Statistical miscellany: causality, networks, and bandits

Arjun Sondhi

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2019

Reading Committee:

Ali Shojaie, Chair

Kenneth M. Rice

Noah R. Simon

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

Statistical miscellany: causality, networks, and bandits

Arjun Sondhi

Chair of the Supervisory Committee:
Dr. Ali Shojaie
Department of Biostatistics

In this dissertation, we make methodological contributions in three separate areas. In Chapter 2, we introduce a new algorithm for learning high-dimensional causal networks from observational data. Our algorithm, which is a simple modification to the well-known PC-Algorithm, provides reductions in both computational and sample complexity, by leveraging properties of common random graph families. In Chapter 3, we develop a penalized regression framework to integrate known network structure into high-dimensional generalized linear models. Our framework is unique in that it considers two-way structured data, where networks connect both the features and the observation units. We also introduce a statistical inference procedure to provide valid confidence intervals and hypothesis tests. Finally, in Chapter 4, we present an improved estimator for counterfactual policy evaluation in contextual bandit problems. This method is based on classifier-based density ratio estimation, and displays state-of-the-art performance for continuous action spaces. We conclude with a discussion in Chapter 5, describing the limitations of the work, and avenues for future research.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

This is going to be long, so if you only care about my research, you should probably skip this section. Personally, I find this to be the most interesting part of peoples' dissertations.

There are a lot of brilliant people who have been involved in my academic journey. First and foremost, I thank Ali Shojaie, my advisor, for guiding me through my PhD. He had the unenviable task of dealing with me for five years — through my unpredictable work ethic, my frequent internship escapes, and my general lack of academic demeanor — and always did so with a steady hand and the words "I'll let you make your own choices". I am extremely grateful to have been allowed the freedom that I had, and to have learned as much as I did about diverse areas in statistics. I am also indebted to the other members of my dissertation committee. I thank Ken Rice for giving me my first research experience (and first publication!) in grad school, and for teaching me the ways of effective writing and visualization. I thank Noah Simon for his enthusiasm and encouragement, and for calling me out when I did dumb things (both academic and non-academic). I thank Jing Ma for her helpful insights, and for always making me feel good about my work.

Although never involved in my research, my grad school experience has been made much better by knowing Marco Carone, who was my initial academic advisor, but has yet to give me any academic advice. I thank Marco for being a friend that I could always complain about things with while reminiscing about Canada and oversharing my personal life. I also acknowledge Jon Wakefield for being in my corner, in addition to Susanne May, Nayak Polissar, Ruth Etzioni, and Mauricio Sadinle for their various positive contributions.

Outside of UW, I have been fortunate to know other academics who have gotten me to this point, and I am thankful for my continued friendship with all of them. I thank

David Arbour, my former intern manager at Facebook, for getting me involved in a cool project within a previously unfamiliar area of statistics that ended up becoming part of my dissertation. I also thank Drew Dimmery for his contributions to this work and look forward to hanging out in New York! Finally, I want to acknowledge the unlikely, unprofessional, and always fulfilling friendships I have with Peter Aronow and Dan Simpson. I thank them both for separately teaching me that not all tenure-track faculty are boring, and for being sources of comfort in some of my worst moments.

As made famous by The Beatles, and also one of the worst scenes in *Gilmore Girls: A Year in the Life*, the phrase "I get by with a little help from my friends" is highly relevant to me. I am thankful to be forever bonded through trauma with my department cohortmates: Natalie Gasca, Kelsey Grinde, Xiaowen Tian, Phuong Vu, Brian Williamson, Fan Xia, Yuxiang Xie, Chaoyu Yu, and Rui Zhuang. I am also `#blessed` to have befriended Katie Wilson, Anu Mishra, Amarise Little, Aaron Hudson, Angela Zhang, and Taylor Okonek, all of whom I've enjoyed drinking, gossiping, and discussing pop culture with. Finally, I acknowledge all the other (bio)statistics students who I've befriended, at UW and elsewhere: Andrew Spieker, David Whitney, Katherine Tan, Asad Haris, Jean Feng, Lucy Gao, Travis Hee Wai, Ernesto Ulloa, Tracy Dong, Subodh Selukar, Adam Elder, Parker Xie, Kendrick Li, Zora Yang, Serge Aleshin-Guendel, Eric Morenz, Sheridan Grant, Nilanjana Laha, Jessica Perry, Sean Jewell, Alex Levis, and Ian Waudby-Smith. I would especially like to thank Jean Feng for being one of my favourite collaborators, Alex Levis for hosting me in Boston, and Katherine Tan for referring me to my first full-time job! To everyone in this paragraph, remember that the future of our field is in great hands with all of us, and that it's never too late to become a data scientist.

I thank Aaron Jaech for being my first friend in Seattle outside of my department and my family, leading to our "many fruitful discussions". I also thank Auberon Lopez, Xinting Li, Jenny Choi, Scott Lippert, Arjuna Hayes, Matt Garnes, Zack Meyer, and Alyssa Wong

for their friendship and for getting me out of the academic bubble sometimes. A huge thanks to my cousins, Shivaas and Megha Gulati, for their immense help in getting me settled when I first moved to Seattle and their constant availability and support afterwards.

Throughout my time in grad school, I've been fortunate enough to have several industry internships where I met some fantastic friends, including Kyra Singh, Georg Goerg, and Erika Lu. I would particularly like to acknowledge members of the summer 2018 Facebook Core Data Science intern group: Benjamin Miller, Brendan McVeigh, Claire Donnat, Jonathan Zhang, Karthik Rajkumar, Molly Offer-Westort, Wesley Lee, Wilson Cai, Xiao Ma, and Zhe Feng, all of whom are super cool party people, and indulged me as I regaled them with facts about Mindy Kaling. You guys, I'm like really smart now. You don't even know. I am also glad to have met Ben Strauber during this internship, who is my favourite globalist due to his excellent taste in Canadian teen dramas.

I owe my biggest thanks to my oldest and closest friends, all of whom I'm thrilled to still have in my life. Thank you to Carrie Xing for being my statistically significant other, to Hanzhen Guo for being the queen of shade, to Tom Peng for his fresh and unique perspective on the world, and to Jamie Murdoch for always being there to talk about everything and anything. I am also grateful that you all indulged my constant need for attention by attending my graduation ceremony. Thank you to my other friends from undergrad and high school: Kira Vimal, Aayush Rajasekaran, Matt Janzen, Priyanka Aggarwal, Ashley Nickle, Edson Iturri, Jennifer Hernes, Brunelle Lewis, and Vanessa Vidal. I may not have seen you as often as I'd have liked, but we are always able to pick up right where we left off. Unlike Erin from *Derry Girls*, I am not planning to "move on friendship-wise" because you are all my dicks.

Finally, I am forever grateful to my mom, dad, and Karan for their constant love and insurmountable faith in me, especially when I didn't have faith in myself. I'm glad I delivered.

If I could have it back

All the time that we wasted

I'd only waste it again

If I could have it back

You know I would love to waste it again

Waste it again and again and again

— "The Suburbs (Continued)" // Arcade Fire

This day is bananas,

B-A-N-A-N-A-S.

This day is bananas,

B-A-N-A-N-A-S!

— Kelly Kapoor // *The Office*

Chapter 1

INTRODUCTION

This dissertation consists of three projects. In the first project, we develop an efficient algorithm for learning the causal network connecting a large set of variables. We then consider incorporating known network structures in the second project, providing a new framework for high-dimensional generalized linear models given networks observed over both features and observation units. The final part of the dissertation develops a new method for off-policy evaluation in contextual bandit settings with an empashis on continuous action spaces. In this chapter, we describe the motivations and provide an overview of the methodological contributions for each project.

1.1 Learning causal networks from observational data

Directed acyclic graphs, or DAGs, are used to represent causal relationships among a set of random variables. They are commonly used as models of complex biological systems; for example, in gene regulatory networks, directed edges represent regulatory interactions among genes, which are represented as nodes of the graph. While causal effects in biological networks can be inferred from perturbation experiments (the gold standard for determining causality) [95, 91, 121], these are costly to run. Therefore, estimating DAGs from observational data is an important exploratory task for generating plausible causal hypotheses and designing more efficient experiments [33, 50].

Methods for estimating DAGs from observational data can be broadly categorized into three classes. The first class, score-based methods, search over the space of all possible graphs, and attempt to maximize a goodness-of-fit score, generally using a greedy algorithm. The second class, constraint-based methods, first estimate the graph skeleton (the undirected

graph obtained by removing the direction of edges) by performing conditional independence tests. Information from conditional independence relations is then used to partially orient the edges of the graph. The most well-known constraint-based method is the PC-Algorithm [98], which was popularized by Kalisch and Bühlmann [54] for high-dimensional estimation. Finally, hybrid methods combine score and constraint-based approaches.

The PC-Algorithm is considered a gold standard for DAG estimation in high-dimensional sparse settings, due to its polynomial time complexity [54]. However, the PC-Algorithm entails several properties that do not scale well to common settings. Specifically, large real-world biological systems are known to commonly be sparse graphs containing a small number of highly connected *hub* nodes [16, 52]. In such graphs, the average node degree will be small, while the maximum tends to be much larger, and increases with the number of nodes. This is particularly problematic for the PC-Algorithm, whose computational and sample complexities scale with the maximum node degree in the graph. Moreover, the recent work by Uhler et al. [111] shows that the distributional assumptions required for high-dimensional consistency of the PC-Algorithm are overly restrictive in practice, and that the class of graphs which do not satisfy these assumptions is large.

In Chapter 2, we address these issues by developing a modified version of the PC-Algorithm, which we refer to as *reduced PC* (rPC). rPC exploits the local separation property of large random networks, which implies that the number of short paths between any two nodes is bounded. This property is observed in many networks, including those which allow for hub node formation [25]. We show that rPC can consistently estimate the skeleton of high-dimensional DAGs by conditioning only on sets of small cardinality. We also show that computational and sample complexities of rPC do not scale with the maximal node degree, unlike with the PC-Algorithm. Moreover, rPC requires a weaker faithfulness conditions on the underlying probability distributions than the PC-Algorithm.

1.2 High-dimensional estimation and inference with two-way network structure

We now consider the setting with known networks, and develop a framework for incorporating structural information in high-dimensional regression models. Our method is motivated by datasets from metabolomics studies, where a common goal is to determine predictive biomarkers for diseases. In mass spectrometry metabolomics, data is often collected using both *targeted* and *untargeted* profiling. With targeted profiling, measurements are obtained on known annotated metabolites which cover selected biochemical pathways. On the other hand, untargeted profiling covers all measurable analytes in a sample, but the resulting data are mostly non-annotated *unnamed* features. For the purposes of biomarker discovery, these features are usually discarded, or only used in exploratory analyses. However, the unnamed features are often correlated with both the named metabolites and the disease outcome of interest. Therefore, using information from the unnamed features can improve the detection of named metabolites which are associated with the disease. We can consider that observation units to be similar if they have similar unnamed feature distributions. Along with the structure induced through named metabolites on known pathways, this is an example of *two-way structured data*, where both the features and observation units are connected over networks.

While many methods have been developed to incorporate network information [64, 43, 94], these usually focus solely on networks over *features*. There are also methods that account for unit network information, but these usually do not extend to the high dimensional setting [67], or to non-Gaussian generalized linear models [83]. Analyses involving two networks simultaneously have received less attention. In Chapter 3, we develop a penalized regression framework to analyze high-dimensional two-way network structured data using generalized linear models. For the feature network, we follow a common strategy in network-adjusted regression, and penalize the distance between regression coefficients which correspond to connected features. Based on a recent proposal by Li et al. [67], we simultaneously fit unit-

level intercept parameters, and enforce smoothness based on the unit network. Our method unifies the incorporation of feature and unit network information in the high-dimensional setting, and also provides a statistical inference procedure for testing the individual association between the outcome of interest and a feature in the model. We show empirically that our framework provides improved predictive accuracy and inferential power compared to existing high-dimensional methods. This holds given fully accurate network information, and also networks which are partially misspecified or uninformative. An application to a real-world metabolomics profiling study suggests that our model can improve classification of diseased subjects, and select relevant metabolites more often than standard methods.

1.3 Balanced off-policy evaluation for contextual bandits

In contextual bandit problems, a state s is observed, for which a *policy* $\pi : s \to a$ produces an action a. Then, a reward r is observed, which is a function of the state and action. The goal of contextual bandits is to develop a policy π that produces actions which maximize the reward. Many applications can be found, such as in medicine, where personalized treatments are designed based on known patient history [107], and internet marketing, where advertisements can be tailored to user interests [65]. In internet applications, learning an optimal policy is typically done online with streaming data, so the policy is being constantly updated. However, this may be prohibitively expensive. Moreover, experimenting with an untested policy could result in unacceptably negative results, such as patient death in the medical setting. Therefore, an important problem in this area is *counterfactual* or *off-policy* policy evaluation, where the average reward that would result from a policy of interest is estimated based on observed data under a different policy. This problem is even more important when attempting to safely deploy a policy for an application that previously used ad-hoc or difficult-to-enumerate rules.

Typical approaches to off-policy evaluation either use a regression model to predict the counterfactual rewards, importance sampling to reweight the observed reward data, or a combination of the two [24, 108, 118]. Importance sampling methods are useful for pro-

ducing unbiased reward estimates, as they correct for the difference in the observed and counterfactual policy distributions. However, most modern methods assume the importance sampling weights, which are ratios of policy densities $\pi(a|s)$, are known exactly [49]. This is usually not the case in practice, and a density estimate is used instead. This is particularly problematic in continuous action spaces, where estimating a continuous conditional density is a difficult problem.

In Chapter 4, we develop a new off-policy evaluation method for contextual bandit problems with arbitrary action spaces. Our proposed method does not require true knowledge or an estimator of either policy density, and can be directly plugged into existing methods instead of inverse propensity scores [56, 118, 24, 28]. A probabilistic classifier is trained on state-action data from both policies, and is used to directly estimate the density ratio. Hence, the method only requires logged data on states, and the actions which would be taken by both observed and target policies at those states. We show that the loss of the classification problem bounds the bias and variance, which essentially reduces the off-policy evaluation problem into a simpler binary classification problem. We demonstrate the efficacy of our method empirically on benchmark datasets, and show that it produces state-of-the-art results in continuous action spaces.

Chapter 2

THE REDUCED PC-ALGORITHM: IMPROVED CAUSAL STRUCTURE LEARNING IN LARGE RANDOM NETWORKS

2.1 Introduction

Directed acyclic graphs, or DAGs, are commonly used to represent causal relationships in complex biological systems. For example, in gene regulatory networks, directed edges represent regulatory interactions among genes, which are represented as nodes of the graph. While causal effects in biological networks can be accurately inferred from perturbation experiments [95]—including single or double gene knockouts [91, 121]—these are costly to run. Estimating DAGs from observational data is thus an important exploratory task for generating causal hypotheses [33, 50], and designing more efficient experiments.

Since the number of possible DAGs grows super-exponentially in the number of nodes, estimation of directed acyclic graphs is an NP-hard problem [18]. Methods of estimating DAGs from observational data can be broadly categorized into three classes. The first class, score-based methods, search over the space of all possible graphs, and attempt to maximize a goodness-of-fit score, generally using a greedy algorithm. Examples include the hill climbing and tabu search algorithms [93], as well as Bayesian approaches [26]. The second class, constraint-based methods, first estimate the graph skeleton by performing conditional independence tests; the skeleton of a directed acyclic graph is the undirected graph obtained by removing the direction of edges. Information from conditional independence relations is then used to partially orient the edges of the graph. The resulting completed partially directed acyclic graph (CPDAG) represents the class of all directed acyclic graphs that are Markov equivalent, and therefore not distinguishable from observational data. The most well-known constraint-based method is the PC-Algorithm [98], which was popularized by Kalisch

and Bühlmann [54] for high-dimensional estimation. Finally, hybrid methods combine score and constraint-based approaches. For example, the Max-Min Hill Climbing algorithm [110] estimates the skeleton using a constraint-based method, and then orients the edges by using a greedy search algorithm. Sparsity-inducing regularization approaches have also been used to develop efficient hybrid methods [92].

Estimating DAGs in high dimensions introduces new computational and statistical challenges. Until recently, graph recovery in high dimensions was only established for the PC-Algorithm [54] and hybrid constraint-based methods [36]. While the recent work of Nandy et al. [76] extends these results to score-based algorithms and their hybrid extensions, the PC-Algorithm is still considered a gold standard in high-dimensional sparse settings, due to its polynomial time complexity [54]. Moreover, constraint-based methods are indeed the building blocks of various hybrid approaches. Therefore, we primarily focus on constraint-based methods in this chapter.

Despite its appealing features, the PC-Algorithm entails several properties that do not scale well to common high-dimensional settings. Specifically, large real-world biological systems are known to commonly be sparse graphs containing a small number of highly connected *hub* nodes [16, 52]. In such graphs, the average node degree will be small, while the maximum tends to be much larger, and increases with the number of nodes. This is particularly problematic for the PC-Algorithm, whose computational and sample complexities scale with the *maximum node degree* in the graph. Moreover, the recent work by Uhler et al. [111] shows that the distributional assumptions required for high-dimensional consistency of the PC-Algorithm are overly restrictive in practice, and that the class of graphs which do not satisfy these assumptions is large. Although work has been done by Peters et al. [80] on estimating DAGs defined over a larger class of probability models, the resulting methods also do not scale to high dimensions.

A common limitation of existing methods for estimating DAGs is that they do not account for structural properties of large networks. For instance, the PC-Algorithm only incorporates the sparsity of the network, by assuming that the maximum node degree in the graph skeleton

is small relative to the sample size. However, real-world networks, particularly those observed in biology, are known to posses a number of other important properties. Of particular interest in estimating DAGs is the so-called *local separation property* of large networks [3], which implies that the number of short paths between any two nodes is bounded. This property is observed in many large sparse networks, including polytrees, Erdős-Rényi, power law, and small world graphs [25]. Power law graphs are of particular interest in many biological applications, as they allow for the presence of hub nodes. In this chapter, we propose a low-complexity constraint-based method for estimating high-dimensional sparse DAGs. The new method, termed *reduced PC* (rPC), exploits the local separation property of large random networks, which was used by Anandkumar et al. [3] in estimation of undirected graphical models. We show that rPC can consistently estimate the skeleton of high-dimensional DAGs by conditioning only on sets of small cardinality. This is in contrast to previous heuristic DAG learning approaches that set an upper bound on the number of parents of each node [1, 45], which is an assumption that cannot be justified in many real-world networks. We also show that computational and sample complexities of rPC only depend on average sparsity of the graph—a notion that is made more precise in Sections 2.3 and 2.4. This leads to considerable advantages over the PC-Algorithm, whose computational and sample complexities scale with the maximal node degree. Moreover, these properties hold for linear structural equation models [96] with arbitrary noise distributions, and require a weaker faithfulness conditions on the underlying probability distributions than the PC-Algorithm. We present two version of the rPC algorithm: a "full" version for which we provide theoretical guarantees, and an approximate version which is much faster and performs almost identically in practice.

The rest of the chapter is organized as follows. In Section 2.2 we review basic properties of graphical models over DAGs, and give a short overview of the PC-Algorithm. Our new algorithm is presented in Section 2.3 and its properties, including consistency in high dimensions are discussed in Section 2.4. Results of simulation studies and a real data example concerning the estimation of gene regulatory networks are presented in Sections 2.5 and 2.6, respectively. We provide a brief discussion in Section 2.7, and end with technical proofs and

additional simulations presented afterwards.

2.2 Preliminaries

In this section, we review relevant properties of graphical models defined over DAGs, and briefly describe the theory and implementation of the PC-Algorithm.

2.2.1 Background

For p random variables $X_1, \ldots, X_p$, we define a graph $G = (V, E)$ with vertices, or nodes, $V = \{1, \ldots, p\}$ such that variable X_j corresponds to node j. The edge set $E \subset V \times V$ contains directed edges; that is, $(j, k) \in E$ implies $(k, j) \notin E$. Furthermore, there are no directed cycles in G. We denote an edge from j to k as $j \to k$ and call j a parent of k and k a child of j. The set of parents of node k is denoted $pa(k)$, while the set of nodes adjacent to it, or all of k's parents and children, is denoted $adj(k)$. These notations are also used for the corresponding random variable X_k. We assume there are no hidden common parents of node pairs (that is, no unmeasured confounders). The degree of node k is defined as the number of nodes which are adjacent to it, $|adj(k)|$; we denote the maximal degree in the graph as d_{max}. A triplet of nodes (i, j, k) is called an *unshielded triple* if i and j are adjacent to k but i and j are not adjacent. An unshielded triple is called a *v-structure* if $i \to j \leftarrow k$.

We assume random variables follow a linear structural equation model (SEM),

$$X_k = \sum_{j \in pa(k)} \rho_{jk} X_j + \epsilon_k, \tag{2.1}$$

where for $k = 1, \ldots, p$, ϵ_k are independent random variables with finite variance, and ρ_{jk} are fixed unknown constants. The directed Markov property, stated below, is usually assumed in order to connect the joint probability distribution of $X_1, \ldots, X_p$ to the structure of the graph G.

Definition 1. *A probability distribution is Markov on a DAG $G = (V, E)$ if every random variable X_k is independent of its non-descendants conditional on its parents; that is, $X_k \perp\!\!\!\perp X_j \mid pa(X_k)$ for all $j \in V$ which are non-descendants of k.*

Although this assumption allows us to connect conditional independence relationships to the DAG structure, there are generally multiple graphs that generate the same distribution under the Markov property. More concretely, DAGs are Markov equivalent if they have the same skeleton and the same set of v-structures. Therefore, constraint-based methods focus primarily on estimating the skeleton of the DAGs from observational data. Conditional independence relations identified when learning the skeleton are then used to orient some of the edges to obtain the CPDAG, which represents the Markov equivalence class of directed graphs [54].

We next define d-separation, a graphical property which is used to read conditional independence relationships from the DAG structure.

Definition 2. *In a DAG G, two nodes k_1 and k_2 are d-separated by a set S if and only if, for all paths π between k_1 and k_2:*
(i) π contains a chain $i \to m \to j$ or a fork $i \leftarrow m \to j$ such that the middle node m is in S, or
(ii) π contains an inverted fork (or collider) $i \to m \leftarrow j$ such that the middle node m is not in S and no descendant of m is in S.

A path π which does not satisfy the requirements in this definition is known as a *d-connecting path*. Using observed data, d-separations in a graph G can be identified based on conditional independence relationships. To this end, we require the following assumption, known as faithfulness, on the probability distribution of random variable on G.

Definition 3. *A probability distribution is faithful to a DAG G if $X_i \perp\!\!\!\perp X_j \mid X_S$ whenever i and j are d-separated by S.*

2.2.2 The PC-Algorithm

Together, d-separation and faithfulness suggest a simple algorithm for recovering the DAG skeleton. If we discover that $X_i \perp\!\!\!\perp X_j \mid S$ for some set S, then there cannot be an edge $(i, j) \in E$. Conversely, if we discover $X_i \not\perp\!\!\!\perp X_j \mid S$ for all possible sets S, then there must be

an edge $(i, j) \in E$. Therefore, under faithfulness, an obvious strategy for skeleton estimation would be to test all possible conditional independence relations for each pair of variables; that is, test whether $X_i \perp\!\!\!\perp X_j \mid S$ for any $S \subset V \setminus \{i, j\}$. While this strategy is computationally infeasible for large p, and statistically problematic when $p > n$, it forms the basis of the PC-Algorithm. The PC-Algorithm starts with a complete undirected graph and deletes edges (i, j) if a set S can be found such that $X_i \perp\!\!\!\perp X_j \mid S$. The algorithm also uses the fact that if such an S exists, then there exists a set S' such that all nodes in S' are adjacent to i or j and $X_i \perp\!\!\!\perp X_j \mid S'$. Thus, at each step of the algorithm, only local neighbourhoods need to be examined in order to find the separating sets.

Although consistent for sufficiently sparse high-dimensional DAGs, the PC-Algorithm's computational and sample complexity scale with the maximal degree of the graph, d_{max}. Specifically, the algorithm's computational complexity is $O\big(p^{d_{max}}\big)$ and its sample complexity is $\Omega\big\{\max(\log p, d_{max}^{1/b})\big\}$ for some $b \in (0, 1]$. This is problematic for graphs with highly connected hub nodes, which are common in real-world networks [16, 52]. In such graphs, d_{max} typically grows with the number of nodes p, leading to poor accuracy and runtime for the PC-Algorithm.

Another limitation of the PC-Algorithm is that it requires partial correlations between adjacent nodes to be bounded away from 0; this requirement, which needs to hold for all conditioning sets S such that $|S| \leq d_{max}$, is known as restricted strong faithfulness [111], and is defined next.

Definition 4. *Given $\lambda \in (0, 1)$, a distribution P is said to be restricted λ-strong-faithful to a DAG $G = (V, E)$ if the following conditions are satisfied:*
(i) $\min\{|\rho(X_i, X_j \mid X_S)| : (i, j) \in E, S \subset V \setminus \{i, j\}, |S| \leq d_{max}\} > \lambda$, *and*
(ii) $\min\{|\rho(X_i, X_j \mid X_S)| : (i, j, S) \in N_G\} > \lambda$, *where N_G is the set of triples (i, j, S) such that i, j are not adjacent, but there exists $k \in V$ making (i, j, k) an unshielded triple, and i, j are not d-separated given S.*

Kalisch and Bühlmann [54] assume that $\lambda = \Omega(n^{-w})$ for $w \in (0, b/2)$ where $b \in (0, 1]$

relates to the scaling of d_{max}. In the low-dimensional setting, it has been shown that the PC-Algorithm achieves uniform consistency with λ converging to zero at rate $n^{1/2}$, which then gives the same condition as ordinary faithfulness [123]. However, in the high-dimensional setting, the set of distributions which are not restricted strong faithful has nonzero measure. In fact, Uhler et al. [111] showed that this assumption is overly restrictive and that the measure of unfaithful distributions converges to 1 exponentially in p. We will revisit the faithfulness assumption in Section 2.4.2.

To address the limitations of the PC-Algorithm, we next propose a new algorithm that takes advantage of the structure of large networks from common random graph families. By doing so, we obtain improved computational and sample complexity; as we will show, these complexities are unaffected by the increase in the maximal degree as the graph becomes larger. We also prove consistency under a weaker faithfulness assumption than that needed for the PC-Algorithm.

2.3 The Reduced PC-Algorithm

As with the PC-Algorithm, our strategy for estimating the graph skeleton is to start with a complete graph, and then delete edges by discovering separating sets. Under a faithfulness assumption on a linear structural equation model, we do so by computing partial correlations and declaring X_i and X_j d-separated by S if $\rho(X_i, X_j \mid S)$ is smaller than some threshold α. Aside from thresholding, the key difference between our proposal and the PC-Algorithm is that we only consider partial correlations conditional on sets S with small cardinality.

We justify our method using two key observations. Our first key observation is based on the decomposition of covariances over *treks*, which are special types of paths in directed graphs.

Definition 5. *A trek between two nodes i and j in a DAG G is either a path from i to j, a path from j to i, or a pair of paths from a third node k to i and j such that the two paths only have k in common.*

In a linear SEM (2.1), the covariance between two random variables is characterized by the treks between them. Denoting a trek from node i to node j as $\pi : i \leftrightarrow j$ with common node or source k, the covariance is given by

$$\text{cov}(X_i, X_j) = \sum_{\pi: i \leftrightarrow j} \sigma_k \prod_{e \in \pi} \rho_e, \tag{2.2}$$

where σ_k is the variance of ϵ_k from Equation (2.1) and ρ_e denotes the weight of an edge along the trek, which is the corresponding coefficient in the SEM. This is shown in full detail by Sullivant et al. [102], who also show that the covariance conditional on a d-separating set S leaves out treks which include any nodes in S. This conditioning effect is illustrated in Figure 2.1, where conditioning on an appropriate set S blocks the treks between non-adjacent nodes, without resulting in any additional d-connecting paths. If all edge weights are bounded by 1 in absolute value, then the contribution of each trek to the covariance decays exponentially in trek length. In practice, we scale the data matrix X so that each column has unit standard deviation. Under some conditions, most of the edge weights in the linear SEM (2.1) then satisfy $|\rho_{ij}| < 1$ (this is discussed and shown empirically in Section 2.9). Hence, the contribution of long treks to the conditional covariance among non-adjacent nodes is negligible. This decay motivates the thresholding of partial correlations in rPC, and is further discussed in Section 2.4.

The above observation suggests a new strategy for learning DAG structures by only considering short treks. Suppose S is the set that blocks all short treks between two nodes j and k. If the correlation over all remaining d-connecting paths (after conditioning on S) between j and k is negligible, then the partial correlation given S, $\rho(X_j, X_k \mid S)$ can be used to determine whether j and k are adjacent.

To determine the size of the conditioning set, S, we need to determine the number of short treks between any two nodes j and k. Our second key observation addresses this question, by utilizing properties of large random graphs. More specifically, motivated by Anandkumar et al.'s proposal for estimating undirected graphs, we consider a key feature of large random networks, known as the the *local separation property*.

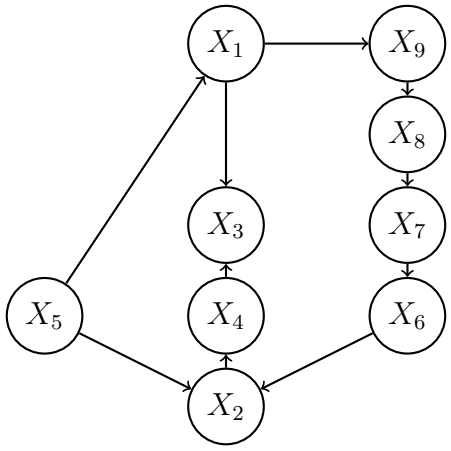

Figure 2.1: Illustration of treks between nodes X_1 and X_2 within a DAG. The middle path involves a collider, X_3, so does not contribute to $\text{cov}(X_1, X_2)$, which is $\rho_{51}\rho_{52}+\rho_{19}\rho_{98}\rho_{87}\rho_{76}\rho_{62}$ The conditional covariance $\text{cov}(X_1, X_2 \mid X_5)$ excludes treks that involve X_5, and is thus $\rho_{19}\rho_{98}\rho_{87}\rho_{76}\rho_{62}$. Here, $S = \{X_5, X_7\}$ is a d-separating set, as it blocks both treks, giving $\text{cov}(X_1, X_2 \mid S) = 0$.

Definition 6. *Given a graph G, a γ-local separator $S_\gamma(i, j) \subset V$ between non-neighbours i and j minimally separates i and j with respect to paths of length at most γ.*

Definition 7. *A family of graphs $\mathcal{G}$ satisfies the (η, γ)-local separation property if, as $p \to \infty$, $Pr(\exists G \in \mathcal{G} : \exists (i, j) \notin E_G, |S_\gamma(i, j)| > \eta) \to 0$.*

Intuitively, under (η, γ)-local separation, with high probability, the number of short treks—of length at most γ—between any two non-neighbouring nodes is bounded above by η. In fact, as the local separation property refers to any type of path, there are likely even fewer than η short treks between any two neighbouring nodes. Therefore, we only need to consider conditioning on sets S of size at most η in order to remove the correlation induced by short treks. Combining this with our first insight, we ignore treks (and other possible d-connecting paths) of length longer than γ, which, for appropriate probability distributions, have a negligible impact on partial correlations. The resulting procedure, called the full reduced PC-Algorithm (rPC-full), is presented in Algorithm 1.

We note a key implementation difference between rPC-full and the ordinary PC-Algorithm:

when searching for a separating set, rPC-full considers all $S \subset V \setminus \{i, j\}$, while PC-Algorithm only searches over the local neighbourhoods of i and j. Recall that if a set S d-separates nodes i and j, then there exists a d-separating set S' such that all nodes in S' are adjacent to i or j. However, $|S'|$ may be larger than $|S|$; because rPC only considers sets of size up to η, a full search is needed to ensure discovery of a d-separating set. We also propose an approximate reduced PC-Algorithm (rPC-approx), which uses the same local neighbourhood search

---

Algorithm 1: The full reduced PC-Algorithm (rPC-full)

INPUT: Observations from random variables $X_1, X_2, \ldots, X_p$; threshold level α; maximum separating set size η.

OUTPUT: Estimated skeleton C.

Set $V = \{1, \ldots, p\}$.

Form the complete undirected graph $\tilde{C}$ on the vertex set V.

Set $l = -1$; $C = \tilde{C}$.

repeat

$l = l + 1$

repeat

Select a (new) ordered pair of nodes i, j that are adjacent in C

repeat

Choose (new) $S \subset V \setminus \{i, j\}$ with $|S| = l$

if $\rho(X_i, X_j \mid S) \leq \alpha$

Delete edge (i, j)

Denote this new graph by C

end if

until edge (i, j) deleted or all $S \subset V \setminus \{i, j\}$ with $|S| = l$ have been chosen

until all ordered pairs of adjacent nodes i and j have been examined for $\rho(X_i, X_j \mid S) \leq \alpha$

until $l > \eta$

---

as the PC-Algorithm, i.e. $S \subset adj(i) \cup adj(j) \setminus \{i, j\}$. In practice, we show that rPC-approx performs almost identically to rPC-full.

