq \chi t \lambda \ \mid \ \lambda_{\min}\left(\hat{\Sigma}(\mathcal{S}_{i,t})\right) \geq \lambda t\right] \mathbb{P}\left[\lambda_{\min}\left(\hat{\Sigma}(\mathcal{S}_{i,t})\right) \geq \lambda t\right] \\ &\quad \leq \mathbb{P}\left[\|\mathbf{X}(\mathcal{S}_{i,t})^\top \varepsilon(\mathcal{S}_{i,t})\|_2 \geq \chi t \lambda\right] \\ &\quad \leq \sum_{r=1}^{d} \mathbb{P}\left[|\varepsilon(\mathcal{S}_{i,t})^\top \mathbf{X}(\mathcal{S}_{i,t})^{(r)}| \geq \frac{\lambda t \cdot \chi}{\sqrt{d}}\right] ,\end{aligned}$$

where $\mathbf{X}^{(r)}$ denotes the r^{th} column of $\mathbf{X}$. We can expand

$$\varepsilon(\mathcal{S}_{i,t})^\top \mathbf{X}(\mathcal{S}_{i,t})^{(r)} = \sum_{j=1}^{t} \varepsilon_j X_{j,r} \mathbb{I}\left[j \in \mathcal{S}_{i,j}\right] .$$

For simplicity, define $D_j = \varepsilon_j X_{j,r} \mathbb{I}\left[j \in \mathcal{S}_{i,j}\right]$. First, note that D_j is $(x_{\max}\sigma)$-subgaussian, since ε_j is σ-subgaussian and $|X_{j,r}| \leq x_{\max}$. Next, note that $X_{j,r}$ and $\mathbb{I}\left[j \in \mathcal{S}_{i,j}\right]$ are both $\mathcal{H}_{j-1}$ measurable; taking the expectation gives $\mathbb{E}[D_j \mid \mathcal{H}_{j-1}] = X_{j,r}\mathbb{I}\left[j \in \mathcal{S}_{i,j}\right] \mathbb{E}[\varepsilon_j \mid \mathcal{H}_{j-1}] = 0$. Thus, the sequence $\{D_j\}_{j=1}^{t}$ is a martingale difference sequence adapted to the filtration $\mathcal{H}_1 \subset \mathcal{H}_2 \subset \cdots \subset \mathcal{H}_t$. Applying a standard Bernstein concentration inequality (see Lemma 13 in Appendix A.2), we can write

$$\mathbb{P}\left[\Big|\sum_{j=1}^{t} D_j\Big| \geq \frac{\lambda t \cdot \chi}{\sqrt{d}}\right] \leq 2\exp\left(-\frac{t^2\lambda^2\chi^2}{2d\sigma^2 x_{\max}^2 n}\right),$$

where n is an upper bound on the number of nonzero terms in above sum, i.e., an upper bound on $|\mathcal{S}_{i,t}|$. This yields the desired result. □

To summarize, Lemma 4 provides a lower bound (with high probability) on the minimum eigenvalue of the sample covariance matrix. Lemma 5 states that if such a bound holds on the minimum eigenvalue of the sample covariance matrix, then the estimated parameter $\hat{\beta}(\mathcal{S}_{i,t})$ is close to the true β_i (with high probability). Having established convergence of the arm parameters under the Greedy Bandit, one can use a standard peeling argument (as in Goldenshluger and Zeevi (2013)) to bound the instantaneous expected regret of the Greedy Bandit algorithm.

Lemma 6. *Define $\mathcal{F}_{i,t}^{\lambda} = \left\{\lambda_{\min}\left(\mathbf{X}(\mathcal{S}_{i,t})^\top \mathbf{X}(\mathcal{S}_{i,t})\right) \geq \lambda t\right\}$. Then, the instantaneous expected regret of the Greedy Bandit at time $t \geq 2$ satisfies*

$$r_t(\pi) \leq \frac{4(K-1)C_0 \bar{C} x_{\max}^2 (\log d)^{3/2}}{C_3} \frac{1}{t-1} + 4(K-1) b_{\max} x_{\max} \left(\max_i \mathbb{P}[\overline{\mathcal{F}_{i,t-1}^{\lambda_0/4}}]\right) ,$$

where $C_3 = \lambda_0^2/(32 d \sigma^2 x_{\max}^2)$, C_0 is defined in Assumption 2, and $\bar{C}$ is defined in Theorem 1.

Note that $\mathbb{P}[\overline{\mathcal{F}_{i,t-1}^{\lambda_0/4}}]$ can be upper bounded using Lemma 4. Substituting this in the upper bound derived on $r_t(\pi)$ in Lemma 6, and using $R_T(\pi) = \sum_{t=1}^T r_t(\pi)$ finishes the proof of Theorem 1.

2.3.4 Generalized Linear Rewards

In this section, we discuss how our results generalize when the arm rewards are given by a generalized linear model (GLM). Now, upon playing arm i after observing context X_t, the decision-maker realizes a reward $Y_{i,t}$ with expectation $\mathbb{E}[Y_{i,t}] = \mu(X_t^\top \beta_i)$, where μ is the inverse link function. For instance, in logistic regression, this would correspond to a binary reward $Y_{i,t}$ with $\mu(z) = 1/(1+\exp(-z))$; in Poisson regression, this would correspond to an integer-valued reward $Y_{i,t}$ with $\mu(z) = \exp(z)$; in linear regression, this would correspond to $\mu(z) = z$.

In order to describe the greedy policy in this setting, we give a brief overview of the exponential family, generalized linear model, and maximum likelihood estimation.

Exponential family. A univariate probability distribution belongs to the *canonical exponential family* if its density with respect to a reference measure (e.g., Lebesgue measure) is given by

$$p_\theta(z) = \exp\left[z\theta - A(\theta) + B(z)\right] , \tag{2.3.3}$$

where θ is the underlying real-valued parameter, $A(\cdot)$ and $B(\cdot)$ are real-valued functions, and $A(\cdot)$ is assumed to be twice continuously differentiable. For simplicity, we assume the reference measure is the Lebesgue measure. It is well known that if Z is distributed according to the above canonical exponential family, then it satisfies $\mathbb{E}[Z] = A'(\theta)$ and $\mathrm{Var}[Z] = A''(\theta)$, where A' and A'' denote the first and second derivatives of the function A with respect to θ, and A is strictly convex (see e.g., Lehmann and Casella 1998).

Generalized linear model (GLM). The natural connection between exponential families and GLMs is provided by assuming that the density of $Y_{i,t}$ for the context X_t and arm i is given by $g_{\beta_i}(Y_{i,t} \mid X_t) = p_{X_t^\top \beta_i}(Y_{i,t})$. where p is defined in (2.3.3). In other words, the reward upon playing arm i for context X_t is $Y_{i,t}$ with density

$$\exp\left[Y_{i,t} X_t^\top \beta_i - A(X_t^\top \beta_i) + B(Y_{i,t})\right] .$$

Using the aforementioned properties of the exponential family, $\mathbb{E}[Y_{i,t}] = A'(X_t^\top \beta_i)$, i.e., the link function $\mu = A'$. This implies that μ is continuously differentiable and its derivative is A''. Thus, μ is strictly increasing since A is strictly convex.

Maximum likelihood estimation. Suppose that we have n samples $\{(X_i, Y_i)\}_{i=1}^n$ from a distribution with density $g_\beta(Y \mid X)$. The maximum likelihood estimator of β based on this sample is given by

$$\arg\max_{\beta} \sum_{\ell=1}^{n} \log g_\beta(Y_\ell \mid X_\ell) = \arg\max_{\beta} \sum_{\ell=1}^{n} \left[Y_\ell X_\ell^\top \beta - A(X_\ell^\top \beta) + B(Y_\ell)\right] . \tag{2.3.4}$$

Since A is strictly convex (so $-A$ is strictly concave), the solution to (2.3.4) can be obtained efficiently (see e.g., McCullagh and Nelder 1989). It is not hard to see that whenever $\mathbf{X}^\top \mathbf{X}$ is positive definite, this solution is unique (see Appendix A.5.1 for a proof). We denote this unique solution by $h_\mu(\mathbf{X}, \mathbf{Y})$.

Now we are ready to generalize the Greedy Bandit algorithm when the arm rewards are given by a GLM. Using similar notation as in the linear reward case, given the estimates $\left\{\hat{\beta}(\mathcal{S}_{i,t-1})\right\}_{i \in [K]}$ at time t, the greedy policy plays the arm that maximizes expected estimated reward, i.e.,

$$\pi_t = \arg\max_{i \in [K]} \mu\left(X_t^\top \hat{\beta}(\mathcal{S}_{i,t-1})\right) .$$

Since μ is a strictly increasing function, this translates to $\pi_t = \arg\max_{i\in[K]} X_t^\top \hat{\beta}(\mathcal{S}_{i,t-1})$.

Algorithm 2 Greedy Bandit for Generalized Linear Models

```
**Input parameters:** inverse link function $\mu$
Initialize $\hat{\beta}(\mathcal{S}_{i,0}) = 0$ for $i \in [K]$
**for** $t \in [T]$ **do**
    Observe $X_t \sim p_X$
    $\pi_t \leftarrow \arg\max_i X_t^\top \hat{\beta}(\mathcal{S}_{i,t-1})$ (break ties randomly)
    Play arm $\pi_t$, observe $Y_{i,t} = \mu(X_t^\top \beta_{\pi_t}) + \varepsilon_{\pi_t,t}$
    Update $\hat{\beta}(\mathcal{S}_{\pi_t,t}) \leftarrow h_\mu\left(\mathbf{X}(\mathcal{S}_{\pi_t,t}), \mathbf{Y}(\mathcal{S}_{\pi_t,t})\right)$, where $h_\mu(\mathbf{X}, \mathbf{Y})$ is the solution to the maximum likelihood estimation in Equation (2.3.4)
**end for**
```

Next, we state the following result (proved in Appendix A.5.1) that Algorithm 2 achieves logarithmic regret when $K = 2$ and the covariate diversity assumption holds.

Proposition 1. *Consider arm rewards given by a GLM with σ-subgaussian noise $\varepsilon_{i,t} = Y_{i,t} - \mu(X_t^\top \beta_i)$. Define $m_\theta = \min\left\{\mu'(z) : z \in [-(b_{\max}+\theta)x_{\max}, (b_{\max}+\theta)x_{\max}]\right\}$. If $K = 2$ and Assumptions 1-3 are satisfied, the cumulative expected regret of Algorithm 2 at time T is at most*

$$\begin{aligned} R_T(\pi) &\leq \frac{128 C_0 \bar{C}_\mu L_\mu x_{\max}^4 \sigma^2 d}{\lambda_0^2} \log T + \bar{C}_\mu L_\mu \left(128 \frac{C_0 x_{\max}^4 \sigma^2 d}{\lambda_0^2} + 160 \frac{b_{\max} x_{\max}^3 d}{\lambda_0} + 2 x_{\max} b_{\max}\right) \\ &= \mathcal{O}\left(\log T\right) , \end{aligned}$$

where the constant C_0 is defined in Assumption 2, L_μ is the Lipschitz constant[1] *of the function $\mu(\cdot)$ on the interval $[-x_{\max}b_{\max}, x_{\max}b_{\max}]$, and $\bar{C}_\mu$ is defined as $\bar{C}_\mu = \frac{1}{3}\left(\frac{\sqrt{\log 4d}}{m_{b_{\max}}} + 1\right)^3 + \frac{3}{2}\left(\frac{\sqrt{\log 4d}}{m_{b_{\max}}} + 1\right)^2 + \frac{8}{3}\left(\frac{\sqrt{\log 4d}}{m_{b_{\max}}} + 1\right) + \frac{1}{m_{b_{\max}}^3}\left(\left(\frac{\sqrt{\log 4d}}{m_{b_{\max}}} + 1\right)\frac{m_{b_{\max}}}{2} + \frac{1}{4}\right) + \frac{1}{m_{b_{\max}}^2} + \frac{1}{2m_{b_{\max}}}$.*

2.3.5 Regret Analysis of Greedy Bandit without Covariate Diversity

Thus far, we have shown that the greedy algorithm is rate optimal when there are only two arms and in the presence of covariate diversity in the observed context distribution. However, when these additional assumptions do not hold, the greedy algorithm may fail to converge to the true arm parameters and achieve linear regret. We now show that a greedy approach achieves rate optimal performance with *some probability* even when these assumptions do not hold. This result will motivate the design of the Greedy-First algorithm in Appendix 2.4.

[1] Exists by continuity of $\mu' = A''$.

Assumptions. For the rest of the chapter, we allow the number of arms $K > 2$, and remove Assumption 3 on covariate diversity. Instead, we will make the following weaker Assumption 4, which is typically made in the contextual bandit literature (see e.g., Goldenshluger and Zeevi 2013, Bastani and Bayati 2015), which allows for multiple arms, and relaxes the assumption on observed contexts (e.g., allowing for intercept terms in the arm parameters).

Assumption 4 (Positive-Definiteness). *Let $\mathcal{K}_{opt}$ and $\mathcal{K}_{sub}$ be mutually exclusive sets that include all K arms. Sub-optimal arms $i \in \mathcal{K}_{sub}$ satisfy $\mathbf{x}^\top \beta_i < \max_{j \neq i} \mathbf{x}^\top \beta_j - h$ for some $h > 0$ and every $\mathbf{x} \in \mathcal{X}$. On the other hand, each optimal arm $i \in \mathcal{K}_{opt}$, has a corresponding set $U_i = \{\mathbf{x} \mid \mathbf{x}^\top \beta_i > \max_{j \neq i} \mathbf{x}^\top \beta_j + h\}$. Define $\Sigma_i \equiv \mathbb{E}\left[XX^\top \mathbb{I}(X \in U_i)\right]$ for all $i \in \mathcal{K}_{opt}$. Then, there exists $\lambda_1 > 0$ such that for all $i \in \mathcal{K}_{opt}$, $\lambda_{\min}(\Sigma_i) \geq \lambda_1 > 0$.*

Remark 2.3.3. *This assumption is slightly different as stated than the assumptions made in prior literature; however, these assumptions are equivalent for bounded context distributions p_X (Assumption 1). We discuss the comparison in Appendix A.4 for completeness.*

Algorithm. We consider a small modification of the Greedy Bandit (Algorithm 1), by initializing each arm parameter estimate with $m > 0$ random samples. Note that OLS requires at least d samples for an arm parameter estimate to be well-defined, and Algorithm 1 does not update the arm parameter estimates from the initial ad-hoc value of 0 until this stage is reached (i.e., the covariance matrix $\mathbf{X}(\mathcal{S}_{i,t})^\top \mathbf{X}(\mathcal{S}_{i,t})$ for a given arm i becomes invertible); thus, all actions up to that point are essentially random. Consequently, we argue that initializing each arm parameter with $m = d$ samples at the beginning is qualitatively no different than Algorithm 1. We consider general values of m to study how the probabilistic guarantees of the greedy algorithm vary with the number of initial samples.

Remark 2.3.4. *We note that there is a class of explore-then-exploit bandit algorithms that follow a similar strategy of randomly sampling each arm for a length of time and using those estimates for the remaining horizon (Bubeck and Cesa-Bianchi 2012). However, (i) m is a function of the horizon length T in these algorithms (typically $m = \sqrt{T}$) while we consider m to be a (small) constant with respect to T, and (ii) these algorithms do not follow a greedy strategy since they do not update the parameter estimates after the initialization phase.*

Result. The following theorem shows that the Greedy Bandit converges to the correct policy and achieves rate optimal performance with at least some problem-specific probability.

Theorem 2. *Under Assumptions 1, 2, and 4, Greedy Bandit achieves logarithmic cumulative regret with probability at least*

$$S^{gb}(m, K, \sigma, x_{\max}, \lambda_1, h) := 1 - \inf_{\gamma \in (0,1), \delta > 0, p \geq Km+1} L(\gamma, \delta, p)\,, \tag{2.3.5}$$

where the function $L(\gamma, \delta, p)$ is defined as

$$\begin{aligned} L(\gamma, \delta, p) &:= 1 - \mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m}) \geq \delta\right]^K + 2Kd\; \mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m}) \geq \delta\right] \exp\left\{-\frac{h^2 \delta}{8d\sigma^2 x_{\max}^2}\right\} \\ &+ \sum_{j=Km+1}^{p-1} 2d \exp\left\{-\frac{h^2\delta^2}{8d(j-(K-1)m)\sigma^2 x_{\max}^4}\right\} + \frac{d \exp\left(-D_1(\gamma)(p - m|\mathcal{K}_{sub}|)\right)}{1 - \exp(-D_1(\gamma))} \\ &+ \frac{2d \exp\left(-D_2(\gamma)(p - m|\mathcal{K}_{sub}|)\right)}{1 - \exp(-D_2(\gamma))}\,. \end{aligned} \tag{2.3.6}$$

Here $\mathbf{X}_{1:m}$ denotes the matrix obtained by drawing m random samples from distribution p_X and the constants are

$$D_1(\gamma) = \frac{\lambda_1(\gamma + (1-\gamma)\log(1-\gamma))}{x_{\max}^2}\,, \tag{2.3.7}$$

$$D_2(\gamma) = \frac{\lambda_1^2 h^2 (1-\gamma)^2}{8d\sigma^2 x_{\max}^4}\,. \tag{2.3.8}$$

Proof Strategy. The proof of Theorem 2 is provided in Appendix A.7. We observe that if all arm parameter estimates remain within a Euclidean distance of $\theta_1 = h/(2x_{\max})$ from their true values for all time periods $t > m$, then the Greedy Bandit converges to the correct policy and is rate optimal. We derive lower bounds on the probability that this event occurs using Lemma 5, after proving suitable lower bounds on the minimum eigenvalue of the covariance matrices. The key steps are as follows:

1. Assuming that the minimum eigenvalue of the sample covariance matrix for each arm is above some threshold value $\delta > 0$, we derive a lower bound on the probability that after initialization, each arm parameter estimates lie within a ball of radius $\theta_1 = h/(2x_{\max})$ centered around the true arm parameter.

2. Next, we derive a lower bound on the probability that these estimates remain within this ball after $p \geq Km + 1$ rounds for some choice of p.

3. We use the concentration result in Lemma 14 to derive a lower bound on the probability that the minimum eigenvalue of the sample covariance matrix of each arm in $\mathcal{K}_{opt}$ is above $(1-\gamma)\lambda_1(t - m|\mathcal{K}_{sub}|)$ for any $t \geq p$.

4. We derive a lower bound on the probability that the estimates ultimately remain inside the ball with radius θ_1. This ensures that no sub-optimal arm is played for any $t \geq Km$.

5. Summing up these probability terms implies Theorem 2. The parameters γ, δ, and p can be chosen arbitrarily and we optimize over their choice.

The following Proposition 2 illustrates some of the properties of the function S^{gb} in Theorem 2 with respect to problem-specific parameters. The proof is provided in Appendix A.7.

Proposition 2. *The function* $S^{gb}(m, K, \sigma, x_{\max}, \lambda_1, h)$ *defined in Equation* (2.3.5) *is non-increasing with respect to* σ *and* K*; it is non-decreasing with respect to* m*,* λ_1 *and* h*. Furthermore, the limit of this function when* σ *goes to zero is*

$$\mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^{\top}\mathbf{X}_{1:m}) > 0\right]^K .$$

In other words, the greedy algorithm is more likely to succeed when there is less noise and when there are fewer arms; it is also more likely to succeed with additional initialization samples, when the optimal arms each have a larger probability of being the best arm under p_X, and when the sub-optimal arms are worse than the optimal arms by a larger margin. Intuitively, these conditions make it easier for the Greedy Bandit to avoid "dropping a good arm" early on, which would result in its convergence to the wrong policy. As the noise goes to zero, the greedy algorithm always succeeds as long as the sample covariance matrix for each of the K arms is positive definite after the initialization periods.

In Corollary 1, we simplify the expression in Theorem 2 for better readability. However, the simplified expression leads to poor tail bounds when m is close to d, while the general expression in Theorem 2 works when $m = d$ as demonstrated later in §2.4.3 (see Figure 2.1).

Corollary 1. *Under the assumptions of Theorem 2, Greedy Bandit achieves logarithmic cumulative regret with probability at least*

$$1 - \frac{3Kd\exp(-D_{\min}|\mathcal{K}_{opt}|m)}{1-\exp(-D_{\min})},$$

where function $D_{\min}$ is defined as $D_{\min} = \min\left\{\frac{0.153\lambda_1}{x_{\max}^2}, \frac{\lambda_1^2 h^2}{32d\sigma^2 x_{\max}^4}\right\}$.

To summarize, these probabilistic guarantees on the success of Greedy Bandit suggest that a greedy approach can be effective and rate optimal in general with at least some probability. Therefore, in the next section, we introduce the Greedy-First algorithm which executes a greedy strategy and only resorts to forced exploration when the observed data suggests that the greedy updates are not converging. This helps eliminate unnecessary exploration with high probability.

2.4 Greedy-First Algorithm

As noted in Theorem 1, the optimality of the Greedy Bandit requires that there are only two arms and that the context distribution satisfies covariate diversity. The latter condition rules out some standard settings, e.g., the arm rewards cannot have an intercept term (since the addition of a one to every context vector would violate Assumption 3). While there are many examples that satisfy these conditions (see §2.2.2), the decision-maker may not know a priori whether a greedy algorithm is appropriate for her particular setting. Thus, we introduce the Greedy-First algorithm (Algorithm 3), which is rate optimal without these additional assumptions, but seeks to use the greedy algorithm without forced exploration when possible.

2.4.1 Algorithm

The Greedy-First algorithm has two inputs λ_0 and t_0. It starts by following the greedy algorithm up to time t_0, after which it iteratively checks whether all the arm parameter estimates are converging to their true values at a suitable rate. A sufficient statistic for checking this is simply the minimum eigenvalue of the sample covariance matrix of each arm; if this value is above the threshold of $\lambda_0 t/4$, then greedy estimates are converging with high probability. On the other hand, if this condition is not met, the algorithm switches to a standard bandit algorithm with forced exploration. We choose the OLS Bandit algorithm (introduced by Goldenshluger and Zeevi (2013) for two arms and extended to the general setting by Bastani and Bayati (2015)), which is provided in Appendix A.4 for completeness.

Remark 2.4.1. *Greedy-First can switch to any contextual bandit algorithm (e.g., OFUL by Abbasi-Yadkori et al. (2011) or Thompson sampling by Agrawal and Goyal (2013), Russo*

Algorithm 3 Greedy-First Bandit

```
Input parameters: $\lambda_0, t_0$
Initialize $\hat{\beta}(\mathcal{S}_{i,0})$ at random for $i \in [K]$
Initialize switch to $R = 0$
for $t \in [T]$ do
    if $R \neq 0$ then break
    end if
    Observe $X_t \sim p_X$
    $\pi_t \leftarrow \arg\max_i X_t^\top \hat{\beta}(\mathcal{S}_{i,t-1})$ (break ties randomly)
    $\mathcal{S}_{\pi_t,t} \leftarrow \mathcal{S}_{\pi_t,t-1} \cup \{t\}$
    Play arm $\pi_t$, observe $Y_{i,t} = X_t^\top \beta_{\pi_t} + \varepsilon_{\pi_t,t}$
    Update arm parameter $\hat{\beta}(\mathcal{S}_{\pi_t,t}) = \left[\mathbf{X}(\mathcal{S}_{\pi_t,t})^\top \mathbf{X}(\mathcal{S}_{\pi_t,t})\right]^{-1} \mathbf{X}(\mathcal{S}_{\pi_t,t})^\top \mathbf{Y}(\mathcal{S}_{\pi_t,t})$
    Compute covariance matrices $\hat{\Sigma}(\mathcal{S}_{i,t}) = \mathbf{X}(\mathcal{S}_{i,t})^\top \mathbf{X}(\mathcal{S}_{i,t})$ for $i \in [K]$
    if $t > t_0$ and $\min_{i \in [K]} \lambda_{\min}\left(\hat{\Sigma}_(\mathcal{S}_{i,t})\right) < \frac{\lambda_0 t}{4}$ then
        Set $R = t$
    end if
end for
Execute OLS Bandit for $t \in [R+1, T]$
```

and Van Roy (2014a)) instead of the OLS Bandit. Then, the assumptions used in the theoretical analysis would be replaced with analogous assumptions required by that algorithm. Our proof naturally generalizes to adopt the assumptions and regret guarantees of the new algorithm when Greedy Bandit fails.

In practice, λ_0 may be an unknown constant. Thus, we suggest the following heuristic routine to estimate this parameter:

1. Execute Greedy Bandit for t_0 time steps.

2. Estimate λ_0 using the observed data via $\hat{\lambda}_0 = \frac{1}{2t_0} \min_{i \in [K]} \lambda_{\min}\left(\hat{\Sigma}(\mathcal{S}_{i,t_0})\right)$.

3. If $\hat{\lambda}_0 = 0$, this suggests that one of the arms is not receiving sufficient samples, and thus, Greedy-First will switch to OLS Bandit immediately. Otherwise, execute Greedy-First for $t \in [t_0 + 1, T]$ with $\lambda_0 = \hat{\lambda}_0$.

The pseudo-code for this heuristic is given in Appendix A.4. The regret guarantees of Greedy-First (given in the next section) are always valid, but the choice of the input parameters may affect the empirical performance of Greedy-First and the probability with which it remains exploration-free. For example, if t_0 is too small, then Greedy-First may incorrectly switch to OLS Bandit even when a greedy algorithm will converge; thus, choosing $t_0 \gg Kd$ is advisable.

2.4.2 Regret Analysis of Greedy-First

As noted in §2.3.5, we replace the more restrictive assumption on covariate diversity (Assumption 3) with a more standard assumption made in the bandit literature (Assumption 2). Theorem 3 establishes an upper bound of $\mathcal{O}(\log T)$ on the expected cumulative regret of Greedy-First. Furthermore, we establish that Greedy-First remains purely greedy with high probability when there are only two arms and covariate diversity is satisfied.

Theorem 3. *The cumulative expected regret of Greedy-First at time T is at most*

$$C \log T + 2t_0 x_{\max} b_{\max} \,,,$$

where $C = (K-1)C_{GB} + C_{OB}$, C_{GB} is the constant defined in Theorem 1, and C_{OB} is the coefficient of $\log(T)$ *in the upper bound of the regret of the OLS Bandit algorithm.*

Furthermore, if Assumption 3 is satisfied (with the specified parameter λ_0) and $K = 2$, then the Greedy-First algorithm will purely execute the greedy policy (and will not switch to the OLS Bandit algorithm) with probability at least $1-\delta$, where $\delta = 2d\exp[-t_0 C_1]/C_1$, and $C_1 = \lambda_0/40x_{\max}^2$. Note that δ can be made arbitrarily small since t_0 is an input parameter to the algorithm.

The key insight to this result is that the proof of Theorem 1 only requires Assumption 3 in the proof of Lemma 4. The remaining steps of the proof hold without the assumption. Thus, if the conclusion of Lemma 4, $\min_{i\in[K]} \lambda_{\min}(\hat{\Sigma}(\mathcal{S}_{i,t})) \geq \frac{\lambda_0 t}{4}$ holds at every $t \in [t_0{+}1, T]$, then we are guaranteed at most $\mathcal{O}\left(\log T\right)$ regret by Theorem 1, regardless of whether Assumption 3 holds.

Proof. fProof of Theorem 3. First, we will show that Greedy-First achieves asymptotically optimal regret. Note that the expected regret during the first t_0 rounds is upper bounded by $2x_{\max}b_{\max}t_0$. For the period $[t_0+1, T]$ we consider two cases: (1) the algorithm pursues a purely greedy strategy, i.e., $R = 0$, or (2) the algorithm switches to the OLS Bandit algorithm, i.e., $R \in [t_0+1, T]$.

Case 1: By construction, we know that $\min_{i\in[K]} \lambda_{\min}\left(\hat{\Sigma}(\mathcal{S}_{i,t})\right) \geq \lambda_0 t/4$, for all $t > t_0$. This is because Greedy-First only switches when the minimum eigenvalue of the sample covariance matrix for some arm is less than $\lambda_0 t/4$. Therefore, if the algorithm does not switch, it implies that the minimum eigenvalue of each arm's sample covariance matrix is greater that or equal to $\lambda_0 t/4$ for all values of $t > t_0$. Then, the conclusion of Lemma 4 holds in this time range ($\mathcal{F}_{i,t}^{\lambda}$ holds for all $i \in [K]$). Consequently, even if Assumption 3 does not hold and $K \neq 2$, Lemma 6 holds and provides an upper bound on the expected regret r_t. This implies that the regret bound of Theorem 1, after multiplying by $(K-1)$, holds for Greedy-First. Therefore, Greedy-First is guaranteed to achieve $(K-1)C_{GB} \log(T - t_0)$ regret in the period $[t_0+1, T]$ for some constant C_{GB} that depends only on p_X, b and σ. Hence, the regret in this case is upper bounded by $2x_{\max} b_{\max} t_0 + (K-1)C_{GB} \log T$.

Case 2: Once again, by construction, we know that $\min_{i \in [K]} \lambda_{\min}\left(\hat{\Sigma}(\mathcal{S}_{i,t})\right) \geq \lambda_0 t/4$ for all $t \in [t_0+1, R]$ before the switch. Then, using the same argument as in Case 1, Theorem 1 guarantees that we achieve at most $(K-1)C_{GB} \log(R - t_0)$ regret for some constant C_{GB} over the interval $[t_0+1, R]$. Next, Theorem 2 of Bastani and Bayati (2015) guarantees that, under Assumptions 1, 2 and 2, the OLS Bandit's cumulative regret in the interval $t \in [R+1, T]$ is upper bounded by $C_{OB} \log(T - R)$ for some constant C_{OB}. Thus, the total regret is at most $2x_{\max} b_{\max} t_0 + ((K-1)C_{GB} + C_{OB}) \log T$. Note that although the switching time R is a random variable, the upper bound on the cumulative regret $2x_{\max} b_{\max} t_0 + ((K-1)C_{GB} + C_{OB}) \log T$ holds uniformly regardless of the value of R.

Thus, the Greedy-First algorithm always achieves $\mathcal{O}(\log T)$ cumulative regret. Next, we prove that when Assumption 3 holds and $K = 2$, the Greedy-First algorithm maintains a purely greedy policy with high probability. In particular, Lemma 4 states that if the specified λ_0 satisfies $\lambda_{\min}\left(\mathbb{E}_X\left[XX^\top \mathbb{I}(X^\top \mathbf{u} \geq 0)\right]\right) \geq \lambda_0$ for each vector $\mathbf{u} \in \mathbb{R}^d$, then at each time t,

$$\mathbb{P}\left[\lambda_{\min}\left(\hat{\Sigma}(\mathcal{S}_{i,t})\right) \geq \frac{\lambda_0 t}{4}\right] \geq 1 - \exp\left[\log d - C_1 t\right] ,$$

where $C_1 = \lambda_0 / 40 x_{\max}^2$. Thus, by using a union bound over all $K = 2$ arms, the probability that the algorithm switches to the OLS Bandit algorithm is at most

$$K \sum_{t=t_0+1}^{T} \exp\left[\log d - C_1 t\right] \leq 2 \int_{t_0}^{\infty} \exp\left[\log d - C_1 t\right] \mathrm{d}t = \frac{2d}{C_1} \exp\left[-t_0 C_1\right] .$$

This concludes the proof. □

2.4.3 Probabilistic Guarantees for Greedy-First Algorithm

The key value proposition of Greedy-First is to reduce forced exploration when possible. Theorem 2 established that Greedy-First eliminates forced exploration entirely with high probability when there are only two arms and when covariate diversity holds. However, a natural question might be the extent to which Greedy-First reduces forced exploration in general problem instances.

To answer this question, we leverage the probabilistic guarantees we derived for the greedy algorithm in §2.3.5. Note that unlike the greedy algorithm, Greedy-First always achieves rate optimal regret. We now study the probability with which Greedy-First is purely greedy under an arbitrary number of arms K and the less restrictive Assumption 2. However, we impose that all K arms are optimal for some set of contexts under p_X, i.e., $\mathcal{K}_{opt} = [K], \mathcal{K}_{sub} = \emptyset$. This is because Greedy-First *always* switches to the OLS Bandit when an arm is sub-optimal across all contexts. In order for any algorithm to achieve logarithmic cumulative regret, sub-optimal arms must be assigned fewer samples over time and thus, the minimum eigenvalue of the sample covariance matrices of those arms cannot grow sufficiently fast; as a result, the Greedy-First algorithm will switch with probability 1. This may be practically desirable as the decision-maker can decide whether to "drop" the arm and proceed greedily or to use an exploration-based algorithm when the switch triggers.

Theorem 4. *Let Assumptions 1, 2, and 4 hold and suppose that $\mathcal{K}_{sub} = \emptyset$. Then, with probability at least*

$$S^{gf}(m, K, \sigma, x_{\max}, \lambda_1, h) = 1 - \inf_{\gamma \le 1-\lambda_0/(4\lambda_1), \delta>0, Km+1 \le p \le t_0} L'(\gamma, \delta, p)\,, \tag{2.4.1}$$

Greedy-First remains purely greedy (does not switch to an exploration-based bandit algorithm) and achieves logarithmic cumulative regret. The function L' is closely related to the function L from Theorem 2, and is defined as

$$L'(\gamma, \delta, p) = L(\gamma, \delta, p) + (K-1)\frac{d \exp(-D_1(\gamma)p)}{1 - \exp(-D_1(\gamma))}\,. \tag{2.4.2}$$

The proof of Theorem 4 is provided in Appendix A.7. The steps followed are similar to that of the proof of Theorem 2. In the third step of the proof strategy of Theorem 2 (see §2.3.5), we used concentration results to derive a lower bound on the probability that the minimum eigenvalue of the sample covariance matrix of all arms in $\mathcal{K}_{opt}$ are above $(1-\gamma)\lambda_1 t$ for any $t \geq p$ (note that we are assuming $\mathcal{K}_{sub} = \emptyset$ in this section). For Greedy Bandit, this result was only required for the *played arm*; in contrast, for Greedy-First to remain greedy, *all arms* are required to have the minimum eigenvalues of their sample covariance matrices above $(1-\gamma)\lambda_1 t$. This causes the difference in L and L' since we need a union bound over all K arms. The additional constraints on p ensure that the Greedy-First algorithm does not switch,

The following Proposition 3 illustrates some of the properties of the function S^{gf} in Theorem 4 with respect to problem-specific parameters. The proof is provided in Appendix A.7.

Proposition 3. *The function $S^{gf}(m, K, \sigma, x_{\max}, \lambda_1, h)$ defined in Equation* (2.4.1) *is non-increasing with respect to σ and K; it is non-decreasing with respect to λ_1 and h. Furthermore, the limit of this function when σ goes to zero is*

$$\mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m}) > 0\right]^K - \frac{Kd\exp(-D_1(\gamma^*)t_0)}{1-\exp(-D_1(\gamma^*))},$$

where $\gamma^* = 1 - \lambda_0/(4\lambda_1)$.

These relationships mirror those in Proposition 2, i.e., Greedy-First is more likely to remain exploration-free when Greedy Bandit is more likely to succeed. In particular, Greedy-First is more likely to avoid exploration entirely when there is less noise and when there are fewer arms; it is also more likely to avoid exploration with additional initialization samples and when the optimal arms each have a larger probability of being the best arm under p_X. Intuitively, these conditions make it easier for the greedy algorithm to avoid "dropping" an arm, so the minimum eigenvalue of each arm's sample covariance matrix grows at a suitable rate over time, allowing Greedy-First to remain greedy.

In Corollary 2, we simplify the expression in Theorem 4 for better readability. However, the simplified expression leads to poor tail bounds when m is close to d, while the general expression in Theorem 4 works when $m = d$ as demonstrated in Figure 2.1.

Corollary 2. *Under the assumptions made in Theorem 4, Greedy-First remains purely greedy and achieves logarithmic cumulative regret with probability at least*

$$1 - \frac{3Kd\exp(-D_{\min}Km)}{1-\exp(-D_{\min})},$$

where the function $D_{\min}$ is defined in Corollary 1.

We now illustrate the probabilistic bounds given in Theorems 2 and 4 through a simple example.

Example 2.4.1. *Let $K = 3$ and $d = 2$. Suppose that arm parameters are given by $\beta_1 = (1, 0), \beta_2 = (-1/2, \sqrt{3}/2)$ and $\beta_3 = (-1/2, -\sqrt{3}/2)$. Furthermore, suppose that the distribution of covariates p_X is the uniform distribution on the unit ball $B_1^2 = \{\mathbf{x} \in \mathbb{R}^2 \mid \|x\| \leq 1\}$, implying $x_{\max} = 1$. The constants h and λ_1 are chosen to satisfy Assumption 4; here, we choose $h = 0.3$, and $\lambda_1 \approx 0.025$. We then numerically plot our lower bounds on the probability of success of the Greedy Bandit (Theorem 2) and on the probability that Greedy-First remains greedy (Theorem 4) via Equations (2.3.5) and (2.4.1) respectively. Figure 2.1 depicts these probabilities as a function of the noise σ for several values of initialization samples m.*

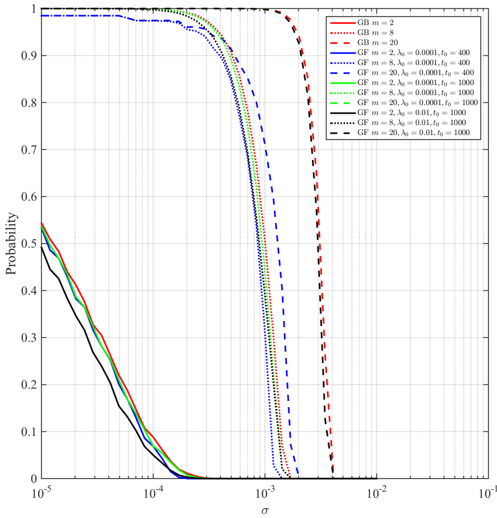


Figure 2.1: Lower (theoretical) bound on the probability of success for Greedy Bandit and Greedy-First. For $m = 20, t_0 = 1000$, the performance of Greedy-First for $\lambda_0 \in \{0.01, 0.0001\}$ are similar and indistinguishable.

We note that our lower bounds are very conservative, and in practice, both Greedy Bandit and Greedy-First succeed and remain exploration-free respectively with much larger probability. For instance, as observed in Example 2.4.1, one can optimize over the choice of λ_1 and h. In the next section, we verify via simulations that both Greedy Bandit and Greedy-First are successful with a higher probability than our lower bounds may suggest.

2.5 Simulations

We now validate our theoretical findings on synthetic and real datasets.

2.5.1 Synthetic Data

Linear Reward. We compare Greedy Bandit and Greedy-First with state-of-the-art contextual bandit algorithms. These include:

1. *OFUL* by Abbasi-Yadkori et al. (2011), which builds on the original upper confidence bound (UCB) approach of Lai and Robbins (1985),

2. *Prior-dependent TS* by Russo and Van Roy (2014b), which builds on the original Thompson sampling approach of Thompson (1933),

3. *Prior-free TS* by Agrawal and Goyal (2013), which builds on the original Thompson sampling approach of Thompson (1933), and

4. *OLS Bandit* by Goldenshluger and Zeevi (2013), which builds on ϵ-greedy methods.

Remark 2.5.1. *Prior-dependent TS requires knowledge of the prior distribution of arm parameters β_i, while prior-free TS does not. All algorithms above require knowledge of an upper bound on the noise variance σ.*

Following the setup of (Russo and Van Roy 2014b), we consider Bayes regret over randomly-generated arm parameters. In particular, for each scenario, we generate 1000 problem instances and sample the true arm parameters $\{\beta_i\}_{i=1}^K$ independently. At each time step within each instance, new context vectors are drawn i.i.d. from a fixed context distribution p_X. We then plot the average Bayes regret across all these instances, along with the 95% confidence interval, as a function of time t with a horizon length $T = 10,000$. We take $K = 2$ and $d = 3$ (see Appendix A.6 for simulations with other values of K and d). The noise variance $\sigma^2 = 0.25$.

We consider four different scenarios, varying (i) whether covariate diversity holds, and (ii) whether algorithms have knowledge of the true prior. The first condition allows us to explore how the performance of Greedy Bandit and Greedy-First compare against benchmark bandit algorithms when conditions are favorable / unfavorable for the greedy approach. The second condition helps us understand how knowledge of the prior distribution and noise variance affects the performance of benchmark algorithms relative to Greedy Bandit and Greedy-First (which do not require this knowledge). When the correct prior is provided, we assume that OFUL and both versions of TS know the noise variance.

Context vectors: For scenarios where covariate diversity holds, we sample the context vectors from a truncated Gaussian distribution, i.e., $0.5 \times \mathsf{N}(\mathbf{0}_d, \mathbf{I}_d)$ truncated to have ℓ_∞ norm at most 1. For scenarios where covariate diversity does not hold, we generate the context vectors the same way but we add an intercept term.

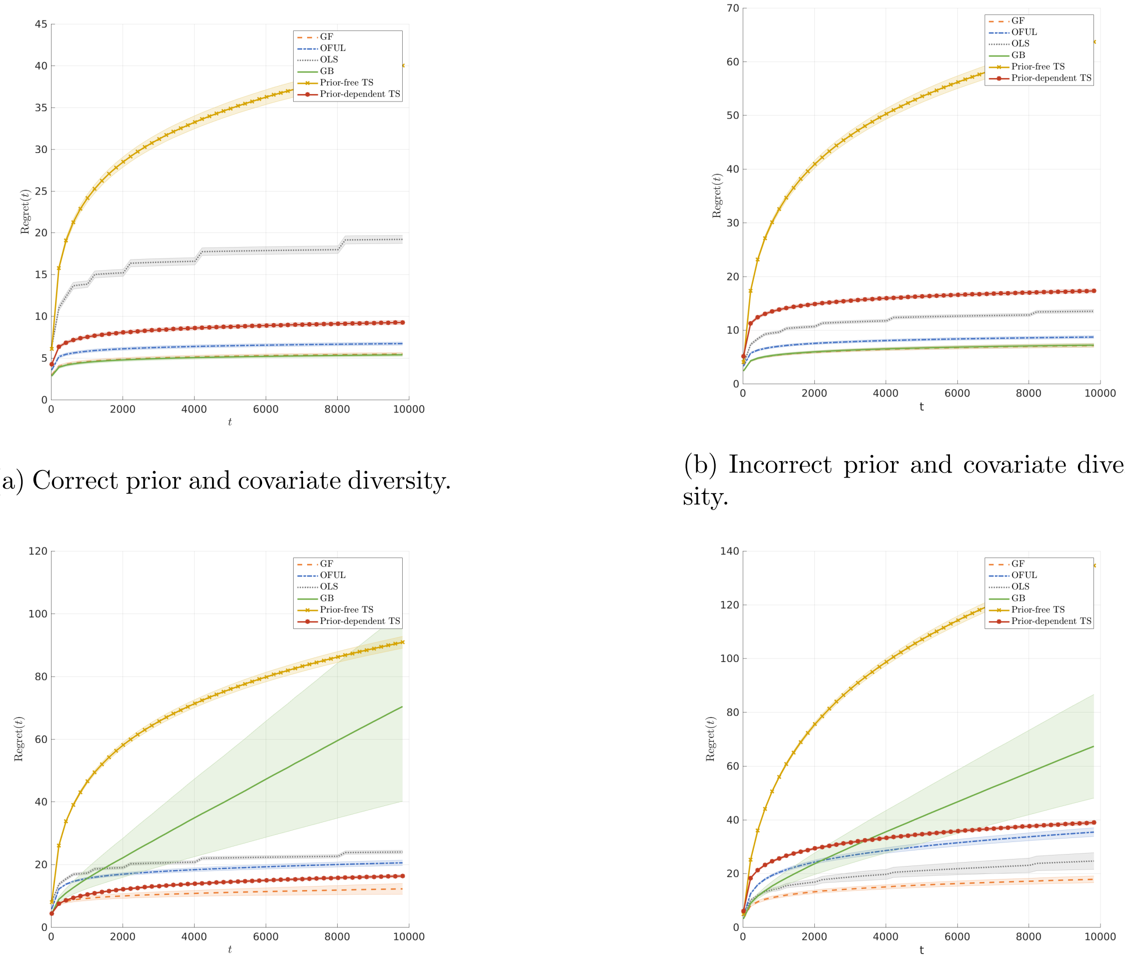
Arm parameters and prior: For scenarios where the algorithms have knowledge of the true prior, we sample the arm parameters $\{\beta_i\}$ independently from $\mathsf{N}(\mathbf{0}_d, \mathbf{I}_d)$, and provide all algorithms with knowledge of σ, and prior-dependent TS with the additional knowledge of the true prior distribution of arm parameters. For scenarios where the algorithms do not have knowledge of the true prior, we sample the arm parameters $\{\beta_i\}$ independently from a mixture of Gaussians, i.e., they are sampled from the distribution $0.5 \times \mathsf{N}(\mathbf{1}_d, \mathbf{I}_d)$ with probability 0.5 and from the distribution $0.5 \times \mathsf{N}(-\mathbf{1}_d, \mathbf{I}_d)$ with probability 0.5. However, prior-dependent TS is given the following incorrect prior distribution over the arm parameters: $10 \times \mathsf{N}(\mathbf{0}_d, \mathbf{I}_d)$. None of the algorithms in this scenario are given knowledge of σ; rather, this parameter is sequentially estimated over time using past data within the algorithm. Parameters of OLS Bandit are chosen according to $h = 5, q = 1$, that are also used for Greedy-First when it switches to OLS Bandit. For Greedy-First, $t_0 = 4Kd$ in all experiments.

Results. Figure 2.2 shows the cumulative Bayes regret of all the algorithms for the four different scenarios discussed above (with and without covariate diversity, with and without the true prior). When covariate diversity holds (a-b), the Greedy Bandit is the clear frontrunner, and Greedy-First achieves the same performance since it never switches to OLS Bandit. However, when covariate diversity does not hold (c-d), we see that the Greedy Bandit performs very poorly (achieving linear regret), but Greedy-First is the clear frontrunner. This is because the greedy algorithm succeeds a significant fraction of the time (Theorem 2), but fails on other instances. Thus, always following the greedy algorithm yields poor performance, but a standard bandit algorithm like the OLS Bandit explores unnecessarily in the instances where a greedy algorithm would have sufficed. Greedy-First leverages this observation by only exploring (switching to OLS Bandit) when the greedy algorithm has failed (with high probability), thereby outperforming both Greedy Bandit and OLS Bandit. Thus, Greedy-First provides a desirable compromise between avoiding exploration and learning the true policy.

Logistic Reward. We now move beyond linear rewards and explore how the performance of Greedy Bandit (Algorithm 2) compares to other bandit algorithms for GLM rewards when covariate diversity holds. We compare to the state-of-the-art GLM-UCB algorithm (Filippi et al. 2010), which is designed to handle GLM reward functions unlike the bandit algorithms from the previous section. Our reward is logistic, i.e, $Y_{it} = 1$ with probability $1/[1 + \exp(-X_t^\top \beta_i)]$ and is 0 otherwise.

We again consider Bayes regret over randomly-generated arm parameters. For each scenario, we generate 10 problem instances (due to the computational burden of solving a maximum likelihood estimation step in each iteration) and sample the true arm parameters $\{\beta_i\}_{i=1}^K$ independently. At each time step within each instance, new context vectors are drawn i.i.d. from a fixed context distribution p_X. We then plot the average Bayes regret across all these instances, along with the 95% confidence interval, as a function of time t with a horizon length $T = 2,000$. Once again, we sample the context vectors from a truncated Gaussian distribution, i.e., $0.5 \times \mathsf{N}(\mathbf{0}_d, \mathbf{I}_d)$ truncated to have ℓ_2 norm at most $x_{\max}$. Note that this context distribution satisfies covariate diversity. We take $K = 2$, and we sample the arm parameters $\{\beta_i\}$ independently from $\mathsf{N}(\mathbf{0}_d, \mathbf{I}_d)$. We consider two different scenarios for d and $x_{\max}$. In the first scenario, we take $d = 3, x_{\max} = 1$; in the second scenario, we take $d = 10, x_{\max} = 5$.

Results: Figure 2.3 shows the cumulative Bayes regret of the Greedy Bandit and GLM-UCB algorithms for the two different scenarios discussed above. As is evident from these results, the Greedy Bandit far outperforms GLM-UCB. We suspect that this is due to the conservative construction of confidence sets in GLM-UCB, particularly for large values of d and $x_{\max}$. In particular, the radius of the confidence set in GLM-UCB is proportional to $(\inf_{z \in C} \mu'(z))^{-1}$ where $C = \{z \mid z \in [-x_{\max} b_{\max}, x_{\max} b_{\max}]\}$. Hence, the radius of the confidence set scales as $\exp(x_{\max} b_{\max})$, which is exponentially large in $x_{\max}$. This can be seen from the difference in Figure 2.3 (a) and (b); in (b),$x_{\max}$ is much larger, causing GLM-UCB's performance to severely degrade. Although the same quantity appears in the theoretical analysis of Greedy Bandit for GLM (Proposition 1), the empirical performance of Greedy Bandit appears much better.



(a) Correct prior and covariate diversity.

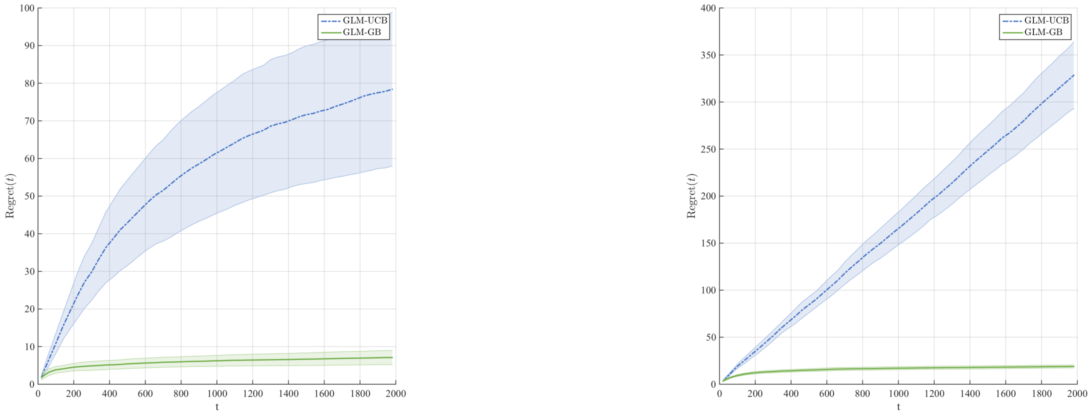
(b) Incorrect prior and covariate diversity.

(c) Correct prior and no covariate diversity.

(d) Incorrect prior and no covariate diversity.

Figure 2.2:]

Expected regret of all algorithms on synthetic data in four different regimes for the covariate diversity condition and whether OFUL and TS are provided with correct or incorrect information on true prior distribution of the parameters. Out of 1000 runs of each simulation Greedy-First never switched in (a) and (b) and switched only 69 times in (c) and 139 times in (d).



(a) $d = 3, x_{\max} = 1$

(b) $d = 10, x_{\max} = 5$

Figure 2.3: Expected regret of GLM-GB and GLM-UCB on synthetic data for logistic reward

Additional Simulations. We explore the performance of Greedy Bandit as a function of K and d; we find that the performance of Greedy Bandit improves dramatically as the dimension d increases, while it degrades with the number of arms K (as predicted by Proposition 2). We also study the dependence of the performance of Greedy-First on the input parameters t_0 (which determines when to switch) and h, q (which are inputs to OLS Bandit after switching); we find that the performance of Greedy-First is quite robust to the choice of inputs. Note that Greedy Bandit is entirely parameter-free. These simulations can be found in Appendix A.6.

2.5.2 Simulations on Real Datasets

We now explore the performance of Greedy and Greedy-First with respect to competing algorithms on real datasets. As mentioned earlier, Bietti et al. (2018) performed an extensive empirical study of contextual bandit algorithms on 524 datasets that are publicly available on the OpenML platform, and found that the greedy algorithm outperforms a wide range of bandit algorithms in cumulative regret on more that 400 datasets. We take a closer look at 3 healthcare-focused datasets ((a) EEG, (b) Eye Movement, and (c) Cardiotocography) among these. We also study the (d) warfarin dosing dataset (Consortium 2009), a publicly available patient dataset that was used by Bastani and Bayati (2015) for analyzing contextual bandit algorithms.

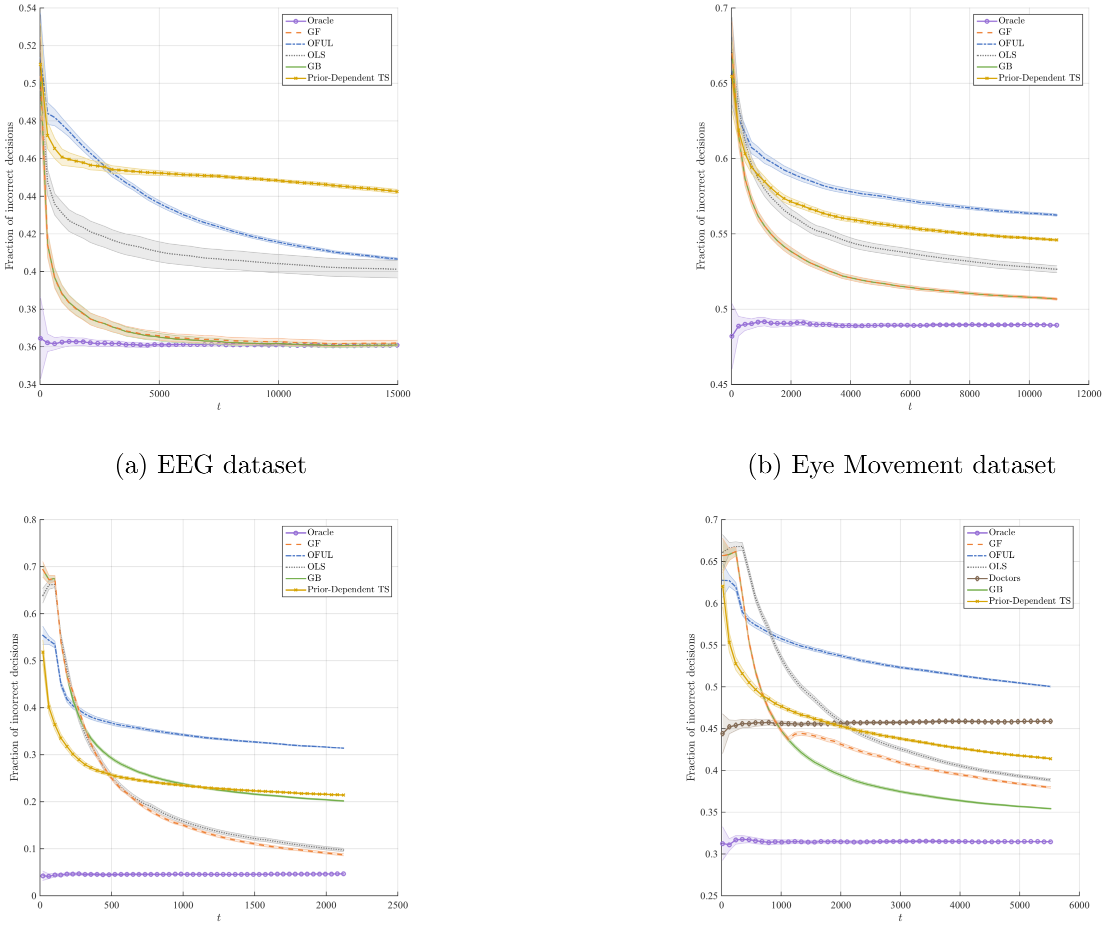
Setup: These datasets all involve classification tasks using patient features. Accordingly, we take the number of decisions K to be the number of classes, and consider a binary reward (1 if we output the correct class, and 0 otherwise). The dimension of the features for datasets (a)-(d) is 14, 27, 35 and 93 respectively; similarly, the number of arms is 2, 3, 3, and 3 respectively.

Remark 2.5.2. *Note that we are now evaluating regret rather than Bayes regret. This is because our arm parameters are given by the true data, and are not simulated from a known prior distribution.*

We compare to the same algorithms as in the previous section, i.e., OFUL, prior-dependent TS, prior-free TS, and OLS Bandit. As an additional benchmark, we also include an oracle policy, which uses the best linear model trained on *all the data* in hindsight; thus, one cannot perform better than the oracle policy using linear models on these datasets.

Results: In Figure 2.4, we plot the regret (averaged over 100 trials with randomly permuted patients) as a function of the number of patients seen so far, along with the 95% confidence intervals. First, in both datasets (a) and (b), we observe that Greedy Bandit and Greedy-First perform the best; Greedy-First recognizes that the greedy algorithm is converging and does not switch to an exploration-based strategy. In dataset (c), the Greedy Bandit gets “stuck” and does not converge to the optimal policy on average. Here, Greedy-First performs the best, followed closely by the OLS Bandit. This result is similar to our results in Fig 2.2 (c-d), but in this case, exploration appears to be necessary in nearly all instances, explaining the extremely close performance of Greedy-First and OLS Bandit. Finally, in dataset (d), we see that the Greedy Bandit performs the best, followed by Greedy-First. An interesting feature of this dataset is that one arm (high dose) is optimal for a very small number of patients; thus, dropping this arm entirely leads to better performance over a short horizon than attempting to learn its parameter. In this case, Greedy Bandit is not converging to the optimal policy since it never assigns any patient the high dose. However, Greedy-First recognizes that the high-dose arm is not getting sufficient samples and switches to an exploration-based algorithm. As a result, Greedy-First performs worse than the Greedy Bandit. However, if the horizon were to be extended[2], Greedy-First and

[2] Our horizon is limited by the number of patients available in the dataset.



(a) EEG dataset

(b) Eye Movement dataset

(c) Cardiotocography dataset

(d) Warfarin dataset

Figure 2.4: Expected regret of all algorithms on four real healthcare datasets

the other bandit algorithms would eventually overtake the Greedy Bandit. Alternatively, for non-binary reward functions (e.g., when cost of a mistake for high-dose patients is larger than for other patients) Greedy Bandit would perform poorly.

Looking at these results as a whole, we see that Greedy-First is a robust frontrunner. When exploration is unnecessary, it matches the performance of the Greedy Bandit; when exploration is necessary, it matches or outperforms competing bandit algorithms.

2.6 Conclusions and Discussions

In this chapter, we prove that a greedy algorithm can be rate optimal in cumulative regret for a two-armed contextual bandit as long as the contexts satisfy *covariate diversity.* Greedy algorithms are significantly preferable when exploration is costly (e.g., result in lost customers for online advertising or A/B testing) or unethical (e.g., personalized medicine or clinical trials). Furthermore, the greedy algorithm is entirely parameter-free, which makes it desirable in settings where tuning is difficult or where there is limited knowledge of problem parameters. Despite its simplicity, we provide empirical evidence that the greedy algorithm can outperform standard contextual bandit algorithms when the contexts satisfy covariate diversity. Even when the contexts do not satisfy covariate diversity, we prove that a greedy algorithm is rate optimal with *some probability*, and provide lower bounds on this probability.

However, in many scenarios, the decision-makers may not know whether their problem instance is amenable to a greedy approach, and may still wish to ensure that their algorithm provably converges to the correct policy. In this case, the decision-maker may under-explore by using a greedy algorithm, while a standard bandit algorithm may over-explore (since the greedy algorithm converges to the correct policy with some probability in general). Consequently, we propose the Greedy-First algorithm, which follows a greedy policy in the beginning and only performs exploration when the observed data indicate that exploration is necessary. Greedy-First is rate optimal without the covariate diversity assumption. More importantly, it remains exploration-free when covariate diversity is satisfied, and may provably reduce exploration even when covariate diversity is not satisfied. Our empirical results suggest that Greedy-First outperforms standard bandit algorithms (e.g., UCB, Thompson Sampling, and ϵ-greedy methods) by striking a balance between avoiding exploration and converging to the correct policy.

Chapter 3

Treatment Effect Estimation in Panel Models

3.1 Introduction

In this chapter we develop new methods for estimating average causal effects in settings with panel or longitudinal data, where a subset of units is exposed to a binary treatment during a subset of periods, and we observe the realized outcome for each unit in each time period. To estimate the (average) effect of the treatment on the treated units in this setting, we focus on imputing the missing potential outcomes. The statistics and econometrics literatures have taken two general approaches to this problem. The literature on unconfoundedness (Rosenbaum and Rubin (1983), Imbens and Rubin (2015)) imputes missing potential outcomes using observed outcomes for units with similar values for observed outcomes in previous periods. The synthetic control literature (Abadie and Gardeazabal (2003), Abadie et al. (2010, 2015), Doudchenko and Imbens (2016)) imputes missing control outcomes for treated units by finding weighted averages of control units that match the treated units in terms of lagged outcomes. Although at first sight similar, the two approaches are conceptually quite different in terms of the patterns in the data they exploit to impute the missing potential outcomes. The unconfoundedness approach estimates patterns over time that are assumed to be stable across units, and the synthetic control approach estimates patterns across units that are assumed to be stable over time. Both sets of methods also primarily focus on settings with different structures on the missing data or assignment mechanism. In the case of the unconfoundedness literature typically the assumption is that the treated units are all treated in the same periods, typically only the last period, and there are a substantial number of control units. The synthetic control literature has primarily focused on the case where one or a small number of treated units are observed prior to the treatment over a substantial number of periods.

In this study we also draw on the econometric literature on factor models and interactive fixed effects, and the computer science and statistics literatures on matrix completion, to take an approach to imputing the missing potential outcomes that is different from the unconfoundedness and synthetic control approaches. In the literature on factor models and interactive effects (Bai and Ng (2002), Bai (2003)) researchers model the observed outcome, in a balanced panel setting, as the sum of a linear function of covariates and an unobserved component that is a low rank matrix plus noise. Estimates are typically based on minimizing the sum of squared errors given the rank of the matrix of unobserved components, sometimes with the rank estimated. Xu (2017) applies this to causal settings where a subset of units is treated from common period onward, so that the complete data methods for estimating the factors and factor loadings can be used. The matrix completion literature (Candès and Recht (2009), Candès and Plan (2010), Mazumder et al. (2010)) focuses on imputing missing elements in a matrix assuming the complete matrix is the sum of a low rank matrix plus noise and the missingness is completely at random. The rank of the matrix is impliclty determined by the regularization through the addition of a penalty term to the objective function. Especially with complex missing data patterns using the nuclear norm as the regularizer is attractive for computational reasons.

We make two contributions in this chapter. First, we generalize the methods from the matrix completetion literature to settings where the missing data patterns are not completely at random. In particular we allow for the possibility of staggered adoption (Athey and Imbens (2018)), where units are treated from some initial adoption date onwards, but the adoption dates vary. Compared to the factor model literature the proposed estimator focuses on nuclear norm regularization to avoid the computational difficulties associated with imputation that would arise for complex missing data patterns with the fixed-rank methods in Bai and Ng (2002) and Xu (2017), similar to the way LASSO (ℓ_1 regularization, Tibshirani (1996)) is computationally attractive relative to subset selection (ℓ_0 regularization) in linear regression models. The second contribution is to show that the synthetic control and unconfoundedness approaches, as well as our proposed method, can all be viewed as matrix completion methods based on matrix factorization, all with the same objective function based on the Fröbenius norm for the difference between the latent matrix and the observed matrix. Given the objective function the unconfoundedness and synthetic control approaches impose different sets of restrictions on the factors in the matrix factorization, whereas the proposed method does not impose any restrictions but uses regularization to define the estimator.

3.2 Set Up

Consider an $N \times T$ matrix $\mathbf{Y}$ of outcomes with typical element Y_{it}. We only observe Y_{it} for some units and some time periods. We define $\mathcal{M}$ to be the set of pairs of indices (i,t), $i \in \{1,\ldots,N\}$, $t \in \{1,\ldots,T\}$, corresponding to the missing outcomes and $\mathcal{O}$ to be the observed outcomes: Y_{it} is missing if $(i,t) \in \mathcal{M}$ and observed if $(i,t) \in \mathcal{O}$. We wish to impute the missing Y_{it}. Our motivation for this problem arises from a causal potential outcome setting (e.g., Rubin (1974), Imbens and Rubin (2015)), where for each of N units and T time periods there exists a pair of potential outcomes, $Y_{it}(0)$ and $Y_{it}(1)$, with unit i exposed in period t to treatment $W_{it} \in \{0,1\}$, and the realized outcome equal to $Y_{it} = Y_{it}(W_{it})$. In that case the primary object of interest may be the average causal effect of the treatment, $\tau = \sum_{i,t}[Y_{it}(1) - Y_{it}(0)]/(NT)$, or some other average treatment effect. In order to estimate such average treatment effects, one approach is to impute the missing potential outcomes. In this chapter, we focus directly on the problem of imputing the missing entries in the $\mathbf{Y}(1)$ matrix for treated units with $W_{it} = 0$.

In addition to partially observing the matrix $\mathbf{Y}$, we may also observe covariate matrices $\mathbf{X} \in \mathbb{R}^{N\times P}$ and $\mathbf{Z} \in \mathbb{R}^{T\times Q}$ where columns of $\mathbf{X}$ are unit-specific covariates, and columns of $\mathbf{Z}$ are time-specific covariates. We may also observe unit/time specific covariates $V_{it} \in \mathbb{R}^J$.

