Inference on Nonparametric Targets and Discrete Structures

HARVARD UNIVERSITY
Graduate School of Arts and Sciences

DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

Department of Statistics

have examined a dissertation entitled

Inference on nonparametric targets and discrete structures

presented by Lu Zhang

candidate for the degree of Doctor of Philosophy and hereby certify that it is worthy of acceptance.

Signature _____

Typed name: Prof. Lucas Janson

Signature _____

Typed name: Prof. Junwei Lu

Signature _____

Typed name: Prof. Jun Liu

Date: July 27, 2022

Inference on Nonparametric Targets and Discrete Structures

A DISSERTATION PRESENTED
BY
LU ZHANG
TO
THE DEPARTMENT OF STATISTICS

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

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
JULY 2022

Dissertation Advisor: Professor Lucas Janson Lu Zhang

Inference on Nonparametric Targets and Discrete Structures

Abstract

Many modern applications seek to understand the relationship between different variables. For example, scientists want to infer the dependence between an outcome variable and a covariate in the presence of a (possibly high-dimensional) confounding variable. In the context of graphs and networks, it is also interesting to learn the underlying discrete structures. This dissertation focuses on designing uncertainty assessment methodologies for nonparametric targets and discrete graph structures to reveal complex patterns in the underlying data-generating distributions. Chapter 1 focuses on the variable importance problem: it proposes a new approach called floodgate and applies it to the minimum mean squared error gap, an interpretable and sensitive model-free measure of variable importance. Floodgate can leverage any working regression function chosen by the user to construct asymptotic lower confidence bounds, and its adaptivity and robustness are also discussed. Chapter 2 delivers a regression inference framework: it uses the mMSE gap with respect to a closed linear subspace or a convex cone to define a diverse range of inferential targets; it utilizes the floodgate idea to conduct inference in a unified way. To demonstrate the generality and flexibility of floodgate, it presents the computation details of implementing floodgate for multiple statistical problems, including nonlinearity, interactions, deviation from shape constraints and many others. Chapter 3 studies the hub, a particular type of discrete structure. It proposes the StarTrek filter to select hub nodes over the networks and establishes FDR control guarantees in high-dimensional models. As core techniques for such FDR control problems, novel probabilistic results, i.e., Cramér-type Gaussian comparison bounds, are developed in this chapter.

Contents

Title Page **i**

Copyright **ii**

Abstract **iii**

Table of Contents **iv**

List of Tables **vi**

List of Figures **vii**

Dedication **xi**

Acknowledgments **xii**

0 Introduction **1**

1 Floodgate: Inference for Model-free Variable Importance **5**
- 1.1 Introduction . . . 6
- 1.2 Methodology . . . 12
- 1.3 Extensions . . . 27
- 1.4 Simulations . . . 33
- 1.5 Application to genomic study of platelet count . . . 42
- 1.6 Discussion . . . 45

2 Floodgate: A Swiss Army Knife for Inference in Regression **47**
- 2.1 Introduction . . . 48
- 2.2 Main Idea . . . 52
- 2.3 Different examples of subspaces . . . 57
- 2.4 Extension from subspaces to convex cones . . . 73
- 2.5 Discussion . . . 77

3 StarTrek: Combinatorial Variable Selection with False Discovery Rate Control 79
3.1 Introduction . . . 81
3.2 Methodology . . . 88
3.3 Cramér-type comparison bounds for Gaussian maxima . . . 93
3.4 Discovering hub responses in multitask regression . . . 97
3.5 Discovering hub nodes in Gaussian graphical models . . . 102
3.6 Numerical results . . . 106

Appendix A Appendix of Chapter 1 113
A.1 Proofs for main text . . . 113
A.2 An example for projection methods . . . 154
A.3 Rate results . . . 154
A.4 Applicability of the Model-X assumption . . . 170
A.5 Robustness . . . 171
A.6 Details of extending the mMSE gap . . . 179
A.7 Transporting inference to other covariate distributions . . . 187
A.8 Algorithm details for inference on the MACM gap . . . 188
A.9 Co-sufficient floodgate details . . . 200
A.10 Further simulation details . . . 222
A.11 Implementation details of genomics application . . . 233

Appendix B Appendix of Chapter 2 238
B.1 Proofs for main text . . . 238
B.2 Methodological details deferred . . . 251

Appendix C Appendix of Chapter 3 258
C.1 Proofs for FDR control . . . 259
C.2 Proofs of Cramér-type comparison bounds . . . 289
C.3 Ancillary propositions for FDR control . . . 326
C.4 Validity and power analysis of single node testing . . . 338
C.5 Tables and plots deferred from the main paper . . . 353

References 356

List of Tables

3.1 Empirical FDR 108
3.2 Power 109

C.1 $q\frac{d_0}{d}$ 354

List of Figures

1.1 Average half-widths for the linear-$\mu^\star$ (left) and nonlinear-$\mu^\star$ (right) simulations of Section 1.4.2. The coefficient amplitude is chosen to be 10 for the left panel and the sample size n equals 3000 in the right panel; see Section 1.4.1 for remaining details. Standard errors are below 0.005 (left) and 0.006 (right). 36

1.2 Average half-widths for the linear-$\mu^\star$ (left) and nonlinear-$\mu^\star$ (right) simulations of Section 1.4.3. OLS is run on the full sample. p is varied on the x-axis; see Section 1.4.1 for remaining details. Standard errors are below 0.002 (left) and 0.008 (right). 37

1.3 Average LCB values (solid lines) for floodgate and W20b in the sine function simulation of Section 1.4.4. The frequency λ is varied on the x-axis, and the solid blue line in the plot shows the true value of $\mathcal{I}$. The dashed lines correspond to the average estimator values of $\mathcal{I}$. The results are averaged over 640 independent replicates, and the standard errors are below 0.01. 38

1.4 Coverage of null (left) and non-null (right) covariates when the covariate distribution is estimated in-sample for the linear-$\mu^\star$ simulations of Section 1.4.5. See Section 1.4.1 for remaining details. Standard errors are below 0.001 (left) and 0.008 (right). . . 40

1.5 Coverage (left) and average half-widths (right) for the binary response simulations of Section 1.4.6. The explained variance proportion is varied over the x-axis. See Section 1.4.1 and 1.4.6 for remaining details. Standard errors are below 0.006 (left) and 0.001 (right). 41

1.6 Coverage (left) and average half-widths (right) for co-sufficient floodgate and original floodgate in the nonlinear-$\mu^\star$ simulations. The number of batches n_1 is varied over the x-axis. See Section 1.4.1 and 1.4.7 for remaining details. Standard errors are below 0.009 (left) and 0.002 (right). 42

1.7 Colored Chicago plot analogous to Figure 1a of Sesia et al. (2020b). The color of each point represents the floodgate LCB for the block that contains the SNP at the location indicated on the x-axis at the resolution (measured by average block width) indicated on the y-axis (note some blocks appearing in the original Chicago plot have an LCB of zero and hence are colored grey). The second panel zooms into the region of the first panel containing the largest floodgate LCB. 44

2.1 A graphical demonstration of Algorithm 4 (left) and Algorithm 5 (right). 71

3.1 Left panel: a graphical demonstration of the definition of S via four examples of a 4-vertex graph; Right panel: four different graph patterns with 6 vertices. Calculating $|S|$ yields $10, 15, 24, 51$ for (a),(b),(c),(d) respectively. 104
3.2 FDP and power plots for the StarTrek filter in the random graph. The connecting probability is varied on the x-axis. The number of samples n is chosen to be 300 and the number of connected components p equals 20. The nominal FDR level is set to be $q = 0.1$; the short blue solid lines correspond to qd_0/d, calculated by averaging over the 64 replicates. For both panels, the box plots are plotted with the black points representing the outliers. Colored points are jittered around, demonstrating how the FDP and power distribute. 110
3.3 FDP and power plots for the StarTrek filter in the random graph. The other setups are the same as Figure 3.3 except for $p = 30$. 111
3.4 The above graphs are based the estimated precision matrices (the left two plots). The adjacency matrices of the other six plots are based on the standardized estimated precision matrices but thresholded at $0.025, 0.05, 0.075$ respectively. Blue vertices represent the selected hub genes. 112
3.5 Plots of the sorted p-values ($\alpha_j, j \in [d]$) in Algorithm 8. Those blue points correspond to selected hub genes. The blue line is the rejection line of the BHq procedure. The coordinates of the plots are flipped. We abbreviate the names of the 100 genes and only show selected ones with blue colored text. 112

A.1 Coverage for the the linear-$\mu^\star$ (left) and nonlinear-$\mu^\star$ (right) simulations of Section 1.4.2. The coefficient amplitude is chosen to be 10 for the left panel and the sample size n equals 3000 in the right panel; see Section 1.4.1 for remaining details. Standard errors are below 0.007 (left) and 0.003 (right). 226
A.2 Average half-widths for the linear-$\mu^\star$ (left) and nonlinear-$\mu^\star$ (right) simulations of Section 1.4.2. The coefficient amplitude is chosen to be 5 for the left panel and the sample size n equals 1000 in the right panel; see Section 1.4.1 for remaining details. Standard errors are below 0.002 (left) and 0.01 (right). 226
A.3 Coverage for the the linear-$\mu^\star$ (left) and nonlinear-$\mu^\star$ (right) simulations of Section 1.4.2. The coefficient amplitude is chosen to be 5 for the left panel and the sample size n equals 1000 in the right panel; see Section 1.4.1 for remaining details. Standard errors are below 0.006 (left) and 0.004 (right). 227
A.4 Coverage for the linear-$\mu^\star$ (left) and nonlinear-$\mu^\star$ (right) simulations of Section 1.4.3. OLS is run on the full sample. p is varied on the x-axis; see Section 1.4.1 for remaining details. Standard errors are below 0.006 (left) and 0.004 (right). 227
A.5 Average half-widths for the linear-$\mu^\star$ (left) and nonlinear-$\mu^\star$ (right) simulations of Section 1.4.3. The splitting proportion is chosen to be 0.25 for the left panel and the sample size n equals 3000 in the right panel. p is varied on the x-axis; see Section 1.4.1 for remaining details. Standard errors are below 0.002 (left) and 0.005 (right). 228

A.6 Coverage for the linear-$\mu^\star$ (left) and nonlinear-$\mu^\star$ (right) simulations of Section 1.4.3. The splitting proportion is chosen to be 0.25 for the left panel and the sample size n equals 3000 in the right panel. p is varied on the x-axis; see Section 1.4.1 for remaining details. Standard errors are below 0.006 (left) and 0.004 (right). 228
A.7 Coverage for floodgate and W20b in the sine function simulation of Section 1.4.4. The frequency λ is varied on the x-axis, and the dotted black line in the plot shows the nominal coverage level $1 - \alpha$. The results are averaged over 640 independent replicates, and the standard errors are below 0.006. 229
A.8 Coverage (left) and average LCB values (right) for floodgate, W20b, and OLS (run on the full sample) in the linear-$\mu^\star$ simulation of Section 1.4.4. p is varied on the x-axis, and the solid blue line in the right-hand plot shows the value of $\mathcal{I}$; see Section 1.4.1 for remaining details. The results are averaged over 640 independent replicates, and the standard errors are below 0.012 (left) and 0.004 (right). 229
A.9 Coverage of null (left) and non-null (right) covariates when the covariate distribution is estimated in-sample for the nonlinear-$\mu^\star$ simulations of Section 1.4.5. See Section 1.4.1 for remaining details. Standard errors are below 0.001 (left) and 0.003 (right). . . 230
A.10 Half-width plot of non-null covariates when the covariate distribution is estimated in-sample for the linear-$\mu^\star$ (left) and nonlinear-$\mu^\star$ (right) simulations of Section 1.4.5. See Section 1.4.1 for remaining details. Standard errors are below 0.002 (left) and 0.007 (right). 230
A.11 Coverage (left) and average half-widths (right) for co-sufficient floodgate and original floodgate in the linear-$\mu^\star$ simulations. The number of batches n_1 is varied over the x-axis. See Section 1.4.1 and 1.4.7 for remaining details. Standard errors are below 0.008 (left) and 0.001 (right). 231
A.12 Average half-widths for the linear-$\mu^\star$ (left) and nonlinear-$\mu^\star$ (right) simulations of Section A.10.4. The covariate dimension $p = 1000$ and the covariate autocorrelation coefficient is varied on the x-axis; see Section 1.4.1 for remaining details. Standard errors are below 0.002 (left) and 0.009 (right). 231
A.13 Coverage for the linear-$\mu^\star$ (left) and nonlinear-$\mu^\star$ (right) simulations of Section A.10.4. The covariate dimension $p = 1000$ and the covariate autocorrelation coefficient is varied on the x-axis; see Section 1.4.1 for remaining details. Standard errors are below 0.006 (left) and 0.003 (right). 232
A.14 Average half-widths for the linear-$\mu^\star$ (left) and nonlinear-$\mu^\star$ (right) simulations of Section A.10.4. The covariate dimension $p = 500$ and the covariate autocorrelation coefficient is varied on the x-axis; see Section 1.4.1 for remaining details. Standard errors are below 0.002 (left) and 0.01(right). 232
A.15 Coverage for the linear-$\mu^\star$ (left) and nonlinear-$\mu^\star$ (right) simulations of Section A.10.4. The covariate dimension $p = 500$ and the covariate autocorrelation coefficient is varied on the x-axis; see Section 1.4.1 for remaining details. Standard errors are below 0.007 (left) and 0.004 (right). 233

A.16 Coverage (left) and average half-widths (right) for the linear-$\mu^\star$ simulations of Section A.10.4. The sample size n is varied on the x-axis; see Section 1.4.1 for remaining details. Standard errors are below 0.007 (left) and 0.003 (right). 233
A.17 Coverage (left) and average half-widths (right) for the nonlinear-$\mu^\star$ simulations of Section A.10.4. The sample size n is varied on the x-axis; see Section 1.4.1 for remaining details. Standard errors are below 0.004 (left) and 0.011 (right). 234

C.1 FDP and power plots with $p = 20$ and the nominal FDR level $q = 0.2$. The other setups are the same as Figure 3.3. 354
C.2 FDP and power plots with $p = 30$ and the nominal FDR level $q = 0.2$. The other setups are the same as Figure 3.3. 355

To my loving parents.

Acknowledgments

First, I would like to thank my advisor Prof. Lucas Janson, for his support and encouragement throughout my PhD. He is a talented researcher and thoughtful advisor. His guidance helps me develop good research tastes, broaden my statistics horizons, improve my writing and presentation skills and eventually become a better researcher. My sincere gratitude also goes to Prof. Junwei Lu, from who I have learned a great deal during our collaborations. Without his constructive suggestions and feedback on our project, my achievements would never be possible. I am extremely honored to have Prof. Jun Liu on my dissertation committee, and I would like to thank him for giving valuable feedback on my dissertation.

I am grateful to all the faculty and administrative staff in the Department of Statistics at Harvard. In particular, I would like to thank Professors Joseph Blitzstein, Mark Glickman, Pierre Jacob, Samuel Kou, Tracy Ke, and Xiao-Li Meng.

I would also like to thank my fellow students as well as friends outside the department, especially Niloy Biswas, Ambarish Chattopadhyay, Chi-Ning Chou, Louis Cammarata, Luis Campos, Chenguang Dai, Siyi Deng, Dongming Huang, Biyonka Liang, Molei Liu, Shengming Luo, Xingmei

Lou, Hansheng Jiang, Phyllis Ju, Yucong Ma, Hanyu Song, Jinjin Tian, Masatoshi Uehara, Zixuan Wang, Shuran Zheng, Shuyi Zhang, and Azeem Zaman.

Finally, I dedicate my deepest thanks to my parents for their unconditional love and care, and for always supporting me in all my choices.

0

Introduction

Canonical statistical methods often assume a parametric model and answer interesting statistical questions based on the parameter estimation and inference results. Many parametric targets are often continuous functionals of the data-generating distribution. Nowadays, lots of real-world problems in genetics, neuroscience, and finance usually come with large-scale datasets Johnstone & Titterington (2009) involving high-dimensionality, nonlinearity, and heterogeneity. Those problems seek a good understanding of the complex patterns in the underlying data-generating distributions.

They can involve a diverse range of targets, including variable importance, interactions, nonlinearity, heterogeneity, deviations from shape constraints, discrete graph structures in the network analysis, and many others. Parametric assumptions can fail in some practical examples; those methods thus suffer from invalidity and inaccuracy (Janson, 2017; Candès et al., 2018a). As for targets encoding discrete structures, existing approaches are not even applicable since they are often designed for continuous (and local) functionals. This dissertation tackles the aforementioned challenges and focuses on two types of problems: inferring nonparametric targets in regression and discovering network hub structures.

Given a response variable and some covariates, the first question people usually ask is the importance of a given covariate or a given group of covariates in the conditional relationship between the response variable and the covariates. This question is answered through conducting hypothesis tests or constructing confidence bounds with some chosen notions of variable importance. When assuming a parametric model, variable importance can be defined in terms of the model parameters. Hypothesis tests and confidence bounds for those parameters have been sufficiently studied in the literature (Bühlmann et al., 2013; Nickl et al., 2013; Zhang & Zhang, 2014; Van de Geer et al., 2014; Javanmard & Montanari, 2014a; Bühlmann et al., 2015; Dezeure et al., 2017; Zhang & Cheng, 2017). More recently, there have been works considering semi-parametric or nonparametric approaches (Berk et al., 2013; Taylor et al., 2014; Robins et al., 2008, 2009; Li et al., 2011; Huang et al., 2020b; Williamson et al., 2019, 2020). The validity of those approaches often relies on the conditional relationship being estimated sufficiently well, e.g., a certain rate consistency of the mean squared error (MSE) of estimating the true regression function (Williamson et al., 2019, 2020). Such requirements limit the generality and flexibility of the methodologies. To avoid the limitations, we introduce a new statistical idea called floodgate to conduct model-free inference. Chapter 1, summarizing the work in Zhang & Janson (2020), considers an interpretable estimand, the minimum mean squared error (mMSE) gap, which measure the variable importance in a way

that does not depends on any modelling assumptions, and is sensitive to nonlinearities and interactions. The application of floodgate to the variable importance problem produces confidence bounds for the mMSE gap, by leveraging a user-chosen working regression function (allowing the use of modern machine learning algorithms or the incorporation of qualitative domain knowledge). It is shown that the validity of floodgate inference does not depend on particular sets of assumptions about the underlying distributions or the quality of the working regression function. For example, two different types of assumptions (the model-X (Candès et al., 2018a) assumption and the double robustness (Chernozhukov et al., 2018a) assumption) can both guarantee the coverage of the floodgate confidence bounds. The adaptivity and robustness of floodgate and some interesting extensions are also discussed. In addition to variable importance, there are also many other targets such as nonlinearity (Kotchoni, 2018), interactions (Egami & Imai, 2018), heterogeneity (Angrist, 2004) and deviations from shape-constraints (Chetverikov, 2019). Inference for them is insufficiently explored, let alone model-free inference. To solve these problems, Chapter 2 presents a new regression inference framework based on the floodgate idea. Specifically, it defines the mMSE gap with respect to a closed linear subspace to characterize a class of interpretable model-free targets and provides the computation details of running floodgate in multiple examples, including nonlinearity, interactions, and many others.

Variable selection on large-scale networks has been studied a lot and applies to fields such as social networks, neuroscience and genetics (Newman et al., 2002; Luscombe et al., 2004; Rubinov & Sporns, 2010). Existing literature mainly focuses on continuous and local functionals especially the graph edges (Cai & Ma, 2013; Cai & Zhang, 2016; Janková & van de Geer, 2017; Yang et al., 2018; Feng & Ning, 2019; Liu, 2013; Xia et al., 2018). However, discrete structures in networks like hubs can arise from many real-world examples such as brain networks (Shaw et al., 2008) and gene co-expression networks (Yuan et al., 2017). Discovering the hubs can provide scientists and practitioners with a better understanding of the underlying patterns in those networks. We for-

mulate this as a combinatorial variable selection problem where we would like to select nodes with degrees larger than a prespecified thresholding level on the graph with false discovery rate (FDR) control guarantees. This problem brings new challenges in two aspects. First, it is unclear how to construct appropriate test statistics for testing the degree of a single node. Second, simultaneously testing all the nodes unavoidably gives rise to complicated dependence that is hard to quantify. This is because the computation of any reasonable test statistic for a single node has to involve the whole graph. Chapter 3, summarizing the work in Zhang & Lu (2021), tackles the challenges. In methodology, it proposes the StarTrek filter involving the maximum statistics and their quantile estimates via the Gaussian multiplier bootstrap (Chernozhukov et al., 2013). In theory, it establishes asymptotic FDR control in high dimensions via proving accurate bounds on the approximation errors of the quantile estimates and characterizing the dependence structures among the test statistics. Such results build on our novel probabilistic results, two different Cramér-type Gaussian comparisons bounds. These comparison bounds differ from the Kolmogorov distance bounds in Chernozhukov et al. (2015) and concern the relative difference between the distribution functions of two high-dimensional Gaussian random vectors. To demonstrate the idea, we also apply the StarTrek filter to two specific high-dimensional settings: Gaussian graphical models and multi-task regression with linear models.

1

Floodgate: Inference for Model-free Variable Importance

CONTRIBUTION

This chapter is based on a manuscript Zhang & Janson (2020), jointly with Prof. Lucas Janson.

Abstract

Many modern applications seek to understand the relationship between an outcome variable Y and a covariate X in the presence of a (possibly high-dimensional) confounding variable Z. Although much attention has been paid to testing *whether* Y depends on X given Z, in this paper we seek to go beyond testing by inferring the *strength* of that dependence. We first define our estimand, the minimum mean squared error (mMSE) gap, which quantifies the conditional relationship between Y and X in a way that is deterministic, model-free, interpretable, and sensitive to nonlinearities and interactions. We then propose a new inferential approach called *floodgate* that can leverage any working regression function chosen by the user (allowing, e.g., it to be fitted by a state-of-the-art machine learning algorithm or be derived from qualitative domain knowledge) to construct asymptotic confidence bounds, and we apply it to the mMSE gap. We additionally show that floodgate's accuracy (distance from confidence bound to estimand) is adaptive to the error of the working regression function. We then show we can apply the same floodgate principle to a different measure of variable importance when Y is binary. Finally, we demonstrate floodgate's performance in a series of simulations and apply it to data from the UK Biobank to infer the strengths of dependence of platelet count on various groups of genetic mutations.

Keywords. Variable importance, effect size, model-X, heterogeneous treatment effects, heritability.

1.1 Introduction

1.1.1 Problem Statement

Scientists looking to better-understand the relationship between a response variable Y of interest and a covariate X in the presence of confounding variables $Z = (Z_1, \ldots, Z_{p-1})$ often start by asking *how important* X is in this relationship. Although this question is sometimes simplified

by statisticians to the binary question of 'is X important or not?', a more informative and useful inferential goal is to provide inference (i.e., confidence bounds) for an interpretable real-valued measure of variable importance (MOVI). The canonical approach of assuming a parametric model for $Y \mid X, Z$ will usually provide obvious MOVI candidates in terms of the model parameters, but the simple models for which it is known how to construct confidence intervals (e.g., low-dimensional or ultra-sparse generalized linear models) often provide at best very coarse approximations to the true $Y \mid X, Z$ (as evidenced by the marked predictive outperformance of nonparametric machine learning methods in many domains), resulting in undercoverage due to violated assumptions *and* lost power due to insufficient capacity to capture complex relationships. This raises the motivating question for this paper: **what is an interpretable, sensitive, and model-free measure of variable importance and how can we provide valid and narrow confidence bounds for it?**

1.1.2 Our contribution

The main contribution of this paper is to introduce *floodgate*, a method for inference of the minimum mean squared error (mMSE) gap, which satisfies the following high-level objectives which we believe are fairly universal for the task at hand.

(Sensitivity) The mMSE gap is strictly positive unless $\mathbb{E}[Y \mid X, Z] \overset{a.s.}{=} \mathbb{E}[Y \mid Z]$, and is large whenever X explains a lot of the variance in Y not already explained by Z alone, making it sensitive to arbitrary nonlinearities and interactions in Y's relationship with X.

(Interpretability) The mMSE gap has simple predictive, explanatory, and causal interpretations for Y's relationship with X, is a functional of *only* the joint distribution of (Y, X, Z), and is exactly zero when $Y \perp\!\!\!\perp X \mid Z$.

(Validity) We first prove floodgate's asymptotic validity assuming the user knows the distribution of $X \mid Z$, but with essentially no other assumptions (in particular we require no smoothness,

sparsity, or other constraints on $\mathbb{E}[Y \mid X, Z]$ that would ensure its learnability at *any* geometric rate). However, to emphasize that the floodgate idea is not tied to such assumptions, we also provide a version of floodgate valid under double-robustness-type assumptions.

(Accuracy) Floodgate derives accuracy from flexibility by allowing the user to estimate $\mathbb{E}[Y \mid X, Z]$ in whatever way they like, and we prove that the accuracy of inference is adaptive to the mean squared error (MSE) of that estimate.

In a bit more detail, we (in Section 1.2) define the mMSE gap as an interpretable and model-free MOVI (Section 1.2.1) and present a method, *floodgate*, to construct asymptotic lower confidence bounds for it that provides the user absolute latitude to leverage any domain knowledge or advanced machine learning algorithms to make those bounds as tight as possible (Section 1.2.2). We consider upper confidence bounds (Section 1.2.3), address computational considerations (Section 1.2.4), theoretically characterize the width of floodgate's confidence bounds (Section 1.2.5), and briefly address some immediate generalizations (Section 1.2.6).

We then proceed to extensions of floodgate (Section 1.3), first presenting an alternative MOVI that we can similarly construct asymptotic confidence bounds for when Y is binary (Section 1.3.1). Second, we present a modification of floodgate that, for certain models, allows asymptotic inference even when X's distribution is only known up to a parametric model (Section 1.3.2) and apply it to multivariate Gaussian (Section 1.3.2) and discrete Markov chain (Section 1.3.2) covariate models.

Finally we demonstrate floodgate's performance and support our theory with simulations (Section 1.4) and an application to data from the UK Biobank (Section 1.5). We end with a discussion of the future research directions opened by this work (Section 1.6). All proofs are deferred to the appendix.

1.1.3 Related work

Many existing works consider *marginal* variable importance, i.e., not accounting for the presence of Z in the relationship between Y and X (Hirschfeld, 1935; Gretton et al., 2005, 2007; Székely et al., 2007; Székely & Rizzo, 2013; Heller et al., 2013; Shao & Zhang, 2014; Wang et al., 2017; Chatterjee, 2021; Deb & Sen, 2021), including some that measure that importance via differences in conditional means in a way resembling our mMSE gap (Shao & Zhang, 2014). Such approaches address a very different statistical question, and so we focus our literature review on works that, like us, consider conditional variable importance.

The standard approach to conditional statistical inference in regression is to assume a parametric model for $Y \mid X, Z$, often a generalized linear model (GLM) or cousin thereof. With $Y \mid X, Z$ so parameterized, it is usually straightforward to define a parametric MOVI and a large body of literature is available to provide asymptotic inference for such parametric MOVIs (see, for example, Bühlmann et al. (2013); Nickl et al. (2013); Zhang & Zhang (2014); Van de Geer et al. (2014); Javanmard & Montanari (2014a); Bühlmann et al. (2015); Dezeure et al. (2017); Zhang & Cheng (2017)). However, when the parametric $Y \mid X, Z$ model is misspecified even slightly, the associated parametric MOVI becomes ill-defined, reducing its interpretability. Furthermore, many $Y \mid X, Z$ models are too simple to capture or detect nonlinearities that may be present in real-world data sets.

One approach to addressing the shortcomings of parametric inference is to generalize the parameters of common parametric models to be well-defined in a much larger nonparametric model class. For example, under mild moment conditions one can generalize the parameters in a linear model for $Y \mid X, Z$ as parameters in the least-squares *projection* to a linear model of any $Y \mid X, Z$ distribution (Berk et al., 2013; Taylor et al., 2014; Buja & Brown, 2014; Buja et al., 2015; Rinaldo et al., 2019a; Lee et al., 2016; Buja et al., 2019a,b). Such a linear projection MOVI can be hard to interpret because it will in general have a non-zero value even when $Y \perp\!\!\!\perp X \mid Z$; see Appendix A.2 for a simple example. Another example of a generalized parameter is the expected conditional covariance functional $\mathbb{E}\left[\mathrm{Cov}\left(Y, X \mid Z\right)\right]$ (see, for example, Robins et al. (2008, 2009); Li et al. (2011); Robins et al.

(2017); Newey & Robins (2018); Shah & Peters (2020); Chernozhukov et al. (2018b); Liu et al. (2019a); Katsevich & Ramdas (2020)), which represents a generalization of the linear coefficient in a *partially* linear model. $\mathbb{E}\left[\mathrm{Cov}\left(Y, X \mid Z\right)\right]$ always equals zero when $Y \perp\!\!\!\perp X \mid Z$, but it shares the shortcoming of linear projection MOVIs that it lacks sensitivity to capture nonlinearities or interactions in Y's relationship with X. That is, both MOVIs mentioned in this paragraph will assign any non-null variable that influences Y nonlinearly or through interactions with other covariates a value that can severely underrate that variable's true importance, and can even assign a variable the MOVI value zero when Y is a deterministic non-constant function of it.

A second approach has been to infer model-free MOVIs defined through machine learning algorithms fitted to part of the data itself (Lei et al., 2018; Fisher et al., 2019; Watson & Wright, 2019). By leveraging the expressiveness of machine learning, such a MOVI can be made sensitive to nonlinearities and interactions but is itself *random* and depends both on the data and the choice of machine learning algorithm. This poses a challenge for interpretability and in particular for replicability, since even *identical* analyses run on two independent data sets that are *identically-distributed* will provide inferences for *different* MOVI values.

Another line of work (Castro et al., 2009; Štrumbelj & Kononenko, 2014; Owen & Prieur, 2017; Lundberg et al., 2020; Covert et al., 2020; Williamson & Feng, 2020) considers MOVIs based on the classical form of the Shapley value (Shapley, 1953; Charnes et al., 1988), which in general assigns a non-zero MOVI value to covariates X with $Y \perp\!\!\!\perp X \mid Z$, making it hard to interpret its value mechanistically or causally (though it has some appealing properties for a *predictive* interpretation).

An interesting new proposal for a model-free MOVI was made in Azadkia & Chatterjee (2019). Their MOVI has the distinction that it equals zero if and only if $Y \perp\!\!\!\perp X \mid Z$ *and* it attains the maximum value 1 if Y is almost surely a measurable function of X given Z. More recently, Huang et al. (2020b) proposed a larger class of MOVIs satisfying the same properties. However, both papers focus on consistent estimators and do not provide confidence bounds for their MOVIs.

As we will detail in Section 1.2.1, the MOVI we provide inference for, the mMSE gap, does not suffer from the drawbacks of the MOVIs described in the previous paragraphs, and indeed the same MOVI has been considered before. In the sensitivity analysis literature it is called the "total-effect index" (Saltelli et al., 2008) but to our knowledge its inference (confidence lower- or upper-bounds) is not considered there. In one of the Shapley value papers (Covert et al., 2020) a generalization of the mMSE gap is used as the input to the Shapley value calculation, but again inferential results (for the mMSE gap or its Shapley version) are not considered in that paper. Otherwise, Williamson et al. (2019) appears to be the first to consider inference for the mMSE gap (this inference is then used with neural networks in Feng et al. (2018)), but the asymptotic normality theory their coverage guarantee relies on fails at the boundary of the parameter space, i.e., the important case of when the mMSE gap is zero, or the variable is unimportant. A recent follow-up work (Williamson et al., 2020) addresses this limitation by combining estimators on two disjoint subsets of the data (though their inference still requires the *group* mMSE gap of the entire covariate vector to be positive). Our different approach avoids altogether this issue when the mMSE gap is zero so that our inference is valid for any value of the mMSE gap (group or otherwise), and although we also use data splitting, we do so in a way that seems to lead to significantly reduced variance (and hence more accurate inference) relative to Williamson et al. (2020), as we show in Section 1.4.4.

1.1.4 Notation

For two random variables A and B defined on the same probability space, let $P_{A \mid B}$ denote the conditional distribution of $A \mid B$. Denote the $(1-\alpha)$th quantile of the standard normal distribution by z_α. Let $\chi^2\left(P \| Q\right)$ denote the χ^2 divergence $\int_\Omega (\frac{dP}{dQ} - 1)^2 dQ$ between two distributions P, Q on the probability space Ω. Let $[n]$ denote the set $\{1, \ldots, n\}$.

1.2 Methodology

1.2.1 Measuring variable importance with the mMSE gap

We begin by defining the MOVI that we will provide inference for in this paper.

Definition 1.2.1 (Minimum mean squared error gap)**.** *The* minimum mean squared error (mMSE) gap *for variable X is defined as*

$$\mathcal{I}^2 = \mathbb{E}\left[(Y - \mathbb{E}[Y \mid Z])^2\right] - \mathbb{E}\left[(Y - \mathbb{E}[Y \mid X, Z])^2\right] \tag{1.2.1}$$

whenever all the above expectations exist.

We will at times refer to either $\mathcal{I}^2$ or $\mathcal{I}$ as the mMSE gap when it causes no confusion. Although the same MOVI has been used before (see Section 1.1.3), we provide here a number of equivalent definitions/interpretations which we have not seen presented together before.

- Equation (1.2.1) has a direct *predictive* interpretation as the increase in the achievable or minimum MSE for predicting Y when X is removed.
- The mMSE gap can also be interpreted as the decrease in the *explainable variance* of Y without X:

$$\mathcal{I}^2 = \mathrm{Var}\left(\mathbb{E}[Y \mid X, Z]\right) - \mathrm{Var}\left(\mathbb{E}[Y \mid Z]\right). \tag{1.2.2}$$

- When X is viewed as a treatment level for Y and Z is a set of measured confounders, $\mathcal{I}$ can be seen as an *expected squared treatment effect*:

$$\mathcal{I}^2 = \frac{1}{2}\mathbb{E}_{x_1,x_2,Z}\left[\left(\mathbb{E}[Y \mid X = x_1, Z] - \mathbb{E}[Y \mid X = x_2, Z]\right)^2\right]. \tag{1.2.3}$$

  where x_1 and x_2 are independently drawn from $P_{X|Z}$ in the outer expectation.

- We can also rewrite the mMSE gap as:

$$\mathcal{I}^2 = \mathbb{E}\left[\left(\mathbb{E}\left[Y \mid Z\right] - \mathbb{E}\left[Y \mid X, Z\right]\right)^2\right] \tag{1.2.4}$$

  and interpret $\mathcal{I}$ as the ℓ_2 distance between the two regression functions $\mathbb{E}\left[Y \mid Z\right]$ and $\mathbb{E}\left[Y \mid X, Z\right]$.

- Lastly, we remark that $\mathcal{I}^2$ also admits a very compact (if less immediately interpretable) expression:

$$\mathcal{I}^2 = \mathbb{E}\left[\mathrm{Var}\left(\mathbb{E}\left[Y \mid X, Z\right] \mid Z\right)\right]. \tag{1.2.5}$$

In light of these multiple alternative expressions, we find the mMSE gap remarkably interpretable. Note that it only requires the existence of some low-order conditional and unconditional moments of Y to be well-defined, and its value is invariant to any fixed translation of Y and to the replacement of X or Z by any fixed bijective function of itself. Furthermore, the mMSE gap is zero if and only if $\mathbb{E}\left[Y \mid X, Z\right] \overset{a.s.}{=} \mathbb{E}\left[Y \mid Z\right]$, and in particular it is exactly zero when $Y \perp\!\!\!\perp X \mid Z$ and strictly positive if $\mathbb{E}\left[Y \mid X, Z\right]$ depends at all on X, allowing it to fully capture arbitrary nonlinearities and interactions in $\mathbb{E}\left[Y \mid X, Z\right]$.

Note that $\mathcal{I}$ has the same units as Y, which can help interpretation when Y's units are meaningful (much like it does for the average treatment effect in causal inference). However, if a unitless quantity is preferred, such as for comparison between MOVIs across Ys with different units, we can also measure variable importance by and extend our methodology to a standardized version of $\mathcal{I}^2$, namely, $\mathcal{I}^2/\mathrm{Var}\left(Y\right)$. In fact, with some more work, we can even extend our inferential results to a version of the mMSE gap which is invariant to transformations of Y, or versions that are zero if *and only if* $Y \perp\!\!\!\perp X \mid Z$; see Section 1.2.6 and Appendix A.6 for details, with Appendix A.6.2 extending our results to the kernel partial correlation of Huang et al. (2020b).

1.2.2 Floodgate: asymptotic lower confidence bounds for the mMSE gap

As can be seen by Equation (1.2.5), the mMSE gap is a nonlinear functional of the true regression function $\mu^\star(x,z) := \mathbb{E}[Y \mid X = x, Z = z]$. Hence if we had a sufficiently-well-behaved estimator $\widehat{\mu}$ for $\mu^\star$ (e.g., asymptotically normal or consistent at a sufficiently-fast geometric rate), there would be a number of existing tools in the literature (e.g., the delta method, influence functions) that we could use to provide inference for the mMSE gap. But such estimation-accuracy assumptions are only known to hold for a very limited class of regression estimators, and in particular preclude most modern machine learning algorithms and methods that integrate hard-to-quantify domain knowledge, which are exactly the types of powerful regression estimators we would most like to leverage for accurate inference.

However, given the centrality of $\mu^\star$ in the definition of the mMSE gap, it seems we need to at least implicitly estimate it with some working regression function μ. And even if we avoid assumptions on μ's accuracy, if we want to provide rigorous inference then we ultimately still need *some* way to relate μ to $\mathcal{I}$, which is a function of $\mu^\star$. We address this issue in the context of constructing a lower confidence bound (LCB) for the mMSE gap. The key idea proposed in this paper is to use a functional, which we call a *floodgate*, to relate *any* μ to $\mathcal{I}$. In particular, we will shortly introduce a $f(\mu)$ such that for *any* μ,

(a) $f(\mu) \leq \mathcal{I}$

(b) we can construct a lower confidence bound L for $f(\mu)$.