To recap, our proposal in Algorithm 1 hinges on two important properties of probability models on large DAGs: (P1) boundedness of the number of short treks between any two nodes, and (P2) negligibility of correlation over the remaining (long) d-connecting paths. The first property, which is characterized by local separation, concerns solely the DAG structure. Anandkumar et al. [3] show that many common graph families satisfy the local separation property with small η. Specifically, sparse, large binary trees, Erdős-Rényi graphs, and graphs with power law degree distributions all satisfy this property with $\eta \leq 2$. Moreover, the sparsity requirement for these graph families is in terms of average node degree, and not the maximum node degree. For these graph families, our algorithm only needs to consider separating sets of size 0, 1, and 2, irrespective of the maximum node degree. Small-world graphs, as generated by the Watts-Strogatz algorithm [61], also satisfy this property, but with $\eta > 2$. In addition, the γ parameter increases with p for these families; thus, as graphs get larger, the local separation property applies to a larger set of paths.

By only considering a bounded number of short paths, our algorithm has computational complexity $O(p^{\eta+2})$, and thus avoids the exponential scaling in d_{max} that the PC-Algorithm suffers from. This is particularly significant in the case of power-law graphs, where $d_{max} = O(p^a)$ for $a > 0$ [75]; in this case, PC-Algorithm has a computational complexity of $O(p^{p^a})$, which is significantly worse than rPC's complexity of $O(p^4)$. While rPC-full might not be faster in practice due to the larger search space, rPC-approx would show a significant speed increase.

Unlike the first property (P1), the second property needed for our algorithm, namely the negligibility of correlation over d-connecting paths, concerns both the structure of the DAG G, and the probability distribution P of variables on the graph. In the next section, we discuss two alternative sufficient conditions that guarantee this property, and allow us to consistently estimate the DAG skeleton.

2.4 Algorithm analysis and asymptotics

In this section, we describe in detail the assumptions required for the rPC-full algorithm to consistently recover the DAG skeleton. We also discuss its computational and statistical properties, particularly in comparison with the PC-Algorithm.

2.4.1 Consistency

As discussed in the previous section, to consistently recover the DAG skeleton, rPC requires that properties (P1) and (P2) hold; namely, that the graph under consideration has a bounded number of short paths and the correlation over d-connecting paths of length greater than γ decays sufficiently quickly. In fact, the trek decomposition (2.2) indicates that the correlation for each long trek is small when most of the edge weights are bounded by 1 in absolute value. However, condition (P2) requires the total correlation over all long treks (and other d-connecting paths) to decay sufficiently quickly. To this end, we consider two alternative sufficient conditions. The first condition is a direct assumption on the boundedness of the conditional correlation. The second is inspired by Anandkumar et al. [3], and assumes the underlying probability model satisfies what we term *directed walk-summability*. This condition mirrors the walk-summability condition for undirected graphical models, which has been well-studied and shown to hold in a large class of models [73].

Definition 8. *A probability model is directed β-walk-summable on a DAG with weighted adjacency matrix A, if $\|A\| \leq \beta < 1$ where $\|\cdot\|$ denotes the spectral norm.*

As an alternative to this condition, we also present a direct assumption, Assumption 4. This condition is less restrictive than directed walk-summability, but has not been characterized in the literature. However, given that γ increases with p in the graph families we are considering, it is intuitive that the sum of edge weight products over treks longer than length γ will be decreasing and asymptotically small. This also holds for non-trek d-connecting paths. These two assumptions lead to two parallel proofs of the consistency of our algorithm, presented in Theorem 1. Before stating the theorem, we discuss our assumptions.

Similar to the PC-Algorithm, our method requires a faithfulness condition. As stated previously, our condition, which we term *path faithfulness* and is defined next, is weaker than PC-Algorithm's λ-strong faithfulness stated in Definition 4 (see Section 2.4.2 for additional details).

Definition 9. *Given $\lambda \in (0,1)$, a distribution P is said to be λ-path-faithful to a DAG $G = (V, E)$ if both of the following conditions hold:*
(i) $\min\{|\rho(X_i, X_j \mid X_S)| : (i,j) \in E, S \subset V \setminus \{i,j\}, |S| \leq \eta\} > \lambda$, *for some η, and*
(ii) $\min\{|\rho(X_i, X_j \mid X_S)| : (i,j,S) \in N_G\} > \lambda$, *where N_G is the set of triples (i,j,S) such that i, j are not adjacent, but there exists $k \in V$ making (i,j,k) an unshielded triple, and i, j are not d-separated given S.*

Part (i) of the assumption only requires partial correlations between true edges conditioned on sets of size up to η to be bounded away from zero, while the PC-Algorithm requires this for conditioning sets of size up to d_{max}. In Section 2.4.2, we discuss how this affects bounds on the true partial correlations, and also empirically show that the above path faithfulness assumption is less restrictive than corresponding assumption for the PC-Algorithm.

Assumption 1 (Path faithfulness and Markov property)**.** *The probability distribution P of random variables corresponds to a linear SEM* (2.1) *with sub-Gaussian errors, and is λ-path-faithful to the DAG G, with $\lambda = \Omega(n^{-c})$ for $c \in (0, 1/2)$.*

Our second assumption ensures that the covariance matrix of the structural equation model and its inverse remain bounded as p grows.

Assumption 2 (Covariance and precision matrix boundedness)**.** *The covariance matrix of the model Σ_G and its inverse Σ_G^{-1} are bounded in spectral norm, that is,* $\max(\|\Sigma_G\|, \|\Sigma_G^{-1}\|) \leq M < \infty$ *for all p.*

The last three assumptions characterize applicable graph families and probability distributions.

Assumption 3 ((η, γ)-local separation)**.** *The DAG G belongs to a family of random graphs $\mathcal{G}$ that satisfies the (η, γ)-local separation property with $\eta = O(1)$ and $\gamma = O(\log p)$.*

Assumption 4 (Bounded long path weight)**.** *For non-adjacent vertices i and j in G, let π denote a d-connecting path between them, and $l(\pi)$ be the length of the path. Then, there exists a conditioning set S such that the total edge weight over d-connecting paths longer than γ satisfies:*

$$\sum_{l=\gamma+1}^{p-1} \sum_{l(\pi)=l} |\rho_{\pi,1} \dots \rho_{\pi,l}| = O(\beta^{\gamma}),$$

for some $\beta \in (0, 1)$.

Assumption 4 guarantees that the sum of weights over long treks between any two nodes i and j is bounded. For a single trek, a sufficient condition is that all edge weights are bounded by 1 in magnitude. In Section 2.9, we provide further discussion on Assumption 4, and empirically investigate its plausibility. In particular, we show that if the data matrix X is scaled so that each column has unit standard deviation, then with high probability all edge weights are bounded by 1 in absolute value. To account for residual correlation induced through conditioning on common descendants of i and j, Assumption 4 also includes non-trek d-connecting paths.

Assumption 5 (Directed β-walk-summability)**.** *The probability distribution P is directed β-walk-summable.*

We are now ready to state our main result. The result is proved in Section 2.8, where the error probabilities are also analyzed.

Theorem 1. *Under Assumptions 1-3 and either Assumption 4 or 5, there exists a parameter α for thresholding partial correlations such that, as $n, p \longrightarrow \infty$ with $n = \Omega\{(\log p)^{1/(1-2c)}\}$, the full reduced PC (rPC-full) procedure, as described in Algorithm 1, consistently learns the skeleton of the DAG G.*

Several theoretical features of our algorithm are attractive. As stated previously, our faithfulness condition is weaker than the corresponding assumption for the PC-Algorithm and related methods. Similar to its computational complexity, the sample complexity of our algorithm also does not scale with the maximal node degree, and is only dependent on the parameter η as p increases. For example, in a power law graph, the sample complexity of the PC-Algorithm is $\Omega\{\max(\log p, p^{ab})\}$ for $0 < a, b < 1$, compared to $\Omega\{(\log p)^{1/1-2c}\}$ with $c \in (0, 1/2)$ for our method. This gain in efficiency is due to fact that the maximum separating set size, η, remains constant in rPC. Finally, our algorithm does not require the data to be jointly Gaussian. The proof of the algorithm's consistency only requires that the population covariance matrix can be well-approximated from the data; for simplicity, we assume a sub-Gaussian distribution.

2.4.2 On faithfulness assumption

As stated in Section 2.2.2, for large biological networks of interest, the maximum node degree, d_{max}, often grows with p. Therefore, the λ-restricted strong faithfulness condition of the PC-Algorithm—Definition 4—becomes exponentially harder to satisfy with increasing network size. A full discussion of this phenomenon can be found in Uhler et al. [111], where it is shown that the measure of strong unfaithful distributions converges to 1 for various graph structures. Although this would also occur with path faithfulness (Definition 9), our condition allows a rate for λ that is independent of d_{max} and p.

The rate for λ in the PC-Algorithm is $\lambda = \Omega(n^{-w})$ for $w \in (0, b/2)$, where $d_{max} = O(n^{1-b})$ for $b \in (0, 1]$. For $b = 1$, or constant d_{max}, the PC-Algorithm's required scaling for λ is identical to that for our method in Assumption 1. This makes sense intuitively, since our method is not affected by the increase in d_{max}. For other values of b, the scaling of λ becomes more restricted for the PC-Algorithm; for example, if $b = 1/2$, then $\lambda = \Omega(n^{-w})$ for $w \in (0, 1/4)$. However, under path faithfulness, we can still achieve a rate of $\lambda = \Omega(n^{-1/2})$; that is, the partial correlations are allowed to be smaller and the condition is weaker.

We report the findings of a simulation study, similar to that in Uhler et al. [111], which

Table 2.1: Empirical probabilities of random DAGs of size $p = 20$ satisfying faithfulness conditions; RSF refers to restricted strong faithfulness of the PC-Algorithm, and PF refers to path faithfulness of reduced PC (rPC).

Graph family	Expected degree	Pr(RSF)	Pr(PF)
Erdős-Rényi	2	0.77	0.92
Erdős-Rényi	5	0	0.08
Power law	2	0.54	0.85
Power law	6	0.003	0.08

examines how often randomly generated DAGs satisfy part (i) of the path faithfulness assumption compared to restricted strong faithfulness. We are primarily interested in part (i), as this part is needed for consistent skeleton estimation; part (ii), on the other hand, is needed to obtain correct separating sets in order to obtain partial orientation of edges. In this simulation, 1000 random DAGs were generated from Erdős-Rényi and power law families, with edge weights drawn independently from a Uniform$(-1, 1)$ distribution. Each DAG had $p = 20$ nodes, with varying expected degrees per node. For each simulation setting, we computed the proportion of DAGs that satisfied part (i) of the λ-restricted-strong-faithfulness and λ-path-faithfulness conditions with $\lambda = 0.001$ and $\eta = 2$. The results are shown in Table 2.1. We see that path faithfulness is much more likely to be satisfied than restricted strong faithfulness, especially for power law graphs. This is to be expected, as the number of constraints required for restricted strong faithfulness grows with d_{max}, but remains constant for path faithfulness. It is, however, difficult for dense graphs to satisfy either condition, although there is a mild advantage for path faithfulness. In Section 2.9, we provide the results of further simulation studies with $p = 10$ and $p = 30$; both of these give similar conclusions and indicate that path faithfulness remains easier to satisfy with increasing network size.

2.4.3 Tuning parameter selection

Our algorithm requires two tuning parameters: the maximum separating set size η, and the threshold level for partial correlations α. The parameter η varies based on the underlying graph family. Thus, given knowledge of a plausible graph structure, η can be pre-specified. Alternatively, η can be selected by maximizing a goodness-of-fit score over a parameter grid, along with α. This may be preferable as the local separation results consider all short paths, not just treks, so better performance may be obtained by specifying a smaller η. Likewise, when using the rPC-approx algorithm, a larger η could be needed to discover appropriate separating sets.

For jointly Gaussian data, we can obtain a modified version of the Bayesian information criterion by fitting the likelihood to the CPDAG obtained based on the estimated DAG skeleton [31]. Following Anandkumar et al. [3], and denoting by X_{obs} the observed data, we use:

$$\text{BIC}(X_{obs}; \hat{G}) = \log f(X_{obs}; \hat{\theta}) - 0.5|E| \log(n) - 2|E| \log(p), \tag{2.3}$$

where $\hat{G}$ denotes one of the DAGs obtained from the estimated CPDAG containing $|E|$ edges. The CPDAG represents the Markov equivalence class of DAGs, so all possible graphs will result in the same fitted Gaussian model with parameters $\hat{\theta}$. We use this BIC for tuning parameter selection; higher scores imply a better fit. For linear SEMs with non-Gaussian noise distributions, the Gaussian likelihood serves as a surrogate goodness-of-fit measure, and BIC can still be used to select the tuning parameters.

2.5 Simulation studies

In this section, we compare the performance of rPC-approx, rPC-full, and the standard PC-Algorithm in multiple simulation settings. We consider both setting a constant value for η in rPC, and tuning it to maximize the BIC.

2.5.1 Pre-specified η parameter

To facilitate the comparison with the PC-Algorithm, we generate data from Gaussian linear SEMs as in Equation 2.1, with the dependency structure specified by a DAG from Erdős-Rényi and power law families. We implement our algorithm with maximum separating set size $\eta = 2$ since these families are known to satisfy (η, γ)-local-separation with $\eta \leq 2$ [3].

We generate a random graph with p nodes using the `igraph` library in R, assigning every edge a weight from a Uniform$(0.1, 1)$ distribution. We then use the `rmvDAG` function from the `pcalg` library to simulate n observations from the DAG. This is repeated 20 times for each thresholding level α; average true positive and true negative rates for both algorithms are reported over the grid of α values. Our grid of α values produces partial receiver operating characteristic (pROC) curves for varying sample sizes and graph structures, which are used to assess the estimation accuracy of the two methods.

For both Erdős-Rényi and power law DAGs, we consider a low-dimensional setting with $p = 100$ nodes and $n = 200$ observations, and two high-dimensional settings with $(p, n) = (200, 100)$ and $(p, n) = (500, 200)$. In all settings, the DAGs are set to have an average degree of 2. The maximum degrees of the Erdős-Rényi graphs range from 5 to 7. The maximum degrees for power law graphs increase with p and are 42, 69, and 71 for the three simulation settings.

Results for Erdős-Rényi DAGs are shown in Figure 2.2. In this setting, all algorithms perform almost identically in both low and high-dimensional cases. This observation suggests that conditioning on larger separating sets by the PC-Algorithm is neither beneficial nor necessary. On the other hand, because of the relatively small maximum degree in Erdős-Rényi graphs, our algorithms do not lead to a considerable improvement over the PC-Algorithm.

Results for power law DAGs are shown in Figure 2.3. In this case, rPC's accuracy outperforms that of the PC-Algorithm, in both low and high-dimensional settings. These results confirms our theoretical findings, and show that our algorithm performs better at estimating DAGs with hub nodes than the PC-Algorithm.

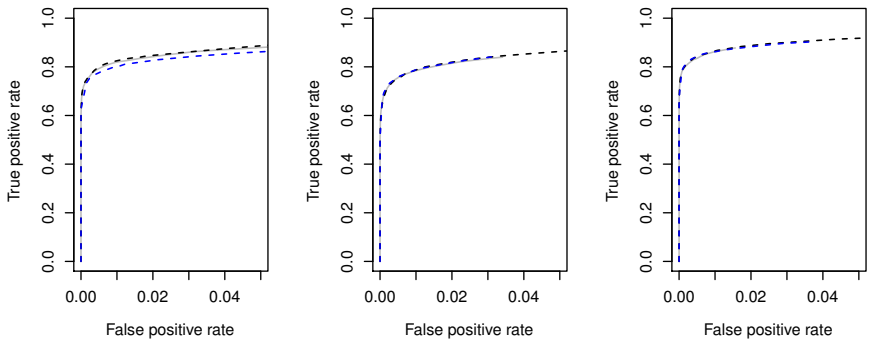

Figure 2.2: Average true vs. false positive rates for PC-Algorithm (grey solid line), rPC-approx (black dashed line), and rPC-full (blue dashed line) estimating Erdős-Rényi DAGs. Left: $p = 100, n = 200$; centre: $p = 200, n = 100$; right: $p = 500, n = 200$.

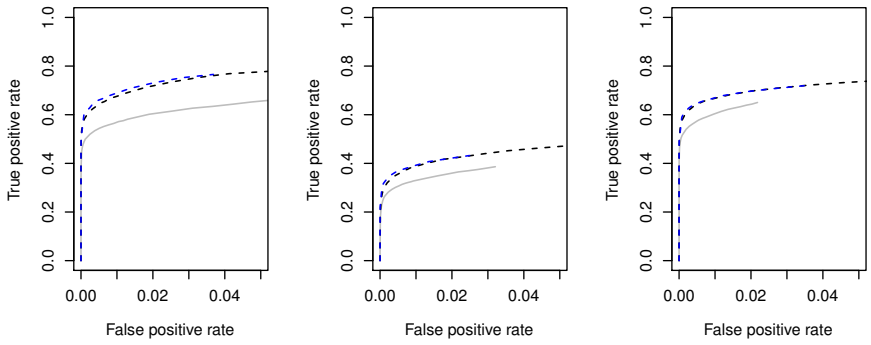

Figure 2.3: Average true vs. false positive rates for PC-Algorithm (grey solid line), rPC-approx (black dashed line), and rPC-full (blue dashed line) estimating power law DAGs. Left: $p = 100, n = 200$; centre: $p = 200, n = 100$; right: $p = 500, n = 200$.

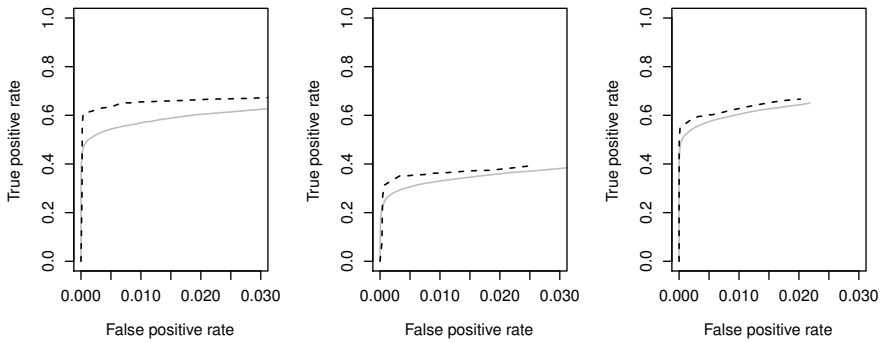

Figure 2.4: Average true vs. false positive rates for PC-Algorithm (grey solid line) and rPC-approx with BIC-tuned η (black dashed line) estimating power law DAGs. Left: $p = 100, n = 200$; centre: $p = 200, n = 100$; right: $p = 500, n = 200$.

Additional simulations in Section 2.9 show similar results in more dense DAGs. As the underlying DAG becomes more dense, both methods perform worse; however, our algorithms maintain an advantage over the PC-Algorithm in the power law setting. All of these simulations also confirm that rPC-full and rPC-approx provide very similar results. This suggests that the situation where conditioning on non-local sets is required occurs with very low probability, indicating that the approximate algorithm will provide suitably good estimation.

We also compare the runtimes of the algorithms for these settings. Over 100 iterations, we generate a random dataset, and apply both algorithms with a range of tuning parameters. Specifically, we set $\eta = 2$ for rPC, and $\alpha = \{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}\}$ for both. We then take the total runtime over all parameters, and compare by considering the mean value of $100\left(\frac{time_{rPC}}{time_{PC}}\right)$ for both rPC-approx and rPC-full. The results are shown in Table 2.2. We observe that rPC-approx is significantly faster than the PC-Algorithm for power law graphs. As expected, our implementation of rPC-full is slower than the PC-Algorithm in all settings, given its exhaustive search over all possible separating sets. Given the results indicating that rPC-approx performs almost identically to rPC-full, we suggest that practitioners use rPC-approx, possibly combined with BIC tuning if a suitable η bound is not known. We present relevant simulation results in the next section.

Table 2.2: Empirical ratio of rPC-approx and rPC-full runtimes to PC-Algorithm runtime.

Graph family	p	n	rPC-approx	rPC-full
Erdős-Rényi	100	200	0.98	3.08
Erdős-Rényi	200	100	1.00	6.16
Erdős-Rényi	500	200	1.00	23.25
Power law	100	200	0.60	1.13
Power law	200	100	0.92	3.81
Power law	500	200	0.36	5.23

2.5.2 BIC-tuned η Parameter

In this section, we consider simulations where the maximum size of the conditioning set for rPC-approx, η, is selected to maximize the BIC score defined in (2.3). To this end, we consider power law DAGs with the same low and high-dimensional settings as before. We select the value of $\eta \in \{1, 2, 3, 4\}$ which maximizes the BIC at each α value. While rPC-full could also be tuned in this way, it would be computationally expensive to consider η beyond 3 for $p > 200$. The results in Figure 2.4 show that our algorithm maintains an advantage over the PC-Algorithm. Interestingly, for values of α which yielded the best estimation accuracy, the optimal η selected was most frequently 1, which confirms our intuition from Section 2.3 that the η parameter for a graph family should be seen as an upper bound for estimating DAGs using our algorithm.

2.6 Application: estimation of gene regulatory networks

We apply our algorithms and the PC-Algorithm to a gene expression data set of $n = 487$ patients with prostate cancer from The Cancer Genome Atlas [12]. We select $p = 272$ genes with known network structure from BioGRID [99], and attempt to recover this network from the data. We choose the tuning parameters for rPC and the p-value threshold for the PC-

Algorithm by searching over a grid of values and selecting those which yielded the largest BIC (2.3). We found that the best rPC algorithm in terms of BIC was rPC-approx with $\eta = 3$ and $\alpha = 0.09$, while the best p-value threshold for the PC-Algorithm was 0.07.

The BioGRID database provides valuable information about known gene regulatory interactions. However, this databse mainly capture genetic interactions in normal cells. Thus, the information from BioGRID may not correctly capture interactions in cancerous cells, which are of interest in our application [47]. Despite this limitation, highly connected hub genes in the BioGRID network, which usually correspond to transcription factors, are expected to stay highly connected in cancer cells. Therefore, to evaluate the performance of the two methods, we focus here on the identification of hub genes, which are often most clinically relevant [42, 53, 14].

The two estimated networks and their hub genes are visualized in Figure 2.5. Here, we define hub genes as nodes with degree at least 8, which corresponds to the 75th percentile in the degree distribution of both estimates. rPC-approx identifies 19 of 57 true hubs, while the PC-Algorithm only identifies 6. Interestingly, several of the hub genes uniquely identified by rPC are known to be associated with prostate cancer, including ACP1, ARHGEF12, CDH1, EGFR, and PLXNB1 [90, 88, 84, 103, 68]. These results suggest that rPC may be a promising alternative for estimating biological networks, where highly-connected nodes are of clinical importance.

Examining the two networks also indicates that for nodes with small degrees, the estimated neighborhoods from rPC are very similar to those from the PC-Algorithm. To assess this observation, we consider the induced subgraph of nodes with degree at most 5 in the PC-Algorithm estimate. The F_1 score—which is a weighted average of precision and recall—between the two estimates of this sparse subnetwork is 0·86. This value indicates that the two algorithms perform very similarly over sparse nodes.

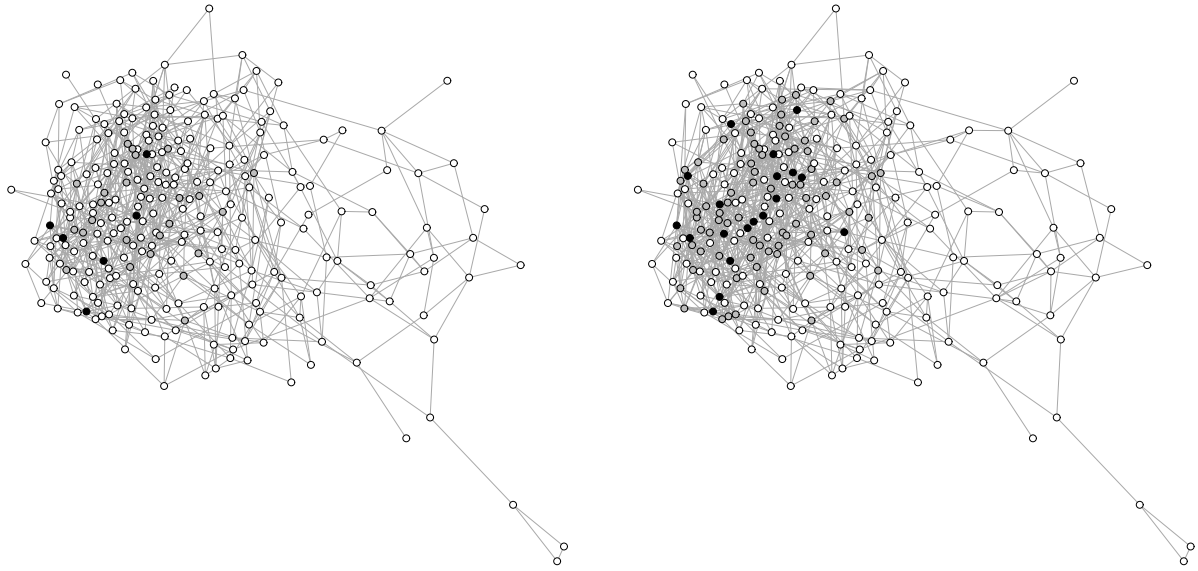

Figure 2.5: Estimated skeletons of gene regulatory networks in prostate cancer subjects. Black nodes are classified as hubs, having estimated degree of at least 8. Grey nodes are identified hubs that are also considered hubs in the BioGRID data. Left: PC-Algorithm; right: rPC-approx.

2.7 Discussion

Our new algorithm for learning directed acyclic graphs (DAGs) by conditioning on small sets leads to more efficient computation and estimation under a less restrictive faithfulness assumption than the PC-Algorithm. However, our weaker faithfulness condition may still not be satisfied for dense DAGs or in structural equation models with edge weights distributed over a larger parameter space. This is shown both geometrically and empirically in Uhler et al. [111], and remains a direction for future research. Generalizing the idea of restricted conditioning to more complex probability models over DAGs, such as nonlinear SEMs [115] would also be of interest. Finally, the idea of conditioning on small sets of variables can also be used to develop more efficient hybrid methods for learning DAGs in high dimensions.

2.8 Technical details

In this section, we prove the consistency of rPC-full for estimating the skeleton of a directed acyclic graph under the assumptions stated in Section 2.4.

We begin by establishing that correlations decay over long paths in the graph, and use this to show that the partial correlation between two non-adjacent nodes, conditional on a suitable set S with small cardinality, is bounded above. We establish this through two possible sufficient conditions: Lemma 1 is based on Assumption 4, which directly assumes that the the total weight over long paths is sufficiently small; on the other hand, Lemma 2 uses Assumption 5, which assumes the underlying model is directed walk-summable. Combining this result with Assumption 1, which says that the relevant partial correlations between two adjacent nodes is bounded below, we have oracle consistency of rPC-full. In Lemma 3, we invoke a concentration inequality for sample partial correlations to bound their deviations from population quantities. Using these results, we then prove that there exists a threshold level that consistently recovers the true skeleton in the finite sample setting.

In our first two lemmas, we make use of the local separation property in Assumption 3. While Assumption 3 does not directly concern *local separating sets*, since a d-separating set is a subset of a general separating set, Assumption 3 asymptotically guarantees the existence of a γ-*local d-separator* of size at most η for any two non-neighbouring nodes.

Definition 10. *Given a graph G, a γ-local d-separator $S_\gamma(i,j) \subset V \setminus \{i,j\}$ between non-neighbours i and j minimally d-separates i and j over paths of length at most γ.*

Lemma 1. *Under Assumptions 1-4, the partial correlation between non-neighbours i and j satisfies*

$$\min_{S \in S_{\eta,\gamma}} |\rho(i,j \mid S)| = O(\beta^\gamma).$$

where $S_{\eta,\gamma}$ is the set of γ-local d-separators of size at most η.

Proof. Recall the form of the linear structural equation model:

$$X_k = \sum_{j \in pa(k)} \rho_{jk} X_j + \epsilon_k,$$

where ϵ_k are independent and $Var(\epsilon_k) = \sigma_k^2 < \infty$ for all k.

Let A_G denote the lower-triangular weighted adjacency matrix for the graph G, obtained by ordering the nodes according to a causal order [96], so that $j \in pa(k)$ implies $j < k$. Then, as shown by Shojaie and Michailidis [96] for the Gaussian case and by Loh and Bühlmann [70] in general linear structural equation models,

$$\Sigma_G = (I - A_G)^{-1} D (I - A_G)^{-T},$$

where $D = \text{diag}(\sigma_1^2, ..., \sigma_p^2)$.

First, suppose that $\sigma_i^2 = 1$ for all i. We consider the conditional covariance $\Sigma_G(i, j \mid S)$ where i and j are non-neighbours, and S is the γ-local d-separator, as defined above. Let π denote a d-connecting path between i and j, $l(\pi)$ denote the length of the path, and $\rho_1, \ldots, \rho_l$ denote the edge weights along the path. Through conditioning on S, we have that covariance is only induced through d-connecting paths of length greater than γ. Then,

$$\Sigma_G(i, j \mid S) = \sum_{\substack{\pi: i \leftrightarrow j \\ \pi \cap S = \emptyset}} \sum_{\pi: l(\pi) = l} \prod_{k=1}^{l} \rho_k = \sum_{\substack{\pi: i \leftrightarrow j \\ l(\pi) = \gamma + 1}}^{p-1} \sum_{\pi: l(\pi) = l} \prod_{k=1}^{l} \rho_k.$$

Therefore:

$$\begin{aligned}
|\Sigma_G(i, j \mid S)| &= \left| \sum_{\substack{\pi: i \leftrightarrow j \\ l(\pi) = \gamma + 1}}^{p-1} \sum_{\pi: l(\pi) = l} \prod_{k=1}^{l} \rho_k \right| \\
&\leq \sum_{\substack{\pi: i \leftrightarrow j \\ l(\pi) = \gamma + 1}}^{p-1} \sum_{\pi: l(\pi) = l} \prod_{k=1}^{l} |\rho_k| \qquad \text{(by triangle inequality)} \\
&= O(\beta^\gamma). \qquad \text{(by Assumption 4)}
\end{aligned}$$

Now, suppose that not all $\sigma_i^2 = 1$. Then, let $\sigma_{max}^2 = \max_i \sigma_i^2$. We have:

$$\begin{aligned}
|\Sigma_G(i,j \mid S)| &\leq \sum_{\substack{\pi: i \leftrightarrow j \\ l(\pi)=\gamma+1}}^{p-1} \sum_{\pi: l(\pi)=l} \sigma_{max}^2 \prod_{k=1}^{l} |\rho_k| \\
&= \sigma_{max}^2 \sum_{\substack{\pi: i \leftrightarrow j \\ l(\pi)=\gamma+1}}^{p-1} \sum_{\pi: l(\pi)=l} \prod_{k=1}^{l} |\rho_k| \\
&= \sigma_{max}^2 O(\beta^\gamma) \\
&= O(\beta^\gamma). \qquad \text{(by Assumption 2)}
\end{aligned}$$

Finally, we have $|\rho(i,j \mid S)| = \dfrac{|\Sigma_G(i,j \mid S)|}{\sqrt{\Sigma_G(i,i \mid S)\Sigma_G(j,j \mid S)}} = O(\beta^\gamma)$ by Assumption 2, since the conditional variances are functions of the marginal variances, which are bounded.

□

Next, we show the same result by assuming directed walk-summability of the model.

Lemma 2. *Under Assumptions 1-3 and Assumption 5, the partial correlation between non-neighbours i and j satisfies*

$$\min_{S \in S_{\eta,\gamma}} |\rho(i,j \mid S)| = O(\beta^\gamma).$$

where $S_{\eta,\gamma}$ is the set of γ-local d-separators of size at most η.

Proof. Recall from the proof of Lemma 1,

$$\Sigma_G = (I - A_G)^{-1} D (I - A_G)^{-T},$$

where $D = \text{diag}(\sigma_1^2, ..., \sigma_p^2)$. First, suppose that $\sigma_i^2 = 1$ for all i. Then, we can write:

$$\begin{aligned}
\Sigma_G &= \left(\sum_{r=0}^{\infty} A_G^r \right) \left(\sum_{r=0}^{\infty} A_G^r \right)^T \\
&= \left(\sum_{r=0}^{\gamma} A_G^r + \sum_{r=\gamma+1}^{\infty} A_G^r \right) \left(\sum_{r=0}^{\gamma} A_G^r + \sum_{r=\gamma+1}^{\infty} A_G^r \right)^T.
\end{aligned}$$

Now, let Σ_H denote the covariance matrix induced by only considering d-connecting paths of length at most γ. For convenience, let $\Lambda_H := \sum_{r=0}^{\gamma} A_G^r$ and $R_\gamma := \sum_{r=\gamma+1}^{\infty} A_G^r$. Considering their spectral norms, denoted by $\|\cdot\|$, we have by walk-summability that $\|\Lambda_H\| \leq \dfrac{1-\beta^{\gamma+1}}{1-\beta}$ and $\|R_\gamma\| \leq \dfrac{\beta^{\gamma+1}}{1-\beta}$. Then,

$$\begin{aligned}
\Sigma_G &= (\Lambda_H + R_\gamma)(\Lambda_H + R_\gamma)^T \\
&= \Lambda_H \Lambda_H^T + \Lambda_H R_\gamma^T + R_\gamma \Lambda_H^T + R_\gamma R_\gamma^T \\
&= \Sigma_H + \Lambda_H R_\gamma^T + R_\gamma \Lambda_H^T + R_\gamma R_\gamma^T
\end{aligned}$$

Now, defining $E_\gamma := \Sigma_G - \Sigma_H$ and taking spectral norms, we get:

$$\begin{aligned}
\|E_\gamma\| = \|\Sigma_G - \Sigma_H\| &= \|\Lambda_H R_\gamma^T + R_\gamma \Lambda_H^T + R_\gamma R_\gamma^T\| \\
&\leq \|\Lambda_H R_\gamma^T\| + \|R_\gamma \Lambda_H^T\| + \|R_\gamma R_\gamma^T\| \\
&\leq 2\|\Lambda_H\|\|R_\gamma\| + \|R_\gamma\|^2 \\
&\leq 2\left(\frac{1-\beta^{\gamma+1}}{1-\beta}\right)\left(\frac{\beta^{\gamma+1}}{1-\beta}\right) + \left(\frac{\beta^{\gamma+1}}{1-\beta}\right)^2 \\
&= \frac{2\beta^{\gamma+1} - 2\beta^{2\gamma+2}}{(1-\beta)^2} + \frac{\beta^{2\gamma+2}}{(1-\beta)^2} \\
&= \frac{2\beta^{\gamma+1} - \beta^{2\gamma+2}}{(1-\beta)^2} \\
&= \frac{\beta^{\gamma+1}(2-\beta^{\gamma+1})}{(1-\beta)^2} \\
&= O(\beta^\gamma). \qquad \text{(S1)}
\end{aligned}$$

Now, suppose that not all $\sigma_i^2 = 1$. Then, following the same expansion of Σ_G as above, we have:

$$\|E_\gamma\| = \|D\|\frac{\beta^{\gamma+1}(2-\beta^{\gamma+1})}{(1-\beta)^2} = \sigma_{max}^2 \frac{\beta^{\gamma+1}(2-\beta^{\gamma+1})}{(1-\beta)^2} = O(\beta^\gamma), \qquad \text{(S2)}$$

by Assumption 2, where $\sigma_{max}^2 = \max_i \sigma_i^2$.