Putting aside the covariates for the time being, the data can be thought of as consisting of two $N \times T$ matrices, one incomplete and one complete,

$$\mathbf{Y} = \begin{pmatrix} Y_{11} & Y_{12} & ? & \dots & Y_{1T} \\ ? & ? & Y_{23} & \dots & ? \\ Y_{31} & ? & Y_{33} & \dots & ? \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Y_{N1} & ? & Y_{N3} & \dots & ? \end{pmatrix}, \qquad \text{and} \;\; \mathbf{W} = \begin{pmatrix} 1 & 1 & 0 & \dots & 1 \\ 0 & 0 & 1 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 1 & \dots & 0 \end{pmatrix},$$

where

$$W_{it} = \begin{cases} 0 & \text{if } (i,t) \in \mathcal{M}, \\ 1 & \text{if } (i,t) \in \mathcal{O}, \end{cases}$$

is an indicator for Y_{it} being observed.

3.3 Panel Configurations

In this section, we discuss a number of particular configurations of the matrices $\mathbf{Y}$ and $\mathbf{W}$ that are the focus of parts of the general literature. This serves to put in context the problem, and to motivate previously developed methods from the literature on causal inference under unconfoundedness, the synthetic control literature, and the interactive fixed effect literature, and subsequently to develop formal connections between all three. First, we consider patterns of missing data. Second, we consider different shapes of the matrices $\mathbf{Y}$ and $\mathbf{W}$. Third, we consider a number of specific analyses that focus on particular combinations of missing data patterns and shapes of the matrices.

3.3.1 Patterns of Missing Data

In the statistics literature on matrix completion the focus is on settings with randomly missing values, allowing for general patterns on the matrix of missing data indicators (Candès and Tao (2010), Recht (2011)). In many social science applications, however, there is a specific structure on the missing data, in the form of restrictions on the values of $\mathbf{W}$.

Block Structure

A leading example is a block structure, with a subset of the units treated during every period from a particular point in time onwards.

$$\mathbf{Y}_{N\times T} = \begin{pmatrix} \checkmark & \checkmark & \checkmark & \checkmark & \dots & \checkmark \\ \checkmark & \checkmark & \checkmark & \checkmark & \dots & \checkmark \\ \checkmark & \checkmark & \checkmark & \checkmark & \dots & \checkmark \\ \checkmark & \checkmark & \checkmark & ? & \dots & ? \\ \checkmark & \checkmark & \checkmark & ? & \dots & ? \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \checkmark & \checkmark & \checkmark & ? & \dots & ? \end{pmatrix}.$$

There are two special cases of the block structure. Much of the literature on estimating average treatment effects under unconfoundedness focuses on the case where $T_0 = T$, so that the only treated units are in the last period. We will refer to this as the single-treated-period block structure. In contrast, the synthetic control literature focuses on the case of with a single treated unit which are treated for a number of periods from period T_0 onwards, the single-treated-unit block structure:

$$\mathbf{Y} = \begin{pmatrix} \checkmark & \checkmark & \checkmark & \dots & \checkmark & \checkmark \\ \checkmark & \checkmark & \checkmark & \dots & \checkmark & \checkmark \\ \checkmark & \checkmark & \checkmark & \dots & \checkmark & ? \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \checkmark & \checkmark & \checkmark & \dots & \checkmark & ? \\ & & & & & \uparrow \\ & & & \multicolumn{3}{r}{\text{treated period}} \end{pmatrix} \quad \text{and} \quad \mathbf{Y} = \begin{pmatrix} \checkmark & \checkmark & \checkmark & \dots & \checkmark & \\ \checkmark & \checkmark & \checkmark & \dots & \checkmark & \\ \checkmark & \checkmark & \checkmark & \dots & \checkmark & \\ \vdots & \vdots & \vdots & \ddots & \vdots & \\ \checkmark & \checkmark & \checkmark & \dots & \checkmark & \\ \checkmark & \checkmark & ? & \dots & ? & \leftarrow \text{treated unit} \end{pmatrix}.$$

Staggered Adoption

Another setting that has received attention is characterized by staggered adoption of the treatment (Athey and Imbens (2018)). Here units may differ in the time they are first exposed to the treatment, but once exposed they remain in the treatment group forever after. This naturally arises in settings where the treatment is some new technology that units can choose to adopt (e.g., Athey and Stern (2002)). Here:

$$\mathbf{Y}_{N\times T} = \begin{pmatrix} \checkmark & \checkmark & \checkmark & \checkmark & \dots & \checkmark & \text{(never adopter)} \\ \checkmark & \checkmark & \checkmark & \checkmark & \dots & ? & \text{(late adopter)} \\ \checkmark & \checkmark & ? & ? & \dots & ? & \\ \checkmark & \checkmark & ? & ? & \dots & ? & \text{(medium adopter)} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \\ \checkmark & ? & ? & ? & \dots & ? & \text{(early adopter)} \end{pmatrix}.$$

3.3.2 Thin and Fat Matrices

A second classification concerns the shape of the matrix $\mathbf{Y}$. Relative to the number of time periods, we may have many units, few units, or a comparable number. These data configurations may make particular analyses more attractive. For example, $\mathbf{Y}$ may be a thin matrix, with $N \gg T$, or a fat matrix, with $N \ll T$, or an approximately square matrix, with $N \approx T$:

$$\mathbf{Y} = \begin{pmatrix} ? & \checkmark & ? \\ \checkmark & ? & \checkmark \\ ? & ? & \checkmark \\ \checkmark & ? & \checkmark \\ ? & ? & ? \\ \vdots & \vdots & \vdots \\ ? & ? & \checkmark \end{pmatrix} \quad (\textbf{thin}) \qquad \mathbf{Y} = \begin{pmatrix} ? & ? & \checkmark & \checkmark & \checkmark & \dots & ? \\ \checkmark & \checkmark & \checkmark & \checkmark & ? & \dots & \checkmark \\ ? & \checkmark & ? & \checkmark & ? & \dots & \checkmark \end{pmatrix} \quad (\textbf{fat}),$$

or

$$\mathbf{Y} = \begin{pmatrix} ? & ? & \checkmark & \checkmark & \dots & ? \\ \checkmark & \checkmark & \checkmark & \checkmark & \dots & \checkmark \\ ? & \checkmark & ? & \checkmark & \dots & \checkmark \\ \checkmark & \checkmark & ? & \checkmark & \dots & \checkmark \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ ? & ? & \checkmark & \checkmark & \dots & \checkmark \end{pmatrix} \quad (\textbf{approximately square}).$$

3.3.3 Horizontal and Vertical Regressions

Two special combinations of missing data patterns and the shape of the matrices deserve particular attention because they are the focus of substantial separate literatures.

Horizontal Regression and the Unconfoundedness Literature

The unconfoundedness literature focuses primarily on the single-treated-period block structure with a thin matrix, and imputes the missing potential outcomes in the last period using control units with similar lagged outcomes. A simple version of that approach is to regress the last period outcome on the lagged outcomes and use the estimated regression to predict the missing potential outcomes. That is, for the units with $(i, T) \in \mathcal{M}$, the predicted outcome is

$$\hat{Y}_{iT} = \hat{\beta}_0 + \sum_{s=1}^{T-1} \hat{\beta}_s Y_{is}, \quad \text{where } \hat{\beta} = \arg\min_{\beta} \sum_{i:(i,T)\in\mathcal{O}} \left(Y_{iT} - \beta_0 - \sum_{s=1}^{T-1} \beta_s Y_{is} \right)^2 . \qquad (3.3.1)$$

We refer to this as a **horizontal** regression, where the rows of the $\mathbf{Y}$ matrix form the units of observation. A more flexible, non-parametric, version of this estimator would correspond to matching where we find for each treated unit i a corresponding control unit j with Y_{jt} approximately equal to Y_{it} for all pre-treatment periods $t = 1, \ldots, T-1$.

Vertical Regression and the Synthetic Control Literature

The synthetic control literature focuses primarily on the single-treated-unit block structure with a fat or approximately square matrix. Doudchenko and Imbens (2016) discuss how the Abadie-Diamond-Hainmueller synthetic control method can be interpreted as regressing the outcomes for the treated unit prior to the treatment on the outcomes for the control units in the same periods. That is, for the treated unit in period t, for $t = T_0, \ldots, T$, the predicted outcome is

$$\hat{Y}_{Nt} = \hat{\gamma}_0 + \sum_{i=1}^{N-1} \hat{\gamma}_i Y_{it}, \quad \text{where} \quad \hat{\gamma} = \arg\min_{\gamma} \sum_{t:(N,t)\in\mathcal{O}} \left(Y_{Nt} - \gamma_0 - \sum_{i=1}^{N-1} \gamma_i Y_{it} \right)^2 . \qquad (3.3.2)$$

We refer to this as a **vertical** regression, where the columns of the $\mathbf{Y}$ matrix form the units of observation. As shown in Doudchenko and Imbens (2016) this is a special case of the Abadie et al. (2015) estimator, without imposing their restrictions that the coefficients are nonnegative and that the intercept is zero.

Although this does not appear to have been pointed out previously, a matching version of this estimator would correspond to finding, for each period t where unit N is treated, a corresponding period $s \in \{1, \ldots, T_0 - 1\}$ such that Y_{is} is approximately equal to Y_{Ns} for all control units $i = 1, \ldots, N-1$. This matching version of the synthetic control estimator clarifies the link between the treatment effect literature under unconfoundedness and the synthetic control literature.

Suppose that there is only a single treated unit/time period combination, i.e. $\mathcal{M} = \{(N,T)\}$. In that case if we estimate the horizontal regression in (3.3.1), it is still the case that $\hat{Y}_{NT}$ is linear in $Y_{1T}, \ldots, Y_{N-1,T}$, just with different weights than those obtained from the vertical regression in (3.3.2). Similarly, if we estimate the vertical regression in (3.3.2), it is still the case that $\hat{Y}_{NT}$ is linear in $Y_{N1}, \ldots, Y_{N,T-1}$, just with different weights from the horizontal regression.

3.3.4 Fixed Effects and Factor Models

The horizontal regression focuses on a pattern in the time path of the outcome Y_{it}, specifically the relation between Y_{iT} and the lagged Y_{it} for $t = 1, \ldots, T-1$, and assumes that is stable across units. The vertical regression focuses on a pattern across units that is stable over time. However, by focusing on only one of these patterns these approaches ignore alternative patterns that may help in imputing the missing values. An alternative is to consider approaches that allow for the exploitation of both stable patterns over time, and stable patterns accross units. Such methods have a long history in the panel data literature, including the literature on fixed effects, and more generally, factor and interactive fixed effect models (e.g., Chamberlain (1984), Arellano and Honoré (2001), Liang and Zeger (1986), Bai (2003, 2009), Pesaran (2006), Moon and Weidner (2015, 2017)). In the absence of covariates (although in this literature the coefficients on these covariates are typically the primary focus of the analyses), such models can be written as

$$Y_{it} = \sum_{r=1}^{R} \gamma_{ir}\delta_{tr} + \varepsilon_{it}, \qquad \text{or} \quad \mathbf{Y} = \mathbf{U}\mathbf{V}^{\top} + \boldsymbol{\varepsilon}, \tag{3.3.3}$$

where $\mathbf{U}$ is $N \times R$ and $\mathbf{V}$ is $T \times R$. Most of the early literature, Anderson (1958) and Goldberger (1972)), focused on the thin matrix case, with $N \gg T$, where asymptotic approximations are based on letting the number of units increase with the number of time periods fixed. In the modern part of this literature researchers allow for more complex asymptotics with both N and T increasing, at rates that allow for consistent estimation of the factors $\mathbf{V}$ and loading s$\mathbf{V}$ after imposing normalizations. In this literature it is typically assumed that the number of factors R is fixed, although not necessarily known. Methods for estimating the rank R are discussed in Bai and Ng (2002) and Moon and Weidner (2015).

Xu (2017) implements this interactive fixed effect approach to the matrix completion problem in the special case with blocked assignment, with additional applications in Gobillon and Magnac (2013), Kim and Oka (2014) and Hsiao et al. (2012). Suppose the first N_C units are in the control group, and the last $N_T = N - N_C$ units are in the treatment group. The treatment group is exposed to the control treatment in the first $T_0 - 1$ pre-treatment periods, and exposed to the active treatment in the post-treatment periods $T_0, \ldots, T$. In that case we can partition $\mathbf{U}$ and $\mathbf{V}$ accordingly and write

$$\mathbf{U}\mathbf{V}^\top = \begin{pmatrix} \mathbf{U}_C \\ \mathbf{U}_T \end{pmatrix} \begin{pmatrix} \mathbf{V}_{\text{pre}} \\ \mathbf{V}_{\text{post}} \end{pmatrix}^\top .$$

Using the data from the control group pre and post, and the pre data only for the treatment group, we have

$$\mathbf{Y}_C = \mathbf{U}_C \begin{pmatrix} \mathbf{V}_{\text{pre}} \\ \mathbf{V}_{\text{post}} \end{pmatrix}^\top + \varepsilon_C, \qquad \text{and} \;\; \mathbf{Y}_{T,\text{pre}} = \mathbf{U}_T \mathbf{V}_{\text{pre}}^\top + \varepsilon_{T,\text{pre}}$$

where the first equation can be used to estimate $\mathbf{U}_C$, $\mathbf{V}_{\text{pre}}$, and $\mathbf{V}_{\text{post}}$, and the second is used to estimate $\mathbf{U}_T$, both by least squares after normalizing $\mathbf{U}$ and $\mathbf{V}$. Note that this is not necessarily efficient, because $\mathbf{Y}_{T,\text{pre}}$ is not used to estimate $\mathbf{V}_{\text{pre}}$.

Independently, a closely related literature has emerged in machine learning and statistics on matrix completion (Srebro et al. (2005), Candès and Recht (2009), Candès and Tao (2010), Keshavan et al. (2010a,b), Gross (2011), Recht (2011), Rohde et al. (2011), Negahban and Wainwright (2011, 2012), Koltchinskii et al. (2011), Klopp (2014)). In this literature the starting point is an incompletely observed matrix, and researchers have proposed matrix-factorization approaches to matrix completion, similar to (3.3.3). The focus is not on estimating $\mathbf{U}$ and $\mathbf{V}$ consistently, only on imputing the missing elements of $\mathbf{Y}$. Instead of fixing the rank of the underlying matrix, estimators rely on regularization, and in particular nuclear norm regularization.

3.4 The Nuclear Norm Matrix Completion Estimator

In the absence of covariates we model the matrix of outcomes $\mathbf{Y}$ as

$$\mathbf{Y} = \mathbf{L}^* + \boldsymbol{\varepsilon}, \qquad \text{where} \quad \mathbb{E}[\boldsymbol{\varepsilon}|\mathbf{L}^*] = \mathbf{0}\,. \tag{3.4.1}$$

The ε_{it} can be thought of as measurement error. The goal is to estimate the $N \times T$ matrix $\mathbf{L}^*$.

To facilitate the characterization of the estimator, define for any matrix $\mathbf{A}$, and given a set of pairs of indices $\mathcal{O}$, the two matrices $\mathbf{P}_{\mathcal{O}}(\mathbf{A})$ and $\mathbf{P}_{\mathcal{O}}^{\perp}(\mathbf{A})$ with typical elements:

$$\mathbf{P}_{\mathcal{O}}(\mathbf{A})_{it} = \begin{cases} A_{it} & \text{if } (i,t) \in \mathcal{O}\,, \\ 0 & \text{if } (i,t) \notin \mathcal{O}\,, \end{cases} \quad \text{and} \quad \mathbf{P}_{\mathcal{O}}^{\perp}(\mathbf{A})_{it} = \begin{cases} 0 & \text{if } (i,t) \in \mathcal{O}\,, \\ A_{it} & \text{if } (i,t) \notin \mathcal{O}\,. \end{cases}$$

A critical role is played by various matrix norms, summarized in Table 3.1. Some of these depend on the singular values, where, given the Singular Value Decomposition (SVD) $\mathbf{L}_{N\times T} = \mathbf{S}_{N\times N}\mathbf{\Sigma}_{N\times T}\mathbf{R}_{T\times T}^{\top}$, the singular values $\sigma_i(\mathbf{L})$ are the ordered diagonal elements of Σ.

Table 3.1: Matrix Norms for Matrix $\mathbf{L}$

Schatten Norm	$\|\mathbf{L}\|_p$	$\left(\sum_i \sigma_i(\mathbf{L})^p\right)^{1/p}$
Fröbenius Norm	$\|\mathbf{L}\|_F$	$\left(\sum_i \sigma_i(\mathbf{L})^2\right)^{1/2} = \left(\sum_{i=1}^N \sum_{t=1}^T L_{it}^2\right)^{1/2}$
Rank Norm	$\|\mathbf{L}\|_0$	$\sum_i \mathbf{1}_{\sigma_i(\mathbf{L})>0}$
Nuclear Norm	$\|\mathbf{L}\|_*$	$\sum_i \sigma_i(\mathbf{L})$
Operator Norm	$\|\mathbf{L}\|_{\text{op}}$	$\max_i \sigma_i(\mathbf{L}) = \sigma_1(\mathbf{L})$
Max Norm	$\|\mathbf{L}\|_{\max}$	$\max_{1\le i\le N, 1\le t\le T} \lvert L_{it}\rvert$

Now consider the problem of estimating $\mathbf{L}$. Directly minimizing the sum of squared differences,

$$\min_{\mathbf{L}} \sum_{(i,t)\in\mathcal{O}} (Y_{it} - L_{it})^2 = \min_{\mathbf{L}} \|\mathbf{P}_{\mathcal{O}}(\mathbf{Y} - \mathbf{L})\|_F^2\,, \tag{3.4.2}$$

does not lead to a useful estimator: if $(i,t) \in \mathcal{M}$ the objective function does not depend on L_{it}, and for other pairs (i,t) the estimator would simply be Y_{it}. Instead, we regularize the problem by adding a penalty term $\lambda\|\mathbf{L}\|$, for some choice of the norm $\|\cdot\|$.

The estimator: The general form of our proposed estimator for $\mathbf{L}^*$ is (Mazumder et al. (2010))

$$\hat{\mathbf{L}} = \arg\min_{\mathbf{L}} \left\{ \|\mathbf{P}_{\mathcal{O}}(\mathbf{Y} - \mathbf{L})\|_F^2 + \lambda\|\mathbf{L}\|_* \right\}\,, \tag{3.4.3}$$

with the penalty factor λ chosen through cross-validation that will be described at the end of this section. We will call this the Matrix-Completion with Nuclear Norm Minimization (MC-NNM) estimator. Some Schatten norms would not work as well. For example, the Fröbenius norm on the penalty term would not have been suitable for estimating $\mathbf{L}^*$ in the case with missing entries because the solution for L_{it} for $(i,t) \in \mathcal{M}$ is always zero (which follows directly from the representation of $\|\mathbf{L}\|_F = \sum_{i,t} L_{it}^2$). The rank norm is not computationally feasible for large N and T if the cardinality of the set $\mathcal{M}$ is substantial. Formally, the problem is NP-hard. In contrast, a major advantage of using the nuclear norm is that the resulting estimator can be computed using fast convex optimization programs, e.g. the SOFT-IMPUTE algorithm by Mazumder et al. (2010) that will be described next.

Calculating the Estimator: The algorithm for calculating our estimator (in the case without additional covariates) goes as follows. Given the SVD for $\mathbf{A}$, $\mathbf{A} = \mathbf{S}\mathbf{\Sigma}\mathbf{R}^\top$, with singular values $\sigma_1(\mathbf{A}), \ldots, \sigma_{\min(N,T)}(\mathbf{A})$, define the matrix shrinkage operator

$$\text{shrink}_\lambda(\mathbf{A}) = \mathbf{S}\tilde{\mathbf{\Sigma}}\mathbf{R}^\top \,, \tag{3.4.4}$$

where $\tilde{\mathbf{\Sigma}}$ is equal to $\mathbf{\Sigma}$ with the i-th singular value $\sigma_i(\mathbf{A})$ replaced by $\max(\sigma_i(\mathbf{A}) - \lambda, 0)$. Now start with the initial choice $\mathbf{L}_1(\lambda, \mathcal{O}) = \mathbf{P}_\mathcal{O}(\mathbf{Y})$. Then for $k = 1, 2, \ldots$, define,

$$\mathbf{L}_{k+1}(\lambda, \mathcal{O}) = \text{shrink}_\lambda \Big\{ \mathbf{P}_\mathcal{O}(\mathbf{Y}) + \mathbf{P}_\mathcal{O}^\perp\Big(\mathbf{L}_k(\lambda)\Big) \Big\} \,, \tag{3.4.5}$$

until the sequence $\{\mathbf{L}_k(\lambda)\}_{k\geq 1}$ converges. The limiting matrix $\hat{\mathbf{L}}(\lambda, \mathcal{O}) = \lim_{k\to\infty} \mathbf{L}_k(\lambda)$ is our estimator given the regularization parameter λ.

Cross-validation: The optimal value of λ is selected through cross-validation. We choose K (e.g., $K = 5$) random subsets $\mathcal{O}_k \subset \mathcal{O}$ with cardinality $\lfloor |\mathcal{O}|^2/NT \rfloor$ to ensure that the fraction of observed data in the cross-validation data sets, $|\mathcal{O}_k/|\mathcal{O}|$, is equal to that in the original sample, $|\mathcal{O}|/(NT)$. We then select a sequence of candidate regularization parameters

$$\lambda_1 > \cdots > \lambda_L = 0 \,.$$

with a large enough λ_1, and for each subset $\mathcal{O}_k$ calculate

$$\hat{\mathbf{L}}(\lambda_1, \mathcal{O}_k), \ldots, \hat{\mathbf{L}}(\lambda_L, \mathcal{O}_k)$$

and evaluate the average squared error on $\mathcal{O} \setminus \mathcal{O}_k$. The value of λ that minimizes the average squared error (among the K produced estimators corresponding to that λ) is the one choosen.

It is worth noting that one can expedite the computation by using $\hat{\mathbf{L}}(\lambda_i, \mathcal{O}_k)$ as a warm-start initialization for calculating $\hat{\mathbf{L}}(\lambda_{i+1}, \mathcal{O}_k)$ for each i and k.

3.5 Theoretical Bounds for the Estimation Error

In this section we focus on the case that there are no covariates and provide theoretical results for the estimation error. Let $L_{\max}$ be a positive constant such that $\|\mathbf{L}^*\|_{\max} \leq L_{\max}$ (recall that $\|\mathbf{L}^*\|_{\max} = \max_{i,t} |\mathbf{L}^*_{it}|$). We also assume that $\mathbf{L}^*$ is a deterministic matrix. Then consider the following estimator for $\mathbf{L}^*$ that is motivated by the low-rank assumption on $\mathbf{L}^*$.

$$\hat{\mathbf{L}} = \underset{\mathbf{L}:\|\mathbf{L}\|_{\max} \leq L_{\max}}{\arg\min} \left\{ \|\mathbf{P}_{\mathcal{O}}(\mathbf{Y} - \mathbf{L})\|_F^2 + \lambda \|\mathbf{L}\|_* \right\} . \tag{3.5.1}$$

3.5.1 Additional Notation

First, we start by introduction some new notation. For each positive integer n let $[n]$ be the set of integers $\{1, 2, \ldots, n\}$. In addition, for any pair of integers i, n with $i \in [n]$ define $e_i(n)$ to be the n dimensional column vector with all of its entries equal to 0 except the i^{th} entry that is equal to 1. In other words, $\{e_1(n), e_2(n), \ldots, e_n(n)\}$ forms the standard basis for $\mathbb{R}^n$. For any two matrices $\mathbf{A}, \mathbf{B}$ of the same dimensions define the inner product

$$\langle \mathbf{A}, \mathbf{B} \rangle \equiv \operatorname{trace}(\mathbf{A}^\top \mathbf{B}) .$$

Note that with this definition, $\langle \mathbf{A}, \mathbf{A} \rangle = \|\mathbf{A}\|_F^2$.

Next, we describe a random observation process that defines the set $\mathcal{O}$. Consider N independent random variables $t_1, \ldots, t_N$ on $[T]$ with distributions $\pi^{(i)}$. Specifically, for each $(i, t) \in [N] \times [T]$, define $\pi_t^{(i)} \equiv \mathbb{P}[t_i = t]$. We also use the short notation $\mathbb{E}_\pi$ when taking expectation with respect to all distributions $\pi^{(1)}, \ldots, \pi^{(N)}$. Now, $\mathcal{O}$ can be written as

$$\mathcal{O} = \bigcup_{i=1}^{N} \Big\{ (i, 1), (i, 2), \ldots, (i, t_i) \Big\} .$$

Also, for each $(i, t) \in \mathcal{O}$, we use the notation $\mathbf{A}_{it}$ to refer to $e_i(N) e_t(T)^\top$ which is a N by T matrix with all entries equal to zero except the (i, t) entry that is equal to 1. The data generating model can now be written as

$$Y_{it} = \langle \mathbf{A}_{it}, \mathbf{L}^* \rangle + \varepsilon_{it} \,, \quad \forall \; (i,t) \in \mathcal{O} \,,$$

where noise variables ε_{it} are independent σ-sub-Gaussian random variables that are also independent of $\mathbf{A}_{it}$. Recall that a random variable ε is σ-sub-Gaussian if for all real numbers t we have $\mathbb{E}[\exp(t\varepsilon)] \leq \exp(\sigma^2 t^2/2)$.

Note that the number of control units (N_{c}) is equal to the number of rows that have all entries observed (i.e., $N_{\mathrm{c}} = \sum_{i=1}^{N} \mathbb{I}_{t_i=T}$). Therefore, the expected number of control units can be written as $\mathbb{E}_\pi[N_{\mathrm{c}}] = \sum_{i=1}^{N} \pi_T^{(i)}$. Defining

$$p_{\mathrm{c}} \equiv \min_{1 \leq i \leq N} \pi_T^{(i)} \,,$$

we expect to have (on average) at least Np_{c} control units. The parameter p_{c} will play an important role in our main theoretical results. In particular, assuming N and T are of the same order, we will show that the average per entry error (i.e., $\|\hat{\mathbf{L}} - \mathbf{L}^*\|_F/\sqrt{NT}$) converges to 0 if p_{c} grows larger than $\log^{3/2}(N)/\sqrt{N}$ up to a constant. To provide some intuition for such assumption on p_{c}, assume $\mathbf{L}^*$ is a matrix that is zero everywhere except in its i^{th} row. Such $\mathbf{L}^*$ is clearly low-rank. But recovering the entry L_{iT}^* is impossible when $i_t < T$. Therefore, $\pi_T^{(i)}$ cannot be too small. Since i is arbitrary, in general p_{c} cannot be too small.

Remark 3.5.1. *It is worth noting that the sources of randomness in our observation process $\mathcal{O}$ are the random variables $\{t_i\}_{i=1}^N$ that are assumed to be independent of each other. But we allow that distributions of these random variables to be functions of $\mathbf{L}^*$. We also assume that the noise variables $\{\varepsilon_{it}\}_{it \in [N]\times[T]}$ are independent of each other and are independent of $\{t_i\}_{i=1}^N$. In §3.9 we discuss how our results could generalize to the cases with correlations among these noise variables.*

Remark 3.5.2. *The estimator* (3.5.1) *penalizes the error terms $(Y_{it} - L_{it})^2$, for $(i,t) \in \mathcal{O}$, equally. But probability of missing entries in each row decreases as t increases. In §3.9.2, we discuss how the estimator can be modified by considering a weighted loss function based on propensity scores for the missing entries.*

3.5.2 Main Result

The main result of this section is the following theorem (proved in §B.1.1) that provides an upper bound for $\|\mathbf{L}^* - \hat{\mathbf{L}}\|_F/\sqrt{NT}$, the root-mean-squared-error (RMSE) of the estimator $\hat{\mathbf{L}}$. In literature on theoretical analysis of empirical risk minimization this type of upper bound is called an *oracle inequality.* The proof is provided in Appendix B.1.

Theorem 5. *If rank of $\mathbf{L}^*$ is R, then there is a constant C such that with probability greater than $1 - 2(N+T)^{-2}$,*

$$\frac{\|\mathbf{L}^* - \hat{\mathbf{L}}\|_F}{\sqrt{NT}} \leq C \max \left[L_{\max}\sqrt{\frac{\log(N+T)}{N\,p_c^2}}, \sigma\sqrt{\frac{R\,\log(N+T)}{T\,p_c^2}}, \sigma\sqrt{\frac{R\,\log^3(N+T)}{N\,p_c^2}} \right], \tag{3.5.2}$$

when the parameter λ is a constant multiple of

$$\frac{\sigma \max\left[\sqrt{N\log(N+T)}, \sqrt{T}\log^{3/2}(N+T)\right]}{|\mathcal{O}|}.$$

Interpretation of Theorem 5: Since our goal is to show that the RMSE of $\hat{\mathbf{L}}$ converges to zero as N and T grow, it is important to see when the right hand side of (3.5.2) converges to zero as N and T grow. One such situation is when $\mathbf{L}^*$ is low-rank (R is constant) and $p_{\mathrm{c}} \gg \log^{3/2}(N+T)/\sqrt{\min(N,T)}$. A sufficient condition for the latter, when N and T are of the same order, is that the lower bound for the average number of control units (Np_{c}) grows larger than a constant times $\sqrt{N}\log^{3/2}(N)$. In §3.9 we will discuss how the estimator $\hat{\mathbf{L}}$ should be modified to obtain a sharper result that would hold for a smaller number of control units.

Comparison with existing theory on matrix-completion: Our estimator and its theoretical analysis are motivated by and generalize existing research on matrix-completion in machine learning and statistics literature Srebro et al. (2005), Mazumder et al. (2010), Candès and Recht (2009), Candès and Tao (2010), Keshavan et al. (2010a,b), Gross (2011), Recht (2011), Rohde et al. (2011), Negahban and Wainwright (2011, 2012), Koltchinskii et al. (2011), Klopp (2014). The main difference is in our observation model $\mathcal{O}$. Existing papers assume that entries $(i,t) \in \mathcal{O}$ are independent random variables whereas we allow for a dependency structure including staggered adoption where if $(i,t) \in \mathcal{O}$ then $(i,t') \in \mathcal{O}$

for all $t' < t$.

3.6 The Relationship with Horizontal and Vertical Regressions

In the second contribution we discuss the relation between the matrix completion estimator and the horizontal (unconfoundedness) and vertical (synthetic control) approaches. To faciliate the discussion, we focus on the case with $\mathcal{M}$ containing a single pair, unit N in period T, so that $\mathcal{M}$ contains a single element, $\mathcal{M} = \{(N,T)\}$. In that case the various previously proposed versions of the vertical and horizontal regressions are directly applicable.

The observed data are $\mathbf{Y}$, an $N \times T$ matrix that can be partitioned as

$$\mathbf{Y} = \begin{pmatrix} \tilde{\mathbf{Y}} & \mathbf{y}_1 \\ \mathbf{y}_2^\top & ? \end{pmatrix},$$

where $\tilde{\mathbf{Y}}$ is $(N-1) \times (T-1)$, $\mathbf{y}_1$ is $(N-1) \times 1$, and $\mathbf{y}_2$ is $(T-1) \times 1$.

The matrix completion solution to imputing $Y_{N,T}$ can be characterized, for a given regularization parameter λ, as

$$\mathbf{L}^{\mathrm{mc-nnm}}(\lambda) = \arg\min_{\mathbf{L}} \left\{ \|P_{\mathcal{O}}\left(\mathbf{Y} - \mathbf{L}\right)\|_F^2 + \lambda \|\mathbf{L}\|_* \right\}. \tag{3.6.1}$$

The predicted value for the missing entry Y_{NT} is then

$$\hat{Y}_{N,T}^{\mathrm{mc-nnm}} = \mathbf{L}_{N,T}^{\mathrm{mc-nnm}}(\lambda). \tag{3.6.2}$$

We are interested in comparing this estimator to horizontal regression estimator. Let us initially assume that the horizontal regression is well defined, without regularization, so that $N > T$. First define

$$\hat{\beta}^{\mathrm{hr}} = \left(\tilde{\mathbf{Y}}^\top \tilde{\mathbf{Y}}\right)^{-1} \left(\tilde{\mathbf{Y}}^\top \mathbf{y}_1\right).$$

Then the horizontal regression based prediction is

$$\hat{Y}_{NT}^{\mathrm{hr}} = \mathbf{y}_2^\top \hat{\beta}^{\mathrm{hr}} = \mathbf{y}_2^\top \left(\tilde{\mathbf{Y}}^\top \tilde{\mathbf{Y}}\right)^{-1} \left(\tilde{\mathbf{Y}}^\top \mathbf{y}_1\right).$$

For the vertical (synthetic control) regression, initially assuming $T > N$, we start with

$$\hat{\gamma}^{\mathrm{vt}} = \left(\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^\top\right)^{-1}\left(\tilde{\mathbf{Y}}\mathbf{y}_2\right),$$

leading to the horizontal regression based prediction

$$\hat{Y}_{NT}^{\mathrm{vt}} = \mathbf{y}_1^\top \hat{\gamma}^{\mathrm{hr}} = \mathbf{y}_1^\top \left(\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^\top\right)^{-1}\left(\tilde{\mathbf{Y}}\mathbf{y}_2\right).$$

The original (Abadie et al. (2010)) synthetic control estimator imposes the additional restrictions $\gamma_i \geq 0$, and $\sum_{i=1}^{N-1} \gamma_i = 1$, leading to

$$\hat{\gamma}^{\mathrm{sc-adh}} = \arg\min_{\gamma} \left\| \mathbf{y}_2 - \tilde{\mathbf{Y}}\gamma \right\|_F^2, \qquad \text{subject to } \forall i \ \gamma_i \geq 0, \ \sum_{i=1}^{N-1} \gamma_i = 1.$$

Then the synthetic control based prediction is

$$\hat{Y}_{NT}^{\mathrm{sc-adh}} = \mathbf{y}_1^\top \hat{\gamma}^{\mathrm{sc-adh}}.$$

The Doudchenko and Imbens (2016) modification allows for the possibility that $N \geq T$ and regularizes the estimator for γ. Focusing here on an elastic net regularization, their proposed estimator is

$$\hat{\gamma}^{\mathrm{vt-en}} = \arg\min_{\gamma} \left\{ \left\| \mathbf{y}_2 - \tilde{\mathbf{Y}}\gamma \right\|_F^2 + \lambda \left(\alpha \left\|\gamma\right\|_1 + \frac{1-\alpha}{2} \left\|\gamma\right\|_F^2 \right) \right\}.$$

Then the elastic net / synthetic control based prediction is

$$\hat{Y}_{NT}^{\mathrm{vt-en}} = \mathbf{y}_1^\top \hat{\gamma}^{\mathrm{vt-en}}.$$

We can modify the horizontal regression in the same way to allow for restrictions on the β, and regularization, although such methods have not been used in practice.

The question in this section concerns the relation between the various predictors, $\hat{Y}_{NT}^{\mathrm{mc-nnm}}$, $\hat{Y}_{NT}^{\mathrm{hr}}$, $\hat{Y}_{NT}^{\mathrm{vt}}$, $\hat{Y}_{NT}^{\mathrm{sc-adh}}$, and $\hat{Y}_{NT}^{\mathrm{vt-en}}$. The first result states that all these estimators can be viewed as particular cases of matrix factorization estimators, with the difference coming in the way the estimation of the components of the matrix factorization is carried out.

Theorem 6. *All five estimators $\hat{Y}_{NT}^{\mathrm{mc-nnm}}$, $\hat{Y}_{NT}^{\mathrm{hr}}$, $\hat{Y}_{NT}^{\mathrm{vt}}$, $\hat{Y}_{NT}^{\mathrm{sc-adh}}$, and $\hat{Y}_{NT}^{\mathrm{vt-en}}$, can be written* *in the form* $\hat{Y}_{NT}^{\text{est}} = \hat{L}_{NT}^{\text{est}}$, *for* $\text{est} \in \{\text{mc} - \text{nnm}, \text{hr}, \text{vt}, \text{sc} - \text{adh}, \text{vt} - \text{en}\}$, *where*

$$\hat{\mathbf{L}}^{\text{est}} = \mathbf{A}^{\text{est}}\mathbf{B}^{\text{est}\top},$$

with $\mathbf{L}$, $\mathbf{A}$, *and* $\mathbf{B}$ $N \times T$, $N \times R$ *and* $T \times R$ *dimensional matrices, and* $\mathbf{A}$ *and* $\mathbf{B}$ *estimated as*

$$\left(\mathbf{A}^{\text{est}}, \mathbf{B}^{\text{est}}\right) = \arg\min_{\mathbf{A},\mathbf{B}} \left\{ \left\| P_{\mathcal{O}}\left(\mathbf{Y} - \mathbf{A}\mathbf{B}^\top\right) \right\|_F^2 + \text{penalty terms on } (\mathbf{A}, \mathbf{B}) \right\},$$

subject to restrictions on $\mathbf{A}$ *and* $\mathbf{B}$, *with the penalty terms and the restrictions specific to the estimator.*

Theorem 6 follows from the following result.

Theorem 7. *We have,*
(i) *(nuclear norm matrix completion)*

$$(\mathbf{A}_\lambda^{\text{mc}-\text{nnm}}, \mathbf{B}_\lambda^{\text{mc}-\text{nnm}}) = \arg\min_{\mathbf{A},\mathbf{B}} \left\| P_{\mathcal{O}}\left(\mathbf{Y} - \mathbf{A}\mathbf{B}^\top\right) \right\|_F^2 + \lambda\|\mathbf{A}\|_F^2 + \lambda\|\mathbf{B}\|_F^2,$$

(ii) *(horizontal regression, defined if* $N > T$*),* $R = T - 1$

$$(\mathbf{A}^{\text{hr}}, \mathbf{B}^{\text{hr}}) = \lim_{\lambda\downarrow 0} \arg\min_{\mathbf{A},\mathbf{B}} \left\{ \left\| P_{\mathcal{O}}\left(\mathbf{Y} - \mathbf{A}\mathbf{B}^\top\right) \right\|_F^2 + \lambda\|\mathbf{A}\|_F^2 + \lambda\|\mathbf{B}\|_F^2 \right\},$$

$$\text{subject to } \mathbf{A}^{\text{hr}} = \begin{pmatrix} \tilde{\mathbf{Y}} \\ \mathbf{y}_2^\top \end{pmatrix},$$

(iii) *(vertical regression, defined if* $T > N$*),* $R = N - 1$

$$(\mathbf{A}^{\text{vt}}, \mathbf{B}^{\text{vt}}) = \lim_{\lambda\downarrow 0} \arg\min_{\mathbf{A},\mathbf{B}} \left\{ \left\| P_{\mathcal{O}}\left(\mathbf{Y} - \mathbf{A}\mathbf{B}^\top\right) \right\|_F^2 + \lambda\|\mathbf{A}\|_F^2 + \lambda\|\mathbf{B}\|_F^2 \right\},$$

subject to

$$\mathbf{B}^{\text{vt}} = \begin{pmatrix} \tilde{\mathbf{Y}}^\top \\ \mathbf{y}_1^\top \end{pmatrix}.$$

(iv) *(synthetic control),* $R = N - 1$

$$(\mathbf{A}^{\text{sc}-\text{adh}}, \mathbf{B}^{\text{sc}-\text{adh}}) = \lim_{\lambda\downarrow 0} \arg\min_{\mathbf{A},\mathbf{B}} \left\{ \left\| P_{\mathcal{O}}\left(\mathbf{Y} - \mathbf{A}\mathbf{B}^\top\right) \right\|_F^2 + \lambda\|\mathbf{A}\|_F^2 + \lambda\|\mathbf{B}\|_F^2 \right\},$$

subject to

$$\mathbf{B}^{\text{sc}-\text{adh}} = \begin{pmatrix} \tilde{\mathbf{Y}}^\top \\ \mathbf{y}_1^\top \end{pmatrix}, \qquad \forall\ i, A_{iT} \geq 0,\ \sum_{i=1}^{N-1} A_{iT} = 1,$$

(v) (elastic net), $R = N - 1$

$$(\mathbf{A}^{\text{vt}-\text{en}}, \mathbf{B}^{\text{vt}-\text{en}}) = \lim_{\lambda \downarrow 0} \arg\min_{\mathbf{A},\mathbf{B}} \left\{ \left\| P_{\mathcal{O}} \left(\mathbf{Y} - \mathbf{A}\mathbf{B}^\top \right) \right\|_F^2 + \lambda \left[\frac{1-\alpha}{2} \left\| \begin{pmatrix} \mathbf{a}_2 \\ \mathbf{a}_3 \end{pmatrix} \right\|_F^2 + \alpha \left\| \begin{pmatrix} \mathbf{a}_2 \\ \mathbf{a}_3 \end{pmatrix} \right\|_1 \right] \right\},$$

subject to

$$\mathbf{B}^{\text{vt}-\text{en}} = \begin{pmatrix} \tilde{\mathbf{Y}}^\top \\ \mathbf{y}_1^\top \end{pmatrix}, \qquad \text{where} \ \ \mathbf{A} = \begin{pmatrix} \tilde{\mathbf{A}} & \mathbf{a}_1 \\ \mathbf{a}_2^\top & \mathbf{a}_3 \end{pmatrix}.$$

Comment 1. For nuclear norm matrix completion, if rank of $\hat{\mathbf{L}}$ is $\hat{R}$, the solution for $\mathbf{A}$ and $\mathbf{B}$ is given by

$$\mathbf{A} = \mathbf{S}\mathbf{\Sigma}^{1/2}\ , \quad \mathbf{B} = \mathbf{R}\mathbf{\Sigma}^{1/2}$$

where $\hat{\mathbf{L}} = \mathbf{S}_{N\times\hat{R}}\mathbf{\Sigma}_{\hat{R}\times\hat{R}}\mathbf{R}_{T\times\hat{R}}^\top$ is singular value decomposition of $\hat{\mathbf{L}}$. The proof of this fact is provided in Mazumder et al. (2010).

Comment 2. For the horizontal regression the solution for $\mathbf{B}$ is

$$\mathbf{B}^{\text{hr}} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1 \\ \hat{\beta}_1 & \hat{\beta}_2 & \dots & \hat{\beta}_{T-1} \end{pmatrix},$$

and similarly for the vertical regression the solution for $\mathbf{A}$ is

$$\mathbf{A}^{\text{vt}} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1 \\ \hat{\gamma}_1 & \hat{\gamma}_2 & \dots & \hat{\gamma}_{N-1} \end{pmatrix}.$$

The regularization in the elastic net version only affects the last row of this matrix, and replaces it with a regularized version of the regression coefficients. □

Comment 3. The horizontal and vertical regressions are fundamentally different approaches, and they cannot easily be nested. Without some form of regularization they cannot be applied in the same setting, because the non-regularized versions require $N > T$ or $N < T$ respectively. As a result there is also no direct way to test the two methods against each other. Given a particular choice for regularization, however, one can use cross-validation methods to compare the two approaches. □

3.7 Three Illustrations

The objective of this section is to compare the accuracy of imputation for the matrix completion method with previously used methods. In particular, in a real data matrix $\mathbf{Y}$ where no unit is treated, we choose a subset of units as hypothetical treatment units and aim to predict their values (for time periods following a randomly selected initial time). Then, we report the average root-mean-squared-error (RMSE) of each algorithm on values for the treated (time, period) pairs. In these cases there is not necessarily a single right algorithm. Rather, we wish to assess which of the algorithms generally performs well, and which ones are robust to a variety of settings, including different adoption regimes and different configurations of the data.

We compare the following estimators:

- **DID**: Difference-in-differences based on regressing the observed outcomes on unit and time fixed effects and a dummy for the treatment.
- **VT-EN**: The vertical regression with elastic net regularization, relaxing the restrictions from the synthetic control estimator.
- **HR-EN**: The horizontal regression with elastic net regularization, similar to unconfoundedness type regressions.
- **SC-ADH**: The original synthetic control approach by Abadie et al. (2010), based on the vertical regression with Abadie-Diamond-Hainmueller restrictions.
- **MC-NNM**: Our proposed matrix completion approached via nuclear norm minimization, explained in Section 2 above.

The comparison between **MC-NNM** and the two versions of the elastic net estimator, **HR-EN** and **VT-EN**, is particularly salient. In much of the literature researchers choose ex ante between vertical and horizontal type regressions. The **MC-NNM** method allows one to sidestep that choice in a data-driven manner.

3.7.1 The Abadie-Diamond-Hainmueller California Smoking Data

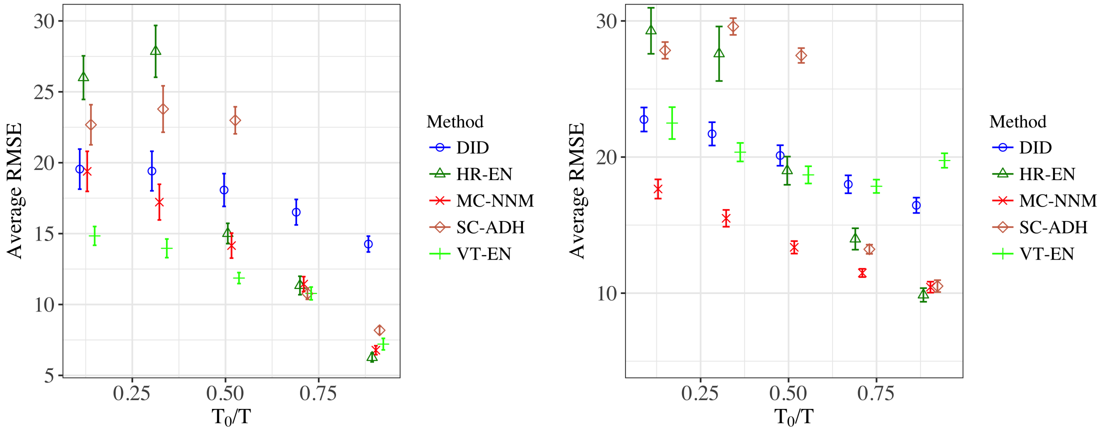
We use the control units from the California smoking data studied in Abadie et al. (2010) with $N = 38, T = 31$. Note that in the original data set there are 39 units but one of them (state of California) is treated which will be removed in this section since the untreated values for that unit are not available. We then artificially designate some units and time periods to be treated, and compare predicted values for those unit/time-periods to the actual values.

We consider two settings for the treatment adoption:

- Case 1: Simultaneous adoption where N_t units adopt the treatment in period $T_0 + 1$, and the remaining units never adopt the treatment.

- Case 2: Staggered adoption where N_t units adopt the treatment in some period after period T, with the actual adoption date varying among these units.

In each case, the average RMSE for different ratios T_0/T is reported in Figure 3.1. For clarity of the figures, for each T_0/T, while all confidence intervals of various methods are calculated using the same ratio T_0/T, in the figure they are slightly jittered to the left or right. In the simultaneous adoption case the VT-EN method is very sensitive to the number of treated periods, with its performance very poor if T_0/T is small, and superior to the others when T_0/T is close to one. DID generally does poorly, suggesting that the data are rich enough to support more complex models. The HR-EN, SC-ADH and MC-NNM methods generally do well in the simultaneous adoption case. With the staggered adoption the EN-T (horizontal regression) method does very poorly. Again our proposed MC-NNM method is always among the top performers, with SC-ADH and DID being competitive with few pre-treatment observations, but not with many pre-treatment observations, and VT-EN being competitive in the setting with many pre-treatment observations but not with few pre-treatment observations.



(a) Simultaneous adoption, $N_t = 8$

(b) Staggered adoption, $N_t = 35$

Figure 3.1: California Smoking Data

3.7.2 Stock Market Data

In the next illustration we use a financial data set – daily returns for 2453 stocks over 10 years (3082 days). Since we only have access to a single instance of the data, in order to observe statistical fluctuations of the RMSE, for each N and T we create 50 sub-samples by looking at the first T daily returns of N randomly sampled stocks for a range of pairs of (N, T), always with $N \times T = 4900$, ranging from very thin to very fat, $(N, T) = (490, 10)$, …, $(N, T) = (70, 70)$, …, $(N, T) = (10, 490)$, with in each case the second half the entries missing for a randomly selected half the units (so 25% of the entries missing overall), in a block design. Here we focus on the comparison between the horizontal and vertical regression and the matrix completion estimator as the shape of the matrix changes. To make the horizontal and vertical estimators well defined we use the elastic net regularized versions. We report the average RMSE. Figure 3.2 shows the results.

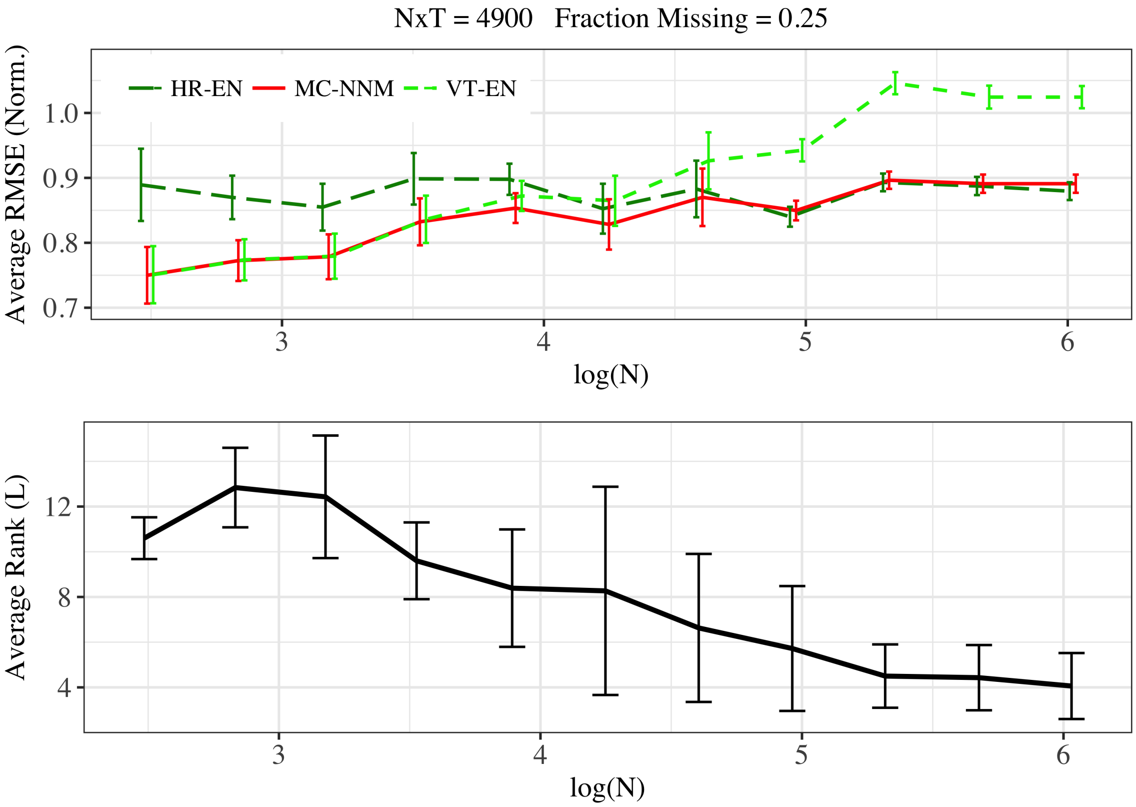


Figure 3.2: Stock Market Data

In the $T \ll N$ case the vertical estimator does poorly, not suprisingly because it attempts to do the vertical regression with too few time periods to estimate that well. When $N \ll T$, the horizontal estimator does poorly. The most interesting finding is that the proposed MC-NNM method adapts well to both regimes and does as well as the best estimator in both settings, and better than both in the approximately square setting.

3.7.3 Synthetic Data: Planted Hidden Confounder

In this illustration, we investigate the performance of different algorithms under the presence of confounding factors. We create (unobserved) dependency between the treatment and outcome and compare the performance of algorithms in estimating the average treatment effect. More precisely, suppose that $N = T = 20$ and $R = 3$, and matrices $\mathbf{Y}(0)$ and $\mathbf{Y}(1)$

are generated according to

$$\mathbf{Y}(0) = \mathbf{R}\mathbf{\Sigma}\mathbf{S}^\top + \epsilon$$
$$\mathbf{Y}(1) = \mathbf{Y}(0) + \mathbf{0.2}_{N\times T}\,,$$

where $\mathbf{R} \in \mathbb{R}^{N\times R} = [\alpha, \frac{\mathbf{1}_{N\times 1}}{\sqrt{N}}, \mathbf{u}_{N\times 1}]$, $\mathbf{S} \in \mathbb{R}^{T\times R} = [\frac{\mathbf{1}_{T\times 1}}{\sqrt{T}}, \gamma, \mathbf{v}_{N\times 1}]$, and $\mathbf{\Sigma} = \mathrm{diag}(10, 10, 5)$ is the SVD of the low rank (rank 3) component. Here, entries of vectors $\alpha, \mathbf{u}, \gamma$ and $\mathbf{v}$ have been generated according to $\mathsf{N}(0, 1)$, normalized and orthogonalized to achieve a valid SVD decomposition (meaning that columns of $\mathbf{R}$ and $\mathbf{S}$ are orthonormal). Also, ϵ has i.i.d. entries generated from $\mathsf{N}(0, 0.001)$. The dependency between outcomes $\mathbf{Y}(\cdot)$ and $\mathbf{W}$ is created as follows: we sort units (rows) based on values of $\mathbf{u}$, pick the 14 largest ones, and treat these units after random starting points. In other words, units with larger values of u_i are treated. Similar to the setting of this paper, we assume to have access to $\mathbf{Y}(\mathbf{W})$ and $\mathbf{W}$. As it can be observed, this creates an unobserved dependency between outcomes and treatments. The following figure depicts the above data generating process.

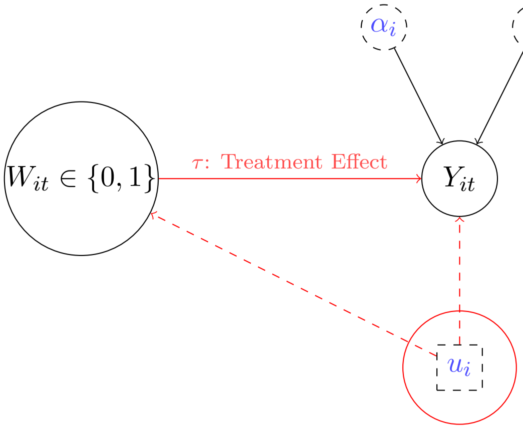


Figure 3.3: Data Generating Process with Confounding Effects

We repeat the above process for 100 times, and compare the performance of DID, SC-ADH, MC-NNM, and VT-EN with the true treatment effect which is $\tau = 0.2$. Figure 3.4 illustrates the achieved empirical distribution of estimates for these four algorithms. According to this figure, MC-NNM is the clear frontrunner and has the smallest bias and variance in estimating τ; it is capable of capturing the hidden factor $\mathbf{u}_i$ and it performs well. In contrast, DID is unable to capture this hidden effect and it leads to negative ATE in almost all problem instances. SC-ADH and VT-EN generally perform better, but they both have large variances.

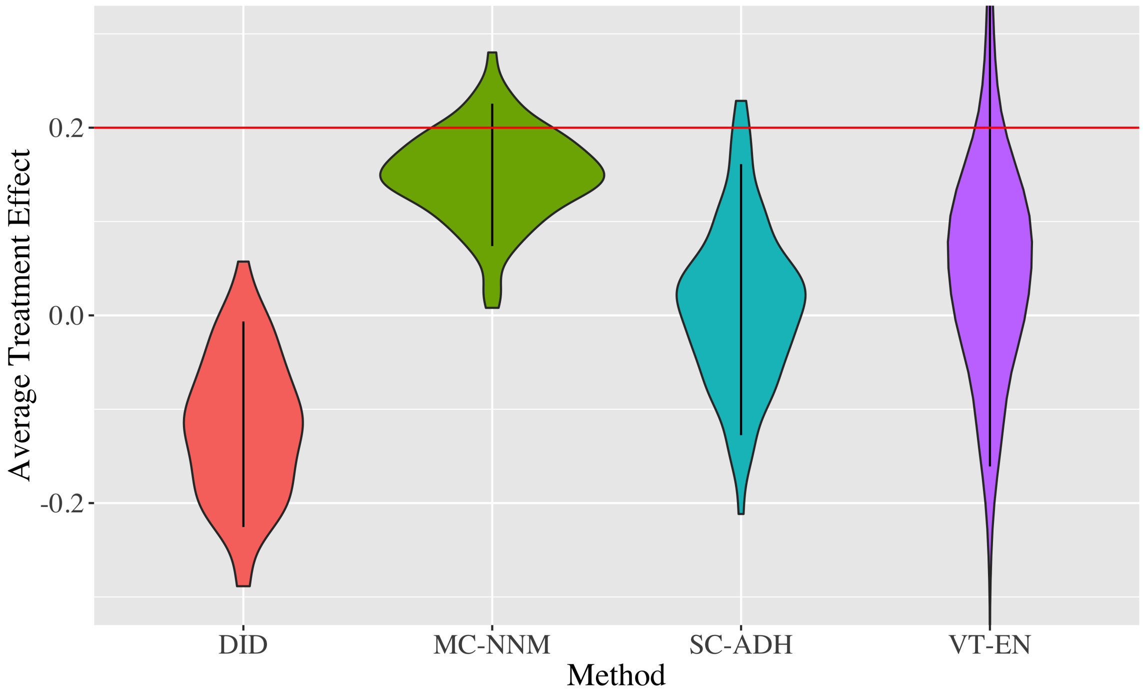


Figure 3.4: Planted Hidden Confounder Simulation Results

3.8 The General Model with Covariates

In Section 3.2 we described the basic model, and discussed the specification and estimation for the case without covariates. In this section we extend that to the case with unit-specific, time-specific, and unit-time specific covariates. For unit i we observe a vector of unit-specific covariates denoted by X_i, and $\mathbf{X}$ denoting the $N \times P$ matrix of covariates with ith row equal to $X_i^\top$. Similarly, Z_t denotes the time-specific covariates for period t, with $\mathbf{Z}$ denoting the $T \times Q$ matrix with t^{th} row equal to $Z_t^\top$. In addition we allow for a unit-time specific J by 1 vector of covariates V_{it}.

The model we consider is

$$Y_{it} = L^*_{it} + \sum_{p=1}^{P} \sum_{q=1}^{Q} X_{ip} H^*_{pq} Z_{qt} + \gamma^*_i + \delta^*_t + V_{it}^\top \beta^* + \varepsilon_{it} \, . \tag{3.8.1}$$

the ε_{it} is random noise. We are interested in estimating the unknown parameters $\mathbf{L}^*$, $\mathbf{H}^*$, γ^*, δ^* and β^*. This model allows for traditional econometric fixed effects for the units (the γ^*_i) and time effects (the δ^*_t). It also allows for fixed covariate (these have time varying coefficients) and time covariates (with individual coefficients) and time varying individual covariates. Note that although we can subsume the unit and time fixed effects into the matrix $\mathbf{L}^*$, we do not do so because we regularize the estimates of $\mathbf{L}^*$, but do not wish to regularize the estimates of the fixed effects.

The model can be rewritten as

$$\mathbf{Y} = \mathbf{L}^* + \mathbf{X}\mathbf{H}^*\mathbf{Z}^\top + \Gamma^* \mathbf{1}_T^\top + \mathbf{1}_N (\Delta^*)^\top + \Big[V_{it}^\top \beta^* \Big]_{it} + \boldsymbol{\varepsilon} \, . \tag{3.8.2}$$

Here $\mathbf{L}^*$ is in $\mathbb{R}^{N \times T}$, $\mathbf{H}^*$ is in $\mathbb{R}^{P \times Q}$, Γ^* is in $\mathbb{R}^{N \times 1}$ and Δ^* is in $\mathbb{R}^{T \times 1}$. An slightly richer version of this model that allows linear terms in covariates can be defined as by

$$\mathbf{Y} = \mathbf{L}^* + \tilde{\mathbf{X}}\tilde{\mathbf{H}}^*\tilde{\mathbf{Z}}^\top + \Gamma^* \mathbf{1}_T^\top + \mathbf{1}_N (\Delta^*)^\top + \Big[V_{it}^\top \beta^* \Big]_{it} + \boldsymbol{\varepsilon} \tag{3.8.3}$$

where $\tilde{\mathbf{X}} = [\mathbf{X} | \mathbf{I}_{N \times N}]$, $\tilde{\mathbf{Z}} = [\mathbf{Z} | \mathbf{I}_{T \times T}]$, and

$$\tilde{\mathbf{H}}^* = \begin{bmatrix} \mathbf{H}^*_{X,Z} & \mathbf{H}^*_X \\ \mathbf{H}^*_Z & \mathbf{0} \end{bmatrix}$$

where $\mathbf{H}^*_{XZ} \in \mathbb{R}^{P \times Q}$, $\mathbf{H}^*_Z \in \mathbb{R}^{N \times Q}$, and $\mathbf{H}^*_X \in \mathbb{R}^{P \times T}$. In particular,

$$\mathbf{Y} = \mathbf{L}^* + \tilde{\mathbf{X}}\tilde{\mathbf{H}}^*_{X,Z}\tilde{\mathbf{Z}}^\top + \tilde{\mathbf{H}}^*_Z\tilde{\mathbf{Z}}^\top + \mathbf{X}\tilde{\mathbf{H}}^*_X + \Gamma^* \mathbf{1}_T^\top + \mathbf{1}_N (\Delta^*)^\top + \Big[V_{it}^\top \beta^* \Big]_{it} + \boldsymbol{\varepsilon} \tag{3.8.4}$$

From now on, we will use the richer model (3.8.4) but abuse the notation and use notation $\mathbf{X}, \mathbf{H}^*, \mathbf{Z}$ instead of $\tilde{\mathbf{X}}, \tilde{\mathbf{H}}^*, \tilde{\mathbf{Z}}$. Therefore, the matrix $\mathbf{H}^*$ will be in $\mathbb{R}^{(N+P) \times (T+Q)}$.

We estimate $\mathbf{H}^*$, $\mathbf{L}^*$, δ^*, γ^*, and β^* by solving the following convex program,

$$\min_{\mathbf{H},\mathbf{L},\delta,\gamma,\beta}\left[\sum_{(i,t)\in\mathcal{O}}\left(Y_{it}-L_{it}-\sum_{p=1}^{P}\sum_{q=1}^{Q}X_{ip}H_{pq}Z_{qt}-\gamma_i-\delta_t-V_{it}\beta\right)^2+\lambda_L\|\mathbf{L}\|_*+\lambda_H\|\mathbf{H}\|_{1,e}\right].$$

Here $\|\mathbf{H}\|_{1,e}=\sum_{i,t}|H_{it}|$ is the element-wise ℓ_1 norm. We choose λ_L and λ_H through cross-validation.

Solving this convex program is similar to the covariate-free case. In particular, by using a similar operator to shrink_λ, defined in §3.2, that performs coordinate descent with respect to $\mathbf{H}$. Then we can apply this operator after each step of using shrink_λ. Coordinate descent with respect to γ, δ, and β is performed similarly but using a simpler operation since the function is smooth with respect to them.

3.9 Generalizations

Here we provide a brief discussion on how our estimator or its analysis should be adapted to more general settings.

3.9.1 Autocorrelated Errors

One drawback of MC-NNM is that it does not take into account the time series nature of the observations. It is likely that the columns of $\boldsymbol{\varepsilon}$ exhibit autocorrelation. We can take this into account by modifying the objective function. Let us consider this in the case without covariates, and, for illustrative purposes, let us use an autoregressive model of order one. Let $\mathbf{Y}_{i\cdot}$ and $\mathbf{L}_{i\cdot}$ be the i^{th} row of $\mathbf{Y}$ and $\mathbf{L}$ respectively. The original objective function for $\mathcal{O}=[N]\times[T]$ is

$$\sum_{i=1}^{N}\sum_{t=1}^{N}(Y_{it}-L_{it})^2+\lambda_L\|\mathbf{L}\|_*=\sum_{i=1}^{N}(Y_{i\cdot}-L_{i\cdot})(Y_{i\cdot}-L_{i\cdot})^\top+\lambda_L\|\mathbf{L}\|_*.$$

We can modify this to

$$\sum_{i=1}^{N}(Y_{i\cdot}-L_{i\cdot})\boldsymbol{\Omega}^{-1}(Y_{i\cdot}-L_{i\cdot})^\top+\lambda_L\|\mathbf{L}\|_*,$$

where the choice for the $T \times T$ matrix $\mathbf{\Omega}$ would reflect the autocorrelation in the $\boldsymbol{\varepsilon}_{it}$. For example, with a first order autoregressive process, we would use

$$\Omega_{ts} = \rho^{|t-s|} \, ,$$

with ρ an estimate of the autoregressive coefficient. Similarly, for the more general version $\mathcal{O} \subset [N] \times [T]$, we can use the function

$$\sum_{(i,t)\in\mathcal{O}} \sum_{(i,s)\in\mathcal{O}} (Y_{it} - L_{it})[\mathbf{\Omega}^{-1}]_{ts}(Y_{is} - L_{is})^\top + \lambda_L \|\mathbf{L}\|_* \, .$$

3.9.2 Weighted Loss Function

Another limitation of MC-NNM is that it puts equal weight on all elements of the difference $\mathbf{Y} - \mathbf{L}$ (ignoring the covariates). Ultimately we care solely about predictions of the model for the missing elements of $\mathbf{Y}$, and for that reason it is natural to emphasize the fit of the model for elements of $\mathbf{Y}$ that are observed, but that are similar to the elements that are missing. In the program evaluation literature this is often achieved by weighting the fit by the propensity score, the probability of outcomes for a unit being missing.