Then by construction L will also constitute a valid LCB for $\mathcal{I}$. The term *floodgate* comes from metaphorically thinking of constructing a LCB as preventing flooding ($L > \mathcal{I}$, i.e., miscoverage) by keeping the water level (L) below a critical threshold ($\mathcal{I}$) under arbitrary weather conditions (μ, or more specifically, μ's error, which we may not expect to be able to control well). Then by con-

trolling L below $\mathcal{I}$ for any μ, f acts as a floodgate, and we also use the same name for the inference procedure we derive from f.

In particular, for any (nonrandom) function $\mu : \mathbb{R}^p \to \mathbb{R}$, define

$$f(\mu) := \frac{\mathbb{E}\left[\mathrm{Cov}(\mu^\star(X, Z), \mu(X, Z) \mid Z)\right]}{\sqrt{\mathbb{E}\left[\mathrm{Var}(\mu(X, Z) \mid Z)\right]}}, \tag{1.2.6}$$

where by convention we define $0/0 = 0$ so that $f(\mu)$ remains well-defined when the denominator of (1.2.6) is zero. It is not hard to see that f tightly satisfies the lower-bounding property (a) and we formalize this in the following lemma which is proved in Appendix A.1.1.

Lemma 1.2.2. *For any μ such that $f(\mu)$ exists, $f(\mu) \leq \mathcal{I}$, with equality when $\mu = \mu^\star$.*

In order to establish property (b) of f, we first take a *model-X* approach (Janson, 2017; Candès et al., 2018a): we assume we know $P_{X|Z}$ but avoid assumptions on $Y \mid X, Z$. We start with such a model-X assumption because its simplicity helps elucidate the key ideas underlying the floodgate method, but floodgate is not tied to such assumptions, and indeed we present alternative versions of floodgate that operate under different assumptions later in the paper (Section 1.3.2's version somewhat relaxes the assumed knowledge of $P_{X|Z}$ without requiring any new assumptions and Remark 1.2.3.1's version relies on a *double-robust* set of assumptions). That said, the model-X assumption is sometimes reasonable and has been used before in a number of applications (see Appendix A.4 for elaboration and examples), including in genomics like in the application presented in Section 1.5, and we theoretically (Appendix A.5) and numerically (Section 1.4.5) characterize model-X floodgate's robustness to misspecification of $P_{X|Z}$. Knowing $P_{X|Z}$ and μ means that, given data $\{(X_i, Z_i, Y_i)\}_{i=1}^n$, we also know $\{V_i := \mathrm{Var}(\mu(X_i, Z_i) \mid Z_i)\}_{i=1}^n$ which are i.i.d. and unbiased for the squared denominator in (1.2.6). And if we rewrite the numerator as

$$\mathbb{E}\left[\mathrm{Cov}(\mu^\star(X, Z), \mu(X, Z) \mid Z)\right] = \mathbb{E}\left[Y\left(\mu(X, Z) - \mathbb{E}\left[\mu(X, Z) \mid Z\right]\right)\right], \tag{1.2.7}$$

then we see we also know $\{R_i := Y_i\big(\mu(X_i, Z_i) - \mathbb{E}\left[\mu(X, Z_i) \mid Z_i\right]\big)\}_{i=1}^n$ which are i.i.d. and unbiased for the numerator. Thus for any given μ, we can use sample means of R_i and V_i to asymptotically-normally estimate both expectations in Equation (1.2.6), and then combine said estimators through the delta method to get an estimator of $f(\mu)$ whose asymptotic normality facilitates an immediate asymptotic LCB. This strategy is spelled out in Algorithm 1 and Theorem 1.2.3 establishes its asymptotic coverage. We pause to mention a simple but important point: when $\mu(X, Z)$ does not depend on X at all, then $f(\mu) = 0$ and all the V_i and R_i are zero with probability 1, making floodgate's LCB computed in Algorithm 1 deterministically zero as well. This implies that when the regression algorithm for obtaining μ is sparse, in the sense that it only depends on a fraction of its inputs, then floodgate will produce LCBs of zero for many of the covariates. For those covariates, coverage will hold *deterministically*, and hence floodgate will have average coverage even higher than the nominal $1 - \alpha$, as observed in some simulations in Section 1.4.

Algorithm 1 Floodgate

Input: Data $\{(Y_i, X_i, Z_i)\}_{i=1}^n$, $P_{X|Z}$, a working regression function $\mu : \mathbb{R}^p \to \mathbb{R}$, and a confidence level $\alpha \in (0, 1)$.
Compute $R_i = Y_i\big(\mu(X_i, Z_i) - \mathbb{E}\left[\mu(X_i, Z_i) \mid Z_i\right]\big)$ and $V_i = \mathrm{Var}\left(\mu(X_i, Z_i) \mid Z_i\right)$ for each $i \in [n]$, and their sample mean $(\bar{R}, \bar{V})$ and sample covariance matrix $\widehat{\Sigma}$, and compute $s^2 = \frac{1}{\bar{V}}\left[\left(\frac{\bar{R}}{2\bar{V}}\right)^2 \widehat{\Sigma}_{22} + \widehat{\Sigma}_{11} - \frac{\bar{R}}{\bar{V}}\widehat{\Sigma}_{12}\right]$.

Output: Lower confidence bound $L_n^\alpha(\mu) = \max\left\{\frac{\bar{R}}{\sqrt{\bar{V}}} - \frac{z_\alpha s}{\sqrt{n}},\, 0\right\}$, with the convention that $0/0 = 0$.

Theorem 1.2.3 (Floodgate validity)**.** *For any given working regression function $\mu : \mathbb{R}^p \to \mathbb{R}$ and i.i.d. data $\{(Y_i, X_i, Z_i)\}_{i=1}^n$, if $\mathbb{E}[Y^4]$, $\mathbb{E}[\mu^4(X, Z)] < \infty$, then $L_n^\alpha(\mu)$ from Algorithm 1 satisfies*

$$\liminf_{n\to\infty} \mathbb{P}\left(L_n^\alpha(\mu) \le \mathcal{I}\right) \ge 1 - \alpha.$$

The proof of Theorem 1.2.3 can be found in Appendix A.1.1. Fourth moments (as opposed to the usual second moments for the CLT) are required because the estimand itself involves the expectations of $Y\mu(X,Z)$ and $\mu^2(X,Z)$. With higher moment conditions, we can apply relatively recent Berry–Esseen-type results for the delta method (Pinelis et al., 2016) to strengthen the pointwise asymptotic coverage of Theorem 1.2.3 to have a rate of $n^{-1/2}$; see Appendix A.3 for details. We note that in both Algorithm 1 and Theorem 1.2.3, Y can be everywhere replaced by $Y - g_0(Z)$ for any non-random function g_0 (e.g., $\mathbb{E}\left[\mu(X,Z) \mid Z = z\right]$ would be a natural choice), which can reduce the variance of the R_i terms and hence improve the LCB.

Remark 1.2.3.1 (Doubly robust floodgate)**.** *Although for ease of exposition we have presented Algorithm 1 and Theorem 1.2.3 under the model-X assumption that $P_{X|Z}$ is known exactly, we emphasize here that the underlying idea of floodgate is* not *tied to this assumption. To reiterate, the key conceptual contribution of this paper is to introduce a lower-bounding functional $f(\mu)$ for $\mathcal{I}$ such that $f(\mu)$ provides a tractable statistical target to obtain a LCB for. To underscore this point, we present here a version of floodgate following the same principle but that is valid under standard double-robust assumptions instead of the aforementioned model-X assumption. Consider the following functional that depends not only on a working regression function $\mu(x,z)$, but also some Q_y estimating the true $P_{Y|Z}$ and some Q_x estimating the true $P_{X|Z}$:*

$$f_{Q_y,Q_x}(\mu) := \frac{\mathbb{E}\left[(Y - \mathbb{E}_{Q_y}\left[Y \mid Z\right])(\mu(X,Z) - \mathbb{E}_{Q_x}\left[\mu(X,Z) \mid Z\right])\right]}{\sqrt{\mathbb{E}\left[(\mu(X,Z) - \mathbb{E}_{Q_x}\left[\mu(X,Z) \mid Z\right])^2\right]}}, \tag{1.2.8}$$

where $\mathbb{E}_{Q_x}$ (resp. $\mathbb{E}_{Q_y}$) denotes expectation with respect to Q_x (resp. Q_y) as opposed to the true data-generating distribution, and by convention we again define $0/0 = 0$. Given Q_y, Q_x, and μ, i.i.d. unbiased estimates analogous to R_i and V_i in Algorithm 1 of the numerator and squared denominator, respectively, of $f_{Q_y,Q_x}(\mu)$ can be computed from each data point under no assumptions whatsoever, thus allowing the exact same kind of LCB as in Algorithm 1 to be computed for $f_{Q_y,Q_x}(\mu)$. It now just

remains to check that $f_{Q_y,Q_x}(\mu)$ lower-bounds $\mathcal{I}$.

Lemma 2.3. *For any μ, Q_y, Q_x such that Q_x is absolutely continuous with respect to $P_{X|Z}$ and $f_{Q_y,Q_x}(\mu)$ exists, we have that $f_{Q_y,Q_x}(\mu) \leq \mathcal{I} + \Delta$, where*

$$\Delta = \sqrt{\mathbb{E}\left[(\mathbb{E}\left[Y \mid Z\right] - \mathbb{E}_{Q_y}\left[Y \mid Z\right])^2\right] \mathbb{E}\left[w_\mu(X,Z)\chi^2\left(Q_x \| P_{X|Z}\right)\right]} \tag{1.2.9}$$

and $w_\mu(X,Z) = \frac{(\mu(X,Z) - \mathbb{E}[\mu(X,Z) \mid Z])^2}{\mathbb{E}[(\mu(X,Z) - \mathbb{E}[\mu(X,Z) \mid Z])^2]}$ is non-negative, has mean 1, and does not depend on Q_y or Q_x, and we again define $0/0 = 0$. Furthermore, $f_{Q_y,P_{X|Z}}(\mu) = f(\mu)$ and thus $f_{Q_y,P_{X|Z}}(\mu^\star) = \mathcal{I}$ (for any Q_y).

The proof can be found in Appendix A.1.1. Lemma 2.3 says that $f_{Q_y,Q_x}(\mu)$ only fails to lower-bound $\mathcal{I}$ to an extent bounded by the square root of the product of two terms: the MSE of $\mathbb{E}_{Q_y}\left[Y \mid Z\right]$ and the weighted χ^2 error of Q_x. The same result also holds if we move $w_\mu(X,Z)$ in Equation (1.2.9) from the second term to the first term; see Equation (A.1.33). As the first term measures the error in modeling $Y \mid Z$ and the second term measures the error in modeling $X \mid Z$, the square root of their product Δ is exactly what we would expect to be bounded as $o(n^{-1/2})$ under standard double-robustness assumptions (see, e.g., Chernozhukov et al. (2018a)). And indeed, since the LCB for $f_{Q_y,Q_x}(\mu)$ will be $\Omega(n^{-1/2})$ below $f_{Q_y,Q_x}(\mu)$, $\Delta = o(n^{-1/2})$ implies asymptotic coverage exactly as in Theorem 1.2.3.

Remark 1.2.3.2 (Floodgate's validity in high dimensions)**.** *Again for ease of exposition, Theorem 1.2.3 establishes floodgate's pointwise asymptotic coverage for a fixed μ and a fixed (and hence fixed-dimensional) distribution for (Y, X, Z). It is certainly of interest to also consider the high-dimensional regime where the data-generating distribution (including the covariate dimension p) and the working regression function μ both depend on n, but it turns out that this setting is actually not very different from the simpler setting of Theorem 1.2.3. To see this, first note that Theorem 1.2.3 relies only*

on Lemma 1.2.2 ($f(\mu) \leq \mathcal{I}$) and a central limit theorem (CLT) applied to the 2-dimensional mean of the i.i.d. pairs (R_i, V_i). But Lemma 1.2.2 is non-asymptotic, and hence $f(\mu) \leq \mathcal{I}$ still holds even if μ varies with n. And the pairs (R_i, V_i) remain i.i.d. and 2-dimensional even as μ and the distribution of (Y, X, Z) vary with n, so all that is needed for floodgate's validity is a 2-dimensional i.i.d. triangular array CLT, which only requires that the 2-dimensional random variables (R_i, V_i) remain "well-behaved". In Appendix A.3 we show in fact an even stronger (non-asymptotic) result, which, similarly to Theorem 1.2.3, only requires certain moments of Y and $\mu(X, Z)$ to remain bounded (although the result in Appendix A.3 requires a bound on higher moments than Theorem 1.2.3 so that recent Berry–Eseen-type results for the delta method can be applied to bound floodgate's undercoverage at a rate of $n^{-1/2}$). In fact, it is even sufficient to replace the bound on $\mu(X, Z)$'s absolute moment with a bound on that of its conditional residual $h(X, Z) := \mu(X, Z) - \mathbb{E}[\mu(X, Z) \mid Z]$. Note that h only really measures the contribution from the single covariate X to the whole working regression function μ, even when Z is high-dimensional. Hence, we believe that assuming that Y's and $h(X, Z)$'s moments do not explode, even in high dimensions (recall Y and $h(X, Z)$ remain 1-dimensional regardless of the dimension of the data), seems quite mild in practice. For instance, if $|Y|$ is a bounded random variable (as it often will be in practice), then as long as μ is winsorized at some level (which, as long as the level is at least as large as $|Y|$'s bound, can only improve μ's performance) (Rinaldo et al., 2019a), then floodgate's asymptotic validity is automatically ensured in the most general high-dimensional regime. Even when Y is unbounded, we would usually not expect the moments of Y or $h(X, Z)$ to diverge. Indeed in Section 1.4.3 we conduct high-dimensional simulations with unbounded Y and μ fitted via various parametric and nonparametric machine learning algorithms, yet floodgate's coverage remains empirically valid regardless of the dimension.

Remark 1.2.3.3 (Choosing μ). *The final missing piece in our LCB procedure is the choice of μ. In terms of how to obtain a working regression function μ, the flexibility of our procedure thus far finally pays off: μ can be chosen in* any *way that does not depend on the data used for inference. Normally we*

expect this to be achieved through data-splitting, i.e., a set of data samples is divided into two independent parts, and one part is used to produce an estimate μ of $\mu^\star$ while floodgate is applied to the other part with input μ; we will explore this strategy in simulations in Section 1.4. But in general, μ can be derived from any independent source, including mechanistic models or data of a completely different type than that used in floodgate (see, for example, Bates et al. (2020) for an example of using a regression model fitted to a separate data set in the context of variable selection). The goal is to allow the user as much latitude as possible in choosing μ so that they can leverage every tool at their disposal, including modern machine learning algorithms and qualitative domain knowledge, to get as close to $\mu^\star$ as possible. We show in Section 1.2.5 that there is a direct relationship between the accuracy of μ and the accuracy of the resulting floodgate LCB.

In fact, an interesting and surprising feature of floodgate (both f and Algorithm 1) is that it is invariant to certain transformations of μ, making floodgate work well even sometimes when μ is quite far from $\mu^\star$. In particular, everything about floodgate remains identical if μ is replaced by any member of the set $S_\mu = \{c\mu(\cdot,\cdot) + g(\cdot,\cdot) : c > 0,\ g(x,\cdot) = g(x',\cdot)\,\forall x, x'\}$. An immediate consequence is that if μ is a partially linear function in x, i.e., $\mu(x,z) = cx + g(z)$ for some c and g, then floodgate only depends on μ through the sign of c, making floodgate particularly forgiving for partially linear working models. To be precise, floodgate using $\mu(x,z) = cx + g(z)$ will perform identically *to floodgate using the* best *partially linear approximation to $\mu^\star$ as long as c has the same sign as the coefficient in that best approximation (regardless of c's magnitude or anything about g).*

1.2.3 Upper confidence bounds for the mMSE gap

Before continuing our study of floodgate LCBs, we first pause to address a natural question: what about an *upper* confidence bound (UCB)? One way to get a UCB is to follow a workflow similar to the previous subsection, as follows. For any working regression function ν for $\mathbb{E}[Y \mid Z]$, consider

the functional

$$f^{\mathrm{UCB}}(\nu) = \mathbb{E}\left[(Y - \nu(Z))^2\right].$$

Then f^{UCB} plays an analogous role to f in the opposite direction, in that for *any* ν, (a) $f^{\mathrm{UCB}}(\nu) \geq \mathcal{I}^2$ and (b) we can construct a level α UCB $U_n^\alpha(\nu)$ for $f^{\mathrm{UCB}}(\nu)$. Property (a) is immediate from the minimality of the first term and non-negativity of the second term in definition (1.2.1), while property (b) can be established without even making model-X assumptions: simply take the CLT-based UCB from the estimator $\frac{1}{n}\sum_{i=1}^n (Y_i - \nu(Z_i))^2$, which is unbiased for $f^{\mathrm{UCB}}(\nu)$.

Unfortunately, there is no value of ν such that $f^{\mathrm{UCB}}(\nu) = \mathcal{I}^2$ except in the noiseless setting where Y is a *deterministic* function of (X, Z). In particular, no matter how well ν is chosen and how large n is, $U_n^\alpha(\nu) - \mathcal{I}^2 \geq \mathbb{E}\left[\mathrm{Var}\left(Y \mid X, Z\right)\right]$ with probability at least $1 - \alpha$. This shortcoming is perhaps foreseeable given that $U_n^\alpha(\nu)$ never even uses the X_i, but it turns out to be unimprovable (even using model-X information), as we now prove in Theorem 1.2.4.

Theorem 1.2.4. *Fix a continuous joint distribution $P_{X,Z}$ for (X, Z), and let $\mathcal{F}$ denote the class of joint distributions F for (Y, X, Z) such that F is compatible with $P_{X,Z}$ and $\mathrm{Var}\left(Y\right) < \infty$. Let $U(D_n)$ denote a scalar-valued function of the n i.i.d. samples $D_n = \{Y_i, X_i, Z_i\}_{i=1}^n$; if $U(D_n)$ outputs a UCB for the mMSE gap that is pointwise asymptotically valid for any $F \in \mathcal{F}$, i.e.,*

$$\inf_{F \in \mathcal{F}} \liminf_{n \to \infty} \mathbb{P}_F(U(D_n) \geq \mathcal{I}_F^2) \geq 1 - \alpha,$$

then

$$\sup_{F \in \mathcal{F}} \limsup_{n \to \infty} \mathbb{P}_F\left(U(D_n) - \mathcal{I}_F^2 < \mathbb{E}_F\left[\mathrm{Var}_F\left(Y \mid X, Z\right)\right]\right) \leq \alpha, \tag{1.2.10}$$

where the subscript F denotes quantities computed with F as the data-generating distribution.

The proof of Theorem 1.2.4 can be found in Appendix A.1.2. Note that since we fix $P_{X,Z}$ at the beginning of the theorem statement, U is allowed to use model-X information. As just mentioned

above, this theorem provides no cause for concern in the noiseless setting when $\mathbb{E}\left[\mathrm{Var}\left(Y \mid X, Z\right)\right] = 0$. However, in many applications we may expect $\mathbb{E}\left[\mathrm{Var}\left(Y \mid X, Z\right)\right]$ to be substantial, and the above theorem guarantees *any* pointwise asymptotically valid UCB must be conservative by this amount. The only way to overcome this problem would be to assume some sort of structure on $Y \mid X, Z$, such as smoothness or sparsity, in contrast to model-X floodgate which requires no information about $Y \mid X, Z$ and can certainly produce nontrivial LCBs and even achieve the parametric rate with sufficiently-accurate μ; see Section 1.2.5. Although it is disappointing that a better UCB is not achievable, we envision MOVI inference often being used to quantify *new* important relationships, in which case we expect it to be more useful to know a variable is *at least as* important as some LCB than to upper-bound its importance with a UCB. Given this perspective and the negative UCB result of Theorem 1.2.4, we return for the remainder of the paper to the study of using floodgate to obtain LCBs.

1.2.4 Computation

Astute readers may have noticed that the quantities R_i and V_i in Algorithm 1 involve conditional expectations/variances which, though in principle known due to the assumed model-X knowledge of $P_{X|Z}$, may be quite hard to compute in practice. In certain cases these conditional expectations can have simple or even closed-form expressions, such as when μ is a generalized linear model and $X \mid Z$ is Gaussian, but otherwise a more general approach is needed. Monte Carlo provides a natural solution: assume that we can sample K copies $\widetilde{X}_i^{(k)}$ of X_i from $P_{X_i|Z_i}$ conditionally independently of X_i and Y_i and thus replace R_i and V_i, respectively, by the sample estimators

$$R_i^K = Y_i \left(\mu(X_i, Z_i) - \frac{1}{K} \sum_{k=1}^{K} \mu(\widetilde{X}_i^{(k)}, Z_i) \right),$$

$$V_i^K = \frac{1}{K-1}\sum_{k=1}^{K}\left(\mu(\widetilde{X}_i^{(k)}, Z_i) - \frac{1}{K}\sum_{k=1}^{K}\mu(\widetilde{X}_i^{(k)}, Z_i)\right)^2.$$

Luckily the same guarantees hold for the Monte Carlo analogue of floodgate, even for fixed K.

Theorem 1.2.5. *Under the conditions of Theorem 1.2.3, for any given $K > 1$, $L^{\alpha}_{n,K}(\mu)$ computed by replacing R_i and V_i with R_i^K and V_i^K, respectively, in Algorithm 1 satisfies*

$$\liminf_{n\to\infty} \mathbb{P}\left(L^{\alpha}_{n,K}(\mu) \leq \mathcal{I}\right) \geq 1 - \alpha.$$

The proof can be found in Appendix A.1.3. In general we expect larger values of K to produce more accurate LCBs, but we found the difference between $K = 2$ and $K = \infty$ to be surprisingly small in our simulations and, of course, it will always be computationally faster to use smaller K. Although Theorem 1.2.5 is a pointwise result holding for any fixed $K > 1$, it can be generalized to a uniform result over all $K > 1$ with miscoverage bounded by a $n^{-1/2}$ rate using higher moment conditions and a variance lower bound assumption; see Appendix A.3 for details.

1.2.5 Accuracy adaptivity to μ's mean squared error

Having established floodgate's validity and computational tractability, the natural next question is: how accurate is it, i.e., how close is the LCB to the mMSE gap? The answer depends on the accuracy of μ—the better that μ approximates $\mu^\star$, the more accurate the floodgate LCB is, as formalized in the following theorem.

Theorem 1.2.6 (Floodgate accuracy and adaptivity)**.** *For i.i.d. data $\{(Y_i, X_i, Z_i)\}_{i=1}^n$ such that $\mathbb{E}[Y^{12}] < \infty$, $\mathrm{Var}(Y \mid X, Z) \geq \tau$ a.s. for some $\tau > 0$, and a sequence of working regression functions $\mu_n : \mathbb{R}^p \to \mathbb{R}$ such that for some C and all n either $\mathbb{E}\left[\mathrm{Var}\left(\mu_n(X, Z) \mid Z\right)\right] = 0$ or*

$\frac{\mathbb{E}\left[\mu_n^{12}(X,Z)\right]}{\mathbb{E}[\mathrm{Var}(\mu_n(X,Z)\,|\,Z)]^6} \leq C$, *the output of Algorithm 1 satisfies*

$$\mathcal{I} - L_n^{\alpha}(\mu_n) = O_p\left(\inf_{\mu \in S_{\mu_n}} \mathbb{E}\left[(\mu(X,Z) - \mu^\star(X,Z))^2\right] + n^{-1/2}\right), \tag{1.2.11}$$

where $S_{\mu_n} = \{c\mu_n(\cdot,\cdot) + g(\cdot,\cdot) : c > 0,\ g(x,\cdot) = g(x',\cdot)\,\forall x, x'\}$ *as defined in Remark 1.2.3.3.*

The proof can be found in Appendix A.1.4. The condition that "$\mathbb{E}\left[\mathrm{Var}\left(\mu_n(X,Z)\,|\,Z\right)\right] = 0$ or $\frac{\mathbb{E}\left[\mu_n^{12}(X,Z)\right]}{\mathbb{E}[\mathrm{Var}(\mu_n(X,Z)\,|\,Z)]^6} \leq C$" is a scale-free moment condition on μ_n which says that $\mu_n(X,Z)$ can have no dependence on Z at all or have a non-vanishing conditional variance (given Z) relative to its higher moments. The high-order moments in our assumptions are likely a technical artifact of our proof (which actually proves a somewhat stronger result than stated in the theorem), and could perhaps be relaxed with a different approach. As it stands, these assumptions allow us to utilize the Berry–Esseen-type results in Appendix A.3.1 to handle the fact that μ_n varies with n.

We call the left-hand side of Equation (1.2.11) the *half-width* (by analogy with the *width* that would measure the accuracy of a two-sided confidence interval) and Theorem 1.2.6 shows it is *adaptive* to the accuracy of μ_n through the MSE of the best element of its equivalence class S_{μ_n}, up to a limit of the parametric or central limit theorem rate of $n^{-1/2}$. So in principle floodgate can achieve $n^{-1/2}$ accuracy if a member of S_{μ_n} converges very quickly to $\mu^\star$, but in general floodgate's accuracy decays gracefully with μ_n's accuracy. Note that the infimum in Equation (1.2.11) means that floodgate is *self-correcting* with respect to μ_n's conditional mean given Z, as explained in the second paragraph of Remark 1.2.3.3.

1.2.6 Straightforward generalizations

Before moving onto extensions, we briefly address a few relatively straightforward generalizations of floodgate.

Extending the mMSE gap The mMSE gap can be very naturally made invariant to the scale of Y and bounded between 0 and 1 by dividing it by $\mathrm{Var}(Y)$. And since $\mathrm{Var}(Y)$ can be easily and asymptotically-normally estimated under weaker conditions than already assumed for floodgate's validity in Theorem 1.2.5, it is straightforward to extend the floodgate procedure and its validity to perform inference on the scale-free version $\mathcal{I}^2_{\text{sf}} = \mathcal{I}^2/\mathrm{Var}(Y)$. We also consider two ways of extending the mMSE gap such that the key property of the MOVI in Azadkia & Chatterjee (2019) is satisfied, i.e., the MOVI equals zero if *and only if* $Y \perp\!\!\!\perp X \mid Z$. Details about defining the MOVIs and providing inference can be found in Appendices A.6.1 and A.6.2.

Inference for group variable importance In applications where a group of variables share a common interpretation or are too correlated to powerfully distinguish, it is often necessary to infer a measure of *group* importance instead of a MOVI. Luckily, when X is multivariate, the mMSE gap remains perfectly well-defined and interpretable and floodgate (both f and Algorithm 1) retain all the same inferential properties. Indeed, we apply floodgate to groups of variables in our genomics application in Section 1.5.

Transporting inference to other covariate distributions In some applications, the samples we collect may not be uniformly drawn from the population we are interested in studying. For instance, our data may come from a lab experiment with covariates randomized according to one distribution, while our interest lies in inference about a population outside the lab whose covariates follow a different distribution. As long as the samples at hand share a common conditional distribution $Y \mid X, Z$ with the target population, it is relatively straightforward to perform an importance-weighted version of floodgate that provides inference for the target population's mMSE gap. We provide the details in Appendix A.7.

Adjusting for selection When inference is required for many variables simultaneously, it is often preferable to focus attention on a subset of variables whose inferences appear particularly interesting. But if we only report the set of LCBs that are, say, farthest from zero, then our coverage guarantees will fail to hold for this set due to selection bias (this is not a defect of floodgate, but a property of nearly every non-selective inferential procedure). One way to address this may be to apply false coverage-statement rate adjustments (Benjamini & Yekutieli, 2005) to floodgate LCBs. The application is straightforward, and floodgate LCBs satisfy the monotone property required by Benjamini & Yekutieli (2005), although they do not in general satisfy the independence or positive regression dependence on a subset (PRDS) condition and hence would require a correction (Benjamini & Yekutieli, 2001) for strict guarantees to hold. We leave a more formal treatment of selection adjustment to future work, but note also some simple ways to perform benign selection.

First, if selection is performed using μ and/or independent data, then no adjustment is needed for validity. For instance, if floodgate is run by data-splitting, we could arbitrarily use the first half of the data (which is also used for choosing μ, but not for running floodgate) for selection, including selecting precisely the subset of variables that μ depends on. In fact, we can even perform a certain type of benign post-hoc data processing based on the floodgate data itself: if the floodgate data are used to construct a *transformation* of the floodgate LCBs such that every transformed LCB either shrinks or remains the same, then the transformed LCBs retain their marginal asymptotic validity. This is because any such transformation, even one depending on the data or LCBs themselves, can only *increase* coverage of each LCB by reducing it or leaving it unchanged; this is related to the screening procedure in Liu et al. (2021). This means, for instance, that if a selection procedure is applied to the floodgate data and used to zero out any unselected LCBs, then as long as the zeroed-out LCBs are reported alongside the rest, the marginal validity of all reported LCBs remains intact even though the same data was used to construct the LCBs and to perform the selection that transformed them.

1.3 Extensions

1.3.1 Beyond the mMSE gap

To demonstrate that the floodgate idea can be used beyond the mMSE gap, we consider the following MOVI.

Definition 1.3.1 (Mean absolute conditional mean gap)**.** *The* mean absolute conditional mean (MACM) gap *for variable X is defined as*

$$\mathcal{I}_{\ell_1} = \mathbb{E}\left[\left|\mathbb{E}\left[Y \mid Z\right] - \mathbb{E}\left[Y \mid X, Z\right]\right|\right] \tag{1.3.1}$$

whenever all the above expectations exist.

The subscript in $\mathcal{I}_{\ell_1}$ reflects its similarity to $\mathcal{I}^2 = \mathbb{E}\left[\left(\mathbb{E}\left[Y \mid Z\right] - \mathbb{E}\left[Y \mid X, Z\right]\right)^2\right]$ except with the square replaced by the absolute value (also known as the ℓ_1 norm). Although we have not found a floodgate function to enable inference for arbitrary Y, the remainder of this subsection shows how to perform floodgate inference when Y is binary (coded as $Y \in \{-1, 1\}$). We note that when Y is binary, $\mathcal{I}_{\ell_1}$ is zero if and *only if* $Y \perp\!\!\!\perp X \mid Z$ holds (the "if" part holds for non-binary Y as well), since the expected value uniquely determines the distribution of a binary random variable.

In particular, for any (nonrandom) function $\mu : \mathbb{R}^p \to \mathbb{R}$, define

$$f_{\ell_1}(\mu) = 2\mathbb{P}\big(Y(\mu(\widetilde{X}, Z) - \mathbb{E}\left[\mu(X, Z) \mid Z\right]) < 0\big) - 2\mathbb{P}\big(Y(\mu(X, Z) - \mathbb{E}\left[\mu(X, Z) \mid Z\right]) < 0\big) \tag{1.3.2}$$

where $\widetilde{X} \sim P_{X|Z}$ and is conditionally independent of X and Y.

Lemma 1.3.2. *If $|Y| \overset{a.s.}{=} 1$, then for any μ such that $f_{\ell_1}(\mu)$ exists, $f_{\ell_1}(\mu) \leq \mathcal{I}_{\ell_1}$, with equality when $\mu = \mu^\star$.*

Obtaining an LCB for $f_{\ell_1}(\mu)$ is even easier than it was for $f(\mu)$ because $f_{\ell_1}(\mu)$ is essentially just one expectation instead of a ratio of expectations, so a straightforward central limit theorem argument suffices; Algorithm 10 (presented in Appendix A.8) formalizes the procedure and Theorem 1.3.3 establishes its asymptotic coverage.

Theorem 1.3.3 (MACM gap floodgate validity). *For any given working regression function* $\mu : \mathbb{R}^p \to \mathbb{R}$ *and i.i.d. data* $\{(Y_i, X_i, Z_i)\}_{i=1}^n$, $L_n^\alpha(\mu)$ *from Algorithm 10 satisfies*

$$\mathbb{P}\left(L_n^\alpha(\mu) \leq \mathcal{I}_{\ell_1}\right) \geq 1 - \alpha - O(n^{-1/2}).$$

Theorem 1.3.3 is proved in Appendix A.1.5, and perhaps its most striking feature is its lack of assumptions, which follows from the boundedness of $f_{\ell_1}(\mu)$ and the R_i. Like f, f_{ℓ_1} is invariant to any transformation of μ that leaves $\operatorname{sign}(\mu(X, Z) - \mathbb{E}[\mu(X, Z) \mid Z])$ unchanged on a set of probability 1, making its validity immediately uniform over large classes of μ.

Although the boundedness of the R_i streamlines the coverage guarantees, their conditional probabilities make it somewhat more complicated to carry out efficient computation of Algorithm 10. In particular, the sharp boundary at zero inside the probabilities requires a certain degree of smoothness in μ and P to be able to estimate the R_i by Monte Carlo samples analogously to Section 1.2.4. We give precise sufficient conditions and a proof of their validity in Appendix A.8, and defer study of Algorithm 10's accuracy and robustness to future work.

1.3.2 Relaxing the assumptions by conditioning

In this section we show that we can relax the model-X assumption that $P_{X|Z}$ be known exactly and apply floodgate when only a *parametric model* is known for $P_{X|Z}$. This is inspired by Huang & Janson (2020) which similarly relaxes the assumptions of model-X knockoffs. We follow the same general principle of conditioning on a sufficient statistic of the parametric model for $P_{X|Z}$, but

doing so in floodgate requires a somewhat different approach than Huang & Janson (2020). Note that this section's method and assumptions are also distinct from the double robust assumptions in Remark 1.2.3.1, further emphasizing that the key ideas underlying floodgate are not tied to any particular set of assumptions.

The approach we take in this section will involve computations on the entire matrix of observations, i.e., $(\boldsymbol{X}, \boldsymbol{Z}) \in \mathbb{R}^{n \times p}$ whose rows are the covariate samples (X_i, Z_i) and $\boldsymbol{y} \in \mathbb{R}^n$ whose entries are the response samples Y_i. Now suppose that we know a model $F_{X|Z}$ for $P_{X|Z}$ with a sufficient statistic functional for n independent (but not necessarily identically distributed) samples $\boldsymbol{X} \mid \boldsymbol{Z}$ given by $\mathcal{T}(\boldsymbol{X}, \boldsymbol{Z})$, whose random value we will denote simply by $\boldsymbol{T}$. We will assume that $\mathcal{T}$ is invariant to permutation of the rows of $(\boldsymbol{X}, \boldsymbol{Z})$ (as we would expect for any reasonable $\mathcal{T}$, since these rows are i.i.d.).

The key idea that allows us to perform floodgate inference without knowing the distribution of $\boldsymbol{X} \mid \boldsymbol{Z}$ is that, by definition of sufficiency, we *do* know the distribution of $\boldsymbol{X} \mid \boldsymbol{Z}, \boldsymbol{T}$. Leveraging this idea requires some adjustment to the floodgate procedure, and we start by defining a conditional analogue of f.

$$f_n^{\mathcal{T}}(\mu) := \frac{\mathbb{E}\left[\mathrm{Cov}(\mu^\star(X_i, Z_i), \mu(X_i, Z_i) \mid \boldsymbol{Z}, \boldsymbol{T})\right]}{\sqrt{\mathbb{E}\left[\mathrm{Var}(\mu(X_i, Z_i) \mid \boldsymbol{Z}, \boldsymbol{T})\right]}}, \tag{1.3.3}$$

again with the convention $0/0 = 0$. Note that $f_n^{\mathcal{T}}(\mu)$ does not depend on the choice of i thanks to $\mathcal{T}$'s permutation invariance, but it *does* depend on the sample size n. Nevertheless, it follows immediately from the proof of Lemma 1.2.2 that $f_n^{\mathcal{T}}(\mu) \leq f_n^{\mathcal{T}}(\mu^\star)$ for any nonrandom μ. On the other hand, $f_n^{\mathcal{T}}(\mu^\star) \neq \mathcal{I}$, but instead a different relationship that is nearly as useful holds:

$$f_n^{\mathcal{T}}(\mu^\star) \leq f(\mu^\star) = \mathcal{I},$$

due to the monotonicity of conditional variance.

With floodgate property (a) ($f_n^{\mathcal{T}}(\mu) \leq \mathcal{I}$) established, we now turn to property (b): the ability to construct a LCB for $f_n^{\mathcal{T}}(\mu)$. In an analogous way as for $f(\mu)$, we can compute n unbiased estimators of the numerator and the squared denominator, but these estimators are no longer i.i.d. because they are linked through $\boldsymbol{T}$, so we cannot immediately apply the central limit theorem or delta method as we did in Section 1.2.2. Our workaround is to split the data into n_2 *batches* of size n_1 and only condition on the sufficient statistic within each batch. This way, there is still independence between batches and we can apply the central limit theorem and delta method across batches. This strategy is spelled out in Algorithm 11 (see Appendix A.9 for details) and Theorem 1.3.4 establishes its asymptotic coverage. We call this procedure *co-sufficient* floodgate because the term "co-sufficiency" describes sampling conditioned on a sufficient statistic (Stephens, 2012).