We now show that $|\rho(i,j \mid S)| = O(\|E_\gamma\|) = O(\beta^\gamma)$ where S is a γ-local d-separator between i and j. Let $A = \{i,j\} \cup S$ and $B = V \setminus A$. Consider the marginal precision matrix, $P := \{\Sigma_G(A,A)\}^{-1}$. Then, using the Schur complement, we can write this as

$$P = \Sigma_G^{-1}(A,A) - \Sigma_G^{-1}(A,B)\{\Sigma_G^{-1}(B,B)\}^{-1}\Sigma_G^{-1}(B,A).$$

Specifically, the partial correlation of X_i and X_j conditional on S is given by $\frac{P_{1,2}}{(P_{1,1}P_{2,2})^{1/2}} = O(P_{1,2})$, by Assumption 2.

Recall from (S1) that $\Sigma_G = \Sigma_H + E_\gamma$. Let F_γ be the matrix such that $\Sigma_G^{-1} = \Sigma_H^{-1} + F_\gamma$. Because Σ_H only considers covariance induced by paths of length at most γ, we have that $\Sigma_H^{-1}(A,A)_{1,2} = 0$.

Thus,

$$\begin{aligned}
|\{\Sigma_G(A,A)\}^{-1}_{1,2}| &= |\Sigma_G^{-1}(A,A)_{1,2} - \Sigma_G^{-1}(A,B)\{\Sigma_G^{-1}(B,B)\}^{-1}\Sigma_G^{-1}(B,A)_{1,2}| \\
&= |\Sigma_H^{-1}(A,A)_{1,2} + F_\gamma(A,A)_{1,2} - \Sigma_G^{-1}(A,B)\{\Sigma_G^{-1}(B,B)\}^{-1}\Sigma_G^{-1}(B,A)_{1,2}| \\
&= |F_\gamma(A,A)_{1,2} - \Sigma_G^{-1}(A,B)\{\Sigma_G^{-1}(B,B)\}^{-1}\Sigma_G^{-1}(B,A)_{1,2}| \\
&\leq \|F_\gamma(A,A) - \Sigma_G^{-1}(A,B)\{\Sigma_G^{-1}(B,B)\}^{-1}\Sigma_G^{-1}(B,A)\|_\infty \\
&\leq \|F_\gamma(A,A) - \Sigma_G^{-1}(A,B)\{\Sigma_G^{-1}(B,B)\}^{-1}\Sigma_G^{-1}(B,A)\|.
\end{aligned}$$

However, since $\Sigma_G^{-1}(A,B)\{\Sigma_G^{-1}(B,B)\}^{-1}\Sigma_G^{-1}(B,A)$ is positive semi-definite, $|\{\Sigma_G(A,A)\}^{-1}_{1,2}| \leq \|F_\gamma(A,A)\|$. We next show that $\|F_\gamma\| = O(\|E_\gamma\|) = O(\beta^\gamma)$. First, note that:

$$\begin{aligned}
F_\gamma &= \Sigma_G^{-1} - \Sigma_H^{-1} \\
&= (\Sigma_H + E_\gamma)^{-1} - \Sigma_H^{-1} \\
&= \Sigma_H^{-1} - \Sigma_H^{-1}(E_\gamma^{-1} + \Sigma_H^{-1})^{-1}\Sigma_H^{-1} - \Sigma_H^{-1} \qquad \text{(by Woodbury)} \\
&= -\Sigma_H^{-1}(E_\gamma^{-1} + \Sigma_H^{-1})^{-1}\Sigma_H^{-1}
\end{aligned}$$

Then, taking spectral norms, and noting that $\Sigma_G = \Sigma_H + E_\gamma$:

$$\begin{aligned}
\|F_\gamma\| &\leq \|\Sigma_H^{-1}\|\|(E_\gamma^{-1} + \Sigma_H^{-1})^{-1}\|\|\Sigma_H^{-1}\| && \text{(by sub-multiplicity)} \\
&= \|\Sigma_H^{-1}\|^2\|E_\gamma - E_\gamma(\Sigma_H + E_\gamma)^{-1}E_\gamma\| && \text{(by Woodbury)} \\
&= \|\Sigma_H^{-1}\|^2\|E_\gamma(I - \Sigma_G^{-1}E_\gamma)\| \\
&\leq \|\Sigma_H^{-1}\|^2\|E_\gamma\|\|I - \Sigma_G^{-1}E_\gamma\| && \text{(by sub-multiplicity)} \\
&\leq \|\Sigma_H^{-1}\|^2\|E_\gamma\|(1 + \|\Sigma_G^{-1}E_\gamma\|) && \text{(by triangle inequality)} \\
&\leq \|\Sigma_H^{-1}\|^2\|E_\gamma\|(1 + \|\Sigma_G^{-1}\|\|E_\gamma\|) && \text{(by sub-multiplicity)} \\
&\leq J\|E_\gamma\|^2, && \text{(by boundedness of } \|\Sigma_G^{-1}\| \geq |\Sigma_H^{-1}\|)
\end{aligned}$$

for some constant J. Then, by walk-summability, $\|F_\gamma\| = O(|E_\gamma\|)$. Hence, $|\rho(i, j \mid S)| = O(\beta^\gamma)$ by (S1).

□

By combining the result from either Lemma 1 or 2 with the λ-path-faithfulness assumption, we achieve oracle consistency for our algorithm given a threshold level α such that $\alpha = O(\lambda)$, $\alpha = \Omega(\beta^\gamma)$.

Next, we consider the finite sample setting, and establish a concentration inequality for sample partial correlations, under sub-Gaussian distributions, using a result from Ravikumar et al. [85].

Lemma 3. *Assume $X = (X_1, ..., X_p)$ is a zero-mean random vector with covariance matrix Σ such that each $X_i/\Sigma_{ii}^{1/2}$ is sub-Gaussian with parameter σ. Assume $\|\Sigma\|_\infty$ and σ are bounded. Then, the empirical partial correlation obtained from n samples satisfies, for some bounded $M > 0$:*

$$P\left(\max_{i \neq j, |S| \leq \eta} |\hat{\rho}(i, j \mid S) - \rho(i, j \mid S)| > \epsilon\right) \leq 4\left(3 + \frac{3}{2}\eta + \frac{1}{2}\eta^2\right)p^{\eta+2}\exp\left(-\frac{n\epsilon^2}{M}\right)$$

for all $\epsilon \leq \max_i(\Sigma_{ii})8(1 + 4\sigma^2)$.

Proof. Using the recursive formula for partial correlation, for any $k \in S$

$$\rho(i,j \mid S) = \frac{\rho(i,j \mid S \setminus k) - \rho(i,k \mid S \setminus k)\rho(k,j \mid S \setminus k)}{(1-\rho^2(i,k \mid S \setminus k))^{1/2}(1-\rho^2(k,j \mid S \setminus k))^{1/2}}.$$

For example, with $S = \{k\}$, we can simplify this to:

$$\rho(i,j \mid S) = \frac{\rho(i,j) - \rho(i,k)\rho(k,j)}{(1-\rho^2(i,k))^{1/2}(1-\rho^2(k,j))^{1/2}},$$

where $\rho(i,j) = \Sigma_{ij}/(\Sigma_{ii}\Sigma_{jj})^{1/2}$.

Rewriting in terms of elements of Σ, we then decompose the empirical partial correlation deviance from the true partial correlation into the deviances of covariance terms. Here, the event of the empirical partial correlation being within ϵ distance of the true partial correlation is contained in the union of the empirical covariance terms being within $C\epsilon$ distance of the true covariance terms for a sufficiently large $C > 0$:

$$\begin{aligned}\Big[|\hat{\rho}(i,j \mid S) - \rho(i,j \mid S)| > \epsilon\Big] \subset &\Big[|\hat{\Sigma}_{ij} - \Sigma_{ij}| > C\epsilon\Big] \bigcup \Big[|\hat{\Sigma}_{ii} - \Sigma_{ii}| > C\epsilon\Big] \bigcup \Big[|\hat{\Sigma}_{jj} - \Sigma_{jj}| > C\epsilon\Big] \\ &\bigcup_{k \in S} \Big[|\hat{\Sigma}_{ik} - \Sigma_{ik}| > C\epsilon\Big] \\ &\bigcup_{k \in S} \Big[|\hat{\Sigma}_{jk} - \Sigma_{jk}| > C\epsilon\Big] \\ &\bigcup_{k \leq k' \in S} \Big[|\hat{\Sigma}_{kk'} - \Sigma_{kk'}| > C\epsilon\Big],\end{aligned}$$

The number of events on the right-hand side is $3 + |S| + |S| + |S|^2 - \binom{|S|}{2}$. For $|S| \leq \eta$, the number of events is then bounded by $3 + \frac{3}{2}\eta + \frac{1}{2}\eta^2$. Then, by applying Lemma 1 in Ravikumar et al. [85], we have that, for any i, j:

$$Pr\Big(|\hat{\rho}(i,j \mid S) - \rho(i,j \mid S)| > \epsilon\Big) \leq 4\Big(3 + \frac{3}{2}\eta + \frac{1}{2}\eta^2\Big) \exp\Big(-\frac{n\epsilon^2}{K}\Big),$$

for some $K > 0$, bounded when $\|\Sigma\|_\infty$ and σ are bounded. From here, the result follows.

□

Combining the results established in Lemmas 1-3, we now prove the consistency of our algorithm in the finite sample setting.

Let α denote the threshold where if $\hat{\rho}_G(i, j \mid S) < \alpha$, the edge (i, j) is deleted. Let G_S denote the true undirected skeleton of G, and let $S_{\eta,\gamma}$ denote the set of γ-local d-separators or size at most η.

For any $(i, j) \notin G_S$, define the false positive event as

$$F_1(i, j) = \Big[\min_{S \in S_{\eta,\gamma}} |\hat{\rho}_G(i, j \mid S)| > \alpha \Big].$$

Define

$$\theta_{max} = \max_{(i,j) \notin G_S} \min_{S \in S_{\eta,\gamma}} |\rho_G(i, j \mid S)|$$

and

$$\hat{\theta}_{max} = \max_{(i,j) \notin G_S} \min_{S \in S_{\eta,\gamma}} |\hat{\rho}_G(i, j \mid S)|.$$

Consider

$$\begin{aligned} Pr\Big\{ \bigcup_{(i,j) \notin G_S} F_1(i, j) \Big\} &= Pr(\hat{\theta}_{max} > \alpha) \\ &= Pr(|\hat{\theta}_{max} - \theta_{max}| > |\alpha - \theta_{max}|) \\ &= O\Big[p^{\eta+2} \exp\Big\{ - \frac{n(\alpha - \theta_{max})^2}{M} \Big\} \Big] \qquad \text{(by Lemma 3)} \end{aligned}$$

where $\theta_{max} = O(\beta^\gamma)$ by Lemma 1 and 2.

For any true edge $(i, j) \in G_S$, define the false negative event as

$$F_2(i, j) = \Big[\min_{S \subset V \setminus \{i,j\}, |S| \leq \eta} |\hat{\rho}_G(i, j \mid S)| < \alpha \Big].$$

Define

$$\theta_{min} = \min_{(i,j) \in G_S} \min_{S \subset V \setminus \{i,j\}, |S| \leq \eta} |\rho_G(i, j \mid S)|$$

and

$$\hat{\theta}_{min} = \min_{(i,j) \in G_S} \min_{S \subset V \setminus \{i,j\}, |S| \leq \eta} |\hat{\rho}_G(i, j \mid S)|.$$

Consider

$$Pr\Big\{ \bigcup_{(i,j)\in G_S} F_2(i,j) \Big\} = Pr(\hat{\theta}_{min} < \alpha)$$
$$= Pr(|\theta_{min} - \hat{\theta}_{min}| > |\theta_{min} - \alpha|)$$
$$= O\Big[p^{\eta+2} \exp\Big\{ - \frac{n(\alpha - \theta_{min})^2}{K} \Big\} \Big] \qquad \text{(by Lemma 3)}$$

where $\theta_{min} = \Omega(\lambda)$ by restricted path-faithfulness assumption.

Under our assumptions, we have that $n = \Omega\{(\log p)^{1/1-2c}\}$, and $\lambda = \Omega(n^{-c})$ with $c \in (0, 1/2)$. Rewriting in terms of λ, we have $n = \Omega(\frac{\log p}{\lambda^2})$. Then, by selecting α such that $\alpha = O(\lambda)$, $\alpha = \Omega(\beta^\gamma)$, we have $Pr\{\bigcup_{(i,j)\notin G_S} F_1(i,j)\} = o(1)$ and $Pr\{\bigcup_{(i,j)\in G_S} F_2(i,j)\} = o(1)$. This completes the proof of Theorem 1.

2.9 Additional simulation results

In this section, we display some simulation results comparing our algorithms to the PC-Algorithm in estimating dense graphs. The simulation setup is otherwise identical to that described in Section 2.5. We consider a low-dimensional setting, with $p = 100$ and $n = 200$, as well as a high-dimensional setting, with $p = 200$ and $n = 100$. For Erdős-Rényi graphs, we use a constant edge probability of 0.05, corresponding to an expected degree of 5 for the low-dimensional graph and 10 for the high-dimensional graph. For power law graphs, we use an expected degree of 6 in both graphs.

While estimation quality is worse in the dense setting for all methods, we observe in Figure 2.6 and Figure 2.7 similar trends as in Section 2.5. Our methods perform as well as the PC-Algorithm in estimating Erdős-Rényi graphs, and shows some improvement for power law graphs. Both versions of rPC give similar results.

We also include more simulation results comparing path faithfulness to the restricted strong faithfulness assumption. As in Section 2.4.2, 1000 random DAGs were generated from Erdős-Rényi and power law families, with edge weights drawn independently from a

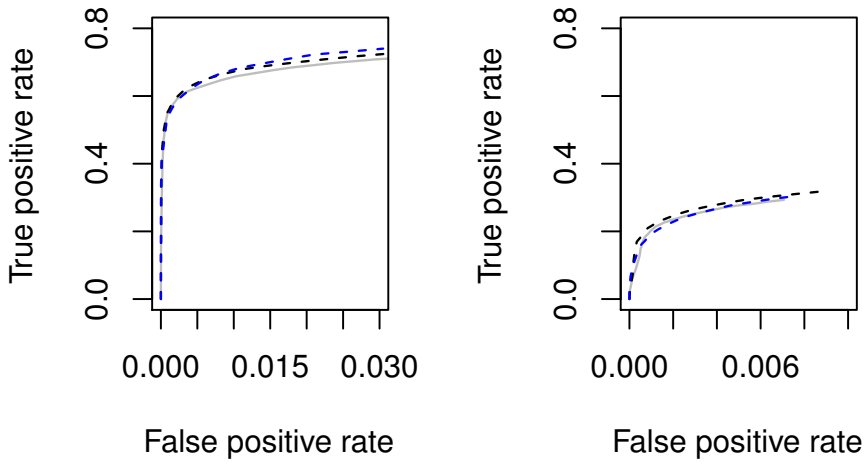

Figure 2.6: Average true vs. false positive rates for PC-Algorithm (grey solid line), rPC-approx (black dashed line), and rPC-full (blue dashed line) estimating Erdős-Rényi graphs. Left: $p = 100, n = 200$, average degree 5; right: $p = 200, n = 100$, average degree 10.

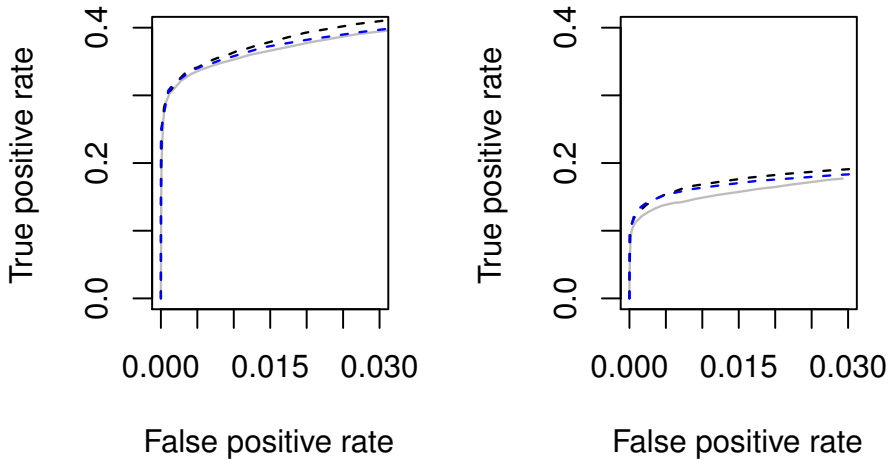

Figure 2.7: Average true vs. false positive rates for PC-Algorithm (grey solid line), rPC-approx (black dashed line), and rPC-full (blue dashed line) estimating power law graphs with average degree 6. Left: $p = 100, n = 200$; right: $p = 200, n = 100$.

Table 2.3: Empirical probabilities of random DAGs of size $p = 10$ satisfying faithfulness conditions; RSF refers to restricted strong faithfulness of the PC-Algorithm, and PF refers to path faithfulness of reduced PC (rPC).

Graph family	Expected degree	Pr(RSF)	Pr(PF)
Erdős-Rényi	2	0.94	0.98
Erdős-Rényi	5	0.10	0.61
Power law	2	0.95	0.97
Power law	6	0.08	0.60

Table 2.4: Empirical probabilities of random DAGs of size $p = 30$ satisfying faithfulness conditions; RSF refers to restricted strong faithfulness of the PC-Algorithm, and PF refers to path faithfulness of reduced PC (rPC).

Graph family	Expected degree	Pr(RSF)	Pr(PF)
Erdős-Rényi	2	0.66	0.86
Erdős-Rényi	5	0	0.01
Power law	2	0.04	0.51
Power law	6	0	0.01

Uniform$(-1, 1)$ distribution. Here, we consider graphs of size $p = 10$ and $p = 20$, with faithfulness condition parameters $\lambda = 0.001$ and $\eta = 2$. The results are shown in Table 2.3 and Table 2.4.

Although it is harder for both conditions to be met as graph size and density increase, we see that path faithfulness is still more likely to be satisfied than restricted strong faithfulness, particularly with power law graphs.

Finally, we consider the plausibility of Assumption 4 when the SEM coefficients are not bounded by 1 in absolute value. Suppose that the data matrix X has been normalized by column-wise standard deviations. We show that this transformation preserves the original network structure, and leads to most edge weights being bounded by 1 in absolute value.

Consider an edge $j \to k$ and let $W_j := \{X_i : i \in pa(k) \setminus \{j\}\}$. Then, taking the conditional covariance:

$$\begin{aligned} Cov(X_j, X_k|W_j) &= Cov\left(X_j, \rho_{jk}X_j + \sum_{i \in pa(k)\setminus j} \rho_{ik}X_i + \epsilon_k \middle| W_j\right) \\ &= \rho_{jk} Var(X_j|W_j). \end{aligned}$$

We can therefore write:

$$\rho_{jk} = \frac{Cov(X_j, X_k|W_j)}{Var(X_j|W_j)}.$$

Now, let $\tilde{X}_k = X_k/sd(X_k)$ for all k. Consider the edge weights $\tilde{\rho}_{jk}$ corresponding to the SEM for this transformed data. We have:

$$\tilde{\rho}_{jk} = \frac{sd(X_j)}{sd(X_k)} \frac{Cov(X_j, X_k|W_j)}{Var(X_j|W_j)} = \frac{sd(X_j)}{sd(X_k)}\rho_{jk}$$

Clearly, $\tilde{\rho}_{jk} = 0$ if and only if $\rho_{jk} = 0$. Therefore, we recover the same network by applying our algorithm to the transformed data. Furthermore,

$$\begin{aligned} |\tilde{\rho}_{jk}| &= \sqrt{\frac{Var(X_j)}{Var(X_k)}\rho_{jk}^2} \\ &= \sqrt{\frac{Var(X_j)\rho_{jk}^2}{Var(X_j)\rho_{jk}^2 + \sigma_k^2 + \sum_{i \in pa(k)\setminus\{j\}} \rho_{ik}^2 Var(X_i) + 2\sum_{i,i' \in pa(k)} \rho_{ik}\rho_{i'k} Cov(X_i, X_i')}} \end{aligned}$$

Thus, $|\tilde{\rho}_{jk}| > 1$ only if

$$\sigma_k^2 + \sum_{i \in pa(k)\setminus\{j\}} \rho_{ik}^2 Var(X_i) + 2 \sum_{i,i' \in pa(k)} \rho_{ik}\rho_{i'k} Cov(X_i, X_i') < 0$$

Rewriting the variance and covariance terms as trek sums, we obtain:

$$\sigma_k^2 + \sum_{i \in pa(k)\setminus\{j\}} \rho_{ik}^2 \sum_{\substack{\pi : u \leftrightarrow i \\ u \in V\setminus\{i\}}} \sigma_w^2 \prod_{\rho_e \in \pi} \rho_e^2 + 2 \sum_{i,i' \in pa(k)} \rho_{ik}\rho_{i'k} \sum_{\pi : i \leftrightarrow i'} \sigma_w \prod_{\rho_e \in \pi} \rho_e < 0$$

where w denotes the source or common node in the trek.

Since the first two terms in this expression will always be positive, $|\tilde{\rho}_{jk}| > 1$ only if the covariance sum is both negative and greater in magnitude than the sum of the first two

terms. A term in the covariance sum will be negative if there is an odd number of negative terms in the product, and the entire sum will be negative only if there are a sufficient number of these to make the whole sum negative. Even then, the sum over squared trek products would need to be smaller than the covariance term sum in absolute value. Moreover, even if some $|\tilde{\rho}_{jk}| > 1$, the product over the entire trek may still be less than 1 if most of the other coefficients along a trek are small. Thus, we may still expect the product over the trek to be small.

Next, we provide simulation results showing that standardized SEM coefficients in graph configurations we consider rarely exceed 1 in absolute value. Specifically, we consider Erdős-Rényi and power law graphs, with $p \in \{200, 500\}$ having average degree 2. We consider edge weights from both a Unif$(-10, 10)$ and $N(0, 3^2)$ distribution. For each graph, we estimate the standardized coefficients using the `fitDag` function in the `ggm` library, and compute the proportion of edge weights that are greater than 1 in absolute value. We repeat this process over 1,000 iterations. The results are shown in Table 2.5. This experiment shows that, with high probability, the covariance induced over long treks between any two nodes is indeed small.

Table 2.5: Empirical average probabilities of standardized coefficients exceeding 1 in absolute value.

Graph family	p	Distribution	$Pr(\|\rho\| > 1)$
Erdős-Rényi	200	Uniform(-10, 10)	0.0036
	200	Normal(0, 3^2)	0.0017
	500	Uniform(-10, 10)	0.0036
	500	Normal(0, 3^2)	0.0012
Power law	200	Uniform(-10, 10)	0.0045
	200	Normal(0, 3^2)	0.0045
	500	Uniform(-10, 10)	0.0034
	500	Normal(0, 3^2)	0.0037

Chapter 3

FEATURE AND UNIT NETWORK KERNEL PENALIZATION FOR HIGH-DIMENSIONAL GENERALIZED LINEAR MODELS

3.1 Introduction

Metabolites provide information about responses of cells to their environment, and metabolic changes are often used as predictive biomarkers for diseases. In mass spectrometry metabolomics, data is usually collected using both targeted and untargeted profiling. With targeted profiling, measurements are obtained on known, *named* metabolites which cover selected biochemical pathways. In contrast, untargeted profiling covers all measurable analytes in a sample, but the resulting spectra are mostly non-annotated, *unnamed* features. For the purposes of biomarker discovery, these unnamed features are usually discarded, or only used in exploratory analyses. However, the unnamed features are often correlated with both the named metabolites and the disease outcome of interest. Therefore, using information from the unnamed features may improve the detection of associations between named metabolites and disease. This information can be incorporated as a distance between observation units with respect to their unnamed feature distributions. That is, we consider that observation units to be similar if they have similar unnamed feature distributions, and use this as the basis for borrowing strength across nearby units. Accounting for the additional structure induced through relating named metabolites on known metabolic pathways, this is an example of *two-way structured data*.

Allen et al. [2] defined two-way structures in the context of neuroimaging data, where features represent spatially-correlated observations in the brain, and observations are collected over distinct time points. Examples can also be found in the analysis of microbiome

data; Randolph et al. [83] considered data where features represent taxa having known relationships, and a phylogenetic distance is computed between observation units. Two-way structures also arise when integrating multi-view omics data [104]; in this case, data from one view can be used to define distances among observation units in order to improve prediction and inference with another view. These structures can be captured as separate *networks* over the features and units, where a connection between features or units represents some scientifically meaningful notion of similarity.

In studies which aim to identify associations with complex diseases, incorporating networks into statistical models can lead to improved predictive power and biomarker discovery [60]. However, while many methods have been developed to incorporate network information [64, 43, 94], these usually focus solely on networks over *features*. Analyses involving two networks has received less attention. In this chapter, motivated by metabolomics studies, we develop a penalized regression framework to analyze high-dimensional two-way network structured data using generalized linear models. For the feature network, we follow a common strategy in network-adjusted regression, and penalize the distance between regression coefficients which correspond to connected features. Based on a recent proposal by Li et al. [67], we simultaneously fit unit-level intercept parameters, and enforce smoothness based on the unit network. In addition, we provide an inference procedure for testing the individual association between the outcome of interest and a feature in the model, with the goal of discovering metabolites associated with the disease.

Our approach is unique among high-dimensional penalized regression methods, as network penalty terms are generally not (semi)norms. As a result, theoretical guarantees are difficult to obtain. In our theoretical work, we consider these non-norm penalty terms to be part of the target loss function, which is distinct from the loss function minimized by the true regression parameters. We control the distance between the target and "true" parameters, and then characterize the asymptotic behaviour of our proposed estimators. This approach allows us to incorporate partially uninformative or misspecified networks and still obtain valid inference for the association of features with the outcome. In simulation studies, we

show that if the networks are fully informative for the true parameters, then our method provides both improved prediction accuracy and inferential power. We also show that our method is robust to partially uninformative networks, and remains competitive with existing methods. An application to real metabolomics data shows that our method has lower estimated prediction error, and that the resulting inference is more likely to identify important features.

The rest of this chapter is organized as follows. In Section 3.2, we introduce relevant concepts for incorporating network information into regression models, and describe existing methods for doing so. We describe our penalized regression framework in Section 3.3, including details of optimization algorithms. In Section 3.4, we state our assumptions and investigate the large sample properties of our estimates. We then describe our inference procedure, and show its validity. We apply our method to simulated data in Section 3.5 and to a metabolomics profiling study in Section 3.6, and provide a discussion in Section 3.7. Technical details follow.

3.2 Background

We assume that the data consists of n observations $(y_1, x_1), \ldots (y_n, x_p)$ where $y_i \in \mathbb{R}$ is the outcome and $x_i \in \mathbb{R}^p$ is the corresponding feature vector. The conditional mean relationship between the outcomes and features is assumed to be

$$\mathbb{E}[Y_i | X_i = x_i] = \mu(\alpha_i + x_i'\beta),$$

where $\alpha_i \in \mathbb{R}$ is an intercept term for unit i, $\beta \in \mathbb{R}^p$ is a vector of common regression coefficients, and μ denotes the inverse link function for the generalized linear model (GLM), which we will also refer to as the mean function. Some examples of inverse canonical link functions used for generalized linear models are $\mu(x) = x$ for Gaussian models and $\mu(x) = \exp(x)(\exp(x) + 1)^{-1}$ for binomial models.

We assume the observation units $i = 1, \ldots, n$ are connected on some known graph $G_n = (V_n, E_n, W_n)$ where $V_n = \{1, \ldots, n\}$ and $E_n \subset V_n \times V_n$ is a set of undirected edges (i, i')

for $i \neq i'$. For the purposes of this chapter, a connection between two units i, i' implies that they are likely to have similar outcomes $y_i, y_{i'}$. W_n is a set of weights $w_{ii'} \in \mathbb{R}^+$ for each edge, which quantify the strength of the similarity. Let A_n be the weighted adjacency matrix of G_n. Then, $L_n = D_n - A_n$ defines the corresponding graph Laplacian [117], where $D_n = \text{diag}(d_1, \ldots, d_n)$ and $d_i = \sum_{j \in V} A_{ij}$. The Laplacian matrix will be used in the network penalties we consider. We further assume that the features $X_j, j = 1, \ldots, p$ are connected on a second known graph $G_p = (V_p, E_p, W_p)$, and define A_p, D_p, and L_p analogously as above. Here, a connection between features implies that they have similar association with the outcome y. We also define J_p as the graph incidence matrix, where each row of J_p corresponds to an edge (j, j') in the graph, with element j of the row having value $w_{jj'}$ and element j' having value $-w_{jj'}$. If all edge weights are equal to 1, then we have that $J_p' J_p = L_p$.

The most common methods for handling correlated outcome data are generalized estimating equations (GEE) and generalized linear mixed models (GLMMs). However, GEEs are generally used for clustered data, where outcomes are independent between clusters and the number of clusters is large. When observation units are observed in a network, these assumptions may not hold. GLMMs can account for more arbitrary structures, but require parametric assumptions and specifications for variance-covariance components [15]. They are also computationally difficult to fit, particularly in large n and large p settings. While generalized least squares (GLS) addresses these issues, it only works for correlated Gaussian outcomes, given a known covariance matrix among units.

The recent proposal of Li et al. [67] accounts for unit network structure by enforcing *cohesion* within the model. It introduces the regression with network cohesion (RNC) model, which estimates unit-level intercepts subject to a cohesion penalty over G_n, by solving the optimization problem,

$$\min_{\alpha \in \mathbb{R}^n, \beta \in \mathbb{R}^p} \left\{ \ell(\alpha + X\beta) + \frac{1}{2} \gamma_n \alpha' (L_n + \delta I_n) \alpha \right\},$$

where ℓ is a loss function (usually the negative log-likelihood), and $\gamma_n > 0$ tunes the strength

of the penalty. The dependence between observations is captured through the n-dimensional intercept term α. The addition of δI_n where $\delta > 0$ guarantees that a solution exists. This problem is easily solved with standard convex optimization algorithms. The penalty term can be written as:

$$\alpha'(L_n + \delta I_n)\alpha = \sum_{(i,j)\in E_n} w_{ij}(\alpha_i - \alpha_j)^2 + \delta \sum_{i\in V} \alpha_i^2.$$

Therefore, units which are more strongly connected are encouraged to have similar intercepts. This cohesion effect implies a similarity independent of the features X; connected units may still differ in their values of $x_i'\beta$. This is similar to incorporating variance components in a generalized linear mixed model; we discuss this equivalence further in Section 3.10. However, the RNC intercepts do not require distributional assumptions, and are specifically fit to optimize prediction power. Li et al. [67] show that the RNC model improves prediction for network-linked observations compared to standard methods, while maintaining the interpretability of the fixed feature effects in standard generalized linear models. However, high-dimensional settings and statistical inference are not considered.

A similar choice for incorporating *feature* network structure is the Grace (graph-constrained estimation) penalty of Li and Li [64], who proposed the estimator,

$$\hat{\beta} = \underset{\beta\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ \ell(X\beta) + \frac{1}{2}\gamma_p \beta' L_p \beta + \lambda\|\beta\|_1 \right\},$$

where ℓ is a loss function as before, and $\gamma_p, \lambda > 0$ are penalty parameters. Unlike the RNC estimator, the Grace penalty is well-defined for high-dimensional settings, due to the regularization applied to the features. This penalized regression encourages cohesion among β coefficients corresponding to connected features. The inclusion of an ℓ_1 penalty also enforces sparsity in the solution $\hat{\beta}$. Zhao and Shojaie [124] developed a significance test for Grace-penalized estimation, but their approach only applies to linear regression models.