We can do so in the current setting by modelling this probability in terms of the covariates and a latent factor structure. Let the propensity score be $e_{it} = \mathbb{P}(W_{it} = 1 | X_i, Z_t, V_{it})$, and let $\mathbf{E}$ be the $N \times T$ matrix with typical element e_{it}. Let us again consider the case without covariates. In that case we may wish to model the assignment $\mathbf{W}$ as

$$\mathbf{W}_{N\times T} = \mathbf{E}_{N\times T} + \boldsymbol{\eta}_{N\times T}.$$

We can estimate this using the same matrix completion methods as before, now without any missing values:

$$\hat{\mathbf{E}} = \arg\min_{\mathbf{E}} \frac{1}{NT} \sum_{(i,t)} (W_{it} - e_{it})^2 + \lambda_L \|\mathbf{E}\|_* \, .$$

Given the estimated propensity score we can then weight the objective function for estimating $\mathbf{L}^*$:

$$\hat{\mathbf{L}} = \arg\min_{\mathbf{L}} \sum_{(i,t)\in\mathcal{O}} \frac{\hat{e}_{it}}{1 - \hat{e}_{it}} (Y_{it} - L_{it})^2 + \lambda_L \|\mathbf{L}\|_* \, .$$

3.9.3 Relaxing the Dependence of Theorem 5 on p_c

Recall from §3.5.1 that the average number of control units is $\sum_{i=1}^{N} \pi_T^{(i)}$. Therefore, the fraction of control units is $\sum_{i=1}^{N} \pi_T^{(i)}/N$. However, the estimation error in Theorem 5 depends on $p_c = \min_{1\le i\le N} \pi_T^{(i)}$ rather than $\sum_{i=1}^{N} \pi_T^{(i)}/N$. The reason for this, as discussed in §3.5.1 is due to special classes of matrices $\mathbf{L}^*$ where most of the rows are nearly zero (e.g, when only one row is non-zero). In order to relax this constraint we would need to restrict the family of matrices $\mathbf{L}^*$. An example of such restriction is given by Negahban and Wainwright (2012) where they assume $\mathbf{L}^*$ is not too spiky. Formally, they assume the ratio $\|\mathbf{L}^*\|_{\max}/\|\mathbf{L}^*\|_F$ should be of order $1/\sqrt{NT}$ up to logarithmic terms. To see the intuition for this, in a matrix with all equal entries this ratio is $1/\sqrt{NT}$ whereas in a matrix where only the $(1, 1)$ entry is non-zero the ratio is 1. While both matrices have rank 1, in the former matrix the value of $\|\mathbf{L}^*\|_F$ is obtained from most of the entries. In such situations, one can extend our results and obtain an upper bound that depends on $\sum_{i=1}^{N} \pi_T^{(i)}/N$.

3.9.4 Nearly Low-rank Matrices

Another possible extension of Theorem 5 is to the cases where $\mathbf{L}^*$ may have high rank, but most of its singular values are small. More formally, if $\sigma_1 \ge \cdots > \sigma_{\min(N,T)}$ are singular values of $\mathbf{L}^*$, one can obtain upper bounds that depend on k and $\sum_{r=k+1}^{\min(N,T)} \sigma_r$ for any $k \in [\min(N, T)]$. One can then optimize the upper bound by selecting the best k. In the low-rank case such optimization leads to selecting k equal to R. This type of more general upper bound has been proved in some of prior matrix completion literature, e.g. Negahban and Wainwright (2012). We expect their analyses would be generalize-able to our setting (when entries of $\mathcal{O}$ are not independent).

3.9.5 Additional Missing Entries

In §3.5.1 we assumed that all entries (i, t) of $\mathbf{Y}$ for $t \le t_i$ are observed. However, it may be possible that some such values are missing due to lack of data collection. This does not mean that any treatment occurred in the pre-treatment period. Rather, such scenario can occur when measuring outcome values is costly and can be missed. In this case, one can extend Theorem 5 to the setting with

$$\mathcal{O} = \left[\bigcup_{i=1}^{N} \Big\{(i,1),(i,2),\ldots,(i,t_i)\Big\}\right] \setminus \mathcal{O}_{\text{miss}}\,.$$

where each $(i,t) \in \cup_{i=1}^{N}\{(i,1),(i,2),\ldots,(i,t_i)\}$ can be in $\mathcal{O}_{\text{miss}}$, independently, with probability p for p that is not too large.

3.10 Conclusions

We develop a new estimator for the interactive fixed effects model in settings where the interest is in average causal effects. The proposed estimator has superior computational properties in settings with large N and T, and allows for a relatively large number of factors. We show how this set up relates to the program evaluation and synthetic control literatures. In illustrations we show that the method adapts well to different configurations of the data, and find that generally it outperforms the synthetic control estimators from Abadie et al. (2010) and the elastic net estimators from Doudchenko and Imbens (2016).

Chapter 4

Non-Parametric Inference in High Dimensions

4.1 Introduction

Many non-parametric estimation problems in econometrics and causal inference can be formulated as finding a parameter vector $\theta(x) \in \mathbb{R}^p$ that is a solution to a set of conditional moment equations:

$$\mathbb{E}[\psi(Z; \theta(x)) | X = x] = 0\,, \tag{4.1.1}$$

when given n i.i.d. samples $(Z_1, \ldots, Z_n)$ from the distribution of Z, where $\psi : \mathcal{Z} \times \mathbb{R}^p \to \mathbb{R}^p$ is a known vector valued moment function, $\mathcal{Z}$ is an arbitrary data space, $X \in \mathcal{X} \subset \mathbb{R}^D$ is the feature vector that is included in Z. Examples include non-parametric regression[1], quantile regression[2], heterogeneous treatment effect estimation[3], instrumental variable regression[4], local maximum likelihood estimation[5] and estimation of structural econometric models (see, e.g., Reiss and Wolak 2007) and examples in Chernozhukov et al. (2016), Chernozhukov et al. (2018)). The study of such conditional moment restriction problems has a long history in econometrics (see, e.g., Newey 1993, Ai and Chen 2003, Chen and Pouzo 2009, Chernozhukov et al. 2015). However, the majority of the literature assumes that the conditioning variable X is low dimensional, i.e. D is a constant as the sample size n grows

[1] $Z = (X, Y)$, where $Y \in \mathbb{R}^p$ is the dependent variable, and $\psi(Z; \theta(x)) = Y - \theta(x)$.

[2] $Z = (X, Y)$ and $\psi(Z; \theta(x)) = 1\{Y \leq \theta(x)\} - \alpha$, for some $\alpha \in [0, 1]$ that denotes the target quantile.

[3] $Z = (X, T, Y)$, where $T \in \mathbb{R}^p$ is a vector of treatments, and $\psi(Z; \theta(x)) = (Y - \langle \theta(x), T \rangle)\, T$.

[4] $Z = (X, T, W, Y)$, where $T \in \mathbb{R}$ is a treatment, $W \in \mathbb{R}$ an instrument and $\psi(Z; \theta(x)) = (Y - \theta(x)\, T)\, W$.

[5] Where the distribution of Z admits a known density $f(z; \theta(x))$ and $\psi(Z; \theta(x)) = \nabla_\theta \log(f(Z; \theta(x)))$.

(see, e.g., Athey et al. 2019). High dimensional variants have primarily been analyzed under parametric assumptions on $\theta(x)$, such as sparse linear forms (see, e.g., Chernozhukov et al. 2018). There are some papers that address the fully non-parametric setup (see, e.g., Lafferty and Wasserman 2008, Dasgupta and Freund 2008, Kpotufe 2011, Biau 2012, Scornet et al. 2015) but those are focused on the estimation problem, and do not address inference (i.e., constructing asymptotically valid confidence intervals).

The goal of this work is to address estimation and inference in conditional moment models with a high-dimensional conditioning variable. As is obvious without any further structural assumptions on the problem, the exponential in dimension rates of approximately $n^{1/D}$ (see, e.g., Stone 1982) cannot be avoided. Thereby, estimation is infeasible even if D grows very slowly with n. Our work, follows a long line of work in machine learning (Dasgupta and Freund 2008, Kpotufe 2011, Kpotufe and Garg 2013), which is founded on the observation that in many practical applications, even though the variable X is high-dimensional (e.g. an image), one typically expects that the coordinates of X are highly correlated. The latter intuition is formally captured by assuming that the distribution of X has a small doubling measure around the target point x.

We refer to the latter notion of dimension, as the intrinsic dimension of the problem. Such a notion has been studied in the statistical machine learning literature, so as to establish fast estimation rates in high-dimensional kernel regression settings (Dasgupta and Freund 2008, Kpotufe 2011, Kpotufe and Garg 2013, Xue and Kpotufe 2018, Chen and Shah 2018, Kim et al. 2018, Jiang 2017). However, these works solely address the problem of estimation and do not characterize the asymptotic distribution of the estimates, so as to enable inference, hypothesis testing and confidence interval construction. Moreover, they only address the regression setting and not the general conditional moment problem and consequently do not extend to quantile regression, instrumental variable regression or treatment effect estimation.

From the econometrics side, pioneering works of Wager and Athey (2018), Athey et al. (2019) address estimation and inference of conditional moment models with all the aforementioned desiderata that are required for the application of such methodologies to social sciences, albeit in the low dimensional regime. In particular, Wager and Athey (2018) consider regression and heterogeneous treatment effect estimation with a scalar $\theta(x)$ and prove $n^{1/D}$-asymptotic normality of a sub-sampled random forest based estimator and Athey et al. (2019) extend it to the general conditional moment settings.

These results have been extended and improved in multiple directions, such as improved estimation rates through local linear smoothing Friedberg et al. (2018), robustness to nuisance parameter estimation error Oprescu et al. (2018) and improved bias analysis via sub-sampled nearest neighbor estimation Fan et al. (2018). However, they all require low dimensional setting and the rate provided by the theoretical analysis is roughly $n^{-1/D}$, i.e. to get a confidence interval of length ϵ or an estimation error of ϵ, one would need to collect $O(\epsilon^{-D})$ samples which is prohibitive in most target applications of machine learning based econometrics.

Hence, there is a strong need to provide theoretical results that justify the success of machine learning estimators for doing inference, via their adaptivity to some low dimensional hidden structure in the data. *Our work makes a first step in this direction and provides estimation and asymptotic normality results for the general conditional moment problem, where the rate of estimation and the asymptotic variance depend only on the intrinsic dimension, independent of the explicit dimension of the conditioning variable.*

Our analysis proceeds in four parts. First, we extend the results by Wager and Athey (2018), Athey et al. (2019) on the asymptotic normality of sub-sampled kernel estimators to the high-dimensional, low intrinsic dimension regime and to vector valued parameters $\theta(x)$. Concretely, when given a sample $S = (Z_1, \ldots, Z_n)$, our estimator is based on the approach proposed in Athey et al. (2019) of solving a locally weighted empirical version of the conditional moment restriction

$$\hat{\theta}(x) \text{ solves} : \sum_{i=1}^{n} K(x, X_i, S)\, \psi(Z_i; \theta) = 0\,, \tag{4.1.2}$$

where $K(x, X_i, S)$ captures proximity of X_i to the target point x. The approach dates back to early work in statistics on local maximum likelihood estimation (Fan et al. 1998, Newey 1994, Stone 1977, Tibshirani and Hastie 1987). As in Athey et al. (2019), we consider weights $K(x, X_i, S)$ that take the form of an average over B base weights: $K(x, X_i, S) = \frac{1}{B}\sum_{b=1}^{B} K(x, X_i, S_b)\, 1\{i \in S_b\}$, where each $K(x, X_i, S_b)$ is calculated based on a randomly drawn sub-sample S_b of size $s < n$ from the original sample. We will typically refer to the function K as the *kernel*. In Wager and Athey (2018), Athey et al. (2019) $K(x, X_i, S_b)$ is calculated by building a tree on the sub-sample, while in Fan et al. (2018) it is calculated based on the 1-NN rule on the sub-sample.

Our main results are general estimation rate and asymptotic normality theorems for the estimator $\hat{\theta}(x)$ (see Theorems 8 and 9), which are stated in terms of two high-level assumptions, specifically an upper bound $\epsilon(s)$ on the rate at which the kernel "shrinks" and a lower bound $\eta(s)$ on the "incrementality" of the kernel. Notably, the explicit dimension of the conditioning variable D does not enter the theorem, so it suffices in what follows to show that $\epsilon(s)$ and $\eta(s)$ depend only on d rather than D.

The shrinkage rate $\epsilon(s)$ is defined as the ℓ_2-distance between the target point x and the furthest point on which the kernel places positive weight X_i, when trained on a data set of s samples, i.e.,

$$\epsilon(s) = \mathbb{E}\left[\sup\{\|X_i - x\|_2 : i \in S_b, K(x, X_i, S_b) > 0, |S_b| = s\}\right] . \tag{4.1.3}$$

The shrinkage rate of the kernel controls the bias of the estimate (small $\epsilon(s)$ implies low bias). The sub-sampling size s is a lever to trade off bias and variance; larger s achieves smaller bias, since $\epsilon(s)$ is smaller, but increases the variance, since for any fixed x the weights $K(x, X_i, S_b)$ will tend to concentrate on the same data points, rather than averaging over observations. Both estimation and asymptotic normality results require the bias to be controlled through the shrinkage rate.

Incrementality of a kernel describes how much information is revealed about the weight of a sample i solely by knowledge of X_i, and is captured by the second moment of the conditional expected weight

$$\eta(s) = \mathbb{E}\left[\mathbb{E}\left[K(x, X_i, S_b) \mid X_i\right]^2\right] . \tag{4.1.4}$$

The incrementality assumption is used in the asymptotic normality proof to argue that the weights have sufficiently high variance that all data points have some influence on the estimate. From the technical side, we use the Hájek projection to analyze our U-statistic estimator. Incrementality ensures that there is sufficiently weak dependence in the weights across a sequence of sub-samples and hence the central limit theorem applies. As discussed, the sub-sampling size s can be used to control the variance of the weights, and so incrementality and shrinkage are related. We make this precise, proving that incrementality can be lower bounded as a function of kernel shrinkage, so that having a sufficiently low shrinkage rate enables both estimation and inference. These general results could be of independent interest beyond the scope of this work.

For the second part of our analysis, we specialize to the case where the base kernel is the

k-NN kernel, for some constant k. We prove that both shrinkage and incrementality depend only on the intrinsic dimension d, rather than the explicit dimension D. In particular, we show that $\epsilon(s) = O(s^{-1/d})$ and $\eta(s) = \Theta(1/s)$. These lead to our main theorem *that the sub-sampled k-NN estimate achieves an estimation rate of order $n^{1/(d+2)}$ and is also $n^{1/(d+2)}$-asymptotically normal.*

In the third part, we provide a closed form characterization of the asymptotic variance of the sub-sampled k-NN estimate, as a function of the conditional variance of the moments, which is defined as $\sigma^2(x) = \mathrm{Var}\left(\psi(Z;\theta) \mid X = x\right)$. For example, for the 1-NN kernel, the asymptotic variance is given by

$$\mathrm{Var}(\hat{\theta}(x)) = \frac{\sigma^2(x)s^2}{n(2s-1)}\,.$$

This strengthens prior results of Fan et al. (2018) and Wager and Athey (2018), which only proved the existence of an asymptotic variance without providing an explicit form (and thereby relied on bootstrap approaches for the construction of confidence intervals). Our tight characterization enables an easy construction of plugin normal-based intervals that only require a preliminary estimate of $\sigma(x)$. Our Monte Carlo study shows that such intervals provide very good finite sample coverage in a high dimensional regression setup (see Figure 4.1)[6].

Finally in the last part, we discuss an adaptive data-driven approach for picking the sub-sample size s so as to achieve estimation or asymptotic normality with rates that only depend on the unknown intrinsic dimension. This allows us to achieve near-optimal rates while adapting to the unknown intrinsic dimension of data (see Propositions 4 and 5). Figure 4.2 depicts the performance of our adaptive approach compared to two benchmarks, one constructed based on theory for intrinsic dimension d which may be unknown, and the other one constructed naïvely based on the known but sub-optimal extrinsic dimension D. As it can be observed from this figure, setting s based on intrinsic dimension d allows us to build more accurate and smaller confidence intervals, which is crucial for drawing inference in the high-dimensional finite sample regime. Our adaptive approach uses samples to pick s very close to the value suggested by our theory and therefore leads to a compelling finite sample coverage[7]. Such estimators address the curse of dimensionality by adapting to a

[6] See Appendix C.1 for detailed explanation of our simulations.

[7] A preliminary implementation of our code is available via http://github.com/khashayarkhv/np_inference_intrinsic.

priori unknown latent structure in the data.

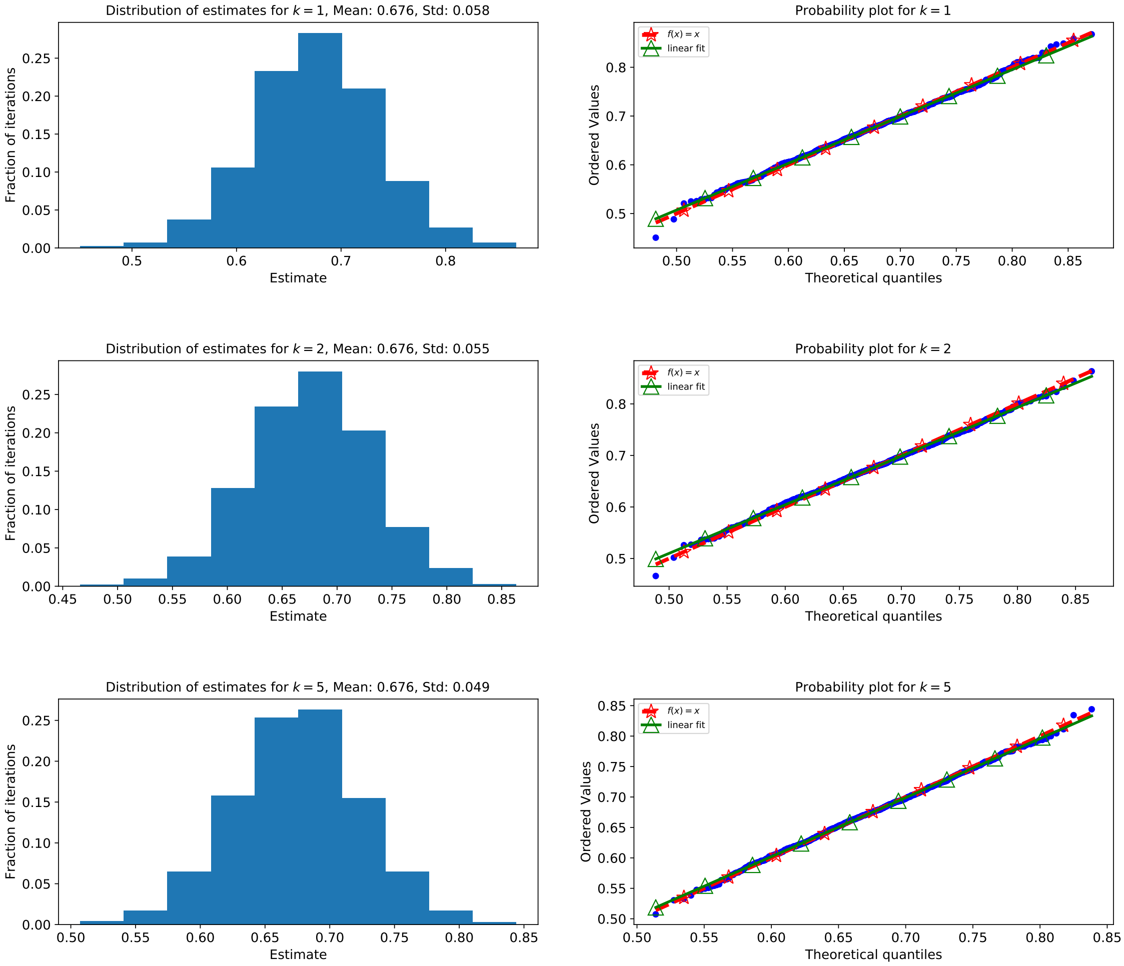


Figure 4.1: Left: distribution of estimates over 1000 Monte Carlo runs for $k = 1, 2, 5$. Right: the quantile-quantile plot when comparing to the theoretical asymptotic normal distribution of estimates stemming from our characterization, whose means are $0.676, 0.676$, and 0.676 for $k = 1, 2, 5$, respectively. Standard deviations are $0.058, 0.055$, and 0.049 for $k = 1, 2, 5$ respectively. $n = 20000$, $D = 20$, $d = 2$, $\mathbb{E}[Y|X] = \frac{1}{1+\exp\{-3X[0]\}}$, $\sigma = 1$. Test point: $x[0] \approx 0.245$, $\mathbb{E}[Y|X = x] \approx 0.676$.

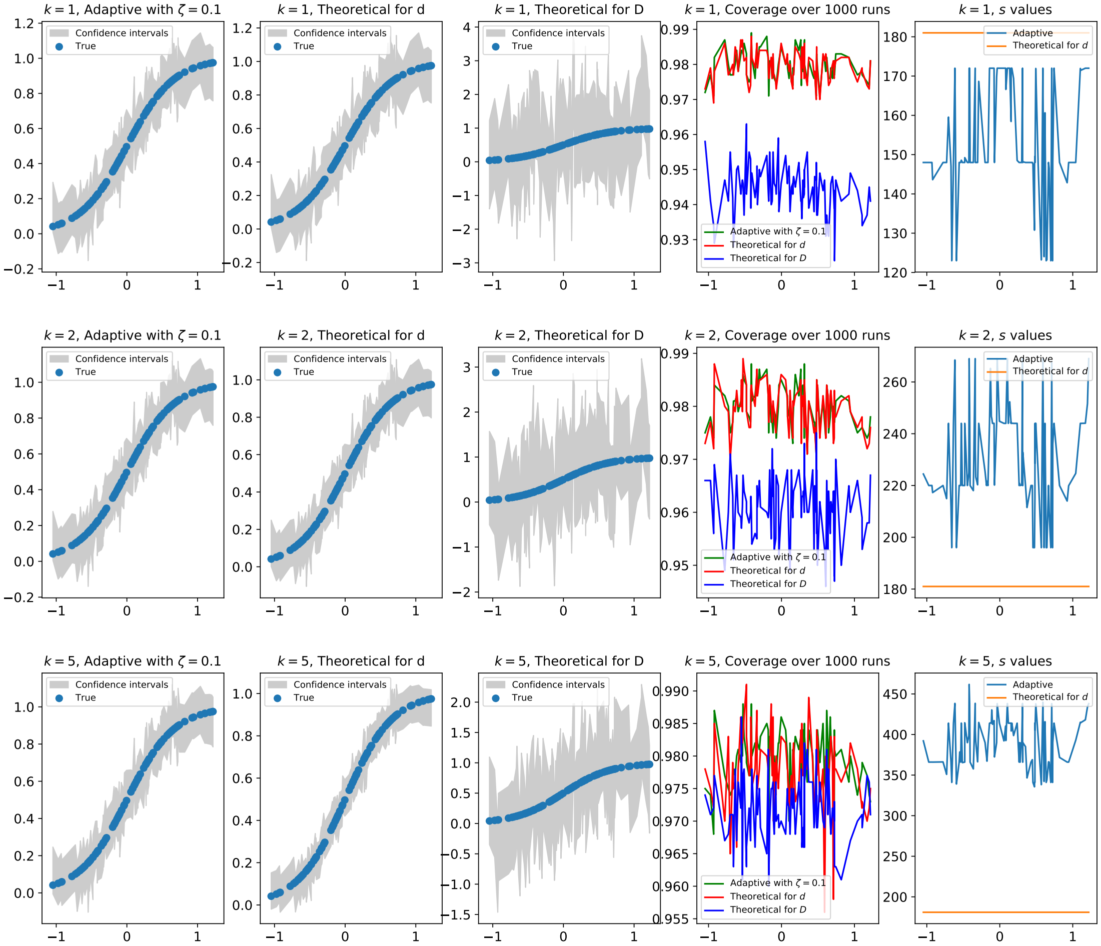


Figure 4.2: Confidence interval and true values for 100 randomly sampled test points on a single run for $k = 1, 2, 5$ and when (1) left: $s = s_\zeta$ is chosen adaptively using Proposition 5 with $\zeta = 0.1$, (2) second from the left: $s = n^{1.05d/(d+2)}$, and (3) middle: $s = n^{1.05D/(D+2)}$. Second from the right: coverage over 1000 runs for three different methods described. Right: average value of s_ζ chosen adaptively using Proposition 5 for $\zeta = 0.1$ for different test points compared to the theoretical value $s = n^{1.05d/(d+2)}$. Here $n = 20000$, $D = 20$, $d = 2$, $\mathbb{E}[Y|X] = \frac{1}{1+\exp\{-3X[0]\}}$, $\sigma = 1$. Nominal coverage: 0.98.

4.1.1 Related Work

Average Treatment Effect Estimation. There exists a vast literature on average treatment effect estimation in high-dimensional settings. The key challenge in such settings is the problem of overfitting which is usually handled by adding regularization terms. However, this leads to a shrinked estimate for the average treatment effect and therefore not desirable. The literature has taken various approaches to solve this issue. For instance, Belloni et al. (2014a,b) used a two-step method for estimating average treatment effect where in the first step feature-selection is accomplished via a lasso and then treatment effect is estimated using selected features. Athey et al. (2018) studied approximate residual balancing where a combination of weight balancing and regression adjustment is used for removing undesired bias and for achieving a double robust estimator. Chernozhukov et al. (2016, 2018) considered a more general semi-parametric framework and studied debiased/double machine learning methods via first order Neyman orthogonality condition. Mackey et al. (2017) extended this result to higher order moments. Please refer to Athey and Imbens (2017), Mullainathan and Spiess (2017), Belloni et al. (2017) for a review on this literature.

Conditional Treatment Effect Estimation. However, in many applications, researchers are interested in estimating conditional treatment effect on various sub-populations. One effective solution is to use one of the methods described in previous paragraph to estimate problem parameters and then project such estimations onto the sub-population of interest. However, these approaches usually perform poorly when there is a model mis-specification, i.e., when the true underlying model does not belong to the parametric search space. Consequently, researchers have studied non-parametric estimators such as k-NN estimators, kernel estimators, and random forests. While these non-parametric estimators are very robust to model mis-specification and work well under mild assumptions on the function of interest, they suffer from the curse of dimensionality (see, e.g., Bellman 1961, Robins and Ritov 1997, Friedman et al. 2001). Therefore, for applying these estimators in high-dimensional settings it is necessary to design and study non-parametric estimators that are able to overcome curse of dimensionality when possible.

The seminal work of Wager and Athey (2018) utilized random forests originally introduced by Breiman (2001) and adapted them nicely for estimating heterogeneous treatment effect. In particular, the authors demonstrated how the recursive partitioning idea, explained in Athey and Imbens (2016) for estimating heterogeneity in causal settings, can be further analyzed to establish asymptotic properties of such estimators. The main premise of random forests is that they are able to adaptively select nearest neighbors and that is very desirable in high-dimensional settings where discarding uninformative features is necessary for combating the curse of dimensionality. In a follow-up work, they extended these results and introduced Generalized Random Forests for more general setting of solving generalized moment equations (Athey et al. 2019). There has been some interesting developments of such ideas to other settings. Fan et al. (2018) introduced Distributional Nearest Neighbor (DNN) where they used 1-NN estimators together with sub-sampling and explained that by precisely combining two of these estimators for different sub-sampling sizes, the first order bias term can be efficiently removed. Friedberg et al. (2018) paired this idea with a local linear regression adjustment and introduced Local Linear Forests in order to improve forest estimations for smooth functions. Oprescu et al. (2018) incorporated the double machine learning methods of Chernozhukov et al. (2018) into GMM framework of Athey et al. (2019) and studied Orthogonal Random Forests in partially linear regression models with high-dimensional controls. Although forest kernels studied in Wager and Athey (2018) and Athey et al. (2019) seem to work well in high-dimensional applications, to the best of our knowledge, there still does not exists a theoretical result supporting it. In fact, all existing theoretical results suffer from the curse of dimensionality as they depend on the dimension of problem D.

Estimation Adaptive to Intrinsic Dimension. The literature on machine learning and non-parametric statistics has recently studied how these worst-case performances can be avoided when the intrinsic dimension of problem is smaller than D. Please refer to Cutler (1993) for different notions of intrinsic dimension in metric spaces. Dasgupta and Freund (2008) studied random projection trees and showed that the structure of these trees do not depend on the actual dimension D, but rather on the intrinsic dimension d. They used the notion of Assouad Dimension, introduced by Assouad (1983), and proved that using random directions for splitting, the number of levels required for halving the diameter of any leaf scales as $O(d \log d)$. The follow-up work (Verma et al. 2009) generalized these results for some other notions of dimension. Kpotufe and Dasgupta (2012) extended this idea to the regression setting and proved integrated risk bounds for random projection trees that were only dependent on intrinsic dimension. Kpotufe (2011), Kpotufe and Garg (2013) studied this in the context of k-NN and kernel estimations and established uniform point-wise risk bounds only depending on the local intrinsic dimension. They also provided data-driven approaches for choosing k in the case of k-NN, and bandwidth in case of kernel estimators, so that they can adapt to the unknown intrinsic dimension and also smoothness of non-parametric function of interest.

k-NN and Generalized Method of Moments. Our work is deeply rooted in the literature on intrinsic dimension explained above, literature on k-NN estimators (see, e.g., Mack 1981, Samworth 2012, Györfi et al. 2006, Biau and Devroye 2015, Berrett et al. 2019, Fan et al. 2018), and generalized method of moments (see, e.g., Tibshirani and Hastie 1987, Staniswalis 1989, Fan et al. 1998, Hansen 1982, Stone 1977, Lewbel 2007, Mackey et al. 2017). We adapt the framework of Athey et al. (2019) and Oprescu et al. (2018) and solve a generalized moment problem using a sub-sampled k-NN estimator, originally studied by Fan et al. (2018). In particular, the authors studied the problem of heterogeneous treatment effect estimation under unconfoundedness using sub-sampled 1-NN estimator, which they refer to as DNN estimator. Our work complements the work of Fan et al. (2018) and extends it to the generalized method of moment setting and also allows for general k-NN estimators. Also, we establish that these DNN estimators are able to adapt to intrinsic dimension of problem d and hence do not suffer from the curse of dimensionality.

Our result differs from existing literature on intrinsic dimension (e,g., Dasgupta and Freund 2008, Kpotufe 2011, Kpotufe and Garg 2013) since in addition to estimation guarantees for the regression setting, we also allow valid inference in solving conditional moment equations. Our asymptotic normality result is different from existing results for k-NN (see, e.g., Mack 1981), generalized method of moments (see, e.g., Lewbel 2007). Indeed, these papers only establish the asymptotic distribution of these estimators without providing a data-driven way for constructing confidence intervals.

We also provide the exact expression for the asymptotic variance of DNN estimator built using a k-NN kernel, which enables plug-in construction of confidence intervals, rather than the bootstrap method of (Efron 1982) which was used by (Wager and Athey 2018, Athey et al. 2019, Fan et al. 2018). While establishing consistency and asymptotic normality of our estimator, we also provide more general bounds on kernel shrinkage rate and also incrementality which can be useful for establishing asymptotic properties in other applications. One such application is given in high-dimensional settings where the exact nearest neighbor search is computationally expensive and Approximate Nearest Neighbor (ANN) search is often replaced in order to reduce this cost. Our flexible result allows us to use the state-of-the-art ANN algorithms (see, e.g., Andoni et al. 2017, 2018) while maintaining consistency and asymptotic normality.

Conditional Stochastic Optimization and Newsvendor Problem. Finally, there are technical parallels between our analysis and the literature on conditional stochastic optimization (see, e.g., Ban and Rudin 2018, Bertsimas and Kallus 2014b, Hannah et al. 2010, Hanasusanto and Kuhn 2013). In particular, this literature focuses on estimating $z^*(x) = \arg\min_z \mathbb{E}[c(z;Y) \mid X = x]$, where z is the decision variable, Y is the uncertain quantity of interest, $X = x$ is the set of observed features, and $c(z;Y)$ is the uncertain cost associated with decision z. Ban and Rudin (2018) study feature-based newsvendor problem and consider two empirical minimization risk (ERM) approaches together with a non-parametric kernel estimation method for solving this problem. Bertsimas and Kallus (2014b) study the problem of conditional stochastic optimization problem with a general cost function, apply various non-parametric machine learning estimators such as k-NN, random forests, and Nadaraya-Watson's kernel regression (Nadaraya 1964, Watson 1964), and provide asymptotic consistency results for them. The main focus of this literature is in providing a decision $\hat{z}(x)$ such that the expected cost under this decision is close to the optimal decision $z^*(x)$ (or $\hat{z}(x)$ itself is close to $z^*(x)$). However, we are mainly interested on the task of inference which in this setting translates to providing valid confidence intervals for $\hat{z}(x)$. While our techniques are mainly designed for solving the conditional moment equations, i.e., finding $\theta(x)$ that solves $\mathbb{E}[\psi(Z;\theta) \mid X = x] = 0$, with a slight change, our techniques are also applicable to the conditional stochastic optimization setting.

4.1.2 Main Contributions and Organization of This Chapter

In §4.2, we explain the problem that we study in this chapter and provide preliminary definitions. In §4.2.1, we explain the general sub-sampled kernel estimation. In particular, given a general kernel K, Algorithm 4 explains how the parameter of interest $\theta(x)$ is estimated. In §4.2.2, we apply this algorithm to the special case of k-NN kernel (see Algorithm 5). In §4.2.3, we explain the notion of intrinsic dimension defined using locally low doubling measures and provide examples of spaces with low intrinsic dimension in §4.2.4. In §4.3, we state other assumptions that we need for our analysis.

Our analysis starts in §4.4, where we provide general estimation (Theorem 8) and inference results (Theorem 9) for kernels that satisfy shrinkage and incrementality conditions. In particular, assuming that the local intrinsic dimension around target point x is equal to d, we prove that the finite sample estimation error of order $n^{-1/(d+2)}$ together with $n^{1/(d+2)}$-asymptotically normality result of general sub-sampled kernel estimator for solving the generalized moment problem regardless of how big the actual dimension D. While an upper bound on the shrinkage rate is sufficient for providing estimation guarantees, for asymptotic normality we also require the incrementality to decay at an appropriate rate. In many situations, it is easier to establish kernel shrinkage. Therefore, for making the asymptotic normality more applicable, in Lemma 7, we prove a lower bound on the incrementality based on the kernel shrinkage.

In §4.5, we establish appropriate shrinkage and incerementality rates for the k-NN kernel and combining this with the results of §4.4, we prove estimation (Theorem 10) and inference rates (Theorem 12) for the k-NN kernel that only depend on the intrinsic dimension d. Along the way of establishing such results, in Theorem 11, we provide the exact expression for the asymptotic variance of sub-sampled k-NN estimator, which enables plug-in construction of confidence intervals.

The sub-sampling size required for achieving these results depends on the intrinsic dimension d, which may be unknown in many applications. In §4.5.3, we explain a data-driven way for choosing the sub-sampling size. In Propositions 4 and 5, we also prove that this method is guaranteed to achieve the optimal rates (depending on d) and therefore it is adaptive. Our simulations (see Figures 4.1 and 4.2) demonstrate that this adaptive algorithm works very well and provides valid finite-sample confidence intervals in high-dimensional, intrinsically low dimensional settings.

Finally, we conclude in §4.6 and defer a discussion on the extension to heterogeneous treatment effect estimation to Appendix C.2 and the technical proofs to Appendix C.3.

4.2 Preliminaries

Suppose we have a data set M of n observations $Z_1, Z_2, \ldots, Z_n$ drawn independently from some distribution $\mathcal{D}$ over the observation domain $\mathcal{Z}$. We focus on the case that $Z_i = (X_i, Y_i)$, where X_i is the vector of features and Y_i is the outcome. In Appendix C.2, we briefly discuss how our results can be extended to the setting where nuisance parameters and treatments are included in the model.

Suppose that the covariate space $\mathcal{X} \subset \mathbb{R}^D$ is contained in a ball with unknown diameter $\Delta_{\mathcal{X}}$. Denote the marginal distribution of X by μ and the empirical distribution of X on n sample points by μ_n. Let $B(x, r) = \left\{z \in \mathbb{R}^D : \|x - z\|_2 < r\right\}$ be the ℓ_2-ball centered at x with radius r and denote the standard basis for $\mathbb{R}^p$ by $\{e_1, e_2, \dots, e_p\}$. Finally, for any integer n, we let $[n] = \{1, 2, \dots, n\}$.

Let $\psi : \mathcal{Z} \times \mathbb{R}^p \to \mathbb{R}^p$ be a score function that maps observation Z and parameter $\theta \in \mathbb{R}^p$ to a p-dimensional score $\psi(Z; \theta)$. For $x \in \mathcal{X}$ and $\theta \in \mathbb{R}^p$ define the expected score as

$$m(x; \theta) = \mathbb{E}[\psi(Z; \theta) \mid X = x] \,.$$

The goal is to estimate the quantity $\theta(x)$ via local moment condition, i.e.

$$\theta(x) \text{ solves: } m(x; \theta) = \mathbb{E}[\psi(Z; \theta) \mid X = x] = 0.$$

4.2.1 Sub-Sampled Kernel Estimation

Base Kernel Learner. Our learner $\mathcal{L}_k$ takes a data set S containing m observations as input and a realization of internal randomness ω, and outputs a kernel weighting function $K_\omega : \mathcal{X} \times \mathcal{X} \times \mathcal{Z}^m \to [0, 1]$. In particular, given any target feature x and the set S, the weight of each observation Z_i in S with feature vector X_i is $K_\omega(x, X_i, S)$. Define the weighted score on a set S with internal randomness ω as $\Psi_S(x; \theta) = \sum_{i \in S} K_\omega(x, X_i, S)\psi(Z_i; \theta)$. When it is clear from context we will omit ω from our notation for succinctness and essentially treat K as a random function. For the rest of this chapter, we are going to use notations $\alpha_{S,\omega}(X_i) = K_\omega(x, X_i, S)$ interchangeably.

Averaging over B Sub-Samples of Size s. Suppose that we consider B random and independent draws from all $\binom{n}{s}$ possible subsets of size s and internal randomness variables ω and look at their average. Index these draws by $b = 1, 2, \dots, B$ where S_b contains samples in bth draw and ω_b is the corresponding draw of internal randomness. We can define the weighted score as

$$\Psi(x; \theta) = \frac{1}{B} \sum_{b=1}^{B} \Psi_{S_b, \omega_b}(x; \theta) = \frac{1}{B} \sum_{b=1}^{B} \sum_{i \in S_b} \alpha_{S_b, \omega_b}(X_i) \psi(Z_i; \theta) \,. \tag{4.2.1}$$

Estimating $\theta(x)$. We estimate $\theta(x)$ as a vanishing point of $\Psi(x;\theta)$. Letting $\hat{\theta}$ be this point, then $\Psi(x;\hat{\theta}) = \frac{1}{B}\sum_{b=1}^{B}\sum_{i=1}^{n}\alpha_{S_b,\omega_b}(X_i)\psi(Z_i;\hat{\theta}) = 0$. This procedure is explained in Algorithm 4.

4.2.2 Sub-Sampled k-NN Estimation

We specially focus on the case that the weights are distributed across the k-NN of x. In other words, given a data set S, the weights are given according to $K_\omega(x, X_i, S) = 1\{X_i \in H_k(x, S)\}/k$, where $H_k(x, S)$ are k-NN of x in the set S. The pseudo-code for this can be found in Algorithm 5.

Complete U-Statistic. The expression in Equation (4.2.1) is an incomplete U-statistic. Complete U-statistic is obtained if we allow each subset of size s from n samples to be included in the model exactly once. In other words, this is achieved if $B = \binom{n}{s}$, all subsets $S_1, S_2, \ldots, S_B$ are distinct, and we also take expectation over the internal randomness ω. Denoting this by $\Psi_0(x;\theta)$, we have

$$\Psi_0(x;\theta) = \binom{n}{s}^{-1}\sum_{S\in[n]:|S|=s}\mathbb{E}_\omega\left[\sum_{i\in S}\alpha_{S,\omega}(X_i)\psi(Z_i;\theta)\right]. \qquad (4.2.2)$$

Note in the case of k-NN estimator we can also represent Ψ_0 in terms of order statistics, i.e., Ψ_0 is an L-statistics (see, e.g., Serfling 2009). By sorting samples in $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$ based on their distance with x as $\|X_{(1)} - x\| \le \|X_{(2)} - x\| \le \cdots \le \|X_{(n)} - x\|$, we can write $\Psi_0(x;\theta) = \sum_{i=1}^{n}\alpha(X_{(i)})\,\psi(Z_{(i)};\theta)$ where the weights are given by

$$\alpha(X_{(i)}) = \begin{cases} \frac{1}{k}\binom{n}{s}^{-1}\binom{n-i}{s-1} & \text{if } i \le k \\ \frac{1}{k}\binom{n}{s}^{-1}\sum_{j=0}^{k-1}\binom{i-1}{j}\binom{n-i}{s-1-j} & \text{if } i \ge k+1\,. \end{cases}$$

Remark 4.2.1. *Note that both algorithms assume that the equation $\Psi(x;\theta) = 0$ is solvable. This has only been made for simplicity and can be replaced with milder assumptions. In fact, similar to Athey et al. (2019), we can allow for settings that $\Psi(x;\theta)$ is only approximately solvable. In such settings, we put an assumption on the existence of an optimization oracle* *that, given weights $\alpha(X_i)$, can solve for $\hat{\theta}$ satisfying*

$$\|\Psi(x;\theta)\|_2 \leq C_{oracle} \max_{i\in[n]} \alpha(X_i)\,,$$

where C_{oracle} is a constant. It is not hard to see that for sub-sampling with size s, if B is large enough, $\alpha(X_i) \leq 2s/n$ for all i with a very high probability. Therefore, the error due to this approximate solution of the optimization oracle is at most $O(s/n)$. We can allow for this amount of error in all our theorems as this is a lower order term compared to the variance which is roughly $O(\sqrt{s/n})$.

Algorithm 4 Sub-Sampled Kernel Estimation

```
Input. Data $\{Z_i = (X_i, Y_i)\}_{i=1}^n$, moment $\psi$, kernel $K$, sub-sampling size $s$, number of iterations $B$
Initialize. $\alpha(X_i) = 0, 1 \leq i \leq n$
for $b \leftarrow 1, B$ do
    Sub-sampling. Draw set $S_b$ by sampling $s$ points from $Z_1, Z_2, \ldots, Z_n$ without replacement.
    Weight Updates. $\alpha(X_i) \leftarrow \alpha(X_i) + K_{\omega_b}(x, X_i, S_b)$
end for
Weight Normalization. $\alpha(X_i) \leftarrow \alpha(X_i)/B$
Estimation. Denote $\hat{\theta}$ as a solution of $\Psi(x;\theta) = \sum_{i=1}^n \alpha(X_i)\psi(Z_i;\theta) = 0$
```

Algorithm 5 Sub-Sampled k-NN Estimation

```
Input. Data $\{Z_i = (X_i, Y_i)\}_{i=1}^n$, moment $\psi$, sub-sampling size $s$, number of iterations $B$, number of neighbors $k$
Initialize. $\alpha(X_i) \leftarrow 0, 1 \leq i \leq n$
for $b \leftarrow 1, B$ do
    Sub-sampling. Draw set $S_b$ by sampling $s$ points from $Z_1, Z_2, \ldots, Z_n$ without replacement
    Weight Updates. $\alpha(X_i) \leftarrow \alpha(X_i) + 1\{X_i \in H_k(x, S_b)\}/k$
end for
Weight Normalization. $\alpha(X_i) \leftarrow \alpha(X_i)/B$
Estimation. Denote $\hat{\theta}$ as a solution of $\Psi(x;\theta) = \sum_{i=1}^n \alpha(X_i)\psi(Z_i;\theta) = 0$
```

4.2.3 Local Intrinsic Dimension

We are interested in settings that the distribution of X has some low dimensional structure on a ball around the target point x. The following notions are adapted from Kpotufe (2011), which we present here for completeness.

Definition 3. *The marginal μ is called* **doubling measure** *if there exists a constant $C_{db} > 0$ such that for any $x \in \mathcal{X}$ and any $r > 0$ we have $\mu(B(x,r)) \leq C_{db}\mu(B(x,r/2))$.*

An equivalent definition of this notion is that, the measure μ is doubling measure if there exists $C, d > 0$ such that for any $x \in \mathcal{X}, r > 0$, and $\theta \in (0,1)$ we have $\mu(B(x,r)) \leq$

$C\theta^{-d}\mu(B(x, \theta r))$. In this definition, d acts as dimension.

A very simple example of doubling measure is Lebesgue measure on the Euclidean space $\mathbb{R}^d$. In particular, for any $r > 0, \theta \in (0, 1)$ we have $\mathrm{vol}(B(x, \theta r)) = \mathrm{vol}(B(x, r))\theta^d$. Building upon this, we can construct doubling probability measures on $\mathbb{R}^D$. Let $\mathcal{X} \in \mathbb{R}^D$ be a subset of d-dimensional hyperplane and suppose that for any ball $B(x, r)$ in $\mathcal{X}$ we have $\mathrm{vol}(B(x, r) \cap \mathcal{X}) = \Theta(r^d)$. If μ is approximately uniform, then we can translate this volume approximation to the probability measure μ. In fact, under this condition, we have $\mu(B(x, \theta r))/\mu(B(x, r)) = \Theta(\theta^d)$.

Unfortunately, the global notion of doubling dimension is very restrictive and many probability measures are globally complex. Rather, once restricted to local neighborhoods, the probability measure becomes lower dimensional and intrinsically less complex. The following definition captures this local notion of dimension more appropriately.

Definition 4. *Fix $x \in \mathcal{X}$ and $r > 0$. The marginal μ is (C, d)-**homogeneous on** $B(x, r)$ if for any $\theta \in (0, 1)$ we have $\mu(B(x, r)) \leq C\theta^{-d}\mu(B(x, \theta r))$.*

Intuitively, this definition requires the marginal μ to have a local support that is intrinsically d-dimensional. This definition covers low-dimensional manifolds, mixture distributions, d-sparse data, and also any combination of these examples.

4.2.4 Examples of Spaces with Small Intrinsic Dimension

In this section we provide examples of metric spaces that have small local intrinsic dimension. Our first example covers the setting where the distribution of data lies on a low-dimensional manifold (see, e.g., Roweis and Saul 2000, Tenenbaum et al. 2000, Belkin and Niyogi 2003). For instance, this happens for image inputs. Even though images are often high-dimensional (e.g., 4096 in the case of 64 by 64 images), all these images belong intrinsically to a 3-dimensional manifold.

Example 4.2.1 (Low-Dimensional Manifold (Adapted from Kpotufe (2011)))**.** *Consider a d-dimensional submanifold $\mathcal{X} \subset \mathbb{R}^D$ and let μ have lower and upper bounded density on $\mathcal{X}$. The local intrinsic dimension of μ on $B(x, r)$ is d, provided that r is chosen small enough and some conditions on curvature hold. In fact, Bishop-Gromov theorem (see, e.g., Carmo 1992) implies that under such conditions, the volume of ball $B(x, r) \cap \mathcal{X}$ is $\Theta(r^d)$. This together with the lower and upper bound on the density implies that $\mu(B(x, r) \cap \mathcal{X})/\mu(B(x, \theta r) \cap \mathcal{X}) = \Theta(\theta^d)$, i.e. μ is (C, d)-homogeneous on $B(x, r)$ for some $C > 0$.*

Another example which happens in many applications, is sparse data. For example, in the bag of words representation of text documents, we usually have a vocabulary consisting of D words. Although D is usually large, each text document contains only a small number of these words. In this application, we expect our data (and measure) to have smaller intrinsic dimension. Before stating this example, let us discuss a more general example about mixture distributions.

Example 4.2.2 (Mixture distributions (adapted from Kpotufe (2011)))**.** *Consider any mixture distribution $\mu = \sum_i \pi_i \mu_i$, with each μ_i defined on $\mathcal{X}$ with potentially different supports. Consider a point x and note that if $x \notin supp(\mu_i)$, then there exists a ball $B(x, r_i)$ such that $\mu_i(B(x, r_i)) = 0$. This is true since the support of any probability measure is always closed, meaning that its complement is an open set. Now suppose that r is chosen small enough such that for any i satisfying $x \in supp(\mu_i)$, μ_i is (C_i, d_i)-homogeneous on $B(x, r)$, while for any i satisfying $x \notin supp(\mu_i)$ we have $\mu_i(B(x, r)) = 0$. Then,*

$$\begin{aligned} \mu(B(x,r)) = & \sum_i \pi_i \mu_i(B(x,r)) = \sum_{i:\mu_i(B(x,r))=0} \pi_i \mu_i(B(x,r)) + \sum_{i:\mu_i(B(x,r))>0} \pi_i \mu_i(B(x,r)) \\ \leq & \ C\theta^{-d} \sum_{i:\mu_i(B(x,r))>0} \pi_i \mu_i(B(x,\theta r)) = \ C\theta^{-d} \sum_i \pi_i \mu_i(B(x,\theta r) = C\theta^{-d} \mu(B(x,\theta r))\,, \end{aligned}$$

where $C = \max_{i:\mu_i(B(x,r))>0} C_i$, $d = \max_{i:\mu_i(B(x,r))>0} d_i$, and we used the fact that if $\mu_i(B(x,r)) = 0$ then $\mu_i(B(x, \theta r)) = 0$. Therefore, μ is (C, d)-homogeneous on $B(x, r)$.

This result applies to the case of d-sparse data and is explained in the following example.

Example 4.2.3 (d-Sparse Data)**.** *Suppose that $\mathcal{X} \subset \mathbb{R}^D$ is defined as*

$$\mathcal{X} = \left\{ (x_1, x_2, \ldots, x_D) \in \mathbb{R}^D : \sum_{i=1}^{D} 1\left\{x_i \neq 0\right\} \leq d \right\}.$$

Let μ be a probability measure on $\mathcal{X}$. In this case, we can write $\mathcal{X}$ as the union of $k = \binom{D}{d}$, d-dimensonal hyperplanes in $\mathbb{R}^D$. In fact,

$$\mathcal{X} = \cup_{1 \leq i_1 < i_2 < \cdots i_d \leq D} \left\{ (x_1, x_2, \cdots, x_D) \in \mathbb{R}^D : x_j = 0,\ j \notin \{i_1, i_2, \ldots, i_d\} \right\}\ .$$

Letting $\mu_{i_1, i_2, \ldots, i_d}$ be the probability measure restricted to the hyperplane defined by $x_j = 0, j \notin \{i_1, i_2, \ldots, i_d\}$, we can express $\mu = \sum_{1 \leq i_1 < i_2 < \cdots i_d \leq D} \pi_{i_1, i_2, \ldots, i_d} \mu_{i_1, i_2, \ldots, i_d}$. Therefore, *the result of Example 4.2.2 implies that for any $x \in \mathcal{X}$, for r that is small enough μ is (C,d)-homogeneous on $B(x,r)$.*

Our final example is about the product measure. This allows us to prove that any concatenation of spaces with small intrinsic dimension has a small intrinsic dimension as well.

Example 4.2.4 (Concatenation under the Product Measure). *Suppose that μ_i is a probability measure on $\mathcal{X}_i \subset \mathbb{R}^{D_i}$, $i = 1,2$. Define $\mathcal{X} = \{(z_1, z_2) \mid z_1 \in \mathcal{X}_1, z_2 \in \mathcal{X}_2\}$ and let $\mu = \mu_1 \times \mu_2$ be the product measure on $\mathcal{X}$, i.e., $\mu(E_1 \times E_2) = \mu_1(E_1) \times \mu_2(E_2)$ for E_i that is μ_i-measurable, $i = 1,2$. Suppose that μ_i is (C_i, d_i)-homogeneous on $B(x_i, r_i)$ and let $x = (x_1, x_2)$. Then, μ is (C,d)-homogeneous on $B(x,r)$, where $d = d_1 + d_2, r = \min\{r_1, r_2\}$ and $C = (C_1\, C_2\, r^{-(d_1+d_2)}\, 2^{(d_1+d_2)/2})/(r_1^{-d_1}\, r_2^{-d_2})$. To establish this, let $r = \min\{r_1, r_2\}$ and note that for any $\theta \in (0,1)$ we have*

$$\begin{aligned}
\mu\left(B(x,r)\right) \leq\ & \mu\left(B(x_1,r) \times B(x_2,r)\right) = \mu_1\left(B(x_1,r)\right) \times \mu_2\left(B(x_2,r)\right) \\
& \leq \mu_1\left(B(x_1,r_1)\right) \times \mu_2\left(B(x_2,r_2)\right) \\
\leq\ & \left[C_1 \left(\frac{r\theta}{r_1\sqrt{2}}\right)^{-d_1} \mu_1\left(B\left(x_1, \frac{r\theta}{\sqrt{2}}\right)\right)\right] \times \left[C_2 \left(\frac{r\theta}{r_2\sqrt{2}}\right)^{-d_2} \mu_2\left(B\left(x_2, \frac{r\theta}{\sqrt{2}}\right)\right)\right] \\
=\ & \frac{C_1\, C_2\, r^{-(d_1+d_2)}}{r_1^{-d_1}\, r_2^{-d_2}\, \sqrt{2}^{-(d_1+d_2)}} \theta^{-d_1-d_2} \mu\left(B(x_1, r\theta/\sqrt{2}) \times B(x_2, r\theta/\sqrt{2})\right) \\
\leq\ & \frac{C_1\, C_2\, r^{-(d_1+d_2)\, 2^{(d_1+d_2)/2}}}{r_1^{-d_1}\, r_2^{-d_2}} \theta^{-(d_1+d_2)} \mu\left(B(x, r\theta)\right),
\end{aligned}$$

where we used two simple inequalities that $\|(z_1,z_2) - (x_1,x_2)\|_2 \leq r$ implies $\|z_i - x_i\|_2 \leq r$, $i = 1,2$, and further $\|z_i - x_i\|_2 \leq r/\sqrt{2}$, $i = 1,2$, implies $\|(z_1,z_2) - (x_1,x_2)\|_2 \leq r$.

4.3 Assumptions

The bias of non-parametric estimator is tightly connected to the kernel shrinkage, as noted by Athey et al. (2019), Wager and Athey (2018), Oprescu et al. (2018).

Definition 5 (Kernel Shrinkage in Expectation). *The kernel weighting function output by learner $\mathcal{L}_k$ when it is given s i.i.d. observations drawn from distribution $\mathcal{D}$ satisfies*

$$\mathbb{E}\left[\sup\left\{\|x - X_i\|_2 : K(x, X_i, S) > 0\right\}\right] = \epsilon(s)\,.$$

Definition 6 (Kernel Shrinkage with High Probability)**.** *The kernel weighting function output by learner $\mathcal{L}_k$ when it is given s i.i.d. observations drawn from distribution $\mathcal{D}$ w.p. $1-\delta$ over the draws of the s samples satisfies*

$$\sup\left\{\|x - X_i\|_2 : K(x, X_i, S) > 0\right\} \le \epsilon(s, \delta)\,.$$

As shown in Wager and Athey (2018), for trees that satisfy some regularity conditions, $\epsilon(s) \le s^{-c/D}$ for a constant c. We are interested in shrinkage rates that scale as $s^{-c/d}$, where d is the local intrinsic dimension of μ on $B(x,r)$. Similar to Oprescu et al. (2018), Athey et al. (2019), we rely on the following assumptions on the moment and score functions. We divide our assumptions into two parts. While the first part is sufficient for establishing estimation guarantees, for asymptotic normality results we require both.

Assumption 5.

1. *The moment $m(x;\theta)$ corresponds to the gradient w.r.t. θ of a λ-strongly convex loss $L(x;\theta)$. This also means that the Jacobian $M_0 = \nabla_\theta m(x;\theta(x))$ has minimum eigenvalue at least λ.*

2. *For any fixed parameters θ, $m(x;\theta)$ is a L_m-Lipschitz function in x for some constant L_m.*

3. *There exists a bound $\psi_{\max}$ such that for any observation z and any θ, $\|\psi(z;\theta)\|_\infty \le \psi_{\max}$.*

4. *The bracketing number $N_{[]}(\mathcal{F}, \epsilon, L_2)$ of the function class: $\mathcal{F} = \{\psi(\cdot;\theta) : \theta \in \Theta\}$, satisfies* $\log(N_{[]}(\mathcal{F}, \epsilon, L_2)) = O(1/\epsilon)$.

Assumption 6.

1. *For any coordinate j of the moment vector m, the Hessian $H_j(x;\theta) = \nabla^2_{\theta\theta} m_j(x;\theta)$ has eigenvalues bounded above by a constant L_H for all θ.*

2. *Maximum eigenvalue of M_0 is upper bounded by L_J.*

3. *Second moment of $\psi(x;\theta)$ defined as* $\mathrm{Var}\left(\psi(Z;\theta) \mid X = x\right)$ *is L_{mm}-Lipschitz in x, i.e.,*

$$\|\,\mathrm{Var}\left(\psi(Z;\theta) \mid X = x\right) - \mathrm{Var}\left(\psi(Z;\theta) \mid X = x'\right)\|_F \le L_{mm}\|x - x'\|_2\,.$$

4. Variogram is Lipschitz: $\sup_{x\in\mathcal{X}} \|\operatorname{Var}(\psi(Z;\theta) - \psi(Z;\theta') \mid X = x)\|_F \leq L_\psi \|\theta - \theta'\|_2$.

Note that our assumption on strong convexity of the moment $m(x;\theta)$ has been made to make the presentation easier. This assumption allows us to establish consistency and convergence rate together in a single analysis. However, once this assumption is removed, the analysis of consistency and establishing the rate of convergence is still feasible, but needs to be divided in two parts (see, e.g., Athey et al. 2019, Oprescu et al. 2018).

The condition on variogram always holds for a ψ that is Lipschitz in θ. This larger class of functions ψ allows estimation in more general settings such as α-quantile regression that involves a ψ which is non-Lipschitz in θ. Similar to Athey and Imbens (2016), Athey et al. (2019), we require kernel K to be *honest* and *symmetric.*

Assumption 7. *The kernel K, built using samples $\{Z_1, Z_2, \ldots, Z_s\}$, is* **honest** *if the weight of sample i given by $K(x, X_i, \{Z_j\}_{j=1}^s)$ is independent of Y_j conditional on X_j for any $j \in [s]$.*

Assumption 8. *The kernel K, built using samples $\{Z_1, Z_2, \ldots, Z_s\}$, is* **symmetric** *if for any permutation $\pi : [s] \to [s]$, the distribution of $K(x, X_i, \{Z_j\}_{j=1}^s)$ and $K(x, X_{\pi(i)}, \left\{Z_{\pi(j)}\right\}_{j=1}^s)$ are equal. In other words, the kernel weighting distribution remains unchanged under permutations.*

For a deterministic kernel K, the above condition implies that $K(x, X_i, \{Z_j\}_{j=1}^s) = K(x, X_i, \{Z_{\pi(j)}\}_{j=1}^s)$, for any $i \in [s]$. In the next section, we provide general estimation and inference results for a general kernel based on the its shrinkage and incrementality rates. Our estimation guarantees require kernel K to be honest (Theorem 8), while for asymptotic normality we also require K to be symmetric (Theorem 9).

4.4 Guarantees for Sub-Sampled Kernel Estimators

Our first result establishes estimation rates, both in expectation and high probability, for kernels based on their shrinkage rates. The proof of this theorem is deferred to Appendix C.3.

Theorem 8 (Finite Sample Estimation Rate)**.** *Let Assumptions 5 and 7 hold. Suppose that Algorithm 4 is executed with $B \geq n/s$. If the base kernel K satisfies kernel shrinkage in expectation, with rate $\epsilon(s)$, then w.p. $1-\delta$*

$$\|\hat{\theta} - \theta(x)\|_2 \leq \frac{2}{\lambda}\left(L_m \epsilon(s) + O\left(\psi_{\max}\sqrt{\frac{p\,s}{n}\left(\log\log(n/s) + \log(p/\delta)\right)}\right)\right). \qquad (4.4.1)$$

Moreover,

$$\sqrt{\mathbb{E}\left[\|\hat{\theta}-\theta(x)\|_2^2\right]} \leq \frac{2}{\lambda}\left(L_m\epsilon(s) + O\left(\psi_{\max}\sqrt{\frac{p\,s}{n}\log\log(p\,n/s)}\right)\right)\,. \tag{4.4.2}$$

The next result establishes asymptotic normality of sub-sampled kernel estimators. In particular, it provides coordinate-wise asymptotic normality of our estimate $\hat{\theta}$ around its true underlying value $\theta(x)$. The proof of this theorem is deferred to Appendix C.3.

Theorem 9 (Asymptotic Normality)**.** *Let Assumptions 5, 6, 7, and 8 hold. Suppose that Algorithm 4 is executed with $B \geq (n/s)^{5/4}$ and the base kernel K satisfies kernel shrinkage, with rate $\epsilon(s,\delta)$ in probability and $\epsilon(s)$ in expectation. Let $\eta(s)$ be the incrementality of kernel K defined in Equation* (4.1.4) *and s grow at a rate such that $s \to \infty$, $n\eta(s) \to \infty$, and $\epsilon(s,\eta(s)^2) \to 0$. Consider any fixed coefficient $\beta \in \mathbb{R}^p$ with $\|\beta\| \leq 1$ and define the variance as*

$$\sigma_{n,\beta}^2(x) = \frac{s^2}{n}\operatorname{Var}\left[\mathbb{E}\left[\sum_{i=1}^{s} K(x, X_i, \{Z_j\}_{j=1}^s)\left\langle \beta, M_0^{-1}\psi(Z_i;\theta(x))\right\rangle \mid Z_1\right]\right]\,.$$

Then it holds that $\sigma_{n,\beta}(x) = \Omega\left(s\sqrt{\eta(s)/n}\right)$. Moreover, suppose that

$$\max\left(\epsilon(s), \epsilon(s)^{1/4}\left(\frac{s}{n}\log\log(n/s)\right)^{1/2}, \left(\frac{s}{n}\log\log(n/s)\right)^{5/8}\right) = o(\sigma_{n,\beta}(x))\,. \tag{4.4.3}$$

Then,

$$\frac{\left\langle \beta, \hat{\theta}-\theta(x)\right\rangle}{\sigma_{n,\beta}(x)} \to_d \mathsf{N}(0,1)\,.$$

Remark 4.4.1. *Our notion of incrementality is slightly different from that of Wager and Athey (2018), as there the incrementality is defined as* $\operatorname{Var}\left[\mathbb{E}\left[K(x, X_1, \{Z_j\}_{j=1}^s) \mid X_1\right]\right]$. *However, using the tower law of expectations*

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{E}[K(x, X_1, \{Z_j\}_{j=1}^s) \mid X_1]^2\right] - \operatorname{Var}\left[\mathbb{E}[K(x, X_1, \{Z_j\}_{j=1}^s) \mid X_1]\right] \\
=\;& \mathbb{E}\left[\mathbb{E}\left[K(x, X_1, \{Z_j\}_{j=1}^s) \mid X_1\right]\right]^2 = \mathbb{E}\left[K(x, X_1, \{Z_j\}_{j=1}^s)\right]^2\,.
\end{aligned}$$

For a symmetric kernel the term $\mathbb{E}\left[K(x, X_1, \{Z_j\}_{j=1}^s)\right]^2$ is equal to $1/s^2$ and is asymptotically negligible compared to $\operatorname{Var}\left[\mathbb{E}\left[K(x, X_1, \{Z_j\}_{j=1}^s) \mid X_1\right]\right]$, *which usually decays at a slower rate.*

Theorems 8 and 9 generalize existing estimation and asymptotic normality results of Athey et al. (2019), Wager and Athey (2018), Fan et al. (2018) to an arbitrary kernel that satisfies appropriate shrinkage and incrementality rates. These general theorems could be of independent interest beyond the scope of this work. Following this approach, the main steps would be deriving shrinkage and incrementality rates of the kernel of interest. The following lemma relates these two and provides a lower bound on the incrementality in terms of kernel shrinkage. The proof uses the Paley-Zygmund inequality and is deferred to Appendix C.3.