Theorem 1.3.4 (Co-sufficient floodgate validity)**.** *For any given working regression function* $\mu : \mathbb{R}^p \to \mathbb{R}$*, i.i.d. data* $\{(X_i, Z_i, Y_i)\}_{i=1}^n$*, and permutation-invariant sufficient statistic functional* $\mathcal{T}$*, if* $\mathbb{E}[Y^4] < \infty$ *and* $\mathbb{E}[\mu^4(X, Z)] < \infty$*, then* $L_n^{\alpha,\mathcal{T}}(\mu)$ *from Algorithm 11 satisfies*

$$\liminf_{n\to\infty} \mathbb{P}\left(L_n^{\alpha,\mathcal{T}}(\mu) \leq \mathcal{I}\right) \geq 1 - \alpha.$$

The proof can be found in Appendix A.1.6. Regarding computation, as in Section 1.2.4, we can replace the conditional expectations in Algorithm 11 with Monte Carlo estimates; see Appendix A.9.1 for details. For a given μ, we may worry that co-sufficient floodgate loses some accuracy relative to regular floodgate due to the gap between $f(\mu)$ and $f_n^{\mathcal{T}}(\mu)$, but in fact this gap is typically $O(n_2^{-1})$ for fixed-dimensional parametric models. We quantify this gap for multivariate Gaussian and discrete Markov chain covariate models in the following two subsections, showing that, at least in these two cases, co-sufficient floodgate relaxes the assumptions of model-X floodgate with only a minimal loss in accuracy.

Low-dimensional multivariate Gaussian model

In this section we let $\mathcal{B}_m = \{(m-1)n_2 + 1, \ldots, mn_2\}$.

Proposition 1.3.5. *Suppose samples $\{X, Z\}_{i=1}^n$ are i.i.d. multivariate Gaussian parameterized as $X_i \mid Z_i \sim \mathcal{N}\left((1, Z_i)\gamma, \sigma^2\right)$ for some $\gamma \in \mathbb{R}^p$ and $\sigma^2 > 0$, and $Z_i \sim \mathcal{N}(\boldsymbol{v}_0, \boldsymbol{\Sigma}_0)$. Assume σ^2 is known and the batch size n_2 satisfies $n_2 > p + 2$. Let $\mathcal{T}$ be the following sufficient statistic functional*

$$\boldsymbol{T}_m := \mathcal{T}(\boldsymbol{X}_m, \boldsymbol{Z}_m) = \left(\sum_{i \in \mathcal{B}_m} X_i, \sum_{i \in \mathcal{B}_m} X_i Z_i \right).$$

Then if $\mathbb{E}\left[\mu^4(X, Z)\right], \mathbb{E}\left[(\mu^\star)^4(X, Z)\right] < \infty$, we have

$$f(\mu) - f_n^{\mathcal{T}}(\mu) = O\left(\frac{p}{n_2 - p - 2}\right). \tag{1.3.4}$$

The proof can be found in Appendix A.9.2. Note the condition $n_2 > p + 2$ is not surprising as when the sample size is smaller than p, the sufficient statistic functional is degenerate, resulting in a zero value of $f_n^{\mathcal{T}}(\mu)$. The bound in (1.3.4) allows p to grow with n in general, but when p is fixed, it gives the rate of $O(n_2^{-1})$, as mentioned earlier in Section 1.3.2.

Discrete Markov chains

To present our second example model, we define some new notation. Consider a random variable W following a discrete Markov chain with K states with $X = W_j$, $Z = W_{-j}$, then the model parameters include the initial probability vector $\pi^{(1)} \in \mathbb{R}^K$ with $\pi_k^{(1)} = \mathbb{P}(W_1 = k)$ and the transition probability matrix $\Pi^{(j)} \in \mathbb{R}^{K \times K}$ (between W_{j-1} and $X = W_j$) with $\Pi_{k,k'}^{(j)} = \mathbb{P}(W_j = k' \mid W_{j-1} = k)$. Further denoting $q(k, k_1, k_2) = \mathbb{P}(W_j = k \mid W_{j-1} = k_1, W_{j+1} = k_2)$,

we have

$$q(k, k_1, k_2) = \frac{\Pi^{(j)}_{k_1,k}\Pi^{(j+1)}_{k,k_2}}{\sum_{k=1}^{K}\Pi^{(j)}_{k_1,k}\Pi^{(j+1)}_{k,k_2}},$$

so that the conditional distribution of $\boldsymbol{X}_m \mid \boldsymbol{Z}_m$ can be compactly written down as

$$\mathbb{P}\left(\boldsymbol{X}_m \mid \boldsymbol{Z}_m\right) = \prod_{k,k_1,k_2\in[K]} \left(q(k, k_1, k_2)\right)^{N(k,k_1,k_2)}, \tag{1.3.5}$$

where $N(k, k_1, k_2) = \sum_{i\in\mathcal{B}_m} \mathbb{1}_{\{X_i=k, W_{i,j-1}=k_1, W_{i,j+1}=k_1\}}$. Thus we finally conclude that $\{N(k, k_1, k_2)\}_{(k,k_1,k_2\in[K])}$ is sufficient, and we proceed with this sufficient statistic.

Proposition 1.3.6. *Consider the above discrete Markov chain model and define the sufficient statistic functional $\mathcal{T}$ as*

$$\boldsymbol{T}_m = \mathcal{T}(\boldsymbol{X}_m, \boldsymbol{Z}_m) = \{N(k, k_1, k_2)\}_{(k,k_1,k_2\in[K])}.$$

Then if for variable $X = W_j$, $K^2 \min\{\mathbb{P}\left(W_{j-1} = k_1, W_{j+1} = k_2\right)\}_{k_1,k_2\in[K]}\} \geq q_0 > 0$ holds and $\mathbb{E}\left[(\mu^\star)^2(X, Z)\right], \mathbb{E}\left[\mu^2(X, Z)\right] < \infty$, we have

$$f(\mu) - f_n^{\mathcal{T}}(\mu) = O\left(\frac{K^3}{n_2}\right).$$

The proof can be found in Appendix A.9.2. Note that $\mathcal{T}$ here is not minimal sufficient and the above rate is cubic in K. The non-minimal sufficient statistic is adopted for the discrete Markov chain model in this paper since it is easier to work with and gives the desired rate in n_2, but we expect the rate in K could be improved by using the minimal sufficient statistic. Again, K is allowed to grow with n in general, but when it is fixed we get a rate of $O(n_2^{-1})$, as mentioned earlier in Section 1.3.2.

1.4 Simulations

Source code for conducting our simulation studies can be found at https://github.com/LuZhangH/floodgate.

1.4.1 Setup

In the following subsections, we conduct simulation studies to complement the main theoretical claims of the paper. We study the effects of the sample-splitting proportion (Section 1.4.2), covariate dimension (Section 1.4.3), and model misspecification (Section 1.4.5) on floodgate. Additional simulation studies on the effect of covariate dependence and sample size can be found in Appendix A.10.4. In Section 1.4.4, we numerically compare floodgate with the method proposed in Williamson et al. (2020). We also study the extensions to floodgate for the MACM gap (Section 1.4.6) and co-sufficient floodgate (Section 1.4.7). Each simulation study generates a set of covariates and performs floodgate inference on each in turn (i.e., treating each covariate as X and the rest as Z) before averaging its results (either coverage or half-width) over the covariates.

This paragraph describes the simulation setup for all but the simulation of Section 1.4.4. The covariates are sampled from a Gaussian autoregressive model of order 1 (AR(1)) with autocorrelation 0.3, except in Section A.10.4 where this value is varied over. The conditional distribution of $Y \mid X, Z$ is given by $\mu^\star(X, Z)$ plus standard Gaussian noise, and in each subsection we perform experiments with both a linear and a highly nonlinear model. The linear model is sparse with non-zero coefficients' locations independently uniformly drawn from among the covariates, and the non-zero coefficients' values having uniform random signs and identical magnitudes (5, unless stated otherwise) divided by $\sqrt{n}$. The nonlinear model combines zero'th-, first-, and second-order interactions between nonlinear (mostly trigonometric and polynomial) transformations of elementwise functions of a subset of covariates, and then multiplies this entire function by an amplitude (50, unless

stated otherwise) divided by $\sqrt{n}$; see Appendix A.10.1 for details. Both models use $n = 1100$, $p = 1000$, and a sparsity of 30 unless stated otherwise.

In our implementations of floodgate, we split the sample into two equal parts (justified by the results of Section 1.4.2) and use the first half to fit μ. In most of the simulations, we consider four fitting algorithms (two linear, two nonlinear): the LASSO (Tibshirani, 1996), Ridge regression, Sparse Additive Models (SAM; (Ravikumar et al., 2009)), and Random Forests (Breiman, 2001); when the response is binary there are two additional fitting algorithms: logistic regression with an L1 penalty and an L2 penalty; see Appendix A.10.2 for implementation details of these algorithms. The Monte Carlo version of floodgate from Section 1.2.4 is not needed for the linear methods, and for the nonlinear methods, $K = 500$ is used.

Given the novelty of considering inference for the mMSE gap, it is challenging to compare floodgate to alternatives except in special cases. For instance, in low-dimensional Gaussian linear models the mMSE gap is a simple function of the coefficient and thus ordinary least squares (OLS) inference can be compared to floodgate; see Appendix A.10.3 for details of how it is made comparable. Thus, in the low-dimensional linear-$\mu^\star$ simulations of Sections 1.4.3 and A.10.4, we compare floodgate's inference to that of OLS, which acts as a sort of oracle since its inference relies on very strong knowledge of $Y \mid X, Z$ which floodgate does not rely on, and OLS is not valid without that knowledge (and does not apply in high dimensions). Another example is when we can assume the group mMSE gap of all of (X, Z) is bounded away from zero, in which case the method of Williamson et al. (2020) applies, so in Section 1.4.4 we compare their method with floodgate in such a setting.

Remark 1.4.1 (Floodgate's connection to conditional independence testing). *Recall that $Y \perp\!\!\!\perp X \mid Z$ implies $\mathcal{I} = 0$, and hence rejecting $Y \perp\!\!\!\perp X \mid Z$ when $L_n^\alpha(\mu) > 0$ constitutes an asymptotically valid level-α conditional independence test (which could then be combined with a multiple testing procedure to perform variable selection). However, floodgate was explicitly designed to solve the harder problem of quantifying strength of dependence, as opposed to the conditional independence problem of*

whether any dependence exists at all. Due to the methodological constraints imposed by the more challenging nature of our problem, especially the need for data splitting, we do not expect this test derived from floodgate to be competitive with (and hence do not compare with) the many excellent conditional independence tests available in the literature (see, e.g., Candès et al. (2018a); Huang & Janson (2020); Berrett et al. (2020); Liu et al. (2021); Barber & Janson (2020); Tansey et al. (2022); Fukumizu et al. (2008); Zhang et al. (2011); Wang et al. (2015); Shah & Peters (2020); Park & Muandet (2020); Huang et al. (2020b)).

We always take the significance level $\alpha = 0.05$, and all results are averaged over 64 independent replicates unless stated otherwise (although in most cases each plotted point is averaged over multiple covariates per replicate as well, since we apply floodgate to each covariate in turn in each replicate).

1.4.2 Effect of sample splitting proportion

As mentioned in Section 1.2.2, we can split a fixed sample size n into a first part of size n_e for estimating $\mu^\star$ and use the remaining $n - n_e$ samples for floodgate inference via Algorithm 1. The choice of n_e represents a tradeoff between higher accuracy in estimating $\mu^\star$ (larger n_e) and having more samples available for inference (smaller n_e).

In Figure 1.1, we vary the sample splitting proportion and plot the average half-widths of floodgate LCBs of non-null covariates under distributions with the linear and the nonlinear $\mu^\star$ described in Section 1.4.1. Corresponding coverage plots and additional plots with different simulation parameters can be found in Appendix A.10.4. Our main takeaway from these plots is that, while the optimal choice of splitting proportion varies between distributions and algorithms, the choice of 0.5 seems to frequently achieve a half-width close to the optimum. Acknowledging that in some circumstances a more informed choice than 0.5 can be made, we nevertheless choose 0.5 as the default

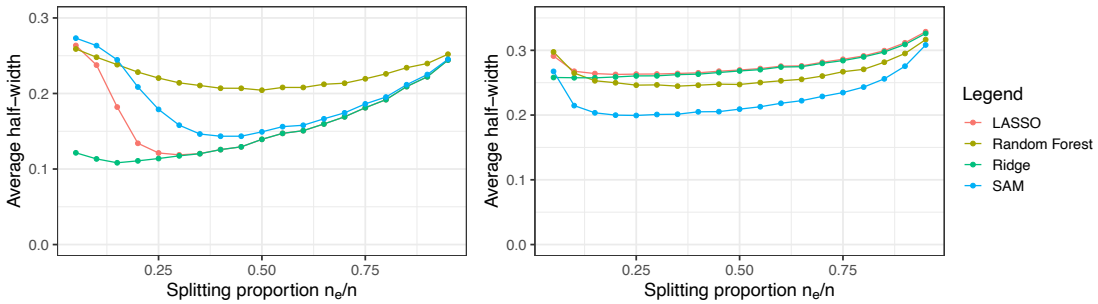


Figure 1.1: Average half-widths for the linear-$\mu^\star$ (left) and nonlinear-$\mu^\star$ (right) simulations of Section 1.4.2. The coefficient amplitude is chosen to be 10 for the left panel and the sample size n equals 3000 in the right panel; see Section 1.4.1 for remaining details. Standard errors are below 0.005 (left) and 0.006 (right).

splitting proportion throughout the rest of our simulations.

In addition to displaying the dynamics of sample splitting proportion, these plots also demonstrate two other phenomena. First, the linear algorithms (LASSO and Ridge) dominate when $\mu^\star$ is linear, and the nonlinear algorithms (SAM and Random Forest) dominate when $\mu^\star$ is nonlinear. Second, Ridge has smaller half-width than LASSO for all sample splitting proportions, which can be explained by floodgate's invariance to (partially-)linear μ: all that matters is getting the sign of the coefficient right, and setting a coefficient to zero guarantees a zero LCB. So the LASSO suffers from being a sparse estimator, although in practice we may still prefer it because of the corresponding computational savings of only having to run floodgate on a subset of covariates.

1.4.3 Effect of covariate dimension

To understand the dependence of dimension on floodgate, we perform simulations varying the dimension. In particular, in the first panel of Figure 1.2, we vary the covariate dimension and plot the average half-widths of floodgate LCBs of non-null covariates when $\mu^\star$ is linear. This setting enables comparison with OLS because it is linear and low-dimensional, so we also include a curve for OLS.

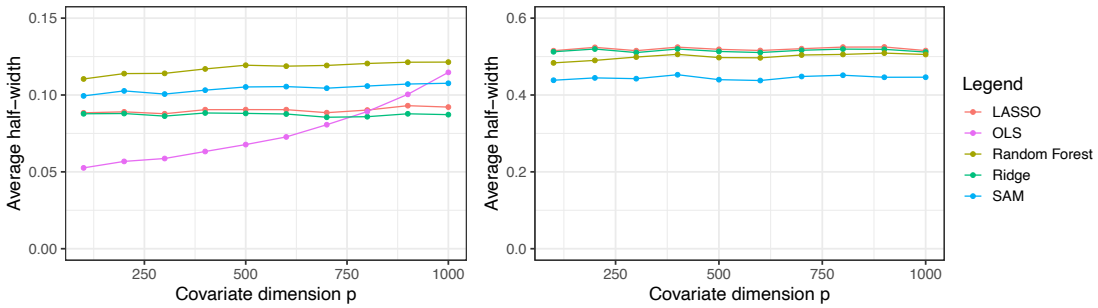


Figure 1.2: Average half-widths for the linear-$\mu^\star$ (left) and nonlinear-$\mu^\star$ (right) simulations of Section 1.4.3. OLS is run on the full sample. p is varied on the x-axis; see Section 1.4.1 for remaining details. Standard errors are below 0.002 (left) and 0.008 (right).

The main takeaway is that floodgate's accuracy is relatively unaffected by dimension, and although for very low dimensions (where OLS is known to be essentially optimal) it is less accurate than OLS, for a good choice of n_e floodgate's half-widths are at most about 50% larger than OLS's and actually narrower than OLS's when $p \approx n/2$. A similar message is found with nonlinear $\mu^\star$ in the second panel of Figure 1.2, except OLS no longer applies and in this case the nonlinear algorithms outperform the linear ones in floodgate. Coverage plots corresponding to Figure 1.2 and additional plots with different simulation parameters can be found in Appendix A.10.4.

1.4.4 Comparison with Williamson et al. (2020)

Although Williamson et al. (2020)'s method (which we refer to as W20b) is only valid when the group mMSE gap of all the covariates is bounded away from zero, we can compare it with floodgate in that setting. We use W20b according to that paper's instructions for ensuring validity for any value of $\mathcal{I}$ (as long as the group mMSE gap for all the variables put together is bounded away from zero), which seems most comparable to floodgate. That is, we implement the sample-split and cross-fitted version using the default function `vimp_rsquared` in the W20b authors' R package `vimp` (version 2.1.0). Since W20b gives confidence intervals for $\mathcal{I}^2/\mathrm{Var}\,(Y)$, we transform its inference into

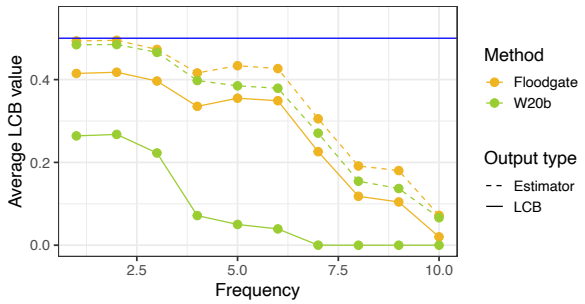


Figure 1.3: Average LCB values (solid lines) for floodgate and W20b in the sine function simulation of Section 1.4.4. The frequency λ is varied on the x-axis, and the solid blue line in the plot shows the true value of $\mathcal{I}$. The dashed lines correspond to the average estimator values of $\mathcal{I}$. The results are averaged over 640 independent replicates, and the standard errors are below 0.01.

a $1 - \alpha$ coverage LCB for $\mathcal{I}$ by taking the lower bound from its $1 - 2\alpha$ confidence interval, multiplying it by $\mathrm{Var}(Y)$, and then taking the square root. Our simulation example uses a sine function of varying frequency for $\mu^\star$. In particular, $p = 2$, the covariates $(X, Z) \in \mathbb{R}^2$ are i.i.d. uniformly distributed on $(-1, 1)$, and Y equals $A(\lambda)\sin(\lambda X)$ plus standard Gaussian noise, where $\lambda > 0$ controls the frequency and $A(\lambda)$ is chosen so that $\mathcal{I} = 0.5$ regardless of λ (thus ensuring the group mMSE gap of (X, Z) is always bounded away from zero, as required by W20b). Both floodgate and W20b must internally fit an estimate of $\mu^\star$, and for both methods we use locally-constant loess smoothing with tuning parameters selected by 5-fold cross-validation, following a different two-dimensional simulation example from Williamson et al. (2019).

The solid curves in Figure 1.3 show the average LCBs of the two methods applied to the non-null variable X as λ varies. Larger λ corresponds to less-smooth $\mathbb{E}[Y \mid X, Z]$ and hence a more challenging estimation problem (for both methods), and both methods become generally more conservative and less accurate as λ grows (both methods achieve at or above nominal coverage throughout this simulation; see Appendix A.10.4 for the coverage plot). Yet floodgate's LCB provides consistently and considerably more accurate inference over the entire range of λ. To better understand this performance difference, we additionally plot as dashed curves the average of the asymptotically normal

estimators of $\mathcal{I}$ each method uses for inference. We see from the plot that the two estimators have similar bias, but the gap between the LCB and the estimator is much smaller for floodgate, reflecting a smaller variance. This is likely due to the form of W20b's estimator, which is the difference of two asymptotically normal test statistics, one computed on each half of the split data. Heuristically, one would expect this to lead to higher variance than an estimator computed on (and hence whose variance comes only from) one half of the data, like floodgate's. This general picture is reinforced by a higher-dimensional simulation given in Appendix A.10.4.

1.4.5 Robustness

In order to study the robustness of floodgate to misspecification of $P_{X|Z}$, we consider a scenario we expect to arise in practice: a data analyst does not know $P_{X|Z}$ exactly, so instead they estimate it using the data they have, and then treat the estimate as the "known" $P_{X|Z}$ and proceed with floodgate. Note that if the analyst splits the data and uses the same subset for estimating μ and for estimating $P_{X|Z}$, then Theorem A.5.1 applies, but if they use *all* of their data to estimate $P_{X|Z}$, then our theory does not apply. Also note we are not studying the performance of co-sufficient floodgate in this subsection.

Note that if the analyst splits the data and uses the same subset for estimating μ and for estimating $P_{X|Z}$, then Theorem A.5.1 applies, but if they use *all* of their data to estimate $P_{X|Z}$, then our theory does not apply. Also note we are not studying the performance of co-sufficient floodgate in this subsection.

Figure 1.4 varies how much in-sample data is used in $P_{X|Z}$-estimation and shows the coverage of floodgate for null and non-null variables in a linear setting. The estimation procedure is to fit the graphical LASSO (GLASSO) with 3-fold cross-validation to a subset of the in-sample data and treat $P_{X|Z}$ as conditionally Gaussian with covariance matrix given by the GLASSO estimate. Since $n = 1100$ in all these simulations and the sample splitting proportion is 0.5, when the x-axis

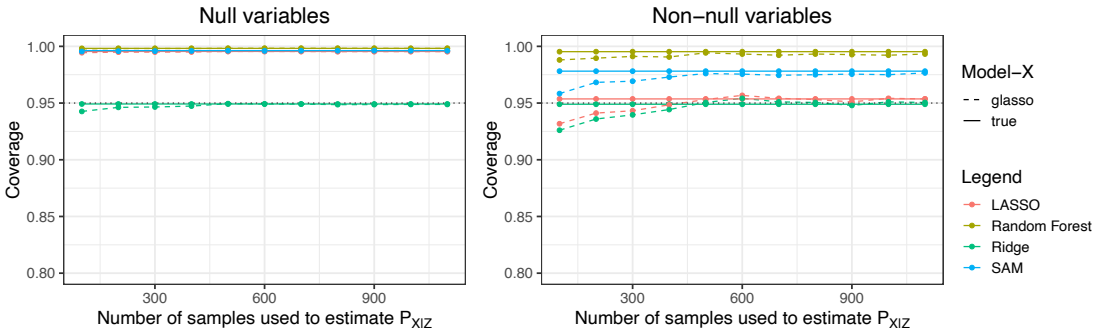


Figure 1.4: Coverage of null (left) and non-null (right) covariates when the covariate distribution is estimated in-sample for the linear-$\mu^\star$ simulations of Section 1.4.5. See Section 1.4.1 for remaining details. Standard errors are below 0.001 (left) and 0.008 (right).

value passes 550 is when the $P_{X|Z}$-estimation and inference sets start to overlap, and at the value 1100, all of the data is being used to estimate $P_{X|Z}$, including the half used for inference (violating Theorem A.5.1's assumptions). Nevertheless, we see the coverage is consistently quite high, only dropping slightly from that with true $P_{X|Z}$ for very low estimation sample sizes (i.e., very bad estimates of the covariance matrix). Note that some μ-fitting algorithms in Figure 1.4 have higher-than-nominal coverage; this is largely because the floodgate procedure will deterministically output a zero LCB (and hence have 100% coverage) when $\mu(x, z)$ does not depend on x. This happens for many covariates when μ is fitted via a sparse regression such as the LASSO and SAM (short for Sparse Additive Models), but also for our version of Random Forests which we effectively sparsify for computational reasons (see Appendix A.10.2 for details). Figures 1.5 and 1.6 show similar overcoverage for the same reason.

Average half-width plots corresponding to Figure 1.4 can be found in Appendix A.10.4. In additional to the linear setting in Figure 1.4, we also observe robust empirical coverage of floodgate when the conditional model of Y is nonlinear; see Appendix A.10.4 for details.

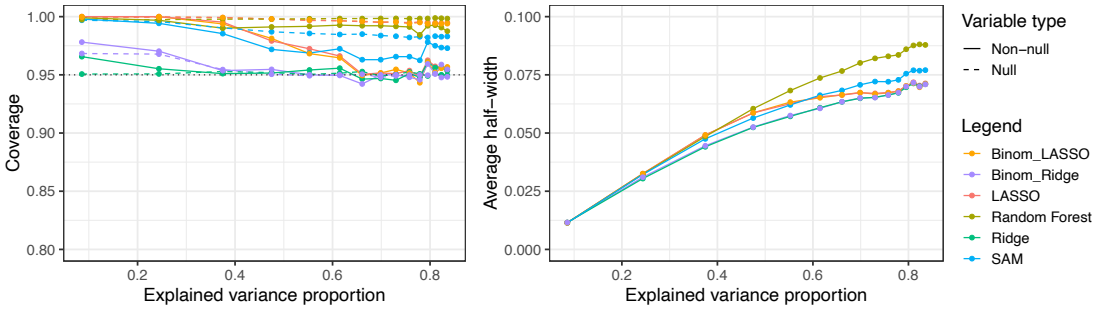


Figure 1.5: Coverage (left) and average half-widths (right) for the binary response simulations of Section 1.4.6. The explained variance proportion is varied over the x-axis. See Section 1.4.1 and 1.4.6 for remaining details. Standard errors are below 0.006 (left) and 0.001 (right).

1.4.6 Floodgate for the MACM gap

Here we study the empirical performance of floodgate applied to the MACM gap as described in Section 1.3.1. Conditional on the covariates, the binary response is generated from a logistic regression with $\frac{\log(\mathbb{P}(Y=1 \mid X,Z))}{\log(\mathbb{P}(Y=-1 \mid X,Z))}$ given by the linear $\mu^\star(X, Z)$ in Section 1.4.1. We set the sample size $n = 1000$, and the remaining simulation parameters to be the values described in Section 1.4.1. Figure 1.5 shows that floodgate has consistent coverage over a range of algorithms for fitting μ, and we see the dynamics of the average half-width as the explained variance proportion in $P_{Y|X,Z}$ increases. Note that R_i in Algorithm 10 needs to in general be estimated by Monte Carlo samples (see Appendix A.8 for details) and in Figure 1.5, we set $K = 100$ and $M = 400$ whenever the Monte Carlo version is used.

1.4.7 Co-sufficient floodgate

Finally, we study the empirical performance of co-sufficient floodgate as described in Section 1.3.2 as compared to the original floodgate method which is given full knowledge of $P_{X|Z}$. We set the covariate dimension $p = 50$, the number of Monte Carlo samples $K = 100$, and the amplitude

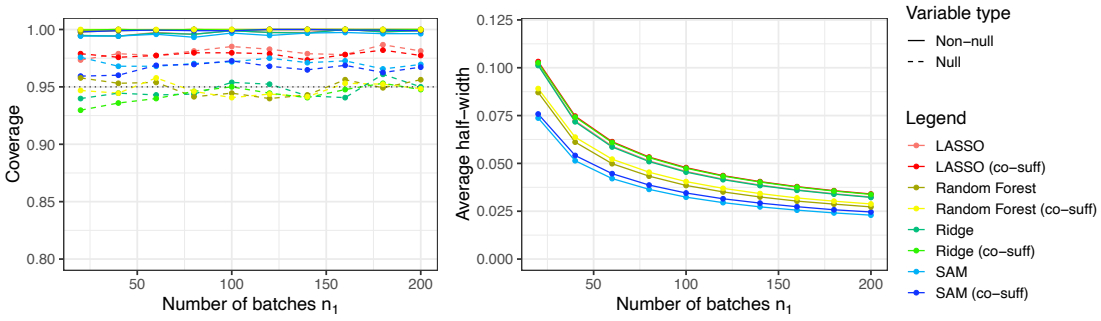


Figure 1.6: Coverage (left) and average half-widths (right) for co-sufficient floodgate and original floodgate in the nonlinear-$\mu^\star$ simulations. The number of batches n_1 is varied over the x-axis. See Section 1.4.1 and 1.4.7 for remaining details. Standard errors are below 0.009 (left) and 0.002 (right).

value for nonlinear-$\mu^\star$ to 30. The remaining simulation parameters are set to the values described in Section 1.4.1. Co-sufficient floodgate and the original floodgate procedure use the same working regression function, fitted from $n_e = 500$ samples, and use the same number of samples $n - n_e$ for inference. The batch size n_2 for co-sufficient floodgate is 300 and we vary the number of batches $n_1 = (n - n_e)/n_2$ on the x-axes. Co-sufficient floodgate is given the conditional variance of the Gaussian distribution of $X \mid Z$, but not its conditional mean, parameterized by a $(p-1)$-dimensional coefficient vector multiplying Z. Figure 1.6 shows that co-sufficient floodgate has satisfying coverage even when the number of batches is small, and has average half-width quite close to the original floodgate procedure which is given the conditional mean of $X \mid Z$ exactly. In additional to the nonlinear setting in Figure 1.6, simulations for a linear $\mu^\star$ lead to similar conclusions; see Appendix A.10.4.

1.5 Application to genomic study of platelet count

The study of genetic *heritability* is the study of how much variance in a trait can be explained by genetics. Precise definitions vary based on modeling assumptions (Zuk et al., 2012), but the fundamental concept is intuitive and central to genomics; indeed the goal of genome-wide association

studies (GWAS) is often precisely to identify single nucleotide polymorphisms (SNPs) or loci that explain the most variance in a trait. To connect heritability with the present paper, suppose Y denotes a trait, X denotes a SNP or group of SNPs, and Z denotes all the remaining SNPs not included in X. Then as can be seen in Equation (1.2.2), the mMSE gap $\mathcal{I}^2$ *exactly* measures the variance in Y that is attributable to X. Thinking of $\mathcal{I}^2$ as a sort of *conditional* heritability also makes it easy to include non-genetic factors such as age in Z, since such factors may influence Y but not be of direct interest to geneticists. Thus $\mathcal{I}^2$ can capture both gene-gene and gene-environment interactions.

Having established $\mathcal{I}^2$ as a quantity of interest, we proceed to infer it for blocks of SNPs at various resolutions of the human genome by applying floodgate to a platelet GWAS from the UK Biobank. Our analysis builds on the work of Sesia et al. (2020b), which carefully applied model-X knockoffs to the same data to perform multi-resolution *selection* of important SNPs , and in doing so require, like floodgate, a model for the SNPs X, Z and a working regression function, both of which we reuse in our own analysis. In particular, we follow the literature on genotype/haplotype modeling (Stephens et al., 2001; Zhang et al., 2002; Li & Stephens, 2003; Scheet & Stephens, 2006; Sesia et al., 2019, 2020a,b) and model the SNPs as following a hidden Markov model, and use the cross-validated Lasso as the algorithm to fit our working regression function μ. Although we use a linear μ to match the existing analysis in Sesia et al. (2020b), we remind the reader that one is in general free to use any μ with floodgate, and we hope that domain experts applying floodgate in the future to GWAS data can tailor μ to be even more powerful. The output of the analysis in Sesia et al. (2020b) is a so-called "Chicago plot", which plots stacked blocks of selected SNPs at a range of block resolutions. The height of the Chicago plot at a given location on the genome reflects the resolution at which the SNP at that location was rejected, with a greater height corresponding to a smaller block of SNPs being rejected. However, since the Chicago plot is derived from a pure selection method, it contains no information about the *strength* of the relationship between the trait and

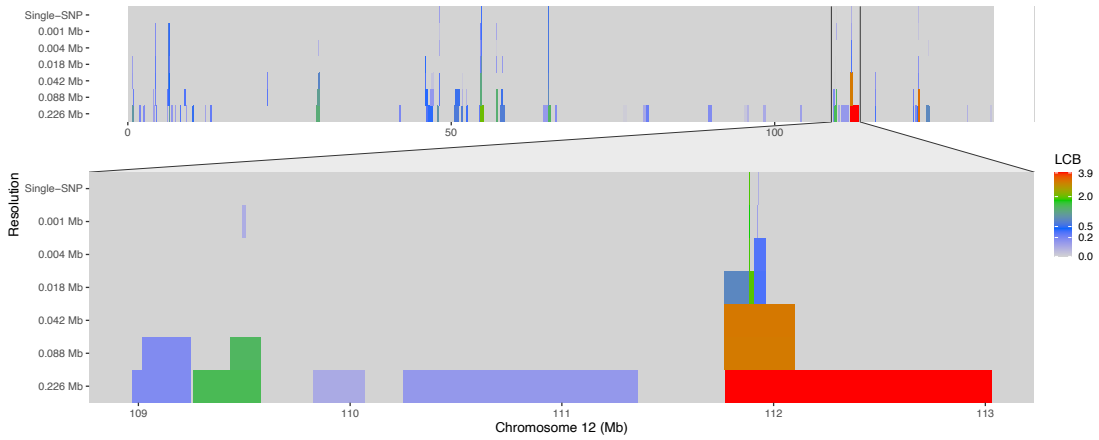


Figure 1.7: Colored Chicago plot analogous to Figure 1a of Sesia et al. (2020b). The color of each point represents the floodgate LCB for the block that contains the SNP at the location indicated on the x-axis at the resolution (measured by average block width) indicated on the y-axis (note some blocks appearing in the original Chicago plot have an LCB of zero and hence are colored grey). The second panel zooms into the region of the first panel containing the largest floodgate LCB.

any of the blocks of SNPs. Floodgate enables us to construct a *colored* Chicago plot by computing an LCB for each selected block of SNPs and reporting an LCB of zero (without computation) for all unselected blocks of SNPs; see Appendix A.11 for implementation details.

In particular, Figure 1.7 is a colored version of Figure 1a of Sesia et al. (2020b), which displayed the genomic regions on chromosome 12 that those authors found to be related to platelet count in the UK Biobank data. Our colored figure shows how informative floodgate LCBs can be over and beyond a pure selection method, as it shows the signal is far from being spread evenly over the SNPs selected by Sesia et al. (2020b). This information is crucial for the *prioritization* of selected regions, as without color the Chicago plot does not give any indication which of the selected SNPs the data indicates are most important (we note that the height of the tallest selected block at a SNP need *not* correspond to its importance, and indeed there are many pairs of locations in the figure such that one has a taller block in the original Chicago plot but the other has a brighter color in Figure 1.7).

1.6 Discussion

Floodgate is a powerful and flexible framework for rigorously inferring the strength of the conditional relationship between Y and X. We prove results about floodgate's validity, accuracy, and robustness and address a number of extensions/generalizations, but a number of questions remain for future work and we highlight two here:

- Floodgate relies on a working regression function that is not estimated from the same data used for inference, which usually will require data splitting. It would be desirable, both from an accuracy standpoint and a derandomization standpoint, to remove the need for data splitting or at least find a way for samples in one or both splits to be recycled between regression estimation and inference.

- The floodgate framework is applied here to the mMSE gap and the MACM gap, but more generally it constitutes a new tool for flexible inference of nonparametric functionals, and we expect it can find use for inferring other MOVIs. The main challenge for its application is the identification of an appropriate floodgate functional, and it would be of interest to better understand principles or even heuristics for finding such functionals for a given MOVI. Indeed we make no claim that the functionals proposed in this paper are unique for their respective MOVIs, and there may be others that lead to better floodgate procedures.

Acknowledgments

L.Z. is partially supported by the NSF-Simons Center for Mathematical and Statistical Analysis of Biology at Harvard, award number #1764269 and the Harvard Quantitative Biology Initiative. L.J. is partially supported by the William F. Milton Fund. We would like to thank the Neale Lab at the Broad Institute of MIT and Harvard for including us in their application to the UK Biobank Re-

source (application 31063), Sam Bryant for the access to on-premise data files, Matteo Sesia, Eugene Katsevich, Asher Spector, Benjamin Spector, Masahiro Kanai and Nikolas Baya for the help with the genomics application, and Dongming Huang for helpful discussions.

2

Floodgate: A Swiss Army Knife for Inference in Regression

CONTRIBUTION

This chapter is based on a manuscript, jointly with Prof. Lucas Janson.

Abstract

To better understand the underlying structures behind massive datasets, scientists and practitioners ask many statistical questions, such as measuring the importance of a given variable, quantifying the interaction effects between a pair of variables, and characterizing how the conditional relationship is away from certain shape constraints, and many others. Traditional statistical approaches often assume parametric models, thus reducing those problems to parameter estimation and inference. In this paper, we aim to avoid parametric constraints and answer those questions in a principled way. We first present the mMSE gap, an interpretable model-free inferential target, and demonstrate how it applies to a diverse range of statistical questions including nonlinearity, interactions, heterogeneity and deviation from shape constraints. To conduct inference for the mMSE gap, we leverage the floodgate idea to construct lower confidence bounds, which can build on any user-chosen working regression function. To illustrate, we provide computational details of our floodgate approaches in multiple examples. Overall, we show how floodgate is a general and useful tool for regression inference.

Keywords. Model-free, mMSE gap, interaction, alternating projection, heterogeneity, non-linearity, shape constraints, isotonic regression, convex regression, privacy, feature engineering, representation learning, floodgate, model-X.

2.1 Introduction

Given a response variable Y and the covariate $X \in \mathbb{R}^p$, scientists and practitioners seek to understand the conditional relationship between Y and X through answering many statistical questions. Variable importance inference is one of the well-studied questions. Beyond variable importance, there are many other appealing inferential questions. In feature engineering and representation

learning, some representation of the covariates is considered for improving the subsequent learning tasks in terms of computation time and power performance. A natural question researchers raise is how to measure the intrinsic predictive accuracy loss from such a representation/transformation. In privacy analysis, people develop procedures to produce synthetic data with principled privacy guarantees but sacrifice a certain amount of information in data. The intrinsic predictive accuracy loss from the privacy mechanism is a critical factor for practitioners to consider in the decision-making of privacy analysis. For $X = (X_1, X_2, Z)$, measuring the interactions between the pair X_1 and X_2 in the presence of Z is a long-standing statistical problem. Regression with shape constraints is broadly used in many real-world problems. When the underlying conditional relationship does not conform to such constraints, quantifying the strength of the deviation is useful. A list of such inferential questions would go on and on. Canonical approaches start by assuming a parametric model, then answer such questions based on the parameter estimation and inferential results. However, such parametric assumptions can largely limit the application to datasets involving complex structures. When parametric assumptions fail in some examples, inferential statements produced by those methods can be invalid or inaccurate. To avoid such limitations and thus provide great generality and flexibility, we seek to deliver model-free solutions by answering the following two questions:

1. how to characterize a class of interpretable model-free targets?
2. how to construct valid confidence bounds for the targets without relying on certain modeling assumptions?