Randolph et al. [83] account for two-way structured data in a *kernel-penalized* linear regression model by solving

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \|y - X\beta\|_H^2 + \lambda\|\beta\|_{Q^{-1}}^2 \right\},$$

where H and Q^{-1} are, respectively, $n \times n$ and $p \times p$ kernel matrices which summarize distances between the units and features. This can be thought of as fitting a generalized least squares model subject to a generalized ridge penalty. We can consider the proposal of Li and Li [64] to fall within this framework, using the graph Laplacian as a kernel. The RNC penalty of Li et al. [67] is also similar to the GLS term. Both methods penalize quantities that capture "left-over" variation from the features with respect to a unit distance matrix; for GLS, the distances between residuals $y - X\beta$ are penalized, while for RNC the intercepts α are. However, the GLS method does not easily extend to non-Gaussian models. Therefore, in this chapter, we apply the idea of kernel penalization to generalized linear models by unifying RNC and Grace-style penalties in high-dimensional settings. We also develop a statistical inference procedure, which allows for valid hypothesis tests and confidence intervals for the regression β coefficients.

3.3 GLMs with feature and unit network kernels

In this section, we describe our framework for generalized linear models with penalization to account for both unit and feature networks. To achieve cohesion with respect to kernels that summarize unit and feature similarity, we propose the following estimator:

$$(\hat{\alpha}, \hat{\beta}) = \underset{\alpha,\beta}{\operatorname{argmin}} \left\{ \ell(\alpha + X\beta) + \frac{1}{2}\gamma_n \alpha'(L_n + \delta I_n)\alpha + \gamma_p P(G_p, \beta) + \lambda \|\beta\|_1 \right\}, \tag{3.1}$$

where γ_n, γ_p, and λ are positive tuning parameters; $\delta > 0$ can also be tuned, but for the purpose of this chapter, we set it to be a small fixed value. Here, $P(G_p, \beta)$ denotes a smoothing penalty over the feature parameters β.

We consider two possible choices of feature smoothing penalty: an ℓ_2 cohesion penalty with a Laplacian kernel $P(G_p, \beta) = \frac{1}{2}\beta' L_p \beta$, and an ℓ_1 fusion penalty with an incidence matrix kernel $P(G_p, \beta) = \|J_p \beta\|_1$. The resulting optimization problems can be thought of, respectively, as generalized ridge and generalized lasso problems. With the generalized ridge solution, the weighted squared distance between connected features' parameters is penalized,

since

$$\beta' L_p \beta = \sum_{(i,j)\in E_p} w_{ij}(\beta_i - \beta_j)^2.$$

On the other hand,

$$\|J_p\beta\|_1 = \sum_{(i,j)\in E_p} w_{ij}|\beta_i - \beta_j|.$$

Using the ℓ_1 penalty results in a similar effect as the ordinary lasso, where the coefficients of graph-connected features will be encouraged to be exactly equal [117]. On the other hand, with the ℓ_2 penalty, the squared difference between the coefficients will be encouraged to be small, but the distance is not shrunk exactly to zero.

To summarize, there are four components in the optimization problem (3.1): (i) the loss function ℓ relates the outcome y to the features X while allowing for a unique intercept for each unit, (ii) the unit network smoothing penalty smooths $\hat{\alpha}$ over G_n, (iii) the feature network smoothing penalty smooths $\hat{\beta}$ over G_p, and (iv) the standard lasso penalty enforces sparsity on the features. The addition of the lasso penalty also allows us to obtain asymptotic consistency for $\hat{\beta}$ at a rate that enables valid inference in high dimensions. We name this framework "generalized linear models with feature and unit network kernels", or `glm-funk`.

3.3.1 Optimization for ℓ_2 feature network smoothing

To describe the algorithm for solving the `glm-funk` problem with ℓ_2 feature network smoothing, we first rewrite the objective function in terms of $\theta := (\alpha, \beta)'$:

$$\hat{\theta} = \underset{\theta\in\mathbb{R}^{n+p}}{\operatorname{argmin}} \left\{ \ell\left(\tilde{X}\theta\right) + \frac{1}{2}\theta'\tilde{L}\theta + \lambda\mathcal{R}(\theta) \right\}, \tag{3.2}$$

where $\tilde{X} = [I_n \quad X]$, $\mathcal{R}(\theta) = \|\beta\|_1$, and

$$\tilde{L} = \begin{bmatrix} \gamma_n(L_n + \delta I_n) & 0 \\ 0 & \gamma_p J_p \end{bmatrix}.$$

Then, (3.2) can be solved using a simple proximal gradient descent algorithm, as given in Algorithm 2. In our simulations and data analysis, we use a fixed step-size of $\eta^t = 0.001$, which provides reasonably fast convergence.

Algorithm 2: Proximal gradient descent for `glm-funk` with ℓ_2 feature smoothing

Define $\mathcal{L}(\theta) := \ell(\tilde{X}\theta) + \frac{1}{2}\theta'\tilde{L}\theta$.

Initialize θ^0.

for $t = 0, 1, \ldots,$ *until convergence of* θ **do**

Compute gradient, $\nabla\mathcal{L}(\theta^t) = \tilde{X}'(\mu(\tilde{X}\theta^t) - y) + \tilde{L}\theta^t$.

Take a gradient step, $\tilde{\theta} = \theta^t - \eta^t\nabla\mathcal{L}(\theta^t)$.

Take a proximal step,

$$\begin{aligned}\theta^{t+1} &= \text{prox}(\tilde{\theta}) \\ &= (\tilde{\theta}_1, \ldots, \tilde{\theta}_n, S_\lambda(\tilde{\theta}_{n+1}), \ldots, S_\lambda(\tilde{\theta}_{n+p}))'\end{aligned}$$

where $S_\lambda(x) = \text{sign}(x)\max(0, |x| - \lambda)$.

end

Result: $\hat{\theta} = \theta^{t+1}$

3.3.2 Optimization for ℓ_1 feature network smoothing

A disadvantage of the generalized lasso penalty is that the elements of β are nonseparable in the penalty function; this leads to computational difficulties when using a non-identity GLM link function. To overcome this challenge, we solve an alternative problem proposed by Chen et al. [17], who replace the generalized lasso penalty with a smooth ℓ_∞ approximation:

$$f_\tau(\beta) = \max_{\|\nu\|_\infty < 1} \left\{ \nu' J_p\beta - \tau\frac{1}{2}\|\nu\|_2^2 \right\}.$$

Here, τ is a parameter that controls the approximation to the original ℓ_1 problem. We note that when $\tau = 0$, $f_\tau(\beta) = \|J_p\beta\|_1$. Chen et al. [17] prove that, for $\tau = \epsilon/|E_p|$, the absolute difference between optimal objective values of the original and approximate problems is upper bounded by ϵ. The gradient of $f_\tau(\beta)$ is $J_p\nu^*$ where $\nu^* = S_\infty\left(\frac{\gamma_p J_p\beta}{\tau}\right)$, and S_∞ is the

element-wise projection operator onto the ℓ_∞ unit ball:

$$S_\infty(x) = \begin{cases} x, & \text{for } -1 \leq x \leq 1 \\ 1, & \text{for } x \geq 1 \\ -1, & \text{for } x \leq -1 \end{cases} .$$

Replacing $\|J_p\beta\|_1$ with $f_\alpha(\beta)$, we can solve the approximate ℓ_1 `glm-funk` problem using an accelerated proximal gradient descent algorithm developed by Beck and Teboulle [5]. An adapted version of the algorithm presented in Chen et al. [17] is given in Algorithm 3. In our simulations, we set $\tau = 0.001$, which provides sensible results, and a fast convergence rate.

Algorithm 3: Accelerated proximal gradient descent for `glm-funk` with approximate ℓ_1 feature smoothing

Define $\mathcal{L}(\theta) := \ell(\tilde{X}\theta) + \frac{1}{2}\gamma_n\alpha'(L_n + \delta I_n)\alpha + \gamma_p f_\tau(\beta)$.

Initialize $\theta^0, w^0 = \theta^0, s^0 = 1$.

for $t = 0, 1, \ldots,$ *until convergence of* θ **do**

Compute gradient,

$$\nabla\mathcal{L}(w^t) := \tilde{X}'(\mu(\tilde{X}w^t) - y) + [\gamma_n L_n w^t \quad \gamma_p J_p'\nu^*]'$$

Compute Lipschitz constant $L := \|\nabla^2\mathcal{L}(w^t)\|_2$.

Take a gradient step, $\tilde{\theta} = w^t - L^{-1}\nabla\mathcal{L}(\theta^t)$.

Take a proximal step,

$$\begin{aligned}\theta^{t+1} &= \text{prox}(\tilde{\theta}) \\ &= (\tilde{\theta}_1, \ldots, \tilde{\theta}_n, S_{\lambda/L}(\tilde{\theta}_{n+1}), \ldots, S_{\lambda/L}(\tilde{\theta}_{n+p}))',\end{aligned}$$

where $S_{\lambda/L}(x) = \text{sign}(x)\max(0, |x| - \frac{\lambda}{L})$.

Set $s^{t+1} = 2/(t+3)$.

Set $w^{t+1} = \theta^{t+1} + \frac{1-s^t}{s^t}s^{t+1}(\theta^{t+1} - \theta^t)$.

end

Result: $\hat{\theta} = \theta^{t+1}$

3.3.3 Prediction and tuning

Now, suppose we use n_{trn} observations for training the `glm-funk` model, and are interested in predicting outcomes for n_{tst} out-of-sample observations. In order to make predictions on out-of-sample data, we require an estimate of the unit-level intercepts α_{tst}. The test sample predictions are then given as $\mu(\hat{\alpha}_{\text{tst}} + X\hat{\beta})$. Assuming we observe the entire network G_{full}

connecting the $n_{\text{trn}} + n_{\text{tst}}$ units, we partition the Laplacian corresponding to G_{full} as:

$$L_{\text{full}} = \begin{bmatrix} L_{\text{trn,trn}} & L_{\text{trn,tst}} \\ L_{\text{tst,trn}} & L_{\text{tst,tst}} \end{bmatrix} = \begin{bmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{bmatrix}$$

Then, we estimate α_{tst} as in Li et al. [67]:

$$\hat{\alpha}_{\text{tst}} = \underset{\alpha_{\text{tst}}}{\text{argmin}} \left\{ (\hat{\alpha}_{\text{trn}}, \alpha_{\text{tst}})' L_{\text{full}} (\hat{\alpha}_{\text{trn}}, \alpha_{\text{tst}}) \right\} = -L_{22}^{-1} L_{21} \hat{\alpha}_{\text{trn}}$$

Note that no network knowledge for the test observations (i.e. the training and test units are disjoint on G_{full}), corresponds to estimating $\hat{\alpha}_{\text{tst}} = 0$.

The `glm-funk` problems involve three tuning parameters γ_n, γ_p, and λ. We tune these using K-fold cross-validation to minimize prediction error. Ideally, the K folds would be determined using non-overlapping connected components of G_n, but this is not always possible for arbitrary networks. Due to the dependence among observation units over G_n, naive cross-validation is not guaranteed to provide a good estimate of out-of-sample prediction error. However, as in Li et al. [67], the procedure works relatively well in practice. While Rabinowicz and Rosset [82] describe a method for cross-validation with correlated data, it requires the population covariance matrix, which we do not directly assume knowledge of. In order to efficiently determine the optimal parameter set, we use coordinate descent [119]. Specifically, we optimize a single parameter (via K-fold cross-validation) while holding the others fixed, and cycle through all three parameters. In practice, this procedure usually converges in a very small number of coordinate descent iterations.

3.4 Asymptotics and inference

We are interested in obtaining valid inference for the association between the features X and the outcome y. Our estimator given in (3.1) is non-standard due to the use of the n-dimensional intercept term and ℓ_2 penalty terms that are not seminorms. In this section, we first investigate the large sample behaviour of $\hat{\beta}$ and $\hat{\alpha}$ estimated using the `glm-funk` model with ℓ_1 smoothing. We defer discussion of the ℓ_2 smoothing model to Section 3.9. These

results are then used to obtain a valid statistical inference procedure for the true regression parameters.

3.4.1 Asymptotics

We begin by describing the assumptions required for our theoretical results to hold. First, we require that the outcomes Y_i satisfy certain tail properties.

Assumption 6 (Tail behaviour)**.** *One of the following holds:*
(i) The centered observed outcomes $Y_i - \mathbb{E}[Y_i|X_i = x_i] = Y_i - \mu_i$ *are uniformly sub-Gaussian, i.e.*

$$\max_{i=1,\dots,n} K^2\mathbb{E}\left[\exp\left(\frac{(Y_i-\mu_i)^2}{K^2}\right)-1\right] \leq \sigma_0^2.$$

for some constants $K, \sigma_0^2 > 0$*, or*
(ii) The centered observed outcomes $Y_i - \mathbb{E}[Y_i|X_i = x_i] = Y_i - \mu_i$ *are uniformly sub-exponential, satisfying*

$$\max_{i\in 1:n} \|Y_i - \mu_i\|_{\psi_1} = K_{\psi_1} < \infty,$$

where

$$\|Y\|_{\psi_1} = \inf\left\{t > 0 : \mathbb{E}\exp(|Y|/t) \leq 2\right\}.$$

These tail conditions cover a large variety of common generalized linear models. Gaussian and binomial data satisfy the sub-Gaussian property, while Poisson and exponential outcomes have sub-exponential tails.

We further require some conditions on the loss function ℓ and the design matrix X.

Assumption 7 (Loss function properties)**.** *The following hold:*
(i) The loss function $\ell : \Theta \times \Omega \to \mathbb{R}$ *is integrable over all* $(X, y) \in \Omega$ *for each* $\theta \in \Theta$*.*
(ii) For almost all $(X, y) \in \Omega$*, the derivative* $\nabla_\alpha \ell$ *exists for all* α*.*
(iii) There exists an integrable function $g : \Omega \to \mathbb{R}$ *such that* $|\ell| \leq g(X, y)$ *for all* $\theta \in \Theta$ *and almost all* $(X, y) \in \Omega$*.*

(iv) The conditional mean function μ and its derivative μ' are Lipschitz continuous with constants $L_\mu < \infty$ and $L_{\mu'} < \infty$.
(v) μ' is uniformly bounded away from zero, that is, $|\mu'(\cdot)|^{-1} \leq U' < \infty$.

Assumption 8 (Design scaling)**.** *The design matrix X satisfies $|X_{ij}| \leq R < \infty$ for all i, j, and also scales as $\|X\|_2 = o_p\left(\sqrt{\frac{n}{\log p}}\right)$.*

The loss function assumptions are fairly mild, and common in the high-dimensional inference literature [113, 51, 10]. The design scaling assumption is equivalent to assuming that the maximum eigenvalue of $X'X$ grows at a rate slower than n, and can be shown to hold for various random designs (see Section 6.4 of Wainwright [116]).

In order to state the remaining assumptions, we first define some quantities of interest. For a generic function f, let $\mathbb{P}f := n^{-1}\sum_{i=1}^n \mathbb{E}[f(y_i, x_i)]$ and $\mathbb{P}_n f := n^{-1}\sum_{i=1}^n f(y_i, x_i)$. Then, we rewrite our optimization problem as:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \left\{\mathbb{P}_n \mathcal{L}(\theta) + \lambda \mathcal{R}(\theta)\right\},$$

where $\mathcal{L}(\theta) = \ell_i(\alpha_i + x_i'\beta) + \frac{1}{2}\gamma_n \alpha'(L_n + \delta I_n)\alpha$ and $\mathcal{R}(\theta) = \|\beta\|_1 + \frac{\gamma_p}{\lambda}\|J_p\beta\|_1$.

Definition 11. *The true parameter is defined as*

$$\theta^0 := \underset{\theta}{\operatorname{argmin}}\, \mathbb{P}\ell(\theta).$$

Definition 12. *The target parameter is defined as*

$$\theta^* := \underset{\theta}{\operatorname{argmin}}\, \mathbb{P}\mathcal{L}(\theta).$$

It is important to note that our theoretical analysis is in a high-dimensional framework, where n and $p(n)$ are both allowed to grow to infinity (and therefore, so are the dimensions of α and β). Hence, θ^0 and θ^* are dependent on n and p, and the following assumptions apply to a sequence of data-generating processes indexed by (n, p). Our theoretical results then hold with high probability for large (n, p). For ease of exposition, we do not include this dependence in our notations.

We make the following assumptions for estimation of the target and true regression parameters.

Assumption 9 (Compatibility condition)**.** *Given a set $S \subset \{1, \ldots, p\}$ with $|S| = s$, for all $c > 0$ and for all $\theta = (\alpha, \beta)'$ satisfying $\|\beta_{S^c}\|_1 + c\|J_p\beta\|_1 \leq \|\alpha\|_1 + 3\|\beta_S\|_1$, it holds that:*

$$\frac{\|\alpha\|_1}{2} + \|\beta_S\|_1 \leq \frac{\|\theta\|\sqrt{s}}{\phi(s)}$$

for some norm $\|\cdot\|$ and constant $\phi(s) > 0$.

Assumption 10 (Restricted strong convexity)**.** *Assume $\bar{\theta} \in \left\{\|\bar{\theta} - \theta^0\| \leq r_0\right\}$ where $r_0 = O_p\left(\sqrt{\frac{\log p}{n}}\right)$. Then, for all $\theta = (\alpha, \beta)'$ satisfying*

$$\|\alpha - \bar{\alpha}\|_1 + \|\beta - \bar{\beta}\|_1 + \frac{\gamma_p}{\lambda}\left\|J_p\left(\beta - \bar{\beta}\right)\right\|_1 \leq M^*,$$

with

$$M^* = \frac{16s\lambda^2}{\rho\phi^2(S)c} + \frac{2\gamma_p\|J_p\bar{\beta}\|_1}{\rho},$$

and $\lambda \geq 8\rho$, it holds that:

$$\mathbb{P}\left(\ell(\theta) - \ell(\bar{\theta})\right) \geq \nabla\mathbb{P}\ell\left(\bar{\theta}\right)'\left(\theta - \bar{\theta}\right) + G\left(\|\theta - \bar{\theta}\|\right),$$

where $G(x) = cx^2$ for some constant $c > 0$.

The compatibility and restricted strong convexity conditions are common in high-dimensional theory [9]. Intuitively, restricted strong convexity at the optimum $\bar{\theta} = \theta^0$ means that the loss function is curved sharply around θ^0. Hence, when $\mathbb{P}(\ell(\theta^0) - \ell(\theta))$ is small, so is $\|\theta^0 - \theta\|$. Negahban et al. [77] proved that restricted strong convexity holds for various common loss functions in sparse high-dimensional regimes, including the least squares loss and logistic regression deviance.

Finally, we make assumptions on components of the true data-generating processes, and their relationship to the penalty parameters in the model.

Assumption 11 (Sparsity)**.** *β^0 is s-sparse, that is, $\|\beta^0\|_0 = s$ with $s = O_p\left(\sqrt{n}/\log p\right)$.*

Assumption 12 (Penalty scaling)**.** *The following hold:*
(i) $\lambda = O_p\left(\sqrt{\frac{\log p}{n}}\right)$,
(ii) $\gamma_p\|J_p\beta^0\|_1 = o_p(\lambda)$, *and*
(iii) $\gamma_n\|(L_n + \delta I)\alpha^0\|_2 = O_p\left(n^c\right)$ *where* $c \in (0, \frac{1}{2})$.

The sparsity assumption and scaling condition of λ are standard rates in the high-dimensional inference literature [113, 78].

Part (ii) of Assumption 12 allows us to not observe fully informative feature networks. It states that the quality of feature network smoothing is inversely proportional to the magnitude of its tuning parameter. That is, if the feature network is truly informative, we expect $\|J_p\beta\|_1 \to 0$ (at a rate faster than λ), and γ_p can be larger. In this scenario, the network structure informs more of the β-penalization than the lasso penalty. However, if the smoothing does not correspond to the true structure of β, then $\|J_p\beta^0\|_1$ will be far from zero, and γ_p should tend to 0 at a rate faster than $\|J_p\beta^0\|_1^{-1}$. In addition, if the feature network is uninformative, then γ_p also needs to go to 0 faster than λ. In this case, most of the penalization is driven by the lasso penalty, rather than the network structure encoded by J_p. In practice, our naive cross-validation approach leads to consistent estimation, as seen empirically in Section 3.5.

Part (iii) of Assumption 12 similarly allows for some degree of non-informativeness in the unit network, and is also necessary to establish control of $\|\alpha^* - \alpha^0\|$. It establishes the trade-off needed between the unit network parameter γ_n and the ridge penalty parameter δ; if the unit network is informative and γ_n is large, then δ should shrink to 0.

Under these assumptions, we can prove that $\hat{\beta}$ and $\hat{\alpha}$ tend to the target parameters β^* and α^* in ℓ_1 norm.

Theorem 2 (Consistency)**.** *Under Assumptions 6-12, we have that*

$$\|\hat{\alpha} - \alpha^*\|_1 + \|\hat{\beta} - \beta^*\|_1 = O_p\left(\lambda + \frac{\gamma_p}{\lambda}\|J_p\beta^*\|_1\right)$$

The proof, given in Section 3.8, follows a similar argument as the proof for generalized sparse additive models in Haris et al. [39]. A key difference in our theory is handling the n-dimensional intercept term α, and its corresponding ℓ_2 penalty. We also prove the result for outcome distributions which are not sub-Gaussian, such as Poisson and exponential data.

3.4.2 Inference

We now describe a statistical inference procedure for the β parameters in the `glm-funk` model. We are specifically interested in testing the individual association between outcome y and feature X_j conditional on the other features X_{-j}. This corresponds to the null hypotheses $H_{0,j} : \beta_j = 0$ for all $j = 1, \ldots, p$. Classical inference theory does not directly apply in the high-dimensional setting. Approaches that have been developed generally either involve sample splitting, or constructing an asymptotically unbiased estimator from the optimal solution $\hat{\beta}$ [20]. We focus on the latter approach, which is known as *debiasing*.

Several de-biasing procedures have been developed, including the low-dimensional projection estimator by Zhang and Zhang [122], the ridge projection estimator by Bühlmann et al. [10], and the desparsified lasso estimator by van de Geer et al. [113]. These methods generally consider regression models with an ordinary ridge or lasso penalty only. The Grace test by Zhao and Shojaie [124] specifically provides inference for linear regression with the ℓ_2 Laplacian penalty. However, this method does not account for unit-level networks, and does not extend to the case of generalized linear models. It also tends to be conservative in terms of power.

We consider the debiased estimator of Javanmard and Montanari [51], which easily extends to generalized linear models. The debiased estimator we consider is defined as:

$$\hat{b} = \hat{\beta} - n^{-1} M X'(\mu(\hat{\alpha} + X\hat{\beta}) - y).$$

where M is an estimate of the inverse of $\hat{\Sigma} := \frac{1}{n}\nabla^2 \ell(\hat{\alpha} + X\hat{\beta})$; M is computed by solving an optimization problem where $\|\hat{\Sigma}M - I_p\|_\infty$ is minimized. Specifically, for each $j = 1, \ldots, p$,

m_j is defined as the solution to

$$\min_{m \in \mathbb{R}^p} m' \hat{\Sigma} m \quad \text{subject to} \quad \|\hat{\Sigma} m - e_i\|_\infty \leq q,$$

where $e_j \in \mathbb{R}^p$ is the j-th basis vector. Then, $M = (m_1, \ldots, m_p)'$. The next result shows that under the assumptions described previously, we obtain an asymptotic distribution suitable for inference.

Theorem 3 (Valid inference)**.** *Under the conditions of Theorem 2, as* $n \longrightarrow \infty$*,*

$$\sqrt{n}\left(\hat{b} - \beta^0\right) \longrightarrow_d N\left(0, n^{-1} M \mathbb{E}\left[\nabla \ell(\alpha^0 + X\beta^0) \nabla \ell(\alpha^0 + X\beta^0)'\right] M\right).$$

Even though our target parameter θ^* is different from θ^0, using Theorem 3, we obtain valid inference for the true regression parameters β_j^0 using the corresponding test statistic:

$$T_j = \frac{\sqrt{n}\hat{b}_j}{[M\hat{\Sigma}M]_{jj}^{1/2}},$$

and 100(1 - α)% confidence interval:

$$\hat{b}_j \pm q_{(1-\frac{\alpha}{2})} n^{-1/2} [M\hat{\Sigma}M]_{jj}^{1/2}.$$

In the proof, given in Section 3.8, we establish that the target β^* is the same as the true β^0, and derive an asymptotic rate for $\|\alpha^* - \alpha^0\|_1$. We then apply Theorem 2 to show the result.

3.5 Simulation studies

In this section, we compare logistic `glm-funk` models with lasso-penalized logistic regression, and a lasso-penalized RNC logistic model. We evaluate model prediction performance, as measured by cross-validated error, and inference performance, as measured by the empirical Type I error rate and power of our debiasing procedure. Note that while Li et al. [67] do not include inference for high-dimensional β parameters, this follows as a special case of our method. In the following, we consider a setting where the feature and unit networks are

fully informative; i.e. the network penalties correspond to the true structure of α^0 and β^0. To assess the robustness of our method, we also evaluate the methods in a setting where uninformative edges are added to the networks, resulting in penalization that does not reflect the true data-generating process.

3.5.1 Fully informative networks

We generate binary data from the following model:

$$P(Y = 1|X = x) = \text{expit}(\alpha + X\beta)$$

with $p = 300, n = 200$, and $s = 20$. We set $\beta_1 = \ldots \beta_{10} = \rho$, $\beta_{11} = \ldots \beta_{20} = -\rho$, and the remaining coefficients to 0. The feature graph G_p is set to have $2p/s$ disconnected components of size $s/2$ each. Each component has a single hub node which is connected to the remaining non-hub nodes that do not have any other connections. In generating the features X, each hub node feature is generated as $x_h \sim N(0, 1)$, with connected non-hub features generated as $x_{nh} \sim N(0.35x_h, 1)$. This setup is similar to that of the simulations in Li and Li [64].

As in Li et al. [67], we set G_n to be a stochastic block model [41]. We divide the observed units into five fully-connected blocks with equal probability of membership. The unit-level intercepts are then generated from normal distributions that differ between blocks. Specifically, the block means considered are -4, -2, 0, 2, and 4, so that the intercepts have meaningful effects on $P(Y = 1|X = x)$. All distributions have a common standard deviation of 0.2.

In our simulations, we vary the effect size ρ, and select tuning parameters via 5-fold cross-validation. We report the average powers and Type I error rates at the 0.05 significance level, and the test set logistic deviance over 100 simulated datasets.

Results for this setting are shown in Figure 3.1. We observe that the `glm-funk` models outperform the models which do not incorporate the feature network information, both in terms of power and test set deviance. As expected, the simple lasso model with a common

intercept for all units performs the worst, while the lasso-penalized RNC model performs slightly better. Comparing `glm-funk` with ℓ_1 and ℓ_2 smoothing, we see that the ℓ_1 model achieves lower test set deviance, which makes sense since the connected β coefficients are exactly equal to each other. All models appropriately control the Type I error rate at the 0.05 level, with the ℓ_2 `glm-funk` model giving the lowest rate.

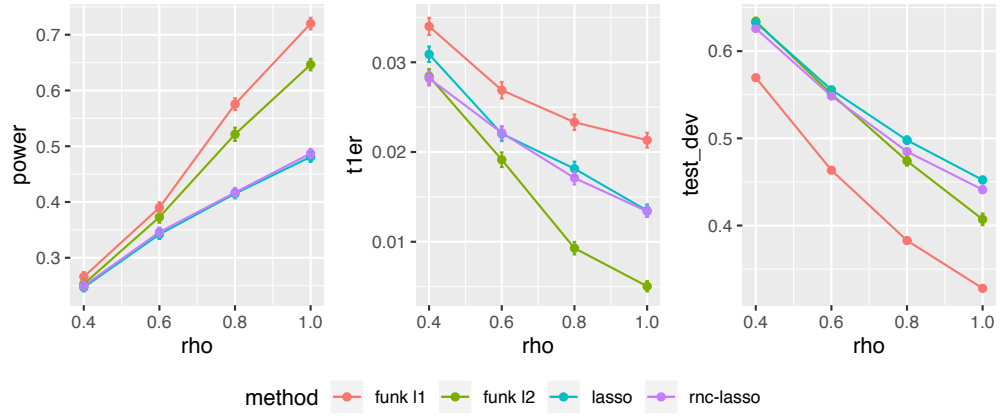

Figure 3.1: Simulation results for fully informative networks. Means over 100 replicates are displayed with standard error bars. Left: power, middle: Type I error rate, right: test set deviance.

3.5.2 Uninformative networks

We now examine the effect of adding uninformative edges to both networks. For G_n, we generate the stochastic block model with a constant uniform probability 0.01 of intra-block edges forming. For G_p, we randomly add intra-component edges to the original network with constant probability 0.003. In our simulations, this corresponds to, on average, 164 additional edges in G_n (4% of edges uninformative) and 135 additional edges in G_p (34% of edges uninformative). These edges encourage intercepts from different blocks or β coefficients from different components to be close to each other.

Results for this setting are shown in Figure 3.2. We observe similar trends as in the fully informative case. However, the ℓ_1 `glm-funk` model now has the worst test set error among

the models. This makes sense, since the ℓ_1 smoothing results in setting coefficients to be exactly equal. Given false edges in G_p, this is more likely to result in inaccurate coefficients than would ℓ_2 smoothing, which maintains predictive ability slightly better than the other methods. As more noise is added to the observed networks, we expect that the models would all perform identically in terms of power, while the ℓ_1 `glm-funk` model may perform worst in terms of test set error.

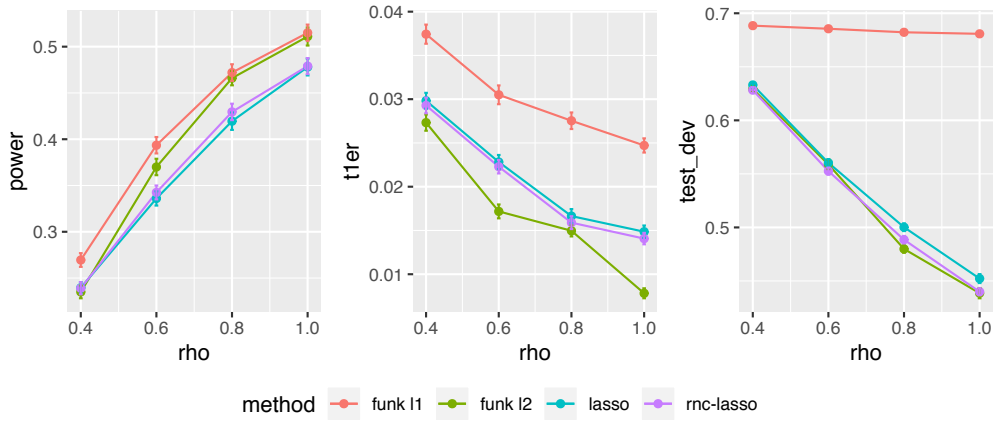

Figure 3.2: Simulation results for uninformative networks. Means over 100 replicates are displayed with standard error bars. Left: power, middle: Type I error rate, right: test set deviance.

3.6 Data analysis

We apply our method, in addition to RNC with a lasso penalty and the standard lasso, to a case-control study of lung cancer [27]. This dataset consists of 162 subjects, on whom plasma measurements were taken and for which there are 137 named metabolites and 153 unnamed features available. The outcome of interest is a binary indicator of whether the subject has a lung cancer diagnosis. There are 94 cases and 74 controls. To identify named metabolites associated with lung cancer status, we fit logistic regression models to predict the probability of a subject having lung cancer, given named metabolite predictors.

Using the unnamed features, we construct a weighted graph for the subjects. To this

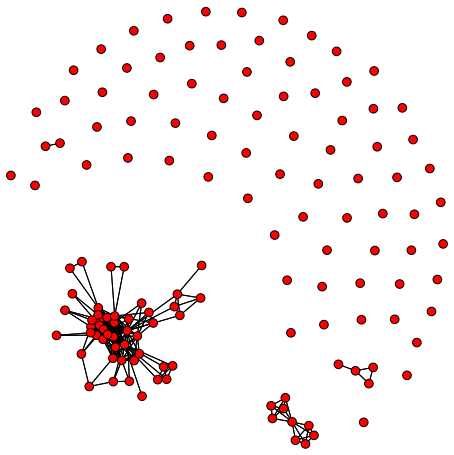

Figure 3.3: Feature network of named metabolites used for lung cancer data analysis.

end, we first apply principal component analysis (PCA) to the unnamed feature matrix, and reduce its dimension to the first 10 principal components (other numbers of PCs gave qualitatively very similar results). For each pair of subjects (i, j), we compute the Euclidean distance d_{ij} with respect to their PCs. The edge weight for subjects (i, j) is defined as $1 - (d_{ij} / \max_{i,j} d_{ij})$. Then, the kernel matrix used is the graph Laplacian.

For the named metabolite features, we use the Kyoto Encyclopedia of Genes and Genomes (KEGG) [58] to determine molecular pathways. We then construct a feature graph where an edge exists between each pair of metabolites that are on the same KEGG pathway. This results in a graph containing 137 nodes and 524 edges. 74 metabolites have no connections, while 63 have at least one. All edge weights are set to 1 (we also considered a feature graph with edge weights set to be the reciprocal of the shortest path length between metabolites, but this provided very similar results). An illustration of this graph can be seen in Figure 3.3.