Lemma 7. *For any symmetric kernel K (Assumption 8) and for any $\delta \in [0,1]$:*

$$\eta_s = \mathbb{E}\left[\mathbb{E}\left[K(x, X_1, \{Z_j\}_{j=1}^s) \mid X_1\right]^2\right] \geq \frac{(1-\delta)^2\ (1/s)^2}{\inf_{\rho>0}\left(\mu(B(x, \epsilon(s,\rho))) + \rho\, s/\delta\right)}\,.$$

Thus if $\mu(B(x, \epsilon(s, 1/(2s^2)))) = O(\log(s)/s)$, then picking $\rho = 1/(2s^2)$ and $\delta = 1/2$ implies that $\mathbb{E}[\mathbb{E}[K(x, X_1, \{Z_j\}_{j=1}^s)|X_1]^2] = \Omega(1/s\log(s))$.

This result has the following immediate corollary. The proof is left for Appendix C.3.

Corollary 3. *If $\epsilon(s,\delta) = O((\log(1/\delta)/s)^{1/d})$ and μ satisfies a two-sided version of the doubling measure property on $B(x,r)$, defined in Definition 4, i.e., the existence of two constants c and C such that $\mu(B(x,\theta r)) \geq C\theta^{-d}\mu(B(x,r))$ and $\mu(B(x,\theta r)) \leq c\theta^{-d}\mu(B(x,r))$, for any $\theta \in (0,1)$. Then, $\mathbb{E}[\mathbb{E}[K(x, X_1, \{Z_j\}_{j=1}^s)|X_1]^2] = \Omega(1/(s\log(s)))$.*

Even without this extra assumption, we can still characterize the incrementality rate of the k-NN estimator, as we observe in the next section.

4.5 Main Theorem: Adaptivity of the Sub-Sampled k-NN Estimator

In this section, we provide estimation guarantees and asymptotic normality of the k-NN estimator by using Theorems 8 and 9. We first establish shrinkage and incrementality rates for this kernel.

4.5.1 Estimation Guarantees for the Sub-Sampled k-NN Estimator

We start by providing shrinkage results for the k-NN kernel. As observed in Theorem 8, shrinkage rates are sufficient for bounding the estimation error. The shrinkage result that we present here would only depend on the local intrinsic dimension of μ on $B(x, r)$.

Lemma 8 (High Probability Shrinkage for the k-NN Kernel)**.** *Suppose that the measure μ is (C, d)-homogeneous on $B(x, r)$. Then, for any δ satisfying $2 \exp(-\mu(B(x, r))s/(8C)) \leq \delta \leq \frac{1}{2}\exp(-k/2)$, w.p. at least $1 - \delta$ we have*

$$\|x - X_{(k)}\|_2 \leq \epsilon_k(s, \delta) = O\left(\frac{\log(1/\delta)}{s}\right)^{1/d} .$$

We can easily turn this into a shrinkage rate in expectation. In fact, by the very convenient choice of $\delta = s^{-1/d}$ combined with the fact that $\mathcal{X}$ has diameter $\Delta_{\mathcal{X}}$, we can establish $O\left((\log(s)/s)^{1/d}\right)$ rate on expected kernel shrinkage. However, a more careful analysis would help us to remove the $\log(s)$ dependency in the bound and is stated in the following corollary:

Corollary 4 (Expected Shrinkage for the k-NN Kernel)**.** *Suppose that the conditions of Lemma 8 hold. Let k be a constant and $\epsilon_k(s)$ be the expected shrinkage for the k-NN kernel. Then, for any s larger than some constant we have $\epsilon_k(s) = \mathbb{E}\left[\|x - X_{(k)}\|_2\right] = O\left(\frac{1}{s}\right)^{1/d}$.*

We are now ready to state our estimation result for the k-NN kernel, which is honest and symmetric. Therefore, we can substitute the expected shrinkage rate, established in Corollary 4, in Theorem 8 to derive estimation rates for this kernel.

Theorem 10 (Estimation Guarantees for the k-NN Kernel)**.** *Suppose that μ is (C, d)-homogeneous on $B(x, r)$, Assumption 5 holds and that Algorithm 5 is executed with $B \geq n/s$. Then, w.p. $1 - \delta$:*

$$\|\hat{\theta} - \theta(x)\|_2 \leq \frac{2}{\lambda}\left(O\left(s^{-1/d}\right) + O\left(\psi_{\max}\sqrt{\frac{p\, s}{n}\left(\log\log(n/s) + \log(p/\delta)\right)}\right)\right) , \tag{4.5.1}$$

and

$$\sqrt{\mathbb{E}\left[\|\hat{\theta} - \theta(x)\|_2^2\right]} \leq \frac{2}{\lambda}\left(O\left(s^{-1/d}\right) + O\left(\psi_{\max}\sqrt{\frac{p\, s\, \log\log(p\, n/s)}{n}}\right)\right) . \tag{4.5.2}$$

By picking $s = \Theta\left(n^{d/(d+2)}\right)$ and $B = \Omega\left(n^{2/(d+2)}\right)$ we get $\sqrt{\mathbb{E}\left[\|\hat{\theta} - \theta(x)\|_2^2\right]} = \tilde{O}\left(n^{-1/(d+2)}\right)$.

4.5.2 Asymptotic Normality of the Sub-sampled k-NN Estimator

In this section we prove asymptotic normality of k-NN estimator. We start by provide bounds on the incrementality of the k-NN kernel.

Lemma 9 (k-NN Incrementality)**.** *Let K be the k-NN kernel and let $\eta_k(s)$ denote the incrementality rate of this kernel. Then, the following holds:*

$$\eta_k(s) = \mathbb{E}\left[\mathbb{E}\left[K(x, X_1, \{Z_j\}_{j=1}^s) \mid X_1\right]^2\right] = \frac{1}{(2s-1)\,k^2}\left(\sum_{t=0}^{2k-2}\frac{a_t}{b_t}\right),$$

where sequences $\{a_t\}_{t=0}^{2k-2}$ and $\{b_t\}_{t=0}^{2k-2}$ are defined as

$$a_t = \sum_{i=\max\{0,t-(k-1)\}}^{\min\{t,k-1\}}\binom{s-1}{i}\binom{s-1}{t-i} \qquad \text{and} \qquad b_t = \sum_{i=0}^{t}\binom{s-1}{i}\binom{s-1}{t-i}.$$

Remark 4.5.1. *Note that $b_t = \binom{2s-2}{t}$ since we can view b_t as follows: how many different subsets of size t can we create from a set of $2s-2$ elements if we pick a number $i = \{0,\dots,t\}$ and then choose i elements from the first half of these elements and $t-i$ elements from the second half. This process creates all possible sets of size t from among the $2s-2$ elements, which is equal to $\binom{2s-2}{t}$.*

Furthermore, for $0 \le t \le k-1$, $a_t = b_t$ and for any $k \le t \le 2k-2$, after some algebra, we have

$$\frac{2k-1-t}{t+1} \le \frac{a_t}{b_t} \le 1\,.$$

This implies that the summation appeared in Lemma 34 satisfies

$$k + \sum_{t=k}^{2k-2}\frac{2k-1-t}{t+1} \le \sum_{t=0}^{2k-2}\frac{a_t}{b_t} \le 2k-1\,.$$

Note that the above remark implies that the summation a_t/b_t that appeared on Lemma 9 always belongs to the interval $[k, 2k-1]$, and therefore it is known up to a factor 2. As we observe later, the same term would appear on the characterization of asymptotic variance of the k-NN estimator. The following lemma shows that when $s \to \infty$, we can exactly characterize this summation up to a lower order term of $1/s$.

Lemma 10. *Suppose that $s \to \infty$ and k is fixed. Then*

$$\sum_{t=0}^{2k-2} \frac{a_t}{b_t} = \zeta_k + O(1/s)\,,$$

where $\zeta_k = k + \sum_{t=k}^{2k-2} 2^{-t} \sum_{i=t-k+1}^{k-1} \binom{t}{i}$.

We can substitute $\eta_k(s)$ in Theorem 9 to prove asymptotic normality of the k-NN estimator. Before doing that, we establish the asymptotic variance of this estimator, i.e. $\sigma_{n,j}(x)$, up to the smaller order terms. The proof of this Lemma is deferred to Appendix C.3.

Theorem 11 (Asymptotic Variance of the Sub-Sampled k-NN Estimator)**.** *Let $j \in [p]$ be one of coordinates. Suppose that k is constant while $s \to \infty$. Then, for the k-NN kernel*

$$\sigma^2_{n,j}(x) = \frac{s^2}{n} \frac{\sigma_j^2(x)}{k^2\,(2s-1)}\,\zeta_k + o(s/n)\,, \tag{4.5.3}$$

where $\sigma_j^2(x) = \mathrm{Var}\left[\left\langle e_j, M_0^{-1}\psi(Z;\theta(x))\right\rangle \mid X = x\right]$ and $\zeta_k = k + \sum_{t=k}^{2k-2} 2^{-t} \sum_{i=t-k+1}^{k-1} \binom{t}{i}$.

Combining results of Theorem 9, Theorem 11, Corollary 4, and Lemma 9 we have:

Theorem 12 (Asymptotic Normality of the Sub-Sampled k-NN Estimator)**.** *Suppose that μ is (C,d)-homogeneous on $B(x,r)$. Let Assumptions 5, 6 hold and suppose that Algorithm 5 is executed with $B \geq (n/s)^{5/4}$ iterations. Suppose that s grows at a rate such that $s \to \infty$, $n/s \to \infty$, and also $s^{-1/d}(n/s)^{1/2} \to 0$. Let $j \in [p]$ be one of coordinates and $\sigma^2_{n,j}(x)$ be defined in Equation (4.5.3). Then,*

$$\frac{\hat{\theta}_j(x) - \theta_j(x)}{\sigma_{n,j}(x)} \to \mathsf{N}(0,1)\,.$$

Finally, if $s = n^{\beta}$ and $B \geq n^{\frac{5}{4}(1-\beta)}$ with $\beta \in (d/(d+2), 1)$. Then,

$$\frac{\hat{\theta}_j(x) - \theta_j(x)}{\sigma_{n,j}(x)} \to \mathsf{N}(0,1)\,.$$

Plug-In Confidence Intervals. Observe that the Theorem 11 implies that if we define $\tilde{\sigma}^2_{n,j}(x) = \frac{s^2}{n}\frac{\sigma_j^2(x)}{2s-1}\frac{\zeta_k}{k^2}$ as the leading term in the variance, then $\frac{\sigma^2_{n,j}(x)}{\tilde{\sigma}^2_{n,j}(x)} \to_p 1$. Thus, due to

Slutsky's theorem

$$\frac{\hat{\theta}_j - \theta_j}{\tilde{\sigma}^2_{n,j}(x)} = \frac{\hat{\theta}_j - \theta_j}{\sigma^2_{n,j}(x)} \frac{\sigma^2_{n,j}(x)}{\tilde{\sigma}^2_{n,j}(x)} \to_d \mathsf{N}(0,1)\,. \tag{4.5.4}$$

Hence, we have a closed form solution to the variance in our asymptotic normality theorem. If we have an estimate $\hat{\sigma}^2_j(x)$ of the variance of the conditional moment around x, then we can build plug-in confidence intervals based on the normal distribution with variance $\frac{s^2}{n}\frac{\hat{\sigma}^2_j(x)}{2s-1}\frac{\zeta_k}{k^2}$. Note that ζ_k can be calculated easily for desired values of k. For instance, we have $\zeta_1 = 1, \zeta_2 = \frac{5}{2}, \zeta_3 = \frac{33}{8}$, and for $k = 1,2,3$ the asymptotic variance becomes $\frac{s^2}{n}\frac{\hat{\sigma}^2_j(x)}{2s-1}, \frac{5}{8}\frac{s^2}{n}\frac{\hat{\sigma}^2_j(x)}{2s-1}, \frac{11}{24}\frac{s^2}{n}\frac{\hat{\sigma}^2_j(x)}{2s-1}$ respectively.

4.5.3 Adaptive Sub-Sample Size Selection

According to Theorem 10, picking $s = \Theta(n^{d/(d+2)})$ would trade-off between bias and variance terms. Also, according to Theorem 12, picking $s = n^\beta$ with $d/(d+2) < \beta < 1$ would result in asymptotic normality of the estimator. However, both choices depend on the unknown intrinsic dimension of μ on the ball $B(x,r)$. Inspired by Kpotufe (2011), we explain a data-driven way for choosing s.

Adaptive Selection for s. Suppose that $\delta > 0$ is given. Let $C_{n,p,\delta} = 2\log(2pn/\delta)$ and pick $\Delta \geq \Delta_{\mathcal{X}}$. For any $k \leq s \leq n$, let $H(s)$ be the U-statistic estimator for $\epsilon(s)$ defined as $H(s) = \sum_{S \in [n]:|S|=s} \max_{X_i \in H_k(x,S)} \|x - X_i\|_2 / \binom{n}{s}$. Each term in the summation computes the distance of x to its k-nearest neighbor on S and $H(s)$ is the average of these numbers over all $\binom{n}{s}$ possible subsets S. Define $G_\delta(s) = \Delta\sqrt{C_{n,p,\delta}ps/n}$. Iterate over $s = n, \cdots, k$. Let s_2 be the smallest s for which we have $H(s) > 2G_\delta(s)$ and let $s_1 = s_2 + 1$. Note that $\epsilon_k(s)$ is decreasing in s and $G_\delta(s)$ is increasing in s. Therefore, there exists a unique $1 \leq s^* \leq n$ such that $\epsilon_k(s^*) \leq G_\delta(s^*)$ and $\epsilon_k(s^*-1) > G_\delta(s^*-1)$. We have the following Lemma.

Lemma 11. *Consider the selection process described above and let s_1 be its output. Then, w.p. $1-\delta$ we have*

$$\frac{s^*-1}{9} \leq s_1 \leq s^*\,.$$

Having this lemma in hand, it is now easy to provide adaptive estimation and asymptotic normality results. In particular, choosing $s_* = 9s_1 + 1$ ensures that $s_* \in [s^*, 10s^*]$ and therefore we are able to trade-off nicely between the bias $\epsilon_k(s)$ and variance (which is roughly at the same order as $G_\delta(s)$). Note that constant numbers would only increase our bounds by constant factors. Furthermore, choosing $s_* = (9s_1 + 1)n^{\zeta}$ for any $\zeta > 0$ ensures that we fall in the region where $\epsilon_k(s) = o(\sigma_{n,j}(x))$ and therefore it leads to the asymptotic normality of this estimator.

Proposition 4 (Adaptive Estimation)**.** *Let Assumptions of Theorem 10 hold. Suppose that s_1 is the output of the above process. Let $s_* = 9s_1 + 1$ and suppose that Algorithm 5 is executed with $s = s_*$ and $B \geq n/s_*$. Then w.p. at least $1 - 2\delta$ we have*

$$\|\hat{\theta} - \theta(x)\|_2 = O(G_\delta(s^*)) = O\left(\left(\frac{n}{p\log(2pn/\delta)}\right)^{-1/(d+2)}\right).$$

Furthermore, for $\delta = 1/n$ we have

$$\sqrt{\mathbb{E}\left[\|\hat{\theta} - \theta(x)\|_2^2\right]} = \tilde{O}\left(n^{-1/(d+2)}\right).$$

Proposition 5 (Adaptive Asymptotic Normality)**.** *Let Assumptions of Theorem 12 hold. Suppose that s_1 is the output of the above process when $\delta = 1/n$ and $s_* = 9s_1 + 1$. For any $\zeta \in (0, (\log(n) - \log(s_1) - \log\log^2(n))/\log(n)))$ define $s_\zeta = s_* n^{\zeta}$. Suppose that Algorithm 5 is executed with $s = s_\zeta$ and $B \geq (n/s_\zeta)^{5/4}$, then for any coordinate $j \in [p]$, we have*

$$\frac{\hat{\theta}_j(x) - \theta_j(x)}{\sigma_{n,j}(x)} \to \mathsf{N}(0, 1).$$

Remark 4.5.2. *Note that although computation of $H(s)$ may look complex as it involves the calculation of distance of x to its k-nearest neighbor on all $\binom{n}{s}$ subsets, there is a closed form expression for $H(s)$ according to its representation based on L-statistic. In fact, by sorting samples $(X_1, X_2, \ldots, X_n)$ based on their distance to x, i.e, $\|x - X_{(1)}\|_2 \leq \|x - X_{(2)}\|_2 \leq \ldots \leq \|x - X_{(n)}\|_2$, we have*

$$H(s) = \binom{n}{s}^{-1} \sum_{i=k}^{n-s+k} \binom{i-1}{k-1}\binom{n-i}{s-k} \|x - X_{(i)}\|_2.$$

Therefore, after sorting all training samples based on their distance with x, we can compute values of $H(s)$ very efficient and fast.

4.6 Conclusions and Discussions

In this chapter, we studied estimation and inference for conditional moment equations in the presence of high-dimensional conditioning variable which has a low intrinsic dimension locally. We proved that by combining sub-sampling techniques and non-parametric estimators, we can achieve both estimation accuracy and also asymptotic normality which is crucial for building confidence intervals and drawing inference about quantities of interest. In particular, letting D and d be the extrinsic and intrinsic dimension respectively, we proved that finely tuned sub-sampled k-NN estimators are able to adapt to unknown intrinsic dimension of the problem and provide $\tilde{O}(n^{-1/(d+2)})$ estimation accuracy and also are $n^{1/(d+2)}$-asymptotically normal.

Our results shed some light on the importance of using adaptive machine learning based estimators, such as nearest neighbor based estimates, when performing estimation and inference in high-dimensional settings. Such estimators address the curse of dimensionality by adapting to a priori unknown latent structure in the data. Moreover, coupled with the powerful sub-sampling based averaging approach, such estimators can maintain their adaptivity, while also satisfying asymptotic normality and thereby enabling asymptotically valid inference; a property that is crucial for embracing such approaches in econometrics and causal inference.

Bibliography

Abadie, A., A. Diamond, and J. Hainmueller (2010). 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.

Abadie, A., A. Diamond, and J. Hainmueller (2015). Comparative politics and the synthetic control method. *American Journal of Political Science*, 495–510.

Abadie, A. and J. Gardeazabal (2003). The economic costs of conflict: A case study of the basque country. *American Economic Review 93*(-), 113–132.

Abbasi-Yadkori, Y., D. Pál, and C. Szepesvári (2011). Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pp. 2312–2320.

Agrawal, S. and N. Goyal (2013). Thompson sampling for contextual bandits with linear payoffs. In *ICML*, pp. 127–135.

Ai, C. and X. Chen (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica 71*(6), 1795–1843.

Anderson, T. W. (1958). *An introduction to multivariate statistical analysis*, Volume 2. Wiley New York.

Andoni, A., P. Indyk, and I. Razenshteyn (2018). Approximate nearest neighbor search in high dimensions. *arXiv preprint arXiv:1806.09823*.

Andoni, A., T. Laarhoven, I. Razenshteyn, and E. Waingarten (2017). Optimal hashing-based time-space trade-offs for approximate near neighbors. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 47–66. Society for Industrial and Applied Mathematics.

Arellano, M. and B. Honoré (2001). Panel data models: some recent developments. *Handbook of econometrics 5*, 3229–3296.

Assouad, P. (1983). Plongements lipschitziens dans $\mathbb{R}^n$. *Bull. Soc. Math. France 111*, 429–448.

Athey, S., M. Bayati, N. Doudchenko, G. Imbens, and K. Khosravi (2018). Matrix completion methods for causal panel data models. Technical report, National Bureau of Economic Research.

Athey, S. and G. Imbens (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences 113*(27), 7353–7360.

Athey, S. and G. W. Imbens (2017). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives 31*(2), 3–32.

Athey, S. and G. W. Imbens (2018). Design-based analysis in difference-in-differences settings with staggered adoption. Technical report, National Bureau of Economic Research.

Athey, S., G. W. Imbens, and S. Wager (2018). Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 80*(4), 597–623.

Athey, S. and S. Stern (2002). The impact of information technology on emergency health care outcomes. *The RAND Journal of Economics 33*(3), 399–432.

Athey, S., J. Tibshirani, and S. Wager (2019). Generalized random forests. *The Annals of Statistics 47*(2), 1148–1178.

Auer, P. (2003). Using confidence bounds for exploitation-exploration trade-offs. *JMLR 3*, 397–422.

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica 71*(1), 135–171.

Bai, J. (2009). Panel data models with interactive fixed effects. *Econometrica 77*(4), 1229–1279.

Bai, J. and S. Ng (2002). Determining the number of factors in approximate factor models. *Econometrica 70*(1), 191–221.

Ban, G.-Y. and N. B. Keskin (2018). Personalized dynamic pricing with machine learning. *Available at SSRN 2972985*.

Ban, G.-Y. and C. Rudin (2014). The big data newsvendor: Practical insights from machine learning. *Working Paper*.

Ban, G.-Y. and C. Rudin (2018). The big data newsvendor: Practical insights from machine learning. *Operations Research 67*(1), 90–108.

Bastani, H. and M. Bayati (2015). Online decision-making with high-dimensional covariates. `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2661896`.

Bastani, H., M. Bayati, and K. Khosravi (2017). Mostly exploration-free algorithms for contextual bandits. *arXiv preprint arXiv:1704.09011*.

Belkin, M. and P. Niyogi (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation 15*(6), 1373–1396.

Bellman, R. (1961). Adaptive control processes princeton. *Press, Princeton, NJ*.

Belloni, A., V. Chernozhukov, I. Fernández-Val, and C. Hansen (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica 85*(1), 233–298.

Belloni, A., V. Chernozhukov, and C. Hansen (2014a). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives 28*(2), 29–50.

Belloni, A., V. Chernozhukov, and C. Hansen (2014b). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies 81*(2), 608–650.

Berrett, T. B., R. J. Samworth, M. Yuan, et al. (2019). Efficient multivariate entropy estimation via k-nearest neighbour distances. *The Annals of Statistics 47*(1), 288–318.

Bertsimas, D. and N. Kallus (2014a). From predictive to prescriptive analytics. *arXiv preprint arXiv:1402.5481*.

Bertsimas, D. and N. Kallus (2014b). From predictive to prescriptive analytics. *arXiv preprint arXiv:1402.5481*.

Biau, G. (2012, April). Analysis of a random forests model. *J. Mach. Learn. Res. 13*(1), 1063–1095.

Biau, G. and L. Devroye (2015). *Lectures on the nearest neighbor method*. Springer.

Bietti, A., A. Agarwal, and J. Langford (2018). A Contextual Bandit Bake-off. *ArXiv e-prints*.

Billingsley, P. (2008). *Probability and measure*. John Wiley & Sons.

Bird, S., S. Barocas, K. Crawford, F. Diaz, and H. Wallach (2016). Exploring or Exploiting? Social and Ethical Implications of Autonomous Experimentation in AI.

Borovkov, A. (2013). *Probability Theory*. Springer London.

Breiman, L. (2001). Random forests. *Machine learning 45*(1), 5–32.

Broder, J. and P. Rusmevichientong (2012, July). Dynamic pricing under a general parametric choice model. *Oper. Res. 60*(4), 965–980.

Bubeck, S. and N. Cesa-Bianchi (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning 5*(1), 1–122.

Bühlmann, P. and S. Van De Geer (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Candès, E. J. and Y. Plan (2010). Matrix completion with noise. *Proceedings of the IEEE 98*(6), 925–936.

Candès, E. J. and B. Recht (2009). Exact matrix completion via convex optimization. *Foundations of Computational mathematics 9*(6), 717.

Candès, E. J. and T. Tao (2010). The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Inf. Theor. 56*(5), 2053–2080.

Carmo, M. P. d. (1992). *Riemannian geometry*. Birkhäuser.

Chamberlain, G. (1984). Panel data. *Handbook of econometrics 2*, 1247–1318.

Chen, G. H. and D. Shah (2018). Explaining the success of nearest neighbor methods in prediction. *Foundations and Trends® in Machine Learning 10*(5-6), 337–588.

Chen, K., I. Hu, and Z. Ying (1999). Strong consistency of maximum quasi-likelihood estimators in generalized linear models with fixed and adaptive designs. *The Annals of Statistics 27*(4), 1155–1163.

Chen, X., Z. Owen, C. Pixton, and D. Simchi-Levi (2015). A statistical learning approach to personalization in revenue management. *Available at SSRN 2579462*.

Chen, X. and D. Pouzo (2009). Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals. *Journal of Econometrics 152*(1), 46 – 60. Recent Adavances in Nonparametric and Semiparametric Econometrics: A Volume Honouring Peter M. Robinson.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and R. James (2016). Double machine learning for treatment and causal parameters. *arXiv preprint arXiv:1608.00060*.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal 21*(1), C1–C68.

Chernozhukov, V., J. C. Escanciano, H. Ichimura, W. K. Newey, and J. M. Robins (2016, July). Locally Robust Semiparametric Estimation. *arXiv e-prints*, arXiv:1608.00033.

Chernozhukov, V., D. Nekipelov, V. Semenova, and V. Syrgkanis (2018). Plug-in regularized estimation of high-dimensional parameters in nonlinear semiparametric models. *arXiv preprint arXiv:1806.04823*.

Chernozhukov, V., W. K. Newey, and A. Santos (2015, September). Constrained Conditional Moment Restriction Models. *arXiv e-prints*, arXiv:1509.06311.

Chu, W., L. Li, L. Reyzin, and R. E. Schapire (2011). Contextual bandits with linear payoff functions. In *AISTATS*, pp. 208–214.

Cohen, M. C., I. Lobel, and R. Paes Leme (2016). Feature-based dynamic pricing. `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2737045`.

Consortium, I. W. P. (2009). Estimation of the warfarin dose with clinical and pharmacogenetic data. *NEJM 360*(8), 753.

Cutler, C. D. (1993). A review of the theory and estimation of fractal dimension. In *Dimension estimation and models*, pp. 1–107. World Scientific.

Dani, V., T. P. Hayes, and S. M. Kakade (2008). Stochastic linear optimization under bandit feedback. pp. 355–366.

Dasgupta, S. and Y. Freund (2008). Random projection trees and low dimensional manifolds. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pp. 537–546. ACM.

den Boer, A. (2015). Tracking the market: Dynamic pricing and learning in a changing environment. *European Journal of Operational Research 247*(3), 914 – 927.

den Boer, A. V. and B. Zwart (2013). Simultaneously learning and optimizing using controlled variance pricing. *Management Science 60*(3), 770–783.

den Boer, A. V. and B. Zwart (2015). Dynamic Pricing and Learning with Finite Inventories. *Operations Research 63*(4), 965–978.

Doudchenko, N. and G. W. Imbens (2016). Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. Technical report, National Bureau of Economic Research.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans.* SIAM.

Fan, J., M. Farmen, and I. Gijbels (1998). Local maximum likelihood estimation and inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 60*(3), 591–608.

Fan, Y., J. Lv, and J. Wang (2018). Dnn: A two-scale distributional tale of heterogeneous treatment effect inference. *arXiv preprint arXiv:1808.08469*.

Filippi, S., O. Cappe, A. Garivier, and C. Szepesvári (2010). Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems*, pp. 586–594.

Friedberg, R., J. Tibshirani, S. Athey, and S. Wager (2018). Local linear forests. *arXiv preprint arXiv:1807.11408*.

Friedman, J., T. Hastie, and R. Tibshirani (2001). *The elements of statistical learning*, Volume 1. Springer Series in Statistics New York, NY, USA:.

Gittins, J. C. (1979). Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society: Series B (Methodological) 41*(2), 148–164.

Gobillon, L. and T. Magnac (2013). Regional policy evaluation: Interactive fixed effects and synthetic controls. *Review of Economics and Statistics* (00).

Goldberger, A. S. (1972). Structural equation methods in the social sciences. *Econometrica: Journal of the Econometric Society*, 979–1001.

Goldenshluger, A. and A. Zeevi (2009). Woodroofe's one-armed bandit problem revisited. *The Annals of Applied Probability 19*(4), 1603–1633.

Goldenshluger, A. and A. Zeevi (2013). A linear response bandit problem. *Stochastic Systems 3*(1), 230–261.

Gross, D. (2011). Recovering low-rank matrices from few coefficients in any basis. *IEEE Trans. Information Theory 57*(3), 1548–1566.

Gutin, E. and V. Farias (2016). Optimistic gittins indices. In *Advances in Neural Information Processing Systems*, pp. 3153–3161.

Györfi, L., M. Kohler, A. Krzyzak, and H. Walk (2006). *A distribution-free theory of nonparametric regression.* Springer Science & Business Media.

Hanasusanto, G. A. and D. Kuhn (2013). Robust data-driven dynamic programming. In *Advances in Neural Information Processing Systems*, pp. 827–835.

Hannah, L., W. Powell, and D. M. Blei (2010). Nonparametric density estimation for stochastic optimization with an observable state variable. In *Advances in Neural Information Processing Systems*, pp. 820–828.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, 1029–1054.

Hoeffding, W. (1994). Probability inequalities for sums of bounded random variables. In *The Collected Works of Wassily Hoeffding*, pp. 409–426. Springer.

Hsiao, C., H. Steve Ching, and S. Ki Wan (2012). A panel data approach for program evaluation: measuring the benefits of political and economic integration of hong kong with mainland china. *Journal of Applied Econometrics 27*(5), 705–740.

Imbens, G. W. and D. B. Rubin (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Javanmard, A. and H. Nazerzadeh (2019). Dynamic pricing in high-dimensions. *The Journal of Machine Learning Research 20*(1), 315–363.

Jiang, H. (2017). Rates of uniform consistency for k-nn regression. *arXiv preprint arXiv:1707.06261*.

Kallus, N. (2016). Learning to personalize from observational data. *arXiv preprint arXiv:1608.08925*.

Kannan, S., J. Morgenstern, A. Roth, B. Waggoner, and Z. S. Wu (2018). A Smoothed Analysis of the Greedy Algorithm for the Linear Contextual Bandit Problem. *ArXiv e-prints*.

Kazerouni, A., M. Ghavamzadeh, Y. Abbasi-Yadkori, and B. V. Roy (2016). Conservative contextual linear bandits. `https://arxiv.org/abs/1611.06426`.

Keshavan, R. H., A. Montanari, and S. Oh (2010a, June). Matrix completion from a few entries. *IEEE Trans. Inf. Theor. 56*(6), 2980–2998.

Keshavan, R. H., A. Montanari, and S. Oh (2010b, August). Matrix completion from noisy entries. *J. Mach. Learn. Res. 11*, 2057–2078.

Keskin, N. B. and A. Zeevi (2014a). Chasing demand: Learning and earning in a changing environment. `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2389750`.

Keskin, N. B. and A. Zeevi (2014b). Dynamic pricing with an unknown demand model: Asymptotically optimal semi-myopic policies. *Operations Research 62*(5), 1142–1167.

Keskin, N. B. and A. Zeevi (2015). On incomplete learning and certainty-equivalence control. *preprint*.

Khosravi, K., G. Lewis, and V. Syrgkanis (2019). Non-parametric inference adaptive to intrinsic dimension. *arXiv preprint arXiv:1901.03719*.

Kim, D. and T. Oka (2014). Divorce law reforms and divorce rates in the usa: An interactive fixed-effects approach. *Journal of Applied Econometrics 29*(2), 231–245.

Kim, E. S., R. S. Herbst, I. I. Wistuba, J. J. Lee, G. R. Blumenschein, A. Tsao, D. J. Stewart, M. E.

Hicks, J. Erasmus, S. Gupta, et al. (2011). The battle trial: personalizing therapy for lung cancer. *Cancer discovery 1*(1), 44–53.

Kim, J., J. Shin, A. Rinaldo, and L. Wasserman (2018, October). Uniform Convergence Rate of the Kernel Density Estimator Adaptive to Intrinsic Dimension. *arXiv e-prints*, arXiv:1810.05935.

Klopp, O. (2014). Noisy low-rank matrix completion with general sampling distribution. *Bernoulli 20*(1), 282–303.

Kohavi, R. and S. H. Thomke (2017). The surprising power of online experiments.

Koltchinskii, V., K. Lounici, A. B. Tsybakov, et al. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics 39*(5), 2302–2329.

Kpotufe, S. (2011). k-nn regression adapts to local intrinsic dimension. In *Advances in Neural Information Processing Systems*, pp. 729–737.

Kpotufe, S. and S. Dasgupta (2012). A tree-based regressor that adapts to intrinsic dimension. *Journal of Computer and System Sciences 78*(5), 1496–1515.

Kpotufe, S. and V. Garg (2013). Adaptivity to local smoothness and dimension in kernel regression. In *Advances in neural information processing systems*, pp. 3075–3083.

Lafferty, J. and L. Wasserman (2008, 02). Rodeo: Sparse, greedy nonparametric regression. *The Annals of Statistics 36*(1), 28–63.

Lai, T. L. and H. Robbins (1985). Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics 6*(1), 4–22.

Langford, J. and T. Zhang (2008). The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in Neural Information Processing Systems*, pp. 817–824.

Lattimore, T. and R. Munos (2014). Bounded regret for finite-armed structured bandits. In *Advances in Neural Information Processing Systems*, pp. 550–558.

Lehmann, E. and G. Casella (1998). *Theory of Point Estimation.* Springer Verlag.

Lewbel, A. (2007). A local generalized method of moments estimator. *Economics Letters 94*(1), 124–128.

Li, L., W. Chu, J. Langford, and R. E. Schapire (2010). A contextual-bandit approach to personalized news article recommendation. In *WWW*, pp. 661–670.

Li, L., Y. Lu, and D. Zhou (2017). Provably optimal algorithms for generalized linear contextual bandits. *arXiv preprint arXiv:1703.00048*.

Liang, K.-Y. and S. L. Zeger (1986). Longitudinal data analysis using generalized linear models. *Biometrika 73*(1), 13–22.

Luca, M. (2014). Were okcupid's and facebook's experiments unethical? *Harvard Business Review Blog Network. http://blogs. hbr. org/2014/07/were-okcupids-and-facebooks-experiments-unethical (visited October 19, 2014).*

Mack, Y.-P. (1981). Local properties of k-nn regression estimates. *SIAM Journal on Algebraic Discrete Methods 2*(3), 311–323.

Mackey, L., V. Syrgkanis, and I. Zadik (2017). Orthogonal machine learning: Power and limitations. *arXiv preprint arXiv:1711.00342*.

Mazumder, R., T. Hastie, and R. Tibshirani (2010). Spectral regularization algorithms for learning large incomplete matrices. *Journal of machine learning research 11*(Aug), 2287–2322.

McCullagh, P. and J. A. Nelder (1989). *Generalized linear models (Second edition)*. London: Chapman & Hall.

Mersereau, A. J., P. Rusmevichientong, and J. N. Tsitsiklis (2009). A structured multiarmed bandit problem and the greedy policy. *IEEE Transactions on Automatic Control 54*(12), 2787–2802.

Moon, H. R. and M. Weidner (2015). Linear regression for panel with unknown number of factors as interactive fixed effects. *Econometrica 83*(4), 1543–1579.

Moon, H. R. and M. Weidner (2017). Dynamic linear panel regression models with interactive fixed effects. *Econometric Theory 33*(1), 158–195.

Mullainathan, S. and J. Spiess (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives 31*(2), 87–106.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability & Its Applications 9*(1), 141–142.

Narendra, K. S. and A. M. Annaswamy (1987). Persistent excitation in adaptive systems. *International Journal of Control 45*(1), 127–160.

Negahban, S. and M. J. Wainwright (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, 1069–1097.

Negahban, S. and M. J. Wainwright (2012). Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *Journal of Machine Learning Research 13*(May), 1665–1697.

Negahban, S. N., P. Ravikumar, M. J. Wainwright, and B. Yu (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science 27*(4), 538–557.

Newey, W. K. (1993). 16 efficient estimation of models with conditional moment restrictions. In *Econometrics*, Volume 11 of *Handbook of Statistics*, pp. 419 – 454. Elsevier.

Newey, W. K. (1994). Kernel estimation of partial means and a general variance estimator. *Econometric Theory 10*(2), 1–21.

Nguyen, N. T. (2018). *Model-reference adaptive control*. Springer.

Oprescu, M., V. Syrgkanis, and Z. S. Wu (2018). Orthogonal random forest for heterogeneous treatment effect estimation. *arXiv preprint arXiv:1806.03467*.

Peel, T., S. Anthoine, and L. Ralaivola (2010). Empirical bernstein inequalities for u-statistics. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta (Eds.), *Advances in Neural Information Processing Systems 23*, pp. 1903–1911. Curran Associates, Inc.

Pesaran, M. H. (2006). Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica 74*(4), 967–1012.

Qiang, S. and M. Bayati (2016). Dynamic pricing with demand covariates. `https://arxiv.org/abs/1604.07463`.

Recht, B. (2011). A simpler approach to matrix completion. *Journal of Machine Learning Research 12*(Dec), 3413–3430.

Reiss, P. C. and F. A. Wolak (2007). Structural econometric modeling: Rationales and examples from industrial organization. *Handbook of econometrics 6*, 4277–4415.

Robins, J. M. and Y. Ritov (1997). Toward a curse of dimensionality appropriate (coda) asymptotic theory for semi-parametric models. *Statistics in medicine 16*(3), 285–319.

Rohde, A., A. B. Tsybakov, et al. (2011). Estimation of high-dimensional low-rank matrices. *The Annals of Statistics 39*(2), 887–930.

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*(1), 41–55.

Roweis, S. T. and L. K. Saul (2000). Nonlinear dimensionality reduction by locally linear embedding. *science 290*(5500), 2323–2326.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology 66*(5), 688–701.

Russo, D. (2019). A note on the equivalence of upper confidence bounds and gittins indices for patient agents. *arXiv preprint arXiv:1904.04732*.

Russo, D. and B. Van Roy (2014a). Learning to optimize via information-directed sampling. In *Advances in Neural Information Processing Systems*, pp. 1583–1591.

Russo, D. and B. Van Roy (2014b). Learning to optimize via posterior sampling. *Mathematics of Operations Research 39*(4), 1221–1243.

Samworth, R. J. (2012). Optimal weighted nearest neighbour classifiers. *The Annals of Statistics 40*(5), 2733–2763.

Sarkar, J. (1991). One-armed bandit problems with covariates. *The Annals of Statistics*, 1978–2002.

Scornet, E., G. Biau, and J.-P. Vert (2015, 08). Consistency of random forests. *The Annals of Statistics 43*(4), 1716–1741.

Serfling, R. J. (2009). *Approximation theorems of mathematical statistics*, Volume 162. John Wiley & Sons.

Sibbald, B. and M. Roland (1998). Understanding controlled trials. why are randomised controlled trials important? *BMJ: British Medical Journal 316*(7126), 201.

Srebro, N., N. Alon, and T. S. Jaakkola (2005). Generalization error bounds for collaborative prediction with low-rank matrices. In L. K. Saul, Y. Weiss, and L. Bottou (Eds.), *Advances in Neural Information Processing Systems 17*, pp. 1321–1328.

Staniswalis, J. G. (1989). The kernel estimate of a regression function in likelihood-based models. *Journal of the American Statistical Association 84*(405), 276–283.

Stone, C. J. (1977). Consistent nonparametric regression. *The Annals of Statistics*, 595–620.

Stone, C. J. (1982, 12). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics 10*(4), 1040–1053.

Tenenbaum, J. B., V. De Silva, and J. C. Langford (2000). A global geometric framework for nonlinear dimensionality reduction. *Science 290*(5500), 2319–2323.

Thompson, W. R. (1933). On the Likelihood that one Unknown Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika 25*, 285–294.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Tibshirani, R. and T. Hastie (1987). Local likelihood estimation. *Journal of the American Statistical Association 82*(398), 559–567.

Tropp, J. A. (2011). User-friendly tail bounds for matrix martingales. Technical report, DTIC Document.

Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics 12*(4), 389–434.

Tsybakov, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, 135–166.

Verma, N., S. Kpotufe, and S. Dasgupta (2009). Which spatial partition trees are adaptive to intrinsic dimension? In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, pp. 565–574. AUAI Press.

Wager, S. and S. Athey (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association 113*(523), 1228–1242.

Wainwright, M. (2016). *High-dimensional statistics: A non-asymptotic viewpoint.* Working Publication.

Wang, C.-C., S. R. Kulkarni, and H. V. Poor (2005a). Arbitrary side observations in bandit problems. *Advances in Applied Mathematics 34*(4), 903 – 938.

Wang, C.-C., S. R. Kulkarni, and H. V. Poor (2005b). Bandit problems with side observations. *IEEE Transactions on Automatic Control 50*(3), 338–355.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, 359–372.

Weed, J., V. Perchet, and P. Rigollet (2015). Online learning in repeated auctions. `https://arxiv.org/abs/1511.05720`.

Woodroofe, M. (1979). A one-armed bandit problem with a concomitant variable. *Journal of the American Statistical Association 74*(368), 799–806.

Wu, Y., R. Shariff, T. Lattimore, and C. Szepesvari (2016). Conservative bandits. In *Proceedings of The 33rd International Conference on Machine Learning*, Volume 48, pp. 1254–1262. PMLR.

Xu, Y. (2017). Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis 25*(1), 57–76.

Xue, L. and S. Kpotufe (2018). Achieving the time of 1-nn, but the accuracy of k-nn. In A. Storkey and F. Perez-Cruz (Eds.), *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, Volume 84 of *Proceedings of Machine Learning Research*, pp. 1628–1636.

Appendix A

Supplementary Materials for Chapter 2

A.1 Properties of Covariate Diversity

Proof of Lemma 1. Since for all $\mathbf{u} \in \mathbb{R}^d$ at least one of $\mathbf{x}^\top \mathbf{u} \geq 0$ or $-\mathbf{x}^\top \mathbf{u} \geq 0$ holds, and using conditions (a), (b), and (c) of Lemma 1 we have:

$$\begin{aligned}
\int \mathbf{x}\mathbf{x}^\top \mathbb{I}(\mathbf{x}^\top \mathbf{u} \geq 0) p_X(\mathbf{x}) \mathrm{d}\mathbf{x} &\succeq \int_W \mathbf{x}\mathbf{x}^\top \mathbb{I}(\mathbf{x}^\top \mathbf{u} \geq 0) p_X(\mathbf{x}) \mathrm{d}\mathbf{x} \\
&= \frac{1}{2} \int_W \mathbf{x}\mathbf{x}^\top \Big[\mathbb{I}(\mathbf{x}^\top \mathbf{u} \geq 0) p_X(\mathbf{x}) + \mathbb{I}(-\mathbf{x}^\top \mathbf{u} \geq 0) p_X(-\mathbf{x})\Big] \mathrm{d}\mathbf{x} \\
&\succeq \frac{1}{2} \int_W \mathbf{x}\mathbf{x}^\top \Big[\mathbb{I}(\mathbf{x}^\top \mathbf{u} \geq 0) + \frac{a}{b} \mathbb{I}(\mathbf{x}^\top \mathbf{u} \leq 0)\Big] p_X(\mathbf{x}) \mathrm{d}\mathbf{x} \\
&\succeq \frac{a}{2b} \int_W \mathbf{x}\mathbf{x}^\top p_X(\mathbf{x}) \mathrm{d}\mathbf{x} \\
&\succeq \frac{a\lambda}{2b} I_d \,.
\end{aligned}$$

Here, the first inequality follows from the fact that $\mathbf{x}\mathbf{x}^\top$ is positive semi-definite, the first equality follows from condition (a) and a change of variable ($\mathbf{x} \to -\mathbf{x}$), the second inequality is by condition (b), the third inequality uses $a \leq b$ which follows from condition (b), and the last inequality uses condition (c). □

We now state the proofs of lemmas that were used in §2.2.2.

Proof of Lemma 2. First note that B_R^d is symmetric with respect to each axis, therefore the off-diagonal entries in $\int_{B_R^d} \mathbf{x}\mathbf{x}^\top \mathrm{d}\mathbf{x}$ are zero. In particular, the (i,j) entry of the integral is equal to $\int_{B_R^d} x_i x_j \mathrm{d}\mathbf{x}$ which is zero when $i \neq j$ using a change of variable $x_i \to -x_i$ that has the identity as its Jacobian and keeps the domain of integral unchanged but changes the sign of $x_i x_j$. Also, by symmetry, all diagonal entry terms are equal. In other words,

$$\int_{B_R^d} \mathbf{x}\mathbf{x}^\top \mathrm{d}\mathbf{x} = \left(\int_{B_R^d} x_1^2 \mathrm{d}\mathbf{x} \right) I_d \,. \tag{A.1.1}$$

Now for computing the right hand side integral, we introduce the spherical coordinate system as

$$\begin{aligned} x_1 &= r\cos\theta_1, \\ x_2 &= r\sin\theta_1\cos\theta_2, \\ &\vdots \\ x_{d-1} &= r\sin\theta_1\sin\theta_2\ldots\sin\theta_{d-2}\cos\theta_{d-1}, \\ x_d &= r\sin\theta_1\sin\theta_2\ldots\sin\theta_{d-2}\sin\theta_{d-1}, \end{aligned}$$

and the determinant of its Jacobian is given by

$$\det J(r,\boldsymbol{\theta}) = \det\left[\frac{\partial \mathbf{x}}{\partial r \partial \boldsymbol{\theta}}\right] = r^{d-1}\sin^{d-2}\theta_1\sin^{d-3}\theta_2\ldots\sin\theta_{d-2}.$$

Now, using symmetry, and summing up equation (A.1.1) with x_i^2 used instead of x_1^2 for all $i \in [d]$, we obtain

$$\begin{aligned} d\int_{B_R^d} \mathbf{x}\mathbf{x}^\top \mathrm{d}\mathbf{x} &= \int_{B_R^d} \left(x_1^2 + x_2^2 + \ldots + x_d^2\right) \mathrm{d}x_1 \mathrm{d}x_2 \ldots \mathrm{d}x_d \\ &= \int_{\theta_1,\ldots,\theta_{d-1}} \int_{r=0}^{R} r^{d+1}\sin^{d-2}\theta_1\sin^{d-3}\theta_2\ldots\sin\theta_{d-2}\,\mathrm{d}r\,\mathrm{d}\theta_1\ldots\mathrm{d}\theta_{d-1}\,. \end{aligned}$$

Comparing this to

$$\mathrm{vol}(B_R^d) = \int_{\theta_1,\ldots,\theta_{d-1}} \int_{r=0}^{R} r^{d-1}\sin^{d-2}\theta_1\sin^{d-3}\theta_2\ldots\sin\theta_{d-2}\,\mathrm{d}r\,\mathrm{d}\theta_1\ldots\mathrm{d}\theta_{d-1}\,,$$

we obtain that

$$\int_{B_R^d} \mathbf{x}\mathbf{x}^\top \mathrm{d}\mathbf{x} = \left[\frac{\int_0^R r^{d+1}\mathrm{d}r}{d\int_0^R r^{d-1}\mathrm{d}r}\mathrm{vol}(B_R^d)\right] I_d$$
$$= \left[\frac{R^2}{d+2}\mathrm{vol}(B_R^d)\right] I_d\,.$$

□

Proof of Lemma 12. We can lower-bound the density $p_{X,\mathrm{trunc}}$ by the uniform density as follows. Note that we have $\mathbf{x}^\top\Sigma^{-1}\mathbf{x} \leq \|\mathbf{x}\|_2^2\lambda_{\max}\left(\Sigma^{-1}\right)$ and as a result for any $\mathbf{x}$ satisfying $\|\mathbf{x}\|_2 \leq x_{\max}$ we have

$$p_{X,\mathrm{trunc}}(\mathbf{x}) \geq p_X(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{d/2}}\exp\left(-\frac{1}{2}\mathbf{x}^\top\Sigma^{-1}\mathbf{x}\right) \geq \frac{\exp\left(-\frac{x_{\max}^2}{2\lambda_{\min}(\Sigma)}\right)}{(2\pi)^{d/2}|\Sigma|^{d/2}} = p_{X,\text{uniform-lb}}\,.$$

Using this we can derive a lower bound on the desired covariance as following

$$\int_{B_{x_{\max}}^d} \mathbf{x}\mathbf{x}^\top p_{X,\mathrm{trunc}}(\mathbf{x})\mathrm{d}\mathbf{x} \succeq \int_{B_{x_{\max}}^d} \mathbf{x}\mathbf{x}^\top p_{X,\text{uniform-lb}}(\mathbf{x})\mathrm{d}\mathbf{x}$$
$$= \frac{1}{(2\pi)^{d/2}|\Sigma|^{d/2}}\exp\left(-\frac{x_{\max}^2}{2\lambda_{\min}(\Sigma)}\right)\int_{B_{x_{\max}}^d}\mathbf{x}\mathbf{x}^\top\mathrm{d}\mathbf{x}$$
$$= \frac{1}{(2\pi)^{d/2}|\Sigma|^{d/2}}\exp\left(-\frac{x_{\max}^2}{2\lambda_{\min}(\Sigma)}\right)\frac{x_{\max}^2}{d+2}\mathrm{vol}(B_{x_{\max}}^d)I_d$$
$$= \lambda_{\mathrm{uni}}I_d\,,$$

where we used Lemma 2 in the third line. This concludes the proof. □

Lemma 12. *The following inequality holds*

$$\int_{B_{x_{\max}}^d} \mathbf{x}\mathbf{x}^\top p_{X,trunc}(\mathbf{x})\mathrm{d}\mathbf{x} \succeq \lambda_{uni}\mathbf{I}_d\,,$$

where $\lambda_{uni} \equiv \frac{1}{(2\pi)^{d/2}|\Sigma|^{d/2}}\exp\left(-\frac{x_{\max}^2}{2\lambda_{\min}(\Sigma)}\right)\frac{x_{\max}^2}{d+2}\mathrm{vol}(B_{x_{\max}}^d)$.

Proof. We can lower-bound the density $p_{X,\mathrm{trunc}}$ by the uniform density as follows. Note that we have $\mathbf{x}^\top\Sigma^{-1}\mathbf{x} \leq \|\mathbf{x}\|_2^2\lambda_{\max}\left(\Sigma^{-1}\right)$ and as a result for any $\mathbf{x}$ satisfying $\|\mathbf{x}\|_2 \leq x_{\max}$

we have

$$p_{X,\text{trunc}}(\mathbf{x}) \geq p_X(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{d/2}} \exp\left(-\frac{1}{2}\mathbf{x}^\top \Sigma^{-1}\mathbf{x}\right) \geq \frac{\exp\left(-\frac{x_{\max}^2}{2\lambda_{\min}(\Sigma)}\right)}{(2\pi)^{d/2}|\Sigma|^{d/2}} = p_{X,\text{uniform-lb}}\,.$$

Using this we can derive a lower bound on the desired covariance as following

$$\begin{aligned}
\int_{B^d_{x_{\max}}} \mathbf{x}\mathbf{x}^\top p_{X,\text{trunc}}(\mathbf{x})\mathrm{d}\mathbf{x} &\succeq \int_{B^d_{x_{\max}}} \mathbf{x}\mathbf{x}^\top p_{X,\text{uniform-lb}}(\mathbf{x})\mathrm{d}\mathbf{x} \\
&= \frac{1}{(2\pi)^{d/2}|\Sigma|^{d/2}} \exp\left(-\frac{x_{\max}^2}{2\lambda_{\min}(\Sigma)}\right) \int_{B^d_{x_{\max}}} \mathbf{x}\mathbf{x}^\top \mathrm{d}\mathbf{x} \\
&= \frac{1}{(2\pi)^{d/2}|\Sigma|^{d/2}} \exp\left(-\frac{x_{\max}^2}{2\lambda_{\min}(\Sigma)}\right) \frac{x_{\max}^2}{d+2}\mathrm{vol}(B^d_{x_{\max}})I_d \\
&= \lambda_{\text{uni}} I_d\,,
\end{aligned}$$

where we used Lemma 2 in the third line. This concludes the proof. □

A.2 Useful Concentration Results

Lemma 13 (Bernstein Concentration). *Let $\{D_k, \mathcal{H}_k\}_{k=1}^\infty$ be a martingale difference sequence, and let D_k be σ_k-subgaussian. Then, for all $t > 0$ we have*

$$\mathbb{P}\left[\Big|\sum_{k=1}^n D_k\Big| \geq t\right] \leq 2\exp\left\{-\frac{t^2}{2\sum_{k=1}^n \sigma_k^2}\right\}.$$

Proof. See Theorem 2.3 of Wainwright (2016) and let $b_k = 0$ and $\nu_k = \sigma_k$ for all k. □

Lemma 14 (Theorem 3.1 of Tropp (2011)). *Let $\mathcal{H}_1 \subset \mathcal{H}_2 \subset \cdots$ be a filtration and consider a finite adapted sequence $\{X_k\}$ of positive semi-definite matrices with dimension d, adapted to this filtration. Suppose that $\lambda_{\max}(X_k) \leq R$ almost surely. Define the series $Y \equiv \sum_k X_k$ and $W \equiv \sum_k \mathbb{E}[X_k \mid \mathcal{H}_{k-1}]$. Then for all $\mu \geq 0, \gamma \in [0,1)$ we have:*

$$\mathbb{P}\left[\lambda_{\min}(Y) \leq (1-\gamma)\mu \quad \text{and} \quad \lambda_{\min}(W) \geq \mu\right] \leq d\left(\frac{e^{-\gamma}}{(1-\gamma)^{1-\gamma}}\right)^{\mu/R}.$$

A.3 Proof of Regret Guarantees for Greedy Bandit

We first prove a lemma on the instantaneous regret of the Greedy Bandit using a standard peeling argument. The proof here is adapted from Bastani and Bayati (2015) with a few modifications; we present it here for completeness.

Notation. We define the following events to simplify notation. For any $\lambda, \chi > 0$, let

$$\mathcal{F}_{i,t}^{\lambda} = \left\{ \lambda_{\min}\left(\mathbf{X}(\mathcal{S}_{i,t})^\top \mathbf{X}(\mathcal{S}_{i,t}) \right) \geq \lambda t \right\} \tag{A.3.1}$$

$$\mathcal{G}_{i,t}^{\chi} = \left\{ \|\hat{\beta}(\mathcal{S}_{i,t}) - \beta_i\|_2 < \chi \right\} . \tag{A.3.2}$$

Proof of Lemma 6. We can decompose the regret as $r_t(\pi) = \mathbb{E}[\text{Regret}_t(\pi)] = \sum_{i=1}^{K} \mathbb{E}[\text{Regret}_t(\pi) \mid X_t \in \mathcal{R}_i] \cdot \mathbb{P}(X_t \in \mathcal{R}_i)$. Now we can expand each term as

$$\mathbb{E}[\text{Regret}_t(\pi) \mid X_t \in \mathcal{R}_l] = \mathbb{E}\left[X_t^\top (\beta_l - \beta_{\pi_t}) \mid X_t \in \mathcal{R}_l \right] ,$$

For each $1 \leq i, l \leq K$ satisfying $i \neq l$, let us define the region where arm i is superior over arm l

$$\hat{\mathcal{R}}_{i \geq l, t} := \left\{ \mathbf{x} \in \mathcal{X} : \mathbf{x}^\top \hat{\beta}(\mathcal{S}_{i,t-1}) \geq \mathbf{x}^\top \hat{\beta}(\mathcal{S}_{l,t-1}) \right\} ,$$

Note that we may incur a nonzero regret if $X_t^\top \hat{\beta}(\mathcal{S}_{\pi_t,t-1}) > X_t^\top \hat{\beta}(\mathcal{S}_{l,t-1})$ or if $X_t^\top \hat{\beta}(\mathcal{S}_{\pi_t,t-1}) = X_t^\top \hat{\beta}(\mathcal{S}_{l,t-1})$ and the tie-breaking random variable W_t indicates an action other than l as the action to be taken. It is worth mentioning that in the case $X_t^\top \hat{\beta}(\mathcal{S}_{\pi_t,t-1}) = X_t^\top \hat{\beta}(\mathcal{S}_{l,t-1})$ we do not incur any regret if W_t indicates arm l as the action to be taken. Nevertheless, as regret is a non-negative quantity, we can write

$$
\begin{aligned}
\mathbb{E}[\text{Regret}_t(\pi)\ |X_t \in \mathcal{R}_l] &\le \mathbb{E}\left[\mathbb{I}(X_t^\top \hat{\beta}(\mathcal{S}_{\pi_t,t-1}) \ge X_t^\top \hat{\beta}(\mathcal{S}_{l,t-1}))X_t^\top(\beta_l - \beta_{\pi_t}) \mid X_t \in \mathcal{R}_l\right] \\
&\le \sum_{i \ne l} \mathbb{E}\left[\mathbb{I}(X_t^\top \hat{\beta}(\mathcal{S}_{i,t-1}) \ge X_t^\top \hat{\beta}(\mathcal{S}_{l,t-1}))X_t^\top(\beta_l - \beta_i) \mid X_t \in \mathcal{R}_l\right] \\
&= \sum_{i \ne l} \mathbb{E}\left[\mathbb{I}(X_t \in \hat{\mathcal{R}}_{i \ge l,t})X_t^\top(\beta_l - \beta_i) \mid X_t \in \mathcal{R}_l\right] \\
&\le \sum_{i \ne l} \left\{ \mathbb{E}\left[\mathbb{I}(\hat{\mathcal{R}}_{i \ge l,t}, \mathcal{F}^{\lambda_0/4}_{l,t-1}, \mathcal{F}^{\lambda_0/4}_{i,t-1})X_t^\top(\beta_l - \beta_i) \mid X_t \in \mathcal{R}_l\right] \right. \\
&\quad + \mathbb{E}\left[\mathbb{I}(X_t \in \hat{\mathcal{R}}_{i \ge l,t}, \overline{\mathcal{F}^{\lambda_0/4}_{l,t-1}})X_t^\top(\beta_l - \beta_i) \mid X_t \in \mathcal{R}_l\right] \\
&\quad \left. + \mathbb{E}\left[\mathbb{I}(X_t \in \hat{\mathcal{R}}_{i \ge l,t}, \overline{\mathcal{F}^{\lambda_0/4}_{i,t-1}})X_t^\top(\beta_l - \beta_i) \mid X_t \in \mathcal{R}_l\right] \right\}
\end{aligned}
$$

where in the second line we used a union bound. Using the fact that $\mathcal{F}^{\lambda_0/4}_{i,t-1}$ and $\mathcal{F}^{\lambda_0/4}_{l,t-1}$ are independent of the event $X_t \in \mathcal{R}_l$ which only depends on X_t, together with the Cauchy-Schwarz inequality implying $X_t^\top(\beta_l - \beta_i) \le 2b_{\max}x_{\max}$, we have

$$
\begin{aligned}
\mathbb{E}[\text{Regret}_t(\pi) \mid X_t \in \mathcal{R}_l] \le & \sum_{i \ne l} \left\{ \mathbb{E}\left[\mathbb{I}(X_t \in \hat{\mathcal{R}}_{i \ge l,t}, \mathcal{F}^{\lambda_0/4}_{l,t-1}, \mathcal{F}^{\lambda_0/4}_{i,t-1})X_t^\top(\beta_l - \beta_i) \mid X_t \in \mathcal{R}_l\right] \right. \\
& \quad \left. + 2b_{\max}x_{\max}\left(\mathbb{P}(\overline{\mathcal{F}^{\lambda_0/4}_{l,t-1}}) + \mathbb{P}(\overline{\mathcal{F}^{\lambda_0/4}_{i,t-1}})\right) \right\} \\
& \sum_{i \ne l} \mathbb{E}\left[\mathbb{I}(X_t \in \hat{\mathcal{R}}_{i \ge l,t}, \mathcal{F}^{\lambda_0/4}_{l,t-1}, \mathcal{F}^{\lambda_0/4}_{i,t-1})X_t^\top(\beta_l - \beta_i) \mid X_t \in \mathcal{R}_l\right] \\
& \quad + 4(K-1)b_{\max}x_{\max} \max_i \mathbb{P}(\overline{\mathcal{F}^{\lambda_0/4}_{i,t-1}})\,. \qquad \text{(A.3.3)}
\end{aligned}
$$

Hence, we need to bound the first term in above. Fix i and note that when we include events $\mathcal{F}^{\lambda_0/4}_{i,t-1}$ and $\mathcal{F}^{\lambda_0/4}_{l,t-1}$, we can use Lemma 5 which proves sharp concentrations for $\hat{\beta}(\mathcal{S}_{l,t-1})$ and $\hat{\beta}(\mathcal{S}_{i,t-1})$. Let us now define the following set

$$
I^h = \{\mathbf{x} \in \mathcal{X} : \mathbf{x}^\top(\beta_l - \beta_i) \in (2\delta x_{\max}h, 2\delta x_{\max}(h+1)]\},
$$

where $\delta = 1/\sqrt{(t-1)}$. Note that since $X_t^\top(\beta_l - \beta_i)$ is bounded above by $2b_{\max}x_{\max}$, the set I^h only needs to be defined for $h \le h^{\max} = \lceil b_{\max}/\delta \rceil$. We can now expand the first term in Equation (A.3.3) for i, by conditioning on $X_t \in I^h$ as following

$$
\begin{aligned}
&\mathbb{E}\left[\mathbb{I}(X_t \in \hat{\mathcal{R}}_{i\geq l,t}, \mathcal{F}^{\lambda_0/4}_{l,t-1}, \mathcal{F}^{\lambda_0/4}_{i,t-1}) X_t^\top(\beta_l - \beta_i) \mid X_t \in \mathcal{R}_l\right] \\
&= \sum_{h=0}^{h^{\max}} \mathbb{E}\left[\mathbb{I}(X_t \in \hat{\mathcal{R}}_{i\geq l,t}, \mathcal{F}^{\lambda_0/4}_{l,t-1}, \mathcal{F}^{\lambda_0/4}_{i,t-1}) X_t^\top(\beta_l - \beta_i) \mid X_t \in \mathcal{R}_l \cap I_h\right] \mathbb{P}[X_t \in I^h] \\
&\leq \sum_{h=0}^{h^{\max}} 2\delta x_{\max}(h+1) \mathbb{E}\left[\mathbb{I}(X_t \in \hat{\mathcal{R}}_{i\geq l,t}, \mathcal{F}^{\lambda_0/4}_{l,t-1}, \mathcal{F}^{\lambda_0/4}_{i,t-1}) \mid X_t \in \mathcal{R}_l \cap I_h\right] \mathbb{P}[X_t \in I^h] \\
&\leq \sum_{h=0}^{h^{\max}} 2\delta x_{\max}(h+1) \mathbb{E}\left[\mathbb{I}(X_t \in \hat{\mathcal{R}}_{i\geq l,t}, \mathcal{F}^{\lambda_0/4}_{l,t-1}, \mathcal{F}^{\lambda_0/4}_{i,t-1}) \mid X_t \in \mathcal{R}_l \cap I_h\right] \\
&\qquad \times \mathbb{P}[X_t^\top(\beta_l - \beta_i) \in (0, 2\delta x_{\max}(h+1)]] \\
&\leq \sum_{h=0}^{h^{\max}} 4C_0\delta^2 x_{\max}^2 (h+1)^2 \mathbb{P}\left[X_t \in \hat{\mathcal{R}}_{i\geq l,t}, \mathcal{F}^{\lambda_0/4}_{l,t-1}, \mathcal{F}^{\lambda_0/4}_{i,t-1} \mid X_t \in \mathcal{R}_l \cap I_h\right],
\end{aligned} \tag{A.3.4}
$$

where in the first inequality we used the fact that conditioning on $X_t \in I^h$, $X_t^\top(\beta_l - \beta_i)$ is bounded above by $2\delta x_{\max}(h+1)$, in the second inequality we used the fact that the event $X_t \in I^h$ is a subset of the event $X_t^\top(\beta_l - \beta_i) \in (0, 2\delta x_{\max}(h+1)]$, and in the last inequality we used the margin condition given in Assumption 2. Now we reach to the final part of the proof, where conditioning on $\mathcal{F}^{\lambda_0/4}_{l,t-1}, \mathcal{F}^{\lambda_0/4}_{i,t-1}$, and $X_t \in I^h$ we want to bound the probability that we pull a wrong arm. Note that conditioning on $X_t \in I^h$, the event $X_t^\top\left(\hat{\beta}(\mathcal{S}_{i,t-1}) - \hat{\beta}(\mathcal{S}_{l,t-1})\right) \geq 0$ happens only when at least one of the following two events: i) $X_t^\top(\beta_l - \hat{\beta}(\mathcal{S}_{l,t-1})) \geq \delta x_{\max} h$ or ii) $X_t^\top(\hat{\beta}(\mathcal{S}_{i,t-1}) - \beta_i) \geq \delta x_{\max} h$ happens. This is true according to

$$
\begin{aligned}
0 &\leq X_t^\top\left(\hat{\beta}(\mathcal{S}_{i,t-1}) - \hat{\beta}(\mathcal{S}_{l,t-1})\right) \\
&= X_t^\top(\hat{\beta}(\mathcal{S}_{i,t-1}) - \beta_i) + X_t^\top(\beta_i - \beta_l) + X_t^\top(\beta_l - \hat{\beta}(\mathcal{S}_{l,t-1})) \\
&\leq X_t^\top(\hat{\beta}(\mathcal{S}_{i,t-1}) - \beta_i) - 2\delta x_{\max} h + X_t^\top(\beta_l - \hat{\beta}(\mathcal{S}_{l,t-1}))\,.
\end{aligned}
$$