2.1.1 Our contribution

In this paper, we introduce a unified regression inference framework: we present the mMSE gap with respect to a subset $\mathcal{S}$ and show how it can characterize a class of interpretable model-free tar-

gets by varying the choice of $\mathcal{S}$; we construct confidence bounds for the mMSE gap based on the floodgate idea. In a bit more detail, we (in Section 2.2.1) define the the mMSE gap with respect to a subset $\mathcal{S}$ and introduce the general floodgate procedure for constructing asymptotic confidence bounds (Section 2.2.2) when $\mathcal{S}$ is a closed linear subspace. To illustrate how floodgate can handle diverse model-free inferential targets, we (in Section 2.3) first consider four different examples of the closed linear subspace and describe the computation details of implementing floodgate. Then we leverage the alternating projection algorithm to handle more complicated cases where the closed linear subspace is the direct sum of subspaces (Section 2.3.5). We also go beyond linear subspaces and consider the mMSE gap with respect to a convex cone (Section 2.4).

2.1.2 Related work

Canonical parametric methods assume the underlying distribution to follow specific models associated with some parameters (Javanmard & Montanari, 2014a; Van de Geer et al., 2014; Zhang & Zhang, 2014; McCullagh & Nelder, 2019). Then one can base on the estimation and inferential results to answer general inferential questions. Such approaches have been popular among practitioners and successful in many fields. However, nonlinearities, interactions and other complexities are ubiquitous in numerous real-world problems. Parametric assumptions thus limit the applications to them.

Partially or entirely removing the model assumption about the conditional distribution of Y given X, semiparametric and nonparametric inference are also studied a lot. Such approaches give rise to many mathematically reliable techniques, thus enabling us to answer statistical questions about the properties of the distribution (Fan, 1992; Delgado, 1993; Yatchew, 1998; Wasserman, 2006). But those rigorous statistical guarantees often count on some smoothness and consistency conditions.

Different from the approaches above, another line of work flexibly exploits modern machine

learning algorithms, and hinges on the fitted result to the dataset at hand to answer inferential questions about the underlying relationship (Lei et al., 2018; Fisher et al., 2019; Watson & Wright, 2019). The targets defined within such an approach are not deterministic quantities. The randomness originating from both the data and the choice of algorithms makes the inferential statements unreliable and uninterpretable.

There is also a line of work taking the so-called model-X methods to conduct hypothesis tests (Barber & Candès, 2015; Candès et al., 2018a; Barber & Janson, 2020; Liu et al., 2021) for conditional independence. They do not require any particular functional form for the conditional model for Y given X but assume complete knowledge of (or some part of) the covariate distribution. Such methods apply to many practical examples such as randomized experiments, fields in which domain scientists have good prior knowledge about the covariates, or cases involving a large amount of unlabelled data.

Recently, Zhang & Janson (2020) introduces a novel method called floodgate to infer the mMSE hap, which is an interpretable model-free measure of variable importance. Floodgate does not rely on any parametric assumption. It can leverage cutting-edge machine learning algorithms to produce rigorous inferential statements about how a given covariate variable (or a group of covariates) is important in the conditional relationship between Y and X.

Zhang & Janson (2020) only focuses on variable importance. This paper will consider multiple statistical problems beyond variable importance, including nonlinearity, interactions, deviations from monotonicity/convexity constraints, and many others. Related work for each concrete problem will be reviewed and discussed separately in those specific subsections.

2.1.3 Notation

For two random variables U and V defined on the same probability space, denote the conditional distribution of $A \mid B$ by $P_{U \mid V}$. Let z_α denote the upper αth quantile (also called as $(1-\alpha)$th

quantile) of the standard normal distribution. Denote $[n] := \{1, \ldots, n\}$. Let $L_2(\Omega, \mathcal{F}, P)$ be the vector space of real-valued random variables with finite second moments, on which we define the inner product and norm: $\langle U, V \rangle = \mathbb{E}[UV]$ and $||U|| = \sqrt{\mathbb{E}[U^2]}$ for given two random variables $U, V \in L_2(\Omega, \mathcal{F}, P)$. Given a random vector $W \in \mathbb{R}^p$, we define a subspace of $L_2(\Omega, \mathcal{F}, P)$ i.e., $L_2(W) := L_2(\Omega, \sigma(W), P)$, where $\sigma(W)$ is the sub σ-algebra generated by X. Let $\mathcal{P}_{\mathcal{S}_0} U$ be the orthogonal projection of a random variable U onto a closed subspace $\mathcal{S}_0 \in L_2(\Omega, \mathcal{F}, P)$ and it satisfies $||U - \mathcal{P}_{\mathcal{S}_0} U||^2 = \inf_{V \in \mathcal{S}_0} \mathbb{E}\left[(U - V)^2\right]$. When $S_0 = L_2(W)$, $\mathcal{P}_{\mathcal{S}_0} U = \mathbb{E}[U \mid W]$. Let $\mathcal{S}^\perp$ be the orthogonal complement of $\mathcal{S}$. Due to the orthogonal decomposition theorem, we have $U = \mathcal{P}_{\mathcal{S}} U + \mathcal{P}_{\mathcal{S}}^\perp U$ hence write $\mathcal{P}_{\mathcal{S}}^\perp = \mathbb{1} - \mathcal{P}_{\mathcal{S}}$ with $\mathbb{1}$ being the identity operator.

2.2 Main Idea

2.2.1 Minimum mean squared error gap

Many inferential questions about the conditional relationship between the response variable Y and the covariates $X = (X_1, \cdots, X_p) \in \mathbb{R}^p$ are essentially about how the true regression function $\mu^\star(x) := \mathbb{E}[Y \mid X = x]$ is far from a function class of interest. For example, the extent of non-linearity can be characterized as discrepancies between $\mu^\star$ and the class of linear functions. The importance of variable X_j for some $j \in [p]$ can be quantified as how far $\mu^\star(x)$ is from the class of functions which do not depend on the x_j coordinate. One natural way to measure the discrepancy is to consider the minimal MSE. Notice that the conditional mean $\mu^\star(X) = \mathbb{E}[Y \mid X]$ is the solution to $\arg\min_{\mu(X) \in L_2(X)} \mathbb{E}\left[(Y - \mu(X))^2\right]$ where $L_2(X) = L_2(\Omega, \sigma(W), P)$ as denoted in Section 2.1.3. We quantify the discrepancy as how much the minimal MSE under no constraint $\min_{\mu(X) \in L_2(X)} \mathbb{E}\left[(Y - \mu(X))^2\right]$ gets increased with μ being subject to a certain constraint. The constraint can be represented as $\mu(X) \in \mathcal{S}$ where $\mathcal{S} \in L_2(X)$. To formalize the above idea, we introduce the following definition.

Definition 2.2.1 (Minimum mean squared error gap). *The minimum mean squared error gap (mMSE gap) with respect to $\mathcal{S} \in L_2(X)$ is defined as*

$$\mathcal{I}^2_{\mathcal{S}} := \inf_{\mu(X)\in\mathcal{S}} \mathbb{E}\left[(Y-\mu(X))^2\right] - \inf_{\mu(X)\in L_2(X)} \mathbb{E}\left[(Y-\mu(X))^2\right] \tag{2.2.1}$$

whenever all the above expectations exist.

We shall mention that $\mathcal{I}^2_{\mathcal{S}}$ with a particular choice of $\mathcal{S}$ reduces to a measure of variable importance, which is studied in Zhang & Janson (2020) and also called the mMSE gap; see such a connection in (2.2.2). Despite the same name, Definition 2.2.1 is a general inferential target and can characterize lots of statistical questions by varying the choice of $\mathcal{S}$.

Now we pause to elaborate on the above definition. $\mathcal{I}^2_{\mathcal{S}}$ is non-negative since $S \in L_2(X)$ and equals zero if and only if $\mu^\star(X) \in \mathcal{S}$. It can be generically interpreted as the increase in the achievable MSE for predicting Y when enforcing a constraint on the working regression function μ through $\mathcal{S}$. Below we consider a few concrete examples of $\mathcal{S}$ and explain the interpretations of $\mathcal{I}^2_{\mathcal{S}}$. Choosing $\mathcal{S} = \{\mu(X) \in L_2(X) : \mu(x) = x\beta\}$, we can view $\mathcal{I}^2_{\mathcal{S}}$ as the extent of nonlinearity in the conditional model of Y given X. For $\mathcal{S} = \{\mu(X) \in L_2(X) : \mu = \sum_{j=1}^p \mu_j(x_j)\}$, we can use $\mathcal{I}^2_{\mathcal{S}}$ to measure $\mu^\star$'s level of non-additivity. Setting S to be $L_2(X_{\text{-}j})$ gives another interesting example: $\mathcal{I}^2_{\mathcal{S}}$ measures the importance of variable X_j via quantifying the increase in minimum MSE for predicting Y when we are constrained not to have access to X_j. Notice $\mathbb{E}[Y \mid X_{\text{-}j}]$ minimizes $\min_{\mu(X)\in L_2(X_{\text{-}j})} \mathbb{E}\left[(Y-\mu(X))^2\right]$ as $\mathbb{E}[Y \mid X]$ does for $\min_{\mu(X)\in L_2(X)} \mathbb{E}\left[(Y-\mu(X))^2\right]$, thus giving an equivalent expression of $\mathcal{I}^2_{\mathcal{S}}$ when $S = L_2(X_{\text{-}j})$:

$$\mathcal{I}^2_{\mathcal{S}} = \mathbb{E}\left[(Y-\mathbb{E}[Y \mid X_{\text{-}j}])^2\right] - \mathbb{E}\left[(Y-\mathbb{E}[Y \mid X])^2\right]. \tag{2.2.2}$$

The above quantity is the same as the measure of variable importance studied in Williamson et al.

(2019); Zhang & Janson (2020); Williamson et al. (2020). Note $\mathcal{I}_{\mathcal{S}}$ also admits a concise expression when $\mathcal{S}$ is a closed linear subspace: it equals the norm of the projection of $\mu^\star(X)$ onto the orthogonal complement of $\mathcal{S}$.

Lemma 2.2.2. *$\mathcal{I}_{\mathcal{S}}$ with respect to a closed linear subspace $\mathcal{S} \in L_2(X)$ satisfies $\mathcal{I}_{\mathcal{S}} = ||\mathcal{P}_{\mathcal{S}}^{\perp}\mu^\star(X)||$.*

The proof can be found in Appendix B.1.1. From now on, we will restrict $\mathcal{S}$ to be a closed linear subspace and consider inference for it. Such a requirement of $\mathcal{S}$ is satisfied in a diverse range of examples. In Section 2.4, we also describe how to extend to the convex cone case.

2.2.2 Floodgate framework

This section describes how to conduct inference on $\mathcal{I}_{\mathcal{S}}$. From the expression of $\mathcal{I}_{\mathcal{S}}$ in Lemma 2.2.2, we know the mMSE gap is a nonlinear functional of $\mu^\star$, which can be estimated using a working regression function μ. Conventional approaches in statistical inference often rely on μ being closed enough to $\mu^\star$ (e.g., with a sufficiently-fast geometric rate). Such requirements limit the choices of modern regression algorithms and the use of hard-to-quantify domain knowledge. To avoid the limitation and thus ensure flexibility, we utilize the idea of floodgate, which is first introduced in Zhang & Janson (2020) and used for inference on the measure of variable importance, as denoted by $\mathcal{I}$ (Zhang & Janson, 2020). The floodgate idea provides confidence lower bounds for $\mathcal{I}$ through the following:

1. construct a functional $f(\mu)$ satisfying $f(\mu) \leq \mathcal{I}$ for any μ and $f(\mu^\star) = \mathcal{I}$;
2. know how to obtain a lower confidence bound $L(\mu)$ of $f(\mu)$ for any μ.

Straightforwardly, $L(\mu)$ is also a valid lower confidence bound for $\mathcal{I}$ no matter the choice of the working regression functions. μ can be fitted by black-box machine learning algorithms or derived from qualitative domain information, as long as it is chosen in a way that is independent from the

data used for obtaining lower confidence bounds. One way to achieve this is data-splitting, i.e., we reserve a proportion of data to produce an estimate μ of $\mu^\star$ and the rest of data is used for constructing confidence bounds for $f(\mu)$.

Zhang & Janson (2020) proposes a choice of the floodgate functional $f(\mu)$. Assuming the knowledge of the covariate distribution P_X, Zhang & Janson (2020) constructs lower confidence bounds for $f(\mu)$ based on the central limit theorem (CLT) and delta method. The resulting floodgate inferential procedure also possesses many nice properties, e.g., its accuracy is adaptive to the MSE of the estimate μ of the true regression function; it is robust to misspecification of the model-X assumption. Zhang & Janson (2020) also presents doubly-robust floodgate, which guarantees inferential validity under some doubly-robustness assumptions.

To utilize the floodgate idea for inference on the mMSE gap $\mathcal{I}_{\mathcal{S}}$, the first step is to construct the floodgate functional. Though it is possible to mimic Zhang & Janson (2020)'s choice, we will focus on an alternative one for ease of exposition:

$$f(\mu) := ||Y - \mathcal{P}_{\mathcal{S}}\mu(X)||^2 - ||Y - \mu(X)||^2 = \langle 2Y - \mu(X), \mu(X) - \mathcal{P}_{\mathcal{S}}\mu(X)\rangle\,. \quad (2.2.3)$$

The above floodgate functional is simply an inner product and tightly satisfies the lower-bounding property as well. Lemma 2.2.3 formalizes this result and its proof can be found in Appendix B.1.1.

Lemma 2.2.3. *For any μ such that $f(\mu)$ exists, $f(\mu) \le \mathcal{I}^2_{\mathcal{S}}$, with equality when $\mu = \mu^\star$.*

Assume for now, we can compute the projection $\mathcal{P}_{\mathcal{S}}$, then straightforwardly we can derive confidence bounds for $f(\mu)$ via the floodgate procedure in Algorithm 2. The coverage validity of $L^\alpha_n(\mu)$ holds as a result of Lemma 2.2.3 and the CLT, as stated in the following theorem:

Theorem 2.2.4 (Validity). *For any given working regression function $\mu : \mathbb{R}^p \to \mathbb{R}$ and i.i.d. data $\{(Y_i, X_i)\}_{i=1}^n$, if $\mathbb{E}[Y^4]$, $\mathbb{E}[\mu^4(X)] < \infty$, then $L^\alpha_n(\mu)$ from Algorithm 2 satisfies $\mathbb{P}\left(L^\alpha_n(\mu) \le f(\mu)\right) \ge$*

Algorithm 2 Floodgate for $\mathcal{I}_{\mathcal{S}}$

Input: Data $\{(Y_i, X_i)\}_{i=1}^n$, $\mathcal{P}_{\mathcal{S}}$, a working regression function $\mu : \mathbb{R}^p \to \mathbb{R}$, and a confidence level $\alpha \in (0, 1)$.
Compute $R_i = \langle 2Y_i - \mu(X_i), \mu(X_i) - \mathcal{P}_{\mathcal{S}}\mu(X_i)\rangle$ for each $i \in [n]$, and its sample mean $\bar{R}$ and sample standard deviation s.

Output: Lower confidence bound $L_n^{\alpha}(\mu) = \max\left\{\bar{R} - \frac{z_\alpha s}{\sqrt{n}}, 0\right\}$.

$1 - \alpha$, *which combined with Lemma 2.2.3 immediately establishes*

$$\liminf_{n\to\infty} \mathbb{P}\left(L_n^{\alpha}(\mu) \leq \mathcal{I}_{\mathcal{S}}^2\right) \geq 1 - \alpha.$$

The proof can be found in Appendix B.1.1. The above 4-th moment conditions are required since $f(\mu)$ is an inner product of the random variables $2Y - \mu(X)$ and $\mu(X) - \mathcal{P}_{\mathcal{S}}\mu(X)$.

From the expression of $f(\mu)$, we know the central part of implementing Algorithm 2 is the evaluation of $\mathcal{P}_{\mathcal{S}}$ for a given choice of $\mathcal{S}$. Ideally, we would like to derive a closed-form expression of $\mathcal{P}_{\mathcal{S}}\mu$ for any μ. Then we can directly compute R_i in Algorithm 2. This option is only possible in some special cases. A more general strategy is to draw null samples to construct i.i.d. unbiased (or asymptotically unbiased) estimates of $f(\mu)$. In this situation, the CLT argument still holds and thus guarantees asymptotic validity as in Theorem 2.2.4. When the above solutions are not tractable, we resort to least squares estimators for approximation. We shall mention that the strategies for all the examples in this paper fall into the above three categories.

The deterministic lower bounding property in Lemma 2.2.3 frees us from assumptions about μ. Therefore, the requirement for implementing floodgate is only in the part of constructing $L(\mu)$ for $f(\mu)$. In this paper, we take a model-X approach to obtain lower confidence bounds, that is, we assume knowledge about the covariate distribution P_X. Such a model-X assumption is sometimes reasonable and has been assumed in many previous work (Barber & Candès, 2015; Candès et al., 2018a; Barber & Janson, 2020; Zhang & Janson, 2020; Liu et al., 2021). We shall emphasize

that the floodgate idea is not tied to a particular set of assumptions. For example, Zhang & Janson (2020) presents doubly-robust floodgate for inference on variable importance. Though in principle we can discuss different sets of assumptions for running floodgate, this paper will still focus on the model-X approach for ease of presentation. How much knowledge about the covariate distribution is required depends on the choice of $\mathcal{S}$, and in some cases, running floodgate requires no such assumptions.

2.3 Different examples of subspaces

As we have presented the general framework for inferring a class of nonparametric targets, let us investigate a few concrete examples. This section will consider some choices of $\mathcal{S}$, which are closed linear subspaces. For those examples, we will interpret Definition 2.2.1 and elaborate on the computational details of Algorithm 2.

2.3.1 Predictive power loss from feature transformations

The mMSE gap applies to context of variable importance, with the closed linear subspace $\mathcal{S}$ chosen as the $\mathcal{S} = \{\mu(X) \in L_2(X) : \mu(X) = \lambda(X_{\text{-}j}),\ \text{for any } \lambda : \mathbb{R}^{p-1} \to \mathbb{R}\}$ for some variable X_j, thus define a measure of X_j's importance. We can also interpret the mMSE gap differently: it quantifies the intrinsic predictive accuracy loss after transforming the original covariate into a particular low-dimensional space. Such a transformation is only a special case in feature engineering and representation learning where people seek to transform the predictors (represent the data) to help supervised learning algorithms better uncover the underlying relationships, thus improving performance (Kuhn & Johnson, 2019). Researchers have been developing techniques that automatically learn the representations from data and eventually contribute to the remarkable success of machine learning (Bengio et al., 2013; Chen et al., 2020). With such a boom in the representation learning commu-

nity, it is important to evaluate the quality of those transformations/representations (Whitney et al., 2020). Existing methods focus on certain "probe architectures" (Alain & Bengio, 2016; Blier & Ollivier, 2018; Bachman et al., 2019; Henaff, 2020; Whitney et al., 2020): specifying a fixed learning algorithm such as neural networks, they train a model on the given representation of data then evaluate the validation accuracy of the fitted model. Such strategies largely depend on the choice (and architecture) of the learning algorithm. Hence, it can not capture the intrinsic properties of the transformation and might lead to misleading conclusions about the quality of the representation. In this section, we take a different approach and propose a measure using the mMSE gap framework. Specifically, for a given transformation $t : x \in \mathbb{R}^p \to t(x) \in \mathbb{R}^r$, we define the linear subspace $\mathcal{S}$ as

$$\mathcal{S} = \{\mu(X) \in L_2(X) : \mu(X) = \lambda(t(X)), \text{ for any } \lambda : \mathbb{R}^r \to \mathbb{R}\}. \tag{2.3.1}$$

and use $\mathcal{I}_{\mathcal{S}}$ (in Definition 2.2.1) to measure the quality of the transformation t. As we can see, $\mathcal{I}_{\mathcal{S}}$ measures the increase in the minimum MSE for predicting Y when the covariates X are transformed according to t. It is a population quantity and only depends on the joint distribution over (Y, X) and the transformation function t, thus provides an intrinsic characterization of the quality of the representation. The remaining question is how to carry out Algorithm 2 for such a $\mathcal{S}$. We have $\mathcal{P}_{\mathcal{S}}\mu(X) = \mathbb{E}[\mu(X) \mid t(X)]$ by noticing that $\mathcal{S}$ in (2.3.1) is exactly $L_2(\mathcal{F}, \sigma(t(X)), P)$. This result is formalized in the following lemma and its proof can be found in Appendix B.1.2.

Lemma 2.3.1. *For* $\mathcal{S} = \{\mu(X) \in L_2(X) : \mu(X) = \lambda(t(X)),$ *for any* $\lambda : \mathbb{R}^r \to \mathbb{R}\}$, $\mathcal{P}_{\mathcal{S}}\mu(X) = \mathbb{E}[\mu(X) \mid t(X)]$.

Analogously to the application of floodgate to variable importance (Zhang & Janson, 2020), we can estimate $\mathbb{E}[\mu(X) \mid t(X)]$ via Monte Carlo sampling conditional on $t(X)$. Define a *null sample*

$\widetilde{X}$ as a random variable satisfying

$$\widetilde{X} \mid t(X) \overset{d}{\sim} X \mid t(X), \qquad \widetilde{X} \perp\!\!\!\perp (X, Y) \mid t(X). \tag{2.3.2}$$

Then we can use $\widetilde{X}$ to unbiasedly estimate $f(\mu)$, as shown in the following result.

Lemma 2.3.2. $\mathbb{E}\left[(2Y - \mu(X))(\mu(X) - \mu(\widetilde{X}))\right] = f(\mu)$.

The proof can be found in Appendix B.1.2. Note $\widetilde{X}$ drew from the conditional distribution of $X \mid t(X)$ (denoted as $P_{X|t(X)}$), conditionally independently from (Y, X), will satisfy (2.3.2). Therefore, we obtain a general version of Algorithm 2 to produce asymptotically valid lower confidence bounds via replacing R_i by its Monte Carlo estimator $R_i^K = (2Y_i - \mu(X_i))(\mu(X_i) - \frac{1}{K}\sum_{k=1}^{K} \mu(\widetilde{X}_i^{(k)}))$ where $\{\widetilde{X}_i^{(k)}\}_{k=1}^K$ are K (conditionally) independent copies from $P_{X|t(X)}$.

The problem of sampling from $P_{X|t(X)}$ requires a case-by-case analysis. For certain X's distribution and transformation function t (see Appendix B.2.1 for an example where X is multivariate Gaussian and t is linear), we can take advantage of their properties to generate exact samples from $P_{X|t(X)}$. In the following, we consider a more complicated example where X is multivariate Gaussian and the transformation function t is defined as a neural network with one hidden layer. Note that such an example is already quite interesting: universal approximation theorems say that any continuous function on a compact set can be approximated arbitrarily well by a neural network with one hidden layer and a finite number of weights (Cybenko, 1989; Hornik et al., 1989; Hornik, 1991; Leshno et al., 1993; Pinkus, 1999). But the nonlinearities in the transformation function t make it difficult to sample from $P_{X|t(X)}$. Despite the challenge, we utilize Markov chain Monte Carlo (MCMC) samplers as well as a modular strategy to generate approximate samples, as described below. And we defer to future work studying more covariate distributions and transformation functions of interest to the feature engineering and representation learning communities.

Example 2.3.3 (Neural network with one hidden layer)**.** Suppose $X \in \mathbb{R}^p$ and $X \sim \mathcal{N}(\mathbf{0}, \Sigma)$. The transformation function $t : \mathbb{R}^p \to \mathbb{R}^r$ is defined as: $t(x) = a(\mathrm{W}^{(2)} a(\mathrm{W}^{(1)} x))$ where $\mathrm{W}^{(1)} \in \mathbb{R}^{r_1 \times p}$, $\mathrm{W}^{(2)} \in \mathbb{R}^{r_2 \times r_1}$ are the matrices of weights and $p > r_1 > r_2$, $a : \mathbb{R} \to \mathbb{R}$ is a 1-to-1 activation function, applied to each element of the input.

To simplify the notation, we assume zero bias terms in the above example. Denote $U^{(1)} = \mathrm{W}^{(1)} X$, $X^{(1)} = a(\mathrm{W}^{(1)} X)$. To construct null sample $\widetilde{X}$ of X given $t(X)$, we will take a modular strategy. Since the activation functions are 1-to-1, we can immediately construct the null sample $\widetilde{U}^{(1)}$ of $U^{(1)}$ by setting $\widetilde{U}^{(1)} = a^{-1}(\widetilde{X}^{(1)})$ if the null sample $\widetilde{X}^{(1)}$ of $X^{(1)}$ is given. Based on such an observation, we break the problem of sampling $\widetilde{X}$ into the following two subproblems.

1. Given $t(X)$ (or equivalently $U^{(2)} := a^{-1}(t(X))$), how to construct a null sample $\widetilde{X}^{(1)}$ of $X^{(1)}$?

2. Given $\widetilde{U}^{(1)}$, how to construct a null sample $\widetilde{X}$ of X?

The second subproblem reduces to the example with multivariate Gaussian covariates and linear transformation function, which can be solved by the procedure in Appendix B.2.1. Regarding the first subproblem, we notice that $X^{(1)}$ is not multivariate Gaussian and the transformation defined by $U^{(2)} = \mathrm{W}^{(2)} X^{(1)}$ is linear. The procedure in Appendix B.2.1 does not apply to such a case. Instead, we introduce the generic MCMC sampler (Zappa et al., 2018) to handle the first subproblem, which involves general un-normalized densities on connected manifolds in Euclidean space defined by equality and inequality constraints. Clearly, in our first subproblem, the constraint $\mathrm{W}^{(2)} X^{(1)} = U^{(2)}$ (i.e., $\mathrm{W}^{(2)} \widetilde{X}^{(1)} = U^{(2)}$) defines a linear subspace (which is a connected manifold). Note the density function of $X^{(1)}$ can also be derived analytically. Therefore, the MCMC sampling methods in Zappa et al. (2018) apply to our first subproblem. We defer the computational details to Appendix B.2.2.

2.3.2 Predictive power loss from privacy mechanisms

Nowadays, companies, government agencies and scientific studies collect and share data which can be sensitive (Isaak & Hanna, 2018; Bowser et al., 2014; Horvitz & Mulligan, 2015; Yu, 2016). The raised privacy concerns motivate recent advances (e.g., Agrawal & Srikant (2000); Sweeney (2002); Dinur & Nissim (2003); Raghunathan et al. (2003)) in generating privatized synthetic data, among which is Differential Privacy (Dwork et al., 2006; McSherry & Talwar, 2007; Nissim et al., 2007; Dwork, 2008; Abadi et al., 2016), a framework to synthesize data with principled privacy guarantees. Differential Privacy mechanisms protect data by adding noises/perturbations to it. Although many works have been devoted to establishing strong privacy guarantees (Dwork & Rothblum, 2016; Bun & Steinke, 2016; Dong et al., 2019), it is crucial to understand the quality of the data since privatized data is only valuable if still preserving curial statistical information (Arnold & Neunhoeffer, 2020; Elliot & Domingo-Ferrer, 2018). In this section, we formulate such quality evaluation problems by leveraging the mMSE gap framework to quantify the intrinsic prediction accuracy loss under the privacy mechanisms that add independent perturbation variables to the original data. Some existing works (e.g., Drechsler (2018); Snoke et al. (2018)) study the quality evaluation problems by focusing on the similarity between the synthetic data and the training data (i.e., the data used for synthesizing privatized data). There are also methods comparing the synthetic data to the underlying population, including quantities that measure the overall distributional similarity (e.g., Arnold & Neunhoeffer (2020)) and quantities that describe the loss of performance for specific prediction or inference tasks (e.g., Chen et al. (2018); Beaulieu-Jones et al. (2019)). Unlike those previous works, the inferential target in our approach is a population quantity that only depends on the underlying data distribution and the privacy mechanisms. It is model-free but also carries appealing predictive interpretations.

For the covariate random vector X, our considered privacy mechanisms will perturb it by adding

independent random variable δ and obtain $X + \delta$ as the synthetic data. δ may follow certain distributions such as Gaussian or Laplace distributions. First we will describe our measures of the loss of prediction accuracy under such privacy mechanisms. Without the perturbations, we know $\mu^\star(X) = \mathbb{E}[Y \mid X]$ achieves the minimum MSE. After masking the original data as $X + \delta$, we immediately know that the MSE is minimized at $\mathbb{E}[Y \mid X + \delta]$. Therefore, we can define

$$\mathbb{E}\left[(Y - \mathbb{E}[Y \mid X + \delta])^2\right] - \mathbb{E}\left[(Y - \mathbb{E}[Y \mid X])^2\right] \tag{2.3.3}$$

to measure the intrinsic predictive power loss under the above perturbation privacy mechanism. To relate (2.3.3) to the mMSE gap, we augment the covariates $X \in \mathbb{R}^p$ to $(X, \delta) \in \mathbb{R}^{2p}$ and define the Hilbert space $L_2(X, \delta)$. By construction, δ are null among the augmented covariates, and thus $\inf_{\mu \in L_2(X)} \mathbb{E}\left[(Y - \mu(X))^2\right] = \inf_{\mu \in L_2(X,\delta)} \mathbb{E}\left[(Y - \mu(X, \delta))^2\right]$. Then we notice (2.3.3) is equivalent to the expression below

$$\begin{aligned} & \inf_{\mu \in L_2(X+\delta)} \mathbb{E}\left[(Y - \mu(X + \delta))^2\right] - \inf_{\mu \in L_2(X,\delta)} \mathbb{E}\left[(Y - \mu(X))^2\right] \\ = \; & \inf_{\mu \in \mathcal{S}} \mathbb{E}\left[(Y - \mu(X, \delta))^2\right] - \inf_{\mu \in L_2(X,\delta)} \mathbb{E}\left[(Y - \mu(X, \delta))^2\right] = \mathcal{I}^2_{\mathcal{S}}, \end{aligned} \tag{2.3.4}$$

where the closed linear subspace $\mathcal{S} = L_2(X + \delta)$. Since $X + \delta$ can be viewed as a transformation of the augmented covariates (X, δ), the floodgate inference problem for (2.3.4) is reduced to the setting studied in Section 2.3.1. Thus it suffices to construct null samples of (X, δ) (i.e., draw independent copies from the conditional distribution of $(X, \delta) \mid (X + \delta)$. We still assume X to be multivariate Gaussian. If the perturbation δ is Gaussian and independent from X, then (X, δ) is a multivariate Gaussian random vector. We can utilize the procedures in Appendix B.2.1 to generate null samples of (X, δ). For the case where δ is Laplace, we describe the connection between the null sampling problem here to a standard Bayesian computation problem. Specifically, we can think of

X as the linear coefficients with Gaussian priors and δ as the Laplacian noise in the Bayesian linear regression model with identity covariates I: $v = \mathrm{I}X + \delta$. Then generating the null samples is equivalent to a well-studied Bayesian computation problem in the literature Hoff (2009); Kruschke (2010); Choi & Hobert (2013); Jung & Hobert (2014); Nevo & Ritov (2016); Yang & Yuan (2017): sampling from the posterior distribution of the linear coefficients given the response variable $v = X + \delta$.

In this section and Section 2.3.1, the definition of $\mathcal{S}$ involves certain transformation to the covariates. Next, we will consider other different types of linear subspaces.

2.3.3 Nonlinearity

Statisticians have been devoted to methodological and theoretical research for statistical estimation and inference under linear assumptions. Practitioners commonly use linear regression models when investigating certain conditional relationships. Such a simplification has wide applicability and notable advantages in many real-world problems. However, there also exist a lot of cases where linear assumptions may fail to capture the underlying conditional structures. In those cases, methods imposing the linearity constraints risk sacrificing statistical accuracy. To mitigate the risk and thus help practitioners' decision-making in modeling, we study a critical problem: measuring the nonlinearity and conducting inference on the measures. On the other hand, understanding how the regression functions are far from linear is by itself a long-standing question in mathematics, statistics, and other scientific fields (Beale, 1960; Williams, 1962; Hamilton et al., 1982; Karolczak & Mickiewicz, 1995).

There are some existing works including Allgöwer (1995); Li (2012) that focus on the nonlinearity of a process/system, curvature-based nonlinearity measures (Guay et al., 1996; Bates & Watts, 1980), Guttman & Meeter (1965); Kotchoni (2018) that consider measuring nonlinearity in the regression model, and references therein. This section focuses on a regression context where we study

a nonlinearity measure of the true regression function of Y given X. Specifically, we quantify nonlinearity via the mMSE gap and conduct inference via floodgate. A key observation is that linear assumptions can be expressed as linear subspace constraints on the regression function. We define

$$\mathcal{S} = \{\mu(X) \in L_2(X) : \mu(X) = X\beta, \text{ for some } \beta \in \mathbb{R}^p\}$$

which a closed linear subspace of $L_2(X)$. The mMSE gap with such a $\mathcal{S}$ is an interpretable nonlinearity measure of $\mu^\star$ as it quantifies the increase in the achievable MSE for predicting Y when enforcing linear regression models for Y given X. In terms of how to do inference, we first notice that $\mathcal{P}_\mathcal{S}\mu(X)$ admits a closed form expression that $\mathcal{P}_\mathcal{S}\mu(X) = \arg\min_{\nu(X)\in\mathcal{S}} ||\mu(X) - \nu(X)|| = X\left(\mathbb{E}\left[XX^\top\right]\right)^{-1}\mathbb{E}\left[X\mu(X)\right]$. This is equivalent to $X\beta^\star$, with $\beta^\star$ being the projection parameter studied in Rinaldo et al. (2019b). In order to run floodgate in Algorithm 2, we need to compute $\mathbb{E}\left[X\mu(X)\right]$ and $\left(\mathbb{E}\left[XX^\top\right]\right)^{-1}$. Given (only) separate covariate data, we can straightforwardly obtain the Monte Carlo estimates of $\mathbb{E}\left[X\mu(X)\right]$ for any μ. Regarding $\left(\mathbb{E}\left[XX^\top\right]\right)^{-1}$, we can estimate it via existing precision matrix estimation methods in the literature (Yuan & Lin, 2007; Banerjee et al., 2008; Friedman et al., 2008; Lam & Fan, 2009; Cai et al., 2011; Zhou et al., 2011; Liu & Luo, 2015; Liu & Wang, 2017).

2.3.4 Deviation from equality of regression functions for multiple outcomes

When different response variables share the same covariates X, understanding the similarity between the conditional structures of responses given covariates can be a natural question. For example, ride-hailing platforms conduct switchback experiments (Kastelman & Ramesh, 2018; Tang & Huang, 2019; Bojinov et al., 2020; Huang et al., 2020a) to test a new pricing algorithm's effectiveness to get around network effects. They randomize based on time-region units. In this example, the geographical characteristics are the covariates X and the numbers of successful deliveries at dif-

ferent time windows are the response variables (e.g., Y_1, Y_2 when considering two time windows). Such time-region randomization implicitly assumes that the conditional relationships between the response variables and the covariates are the same across different time units. Testing whether such an assumption is true or inferring the extent of the violation of the equality assumption is useful for practitioners. In the following, we describe how floodgate can handle such problems. Given two response variables Y_1, Y_2 at two different time units, we seek to infer the extent of inequality between their true regression functions $\mu_1^\star, \mu_2^\star$. To formalize the problem, we introduce new notations. Let $L_2^2(\Omega, \mathcal{F}, P)$ denote the vector space of real-valued 2-dimensional random vectors with finite second moments, on which we can similarly define inner product and norm. For the random vector $X \in \mathbb{R}^p$, we define $L_2^2(\Omega, \mathcal{F}, P)$'s subspace $L_2^2(X) := L_2^2(\Omega, \sigma(X), P)$. Choosing the closed linear subspace $\mathcal{S}$ to be

$$\mathcal{S} = \{(\mu_1(X), \mu_2(X)) \in L_2^2(X) : \mu_1(X) = \mu_2(X)\}, \tag{2.3.5}$$

we measure the extent of inequality between $\mu_1^\star$ and $\mu_2^\star$ using the mMSE gap $\mathcal{I}_{\mathcal{S}}^2$

$$\mathcal{I}_{\mathcal{S}}^2 = \inf_{(\mu_1(X), \mu_2(X)) \in \mathcal{S}} \mathbb{E}\left[\left(\begin{bmatrix} Y_1 - \mu_1(X) \\ Y_2 - \mu_2(X) \end{bmatrix}\right)^2\right] - \inf_{(\mu_1(X), \mu_2(X)) \in L_2^2(X)} \mathbb{E}\left[\left(\begin{bmatrix} Y_1 - \mu_1(X) \\ Y_2 - \mu_2(X) \end{bmatrix}\right)^2\right].$$

For the above $\mathcal{S}$, we can derive the closed form expression of $\mathcal{P}_{\mathcal{S}}$:

$$\mathcal{P}_{\mathcal{S}} \begin{pmatrix} \mu_1(X) \\ \mu_2(X) \end{pmatrix} = \begin{pmatrix} \frac{\mu_1(X)+\mu_2(X)}{2} \\ \frac{\mu_1(X)+\mu_2(X)}{2} \end{pmatrix}.$$

Then we can immediately derive the floodgate functional and run the inference procedure in Algorithm 2. As alluded earlier to in Section 2.2.2, conducting floodgate inference for the problem in this section does not require any knowledge about the covariate distribution P_X.

2.3.5 Direct sum of linear subspaces

In the above examples, the linear subspace $\mathcal{S}$ admits a relatively simple expression. There are also examples where $\mathcal{S}$ is the direct sum of linear subspaces or the intersection of linear subspaces. To run floodgate for those examples, we present a new technique called alternating floodgate, and give the computation details for the example of interactions.