In conducting our analysis, we randomly split the data into 100 subjects for training and 62 subjects as the test set, on which we evaluate the misclassification rate. We average over 100 random splits in order to obtain more stable estimates of the test set error. The

Method	Misclassification rate
`glm-funk` ℓ_1	**0.3657**
`glm-funk` ℓ_2	0.3718
`lasso`	0.3851
`RNC-lasso`	0.3772

Table 3.1: Estimated misclassification errors of models fit to predict lung cancer status given metabolomics data

results are reported in Table 3.1. We see that the `glm-funk` models perform better than the lasso-based models. Using the unnamed features as similarity measures between units also improves prediction compared to the model with only the lasso penalty. In particular, the model with ℓ_1 feature network smoothing has the lowest average misclassification rate.

To investigate the differences between the features selected by the methods, we look at the average fitted β coefficients, and how often each feature would be declared significant (at the 5% level) over the 100 training/testing splits. These results are summarized in Table 3.2 for some of the named metabolite features. The metabolite *pyrophosphate* is found to be significant most often across all methods. Pyrophosphate was also found to be the best predictor of lung cancer in the original investigation by Fahrmann et al. [27]. In our analysis, the `glm-funk` model with ℓ_1 smoothing declares this feature to be significant the most often, and also fits the largest β coefficient on average. This also applies for the metabolite *pyruvic acid*, which was the second-most significant feature across all methods, and has also been found previously to be associated with lung cancer [62]. Therefore, it appears that incorporating the KEGG feature network can result in better detection of important metabolites, compared to methods that do not use this information.

A notable difference between the `glm-funk` and lasso-based methods is seen in the metabolite *taurine*, for which the `glm-funk` fits have smaller estimated coefficients and significance probabilities. The metabolite's neighbours in the KEGG graph (*glycine* and

Metabolite	Method	Average β	P(sig)
pyrophosphate	`glm-funk` ℓ_1	0.376	0.94
	`glm-funk` ℓ_2	0.314	0.80
	`lasso`	0.241	0.88
	`RNC-lasso`	0.277	0.78
pyruvic acid	`glm-funk` ℓ_1	0.157	0.62
	`glm-funk` ℓ_2	0.097	0.47
	`lasso`	0.095	0.56
	`RNC-lasso`	0.085	0.44
taurine	`glm-funk` ℓ_1	0.018	0.20
	`glm-funk` ℓ_2	0.054	0.23
	`lasso`	0.058	0.40
	`RNC-lasso`	0.073	0.27

Table 3.2: Model coefficient fits and significance for select named metabolites

cholesterol) usually have null coefficients fit by all four methods. This does not affect the lasso or RNC, but for the `glm-funk` fits, this leads to more penalization being applied to the coefficient for taurine. This is visualized in Figure 3.4, which compares the average magnitudes of coefficients under the ordinary lasso and the `glm-funk`ℓ_1 fits. We observe that coefficients which are non-null in the lasso fit, but are connected to null coefficients, tend to be shrunken in the `glm-funk` fit.

3.7 Discussion

Motivated by applications in metabolomics, we have developed a new framework for analysis of two-way network-structured data using penalized generalized linear models. We also propose valid high-dimensional inference for our model parameters, under potentially uninformative network structure. This methodology can also be used in other applications, such

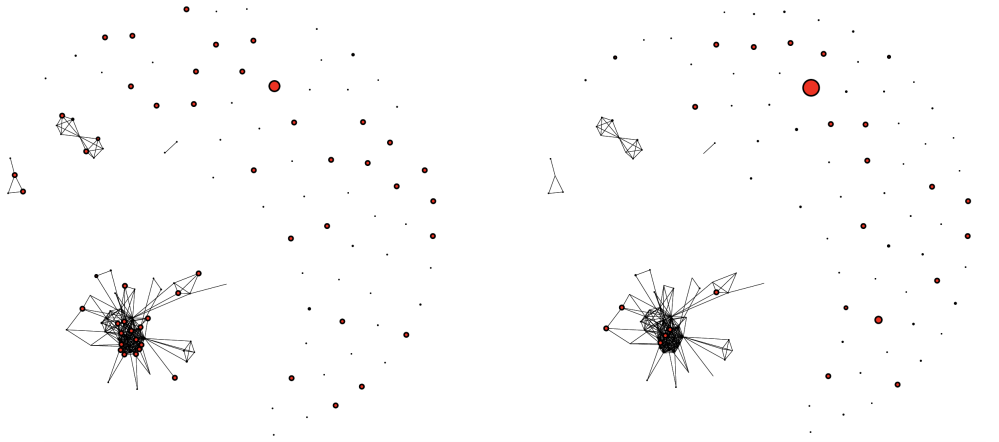

Figure 3.4: Feature networks under `lasso` (left) and `glm-funk` ℓ_1 (right) fits. Size of nodes corresponds to average magnitude of corresponding β coefficient over 100 training/testing splits.

as multi-view data integration.

When the networks are informative for the true data-generating process, we expect our method to show improved prediction and inference compared to standard high-dimensional methods. When the networks are misspecified or uninformative, our method should still theoretically achieve consistent estimation and valid inference for the true regression parameters. However, tuning the network penalty parameters to achieve this in practice may be difficult, as naive cross-validation is not guaranteed to be successful, given the dependencies among the obervation units. This is shown in our empirical studies, where the `glm-funk` method with ℓ_1 feature network smoothing has worse prediction (though still the highest inferential power while controlling type I error) than other methods given an uninformative network.

Although we only considered Laplacian and incidence matrices, different kernels can easily be used within the `glm-funk` model. Another possibility would be to avoid adding unit-level intercepts, and instead penalizing the fitted values $X\beta$ directly; that is, using the penalty $\gamma_n \beta' X' L_n X \beta$. An interesting direction for future research would be to parameterize

the model to have many penalty parameters over each network, rather than a single γ_p or γ_n. By data-adaptively tuning these parameters, the model would ideally incorporate the information from informative edges only. Feng and Simon [29] show that tuning a large number of penalty parameters via cross-validation is feasible and can result in improved prediction performance.

3.8 Technical proofs

Here, we prove our results in Section 3.4. We begin by comparing the target parameters θ^* to the true parameters θ^0, showing that $\beta^* = \beta^0$ and that $\|\alpha^* - \alpha^0\|_1$ decays to 0 asymptotically under our assumptions. We then focus on estimation of the target θ^* using the ℓ_1-regularized $\hat{\theta}$. We derive tail bounds on the empirical process term under both sub-Gaussian and sub-exponential outcomes, which are then used in showing that $\hat{\theta} \to \theta^*$ in ℓ_1 norm. Our proof of Theorem 2 is very similar to that of Theorem 3 from Haris et al. [39], which provides fast rate bounds for generalized sparse additive models. A key step in the proof by Haris et al. [39], which differentiates it from a similar one in Bühlmann and van de Geer [9], is handling an intercept term which is not penalized. We extend this to our setting, with an n-dimensional intercept term that is ℓ_2 penalized. We then prove Theorem 3, which shows the validity of our debiased estimator $\hat{b}$ for inference on the true β^0.

Target vs. true parameters

We first compare the target parameters θ^* to the true parameters of interest θ^0. We start by noting that the target β^* is the same as the true β^0. Using this fact, we characterize the difference in the target and true intercepts; that is, $\|\alpha^* - \alpha^0\|$.

Lemma 4. *The target parameter β^* is equal to the true parameter β^0.*

Proof. This follows immediately by examining the β-optimality conditions for both objective functions,

$$0 = \nabla_\beta \mathbb{E}[\ell(\alpha^* + X\beta^*)]$$
$$0 = \nabla_\beta \mathbb{E}[\ell(\alpha^0 + X\beta^0)]$$

and by convexity of the loss function. □

Lemma 5. *Under Assumptions 7 and 12, we have that $\|\alpha^* - \alpha^0\|_1 = O_p(n^c)$ where $c \in (0, \frac{1}{2})$.*

Proof. Considering the α-optimality condition for the target objective function, we have that

$$\begin{aligned}
0 &= \nabla_\alpha \mathbb{E}[\ell(\alpha^* + X\beta^*)] + \gamma_n(L_n + \delta I)\alpha^* \\
0 &= \mathbb{E}[\nabla_\alpha \ell(\alpha^* + X\beta^*)] + \gamma_n(L_n + \delta I)\alpha^* \quad \text{(by Assumption 7)} \\
0 &= \mathbb{E}[y - \mu(\alpha^* + X\beta^*)] + \gamma_n(L_n + \delta I)\alpha^*
\end{aligned}$$

Taking a first order Taylor expansion around the true parameter α^0, we obtain

$$0 = \mathbb{E}[y - \mu(\alpha^0 + X\beta^*)] + \gamma_n(L_n + \delta I)\alpha^0 + \left[W_P(\tilde{\alpha} + X\beta^*) + \gamma_n(L_n + \delta I)\right](\alpha^* - \alpha^0),$$

where $\tilde{\alpha}$ is an intermediate point on the line segment between α^* and α^0 and W_P is the diagonal matrix of the derivative μ' over the true data-generating distribution $(y, X) \sim P$ at mean $\mu(\tilde{\alpha} + X\beta^*)$.

Then, since $\beta^* = \beta^0$, $\mathbb{E}[y - \mu(\alpha^0 + X\beta^*)] = 0$. Therefore,

$$\alpha^0 - \alpha^* = \left[W_P(\tilde{\alpha} + X\beta^*) + \gamma_n(L_n + \delta I)\right]^{-1} \gamma_n(L_n + \delta I)\alpha^0$$

Taking ℓ_2 norms, we have:

$$\begin{aligned}
\|\alpha^0 - \alpha^*\|_2 &\leq \gamma_n \| \left[W_P(\tilde{\alpha} + X\beta^*) + \gamma_n(L_n + \delta I)\right]^{-1} \|_2 \|(L_n + \delta I)\alpha^0\|_2 \\
&\leq \gamma_n \lambda_{min}[W_P(\tilde{\alpha} + X\beta^*) + \gamma_n(L_n + \delta I)]^{-1} \|(L_n + \delta I)\alpha^0\|_2 \\
&\leq \gamma_n \lambda_{min}[W_P(\tilde{\alpha} + X\beta^*)]^{-1} \|(L_n + \delta I)\alpha^0\|_2 \\
&\leq U' \gamma_n \|(L_n + \delta I)\alpha^0\|_2
\end{aligned}$$

where the final inequality follows from Assumption 7.

The result then follows from part (iii) of Assumption 12. □

Control of empirical process

Recall that our optimization problem can be rewritten as:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \left\{ \mathbb{P}_n \mathcal{L}(\theta) + \lambda \mathcal{R}(\theta) \right\},$$

where $\mathcal{L}(\theta) = \ell_i(\alpha_i + x_i'\beta) + \frac{1}{2}\gamma_n\alpha'(L_n + \delta I_n)\alpha$ and $\mathcal{R}(\theta) = \|\beta\|_1 + \frac{\gamma_p}{\lambda}\|J_p\beta\|_1$.

We define the *empirical process term* as

$$\begin{aligned}\nu_n(\theta) &:= (\mathbb{P}_n - \mathbb{P})\mathcal{L}(\theta) \\ &= (\mathbb{P}_n - \mathbb{P})\ell(\theta),\end{aligned}$$

since $\mathbb{P}_n(\alpha'(L_n + \delta I_n)\alpha) = \mathbb{P}(\alpha'(L_n + \delta I_n)\alpha)$. The *excess risk* is defined as $\mathcal{E}(\theta) = \mathbb{P}(\mathcal{L}(\theta^*) - \mathcal{L}(\theta))$.

Following Haris et al. [39], we have the following basic inequality:

$$\mathcal{E}(\hat{\theta}) + \lambda\mathcal{R}(\hat{\theta}) \leq -[\nu_n(\hat{\theta}) - \nu_n(\theta^*)] + \lambda\mathcal{R}(\theta^*),$$

which also holds if $\hat{\theta}$ is replaced by $\tilde{\theta} := t\hat{\theta} + (1-t)\theta^*$ where $t \in (0,1)$.

In order to prove that $\hat{\theta} \to \theta^*$, we require control of the empirical process term. We first consider Assumption 6, and show the following lemma.

Lemma 6. *Under part (i) of Assumption 6, with probability at least* $1 - 2\exp(-n\rho^2 C_1) - C\exp(-n\rho^2 C_2)$*, the following inequality holds:*

$$\nu_n(\theta) - \nu_n(\theta^*) \leq \rho\left[\|\alpha - \alpha^*\|_1 + \|\beta - \beta^*\|_1 + \frac{\gamma_p}{\lambda}\|J_p(\beta - \beta^*)\|_1\right],$$

where $\rho = O\left(\sqrt{\frac{\log p}{n}}\right)$ *and* C, C_1, C_2 *are positive constants independent of* n *and* p*.*

Proof. This result can be proved almost identically as in Haris et al. [39], with the exception of handling the $n-$dimensional intercept α.

Let $x_i \in \mathbb{R}^p$ and $Y_i \in \mathbb{R}$ denote the fixed covariates and response respectively, for $i = 1, \ldots, n$. Write the loss for a single observation as

$$\ell(\theta) = aY_i(\alpha_i + x_i'\beta) + h(\alpha_i + x_i'\beta),$$

for some $a \in \mathbb{R} \setminus \{0\}$ and function $h : \mathbb{R} \to \mathbb{R}$.

Assume $a = 1$, without loss of generality, since this constant will be absorbed into the probability bounds later. Then, since x_i are assumed fixed, and denoting $\mu_i := \mathbb{E}[Y_i]$.

$$\nu_n(\theta) = n^{-1}\sum_{i=1}^{n}(Y_i - \mu_i)(\alpha_i + x_i'\beta).$$

Then, we can write:

$$\begin{aligned}\nu_n(\theta) - \nu_n(\theta^*) &= n^{-1}\sum_{i=1}^{n}(Y_i - \mu_i)\left[(\alpha_i - \alpha_i^*) + \sum_{j=1}^{p}(\beta_j x_{ij} - \beta_j^* x_{ij})\right] \\ &= n^{-1}\sum_{i=1}^{n}(Y_i - \mu_i)(\alpha_i - \alpha_i^*) + n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{p}(\beta_j x_{ij} - \beta_j^* x_{ij})(Y_i - \mu_i).\end{aligned}$$

We now want to bound the probability that $\nu_n(\theta) - \nu_n(\theta^*)$ exceeds

$$\rho\left[\|\alpha - \alpha^*\|_1 + \|\beta - \beta^*\|_1 + \frac{\gamma_p}{\lambda}\|J_p(\beta - \beta^*)\|_1\right]$$

Consider the following probability involving the first term:

$$P\left(\frac{n^{-1}\sum_{i=1}^{n}(Y_i - \mu_i)(\alpha_i - \alpha_i^*)}{\|\alpha - \alpha^*\|_1} \geq \rho\right).$$

Applying the sub-Gaussian concentration inequality from Lemma 8.2 of van de Geer [112],

$$P\left(\left|\frac{n^{-1}\sum_{i=1}^{n}(Y_i - \mu_i)(\alpha_i - \alpha_i^*)}{\|\alpha - \alpha^*\|_1}\right| \geq \rho\right) \leq 2\exp\left[-\frac{\rho^2}{8(K^2 + \sigma_0^2)\sum_{i=1}^{n}\gamma_i^2}\right],$$

where

$$\begin{aligned}\sum_{i=1}^{n}\gamma_i^2 &= \sum_{i=1}^{n}\left(\frac{\alpha_i - \alpha_i^*}{n\|\alpha - \alpha^*\|_1}\right)^2 \\ &= \frac{1}{n^2}\frac{\|\alpha - \alpha^*\|_2^2}{\|\alpha - \alpha^*\|_1^2} \leq \frac{1}{n^2}.\end{aligned}$$

Therefore,

$$\begin{aligned}P\left(\left|\frac{n^{-1}\sum_{i=1}^{n}(Y_i - \mu_i)(\alpha_i - \alpha_i^*)}{\|\alpha_i - \alpha_i^*\|_1}\right| \geq \rho\right) &\leq 2\exp\left[-\frac{n^2\rho^2}{8(K^2 + \sigma_0^2)}\right] \\ &= 2\exp\left(-C_1 n^2\rho^2\right),\end{aligned}$$

where $C_1 = C_1(K, \sigma_0^2)$.

The rest of the proof follows Haris et al. [39], showing

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}(\beta_j x_{ij} - \beta_j^* x_{ij})(Y_i - \mu_i) \leq \rho\left[\|\beta - \beta^*\|_1 + \frac{\gamma_p}{\lambda}\|J_p(\beta - \beta^*)\|_1\right]$$

with high probability.

More specifically, we use a logarithmic entropy bound on the parametric GLM family of functions. For each j, the following bound on δ-entropy holds with some constant A_0 and $T_n = 1$:

$$\log N(\delta, \mathcal{F}, \|\cdot\|_Q) \leq A_0 T_n \log(1/\delta + 1),$$

where

$$\mathcal{F} = \left\{\beta_j x : |\beta_j| + \frac{\gamma_p}{\lambda}|(J_p\beta)_j| \leq 1\right\},$$

and Q denotes the empirical measure of x_j. Therefore, the same bound holds up to a constant for

$$\left\{\frac{\beta_j x - \beta_j^* x}{|\beta_j - \beta_j^*| + \frac{\gamma_p}{\lambda}|(J_p(\beta - \beta^*))_j|} : |\beta_j| + \frac{\gamma_p}{\lambda}|(J_p\beta)_j| \leq 1\right\},$$

for all $j = 1, \ldots, p$. Note that this function class is bounded in absolute value by $|x_{ij}| \leq R$ (by Assumption 6). Then, using Dudley's integral bound, that is,

$$A_0^{1/2} T_n^{1/2} \int_0^R \log^{1/2}\left(\frac{1}{u} + 1\right) du \leq \tilde{A}_0 T_n^{1/2},$$

by Corollary 8.3 of van de Geer [112], we have for all $\delta \geq 2C\tilde{A}_0\sqrt{\frac{T_n}{n}}$,

$$P\left(\sup_{\beta_j x \in \mathcal{F}}\left|\frac{n^{-1}\sum_{i=1}^{n}(Y_i - \mu_i)f(\beta_j x_{ij} - \beta_j^* x_{ij})}{|\beta_j - \beta_j^*| + \frac{\gamma_p}{\lambda}|(J_p(\beta - \beta^*))_j|}\right| \geq \delta\right) \leq C\exp\left(-\frac{n\delta^2}{4C^2R}\right). \tag{3.3}$$

Let $\delta = \rho = \kappa\sqrt{\frac{\log p}{n}}$. Then, $\rho \geq 2C\tilde{A}_0\sqrt{\frac{\log p}{n}} \geq 2C\tilde{A}_0\sqrt{\frac{T_n}{n}}$ since $T_n = 1$. Thus, applying

(3.3) with a union bound yields

$$
\begin{aligned}
&P\left(\max_{j=1,\ldots,p}\sup_{\beta_j x\in\mathcal{F}}\left|\frac{n^{-1}\sum_{i=1}^n(Y_i-\mu_i)(\beta_j x_{ij}-\beta_j^* x_{ij})}{|\beta_j-\beta_j^*|+\frac{\gamma_p}{\lambda}|(J_p(\beta-\beta^*))_j|}\right|\geq\rho\right)\\
&\leq pC\exp\left(-\frac{n\rho^2}{4C^2R}\right)\\
&=C\exp\left(-\frac{n\rho^2}{4C^2R}+\log p\right)\\
&=C\exp\left[-n\rho^2\left(\frac{1}{4C^2R}-\frac{\log p}{n\rho^2}\right)\right].
\end{aligned}
$$

Now,

$$
\frac{1}{4C^2R}-\frac{\log p}{n\rho^2}=\frac{1}{4C^2R}-\frac{1}{\kappa^2},
$$

is positive if $\kappa>\max(2C\sqrt{R},2C\tilde{A}_0)$.

Thus,

$$
P\left(\max_{j=1,\ldots,p}\sup_{\beta_j x\in\mathcal{F}}\left|\frac{n^{-1}\sum_{i=1}^n(Y_i-\mu_i)(\beta_j x_{ij}-\beta_j^* x_{ij})}{|\beta_j-\beta_j^*|+\frac{\gamma_p}{\lambda}|(J_p(\beta-\beta^*))_j|}\right|\geq\rho\right)\leq C\exp[-n\rho^2C_2]
$$

where $C=C(K,\sigma_0^2)$ and $C_2=C_2(C,R,\kappa)>0$.

Therefore, we have

$$
P\left(\left|n^{-1}\sum_{i=1}^n(Y_i-\mu_i)(\alpha_i-\alpha_i^*)\right|\geq\rho\|\alpha-\alpha^*\|_1\right)\leq 2\exp\left(-C_1n^2\rho^2\right),
$$

and

$$
P\left(\left|n^{-1}\sum_{i=1}^n\sum_{j=1}^p(\beta_j x_{ij}-\beta_j^* x_{ij})(Y_i-\mu_i)\right|\geq\rho\left[\|\beta-\beta^*\|_1+\frac{\gamma_p}{\lambda}\|J_p(\beta-\beta^*)\|_1\right]\right)\leq C\exp[-n\rho^2C_2],
$$

so the result follows. □

The next lemma establishes control of the empirical process term for cases where $Y_i-\mu_i$ is not sub-Gaussian (e.g. for Poisson or exponential GLMs). In such cases, we require part (ii) of Assumption 6 instead.

Lemma 7. *Under part (ii) of Assumption 6, with probability at least* $1 - 2\exp(-n\rho C_3) - 2\exp(-n\rho^2 C_4)$*, the following inequality holds:*

$$\nu_n(\theta) - \nu_n(\theta^*) \leq \rho \left[\|\alpha - \alpha^*\|_1 + \|\beta - \beta^*\|_1 + \frac{\gamma_p}{\lambda} \|J_p(\beta - \beta^*)\|_1 \right],$$

where $\rho = O\left(\sqrt{\frac{\log p}{n}}\right)$ *and* C_3, C_4 *are positive constants independent of* n *and* p*.*

Proof. As in Lemma 6, we can write

$$\begin{aligned}
\nu_n(\theta) - \nu_n(\theta^*) &= n^{-1} \sum_{i=1}^{n} (Y_i - \mu_i) \left[(\alpha_i - \alpha_i^*) + \sum_{j=1}^{p} (\beta_j x_{ij} - \beta_j^* x_{ij}) \right] \\
&= n^{-1} \sum_{i=1}^{n} (Y_i - \mu_i)(\alpha_i - \alpha_i^*) + n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{p} (\beta_j x_{ij} - \beta_j^* x_{ij})(Y_i - \mu_i).
\end{aligned}$$

We now want to bound the probability that $\nu_n(\theta) - \nu_n(\theta^*)$ exceeds

$$\rho \left[\|\alpha - \alpha^*\|_1 + \|\beta - \beta^*\|_1 + \frac{\gamma_p}{\lambda} \|J_p(\beta - \beta^*)\|_1 \right].$$

Consider the following probability involving the first term,

$$P\left(\left| \frac{n^{-1} \sum_{i=1}^{n} (Y_i - \mu_i)(\alpha_i - \alpha_i^*)}{\|\alpha - \alpha^*\|_1} \right| \geq \rho \right).$$

By Bernstein's inequality for sub-exponential random variables (Theorem 2.8.2 in [114]), we have

$$P\left(\left| \sum_{i=1}^{n} \gamma_i (Y_i - \mu_i) \right| \geq \rho \right) \leq 2\exp\left[-c \min\left(\frac{\rho^2}{K_{\psi_1}^2 \|\gamma\|_2^2}, \frac{\rho}{K_{\psi_1} \|\gamma\|_\infty} \right) \right],$$

for some constant c, where $K_{\psi_1} = \max_{i \in 1:n} \|Y_i - \mu_i\|_{\psi_1}$ and

$$\gamma_i = \frac{\alpha_i - \alpha_i^*}{n\|\alpha - \alpha^*\|_1}.$$

Noting that $\|\gamma\|_2^2 \leq n^{-2}$ and $\|\gamma\|_\infty \leq n^{-1}$,

$$P\left(\frac{n^{-1}\sum_{i=1}^n a(Y_i-\mu_i)(\alpha_i-\alpha_i^*)}{\|\alpha-\alpha^*\|_1} \geq \rho\right) \leq 2\exp\left[-c\min\left(\frac{\rho^2 n^2}{K_{\psi_1}^2}, \frac{\rho n}{K_{\psi_1}}\right)\right]$$
$$\leq 2\exp(-n\rho C_3),$$

where $C_3 > 0$.

Now, considering the second term in $\nu_n(\theta) - \nu_n(\theta^*)$, we want to bound, for a specific $j \in \{1\ldots,p\}$, the probability

$$P\left(\left|\frac{n^{-1}\sum_{i=1}^n (Y_i-\mu_i)(\beta_j x_{ij} - \beta_j^* x_{ij})}{|\beta_j-\beta_j^*| + \frac{\gamma_p}{\lambda}|(J_p(\beta-\beta^*))_j|}\right| \geq \rho\right). \tag{3.4}$$

Applying Bernstein's inequality again, with

$$\gamma_i = \frac{\beta_j x_{ij} - \beta_j^* x_{ij}}{n\left(|\beta_j-\beta_j^*| + \frac{\gamma_p}{\lambda}|(J_p(\beta-\beta^*))_j|\right)},$$

we can see that

$$\|\gamma\|_2^2 \leq \frac{\sum_{i=1}^n X_{ij}^2}{n^2} \leq \frac{R^2}{n},$$

and

$$\|\gamma\|_\infty \leq \frac{\|X_j\|_\infty}{n} \leq \frac{R}{n}.$$

Thus,

$$P\left(\left|\frac{n^{-1}\sum_{i=1}^n (Y_i-\mu_i)(\beta_j x_{ij} - \beta_j^* x_{ij})}{|\beta_j-\beta_j^*| + \frac{\gamma_p}{\lambda}|(J_p(\beta-\beta^*))_j|}\right| \geq \rho\right) \leq 2\exp\left[-c\min\left(\frac{\rho^2 n}{K_{\psi_1}^2 R^2}, \frac{\rho n}{K_{\psi_1} R}\right)\right].$$

Now, let $\rho = \kappa\sqrt{\frac{\log p}{n}}$, where

$$K_{\psi_1} R \leq \kappa \leq \frac{\sqrt{n} K_{\psi_1} R}{\sqrt{\log p}}$$

Therefore, $\rho \leq K_{\psi_1} R$, which implies $\dfrac{\rho}{K_{\psi_1} R} \leq 1$, or $\dfrac{\rho^2 n}{K_{\psi_1}^2 R^2} \leq \dfrac{\rho n}{K_{\psi_1} R}$.

Applying a union bound, we have

$$
\begin{aligned}
&P\left(\max_{j\in 1:p}\left|\frac{n^{-1}\sum_{i=1}^{n}(Y_i-\mu_i)(\beta_j x_{ij}-\beta_j^* x_{ij})}{|\beta_j-\beta_j^*|+\frac{\gamma_p}{\lambda}|(J_p(\beta-\beta^*))_j|}\right|\geq\rho\right)\\
&\leq p2\exp\left(-c\frac{\rho^2 n}{K_{\psi_1}^2R^2}\right)\\
&\leq 2\exp\left(-c\frac{\rho^2 n}{K_{\psi_1}^2R^2}+\log p\right)\\
&\leq 2\exp\left[-cn\rho^2\left(\frac{1}{K_{\psi_1}^2R^2}-\frac{\log p}{n\rho^2}\right)\right]\\
&\leq 2\exp\left[-cn\rho^2\left(\frac{1}{K_{\psi_1}^2R^2}-\frac{1}{\kappa^2}\right)\right]\\
&\leq 2\exp\left(-C_4 n\rho^2\right),
\end{aligned}
$$

where $C_4>0$ and the last inequality follows since $\frac{1}{K_{\psi_1}^2R^2}\geq\frac{1}{\kappa^2}$ by the lower bound on κ.

□

Margin condition

Before proving Theorem 2, we show that Assumption 10 implies a margin condition for the loss function $\mathcal{L}(\theta)=\ell(\alpha+X\beta)+\frac{1}{2}\gamma_n\alpha'(L_n+\delta I_n)\alpha$ around the target θ^*.

Lemma 8. *Under Assumption 10, the quadratic margin condition,*

$$\mathcal{E}(\theta)\geq G(\|\theta-\theta^*\|),$$

holds for all θ satisfying

$$\|\alpha-\alpha^*\|_1+\|\beta-\beta^*\|_1+\frac{\gamma_p}{\lambda}\|J_p(\beta-\beta^*)\|_1\leq\frac{16s\lambda^2}{\rho\phi^2(S)c}+\frac{2\gamma_p\|J_p\beta^*\|_1}{\rho},$$

where $\lambda\geq 8\rho$.

Proof. From Lemma 5, we have that $\|\alpha - \alpha^*\|_2 = O_p(\sqrt{\frac{\log p}{n}})$. Therefore, $\|\theta^* - \theta^0\|_2 = O_p(\sqrt{\frac{\log p}{n}})$, and so we have that $\mathbb{P}\ell(\theta)$ satisfies strong convexity for all θ in the local neighbourhood of θ^* defined in Assumption 10. Then, rewriting

$$\frac{1}{2}\gamma_n\alpha'(L_n + \delta I_n)\alpha = \frac{1}{2}\gamma_n\theta'\tilde{L}\theta,$$

we can see that $\nabla^2\frac{1}{2}\gamma_n\alpha'(L_n + \delta I_n)\alpha = \tilde{L}$ is positive semi-definite. Therefore,

$$\mathbb{P}\mathcal{L}(\theta) = \mathbb{P}\ell(\theta) + \frac{1}{2}\gamma_n\alpha'(L_n + \delta I_n)\alpha$$

satisfies restricted strong convexity for all θ in the local neighbourhood around θ^*. Hence,

$$\mathbb{P}(\mathcal{L}(\theta) - \mathcal{L}(\theta^*)) \geq \nabla\mathbb{P}\mathcal{L}(\theta^*)'(\theta - \theta^*) + G(\|\theta - \theta^*\|).$$

Since θ^* minimizes $\mathbb{P}\mathcal{L}(\theta)$, we have

$$\mathbb{P}(\mathcal{L}(\theta) - \mathcal{L}(\theta^*)) = \mathcal{E}(\theta) \geq G(\|\theta - \theta^*\|).$$

□

Proof of Theorem 2

As a result of Lemma 6 or 7, we have with high probability that

$$\begin{aligned}
Z_{M^*} &:= \sup_{\|\alpha-\alpha^*\|_1+\mathcal{R}(\theta-\theta^*)\leq M^*} |\nu_n(\theta) - \nu_n(\theta^*)| \\
&\leq \sup_{\|\alpha-\alpha^*\|_1+\mathcal{R}(\theta-\theta^*)\leq M^*} \rho[\|\alpha - \alpha^*\|_1 + \|\beta - \beta^*\|_1 + \frac{\gamma_p}{\lambda}\|J_p(\beta - \beta^*)\|_1] \\
&= \sup_{\|\alpha-\alpha^*\|_1+\mathcal{R}(\theta-\theta^*)\leq M^*} \rho[\|\alpha - \alpha^*\|_1 + \mathcal{R}(\theta - \theta^*)] \\
&\leq \rho M^*
\end{aligned}$$

Set $t = \dfrac{M^*}{M^* + \|\hat{\alpha} - \alpha^*\|_1 + \mathcal{R}(\hat{\theta} - \theta^*)}$ and take $\tilde{\theta} := t\hat{\theta} + (1-t)\theta^*$. Then,

$$\|\tilde{\alpha} - \alpha^*\|_1 + \mathcal{R}(\tilde{\theta} - \theta^*) \leq M^*$$

by construction.

Note that we can write

$$\|\hat{\alpha} - \alpha^*\|_1 + \mathcal{R}(\hat{\theta} - \theta^*) = \frac{1}{t}\left[\|\tilde{\alpha} - \alpha^*\|_1 + \mathcal{R}(\tilde{\theta} - \theta^*)\right], \tag{3.5}$$

which we will use later to bound $\|\hat{\alpha} - \alpha^*\|_1 + \mathcal{R}(\hat{\theta} - \theta^*)$.