Therefore,

$$\begin{aligned}
&\mathbb{P}\left[\mathbb{I}(X_t \in \hat{\mathcal{R}}_{i\geq l,t}, \mathcal{F}^{\lambda_0/4}_{l,t-1}, \mathcal{F}^{\lambda_0/4}_{i,t-1}) \mid X_t \in \mathcal{R}_l \cap I^h\right] \\
&\quad \leq \mathbb{P}\left[X_t^\top(\beta_l - \hat{\beta}(\mathcal{S}_{l,t-1})) \geq \delta x_{\max} h, \mathcal{F}^{\lambda_0/4}_{l,t-1}, \mathcal{F}^{\lambda_0/4}_{i,t-1} \mid X_t \in \mathcal{R}_l \cap I^h\right] \\
&\qquad + \mathbb{P}\left[X_t^\top(\hat{\beta}(\mathcal{S}_{i,t-1}) - \beta_i) \geq \delta x_{\max} h, \mathcal{F}^{\lambda_0/4}_{l,t-1}, \mathcal{F}^{\lambda_0/4}_{i,t-1} \mid X_t \in \mathcal{R}_l \cap I^h\right] \\
&\quad \leq \mathbb{P}\left[X_t^\top(\beta_l - \hat{\beta}(\mathcal{S}_{l,t-1})) \geq \delta x_{\max} h, \mathcal{F}^{\lambda_0/4}_{l,t-1} \mid X_t \in \mathcal{R}_l \cap I^h\right] \\
&\qquad + \mathbb{P}\left[X_t^\top(\hat{\beta}(\mathcal{S}_{i,t-1}) - \beta_i) \geq \delta x_{\max} h, \mathcal{F}^{\lambda_0/4}_{i,t-1} \mid X_t \in \mathcal{R}_l \cap I^h\right] \\
&\quad \leq \mathbb{P}\left[\|\beta_l - \hat{\beta}(\mathcal{S}_{l,t-1})\|_2 \geq \delta h, \mathcal{F}^{\lambda_0/4}_{l,t-1} \mid X_t \in \mathcal{R}_l \cap I^h\right] \\
&\qquad + \mathbb{P}\left[\|\hat{\beta}(\mathcal{S}_{i,t-1}) - \beta_i\|_2 \geq \delta h, \mathcal{F}^{\lambda_0/4}_{i,t-1} \mid X_t \in \mathcal{R}_l \cap I^h\right],
\end{aligned} \tag{A.3.5}$$

where in the third line we used $P(A, B \mid C) \leq P(A \mid C)$, in the fourth line we used Cauchy-Schwarz inequality. Now using the notation described in Equation (A.3.2) this can be rewritten as

$$\begin{aligned}
&\mathbb{P}\left[\overline{\mathcal{G}^{\delta h}_{l,t-1}}, \mathcal{F}^{\lambda_0/4}_{l,t-1} \mid X_t \in \mathcal{R}_l \cap I^h\right] + \mathbb{P}\left[\overline{\mathcal{G}^{\delta h}_{i,t-1}}, \mathcal{F}^{\lambda_0/4}_{i,t-1} \mid X_t \in \mathcal{R}_l \cap I^h\right] \\
&\quad = \mathbb{P}\left[\overline{\mathcal{G}^{\delta h}_{l,t-1}}, \mathcal{F}^{\lambda_0/4}_{l,t-1}\right] + \mathbb{P}\left[\overline{\mathcal{G}^{\delta h}_{i,t-1}}, \mathcal{F}^{\lambda_0/4}_{i,t-1}\right] \\
&\quad \leq 4d \exp\left(-C_3(t-1)(\delta h)^2\right) = 4d\exp(-h^2),
\end{aligned}$$

in the fifth line we used the fact that both $\mathcal{R}_l$ and I^h only depend on X_t which is independent of $\hat{\beta}(\mathcal{S}_{q,t-1})$ for all q, and in the sixth line we used Lemma 5. We can also bound this probability by 1, which is better than $4d\exp(-h^2)$ for small values of h. Hence, using

$\sum_{l=1}^{K} \mathbb{P}[\mathcal{R}_l] = 1$ we can write the regret as

$$\begin{aligned}
\mathbb{E}[\text{Regret}_t(\pi)] &= \sum_{l=1}^{K} \mathbb{E}[\text{Regret}_t(\pi) \mid X_t \in \mathcal{R}_l] \cdot \mathbb{P}(X_t \in \mathcal{R}_l) \\
&\le \sum_{l=1}^{K} \left(\sum_{i \neq l} \sum_{h=0}^{h^{\max}} \left[4C_0 \delta^2 x_{\max}^2 (h+1)^2 \min\{1, 4d \exp(-h^2)\} \right] + 4(K-1) b_{\max} x_{\max} \max_i \mathbb{P}(\overline{\mathcal{F}_{i,t-1}^{\lambda_0/4}}) \right) \\
&\quad \times \mathbb{P}(X_t \in \mathcal{R}_l) \\
&\le 4(K-1) C_0 \delta^2 x_{\max}^2 \left(\sum_{h=0}^{h^{\max}} (h+1)^2 \min\{1, 4d \exp(-h^2)\} \right) + 4(K-1) b_{\max} x_{\max} \max_i \mathbb{P}(\overline{\mathcal{F}_{i,t-1}^{\lambda_0/4}}) \\
&\le 4(K-1) \left(C_0 \delta^2 x_{\max}^2 \left(\sum_{h=0}^{h_0} (h+1)^2 + \sum_{h=h_0+1}^{h^{\max}} 4d (h+1)^2 \exp(-h^2) \right) + b_{\max} x_{\max} \max_i \mathbb{P}(\overline{\mathcal{F}_{i,t-1}^{\lambda_0/4}}) \right),
\end{aligned} \tag{A.3.6}$$

where we take $h_0 = \lfloor \sqrt{\log 4d} \rfloor + 1$. Note that functions $f(x) = x^2 \exp(-x^2)$ and $g(x) = x \exp(-x^2)$ are both decreasing for $x \ge 1$ and therefore

$$\begin{aligned}
\sum_{h=h_0+1}^{h^{\max}} (h+1)^2 \exp(-h^2) &= \sum_{h=h_0+1}^{h^{\max}} (h^2 + 2h + 1) \exp(-h^2) \\
&= \sum_{h=h_0+1}^{h^{\max}} h^2 \exp(-h^2) + 2 \sum_{h=h_0+1}^{h^{\max}} h \exp(-h^2) + \sum_{h=h_0+1}^{h^{\max}} \exp(-h^2) \\
&\le \int_{h_0}^{\infty} h^2 \exp(-h^2) \mathrm{d}h + \int_{h_0}^{\infty} 2h \exp(-h^2) \mathrm{d}h + \int_{h_0}^{\infty} \exp(-h^2) \mathrm{d}h.
\end{aligned} \tag{A.3.7}$$

Computing the above terms using integration by parts and using the inequality $\int_t^\infty \exp(-x^2) \mathrm{d}x \le$

$\exp(-t^2)/(t+\sqrt{t^2+4/\pi})$ yields

$$\begin{aligned}
&\sum_{h=0}^{h_0}(h+1)^2 + 4d\sum_{h=h_0+1}^{h^{\max}}(h+1)^2\exp(-h^2) = \frac{(h_0+1)(h_0+2)(2h_0+3)}{6} + d(2h_0+7)\exp(-h_0^2)\\
&\quad\le \frac{1}{3}h_0^3 + \frac{3}{2}h_0^2 + \frac{13}{6}h_0 + 1 + d(2h_0+7)\frac{1}{4d}\\
&\quad\le \frac{1}{3}\left(\sqrt{\log 4d}+1\right)^3 + \frac{3}{2}\left(\sqrt{\log 4d}+1\right)^2 + \frac{8}{3}\left(\sqrt{\log 4d}+1\right) + \frac{11}{4}\\
&\quad\le \left(\sqrt{\log d}+2\right)^3 + \frac{3}{2}\left(\sqrt{\log d}+2\right)^2 + \frac{8}{3}\left(\sqrt{\log d}+2\right) + \frac{11}{4}\\
&\quad= \frac{1}{3}\left(\log d\right)^{3/2} + \frac{7}{2}\log d + \frac{38}{3}(\log d)^{1/2} + \frac{67}{4} = (\log d)^{3/2}\bar{C}
\end{aligned}$$

where $\bar{C}$ is defined as (2.3.2). Plugging $\delta = 1/\sqrt{(t-1)C_3}$ and substituting in (A.3.6) implies

$$r_t(\pi) = \mathbb{E}[\mathrm{Regret}_t(\pi)] \le \frac{4(K-1)C_0\bar{C}x_{\max}^2(\log d)^{3/2}}{C_3}\frac{1}{t-1} + 4(K-1)b_{\max}x_{\max}\left(\max_i \mathbb{P}[\overline{\mathcal{F}_{i,t-1}^{\lambda_0/4}}]\right)$$

as desired. □

Proof of Theorem 1. The expected cumulative regret is the sum of expected regret for times up to time T. As the regret term at time $t=1$ is upper bounded by $2x_{\max}b_{\max}$ and as $K=2$, by using Lemma 4 and Lemma 6 we can write

$$\begin{aligned}
R_T(\pi) &= \sum_{t=1}^{T} r_t(\pi)\\
&\le 2x_{\max}b_{\max} + \sum_{t=2}^{T}\left[\frac{4C_0\bar{C}x_{\max}^2(\log d)^{3/2}}{C_3}\frac{1}{t-1} + 4b_{\max}x_{\max}d\exp(-C_1(t-1))\right]\\
&= 2x_{\max}b_{\max} + \sum_{t=1}^{T-1}\left[\frac{4C_0\bar{C}x_{\max}^2(\log d)^{3/2}}{C_3}\frac{1}{t} + 4b_{\max}x_{\max}d\exp(-C_1 t)\right]\\
&\le 2x_{\max}b_{\max} + \frac{4C_0\bar{C}x_{\max}^2(\log d)^{3/2}}{C_3}\left(1+\int_1^T\frac{1}{t}\mathrm{d}t\right) + 4b_{\max}x_{\max}d\int_1^\infty \exp(-C_1 t)\mathrm{d}t\\
&= 2x_{\max}b_{\max} + \frac{4C_0\bar{C}x_{\max}^2(\log d)^{3/2}}{C_3}(1+\log T) + \frac{4b_{\max}x_{\max}d}{C_1}\\
&= \frac{128C_0\bar{C}x_{\max}^4\sigma^2 d(\log d)^{3/2}}{\lambda_0^2}\log T + \left(2x_{\max}b_{\max} + \frac{128C_0\bar{C}x_{\max}^4\sigma^2 d(\log d)^{3/2}}{\lambda_0^2} + \frac{160b_{\max}x_{\max}^3 d}{\lambda_0}\right),
\end{aligned}$$

finishing up the proof. □

A.4 Greedy-First Assumptions and Its Heuristic Implementation

We present the pseudo-code for OLS-Bandit and also the the heuristic for Greedy-First that were not presented in §2.4 due to space limitations. The OLS bandit algorithm is introduced by Goldenshluger and Zeevi (2013) and generalized by Bastani and Bayati (2015). Here, we describe the more general version that applies to more than two arms where some arms may be uniformly sub-optimal. For more details, we defer to the aforementioned papers. As mentioned earlier, in addition to Assumptions 1 and 2, OLS bandit needs two additional assumptions as follows:

Assumption 9 (Arm optimality)**.** . *Let $\mathcal{K}_{opt}$ and $\mathcal{K}_{sub}$ be mutually exclusive sets that include all K arms. Sub-optimal arms $i \in \mathcal{K}_{sub}$ satisfy $X^\top \beta_i < \max_{j \neq i} X^\top \beta_j - h$ for some $h > 0$ and every $X \in \mathcal{X}$. On the other hand, each optimal arm $i \in \mathcal{K}_{opt}$, has a corresponding set $U_i = \{X \mid X^\top \beta_i > \max_{j \neq i} X^\top \beta_j + h\}$ We assume there exists $p_* > 0$ such that $\min_{i \in \mathcal{K}_{opt}} \Pr[U_i] \geq p^*$.*

Assumption 10 (Conditional Positive-Definiteness)**.** *Define $\Sigma_i \equiv \mathbb{E}\left[XX^\top \mid X \in U_i\right]$ for all $i \in \mathcal{K}_{opt}$. Then, there exists $\lambda_1 > 0$ such that for all $i \in \mathcal{K}_{opt}, \lambda_{\min}(\Sigma_i) \geq \lambda_1 > 0$.*

The OLS Bandit algorithm requires definition of *forced-sample sets*. In particular, let us prescribe a set of times when we forced-sample arm i (regardless of the observed covariates X_t):

$$\mathcal{T}_i \equiv \left\{ (2^n - 1) \cdot Kq + j \;\middle|\; n \in \{0, 1, 2, ...\} \text{ and } j \in \{q(i-1)+1, q(i-1)+2, ..., iq\} \right\} . \tag{A.4.1}$$

Thus, the set of forced samples from arm i up to time t is $\mathcal{T}_{i,t} \equiv \mathcal{T}_i \cap [t] = \mathcal{O}(q \log t)$.

We also need to define *all-sample sets* $\mathcal{S}_{i,t} = \left\{ t' \;\middle|\; \pi_{t'} = i \text{ and } 1 \leq t' \leq t \right\}$ that are the set of times we play arm i up to time t. Note that by definition $\mathcal{T}_{i,t} \subset \mathcal{S}_{i,t}$. The algorithm proceeds as follows. During any forced sampling time $t \in \mathcal{T}_i$, the corresponding arm (arm i) is played regardless of observed covariates X_t. However, for other times, the algorithm uses two different estimations of arm parameters in order to make decision. First, it estimates arm parameters via OLS applied only on the forced samples set and discards each arm that is sub-optimal by a margin at least equal to $h/2$. Then, it applies OLS to all-sample sets and picks the arm with the highest estimated reward among the remaining arms. Algorithm 6 explains the pseudo-code for OLS Bandit.

Algorithm 6 OLS Bandit

```
**Input parameters:** $q, h$
Initialize $\hat{\beta}(\mathcal{T}_{i,0})$ and $\hat{\beta}(\mathcal{S}_{i,0})$ by 0 for all $i$ in $[K]$
Use $q$ to construct force-sample sets $\mathcal{T}_i$ using Eq. (A.4.1) for all $i$ in $[K]$
for $t \in [T]$ do
    Observe $X_t \in \mathcal{P}_X$
    if $t \in \mathcal{T}_i$ for any $i$ then
        $\pi_t \leftarrow i$
    else
        $\hat{\mathcal{K}} = \left\{ i \in K \mid X_t^T \hat{\beta}(\mathcal{T}_{i,t-1}) \geq \max_{j \in K} X_t^T \hat{\beta}(\mathcal{T}_{j,t-1}) - h/2 \right\}$
        $\pi_t \leftarrow \arg\max_{i \in \hat{\mathcal{K}}} X_t^T \hat{\beta}(\mathcal{S}_{i,t-1})$
    end if
    $\mathcal{S}_{\pi_t,t} \leftarrow \mathcal{S}_{\pi_t,t-1} \cup \{t\}$
    Play arm $\pi_t$, observe $Y_{i,t} = X_t^T \beta_{\pi_t} + \varepsilon_{i,t}$
end for
```

The pseudo-code for Heuristic Greedy-First bandit is as follows.

Algorithm 7 Heuristic Greedy-First Bandit

```
**Input parameters:** $t_0$
Execute Greedy Bandit for $t \in [t_0]$
Set $\hat{\lambda}_0 = \frac{1}{2t_0} \min_{i \in [K]} \lambda_{\min} \left( \hat{\Sigma}(\mathcal{S}_{i,t_0}) \right)$
if $\hat{\lambda}_0 \neq 0$ then
    Execute Greedy-First Bandit for $t \in [t_0 + 1, T]$ with $\lambda_0 = \hat{\lambda}_0$
else
    Execute OLS Bandit for $t \in [t_0 + 1, T]$
end if
```

A.5 Extensions to Generalized Linear Rewards and α-margin Conditions

A.5.1 Generalized Linear Rewards

Uniqueness of solution of Equation (2.3.4)**.** We first prove that the solution to maximum likelihood equation in Equation (2.3.4) is unique whenever the design matrix $\mathbf{X}^\top \mathbf{X}$ is positive definite. The first order optimality condition in Equation (2.3.4) implies that

$$\sum_{\ell=1}^{n} X_\ell \left(Y_\ell - A'(X_\ell^\top \hat{\beta})\right) = \sum_{\ell=1}^{n} X_\ell \left(Y_\ell - \mu(X_\ell^\top \hat{\beta})\right) = 0\,. \tag{A.5.1}$$

Now suppose that there are two solutions to the above equation, namely $\hat{\beta}_1$ and $\hat{\beta}_2$. Then, we can write

$$\sum_{\ell=1}^{n} X_\ell \left(\mu(X_\ell^\top \hat{\beta}_1) - \mu(X_\ell^\top \hat{\beta}_2)\right) = 0.$$

Using the mean-value theorem, for each $1 \le i \le n$ we have

$$\mu(X_\ell^\top \hat{\beta}_2) - \mu(X_\ell^\top \hat{\beta}_1) = \mu'(X_\ell^\top \tilde{\beta}_\ell) \left(X_\ell^\top (\hat{\beta}_2 - \hat{\beta}_1)\right),$$

where $\tilde{\beta}_\ell$ belongs to the line connecting $\hat{\beta}_1, \hat{\beta}_2$. Replacing this in above equation implies that

$$\sum_{\ell=1}^{n} X_\ell \left(\mu'(X_\ell^\top \tilde{\beta}_\ell) \left(X_\ell^\top (\hat{\beta}_2 - \hat{\beta}_1)\right)\right) = \left(\sum_{\ell=1}^{n} \mu'(X_\ell^\top \tilde{\beta}_\ell) X_\ell X_\ell^\top\right) (\hat{\beta}_2 - \hat{\beta}_1) = 0. \tag{A.5.2}$$

Note that μ is strictly increasing meaning that μ' is always positive. Therefore, letting $m = \min_{1 \le l \le n} \left\{\mu'(X_\ell^\top \tilde{\beta}_\ell)\right\}$, we have that

$$\sum_{\ell=1}^{n} \mu'(X_\ell^\top \tilde{\beta}_\ell) X_\ell X_\ell^\top \succeq m\mathbf{X}\mathbf{X}^\top.$$

Therefore, if the design matrix $\mathbf{X}\mathbf{X}^\top$ is positive definite, then so is $\sum_{\ell=1}^{n} \mu'(X_\ell^\top \tilde{\beta}_\ell) X_\ell X_\ell^\top$. Hence, Equation (A.5.2) implies that $\hat{\beta}_1 = \hat{\beta}_2$.

Proof of Proposition 1. For proving this, we first state and prove a Lemma that will be used later to prove this result.

Lemma 15. *Consider the generalized linear model with the inverse link function μ. Suppose that we have samples $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$, where $Y_i = \mu(X_i^\top \beta_0) + \varepsilon_i$, where $\|X_i\|_2 \le x_{\max}$ and $\|\beta_0\|_2 \le b_{\max}$. Furthermore, assume that the design matrix $\mathbf{X}^\top \mathbf{X} = \sum_{i=1}^{n} X_i X_i^\top$ is positive definite. Let $\hat{\beta} = h_\mu(\mathbf{X}, \mathbf{Y})$ be the (unique) solution to the Equation* (A.5.1)*. Let $\theta > 0$ be arbitrary and define $m_\theta := \min\{\mu'(z) : |z| \le (\theta + b_{\max}) x_{\max}\}$.*

Suppose $\|(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \varepsilon\|_2 \leq \theta m_\theta$, *then*

$$\|\hat{\beta} - \beta_0\|_2 \leq \frac{\|(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \varepsilon\|_2}{m_\theta}.$$

Proving the above Lemma is adapted from Chen et al. (1999). For completeness, we provide a proof here as well. We need the following Lemma which was proved in Chen et al. (1999).

Lemma 16. *Let H be a smooth injection from $\mathbb{R}^d$ to $\mathbb{R}^d$ with $H(\mathbf{x}_0) = \mathbf{y}_0$. Define $B_\delta(\mathbf{x}_0) = \left\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x} - \mathbf{x}_0\| \leq \delta\right\}$ and $S_\delta(\mathbf{x}_0) = \partial B_\delta(\mathbf{x}_0) = \left\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x} - \mathbf{x}_0\| = \delta\right\}$. Then, $\inf_{\mathbf{x} \in S_\delta(\mathbf{x}_0)} \|H(\mathbf{x}) - \mathbf{y}_0\| \geq r$ implies that*

(i) $B_r(\mathbf{y}_0) = \left\{\mathbf{y} \in \mathbb{R}^d : \|\mathbf{y} - \mathbf{y}_0\| \leq r\right\} \subset H(B_\delta(\mathbf{x}_0))$,

(ii) $H^{-1}(B_r(\mathbf{y}_0)) \subset B_\delta(\mathbf{x}_0)$

Proof of Lemma 15. Note that $\hat{\beta}$ is the solution to the Equation (A.5.1) and therefore

$$\sum_{i=1}^n \left(\mu(X_i^\top \hat{\beta}) - \mu(X_i^\top \beta_0)\right) X_i = \sum_{i=1}^n X_i \varepsilon_i. \tag{A.5.3}$$

Using the mean-value theorem for any $\beta \in \mathbb{R}^d$ and $1 \leq i \leq n$ we have

$$\mu(X_i^\top \beta) - \mu(X_i^\top \beta_0) = \mu'(X_i^\top \beta_i') \left(X_i^\top (\beta - \beta_0)\right),$$

where β_i' is a point that lies on the line segment between β and β_0. Define

$$\begin{aligned} G(\beta) &= \left(\sum_{i=1}^n X_i X_i^\top\right)^{-1} \left(\sum_{i=1}^n \left(\mu(X_i^\top \beta) - \mu(X_i^\top \beta_0)\right) X_i\right) \\ &= \left(\sum_{i=1}^n X_i X_i^\top\right)^{-1} \left(\sum_{i=1}^n \mu'(X_i^\top \beta_i') \left(X_i^\top (\beta - \beta_0)\right) X_i\right) \\ &= \left(\sum_{i=1}^n X_i X_i^\top\right)^{-1} \left(\sum_{i=1}^n \mu'(X_i^\top \beta_i') X_i X_i^\top\right) (\beta - \beta_0) \end{aligned}$$

As $\mu'(\cdot) > 0$, $G(\beta)$ is an injection from $\mathbb{R}^d$ to $\mathbb{R}^d$ satisfying $G(\beta_0) = 0$. Consider the sets $B_\theta(\beta_0) = \left\{\beta \in \mathbb{R}^d : \|\beta - \beta_0\|_2 \leq \theta\right\}$ and $S_\theta(\beta_0) = \left\{\beta \in \mathbb{R}^d : \|\beta - \beta_0\| = \theta\right\}$. If $\beta \in B_\theta(\beta_0)$, for each i, β_i' lies on the line segment between β and β_0 and therefore we have $|X_i^\top \beta_i'| \leq$

$\max\left(X_i^\top \beta_0, X_i^\top \beta\right) \le x_{\max}(b_{\max} + \theta)$ according to the Cauchy-Schwarz inequality. Then for each $\beta \in B_\theta(\beta_0)$

$$\begin{aligned}
\|G(\beta)\|_2^2 &= \|G(\beta) - G(\beta_0)\|_2^2 \\
&= (\beta - \beta_0)^\top \left(\sum_{i=1}^n \mu'(X_i^\top \beta_i') X_i X_i^\top\right) \left(\sum_{i=1}^n X_i X_i^\top\right)^{-2} \left(\sum_{i=1}^n \mu'(X_i^\top \beta_i') X_i X_i^\top\right) (\beta - \beta_0) \\
&= m_\theta^2 (\beta - \beta_0)^\top \left(\sum_{i=1}^n \frac{\mu'(X_i^\top \beta_i')}{m_\theta} X_i X_i^\top\right) \left(\sum_{i=1}^n X_i X_i^\top\right)^{-2} \left(\sum_{i=1}^n \frac{\mu'(X_i^\top \beta_i')}{m_\theta} X_i X_i^\top\right) (\beta - \beta_0) \\
&\ge m_\theta^2 (\beta - \beta_0)^\top \left(\sum_{i=1}^n X_i X_i^\top\right) \left(\sum_{i=1}^n X_i X_i^\top\right)^{-2} \left(\sum_{i=1}^n X_i X_i^\top\right) (\beta - \beta_0) \\
&= m_\theta^2 \|(\beta - \beta_0)\|_2^2 \,, \qquad\qquad \text{(A.5.4)}
\end{aligned}$$

or in other words $\|G(\beta)\|_2 \ge m_\theta \|\beta - \beta_0\|_2$. In particular, for any $\beta \in S_\theta(\beta_0)$ we have $G(\beta) \ge \theta m_\theta$. Therefore, letting $\gamma = \theta m_\theta$, Lemma 16 implies that $G^{-1}\left(B_\gamma(0)\right) \subset B_\theta(\beta_0)$. Note that if we let $\mathbf{z} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \varepsilon$, then by the assumption of lemma $\mathbf{z} \in B_\gamma(0)$ and hence there exists $\tilde{\beta}, \|\tilde{\beta} - \beta_0\| \le \theta$ satisfying $G^{-1}(\mathbf{z}) = \tilde{\beta}$, i.e., $G(\tilde{\beta}) = \mathbf{z}$. Now we claim that $\tilde{\beta} = \hat{\beta}$. The is not very difficult to prove. In particular, according to Equation (A.5.3) we know that

$$\sum_{i=1}^n \left(\mu(X_i^\top \hat{\beta}) - \mu(X_i^\top \beta_0)\right) X_i = \sum_{i=1}^n X_i \varepsilon_i \Longrightarrow G(\hat{\beta}) = \left(\sum_{i=1}^n X_i X_i^\top\right)^{-1} \left(\sum_{i=1}^n X_i \varepsilon_i\right) = \mathbf{z}.$$

Since the function $G(\cdot)$ is injective, it implies that $\hat{\beta} = \tilde{\beta}$. As a result, $\hat{\beta} \in B_\theta(\beta_0)$ and $G(\hat{\beta}) = \mathbf{z}$. The desired inequality follows according to Equation (A.5.4). □

Having this we can prove a Corollary of Lemma 5 for the generalized linear models.

Corollary 5. *Consider the generalized linear model with the link function μ. Consider the contextual multi-armed bandit problem, in which upon playing arm i for the context X_t, we observe a reward equal to Y_t satisfying $\mathbb{E}[Y_t] = \mu(X_t^\top \beta_i)$. Furthermore, suppose that the noise terms $\varepsilon_{it} = Y_t - \mu(X_t^\top \beta_i)$ are σ-subgaussian for some $\sigma > 0$. Let $\hat{\beta}(\mathcal{S}_{i,t}) = h_\mu\left(\mathbf{X}(\mathcal{S}_{i,t}), \mathbf{Y}(\mathcal{S}_{i,t})\right)$ be the estimated parameter of arm i. Taking $C_2 = \lambda^2/(2d\sigma^2 x_{\max}^2)$ and $n \ge |S_{i,t}|$, we have for all $\lambda, \chi > 0$,*

$$\mathbb{P}\left[\|\hat{\beta}(\mathcal{S}_{i,t}) - \beta_i\|_2 \ge \chi \;\; and \;\; \lambda_{\min}\left(\hat{\Sigma}(\mathcal{S}_{i,t})\right) \ge \lambda t\right] \le 2d \exp\left(-C_2 t^2 (\chi m_\chi)^2/n\right).$$

Proof. Note that if the design matrix $\hat{\Sigma}(\mathcal{S}_{i,t}) = \mathbf{X}(\mathcal{S}_{i,t})^\top \mathbf{X}(\mathcal{S}_{i,t})$ is positive definite, then the event $\left\{\|\hat{\beta}(\mathcal{S}_{i,t}) - \beta_i\|_2 \geq \chi\right\}$ is the subset of the event

$$\left\{\|\hat{\Sigma}(\mathcal{S}_{i,t})^{-1}\mathbf{X}(\mathcal{S}_{i,t})^\top \varepsilon(\mathcal{S}_{i,t})\| \geq \chi m_\chi\right\}.$$

The reason is very simple. Suppose the contrary, i.e., the possibility of having $\|\hat{\beta}(\mathcal{S}_{i,t}) - \beta_i\|_2 \geq \chi$ while $\|\hat{\Sigma}(\mathcal{S}_{i,t})^{-1}\mathbf{X}(\mathcal{S}_{i,t})^\top \varepsilon(\mathcal{S}_{i,t})\|_2 < \chi m_\chi$. By using the Lemma 16 for $\theta = \chi$ we achieve that

$$\|\hat{\beta}(\mathcal{S}_{i,t}) - \beta_i\|_2 \ \leq \frac{\|\hat{\Sigma}(\mathcal{S}_{i,t})^{-1}\mathbf{X}(\mathcal{S}_{i,t})^\top \varepsilon(\mathcal{S}_{i,t})\|_2}{m_\chi} < \frac{\chi m_\chi}{m_\chi} = \chi,$$

which is a contradiction. Therefore,

$$\begin{aligned}
&\mathbb{P}\left[\|\hat{\beta}(\mathcal{S}_{i,t}) - \beta_i\|_2 \geq \chi \text{ and } \lambda_{\min}\left(\hat{\Sigma}(\mathcal{S}_{i,t})\right) \geq \lambda t\right] \\
&\quad \leq \mathbb{P}\left[\|\hat{\Sigma}(\mathcal{S}_{i,t})^{-1}\mathbf{X}(\mathcal{S}_{i,t})^\top \varepsilon(\mathcal{S}_{i,t})\|_2 \geq \chi m_\chi \text{ and } \lambda_{\min}\left(\hat{\Sigma}(\mathcal{S}_{i,t})\right) \geq \lambda t\right] \\
&\quad \leq 2d \exp\left(-C_2 t^2 (\chi m_\chi)^2/n\right),
\end{aligned}$$

where the last inequality follows from the Lemma 5. □

Now we are ready to prove a Lemma following the same lines of idea as Lemma 6. This lemma can help us to prove the result for the generalized linear models.

Lemma 17. *Recall that $\mathcal{F}_{i,t}^{\lambda} = \left\{\lambda_{\min}\left(\mathbf{X}(\mathcal{S}_{i,t})^\top \mathbf{X}(\mathcal{S}_{i,t})\right) \geq \lambda t\right\}$. Suppose that Assumptions 1 and 2 hold. Then, the instantaneous expected regret of the Greedy Bandit for GLMs (Algorithm 2) at time $t \geq 2$ satisfies*

$$r_t(\pi) \leq \frac{4(K-1)L_\mu C_0 \bar{C}_\mu x_{\max}^2}{C_3} \frac{1}{t-1} + 4(K-1) b_{\max} x_{\max} \left(\max_i \mathbb{P}[\overline{\mathcal{F}_{i,t-1}^{\lambda_0/4}}]\right),$$

where $C_3 = \lambda_0^2/(32 d\sigma^2 x_{\max}^2)$, C_0 is defined in Assumption 2, L_μ is the Lipschitz constant of the function $\mu(\cdot)$ on the interval $[-x_{\max}b_{\max}, x_{\max}b_{\max}]$, and $\bar{C}_\mu$ is defined in Proposition 1.

Proof. The proof is very similar to the proof of Lemma 6. We can decompose the regret as $r_t(\pi) = \mathbb{E}[\text{Regret}_t(\pi)] = \sum_{i=1}^K \mathbb{E}[\text{Regret}_t(\pi) \mid X_t \in \mathcal{R}_i] \cdot \mathbb{P}(X_t \in \mathcal{R}_i)$. Now we can expand each term as

$$\begin{aligned}\mathbb{E}[\text{Regret}_t(\pi) \mid X_t \in \mathcal{R}_l] &= \mathbb{E}\left[\mu\left(X_t^\top \beta_l\right) - \mu\left(X_t^\top \beta_{\pi_t}\right) \mid X_t \in \mathcal{R}_l\right] \\ &\le L_\mu \mathbb{E}\left[X_t^\top(\beta_l - \beta_{\pi_t}) \mid X_t \in \mathcal{R}_l\right],\end{aligned}$$

as μ is L_μ Lipschitz over the interval $[-x_{\max}b_{\max}, x_{\max}b_{\max}]$ and $|X_t^\top \beta_j| \le x_{\max}b_{\max}$ for all $j \in [K]$. Now one can follow all the arguments in Lemma 6 up to the point that we use concentration results for $\beta_j - \hat{\beta}_j$. In particular, Equation (A.3.5) reads as

$$\begin{aligned}&\mathbb{P}\left[\mathbb{I}(X_t \in \hat{\mathcal{R}}_{i\ge l,t}, \mathcal{F}_{l,t-1}^{\lambda_0/4}, \mathcal{F}_{i,t-1}^{\lambda_0/4}) \mid X_t \in \mathcal{R}_l \cap I^h\right] \\ &\quad \le \mathbb{P}\left[\|\beta_l - \hat{\beta}(\mathcal{S}_{l,t-1})\|_2 \ge \delta h, \mathcal{F}_{l,t-1}^{\lambda_0/4} \mid X_t \in \mathcal{R}_l \cap I^h\right] \\ &\quad + \mathbb{P}\left[\|\hat{\beta}(\mathcal{S}_{i,t-1}) - \beta_i\|_2 \ge \delta h, \mathcal{F}_{i,t-1}^{\lambda_0/4} \mid X_t \in \mathcal{R}_l \cap I^h\right].\end{aligned}$$

Using the concentration result on Corollary 5, and noting that X_t is independent of $\hat{\beta}(\mathcal{S}_{j,t-1})$ for all j, the right hand side of above equation turns into

$$\begin{aligned}\mathbb{P}\left[\|\beta_l - \hat{\beta}(\mathcal{S}_{l,t-1})\|_2 \ge \delta h, \mathcal{F}_{l,t-1}^{\lambda_0/4}\right] + \mathbb{P}\left[\|\hat{\beta}(\mathcal{S}_{i,t-1}) - \beta_i\|_2 \ge \delta h, \mathcal{F}_{i,t-1}^{\lambda_0/4}\right] \\ \le 4d \exp\left(-C_3(t-1)(\delta h)^2 m_{\delta h}^2\right) \\ = 4d \exp(-h^2 m_{\delta h}^2).\end{aligned}$$

Now note that δh is at most equal to $b_{\max}$ (since $\mathbf{x}^\top(\beta_i - \beta_l)$ is upper bounded by $2x_{\max}b_{\max}$). As $m_\theta := \min\{\mu'(z) : z \in [-(b_{\max}+\theta)x_{\max}, (b_{\max}+\theta)x_{\max}]\}$, therefore if $\theta_2 > \theta_1$, then $m_{\theta_2} \le m_{\theta_1}$. Hence, for all values of $0 \le h \le h_{\max}$.

$$4d \exp(-h^2 m_{\delta h}^2) \le 4d \exp(-h^2 m_{b_{\max}}^2).$$

We can simply use 1 whenever this number is larger than one as this is the probability of an event. Therefore,

$$
\begin{aligned}
\mathbb{E}[\mathrm{Regret}_t(\pi)] &\le \sum_{l=1}^{K} L_\mu \mathbb{E}\left[X_t^\top(\beta_l - \beta_{\pi_t}) \mid X_t \in \mathcal{R}_l\right] \times \mathbb{P}(X_t \in \mathcal{R}_l) \\
&\le \sum_{l=1}^{K} L_\mu \Bigg(\sum_{i \neq l} \sum_{h=0}^{h^{\max}} \left[4C_0\delta^2 x_{\max}^2 (h+1)^2 \min\{1, 4d\exp(-h^2 m_{b_{\max}}^2)\}\right] \\
&\qquad\qquad + 4(K-1) b_{\max} x_{\max} \max_i \mathbb{P}(\overline{\mathcal{F}_{i,t-1}^{\lambda_0/4}}) \Bigg) \\
&\le 4(K-1)L_\mu \Bigg(C_0 \delta^2 x_{\max}^2 \left(\sum_{h=0}^{h^{\max}} (h+1)^2 \min\{1, 4d\exp(-h^2 m_{b_{\max}}^2)\} \right) \\
&\qquad\qquad + b_{\max} x_{\max} \max_i \mathbb{P}(\overline{\mathcal{F}_{i,t-1}^{\lambda_0/4}}) \Bigg) \\
&\le 4(K-1)L_\mu \Bigg(C_0 \delta^2 x_{\max}^2 \left(\sum_{h=0}^{h_0} (h+1)^2 + \sum_{h=h_0+1}^{h^{\max}} 4d(h+1)^2 \exp(-h^2 m_{b_{\max}}^2) \right) \\
&\qquad\qquad + b_{\max} x_{\max} \max_i \mathbb{P}(\overline{\mathcal{F}_{i,t-1}^{\lambda_0/4}}) \Bigg),
\end{aligned}
$$

where we take $h_0 = \lfloor \frac{\sqrt{\log 4d}}{m_{b_{\max}}} \rfloor + 1$. Note that functions $f(x) = x^2 \exp(-m_{b_{\max}}^2 x^2)$ and $g(x) = x\exp(-m_{b_{\max}}^2 x^2)$ are both decreasing for $x \ge 1/m_{b_{\max}}$ and therefore

$$
\begin{aligned}
\sum_{h=h_0+1}^{h^{\max}} (h+1)^2 \exp(-h^2 m_{b_{\max}}^2) &\le \int_{h_0}^{\infty} h^2 \exp(-h^2 m_{b_{\max}}^2) \mathrm{d}h \\
&\quad + \int_{h_0}^{\infty} 2h \exp(-h^2 m_{b_{\max}}^2) \mathrm{d}h + \int_{h_0}^{\infty} \exp(-h^2 m_{b_{\max}}^2) \mathrm{d}h.
\end{aligned}
$$

Using the change of variable $h' = m_{b_{\max}} h$, integration by parts, and the inequality $\int_t^\infty \exp(-x^2)\mathrm{d}x \le$

$\exp(-t^2)/(t+\sqrt{t^2+4/\pi})$, we obtain that

$$
\begin{aligned}
&\sum_{h=0}^{h_0}(h+1)^2 + 4d\sum_{h=h_0+1}^{h^{\max}}(h+1)^2\exp(-h^2)\\
&\quad= \frac{(h_0+1)(h_0+2)(2h_0+3)}{6} + 4d\left(\frac{h_0\frac{m_{b_{\max}}}{2}+\frac{1}{4}}{m_{b_{\max}}^3} + \frac{1}{m_{b_{\max}}^2} + \frac{1}{2m_{b_{\max}}}\right)\exp(-h_0^2 m_{b_{\max}}^2)\\
&\quad\le \frac{1}{3}h_0^3 + \frac{3}{2}h_0^2 + \frac{13}{6}h_0 + 1 + 4d\left(\frac{h_0\frac{m_{b_{\max}}}{2}+\frac{1}{4}}{m_{b_{\max}}^3} + \frac{1}{m_{b_{\max}}^2} + \frac{1}{2m_{b_{\max}}}\right)\frac{1}{4d}\\
&\quad\le \frac{1}{3}\left(\frac{\sqrt{\log 4d}}{m_{b_{\max}}}+1\right)^3 + \frac{3}{2}\left(\frac{\sqrt{\log 4d}}{m_{b_{\max}}}+1\right)^2 + \frac{8}{3}\left(\frac{\sqrt{\log 4d}}{m_{b_{\max}}}+1\right)\\
&\quad+ \frac{1}{m_{b_{\max}}^3}\left(\left(\frac{\sqrt{\log 4d}}{m_{b_{\max}}}+1\right)\frac{m_{b_{\max}}}{2}+\frac{1}{4}\right) + \frac{1}{m_{b_{\max}}^2} + \frac{1}{2m_{b_{\max}}} = \bar{C}_\mu
\end{aligned}
$$

By replacing this in the regret equation above and substituting $\delta = 1/\sqrt{(t-1)C_3}$ we get

$$
r_t(\pi) = \mathbb{E}[\mathrm{Regret}_t(\pi)] \le \frac{4(K-1)L_\mu C_0 \bar{C}_\mu x_{\max}^2}{C_3}\frac{1}{t-1} + 4(K-1)L_\mu b_{\max} x_{\max}\left(\max_i \mathbb{P}[\overline{\mathcal{F}_{i,t-1}^{\lambda_0/4}}]\right)
$$

as desired. □

Now we are ready to finish up the proof of Proposition 1. The only other result that we need is an upper bound on the probability terms $\mathbb{P}[\overline{\mathcal{F}_{i,t-1}^{\lambda_0/4}}]$. The key here is again Lemma 4. Note that in the case of GLMs this lemma again holds. The reason is simply because of the fact that the greedy decision does not change in the presence of the inverse link function μ. In other words, as $\arg\max_{i\in[K]} \mu'(X_t^\top \beta_i) = \arg\max_{i\in[K]} X_t^\top \beta_i$, the minimum eigenvalue of each of the covariance matrices is above $t\lambda_0/4$ with a high probability and that implies what we exactly want.

Remark A.5.1. *The result of Lemma 4 remains true for the generalized linear models.*

Therefore, we can use this observation to finish the proof of Proposition 1. This consists of summing up the regret terms up to time T.

Proof of Proposition 1. The expected cumulative regret is the sum of expected regret for times up to time T. As the regret term at time $t = 1$ is upper bounded by $2L_\mu x_{\max} b_{\max}$

and as $K = 2$, by using Lemma 4 and Lemma 17 we can write

$$
\begin{aligned}
R_T(\pi) &= \sum_{t=1}^{T} r_t(\pi) \\
&\leq 2L_\mu x_{\max} b_{\max} + \sum_{t=2}^{T} L_\mu \left[\frac{4C_0 \bar{C}_\mu x_{\max}^2}{C_3} \frac{1}{t-1} + 4b_{\max} x_{\max} d \exp(-C_1(t-1)) \right] \\
&= 2L_\mu x_{\max} b_{\max} + \sum_{t=1}^{T-1} L_\mu \left[\frac{4C_0 \bar{C}_\mu x_{\max}^2}{C_3} \frac{1}{t} + 4b_{\max} x_{\max} d \exp(-C_1 t) \right] \\
&\leq 2L_\mu x_{\max} b_{\max} + L_\mu \frac{4C_0 \bar{C}_\mu x_{\max}^2}{C_3} (1 + \int_1^T \frac{1}{t} \mathrm{d}t) + 4L_\mu b_{\max} x_{\max} d \int_1^\infty \exp(-C_1 t) \mathrm{d}t \\
&= 2L_\mu x_{\max} b_{\max} + L_\mu \frac{4C_0 \bar{C}_\mu x_{\max}^2}{C_3} (1 + \log T) + L_\mu \frac{4b_{\max} x_{\max} d}{C_1} \\
&= L_\mu \left(\frac{128 C_0 \bar{C}_\mu x_{\max}^4 \sigma^2 d}{\lambda_0^2} \log T + \left(2x_{\max} b_{\max} + \frac{128 C_0 \bar{C}_\mu x_{\max}^4 \sigma^2 d}{\lambda_0^2} + \frac{160 b_{\max} x_{\max}^3 d}{\lambda_0} \right) \right) ,
\end{aligned}
$$

finishing up the proof. □

A.5.2 Regret bounds for More General Margin Conditions

While the assumed margin condition in Assumption 2 holds for many well-known distributions, one can construct a distribution with a growing density near the decision boundary that violates Assumption 2. Therefore, it is interesting to see how regret bounds would change if we assume other type of margin conditions. Similar to what proposed in Weed et al. (2015), we assume that the distribution of contexts p_X satisfies a more general α-margin condition as following.

Assumption 11 (α-Margin Condition)**.** *For $\alpha \geq 0$, we say that the distribution p_X satisfies the α-margin condition, if there exists a constant $C_0 > 0$ such that for each $\kappa > 0$:*

$$
\forall\ i \neq j : \qquad \mathbb{P}_X \Big[0 < |X^\top (\beta_i - \beta_j)| \leq \kappa \Big] \leq C_0 \kappa^\alpha .
$$

Although it is straightforward to verify that any distribution p_X satisfies the 0-margin condition, it is easy to construct a distribution violating the α-margin condition, for an arbitrary $\alpha > 0$. In addition, if p_X satisfies the α-margin condition, then for any $\alpha' < \alpha$ it also satisfies the α'-margin condition. In the case that there exists some gap between arm rewards, meaning the existence of $\kappa_0 > 0$ such that

$$\forall\ i \neq j: \qquad \mathbb{P}_X\Big[0 < |X^\top(\beta_i - \beta_j)| \leq \kappa_0\Big] = 0,$$

the distribution p_X satisfies the α-margin condition for all $\alpha \geq 0$.

Having this definition in mind, we can prove the following result on the regret of Greedy Bandit algorithm when p_X satisfies the α-margin condition:

Corollary 6. *Let $K = 2$ and suppose that p_X satisfies the α-margin condition. Furthermore, assume that Assumptions 1 and 3 hold, then we have the following asymptotic bound on the expected cumulative regret of Greedy Bandit algorithm*

$$R_T(\pi) = \begin{cases} \mathcal{O}\left(T^{(1-\alpha)/2}\right) & \textit{if } 0 \leq \alpha < 1, \\ \mathcal{O}\left(\log T\right) & \textit{if } \alpha = 1, \\ \mathcal{O}(1) & \textit{if } \alpha > 1, \end{cases} \tag{A.5.5}$$

This result shows that if the distribution p_X satisfies the α-margin condition for $\alpha > 1$, then the Greedy Bandit algorithm is capable of learning the parameters β_i while incurring a constant regret in expectation.

Proof. This corollary can be easily implied from Lemma 6 and Theorem 1 with a very slight modification. Note that all the arguments in Lemma 6 hold and the only difference is where we want to bound the probability $\mathbb{P}[X_t \in I^h]$ in Equation (A.3.4). In this Equation, if we use the α-margin bound as

$$\mathbb{P}[X_t^\top(\beta_l - \beta_i) \in (0, 2\delta x_{\max}(h+1)]] \leq C_0\left(2\delta x_{\max}(h+1)\right)^\alpha,$$

we obtain that

$$\begin{aligned} &\mathbb{E}\left[\mathbb{I}(X_t \in \hat{\mathcal{R}}_{i\geq l,t}, \mathcal{F}_{l,t-1}^{\lambda_0/4}, \mathcal{F}_{i,t-1}^{\lambda_0/4}) X_t^\top(\beta_l - \beta_i) \mid X_t \in \mathcal{R}_l\right] \\ &\quad \leq \sum_{h=0}^{h^{\max}} 2^{1+\alpha} C_0 \delta^{1+\alpha} x_{\max}^{1+\alpha}(h+1)^{1+\alpha} + \mathbb{P}\left[X_t \in \hat{\mathcal{R}}_{i\geq l,t}, \mathcal{F}_{l,t-1}^{\lambda_0/4}, \mathcal{F}_{i,t-1}^{\lambda_0/4} \mid X_t \in \mathcal{R}_l \cap I^h\right], \end{aligned}$$

which turns the regret bound in Equation (A.3.6) into

$$r_t(\pi) \le (K-1)\Big[C_0 2^{1+\alpha}\delta^{1+\alpha}x_{\max}^{1+\alpha}\Big(\sum_{h=0}^{h_0}(h+1)^{1+\alpha} + \sum_{h=h_0+1}^{h^{\max}} 4d(h+1)^{1+\alpha}\exp(-h^2)\Big)\Big] \tag{A.5.6}$$

$$+ 4(K-1)b_{\max}x_{\max}\max_i \mathbb{P}(\overline{\mathcal{F}_{i,t-1}^{\lambda_0}}),$$

Now we claim that the above summation has an upper bound that only depends on d and α. If we prove this claim, the dependency of the regret bound with respect to t can only come from the term $\delta^{1+\alpha}$ and therefore we can prove the desired asymptotic bounds. For proving this claim, consider the summation above and let $h_1 = \lceil\sqrt{3+\alpha}\rceil$. Note that for each $h \ge h_2 = \max(h_0, h_1)$ using $h^2 \ge (3+\alpha)h \ge (3+\alpha)\log h$ we have

$$(h+1)^{1+\alpha}\exp(-h^2) \le (2h)^{1+\alpha}\exp(-h^2) \le 2^{1+\alpha}\exp(-h^2 + (1+\alpha)\log h) \le \frac{2^{1+\alpha}}{h^2}.$$

Furthermore, all the terms corresponding to $h \le h_2 = \max(h_0, h_1)$ have an upper bound equal to $(h+1)^{1+\alpha}$ (remember that for $h \ge h_0 + 1$ we have $4d\exp(-h^2) \le 1$). Therefore, the summation in (A.5.6) is bounded above by

$$\sum_{h=0}^{h_0}(h+1)^{1+\alpha} + \sum_{h=h_0+1}^{h^{\max}} 4d(h+1)^{1+\alpha}\exp(-h^2) \le \sum_{h=0}^{h_2}(h+1)^{1+\alpha} + \sum_{h=h_2+1}^{\infty}\frac{1}{h^2}$$

$$\le (1+h_2)^{2+\alpha} + \frac{\pi^2}{6} := g(d,\alpha)$$

for some function g. This is true according to the fact that h_2 is the maximum of h_0, that only depends on d, and h_1 that only depends on α. Replacing $\delta = 1/\sqrt{(t-1)C_3}$ in the Equation (A.5.6) and absorbing all the constants we reach to

$$r_t(\pi) = (K-1)g_1(d,\alpha,C_0,x_{\max},\sigma,\lambda_0)(t-1)^{-(1+\alpha)/2} + 4(K-1)b_{\max}x_{\max}\left(\max_i \mathbb{P}[\overline{\mathcal{F}_{i,t}^{\lambda_0}}]\right)$$

for some function g_1.

The last part of the proof is summing up the instantaneous regret terms for $t = 1, 2, \ldots, T$. Note that $K = 2$, and using Lemma 4 for $i = 1, 2$, we can bound the probabilities

$\mathbb{P}[\overline{\mathcal{F}_{i,t-1}^{\lambda_0}}]$ by $d\exp(-C_1(t-1))$ and therefore

$$\begin{aligned}
R_T(\pi) &\leq 2x_{\max}b_{\max} + \sum_{t=2}^{T} g_1(d,\alpha,C_0,x_{\max},\sigma,\lambda_0)(t-1)^{-(1+\alpha)/2} + 4b_{\max}x_{\max}d\exp(-C_1(t-1)) \\
&\leq 2x_{\max}b_{\max} + \sum_{t=1}^{T-1} g_1(d,\alpha,C_0,x_{\max},\sigma,\lambda_0)t^{-(1+\alpha)/2} + 4b_{\max}x_{\max}d\exp(-C_1 t) \\
&\leq 2x_{\max}b_{\max} + g_1(d,\alpha,C_0,x_{\max},\sigma,\lambda_0)\left[1 + \left(\int_{t=1}^{T} t^{-(1+\alpha)/2}\mathrm{d}t\right)\right] \\
&\quad + 4db_{\max}x_{\max}\int_0^{\infty}\exp(-C_1 t)\mathrm{d}t \\
&= 2x_{\max}b_{\max} + g_1(d,\alpha,C_0,x_{\max},\sigma,\lambda_0)\left[1 + \left(\int_{t=1}^{T} t^{-(1+\alpha)/2}\mathrm{d}t\right)\right] \\
&\quad + \frac{4b_{\max}x_{\max}d}{C_1}.
\end{aligned}$$

Now note that the integral of $t^{-(1+\alpha)/2}$ over the interval $[1,T]$ satisfies

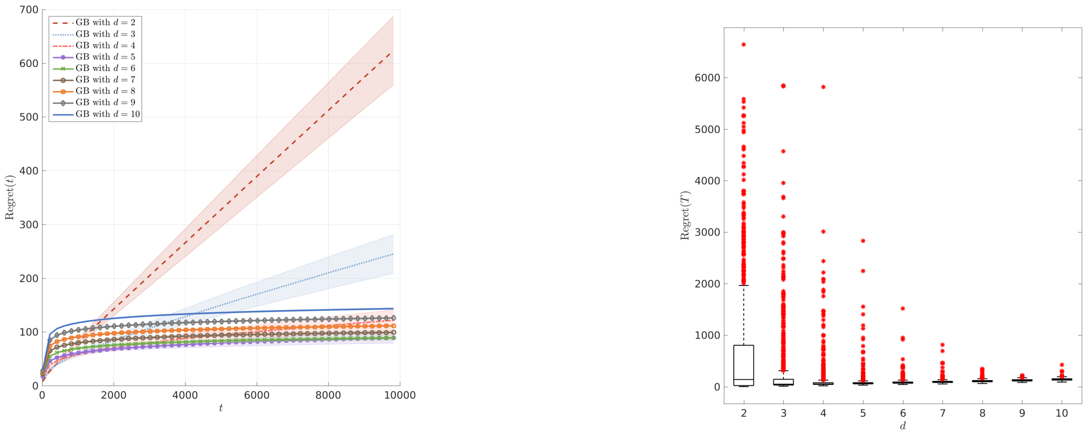
$$\int_{t=1}^{T} t^{-(1+\alpha)/2} \leq \begin{cases} \frac{T^{(1-\alpha)/2}}{(1-\alpha)/2} & \text{if } 0 \leq \alpha < 1, \\ \log T & \text{if } \alpha = 1, \\ \frac{1}{(\alpha-1)/2} & \text{if } \alpha > 1, \end{cases}$$

which yields the desired result. □

A.6 Additional Simulations

A.6.1 More than Two Arms ($K > 2$)

For investigating the performance of the Greedy-Bandit algorithm in presence of more than two arms, we run Greedy Bandit algorithm for $K = 5$ and $d = 2, 3, \ldots, 10$ while keeping the distribution of covariates as $0.5 \times \mathsf{N}(\mathbf{0}_d, \mathbf{I}_d)$ truncated at 1. We assume that β_i is again drawn from $\mathsf{N}(0_d, \mathbf{I}_d)$. For having a fair comparison, we scale the noise variance by d so as to keep the signal-to-noise ratio fixed (i.e., $\sigma = 0.25\sqrt{d}$). For small values of d, it is likely that Greedy Bandit algorithm drops an arm due to the poor estimations and as a result its regret becomes linear. However, for large values of d this issue is resolved and Greedy Bandit starts to perform very well.



(a) Regret for $t = 1, \ldots, 10000$.

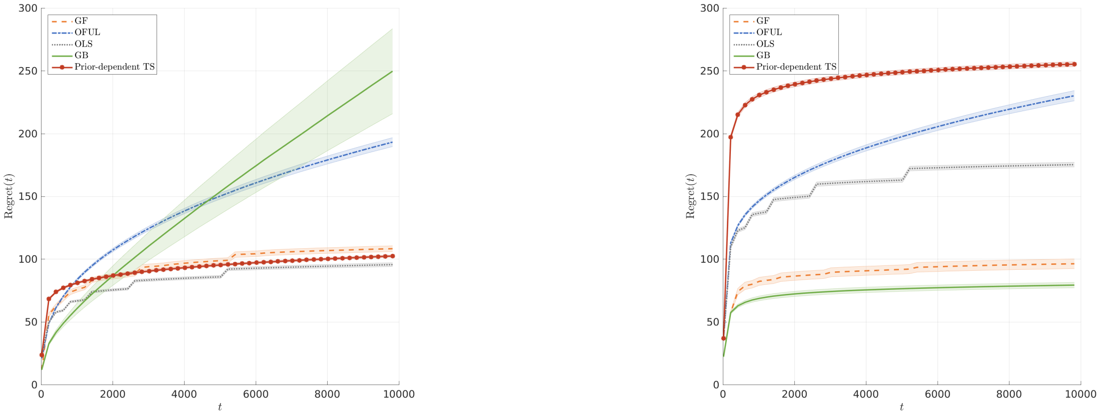
(b) Distribution of regret at $T = 10000$.

Figure A.1: These figures show a sharp change in the performance of Greedy Bandit for $K = 5$ arms as d increases.

We then repeat the simulations of §3.7 for $K = 5$ and $d \in \{3, 7\}$ while keeping the other parameters as in §3.7. In other words, we assume that β_i is drawn from $\mathsf{N}(\mathbf{0}_d, \mathbf{I}_d)$. Also, X is drawn from $0.5 \times \mathsf{N}(\mathbf{0}_d, \mathbf{I}_d)$ truncated to have its ℓ_∞ norm at most one. We create 1000 problem instances and plot the average cumulative regret of algorithms for $T \in \{1, 2, \ldots, 10000\}$. We use the correct prior regime for OFUL and TS. The results, as shown in Figure A.2, demonstrate that Greedy-First nearly ties with Greedy Bandit as the winner when $d = 7$. However for $d = 3$ that Greedy Bandit performs poorly, while Greedy-First performs very close to the best algorithms.

A.6.2 Sensitivity to Parameters

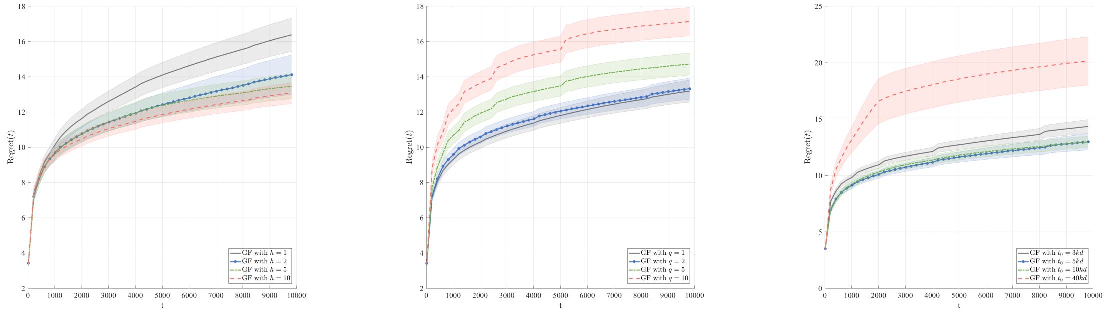
In this section, we will perform a sensitivity analysis to demonstrate that the choice of parameters h, q, and t_0 has a small impact on performance of Greedy First. The sensitivity analysis is performed with the same problem parameters as in Figure 2.2 for the case that covariate diversity does not hold. As it can be observed from Figure A.3, the choice of parameters h, q, and t_0 does have a very small impact on the performance of the Greedy-First algorithm, which verifies the robustness of Greedy-First algorithm to the choice of parameters.



(a) $K = 5, d = 3$

(b) $K = 5, d = 7$

Figure A.2: Simulations for $K > 2$ arms.



(a) Sensitivity with respect to h.

(b) Sensitivity with respect to q.

(c) Sensitivity with respect to t_0.

Figure A.3: Sensitivity analysis for the expected regret of Greedy-First algorithm with respect to the input parameters h, q, and t_0.

A.7 Proofs of Probabilistic Results

Proof of Proposition 2. We first start by proving monotonicity results:

- Let $\sigma_1 < \sigma_2$. Note that only the second, the third, and the last term of $L(\gamma, \delta, p)$, defined in Equation (2.3.6), depend on σ. As for any positive number χ, the function $\exp(-\chi/\sigma^2)$ is increasing with respect to σ, second and third terms are increasing with respect to σ. Furthermore, the last term can be expressed as

$$\frac{2d \exp\left(-D_2(\gamma)(p - m|\mathcal{K}_{sub}|)\right)}{1 - \exp(-D_2(\gamma))} = 2d \sum_{t=p-m|\mathcal{K}_{sub}|}^{\infty} \exp\left(-\frac{\lambda_1^2 h^2 (1-\gamma)^2}{8d\sigma^2 x_{\max}^4} t\right).$$

  Each term in above sum is increasing with respect to σ. Therefore, the function L is increasing with respect to σ. As S^{gb} is one minus the infimum of L taken over the possible parameter space of γ, δ, and p, that is also non-increasing with respect to σ, yielding the desired result.

- Let $m_1 < m_2$ and suppose that we use the superscript $L^{(i)}$ for the function $L(\cdot, \cdot, \cdot)$ when $m = m_i, i = 1, 2$. We claim that for all $\gamma \in (0, 1), \delta > 0$, and $p \geq Km_1 + 1$, conditioning on $L^{(1)}(\gamma, \delta, p) \leq 1$ we have $L^{(1)}(\gamma, \delta, p) \geq L^{(2)}(\gamma, \delta, p + K(m_2 - m_1))$. Note that the region for which $L^{(1)}(\gamma, \delta, p) > 1$ does not matter as it leads to a negative probability of success in the formula $S^{\text{gb}} = 1 - \inf_{\gamma,\delta,p} L(\gamma, \delta, p)$, and we can only restrict our attention to the region for which $L^{(1)}(\gamma, \delta, p) \leq 1$. To prove the claim let $\theta_i = \mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m_i}^\top \mathbf{X}_{1:m_i}) \geq \delta\right],\ i = 1, 2$ and define $f(\theta) = 1 - \theta^K + QK\theta$ for the constant $Q = 2d \exp\left(-(h^2\delta)/(8d\sigma^2 x_{\max}^2)\right)$. Note that $f(\theta_i)$ is equal to to the first two terms of $L^{(i)}(\gamma, \delta, p)$ in Equation (2.3.6). As we later going to replace $\theta = \theta_i$ we only restrict our attention to $\theta \geq 0$. The derivative of f is equal to $f'(\theta) = -K\theta^{K-1} + QK$ which is negative when $\theta^{K-1} > Q$. Note that if $\theta^{K-1} \leq Q$ and if we drop the third, fourth, and fifth term in L (see Equation (2.3.6)) that are all positive, we obtain $L^{(i)}(\gamma, \delta, p) > 1 - \theta^K + QK\theta > 1 - \theta^K + Q\theta \geq 1$, leaving us in the unimportant regime. Therefore, on the important regime the derivative is negative and f is decreasing. It is not very difficult to see that $\theta_1 \leq \theta_2$. Returning to our original claim, if we calculate $L^{(1)}(\gamma, \delta, p) - L^{(2)}(\gamma, \delta, p + K(m_2 - m_1))$ it is easy to observe that the third term cancels out and we end up with

$$\begin{aligned}L^{(1)}(\gamma,\delta,p)-L^{(2)}(\gamma,\delta,p+K(m_2-m_1)) &= f(\theta_1)-f(\theta_2)\\ &+\frac{\exp\left(-D_1(\gamma)(p-m_1|\mathcal{K}_{sub}|)\right)-\exp\left(-D_1(\gamma)(p-m_2|\mathcal{K}_{sub}|+K(m_2-m_1))\right)}{1-\exp(-D_1(\gamma))}\\ &+\frac{\exp\left(-D_2(\gamma)(p-m_1|\mathcal{K}_{sub}|)\right)-\exp\left(-D_2(\gamma)(p-m_2|\mathcal{K}_{sub}|+K(m_2-m_1))\right)}{1-\exp(-D_2(\gamma))}\geq 0\,,\end{aligned}$$

where we used the inequality $(p-m_1|\mathcal{K}_{sub}|)-(p-m_2|\mathcal{K}_{sub}|+K(m_2-m_1))=|\mathcal{K}_{opt}|(m_2-m_1)\geq 0$. This proves our claim. Note that whenever when p varies in the range $[Km_1+1,\infty)$, the quantity $p+K(m_2-m_1)$ covers the range $[Km_2+1,\infty)$. Therefore, we can write that

$$\begin{aligned}S^{\mathrm{gb}}(m_1,K,\sigma,x_{\max},\lambda_1,h) &= 1-\inf_{\gamma\in(0,1),\delta,p\geq Km_1+1}L^{(1)}(\gamma,\delta,p)\\ &\leq 1-\inf_{\gamma\in(0,1),\delta,p\geq Km_1+1}L^{(1)}(\gamma,\delta,p+K(m_2-m_1))\\ &= 1-\inf_{\gamma\in(0,1),\delta,p'\geq Km_2+1}L^{(2)}(\gamma,\delta,p')\\ &= S^{\mathrm{gb}}(m_2,K,\sigma,x_{\max},\lambda_1,h),\end{aligned}$$

as desired.

- Let $h_1<h_2$. In this case it is very easy to check that the first, fourth and fifth terms in L (see Equation (2.3.6)) do not depend on h. Dependency of second and third terms are in the form $\exp(-Qh^2)$ for some constant Q, which is decreasing with respect h. Therefore, if we use the superscript $L^{(i)}$ for the function $L(\cdot,\cdot,\cdot)$ when $h=h_i, i=1,2$, we have that $L^{(1)}(\gamma,\delta,p)\geq L^{(2)}(\gamma,\delta,p)$ which implies

$$\begin{aligned}S^{\mathrm{gb}}(m,K,\sigma,x_{\max},\lambda_1,h_1) &= 1-\inf_{\gamma\in(0,1),\delta,p\geq Km+1}L^{(1)}(\gamma,\delta,p)\\ &\leq 1-\inf_{\gamma\in(0,1),\delta,p\geq Km+1}L^{(2)}(\gamma,\delta,p)\\ &= 1-\inf_{\gamma\in(0,1),\delta,p'\geq Km+1}L^{(2)}(\gamma,\delta,p')\\ &= S^{\mathrm{gb}}(m,K,\sigma,x_{\max},\lambda_1,h_2),\end{aligned}$$

as desired.

- Similar to the previous part, it is easy to observe that the first, second, and third term in L, defined in Equation (2.3.6) do not depend on λ_1. The dependency of last two terms with respect to λ_1 is of the form $\exp(-Q_1\lambda_1)$ and $\exp(-Q_2\lambda_1^2)$ which both are decreasing functions of λ_1. The rest of argument is similar to the previous part and by replicating it with reach to the conclusion that S^{gb} is non-increasing with respect to λ_1.