Suppose $\mathcal{S}_1$ and $\mathcal{S}_2$ are two different closed linear subspaces. Their direct sum $\mathcal{S} = \mathcal{S}_1 \bigoplus \mathcal{S}_2$ is still a closed linear subspace. In this section, we study the mMSE gap with respect to such a $\mathcal{S}$. Note there are also some examples where $\mathcal{S}$ is the direct sum of finite number of closed linear subspaces (e.g., Example 2.3.6). For ease of exposition, we mainly focus on the case of $\mathcal{S} = \mathcal{S}_1 \bigoplus \mathcal{S}_2$. Below we give some concrete examples of $\mathcal{S}$ and describe how the corresponding $\mathcal{I}^2_{\mathcal{S}}$ are interpretable and interesting inferential targets.

Example 2.3.4 (Interactions)**.** For the covariate vector $X = (X_1, X_2, Z)$, we are interested in the interactions between X_1 and X_2 in the presence of Z. We consider $\mathcal{S} = \mathcal{S}_1 \bigoplus \mathcal{S}_2$ with $\mathcal{S}_1 = L_2(X_1, Z), \mathcal{S}_2 = L_2(X_2, Z)$.

We rewrite $\mathcal{S} = \{\mu(X) \in L_2(X) : \mu(x_1, x_2, z) = \mu_1(x_1, z) + \mu_2(x_2, z), \text{ for some } \mu_1, \mu_2\}$, which is a class of additive functions (in (x_1, z) and (x_2, z)). Clearly, $\mathcal{I}^2_{\mathcal{S}}$ is a model-free measure of interactions since it quantifies the intrinsic predictive power loss when restricting the regression function to be additive in the two covariates X_1 and X_2.

Example 2.3.5 (Heterogeneity)**.** For the covariate vector $X = (A, Z)$, we are interested in the heterogeneity of A's effects. We consider $\mathcal{S} = \mathcal{S}_1 \bigoplus \mathcal{S}_2$ with $\mathcal{S}_1 = \{\mu(X) \in L_2(X) : \mu(a, z) = a\beta \text{ for some } \beta \in \mathbb{R}\}$ and $\mathcal{S}_2 = L_2(Z)$.

We rewrite $\mathcal{S} = \{\mu(X) \in L_2(X) : \mu(a, z) = a\beta + g(z), \text{ for some } \beta \in \mathbb{R} \text{ and } g\}$ which is a class of partial linear functions (in a). Then $\mathcal{I}^2_{\mathcal{S}}$ is a model-free measure of heterogeneity since it

quantifies the intrinsic predictive power loss when restricting the regression function to be free of A's heterogeneous effects.

Example 2.3.6 (Non-additivity). For the covariate vector $X = (X_1, X_2, \cdots, X_p)$, we are interested in the non-additivity of the true regression function $\mathbb{E}[Y \mid X]$. We consider $\mathcal{S} = \bigoplus_{j=1}^{p} \mathcal{S}_j$ with $\mathcal{S}_j = L_2(X_j)$.

We rewrite $\mathcal{S} = \{\mu(X) \in L_2(X) : \mu(x) = \sum_j^p \mu_j(x_j), \text{ for some } \mu_j\}$ which is a class of additive functions. $\mathcal{I}_{\mathcal{S}}^2$ can be viewed as a model-free measure of non-additivity since it quantifies the intrinsic predictive power loss when restricting the regression function to be additive in all the covariates.

Recall that $f(\mu) = \langle 2Y - \mu(X), \mu(X) - \mathcal{P}_{\mathcal{S}}\mu(X)\rangle = \langle 2Y - \mu(X), \mathcal{P}_{\mathcal{S}^\perp}\mu(X)\rangle$. To conduct floodgate inference, we need to evaluate the projection $\mathcal{P}_{\mathcal{S}}$ (or $\mathcal{P}_{\mathcal{S}^\perp}$). Notice that the orthogonal complement of $\mathcal{S} = \mathcal{S}_1 \bigoplus \mathcal{S}_2$ is the intersection of two closed linear subspaces, i.e., $\mathcal{S}^\perp = \mathcal{S}_1^\perp \cap \mathcal{S}_2^\perp$. Inspired by this observation, we utilize the idea of alternative projection to evaluate $\mathcal{P}_{\mathcal{S}^\perp}$). First, we introduce a key result about alternating projection, von Neumann's theorem (Von Neumann, 1949).

Theorem 2.3.7 (von Neumann). *Let $\mathcal{P}_1, \mathcal{P}_2$ be the orthogonal projections onto closed subspaces $\mathcal{M}_1, \mathcal{M}_2$ of a Hilbert space $\mathcal{H}$. Let $\mathcal{P}_{\mathcal{M}}$ be the orthogonal projection onto the intersection $\mathcal{M} = \mathcal{M}_1 \cap \mathcal{M}_2$. If $\mathcal{P}_{12} = \mathcal{P}_2\mathcal{P}_1$, then $\mathcal{P}_{12}^N \to \mathcal{P}_{\mathcal{M}}$ as $N \to \infty$. That is, for each $h \in \mathcal{H}$, $||\mathcal{P}_{12}^N(h) - \mathcal{P}_{\mathcal{M}}(h)|| \to 0$ as $N \to \infty$.*

There is also some classical work studying the convergence rate in von Neumann's theorem (Aronszajn, 1950; Kayalar & Weinert, 1988). For example, Aronszajn (1950) established that

$$||\mathcal{P}_{12}^N(h) - \mathcal{P}_{\mathcal{M}}(h)|| \leq \rho^{2N-1}||h||, \text{ for all } h \in \mathcal{H}, \tag{2.3.6}$$

where ρ is the cosine of the "angle" between $\mathcal{M}_1$ and $\mathcal{M}_2$, formally $\rho = \sup\{\langle v_1, v_2\rangle : v_j \in \mathcal{M}_j \cap (\mathcal{M})^\perp, ||v_j|| \leq 1, j = 1, 2\}$. Note $\rho \leq 1$ by definition and $\rho < 1$ in most cases except some pathological ones. Von Neumann's theorem says we can utilize alternating projections to handle $\mathcal{P}_{\mathcal{M}_1 \cap \mathcal{M}_2}$ as long as we know how to evaluate the projections onto M_1, M_2. Back to our context, we let $\mathcal{M}_1 = \mathcal{S}_1^\perp, \mathcal{M}_2 = \mathcal{S}_2^\perp$ and consider the alternating projection $\mathcal{P}_{12} = \mathcal{P}_{\mathcal{S}_2^\perp}\mathcal{P}_{\mathcal{S}_1^\perp} = (\mathbb{I} - \mathcal{P}_{\mathcal{S}_2})(\mathbb{I} - \mathcal{P}_{\mathcal{S}_1})$, where $\mathbb{I}$ is the identity operator. The strategy is spelled out in the alternating floodgate algorithm (Algorithm 3) with asymptotic coverage validity established in Theorem 2.3.8.

Algorithm 3 Alternating floodgate

Input: Data $\{(Y_i, X_i)\}_{i=1}^n, \mathcal{P}_{\mathcal{S}_1}, \mathcal{P}_{\mathcal{S}_2}$, number of alternating steps N, a working regression function $\mu : \mathbb{R}^p \to \mathbb{R}$, and a confidence level $\alpha \in (0, 1)$.
Compute $R_i = (2Y_i - \mu(X_i))\mathcal{P}_{12}^N \mu(X_i)$ for each $i \in [n]$, and its sample mean $\bar{R}$ and sample standard deviation s.

Output: Lower confidence bound $L_n^\alpha(\mu, N) = \max\left\{\bar{R} - \frac{z_\alpha s}{\sqrt{n}}, 0\right\}$.

Theorem 2.3.8 (Validity of alternating floodgate)**.** *For any given working regression function* $\mu : \mathbb{R}^p \to \mathbb{R}$ *and i.i.d. data* $\{(Y_i, X_i)\}_{i=1}^n$*, if* $\mathbb{E}[Y^4], \mathbb{E}[\mu^4(X)] < \infty$*, and for a given tolerance level* $\epsilon > 0$*, then* $L_n^\alpha(\mu, N)$ *from Algorithm 3 with* $2N - 1 \geq \frac{\log(\epsilon/3c_0)}{\log(\rho)}$ *where* $c_0 = \max\{\mathbb{E}\left[Y^2\right], \mathbb{E}\left[\mu^2(X)\right]\}$*, satisfies*

$$\liminf_{n\to\infty} \mathbb{P}\left(L_n^\alpha(\mu, N) \leq \mathcal{I}_{\mathcal{S}}^2 + \epsilon\right) \geq 1 - \alpha. \tag{2.3.7}$$

The proof can be found in Appendix B.1.3. In Algorithm 3, the key is to evaluate $\mathcal{P}_{12}^N$ where $\mathcal{P}_{12} = (\mathbb{I} - \mathcal{P}_{\mathcal{S}_2})(\mathbb{I} - \mathcal{P}_{\mathcal{S}_1})$, but it is unclear how to directly evaluate the alternating projection given some iteration number N. To this end, we figure out a nice expanded expression of $\mathcal{P}_{12}^N$.

Lemma 2.3.9. *Recursively define* $A(s, \mathcal{P}_1, \mathcal{P}_0), A(s-1, \mathcal{P}_0, \mathcal{P}_0)$ *through*

$$\begin{aligned} A(s, \mathcal{P}_1, \mathcal{P}_0) &= \mathcal{P}_1 A(s-1, \mathcal{P}_0, \mathcal{P}_0), \quad A(s-1, \mathcal{P}_0, \mathcal{P}_0) = \mathcal{P}_0 A(s-2, \mathcal{P}_1, \mathcal{P}_0), \quad s > 2, \\ A(3, \mathcal{P}_0, \mathcal{P}_0) &= \mathcal{P}_0 \mathcal{P}_1 \mathcal{P}_0, \quad A(2, \mathcal{P}_1, \mathcal{P}_0) = \mathcal{P}_1 \mathcal{P}_0, \quad A(1, \mathcal{P}_0, \mathcal{P}_0) = \mathcal{P}_0, \end{aligned}$$

then we have $\mathcal{P}_{12}^N$ *admits the following expression*

$$\begin{aligned} \mathcal{P}_{12}^N = \mathbb{1} &+ \sum_{s=1}^{N-1} \left(A(2s, \mathcal{P}_{\mathcal{S}_1}, \mathcal{P}_{\mathcal{S}_2}) + A(2s, \mathcal{P}_{\mathcal{S}_2}, \mathcal{P}_{\mathcal{S}_1}) \right) \\ &- \sum_{s=1}^{N} \left(A(2s-1, \mathcal{P}_{\mathcal{S}_1}, \mathcal{P}_{\mathcal{S}_1}) + A(2s-1, \mathcal{P}_{\mathcal{S}_2}, \mathcal{P}_{\mathcal{S}_2}) \right) + (\mathcal{P}_{\mathcal{S}_2} \mathcal{P}_{\mathcal{S}_1})^N. \end{aligned} \tag{2.3.8}$$

The proof of Lemma 2.3.9 can be found in Appendix B.1.3. Now the problem is reduced to evaluating each single term in (2.3.8), whose solution depends on the specific choices of $\mathcal{S}_1, \mathcal{S}_2$.

Computational details for interactions

This section describes how to apply alternating floodgate to infer the interactions (Example 2.3.4). For general working regression function μ and covariate distribution P_X, instead of trying to analytically compute $R_i = (2Y_i - \mu(X_i))\mathcal{P}_{12}^N \mu(X_i)$, we will generate null samples to construct i.i.d. unbiased estimates of $\mathbb{E}[R_i]$. We shall note that Theorem 2.3.8 remains true when R_i in Algorithm 3 gets replaced with such i.i.d. unbiased estimates of $\mathbb{E}[R_i]$. This is because the i.i.d. estimates are unbiased and the CLT argument can still be applied. Therefore, now it suffices to figure out the computation details of generating null samples. Due to the expansion in Lemma 2.3.9, it remains to consider $\mathbb{E}[(2Y - \mu(X))A(\cdot, \cdot, \cdot)\mu(X)]$ for each $A(\cdot, \cdot, \cdot)$ term in (2.3.8). Recall Example 2.3.4 says $\mathcal{S}_1 = L_2(X_1, Z), \mathcal{S}_2 = L_2(X_2, Z)$. It is immediate that $\mathcal{P}_{\mathcal{S}_1}\mu(X) = \mathbb{E}[\mu(X) \mid X_2, Z], \mathcal{P}_{\mathcal{S}_2}\mu(X) = \mathbb{E}[\mu(X) \mid X_1, Z]$. Notice that if we independently

sample $\widetilde{X}_1$ from $P_{X_1|X_2,Z}$ and $\widetilde{X}_2$ from $P_{X_2|X_1,Z}$, we have

$$\mathbb{E}\left[(2Y-\mu(X))\mu(\widetilde{X}_1,X_2,Z)\right] = \mathbb{E}\left[(2Y-\mu(X))A(1,\mathcal{P}_{\mathcal{S}_1},\mathcal{P}_{\mathcal{S}_1})\mu(X)\right],$$
$$\mathbb{E}\left[(2Y-\mu(X))\mu(X_1,\widetilde{X}_2,Z)\right] = \mathbb{E}\left[(2Y-\mu(X))A(1,\mathcal{P}_{\mathcal{S}_2},\mathcal{P}_{\mathcal{S}_2})\mu(X)\right].$$

To estimate $\mathbb{E}\left[(2Y-\mu(X))A(\cdot,\cdot,\cdot)\mu(X)\right]$ for general $A(\cdot,\cdot,\cdot)$, we construct null samples via the following two algorithms. A graphical demonstration of Algorithms 4 and 5 is given in Figure

Algorithm 4 Alternating Sampler

Input: (X_1, X_2, Z) (and denote $\widetilde{X}_1^{(1,0)} = X_1, \widetilde{X}_2^{(1,0)} = X_2$), number of alternating steps N.
  for t from 1 to N **do**
    Sample $\widetilde{X}_1^{(1,t)}$ conditional on $(\widetilde{X}_2^{(1,t-1)}, Z)$.
    Sample $\widetilde{X}_2^{(1,t)}$ conditional on $(\widetilde{X}_1^{(1,t)}, Z)$.
  end for
Output: $\{\widetilde{X}_1^{(1,t)}, \widetilde{X}_2^{(1,t)}\}_{t=1}^N$.

Algorithm 5 Alternating Sampler

Input: (X_1, X_2, Z) (and denote $\widetilde{X}_1^{(2,0)} = X_1, \widetilde{X}_2^{(2,0)} = X_2$), number of alternating steps N.
  for t from 1 to N **do**
    Sample $\widetilde{X}_2^{(2,t)}$ conditional on $(\widetilde{X}_1^{(2,t-1)}, Z)$.
    Sample $\widetilde{X}_1^{(2,t)}$ conditional on $(\widetilde{X}_2^{(2,t)}, Z)$.
  end for
Output: $\{\widetilde{X}_1^{(2,t)}, \widetilde{X}_2^{(2,t)}\}_{t=1}^N$.

2.1, from which we can see the analogy to the Gibbs sampling algorithm. Such a connection is not a coincidence: the Gibbs sampling algorithm has been regarded as alternating projections; see Amit (1996); Diaconis et al. (2010). Those null samples $\{\widetilde{X}_1^{(1,t)}, \widetilde{X}_2^{(1,t)}\}_{t=1}^N$, $\{\widetilde{X}_1^{(2,t)}, \widetilde{X}_2^{(2,t)}\}_{t=1}^N$ from Algorithms 4 and 5 can be used to construct unbiased estimators of $\mathbb{E}\left[(2Y-\mu(X))A(\cdot,\cdot,\cdot)\mu(X)\right]$

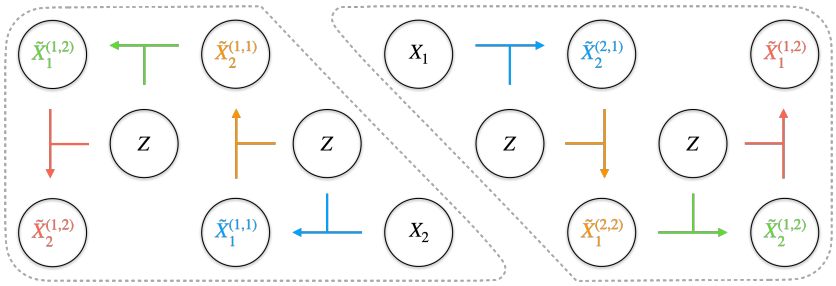


Figure 2.1: A graphical demonstration of Algorithm 4 (left) and Algorithm 5 (right).

for general $A(\cdot,\cdot,\cdot)$ defined in (2.3.8), as formalized in Lemma 2.3.10.

Lemma 2.3.10. *The null samples constructed from Algorithm 4 and 5 satisfy the following,*

$$\begin{aligned}
\mathbb{E}\left[(2Y-\mu(X))\mu(\widetilde{X}_1^{(1,t)},\widetilde{X}_2^{(1,t)},Z)\right] &= \mathbb{E}\left[(2Y-\mu(X))A(2t,\mathcal{P}_{\mathcal{S}_2},\mathcal{P}_{\mathcal{S}_1})\mu(X)\right],\\
\mathbb{E}\left[(2Y-\mu(X))\mu(\widetilde{X}_1^{(1,t)},\widetilde{X}_2^{(1,t-1)},Z)\right] &= \mathbb{E}\left[(2Y-\mu(X))A(2t-1,\mathcal{P}_{\mathcal{S}_2},\mathcal{P}_{\mathcal{S}_2})\mu(X)\right],\\
\mathbb{E}\left[(2Y-\mu(X))\mu(\widetilde{X}_1^{(2,t)},\widetilde{X}_2^{(2,t)},Z)\right] &= \mathbb{E}\left[(2Y-\mu(X))A(2t,\mathcal{P}_{\mathcal{S}_1},\mathcal{P}_{\mathcal{S}_2})\mu(X)\right],\\
\mathbb{E}\left[(2Y-\mu(X))\mu(\widetilde{X}_1^{(2,t-1)},\widetilde{X}_2^{(2,t)},Z)\right] &= \mathbb{E}\left[(2Y-\mu(X))A(2t-1,\mathcal{P}_{\mathcal{S}_1},\mathcal{P}_{\mathcal{S}_1})\mu(X)\right].
\end{aligned} \tag{2.3.9}$$

The proof of Lemma 2.3.10 can be found in Appendix B.1.3. Algorithms 4 and 5 enable us to implement the alternating projection idea, i.e., unbiasedly estimate $\langle 2Y-\mu(X),\mathcal{P}_{12}^N\mu(X)\rangle$. For simplicity, we only proceed with one chain in Algorithm 4 (also in Algorithm 5), which can be clearly seen from Figure 2.1. In practice, we should (conditionally) independently run K chains to generate null samples and thus construct the Monte Carlo estimate of $\langle 2Y-\mu(X),\mathcal{P}_{12}^N\mu(X)\rangle$. Specifically, we initialize $\widetilde{X}_1^{(2,0,k)}=X_1,\widetilde{X}_2^{(2,0,k)}=X_2$ for all $k\in[K]$. Then in each alternating step of Algorithm 4, we sample $\widetilde{X}_1^{(1,t,k)}$ conditional on $(\widetilde{X}_2^{(1,t-1,k)},Z)$, independently for each $k\in[K]$ and sample $\widetilde{X}_2^{(1,t,k)}$ conditional on $(\widetilde{X}_1^{(1,t,k)},Z)$, independently for each $k\in[K]$. Similarly running K chains for Algorithm 5 will produce $\widetilde{X}_1^{(2,t,k)},\widetilde{X}_1^{(2,t,k)}$ in each alternating step. For completeness, the whole procedures are given in Appendix B.2.4. Now we describe how to compute

R_i in Algorithm 3 such that $\mathbb{E}[R_i] = \langle 2Y - \mu(X), \mathcal{P}_{12}^N \mu(X) \rangle$ based on the null samples. For notation simplicity, we drop all the i subscripts and present the Monte Carlo estimate as

$$\begin{aligned} R = (2Y - \mu(X)) \Big(& \mu(X_1, X_2, Z) + \frac{1}{K} \sum_{k=1}^{K} \mu(\widetilde{X}_1^{(1,N,k)}, \widetilde{X}_2^{(1,N,k)}, Z) \\ & + \frac{1}{K} \sum_{k=1}^{K} \sum_{s=1}^{N-1} \big(\mu(\widetilde{X}_1^{(1,s,k)}, \widetilde{X}_2^{(1,s,k)}, Z) + \mu(\widetilde{X}_1^{(2,s,k)}, \widetilde{X}_2^{(2,s,k)}, Z) \big) \\ & - \frac{1}{K} \sum_{k=1}^{K} \sum_{s=1}^{N} \big(\mu(\widetilde{X}_1^{(1,s,k)}, \widetilde{X}_2^{(1,s-1,k)}, Z) + \mu(\widetilde{X}_1^{(2,s-1,k)}, \widetilde{X}_2^{(2,s,k)}, Z) \big) \Big). \end{aligned}$$

For general μ and covariate distribution P_X, Lemmas 2.3.10 and 2.3.9 enable us to obtain i.i.d. unbiased estimates of $\langle 2Y - \mu(X), \mathcal{P}_{12}^N \mu(X) \rangle$, which converges to the floodgate functional $f(\mu) = \langle 2Y - \mu(X), \mathcal{P}_{\mathcal{S}^\perp} \mu(X) \rangle$ as the number of alternating steps N goes to ∞. We shall mention in some special cases, the term $\mathcal{P}_{\mathcal{S}^\perp} \mu(X)$ admits an analytical expression and it suffices to only run one alternating step in Algorithm 3.

Lemma 2.3.11. *If $X_1 \perp\!\!\!\perp X_2 \mid Z$ and $\mu(X) = \mu_1(X_1, Z)\mu_2(X_2, Z)$, we have*

$$\mathcal{P}_{\mathcal{S}^\perp} \mu(X) = (\mu_1(X_1, Z) - \mathbb{E}[\mu_1(X_1, Z) \mid Z]) (\mu_2(X_2, Z) - \mathbb{E}[\mu_2(X_2, Z) \mid Z]). \quad (2.3.10)$$

When $\mu_1(X_2, Z) = X_1, \mu_2(X_2, Z) = X_2$, simply $\mathcal{P}_{\mathcal{S}^\perp} \mu(X) = (X_1 - \mathbb{E}[X_1 \mid Z])(X_2 - \mathbb{E}[X_2 \mid Z])$.

The proof of Lemma 2.3.11 can be found in Appendix B.1.3. Note (2.3.10) can be written as $\mathcal{P}_{\mathcal{S}^\perp} \mu(X) = (\mathbb{1} - A(1, \mathcal{P}_{\mathcal{S}_1}, \mathcal{P}_{\mathcal{S}_1}) - A(1, \mathcal{P}_{\mathcal{S}_2}, \mathcal{P}_{\mathcal{S}_2}) + \mathcal{P}_{\mathcal{S}_2} \mathcal{P}_{\mathcal{S}_1}) \mu(X)$, which equals $\mathcal{P}_{12}^N \mu(X)$ with $N = 1$. Comparing this equation with (2.3.8), we know that under the condition $X_1 \perp\!\!\!\perp X_2 \mid Z$ and $\mu(X) = \mu_1(X_1, Z)\mu_2(X_2, Z)$, the sampling procedures (Algorithms 4 and 5) only need to iterate one step. For illustration purposes, we describe a concrete example such that the condition

holds: X_1 and X_2 do not have a direct edge on the associated graphical model of X and the working regression function μ is fitted from linear models with conventional interaction terms.

2.4 Extension from subspaces to convex cones

In the previous sections, we apply floodgate to the mMSE gap with $\mathcal{S}$ being a closed linear subspace. In the current section, we extend our framework by allowing $\mathcal{S}$ to be a closed convex cone $\mathcal{G}$ and thus focusing on its associated inferential target $\mathcal{I}^2_{\mathcal{G}} := \inf_{\mu\in\mathcal{G}} \mathbb{E}\left[(Y-\mu(X))^2\right] - \inf_{\mu\in L_2(X)} \mathbb{E}\left[(Y-\mu(X))^2\right]$. Since every closed convex set in a Hilbert space is a Chebyshev set, $\inf_{\mu\in\mathcal{G}} \mathbb{E}\left[(Y-\mu(X))^2\right]$ admits a unique minimizer, which will be denoted by $\mathcal{P}_{\mathcal{G}}Y$. We can rewrite $\mathcal{I}^2_{\mathcal{G}} = ||\mu^\star(X) - \mathcal{P}_{\mathcal{G}}\mu^\star(X)||^2$. Similarly as in Section 2.2, the floodgate functional for $\mathcal{I}_{\mathcal{G}}$ is chosen as

$$f(\mu) := \langle 2Y - \mu(X), \mu(X) - \mathcal{P}_{\mathcal{G}}\mu(X)\rangle$$

for any (nonrandom) function $\mu : \mathbb{R}^p \to \mathbb{R}$ and by convention we define $0/0 = 0$. Using a property of the convex cone, we prove the above floodgate functional tightly satisfies the lower-bounding property as well.

Lemma 2.4.1. *For any μ such that $f(\mu)$ exists, $f(\mu) \le \mathcal{I}^2_{\mathcal{G}}$, with equality when $\mu = \mu^\star$.*

The proof can be found in Appendix B.1.4. Given Lemma 2.4.1 holds, we can similarly derive confidence bounds as in Algorithm 2 with $\mathcal{P}_{\mathcal{S}}$ replaced by $\mathcal{P}_{\mathcal{G}}$. Now we pause to give a few examples of $\mathcal{G}$ and elaborate on how the mMSE gap with respect to $\mathcal{G}$ can define interesting and interpretable inferential targets.

Example 2.4.2 (Maximum and minimum constraints)**.** For two given constant values $C_{\max}, C_{\min}$, we define $\mathcal{G}_{\max} := \{\mu(X) - C_{\max} : \mu(X) \in L_2(X), \mu \le C_{\max}\}$ and $\mathcal{G}_{\min} := \{\mu(X) - C_{\min} : \mu(X) \in L_2(X), \mu \ge C_{\min}\}$.

In this example, we are essentially considering function classes under certain maximum/minimum constraints. Note in the above definitions, we subtract the maximum/minimum bound values from $\mu(X)$ to ensure $\mathcal{G}_{\max}, \mathcal{G}_{\min}$ satisfy the requirement of convex cones. Accordingly, we replace Y by $Y - C_{\max}$ (or $Y - C_{\min}$) in the definitions of $\mathcal{I}_{\mathcal{G}_{\max}}$ (or $\mathcal{I}_{\mathcal{G}_{\min}}$). Then we see the mMSE gap naturally measures how the true regression function $\mu^\star$ deviates from the maximum and minimum constraints.

Example 2.4.3 (Non-negative least squares)**.** Let $\mathcal{G} = \{\mu(X) \in L_2(X) : \mu(X) = X\beta, \text{ for some } \beta \geq 0\}$.

Non-negative least squares (NNLS) fit linear models for Y given X but with non-negative constraints on the linear coefficients β. Non-negative constraints and NNLS methods have been applied in many fields such as acoustics (Lin et al., 2004), chemometrics (Zhang et al., 2014), economics (Lee & Pitt, 1986) and proteomics (Slawski et al., 2014). Our mMSE gap $\mathcal{I}_{\mathcal{G}}$ measures the deviation from such constraints by quantifying the increase in the achievable MSE for predicting Y when enforcing NNLS models.

Example 2.4.4 (Monotonicity constraints/isotonic regression)**.** Consider $\mathcal{G} = \{\mu(X) \in L_2(X) : \mu(x_1, \cdots, x_p) \leq \mu(z_1, \cdots, z_p) \text{ when } x_j \leq z_j, \forall j \in [p]\}$.

Monotonicity naturally defines a function class which is a convex cone. $\mathcal{I}_{\mathcal{G}}$ with the above $\mathcal{G}$ quantifies the deviation of the true regression function from the class of monotone functions via loss in predictive power. Isotonic regression enforcing the monotonicity constraints is one of the simplest form of shape-constrained regression. It has been studied a lot in the literature (Barlow & Brunk, 1972; Brunk et al., 1972; Zhang, 2002; Chatterjee et al., 2015; Han et al., 2019; Banerjee et al., 2019; Dai et al., 2020c) and widely applied to many real-world problems, including education Dykstra & Robertson (1982), genetics Luss et al. (2012), protein analysis Wang et al. (2021),

psychology Kruskal (1964) and trend analysis Neelon & Dunson (2004). When the monotonicity assumption is not plausible, it is natural to think of quantifying the non-monotonicity.

Example 2.4.5 (Convexity constraints/convex regression)**.** Consider $\mathcal{G} = \{\mu(X) \in L_2(X) : \mu \text{ is convex.}\}$

Convexity constraints and convex approximations occur in various problems, including demand analysis Varian (1982), option pricing Aıt-Sahalia & Duarte (2003), and geometric programming Boyd et al. (2007). Researchers also devote efforts to such problems (Seijo & Sen, 2011; Lim & Glynn, 2012; Hannah & Dunson, 2013; Guntuboyina & Sen, 2015; Mazumder et al., 2019). On the other hand, quantifying the extent of non-convexity is interesting and useful. Note that the class of convex function is a convex cone since non-negative weighted sums preserve convexity. Therefore, the mMSE gap with the above $\mathcal{G}$ measures non-convexity as the intrinsic prediction accuracy loss under convexity constraints.

In addition to the above choices of $\mathcal{G}$, many other convex cone examples are studied in the literature; see, e.g., Guntuboyina et al. (2018); Wei et al. (2019). In the following, we will focus on two particular shape constraints, i.e., provide the computational details of floodgate on $\mathcal{I}_{\mathcal{G}}$ with respect to the convex cones defined in Examples 2.4.4 and 2.4.5.

2.4.1 Application to monotonicity and convexity constraints

There are existing work considering hypothesis testing problems on shape constraints especially monotonicity and convexity constraints: early works Bowman et al. (1998); Gijbels et al. (2000); Hall & Heckman (2000) study the monotonicity testing problem; Chetverikov (2019) develops a nonparametric framework for testing monotonicity in regression; Meyer (2003) provides a test for simple linear regression models versus convex alternatives; Sen & Meyer (2017) studies similar problems but in a multivariate sense and proposes a likelihood ratio hypothesis test for a linear

model versus a convex or concave alternative; Wei et al. (2019) considers hypothesis testing within the Gaussian sequence model in which the null and alternative are specified by a pair of closed, convex cones such as the non-negative orthant cone, the monotone cone and the ray cone. Unlike the literature, we seek to measure the deviations from the shape constraints and conduct inference on those measures. As mentioned previously, the mMSE gap $\mathcal{I}_{\mathcal{G}}$ with respect to $\mathcal{G}$ in Examples 2.4.4 and 2.4.5 are appealing interpretable measures of deviation from monotonicity and convexity constraints. In the following, we show how to use floodgate to conduct inference on $\mathcal{I}_{\mathcal{G}}$.

The key is to evaluate the projection onto the convex cone $\mathcal{G}$. Given a working regression function μ, computing the projection $\mathcal{P}_{\mathcal{G}}$ is equivalent to solving

$$\arg\min_{g(X)\in\mathcal{G}} ||\mu - g|| = \arg\min_{g(X)\in\mathcal{G}} \mathbb{E}\left[(\mu(X) - g(X))^2\right]$$

for some $\mathcal{G}$ defined by the shape constraints. The above problem (i.e., least squares estimation in shape-restricted regression) has been studied a lot in the literature; see e.g., Guntuboyina & Sen (2015); Chatterjee et al. (2015); Guntuboyina et al. (2018); Fang & Guntuboyina (2019); Lim et al. (2020); Kur et al. (2020). Therefore, we can leverage existing least squares estimation algorithms to compute $\mathcal{P}_{\mathcal{G}}\mu(X)$. The full details of running floodgate is spell out in Algorithm 6.

Algorithm 6 Floodgate for shape constraints

Input: Data $\{(Y_i, X_i)\}_{i=1}^n$, a working regression function $\mu : \mathbb{R}^p \to \mathbb{R}$, separate covariate dataset $\{\widetilde{X}_m\}_{m=1}^N$, $\mathcal{G}$ defined by monotonicity or convexity constraints, and a confidence level $\alpha \in (0, 1)$.

Let $\widetilde{Y}_m = \mu(\widetilde{X}_m)$ and compute $g_N = \arg\min_{g\in\mathcal{G}} \frac{1}{N}\sum_{m=1}^N \left(\widetilde{Y}_m - \mu(\widetilde{X}_m)\right)^2$.

Compute $R_i = (2Y_i - \mu(X_i))\big(\mu(X_i) - g_N(X_i)\big)$ for each $i \in [n]$, and its sample mean $\bar{R}$ and sample standard deviation s.

Output: Lower confidence bound $L_n^{\alpha}(\mu, N) = \max\left\{\bar{R} - \frac{z_\alpha s}{\sqrt{n}}, 0\right\}$.

In the above algorithm, we run least squares algorithms on $\{\widetilde{X}_m\}_{m=1}^N$ to estimate the projection

of μ onto $\mathcal{G}$ by g_N. $\{\widetilde{X}_m\}_{m=1}^N$ can be some separate unlabelled dataset or covariate dataset generated by independently sampling from P_X. The coverage validity of floodgate confidence bounds depends on g_N's accuracy of estimating $\mathcal{P}_\mathcal{G}$. We formalize this in Theorem 2.4.6.

Theorem 2.4.6. *For any given working regression function $\mu : \mathbb{R}^p \to \mathbb{R}$ and i.i.d. data $\{(Y_i, X_i)\}_{i=1}^n$, if $\mathbb{E}[Y^4]$, $\mathbb{E}[\mu^4(X)] < \infty$, then $L_n^\alpha(\mu, N)$ from Algorithm 6 satisfies*

$$\liminf_{n\to\infty} \mathbb{P}\left(L_n^\alpha(\mu, N) \leq \mathcal{I}_\mathcal{G}^2 + \epsilon_N\right) \geq 1 - \alpha. \tag{2.4.1}$$

where $\epsilon_N = 3\sqrt{c_0}(\mathbb{E}\left[(g_N(X) - \mathcal{P}_\mathcal{G}\mu(X))^2\right])^{1/2}$ with $c_0 = \max\{\mathbb{E}\left[Y^2\right], \mathbb{E}\left[\mu^2(X)\right]\}$.

The proof of Theorem 2.4.6 can be found in Appendix B.1.4. In the above result, ϵ_N quantifies the accuracy of least squares estimation through the MSE term $\mathbb{E}\left[(g_N(X) - \mathcal{P}_\mathcal{G}\mu(X))^2\right]$. There is some existing literature studying limiting behaviors of the least squares estimators. For example, Corollary 4.1 and Corollary 4.2 in Lim et al. (2020) derive the rate dependence of ϵ_N on N and p for isotonic and convex regression estimators respectively under some regularity assumptions. When the sample size N of separate covariate dataset $\{\widetilde{X}_m\}_{m=1}^N$ is large enough, the error term ϵ_N in (2.4.1) will vanish in general.

2.5 Discussion

Floodgate is a general and flexible inferential approach for a class of model-free targets. We apply it to many interesting problems, including representation learning, privacy analysis, nonlinearity, interactions, heterogeneity, and shape constraints. But there are some remaining questions for future study:

- In this paper, we can handle the mMSE gap with respect to a closed linear subspace (and extend to the convex cone case). The current floodgate framework relies on the properties

of the Hilbert space and ℓ_2 norm. It would be desirable to extend to more general situations where we can go beyond Hilbert spaces and infer model-free targets involving different norms or probability functions. Extensions to such problems can be useful as they apply to quantile regression and survival analysis.

- The floodgate functional choice differs from the one used in Zhang & Janson (2020) for ease of exposition. However, from a power/accuracy standpoint, it would still be of interest to systematically study the construction of floodgate functionals and conduct thorough comparisons to guide how practitioners choose or design optimal floodgate inferential procedures for given inferential tasks at hand.

3

StarTrek: Combinatorial Variable Selection with False Discovery Rate Control

CONTRIBUTION

This chapter is based on a manuscript Zhang & Lu (2021), jointly with Prof. Junwei Lu.

Abstract

Variable selection on the large-scale networks has been extensively studied in the literature. While most of the existing methods are limited to the local functionals especially the graph edges, this paper focuses on selecting the discrete hub structures of the networks. Specifically, we propose an inferential method, called StarTrek filter, to select the hub nodes with degrees larger than a certain thresholding level in the high dimensional graphical models and control the false discovery rate (FDR). Discovering hub nodes in the networks is challenging: there is no straightforward statistic for testing the degree of a node due to the combinatorial structures; complicated dependence in the multiple testing problem is hard to characterize and control. In methodology, the StarTrek filter overcomes this by constructing p-values based on the maximum test statistics via the Gaussian multiplier bootstrap. In theory, we show that the StarTrek filter can control the FDR by providing accurate bounds on the approximation errors of the quantile estimation and addressing the dependence structures among the maximal statistics. To this end, we establish novel Cramér-type comparison bounds for the high dimensional Gaussian random vectors. Comparing to the Gaussian comparison bound via the Kolmogorov distance established by Chernozhukov et al. (2014), our Cramér-type comparison bounds establish the relative difference between the distribution functions of two high dimensional Gaussian random vectors, which is essential in the theoretical analysis of FDR control. Moreover, the StarTrek filter can be applied to general statistical models for FDR control of discovering discrete structures such as simultaneously testing the sparsity levels of multiple high dimensional linear models. We illustrate the validity of the StarTrek filter in a series of numerical experiments and apply it to the genotype-tissue expression dataset to discover central regulator genes.

Keywords. Graphical models, multiple testing, false discovery rate control, combinatorial inference, Gaussian multiplier bootstrap, comparison bounds.