Then, starting from the basic inequality:

$$\begin{aligned}
\mathcal{E}(\tilde{\theta}) + \lambda\mathcal{R}(\tilde{\theta}) &\leq -[\nu_n(\tilde{\theta}) - \nu_n(\theta^*)] + \lambda\mathcal{R}(\theta^*) \\
&\leq Z_{M^*} + \lambda\mathcal{R}(\theta^*) \\
&\leq \rho M^* + \lambda\mathcal{R}(\theta^*).
\end{aligned}$$

Now,

$$\begin{aligned}
\lambda\mathcal{R}(\theta^*) &= \lambda\left[\|\beta_S^*\|_1 + \frac{\gamma_p}{\lambda}\|J_p\beta^*\|_1\right] \\
&\leq \lambda\left[\|\beta_S^* - \tilde{\beta}_S\|_1 + \|\tilde{\beta}_S\|_1 + \frac{\gamma_p}{\lambda}\|J_p\beta^*\|_1\right]
\end{aligned}$$

and

$$\begin{aligned}
\lambda\mathcal{R}(\tilde{\theta}) &= \lambda\left[\|\tilde{\beta}_S\|_1 + \|\tilde{\beta}_{S^c}\|_1 + \frac{\gamma_p}{\lambda}\|J_p\tilde{\beta}\|_1\right] \\
&\geq \lambda\left[\|\tilde{\beta}_S\|_1 + \|(\tilde{\beta} - \beta^*)_{S^c}\|_1 + \frac{\gamma_p}{\lambda}\|J_p(\tilde{\beta} - \beta^*)\|_1 - \frac{\gamma_p}{\lambda}\|J_p\beta^*\|_1\right].
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathcal{E}(\tilde{\theta}) + \lambda\left[\|\tilde{\beta}_S\|_1 + \|(\tilde{\beta} - \beta^*)_{S^c}\|_1 + \frac{\gamma_p}{\lambda}\|J_p(\tilde{\beta} - \beta^*)\|_1 - \frac{\gamma_p}{\lambda}\|J_p\beta^*\|_1\right] \\
&\leq \lambda\left[\|\beta_S^* - \tilde{\beta}_S\|_1 + \|\tilde{\beta}_S\|_1 + \frac{\gamma_p}{\lambda}\|J_p\beta^*\|_1\right] + \rho M^*.
\end{aligned}$$

Rearranging yields:

$$\mathcal{E}(\tilde{\theta}) + \lambda\left[\|(\tilde{\beta} - \beta^*)_{S^c}\|_1 + \frac{\gamma_p}{\lambda}\|J_p(\tilde{\beta} - \beta^*)\|_1\right] \leq 2\lambda\|(\tilde{\beta} - \beta^*)_S\|_1 + 2\gamma_p\|J_p\beta^*\|_1 + \rho M^*.$$

Adding $\lambda\left[\|\tilde{\alpha} - \alpha^*\|_1 + \|(\tilde{\beta} - \beta^*)_S\|_1\right]$ to both sides:

$$\mathcal{E}(\tilde{\theta}) + \lambda[\|\tilde{\alpha} - \alpha^*\|_1 + \mathcal{R}(\tilde{\theta} - \theta^*)] \leq 2\lambda\|(\tilde{\beta} - \beta^*)_S\|_1 + \lambda\|\tilde{\alpha} - \alpha^*\|_1 + 2\gamma_p\|J_p\beta^*\|_1 + \rho M^*. \tag{3.6}$$

Now, we have two possible cases for the RHS of (3.6).

Case I

First, consider the case when $2\lambda\|(\tilde{\beta}-\beta^*)_S\|_1 + \lambda\|\tilde{\alpha}-\alpha^*\|_1 \leq 2\gamma_p\|J_p\beta^*\|_1 + \rho M^*$. Then:

$$\begin{aligned}
\mathcal{E}(\tilde{\theta}) + \lambda[\|\tilde{\alpha}-\alpha^*\|_1 + \mathcal{R}(\tilde{\theta}-\theta^*)] &\leq 4\gamma_p\|J_p\beta^*\|_1 + 2\rho M^* \\
&\leq 2\rho M^* + 2\rho M^* \quad \text{(by definition of } M^*) \\
&= 4\rho M^* \\
&\leq 4\frac{\lambda}{8}M^* \\
&= \lambda\frac{M^*}{2}
\end{aligned}$$

Therefore, since $\mathcal{E}(\tilde{\theta}) \geq 0$,

$$\|\tilde{\alpha}-\alpha^*\|_1 + \mathcal{R}(\tilde{\theta}-\theta^*) \leq \frac{M^*}{2}.$$

and by (3.5),

$$\begin{aligned}
\|\hat{\alpha}-\alpha^*\|_1 + \mathcal{R}(\hat{\theta}-\theta^*) &= \frac{1}{t}\left[\|\tilde{\alpha}-\alpha^*\|_1 + \mathcal{R}(\tilde{\theta}-\theta^*)\right] \\
&\leq \left[1 + \frac{\|\hat{\alpha}-\alpha^*\|_1 + \mathcal{R}(\hat{\theta}-\theta^*)}{M^*}\right]\frac{M^*}{2} \\
&= \frac{M^*}{2} + \frac{\|\hat{\alpha}-\alpha^*\|_1 + \mathcal{R}(\hat{\theta}-\theta^*)}{2}.
\end{aligned}$$

Hence, we can show

$$\|\hat{\alpha}-\alpha^*\|_1 + \mathcal{R}(\hat{\theta}-\theta^*) \leq M^*.$$

As a result, we can redo the above arguments replacing $\tilde{\alpha}, \tilde{\beta}$ with $\hat{\alpha}, \hat{\beta}$.

Case II

Next, we consider the case of $2\lambda\|(\tilde{\beta}-\beta^*)_S\|_1 + \lambda\|\tilde{\alpha}-\alpha^*\|_1 \geq 2\gamma_p\|J_p\beta^*\|_1 + \rho M^*$, and show the same result.

We can bound the RHS of (3.6) as:

$$\mathcal{E}(\tilde{\theta}) + \lambda \left[\|\tilde{\alpha} - \alpha^*\|_1 + \mathcal{R}(\tilde{\beta} - \beta^*) \right] \leq 4\lambda \|(\tilde{\beta} - \beta^*)_S\|_1 + 2\lambda \|\tilde{\alpha} - \alpha^*\|_1. \tag{3.7}$$

Then,

$$\mathcal{E}(\tilde{\theta}) + \lambda \|\tilde{\beta} - \beta^*\|_1 + \gamma_p \|J_p(\tilde{\beta} - \beta^*)\|_1 - \lambda \|(\tilde{\beta} - \beta^*)_S\|_1 + \leq 3\lambda \|(\tilde{\beta} - \beta^*)_S\|_1 + \lambda \|\tilde{\alpha} - \alpha^*\|_1$$

$$\mathcal{E}(\tilde{\theta}) + \lambda \|(\tilde{\beta} - \beta^*)_{S^c}\|_1 + \gamma_p \|J_p(\tilde{\beta} - \beta^*)\|_1 \leq 3\lambda \|(\tilde{\beta} - \beta^*)_S\|_1 + \lambda \|\tilde{\alpha} - \alpha^*\|_1$$

$$\lambda \|(\tilde{\beta} - \beta^*)_{S^c}\|_1 + \gamma_p \|J_p(\tilde{\beta} - \beta^*)\|_1 \leq 3\lambda \|(\tilde{\beta} - \beta^*)_S\|_1 + \lambda \|\tilde{\alpha} - \alpha^*\|_1$$

$$\|(\tilde{\beta} - \beta^*)_{S^c}\|_1 + \frac{\gamma_p}{\lambda} \|J_p(\tilde{\beta} - \beta^*)\|_1 \leq 3\|(\tilde{\beta} - \beta^*)_S\|_1 + \|\tilde{\alpha} - \alpha^*\|_1$$

Therefore, the condition in Assumption 9 is satisfied for $(\tilde{\alpha} - \alpha^*, \tilde{\beta} - \beta^*)$, so we can use the compatibility condition:

$$\frac{\|\tilde{\alpha} - \alpha^*\|_1}{2} + \|(\tilde{\beta} - \beta^*)_S\|_1 \leq \frac{\|\tilde{\theta} - \theta^*\|\sqrt{s}}{\phi(s)}$$

Plugging this into (3.7), and denoting the convex conjugate of G by H, we have:

$$\begin{aligned}
\mathcal{E}(\tilde{\theta}) + \lambda[\|\tilde{\alpha} - \alpha^*\|_1 + \mathcal{R}(\tilde{\theta} - \theta^*)] &\leq 4\lambda \frac{\|\tilde{\theta} - \theta^*\|\sqrt{s}}{\phi(s)} \\
&\leq H\left(\frac{4\lambda\sqrt{s}}{\phi(s)}\right) + G(\|\tilde{\theta} - \theta^*\|) \\
&\leq H\left(\frac{4\lambda\sqrt{s}}{\phi(s)}\right) + \mathcal{E}(\tilde{\theta}) \\
&= \frac{16s\lambda^2}{4c\phi^2(s)} + \mathcal{E}(\tilde{\theta}) \\
&\leq \rho M^* + \mathcal{E}(\tilde{\theta}) \\
&\leq \frac{\lambda M^*}{8} + \mathcal{E}(\tilde{\theta})
\end{aligned}$$

Finally:

$$\|\tilde{\alpha} - \alpha^*\|_1 + \mathcal{R}(\tilde{\beta} - \beta^*) \leq \frac{M^*}{8} \leq \frac{M^*}{2}$$

Hence, as in Case I, we can show

$$\|\hat{\alpha} - \alpha^*\|_1 + \mathcal{R}(\hat{\beta} - \beta^*) \leq M^*.$$

As a result, we can redo the above arguments replacing $\tilde{\alpha}, \tilde{\beta}$ with $\hat{\alpha}, \hat{\beta}$.

Therefore, we can obtain, in both cases::

$$\begin{aligned}\mathcal{E}(\hat{\beta}) + \lambda[\|\hat{\alpha} - \alpha^*\|_1 + \mathcal{R}(\hat{\beta} - \beta^*)] &\leq 4\rho M^* \\ &\leq \frac{64 s\lambda^2}{c\phi^2(s)} + 8\gamma_p \|J_p \beta^*\|_1.\end{aligned}$$

Thus, since $\mathcal{E}(\tilde{\theta}) \geq 0$,

$$\|\hat{\alpha} - \alpha^*\|_1 + \|\hat{\beta} - \beta^*\|_1 = O\left(\lambda + \frac{\gamma_p}{\lambda}\|J_p \beta^*\|_1\right),$$

with probability $1 - 2\exp(-n^2\rho^2 C_1) - C\exp(-n\rho^2 C_2)$ by Lemma 6 or 7.

Taking $\lambda = O_p\left(\sqrt{\frac{\log p}{n}}\right)$ and $\gamma_p \|J_p \beta^*\|_1 = o_p(\lambda)$, we have that $\hat{\alpha}$ and $\hat{\theta}$ are ℓ_1-consistent for α^* and β^*.

Proof of Theorem 3

Next, we prove Theorem 3, which shows the validity of our debiasing inference procedure. This proof is very similar to that of Theorem 3.1 in van de Geer et al. [113]. For ease of exposition, we assume the case of GLM families with known scale parameter $\phi = 1$. However, these results trivially extend to the case with finite known ϕ, or finite unknown ϕ with a consistent estimator $\hat{\phi}$.

Recall that we use the approach of Javanmard and Montanari for GLMs, defining the debiased estimator:

$$\hat{b} := \hat{\beta} - M\frac{1}{n}\nabla_\beta \ell(\hat{\alpha} + X\hat{\beta}),$$

where $\hat{\Theta}$ is an estimator of the inverse of $\hat{\Sigma} := \frac{1}{n}\nabla^2_\beta \ell(\hat{\alpha} + X\hat{\beta})$.

Taking a first-order Taylor expansion of the gradient at $\hat{\alpha} + X\hat{\beta}$ around $\hat{\alpha} + X\beta^0$, we have:

$$\nabla_\beta \ell(\hat{\alpha} + X\hat{\beta}) = X'(\mu(\hat{\alpha} + X\beta^0) - y) + X'W(\tilde{q})(X\hat{\beta} - X\beta^0)$$

where W is a diagonal matrix of $\frac{\partial \mu}{\partial(\alpha + X_i\beta)}$, and $\tilde{q}$ is an intermediate point in between $\hat{\alpha} + X\hat{\beta}$ and $\hat{\alpha} + X\beta^0$. Then,

$$\begin{aligned}\nabla_\beta \ell(\hat{\alpha} + X\hat{\beta}) = X'(\mu(\hat{\alpha} + X\beta^0) - y) + X'W(\hat{\alpha} + X\hat{\beta})X(\hat{\beta} - \beta^0) + X'W(\tilde{q})X(\hat{\beta} - \beta^0) \\ - X'W(\hat{\alpha} + X\hat{\beta})X(\hat{\beta} - \beta^0).\end{aligned}$$

Defining Rem_1 as $X'W(\tilde{q})X(\hat{\beta} - \beta^0) - X'W(\hat{\alpha} + X\hat{\beta})X(\hat{\beta} - \beta^0)$, we consider its ℓ_2 norm:

$$\begin{aligned}\|Rem_1\|_2 &= \|X'W(\tilde{q})X(\hat{\beta} - \beta^0) - X'W(\hat{\alpha} + X\hat{\beta})X(\hat{\beta} - \beta^0)\|_2 \\ &\leq \|X'W(\tilde{q}) - X'W(\hat{\alpha} + X\hat{\beta})\|_2 \|X(\hat{\beta} - \beta^0)\|_2 \\ &\leq \|X\|_2 \|W(\tilde{q}) - W(\hat{\alpha} + X\hat{\beta})\|_2 \|X(\hat{\beta} - \beta^0)\|_2 \\ &\leq \|X\|_2 L_{\mu'} \|\tilde{q} - \hat{\alpha} - X\hat{\beta}\|_2 \|X(\hat{\beta} - \beta^0)\|_2 \quad \text{(by Lipschitz condition)} \\ &\leq L_{\mu'} \|X\|_2 \|X(\hat{\beta} - \beta)\|_2 \|X(\hat{\beta} - \beta^0)\|_2 \quad \text{(by definition of } \tilde{q}) \\ &\leq L_{\mu'} \|X\|_2^2 \|\hat{\beta} - \beta^0\|_2^2\end{aligned}$$

Then, we have:

$$\begin{aligned}\hat{b} - \beta^0 &= (\hat{\beta} - \beta^0) - M\frac{1}{n}\nabla_\beta \ell(\hat{\alpha} + X\hat{\beta}) \\ &= (\hat{\beta} - \beta^0) - M\frac{1}{n}X'(\mu(\hat{\alpha} + X\beta^0) - y) - M\frac{1}{n}X'W(\hat{\alpha} + X\hat{\beta})X(\hat{\beta} - \beta^0) - M\frac{1}{n}Rem_1 \\ &= -\frac{1}{n}MX'(\mu(\hat{\alpha} + X\beta^0) - y) - [M\hat{\Sigma} - I](\hat{\beta} - \beta^0) - M\frac{1}{n}Rem_1 \\ &= \frac{1}{n}MZ_n - Rem_2 - M\frac{1}{n}Rem_1\end{aligned}$$

where $Rem_2 := [M\hat{\Sigma} - I](\hat{\beta} - \beta^0)$ and $Z_n := X'(y - \mu(\hat{\alpha} + X\beta^0))$.

Now, by Theorem 2 and Lemma 4, we have that $\|Mn^{-1}Rem_1\|_2 = o_p(1)$. We also have that $\|Rem_2\|_2 = o_p(1)$ by Theorem 2, Lemma 4, and construction of M.

Also, we can write

$$\begin{aligned} Z_n &= X' \left(y - \mu(\hat{\alpha} + X\beta^0) + \mu(\alpha^0 + X\beta^0) - \mu(\alpha^0 + X\beta^0)\right) \\ &= X' \left(y - \mu(\alpha^0 + X\beta^0)\right) + X'(\mu(\hat{\alpha} + X\beta^0) - \mu(\alpha^0 + X\beta^0)\big) \\ &= X' \left(y - \mu(\alpha^0 + X\beta^0)\right) + Rem_3 \end{aligned}$$

where $Rem_3 := X' \left(\mu(\hat{\alpha} + X\beta^0) - \mu(\alpha^0 + X\beta^0)\right)$.

Then,

$$\begin{aligned} \|Rem_3\|_2 &\leq \|X\|_2 \|\mu(\hat{\alpha} + X\beta^0) - \mu(\alpha^0 + X\beta^0)\|_2 \\ &\leq \|X\|_2 L_\mu \|\hat{\alpha} - \alpha^0\|_2 \\ &\leq \|X\|_2 L_\mu \left(\|\hat{\alpha} - \alpha^*\|_2 + \|\alpha^* - \alpha^0\|_2\right) \\ &\leq \|X\|_2 L_\mu \left(\|\hat{\alpha} - \alpha^*\|_1 + \|\alpha^* - \alpha^0\|_2\right) \\ &= o_p\left(\frac{n^{c_1}}{\sqrt{\log p}}\right), \end{aligned}$$

where $c_1 < 1$, by Lemma 5 and Theorem 2.

Hence, $n^{-1} M Rem_3 = o_p(1)$, and

$$\sqrt{n}(\hat{b} - \beta^0) \longrightarrow_d N\left(0, M\frac{1}{n}E[\nabla\ell(\alpha^0 + X\beta^0)\nabla\ell(\alpha^0 + X\beta^0)']M\right) + o_p(1)$$

3.9 ℓ_2 *feature network smoothing*

In this section, we briefly discuss how the theory given previously can be applied to the `glm-funk` model with ℓ_2 feature network smoothing. Because the generalized ridge penalty $\beta' J_p \beta$ does not represent a norm, it is difficult to work with as part of the regularizer term. Therefore, we consider it part of the loss function instead, and write the objective function as

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \left\{ \mathbb{P}_n \mathcal{L}_i(\theta) + \lambda \mathcal{R}(\theta) \right\}$$

where $\mathcal{L}_i(\theta) = \ell_i(\alpha_i + x_i'\beta) + \frac{1}{2}\gamma_n \alpha'(L_n + \delta I_n)\alpha + \frac{1}{2}\gamma_p \beta' L_p \beta$ and $\mathcal{R}(\theta) = \|\beta\|_1$.

Then, since the loss function $\mathbb{P}_n \mathcal{L}_i$ is convex and differentiable in θ, we can apply a similar proof as in Theorem 2 to show $\hat{\theta} \to \theta^*$, where $\theta^* = \operatorname{argmin}_\theta \{\mathbb{P}\mathcal{L}(\theta)\}$. In order to show that $\|\theta^* - \theta^0\|$ is negligible, we would need to make a stronger assumption on the target parameters. That is, if $\gamma_p \|L_p \beta^*\|_2$ and $\gamma_n \|(L_n + \delta I_n)\alpha^*\|_2$ are $o_p(1)$, we can conclude that the target parameter θ^* asymptotically tends to the true parameter θ^0. From here, the validity of our inference procedure given in Theorem 3 would follow.

3.10 Equivalence of RNC and linear mixed models

In this section, we briefly discuss the equivalence of the RNC estimator to a linear mixed effects model. For simplicity, we consider a low-dimensional setting here, where $n < p$, and do not incorporate any feature network information.

We assume a linear model,

$$\begin{aligned} Y &= \alpha + X\beta + \epsilon \\ \alpha &\sim N(0, \phi(L_n + \delta I)^{-1}) \\ \epsilon &\sim N(0, \sigma^2 I), \end{aligned}$$

where $\alpha \perp\!\!\!\perp \epsilon$. We first consider *conditional* estimation of β. It is easy to see that $Y|\alpha \sim N(\alpha + X\beta, \sigma^2 I)$. Then, maximizing the log-likelihood $\log \pi(y|\alpha) + \log \pi(\alpha)$ is equivalent to minimizing the objective function

$$\frac{1}{\sigma^2}(y - \alpha - X\beta)'(y - \alpha - X\beta) + \frac{1}{\phi}\alpha'(L_n + \delta I)\alpha.$$

With known σ and setting $\phi = \gamma_n^{-1}$, we obtain the RNC objective function

$$(y - \alpha - X\beta)'(y - \alpha - X\beta) + \gamma_n \alpha'(L_n + \delta I)\alpha.$$

This relationship holds for other generalized linear models. Therefore, we can interpret RNC as estimating conditional associations between y and X, given correlation induced through random intercepts α.

In the linear model case, we can also use the equivalence of marginal and conditional models due to the additivity of the random effects, and α having mean zero [87]. With known variance, maximizing the likelihood $\pi(y)$ directly is equivalent to minimizing

$$(y - X\beta)'[\phi(L_n + \delta I)^{-1} + \sigma^2 I]^{-1}(y - X\beta).$$

Hence, RNC can also be interpreted as marginal estimation of β in a mixed model using a generalized least squares estimator.

Chapter 4

BALANCING OFF-POLICY EVALUATION IN GENERAL ACTION SPACES

4.1 Introduction

In contextual bandit problems, algorithms make decisions about actions to take under uncertainty, with the goal of optimizing some reward. This is done through implementing a *policy*, which chooses actions based on observed states [63]. Applications abound in medicine, where personalized treatments are designed based on known patient history [107], and internet marketing, where advertisements can be tailored to user interests [65]. Learning an optimal policy may be prohibitively expensive, and experimenting with an untested policy could result in unacceptably negative results, such as patient death or user churn [1]. Therefore, an important problem in this area is *counterfactual* or *off-policy* policy evaluation, where the value of the policy of interest is estimated based on observed historical data. This problem is even more important when attempting to safely deploy a policy for an application that previously used ad-hoc or difficult-to-enumerate rules.

Typical approaches to off-policy policy evaluation either use a regression model to predict the counterfactual rewards, importance sampling to reweight the observed reward data, or a combination of the two [24, 108, 118]. Because regression models can give biased results in off-policy settings, current methods usually incorporate importance sampling. However, their primary focus is on settings with discrete or parametric action spaces. While these methods can extend to arbitrary continuous action spaces, doing so requires true knowledge or at least a good estimate of the importance sampling weights, which are ratios of policy densities [49]. In the existing literature, these ratios are typically assumed to be known

[1] This approach also carries a non-trivial probability of prison time.

exactly, which is unlikely to hold in practice, particularly if a policy is drawn from an unknown continuous density. This would likely be a problem, for example, with a policy that delivers personalized user advertisements based on a black-box machine learning system. The resulting policy densities may be difficult or impossible to estimate, particularly in high dimensions. Methods based on kernel-based rejection sampling have also been proposed as a possible solution [56], but these still require true knowledge or good estimates of the observed policy density.

In this chapter, we develop a new counterfactual policy evaluation method for contextual bandit problems with arbitrary action spaces. Our proposed method, which we call *balancing off-policy evaluation* (BOP-e), does not require true knowledge or an estimator of either policy density. BOP-e is an importance sampler which can be directly plugged into existing methods instead of inverse propensity scores [56, 118, 24, 28]. A probabilistic classifier is trained on state-action data from both policies, and is used to directly estimate the density ratio. Hence, the method only requires logged data on states, and the actions which would be taken by both observed and target policies at those states. In contrast to prior work on balancing weights which focuses on minimizing pre-specified statistical distances between the proposed and observed policy, e.g. the maximum mean discrepancy [55], BOP-e is defined more generally with respect to a Bregman divergence [8] implied by choice of classification loss. We show how BOP-e explicitly optimizes balance along the same lines. We also show that the loss of the classification problem bounds the bias and variance which allows practitioners to discriminate amongst losses by using standard model selection methodology from the supervised learning literature.

The rest of the chapter is structured as follows. We provide an overview of OPE in Section 4.2. We then describe BOP-e in Section 4.3, and explain how classifier probabilities can be used to directly obtain importance sampling weights. In Section 4.4, we provide a theoretical analysis of our estimators and prove consistency for the counterfactual policy value. We summarize and discuss related work in Section 4.5. In Section 4.6, we evaluate our estimators in numerical experiments, considering both discrete and continuous action

spaces. The latter experiments provide a novel extension of the "classifier trick" of Dudík et al. [23] to continuous action spaces.

4.2 Problem description and background

We will assume a contextual bandit setup, where our data consists of n independent observations of (s_i, a_i, r_i). For each unit, a *state* s_i is observed, an *action* a_i is taken in accordance with some *policy* π, and a *reward* r_i is observed in response. We use the notation π to refer to both a policy and its density, and use $\pi(s)$ to denote the action that would be taken under policy π for a state s.

The problem addressed is as follows: given a proposed policy π_1 and observed data (s, a, r) collected following a policy π_0, estimate the expected reward of instead following π_1 on the observed states. We denote the reward function as $r(a, s)$, and an estimated reward function as $\hat{r}(a, s)$.

We assume the following throughout:

Assumption 13. $\pi_1(a, s) > 0 \iff \pi_0(a, s) > 0 \; \forall a \in A, s \in S$

Assumption 14. $0 < \frac{\pi_1(a,s)}{\pi_0(a,s)} \leq \alpha_1 < \infty$

Assumption 15. $0 \leq r(a, s) \leq \alpha_2 < \infty$, $\forall s, a \in S \times A$

Assumption 16. *The distribution of rewards across potential actions is independent of policy, conditional on state.*

4.2.1 Off-policy estimation

We now briefly review the three broad classes of off-policy estimation: direct modeling, importance sampling, and doubly robust estimation. Throughout this section we assume that (s_i, a_i, r_i) are data collected under observed policy π_0, and a_i' is an action that would be taken under the proposed policy π_1.

The *direct method* approach to this problem fits a regression model $\hat{r}(a, s)$ to approximate the reward function $r(a, s)$ under the observed policy π_0. The counterfactual policy value, $V_{\pi_1} := \mathbb{E}_{\pi_1}[r]$, is estimated by predicting the rewards that would have been observed under the actions of policy π_1, i.e.

$$\hat{V}^{DM} = \frac{1}{n} \sum_{i=1}^{n} \hat{r}(s_i, a'_i)$$

In order for the resulting estimate to be consistent, the reward model $\hat{r}$ needs to generalize well to the reward distribution that would be observed under policy π_1. In practice, this method give badly biased results if the observed state-action data is very different from the counterfactual distribution [23].

Importance sampling is another approach which reweights the observed rewards by an inverse propensity score (IPS), and a rejection sampling term, i.e.

$$\hat{V}^{IPS} = \frac{1}{n} \sum_{i=1}^{n} r_i \frac{\mathbb{1}_a(a'_i)}{\hat{\pi}_0(a_i|s_i)}$$

Importance sampling, while unbiased for V_{π_1}, often suffers from high variance. The *weighted importance sampling estimator* (also called the "self-normalized" or Hájek estimator) has been used to reduce variance, at the cost of small bias, while maintaining consistency [105, 19]:

$$\hat{V}^{WIS} = \frac{\sum_{i=1}^{n} r_i \frac{\mathbb{1}_a(a'_i)}{\hat{\pi}_0(a_i|s_i)}}{\sum_{i=1}^{n} \frac{\mathbb{1}_a(a'_i)}{\hat{\pi}_0(a_i|s_i)}}$$

For continuous action spaces, Kallus and Zhou [56] recently proposed an IPS-based method that replaces the indicator function $\mathbb{1}_a(\cdot)$ with a kernel smoothing term K, i.e.,

$$V^{KIS} = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{a'_i - a_i}{h}\right) \frac{r_i}{\hat{\pi}_0(a_i|s_i)}$$

. The corresponding weighted importance sampling estimator is defined analogously.

Finally, *doubly robust* estimators combine the direct method and importance sampling. These tend to have lower variance, and are consistent if either the direct method regression

model or the importance sampling weights are correctly specified [24, 108]. For discrete or continuous action spaces, the reward is estimated as [24]

$$\hat{V}^{DR} = \frac{1}{n}\sum_{i=1}^{n}\left(r_i - \hat{r}(s_i, a_i)\right)\frac{J(a_i, a_i'))}{\hat{\pi}_0(a_i|s_i)} + \hat{r}(s_i, a_i'),$$

where $J(a_i, a_i')$ is a suitable rejection sampling term. The SWITCH estimator of Wang et al. [118] adaptively, over all observations, uses IPS unless the weight is too large, in which case it uses the direct method.

4.3 Balanced importance sampling

There are several weaknesses with existing approaches that leverage importance sampling with inverse propensity scores. First, the probability of some observed actions for some observed states may be very close to 0 to 1, leading to instability and small sample bias of the propensity score model [72]. Second, the propensity score model must be correctly specified. In the absence of this, prior work has shown that the performance of IPS can be arbitrarily bad, because there is no guarantee of *balance* under a misspecified propensity score [59, 97, 48]. In particular, a misspecified IPS will not, even in large samples, ensure that the weighted state-action distribution of the observed policy will match that of the proposed policy [48, 37]. This implies that policy evaluation will be incorrect, as it reflects the performance of a policy *on the wrong state distribution*.

Using doubly robust estimation partially addresses the case of a misspecified propensity model. However, while they provide consistent estimates when either the direct method or the propensity score model is unbiased, they do not protect against failure of both. To address this weakness, recent work has focused on weighting estimators that explicitly seek to optimize for balance, seeking weighting functions that make the choice of action independent from the observed contexts [69, 55]. These estimators have been shown to provide strong results in their respective applications even under misspecification. However, their use is limited to discrete action spaces, and often involve hyperparameters that need to be set

by heuristics, or are computationally intractable to tune. To remedy this, we now describe BOP-e which defines a class of balancing importance samplers for off-policy evaluation.

BOP-e leverages classifier-based density ratio estimation [101, 74] to learn importance sampling ratios. Specifically, off policy evaluation using BOP-e consists of four steps:

1. Create a supervised learning problem using the concatenated proposed policy instances (s, a') and observed policy instances (s, a), as covariates and giving a label (C) of 0 to the observed policy and 1 to the proposed policy.

2. Learn a classifier to distinguish between the observed and proposed policy.

3. Take the importance sampling ratio as

$$\hat{\rho}(a_i, s_i) = \frac{\hat{p}(C = 1|a_i, s_i)}{\hat{p}(C = 0|a_i, s_i)}$$

4. Take the off policy estimate as

$$\hat{V}^{BOP-e} = \frac{\sum_{i=1}^{n} J(a_i, a_i')\hat{\rho}(a_i, s_i)r_i}{\sum_{i=1}^{n} J(a_i, a_i')\hat{\rho}(a_i, s_i)}$$

where J defines a rejection sampler term between the observed action a_i and the proposed action a_i'. For discrete action spaces, this is simply $\mathbb{1}_a(a_i')$. For continuous actions, we use the kernel term of Kallus and Zhou [56], that is $J(a_i, a_i') = \frac{1}{h}K(\frac{a_i' - a_i}{h})$, where K is some kernel function, and h is a bandwidth parameter. A corresponding doubly robust estimator can also be constructed.

We can see how step three arrives at the importance sampler through an application of Bayes rule [7]:

$$\begin{aligned}
\frac{P(C = 1|a, s)}{P(C = 0|a, s)} &= \frac{\frac{\pi(a,s|C=1)P(C=1)}{\pi(a,s|C=1)P(C=1)+\pi(a,s|C=0)P(C=0)}}{\frac{\pi(a,s|C=0)P(y=0)}{\pi(a,s|C=1)P(C=1)+\pi(a,s|C=0)P(C=0)}} \\
&= \frac{\pi(a, s|C = 1)P(C = 1)}{\pi(a, s|C = 0)P(C = 0)} \\
&= \frac{\pi_1(a, s)}{\pi_0(a, s)}
\end{aligned}$$

where $\frac{P(C=1)}{P(C=0)} = 1$ by design.

As described, this procedure provides a large degree of flexibility to practitioners, requiring only that a classification model be learned. The question as to which classifiers fit within this framework is given by the additional assumption.

Assumption 17. *The classifier is trained using a strictly proper composite loss*[2]*, ℓ, with a twice differentiable Bayes risk, f.*

This assumption allows for a large number of widely used loss functions, such as logistic, exponential, and mean squared error, as well as models commonly used for distribution comparison such as the kernel based density ratio estimators of Sugiyama et al. [101], and maximum mean discrepancy [55].

Given that BOP-e targets the policy density ratio, it optimizes a measure of balance as described in the following proposition, which is the difference between the reweighted source and target distributions:

Proposition 1. *Let ϕ and ψ be real-valued functions of a and s, respectively. The L_1 functional discrepancy between the observed policy π_0 and the proposed policy π_1 under BOP-e is given by* $\|\mathbb{E}_{\pi_0}[\phi(a) \otimes \psi(s)\hat{\rho}(a,s)] - \mathbb{E}_{\pi_1}[\phi(a) \otimes \psi(s)]\|_1 = \|\mathbb{E}_{\pi_0}[\phi(a) \otimes \psi(s)(\hat{\rho}(a,s) - \rho(a,s))]\|_1 \leq \|\mathbb{E}_{\pi_0}[\phi(a) \otimes \psi(s)B(\hat{\rho}, \rho)]\|_1$ *where B is a Bregman divergence.*

When $\hat{\rho} = \rho$, trivially reduces this discrepancy to 0. Thus, the degree to which balance is attained is implied by the quality of the approximation of $\hat{\rho}$ to ρ. The upper bound involves a Bregman divergence B which depends on the classifier $\hat{p}$ used. We discuss this more in the next section, when connecting the minimization of imbalance to the bias and variance of BOP-e.

While the BOP-e procedure as described above gives an importance sampling estimator, the resulting weights $\hat{\rho}$ can be used in any off-policy method which uses importance weights.

[2]A loss is strictly composite if the Bayes-optimal score is given by $\bar{s}^* = \Psi \circ \hat{p}(C = 1|s, a)$ where Ψ is a link function $\Psi\ [0, 1] \to \mathbb{R}$. Readers should see Buja et al. [11] and Reid and Williamson [86] for complete treatments of strictly proper composite losses.

Extension to doubly robust estimation is trivial, as well as methods which adaptively combine direct method predictions and importance sampling weights, such as the SWITCH estimator of Wang et al. [118]. In Section 4.6, we implement BOP-e within both these frameworks, and compare to using inverse propensity score (IPS) weights.

4.4 Estimator analysis and asymptotics

In this section, we describe the statistical properties of our estimator, and prove consistency for the target policy value.

Let $p(a,s) := \frac{\pi_1(a,s)}{\pi_1(a,s)+\pi_0(a,s)}$ denote the true class probability of observing data (a,s) under the target policy π_1 instead of the behaviour policy π_0. This is estimated with a probabilistic classifier $\hat{p}(a,s)$ on labelled state-action data. Additionally, let $\rho(a,s) := \frac{\pi_1(a,s)}{\pi_0(a,s)} = \frac{p(a,s)}{1-p(a,s)}$ denote the true policy density ratio, with estimator $\hat{\rho}$. We assume the classifier has regret that decays with increasing n.

Assumption 18. *Let $\hat{p}(a,s)$ be a probabilistic classifier such that regret$(\hat{p};\mathcal{D},\ell) = O(n^{-\epsilon})$ for some constant $\epsilon \in (0,1)$.*

Next, we require that our importance sampling weight estimator, $\hat{\rho}$, is independent of the observed rewards r. This can be easily achieved through sample splitting, training the classifier $\hat{p}$ and applying BOP-e on independent datasets.