- Let us suppose that $K_1 m_1 = K_2 m_2, |\mathcal{K}_{1_{sub}}| m_1 = |\mathcal{K}_{2_{sub}}| m_2$, and $K_1 < K_2$. Similar to before, we use superscript $L^{(i)}$ to denote the function $L(\cdot,\cdot,\cdot)$ when $m = m_i, K = K_i, \mathcal{K}_{sub} = \mathcal{K}_{i_{sub}}$. Then it is easy to check that the last three terms in $L^{(1)}$ and $L^{(2)}$ are the same. Therefore, for comparing $S^{\mathrm{gb}}(m_1, K_1, \sigma, x_{\max}, \lambda_1)$ and $S^{\mathrm{gb}}(m_2, K_2, \sigma, x_{\max}, \lambda_1)$ one only needs to compare the first two terms. Letting $\mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m_i}^\top \mathbf{X}_{1:m_i}) \geq \delta\right] = \theta_i,\ i = 1,2$ and $Q = 2d\exp\left(-\frac{h^2\delta}{8d\sigma^2 x_{\max}^2}\right)$ we have

  $$L^{(1)}(\gamma,\delta,p) - L^{(2)}(\gamma,\delta,p) = \theta_2^{K_2} - \theta_1^{K_1} + QK_1\theta_1 - QK_2\theta_2.$$

  Similar to the proof of second part, it is not very hard to prove that on the reasonable regime for the parameters the function $g(\theta) = -\theta^{K_1} + QK_1\theta$ is decreasing and therefore

  $$L^{(1)}(\gamma,\delta,p) - L^{(2)}(\gamma,\delta,p) = \theta_2^{K_2} - \theta_1^{K_1} + QK_1\theta_1 - QK_2\theta_2 \leq \theta_2^{K_2} - \theta_2^{K_1} + QK_1\theta_2 - QK_2\theta_2 < 0,$$

  as $\theta_1 \geq \theta_2 \in [0,1]$ and $K_2 > K_1$. Taking the infimum implies the desired result.

Now let us derive the limit of L when $\sigma \to 0$. For each $\sigma < (1/Km)^2$, define $\gamma(\sigma) = 1/2,\ \delta(\sigma) = \sqrt{\sigma}$, and $p(\sigma) = \lceil 1/\sqrt{\sigma} \rceil$. Then, by computing the function L for these specific choices of parameters and upper bounding the summation in Equation (2.3.6) with its maximum times the number of terms we get

$$\begin{aligned}
L(\gamma(\sigma),\delta(\sigma),p(\sigma)) \leq 1 &- \left(\mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m}) \geq \sqrt{\sigma}\right]\right)^K \\
&+ 2Kd\mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m}) \geq \sqrt{\sigma}\right] \exp\left(-Q_1/\sigma^{3/2}\right) + \frac{2d}{\sqrt{\sigma}} \exp\left(-Q_2/\sqrt{\sigma}\right) \\
&+ d\frac{\exp\left(-Q_3/\sqrt{\sigma}\right)}{1-\exp(-Q_3)} + 2d\frac{\exp\left(-Q_4/\sigma^{5/2}\right)}{1-\exp\left(-Q_4/\sigma^2\right)} := J(\sigma),
\end{aligned}$$

for positive constants Q_1, Q_2, Q_3, and Q_4 that do not depend on σ. Note that for $\sigma > 0$,

$$\inf_{\gamma \in (0,1), \delta > 0, p \geq Km+1} L(\gamma, \delta, p) \leq J(\sigma).$$

Therefore, by taking limit with respect to σ we get

$$\begin{aligned}
\lim_{\sigma \downarrow 0} S^{\mathrm{gb}}(m, K, \sigma, x_{\max}, \lambda_1, h) &= 1 - \lim_{\sigma \downarrow 0} L(\gamma, \delta, p) \\
&\geq \lim_{\sigma \downarrow 0} (1 - J(\sigma)) = 1 - \left\{ 1 - \left(\mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^{\top} \mathbf{X}_{1:m}) > 0 \right] \right)^K \right\} \\
&= \mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^{\top} \mathbf{X}_{1:m}) > 0 \right]^K,
\end{aligned}$$

proving one side of the result. For achieving the desired result we need to prove that $\mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^{\top} \mathbf{X}_{1:m}) > 0 \right]^K \geq \lim_{\sigma \downarrow 0} S^{\mathrm{gb}}(m, K, \sigma, x_{\max}, \lambda_1, h)$ which is the easier way. Note that the function L always satisfies

$$L(\gamma, \delta, p) \geq 1 - \left(\mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^{\top} \mathbf{X}_{1:m}) \geq \delta \right] \right)^K \geq 1 - \left(\mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^{\top} \mathbf{X}_{1:m}) > 0 \right] \right)^K.$$

As a result, for any $\sigma > 0$ we have

$$S^{\mathrm{gb}}(m, K, \sigma, x_{\max}, \lambda_1, h) \leq 1 - \left(1 - \mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^{\top} \mathbf{X}_{1:m}) > 0 \right] \right)^K = \mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^{\top} \mathbf{X}_{1:m}) > 0 \right]^K.$$

By taking limits we reach to the desired conclusion. □

Proof of Proposition 3. We omit proofs regarding to the monotonicity results as they are very similar to those provided in Proposition 2.

For deriving the limit when $\sigma \to 0$, define $\gamma(\sigma) = \gamma^*$, $\delta(\sigma) = \sqrt{\sigma}$, and $p(\sigma) = t_0$. Then, by computing the function L' for these specific values we have

$$\begin{aligned}
L'(\gamma(\sigma), \delta(\sigma), p(\sigma)) \leq 1 &- \left(\mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^{\top} \mathbf{X}_{1:m}) \geq \sqrt{\sigma} \right] \right)^K \\
&+ 2Kd\mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^{\top} \mathbf{X}_{1:m}) \geq \sqrt{\sigma} \right] \exp\left(\frac{-Q_1'}{\sigma^{3/2}} \right) + 2dt_0 \exp\left\{ -\frac{Q_2'}{\sigma} \right\} \\
&+ \frac{Kd \exp(-D_1(\gamma^*) t_0)}{1 - \exp(-D_1(\gamma^*))} + 2d \frac{\exp\left(-Q_3' t_0 / \sigma^2 \right)}{1 - \exp\left(-Q_3' / \sigma^2 \right)} := J'(\sigma),
\end{aligned}$$

for positive constants Q_1', Q_2', and Q_3' that do not depend on σ. Note that for $\sigma > 0$,

$$\inf_{\gamma \le \gamma^*, \delta > 0, Km+1 \le p \le t_0} L'(\gamma, \delta, p) \le J'(\sigma).$$

Therefore, by taking limit with respect to σ we get

$$\begin{aligned}\lim_{\sigma \downarrow 0} S^{\mathrm{gf}}(m, K, \sigma, x_{\max}, \lambda_1, h) &= 1 - \lim_{\sigma \downarrow 0} L'(\gamma, \delta, p) \\ &\ge \lim_{\sigma \downarrow 0} \big(1 - J'(\sigma)\big) \\ &= \mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m}) > 0\right]^K - \frac{Kd \exp(-D_1(\gamma^*) t_0)}{1 - \exp(-D_1(\gamma^*))},\end{aligned}$$

proving one side of the result. For achieving the desired result we need to prove that the other side of this inequality. Note that the function L' always satisfies

$$L'(\gamma, \delta, p) \ge 1 - \left(\mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m}) \ge \delta\right]\right)^K + \frac{Kd \exp(-D_1(\gamma) p)}{1 - \exp(-D_1(\gamma))}. \tag{A.7.1}$$

Note that the function $D_1(\gamma)$ is increasing with respect to γ. This is easy to verify as the first derivative of $D_1(\gamma)$ with respect to γ is equal to

$$\frac{\partial D_1}{\partial \gamma} = \frac{\lambda_1}{x_{\max}^2} \left\{1 - \log(1 - \gamma) - 1\right\} = -\frac{\lambda_1}{x_{\max}^2} \log(1 - \gamma),$$

which is increasing for $\gamma \in [0, 1)$. Therefore, by using $p \le t_0$ and $\gamma \le \gamma^*$ we have

$$\frac{Kd \exp(-D_1(\gamma) p)}{1 - \exp(-D_1(\gamma))} \ge \frac{Kd \exp(-D_1(\gamma^*) t_0)}{1 - \exp(-D_1(\gamma^*))}.$$

Substituting this in Equation (A.7.1) implies that

$$\begin{aligned}S^{\mathrm{gf}}(m, K, \sigma, x_{\max}, \lambda_1, h) &\le 1 - \left\{\left(1 - \mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m}) > 0\right]\right)^K + \frac{Kd \exp(-D_1(\gamma^*) t_0)}{1 - \exp(-D_1(\gamma^*))}\right\} \\ &= \mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m}) > 0\right]^K - \frac{Kd \exp(-D_1(\gamma^*) t_0)}{1 - \exp(-D_1(\gamma^*))}.\end{aligned}$$

By taking limits we reach to the desired conclusion. □

Proofs of Theorems 2 and 4

Let us first start by introducing two new notations and recalling some others. For each $\delta > 0$ define

$$\begin{aligned}\mathcal{H}_i^\delta &:= \left\{\lambda_{\min}\left(\mathbf{X}(\mathcal{S}_{i,Km})^\top \mathbf{X}(\mathcal{S}_{i,Km})\right) \geq \delta\right\}\\ \mathcal{J}_{i,t}^\lambda &= \left\{\lambda_{\min}\left(\mathbf{X}(\mathcal{S}_{i,t})^\top \mathbf{X}(\mathcal{S}_{i,t})\right) \geq \lambda t - m|\mathcal{K}_{sub}|\right\},\end{aligned}$$

and recall that

$$\begin{aligned}\mathcal{F}_{i,t}^\lambda &= \left\{\lambda_{\min}\left(\mathbf{X}(\mathcal{S}_{i,t})^\top \mathbf{X}(\mathcal{S}_{i,t})\right) \geq \lambda t\right\}\\ \mathcal{G}_{i,t}^\chi &= \left\{\|\hat{\beta}(\mathcal{S}_{i,t}) - \beta_i\|_2 < \chi\right\}.\end{aligned}$$

Note that whenever $|\mathcal{K}_{sub}| = 0$, the sets $\mathcal{J}$ and $\mathcal{F}$ coincide. We first start by proving some lemmas that will be used later to prove Theorems 2 and 4.

Lemma 18. *Let $i \in [K]$ be arbitrary. Then*

$$\mathbb{P}\left[\mathcal{H}_i^\delta \cap \overline{\mathcal{G}_{i,Km}^{\theta_1}}\right] \leq 2d\mathbb{P}\left\{\lambda_{\min}\left(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m}\right) \geq \delta\right\} \exp\left\{-\frac{\theta_1^2 \delta}{2d\sigma^2}\right\}$$

Remark A.7.1. *Note that Lemma 5 provides an upper bound on the same probability event described above. However, those results are addressing the case that samples are highly correlated due to greedy decisions. In the first Km rounds that m rounds of random sampling are executed for each arm, samples are independent and we can use sharper tail bounds. This would help us to get better probability guarantees for the Greedy Bandit algorithm.*

Proof. Note that we can write

$$\mathbb{P}\left[\mathcal{H}_i^\delta \cap \overline{\mathcal{G}_{i,Km}^{\theta_1}}\right] = \mathbb{P}\left[\lambda_{\min}\left(\mathbf{X}(\mathcal{S}_{i,Km})^\top \mathbf{X}(\mathcal{S}_{i,Km})\right) \geq \delta, \|\hat{\beta}(\mathcal{S}_{Km,t}) - \beta_i\|_2 \geq \theta_1\right]. \quad \text{(A.7.2)}$$

Note that if $\lambda_{\min}\left(\mathbf{X}(\mathcal{S}_{i,Km})^\top \mathbf{X}(\mathcal{S}_{i,Km})\right) \geq \delta > 0$, this means that the covariance matrix is invertible. Therefore, we can write

$$
\begin{aligned}
\hat{\beta}(\mathcal{S}_{Km,t}) - \beta_i &= \Big[\mathbf{X}(\mathcal{S}_{i,Km})^\top \mathbf{X}(\mathcal{S}_{i,Km})\Big]^{-1} \mathbf{X}(\mathcal{S}_{i,Km})^\top Y(\mathcal{S}_{i,Km}) - \beta_i \\
&= \Big[\mathbf{X}(\mathcal{S}_{i,Km})^\top \mathbf{X}(\mathcal{S}_{i,Km})\Big]^{-1} \mathbf{X}(\mathcal{S}_{i,Km})^\top \left[\mathbf{X}(\mathcal{S}_{i,Km})\beta_i + \varepsilon(\mathcal{S}_{i,Km})\right] - \beta_i \\
&= \Big[\mathbf{X}(\mathcal{S}_{i,Km})^\top \mathbf{X}(\mathcal{S}_{i,Km})\Big]^{-1} \mathbf{X}(\mathcal{S}_{i,Km})^\top \varepsilon(\mathcal{S}_{i,Km})\,.
\end{aligned}
$$

To avoid clutter, we drop the term $\mathcal{S}_{i,Km}$ in equations from here onwards. Letting $M = \left[\mathbf{X}(\mathcal{S}_{i,Km})^\top \mathbf{X}(\mathcal{S}_{i,Km})\right]^{-1} \mathbf{X}(\mathcal{S}_{i,Km})$ the probability in Equation (A.7.2) turns into

$$
\begin{aligned}
\mathbb{P}\left[\mathcal{H}_i^\delta \cap \overline{\mathcal{G}_{i,Km}^{\theta_1}}\right] &= \mathbb{P}\left[\lambda_{\min}\left(\mathbf{X}^\top \mathbf{X}\right) \geq \delta, \|M\varepsilon\|_2 \geq \theta_1\right] \\
&= \mathbb{P}\left[\lambda_{\min}\left(\mathbf{X}^\top \mathbf{X}\right) \geq \delta, \sum_{j=1}^{d} |m_j^\top \varepsilon| \geq \theta_1\right] \\
&\leq \mathbb{P}\left[\lambda_{\min}\left(\mathbf{X}^\top \mathbf{X}\right) \geq \delta, \exists j \in [d], |m_j^\top \varepsilon| \geq \theta_1/\sqrt{d}\right] \\
&\leq \sum_{j=1}^{d} \mathbb{P}\left[\lambda_{\min}\left(\mathbf{X}^\top \mathbf{X}\right) \geq \delta, |m_j^\top \varepsilon| \geq \theta_1/\sqrt{d}\right] \\
&= \sum_{j=1}^{d} \mathbb{P}_{\mathbf{X}} \mathbb{P}_{\varepsilon|\mathbf{X}} \left[\lambda_{\min}\left(\mathbf{X}^\top \mathbf{X}\right) \geq \delta, |m_j^\top \varepsilon| \geq \theta_1/\sqrt{d} \mid \mathbf{X} = \mathbf{X}_0\right], \qquad \text{(A.7.3)}
\end{aligned}
$$

where in the second inequality we used a union bound. Note that in above $\mathbb{P}_{\mathbf{X}}$ means the probability distribution over the matrix $\mathbf{X}$, which can also be thought as the multi-dimensional probability distribution of p_X, or alternatively p_X^m. Now fixing $\mathbf{X} = \mathbf{X}_0$, the matrix M only depends on $\mathbf{X}_0$ and we can use the well-known Chernoff bound for subgaussian random variables to achieve

$$
\begin{aligned}
\mathbb{P}[\lambda_{\min}\left(\mathbf{X}_0^\top \mathbf{X}_0\right) \geq \delta, |m_j^\top \varepsilon| \geq \frac{\theta_1}{\sqrt{d}} \mid \mathbf{X} = \mathbf{X}_0] &= \mathbb{I}\left[\lambda_{\min}\left(\mathbf{X}_0^\top \mathbf{X}_0\right) \geq \delta\right] \mathbb{P}[|m_j^\top \varepsilon| \geq \frac{\theta_1}{\sqrt{d}} \mid \mathbf{X} = \mathbf{X}_0] \\
&\leq 2\mathbb{I}\left[\lambda_{\min}\left(\mathbf{X}_0^\top \mathbf{X}_0\right) \geq \delta\right] \exp\left\{-\frac{\theta_1^2}{2d\sigma^2\|m_j\|_2^2}\right\}
\end{aligned}
$$

Now note that when $\lambda_{\min}\left(\mathbf{X}_0^\top \mathbf{X}_0\right) \geq \delta$ we have

$$
\max_{j \in [d]} \|m_j\|_2^2 = \max\left(\operatorname{diag}\left(MM^\top\right)\right) = \max\left(\operatorname{diag}\left(\mathbf{X}^\top \mathbf{X}^{-1}\right)\right) \leq \lambda_{\max}\left(\mathbf{X}^\top \mathbf{X}^{-1}\right) \leq \frac{1}{\delta},
$$

Hence,

$$\mathbb{P}_{\varepsilon|\mathbf{X}}\left[\lambda_{\min}\left(\mathbf{X}^\top\mathbf{X}\right) \geq \delta, |m_j^\top\varepsilon| \geq \theta_1/\sqrt{d} \mid \mathbf{X} = \mathbf{X}_0\right] \leq 2\mathbb{I}\left[\lambda_{\min}\left(\mathbf{X}_0^\top\mathbf{X}_0\right) \geq \delta\right]\exp\left\{-\frac{\theta_1^2\delta}{2d\sigma^2}\right\}.$$

Putting this back in Equation (A.7.3) gives

$$\begin{aligned}\mathbb{P}\left[\mathcal{H}_i^\delta \cap \overline{\mathcal{G}_{i,Km}^{\theta_1}}\right] &\leq 2d\mathbb{P}_{\mathbf{X}}\left[\left(\lambda_{\min}\left(\mathbf{X}^\top\mathbf{X}\right)\right) \geq \delta\right]\exp\left\{-\frac{\theta_1^2\delta}{2d\sigma^2}\right\} \\ &= 2d\mathbb{P}\left\{\lambda_{\min}\left(\mathbf{X}_{1:m}^\top\mathbf{X}_{1:m}\right) \geq \delta\right\}\exp\left\{-\frac{\theta_1^2\delta}{2d\sigma^2}\right\},\end{aligned}$$

as desired. In above we use the fact that $\mathbb{P}_{\mathbf{X}}\left[\lambda_{\min}\left(\mathbf{X}^\top\mathbf{X}\right) \geq \delta\right]$ is equal to $\mathbb{P}\left\{\lambda_{\min}\left(\mathbf{X}_{1:m}^\top\mathbf{X}_{1:m}\right) \geq \delta\right\}$ as they both describe the probability that the minimum eigenvalue of a matrix derived from m random samples from p_X is not smaller than δ. □

Lemma 19. *For an arbitrary $Km + 1 \leq t \leq p - 1$ and $i \in [K]$ we have*

$$\mathbb{P}\left[\mathcal{H}_i^\delta \cap \overline{\mathcal{G}_{i,t}^{\theta_1}}\right] \leq 2d\exp\left\{-\frac{\theta_1^2\delta^2}{2d(t-(K-1)m)\sigma^2 x_{\max}^2}\right\}$$

Proof. This is an immediate consequence of Lemma 5. Replace $\chi = \theta_1, \lambda = \delta/t$ and note that $|\mathcal{S}_{i,t}| \leq t - (K-1)m$ always holds as $(K-1)m$ rounds of random sampling for arms other than i exist in algorithm. □

The next step is proving that if all arm estimates are within the ball of radius θ_1 around their true values, the minimum eigenvalue of arms in $\mathcal{K}_{opt}$ grow linearly, while sub-optimal arms are not picked by Greedy Bandit algorithm. The proof is a general extension of Lemma 4.

Lemma 20. *For each $t \geq p, i \in \mathcal{K}_{opt}$*

$$\mathbb{P}\left[\overline{\mathcal{J}_{i,t}^{\lambda_1(1-\gamma)}} \cap \left(\cap_{l=1}^K \cap_{j=Km}^{t-1} \mathcal{G}_{l,j}^{\theta_1}\right)\right] \leq d\exp\left(-D_1(\gamma)(t - m|\mathcal{K}_{sub}|)\right).$$

Furthermore, for each $t \geq Km + 1$ and $i \in \mathcal{K}_{sub}$ conditioning on the event $\cap_{l=1}^K \mathcal{G}_{l,t-1}^{\theta_1}$, arm i would not be played at time t under greedy policy.

Proof. The idea is again using concentration inequality in Lemma 14. Let $i \in \mathcal{K}_{opt}$ and recall that

$$\tilde{\Sigma}_{i,t} = \sum_{k=1}^{t} \mathbb{E}\left[X_k X_k^\top \mathbb{I}\left[X_k \in \hat{\mathcal{R}}_{i,k}^{\pi}\right] \mid \mathcal{H}_{k-1}^{-}\right]$$
$$\hat{\Sigma}_{i,t} = \sum_{k=1}^{t} X_k X_k^\top \mathbb{I}\left[X_k \in \hat{\mathcal{R}}_{i,k}^{\pi}\right],$$

denote the expected and sample covariance matrices of arm i at time t respectively. The aim is deriving an upper bound on the probability that minimum eigenvalue of $\hat{\Sigma}_{i,t}$ is less than the threshold $t\lambda_1(1-\gamma) - m|\mathcal{K}_{sub}|$. Note that $\hat{\Sigma}_{i,t}$ consists of two different types of terms: 1) random sampling rounds $1 \le k \le Km$ and 2) greedy action rounds $Km+1 \le k \le t$. We analyze these two types separately as following:

- $k \le Km$. Note that during the first Km periods, each arm receives m random samples from the distribution p_X and therefore using concavity of the function $\lambda_{\min}(\cdot)$ we have

$$\begin{aligned}
&\lambda_{\min}\left(\sum_{k=1}^{Km} \mathbb{E}\left[X_k X_k^\top \mathbb{I}\left[X_k \in \hat{\mathcal{R}}_{i,k}^{\pi}\right] \mid \mathcal{H}_{k-1}^{-}\right]\right) \\
&\qquad \ge m\lambda_{\min}\mathbb{E}\left(XX^\top\right) \\
&\qquad \ge m\lambda_{\min}\left(\sum_{j\in\mathcal{K}_{opt}} \mathbb{E}\left(XX^\top \mathbb{I}\left(X^\top\beta_j > \max_{l\ne j} X^\top\beta_l + h\right)\right)\right) \\
&\qquad \ge m|\mathcal{K}_{opt}|\lambda_1,
\end{aligned}$$

  where X is a random sample from distribution p_X.

- $k \ge Km+1$. If $\mathcal{G}_{l,j}^{\theta_1}$ holds for all $l \in [K]$, then

$$\mathbb{E}\left[X_k X_k^\top \mathbb{I}\left(X_k \in \hat{\mathcal{R}}_{i,k}^{\pi}\right) \mid \mathcal{H}_{k-1}^{-}\right] \succeq \mathbb{E}\left[XX^\top \mathbb{I}\left(X^\top \hat{\beta}(\mathcal{S}_{i,k}) > \max_{l\ne i} X^\top \hat{\beta}(\mathcal{S}_{l,k})\right)\right] \succeq \lambda_1 \mathbf{I}\,.$$

  The reason is very simple; basically having $\cap_{l=1}^{K}\mathcal{G}_{l,j}^{\theta_1}$ means that $\|\hat{\beta}(\mathcal{S}_{l,k}) - \beta_l\| < \theta_1$ and therefore for each $\mathbf{x}$ satisfying $\mathbf{x}^\top\beta_i \ge \max_{l\ne i}\mathbf{x}^\top\beta_l + h$, using two Cauchy-Schwarz inequalities we can write

$$\mathbf{x}^\top\hat{\beta}(\mathcal{S}_{i,j}) - \mathbf{x}^\top\hat{\beta}(\mathcal{S}_{l,j}) > \mathbf{x}^\top(\beta_i - \beta_l) - 2x_{\max}\theta_1 = \mathbf{x}^\top(\beta_i - \beta_l) - h \ge 0,$$

for each $l \neq i$. Therefore, by taking a maximum over l we obtain $\mathbf{x}^\top \hat{\beta}(\mathcal{S}_{i,j}) - \max_{i\neq l} \mathbf{x}^\top \hat{\beta}(\mathcal{S}_{l,j}) > 0$. Hence,

$$\begin{aligned}\mathbb{E}\left[X_k X_k^\top \mathbb{I}\left(X_k^\top \hat{\beta}(\mathcal{S}_{i,k}) > \max_{l\neq i} X_k^\top \hat{\beta}(\mathcal{S}_{l,j})\right) \mid \mathcal{H}_{k-1}^-\right] &\succeq \mathbb{E}\left[XX^\top \mathbb{I}\left(X^\top \beta_i > \max_{l\neq i} X^\top \beta_l + h\right)\right] \\ &\succeq \lambda_1 \mathbf{I},\end{aligned}$$

using Assumption 4, which holds for all optimal arms, i.e, $i \in \mathcal{K}_{opt}$.

Putting these two results together and using concavity of $\lambda_{\min}(\cdot)$ over positive semi-definite matrices we have

$$\begin{aligned}\lambda_{\min}\left(\tilde{\Sigma}_{i,t}\right) &= \lambda_{\min}\left(\sum_{k=1}^{t} \mathbb{E}\left[X_k X_k^\top \mathbb{I}\left[X_k \in \hat{\mathcal{R}}_{i,k}^\pi\right] \mid \mathcal{H}_{k-1}^-\right]\right) \\ &\geq \sum_{k=1}^{Km} \lambda_{\min}\left(\mathbb{E}\left[X_k X_k^\top \mathbb{I}\left[X_k \in \hat{\mathcal{R}}_{i,k}^\pi\right] \mid \mathcal{H}_{k-1}^-\right]\right) \\ &\quad + \sum_{k=Km+1}^{t} \lambda_{\min}\left(\mathbb{E}\left[X_k X_k^\top \mathbb{I}\left[X_k \in \hat{\mathcal{R}}_{i,k}^\pi\right] \mid \mathcal{H}_{k-1}^-\right]\right) \\ &\geq m|\mathcal{K}_{opt}|\lambda_1 + (t - Km)\lambda_1 = (t - m|\mathcal{K}_{sub}|)\lambda_1.\end{aligned}$$

Now the rest of the argument is similar to Lemma 4. Note that in the proof of Lemma 4, we simply put $\gamma = 0.5$, however if use an arbitrary $\gamma \in (0,1)$ together with $X_k X_k^\top \preceq x_{\max}^2 \mathbf{I}$, which is the result of Cauchy-Schwarz inequality, then Lemma 14 implies that

$$\begin{aligned}&\mathbb{P}\left[\lambda_{\min}\left(\hat{\Sigma}_{i,t}\right) \leq (t - m|\mathcal{K}_{sub}|)\lambda_1(1-\gamma) \text{ and } \lambda_{\min}\left(\tilde{\Sigma}_{i,t}\right) \geq (t - m|\mathcal{K}_{sub}|)\lambda_1\right] \\ &\qquad \leq d \exp\left(-D_1(\gamma)(t - m|\mathcal{K}_{sub}|)\right).\end{aligned}$$

The second event inside the probability event can be removed, as it always holds under $\left(\cap_{l=1}^K \cap_{j=Km}^{t-1} \mathcal{G}_{l,j}^{\theta_1}\right)$. The first event also can be translated to $\overline{\mathcal{J}_{i,t}^{\lambda_1(1-\gamma)}}$ and therefore for all $i \in \mathcal{K}_{opt}$ we have

$$\mathbb{P}\left[\overline{\mathcal{J}_{i,t}^{\lambda_1(1-\gamma)}} \cap \left(\cap_{l=1}^K \cap_{j=Km}^{t-1} \mathcal{G}_{l,j}^{\theta_1}\right)\right] \leq d \exp\left(-D_1(\gamma)(t - m|\mathcal{K}_{sub}|)\right),$$

as desired.

For a sub-optimal arm $i \in \mathcal{K}_{sub}$ using Assumption 4, for each $\mathbf{x} \in \mathcal{X}$ there exist $l \in [K]$

such that $\mathbf{x}^\top \beta_i \le \mathbf{x}^\top \beta_l - h$ and as a result conditioning on $\cap_{l=1}^K \mathcal{G}_{l,t-1}^{\theta_1}$ by using a Cauchy-Schwarz inequality we have

$$\mathbf{x}^\top \hat{\beta}(\mathcal{S}_{l,t-1}) - \mathbf{x}^\top \hat{\beta}(\mathcal{S}_{i,t-1}) > \mathbf{x}^\top (\beta_l - \beta_i) - 2x_{\max}\theta_1 = \mathbf{x}^\top (\beta_l - \beta_i) - h > 0.$$

This implies that $i \notin \arg\max_{l \in [K]} \mathbf{x}^\top \hat{\beta}(\mathcal{S}_{l,t-1})$ and therefore arm i is not played for $\mathbf{x}$ at time t (Note that once Km rounds of random sampling are finished the algorithm executes greedy algorithm). As this result holds for all choices of $\mathbf{x} \in \mathcal{X}$, arm i becomes sub-optimal at time t, as desired. □

Here, we state the final Lemma, which bounds the probability that the event $\overline{\mathcal{G}_{i,t}^{\theta_1}}$ occurs whenever $\mathcal{J}_{i,t}^{\lambda_1(1-\gamma)}$ holds for any $t \ge p$.

Lemma 21. *For each $t \ge p, i \in [K]$*

$$\mathbb{P}\left[\ \overline{\mathcal{G}_{i,t}^{\theta_1}} \cap \mathcal{J}_{i,t}^{\lambda_1(1-\gamma)}\right] \le 2d \exp\left(-D_2(\gamma)(t - m|\mathcal{K}_{sub}|)\right)\ .$$

Proof. This is again obvious using Lemma 5. □

Now we are ready to prove Theorems 2 and 4. As the proofs of these two theorems are very similar we state and prove a lemma that implies both theorems.

Lemma 22. *Let Assumption and 4 hold. Suppose that Greedy Bandit algorithm with m-rounds of forced sampling in the beginning is executed. Let $\gamma \in (0,1), \delta > 0, p \ge Km + 1$. Suppose that $\mathcal{W}$ is an event which can be decomposed as $\mathcal{W} = \cap_{t \ge p} \mathcal{W}_t$, then event*

$$\left(\cap_{i=1}^K \cap_{t \ge Km} \mathcal{G}_{i,t}^{\theta_1}\right) \cap \mathcal{W}$$

holds with probability at least

$$\begin{aligned}1 - &\left(\mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m}) \ge \delta\right]\right)^K + 2Kd\ \mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m}) \ge \delta\right] \exp\left\{-\frac{h^2\delta}{8d\sigma^2 x_{\max}^2}\right\} \\ &+ \sum_{j=Km+1}^{p-1} 2d \exp\left\{-\frac{h^2\delta^2}{8d(j-(K-1)m)\sigma^2 x_{\max}^4}\right\} \\ &+ \sum_{t \ge p} \mathbb{P}\left[\left(\cap_{i=1}^K \cap_{k=Km}^{t-1} \mathcal{G}_{i,k}^{\theta_1}\right) \cap \left(\overline{\mathcal{G}_{\pi_t,t}^{\theta_1}} \cup \overline{\mathcal{W}_t}\right)\right]\ .\end{aligned}$$

In above, $\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m})$ *denotes the minimum eigenvalue of a matrix obtained from* m *random samples from the distribution* p_X *and constants are defined in Equations* (A.3.1) *and* (A.3.2)*.*

Proof. One important property to note is the following result on the events:

$$\left\{ \left(\cap_{i=1}^K \mathcal{G}_{i,t-1}^{\theta_1} \right) \cap \left(\cup_{i=1}^K \overline{\mathcal{G}_{i,t}^{\theta_1}} \right) \right\} = \left\{ \left(\cap_{i=1}^K \mathcal{G}_{i,t-1}^{\theta_1} \right) \cap \overline{\mathcal{G}_{\pi_t,t}^{\theta_1}} \right\} . \tag{A.7.4}$$

The reason is that the estimates for arms other than arm π_t do not change at time t, meaning that for each $i \neq \pi_t, \mathcal{G}_{i,t-1}^{\theta_1} = \mathcal{G}_{i,t}^{\theta_1}$. Therefore, the above equality is obvious. This observation comes handy when we want to avoid using a union bound over different arms for the probability of undesired event. For deriving a lower bound on the probability of desired event we have

$$\mathbb{P}\left[\left(\cap_{i=1}^K \cap_{t \geq Km} \mathcal{G}_{i,t}^{\theta_1} \right) \cap \mathcal{W} \right] = 1 - \mathbb{P}\left[\left(\cup_{i=1}^K \cup_{t \geq Km} \overline{\mathcal{G}_{i,t}^{\theta_1}} \right) \cup \overline{\mathcal{W}} \right] .$$

Therefore, we can write

$$\mathbb{P}\left[\left(\cup_{i=1}^K \cup_{t \geq Km} \overline{\mathcal{G}_{i,t}^{\theta_1}} \right) \cup \overline{\mathcal{W}} \right] \leq \mathbb{P}\left[\cup_{i=1}^K \overline{\mathcal{H}_i^\delta} \right] + \mathbb{P}\left[\left(\cap_{i=1}^K \mathcal{H}_i^\delta \right) \cap \left[\left(\cup_{i=1}^K \cup_{t \geq Km} \overline{\mathcal{G}_{i,t}^{\theta_1}} \right) \cup \overline{\mathcal{W}} \right] \right] .$$

The first term is equal to $1 - \left(\mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m}) \geq \delta \right] \right)^K$. The reason is simple; probability of each $\mathcal{H}_i^\delta, i \in [K]$ is given by $\mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m}) \geq \delta \right]$ and these events are all independent due to the random sampling. Therefore, the probability that at least one of them does not happen is given by the mentioned expression. In addition, the probability of the second event can be upper bounded by

$$\begin{aligned}
&\mathbb{P}\left[\left(\cap_{i=1}^K \mathcal{H}_i^\delta \right) \cap \left[\left(\cup_{i=1}^K \cup_{t \geq Km} \overline{\mathcal{G}_{i,t}^{\theta_1}} \right) \cup \overline{\mathcal{W}} \right] \right] \\
&\quad \leq \sum_{l=1}^K \mathbb{P}\left[\left(\cap_{i=1}^K \mathcal{H}_i^\delta \right) \cap \overline{\mathcal{G}_{l,Km}^{\theta_1}} \right] + \mathbb{P}\left[\left(\cap_{i=1}^K \mathcal{H}_i^\delta \right) \cap \left(\cap_{i=1}^K \mathcal{G}_{i,Km}^{\theta_1} \right) \cap \left[\left(\cup_{i=1}^K \cup_{t \geq Km} \overline{\mathcal{G}_{i,t}^{\theta_1}} \right) \cup \overline{\mathcal{W}} \right] \right] \\
&\quad \leq \sum_{l=1}^K \mathbb{P}\left[\mathcal{H}_l^\delta \cap \overline{\mathcal{G}_{l,Km}^{\theta_1}} \right] + \mathbb{P}\left[\left(\cap_{i=1}^K \mathcal{H}_i^\delta \right) \cap \left(\cap_{i=1}^K \mathcal{G}_{i,Km}^{\theta_1} \right) \cap \left[\left(\cup_{i=1}^K \cup_{t \geq Km} \overline{\mathcal{G}_{i,t}^{\theta_1}} \right) \cup \overline{\mathcal{W}} \right] \right] \\
&\quad \leq 2Kd\mathbb{P}\left\{ \lambda_{\min}\left(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m} \right) \geq \delta \right\} \exp\left\{ -\frac{\theta_1^2 \delta}{2d\sigma^2} \right\} \\
&\qquad\qquad + \mathbb{P}\left[\left(\cap_{i=1}^K \mathcal{H}_i^\delta \right) \cap \left(\cap_{i=1}^K \mathcal{G}_{i,Km}^{\theta_1} \right) \cap \left[\left(\cup_{i=1}^K \cup_{t \geq Km} \overline{\mathcal{G}_{i,t}^{\theta_1}} \right) \cup \overline{\mathcal{W}} \right] \right] ,
\end{aligned}$$

where we used Lemma 18 together with a union bound. For finding an upper bound on the the second probability, we treat terms $t \in [Km+1, p-1]$ and $t \geq p$ differently. Basically, for the first interval we have guarantees when $\cap_{i=1}^{K} \mathcal{H}_i^{\delta}$ holds (Lemma 19) and for the second interval the guarantee comes from having the event $\cap_{l=1}^{K} \cap_{j=Km}^{t-1} \mathcal{G}_{l,j}^{\theta_1}$ (Lemma 20). Following this path leads to

$$
\begin{aligned}
&\mathbb{P}\left[\left(\cap_{i=1}^{K}\mathcal{H}_i^{\delta}\right) \cap \left(\cap_{i=1}^{K}\mathcal{G}_{i,Km}^{\theta_1}\right) \cap \left[\left(\cup_{i=1}^{K} \cup_{t\geq Km} \overline{\mathcal{G}_{i,t}^{\theta_1}}\right) \cup \overline{\mathcal{W}}\right]\right] \\
&\quad \leq \sum_{t=Km+1}^{p-1} \mathbb{P}\left[\left(\cap_{i=1}^{K}\mathcal{H}_i^{\delta}\right) \cap \left(\cap_{i=1}^{K} \cap_{k=Km}^{t-1} \mathcal{G}_{i,k}^{\theta_1}\right) \cap \left(\cup_{i=1}^{K}\overline{\mathcal{G}_{i,t}^{\theta_1}}\right)\right] \\
&\quad + \sum_{t\geq p} \mathbb{P}\left[\left(\cap_{i=1}^{K}\mathcal{H}_i^{\delta}\right) \cap \left(\cap_{i=1}^{K} \cap_{k=Km}^{t-1} \mathcal{G}_{i,k}^{\theta_1}\right) \cap \left(\cup_{i=1}^{K}\overline{\mathcal{G}_{i,t}^{\theta_1}} \cup \overline{\mathcal{W}_t}\right)\right] \\
&\quad \leq \sum_{t=Km+1}^{p-1} \mathbb{P}\left[\left(\cap_{i=1}^{K}\mathcal{H}_i^{\delta}\right) \cap \left(\cap_{i=1}^{K}\mathcal{G}_{i,t-1}^{\theta_1}\right) \cap \overline{\mathcal{G}_{\pi_t,t}^{\theta_1}}\right] \\
&\qquad\qquad + \sum_{t\geq p} \mathbb{P}\left[\left(\cap_{i=1}^{K}\mathcal{H}_i^{\delta}\right) \cap \left(\cap_{i=1}^{K} \cap_{k=Km}^{t-1} \mathcal{G}_{i,k}^{\theta_1}\right) \cap \left(\overline{\mathcal{G}_{\pi_t,t}^{\theta_1}} \cup \overline{\mathcal{W}_t}\right)\right] \\
&\quad \leq \sum_{t=Km+1}^{p-1} \mathbb{P}\left[\left(\cap_{i=1}^{K}\mathcal{H}_i^{\delta}\right) \cap \overline{\mathcal{G}_{\pi_t,t}^{\theta_1}}\right] + \sum_{t\geq p} \mathbb{P}\left[\left(\cap_{i=1}^{K} \cap_{k=Km}^{t-1} \mathcal{G}_{i,k}^{\theta_1}\right) \cap \left(\overline{\mathcal{G}_{\pi_t,t}^{\theta_1}} \cup \overline{\mathcal{W}_t}\right)\right].
\end{aligned}
$$

using Equation (A.7.4) and carefully breaking down the event $\left[\left(\cup_{i=1}^{K} \cup_{t\geq Km} \overline{\mathcal{G}_{i,t}^{\theta_1}}\right) \cup \overline{\mathcal{W}}\right]$. Note that by using the second part of Lemma 20, if the event $\cap_{i=1}^{K}\mathcal{G}_{i,t-1}^{\theta_1}$ holds, then π is equal to one of the elements in $\mathcal{K}_{opt}$ and sub-optimal arms in $\mathcal{K}_{sub}$ will not be pulled. Therefore, with further reduction the first term is upper bounded by

$$
\begin{aligned}
&\sum_{t=Km+1}^{p-1} \sum_{l\in\mathcal{K}_{opt}} \mathbb{P}\left[\pi_t = l\right]\mathbb{P}\left[\left(\cap_{i=1}^{K}\mathcal{H}_i^{\delta}\right) \cap \overline{\mathcal{G}_{l,t}^{\theta_1}}\right] \\
&\qquad\qquad\qquad \leq \sum_{t=Km+1}^{p-1} \sum_{l\in\mathcal{K}_{opt}} \mathbb{P}\left[\pi_t = l\right] 2d \exp\left\{-\frac{\theta_1^2\delta^2}{2d(t-(K-1)m)\sigma^2 x_{\max}^2}\right\} \\
&\qquad\qquad\qquad \leq \sum_{t=Km+1}^{p-1} 2d \exp\left\{-\frac{\theta_1^2\delta^2}{2d(t-(K-1)m)\sigma^2 x_{\max}^2}\right\},
\end{aligned}
$$

using uniform upper bound provided in Lemma 19 and $\sum_{l\in\mathcal{K}_{opt}} \mathbb{P}\left[\pi_t = l\right] = 1$. This concludes the proof. □

Proof of Theorem 2. The proof consists of using Lemma 22. Basically, if we know that the events $\mathcal{G}_{i,t}^{\theta_1}$ for $i \in [K]$ and $t \geq Km$ all hold, we have derived a lower bound on the probability that greedy succeeds. The reason is pretty simple here, if the distance of true parameters β_i and $\hat{\beta}_i$ is at most θ_1 for each t, we can easily ensure that the minimum eigenvalue of covariance matrices of optimal arms are growing linearly, and sub-optimal arms remain sub-optimal for all $t \geq Km+1$ using Lemma 20. Therefore, we can prove the optimality of Greedy Bandit algorithm and also establish its logarithmic regret. Therefore, in this case we need not use any $\mathcal{W}$ in Lemma 22, we simply put $\mathcal{W}_t = \mathcal{W} = \Omega$, where Ω is the whole probability space. Then we have

$$\begin{aligned}
\mathbb{P}\left[\cap_{i=1}^{K} \cap_{t \geq Km} \mathcal{G}_{i,t}^{\theta_1}\right] \geq 1 - &\left(\mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m}) \geq \delta\right]\right)^K \\
&+ 2Kd\, \mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m}) \geq \delta\right] \exp\left\{-\frac{h^2\delta}{8d\sigma^2 x_{\max}^2}\right\} \\
&+ \sum_{j=Km+1}^{p-1} 2d \exp\left\{-\frac{h^2\delta^2}{8d(j-(K-1)m)\sigma^2 x_{\max}^4}\right\} \\
&+ \sum_{t \geq p} \mathbb{P}\left[\left(\cap_{i=1}^{K} \cap_{k=Km}^{t-1} \mathcal{G}_{i,k}^{\theta_1}\right) \cap \overline{\mathcal{G}_{\pi_t,t}^{\theta_1}}\right].
\end{aligned}$$

The upper bound on the last term can be derived as following

$$\begin{aligned}
&\sum_{t \geq p} \mathbb{P}\left[\left(\cap_{i=1}^{K} \cap_{k=Km}^{t-1} \mathcal{G}_{i,k}^{\theta_1}\right) \cap \left(\cup_{i=1}^{K} \overline{\mathcal{G}_{\pi_t,t}^{\theta_1}}\right)\right] \\
&\quad = \sum_{t \geq p} \sum_{l \in \mathcal{K}_{opt}} \mathbb{P}[\pi_t = l] \mathbb{P}\left[\left(\cap_{i=1}^{K} \cap_{k=Km}^{t-1} \mathcal{G}_{i,k}^{\theta_1}\right) \cap \left(\cup_{i=1}^{K} \overline{\mathcal{G}_{l,t}^{\theta_1}}\right)\right] \\
&\quad \leq \sum_{t \geq p} \sum_{l \in \mathcal{K}_{opt}} \mathbb{P}[\pi_t = l] \left\{\mathbb{P}\left[\overline{\mathcal{J}_{l,t}^{\lambda_1(1-\gamma)}} \cap \left(\cap_{i=1}^{K} \cap_{j=Km}^{t-1} \mathcal{G}_{i,j}^{\theta_1}\right)\right] + \mathbb{P}\left[\overline{\mathcal{G}_{l,t}^{\theta_1}} \cap \mathcal{J}_{l,t}^{\lambda_1(1-\gamma)}\right]\right\},
\end{aligned}$$

which by using Lemmas 20 and 21 can be upper bounded by

$$\begin{aligned}
&\sum_{t \geq p} \sum_{l \in \mathcal{K}_{opt}} \mathbb{P}[\pi_t = l] \left\{d \exp\left(-D_1(\gamma)(t - m|\mathcal{K}_{sub}|)\right) + 2d \exp\left(-D_2(\gamma)(t - m|\mathcal{K}_{sub}|)\right)\right\} \\
&\qquad\qquad = \sum_{t \geq p} \exp\left(-D_1(\gamma)(t - m|\mathcal{K}_{sub}|)\right) + \sum_{t \geq p} 2d \exp\left(-D_2(\gamma)(t - m|\mathcal{K}_{sub}|)\right) \\
&\qquad\qquad = \frac{d \exp\left(-D_1(\gamma)(p - m|\mathcal{K}_{sub}|)\right)}{1 - \exp(-D_1(\gamma))} + \frac{2d \exp\left(-D_2(\gamma)(p - |\mathcal{K}_{sub}|)\right)}{1 - \exp(-D_2(\gamma))}.
\end{aligned}$$

Summing up all these term yields the desired upper bound. Now note that this upper bound is algorithm-independent and holds for all values of $\gamma \in (0,1), \delta \geq 0$, and $p \geq Km$ and therefore we can take the supremum over these values for our desired event (or infimum over undesired event). This concludes the proof. $\square$

For proving Theorem 4 the steps are very similar, the only difference is that the desired event happens if all events $\mathcal{G}_{i,t}^{\theta_1}$, $i \in [K], t \geq Km$ hold, and in addition to that, events $\mathcal{F}_{i,t}^{\lambda}, i \in [K], t \geq t_0$ all need to hold for some $\lambda > \lambda_0/4$. Recall that in Theorem 4, $\mathcal{K}_{sub} = \emptyset$ and therefore we can use the notations $\mathcal{J}$ and $\mathcal{F}$ interchangeably. For Greedy-First, we define $\mathcal{W} = \cap_{i \in [K]} \cap_{t \geq p} \mathcal{F}_{i,t}^{\lambda}$ for some λ. This basically, means we need to take $\mathcal{W}_t = \cap_{i \in [K]} \mathcal{F}_{i,t}^{\lambda}$ for some λ.

Proof of Theorem 4. The proof is very similar to proof of Theorem 2. For arbitrary γ, δ, p we want to derive a bound on the probability of the event

$$\mathbb{P}\left[\left(\cap_{i=1}^{K} \cap_{t \geq Km} \mathcal{G}_{i,t}^{\theta_1}\right) \cap \left(\cap_{i=1}^{K} \cap_{t \geq p} \mathcal{F}_{i,t}^{\lambda_1(1-\gamma)}\right)\right].$$

Note that if $p \leq t_0$ and $\gamma \leq 1 - \lambda_0/(4\lambda_1)$, then having events $\mathcal{F}_{i,t}^{\lambda_1(1-\gamma)}, i \in [K], t \geq p$ implies that the events $\mathcal{F}_{i,t}^{\lambda_0/4}, i \in [K], t \geq t_0$ all hold. In other words, Greedy-First does not switch to the exploration-based algorithm and is able to achieve logarithmic regret. Let us substitute $\mathcal{W}_t = \cap_{i=1}^{K} \mathcal{F}_{i,t}^{\lambda_1(1-\gamma)}$ which implies that $\mathcal{W} = \cap_{i=1}^{K} \cap_{t \geq p} \mathcal{F}_{i,t}^{\lambda_1(1-\gamma)}$. Lemma 22 can be used to establish a lower bound on the probability of this event as

$$\begin{aligned}
&\mathbb{P}\left[\left(\cap_{i=1}^{K} \cap_{t \geq Km} \mathcal{G}_{i,t}^{\theta_1}\right) \cap \left(\cap_{i=1}^{K} \cap_{t \geq p} \mathcal{F}_{i,t}^{\lambda_1(1-\gamma)}\right)\right] \\
&\qquad \geq 1 - \left(\mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m}) \geq \delta\right]\right)^K \\
&\qquad + 2Kd\ \mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m}) \geq \delta\right] \exp\left\{-\frac{h^2\delta}{8d\sigma^2 x_{\max}^2}\right\} \\
&\qquad + \sum_{j=Km+1}^{p-1} 2d \exp\left\{-\frac{h^2\delta^2}{8d(j-(K-1)m)\sigma^2 x_{\max}^4}\right\} \\
&\qquad + \sum_{t \geq p} \mathbb{P}\left[\left(\cap_{i=1}^{K} \cap_{k=Km}^{t-1} \mathcal{G}_{i,k}^{\theta_1}\right) \cap \left(\overline{\mathcal{G}_{\pi_t,t}^{\theta_1}} \cup \left(\overline{\cap_{i=1}^{K} \mathcal{F}_{i,t}^{\lambda_1(1-\gamma)}}\right)\right)\right].
\end{aligned}$$

Hence, we only need to derive an upper bound on the last term. By expanding this based on the value of π_t we have

$$\begin{aligned}
&\sum_{t\geq p}\mathbb{P}\left[\left(\cap_{i=1}^{K}\cap_{k=Km}^{t-1}\mathcal{G}_{i,k}^{\theta_1}\right)\cap\left(\overline{\mathcal{G}_{\pi_t,t}^{\theta_1}}\cup\left(\overline{\cap_{i=1}^{K}\mathcal{F}_{i,t}^{\lambda_1(1-\gamma)}}\right)\right)\right]\\
&\quad=\sum_{t\geq p}\sum_{l=1}^{K}\mathbb{P}[\pi_t=l]\mathbb{P}\left[\left(\cap_{i=1}^{K}\cap_{k=Km}^{t-1}\mathcal{G}_{i,k}^{\theta_1}\right)\cap\left(\overline{\mathcal{G}_{l,t}^{\theta_1}}\cup\left(\cup_{i=1}^{K}\overline{\mathcal{F}_{i,t}^{\lambda_1(1-\gamma)}}\right)\right)\right]\\
&\quad\leq\sum_{t\geq p}\sum_{l=1}^{K}\mathbb{P}[\pi_t=l]\left\{\sum_{w=1}^{K}\left(\mathbb{P}\left[\left(\cap_{i=1}^{K}\cap_{j=Km}^{t-1}\mathcal{G}_{i,j}^{\theta_1}\right)\cap\overline{\mathcal{F}_{w,t}^{\lambda_1(1-\gamma)}}\right]\right)+\mathbb{P}\left[\,\overline{\mathcal{G}_{l,t}^{\theta_1}}\cap\mathcal{F}_{l,t}^{\lambda_1(1-\gamma)}\right]\right\},
\end{aligned}$$

using a union bound and the fact that the space $\overline{\mathcal{F}_{l,t}^{\lambda_1(1-\gamma)}}$ has already been included in the first term, so its complement can be included in the second term. Now, using Lemmas 20 and 21 this can be upper bounded by

$$\begin{aligned}
&\sum_{t\geq p}\sum_{l\in\mathcal{K}_{opt}}\mathbb{P}[\pi_t=l]\left\{Kd\exp(-D_1(\gamma)t)+2d\exp(-D_2(\gamma)t)\right\}\\
&\qquad\qquad=\sum_{t\geq p}Kd\exp(-D_1(\gamma)t)+\sum_{t\geq p}2d\exp(-D_2(\gamma)t)\\
&\qquad\qquad=\frac{Kd\exp(-D_1(\gamma)p)}{1-\exp(-D_1(\gamma))}+\frac{2d\exp(-D_2(\gamma)p)}{1-\exp(-D_2(\gamma))}.
\end{aligned}$$

As mentioned earlier, we can take supremum on parameters p,γ,δ as long as they satisfy $p\leq t_0,\gamma\leq 1-\lambda_0/(4\lambda_1)$, and $\delta>0$. They would lead to the same result only with the difference that the infimum over L should be replaced by L' and these two functions satisfy

$$L'(\gamma,\delta,p)=L(\gamma,\delta,p)+(K-1)\frac{d\exp(-D_1(\gamma)p)}{1-\exp(-D_1(\gamma))},$$

which yields the desired result. □

Proof of Corollary 1. We want to use the result of Theorem 2. In this theorem, let us substitute $\gamma=0.5,p=Km+1$, and $\delta=0.5\lambda_1 m|\mathcal{K}_{opt}|$. After this substitution, Theorem 2 implies that the Greedy Bandit algorithm succeeds with probability at least

$$\begin{aligned}
&\mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m}) \geq 0.5\lambda_1 m|\mathcal{K}_{opt}|\right]^K \\
&\qquad - 2Kd\ \mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m}) \geq 0.5\lambda_1 m|\mathcal{K}_{opt}|\right] \exp\left\{-\frac{0.5h^2\lambda_1 m|\mathcal{K}_{opt}|}{8d\sigma^2 x_{\max}^2}\right\} \\
&\qquad - \frac{d\exp\left\{-D_1(0.5)(Km+1-m|\mathcal{K}_{sub}|)\right\}}{1-\exp\left\{-D_1(0.5)\right\}} \\
&\qquad - \frac{2d\exp\left\{-D_2(0.5)(Km+1-m|\mathcal{K}_{sub}|)\right\}}{1-\exp\left\{-D_2(0.5)\right\}}.
\end{aligned}$$

For deriving a lower bound on the first term let us use the concentration inequality in Lemma 14. Note that here the samples are drawn i.i.d. from the same distribution p_X. Therefore, by applying this Lemma we have

$$\begin{aligned}
&\mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m}) \leq 0.5\lambda_1 m|\mathcal{K}_{opt}|) \ \text{ and } \ \mathbb{E}[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m})] \geq \lambda_1 m|\mathcal{K}_{opt}|\right] \\
&\qquad \leq d\left(\frac{e^{-0.5}}{0.5^{0.5}}\right)^{\lambda_1 m|\mathcal{K}_{opt}|/x_{\max}^2} \\
&\qquad = d\exp\left\{-\frac{\lambda_1 m|\mathcal{K}_{opt}|}{x_{\max}^2}\left(-0.5-0.5\log(0.5)\right)\right\} \\
&\qquad \geq d\exp\left(-0.153\frac{\lambda_1 m|\mathcal{K}_{opt}|}{x_{\max}^2}\right).
\end{aligned}$$

Note that the second event, i.e. $\mathbb{E}[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m})] \geq \lambda_1 m|\mathcal{K}_{opt}|$ happens with probability one. This is true according to

$$\begin{aligned}
\mathbb{E}[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m})] &= \mathbb{E}[\lambda_{\min}(\sum_{l=1}^m X_l X_l^\top)] \\
&\geq \mathbb{E}[\sum_{l=1}^m \lambda_{\min}(X_l X_l^\top)] = \sum_{l=1}^m \mathbb{E}[\lambda_{\min}(X_l X_l^\top)] \\
&= m\mathbb{E}[\lambda_{\min}(XX^\top)],
\end{aligned}$$

where $X \sim p_X$ and the inequality is true according to the Jensen's inequality for the concave function $\lambda_{\min}(\cdot)$. Now note that, this expectation can be bounded by

$$\begin{aligned}\mathbb{E}[\lambda_{\min}(XX^\top)] &\geq \mathbb{E}\left[\lambda_{\min}\left(\sum_{i=1}^{K} XX^\top \mathbb{I}(X^\top\beta_i \geq \max_{j\neq i} X^\top\beta_j + h)\right)\right] \\ &\geq \sum_{i=1}^{K} \mathbb{E}\left[\lambda_{\min}\left(XX^\top \mathbb{I}(X^\top\beta_i \geq \max_{j\neq i} X^\top\beta_j + h)\right)\right] \\ &\geq |\mathcal{K}_{opt}|\lambda_1,\end{aligned}$$

according to Assumption 4 and another use of Jensen's inequality for the function $\lambda_{\min}(\cdot)$. Note that this part of proof was very similar to Lemma 20. Thus, with a slight modification we get

$$\mathbb{P}\left[\lambda_{\min}(\mathbf{X}_{1:m}^\top \mathbf{X}_{1:m}) \geq 0.5\lambda_1 m|\mathcal{K}_{opt}|\right] \geq 1 - d\exp\left(-0.153\frac{\lambda_1 m|\mathcal{K}_{opt}|}{x_{\max}^2}\right).$$

After using this inequality together with the inequality $(1-x)^K \geq 1 - Kx$, and after replacing values of $D_1(0.5)$ and $D_2(0.5)$, the lower bound on the probability of success of Greedy Bandit reduces to

$$\begin{aligned}&1 - Kd\exp\left(\frac{-0.153\lambda_1 m|\mathcal{K}_{opt}|}{x_{\max}^2}\right) - 2Kd\exp\left(-\frac{h^2\lambda_1 m|\mathcal{K}_{opt}|}{16d\sigma^2 x_{\max}^2}\right) \\ &- d\sum_{l=(K-|\mathcal{K}_{sub}|)m+1}^{\infty} \exp\left(\frac{-0.153\lambda_1}{x_{\max}^2}l\right) - 2d\sum_{l=(K-|\mathcal{K}_{sub}|)m+1}^{\infty} \exp\left(-\frac{\lambda_1^2 h^2}{32d\sigma^2 x_{\max}^4}l\right).\end{aligned}$$

In above we used the expansion $1/(1-x) = \sum_{l=0}^{\infty} x^l$. In order to finish the proof note that by a Cauchy-Schwarz inequality $\lambda_1 \leq x_{\max}^2$. Furthermore, $K - |\mathcal{K}_{sub}| = |\mathcal{K}_{opt}|$ and therefore the above bound is greater than or equal to

$$\begin{aligned}1 - Kd\sum_{l=m|\mathcal{K}_{opt}|}^{\infty} \exp\left(\frac{-0.153\lambda_1}{x_{\max}^2}l\right) - 2Kd\sum_{l=m|\mathcal{K}_{opt}|}^{\infty} \exp\left(-\frac{\lambda_1^2 h^2}{32d\sigma^2 x_{\max}^4}l\right) \\ \geq 1 - \frac{3Kd\exp(-D_{\min} m|\mathcal{K}_{opt}|)}{1-\exp(-D_{\min})},\end{aligned}$$

as desired.

□

Proof of Corollary 2. Proof of this corollary is very similar to the previous corollary. Extra conditions of the corollary ensure that both $\gamma = 0.5, p = Km + 1$ lie on their accepted region. For avoiding clutter, we skip the proof.

$\square$

Appendix B

Supplementary Materials for Chapter 3

B.1 Proofs

B.1.1 Proof of Theorem 5

First, we will discuss three main steps that are needed for the proof.

Step 1: We show an upper bound for the sum of squared errors for all $(i,t) \in \mathcal{O}$ in terms of the regularization parameter λ, rank of $\mathbf{L}^*$, $\|\mathbf{L}^* - \hat{\mathbf{L}}\|_F$, and $\|\mathfrak{E}\|_{\text{op}}$ where $\mathfrak{E} \equiv \sum_{(i,t)\in\mathcal{O}} \varepsilon_{it}\mathbf{A}_{it}$.

Lemma 23 (Adapted from Negahban and Wainwright (2011))**.** *For all* $\lambda \geq 3\|\mathfrak{E}\|_{\text{op}}/|\mathcal{O}|$,

$$\sum_{(i,t)\in\mathcal{O}} \frac{\langle \mathbf{A}_{it}, \mathbf{L}^* - \hat{\mathbf{L}}\rangle^2}{|\mathcal{O}|} \leq 10\lambda\sqrt{R}\,\|\mathbf{L}^* - \hat{\mathbf{L}}\|_F\,. \tag{B.1.1}$$

This type of result has been shown before by Recht (2011), Negahban and Wainwright (2011), Koltchinskii et al. (2011), Klopp (2014). Similar results also appear in the analysis of LASSO type estimators (for example see Bühlmann and Van De Geer (2011) and references therein).

Step 2: The upper bound provided by Lemma 23 contains λ and also requires the condition $\lambda \geq 3\|\mathfrak{E}\|_{\text{op}}/|\mathcal{O}|$. Therefore, in order to have a tight bound, it is important to show an upper bound for $\|\mathfrak{E}\|_{\text{op}}$ that holds with high probability. Next lemma provides one such result.

Lemma 24. *There exist a constant C_1 such that*

$$\|\mathfrak{E}\|_{\text{op}} \leq C_1\sigma \max\left[\sqrt{N\log(N+T)}, \sqrt{T}\log^{3/2}(N+T)\right] ,$$

with probability greater than $1-(N+T)^{-2}$.

This result uses a concentration inequality for sum of random matrices to find a bound for $\|\mathfrak{E}\|_{\text{op}}$. We note that existing papers Recht (2011), Negahban and Wainwright (2011), Koltchinskii et al. (2011), Klopp (2014), contain a similar step but in their case $\mathcal{O}$ is obtained by independently sampling elements of $[N]\times[T]$. However, in our case observations from each row of the matrix are correlated. Therefore, prior results do not apply. In fact, the correlation structure deteriorates the type of upper bound that can be obtained for $\|\mathfrak{E}\|_{\text{op}}$.

Step 3: The last main step is to show that, with high probability, the random variable on the left hand side of (B.1.1) is larger than a constant fraction of $\|\hat{\mathbf{L}}-\mathbf{L}^*\|_F^2$. In high-dimensional statistics literature this property is also referred to as *Restricted Strong Convexity*, Negahban et al. (2012), Negahban and Wainwright (2011, 2012). The following Lemma states this property for our setting and its proof that is similar to the proof of Theorem 1 in (Negahban and Wainwright 2012) or Lemma 12 in (Klopp 2014) is omitted.

Lemma 25. *If the estimator $\hat{\mathbf{L}}$ defined above satisfies $\|\hat{\mathbf{L}}-\mathbf{L}^*\|_F \geq \kappa$ for a positive number κ, then,*

$$\mathbb{P}_\pi\left\{\frac{p_c}{2}\,\|\hat{\mathbf{L}}-\mathbf{L}^*\|_F^2 \leq \sum_{(i,t)\in\mathcal{O}} \langle \mathbf{A}_{it}, \hat{\mathbf{L}}-\mathbf{L}^*\rangle^2\right\} \geq 1-\exp\left(-\frac{p_c^2\kappa^2}{32\,T\,L_{\max}^2}\right) .$$

Now we are equipped to prove the main theorem.

Proof of Theorem 5. Let $\mathbf{\Delta} = \mathbf{L}^* - \hat{\mathbf{L}}$. Then using Lemma 24 and selecting λ equal to $3\|\mathfrak{E}\|_{\text{op}}/|\mathcal{O}|$ in Lemma 23, with probability greater than $1-(N+T)^{-2}$, we have

$$\sum_{(i,t)\in\mathcal{O}} \frac{\langle \mathbf{A}_{it}, \mathbf{\Delta}\rangle^2}{|\mathcal{O}|} \leq \frac{30C_1\sigma\sqrt{R}\max\left[\sqrt{N\,\log(N+T)}, \sqrt{T}\,\log^{3/2}(N+T)\right]}{|\mathcal{O}|}\,\|\mathbf{\Delta}\|_F\,. \quad \text{(B.1.2)}$$

Now, we use Lemma 25 to find a lower bound for the left hand side of (B.1.2). But first note that if $p_{\rm c}^2\|\mathbf{\Delta}\|_F^2/(32\,T\,L_{\max}^2) \le 2\log(N+T)$ then

$$\frac{\|\mathbf{\Delta}\|_F}{\sqrt{NT}} \le 8L_{\max}\sqrt{\frac{\log(N+T)}{N\,p_{\rm c}^2}}$$

holds which proves Theorem 5. Otherwise, using Lemma 25 for $\kappa = (8L_{\max}/p_{\rm c})\sqrt{T\log(N+T)}$,

$$\mathbb{P}\left\{\frac{1}{2}p_{\rm c}\|\mathbf{\Delta}\|_F^2 \le \sum_{(i,t)\in\mathcal{O}}\langle \mathbf{A}_{it}, \mathbf{\Delta}\rangle^2\right\} \ge 1-\frac{1}{(N+T)^2}\,. \tag{B.1.3}$$

Combining this result, (B.1.2), and union bound we have, with probability greater than $1-2(N+T)^{-2}$,

$$\|\mathbf{\Delta}\|_F^2 \le 60C_1\sigma\sqrt{R}\max\left(\sigma\sqrt{\frac{N\,\log(N+T)}{p_{\rm c}^2}},\sqrt{\frac{T}{p_{\rm c}^2}}\,\log^{3/2}(N+T)\right)\|\mathbf{\Delta}\|_F\,.$$

The main result now follows after dividing both sides with $\sqrt{NT}\|\mathbf{\Delta}\|_F$. □

B.1.2 Proof of Lemma 23

Variants of this Lemma for similar models have been proved before. But for completeness we include its proof that is adapted from Negahban and Wainwright (2011).