3.1 Introduction

Graphical models are widely used for real-world problems in a broad range of fields, including social science, economics, genetics, and computational neuroscience (Newman et al., 2002; Luscombe et al., 2004; Rubinov & Sporns, 2010). Scientists and practitioners aim to understand the underlying network structure behind large-scale datasets. For a high-dimensional random vector $\boldsymbol{X} = (\boldsymbol{X}_1, \cdots, \boldsymbol{X}_d) \in \mathbb{R}^d$, we let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an undirected graph, which encodes the conditional dependence structure among $\boldsymbol{X}$. Specifically, each component of $\boldsymbol{X}$ corresponds to some vertex in $\mathcal{V} = \{1, 2 \cdots, d\}$, and $(j, k) \notin \mathcal{E}$ if and only if $\boldsymbol{X}_j$ and $\boldsymbol{X}_k$ are conditionally independent given the rest of variables. We denote the associated weight matrix by $\boldsymbol{\Theta}$. Many existing works in the literature seek to learn the structure of $\mathcal{G}$ via estimating the weight matrix $\boldsymbol{\Theta}$. For example, Meinshausen & Bühlmann (2006); Yuan & Lin (2007); Friedman et al. (2008); Rothman et al. (2008); Peng et al. (2009); Lam & Fan (2009); Ravikumar et al. (2011); Cai et al. (2011); Shen et al. (2012) focus on estimating the precision matrix in a Gaussian graphical model. Further, there is also a line of work developing methodology and theory to assess the uncertainty of edge estimation, i.e., constructing hypothesis tests and confidence intervals on the network edges, see Cai & Ma (2013); Gu et al. (2015); Ren et al. (2015); Cai & Zhang (2016); Janková & van de Geer (2017); Yang et al. (2018); Feng & Ning (2019); Ding & Zhou (2020). Recently, simultaneously testing multiple hypotheses on edges of the graphical models has received increasing attention (Liu, 2013; Cai et al., 2013; Xia et al., 2015, 2018; Li & Maathuis, 2019; Eisenach et al., 2020).

Most of the aforementioned works formulate the testing problems based on continuous parameters and local properties. For example, Liu (2013) proposes a method to select edges in Gaussian graphical models with asymptotic FDR control guarantees. Testing the existence of edges concerns the local structure of the graph. Under certain modeling assumptions, its null hypothesis can be translated into a single point in the continuous parameter space, for example, $\boldsymbol{\Theta}_{jk} = 0$ where $\boldsymbol{\Theta}$ is

the precision matrix or the general weight matrix. However, for many scientific questions involving network structures, we need to detect and infer discrete and combinatorial signals in the networks, which does not follow from single edge testing. For example, in the study of social networks, it is interesting to discover active and impactful users, usually called "hub users," as they are connected to many other nodes in the social network (Ilyas et al., 2011; Lee et al., 2019). In gene co-expression network analysis, identifying central regulators/hub genes (Yuan et al., 2017; Liu et al., 2019c,b) is known to be extremely useful to the study of progression and prognosis of certain cancers and can support the treatment in the future. In neuroscience, researchers are interested in identifying the cerebral areas which are intensively connected to other regions (Shaw et al., 2008; van den Heuvel & Sporns, 2013; Power et al., 2013) during certain cognitive processes. The discovery of such central/hub areas can provide scientists with a better understanding of the mechanisms of human cognition.

Motivated by these applications in various areas, in this paper, we consider the hub node selection problem from the network models. In specific, given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the vertex set and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the edge set, we consider multiple hypotheses on whether the degree of some node $j \in \mathcal{V}$ exceeds a given threshold k_τ:

$$H_{0j} : \text{degree of node } j < k_\tau \text{ v.s. } H_{1j} : \text{degree of node } j \geq k_\tau,$$

based on i.i.d. samples $\boldsymbol{X}_1, \cdots \boldsymbol{X}_n \overset{i.i.d.}{\sim} \boldsymbol{X} \in \mathbb{R}^d$. Throughout the paper, these nodes with large degrees will be called hub nodes. For each $j \in [d]$, let $\psi_j = 1$ if H_{0j} is rejected and $\psi_j = 0$ otherwise. When selecting hub nodes, we would like to control the false discovery rate, as defined below:

$$\text{FDR} = \mathbb{E}\left[\frac{\sum_{j \in \mathcal{H}_0} \psi_j}{\max\left\{\sum_{j=1}^{d} \psi_j, 1\right\}}\right],$$

where $\mathcal{H}_0 = \{j \mid \text{degree of node } j < k_\tau\}$. Remark the hypotheses $H_{0j}, j \in [d]$ are not based

on continuous parameters. They instead involve the degrees of the nodes, which are intrinsically discrete/combinatorial functionals. To the best of our knowledge, there is no existing literature studying such combinatorial variable selection problems. The most relevant work turns out to be Lu et al. (2017), which proposes a general framework for inference about graph invariants/combinatorial quantities on undirected graphical models. However, they study single hypothesis testing and have to decide which subgraph to be tested before running the procedure.

The combinatorial variable selection problems bring many new challenges. First, most of the existing work focus on testing continuous parameters (Liu, 2013; Javanmard & Montanari, 2013, 2014a,b; Belloni et al., 2014; Van de Geer et al., 2014; Xia et al., 2015, 2018; Javanmard & Javadi, 2019; Sur & Candès, 2019; Zhao et al., 2020). For discrete functionals, it is more difficult to construct appropriate test statistics and estimate its quantile accurately, especially in high dimensions. Second, many multiple testing procedures rely on an independence assumption (or certain dependence assumptions) on the null p-values (Benjamini & Hochberg, 1995; Benjamini & Yekutieli, 2001; Benjamini, 2010). However, the single hypothesis here is about the global property of the graph, which means that any reasonable test statistic has to involve the whole graph. Therefore, complicated dependence structures exist inevitably, which presents another layer of difficulty for controlling the false discoveries. Now we summarize the motivating question for this paper: how to develop a combinatorial selection procedure to discover nodes with large degrees on a graph with FDR control guarantees?

This paper introduces the StarTrek filter to select hub nodes. The filter is based on the maximum statistics, whose quantiles are approximated by the Gaussian multiplier bootstrap procedure. Briefly speaking, the Gaussian multiplier bootstrap procedure estimates the distribution of a given maximum statistic of general random vectors with unknown covariance matrices by the distribution of the maximum of a sum of the conditional Gaussian random vectors. The validity of high dimensional testing problems, such as family-wise error rate (FWER) control, relies on the non-asymptotic

bounds of the Kolmogorov distance between the true distribution of the maximum statistics and the Gaussian multiplier bootstrap approximation, which is established in Chernozhukov et al. (2013). However, in order to control the FDR in the context of combinatorial variable selection, a more refined characterization of the quantile approximation errors is required. In specific, we need the so called Cramér-type comparison bounds quantifying the accuracy of the p-values in order to control the FDR in the simultaneous testing procedures (Chang et al., 2016). In our context, consider two centered Gaussian random vectors $U, V \in \mathbb{R}^d$ with different covariance matrices $\mathbf{\Sigma}^U, \mathbf{\Sigma}^V$ and denote the ℓ_∞ norms of U, V by $||U||_\infty, ||V||_\infty$ respectively, then the Cramér-type comparison bounds aim to control the relative error $\left|\frac{\mathbb{P}(||U||_\infty > t)}{\mathbb{P}(||V||_\infty > t)} - 1\right|$ for certain range of t. Comparing to the Kolmogorov distance $\sup_{t\in\mathbb{R}} |\mathbb{P}(||U||_\infty > t) - \mathbb{P}(||V||_\infty > t)|$ (Chernozhukov et al., 2015), the Cramér-type comparison bound leads to the relative error between two cumulative density functions, which is necessary to guarantee the FDR control. In specific, we show in this paper a novel Cramér-type Gaussian comparison bound

$$\sup_{0\le t\le C_0\sqrt{\log d}} \left|\frac{\mathbb{P}(||U||_\infty > t)}{\mathbb{P}(||V||_\infty > t)} - 1\right| = O\left(\min\left\{(\log d)^{5/2}\Delta_\infty^{1/2}, \frac{\Delta_0 \log d}{\mathfrak{p}}\right\}\right), \qquad (3.1.1)$$

for some constant $C_0 > 0$, where $\Delta_\infty := ||\mathbf{\Sigma}^U - \mathbf{\Sigma}^V||_{\max}$ is the entrywise maximum norm difference between the two covariance matrices, $\Delta_0 := ||\mathbf{\Sigma}^U - \mathbf{\Sigma}^V||_0$ with $\|\cdot\|_0$ being the entrywise ℓ_0-norm of the matrix, and $\mathfrak{p}$ is the number of connected subgraphs in the graph whose edge set $\mathcal{E} = \{(j,k) : \mathbf{\Sigma}^U_{jk} \neq 0 \text{ or } \mathbf{\Sigma}^V_{jk} \neq 0\}$. This comparison bound in (3.1.1) characterizes the relative errors between Gaussian maxima via two types of rates: the ℓ_∞-norm Δ_∞ and the ℓ_0-norm Δ_0. This implies a new insight that the Cramér type bound between two Gaussian maxima is small as long as either their covariance matrices are uniformly close or only sparse entries of the two covariance matrices differ. As far as we know, the second type of rate in (3.1.1) has not been developed even in Kolmogorov distance results of high dimensional Gaussian maxima. In the study of FDR

control, we need both types of rates: the Δ_∞ rate is used to show that the Gaussian multiplier bootstrap procedure is an accurate approximation for the maximum statistic quantiles and the Δ_0 rate is used to quantify the complicated dependence structure of the p-values for the single tests on the degree of graph nodes. In order to prove the Cramér-type comparison bound in (3.1.1), we develop two novel theoretic techniques to prove the two types of rates separately. For the Δ_∞ rate, we reformulate the Slepian's interpolation (Slepian, 1962) into an ordinary differential inequality such that the relative error can be controlled via the Grönwall's inequality (Grönwall, 1919). To control the Δ_0 rate, the anti-concentration inequality of Gaussian maxima developed in Chernozhukov et al. (2015) is no longer sufficient, we establish a new type of anti-concentration inequality for the derivatives of the soft-max of high dimensional Gaussian vectors. The existing works on the Cramér type comparison bounds such as Liu & Shao (2010, 2014); Chang et al. (2016) does not cover the high dimensional maximum statistics. Therefore, their techniques can not be directly extended to our case. To the best of our knowledge, it is the first time in our paper to prove the Cramér-type Gaussian comparison bounds (3.1.1) for high dimensional Gaussian maxima.

In summary, our paper makes the following major contributions. First, we develop a novel StarTrek filter to select combinatorial statistical signals: the hub nodes with the FDR control. This procedure involves maximum statistic and Gaussian multiplier bootstrap for quantile estimation. Second, in theory, the proposed method is shown to be valid for many different models with the network structures. In this paper, we provide two examples, the Gaussian graphical model and the bipartite network in the multiple linear models. Third, we prove a new Cramér-type Gaussian comparison bound with two types of rates: the maximum norm difference and ℓ_0 norm difference. These results are quite generic and has its own significance in the probability theory.

3.1.1 Related work

Canonical approaches to FDR control and multiple testing (Benjamini & Hochberg, 1995; Benjamini & Yekutieli, 2001; Benjamini, 2010) require that valid p-values are available, and they only allow for certain forms of dependence between these p-values. However, obtaining asymptotic p-values with sufficient accuracy is generally non-trivial for high dimensional hypothesis testing problems concerning continuous parameters (Javanmard & Montanari, 2013, 2014a,b; Belloni et al., 2014; Van de Geer et al., 2014; Sur & Candès, 2019; Zhao et al., 2020), not even to mention discrete/combinatorial functionals.

Recently, there is a line of work conducting variable selection without needing to act on a set of valid p-values, including Barber & Candès (2015, 2019); Candès et al. (2018b); Xing et al. (2019); Dai et al. (2020a,b). These approaches take advantage of the symmetry of the null test statistics and establish FDR control guarantee. As their single hypothesis is often formulated as conditional independence testing, it is challenging to apply those techniques to select discrete signals for the problem studied in this paper.

Another line of work develops multiple testing procedures based on asymptotic p-values for specific high dimensional models (Liu, 2013; Liu & Luo, 2014; Javanmard & Javadi, 2019; Xia et al., 2015, 2018; Liu et al., 2020). Among them, Liu (2013) studies the edge selection problem on Gaussian graphical models, which turns out to be the most relevant work to our paper. However, their single hypothesis is about the local property of the graph. Our problem of discovering nodes with large degrees concerns the global property of the whole network, therefore requiring far more work.

There exists some recent work inferring combinatorial functionals. For example, the method proposed in Ke et al. (2020) provides a confidence interval for the number of spiked eigenvalues in a covariance matrix. Jin et al. (2020) focuses on estimating the number of communities in a network and yields confidence lower bounds. Neykov et al. (2019); Lu et al. (2017) propose a general frame-

work for conducting inference on graph invariants/combinatorial quantities, such as the maximum degree, the negative number of connected subgraphs, and the size of the longest chain of a given graph. Shen & Lu (2020) develops methods for testing the general community combinatorial properties of the stochastic block model. Regarding the hypothesis testing problem, all these works only deal with a single hypothesis and establish asymptotic type-I error rate control. While simultaneously testing those combinatorial hypotheses is also very interesting and naturally arises from many practical problems.

3.1.2 Outline

In Section 3.2, we set up the general testing framework and introduce the StarTrek filter for selecting hub nodes. In Section 3.3, we present our core probabilistic tools: Cramér-type Gaussian comparison bounds in terms of maximum norm difference and ℓ_0 norm difference. To offer a relatively simpler illustration of our generic theoretical results, we first consider the hub selection problem on a bipartite network (multitask regression with linear models). Specifically, the input of the general StarTrek filter is chosen to be the estimators and quantile estimates described in Section 3.4. Applying the probabilistic results under this model, we establish FDR control guarantees under certain conditions. Then we move to the Gaussian graphical model in Section 3.5. In Section 3.6, we demonstrate StarTrek's performance through empirical simulations and a real data application.

3.1.3 Notations

Let $\phi(x), \Phi(x)$ be the probability density function (PDF) and the cumulative distribution function (CDF) respectively of the standard Gaussian distribution and denote $\bar{\Phi}(x) = 1 - \Phi(x)$. Let $\mathbf{1}_d$ be the vector of ones of dimension d. We use $\mathbb{1}(\cdot)$ to denote the indicator function of a set and $|\cdot|$ to denote the cardinality of a set. For two sets A and B, denote their symmetric difference by

$A \ominus B$, i.e., $A \ominus B = (A \setminus B) \cup (B \setminus A)$; let $A \times B$ be the Cartesian product. For two positive sequences $\{x_n\}_{n=1}^{\infty}$ and $\{y_n\}_{n=1}^{\infty}$, we say $x_n = O(y_n)$ if $x_n \leq Cy_n$ holds for any n with some large enough $C > 0$. And we say $x_n = o(y_n)$ if $x_n/y_n \to 0$ as $n \to \infty$. For a sequence of random variables $\{X_n\}_{n=1}^{\infty}$ and a scalar a, we say $X_n \leq a + o_{\mathbb{P}}(1)$ if for all $\epsilon > 0$, $\lim_{n\to\infty} \mathbb{P}(X_n - a > \epsilon) = 0$. Let $[d]$ denote the set $\{1, \ldots, d\}$. The ℓ_∞ norm and the ℓ_1 norm on $\mathbb{R}^d$ are denoted by $||\cdot||_\infty$ and $||\cdot||_1$ respectively. For a random vector X, let $||X||_\infty$ be its ℓ_∞ norm. For a matrix $\mathbf{A} \in \mathbb{R}^{d_1 \times d_2}$, we denote its minimal and maximal eigenvalues by $\lambda_{\min}(\mathbf{A}), \lambda_{\max}(\mathbf{A})$ respectively, the elementwise max norm by $\|\mathbf{A}\|_{\max} = \max_{i\in[d_1], j\in[d_2]} |\mathbf{A}_{ij}|$ and the elementwise ℓ_0 norm by $\|\mathbf{A}\|_0 = \sum_{i\in[d_1], j\in[d_2]} \mathbb{1}(\mathbf{A}_{ij} \neq 0)$. Throughout this paper, $C, C', C'', C_0, C_1, C_2, \ldots$ are used as generic constants whose values may vary across different places.

3.2 Methodology

Before introducing our method, we set up the problem with more details. Specifically, we consider a graph $\mathcal{G} = (\mathcal{V}_1, \mathcal{V}_2, \mathcal{E})$ with the node sets $\mathcal{V}_1, \mathcal{V}_2$ and the edge set $\mathcal{E}$. Let $d_1 = |\mathcal{V}_1|$, $d_2 = |\mathcal{V}_2|$ and denote its weight matrix by $\mathbf{\Theta} \in \mathbb{R}^{d_1 \times d_2}$. In the undirected graph where $\mathcal{V}_1 = \mathcal{V}_2 := \mathcal{V}$, $\mathbf{\Theta}$ is a square matrix and its element $\mathbf{\Theta}_{jk}$ is nonzero when there is an edge between node j and node k, zero when there is no edge. In a bipartite graph where $\mathcal{V}_1 \neq \mathcal{V}_2$, elements of $\mathbf{\Theta}$ describe the existence of an edge between node j in $\mathcal{V}_1$ and node k in $\mathcal{V}_2$. Without loss of generality, we focus on one of the node sets and denote it by $\mathcal{V}$ with $|\mathcal{V}| := d$. We would like to select those nodes among $\mathcal{V}$ whose degree exceeds a certain threshold k_τ, based on the n data samples $\boldsymbol{X}_1, \cdots \boldsymbol{X}_n \overset{i.i.d.}{\sim} \boldsymbol{X} \in \mathbb{R}^d$. And the selection problem is equivalent to simultaneously testing d hypotheses:

$$H_{0j} : \text{degree of node } j < k_\tau \text{ v.s. } H_{1j} : \text{degree of node } j \geq k_\tau, \tag{3.2.1}$$

for $j \in [d]$. Let $\psi_j = 1$ if H_{0j} is rejected and $\psi_j = 0$ otherwise, then for some multiple testing procedure with output $\{\psi_j\}_{j\in[d]}$, the false discovery proportion (FDP) and FDR can be defined as below:

$$\mathrm{FDP} = \frac{\sum_{j\in\mathcal{H}_0}^{d} \psi_j}{\max\left\{1, \sum_{j=1}^{d} \psi_j\right\}}, \quad \mathrm{FDR} := \mathbb{E}[\mathrm{FDP}],$$

where $\mathcal{H}_0 = \{j \mid \text{degree of node } j < k_\tau\}$. Given the data $\boldsymbol{X}_1, \cdots \boldsymbol{X}_n$ from the graphical model, we aim to propose a multiple testing procedure such that the FDP or FDR can be controlled at a given level $0 < q < 1$.

We illustrate the above general setup in two specific examples. In multitask regression with linear models, we are working with the bipartite graph case, then the weight matrix $\boldsymbol{\Theta}$ corresponds to the parameter matrix whose row represents the linear coefficients for one given response variable. Given a threshold k_τ, we want to select those rows (response variables) with ℓ_0 norm being at least k_τ. In the context of Gaussian graphical models where $\mathcal{V}_1 = \mathcal{V}_2$, $\boldsymbol{\Theta}$ represents the precision matrix, and we want to select those hub nodes i.e., whose degree is larger than or equal to k_τ.

3.2.1 StarTrek filter

Letting $\boldsymbol{\Theta}_j$ be the j-th row of $\boldsymbol{\Theta}$ and $\boldsymbol{\Theta}_{j,-j}$ be the vector $\boldsymbol{\Theta}_j$ excluding its j-th element, we formulate the testing problem for each single node as below,

$$H_{0j} : \|\boldsymbol{\Theta}_{j,-j}\|_0 < k_\tau \text{ v.s. } H_{1j} : \|\boldsymbol{\Theta}_{j,-j}\|_0 \geq k_\tau.$$

To test the above hypothesis, we need some estimator of the weight matrix $\boldsymbol{\Theta}$. In Gaussian graphical model, it is natural to use the estimator of a precision matrix. In the bipartite graph (multiple response model), estimated parameter matrix will suffice. Denote this generic estimator by $\widetilde{\boldsymbol{\Theta}}$ (without causing confusion in notation), the maximum test statistic over a given subset E of $\mathcal{V} \times \mathcal{V}$ will

be

$$T_E := \max_{(j,k)\in E} \sqrt{n} \left| \widetilde{\mathbf{\Theta}}_{jk} \right|$$

and its quantile is defined as $c(\alpha, E) = \inf \{t \in \mathbb{R} \mid \mathbb{P}(T_E \leq t) \geq 1 - \alpha\}$, which is often unknown. Assume it can be estimated by $\widehat{c}(\alpha, E)$ from some procedure such as the Gaussian multiplier bootstrap, a generic method called skip-down procedure can be used, which was originally proposed in Lu et al. (2017) for testing a family of monotone graph invariants. When applied to the specific degree testing problem, it leads to the following algorithm.

Algorithm 7 Skip-down Method in Lu et al. (2017) (for testing the degree of node j)

Input: $\{\widetilde{\mathbf{\Theta}}_e\}_{e\in\mathcal{V}\times\mathcal{V}}$, significance level α.
Initialize $t = 0, E_0 = \{(j,k) : k \in [d], k \neq j\}$.
repeat
  $t \leftarrow t + 1$;
  Select the rejected edges $\mathcal{R} \leftarrow \{(j,k) \in E_{t-1} \mid \sqrt{n}|\widetilde{\mathbf{\Theta}}_{jk}| > \widehat{c}(\alpha, E_{t-1})\}$;
  $E_t \leftarrow E_{t-1} \backslash \mathcal{R}$;
until $|E_t^c| \geq k$ or $\mathcal{R} = \varnothing$
Output: $\psi_{j,\alpha} = 1$ if $|E_t^c| \geq k$ and $\psi_{j,\alpha} = 0$ otherwise.

To conduct the node selection over the whole graph, we need to determine an appropriate threshold $\widehat{\alpha}$ then reject H_{0j} if $\psi_{j,\widehat{\alpha}} = 1$. A desirable choice of $\widehat{\alpha}$ should be able to discover as many as hub nodes with the FDR remaining controlled under the nominal level q. For example, if the BHq procedure is considered, $\widehat{\alpha}$ can be defined as follows:

$$\widehat{\alpha} = \sup \left\{ \alpha \in (0,1) : \frac{\alpha d}{\max\left\{1, \sum_{j\in[d]} \psi_{j,\alpha}\right\}} \leq q \right\}. \tag{3.2.2}$$

The above range of α is $(0,1)$, it will be very computationally expensive if we do an exhaustive search since for each α, we have to recompute the quantiles $\widehat{c}(\alpha, E)$ for a lot of sets E.

We overcome the computational difficulty and propose a efficient procedure called StarTrek fil-

ter, which is presented in Algorithm 8. Remark it only involves estimating k_τ different quantiles of

Algorithm 8 StarTrek Filter

Input: $\{\widetilde{\boldsymbol{\Theta}}_e\}_{e\in\mathcal{V}\times\mathcal{V}}$, nominal FDR level q.
for $j \in [d]$ **do**
We order the elements in $\{|\widetilde{\boldsymbol{\Theta}}_{j\ell}| : \ell \neq j\}$ as $|\widetilde{\boldsymbol{\Theta}}_{j,(1)}| \geq |\widetilde{\boldsymbol{\Theta}}_{j,(2)}| \geq \ldots \geq |\widetilde{\boldsymbol{\Theta}}_{j,(d-1)}|$, where $|\widetilde{\boldsymbol{\Theta}}_{j,(\ell)}|$ is the ℓth largest entry. Compute $\alpha_j = \max_{1\leq s\leq k_\tau} \widehat{c}^{-1}(\sqrt{n}|\widetilde{\boldsymbol{\Theta}}_{j,(s)}|, E_j^{(s)})$ where $E_j^{(s)} := \{(j,\ell) : \ell \neq j, |\widetilde{\boldsymbol{\Theta}}_{j\ell}| \leq |\widetilde{\boldsymbol{\Theta}}_{j,(s)}|\}$.
end for
Order α_j as $\alpha_{(1)} \leq \alpha_{(2)} \leq \cdots \leq \alpha_{(d)}$ and set $\alpha_{(0)} = 0$, let $j_{\max} = \max\{0 \leq j \leq d : \alpha_{(j)} \leq qj/d\}$.
Output: $S = \{j : \alpha_j \leq \alpha_{(j_{\max})}\}$ if $j_{\max} > 0$; $S = \varnothing$ otherwise.

some maximum statistics per node, which is more efficient than the Skip-down procedure (Lu et al., 2017) in terms of computation.

3.2.2 Accuracy of approximate quantiles

Before diving into the theoretical results, we pause to give specific forms of the estimator of $\boldsymbol{\Theta}$ and how to compute the estimated quantiles of the maximum statistic. Take the Gaussian graphical model as an example, suppose that $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \overset{\text{i.i.d.}}{\sim} N_d(0, \boldsymbol{\Sigma})$. Let $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$, which will have the same ℓ_0 elementwise norm as the adjacency matrix $\boldsymbol{\Theta}$. Denote $\mathbf{e}_k$ be the kth canonical basis in $\mathbb{R}^d$, we consider the following one-step estimator of $\boldsymbol{\Theta}_{jk}$,

$$\widehat{\boldsymbol{\Theta}}^{\mathrm{d}}_{jk} := \widehat{\boldsymbol{\Theta}}_{jk} - \frac{\widehat{\boldsymbol{\Theta}}_j^\top \left(\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\Theta}}_k - \mathbf{e}_k\right)}{\widehat{\boldsymbol{\Theta}}_j^\top \widehat{\boldsymbol{\Sigma}}_j}, \tag{3.2.3}$$

where $\widehat{\boldsymbol{\Theta}}$ could be either the graphical Lasso (GLasso) estimator (Friedman et al., 2008) or the CLIME estimator (Cai et al., 2011). Let $\widetilde{\boldsymbol{\Theta}}^{\mathrm{d}}_{jk} := \widehat{\boldsymbol{\Theta}}^{\mathrm{d}}_{jk}/\sqrt{\widehat{\boldsymbol{\Theta}}^{\mathrm{d}}_{jj}\widehat{\boldsymbol{\Theta}}^{\mathrm{d}}_{kk}}$ and the standardized version $\{\widetilde{\boldsymbol{\Theta}}^{\mathrm{d}}_e\}_{\mathcal{V}\times\mathcal{V}}$ will be the input $\{\widetilde{\boldsymbol{\Theta}}_e\}_{\mathcal{V}\times\mathcal{V}}$ of Algorithm 8. Then the maximum test statistics (over the

subset E) is defined as $T_E = \max_{(j,k)\in E} \sqrt{n}|\widetilde{\mathbf{\Theta}}^{\mathrm{d}}_{jk}|$. To estimate its quantile, we construct the following Gaussian multiplier bootstrap

$$T_E^{\mathcal{B}} := \max_{(j,k)\in E} \frac{1}{\sqrt{n\,\widehat{\mathbf{\Theta}}_{jj}\widehat{\mathbf{\Theta}}_{kk}}} \left| \sum_{i=1}^{n} \widehat{\mathbf{\Theta}}_j^\top \left(\boldsymbol{X}_i \boldsymbol{X}_i^\top \widehat{\mathbf{\Theta}}_k - \mathbf{e}_k \right) \xi_i \right|, \tag{3.2.4}$$

where $\xi_i \overset{\text{i.i.d.}}{\sim} N(0,1)$, which produces $\widehat{c}(\alpha, E) = \inf\left\{t \in \mathbb{R} : \mathbb{P}_\xi\left(T_E^{\mathcal{B}} \le t\right) \ge 1-\alpha\right\}$ as the quantile estimate. We also denote the standardized true precision matrix $(\mathbf{\Theta}_{jk}/\sqrt{\mathbf{\Theta}_{jj}\mathbf{\Theta}_{kk}})_{j,k\in[d]}$ by $\mathbf{\Theta}^\star$. The theoretical results for Gaussian multiplier bootstrap developed in Chernozhukov et al. (2013) basically imply the above quantile estimates are accurate in the following sense:

Lemma 3.2.1. *Suppose* $\mathbf{\Theta} \in \mathcal{U}(M, s, r_0)$ *and* $(\log(dn))^7/n + s^2(\log dn)^4/n = o(1)$, *for any edge set* $E \subseteq \mathcal{V} \times \mathcal{V}$, *we have*

$$\lim_{(n,d)\to\infty} \sup_{\mathbf{\Theta}\in\mathcal{U}(M,s,r_0)} \sup_{\alpha\in(0,1)} \left| \mathbb{P}\left(\max_{e\in E} \sqrt{n}|\widetilde{\mathbf{\Theta}}^d_e - \mathbf{\Theta}^\star_e| > \widehat{c}(\alpha, E) \right) - \alpha \right| = 0. \tag{3.2.5}$$

where $\widetilde{\mathbf{\Theta}}^d_e$ *is the standardized version of the one-step estimator* (3.2.3).

Note that $\mathcal{U}(M, s, r_0)$ denotes the parameter space of precision matrices and is defined as below:

$$\mathcal{U}(M, s, r_0) = \left\{ \mathbf{\Theta} \in \mathbb{R}^{d\times d} \,\middle|\, \lambda_{\min}(\mathbf{\Theta}) \ge 1/r_0, \lambda_{\max}(\mathbf{\Theta}) \le r_0, \max_{j\in[d]} \|\mathbf{\Theta}_j\|_0 \le s, \|\mathbf{\Theta}\|_1 \le M \right\}.$$

The proof of Lemma 3.2.1 can be found in Appendix C.4.2. However, Lemma 3.2.1 is not sufficient for our multiple testing problem. Generally speaking, the probabilistic bounds in Chernozhukov et al. (2013) are in terms of Kolmogorov distance, which only provides a uniform characterization for the deviation behaviors. Their results can be used to establish FWER control for global testing problems based on the maximum test statistics. However, in order to establish FDR

control, we have to show that the estimation of number of false discoveries is sufficiently accurate enough in the following sense, i.e., uniformly over certain range of α,

$$\frac{\alpha d_0}{\sum_{j\in\mathcal{H}_0}\psi_{j,\alpha}} \to 1, \quad \text{in probability}$$

where $\mathcal{H}_0 = \{j : \|\mathbf{\Theta}_{j,-j}\|_0 < k_\tau\}$. The above result is different from the one needed for FWER control: $\mathbb{E}[\psi_{j,\alpha}] = \alpha + o(1), j \in \mathcal{H}_0$. In the context of our node selection problem, it can be reduced to the following,

$$\left|\frac{\sum_{j\in\mathcal{H}_0}\mathbb{1}(\max_{e\in E}\sqrt{n}|\widetilde{\mathbf{\Theta}}^{\mathrm{d}}_e - \mathbf{\Theta}^{\star}_e| \geq \widehat{c}(\alpha, E))}{d_0\alpha} - 1\right| \to 0 \quad \text{in probability}$$

uniformly over certain range of α for some subset E. The above ratio is closedly related to the ratio in Cramér-type moderation deviation results (Liu & Shao, 2010, 2014; Liu, 2013). To this end, we establish the Cramér-type deviation bounds for the Gaussian multiplier bootstrap procedure. This type of results is built on two types of Cramér-type Gaussian comparison bounds, which are presented in Section 3.3.

3.3 Cramér-type comparison bounds for Gaussian maxima

In this section, we present the theoretic results on the Cramér-type comparison bounds for Gaussian maxima. Let $U, V \in \mathbb{R}^d$ be two centered Gaussian random vectors with different covariance matrices $\mathbf{\Sigma}^U = (\sigma^U_{jk})_{1\leq j,k\leq d}, \mathbf{\Sigma}^V = (\sigma^V_{jk})_{1\leq j,k\leq d}$. Recall that the maximal difference of the covariance matrices is $\Delta_\infty := ||\mathbf{\Sigma}^U - \mathbf{\Sigma}^V||_{\max}$ and the elementwise ℓ_0 norm difference of the covariance matrices is denoted by $\Delta_0 := \left\|\mathbf{\Sigma}^U - \mathbf{\Sigma}^V\right\|_0 = \sum_{j,k\in[d]}\mathbb{1}(\sigma^U_{jk} \neq \sigma^U_{jk})$. The Gaussian maxima of U and V are denoted as $||U||_\infty$ and $||V||_\infty$. Now we present a Cramér-type comparison bound (CCB) between Gaussian maxima in terms of the maximum norm difference Δ_∞.

Theorem 3.3.1 (CCB with maximum norm difference). *Suppose* $(\log d)^5 \Delta_\infty = O(1)$, *then we have*

$$\sup_{0 \le t \le C_0 \sqrt{\log d}} \left| \frac{\mathbb{P}(||U||_\infty > t)}{\mathbb{P}(||V||_\infty > t)} - 1 \right| = O\left((\log d)^{5/2} \Delta_\infty^{1/2} \right), \tag{3.3.1}$$

for some constant $C_0 > 0$.

Remark 3.3.1.1. *We can actually prove a more general form (see Theorem C.2.2 in the appendix) of the upper bound on the above term, without the assumption on* Δ_∞. *In fact, we bound the right hand side of* (3.3.1) *as* $M_3 (\log d)^{3/2} A(\Delta_\infty) e^{M_3 (\log d)^{3/2} A(\Delta_\infty)}$, *where* $A(\Delta_\infty) = M_1 \log d \Delta_\infty^{1/2} \exp\left(M_2 \log^2 d \Delta_\infty^{1/2}\right)$ *with the constants* M_1, M_2 *only depending on the variance terms* $\min_{1 \le j \le d} \{\sigma_{jj}^U, \sigma_{jj}^V\}, \max_{1 \le j \le d} \{\sigma_{jj}^U, \sigma_{jj}^V\}$ *and* M_3 *being a universal constant.*

When applying Theorem 3.3.1 to Gaussian multiplier bootstrap, $||\Delta||_\infty$ actually controls the maximum differences between the true covariance matrix and the empirical covariance matrix. Based on the bound of $||\Delta||_\infty$, we can show that the Cramér-type comparison bound in (3.3.1) will be $O((\log d)^{3/2} n^{-1/4})$ with high probability.

The proof can be found in Appendix C.2.1. The above result bounds the relative difference between the distribution functions of the two Gaussian maxima. Compared with the bound in terms of Kolmogorov distance, it has more refined characterization when t is large, which benefits from our iterative use of the Slepian interpolation. We denote the interpolation between U and V as $W(s) = \sqrt{s}U + \sqrt{1-s}V, s \in [0,1]$ and let $Q_t(s) = \mathbb{P}(||W(s)||_\infty > t)$. Existing results (Chernozhukov et al., 2013, 2014) quantify the difference between $Q_t(1)$ and $Q_t(0)$ uniformly over $t \in \mathbb{R}$, which leads to a bound on the Kolmogorov distance between Gaussian maxima. Our main innovation is to consider $R_t(s) = Q_t(s)/Q_t(0) - 1$ and show that for any given t, $\mathcal{R}_t : s \in [0,1] \mapsto |R_t(s)|$ is a contraction mapping with 0 being its fixed point. Specifically, we

have the following upper bound on $|R_t(s)|$,

$$|R_t(s)| \leq AB \int_0^s |R_t(\mu)| d\mu + AB \cdot s + A,$$

where AB and A can be controlled via the bound on the maximal difference of the covariance matrices Δ_∞ and converge to 0 under certain conditions. By Grönwall's inequality (Grönwall, 1919), we then derive the bound on $R_t(1)$ explicitly in terms of A and B, which finally lead to the desired Cramér-type comparison bound in (3.3.1).

The above theorem is a key ingredient for deriving Cramér-type deviation results for the Gaussian multiplier bootstrap procedure. However, in certain situations, comparison bounds in terms of maximum norm difference may not be appropriate. There exist cases where the covariance matrices of two Gaussian random vectors are not uniformly closed to each other, but have lots of identical entries. In particular, for the combinatorial variable selection problem in this paper, there exist complicated dependence structures between the maximum statistic for different nodes, since each time when the degree of one single node is tested, the statistic is computed based on the whole graph. Again, this highlights the challenge of the multiple testing problem in our paper. To establish FDR control, we need to deal with the dependence between the maximum statistic of pairs of non-hub nodes. By the definition of non-hub nodes, the covariance matrix difference between each pair of the involving Gaussian vectors actually has lots of zero entries. We would like to take advantage of this sparsity pattern when applying the comparison bound. However, the bound in (3.3.1) is not sharp when Δ_∞ is not negligible but Δ_0 is small. To this end, we develop a different version of the Cramér-type comparison bound as below.

Theorem 3.3.2 (CCB with elementwise ℓ_0-norm difference)**.** *Assume the Gaussian random vectors U and V have unit variances, i.e., $\sigma^U_{jj} = \sigma^V_{jj} = 1, j \in [d]$ and there exists some $\sigma_0 < 1$ such that $|\sigma^V_{jk}| \leq \sigma_0, |\sigma^U_{jk}| \leq \sigma_0$ for any $j \neq k$. Suppose there exists a disjoint $\mathfrak{p}$-partition of nodes*

$\cup_{\ell=1}^{\mathfrak{p}} \mathcal{C}_\ell = [d]$ *such that* $\sigma^U_{jk} = \sigma^V_{jk} = 0$ *when* $j \in \mathcal{C}_\ell$ *and* $k \in \mathcal{C}_{\ell'}$ *for some* $\ell \neq \ell'$. *We have*

$$\sup_{0 \leq t \leq C_0 \sqrt{\log d}} \left| \frac{\mathbb{P}(||U||_\infty > t)}{\mathbb{P}(||V||_\infty > t)} - 1 \right| \leq O\left(\frac{\Delta_0 \log d}{\mathfrak{p}} \right), \tag{3.3.2}$$

for some constant $C_0 > 0$.