Assumption 19. *Given observed state-action data, the density ratio estimator $\hat{\rho}$ is independent of the observed rewards $r(\pi_0(s),s)$.*

Finally, we require certain regularity conditions and rates to use in our theoretical results.

Assumption 20. *(i) The functions $\pi_0(a,s), \pi_1(a,s), \rho(a,s)$, and $\hat{\rho}(a,s)$ have bounded second derivatives with respect to a, and (ii) In the continuous action domain, the bandwidth parameter $h = O(n^{-1/5})$.*

We now show that the BOP-e estimator is asymptotically unbiased, and derive a bound for its variance. We accomplish this by characterizing the asymptotic quantities in terms of

the Bregman divergence between the estimated and true density ratios. In the propositions below, we use r_{π_1} to denote $r(\pi_1(s), s)$ and ρ_{π_1} to denote $\rho(\pi_1(s), s)$.

Proposition 2. *(i) In discrete action spaces, the bias of $\hat{V}^{BOP-e}$ obeys the following bound*

$$\begin{aligned}&|\mathbb{E}_{\pi_1}[r] - \mathbb{E}_{\pi_0}\left[\mathbb{1}_a(\pi_1(s))\hat{\rho}(a,s)r(a,s)\right]| \\ &\leq \mathbb{E}_{\pi_0}[B\left(\rho, \hat{\rho}\right) r_{\pi_1}]\end{aligned}$$

(ii) In continuous action spaces, the expected bias of $\hat{V}^{BOP-e}$ obeys the following bound

$$\begin{aligned}&\left|\mathbb{E}_{\pi_1}[r] - \mathbb{E}_{\pi_0}\left[\frac{1}{h}K\left(\frac{a-\pi_1(s)}{h}\right)\hat{\rho}(a,s)r(a,s)\right]\right| \\ &\leq \mathbb{E}_{\pi_0}[B\left(\rho, \hat{\rho}\right) r_{\pi_1}] + o(h^2)\end{aligned}$$

Proposition 3. *(i) In discrete action spaces, the variance of $\hat{V}^{BOP-e}$ obeys the following bound*

$$\begin{aligned}&Var_{\pi_0}\left[\hat{V}^{BOP-e}\right] \\ &\leq \frac{1}{n}\Big(\mathbb{E}_{\pi_1}[\rho(\pi_1(s),s)r_{\pi_1}^2] + \mathbb{E}_{\pi_0}[B(\rho,\hat{\rho})^2 r_{\pi_1}^2] \\ &+ \mathbb{E}_{\pi_0}[2B(\rho,\hat{\rho})\rho(\pi_1(s),s)r_{\pi_1}^2]\Big)\end{aligned}$$

(ii) In continuous action spaces, the variance of $\hat{V}^{BOP-e}$ obeys the following bound

$$\begin{aligned}&Var_{\pi_0}\left[\hat{V}^{BOP-e}\right] \\ &\leq \frac{R(K)}{nh}\Big(\mathbb{E}_{\pi_1}[\rho(\pi_1(s),s)r_{\pi_1}^2] + \mathbb{E}_{\pi_0}[B(\rho,\hat{\rho})^2 r_{\pi_1}^2] \\ &+ \mathbb{E}_{\pi_0}[2B(\rho,\hat{\rho})\rho(\pi_1(s),s)r_{\pi_1}^2]\Big) + o\left(\frac{1}{nh}\right)\end{aligned}$$

where $R(K) = \int K(u)^2 du$.

The implication of Proposition 2 is that the expected bias of BOP-e is bounded from above by the Bregman divergence between the true density ratio between the observed and proposed policy and the model estimate of the density ratio. The specific Bregman divergence depends on the choice of classifier $\hat{p}$; for example, a logistic regression classifier would imply a KL-divergence. We note that Bregman divergences define a wide variety of divergences including KL-divergence and maximum mean discrepancy [46] that are often considered in the analysis of off-policy evaluation and covariate shift [55, 7, 34]. We can then appeal to Proposition 3 of Menon and Ong [74] that provides an explicit link between the risk of the classifier and the Bregman divergence between $\rho(a, s)$ and $\hat{\rho}(a, s)$.

Proposition 4. *[74] Let P be the class conditional $p(C = 1|s, a)$ and Q be the class conditional $p(C = 0|s, a)$ with marginal class probability $\frac{1}{2}$. Let $\mathcal{D}(P, Q, \frac{1}{2})$ be the joint distribution over C, S, A decomposed into P and Q and the marginal $p(C) = \frac{1}{2}$. Under assumption A17, for any scorer $\bar{s} : \mathcal{X} \to \mathbb{R}$,*

$$\textit{regret}(\bar{s}; \mathcal{D}, \ell) = \frac{1}{2}\mathbb{E}_{X\sim Q}\left[B_{f^{\circledast}}(\rho, \hat{\rho})\right],$$

where $f^{\circledast}(z) = (1 + z)f\left(\frac{z}{1+z}\right)$.

The proof can be found in Menon and Ong [74]. Which Bregman divergence is minimized is a consequence of the choice of loss used in BOP-e.

We now prove our main result below:

Proposition 5. *Under Assumptions 13-20, and with bounded variance of the Bregman divergence, the BOP-e estimator is consistent for the counterfactual policy value, that is, as $n \longrightarrow \infty$, $\hat{V}^{BOP-e} \longrightarrow \mathbb{E}_{\pi_1}[r]$.*

Proof sketch. This result follows by leveraging Propositions 2 and 3 to characterize the asymptotic bias and variance of $\hat{V}^{BOP-e}$. Then, we use Proposition 4 to connect the Bregman loss to the error of the classifier. Therefore, under Assumption 18, the bias and variance vanish as $n \longrightarrow \infty$, and the mean squared error of $\hat{V}^{BOP-e}$ tends to zero. □

The full proof and technical details for these results can be found in Section 4.8.

It is worth briefly discussing the implications of Propositions 2 and 3 combined with Proposition 4. Proposition 4 implies that optimizing classifier performance directly translates into optimizing the quality of the importance sampler. This provides a powerful property for BOP-e: the bias and variance of the estimated policy evaluation can be minimized by optimizing for classifier performance. Because the classifier risk is directly tied to the quality of the off-policy estimate, the problem is essentially reduced to model selection for supervised learning.

4.5 Related work

Related work can roughly be divided into three categories: off-policy evaluation of contextual bandits, balancing estimators, and density ratio estimation.

4.5.1 Off-policy evaluation of contextual bandits

Li et al. [66] introduced the use of rejection sampling for offline evaluation of contextual bandit problems. Within the causal inference community there is a long literature on the use of double-robust estimators (c.f. Bang and Robins [4], Kang et al. [59], Tan [106], Cao et al. [13]). Dudík et al. [23] later proposed the use of double-robust estimation for off-policy evaluation of contextual bandits, combining the double robust estimator of causal effects with rejection sampler. Since then, several works have sought to minimize the variance and improve robustness of the doubly robust estimator. Farajtabar et al. [28] and Wang et al. [118] present work to minimize the variance of the estimators by reducing the dependence on the inverse propensity score in high variance settings. Swaminathan and Joachims [105] use a Hájek style estimator [38]. Later work from Thomas [109] and Swaminathan and Joachims [105] builds on this work to improve estimation.

4.5.2 Balancing estimators

Balancing estimators have a long history within the causal inference community. Under correct specification of the conditional model Rosenbaum and Rubin [89] show balance of the propensity score. More recently, a growing literature seeks to develop balancing estimators which are robust to mis-specification. Hainmueller [37] and Zubizarreta [125] provide optimization-based procedures which define weights that are balancing but are not necessarily valid propensity scores. Imai and Ratkovic [48] later defined an estimator which strives to find a valid propensity score subject to balancing constraints. This was extended to general treatment regimes by Fong et al. [30]. However, none of these directly address the problem of off-policy evaluation for contextual bandits.

There is a much smaller literature explicitly on balancing estimators for the contextual bandit problem. Kallus [55] introduces a method for balanced policy evaluation that relies on a regularized estimator that seeks to minimize the maximum mean discrepancy [35]. Calculation of weights is achieved through a quadratic program, which presents computational challenges as sample size grows large. It is interesting to note that the proposed evaluation optimization of Kallus [55] fits within the assumptions of BOP-e where the scoring rule is maximum mean discrepancy (a strictly proper scoring rule) and the model is learned with variance regularization. The accompanying classifier can also be defined via a modification of support vector machine classification [7]. Dimakopoulou et al. [21] propose balancing in the context of online learning linear contextual bandits by reweighting based on the propensity score. This differs from this work in the focus on online learning rather than policy evaluation and the use of a linear model-based propensity score which provides mean balance only in the case of correct specification. Wu and Wang [120] propose a method which seeks to minimize an f-divergence to minimize regret, similar to the target in this work. However in the setting of Wu and Wang [120] access to the true propensities are assumed, whereas BOP-e estimates the density ratio directly from observed and proposed state action pairs.

4.5.3 Density ratio estimation

The use of classification for density ratio estimation dates back to at least Qin [81]. Later work leverages classification for covariate shift adaptation [6, 7] and two-sample testing [32, 71]. However, this work represents the first time classifier-based density estimation has been used for off-policy evaluation. There is also a growing literature on density ratio estimation that is defined outside of the framework of classification. These methods largely rely on kernels to perform estimation [44, 101]. KL importance estimation (KLIEP) [100], and least squares importance fitting (LSIF) [57] are the most directly relevant, given their ability to optimize hyper-parameters via cross validation. Interestingly, Menon and Ong [74] provides a loss for classification-based density ratio estimation that produces KLIEP and LSIF. Thus, these estimators can also be considered special cases of BOP-e by considering the corresponding loss functions for the classifier.

4.6 Experiments

In the experiments that follow, we evaluate direct method, importance sampling, and SWITCH estimators for off-policy evaluation. For the latter two methods, we compare inverse propensity score and BOP-e weights, and use the self-normalized versions of the estimators given in section 4.2.

The direct method, propensity score, and BOP-e estimators are all trained as gradient boosted tree classifiers (or regressors for the continuous evaluations). These boosting models correspond to an exponential scoring rule (satisfying A 17) which sharply penalizes overconfidence about the true class label [74].

4.6.1 Discrete action spaces

In this section, we evaluate the accuracy of our estimator for the value of an unobserved policy in the discrete reward setting. We employ the method of Dudík et al. [23] to turn a k-class classification problem into a k-armed contextual bandit problem. We split our data,

Table 4.1: Summary of datasets used in discrete reward experiments

Dataset	ecoli	glass	letters	optdigits	page-blocks	pendigits	satimage	vehicle	yeast
Classes (k)	5	6	26	10	5	10	6	4	9
Observations (n)	327	214	20000	5620	5473	10992	6435	846	1479
Covariates (p)	7	9	16	64	10	16	36	18	8

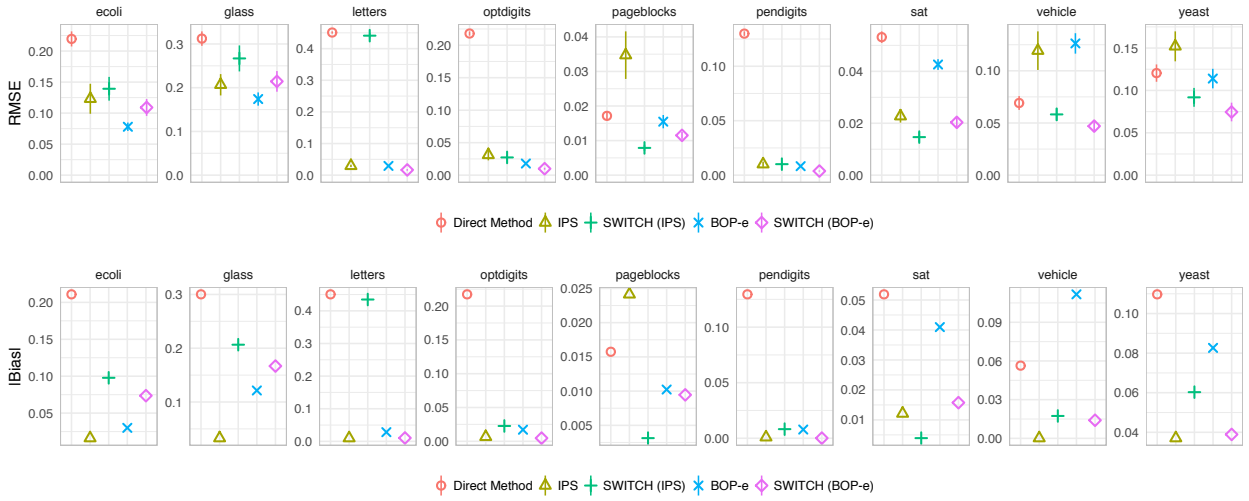

Figure 4.1: Root mean-squared error (RMSE) and bias plots for discrete action spaces using the classifier trick of Dudík et al. [24]. Full dataset descriptions are provided in table 4.1.

training a classifier on one half of the data (`train`). This classifier defines our target policy, wherein the action taken is the label predicted. The reward is defined as an indicator of whether the predicted label is the true label. The optimal policy, then, is to take an action equal to the true label in the original data. Evaluating this policy corresponds to estimating the classifier's accuracy.

In the second half of the dataset (`test`) we retain only a 'partially labeled' dataset wherein we uniformly sample actions (labels) and observe the resulting rewards. The `train` half of the data is also used to train direct method, propensity score, and BOP-e models. These are then applied to the `test` data to estimate the relevant quantities for off-policy

evaluation methods. We compare the expected reward estimates to the true mean reward of the target policy applied to the `test` data. For each dataset, this process is repeated over 100 iterations, where we vary the actions under the observed uniform policy.

Our policy models are trained as random forest classifiers. These models use the default hyperparameter values from scikit-learn with the exception of the number of trees [79]. In order to provide increasingly complex policies to evaluate, we increase the number of trees as a function of sample size: $\left\lfloor 10 \times n^{\frac{1}{4}} \right\rfloor$.

The propensity score, BOP-e and direct method (one-vs-rest) models are gradient boosted decision trees with default XGBoost hyperparameters with the exception of the number of boosting iterations. In order to adapt the estimator to the size of the dataset, the number of iterations is set as a function of sample size: $\lceil 20 \times \sqrt{n} \rceil$.

We use the same datasets from the UCI repository [22] used by Dudík et al. [23], and summarize their characteristics in Table 4.1. For some datasets, we removed classes with low frequencies to avoid issues when data splitting. The policy we evaluate is given by training a multi-class random forest model. We compare BOP-e to importance sampling with a standard inverse propensity score (IPS), including the corresponding doubly robust estimators and the direct method. The performance results of the estimators are summarized in Figure 4.1, where we plot the root mean squared error and bias averaged over 100 iterations. We see that the direct method estimator tends to be heavily biased for the true policy value, compared to BOP-e and IPS. Moreover, in these simulations, the direct method perfoms generally quite poorly in terms of overall accuracy. The standard BOP-e estimator performs at least as well as and typically better than the IPS estimator. This also holds for the corresponding SWITCH estimators. While BOP-e often has slightly higher bias than IPS, it strikes a better balance between bias and variance, leading to substantially improved accuracy in most cases.

Table 4.2: Summary of datasets used in continuous reward experiments

Dataset	abalone	admissions	airfoil	auto	housing	power	wine
Observations (n)	4177	400	1503	392	10000	9568	1599
Covariates (p)	10	7	5	7	14	4	11

4.6.2 Continuous action spaces

For the continuous action case, we provide a novel extension of the same transformation employed in the previous section for evaluation of discrete actions. We take a selection of datasets with continuous outcomes, and train a predictive model on the `train` half of the data, which constitutes our target policy. The reward of a prediction (defined to be an action in our evaluation) is the negative of the Euclidean distance to the true outcome[3]. Thus, the optimal action is to choose actions equal to the true outcome as in the discrete evaluation. Evaluating the behavior policy is equivalent to estimating the mean squared error of the predictive model.

As before, we retain the `test` data for evaluation, while using the `train` data to train direct method, propensity score, and BOP-e models. For our observed policy, we sample actions from the empirical distribution of `train` outcomes, and compute the corresponding rewards. We then estimate the target policy value, repeating this over 300 iterations. We retain the same basic models from the previous section for this evaluation, swapping out classifiers for regressors as appropriate.

We use datasets from the UCI repository [22] and Kaggle, and summarize their characteristics in Table 4.2. The policy we evaluate is given by training a random forest regression to predict the continuous outcome. We also use gradient boosted regression trees for training direct method, propensity score, and BOP-e models. Specifically, to obtain a continuous propensity score, we apply our observed policy to the `train` data, and train a model $\hat{g}$ to

[3]This formulation of the problem makes clear that there is a class of reasonable evaluations on mixed or continuous action spaces so long as a suitable distance metric is used.

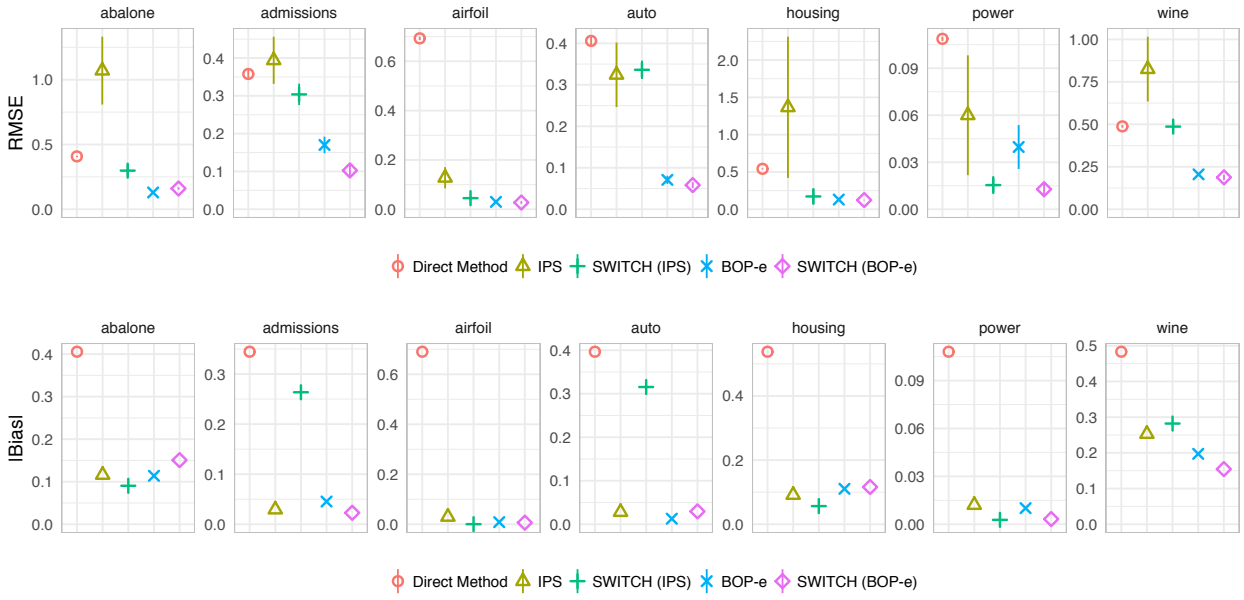

Figure 4.2: Root mean-squared error (RMSE) and bias plots for continous action spaces using a modification of the classifier trick of Dudík et al. [24] for regression detailed in section 4.6.2. Full dataset descriptions are provided in table 4.2.

predict actions from state features. Then, conditional on state s, the action is assumed to come from a normal distribution with mean $\hat{g}(s)$ and variance $MSE(\hat{g})$ as is standard practice [40]. For each state-action pair (s, a) in the `test` data, the propensity score is then the density of this distribution at a.

As in the previous section, we compare BOP-e to IPS (with the Kallus and Zhou [56] kernel) and the direct method, including the relevant SWITCH estimators. These results are displayed in Figure 4.2. We see that BOP-e outperforms the other methods uniformly across all datasets. In contrast to the binary setting, BOP-e does a better job of correcting for bias than the naïve IPS method. This is not surprising as the IPS is forced to make strong assumptions about the conditional distribution of action given state which BOP-e does not require. Specifically, IPS must assume conditional normality; BOP-e does not. Given that real data rarely conforms to ideal theoretical distributions, this provides major

benefits. In our simulations, the true distribution of the behavior policy matches the marginal distribution of labels in the dataset and is thus not conditionally normal (unless the marginal is).

In addition to reducing bias, BOP-e greatly reduces RMSE in most datasets. The BOP-e SWITCH estimator improves on the IPS version in both RMSE and bias in almost all cases. On the `power` dataset, BOP-e provides half the RMSE of IPS when used within the SWITCH estimator. On `admissions` and `auto`, BOP-e incurs less than one-third of the RMSE than does standard IPS.

4.7 Conclusions

In this work, we introduced BOP-e, a simple, flexible, and powerful method for off-policy evaluation of contextual bandits. BOP-e is easily implemented using off the shelf classifiers and trivially generalizes to arbitrary action types, e.g. continuous, multi-valued. In section 4.4 we connect the bias and variance of our estimator with the risk of the classification task, and show that BOP-e is inherently balance-seeking. As a consequence of the theoretical results, hyperparameter tuning and model selection can be performed by minimizing classification error using well-known strategies from supervised learning. Experimental evidence indicates that BOP-e provides state of the art performance for discrete and continuous actions spaces. A natural direction for future work is considering the case of evaluation with sequential decision making and structured action spaces. Our method could also be extended to perform policy optimization in all of these settings. It would also be interesting to consider the integration of BOP-e with methods for variance reduction, e.g. Thomas and Brunskill [108] and Farajtabar et al. [28], to further improve performance.

4.8 Proofs of technical results

Here, we provide technical proofs of the propositions in Section 4.4.

4.8.1 Proof of Proposition 2

Because the weights in the denominator $\hat{V}^{BOP-e}$ are each consistent for 1, we have that the sum is consistent for n. Therefore, by the continuous mapping theorem, we can consider the expectation of a single term in the $\hat{V}^{BOP-e}$ numerator.

Recall that $\rho(a,s) = \frac{\pi_1(a,s)}{\pi_0(a,s)}$ denotes the true density ratio and $\hat{\rho}(a,s)$ is the estimated density ratio. Further let $\delta(a,s) = \hat{\rho}(a,s) - \rho(a,s)$. First, we consider the discrete action setting. We can express the expectation as:

$$\begin{aligned}\mathbb{E}_{\pi_0}\left[\mathbb{1}_a(\pi_1(s))\hat{\rho}(a,s)r(a,s)\right] &= \mathbb{E}_{\pi_0}\left[\mathbb{1}_a(\pi_1(s))(\rho(a,s)+\delta(a,s))r(a,s)\right] \\ &= \mathbb{E}_{\pi_0}\left[\mathbb{1}_a(\pi_1(s))\rho(a,s)r(a,s)\right] + \mathbb{E}_{\pi_0}\left[\mathbb{1}_a(\pi_1(s))\delta(a,s))r(a,s)\right]\end{aligned}$$

We can show that the first term is equal to the policy value of π_1, while the second term provides the estimator's bias. Considering the first term, we have:

$$\begin{aligned}\mathbb{E}_{\pi_0}\left[\mathbb{1}_a(\pi_1(s))\rho(a,s)r(a,s)\right] &= \sum_{(a,s)} \mathbb{1}_a(\pi_1(s))\rho(a,s)r(a,s)\pi_0(a,s) \\ &= \sum_{(a,s)} \mathbb{1}_a(\pi_1(s))r(a,s)\pi_1(a,s) \\ &= \sum_{s} r(\pi_1(s),s)\pi_1(\pi_1(s),s) \\ &= \mathbb{E}_{\pi_1}\left[r_{\pi_1}\right],\end{aligned}$$

where r_{π_1} denotes $r(\pi_1(s),s)$.

Now, considering the bias term, and bounding δ with the Bregman divergence between

ρ and $\hat{\rho}$, we have:

$$\begin{aligned}
\mathbb{E}_{\pi_0}\left[\mathbb{1}_a(\pi_1(s))\delta(a,s))r(a,s)\right] &= \sum_{(a,s)} \mathbb{1}_a(\pi_1(s))\delta(a,s)r(a,s)\pi_0(a,s) \\
&\leq \sum_{(a,s)} \mathbb{1}_a(\pi_1(s))B(\rho,\hat{\rho})r(a,s)\pi_0(a,s) \\
&= \sum_{s} B(\rho,\hat{\rho})r(\pi_1(s),s)\pi_0(\pi_1(s),s) \\
&= \mathbb{E}_{\pi_0}\left[B(\rho,\hat{\rho})r_{\pi_1}\right]
\end{aligned}$$

We now move on to the continuous action setting. We can express the expectation as:

$$\begin{aligned}
&\mathbb{E}_{\pi_0}\left[\frac{1}{h}K\left(\frac{a-\pi_1(s)}{h}\right)\hat{\rho}(a,s)r(a,s)\right] \\
&= \int \frac{1}{h}K\left(\frac{a-\pi_1(s)}{h}\right)\left(\rho(a,s)+\delta(a,s)\right)r(a,s)\pi_0(a,s)d(a,s) \\
&= \int \frac{1}{h}K\left(\frac{a-\pi_1(s)}{h}\right)\rho(a,s)r(a,s)\pi_0(a,s)d(a,s) \\
&\quad + \int \frac{1}{h}K\left(\frac{a-\pi_1(s)}{h}\right)\delta(a,s)r(a,s)\pi_0(a,s)d(a,s)
\end{aligned}$$

We can show that the first term is equal to the true counterfactual policy value, while the second term describes the bias induced from estimating the density ratio. Considering the first term, we have:

$$\int \frac{1}{h}K\left(\frac{a-\pi_1(s)}{h}\right)\frac{\pi_1(a,s)}{\pi_0(a,s)}r(a,s)\pi_0(s,a)d(s,a) = \int \frac{1}{h}K\left(\frac{a-\pi_1(s)}{h}\right)r(a,s)\pi_1(a,s)d(s,a)$$

Let $u = \dfrac{a-\pi_1(s)}{h}$. Thus, $a = \pi_1(s) + hu$ and $da = hdu$. Then, taking a second-order Taylor expansion of π_1 around $\pi_1(s)$:

$$
\begin{aligned}
&\int \frac{1}{h} K\left(\frac{a-\pi_1(s)}{h}\right) \frac{\pi_1(a,s)}{\pi_0(a,s)} r(a,s)\pi_0(s,a)d(s,a) \\
&= \int K(u)\, r(\pi_1(s)+hu,s)\pi_1(\pi_1(s)+hu,s)d(s,u) \\
&= \int K(u)\, r(\pi_1(s),s)\pi_1(\pi_1(s),s)d(s,u) \\
&\quad + \int K(u)\, r(\pi_1(s),s)\pi_1'(\pi_1(s),s)(hu)d(s,u) \\
&\quad + \int K(u)\, r(\pi_1(s),s)\pi_1''(\pi_1(s),s)\frac{(hu)^2}{2}d(s,u) \\
&\quad + \int K(u)\, o(h^2) r(\pi_1(s),s)d(s,u) \\
&= \int K(u)\, du \int r(\pi_1(s),s)\pi_1(\pi_1(s),s)ds \\
&\quad + \int uK(u)\, du \int r(\pi_1(s),s)\pi_1'(\pi_1(s),s)hd(s,u) \\
&\quad + \int u^2 K(u)\, du \int \frac{h^2}{2} r(\pi_1(s),s)\pi_1''(\pi_1(s),s)ds \\
&\quad + \int K(u)\, du \int o(h^2) r(\pi_1(s),s)ds \\
&= \int r(\pi_1(s),s)\pi_1(\pi_1(s),s)ds + o(h^2) \\
&= E_{\pi_1}[r_{\pi_1}] + o(h^2).
\end{aligned}
$$

This result follows similarly to those in Kallus and Zhou [56], by properties of kernels, bounded rewards, and since $\pi_1(a,s)$ has a bounded second derivative with respect to a.

Now, considering the bias term, we use the same $u-$substitution and Taylor expansion

as before. We also bound δ by the Bregman divergence between ρ and $\hat{\rho}$, yielding:

$$
\begin{aligned}
&\int \frac{1}{h} K\left(\frac{a-\pi_1(s)}{h}\right) \delta(a,s) r(a,s) \pi_0(a,s) d(a,s) \\
\leq &\int \frac{1}{h} K\left(\frac{a-\pi_1(s)}{h}\right) B(\rho,\hat{\rho}) r(a,s) \pi_0(a,s) d(a,s) \\
= &\int K(u)\, B(\rho,\hat{\rho}) r(\pi_1(s)+hu) \pi_0(\pi_1(s)+hu,s) d(u,s) \\
= &\int K(u)\, du \int B(\rho,\hat{\rho}) r(\pi_1(s),s) \pi_0(\pi_1(s),s) ds + Rem(h) \\
= &\int B(\rho,\hat{\rho}) r(\pi_1(s),s) \pi_0(\pi_1(s),s) ds + o(h^2) \\
= &E_{\pi_0}[B(\rho,\hat{\rho}) r_{\pi_1}] + o(h^2)
\end{aligned}
$$

4.8.2 Proof of Proposition 3

We consider the second moment of a single numerator term, and write the estimator in terms of ρ and δ as above. We first consider the discrete action setting.

$$\begin{aligned}
\mathbb{E}_{\pi_0}\left[\mathbb{1}_a(\pi_1(s))^2\hat{\rho}(a,s)^2 r(a,s)^2\right] &= \mathbb{E}_{\pi_0}\left[\mathbb{1}_a(\pi_1(s))(\rho(a,s)+\delta(a,s))^2 r(a,s)^2\right] \\
&= \mathbb{E}_{\pi_0}\left[\mathbb{1}_a(\pi_1(s))(\rho(a,s)^2+\delta(a,s)^2+2\rho(a,s)\delta(a,s))r(a,s)^2\right] \\
&= \sum_{(a,s)} \mathbb{1}_a(\pi_1(s))\rho(a,s)^2 r(a,s)^2\pi_0(a,s) \\
&+ \sum_{(a,s)} \mathbb{1}_a(\pi_1(s))\delta(a,s)^2 r(a,s)^2\pi_0(a,s) \\
&+ \sum_{(a,s)} \mathbb{1}_a(\pi_1(s))2\rho(a,s)\delta(a,s) r(a,s)^2\pi_0(a,s) \\
&\leq \sum_{s} \rho(\pi_1(s),s) r(\pi_1(s),s)^2\pi_1(\pi_1(s),s) \\
&+ \sum_{s} 2B(\rho,\hat{\rho})\rho(\pi_1(s),s) r(\pi_1(s),s)^2\pi_0(\pi_1(s),s) \\
&+ \sum_{s} B(\rho,\hat{\rho})^2 r(\pi_1(s),s)^2\pi_0(\pi_1(s),s) \\
&= \mathbb{E}_{\pi_1}[\rho(\pi_1(s),s)r_{\pi_1}^2] + \mathbb{E}_{\pi_0}[B(\rho,\hat{\rho})^2 r_{\pi_1}^2] + \mathbb{E}_{\pi_0}[2B(\rho,\hat{\rho})\rho(\pi_1(s),s)r_{\pi_1}^2]
\end{aligned}$$

Therefore, the variance of the estimator is bounded by:

$$\frac{1}{n}\left(\mathbb{E}_{\pi_1}[\rho(\pi_1(s),s)r_{\pi_1}^2] + \mathbb{E}_{\pi_0}[B(\rho,\hat{\rho})^2 r_{\pi_1}^2] + \mathbb{E}_{\pi_0}[2B(\rho,\hat{\rho})\rho(\pi_1(s),s)r_{\pi_1}^2]\right)$$

Next, we consider the second moment of a term in the estimator in the continuous action setting:

$$\begin{aligned}
&\mathbb{E}_{\pi_0}\left[\left(\frac{1}{h}K\left(\frac{a-\pi_1(s)}{h}\right)\hat{\rho}(a,s)r(a,s)\right)^2\right] \\
&= \int \frac{1}{h^2}K\left(\frac{a-\pi_1(s)}{h}\right)^2 \left(\rho(a,s)+\delta(a,s)\right)^2 r(a,s)^2\pi_0(a,s)d(a,s)
\end{aligned}$$

We substitute $u = \dfrac{a - \pi_1(s)}{h}$ as before. Then, $a = \pi_1(s) + hu$ and $da = hdu$.

$$\mathbb{E}_{\pi_0}\left[\left(\frac{1}{h}K\left(\frac{a-\pi_1(s)}{h}\right)\hat{\rho}(a,s)r(a,s)\right)^2\right]$$

$$= \int \frac{1}{h}K\left(u\right)^2\left(\rho(\pi_1(s)+hu,s)+\delta(\pi_1(s)+hu,s)\right)^2 r(\pi_1(s)+hu)^2\pi_0(\pi_1(s)+hu,s)d(s,u)$$

Next, we apply a second-order Taylor series expansion of ρ, δ, and π_0 around $\pi_1(s)$. Given that these functions have bounded second derivatives, we can bound the remainder by $o(h^{-1})$, as in Kallus and Zhou [56]. This yields:

$$\mathbb{E}_{\pi_0}\left[\left(\frac{1}{h}K\left(\frac{a-\pi_1(s)}{h}\right)\hat{\rho}(a,s)r(a,s)\right)^2\right]$$

$$= \int \frac{1}{h}K\left(u\right)^2 du \int \left(\rho(\pi_1(s),s)+\delta(\pi_1(s),s)\right)^2 r(\pi_1(s),s)^2\pi_0(\pi_1(s),s)ds + o(h^{-1})$$

$$= \frac{R(K)}{h}\int \left(\rho(\pi_1(s),s)+\delta(\pi_1(s),s)\right)^2 r(\pi_1(s),a)^2\pi_0(\pi_1(s),s)ds + o(h^{-1})$$

$$= \frac{R(K)}{h}\int \left(\rho(\pi_1(s),s)^2+\delta(\pi_1(s),s)^2+2\rho(\pi_1(s),s)\delta(\pi_1(s),s)\right) r(\pi_1(s),a)^2\pi_0(\pi_1(s),s)ds + o(h^{-1})$$

$$= \frac{R(K)}{h}\Big[\int \rho(\pi_1(s),s)^2 r_{\pi_1}^2\pi_0(\pi_1(s),s)ds + \int \delta(\pi_1(s),s)^2 r_{\pi_1}^2\pi_0(\pi_1(s),s)ds$$

$$+ \int 2\rho(\pi_1(s),s)\delta(\pi_1(s),s) r_{\pi_1}^2\pi_0(\pi_1(s),s)ds\Big]$$

$$+ o(h^{-1})$$

$$= \frac{R(K)}{h}\Big[\int \rho(\pi_1(s),s) r_{\pi_1}^2\pi_1(\pi_1(s),s)ds + \int \delta(\pi_1(s),s)^2 r_{\pi_1}^2\pi_0(\pi_1(s),s)ds$$

$$+ \int 2\delta(\pi_1(s),s) r_{\pi_1}^2\pi_1(\pi_1(s),s)ds\Big]$$

$$+ o(h^{-1})$$

where $R(K) := \int K(u)^2 du$ is some constant.