Proof of Lemma 23. Let

$$f(\mathbf{L}) \equiv \sum_{(i,t)\in\mathcal{O}}\frac{(Y_{it}-L_{it})^2}{|\mathcal{O}|} + \lambda\|\mathbf{L}\|_*\,.$$

Now, using the definition of $\hat{\mathbf{L}}$,

$$f(\hat{\mathbf{L}}) \le f(\mathbf{L}^*)\,,$$

which is equivalent to

$$\sum_{(i,t)\in\mathcal{O}}\frac{\langle \mathbf{L}^*-\hat{\mathbf{L}}, \mathbf{A}_{it}\rangle^2}{|\mathcal{O}|} + 2\sum_{(i,t)\in\mathcal{O}}\frac{\varepsilon_{it}\langle \mathbf{L}^*-\hat{\mathbf{L}}, \mathbf{A}_{it}\rangle}{|\mathcal{O}|} + \lambda\|\hat{\mathbf{L}}\|_* \le \lambda\|\mathbf{L}^*\|_*\,. \tag{B.1.4}$$

Now, defining $\mathbf{\Delta} \equiv \mathbf{L}^* - \hat{\mathbf{L}}$ and using the definition of $\mathfrak{E}$, the above equation gives

$$\sum_{(i,t)\in\mathcal{O}} \frac{\langle \mathbf{\Delta}, \mathbf{A}_{it}\rangle^2}{|\mathcal{O}|} \leq -\frac{2}{|\mathcal{O}|}\langle \mathbf{\Delta}, \mathfrak{E}\rangle + \lambda\|\mathbf{L}^*\|_* - \lambda\|\hat{\mathbf{L}}\|_* \tag{B.1.5}$$

$$\overset{(a)}{\leq} \frac{2}{|\mathcal{O}|}\|\mathbf{\Delta}\|_*\|\mathfrak{E}\|_{\text{op}} + \lambda\|\mathbf{L}^*\|_* - \lambda\|\hat{\mathbf{L}}\|_* \tag{B.1.6}$$

$$\leq \frac{2}{|\mathcal{O}|}\|\mathbf{\Delta}\|_*\|\mathfrak{E}\|_{\text{op}} + \lambda\|\mathbf{\Delta}\|_* \tag{B.1.7}$$

$$\overset{(b)}{\leq} \frac{5}{3}\lambda\|\mathbf{\Delta}\|_* \,. \tag{B.1.8}$$

Here, (a) uses inequality $|\langle \mathbf{A}, \mathbf{B}\rangle| \leq \|\mathbf{A}\|_{\text{op}}\|\mathbf{B}\|_{\max}$ which is due to the fact that operator norm is dual norm to nuclear norm, and (b) uses the assumption $\lambda \geq 3\|\mathfrak{E}\|_{\text{op}}/|\mathcal{O}|$. Before continuing with the proof of Lemma 23 we state the following Lemma that is proved later in this section.

Lemma 26. *Let* $\mathbf{\Delta} \equiv \mathbf{L}^* - \hat{\mathbf{L}}$ *for* $\lambda \geq 3\|\mathfrak{E}\|_{\text{op}}/|\mathcal{O}|$ *Then there exist a decomposition* $\mathbf{\Delta} = \mathbf{\Delta}_1 + \mathbf{\Delta}_2$ *such that*

(i) $\langle \mathbf{\Delta}_1, \mathbf{\Delta}_2\rangle = 0$*.*

(ii) rank$(\mathbf{\Delta}_1) \leq 2r$*.*

(iii) $\|\mathbf{\Delta}_2\|_* \leq 3\|\mathbf{\Delta}_1\|_*$*.*

Now, invoking the decomposition $\mathbf{\Delta} = \mathbf{\Delta}_1 + \mathbf{\Delta}_2$ from Lemma 26 and using the triangle inequality, we obtain

$$\|\mathbf{\Delta}\|_* \overset{(c)}{\leq} 4\|\mathbf{\Delta}_1\|_* \overset{(d)}{\leq} 4\sqrt{2r}\|\mathbf{\Delta}_1\|_F \overset{(e)}{\leq} 4\sqrt{2r}\|\mathbf{\Delta}\|_F \,. \tag{B.1.9}$$

where (c) uses Lemma 26(iii), (d) uses Lemma 26(ii) and Cauchy-Schwarz inequality, and (e) uses Lemma 26(i). Combining this with (B.1.8) we obtain

$$\mathbf{B}\sum_{(i,t)\in\mathcal{O}} \frac{\langle \mathbf{\Delta}, \mathbf{A}_{it}\rangle^2}{|\mathcal{O}|} \leq 10\lambda\sqrt{r}\,\|\mathbf{\Delta}\|_F \,, \tag{B.1.10}$$

which finishes the proof of Lemma 23. □

Proof of Lemma 26. Let $\mathbf{L}^* = \mathbf{U}_{N\times r}\mathbf{S}_{r\times r}(\mathbf{V}_{T\times r})^\top$ be the singular value decomposition for the rank r matrix $\mathbf{L}^*$. Let $\mathbf{P}_U$ be the projection operator onto column space of $\mathbf{U}$ and let

$\mathbf{P}_{U^\perp}$ be the projection operator onto the orthogonal complement of the column space of $\mathbf{U}$. Let us recall a few linear algebra facts about these projection operators. If columns of $\mathbf{U}$ are denoted by $u_1, \dots, u_0$, since $\mathbf{U}$ is unitary, $\mathbf{P}_U = \sum_{i=1}^{r} u_i u_i^\top$. Similarly, $\mathbf{P}_{U^\perp} = \sum_{i=r+1}^{N} u_i u_i^\top$ where $u_1, \dots, u_0, u_{r+1}, \dots, u_N$ forms an orthonormal basis for $\mathbb{R}^N$. In addition, the projector operators are idempotent (i.e., $\mathbf{P}_U^2 = \mathbf{P}_U, \mathbf{P}_{U^\perp}^2 = \mathbf{P}_{U^\perp}$), $\mathbf{P}_U + \mathbf{P}_{U^\perp} = \mathbf{I}_{N\times N}$.

Define $\mathbf{P}_V$ and $\mathbf{P}_{V^\perp}$ similarly. Now, we define $\mathbf{\Delta}_1$ and $\mathbf{\Delta}_2$ as follows:

$$\mathbf{\Delta}_2 \equiv \mathbf{P}_{U^\perp}\mathbf{\Delta}\mathbf{P}_{V^\perp} \quad , \qquad \mathbf{\Delta}_1 \equiv \mathbf{\Delta} - \mathbf{\Delta}_2 \,.$$

It is easy to see that

$$\begin{aligned} \mathbf{\Delta}_1 &= (\mathbf{P}_U + \mathbf{P}_{U^\perp})\mathbf{\Delta}(\mathbf{P}_V + \mathbf{P}_{V^\perp}) - \mathbf{P}_{U^\perp}\mathbf{\Delta}\mathbf{P}_{V^\perp} && \text{(B.1.11)} \\ &= \mathbf{P}_U\mathbf{\Delta} + \mathbf{P}_{U^\perp}\mathbf{\Delta}\mathbf{P}_V \,. && \text{(B.1.12)} \end{aligned}$$

Using this fact we have

$$\begin{aligned} \langle \mathbf{\Delta}_1, \mathbf{\Delta}_2 \rangle &= \text{trace}\Big(\mathbf{\Delta}^\top \mathbf{P}_U \mathbf{P}_{U^\perp}\mathbf{\Delta}\mathbf{P}_{V^\perp} + \mathbf{P}_V \mathbf{\Delta}^\top \mathbf{P}_{U^\perp}\mathbf{P}_{U^\perp}\mathbf{\Delta}\mathbf{P}_{V^\perp}\Big) && \text{(B.1.13)} \\ &= \text{trace}\Big(\mathbf{P}_V \mathbf{\Delta}^\top \mathbf{P}_{U^\perp}\mathbf{\Delta}\mathbf{P}_{V^\perp}\Big) && \text{(B.1.14)} \\ &= \text{trace}\Big(\mathbf{\Delta}^\top \mathbf{P}_{U^\perp}\mathbf{\Delta}\mathbf{P}_{V^\perp}\mathbf{P}_V\Big) = 0 && \text{(B.1.15)} \end{aligned}$$

that gives part (i). Note that we used trace$(\mathbf{AB})$ = trace$(\mathbf{BA})$.

Looking at (B.1.12), part (ii) also follows since both $\mathbf{P}_U$ and $\mathbf{P}_V$ have rank r and sum of two rank r matrices has rank at most $2r$.

Before moving to part (iii), we note another property of the above decomposition of $\mathbf{\Delta}$ that will be needed next. Since the two matrices $\mathbf{L}^*$ and $\mathbf{\Delta}_2$ have orthogonal singular vectors to each other,

$$\|\mathbf{L}^* + \mathbf{\Delta}_2\|_* = \|\mathbf{L}^*\|_* + \|\mathbf{\Delta}_2\|_* \,. \qquad \text{(B.1.16)}$$

On the other hand, using inequality (B.1.6), for $\lambda \geq 3\|\mathbf{\mathfrak{E}}\|_{\text{op}}/|\mathcal{O}|$ we have

$$\begin{aligned}\lambda\left(\|\hat{\mathbf{L}}\|_* - \|\mathbf{L}^*\|_*\right) &\leq \frac{2}{|\mathcal{O}|}\|\mathbf{\Delta}\|_*\|\mathbf{\mathfrak{E}}\|_{\text{op}} \\ &\leq \frac{2}{3}\lambda\|\mathbf{\Delta}\|_* \\ &\leq \frac{2}{3}\lambda\left(\|\mathbf{\Delta}_1\|_* + \|\mathbf{\Delta}_2\|_*\right). \qquad \text{(B.1.17)}\end{aligned}$$

Now, we can use the following for the left hand side

$$\begin{aligned}\|\hat{\mathbf{L}}\|_* - \|\mathbf{L}^*\|_* &= \|\mathbf{L}^* + \mathbf{\Delta}_1 + \mathbf{\Delta}_2\|_* - \|\mathbf{L}^*\|_* \\ &\geq \|\mathbf{L}^* + \mathbf{\Delta}_2\|_* - \|\mathbf{\Delta}_1\|_* - \|\mathbf{L}^*\|_* \\ &\overset{(f)}{=} \|\mathbf{L}^*\|_* + \|\mathbf{\Delta}_2\|_* - \|\mathbf{\Delta}_1\|_* - \|\mathbf{L}^*\|_* \\ &= \|\mathbf{\Delta}_2\|_* - \|\mathbf{\Delta}_1\|_*.\end{aligned}$$

Here (f) follows from (B.1.16). Now, combining the last inequality with (B.1.17) we get

$$\|\mathbf{\Delta}_2\|_* - \|\mathbf{\Delta}_1\|_* \leq \frac{2}{3}\left(\|\mathbf{\Delta}_1\|_* + \|\mathbf{\Delta}_2\|_*\right).$$

That finishes proof of part (iii). □

B.1.3 Proof of Lemma 24

First we state the matrix version of Bernstein inequality for rectangular matrices (see Tropp (2012) for a derivation of it).

Proposition 6 (Matrix Bernstein Inequality)**.** *Let $\mathbf{Z}_1, \ldots, \mathbf{Z}_N$ be independent matrices in $\mathbb{R}^{d_1 \times d_2}$ such that $\mathbb{E}[\mathbf{Z}_i] = \mathbf{0}$ and $\|\mathbf{Z}_i\|_{\text{op}} \leq D$ almost surely for all $i \in [N]$ and a constant R. Let σ_Z be such that*

$$\sigma_Z^2 \geq \max\left\{ \left\|\sum_{i=1}^N \mathbb{E}[\mathbf{Z}_i\mathbf{Z}_i^\top]\right\|_{\text{op}}, \left\|\sum_{i=1}^N \mathbb{E}[\mathbf{Z}_i^\top\mathbf{Z}_i\right\|_{\text{op}} \right\}.$$

Then, for any $\alpha \geq 0$

$$\mathbb{P}\left\{ \left\|\sum_{i=1}^N \mathbf{Z}_i\right\|_{\text{op}} \geq \alpha \right\} \leq (d_1 + d_2)\exp\left[\frac{-\alpha^2}{2\sigma_Z^2 + (2D\alpha)/3}\right]. \qquad \text{(B.1.18)}$$

Proof of Lemma 24. Our goal is to use Proposition 6. Define the sequence of independent random matrices $\mathbf{B}_1, \ldots, \mathbf{B}_N$ as follows. For every $i \in [N]$, define

$$\mathbf{B}_i = \sum_{t=1}^{t_i} \varepsilon_{it} \mathbf{A}_{it} \,.$$

By definition, $\mathfrak{E} = \sum_{i=1}^{N} \mathbf{B}_i$ and $\mathbb{E}[\mathbf{B}_i] = \mathbf{0}$ for all $i \in [N]$. Define the bound $D \equiv C_2 \sigma \sqrt{\log(N+T)}$ for a large enough constant C_2. For each $(i,t) \in \mathcal{O}$ define $\bar{\varepsilon}_{it} = \varepsilon_{it} \mathbb{I}_{|\varepsilon_{it}| \le D}$. Also define $\overline{\mathbf{B}}_i = \sum_{t=1}^{t_i} \bar{\varepsilon}_{it} \mathbf{A}_{it}$ for all $i \in [N]$.

Using union bound and the fact that for σ-sub-Gaussian random variables ε_{it} we have $\mathbb{P}(|\varepsilon_{it}| \ge t) \le 2\exp\{-t^2/(2\sigma^2)\}$ gives, for each $\alpha \ge 0$,

$$\begin{aligned}
\mathbb{P}\{\ \|\mathfrak{E}\|_{\text{op}} \ge \alpha\} &\le \mathbb{P}\left\{ \left\| \sum_{i=1}^{N} \overline{\mathbf{B}}_i \right\|_{\text{op}} \ge \alpha \right\} + \sum_{(i,t)\in\mathcal{O}} \mathbb{P}\{|\varepsilon_{it}| \ge D\} \\
&\le \mathbb{P}\left\{ \left\| \sum_{i=1}^{N} \overline{\mathbf{B}}_i \right\|_{\text{op}} \ge \alpha \right\} + 2|\mathcal{O}| \exp\left\{ \frac{-D^2}{2\sigma^2} \right\} \\
&\le \mathbb{P}\left\{ \left\| \sum_{i=1}^{N} \overline{\mathbf{B}}_i \right\|_{\text{op}} \ge \alpha \right\} + \frac{1}{(N+T)^3} \,. \qquad \text{(B.1.19)}
\end{aligned}$$

Now, for each $i \in [N]$, define $\mathbf{Z}_i \equiv \overline{\mathbf{B}}_i - \mathbb{E}[\overline{\mathbf{B}}_i]$. Then,

$$\begin{aligned}
\left\| \sum_{i=1}^{N} \overline{\mathbf{B}}_i \right\|_{\text{op}} &\le \left\| \sum_{i=1}^{N} \mathbf{Z}_i \right\|_{\text{op}} + \left\| \mathbb{E}\left[\sum_{1 \le i \le N} \overline{\mathbf{B}}_i \right] \right\|_{\text{op}} \\
&\le \left\| \sum_{i=1}^{N} \mathbf{Z}_i \right\|_{\text{op}} + \left\| \mathbb{E}\left[\sum_{1 \le i \le N} \overline{\mathbf{B}}_i \right] \right\|_{F} \le \left\| \sum_{i=1}^{N} \mathbf{Z}_i \right\|_{\text{op}} + \sqrt{NT} \left\| \mathbb{E}\left[\sum_{1 \le i \le N} \overline{\mathbf{B}}_i \right] \right\|_{\max} .
\end{aligned}$$

But since each ε_{it} has mean zero,

$$\begin{aligned}
|\mathbb{E}[\bar{\varepsilon}_{it}]| = |\mathbb{E}[\varepsilon_{it} \mathbb{I}_{|\varepsilon_{it}| \le D}]| = |\mathbb{E}[\varepsilon_{it} \mathbb{I}_{|\varepsilon_{it}| \ge D}]| &\le \sqrt{\mathbb{E}[\varepsilon_{it}^2]\, \mathbb{P}(|\varepsilon_{it}| \ge D)} \\
&\le \sqrt{2\sigma^2 \, \exp[-D^2/(2\sigma^2)]} \\
&\le \frac{\sigma}{(N+T)^4} \,.
\end{aligned}$$

Therefore,

$$\sqrt{NT}\left\|\mathbb{E}\left[\sum_{1\le i\le N}\overline{\mathbf{B}}_i\right]\right\|_{\max}\le\frac{\sigma\sqrt{NT}}{(N+T)^4}\le\frac{\sigma}{(N+T)^3}\,,$$

which gives

$$\left\|\sum_{i=1}^{N}\overline{\mathbf{B}}_i\right\|_{\text{op}}\le\left\|\sum_{i=1}^{N}\mathbf{Z}_i\right\|_{\text{op}}+\frac{\sigma}{(N+T)^3}\,. \tag{B.1.20}$$

We also note that $\|\mathbf{Z}_i\|_{\text{op}}\le 2D\sqrt{T}$ for all $i\in[N]$. The next step is to calculate σ_Z defined in the Proposition 6. We have,

$$\begin{aligned}\left\|\sum_{i=1}^{N}\mathbb{E}[\mathbf{Z}_i\mathbf{Z}_i^\top]\right\|_{\text{op}} &\le \max_{(i,t)\in\mathcal{O}}\left\{\mathbb{E}\left[(\bar{\varepsilon}_{it}-E[\bar{\varepsilon}_{it}])^2\right]\right\}\ \left\|\sum_{i=1}^{N}\mathbb{E}\left[\sum_{t=1}^{t_i}e_i(N)e_i(N)^\top\right]\right\|_{\text{op}} && \text{(B.1.21)}\\ &\le 2\sigma^2\max_{i\in[N]}\left(\sum_{t\in[T]}t\pi_t^{(i)}\right)\le 2T\sigma^2 && \text{(B.1.22)}\end{aligned}$$

and

$$\begin{aligned}\left\|\sum_{i=1}^{N}\mathbb{E}[\mathbf{Z}_i^\top\mathbf{Z}_i]\right\|_{\text{op}} &\le 2\sigma^2\left\|\sum_{i=1}^{N}\mathbb{E}\left[\sum_{t=1}^{t_i}e_t(T)e_t(T)^\top\right]\right\|_{\text{op}} && \text{(B.1.23)}\\ &= 2\sigma^2\max_{t\in[T]}\left(\sum_{i\in[N]}\sum_{t'=t}^{T}\pi_{t'}^{(i)}\right)=2N\sigma^2\,. && \text{(B.1.24)}\end{aligned}$$

Note that here we used the fact that random variables $\bar{\varepsilon}_{it}-E[\bar{\varepsilon}_{it}]$ are independent of each other and centered which means all cross terms of the type $\mathbb{E}\{(\bar{\varepsilon}_{it}-E[\bar{\varepsilon}_{it}])(\bar{\varepsilon}_{js}-E[\bar{\varepsilon}_{js}])\}$ are zero for $(i,t)\neq(j,s)$. Therefore, $\sigma_Z^2=2\sigma^2\max(N,T)$ works. Applying Proposition 6, we obtain

$$\mathbb{P}\left\{\left\|\sum_{i=1}^{N}\mathbf{Z}_i\right\|_{\text{op}} \geq \alpha\right\} \leq (N+T)\exp\left[-\frac{\alpha^2}{4\sigma^2\max(N,T)+(4D\alpha\sqrt{T})/3}\right]$$

$$\leq (N+T)\exp\left[-\frac{3}{16}\min\left(\frac{\alpha^2}{\sigma^2\max(N,T)}, \frac{\alpha}{D\sqrt{T})}\right)\right] .$$

Therefore, there is a constant C_3 such that with probability greater than $1-\exp(-t)$,

$$\left\|\sum_{i=1}^{N}\mathbf{Z}_i\right\|_{\text{op}} \leq C_3\sigma\max\left(\sqrt{\max(N,T)[t+\log(N+T)]}, \sqrt{T\log(N+T)}[t+\log(N+T)]\right) .$$

Using this for a t that is a large enough constant times $\log(N+T)$, together with (B.1.19) and (B.1.20), shows with probability larger than $1-2(N+T)^{-3}$

$$\begin{aligned}\|\mathfrak{E}\|_{\text{op}} &\leq C_1\sigma\max\left[\sqrt{\max(N,T)\log(N+T)}, \sqrt{T}\log^{3/2}(N+T)\right] \\ &= C_1\sigma\max\left[\sqrt{N\log(N+T)}, \sqrt{T}\log^{3/2}(N+T)\right] ,\end{aligned}$$

for a constant C_1. □

Appendix C

Supplementary Materials for Chapter 4

C.1 Simulation Setting

Here we explain the settings for simulations shown in Figures 4.1 and 4.2.

C.1.1 Single Test Point

The data for single test point simulation, shown in Figure 4.1, has been generated as follows. Here $p = 1$, $D = 20$ and $d = 2$. All the points are generated using $X_i = AX_i^{\text{low}}$, where $A \in \mathbb{R}^{D\times d}$ and entries of A are independently sampled from $U[-1, 1]$. Components of each X_i^{low} are also generated independently from $U[-1, 1]$. We generate a fix test point $x_{\text{test}} = Ax_{\text{test}}^{\text{low}}$ and keep the matrix A throughout all Monte-Carlo iterations fixed. In each Monte-Carlo iteration, we generate $n = 20000$ training points as mentioned before. The values of Y_i are generated according to $Y_i = f(X_i) + \varepsilon_i$, where $f(X) = \frac{1}{1+\exp(-3X[0])}$, and $\varepsilon_i \sim \mathsf{N}(0, \sigma_e^2)$ with $\sigma_e = 1$. We are interested in estimating and drawing inference about $f(x_{\text{test}})$ which is equivalent to solving for $\mathbb{E}[\psi(Z; \theta(x)) \mid X = x] = 0$ with $\psi(Z; \theta(x)) = Y - \theta(x)$ at $x = x_{\text{test}}$. We run the sub-sampled k-NN estimation (Algorithm 5) for $k = 1, 2$ and 5 with parameter $s = s_\zeta$ chosen using Proposition 5 with $\zeta = 0.1$ over 1000 Monte-Carlo iterations and report the histogram and quantile-quantile plot of estimates compared to theoretical asymptotic normal distribution of estimates stemming from our characterization. In our simulations, we considered the complete U-statistic case, i.e., $B = \binom{n}{s}$.

C.1.2 Multiple Test Points

The data for the multiple test point simulation, shown in Figure 4.2, has been generated very similarly to the single test point setting. The only difference is that instead of generating a single test point we generate 100 test points. These test points together with matrix A are kept fixed throughout all 1000 Monte-Carlo iterations. We compare the performance of sub-sampled k-NN estimator (Algorithm 5) with parameter $s = s_\zeta$ chosen using Proposition 5 with $\zeta = 0.1$ with two benchmarks that set $s_d = n^{1.05d/(d+2)}$ and $s_D = n^{1.05D/(D+2)}$. This process has been repeated for $k = 1, 2$ and 5 and the coverage over a single run for all test points, the empirical coverage over 1000 runs, and chosen s_ζ versus s_d are depicted.

C.2 Nuisance Parameters and Heterogeneous Treatment Effects

Using the techniques of Oprescu et al. (2018), our work also easily extends to the case where the moments depend on, potentially infinite dimensional, nuisance components h_0, that also need to be estimated, i.e.,

$$\theta(x) \text{ solves: } m(x;\theta,h_0) = \mathbb{E}[\psi(Z;\theta,h_0) \mid x] = 0. \tag{C.2.1}$$

If the moment m is orthogonal with respect to h and assuming that h_0 can be estimated on a separate sample with a conditional MSE rate of

$$\mathbb{E}[(\hat{h}(z) - h_0(z))^2 | X = x] = o_p(\epsilon(s) + \sqrt{s/n})\,, \tag{C.2.2}$$

then using the techniques of Oprescu et al. (2018), we can argue that both our finite sample estimation rate and our asymptotic normality rate, remain unchanged, as the estimation error only impacts lower order terms. This extension allows us to capture settings like heterogeneous treatment effects, where the treatment model also needs to be estimated when using the orthogonal moment as

$$\psi(z;\theta,h_0) = (y - q_0(x,w) - \theta(t - p_0(x,w)))\,(t - p_0(x,w))\,, \tag{C.2.3}$$

where y is the outcome of interest, t is a treatment, x, w are confounding variables, $q_0(x,w) = \mathbb{E}[Y|X = x, W = w]$ and $p_0(x,w) = E[T|X = x, W = w]$. The latter two nuisance functions can be estimated via separate non-parametric regressions. In particular, if we assume that these functions are sparse linear in w, i.e.:

$$q_0(x, w) = \langle \beta(x), w \rangle \,, \qquad\qquad p_0(x, w) = \langle \gamma(x), w \rangle \,. \tag{C.2.4}$$

Then we can achieve a conditional mean-squared-error rate of the required order by using the kernel lasso estimator of Oprescu et al. (2018), where the kernel is the sub-sampled k-NN kernel, assuming the sparsity does not grow fast with n.

C.3 Proofs

C.3.1 Proof of Theorem 8

Lemma 27. *For any $\theta \in \Theta$:*

$$\|\theta - \theta(x)\|_2 \leq \frac{2}{\lambda} \|m(x; \theta)\|_2 \,. \tag{C.3.1}$$

Proof. By strong convexity of the loss $L(x; \theta)$ and the fact that $m(x; \theta(x)) = 0$, we have:

$$L(x; \theta) - L(x; \theta(x)) \geq \langle m(x; \theta(x)), \theta - \theta(x) \rangle + \frac{\lambda}{2} \cdot \|\theta - \theta(x)\|_2^2 = \frac{\lambda}{2} \cdot \|\theta - \theta(x)\|_2^2 \,.$$

By convexity of the loss $L(x; \theta)$ we have:

$$L(x; \theta(x)) - L(x; \theta) \geq \langle m(x; \theta), \theta(x) - \theta \rangle \,.$$

Combining the latter two inequalities we get:

$$\frac{\lambda}{2} \cdot \|\theta - \theta(x)\|_2^2 \leq \langle m(x; \theta), \theta - \theta(x) \rangle \leq \|m(x; \theta)\|_2 \cdot \|\theta - \theta(x)\|_2 \,.$$

Note that if $\|\theta - \theta(x)\|_2 = 0$, then the result is obvious. Otherwise, dividing over by $\|\theta - \theta(x)\|_2$ completes the proof of the lemma. □

Lemma 28. *Let $\Lambda(x; \theta) = m(x; \theta) - \Psi(x; \theta)$. Then the estimate $\hat{\theta}$ satisfies:*

$$\|m(x; \hat{\theta})\|_2 \leq \sup_{\theta \in \Theta} \|\Lambda(x; \theta)\|_2 \,. \tag{C.3.2}$$

Proof. Observe that $\hat{\theta}$, by definition, satisfies $\Psi(x;\hat{\theta}) = 0$. Thus:

$$\|m(x;\hat{\theta})\|_2 = \|m(x;\hat{\theta}) - \Psi(x;\hat{\theta})\|_2 = \|\Lambda(x;\hat{\theta})\|_2 \leq \sup_{\theta\in\Theta} \|\Lambda(x;\theta)\|_2 \,.$$

□

Lemma 29. *Suppose that the kernel is built with sub-sampling at rate s, in an honest manner (Assumption 7) and with at least $B \geq n/s$ sub-samples. If the base kernel satisfies kernel shrinkage in expectation, with rate $\epsilon(s)$, then w.p. $1-\delta$:*

$$\sup_{\theta\in\Theta} \|\Lambda(x;\theta)\|_2 \leq L_m \epsilon(s) + O\left(\psi_{\max}\sqrt{\frac{p\, s}{n}\left(\log\log(n/s) + \log(p/\delta)\right)}\right)\,. \tag{C.3.3}$$

Proof. Define

$$\mu_0(x;\theta) = \mathbb{E}\left[\Psi_0(x;\theta)\right]\,,$$

where we remind that Ψ_0 denotes the complete U-statistic:

$$\Psi_0(x;\theta) = \binom{n}{s}^{-1} \sum_{S_b\subset[n]:|S_b|=s} \mathbb{E}_{\omega_b}\left[\sum_{i\in S_b} \alpha_{S_b,\omega_b}(X_i)\psi(Z_i;\theta)\right]\,.$$

Here the expectation is taken with respect to the random draws of n samples. Then, the following result which is due to Oprescu et al. (2018) holds.

Lemma 30 (Adapted from Oprescu et al. (2018))**.** *For any θ and target x*

$$\mu_0(x;\theta) = \binom{n}{s}^{-1} \sum_{S_b\subset[n]:|S_b|=s} \mathbb{E}\left[\sum_{i\in S_b} \alpha_{S_b,\omega_b}(X_i)m(X_i;\theta)\right]\,.$$

In other words, Lemma 30 states that, in the expression for μ_0 we can simply replace $\psi(Z_i;\theta)$ with its expectation which is $m(X_i;\theta)$. We can then express $\Lambda(x;\theta)$ as sum of kernel error, sampling error, and sub-sampling error, by adding and subtracting appropriate terms, as follows:

$$\begin{aligned}\Lambda(x;\theta) =\ & m(x;\theta) - \Psi(x;\theta)\\ =\ & \underbrace{m(x;\theta) - \mu_0(x;\theta)}_{\Gamma(x,\theta)=\text{Kernel error}} + \underbrace{\mu_0(x;\theta) - \Psi_0(x;\theta)}_{\Delta(x,\theta)=\text{Sampling error}} + \underbrace{\Psi_0(x;\theta) - \Psi(x;\theta)}_{\Upsilon(x,\theta)=\text{Sub-sampling error}}\end{aligned}$$

The parameters should be chosen to trade-off these error terms nicely. We will now bound each of these three terms separately and then combine them to get the final bound.

Bounding the Kernel error. By Lipschitzness of m with respect to x and triangle inequality, we have:

$$\begin{aligned}
\|\Gamma(x;\theta)\|_2 \leq\ & \binom{n}{s}^{-1} \sum_{S_b\subset[n]:|S_b|=s} \mathbb{E}\left[\sum_{i\in S_b} \alpha_{S_b,\omega_b}(X_i)\|m(x;\theta)-m(X_i;\theta)\|\right] \\
\leq\ & L_m \binom{n}{s}^{-1} \sum_{S_b\subset[n]:|S_b|=s} \mathbb{E}\left[\sum_{i\in S_b} \alpha_{S_b,\omega_b}(X_i)\|x-X_i\|\right] \\
\leq\ & L_m \binom{n}{s}^{-1} \sum_{S_b\subset[n]:|S_b|=s} \mathbb{E}\left[\sup\{\|x-X_i\| : \alpha_{S_b,\omega_b}(X_i)>0\}\right] \\
\leq\ & L_m\, \epsilon(s)\,,
\end{aligned}$$

where the second to last inequality follows from the fact that $\sum_i |\alpha_{S_b}(X_i)| = 1$.

Bounding the Sampling Error. For bounding the sampling error we rely on Lemma 38 and in particular Corollary 8. Observe that for each $j \in \{1,\ldots,p\}$, $\Psi_{0j}(x;\theta)$ is a complete U-statistic for each θ. Thus the sampling error defines a U-process over the class of symmetric functions $\text{conv}(\mathcal{F}_j) = \{f_j(\cdot;\theta) : \theta \in \Theta\}$, with $f_j(Z_1,\ldots,Z_s;\theta) = \mathbb{E}_\omega\left[\sum_{i=1}^s \alpha_{Z_{1:s},\omega}(X_i)\psi_j(Z_i;\theta)\right]$. Observe that since $f_j \in \text{conv}(\mathcal{F}_j)$ is a convex combination of functions in $\mathcal{F}_j = \{\psi_j(\cdot;\theta) : \theta\in\Theta\}$, the bracketing number of functions in $\text{conv}(\mathcal{F}_j)$ is upper bounded by the bracketing number of $\mathcal{F}_j$, which by our assumption, satisfies $\log(N_{[]}(\mathcal{F}_j,\epsilon,L_2)) = O(1/\epsilon)$. Moreover, by our assumptions on the upper bound $\psi_{\max}$ of $\psi_j(z;\theta)$, we have that $\sup_{f_j\in\text{conv}(\mathcal{F}_j)}\|f_j\|_2, \sup_{f_j\in\text{conv}(\mathcal{F}_j)}\|f_j\|_\infty \leq \psi_{\max}$. Thus all conditions of Corollary 8 are satisfied, with $\eta = G = \psi_{\max}$ and we get that w.p. $1-\delta/2p$:

$$\sup_{\theta\in\Theta}|\Delta_j(x,\theta)| = O\left(\psi_{\max}\sqrt{\frac{s}{n}\left(\log\log(n/s)+\log(2p/\delta)\right)}\right)\,. \tag{C.3.4}$$

By a union bound over j, we get that w.p. $1-\delta/2$:

$$\sup_{\theta\in\Theta}\|\Delta_j(x,\theta)\|_2 \leq \sqrt{p}\max_{j\in[p]}\sup_{\theta\in\Theta}|\Delta_j(x,\theta)| = O\left(\psi_{\max}\sqrt{\frac{p\,s}{n}\left(\log\log(n/s)+\log(p/\delta)\right)}\right)\,. \tag{C.3.5}$$

Bounding the Sub-Sampling Error. Sub-sampling error decays as B is increased. Note that for a fixed set of samples $\{Z_1, Z_2, \ldots, Z_n\}$, for a set S_b randomly chosen among all $\binom{n}{s}$ subsets of size s from the n samples, we have:

$$\mathbb{E}_{S_b,\omega_b}\left[\sum_{i\in S_b} \alpha_{S_b,\omega_b}(X_i)\psi(Z_i;\theta)\right] = \Psi_0(x;\theta)\,.$$

Therefore, $\Psi(x;\theta)$ can be thought as the sum of B i.i.d. random variables each with expectation equal to $\Psi_0(x;\theta)$, where expectation is taken over B draws of sub-samples, each with size s. Thus one can invoke standard results on empirical processes for function classes as a function of the bracketing entropy. For simplicity, we can simply invoke Corollary 8 in the Appendix C.3.12 for the case of a trivial U-process, with $s = 1$ and $n = B$ to get that w.p. $1 - \delta/2$:

$$\sup_{\theta\in\Theta} |\Upsilon(x;\theta)| = O\left(\psi_{\max}\sqrt{\frac{\log\log(B) + \log(2/\delta)}{B}}\right)$$

Thus for $B \geq n/s$, the sub-sampling error is of lower order than the sampling error and can be asymptotically ignored. Putting together the upper bounds on sampling, sub-sampling and kernel error finishes the proof of the Lemma. □

The probabilistic statement of the proof follows by combining the inequalities in the above three lemmas. The in expectation statement follows by simply integrating the exponential tail bound of the probabilistic statement.

C.3.2 Proof of Theorem 9

We will show asymptotic normality of $\hat{\alpha} = \left\langle \beta, \hat{\theta} \right\rangle$ for some arbitrary direction $\beta \in \mathbb{R}^p$, with $\|\beta\|_2 \leq R$. Consider the complete multi-dimensional U-statistic:

$$\Psi_0(x;\theta) = \binom{n}{s}^{-1} \sum_{S_b\subset[n]:|S_b|=s} \mathbb{E}_{\omega_b}\left[\sum_{i\in S_b} \alpha_{S_b,\omega_b}(X_i)\psi(Z_i;\theta)\right]. \tag{C.3.6}$$

Let

$$\Delta(x;\theta) = \Psi_0(x;\theta) - \mu_0(x;\theta) \tag{C.3.7}$$

where $\mu_0(x;\theta) = \mathbb{E}[\Psi_0(x;\theta)]$ (as in the proof of Theorem 8) and

$$\tilde{\theta} = \theta(x) - M_0^{-1}\Delta(x;\theta(x)) \tag{C.3.8}$$

Finally, let

$$\tilde{\alpha} \triangleq \left\langle \beta, \tilde{\theta} \right\rangle = \langle \beta, \theta(x)\rangle - \left\langle \beta, M_0^{-1}\,\Delta(x;\theta(x))\right\rangle \tag{C.3.9}$$

For shorthand notation let $\alpha_0 = \langle \beta, \theta(x)\rangle$, $\psi_\beta(Z;\theta) = \left\langle \beta, M_0^{-1}(\psi(Z;\theta) - m(X;\theta))\right\rangle$ and

$$\begin{aligned}\Psi_{0,\beta}(x;\theta) = & \left\langle \beta, M_0^{-1}\,\Delta(x;\theta(x))\right\rangle \\ = & \binom{n}{s}^{-1} \sum_{S_b \subset [n]:|S_b|=s} \mathbb{E}_{\omega_b}\left[\sum_{i\in S_b} \alpha_{S_b,\omega_b}(X_i)\psi_\beta(Z_i;\theta)\right]\end{aligned}$$

be a single dimensional complete U-statistic. Thus we can re-write:

$$\tilde{\alpha} = \alpha_0 - \Psi_{0,\beta}(x;\theta(x))$$

We then have the following lemma which its proof is provided in Appendix C.3.10:

Lemma 31. *Under the conditions of Theorem 9:*

$$\frac{\Psi_{0,\beta}(x;\theta(x))}{\sigma_n(x)} \to \mathsf{N}(0,1)\,,$$

for $\sigma_n^2(x) = \frac{s^2}{n}\operatorname{Var}\left[\mathbb{E}\left[\sum_{i=1}^s K(x, X_i, \{X_j\}_{j=1}^s)\psi_\beta(Z_i;\theta) \mid X_1\right]\right] = \Omega(\frac{s^2}{n}\eta(s))$.

Invoking Lemma 31 and using our assumptions on the kernel, we conclude that:

$$\frac{\tilde{\alpha} - \alpha_0(x)}{\sigma_n(x)} \to \mathsf{N}(0,1). \tag{C.3.10}$$

For some sequence σ_n^2 which decays at least as slow as $s^2\eta(s)/n$. Hence, since

$$\frac{\hat{\alpha} - \alpha_0}{\sigma_n(x)} = \frac{\tilde{\alpha} - \theta(x)}{\sigma_n(x)} + \frac{\hat{\alpha} - \tilde{\alpha}}{\sigma_n(x)},$$

if we show that $\frac{\hat{\alpha}-\tilde{\alpha}}{\sigma_n(x)} \to_p 0$, then by Slutsky's theorem we also have that:

$$\frac{\hat{\alpha} - \alpha_0}{\sigma_n(x)} \to \mathsf{N}(0,1), \tag{C.3.11}$$

as desired. Thus, it suffices to show that:

$$\frac{\|\hat{\alpha} - \tilde{\alpha}\|_2}{\sigma_n(x)} \to_p 0. \tag{C.3.12}$$

Observe that since $\|\beta\|_2 \leq R$, we have $\|\hat{\alpha} - \tilde{\alpha}\|_2 \leq R\|\hat{\theta} - \tilde{\theta}\|_2$. Thus it suffices to show that:

$$\frac{\|\hat{\theta} - \tilde{\theta}\|}{\sigma_n(x)} \to_p 0.$$

Lemma 32. *Under the conditions of Theorem 9, for* $\sigma_n^2(x) = \Omega\left(\frac{s^2}{n}\eta(s)\right)$*:*

$$\frac{\|\hat{\theta} - \tilde{\theta}\|}{\sigma_n(x)} \to_p 0. \tag{C.3.13}$$

Proof. Performing a second-order Taylor expansion of $m_j(x;\theta)$ around $\theta(x)$ and observing that $m_j(x;\theta(x)) = 0$, we have that for some $\bar{\theta}_j \in \Theta$:

$$m_j(x;\hat{\theta}) = \left\langle \nabla_\theta m_j(x;\theta(x)), \hat{\theta} - \theta(x) \right\rangle + \underbrace{(\hat{\theta} - \theta(x))^\top H_j(x;\bar{\theta}_j)(\hat{\theta} - \theta(x))^\top}_{\rho_j}.$$

Letting $\rho = (\rho_1, \ldots, \rho_p)$, writing the latter set of equalities for each j in matrix form, multiplying both sides by M_0^{-1} and re-arranging, we get that:

$$\hat{\theta} = \theta(x) + M_0^{-1} m(x;\hat{\theta}) - M_0^{-1}\rho\,.$$

Thus by the definition of $\tilde{\theta}$ we have:

$$\hat{\theta} - \tilde{\theta} = M_0^{-1} \cdot (m(x;\hat{\theta}) + \Delta(x;\theta(x))) - M_0^{-1}\rho\,.$$

By the bounds on the eigenvalues of $H_j(x;\theta)$ and M_0^{-1}, we have that:

$$\|M_0^{-1}\rho\|_2 \leq \frac{L_H}{\lambda}\|\hat{\theta} - \theta(x)\|_2^2\,. \tag{C.3.14}$$

Thus we have:

$$\|\hat{\theta} - \tilde{\theta}\|_2 = \frac{1}{\lambda}\|m(x;\hat{\theta}) + \Delta(x;\theta(x))\|_2 + \frac{L_H}{\lambda}\|\hat{\theta} - \theta(x)\|_2^2\,.$$

By our estimation error Theorem 8, we have that the expected value of the second term on the right hand side is of order $O\left(\epsilon(s)^2, \frac{s}{n}\log\log(n/s)\right)$. Thus by the assumptions of the theorem, both are $o(\sigma_n)$. Hence, the second term is $o_p(\sigma_n)$.

We now argue about the convergence rate of the first term on the right hand side. Similar to the proof of Theorem 8, since $\Psi(x;\hat{\theta}) = 0$ we have:

$$m(x;\hat{\theta}) = m(x;\hat{\theta}) - \Psi(x;\hat{\theta}) = m(x;\hat{\theta}) - \Psi_0(x;\hat{\theta}) + \underbrace{\Psi_0(x;\hat{\theta}) - \Psi(x;\hat{\theta})}_{\text{Sub-sampling error}} .$$

We can further add and subtract μ_0 from $m(x;\hat{\theta})$.

$$\begin{aligned} m(x;\hat{\theta}) = &\ m(x;\hat{\theta}) - \mu_0(x;\hat{\theta}) + \mu_0(x;\hat{\theta}) - \Psi_0(x;\hat{\theta}) + \Psi_0(x;\hat{\theta}) - \Psi(x;\hat{\theta}) \\ = &\ m(x;\hat{\theta}) - \mu_0(x;\hat{\theta}) - \Delta(x;\hat{\theta}) + \Psi_0(x;\hat{\theta}) - \Psi(x;\hat{\theta}) \,. \end{aligned}$$

Combining we have:

$$m(x;\hat{\theta}) + \Delta(x;\theta(x)) = \underbrace{m(x;\hat{\theta}) - \mu_0(x;\hat{\theta})}_{C=\text{Kernel error}} + \underbrace{\Delta(x;\theta(x)) - \Delta(x;\hat{\theta})}_{F=\text{Sampling error}} + \underbrace{\Psi_0(x;\hat{\theta}) - \Psi(x;\hat{\theta})}_{E=\text{Sub-sampling error}} .$$

Now similar to proof of Theorem 8 we bound different terms separately and combine the results.

Kernel Error. Term C is a kernel error and hence is upper bounded by $\epsilon(s)$ in expectation. Since, by assumption s is chosen such that $\epsilon(s) = o(\sigma_n(x))$, we ge that $\|C\|_2/\sigma_n(x) \to_p 0$.

Sub-Sampling Error. Term E is a sub-sampling error, which can be made arbitrarily small if the number of drawn sub-samples is large enough and hence $\|E\|_2/\sigma_n(x) \to_p 0$. In fact, similar to the part about bounding sub-sampling error in Lemma 29 we have that that:

$$\mathbb{E}_{S_b}\left[\sum_{i \in S_b} \alpha_{S_b}(X_i)\psi(Z_i;\theta)\right] = \Psi_0(x;\theta) \,,$$

Therefore, $\Psi(x;\theta)$ can be thought as the sum of B independent random variables each with expectation equal to $\Psi_0(x;\theta)$. Now we can invoke Corollary 8 in Appendix C.3.12 for the trivial U-process, with $s = 1, n = B$ to get that w.p. $1 - \delta_1$:

$$\sup_{\theta \in \Theta} \|\Psi_0(x;\theta) - \Psi(x;\theta)\| \leq O\left(\Psi_{\max}\sqrt{\frac{\log\log(B) + \log(1/\delta_1)}{B}}\right).$$

Hence, for $B \geq (n/s)^{5/4}$, due to our assumption that $(s/n \log\log(n/s))^{5/8} = o(\sigma_n(x))$ we get $\|E\|_2/\sigma_n(x) \to_p 0$.

Sampling Error. Thus it suffices that show that $\|F\|_2/\sigma_n(x) \to_p 0$, in order to conclude that $\frac{\|m(x;\hat{\theta})+\Psi_0(x;\theta(x))\|_2}{\sigma_n(x)} \to_p 0$. Term F can be re-written as:

$$F = \Psi_0(x;\theta(x)) - \Psi_0(x;\hat{\theta}) - \mathbb{E}\left[\Psi_0(x;\theta(x)) - \Psi_0(x;\hat{\theta})\right]. \tag{C.3.15}$$

Observe that each coordinate j of F, is a stochastic equicontinuity term for U-processes over the class of symmetric functions $\text{conv}(\mathcal{F}_j) = \{f_j(\cdot;\theta) : \theta \in \Theta\}$, with $f_j(Z_1, \ldots, Z_s; \theta) = \mathbb{E}_\omega\left[\sum_{i=1}^s \alpha_{Z_{1:s},\omega}(X_i)(\psi_j(Z_i;\theta(x)) - \psi_j(Z_i;\theta))\right]$. Observe that since $f_j \in \text{conv}(\mathcal{F}_j)$ is a convex combination of functions in $\mathcal{F}_j = \{\psi_j(\cdot;\theta(x)) - \psi_j(\cdot;\theta) : \theta \in \Theta\}$, the bracketing number of functions in $\text{conv}(\mathcal{F}_j)$ is upper bounded by the bracketing number of $\mathcal{F}_j$, which in turn is upper bounded by the bracketing number of the function class $\{\psi_j(\cdot;\theta) : \theta \in \Theta\}$, which by our assumption, satisfies $\log(N_{[]}(\mathcal{F}_j, \epsilon, L_2)) = O(1/\epsilon)$. Moreover, under the variogram assumption and the lipschitz moment assumption we have that if $\|\theta - \theta(x)\| \leq r \leq 1$, then:

$$\begin{aligned}
\|f_j(\cdot;\theta)\|_{P,2}^2 &= \mathbb{E}\left[\left(\sum_{i=1}^s \alpha_{Z_{1:s}}(X_i)(\psi_j(Z_i;\theta(x)) - \psi_j(Z_i;\theta)\right)^2\right] \\
&\leq \mathbb{E}\left[\sum_{i=1}^s \alpha_{Z_{1:s}}(X_i)\left(\psi_j(Z_i;\theta(x)) - \psi_j(Z_i;\theta)\right)^2\right] && \text{(Jensen's inequality)} \\
&= \mathbb{E}\left[\sum_{i=1}^s \alpha_{Z_{1:s}}(X_i)\mathbb{E}\left[\psi_j(Z_i;\theta(x)) - \psi_j(Z_i;\theta)\right]^2 |X_i]\right] && \text{(honesty of kernel)} \\
&= \mathbb{E}\left[\sum_{i=1}^s \alpha_{Z_{1:s}}(X_i)\left(\text{Var}(\psi(Z;\theta(x)) - \psi(Z;\theta)|X_i) + (m(X_i;\theta(x)) - m(X_i;\theta))^2\right)\right] \\
&\leq L_\psi\|\theta - \theta(x)\| + L_J^2\|\theta - \theta(x)\|^2 \leq L_\psi r + L_J^2 r^2 = O(r).
\end{aligned}$$

Moreover, $\|f_j\|_\infty \leq 2\psi_{\max}$. Thus we can apply Corollary 8, with $\eta = \sqrt{L_\psi r + L_J^2 r^2} = O(\sqrt{r})$ and $G = 2\psi_{\max}$ to get that if $\|\hat{\theta} - \theta(x)\| \leq r$, then w.p. $1 - \delta/p$:

$$\begin{aligned} |F_j| \leq & \sup_{\theta:\|\theta-\theta(x)\|\leq r} \left|\Psi_0(x;\theta(x)) - \Psi_0(x;\hat{\theta}) - \mathbb{E}\left[\Psi_0(x;\theta(x)) - \Psi_0(x;\hat{\theta})\right]\right| \\ = & \; O\left(\left(r^{1/4} + \sqrt{r}\sqrt{\log(p/\delta) + \log\log(n/(s\,r))}\right)\sqrt{\frac{s}{n}}\right) \\ = & \; O\left(\left(r^{1/4}\sqrt{\log(p/\delta) + \log\log(n/s)}\right)\sqrt{\frac{s}{n}}\right) \triangleq \kappa(r,s,n,\delta)\,. \end{aligned}$$

Using a union bound this implies that w.p. $1 - \delta$ we have

$$\max_j |F_j| \leq \kappa(r,s,n,\delta)\,.$$

By our MSE theorem and also Markov's inequality, w.p. $1-\delta'$: $\|\hat{\theta}-\theta(x)\| \leq \nu(s)/\delta'$, where:

$$\nu(s) = \frac{1}{\lambda}\left(L_m\epsilon(s) + O\left(\psi_{\max}\sqrt{\frac{p\,s}{n}\log\log(p\,s/n)}\right)\right)$$

Thus using a union bound w.p. $1 - \delta - \delta'$, we have:

$$\max_j |F_j| = O\left(\kappa(\nu(s)/\delta', s, n, \delta)\right)$$

To improve readability from here we ignore all the constants in our analysis, while we keep all terms (even log or log log terms) that depend on s and n. Note that we can even ignore δ and δ', because they can go to zero at very slow rate such that terms $\log(1/\delta)$ or even $\delta'^{1/4}$ appearing in the analysis grow slower than log log terms. Now, by the definition of $\nu(s)$ and $\kappa(r,s,n,\delta')$, as well as invoking the inequality $(a+b)^{1/4} \leq a^{1/4} + b^{1/4}$ for $a, b > 0$ we have:

$$\max_j |F_j| \leq O(\kappa(\nu(s)/\delta', s, n, \delta)) \leq O\left(\epsilon(s)^{1/4}\left(\frac{s}{n}\log\log(n/s)\right)^{1/2} + \left(\frac{s}{n}\log\log(n/s)\right)^{5/8}\right), \tag{C.3.16}$$

Hence, using our Assumption on the rates in the statement of Theorem 9 we get that both of the terms above are $o(\sigma_n(x))$. Therefore, $\|F\|_2/\sigma_n(x) \to_p 0$. Thus, combining all of the above, we get that:

$$\frac{\|\tilde{\theta} - \hat{\theta}\|}{\sigma_n(x)} = o_p(1)$$

as desired. □

C.3.3 Proof of Lemma 7

We give a generic lower bound on the quantity $\mathbb{E}[\mathbb{E}[K(x, X_1, \{Z_j\}_{j=1}^s)|X_1]^2]$ that depends only on the kernel shrinkage. The bound essentially implies that if we know that the probability that the distribution of X's assigns to a ball of radius $\epsilon(s, 1/2s)$ around the target x is of order $1/s$, i.e. we should expect at most a constant number of samples to fall in the kernel shrinkage ball, then the main condition on incrementality of the kernel, required for asymptotic normality, holds. In some sense, this property states that the kernel shrinkage behavior is tight in the following sense. Suppose that the kernel was assigning positive weight to at most a constant number of k samples. Then kernel shrinkage property states that with high probability we expect to see at least k samples in a ball of radius $\epsilon(s, \delta)$ around x. The above assumption says that we should also not expect to see too many samples in that radius, i.e. we should also expect to see at most a constant number $K > k$ of samples in that radius. Typically, the latter should hold, if the characterization of $\epsilon(s, \delta)$ is tight, in the sense that if we expected to see too many samples in the radius, then most probably we could have improved our analysis on kernel shrinkage and given a better bound that shrinks faster. The proof of Lemma 7 is as follows.

Proof. By the Paley-Zygmund inequality, for any random variable $Z \geq 0$ and for any $\delta \in [0, 1]$:

$$\mathbb{E}[Z^2] \geq (1-\delta)^2 \frac{\mathbb{E}[Z]^2}{\Pr[Z \geq \delta \mathbb{E}[Z]]}\,.$$

Let $W_1 = K(x, X_1, \{Z_j\}_{j=1}^s)$. Then, applying the latter to the random variable $Z = \mathbb{E}[W_1|X_1]$ and observing that by symmetry $\mathbb{E}[Z] = \mathbb{E}[W_1] = 1/s$, yields:

$$\mathbb{E}\left[\mathbb{E}[W_1|X_1]^2\right] \geq \frac{(1-\delta)^2\, \mathbb{E}[W_1]^2}{\Pr[\mathbb{E}[W_1|X_1] > \delta \mathbb{E}[W_1]]} = \frac{(1-\delta)^2\, (1/s)^2}{\Pr[\mathbb{E}[W_1|X_1] > \delta/s]}\,.$$

Moreover, observe that by the definition of $\epsilon(s, \rho)$ for some $\rho > 0$ we have

$$\Pr[W_1 > 0 \wedge \|X_1 - x\| \geq \epsilon(s, \rho)] \leq \rho\,.$$

This means that at most a mass $\rho\, s/\delta$ of the support of X_1 in the region $\|X_1 - x\| \geq \epsilon(s, \rho)$

can have $\Pr[W_1 > 0|X_1] \geq \delta/s$. Otherwise the overall probability that $W_1 > 0$ in the region of $\|X_1 - x\| \geq \epsilon(s,\rho)$ would be more than ρ. Thus we have that except for a region of mass $\rho\, s/\delta$, for each X_1 in the region $\|X_1 - x\| \geq \epsilon(s,\rho)$: $\mathbb{E}[W_1|X_1] \leq \delta/s$. Combining the above we get:

$$\Pr[\mathbb{E}[W_1|X_1] \leq \delta/s] \geq \Pr[\|X_1 - x\| \geq \epsilon(s,\rho)] - \rho\, s/\delta\,.$$

Thus,

$$\Pr[\mathbb{E}[W_1|X_1] > \delta/s] \leq \Pr[\|X_1 - x\| \leq \epsilon(s,\rho)] + \rho\, s/\delta = \mu(B(x,\epsilon(s,\delta))) + \rho\, s/\delta\,.$$

Since ρ was arbitrarily chosen, the latter upper bound holds for any ρ, which yields the result.

C.3.4 Proof of Corollary 3

Applying Lemma 7 with $\delta = 1/2$ yields

$$\mathbb{E}[\mathbb{E}[K(x, X_1, \{Z_j\}_{j=1}^s)|X_1]^2] \geq \frac{(1/2s)^2}{\inf_{\rho>0}\left(\mu(B(x,\epsilon(s,\rho))) + 2\rho\, s\right)}\,.$$

Observe that

$$\mu(B(x,\epsilon(s,\rho))) \leq \frac{1}{cr^d}\epsilon(s,\rho)^d \mu(B(x,r)) = O\left(\frac{\log(1/\rho)}{s}\right)\,.$$

Hence,

$$\inf_{\rho>0}\left(\mu(B(x,\epsilon(s,\rho))) + 2\rho\, s\right) = O\left(\inf_{\rho>0}\left(\frac{\log(1/\rho)}{s} + 2\rho\, s\right)\right) = O\left(\frac{\log(s)}{s}\right)\,,$$

where the last follows by choosing $\rho = 1/s^2$. Combining all the above yields

$$\mathbb{E}[\mathbb{E}[K(x, X_1, \{Z_j\}_{j=1}^s)|X_1]^2] = \Omega\left(\frac{1}{s\log(s)}\right)\,,$$

as desired. □

C.3.5 Proof of Lemma 8

For proving this lemma, we rely on Bernstein's inequality which is stated below:

Proposition 7 (Bernstein's Inequality)**.** *Suppose that random variables $Z_1, Z_2, \ldots, Z_n$ are i.i.d., belong to $[-c, c]$ and $\mathbb{E}[Z_i] = \mu$. Let $\bar{Z}_n = \frac{1}{n}\sum_{i=1}^n Z_i$ and $\sigma^2 = \mathrm{Var}(Z_i)$. Then, for any $\theta > 0$,*

$$\Pr\left(|\bar{Z}_n - \mu| > \theta\right) \le 2\exp\left(\frac{-n\theta^2}{2\sigma^2 + 2c\theta/3}\right).$$

This also implies that w.p. at least $1-\delta$ the following holds:

$$|\bar{Z}_n - \mu| \le \sqrt{\frac{2\sigma^2 \log(2/\delta)}{n}} + \frac{2c\log(2/\delta)}{3n}. \tag{C.3.17}$$

Let A be any μ-measurable set. An immediate application of Bernstein's inequality to random variables $Z_i = 1\{X_i \in A\}$, implies that w.p. $1-\delta$ over the choice of features $(X_i)_{i=1}^s$, we have:

$$|\mu_s(A) - \mu(A)| \le \sqrt{\frac{2\mu(A)\log(2/\delta)}{s}} + \frac{2\log(2/\delta)}{3s}.$$

In above, we used the fact that $\mathrm{Var}(Z_i) = \mu(A)(1-\mu(A)) \le \mu(A)$. This result has the following corollary.

Corollary 7. *Define $U = 2\log(2/\delta)/s$ and let A be an arbitrary μ-measurable set. Then, w.p. $1-\delta$ over the choice of training samples, $\mu(A) \ge 4U$ implies $\mu_s(A) \ge U$.*

Proof. Define $U = 2\log(2/\delta)/s$. Then, Bernstein's inequality in Proposition 7 implies that w.p. $1-\delta$ we have

$$|\mu_s(A) - \mu(A)| \le \sqrt{U\mu(A)} + \frac{U}{3}\,.$$

Assume that $\mu(A) \ge 4U$, we want to prove that $\mu_s(A) \ge U$. Suppose, the contrary, i.e., $\mu_s(A) < U$. Then, by dividing the above equation by $\mu(A)$ we get

$$\left|\frac{\mu_s(A)}{\mu(A)} - 1\right| \le \sqrt{\frac{U}{\mu(A)}} + \frac{1}{3}\frac{U}{\mu(A)},.$$

Note that since $\mu_s(A) < U < \mu(A)$, by letting $z = U/\mu(A) \le 1/4$ the above implies that

$$1 - z \le \sqrt{z} + \frac{z}{3} \Rightarrow \frac{4}{3}z + \sqrt{z} - 1 \ge 0\,,$$

which as $z > 0$ only holds for

$$\sqrt{z} \geq \frac{-3 + \sqrt{57}}{8} \Rightarrow z \geq 0.3234 \,.$$

This contradicts with $z \leq 1/4$, implying the result. $\square$

Now we are ready to finish the proof of Lemma 8. First, note that using the definition of (C, d)-homogeneous measure, for any $\theta \in (0, 1)$ we have $\mu(B(x, \theta r)) \geq (1/C)\theta^d \mu(B(x, r))$. Replace $\theta r = \epsilon$ in above. It implies that for any $\epsilon \in (0, r)$

$$\mu(B(x, \epsilon)) \geq \frac{1}{C\, r^d} \epsilon^d \mu(B(x, r)) \,. \tag{C.3.18}$$

Pick $\epsilon_k(s, \delta)$ according to

$$\epsilon_k(s, \delta) = r \left(\frac{8C\, \log(2/\delta)}{\mu(B(x, r))s} \right)^{1/d} \,.$$

Note that for having $\epsilon_k(s, \delta) \in (0, r)$ we need

$$\log(2/\delta) \leq \frac{1}{8\, C} \mu(B(x, r))s \Rightarrow \delta \geq 2 \exp\left(-\frac{1}{8\, C} \mu(B(x, r)s \right) \,.$$

Therefore, replacing this choice of $\epsilon_k(s, \delta)$ in Equation (C.3.18) implies that $\mu(B(x, \epsilon_k(s, \delta))) \geq \frac{8 \log(2/\delta)}{s}$. Now we can use the result of Corollary 7 for the choice $A = B(x, \epsilon_k(s, \delta))$. It implies that w.p. $1 - \delta$ over the choice of s training samples, we have

$$\mu_s(B(x, \epsilon_k(s, \delta))) \geq \frac{2 \log(2/\delta)}{s} \,.$$

Note that whenever $\delta \leq \exp(-k/2)/2$ we have

$$\frac{2 \log(2/\delta)}{s} \geq \frac{k}{s} \,.$$

Therefore, w.p. $1 - \delta$ we have

$$\|x - X_{(k)}\| \leq \epsilon_k(s, \delta) = O\left(\frac{\log(1/\delta)}{s} \right)^{1/d} \,.$$

C.3.6 Proof of Corollary 4

Lemma 8 shows that for any $t = \epsilon_k(s,\delta) = r\left(\frac{8C\log(2/\delta)}{\mu(B(x,r))s}\right)^{1/d}$, such that $t \leq r$ and $t \geq r\left(\frac{4k\,C}{\mu(B(x,r))s}\right)^{1/d}$, we have that:

$$\Pr[\|x - X_{(k)}\|_2 \geq \epsilon_k(s,\delta)] \leq \delta\,.$$

Let $\rho = \frac{1}{r}\left(\frac{\mu(B(x,r))}{8C}\right)^{1/d}$, which is a constant. Solving for δ in terms of t we get:

$$\Pr[\|x - X_{(k)}\|_2 \geq t] \leq 2\exp\left(-\rho^d\, s\, t^d\right)\,,$$

for any $t \in \left[\frac{(s/k)^{-1/d}}{\rho}, r\right]$. Thus, noting that X_i's and target x both belong to $\mathcal{X}$ that has diameter $\Delta_{\mathcal{X}}$, we can upper bound the expected value of $[\|x - X_{(k)}\|_2$ as:

$$\begin{aligned}\mathbb{E}\left[\|x - X_{(k)}\|_2\right] &= \int_0^{\Delta_{\mathcal{X}}} \Pr\left[\|x - X_{(k)}\|_2 \geq t\right] dt\\ &\leq \frac{(s/k)^{-1/d}}{\rho} + \int_{\rho\,(s/k)^{-1/d}}^{r} \Pr\left[\|x - X_{(k)}\|_2 \geq t\right] dt + \Pr\left[\|x - X_{(k)}\|_2 \geq r\right](\Delta_{\mathcal{X}} - r)\\ &\leq \frac{(s/k)^{-1/d}}{\rho} + \int_{\rho\,(s/k)^{-1/d}}^{r} 2\exp\left\{-\rho^d\, s\, t^d\right\} dt + 2\exp\left\{-\rho^d\, r^d\, s\right\}(\Delta_{\mathcal{X}} - r)\,.\end{aligned}$$

Note that for s larger than some constant, we have $\exp\left\{-\rho^d\, r^d\, s\right\} \leq s^{-1/d}$. Thus the first and last terms in the latter summation are of order $\left(\frac{1}{s}\right)^{1/d}$. We now show that the same holds for the middle term, which would complete the proof. By setting $u = \rho^d\, s\, t^d$ and doing a change of variables in the integral we get:

$$\begin{aligned}\int_{\rho\,(s/k)^{1/d}}^{r} 2\exp\left\{-\rho^d\, s\, t^d\right\} dt \leq \int_0^{\infty} 2\exp\left\{-\rho^d\, s\, t^d\right\} dt\\ = \frac{1}{d\,\rho\, s^{1/d}}\int_0^{\infty} u^{1/d-1}\exp\{-u\}\, du = \frac{s^{-1/d}}{\rho}\frac{1}{d}\Gamma(1/d)\,.\end{aligned}$$

where Γ is the Gamma function. Since by the properties of the Gamma function $z\Gamma(z) = \Gamma(z+1)$, the latter evaluates to: $\frac{s^{-1/d}}{\rho}\Gamma((d+1)/d)$. Since $(d+1)/d \in [1,2]$, we have that $\Gamma((d+1)/d) \leq 2$. Thus the middle term is upper bounded by $\frac{2s^{-1/d}}{\rho}$, which is also of order $\left(\frac{1}{s}\right)^{1/d}$.

C.3.7 Proof of Lemma 9

Before proving this lemma we state and prove and auxiliary lemma which comes in handy in our proof.