When applying the above result to our multiple degree testing problem, specifically the covariance of maximum test statistics for pairs of non-hub nodes, Δ_0 can be controlled as k_τ^2 which is in a constant order. In Theorem 3.3.2, the quantity $\mathfrak{p}$ represents the number of connected subgraphs shared by the coviarance matrix networks of U and V. We refer to Theorem C.2.4 in the appendix for a generalized definition of $\mathfrak{p}$ to strengthen the results in (3.3.2). The $\mathfrak{p}$ in the denominator of the right hand side of Cramér-type comparison bound in (3.3.2) is necessary: it is possible that even if Δ_0 is small, when $\mathfrak{p}$ is large, the Camér-type Gaussian comparison bound is not converging to zero. For example, consider Gaussian vectors with unit variances $U = (X_1, X_2, Z, \ldots, Z) \in \mathbb{R}^d$, $V = (Y_1, Y_2, Z, \ldots, Z) \in \mathbb{R}^d$, where $\text{corr}(X_1, X_2) = 0.9$, $\text{corr}(Y_1, Y_2) = 0$ and $(X_1, X_2) \perp\!\!\!\perp Z$, $(Y_1, Y_2) \perp\!\!\!\perp Z$. For this case, the Camér-type Gaussian comparison bound

$$\sup_{0 \leq t \leq C_0 \sqrt{\log d}} \left| \frac{\mathbb{P}(||U||_\infty > t)}{\mathbb{P}(||V||_\infty > t)} - 1 \right| = \sup_{0 \leq t \leq C_0 \sqrt{\log d}} \left| \frac{\mathbb{P}(\max\{|X_1|, |X_2|, |Z|\} > t)}{\mathbb{P}(\max\{|Y_1|, |Y_2|, |Z|\} > t)} - 1 \right|$$

is not converging to zero as d goes to infinity even if the corresponding Δ_0 is 1 but $\mathfrak{p} = 2$.

Compared with Theorem 3.3.1, the above theorem provides a sharper comparison bound for large $\mathfrak{p}$ and small Δ_0. The two theorems together describe a interesting phase transition phenomenon, i.e., the dependence on $\mathbf{\Sigma}^U - \mathbf{\Sigma}^V$ of the Cramér-type comparison bound exhibits a difference behavior in the regime of large $\mathfrak{p}$ and small Δ_0 versus the regime of small Δ_∞.

The proof of Theorem 3.3.2 can be found in Appendix C.2.2. Our main technical innovation is to establish a new type of anti-concentration bound for "derivatives" of Gaussian maxima. Since

both the indicator function and maximum function are discontinuous, we follows the idea of using smoothing approximation as in the proof of Theorem 3.3.1, specifically, we bound the following term

$$\mathbb{E}[|\partial_j \partial_k \varphi(W(s))| \cdot \mathbb{1}(t - \epsilon \leq ||W(s)||_\infty \leq t + \epsilon)], \tag{3.3.3}$$

where φ is the same approximation function of the indicator of ℓ_∞ norm with certain smoothing parameter β. Note that $\mathbb{E}[\mathbb{1}(t - \epsilon \leq ||W(s)||_\infty \leq t + \epsilon)]$ is the anti-concentration bound for Gaussian maxima (Chernozhukov et al., 2014). A non-uniform version is also established in Kuchibhotla et al. (2021). (3.3.3) can be viewed as the anti-concentration bound on the second order partial derivatives of the smooth approximation function. When deriving the comparison bound in terms of ℓ_0 norm difference, we have to deal with such terms as (3.3.3) when $\sigma^U_{jk} \neq \sigma^V_{jk}$. We show (3.3.3) can be controlled as

$$\mathbb{E}[|\partial_j \partial_k \varphi(W(s))| \cdot \mathbb{1}(t - \epsilon \leq ||W(s)||_\infty \leq t + \epsilon)] \lesssim \frac{\mathbb{P}\left(||V||_\infty > t\right) (\log d)^2}{\epsilon \beta \mathfrak{p}}.$$

The above anti-concentration bound is non-uniform and has only a logarithm dependence on the dimension d. It provides a relatively sharp characterization when t is large and the Gaussian graphical model is not highly connected (i.e., the number of connected components $\mathfrak{p}$ being large).

3.4 Discovering hub responses in multitask regression

The theoretical results presented in Section 3.3 will be the cornerstone for establishing FDR control of the multiple testing problem described in Section 3.2. As seen previously, the testing problem (3.2.1) is set up in a quite general way: $\boldsymbol{\Theta}$ is a weight matrix, and we would like to select rows whose ℓ_0 norm exceeds some threshold. This section considers the specific application to multitask/multiple response regression, which turns out to be less involved. We take advantage of it and

demonstrate how to utilize the probabilistic tools in Section 3.3. After that, the theoretical results on FDR control for the Gaussian graphical models are presented and discussed in Section 3.5.

In multitask regression problem, multiple response variables are regressed on a common set of predictors. We can view this example as a bipartite graph $\mathcal{G} = (\mathcal{V}_1, \mathcal{V}_2, \mathcal{E}), |\mathcal{V}_1| = d_1, |\mathcal{V}_2| = d_2$, where $\mathcal{V}_1$ contains the response variables and $\mathcal{V}_2$ represents the common set of predictors. Each entry of the weight matrix $\boldsymbol{\Theta}$ indicates whether a given predictor is non-null or not for a given response variable. In the case of parametric model, $\boldsymbol{\Theta} \in \mathbb{R}^{d_1 \times d_2}$ corresponds to the parameter matrix. One might be interested in identifying shared sparsity patterns across different response variables. It can be solved by selecting a set of predictors being non-null for all response variables (Obozinski et al., 2006; Dai & Barber, 2016). This section problem is column-wise in the sense that we want to select columns of $\boldsymbol{\Theta}$, denoted by $\boldsymbol{\Theta}_{\cdot j}$, such that $||\boldsymbol{\Theta}_{\cdot j}||_0 = d_1$. It is also interesting to consider a row-wise selection problem formalized in (3.2.1). Under the multitask regression setup, we would like to select response variables with at least a certain amount of non-null predictors. We will call this type of response variables hub responses throughout the section. This has practical applications in real-world problems such as the gene-disease network.

Consider the multitask regression problem with linear models, we have n i.i.d. pairs of the response vector and the predictor vector, denoted by $(\boldsymbol{Y}_1, \boldsymbol{X}_1), (\boldsymbol{Y}_2, \boldsymbol{X}_2), \ldots, (\boldsymbol{Y}_n, \boldsymbol{X}_n)$, where $\boldsymbol{Y}_i \in \mathbb{R}^{d_1}, \boldsymbol{X}_i \in \mathbb{R}^{d_2}$ satisfy the following relationship,

$$\boldsymbol{Y}_i = \boldsymbol{\Theta}\boldsymbol{X}_i + \boldsymbol{E}_i, \text{ where } \boldsymbol{E}_i \sim \mathcal{N}(0, \mathbf{D}_{d_1 \times d_1}) \text{ and } \boldsymbol{X}_i \perp\!\!\!\perp \boldsymbol{E}_i, \tag{3.4.1}$$

where $\boldsymbol{\Theta} \in \mathbb{R}^{d_1 \times d_2}$ is the parameter matrix and $\mathbf{D}$ is a d_1 by d_1 diagonal matrix whose diagonal elements σ_j^2 is the noise variance for response variable $\boldsymbol{Y}^{(j)}$. Let $\boldsymbol{X}$ be the design matrix with rows $\boldsymbol{X}_1^\top, \ldots, \boldsymbol{X}_n^\top$, shared by different response variables, and assume the noise variables are independent conditional on the design matrix $\boldsymbol{X}$. Let $s = \max_{j \in [d_1]} ||\boldsymbol{\Theta}_j||_0$ be the sparsity level of the pa-

rameter matrix $\mathbf{\Theta}$, we want to select columns of the parameter matrix which has at least k_τ nonzero entries, i.e., select nodes with large degree among $[d_1]$ in the bipartite graph $\mathcal{G} = (\mathcal{V}_1, \mathcal{V}_2, \mathcal{E})$.

As mentioned in Section 3.2, some estimator of the parameter matrix is needed to conduct hypothesis testing. Debiased Lasso is widely used for parameter estimation and statistical inference in high dimensional linear models (Javanmard & Montanari, 2014a,b). For each response variable $\boldsymbol{Y}^{(j)}, j \in [d_1]$, we compute the debiased Lasso estimator, denoted by $\widetilde{\mathbf{\Theta}}_j^{\mathrm{d}}$ as

$$\widetilde{\mathbf{\Theta}}_j^{\mathrm{d}} = \widehat{\mathbf{\Theta}}_j + \frac{1}{n}\mathbf{M}\boldsymbol{X}^\top(\boldsymbol{Y}^{(j)} - \boldsymbol{X}\widehat{\mathbf{\Theta}}_j), \text{ where } \widehat{\mathbf{\Theta}}_j = \arg\min_{\beta\in\mathbb{R}^{d_2}} \left\{\frac{1}{2n}\|\boldsymbol{Y}^{(j)} - \boldsymbol{X}\beta\|_2^2 + \lambda\|\beta\|_1\right\}. \tag{3.4.2}$$

Note the above $\mathbf{M}$ is defined as $\mathbf{M} = (m_1, \ldots, m_{d_2})^\top$ where

$$m_i = \arg\min_m m^\top \widehat{\mathbf{\Sigma}} m, \quad \text{s.t. } \|\widehat{\mathbf{\Sigma}} m - e_i\|_\infty \le \mu, \tag{3.4.3}$$

and here $\widehat{\mathbf{\Sigma}} = (\boldsymbol{X}^\top\boldsymbol{X})/n$.

Then the debiased estimator of the parameter matrix, defined by $\widetilde{\mathbf{\Theta}}^{\mathrm{d}} := (\widetilde{\mathbf{\Theta}}_1^{\mathrm{d}}, \cdots, \widetilde{\mathbf{\Theta}}_{d_1}^{\mathrm{d}})^\top$, will be used the input $\{\widetilde{\mathbf{\Theta}}_e\}_{e\in\mathcal{V}_1\times\mathcal{V}_2}$ of Algorithm 8. In addition, we also need to compute the quantile of the maximum statistics. There exist many work studying the asymptotic distribution of the debiased Lasso estimator. Among them, the results in Javanmard & Montanari (2014a) (when translated into our multitask regression setup) imply, for each response variable $\boldsymbol{Y}^{(j)}, j \in [d_1]$,

$$\sqrt{n}(\widetilde{\mathbf{\Theta}}_j^{\mathrm{d}} - \mathbf{\Theta}_j) = Z + \Xi, \quad Z|\boldsymbol{X} \sim \mathcal{N}(0, \sigma_j^2 M\widehat{\mathbf{\Sigma}}M^\top), \tag{3.4.4}$$

under proper assumptions. Additionally with a natural probabilistic model of the design matrix, the bias term can be showed to be $||\Xi||_\infty = O(\frac{s\log d_2}{\sqrt{n}})$ with high probability. As discussed in (Ja-

vanmard & Montanari, 2014a), the asymptotic normality result can be used for deriving confidence intervals and statistical hypothesis tests. As the noise variance σ_j is unknown, the scaled Lasso is used for its estimation (Javanmard & Montanari, 2014a; Sun & Zhang, 2012), given by the following joint optimization problem,

$$\{\widehat{\boldsymbol{\Theta}}_j, \widehat{\sigma}_j\} = \arg \min_{\beta \in \mathbb{R}^{d_2}, \sigma > 0} \left\{ \frac{1}{2\sigma n} \|\boldsymbol{Y}^{(j)} - \boldsymbol{X}\beta\|_2^2 + \frac{\sigma}{2} + \lambda \|\beta\|_1 \right\}. \tag{3.4.5}$$

Regarding our testing problem, intuitively we can use the quantile of the Gaussian maxima of $\mathcal{N}(0, \widehat{\sigma}_j^2 M \widehat{\boldsymbol{\Sigma}} M^\top)$ to approximate the quantile of maximum statistic $T_E = \max_{(j,k) \in E} \sqrt{n} |\widetilde{\boldsymbol{\Theta}}^{\mathrm{d}}_{jk}|$ for some given subset E. Specifically, let $Z_j \mid \boldsymbol{X}, \boldsymbol{Y}^{(j)} \sim \mathcal{N}(0, \widehat{\sigma}_j^2 M \widehat{\boldsymbol{\Sigma}} M^\top)$ where $Z_j \in \mathbb{R}^{d_2}$ and consider the subset $E \subset \{j\} \times \mathcal{V}_2$, we approximate the quantile of T_E by the following

$$T_E^{\mathcal{N}} := \max_{(j,k) \in E} |Z_{jk}|, \quad \widehat{c}(\alpha, E) = \inf \left\{ t \in \mathbb{R} : \mathbb{P}_Z \left(T_E^{\mathcal{N}} \leq t \right) \geq 1 - \alpha \right\}. \tag{3.4.6}$$

Indeed, under proper scaling conditions, similar results as (3.2.5) can be established, i.e., as $n, d \to \infty$,

$$\sup_{\alpha \in (0,1)} \left| \mathbb{P} \left(\max_{(j,k) \in E} \sqrt{n} |\widetilde{\boldsymbol{\Theta}}^{\mathrm{d}}_{jk} - \boldsymbol{\Theta}_{jk}| > \widehat{c}(\alpha, E) \right) - \alpha \right| \to 0. \tag{3.4.7}$$

The above result is based on two ingredients: the asymptotic normality result and the control of the bias term Ξ. Below we list the required assumptions for those two ingredients, i.e., (3.4.4) and $||\Xi||_\infty = O(\frac{s \log d_2}{\sqrt{n}})$.

Assumption 3.4.1 (Debiased Lasso with random designs)**.** The following assumptions are from the ones of Theorems 7 and 8 in Javanmard & Montanari (2014a).

- Let $\boldsymbol{\Sigma} = \mathbb{E}\left[\boldsymbol{X}_1 \boldsymbol{X}_1^\top\right] \in \mathbb{R}^{d_2 \times d_2}$ be such that $\sigma_{\min}(\boldsymbol{\Sigma}) \geq C_{\min} > 0$, and $\sigma_{\max}(\boldsymbol{\Sigma}) \leq C_{\max} < \infty$, and $\max_{j \in [d_2]} \boldsymbol{\Sigma}_{jj} \leq 1$. Assume $\boldsymbol{X}\boldsymbol{\Sigma}^{-1/2}$ have independent subgaussian

rows, with zero mean and subgaussian norm $\|\mathbf{\Sigma}^{-1/2}\boldsymbol{X}_i\|_{\psi_2} = \kappa$, for some constant $\kappa \in (0, \infty)$.

- $\mu = a\sqrt{(\log d_2)/n}$, and $n \geq \max(\nu_0 s \log(d_2/s), \nu_1 \log d_2), \nu_1 = \max(1600\kappa^4, a/4)$, and $\lambda = \sigma\sqrt{(c^2 \log d_2)/n}$.

Remark that there may exist other ways of obtaining a consistent estimator of $\mathbf{\Theta}$ and sufficiently accurate quantile estimates under different assumptions. Since it is not the main focus of this paper, we will not elaborate on it. As mentioned before, the Kolmogorov type result in (3.4.7) can be immediately applied to the global testing problem to guarantee FWER control. However, it is not sufficient for FDR control of the multiple testing problem in this paper. And this is when the Cramér-type comparison bound for Gaussian maxima established in Section 3.3 play its role. In addition, signal strength condition is needed. Recall that $\mathcal{H}_0 = \{j \in [d_1] : ||\mathbf{\Theta}_j||_0 < k_\tau\}$ with $d_0 = |\mathcal{H}_0|$, we consider the following rows of $\mathbf{\Theta}$,

$$\mathcal{B} := \{j \in \mathcal{H}_0^c : \forall k \in \text{supp}(\mathbf{\Theta}_j), |\mathbf{\Theta}_{jk}| > c\sqrt{\log d_2/n}\}, \tag{3.4.8}$$

and define the proportion of such rows as $\rho = |\mathcal{B}|/d_1$. In the context of multitask regression, ρ measures the proportion of hub response variables whose non-null parameter coefficients all exceed certain thresholds, thus characterizes the overall signal strength. Below we present our result on FDP/FDR control under appropriate assumptions.

Theorem 3.4.2 (FDP/FDR control)**.** *Under Assumption 3.4.1 and the scaling condition* $\frac{d_2 \log d_2 + d_0}{d_0 d_2 \rho} + \frac{s \log^2 d_2}{n^{1/2}} + \frac{\log^2 d_2}{(n\rho)^{1/5}} = o(1)$*, if we implement the StarTrek procedure in Algorithm 8 with* $\mathbf{\Theta}$ *estimated by* (3.4.2) *and the quantiles approximated by* (3.4.6)*, as* $(n, d_1, d_2) \to \infty$*, we have*

$$\mathit{FDP} \leq q\frac{d_0}{d_1} + o_{\mathbb{P}}(1) \quad \textit{and} \quad \lim_{(n,d_1,d_2)\to\infty} \mathit{FDR} \leq q\frac{d_0}{d_1}. \tag{3.4.9}$$

The proof of Theorem 3.4.2 can be found in Appendix C.1.3. Note that signal strength conditions which require some entries of parameter matrix $\mathbf{\Theta}$ have magnitudes exceeding $c\sqrt{\log d_2/n}$ are usually assumed in existing work studying FDR control problem for high dimensional models (Liu, 2013; Liu & Shao, 2014; Liu & Luo, 2014; Xia et al., 2015, 2018; Javanmard & Javadi, 2019).

3.5 Discovering hub nodes in Gaussian graphical models

This section focuses on the hub node selection problem on Gaussian graphical models. Recall in Section 3.2, we first compute the one-step estimator $\{\widehat{\mathbf{\Theta}}^{\mathrm{d}}_e\}_{e\in\mathcal{V}\times\mathcal{V}}$ in (3.2.3) then take its standardized version $\{\widetilde{\mathbf{\Theta}}^{\mathrm{d}}_e\}_{e\in\mathcal{V}\times\mathcal{V}}$ as the input of Algorithm 8 i.e.,

$$\widehat{\mathbf{\Theta}}^{\mathrm{d}}_{jk} := \widehat{\mathbf{\Theta}}_{jk} - \frac{\widehat{\mathbf{\Theta}}_j^\top\left(\widehat{\mathbf{\Sigma}}\widehat{\mathbf{\Theta}}_k - \mathbf{e}_k\right)}{\widehat{\mathbf{\Theta}}_j^\top\widehat{\mathbf{\Sigma}}_j}, \quad \widetilde{\mathbf{\Theta}}^{\mathrm{d}}_{jk} := \widehat{\mathbf{\Theta}}^{\mathrm{d}}_{jk}/\sqrt{\widehat{\mathbf{\Theta}}^{\mathrm{d}}_{jj}\widehat{\mathbf{\Theta}}^{\mathrm{d}}_{kk}}. \tag{3.5.1}$$

Our StarTrek filter selects nodes with large degrees based on the maximum statistics $T_E = \max_{(j,k)\in E}\sqrt{n}|\widetilde{\mathbf{\Theta}}^{\mathrm{d}}_{jk}|$ over certain subset E. We use the Gaussian multiplier bootstrap (3.2.4) to approximate the quantiles, specifically,

$$\widehat{c}(\alpha, E) = \inf\left\{t\in\mathbb{R} : \mathbb{P}_\xi\left(T_E^{\mathcal{B}} \le t\right) \ge 1-\alpha\right\}. \tag{3.5.2}$$

Chernozhukov et al. (2013) shows that this quantile approximation is accurate enough for FWER control in modern high dimensional simultaneous testing problems. Their results are based on the control of the non-asymptotic bounds in a Kolmogorov distance sense. Lu et al. (2017) also takes advantage of this result to test single hypothesis of graph properties or derive confidence bounds on graph invariants.

However, in order to conduct combinatorial variable selection with FDR control guarantees, we need more refined studies about the accuracy of the quantile approximation. This is due to the ratio nature of the definition of FDR, as explained in Section 3.2.2. Compared with the results in

Chernozhukov et al. (2013), we provide a Cramér-type control on the approximation errors of the Gaussian multiplier bootstrap procedure. This is built on the probabilistic tools in Section 3.3, in particular, the Cramér-type Gaussian comparison bound with max norm difference in Theorem 3.3.1. Due to the dependence structure behind the hub selection problem in Graphical models, we also have to utilize Theorem 3.3.2. In a bit more detail, computing the maximum test statistic for testing node node actually involves the whole graph, resulting complicated dependence among the test statistics. The non-differentiability of the maximum function makes it very difficult to track this dependence. Also note that, this type of difficulty can not be easily circumvented by alternative methods, due to the discrete nature of the combinatorial inference problem. However, we figure out that the Cramér-type Gaussian comparison bound with ℓ_0 norm difference plays an important role in handling this challenge.

In general, the sparsity/density of the graph is closed related to the dependence level of multiple testing problem on graphical models. For example, Liu (2013); Xia et al. (2015, 2018) make certain assumptions on the sparsity level and control the dependence of test statistics when testing multiple hypotheses on graphical models/networks. For the hub node selection problem in this paper, a new quantity is introduced, and we will explain why it is suitable. Recall that we define the set of non-hub response variables in Section 3.4. Similarly, the set of non-hub nodes is denoted by $\mathcal{H}_0 = \{j \in [d] : ||\mathbf{\Theta}_j||_0 < k_\tau\}$ with $d_0 = |\mathcal{H}_0|$. Now we consider the following set,

$$S = \{(j_1, j_2, k_1, k_2) : j_1, j_2 \in \mathcal{H}_0, j_1 \neq j_2, k_1 \neq k_2, \mathbf{\Theta}_{j_1 j_2} = \mathbf{\Theta}_{j_1 k_1} = \mathbf{\Theta}_{j_2 k_2} = 0, \mathbf{\Theta}_{j_1 k_2} \neq 0, \mathbf{\Theta}_{j_2 k_1} \neq 0\}. \tag{3.5.3}$$

Remark that in the above definition, k_1 can be the same as j_2 and k_2 can be the same as j_1. If there exists a large number of nodes which are neither connected to j_1 nor j_2, we then do not need to worry much about the dependence between the test statistics for non-hub nodes. Therefore, $|S|$ actually measures the dependence level via checking how a pair of non-hub nodes interact through

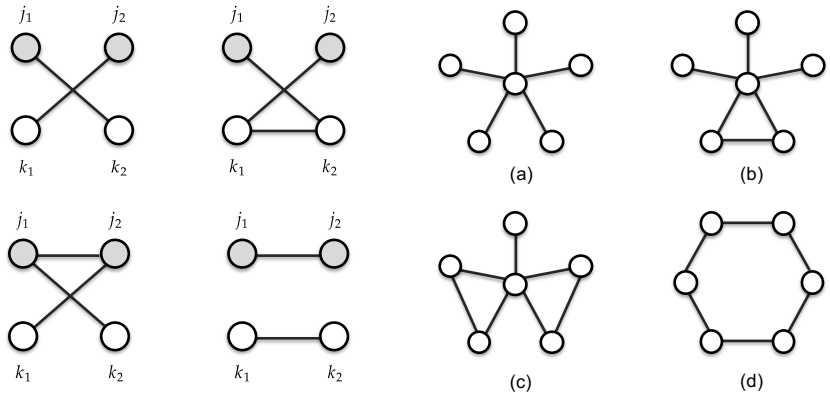


Figure 3.1: Left panel: a graphical demonstration of the definition of S via four examples of a 4-vertex graph; Right panel: four different graph patterns with 6 vertices. Calculating $|S|$ yields $10, 15, 24, 51$ for (a),(b),(c),(d) respectively.

other nodes. Liu (2013); Cai et al. (2013) also examine the connection structures in the 4-vertex graph and control the dependence level by carefully bounding the number of the 4-vertex graphs with different numbers of edges.

We provide a graphical demonstration of S and show how $|S|$ looks like in certain types of graph patterns via some simple examples. Though the definition of S does not exclude the possibility of (j_1, j_2, k_1, k_2) being a graph with 2 or 3 vertices, we only draw 4-vertex graph in Figure 3.1 for convenience. In the left panel of Figure 3.1, we consider four different cases of the 4-vertex graph. The upper two belong to the set S, while the lower three do not. In the right panel, we consider four graphs which all have 6 vertices. They have different graph patterns. For example, (a) clearly has a hub structure. All of the non-hub nodes are only connected to the hub node. While in (d), the edges are evenly distributed and each node are connected to its two nearest neighbours. For each graph, we count the value of $|S|$ and obtain $10, 15, 24, 51$ respectively, which show a increasing trend of $|S|$. This sort of matches our intuition that it is relatively easier to discover hub nodes on graph (a) compared with graph (d). See more evidence in the empirical results of Section 3.6.

In addition to $|S|$, we also characterize the dependence level via the connectivity of the graph, specifically let p be the number of connected components. And similarly as in Section 3.4, we define

ρ to measure the signal strength, i.e., $\rho = |\mathcal{B}|/d$, where $\mathcal{B} := \{j \in \mathcal{H}_0^c : \forall k \in \text{supp}(\mathbf{\Theta}_j), |\mathbf{\Theta}_{jk}| > c\sqrt{\log d/n}\}$. In the following, we list our assumptions needed for FDR control.

Assumption 3.5.1. Suppose that $\mathbf{\Theta} \in \mathcal{U}(M, s, r_0)$ and the following conditions hold:

(i) Signal strength and scaling condition.

$$\frac{\log d}{\rho}\left(\frac{(\log d)^{19/6}}{n^{1/6}} + \frac{(\log d)^{11/6}}{\rho^{1/3}n^{1/6}} + \frac{s(\log d)^3}{n^{1/2}}\right) = o(1). \tag{3.5.4}$$

(ii) Dependency and connectivity condition.

$$\frac{\log d}{\rho d_0} + \frac{(\log d)^2|S|}{\rho d_0^2 p} = o(1). \tag{3.5.5}$$

In the above assumption, (3.5.4) places conditions on the signal strength and scaling. The first and the second term come from the Cramér-type large deviation bounds in the high dimensional CLT setting (Kuchibhotla et al., 2021) and the Cramér-type Gaussian comparison bound established in Theorem 3.3.1. And the third term comes from the fact that the relevant test statistics arise as maxima of approximate averages instead of the exact averages and thus the approximation error needs to be controlled. See similar discussions about this in (Chernozhukov et al., 2013). Remark that the signal strength condition is mild here, due to similar reasons as the discussion in Section 3.4. Regarding (3.5.5), there is a trade-off between the dependence level and connectivity level of the topological structure. $|S|/d_0^2$ characterizes how the test statistics of non-hub nodes are correlated to each other in average. p by definition describes the level of connectivity. Due to the condition (3.5.5), larger signal strength generally makes the hub selection problem easier. And when $|S|/d_0^2$ is small, the graph is allowed to be more connected. When there exist more sub-graphs, we allow higher correlations between the non-hub nodes. Note that the cardinality of S is directly related to the ℓ_0 norm covariance matrix difference term Δ_0, and arises from the application of Theorem

3.3.2. In the following, we present our core theoretical result on FDP/FDR control for hub selection using the StarTrek filter on Gaussian graphical models.

Theorem 3.5.2 (FDP/FDR control)**.** *Under Assumption 3.5.1, the StarTrek procedure in Algorithm 8 with* (3.5.1) *as input and the quantiles approximated by* (3.5.2) *satisfies: as* $(n, d) \to \infty$*,*

$$FDP \leq q\frac{d_0}{d} + o_{\mathbb{P}}(1) \quad \textit{and} \quad \lim_{(n,d)\to\infty} FDR \leq q\frac{d_0}{d}. \tag{3.5.6}$$

The proof can be found in Appendix C.1.1. Remark that control of the FDR does not prohibit the FDP from varying. Therefore our result on FDP provides a stronger guarantee on controlling the false discoveries. See clear empirical evidence in Section 3.6.1. To the best of our knowledge, the proposed StarTrek filter in Section 3.2 and the above FDP/FDR control result are the first Algorithm and theoretical guarantee for the problem of simultaneously selecting hub nodes. Existing work like Liu (2013); Liu & Luo (2014); Xia et al. (2015, 2018); Javanmard & Javadi (2019) focus on the discovery of continuous signals and their tools are not applicable to the problem here.

3.6 Numerical results

3.6.1 Synthetic data

In this section, we apply the StarTrek filter to synthetic data and demonstrate the performance of our method. The synthetic datasets are generated from Gaussian graphical models. The corresponding precision matrices are specified based on four different types of graphs. Given the number of nodes d and the number of connected components p, we will randomly assign those nodes into p groups. Within each group (sub-graph), the way of assigning edges for different graph types will be explained below in detail. After determinning the adjacency matrix of the graph, we follow Zhao et al. (2012) to construct the precision matrix, more specifically, we set the off-diagonal elements to

be of value v which control the magnitude of partial correlations and is closely related to the signal strength. In order to ensure positive-definiteness, we add some value v together with the absolute value of the minimal eigenvalues to the diagonal terms. In the following simulations, v and u are set to be 0.4 and 0.1 respectively. Now we explain how to determine the edges within each group (sub-graph) for four different graph patterns.

- **Hub graph.** We randomly pick one node as the hub node of the sub-graph, then the rest of the nodes are made to connect with this hub node. There is no edge between the non-hub nodes.

- **Random graph.** This is the Erdös-Rényi random graph. There is an edge between each pair of nodes with certain probability independently. In the following simulations, we will set this probability to be 0.15 unless stated otherwise.

- **Scale-free graph.** In this type of graphs, the degree distribution follows a power law. We construct it by the Barabási-Albert algorithm: starting with two connected nodes, then adding each new node to be connected with only one node in the existing graph; and the probability is proportional to the degree of the each node in the existing graph. The number of the edges will be the same as the number of nodes.

- **K-nearest-neighbor (knn) graph.** For a given number of k, we add edges such that each node is connected to another k nodes. In our simulations, k is sampled from $\{1, 2, 3, 4\}$ with probability mass $\{0.4, 0.3, 0.2, 0.1\}$.

See a visual demonstration of the above four different graph patterns in Appendix C.5.1. Throughout the simulated examples, we fix the number of nodes d to be 300 and vary other quantities such as sample size n or the number of connected components p. To estimate the precision matrix, we run the graphical Lasso algorithm with 5-fold cross-validation. Then we obtain the standardized

debiased estimator as described in (3.2.3). To obtain the quantile estimates, we use the Gaussian multiplier bootstrap with 4000 bootstrap samples. The threshold k_τ for determining hub nodes is set to be 3. And all results (of FDR and power) are averaged over 64 independent replicates.

As we can see from Table 3.1, the FDRs of StarTrek filter for different types of graph are well controlled below the nominal levels. In hub graph, the FDRs are relatively small but the power is still pretty good. Similar phenomenon for multiple edge testing problem is observed (Liu, 2013). In the context of node testing, it is also unsurprising. These empirical results actually match our demonstration about $|S|$ in Figure 3.1: hub graphs have a relatively weaker dependence structure (smaller S values) and make it is easier to discover true hub nodes without making many errors.

Table 3.1: Empirical FDR

$d = 300$	$q = 0.1$			$q = 0.2$		
n	200	300	400	200	300	400
			$p = 20$			
hub	0.0000	0.0000	0.0007	0.0000	0.0000	0.0029
random	0.0255	0.0383	0.0467	0.0521	0.0770	0.0833
scale-free	0.0093	0.0211	0.0282	0.0352	0.0486	0.0581
knn	0.0101	0.0296	0.0370	0.0228	0.0620	0.0769
			$p = 30$			
hub	0.0013	0.0000	0.0016	0.0027	0.0054	0.0036
random	0.0347	0.0359	0.0568	0.0725	0.0753	0.0963
scale-free	0.0215	0.0335	0.0317	0.0521	0.0624	0.0584
knn	0.0297	0.0420	0.0563	0.0504	0.0857	0.1030

The power performance of the StarTrek filter is showed in Table 3.2. As the sample size grows, we see the power is increasing for all four different types of graphs. When p is larger, there are more hub nodes in general due to the way of constructing the graphs, and we find the power is higher. Among different types of graphs, the power in hub graph and scale-free graph is higher than that in random and knn graph since the latter two are relatively denser and have more complicated topological structures.

Table 3.2: Power

$d = 300$	$q = 0.1$			$q = 0.2$		
n	200	300	400	200	300	400
			$p = 20$			
hub	0.7109	0.9453	0.9898	0.7805	0.9648	0.9938
random	0.3343	0.7815	0.9408	0.4520	0.8514	0.9604
scale-free	0.4524	0.8145	0.9363	0.5281	0.8614	0.9568
knn	0.0905	0.5306	0.8067	0.1634	0.6511	0.8630
			$p = 30$			
hub	0.6848	0.9244	0.9706	0.7588	0.9459	0.9784
random	0.4882	0.8863	0.9790	0.5770	0.9225	0.9870
scale-free	0.6472	0.9047	0.9810	0.7197	0.9331	0.9870
knn	0.2409	0.6841	0.8922	0.3298	0.7706	0.9241

In Figure 3.2 and 3.3, we demonstrate the performance of our method in the random graph with different parameters. Specifically, we vary the connecting probability changing from 0.1 to 0.3 in the x-axis. In those plots, we see the FDRs are all well controlled below the nominal level $q = 0.1$. As the connecting probability of the random graph grows, the graph gets denser, resulting more hub nodes. Thus we can see the height of the short blue solids lines (representing qd_0/d) is decreasing. Based on our results in Theorem 3.5.2, the target level of FDP/FDR control is qd_0/d. This is why we find the mean and median of each box-plot is getting smaller as the connecting probability increases (hence d_0 decreases).

The box-plots and the jittering points show that our StarTrek procedure not only controls the FDR but also prohibit it from varying too much, as implied by the theoretical results on FDP control in Section 3.5. Regarding the power plots, we see that the power is smaller when the graph is denser since the hub selection problem becomes more difficult with more disturbing factors. Plots with nominal FDR level $q = 0.2$ are deferred to Appendix C.5.3.

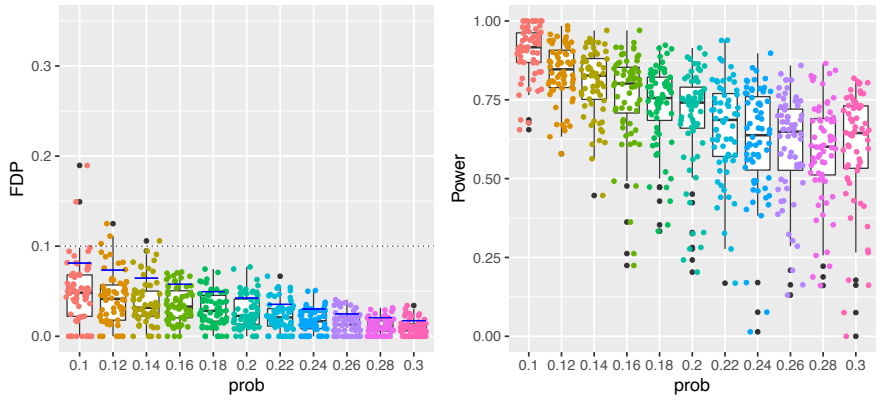


Figure 3.2: FDP and power plots for the StarTrek filter in the random graph. The connecting probability is varied on the x-axis. The number of samples n is chosen to be 300 and the number of connected components p equals 20. The nominal FDR level is set to be $q = 0.1$; the short blue solid lines correspond to qd_0/d, calculated by averaging over the 64 replicates. For both panels, the box plots are plotted with the black points representing the outliers. Colored points are jittered around, demonstrating how the FDP and power distribute.

3.6.2 Application to gene expression data

We also apply our method to the Genotype-Tissue Expression (GTEx) data studied in Lonsdale et al. (2013). Beginning with a 2.5-year pilot phase, the GTEx project establishes a great database and associated tissue bank for studying the relationship between certain genetic variations and gene expressions in human tissues. The original dataset involves 54 non-diseased tissue sites across 549 research subjects. Here we only focus on analyzing the breast mammary tissues. It is of great interest to identify hub genes over the gene expression network.

First we calculate the variances of the gene expression data and focus on the top 100 genes in the following analysis. The data involves $n = 291$ samples for male individuals and $n = 168$ samples for female individuals. The original count data is log-transformed and scaled. We then obtain the estimator of the precision matrix by the Graphical Lasso with 2-fold cross-validation. As for the hub node criterion, we set k_τ as the 50% quantile of the node degrees in the estimated precision matrix. We run StarTrek filter with 2000 bootstrap samples and nominal FDR level $q = 0.1$ to select hub

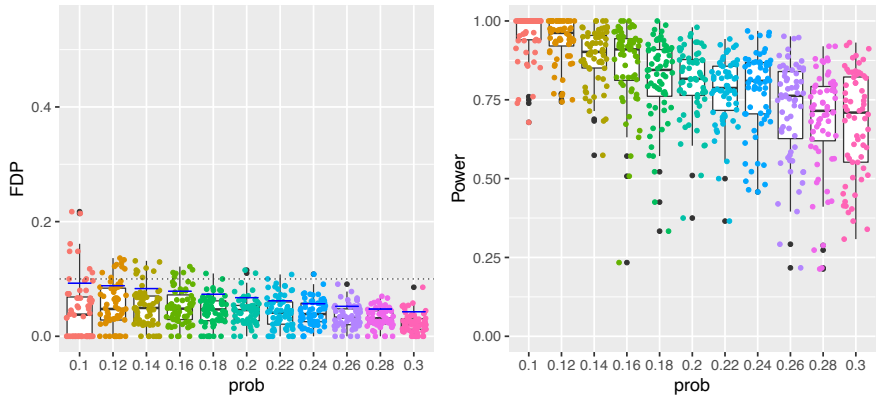


Figure 3.3: FDP and power plots for the StarTrek filter in the random graph. The other setups are the same as Figure 3.3 except for $p = 30$.

genes for both the male and female datasets.