Then, bounding δ by the Bregman divergence B,

$$\mathbb{E}_{\pi_0}\left[\left(\frac{1}{h}K\left(\frac{a-\pi_1(s)}{h}\right)\hat{\rho}(a,s)r\right)^2\right]$$
$$\leq \frac{R(K)}{h}\left[\mathbb{E}_{\pi_1}[\rho(\pi_1(s),s)r_{\pi_1}^2]+\mathbb{E}_{\pi_0}[B(\rho,\hat{\rho})^2 r_{\pi_1}^2]+\mathbb{E}_{\pi_1}[2B(\rho,\hat{\rho})r_{\pi_1}^2]\right]+o(h^{-1})$$

Therefore, the variance of our estimator is bounded by:

$$\frac{R(K)}{nh}\left(\mathbb{E}_{\pi_1}[\rho(\pi_1(s),s)r_{\pi_1}^2]+\mathbb{E}_{\pi_0}[B(\rho,\hat{\rho})^2 r_{\pi_1}^2]+\mathbb{E}_{\pi_1}[2B(\rho,\hat{\rho})r_{\pi_1}^2]\right)+o\left(\frac{1}{nh}\right)$$

4.8.3 Proof of Proposition 5

Based on Propositions 2 and 3, by selecting a Bregman divergence of the form in Proposition 4, we can bound the bias and variance in terms of the classifier $\hat{\rho}$ regret. Recall from Assumption 18, this regret scales as $O(n^{-\epsilon})$ for $\epsilon \in (0,1)$. Then, since rewards r are bounded, and $h = O(n^{-1/5})$ we have that the bias tends to 0 as $n \to \infty$.

We can apply a similar argument for the variance, by decomposing $\mathbb{E}_{\pi_0}[B(\rho,\hat{\rho})^2] = Var_{\pi_0}[B(\rho,\hat{\rho})] + \mathbb{E}_{\pi_0}[B(\rho,\hat{\rho})]^2$. Then, given that $Var_{\pi_0}[B(\rho,\hat{\rho})]$, ρ, and r are bounded, we have that the variance bound in Proposition 3 also goes to 0 as $n \to \infty$.

4.9 Additional experimental results

Here, we include additional evaluations of the datasets considered in Section 4.6, where we consider doubly robust implementations of IPS and BOP-e. Overall, we see very similar trends as those in the experiments given previously.

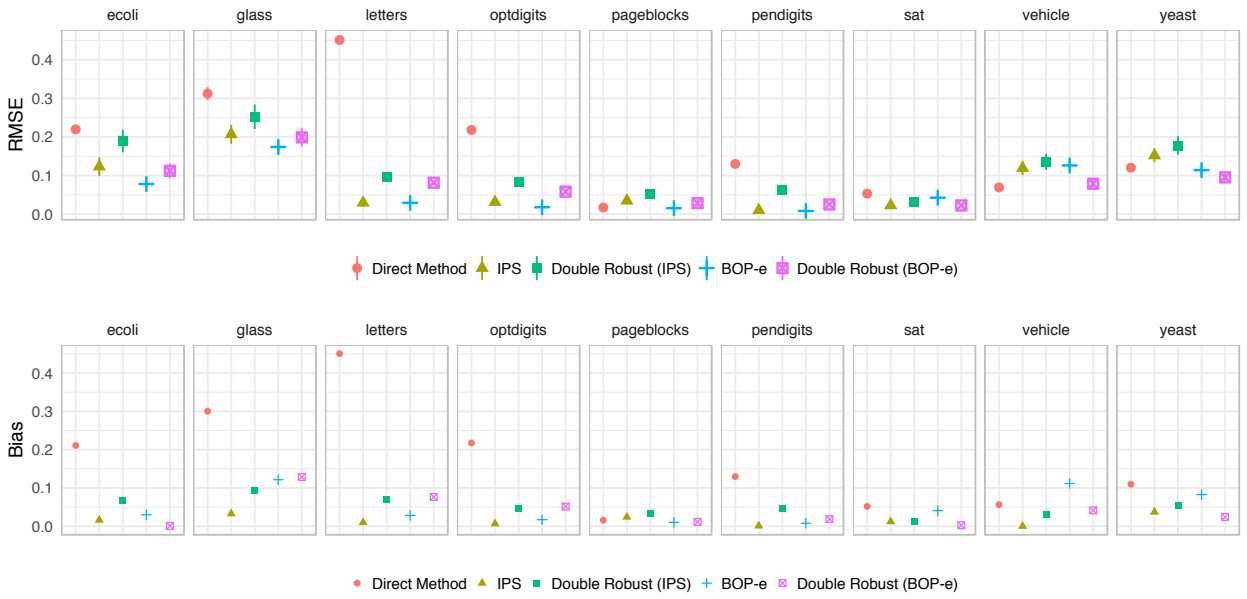

Figure 4.3: Additional root mean-squared error (RMSE) and bias plots for discrete action spaces using the classifier trick of Dudík et al. [24]. Full dataset descriptions are provided in table 4.1.

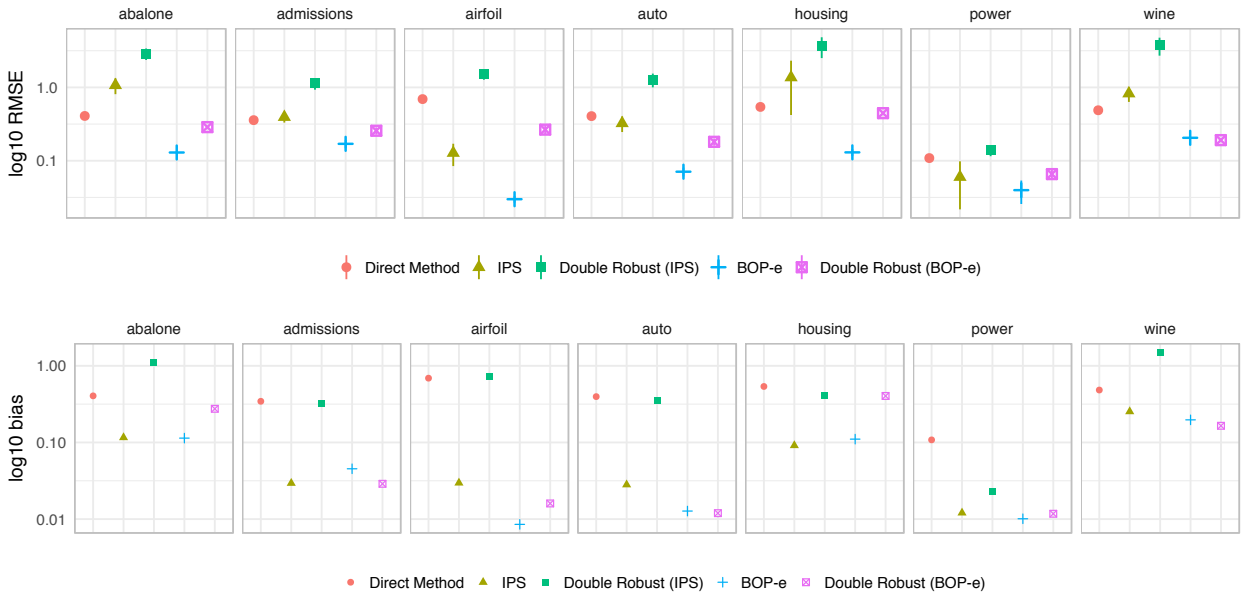

Figure 4.4: Additional root mean-squared error (RMSE) and bias plots for continous action spaces using a modification of the classifier trick of Dudík et al. [24] for regression detailed in section 4.6.2. Reports for each are provided in log-scale due to the poor performance of the inverse propensity score based methods. Full dataset descriptions are provided in table 4.2.

Chapter 5
DISCUSSION

5.1 Summary

In this dissertation, we proposed three novel methodological contributions in separate-but-related areas. In Chapter 2, we develop a new method to estimate directed acyclic graphs, which represent causal networks, in high dimensions. This method relies on leveraging structural information—specifically, the local separation property found in common random graph families—in order to gain computational and statistical efficiency. Compared to the standard PC-Algorithm for DAG estimation, we require weaker theoretical assumptions for consistency of our algorithm. We also empirically demonstrate that our algorithm provides more accurate estimation and a faster runtime than the PC-Algorithm for graphs with hub nodes.

In Chapter 3, we now assume knowledge of networks among both observation units and features, and incorporate their information to improve prediction and variable selection in the high-dimensional setting. This work follows similar themes as Chapter 2, in that we are using auxiliary information to improve a method. However, the concept of networks here is more general than in Chapter 2; the unit and feature networks do not necessarily need to imply causal relationships, and can instead be based on any type of association or distance between nodes. Our framework fills a gap in the literature by allowing for analysis of two-way structured data with penalized generalized linear models. If the observed networks are fully informative, we show that this method leads to improved prediction and inference, and remains competitive even if they are misspecified.

Finally, in Chapter 4, we work in the very different area of contextual bandits, and construct a new class of estimators for off-policy evaluation. Although unrelated to networks,

our use of balancing estimators and importance sampling are closely related to propensity score methods in observational causal inference, which ties to the goal of Chapter 2. Using classifier-based density ratio estimation, we develop an improved estimator for importance sampling weights, which is particularly effective in continuous action domains. These weights can easily be combined with modern methods that adaptively combine importance sampling with regression models. We show that minimizing the classifier loss is equivalent to minimizing a distance between the observed and target policy distributions. Essentially, this translates the off-policy evaluation problem into a simple binary classification task. We demonstrate that our method greatly outperforms existing approaches in continuous action spaces, and remains competitive under discrete actions.

5.2 Limitations and future research

As we described in Chapter 2, our reduced PC algorithm (rPC) requires a weaker faithfulness condition than that of the PC-Algorithm. This follows from the algorithm only conditioning on sets of small cardinally. However, our path faithfulness condition still has a low probability of being satisfied in certain settings, such as with dense DAGs having a large average degree, or with edge weights distributed over a larger parameter space. We considered both these settings in simulation, and showed that the probability of both conditions tends to zero. Therefore, an interesting future research direction would be to weaken this assumption further, possibly by incorporating more graph properties. In addition, like most DAG estimation methods, rPC considers linear structural equation models only. Generalizing the idea of restricted conditioning to more complex probability models over DAGs, such as nonlinear SEMs [115] would also be of interest. Our idea can also be used to develop more efficient hybrid methods, which combine conditional independence testing and score-based optimization, for learning DAGs in high dimensions.

The `glm-funk` framework that we developed in Chapter 3 accounts for unit and feature network information through penalization involving a kernel matrix for each network. The kernel matrix summarizes edge and distance information between units and features. We only

considered Laplacian and incidence matrices as kernels; however, different kernels can easily be used within the `glm-funk` framework, as long as the corresponding penalty term is zero given perfect information. Another way to incorporate unit network information would be to penalize the fitted values $X\beta$ directly instead of adding intercepts α to the model; that is, using the penalty term $\gamma_n \beta' X' L_n X \beta$. This implies a different type of network cohesion than the RNC penalty, which penalizes the extra variability in y not captured by the features X. Conversely, penalizing $X\beta$ directly incorporates the network G_n into the association between y and X.

Regarding the penalized regression framework, an interesting direction for future research would be to parameterize the model to have many penalty parameters over each network, rather than a single γ_p or γ_n. By data-adaptively tuning these parameters, the model would ideally incorporate the information from informative edges only. Feng and Simon [29] show that tuning a large number of penalty parameters via cross-validation is feasible and can improve model prediction power. Another important contribution would be to incorporate methods for cross-validation with correlated data, which would improve our tuning procedure and estimates of generalization error.

BOP-e, our new method for off-policy evaluation described in Chapter 4, applies to contextual bandit settings, where each observation consists of a single state, action, and reward. A natural direction for future research is extending BOP-e to the reinforcement learning setting, where a policy conducts sequential decision making. Here, each observation is a sequence of (state, action, reward) tuples, until the process reaches some terminating state. It would also be very fruitful to extend our method to policy optimization, for both contextual bandit and reinforcement learning settings. Finally, it would be interesting to conduct an empirical study of BOP-e and other policy evaluation methods on real-world datasets, such as those involving dynamic treatment regimes.

BIBLIOGRAPHY

[1] Silvia Acid and Luis M de Campos. Approximations of causal networks by polytrees: an empirical study. In *Advances in Intelligent ComputingIPMU'94*, pages 149–158. Springer, 1994.

[2] Genevera I Allen, Logan Grosenick, and Jonathan Taylor. A generalized least-square matrix decomposition. *Journal of the American Statistical Association*, 109(505):145–159, 2014.

[3] Animashree Anandkumar, Vincent YF Tan, Furong Huang, and Alan S Willsky. High-dimensional gaussian graphical model selection: Walk summability and local separation criterion. *The Journal of Machine Learning Research*, 13(1):2293–2337, 2012.

[4] Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

[5] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[6] Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning for differing training and test distributions. In *Proceedings of the 24th international conference on Machine learning*, pages 81–88. ACM, 2007.

[7] Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning under covariate shift. *Journal of Machine Learning Research*, 10(Sep):2137–2155, 2009.

[8] L.M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR*

Computational Mathematics and Mathematical Physics, 7(3):200 – 217, 1967. ISSN 0041-5553. doi: https://doi.org/10.1016/0041-5553(67)90040-7. URL `http://www.sciencedirect.com/science/article/pii/0041555367900407`.

[9] Peter Bühlmann and Sara van de Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

[10] Peter Bühlmann et al. Statistical significance in high-dimensional linear models. *Bernoulli*, 19(4):1212–1242, 2013.

[11] Andreas Buja, Werner Stuetzle, and Yi Shen. Loss functions for binary class probability estimation and classification: Structure and applications. *Working draft, November*, 3, 2005.

[12] Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, 489(7417):519–525, 2012.

[13] Weihua Cao, Anastasios A Tsiatis, and Marie Davidian. Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*, 96(3):723–734, 2009.

[14] Marc RJ Carlson, Bin Zhang, Zixing Fang, Paul S Mischel, Steve Horvath, and Stanley F Nelson. Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks. *BMC genomics*, 7(1):40, 2006.

[15] Han Chen, Chaolong Wang, Matthew P Conomos, Adrienne M Stilp, Zilin Li, Tamar Sofer, Adam A Szpiro, Wei Chen, John M Brehm, Juan C Celedón, et al. Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *The American Journal of Human Genetics*, 98(4):653–666, 2016.

[16] Hao Chen and Burt M Sharp. Content-rich biological network constructed by mining pubmed abstracts. *BMC bioinformatics*, 5(1):147, 2004.

[17] Xi Chen, Qihang Lin, Seyoung Kim, Jaime G Carbonell, Eric P Xing, et al. Smoothing proximal gradient method for general structured sparse regression. *The Annals of Applied Statistics*, 6(2):719–752, 2012.

[18] David Maxwell Chickering. Learning bayesian networks is np-complete. In *Learning from data*, pages 121–130. Springer, 1996.

[19] William G Cochran. *Sampling techniques*. Wiley, 1977.

[20] Ruben Dezeure, Peter Bühlmann, Lukas Meier, and Nicolai Meinshausen. High-dimensional inference: Confidence intervals, p-values and r-software hdi. *Statistical science*, pages 533–558, 2015.

[21] Maria Dimakopoulou, Zhengyuan Zhou, Susan Athey, and Guido Imbens. Balanced linear contextual bandits. *arXiv preprint arXiv:1812.06227*, 2018.

[22] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

[23] Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 1097–1104, 2011.

[24] Miroslav Dudík, Dumitru Erhan, John Langford, Lihong Li, et al. Doubly robust policy evaluation and optimization. *Statistical Science*, 29(4):485–511, 2014.

[25] Richard Durrett. *Random graph dynamics*, volume 200. Cambridge university press Cambridge, 2007.

[26] Daniel Eaton and Kevin Murphy. Bayesian structure learning using dynamic programming and mcmc. *arXiv preprint arXiv:1206.5247*, 2012.

[27] Johannes F Fahrmann, Kyoungmi Kim, Brian C DeFelice, Sandra L Taylor, David R Gandara, Ken Y Yoneda, David T Cooke, Oliver Fiehn, Karen Kelly, and Suzanne

Miyamoto. Investigation of metabolomic blood biomarkers for detection of adenocarcinoma lung cancer. *Cancer Epidemiology and Prevention Biomarkers*, 24(11):1716–1723, 2015.

[28] Mehrdad Farajtabar, Yinlam Chow, and Mohammad Ghavamzadeh. More robust doubly robust off-policy evaluation. In *International Conference on Machine Learning*, pages 1446–1455, 2018.

[29] Jean Feng and Noah Simon. An analysis of the cost of hyper-parameter selection via split-sample validation, with applications to penalized regression. *arXiv preprint arXiv:1903.12297*, 2019.

[30] Christian Fong, Chad Hazlett, Kosuke Imai, et al. Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements. *The Annals of Applied Statistics*, 12(1):156–177, 2018.

[31] Rina Foygel and Mathias Drton. Extended bayesian information criteria for gaussian graphical models. In *Advances in Neural Information Processing Systems*, pages 604–612, 2010.

[32] Jerome Friedman. On multivariate goodness-of-fit and two-sample testing. Technical report, Stanford Linear Accelerator Center, Menlo Park, CA (US), 2004.

[33] Nir Friedman, Michal Linial, and Iftach Nachman. Using bayesian networks to analyze expression data. *Journal of Computational Biology*, 7:601–620, 2000.

[34] Arthur Gretton, Alex Smola, Jiayuan Huang, Marcel Schmittfull, Karsten Borgwardt, and Bernhard Schölkopf. Covariate shift by kernel mean matching. *Dataset shift in machine learning*, 3(4):5, 2009.

[35] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

[36] Min Jin Ha, Wei Sun, and Jichun Xie. Penpc: A two-step approach to estimate the skeletons of high-dimensional directed acyclic graphs. *Biometrics*, 72(1):146–155, 2016.

[37] Jens Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20 (1):25–46, 2012.

[38] Jaroslav Hájek et al. Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, 35(4):1491–1523, 1964.

[39] Asad Haris, Noah Simon, and Ali Shojaie. Generalized Sparse Additive Models. *arXiv e-prints*, art. arXiv:1903.04641, Mar 2019.

[40] Keisuke Hirano and Guido W Imbens. The propensity score with continuous treatments. *Applied Bayesian modeling and causal inference from incomplete-data perspectives*, 226164:73–84, 2004.

[41] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

[42] S Horvath, B Zhang, M Carlson, KV Lu, S Zhu, RM Felciano, MF Laurance, W Zhao, S Qi, Z Chen, et al. Analysis of oncogenic signaling networks in glioblastoma identifies aspm as a molecular target. *Proceedings of the National Academy of Sciences*, 103(46): 17402–17407, 2006.

[43] Jian Huang, Shuangge Ma, Hongzhe Li, and Cun-Hui Zhang. The sparse laplacian shrinkage estimator for high-dimensional regression. *Annals of statistics*, 39(4):2021, 2011.

[44] Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex J Smola. Correcting sample selection bias by unlabeled data. In *Advances in neural information processing systems*, pages 601–608, 2007.

[45] Juan F. Huete and Luis M. Campos. *Symbolic and Quantitative Approaches to Reasoning and Uncertainty: European Conference ECSQARU '93 Granada, Spain, November 8–10, 1993 Proceedings*, chapter Learning causal polytrees, pages 180–185. Springer Berlin Heidelberg, Berlin, Heidelberg, 1993. ISBN 978-3-540-48130-0. doi: 10.1007/BFb0028199. URL http://dx.doi.org/10.1007/BFb0028199.

[46] Ferenc Huszar. *Scoring rules, divergences and information in Bayesian machine learning*. PhD thesis, University of Cambridge, 2013.

[47] Trey Ideker and Nevan J Krogan. Differential network biology. *Molecular systems biology*, 8(1):565, 2012.

[48] Kosuke Imai and Marc Ratkovic. Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):243–263, 2014.

[49] Guido W Imbens. The role of the propensity score in estimating dose-response functions. *Biometrika*, 87(3):706–710, 2000.

[50] Ronald Jansen, Haiyuan Yu, Dov Greenbaum, Yuval Kluger, Nevan J. Krogan, Sambath Chung, Andrew Emili, Michael Snyder, Jack F. Greenblatt, and Mark Gerstein. A bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, 302(5644):449–453, 2003. ISSN 0036-8075. doi: 10.1126/science.1087361. URL http://science.sciencemag.org/content/302/5644/449.

[51] Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional statistical models. In *Advances in Neural Information Processing Systems*, pages 1187–1195, 2013.

[52] Hawoong Jeong, Bálint Tombor, Réka Albert, Zoltan N Oltvai, and A-L Barabási. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, 2000.

[53] Hawoong Jeong, Sean P Mason, A-L Barabási, and Zoltan N Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–42, 2001.

[54] Markus Kalisch and Peter Bühlmann. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *The Journal of Machine Learning Research*, 8:613–636, May 2007. ISSN 1532-4435. URL `http://dl.acm.org/citation.cfm?id=1248659.1248681`.

[55] Nathan Kallus. Balanced policy evaluation and learning. In *Advances in Neural Information Processing Systems*, pages 8909–8920, 2018.

[56] Nathan Kallus and Angela Zhou. Policy evaluation and optimization with continuous treatments. In *International Conference on Artificial Intelligence and Statistics*, pages 1243–1251, 2018.

[57] Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 10(Jul):1391–1445, 2009.

[58] Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

[59] Joseph DY Kang, Joseph L Schafer, et al. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, 22(4):523–539, 2007.

[60] Purvesh Khatri, Marina Sirota, and Atul J Butte. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS computational biology*, 8(2): e1002375, 2012.

[61] Jon Kleinberg. The small-world phenomenon: An algorithmic perspective. In *Proceedings of the thirty-second annual ACM symposium on Theory of computing*, pages 163–170. ACM, 2000.

[62] Nishith Kumar, Md Shahjaman, Md Nurul Haque Mollah, SM Shahinul Islam, and Md Aminul Hoque. Serum and plasma metabolomic biomarkers for lung cancer. *Bioinformation*, 13(6):202, 2017.

[63] John Langford and Tong Zhang. The epoch-greedy algorithm for contextual multi-armed bandits. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, pages 817–824. Citeseer, 2007.

[64] Caiyan Li and Hongzhe Li. Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, 24(9):1175–1182, 2008.

[65] Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 661–670, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-799-8. doi: 10.1145/1772690.1772758. URL `http://doi.acm.org/10.1145/1772690.1772758`.

[66] Lihong Li, Wei Chu, John Langford, and Xuanhui Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 297–306. ACM, 2011.

[67] Tianxi Li, Elizaveta Levina, Ji Zhu, et al. Prediction models for network-linked data. *The Annals of Applied Statistics*, 13(1):132–164, 2019.

[68] Bide Liu, Xiao Gu, Tianbao Huang, Yang Luan, and Xuefei Ding. Identification of tmprss2-erg mechanisms in prostate cancer invasiveness: Involvement of mmp-9 and plexin b1. *Oncology Reports*, 37(1):201–208, 2017.

[69] Yao Liu, Omer Gottesman, Aniruddh Raghu, Matthieu Komorowski, Aldo A Faisal, Finale Doshi-Velez, and Emma Brunskill. Representation balancing mdps for off-policy

policy evaluation. In *Advances in Neural Information Processing Systems*, pages 2649–2658, 2018.

[70] Po-Ling Loh and Peter Bühlmann. High-dimensional learning of linear causal networks via inverse covariance estimation. *Journal of Machine Learning Research*, 15(1):3065–3105, 2014.

[71] David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. In *International Conference on Learning Representations*, 2017.

[72] Xinwei Ma and Jingshen Wang. Robust inference using inverse probability weighting. *arXiv preprint arXiv:1810.11397*, 2018.

[73] Dmitry M Malioutov, Jason K Johnson, and Alan S Willsky. Walk-sums and belief propagation in gaussian graphical models. *Journal of Machine Learning Research*, 7 (Oct):2031–2064, 2006.

[74] Aditya Menon and Cheng Soon Ong. Linking losses for density ratio and class-probability estimation. In *International Conference on Machine Learning*, pages 304–313, 2016.

[75] Michael Molloy and Bruce Reed. A critical point for random graphs with a given degree sequence. *Random structures & algorithms*, 6(2-3):161–180, 1995.

[76] Preetam Nandy, Alain Hauser, and Marloes H Maathuis. High-dimensional consistency in score-based and hybrid structure learning. *arXiv preprint arXiv:1507.02608*, 2015.

[77] Sahand Negahban, Pradeep Ravikumar, Martin J Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Technical report*, 2010.

[78] Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, Bin Yu, et al. A

unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

[79] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[80] Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15:2009–2053, 2014.

[81] Jing Qin. Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika*, 85(3):619–630, 1998.

[82] Assaf Rabinowicz and Saharon Rosset. Cross-Validation for Correlated Data. *arXiv e-prints*, art. arXiv:1904.02438, Apr 2019.

[83] Timothy W Randolph, Sen Zhao, Wade Copeland, Meredith Hullar, and Ali Shojaie. Kernel-penalized regression for analysis of microbiome data. *The Annals of Applied Statistics*, 12(1):540–566, 2018.

[84] Leonie Ratz, Mark Laible, Lukasz A Kacprzyk, Stephanie M Wittig-Blaich, Yanis Tolstov, Stefan Duensing, Peter Altevogt, Sabine M Klauck, and Holger Sültmann. Tmprss2: Erg gene fusion variants induce tgf-β signaling and epithelial to mesenchymal transition in human prostate cancer cells. *Oncotarget*, 8(15):25115, 2017.

[85] Pradeep Ravikumar, Martin J Wainwright, Garvesh Raskutti, Bin Yu, et al. High-dimensional covariance estimation by minimizing 1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

[86] Mark D Reid and Robert C Williamson. Composite binary losses. *Journal of Machine Learning Research*, 11(Sep):2387–2422, 2010.

[87] John Ritz and Donna Spiegelman. Equivalence of conditional and marginal regression models for clustered and longitudinal data. *Statistical Methods in Medical Research*, 13(4):309–323, 2004.

[88] Christiane M Robbins, Waibov A Tembe, Angela Baker, Shripad Sinari, Tracy Y Moses, Stephen Beckstrom-Sternberg, James Beckstrom-Sternberg, Michael Barrett, James Long, Arul Chinnaiyan, et al. Copy number and targeted mutational analysis reveals novel somatic events in metastatic prostate tumors. *Genome research*, 21(1): 47–55, 2011.

[89] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[90] Roberta R Ruela-de Sousa, Elmer Hoekstra, A Marije Hoogland, Karla C Souza Queiroz, Maikel P Peppelenbosch, Andrew P Stubbs, Karin Pelizzaro-Rocha, Geert JLH van Leenders, Guido Jenster, Hiroshi Aoyama, et al. Low-molecular-weight protein tyrosine phosphatase predicts prostate cancer outcome by increasing the metastatic potential. *European urology*, 69(4):710–719, 2016.

[91] Eric E Schadt, John Lamb, Xia Yang, Jun Zhu, Steve Edwards, Debraj GuhaThakurta, Solveig K Sieberts, Stephanie Monks, Marc Reitman, Chunsheng Zhang, et al. An integrative genomics approach to infer causal associations between gene expression and disease. *Nature genetics*, 37(7):710–717, 2005.

[92] Mark Schmidt, Alexandru Niculescu-Mizil, Kevin Murphy, et al. Learning graphical model structure using l1-regularization paths. In *AAAI*, volume 7, pages 1278–1283, 2007.

[93] Marco Scutari. Learning bayesian networks with the bnlearn r package. *Journal of Statistical Software*, 35(1):1–22, 2010. ISSN 1548-7660. doi: 10.18637/jss.v035.i03. URL `https://www.jstatsoft.org/index.php/jss/article/view/v035i03`.

[94] Xiaotong Shen, Hsin-Cheng Huang, and Wei Pan. Simultaneous supervised clustering and feature selection over a graph. *Biometrika*, 99(4):899–914, 2012.

[95] A. Shojaie, A. Jauhiainen, M. Kallitsis, and G. Michailidis. Inferring Regulatory Networks by Combining Perturbation Screens and Steady State Gene Expression Profiles. *PLoS One*, 9(2):e82393, 2014.

[96] Ali Shojaie and George Michailidis. Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. *Biometrika*, 97(3):519–538, 2010.

[97] Jeffrey A Smith and Petra E Todd. Does matching overcome lalonde's critique of nonexperimental estimators? *Journal of econometrics*, 125(1-2):305–353, 2005.

[98] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT press, 2nd edition, 2000.

[99] Chris Stark, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, and Mike Tyers. Biogrid: a general repository for interaction datasets. *Nucleic acids research*, 34(suppl 1):D535–D539, 2006.

[100] Masashi Sugiyama, Taiji Suzuki, Shinichi Nakajima, Hisashi Kashima, Paul von Bünau, and Motoaki Kawanabe. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60(4):699–746, 2008.

[101] Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density ratio estimation in machine learning*. Cambridge University Press, 2012.

[102] Seth Sullivant, Kelli Talaska, and Jan Draisma. Trek separation for gaussian graphical models. *The Annals of Statistics*, pages 1665–1685, 2010.

[103] Liang Sun, Jiaju Lü, Sentai Ding, Dongbin Bi, Kejia Ding, Zhihong Niu, and Ping Liu. Hcrp1 regulates proliferation, invasion, and drug resistance via egfr signaling in prostate cancer. *Biomedicine & Pharmacotherapy*, 91:202–207, 2017.

[104] Shiliang Sun. A survey of multi-view machine learning. *Neural computing and applications*, 23(7-8):2031–2038, 2013.

[105] Adith Swaminathan and Thorsten Joachims. The self-normalized estimator for counterfactual learning. In *advances in neural information processing systems*, pages 3231–3239, 2015.

[106] Zhiqiang Tan. Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika*, 97(3):661–682, 2010.

[107] Ambuj Tewari and Susan A Murphy. From ads to interventions: Contextual bandits in mobile health. In *Mobile Health*, pages 495–517. Springer, 2017.

[108] Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 2139–2148, 2016.

[109] Philip S Thomas. *Safe reinforcement learning*. PhD thesis, University of Massachusetts Libraries, 2015.

[110] Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning*, 65(1):31–78, 2006.

[111] Caroline Uhler, Garvesh Raskutti, Peter Bühlmann, Bin Yu, et al. Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics*, 41(2):436–463, 2013.

[112] Sara van de Geer. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

[113] Sara van de Geer, Peter Bühlmann, Yaacov Ritov, Ruben Dezeure, et al. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

[114] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

[115] Arend Voorman, Ali Shojaie, and Daniela Witten. Graph estimation with joint additive models. *Biometrika*, 101(1):85, 2014.

[116] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

[117] Yu-Xiang Wang, James Sharpnack, Alexander J Smola, and Ryan J Tibshirani. Trend filtering on graphs. *The Journal of Machine Learning Research*, 17(1):3651–3691, 2016.

[118] Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudık. Optimal and adaptive off-policy evaluation in contextual bandits. In *International Conference on Machine Learning*, pages 3589–3597, 2017.

[119] Stephen J Wright. Coordinate descent algorithms. *Mathematical Programming*, 151 (1):3–34, 2015.

[120] Hang Wu and May Wang. Variance regularized counterfactual risk minimization via variational divergence minimization. In *International Conference on Machine Learning*, pages 5349–5358, 2018.

[121] Changwon Yoo, Vesteinn Thorsson, and Gregory F Cooper. Discovery of causal relationships in a gene-regulation pathway from a mixture of experimental and observational dna microarray data. In *Pacific Symposium on Biocomputing*, volume 7, pages 498–509, 2002.

[122] Cun-Hui Zhang and Stephanie S Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.

[123] Jiji Zhang and Peter Spirtes. Strong faithfulness and uniform consistency in causal inference. In *Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence*, pages 632–639. Morgan Kaufmann Publishers Inc., 2002.

[124] Sen Zhao and Ali Shojaie. A significance test for graph-constrained estimation. *Biometrics*, 72(2):484–493, 2016.

[125] José R Zubizarreta. Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511):910–922, 2015.