Lemma 33. *Let P_1 denote the mass that the density of the distribution of X_i puts on the ball around x with radius $\|x - X_1\|_2$, which is a random variable as it depends on X_1. Then, for any $s \geq k$ the following holds:*

$$\mathbb{E}\left[\sum_{i=0}^{k-1} \binom{s-1}{i} (1-P_1)^{s-1-i} P_1^i\right] = \mathbb{E}\left[\mathbb{E}\left[S_1 \mid X_1\right]\right] = \frac{k}{s}\,.$$

Proof. Let $S_1 = 1\{\text{sample 1 is among } k \text{ nearest neighbors}\}$, then we can write

$$\mathbb{E}\left[\mathbb{E}[S_1 \mid X_1]\right] = \mathbb{E}\left[\sum_{i=0}^{k-1} \binom{s-1}{i} (1-P_1)^{s-1-i} P_1^i\right]\,,$$

which simply computes the probability that there are at most $k-1$ other points in the ball with radius $\|x - X_1\|$. Now, by using the tower law of expectations

$$\mathbb{E}\left[\mathbb{E}[S_1 \mid X_1]\right] = \mathbb{E}[S_1] = \frac{k}{s}\,,$$

which holds because of the symmetry. In other words, the probability that sample 1 is among the k-NN is equal to k/s. Hence, the conclusion follows. □

We can finish the proof of Lemma 9. Define $S_1 = 1\{\text{sample 1 is among } k \text{ nearest neighbors}\}$, then we can write

$$\mathbb{E}\left[\mathbb{E}\left[K(x, X_1, \{Z_j\}_{j=1}^s) \mid X_1\right]^2\right] = \frac{1}{k^2}\mathbb{E}\left[\mathbb{E}\left[S_1 \mid X_1\right]^2\right]\,.$$

Recall that P_1 denotes the mass that the density of the distribution of X_i puts on the ball around x with radius $\|x - X_1\|_2$, which is a random variable depending on X_1. Therefore,

$$\mathbb{E}\left[S_1 \mid X_1\right] = \sum_{i=0}^{k-1} \binom{s-1}{i} (1-P_1)^{s-1-i} P_1^i\,.$$

Now we can write

$$\begin{aligned}
\mathbb{E}\left[\mathbb{E}\left[S_1 \mid X_1\right]^2\right] &= \mathbb{E}\left[\left(\sum_{i=0}^{k-1}\binom{s-1}{i}(1-P_1)^{s-1-i}P_1^i\right)^2\right] \\
&= \mathbb{E}\left[\sum_{i=0}^{k-1}\sum_{j=0}^{k-1}\binom{s-1}{i}\binom{s-1}{j}(1-P_1)^{2s-2-i-j}P_1^{i+j}\right] \\
&= \mathbb{E}\left[\sum_{t=0}^{2k-2}(1-P_1)^{2s-2-t}P_1^t\sum_{i=0}^{k-1}\sum_{j=0}^{k-1}\binom{s-1}{i}\binom{s-1}{j}1\left\{i+j=t\right\}\right] \\
&= \mathbb{E}\left[\sum_{t=0}^{2k-2}(1-P_1)^{2s-2-t}P_1^t\sum_{i=\max\{0,t-(k-1)\}}^{\min\{t,k-1\}}\binom{s-1}{i}\binom{s-1}{t-i}\right] \\
&= \mathbb{E}\left[\sum_{t=0}^{2k-2}a_t\,(1-P_1)^{2s-2-t}P_1^t\right]
\end{aligned}$$

Now using Lemma 33 (where s is replaced by $2s-1$) we know that for any value of $0 \leq r \leq 2s-2$ we have

$$\mathbb{E}\left[\sum_{t=0}^{r}b_t\,(1-P_1)^{2s-2-r}P_1^r\right] = \mathbb{E}\left[\sum_{t=0}^{r}\binom{2s-2}{t}(1-P_1)^{2s-2-t}P_1^t\right] = \frac{r+1}{2s-1}\,. \qquad \text{(C.3.19)}$$

This implies that for any value of r we have $\mathbb{E}\left[b_r(1-P_1)^{2s-2-r}P_1^r\right] = 1/(2s-1)$. The reason is simple. Note that the above is obvious for $r = 0$ using Equation (C.3.19). For other values of $r \geq 1$, we can write Equation (C.3.19) for values r and $r-1$. Taking their difference implies the result. Note that this further implies that $\mathbb{E}\left[(1-P_1)^{2s-2-r}P_1^r\right] = 1/(b_r\,(2s-1))$, as b_r is a constant. Therefore, by plugging this back into the expression of $\mathbb{E}[\mathbb{E}[S_1 \mid X_1]^2]$ we have

$$\mathbb{E}[\mathbb{E}[S_1|X_1]^2] = \mathbb{E}\left[\sum_{t=0}^{2k-2}a_t\,(1-P_1)^{2s-2-t}P_1^t\right] = \frac{1}{2s-1}\left(\sum_{t=0}^{2k-2}\frac{a_t}{b_t}\right),$$

which implies the desired result.

C.3.8 Proof of Theorem 11

Note that according to Lemma 37, the asymptotic variance $\sigma_{n,j}^2(x) = \frac{s^2}{n}\operatorname{Var}[\Phi_1(Z_1)]$, where $\Phi_1(Z_1) = \frac{1}{k}\mathbb{E}[\sum_{i\in H_k(x,s)}\langle e_j, M_0^{-1}\psi(Z_i;\theta(x))\rangle \mid Z_1]$. Therefore, once we establish an expression for $\operatorname{Var}[\Phi_1(Z_1)]$ we can finish the proof of this theorem. The following lemma provides such a result.

Lemma 34. *Let K be the k-NN kernel and $\sigma_j^2(x) = \operatorname{Var}\left(\langle e_j, M_0^{-1}\psi(z;\theta(x))\rangle \mid X = x\right)$. Moreover, suppose that $\epsilon_k(s, 1/s^2) \to 0$ for any constant k. Then:*

$$\begin{aligned}\operatorname{Var}[\Phi_1(Z_1)] &= \sigma_j(x)^2\,\mathbb{E}\left[\mathbb{E}\left[K(x, X_1, \{Z_j\}_{j=1}^s) \mid X_1\right]^2\right] + o(1/s)\\ &= \frac{\sigma_j^2(x)}{(2s-1)k^2}\left(\sum_{t=0}^{2k-2}\frac{a_t}{b_t}\right) + o(1/s)\end{aligned}$$

where the second equality above holds due to Lemma 9 and sequences a_t and b_t, for $0 \le t \le 2k-2$, are defined in Lemma 9.

Proof. In this proof for simplicity we let $Y_i = \langle e_j, M_0^{-1}\psi(Z_i;\theta(x))\rangle$ and $\mu(X_i) = \mathbb{E}[Y_i] = \langle e_j, M_0^{-1}m(X_i;\theta(x))\rangle$. Let $Z_{(i)}$ denote the random variable of the i-th closest sample to x. For the case of k-NN we have that:

$$k\,\Phi_1(Z_1) = \mathbb{E}\left[\sum_{i=1}^k Y_{(i)} \mid Z_1\right].$$

Let $S_1 = \mathbf{1}\{\text{sample 1 is among } k \text{ nearest neighbors}\}$. Then we have:

$$k\,\Phi_1(Z_1) = \mathbb{E}\left[S_1\sum_{i=1}^k Y_{(i)} \mid Z_1\right] + \mathbb{E}\left[(1-S_1)\sum_{i=1}^k Y_{(i)} \mid Z_1\right].$$

Let $\tilde{Y}_{(i)}$ denote the label of the i-th closest point to x, excluding sample 1. Then:

$$\begin{aligned}k\,\Phi_1(Z_1) =&\ \mathbb{E}\left[S_1\sum_{i=1}^k Y_{(i)} \mid Z_1\right] + \mathbb{E}\left[(1-S_1)\sum_{i=1}^k \tilde{Y}_{(i)} \mid Z_1\right]\\ =&\ \mathbb{E}\left[S_1\sum_{i=1}^k \left(Y_{(i)} - \tilde{Y}_{(i)}\right) \mid Z_1\right] + \mathbb{E}\left[\sum_{i=1}^k \tilde{Y}_{(i)} \mid Z_1\right].\end{aligned}$$

Observe that $\tilde{Y}_{(i)}$ are all independent of Z_1. Hence:

$$k\,\Phi_1(Z_1) = \mathbb{E}\left[S_1 \sum_{i=1}^{k}\left(Y_{(i)} - \tilde{Y}_{(i)}\right) \mid Z_1\right] + \mathbb{E}\left[\sum_{i=1}^{k} \tilde{Y}_{(i)}\right].$$

Therefore the variance of $\Phi(Z_1)$ is equal to the variance of the first term on the right hand side. Hence:

$$\begin{aligned} k^2 \operatorname{Var}\left[\Phi_1(Z_1)\right] &= \mathbb{E}\left[\mathbb{E}\left[S_1 \sum_{i=1}^{k}\left(Y_{(i)} - \tilde{Y}_{(i)}\right) \mid Z_1\right]^2\right] - \mathbb{E}\left[S_1 \sum_{i=1}^{k}\left(Y_{(i)} - \tilde{Y}_{(i)}\right)\right]^2 \\ &= \mathbb{E}\left[\mathbb{E}\left[S_1 \sum_{i=1}^{k}\left(Y_{(i)} - \tilde{Y}_{(i)}\right) \mid Z_1\right]^2\right] + o(1/s). \end{aligned}$$

Where we used the fact that:

$$\left|\mathbb{E}\left[S_1 \sum_{i=1}^{k}\left(Y_{(i)} - \tilde{Y}_{(i)}\right)\right]\right| \leq \mathbb{E}\left[S_1\right] 2k\psi_{\max} = \frac{2k^2\psi_{\max}}{s}. \tag{C.3.20}$$

Moreover, observe that under the event that $S_1 = 1$, we know that the difference between the closest k values and the closest k values excluding 1 is equal to the difference between the Y_1 and $Y_{(k+1)}$. Hence:

$$\mathbb{E}\left[S_1 \sum_{i=1}^{k}\left(Y_{(i)} - \tilde{Y}_{(i)}\right) \mid Z_1\right] = \mathbb{E}\left[S_1\left(Y_1 - Y_{(k+1)}\right) \mid Z_1\right] = \mathbb{E}\left[S_1\left(Y_1 - \mu(X_{(k+1)})\right) \mid Z_1\right].$$

where the last equation holds from the fact that for any $j \neq 1$, conditional on X_j, the random variable Y_j is independent of Z_1 and is equal to $\mu(X_j)$ in expectation. Under the event $S_1 = 1$, we know that the $(k+1)$-th closest point is different from sample 1. We now argue that up to lower order terms, we can replace $\mu(X_{(k+1)})$ with $\mu(X_1)$ in the last equality:

$$\mathbb{E}\left[S_1\left(Y_1 - \mu(X_{(k+1)})\right) \mid Z_1\right] = \underbrace{\mathbb{E}\left[S_1\left(Y_1 - \mu(X_1)\right) \mid Z_1\right]}_{A} + \underbrace{\mathbb{E}\left[S_1\left(\mu(X_1) - \mu(X_{(k+1)})\right) \mid Z_1\right]}_{\rho}.$$

Observe that:

$$\mathbb{E}\left[\mathbb{E}\left[S_1\left(Y_1-\mu(X_{(k+1)})\right)\mid Z_1\right]^2\right]=\mathbb{E}[A^2]+\mathbb{E}[\rho^2]+2\mathbb{E}[A\rho]\,.$$

Moreover, by Jensen's inequality, Lipschitzness of the first moments and kernel shrinkage:

$$\begin{aligned}\left|\mathbb{E}[\rho^2]\right| = {} & \mathbb{E}\left[\mathbb{E}\left[S_1\left(\mu(X_1)-\mu(X_{(k+1)})\right)\mid Z_1\right]^2\right]\leq\mathbb{E}\left[S_1\left(\mu(X_1)-\mu(X_{(k+1)})\right)^2\right]\\ \leq {} & 4L_m^2\epsilon_{k+1}(s,\delta)^2\mathbb{E}[\mathbb{E}[S_1|X_1]]+4\delta\psi_{\max}^2\leq 4L_m^2\epsilon_{k+1}(s,\delta)^2\frac{k}{s}+4\delta\psi_{\max}^2\,.\end{aligned}$$

Hence, for $\delta=1/s^2$, the latter is $o(1/s)$. Similarly:

$$\begin{aligned}|\mathbb{E}[A\rho]|\leq\mathbb{E}[|A|\,|\rho|]\leq {} & \psi_{\max}\mathbb{E}\left[\mathbb{E}\left[S_1\left|\mu(X_1)-\mu(X_{(k+1)})\right|\mid Z_1\right]\right]=\psi_{\max}\mathbb{E}\left[S_1\left|\mu(X_1)-\mu(X_{(k+1)})\right|\right]\\ \leq {} & \psi_{\max}\mathbb{E}[S_1]\epsilon_{k+1}(s,\delta)+2\delta\psi_{\max}=\psi_{\max}\epsilon_{k+1}(s,\delta)\frac{k}{s}+2\delta\psi_{\max}\,.\end{aligned}$$

which for $\delta=1/s^2$ is also of order $o(1/s)$. Combining all the above we thus have:

$$\begin{aligned}k^2\ \mathrm{Var}\left[\Phi_1(Z_1)\right] = {} & \mathbb{E}\left[\mathbb{E}\left[S_1\left(Y_1-\mu(X_1)\right)\mid Z_1\right]^2\right]+o(1/s)\\ = {} & \mathbb{E}\left[\mathbb{E}\left[S_1\mid X_1\right]^2\left(Y_1-\mu(X_1)\right)^2\right]+o(1/s)\,.\end{aligned}$$

We now work with the first term on the right hand side. By the tower law of expectations

$$\begin{aligned}\mathbb{E}\left[\mathbb{E}[S_1\mid X_1]^2\left(Y_1-\mu(X_1)\right)^2\right] = {} & \mathbb{E}\left[\mathbb{E}[S_1\mid X_1]^2\mathbb{E}\left[Y_1-\mu(X_1)^2\mid X_1\right]\right]=\mathbb{E}\left[\mathbb{E}[S_1\mid X_1]^2\sigma_j^2(X_1)\right]\\ = {} & \mathbb{E}\left[\mathbb{E}[S_1\mid X_1]^2\sigma_j^2(x)\right]+\mathbb{E}\left[\mathbb{E}[S_1\mid X_1]^2\left(\sigma_j^2(X_1)-\sigma_j^2(x)\right)\right]\,.\end{aligned}$$

By Lipschitzness of the second moments, we know that the second part is upper bounded as:

$$\begin{aligned}\left|\mathbb{E}\left[\mathbb{E}[S_1\mid X_1]^2\left(\sigma_j^2(X_1)-\sigma_j^2(x)\right)\right]\right|\leq {} & \left|\mathbb{E}\left[\mathbb{E}[S_1\mid X_1]\left(\sigma_j^2(X_1)-\sigma_j^2(x)\right)\right]\right|\\ \leq {} & \left|\mathbb{E}\left[S_1\left(\sigma_j^2(X_1)-\sigma_j^2(x)\right)\right]\right|\\ = {} & \left|\mathbb{E}\left[S_1\left(\sigma_j^2(X_{(k)})-\sigma_j^2(x)\right)\right]\right|\\ \leq {} & L_{mm}\mathbb{E}\left[S_1\right]\epsilon_k(s,\delta)+\delta\psi_{\max}^2\\ = {} & \frac{L_{mm}\,\epsilon_k(s,\delta)\,k}{s}+\delta\psi_{\max}^2\,.\end{aligned}$$

For $\delta = 1/s^2$ it is of $o(1/s)$. Thus:

$$k^2 \ \mathrm{Var}\left[\Phi_1(Z_1)\right] = \mathbb{E}\left[\mathbb{E}[S_1 \mid X_1]^2\right] \sigma_j^2(x) + o(1/s)\,.$$

Note that Lemma 9 provides an expression for $\mathbb{E}\left[\mathbb{E}[S_1 \mid X_1]^2\right]$ which finishes the proof. □

For finishing the proof of Theorem 11 we need to prove that $\sum_{t=0}^{2k-2} \frac{a_t}{b_t}$ is equal to ζ_k plus lower order terms. This is stated in Lemma 10 and the proof of this lemma is provided below.

Proof of Lemma 10. Note that for any $0 \leq t \leq k-1$ we have $a_t = b_t$ according to Remark 4.5.1. For any $k \leq t \leq 2k-2$ we have

$$\begin{aligned}
\frac{a_t}{b_t} &= \sum_{i=t-k+1}^{k-1} \frac{\binom{s-1}{i}\binom{s-1}{t-i}}{\binom{2s-2}{t}} = \sum_{i=t-k+1}^{k-1} \frac{\frac{(s-1)(s-2)\dots(s-i)}{i\,!}\,\frac{(s-1)(s-2)\dots(s-t+i)}{(t-i)\,!}}{\frac{(2s-2)(2s-3)\dots(2s-1-t)}{t\,!}} \\
&= \sum_{i=t-k+1}^{k-1} \binom{t}{i} \frac{(s-1)(s-2)\dots(s-i)\ (s-1)(s-2)\dots(s-t+i)}{(2s-2)(2s-3)\dots(2s-1-t)} \\
&= \sum_{i=t-k+1}^{k-1} \binom{t}{i} \frac{s-1}{2s-2}\frac{s-2}{2s-3}\cdots\frac{s-i}{2s-1-i}\,\frac{s-1}{2s-i}\frac{s-2}{2s-i-1}\cdots\frac{s-t+i}{2s-1-t} \\
&= \sum_{i=t-k+1}^{k-1} 2^{-t}\binom{t}{i}\left(1-\frac{1}{2s-3}\right)\cdots\left(1-\frac{i-1}{2s-1-i}\right)\left(1+\frac{i-2}{2s-i}\right)\cdots\left(1+\frac{i-(i-t+1)}{2s-1-t}\right) \\
&= 2^{-t}\sum_{i=t-k+1}^{k-1}\binom{t}{i}(1+O(1/s)) \\
&= 2^{-t}\sum_{i=t-k+1}^{k-1}\binom{t}{i} + O(1/s)\,,
\end{aligned}$$

where we used the fact that t and i are both bounded above by $2k-2$ which is a constant. Hence,

$$\sum_{t=0}^{2k-2}\frac{a_t}{b_t} = k + \sum_{t=k}^{2k-2} 2^{-t}\sum_{i=t-k+1}^{k-1}\binom{t}{i} + O(1/s) = \zeta_k + O(1/s)\,,$$

as desired. □

C.3.9 Proof of Theorem 12

The goal is to apply Theorem 9. Note that k-NN kernel is both honest and symmetric. According to Lemma 8, we have that $\epsilon_k(s,\delta) = O\left((\log(1/\delta)/s)^{1/d}\right)$ for $\exp(-Cs) \leq \delta \leq D$, where C and D are constants. Corollary 4 also implies that $\epsilon_k(s) = O((1/s)^{1/d})$. Furthermore, according to Lemma 9, the incrementality $\eta_k(s)$ is $\Theta(1/s)$. Therefore, as s goes to ∞ we have $\epsilon_k(s,\eta_k(s)) = O\left((\log(s)/s)^{1/d}\right) \to 0$. Moreover, as $\eta_k(s) = \Theta(1/s)$, we also get that $n\eta_k(s) = O(n/s) \to \infty$. We only need to ensure that Equation (4.4.3) is satisfied. Note that $\sigma_{n,j}(x) = \Theta(\sqrt{s/n})$. Therefore, by dividing terms in Equation (4.4.3) it suffices that

$$\max\left(s^{-1/d}\left(\frac{n}{s}\right)^{1/2}, s^{-1/4d}\left(\log\log(n/s)\right)^{1/2}, \left(\frac{n}{s}\right)^{-1/8}\left(\log\log(n/s)\right)^{5/8}\right) = o(1)\,.$$

Note that due to our Assumption $n/s \to \infty$, the last term obviously goes to zero. Also, because of the assumption made in the statement of theorem, the first term also goes to zero. We claim that if the first term goes to zero, the same also holds for the second term. Note that we can write

$$s^{-1/4d}\left(\log\log(n/s)\right)^{1/2} = \left(s^{-1/d}\left(\frac{n}{s}\right)^{1/2}\right)^{1/4} \cdot \left[\left(\frac{n}{s}\right)^{-1/8}\left(\log\log(n/s)\right)^{1/2}\right],$$

and since $n/s \to \infty$, our claim follows. Therefore, all the conditions of Theorem 9 are satisfied and the result follows.

The second part of result is implied by the first part since if $s = n^{\beta}$ and $\beta \in (d/(d+2), 1)$ then $s^{-1/d}\sqrt{\frac{n}{s}} \to 0$.

Proofs of Propositions 4 and 5. We first start by proving Lemma 11. For proving this result, we need one auxiliary lemma, which can be derived from Hoeffding's inequality for U-statistics (Hoeffding 1994) that is stated as follows.

Proposition 8. *Suppose that* $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ *are i.i.d. and* q *is a function that has range* $[0,1]$*. Define* $U_s = \binom{n}{s}^{-1}\sum_{i_1<i_2<\ldots<i_s} q(X_{i_1}, X_{i_2}, \ldots, X_{i_s})$*. Then, for any* $\epsilon > 0$

$$\Pr\left[|U_s - \mathbb{E}[U_s]| \geq \epsilon\right] \leq 2\exp\left(-\lfloor n/s\rfloor \epsilon^2\right)\,.$$

Furthermore, for any $\delta > 0$, w.p. $1 - \delta$ we have

$$|U_s - \mathbb{E}[U_s]| \leq \sqrt{\frac{1}{\lfloor n/s \rfloor} \log(2/\delta)}\,.$$

Lemma 35. *Recall $H(s)$ defined in Section 4.5.3 and let $G_\delta(s) = \Delta\sqrt{2ps/n \log(2np/\delta)}$. Then, w.p. $1 - \delta$, for all values of $k \leq s \leq n$ we have*

$$|H(s) - \epsilon_k(s)| \leq G_\delta(s)\,.$$

Proof. Note that $H(s)$ is the complete U-statistic estimator for $\epsilon_k(s)$. For each subset S of size s of $\{1, 2, \ldots, n\}$ we have

$$\mathbb{E}\left[\max_{X_i \in H_k(x,S)} \|x - X_i\|_2\right] = \epsilon_k(s)\,.$$

Further, $\|x - x'\|_2 \leq \Delta_{\mathcal{X}} \leq \Delta$ holds for any $x' \in \mathcal{X}$. Therefore, using Hoeffding's inequality for U-statistics stated in Proposition 8, for any fixed s, w.p. $1 - \delta$ we have

$$|H(s) - \epsilon_k(s)| \leq \Delta\sqrt{\frac{1}{\lfloor n/s \rfloor} \log(2/\delta)}\,.$$

Note that $\lfloor z \rfloor \geq z/2$ for $z \geq 1$ and therefore the above translates to

$$|H(s) - \epsilon_k(s)| \leq \Delta\sqrt{\frac{2s}{n} \log(2/\delta)}\,.$$

Taking a union bound over $s = k, k+1, \ldots, n$, replacing $\delta = \delta/n$, and using $p \geq 1$, implies the result. □

Proof of Lemma 11. Note that using Lemma 35, w.p. $1 - \delta$, for all values of s we have $|H(s) - \epsilon_k(s)| \leq G_\delta(s)$. Now consider three different cases:

- $s_1 \geq s_2 \geq s^*$: Note that based on the choice of s_1, s_2, we have $H(s_2) > 2G_\delta(s_2)$. However, $H(s_2) \leq \epsilon_k(s_2) + G_\delta(s_2)$. Hence, $\epsilon_k(s_2) > G_\delta(s_2)$ which contradicts with the assumption that $s_2 \geq s^*$. Note that this is true since $\epsilon_k(s) - G_\delta(s)$ is non-positive for $s \geq s^*$.

- $s_1 = s^*, s_2 = s^* - 1$: Obviously $s_1 \leq s^*$.

- $s_2 \leq s_1 \leq s^* - 1$: Note that we have

  $$\epsilon_k(s_1) - G_\delta(s_1) \leq H(s_1) \leq 2G_\delta(s_1)\,.$$

  Hence, $G_\delta(s^* - 1) < \epsilon_k(s^* - 1) \leq \epsilon_k(s_1) \leq 3G_\delta(s_1)$. This means that $G_\delta(s^* - 1)/G_\delta(s_1) \leq 3$ which implies $\sqrt{(s^* - 1)/s_1} \leq 3$. Therefore, $s_1 \geq (s^* - 1)/9$.

Putting together all these cases finishes the proof. □

Now we are ready to finalize the proof of Proposition 4.

Proof of Proposition 4. Note that using the result of Lemma 11, w.p. $1 - \delta$, we have

$$\frac{s^* - 1}{9} \leq s_1 \leq s^*\,.$$

This basically means that if $s_* = 9s_1 + 1$, then s_* belongs to $[s^*, 10s^*]$. Hence, we have $\epsilon_k(s_*) \leq \epsilon_k(s^*) \leq G_\delta(s^*)$ and $G_\delta(s_*) \leq G_\delta(10s^*) = \sqrt{10}G_\delta(s^*)$. Now using Theorem 8, for $B \geq n/s_*$ w.p. $1 - \delta$ we have

$$\|\hat{\theta} - \theta(x)\|_2 \leq \frac{2}{\lambda}\left(L_m \epsilon(s_*) + O\left(\psi_{\max}\sqrt{\frac{p\, s_*}{n}\left(\log\log(n/s_*) + \log(p/\delta)\right)}\right)\right)\,.$$

Note that $G_\delta(s_*) = \Delta\sqrt{\frac{2ps_*}{n}\log(2pn/\delta)}$. Therefore,

$$\sqrt{\frac{p\, s_*}{n}\left(\log\log(n/s_*) + \log(p/\delta)\right)} \leq G_\delta(s_*) \leq \sqrt{10}G_\delta(s^*)\,.$$

Replacing this in above equation together with a union bound implies that w.p. at least $1 - 2\delta$ we have

$$\|\hat{\theta} - \theta(x)\|_2 = O(G_\delta(s^*))\,,$$

which finishes the first part of the proof. For the second part, note that according to Corollary 4, for the k-NN kernel $\epsilon(s) \leq Cs^{-1/d}$, for a constant C. Note that according to the definition of s^*, for $s = s^* - 1$ we have

$$\Delta\sqrt{\frac{2ps}{n}\log(2np/\delta)} = \epsilon_k(s) \leq Cs^{-1/d}\,,$$

for a constant C. The above implies that

$$s^* \leq 1 + \left(\frac{C}{\Delta}\right)^{2d/(d+2)} \left(\frac{n}{2p\log(2np/\delta)}\right)^{d/(d+2)} \leq 2\left(\frac{C}{\Delta}\right)^{2d/(d+2)} \left(\frac{n}{2p\log(2np/\delta)}\right)^{d/(d+2)} .$$

Hence,

$$G_\delta(s^*) \leq \sqrt{2}\Delta^{2/(d+2)} C^{d/(d+2)} \left(\frac{n}{2p\log(2np/\delta)}\right)^{-1/(d+2)} ,$$

which concludes the proof. □

Finally, we can use Lemma 11 to also prove Proposition 5.

Proof of Proposition 5. Note that according to Lemma 11, w.p. $1 - 1/n$, the output of process, s_1 satisfies

$$\frac{s^* - 1}{9} \leq s_1 \leq s^* ,$$

where s^* is the point for which we have $\epsilon_k(s^*) = G_{1/n}(s^*)$. This basically means that $s_* = 9s_1 + 1 \geq s^*$. Note that for the k-NN kernel we have $\eta_k(s) = \Theta(1/s)$. As $s_\zeta \geq n^\zeta$, this also implies that $\epsilon_k(s_\zeta, \eta_k(s_\zeta)) = O((\log(s_\zeta)/s_\zeta)^{1/d}) \to 0$. Also, according to the inequality $\zeta < \frac{\log(n)-\log(s_*)-\log\log^2(n)}{\log(n)}$ we have $1 - \zeta > (\log(s_*) + \log\log^2(n))/\log(n)$ and therefore

$$n^{1-\zeta} \geq s_\zeta \log^2(n) \to \frac{s_\zeta}{n} \leq \frac{1}{\log^2(n)} ,$$

and hence $n\eta_k(s_\zeta) \to 0$. Finally, note that $\sigma_{n,j}(x) = \Theta(\sqrt{s/n})$ and according to Theorem 9 it suffices that

$$\max\left(\epsilon_k(s_\zeta)\left(\frac{s_\zeta}{n}\right)^{-1/2}, \epsilon_k(s_\zeta)^{1/4}(\log\log(n/s_\zeta))^{1/2}, \left(\frac{s_\zeta}{n}\right)^{1/8}(\log\log(n/s_\zeta))^{5/8}\right) = o(1) .$$

Note that for any $\zeta > 0, s_\zeta \geq s^*$ and therefore $\epsilon_k(s_\zeta) \leq \epsilon_k(s^*) = G_{1/n}(s^*)$. For the first term,

$$\begin{aligned}\epsilon_k(s_\zeta)\left(\frac{s_\zeta}{n}\right)^{-1/2} &\leq G_{1/n}(s^*)\left(\frac{s_\zeta}{n}\right)^{-1/2} \\ &= \Delta\sqrt{\frac{2p\, s^*}{n}\log(2n^2/p)}\,\left(\frac{s_\zeta}{n}\right)^{-1/2} \\ &= O\left(\sqrt{\frac{s^*}{s_\zeta}}\log(n)\right) .\end{aligned}$$

Now note that $s_\zeta = s_* n^\zeta \geq s^* n^\zeta$ and hence $\sqrt{s^*/s_\zeta}\log(n) = O(n^{-\zeta/2}\log(n)) \to 0$. For the second term, note that again $s_\zeta \geq s^*$ and therefore $\epsilon_k(s_\zeta) \leq \epsilon_k(s^*) = G_{1/n}(s^*) \leq G_{1/n}(s_\zeta)$. Now note that since $s_\zeta/n \leq 1/\log^2(n)$ hence

$$\epsilon_k(s_\zeta)^{1/4}\log\log(n/s_\zeta)^{1/2} \leq G_{1/n}(s_\zeta)\log\log(n) = O\left(\left(\frac{\log(n)}{\log^2(n)}\right)^{1/8}\log\log(n)\right) \to 0\,.$$

Finally, for the last term we have $s_\zeta/n \leq 1/\log^2(n)$ and hence

$$\left(\frac{s_\zeta}{n}\right)^{1/8}(\log\log(n/s_\zeta))^{5/8} \leq \left(\frac{1}{\log(n)}\right)^{1/4}\log\log(n) \to 0.$$

This basically means w.p. $1 - 1/n$, s_ζ belongs to the interval for which the asymptotic normality result in Theorem 9 holds. As $n \to \infty$, the conclusion follows. □

C.3.10 Proof of Lemma 31

We will argue asymptotic normality of the U-statistic defined as:

$$\Psi_{0,\beta}(x;\theta(x)) = \binom{n}{s}^{-1}\sum_{b\subset[n]:|b|=s}\mathbb{E}_{\omega_b}\left[\sum_{i\in S_b}\alpha_{S_b,\omega_b}(X_i)\psi_\beta(Z_i;\theta(x))\right]$$

under the assumption that for any subset of indices S_b of size s: $\mathbb{E}\left[\mathbb{E}[\alpha_{S_b,\omega_b}(X_1)|X_1]^2\right] = \eta(s)$ and that the kernel satisfies shrinkage in probability with rate $\epsilon(s,\delta)$ such that $\epsilon(s,\eta(s)^2) \to 0$ and $n\eta(s) \to \infty$. For simplicity of notation we let:

$$Y_i = \psi_\beta(Z_i;\theta(x)) \tag{C.3.21}$$

and we then denote:

$$\Phi(Z_1,\ldots,Z_s) = \mathbb{E}_\omega\left[\sum_{i=1}^s K_\omega(x,X_i,\{Z_j\}_{j=1}^s)\,Y_i\right]\,. \tag{C.3.22}$$

Observe that we can then re-write our U-statistic as:

$$\Psi_{0,\beta}(x;\theta(x)) = \binom{n}{s}^{-1}\sum_{1\leq i_1\leq\ldots\leq i_s\leq n}\Phi(Z_{i_1},\ldots,Z_{i_s})\,.$$

Moreover, observe that by the definition of Y_i, $\mathbb{E}[Y_i \mid X_i] = 0$ and also

$$|Y_i| \leq \|\beta\|_2 \|M_0^{-1}(\psi(Z_i;\theta(x)) - m(X_i;\theta(x))\|_2 2 \leq \frac{R}{\lambda}\|\psi(Z_i;\theta(x))\|_2 \leq 2\frac{R\sqrt{p}}{\lambda}\psi_{\max} \triangleq y_{\max}\,.$$

Invoking Lemma 37, it suffices to show that: $\mathrm{Var}\left[\Phi_1(Z_1)\right] = \Omega(\eta(s))$, where $\Phi_1(z_1) = \mathbb{E}[\Phi(z_1, Z_2, \ldots, Z_s)]$. The following lemma shows that under our conditions on the kernel, the latter property holds.

Lemma 36. *Suppose that the kernel K is symmetric (Assumption 8), has been built in an honest manner (Assumption 7) and satisfies:*

$$\mathbb{E}\left[\mathbb{E}\left[K(x, X_1, \{Z_j\}_{j=1}^s) \mid X_1\right]^2\right] = \eta(s) \leq 1 \qquad \textit{and} \qquad \epsilon(s, \eta(s)^2) \to 0\,.$$

Then, the following holds

$$\mathrm{Var}\left[\Phi_1(Z_1)\right] \geq \mathrm{Var}(Y \mid X = x)\,\eta(s) + o(\eta(s)) = \Omega\left(\eta(s)\right)\,.$$

Proof. Note we can write

$$\Phi_1(Z_1) = \underbrace{\mathbb{E}\left[\Phi(Z_1, \ldots, Z_s) \mid X_1\right]}_{A} + \underbrace{\mathbb{E}\left[\Phi(Z_1, \ldots, Z_s) \mid X_1, Y_1\right] - \mathbb{E}\left[\Phi(Z_1, \ldots, Z_s) \mid X_1\right]}_{B}.$$

Here, B is zero mean conditional on X_1 and also A and B are uncorrelated, i.e., $\mathbb{E}[AB] = \mathbb{E}[A]\mathbb{E}[B] = 0$. Therefore:

$$\begin{aligned}\mathrm{Var}\left[\Phi_1(Z_1)\right] &\geq \mathrm{Var}\left[B\right] \\ &= \mathrm{Var}\left[\sum_{i=1}^s \left(\mathbb{E}\left[K(x, X_i, \{Z_j\}_{j=1}^s)Y_i \mid X_1, Y_1\right] - \mathbb{E}[K(x, X_i, \{Z_j\}_{j=1}^s)Y_i \mid X_1]\right)\right].\end{aligned}$$

For simplicity of notation let $W_i = K(x, X_1, \{Z_j\}_{j=1}^s)$ denote the random variable which corresponds to the weight of sample i. Note that thanks to the honesty of kernel defined in Assumption 7, W_i is independent of Y_1 conditional on X_1, for $i \geq 2$. Hence all the corresponding terms in the summation are zero. Therefore, the expression inside the variance above simplifies to

$$\mathbb{E}[W_1 Y_1 \mid X_1, Y_1] - \mathbb{E}[W_1 Y_1 \mid X_1]\,.$$

Moreover, by honesty W_1 is independent of Y_1 conditional on X_1. Thus, the above further simplifies to:

$$\mathbb{E}[W_1 \mid X_1]\ (Y_1 - \mathbb{E}[Y_1 \mid X_1])\ .$$

Using $\mathrm{Var}(G) = \mathbb{E}[G^2] - \mathbb{E}[G]^2$, this can be further rewritten as

$$\mathrm{Var}\left[\Phi_1(Z_1)\right] \geq \mathbb{E}\left[\mathbb{E}[W_1 \mid X_1]^2 \left(Y_1 - \mathbb{E}[Y_1 \mid X_1]\right)^2\right] - \mathbb{E}\left[\mathbb{E}[W_1 \mid X_1](Y_1 - \mathbb{E}[Y_1 \mid X_1])\right]^2\ .$$

Note that $Y_1 - \mathbb{E}[Y_1 \mid X_1]$ is uniformly upper bounded by some $\psi_{\max}$. Furthermore, by the symmetry of the kernel we have $\mathbb{E}\left[\mathbb{E}[W_1 \mid X_1]\right] = \mathbb{E}[W_1] = 1/s$.[1] Thus the second term in the latter is of order $1/s^2$. Hence:

$$\mathrm{Var}\left[\Phi(Z_1)\right] \geq \mathbb{E}\left[\mathbb{E}[\alpha_b(X_1) \mid X_1]^2 \left(Y_1 - \mathbb{E}[Y_1 \mid X_1]\right)^2\right] + o(1/s)\,.$$

Focusing on the first term and letting $\sigma^2(x) = \mathrm{Var}(Y|X=x)$, we have:

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{E}[W_1 \mid X_1]^2 \left(Y_1 - \mathbb{E}[Y_1 \mid X_1]\right)^2\right] \\
=&\mathbb{E}\left[\mathbb{E}[W_1 \mid X_1]^2 \sigma^2(X_1)\right] \\
=&\mathbb{E}\left[\mathbb{E}[W_1 \mid X_1]^2\right]\sigma^2(x) + \mathbb{E}\left[\mathbb{E}[W_1 \mid X_1]^2\left(\sigma^2(X_1) - \sigma^2(x)\right)\right]\ .
\end{aligned}$$

The goal is to prove that the second term is $o(1/s)$. For ease of notation let $V_1 = \mathbb{E}\left[W_1 \mid X_1\right]$. Then we can bound the second term as:

$$\begin{aligned}
\left|\mathbb{E}\left[V_1^2\left(\sigma^2(X_1) - \sigma^2(x)\right)\right]\right| \leq\ & L_{mm}\epsilon(s,\delta)\,\mathbb{E}\left[V_1^2\ 1\left\{\|x - X_1\|_2 \leq \epsilon(s,\delta)\right\}\right] \\
& \qquad + 2y_{\max}^2\mathbb{E}\left[V_1^2\,1\left\{\|x - X_1\|_2 > \epsilon(s,\delta)\right\}\right] \\
\leq\ & L_{mm}\epsilon(s,\delta)\,\mathbb{E}\left[V_1^2\right] + 2y_{\max}^2\mathbb{E}\left[V_1^2\,1\left\{\|x - X_1\|_2 > \epsilon(s,\delta)\right\}\right] \\
\leq\ & L_{mm}\epsilon(s,\delta)\eta(s) + 2y_{\max}^2\mathbb{E}\left[V_1\,1\left\{\|x - X_1\|_2 > \epsilon(s,\delta)\right\}\right] \\
\leq\ & L_{mm}\epsilon(s,\delta)\eta(s) + 2y_{\max}^2\mathbb{E}\left[W_1\,1\left\{\|x - X_1\|_2 > \epsilon(s,\delta)\right\}\right]\ ,
\end{aligned}$$

where we used the fact that $V_1 \leq 1$, the assumption that $\sigma^2(\cdot)$ is L_{mm}-Lipschitz, the tower

[1] Since $\mathbb{E}[W_i]$ are all equal to the same value κ and $\sum_i \mathbb{E}[W_i] = 1$, we get $\kappa = 1/s$.

rule and the definition of $\eta(s)$. Furthermore,

$$\begin{aligned}\mathbb{E}\left[W_1\, 1\left\{\|x - X_1\|_2 > \epsilon(s,\delta)\right\}\right] &\leq \Pr\left[\|x - X_1\|_2 \geq \epsilon(s,\delta) \text{ and } W_1 > 0\right]\\ &\leq \Pr\left[\sup_i \left\{\|x - X_i\|_2 : W_i > 0\right\} \geq \epsilon(s,\delta)\right]\,,\end{aligned}$$

which by definition is at most δ. By putting $\delta = \eta(s)^2$ we obtain

$$\left|\mathbb{E}\left[\mathbb{E}[W_1 \mid X_1]^2\left(\sigma^2(X_1) - \sigma^2(x)\right)\right]\right| \leq L_{mm}\epsilon(s,\eta(s)^2)\eta(s) + 2y_{\max}^2\eta(s)^2 = o(\eta(s))\,,$$

where we invoked our assumption that $\epsilon(s,\eta(s)^2) \to 0$. Thus we have obtained that:

$$\operatorname{Var}\left[\Phi_1(Z_1)\right] \geq \mathbb{E}\left[\mathbb{E}[W_1 \mid X_1]^2\right]\sigma^2(x) + o(\eta(s))\,,$$

which is exactly the form of the lower bound claimed in the statement of the lemma. This concludes the proof. □

C.3.11 Hájek Projection Lemma for Infinite Order U-Statistics

The following is a small adaptation of Theorem 2 of Fan et al. (2018), which we present here for completeness.

Lemma 37 (Fan et al. (2018)). *Consider a U-statistic defined via a symmetric kernel Φ:*

$$U(Z_1,\ldots,Z_n) = \binom{n}{s}^{-1}\sum_{1\leq i_1\leq\ldots\leq i_s\leq n}\Phi(Z_{i_1},\ldots,Z_{i_s})\,, \tag{C.3.23}$$

where Z_i are i.i.d. random vectors and s can be a function of n. Let $\Phi_1(z_1) = \mathbb{E}[\Phi(z_1, Z_2,\ldots,Z_s)]$ and $\eta_1(s) = \operatorname{Var}_{z_1}\left[\Phi_1(z_1)\right]$. Suppose that $\operatorname{Var}\Phi$ *is bounded, $n\,\eta_1(s) \to \infty$. Then:*

$$\frac{U(Z_1,\ldots,Z_n) - \mathbb{E}\left[U\right]}{\sigma_n} \to_d \mathsf{N}(0,1)\,, \tag{C.3.24}$$

where $\sigma_n^2 = \frac{s^2}{n}\eta_1(s)$.

Proof. The proof follows identical steps as the one in Fan et al. (2018). We argue about the asymptotic normality of a U-statistic:

$$U(Z_1, \ldots, Z_n) = \binom{n}{s}^{-1} \sum_{1 \leq i_1 \leq \ldots \leq i_s \leq n} \Phi(Z_{i_1}, \ldots, Z_{i_s}) \,. \tag{C.3.25}$$

Consider the following projection functions:

$$\begin{aligned}
\Phi_1(z_1) =&\ \mathbb{E}[\Phi(z_1, Z_2, \ldots, Z_s)]\,, & \tilde{\Phi}_1(z_1) =&\ \Phi_1(z_1) - \mathbb{E}\left[\Phi\right]\,, \\
\Phi_2(z_1, z_2) =&\ \mathbb{E}[\Phi(z_1, z_2, Z_3, \ldots, Z_s)\,, & \tilde{\Phi}_2(z_1, z_2) =&\ \Phi_2(z_1, z_2) - \mathbb{E}\left[\Phi\right]\,, \\
&\ \vdots & & \\
\Phi_s(z_1, z_2, \ldots, z_s) =&\ \mathbb{E}[\Phi(z_1, z_2, Z_3, \ldots, Z_s)\,, & \tilde{\Phi}_s(z_1, z_2, \ldots, z_s) =&\ \Phi_s(z_1, z_2, \ldots, z_s) - \mathbb{E}\left[\Phi\right]\,,
\end{aligned}$$

where $\mathbb{E}\left[\Phi\right] = \mathbb{E}\left[\Phi(Z_1, \ldots, Z_s)\right]$. Then we define the canonical terms of Hoeffding's U-statistic decomposition as:

$$\begin{aligned}
g_1(z_1) =&\ \tilde{\Phi}_1(z_1)\,, \\
g_2(z_1, z_2) =&\ \tilde{\Phi}_2(z_1, z_2) - g_1(z_1) - g_2(z_2)\,, \\
g_3(z_1, z_2, z_3) =&\ \tilde{\Phi}_2(z_1, z_2, Z_3) - \sum_{i=1}^{3} g_1(z_i) - \sum_{1 \leq i < j \leq 3} g_2(z_i, z_j)\,, \\
&\ \vdots \\
g_s(z_1, z_2, \ldots, z_s) =&\ \tilde{\Phi}_s(z_1, z_2, \ldots, z_s) - \sum_{i=1}^{s} g_1(z_i) - \sum_{1 \leq i < j \leq s} g_2(z_i, z_j) - \ldots \\
&\qquad \ldots - \sum_{1 \leq i_1 < i_2 < \ldots < i_{s-1} \leq s} g_{s-1}(z_{i_1}, z_{i_2}, \ldots, z_{i_{s-1}})\,.
\end{aligned}$$

Subsequently the kernel of the U-statistic can be re-written as a function of the canonical terms:

$$\tilde{\Phi}(z_1, \ldots, z_s) = \Phi(z_1, \ldots, z_s) - \mathbb{E}\left[\Phi\right] = \sum_{i=1}^{s} g_1(z_i) + \sum_{1 \leq i < j \leq s} g_2(z_i, z_j) + \ldots + g_s(z_1, \ldots, z_s)\,. \tag{C.3.26}$$

Moreover, observe that all the canonical terms in the latter expression are un-correlated.

Hence, we have:

$$\mathrm{Var}\left[\Phi(Z_1,\dots,Z_n)\right] = \binom{s}{1}\mathbb{E}\left[g_1^2\right] + \binom{s}{2}\mathbb{E}\left[g_2^2\right] + \ldots + \binom{s}{s}\mathbb{E}\left[g_s^2\right] . \tag{C.3.27}$$

We can now re-write the U statistic also as a function of canonical terms:

$$\begin{aligned} U(Z_1,\dots,Z_n) - \mathbb{E}\left[U\right] = &\binom{n}{s}^{-1} \sum_{1\le i_1<i_2<\ldots<i_s\le n} \tilde{\Phi}(Z_{i_1},\dots,Z_{i_s}) \\ = &\binom{n}{s}^{-1}\left(\binom{n-1}{s-1}\sum_{i=1}^{n} g_1(Z_i) + \binom{n-2}{s-2}\sum_{1\le i<j\le n} g_2(Z_i,Z_j) + \ldots \right. \\ &\left. + \binom{n-s}{s-s}\sum_{1\le i_1<i_2<\ldots<i_s\le n} g_s(Z_{i_1},\dots,Z_{i_s})\right) . \end{aligned}$$

Now we define the Hájek projection to be the leading term in the latter decomposition:

$$\hat{U}(Z_1,\dots,Z_n) = \binom{n}{s}^{-1}\binom{n-1}{s-1}\sum_{i=1}^{n} g_1(Z_i) . \tag{C.3.28}$$

The variance of the Hajek projection is:

$$\sigma_n^2 = \mathrm{Var}\left[\hat{U}(Z_1,\dots,Z_n)\right] = \frac{s^2}{n}\,\mathrm{Var}\left[\Phi_1(z_1)\right] = \frac{s^2}{n}\eta_1(s) . \tag{C.3.29}$$

The Hájek projection is the sum of independent and identically distributed terms and hence by the Lindeberg-Levy Central Limit Theorem (see, e.g., Billingsley 2008, Borovkov 2013):

$$\frac{\hat{U}(Z_1,\dots,Z_n)}{\sigma_n} \to_d \mathsf{N}(0,1) . \tag{C.3.30}$$

We now argue that the remainder term: $\frac{U-\mathbb{E}[U]-\hat{U}}{\sigma_n}$ vanishes to zero in probability. The latter then implies that $\frac{U-\mathbb{E}[U]}{\sigma_n} \to_d \mathsf{N}(0,1)$ as desired. We will show the sufficient condition of convergence in mean square: $\frac{\mathbb{E}\left[\left(U-\mathbb{E}[U]-\hat{U}\right)^2\right]}{\sigma_n^2} \to 0$. From an inequality due to Wager and

Athey (2018):

$$\begin{aligned}
\mathbb{E}\left[\left(U-\mathbb{E}\left[U\right]-\hat{U}\right)^2\right] &= \binom{n}{s}^{-2}\left\{\binom{n-2}{s-2}^2\binom{n}{2}\mathbb{E}[g_2^2]+\ldots+\binom{n-s}{s-s}^2\binom{n}{s}\mathbb{E}[g_s^2]\right\}\\
&= \sum_{r=2}^{s}\left\{\binom{n}{s}^{-2}\binom{n-r}{s-r}^2\binom{n}{r}\mathbb{E}[g_r^2]\right\}\\
&= \sum_{r=2}^{s}\left\{\frac{s!(n-r)!}{n!(s-r)!}\binom{s}{r}\mathbb{E}[g_r^2]\right\}\\
&\leq \frac{s(s-1)}{n(n-1)}\sum_{r=2}^{s}\binom{s}{r}\mathbb{E}[g_r^2]\\
&\leq \frac{s^2}{n^2}\operatorname{Var}\left[\Phi(Z_1,\ldots,Z_s)\right]\,.
\end{aligned}$$

Since $\operatorname{Var}\left[\Phi(Z_1,\ldots,Z_n)\right]$ is bounded by a constant V^* and $n\,\eta_1(s)\to\infty$, by our assumption, we have:

$$\frac{\mathbb{E}\left[\left(U-\mathbb{E}\left[U\right]-\hat{U}\right)^2\right]}{\sigma_n^2}\leq\frac{\frac{s^2}{n^2}V^*}{\frac{s^2}{n}\eta_1}=\frac{V^*}{n\,\eta_1(s)}\to 0\,. \tag{C.3.31}$$

□

C.3.12 Stochastic Equicontinuity of U-Statistics via Bracketing

We define here some standard terminology on bracketing numbers in empirical process theory. Consider an arbitrary function space $\mathcal{F}$ of functions from a data space $\mathcal{Z}$ to $\mathbb{R}$, equipped with some norm $\|\cdot\|$. A *bracket* $[a,b]\subseteq\mathcal{F}$, where $a,b:\mathcal{Z}\to\mathbb{R}$ consists of all functions $f\in\mathcal{F}$, such that $a\leq f\leq b$. An *ϵ-bracket* is a bracket $[a,b]$ such that $\|a-b\|\leq\epsilon$. The *bracketing number* $N_{[]}(\epsilon,\mathcal{F},\|\cdot\|)$ is the minimum number of ϵ-brackets needed to cover $\mathcal{F}$. The functions $[a,b]$ used in the definition of the brackets need not belong to $\mathcal{F}$ but satisfy the same norm constraints as functions in $\mathcal{F}$. Finally, for an arbitrary measure P on $\mathcal{Z}$, let

$$\|f\|_{P,2}=\sqrt{\mathbb{E}_{Z\sim P}[f(Z)^2]}\qquad\qquad \|f\|_{P,\infty}=\sup_{z\,\in\,\text{support}(P)}|f(z)| \tag{C.3.32}$$

Lemma 38 (Stochastic Equicontinuity for U-Statistics via Bracketing)**.** *Consider a function space $\mathcal{F}$ of symmetric functions from some data space $\mathcal{Z}^s$ to $\mathbb{R}$ and consider the U-statistic of order s, with kernel f over n samples:*

$$\Psi_s(f, z_{1:n}) = \binom{n}{s}^{-1} \sum_{1\leq i_1\leq \ldots \leq i_s \leq n} f(z_{i_1}, \ldots, z_{i_s}) \tag{C.3.33}$$

Suppose $\sup_{f\in\mathcal{F}} \|f\|_{P,2} \leq \eta$, $\sup_{f\in\mathcal{F}} \|f\|_{P,\infty} \leq G$ *and let* $\kappa = n/s$*. Then for* $\kappa \geq \frac{G^2}{\log N_{[]}(1/2, \mathcal{F}, \|\cdot\|_{P,2})}$*, w.p.* $1-\delta$*:*

$$\sup_{f\in\mathcal{F}} |\Psi_s(f, Z_{1:n}) - \mathbb{E}[f(Z_{1:s})]| \\ = O\left(\inf_{\rho>0} \frac{1}{\sqrt{\kappa}} \int_{\rho}^{2\eta} \sqrt{\log(N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|_{P,2})} + \eta\sqrt{\frac{\log(1/\delta) + \log\log(\eta/\rho)}{\kappa}} + \rho\right)$$

Proof. Let $\kappa = n/s$. Moreover, wlog we will assume that $\mathcal{F}$ contains the zero function, as we can always augment $\mathcal{F}$ with the zero function without changing the order of its bracketing number. For $q = 1, \ldots, M$, let $\mathcal{F}_q = \cup_{i=1}^{N_q} \mathcal{F}_{qi}$ be a partition of $\mathcal{F}$ into brackets of diameter at most $\epsilon_q = 2\eta/2^q$, with $\mathcal{F}_0$ containing a single partition of all the functions. Moreover, we assume that $\mathcal{F}_q$ are nested partitions. We can achieve the latter as follows: i) consider a minimal bracketing cover of $\mathcal{F}$ of diameter ϵ_q, ii) assign each $f \in \mathcal{F}$ to one of the brackets that it is contained arbitrarily and define the partition $\bar{\mathcal{F}}_q$ of size $\bar{N}_q = N_{[]}(\epsilon_q, \mathcal{F}, \|\cdot\|_{P,2})$, by taking $\mathcal{F}_{qi}$ to be the functions assigned to bracket i, iii) let $\mathcal{F}_q$ be the common refinement of all partitions $\bar{\mathcal{F}}_0, \ldots, \bar{\mathcal{F}}_q$. The latter will have size at most $N_q \leq \prod_{q=0}^{M} \bar{N}_q$. Moreover, assign a representative function f_{qi} to each partition $\mathcal{F}_{qi}$, with the representative for the single partition at level $q = 0$ is the zero function.

Chaining Definitions. Consider the following random variables, where the dependence on the random input Z is hidden:

$$\begin{aligned} \pi_q f =&\ f_{qi}, \quad \text{if } f \in \mathcal{F}_{qi} \\ \Delta_q f =&\ \sup_{g,h\in\mathcal{F}_{qi}} |g-h|, \quad \text{if } f \in \mathcal{F}_{qi} \\ B_q f =&\ \{\Delta_0 f \leq \alpha_0, \ldots, \Delta_{q-1} f \leq \alpha_{q-1}, \Delta_q f > \alpha_q\} \\ A_q f =&\ \{\Delta_0 f \leq \alpha_0, \ldots, \Delta_q f \leq \alpha_q\}, \end{aligned}$$

for some sequence of numbers $\alpha_0, \ldots, \alpha_M$, to be chosen later. By noting that $A_{q-1}f = A_q f + B_q f$ and continuously expanding terms by adding and subtracting finer approximations to f, we can write the telescoping sum:

$$\begin{aligned} f - \pi_0 f = &\ (f - \pi_0 f)\, B_0 f + (f - \pi_0 f) A_0 f \\ = &\ (f - \pi_0 f)\, B_0 f + (f - \pi_1 f) A_0 f + (\pi_1 f - \pi_0 f) A_0 f \\ = &\ (f - \pi_0 f)\, B_0 f + (f - \pi_1 f) B_1 f + (f - \pi_1 f) A_1 f + (\pi_1 f - \pi_0 f) A_0 f \\ & \ldots \\ = &\ \sum_{q=0}^{M} (f - \pi_q f) B_q f + \sum_{q=1}^{M} (\pi_q f - \pi_{q-1} f) A_{q-1} f + (f - \pi_M f) A_M f\,. \end{aligned}$$

For simplicity let $\mathbb{P}_{s,n} f = \Psi(f, Z_{1:n})$, $\mathbb{P} f = E[f(Z_{1:s})]$ and $\mathbb{G}_{s,n}$ denote the U-process:

$$\mathbb{G}_{s,n} f = \mathbb{P}_{s,n} f - \mathbb{P} f\,. \tag{C.3.34}$$

Our goal is to bound $\|\mathbb{P}_{s,n} f\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |\mathbb{P}_{s,n} f|$, with high probability. Observe that since $\mathcal{F}_0$ contains only the zero function, then $\mathbb{G}_{s,n} f_0 = 0$. Moreover, the operator $\mathbb{G}_{s,n}$ is linear. Thus:

$$\begin{aligned} \mathbb{G}_{s,n} f &= \mathbb{G}_{s,n}(f - \pi_0 f) \\ &= \sum_{q=0}^{M} \mathbb{G}_{s,n}(f - \pi_q f) B_q f + \sum_{q=1}^{M} \mathbb{G}_{s,n}(\pi_q f - \pi_{q-1} f) A_{q-1} f + \mathbb{G}_{s,n}(f - \pi_M f) A_M f\,. \end{aligned}$$

Moreover, by triangle inequality:

$$\begin{aligned} \|\mathbb{G}_{s,n} f\|_{\mathcal{F}} \leq & \sum_{q=0}^{M} \|\mathbb{G}_{s,n}(f - \pi_q f) B_q f\|_{\mathcal{F}} \\ & \qquad + \sum_{q=1}^{M} \|\mathbb{G}_{s,n}(\pi_q f - \pi_{q-1} f) A_{q-1} f\|_{\mathcal{F}} + \|\mathbb{G}_{s,n}(f - \pi_M f) A_M f\|_{\mathcal{F}}\,. \end{aligned}$$

We will bound each term in each summand separately.

Edge Cases. The final term we will simply bound it by $2\alpha_M$, since $|(f - \pi_M f) A_M f| \leq \alpha_M$, almost surely. Moreover, the summand in the first term for $q = 0$, we bound as follows. Observe that $B_0 f = 1\{\sup_f |f| > \alpha_0\}$. But we know that $\sup_f |f| \leq G$, hence: $B_0 f \leq 1\{G > \alpha_0\}$.

$$\mathbb{G}_{s,n}(f - \pi_0 f) B_0 f = \mathbb{G}_{s,n} f B_0 f \leq |\mathbb{P}_{s,n} f B_0 f| + |\mathbb{P} f B_0 f| \leq 2G\, 1\{G > \alpha_0\}\,.$$

Hence, if we assume that α_0 is large enough such that $\alpha_0 > G$, then the latter term is zero. By the setting of α_0 that we will describe at the end, the latter would be satisfied if $\kappa \geq \frac{G^2}{\log N_{[]}(1/2, \mathcal{F}, \|\cdot\|_{P,2})}$.

B_q **Terms.** For the terms in the first summand we have by triangle inequality:

$$\begin{aligned}
|\mathbb{G}_{s,n}(f - \pi_q f) B_q f| \leq&\ \mathbb{P}_{s,n}|f - \pi_q f| B_q f + \mathbb{P}|f - \pi_q f| B_q f\\
\leq&\ \mathbb{P}_{s,n}\Delta_q f B_q f + \mathbb{P}\Delta_q f B_q f\\
\leq&\ \mathbb{G}_{s,n}\Delta_q f B_q f + 2\mathbb{P}\Delta_q f B_q f\,.
\end{aligned}$$

Moreover, observe that:

$$\begin{aligned}
\mathbb{P}\Delta_q f B_q f \leq \mathbb{P}\Delta_q f 1\{\Delta_q f > \alpha_q\} \leq \frac{1}{\alpha_q}\mathbb{P}(\Delta_q f)^2 1\{\Delta_q f > \alpha_q\}\\
\leq \frac{1}{\alpha_q}\mathbb{P}(\Delta_q f)^2 = \frac{1}{\alpha_q}\|\Delta_q f\|_{P,2}^2 \leq \frac{\epsilon_q^2}{\alpha_q}\,,
\end{aligned}$$

where we used the fact that the partitions in $\mathcal{F}_q$, have diameter at most ϵ_q, with respect to the $\|\cdot\|_{P,2}$ norm. Now observe that because the partitions $\mathcal{F}_q$ are nested, $\Delta_q f \leq \Delta_{q-1} f$. Therefore, $\Delta_q f B_q f \leq \Delta_{q-1} f B_q f \leq \alpha_{q-1}$, almost surely. Moreover, $\|\Delta_q f B_q f\|_{P,2} \leq \|\Delta_q f\|_{P,2} \leq \epsilon_q$. By Bernstein's inequality for U statistics (see, e.g., Peel et al. 2010) for any fixed f, w.p. $1 - \delta$:

$$|\mathbb{G}_{s,n}\Delta_q f B_q f| \leq\ \epsilon_q\sqrt{\frac{2\log(2/\delta)}{\kappa}} + \alpha_{q-1}\frac{2\log(2/\delta)}{3\kappa}\,.$$

Taking a union bound over the N_q members of the partition, and combining with the bound on $\mathbb{P}\Delta_q f B_q f$, we have w.p. $1 - \delta$:

$$\|\mathbb{G}_{s,n}(f - \pi_q f) B_q f\|_{\mathcal{F}} \leq \epsilon_q\sqrt{\frac{2\log(2N_q/\delta)}{\kappa}} + \alpha_{q-1}\frac{2\log(2N_q/\delta)}{3\kappa} + \frac{2\epsilon_q^2}{\alpha_q}\,. \tag{C.3.35}$$

A_q **Terms.** For the terms in the second summand, we have that since the partitions are nested, $|(\pi_q f - \pi_{q-1} f) A_{q-1} f| \leq \Delta_{q-1} f A_{q-1} f \leq \alpha_{q-1}$. Moreover, $\|(\pi_q f - \pi_{q-1} f) A_{q-1} f\|_{P,2} \leq \|\Delta_{q-1} f\|_{P,2} \leq \epsilon_{q-1} \leq 2\epsilon_q$. Thus, by similar application of Bernstein's inequality for U-statistics, we have for a fixed f, w.p. $1-\delta$:

$$|\mathbb{G}_{s,n}(\pi_q f - \pi_{q-1} f) A_{q-1} f| \leq \epsilon_q \sqrt{\frac{8\log(2/\delta)}{\kappa}} + \alpha_{q-1} \frac{2\log(2/\delta)}{3\kappa}\,.$$

As f ranges there are at most $N_{q-1}\, N_q \leq N_q^2$ different functions $(\pi_q f - \pi_{q-1} f) A_{q-1} f$. Thus taking a union bound, we have that w.p. $1-\delta$:

$$\|\mathbb{G}_{s,n}(\pi_q f - \pi_{q-1} f) A_{q-1} f\|_{\mathcal{F}} \leq \epsilon_q \sqrt{\frac{16\log(2N_q/\delta)}{\kappa}} + \alpha_{q-1} \frac{4\log(2N_q/\delta)}{3\kappa}\,.$$

Taking also a union bound over the $2M$ summands and combining all the above inequalities, we have that w.p. $1-\delta$:

$$\|\mathbb{G}_{s,n} f\|_{\mathcal{F}} \leq \sum_{q=1}^{M} \epsilon_q \sqrt{\frac{32\log(2N_q M/\delta)}{\kappa}} + \alpha_{q-1} \frac{6\log(2N_q M/\delta)}{3\kappa} + \frac{2\epsilon_q^2}{\alpha_q}\,.$$

Choosing $\alpha_q = \epsilon_q \sqrt{\kappa}/\sqrt{\log(2N_{q+1}M/\delta)}$ for $q < M$ and $\alpha_M = \epsilon_M$, we have for some constant C:

$$\begin{aligned}
\|\mathbb{G}_{s,n} f\|_{\mathcal{F}} \leq\ & C \sum_{q=1}^{M} \epsilon_q \sqrt{\frac{\log(2N_q M/\delta)}{\kappa}} + 3\epsilon_M \\
\leq\ & C \sum_{q=1}^{M} \epsilon_q \sqrt{\frac{\log(N_q)}{\kappa}} + C \sum_{q=1}^{M} \epsilon_q \sqrt{\frac{\log(2M/\delta)}{\kappa}} + 3\epsilon_M \\
\leq\ & C \sum_{q=1}^{M} \epsilon_q \sqrt{\frac{\log(N_q)}{\kappa}} + 2C\eta \sqrt{\frac{\log(2M/\delta)}{\kappa}} + 3\epsilon_M\,.
\end{aligned}$$

Moreover, since $\log(N_q) \leq \sum_{t=0}^{q} \log(N_{[]}(\epsilon_q, \mathcal{F}, \|\cdot\|_{P,2})$, we have:

$$
\begin{aligned}
\sum_{q=1}^{M} \epsilon_q \sqrt{\log(N_q)} \leq\ & \sum_{q=1}^{M} \epsilon_q \sum_{t=0}^{q} \sqrt{\log(N_{[]}(\epsilon_t, \mathcal{F}, \|\cdot\|_{P,2})} = \sum_{t=0}^{M} \sqrt{\log(N_{[]}(\epsilon_t, \mathcal{F}, \|\cdot\|_{P,2})} \sum_{q=t}^{M} \epsilon_q \\
\leq\ & 2 \sum_{t=0}^{M} \epsilon_t \sqrt{\log(N_{[]}(\epsilon_t, \mathcal{F}, \|\cdot\|_{P,2})} \\
\leq\ & 4 \sum_{t=0}^{M} (\epsilon_t - \epsilon_{t+1}) \sqrt{\log(N_{[]}(\epsilon_t, \mathcal{F}, \|\cdot\|_{P,2})} \\
\leq\ & 4 \int_{\epsilon_M}^{\epsilon_0} \sqrt{\log(N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|_{P,2})}\,.
\end{aligned}
$$

Combining all the above yields the result. □

Corollary 8. *Consider a function space $\mathcal{F}$ of symmetric functions. Suppose that $\sup_{f\in\mathcal{F}} \|f\|_{P,2} \leq \eta$ and $\log(N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|_{P,2}) = O(1/\epsilon)$. Then for $\kappa \geq O(G^2)$, w.p. $1-\delta$:*

$$
\sup_{f\in\mathcal{F}} |\Psi_s(f, Z_{1:n}) - \mathbb{E}[f(Z)]| = O\left(\sqrt{\frac{\eta}{\kappa}} + \eta\sqrt{\frac{\log(1/\delta) + \log\log(\kappa/\eta)}{\kappa}}\right). \tag{C.3.36}
$$

Proof. Applying Lemma 38, we get for every $\rho > 0$, the desired quantity is upper bounded by:

$$
\begin{aligned}
& O\left(\frac{1}{\sqrt{\kappa}} \int_{\rho}^{\eta} \frac{1}{\sqrt{\epsilon}} + \eta\sqrt{\frac{\log(1/\delta) + \log\log(\eta/\rho)}{\kappa}} + \rho\right) \\
= & O\left(\frac{\sqrt{\eta} - \sqrt{\rho}}{\sqrt{\kappa}} + \eta\sqrt{\frac{\log(1/\delta) + \log\log(\eta/\rho)}{\kappa}} + \rho\right).
\end{aligned}
$$

Choosing $\rho = \sqrt{\eta}/\sqrt{\kappa}$, yields the desired bound. □