Figure 3.4 shows that the selected hub genes by the StarTrek filter also have large degrees on the estimated gene networks (based on the estimated precision matrices). In Figure 3.5, the results for male and female dataset agree with each other except that the number of selected hub genes using female dataset is smaller due to a much smaller sample size. The selected hub genes are found to play an important role in breast-related molecular processes, either as central regulators or their abnormal expressions are considered as the causes of breast cancer initiation and progression, see relevant literature in genetic research such as Hellwig et al. (2016); Blein et al. (2015); Chen et al. (2016); Li et al. (2019); Lou et al. (2020); Mohamed et al. (2014); Bai et al. (2019); Sirois et al. (2019); Marino et al. (2020); Malvia et al. (2019). Therefore, our proposed method for selecting hub nodes can be applied to the hub gene identification problem. It may improve our understanding of the mechanisms of breast cancer and provide valuable prognosis and treatment signature.

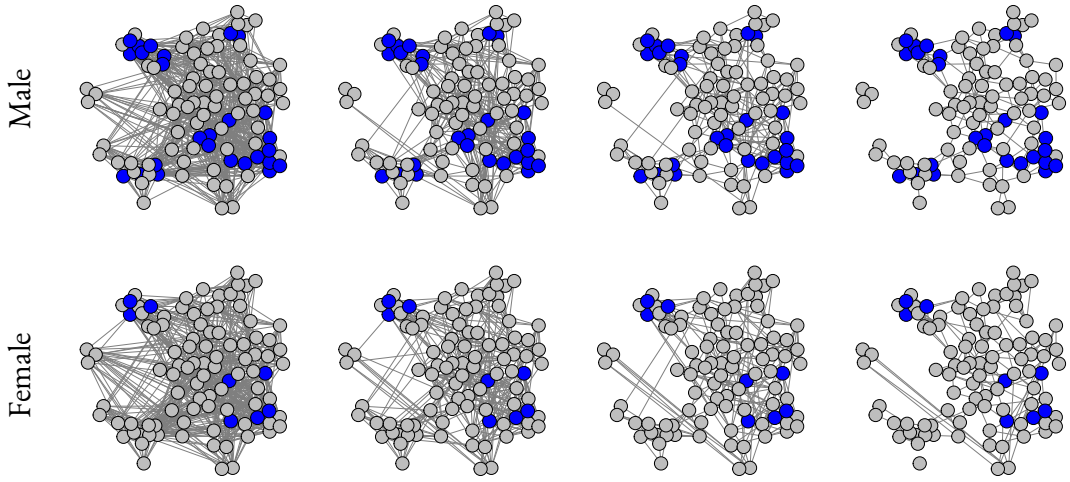


Figure 3.4: The above graphs are based the estimated precision matrices (the left two plots). The adjacency matrices of the other six plots are based on the standardized estimated precision matrices but thresholded at $0.025, 0.05, 0.075$ respectively. Blue vertices represent the selected hub genes.

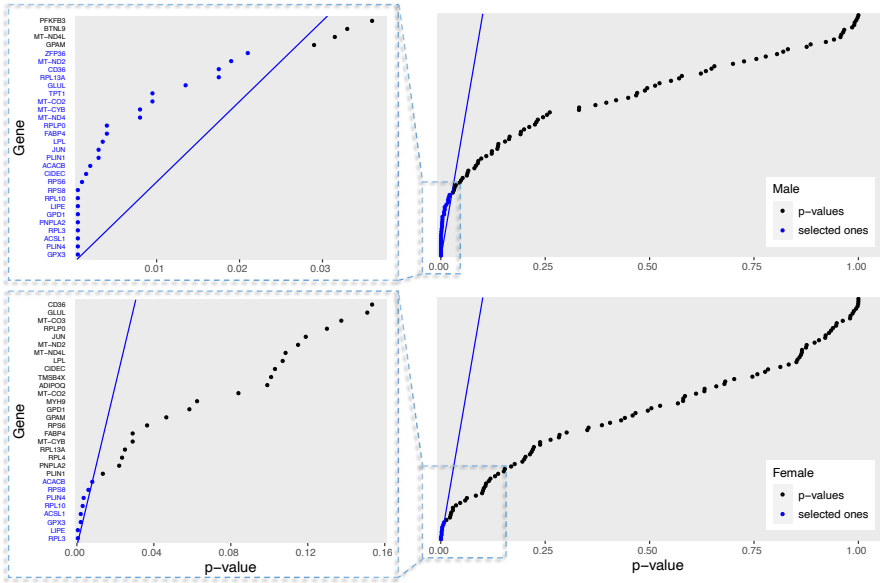


Figure 3.5: Plots of the sorted p-values ($\alpha_j, j \in [d]$) in Algorithm 8. Those blue points correspond to selected hub genes. The blue line is the rejection line of the BHq procedure. The coordinates of the plots are flipped. We abbreviate the names of the 100 genes and only show selected ones with blue colored text.

A

Appendix of Chapter 1

A.1 Proofs for main text

Throughout the proofs, we will abbreviate $(X, Z) = W$, $(\widetilde{X}, Z) = \widetilde{W}$ for simplicity and write $w = (x, z)$. And $g^\star, g : \mathbb{R}^{p-1} \to \mathbb{R}$; $h^\star, h : \mathbb{R}^p \to \mathbb{R}$ are defined as below:

$$g^\star(z) = \mathbb{E}[\mu^\star(W) \,|\, Z = z],\ g(z) = \mathbb{E}[\mu(W) \,|\, Z = z], \tag{A.1.1}$$

$$h^\star(w) = \mu^\star(w) - g^\star(z),\ h(w) = \mu(w) - g(z). \tag{A.1.2}$$

And we can further decompose Y:

$$Y = \mathbb{E}\left[Y \mid X, Z\right] + \epsilon(Y, X, Z) = \mu^\star(W) + \epsilon(Y, W) = g^\star(Z) + h^\star(W) + \epsilon(Y, W). \quad \text{(A.1.3)}$$

Let $L_2(\Omega, \mathcal{F}, P)$ denote the vector space of real-valued random variables with finite second moments, which is a Hilbert space, and define its subspace $L_2(W) := L_2(\Omega, \mathscr{A}(W), P)$, where $\mathscr{A}(W)$ is the sub σ-algebra generated by $W = (X, Z)$. ($L_2(Z) := L_2(\Omega, \mathscr{A}(Z), P)$ is defined analogously). Then $\mu^\star(W)$ and $g^\star(Z)$ can be interpreted as the projections of Y onto the subspaces $L_2(W)$ and $L_2(Z)$, respectively. Y and $\mu^\star(W)$ admit the orthogonal decompositions $Y = \mu^\star(W) + \epsilon(Y, W)$ and $\mu^\star(W) = g^\star(Z) + h^\star(W)$, respectively. Similarly note the projection of $\mu(W)$ onto $L_2(Z)$ and the decomposition $\mu(W) = g(Z) + h(W)$. We remark these imply the following facts:

$$\begin{gathered} \mathbb{E}[\epsilon(Y, W) \mid W] = 0,\ \mathbb{E}\left[\epsilon(Y, W)\lambda(W)\right] = 0, \\ \mathbb{E}[h^\star(W) \mid Z] = 0,\ \mathbb{E}\left[h^\star(W)\gamma(Z)\right] = 0, \mathbb{E}[h(W) \mid Z] = 0,\ \mathbb{E}\left[h(W)\gamma(Z)\right] = 0. \end{gathered} \quad \text{(A.1.4)}$$

for any function $\lambda(w)$ and any function $\gamma(z)$. Thus we can rewrite the denominator of $f(\mu)$ by noticing the equivalence below:

$$\mathbb{E}\left[\mathrm{Var}(\mu(X, Z) \mid Z)\right] = \mathbb{E}\left[\mathrm{Var}(h(W) \mid Z)\right] = \mathbb{E}\left[\mathbb{E}\left[h^2(W) \mid Z\right]\right] = \mathbb{E}\left[h^2(W)\right]. \quad \text{(A.1.5)}$$

As for the numerator of $f(\mu)$, (1.2.7) mentions the rewritten expression. Here we formally derive the following equivalent expressions of $f(\mu)$,

$$\begin{aligned} f(\mu) &:= \frac{\mathbb{E}\left[\mathrm{Cov}(\mu^\star(X,Z),\mu(X,Z)\mid Z)\right]}{\sqrt{\mathbb{E}\left[\mathrm{Var}(\mu(X,Z)\mid Z)\right]}} \\ &= \frac{\mathbb{E}\left[\mathrm{Cov}(h^\star(W),h(W)\mid Z)\right]}{\sqrt{\mathbb{E}\left[h^2(W)\right]}} \\ &= \frac{\mathbb{E}\left[h^\star(W)h(W)\right]}{\sqrt{\mathbb{E}\left[h^2(W)\right]}} \qquad \text{(A.1.6)} \\ &= \frac{\mathbb{E}\left[Yh(W)\right]}{\sqrt{\mathbb{E}\left[h^2(W)\right]}} - \frac{\mathbb{E}\left[\epsilon(Y,W)h(W)\right]}{\sqrt{\mathbb{E}\left[h^2(W)\right]}} - \frac{\mathbb{E}\left[g^\star(Z)h(W)\right]}{\sqrt{\mathbb{E}\left[h^2(W)\right]}} \\ &= \frac{\mathbb{E}\left[Yh(W)\right]}{\sqrt{\mathbb{E}\left[h^2(W)\right]}} \qquad \text{(A.1.7)} \end{aligned}$$

where the second equality is by (A.1.5) and the definitions of $h^\star(W)$, $h(W)$, the third equality holds by the total law of conditional expectation and (A.1.4), the fourth equality comes from (A.1.3), and the last equality holds due to (A.1.4) and the total law of conditional expectation. As (A.1.7) is very concise, we will work with this expression of $f(\mu)$ throughout the following proof. Also note we have an equivalent expression of $\mathcal{I}$.

$$\sqrt{\mathbb{E}\left[(h^\star)^2(W)\right]} = \sqrt{\mathbb{E}\left[\mathbb{E}\left[\left(\mu^\star(W)-\mathbb{E}\left[\mu^\star(W)\mid Z\right]\right)^2\,\middle|\, Z\right]\right]} = \sqrt{\mathbb{E}\left[\mathrm{Var}\left(\mathbb{E}\left[Y\mid X,Z\right]\mid Z\right)\right]} = \mathcal{I}. \tag{A.1.8}$$

Note that the proofs of Theorems 1.2.3 and 1.2.5 only require moment conditions on $h(W)$, which will hold under the corresponding moment conditions on $\mu(X,Z)$. This can be seen from the following example where the finiteness of $\mathbb{E}\left[\mu^r(W)\right]$ implies that of $\mathbb{E}\left[h^r(W)\right]$ for some posi-

tive integer r:

$$\begin{aligned}\mathbb{E}\left[h^r(W)\right] &= \mathbb{E}\left[(\mu(W) - \mathbb{E}\left[\mu(W) \mid Z\right])^r\right] \\ &\leq 2^{r-1}(\mathbb{E}\left[\mu^r(W)\right] + \mathbb{E}\left[(\mathbb{E}\left[\mu(W) \mid Z\right])^r\right]) \\ &\leq 2^{r-1}(\mathbb{E}\left[\mu^r(W)\right] + \mathbb{E}\left[\mathbb{E}\left[\mu^r(W) \mid Z\right]\right]) = 2^r\mathbb{E}\left[\mu^r(X,Z)\right],\end{aligned} \tag{A.1.9}$$

where the first inequality holds by the C_r inequality (which states that $\mathbb{E}\left[|X+Y|^r\right] \leq C_r(\mathbb{E}\left[|X|^r\right] + \mathbb{E}\left[|Y|^r\right])$ with $C_r = 1$ for $0 < r \leq 1$ and $C_r = 2^{r-1}$ for $r \geq 1$), the second inequality holds by Jensen's inequality, and the last equality holds due to the tower property of conditional expectation.

In the proofs of Theorems 1.2.3 and 1.2.5, we will use a key fact to simplify exposition: when $\mathbb{E}\left[h^2(W)\right] > 0, \mathbb{E}\left[h^2(W)\right] = 1$ can be assumed without loss of generality. This is because (A.1.7) says $f(\mu) = \frac{\mathbb{E}[Yh(W)]}{\sqrt{\mathbb{E}[h^2(W)]}}$ and R_i, V_i in Algorithm 1 can be rewritten as

$$\begin{aligned}R_i &= Y_i(\mu(X_i, Z_i) - \mathbb{E}\left[\mu(X_i, Z_i \mid Z_i\right]) = Y_i h(W_i), \\ V_i &= \mathrm{Var}\left(\mu(X_i, Z_i) \mid Z_i\right) = \mathrm{Var}\left(h(X_i, Z_i) \mid Z_i\right)\end{aligned}$$

by definition of h. Regarding Theorem 1.2.5, R_i^K, V_i^K can be rewritten as

$$\begin{aligned}R_i^K &= Y_i\left(h(W_i) - \frac{1}{K}\sum_{k=1}^{K} h(X_i^{(k)}, Z_i)\right), \\ V_i^K &= \frac{1}{K-1}\sum_{k=1}^{K}\left(h(X_i^{(k)}, Z_i) - \frac{1}{K}\sum_{k=1}^{K} h(X_i^{(k)}, Z_i)\right)^2\end{aligned}$$

due to (A.1.51), (A.1.53). It is immediate that the floodgate procedure is invariant to positive scaling thus we assume $\mathbb{E}\left[h^2(W)\right] = 1$ without loss of generality.

A.1.1 Proofs in Section 1.2.2

Lemma 1.2.2

Proof of Lemma 1.2.2. When $\mathbb{E}\left[\mathrm{Var}(\mu(X,Z)\mid Z)\right] = 0$, the numerator must also be zero, and hence the ratio is 0 by convention and $f(\mu) \leq \mathcal{I}$. Now assuming $\mathbb{E}\left[\mathrm{Var}(\mu(X,Z)\mid Z)\right] > 0$,

$$\begin{aligned}
f(\mu) &= \frac{\mathbb{E}\left[\mathrm{Cov}(\mu(X,Z),\mu^\star(X,Z)\mid Z)\right]}{\sqrt{\mathbb{E}\left[\mathrm{Var}(\mu(X,Z)\mid Z)\right]}} \\
&= \frac{\mathbb{E}\left[\sqrt{\mathrm{Var}(\mu(X,Z)\mid Z)}\sqrt{\mathrm{Var}(\mu^\star(X,Z)\mid Z)}\mathrm{Cor}\left(\mu(X,Z),\mu^\star(X,Z)\mid Z\right)\right]}{\sqrt{\mathbb{E}\left[\mathrm{Var}(\mu(X,Z)\mid Z)\right]}} \\
&\leq \frac{\mathbb{E}\left[\sqrt{\mathrm{Var}(\mu(X,Z)\mid Z)}\sqrt{\mathrm{Var}(\mu^\star(X,Z)\mid Z)}\right]}{\sqrt{\mathbb{E}\left[\mathrm{Var}(\mu(X,Z)\mid Z)\right]}} \\
&\leq \frac{\sqrt{\mathbb{E}\left[\mathrm{Var}(\mu(X,Z)\mid Z)\right]}\sqrt{\mathbb{E}\left[\mathrm{Var}(\mu^\star(X,Z)\mid Z)\right]}}{\sqrt{\mathbb{E}\left[\mathrm{Var}(\mu(X,Z)\mid Z)\right]}} = \mathcal{I},
\end{aligned}$$

where the first inequality uses the fact that correlation is bounded by 1, and the second inequality uses Cauchy–Schwarz. Finally, it is immediate that $f(\mu^\star) = \mathcal{I}$. □

Theorem 1.2.3

Proof of Theorem 1.2.3. Due to (A.1.9), $\mathbb{E}\left[\mu^4(X,Z)\right] < \infty$ implies $\mathbb{E}\left[h^4(W)\right] < \infty$. In the following proof, we will only assume the weaker moment conditions $\mathbb{E}\left[Y^4\right], \mathbb{E}\left[h^4(W)\right] < \infty$. Under such moment conditions, we also have $\mathbb{E}\left[Yh(W)\right] \leq \sqrt{\mathbb{E}\left[Y^2\right]}\sqrt{\mathbb{E}\left[h^2(W)\right]}$ and $\mathbb{E}\left[h^2(W)\right] < \infty$ since the finiteness of higher moments implies that of lower moments.

When $\mu(X,Z) \in \mathscr{A}(Z)$, i.e., $\mathbb{E}\left[\mathrm{Var}(\mu(X,Z)\mid Z)\right] = 0$, we immediately have coverage since $L_n^\alpha(\mu) = 0$ by construction and $\mathcal{I} \geq 0$ by its definition. Regarding the case where $\mathbb{E}\left[\mathrm{Var}(\mu(X,Z)\mid Z)\right] \neq 0$, we have $\mathbb{E}\left[h^2(W)\right] = \mathbb{E}\left[\mathrm{Var}(\mu(X,Z)\mid Z)\right] > 0$ due to (A.1.5).

Based on the discussions in the part after (A.1.9), we can assume $\mathbb{E}\left[h^2(W)\right] = 1$ without loss of generality.

Recall in Algorithm 1, we denote $R_i = Y_i\big(\mu(X_i, Z_i) - \mathbb{E}\left[\mu(X_i, Z_i) \mid Z_i\right]\big)$ and $V_i = \mathrm{Var}\left(\mu(X_i, Z_i) \mid Z_i\right)$ for each $i \in [n]$, and compute their sample mean $(\bar{R}, \bar{V})$ and sample covariance matrix $\widehat{\Sigma}$. The LCB is constructed as

$$L_n^\alpha(\mu) = \max\left\{\frac{\bar{R}}{\sqrt{\bar{V}}} - \frac{z_\alpha s}{\sqrt{n}}, 0\right\}, \text{ where } s^2 = \frac{1}{\bar{V}}\left[\left(\frac{\bar{R}}{2\bar{V}}\right)^2 \widehat{\Sigma}_{22} + \widehat{\Sigma}_{11} - \frac{\bar{R}}{\bar{V}}\widehat{\Sigma}_{12}\right].$$

And we have

$$\{L_n^\alpha(\mu) \le \mathcal{I}\} = \left\{\frac{\bar{R}}{\sqrt{\bar{V}}} - \frac{z_\alpha s}{\sqrt{n}} \le \mathcal{I}\right\} \supset \left\{\frac{\bar{R}}{\sqrt{\bar{V}}} - \frac{z_\alpha s}{\sqrt{n}} \le f(\mu)\right\},$$

where the first equality holds since $\mathcal{I} \ge 0$ and the subset relation holds due to Lemma 1.2.2. Hence it suffices to show that

$$\mathbb{P}\left(\frac{\bar{R}}{\sqrt{\bar{V}}} - \frac{z_\alpha s}{\sqrt{n}} \le f(\mu)\right) \ge 1 - \alpha - o(1). \tag{A.1.10}$$

We will utilize the central limit theorem (CLT) and the delta method to prove the above result. Now we consider four different cases.

(I) $\mathrm{Var}\left(Yh(W)\right) = 0$ and $\mathrm{Var}\left(\mathrm{Var}\left(h(W) \mid Z\right)\right) = 0$.

(II) $\mathrm{Var}\left(Yh(W)\right) > 0$ and $\mathrm{Var}\left(\mathrm{Var}\left(h(W) \mid Z\right)\right) = 0$.

(III) $\mathrm{Var}\left(Yh(W)\right) = 0$ and $\mathrm{Var}\left(\mathrm{Var}\left(h(W) \mid Z\right)\right) > 0$.

(IV) $\mathrm{Var}\left(Yh(W)\right) > 0$ and $\mathrm{Var}\left(\mathrm{Var}\left(h(W) \mid Z\right)\right) > 0$.

Note that assuming $\mathbb{E}\left[Y^4\right]$ and $\mathbb{E}\left[h^4(W)\right] < \infty$ ensures all the above variances exist; the bounding strategy is the same as (A.1.9), thus we omit the proof. For the first case where $\mathrm{Var}\left(Yh(W)\right) =$

0 and $\mathrm{Var}\left(\mathrm{Var}\left(h(W) \mid Z\right)\right) = 0$, we have the following facts.

$$\mathrm{Var}\left(Yh(W)\right) = 0 \;\Rightarrow\; R_i = \mathbb{E}\left[Yh(W)\right], \forall\, i \in [n], \bar{R} = \mathbb{E}\left[Yh(W)\right],\; \widehat{\Sigma}_{11} = \widehat{\Sigma}_{12} = 0, \tag{A.1.11}$$

$$\mathrm{Var}\left(\mathrm{Var}\left(h(W) \mid Z\right)\right) = 0 \;\Rightarrow\; V_i = \mathbb{E}\left[h^2(W)\right], \forall\, i \in [n], \bar{V} = \mathbb{E}\left[h^2(W)\right],\; \widehat{\Sigma}_{22} = \widehat{\Sigma}_{12} = 0. \tag{A.1.12}$$

Case (I): due to (A.1.11) and (A.1.12), we simply have $\frac{\bar{R}}{\sqrt{\bar{V}}} = \mathbb{E}\left[Yh(W)\right] / \sqrt{\mathbb{E}\left[h^2(W)\right]} = f(\mu)$ and $s = 0$, thus (A.1.10) holds.

Case (II): due to (A.1.12), $s^2 = \widehat{\Sigma}_{11}/\bar{V} = \widehat{\Sigma}_{11}/\mathbb{E}\left[h^2(W)\right]$, hence we have the following equivalence

$$\left\{\frac{\bar{R}}{\sqrt{\bar{V}}} - \frac{z_\alpha s}{\sqrt{n}} \le f(\mu)\right\} = \left\{\bar{R} - \frac{z_\alpha(\widehat{\Sigma}_{11})^{1/2}}{\sqrt{n}} \le \mathbb{E}\left[Yh(W)\right]\right\}.$$

Thus the problem is reduced to showing that

$$\mathbb{P}\left(\bar{R} - \frac{z_\alpha(\widehat{\Sigma}_{11})^{1/2}}{\sqrt{n}} \le \mathbb{E}\left[Yh(W)\right]\right) \ge 1 - \alpha - o(1). \tag{A.1.13}$$

Notice $\bar{R}$ is simply the sample mean estimator of the quantity $\mathbb{E}\left[Yh(W)\right]$ and $\widehat{\Sigma}_{11}$ is the corresponding sample variance. (A.1.13) is an immediate result of the central limit theorem and Slutsky's theorem.

Case (III): due to (A.1.11), we have

$$\frac{\bar{R}}{\sqrt{\bar{V}}} - \frac{z_\alpha s}{\sqrt{n}} = \frac{\mathbb{E}\left[Yh(W)\right]}{\sqrt{\bar{V}}} - \frac{z_\alpha s}{\sqrt{n}}, \text{ where } s^2 = \frac{1}{\bar{V}}\left(\frac{\mathbb{E}\left[Yh(W)\right]}{2\bar{V}}\right)^2 \widehat{\Sigma}_{22}.$$

$\frac{\mathbb{E}[Yh(W)]}{\sqrt{\bar{V}}}$ is a nonlinear function of the moment estimators. We will use the delta method to estab-

lish the asymptotic normality result. In case (IV), $\mathbb{E}[Yh(W)]$ is further replaced by its moment estimator, and we are dealing with a bit more complicated nonlinear statistic than $1/\sqrt{\bar{V}}$. Hence we focus on case (IV) and omit the very similar proof for case (III).

Case (IV): since $\mathrm{Var}(Yh(W)) > 0$ and $\mathrm{Var}(\mathrm{Var}(h(W) \mid Z)) > 0$, we have as $n \to \infty$,

$$\sqrt{n} \begin{pmatrix} \bar{R} - \mathbb{E}[Yh(W)] \\ \bar{V} - \mathbb{E}\left[h^2(W)\right] \end{pmatrix} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma) \tag{A.1.14}$$

by the multivariate central limit theorem, where the covariance matrix of the random vector $(R_i, V_i) \in \mathbb{R}^2$ is denoted by Σ with

$$\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} = \begin{pmatrix} \mathrm{Var}(Yh(W)) & \mathrm{Cov}(Yh(W), \mathrm{Var}(h(W) \mid Z)) \\ \mathrm{Cov}(Yh(W), \mathrm{Var}(h(W) \mid Z)) & \mathrm{Var}(\mathrm{Var}(h(W) \mid Z)) \end{pmatrix}.$$

$\mathbb{E}\left[Y^4\right], \mathbb{E}\left[h^4(W)\right] < \infty$ ensures the finiteness of $\Sigma_{11}, \Sigma_{12}, \Sigma_{22}$. Denote

$$\widetilde{\sigma}_0^2 = \frac{1}{\mathbb{E}[h^2(W)]} \left[\left(\frac{\mathbb{E}[Yh(W)]}{2\mathbb{E}[h^2(W)]} \right)^2 \Sigma_{22} + \Sigma_{11} - \frac{\mathbb{E}[Yh(W)]}{\mathbb{E}[h^2(W)]} \Sigma_{12} \right], \tag{A.1.15}$$

and we will show $\widetilde{\sigma}_0 > 0$ over the course of derivations from (A.1.20) to the end of the proof. Now consider

$$\left(\frac{\bar{R}}{\sqrt{\bar{V}}} - f(\mu) \right) / s = \left(\frac{\bar{R}}{\sqrt{\bar{V}}} - f(\mu) \right) / \widetilde{\sigma}_0 \cdot \frac{\widetilde{\sigma}_0}{s} := \frac{H(\bar{R}, \bar{V}) - f(\mu)}{\widetilde{\sigma_0}} \cdot \left(\frac{s}{\widetilde{\sigma_0}} \right)^{-1}, \tag{A.1.16}$$

where $H(x_1, x_2) : \mathbb{R}^2 \to \mathbb{R}$ is defined as $H(x_1, x_2) = x_1/\sqrt{x_2}$ for $x_2 > 0$ and its gradient equals $\nabla H(x_1, x_2) = (\frac{\partial H}{\partial x_1}, \frac{\partial H}{\partial x_2}) = \frac{1}{\sqrt{x_2}}(1, -\frac{x_1}{2x_2})$. Let $\theta = (\mathbb{E}[Yh(W)], \mathbb{E}\left[h^2(W)\right])$, then

$$\nabla H(\theta) = \frac{1}{\sqrt{\mathbb{E}[h^2(W)]}} \left(1, -\frac{\mathbb{E}[Yh(W)]}{2\mathbb{E}[h^2(W)]} \right), \tag{A.1.17}$$

and we obtain

$$\operatorname{Var}\left(\nabla H(\theta)^{\top}\begin{pmatrix}\sqrt{n}\bar{R}\\ \sqrt{n}\bar{V}\end{pmatrix}\right)=\operatorname{Var}\left(\nabla H(\theta)^{\top}\begin{pmatrix}R_i\\ V_i\end{pmatrix}\right)=\nabla H(\theta)^{\top}\,\Sigma\,\nabla H(\theta)=\widetilde{\sigma}_0^2 \tag{A.1.18}$$

where the second equality holds by the definition of Σ and the last equality holds by elementary calculation. Therefore, by applying the multivariate delta method to (A.1.14), we have $\sqrt{n}(H(\bar{R},\bar{V})-H(\theta)) \xrightarrow{d} \mathcal{N}\left(0,\nabla H(\theta)^{\top}\,\Sigma\,\nabla H(\theta)\right)$, i.e.,

$$\sqrt{n}(H(\bar{R},\bar{V})-f(\mu))/\widetilde{\sigma}_0 \xrightarrow{d} \mathcal{N}(0,1)\,. \tag{A.1.19}$$

Replacing the means, variances and covariances in $\widetilde{\sigma}_0^2$ by their moment estimators, we obtain

$$\frac{1}{\bar{V}}\left[\left(\frac{\bar{R}}{2\bar{V}}\right)^2\widehat{\Sigma}_{22}+\widehat{\Sigma}_{11}-\frac{\bar{R}}{\bar{V}}\widehat{\Sigma}_{12}\right],$$

which equals s^2 by its definition. Due to the finiteness of $\mathbb{E}\left[Yh(W)\right], \mathbb{E}\left[h^2(W)\right], \Sigma_{11}, \Sigma_{12}, \Sigma_{22}$, we have

$$(\bar{R},\bar{V},\widehat{\Sigma}_{11},\widehat{\Sigma}_{12},\widehat{\Sigma}_{22}) \xrightarrow{p} (\mathbb{E}\left[Yh(W)\right],\mathbb{E}\left[h^2(W)\right],\Sigma_{11},\Sigma_{12},\Sigma_{22})$$

by the law of large numbers. Then by the continuous mapping theorem, we have $s \xrightarrow{p} \widetilde{\sigma}_0$ as $n \to \infty$. Combining this with (A.1.16) and (A.1.19), we have

$$\sqrt{n}\left(\frac{\bar{R}}{\sqrt{\bar{V}}}-f(\mu)\right)/s \xrightarrow{d} \mathcal{N}(0,1)\,,$$

as $n \to \infty$, which establishes (A.1.10).

Now we will verify the positiveness of $\widetilde{\sigma}_0$. Recall $\mathbb{E}\left[h^2(W)\right] = 1$ as assumed without loss of

generality; we rewrite $\widetilde{\sigma}_0^2$

$$\begin{aligned}
\widetilde{\sigma}_0^2 &= \operatorname{Var}\left(\nabla H(\theta)^\top \begin{pmatrix} R_i - \mathbb{E}\left[Yh(W)\right] \\ V_i - \mathbb{E}\left[h^2(W)\right] \end{pmatrix}\right) && \text{(A.1.20)} \\
&= \operatorname{Var}\left(\left(1, -\frac{\mathbb{E}\left[Yh(W)\right]}{2}\right)^\top \begin{pmatrix} R_i - \mathbb{E}\left[Yh(W)\right] \\ V_i - \mathbb{E}\left[h^2(W)\right] \end{pmatrix}\right) \\
&= \mathbb{E}\left[(R_i - \mathbb{E}\left[Yh(W)\right] - 0.5\,\mathbb{E}\left[Yh(W)\right](V_i - 1))^2\right] && \text{(A.1.21)} \\
&:= \mathbb{E}\left[(A+B)^2\right]
\end{aligned}$$

where the first equality holds due to (A.1.18) and the basic property of variance, the second equality holds due to (A.1.17), and the last equality is by rearranging and the terms A, B are defined as below:

$$A := R_i - \mathbb{E}\left[Y_i h(W_i) \mid Z_i\right] = Y_i h(W_i) - \mathbb{E}\left[Y_i h(W_i) \mid Z_i\right], \tag{A.1.22}$$

$$B := \mathbb{E}\left[Y_i h(W_i) \mid Z_i\right] - \mathbb{E}\left[Yh(W)\right] - 0.5\,\mathbb{E}\left[Yh(W)\right]\left(\operatorname{Var}\left(h(W_i) \mid Z_i\right) - 1\right). \tag{A.1.23}$$

Now we can expand (A.1.20) as

$$\begin{aligned}
\widetilde{\sigma}_0^2 = \mathbb{E}\left[(A+B)^2\right] &= \mathbb{E}\left[\mathbb{E}\left[(A+B)^2 \mid Z_i\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[A^2 \mid Z_i\right] - 2B\,\mathbb{E}\left[A \mid Z_i\right] + B^2\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[A^2 \mid Z_i\right] + B^2\right] \\
&\geq \mathbb{E}\left[\operatorname{Var}\left(Yh(W) \mid Z\right)\right], && \text{(A.1.24)}
\end{aligned}$$

where the first equality comes from the tower property of conditional expectation, the second equality holds since $B \in \mathscr{A}(Z_i)$ and the third equality holds due to $\mathbb{E}\left[A \mid Z_i\right] = 0$.

Since (A.1.24) gives a lower bound for $\widetilde{\sigma}_0^2$, we are done when $\mathbb{E}\left[\mathrm{Var}\left(Yh(W) \mid Z\right)\right] > 0$. Otherwise, we assume $\mathbb{E}\left[\mathrm{Var}\left(Yh(W) \mid Z\right)\right] = 0$, then $\widetilde{\sigma}_0^2 = \nabla H(\theta)^\top \Sigma \nabla H(\theta) = 0$ implies the degeneracy of Σ since the vector $\nabla H(\theta) = (1, -0.5\,\mathbb{E}\left[Yh(W)\right])$ is nonzero. It suffices to show it is impossible to have Σ degenerate when $\mathbb{E}\left[\mathrm{Var}\left(Yh(W) \mid Z\right)\right] = 0$. According to the definition of Σ, we have that $Yh(W)$ is a linear function of $\mathrm{Var}\left(h(W) \mid Z\right)$ in the degenerate case. This means $Yh(W) = c\mathrm{Var}\left(h(W) \mid Z\right) + d$ for some constants c, d. Then we obtain

$$\mathrm{Var}\left(Yh(W) \mid Z\right) = \mathrm{Var}\left(c\mathrm{Var}\left(h(W) \mid Z\right) + d \mid Z\right) = c^2\mathrm{Var}\left(\mathrm{Var}\left(h(W) \mid Z\right)\right) > 0,$$

since we are dealing with case (IV) where $\mathrm{Var}\left(\mathrm{Var}\left(h(W) \mid Z\right)\right) > 0$ and $\mathrm{Var}\left(Yh(W)\right) > 0$ (thus $c^2 > 0$). The above result contradicts the assumption $\mathbb{E}\left[\mathrm{Var}\left(Yh(W) \mid Z\right)\right] = 0$. This finishes showing the positiveness of $\widetilde{\sigma}_0$, □

Lemma 2.3

Proof of Lemma 2.3. Recall the notations $g(z) = \mathbb{E}\left[\mu(X, Z) \mid Z = z\right]$ and $h(w) = h(x, z) = \mu(x, z) - g(z)$ introduced in (A.1.1) and (A.1.2). When $Q_x = P_{X|Z}$, we immediately have $\mu(X, Z) - \mathbb{E}_{Q_x}\left[\mu(X, Z) \mid Z\right] = \mu(X, Z) - \mathbb{E}\left[\mu(X, Z) \mid Z\right] = h(W)$, thus

$$f_{Q_y, Q_x}(\mu) = \frac{\mathbb{E}\left[(Y - \mathbb{E}_{Q_y}\left[Y \mid Z\right])h(W)\right]}{\sqrt{\mathbb{E}\left[h^2(W)\right]}} = \frac{\mathbb{E}\left[Yh(W)\right]}{\sqrt{\mathbb{E}\left[h^2(W)\right]}} = f(\mu)$$

where the second equality holds since $\mathbb{E}\left[\mathbb{E}_{Q_y}\left[Y \mid Z\right] h(W)\right] = 0$ by (A.1.4) and the last equality holds by (A.1.7). Hence $f_{Q_y, P_{X|Z}}(\mu) = f(\mu)$ is proved. For convenience, we also use the following notations throughout this proof: $P_x := P_{X|Z}$, $g_y(Z) := \mathbb{E}_{Q_y}\left[Y \mid Z\right]$, $g_x(Z) :=$

$\mathbb{E}_{Q_x}[\mu(X,Z) \mid Z]$. Thus we rewrite $f_{Q_y,Q_x}(\mu)$ in (1.2.8) as

$$f_{Q_y,Q_x}(\mu) = \frac{\mathbb{E}\left[(Y - g_y(Z))(\mu(X,Z) - g_x(Z))\right]}{\sqrt{\mathbb{E}\left[(\mu(X,Z) - g_x(Z))^2\right]}} \leq \frac{\sqrt{\mathbb{E}\left[(Y - g_y(Z))^2\right]}\sqrt{\mathbb{E}\left[(\mu(X,Z) - g_x(Z))^2\right]}}{\sqrt{\mathbb{E}\left[(\mu(X,Z) - g_x(Z))^2\right]}}, \tag{A.1.25}$$

where the inequality holds by the Cauchy–Schwarz inequality. If $\mathbb{E}\left[(\mu(X,Z) - g_x(Z))^2\right] = 0$, $f_{Q_y,Q_x}(\mu)$ is $0/0 = 0$ by convention and thus $f_{g_y,g_x}(\mu) \leq \mathcal{I} + \Delta$ automatically holds due to the non-negativeness of Δ and $\mathcal{I}$. Otherwise, we notice that

$$\begin{aligned}
\mathbb{E}\left[(\mu(X,Z) - g_x(Z))^2\right] &= \mathbb{E}\left[(\mu(X,Z) - \mathbb{E}\left[\mu(X,Z) \mid Z\right] + \mathbb{E}\left[\mu(X,Z) \mid Z\right] - g_x(Z))^2\right] \\
&= \mathbb{E}\left[(\mu(X,Z) - g(Z))^2\right] + \mathbb{E}\left[(g(Z) - g_x(Z))^2\right] \\
&\geq \mathbb{E}\left[h^2(W)\right], \qquad \text{(A.1.26)}
\end{aligned}$$

where the first equality holds due to rearranging, the second equality holds since

$$\mathbb{E}\left[(\mu(X,Z) - \mathbb{E}\left[\mu(X,Z) \mid Z\right])(\mathbb{E}\left[\mu(X,Z) \mid Z\right] - g_x(Z))\right] = \mathbb{E}\left[h(W)(g(Z) - g_x(Z))\right] = 0$$

by (A.1.4), and the last inequality holds by the definition of $h(w)$ and the non-negativeness of $\mathbb{E}\left[(g(Z) - g_x(Z))^2\right]$. When $\mathbb{E}\left[(Y - g_y(Z))(\mu(X,Z) - g_x(Z))\right] \leq 0$, we note that $f_{Q_y,Q_x}(\mu) \leq 0 \leq \mathcal{I} + \Delta$. Thus it remains to deal with the case where $\mathbb{E}\left[(Y - g_y(Z))(\mu(X,Z) - g_x(Z))\right] > 0$.

Now we expand $f_{Q_y,Q_x}(\mu)$ and bound it as below:

$$
\begin{aligned}
f_{Q_y,Q_x}(\mu) &= \frac{\mathbb{E}\left[(Y - g_y(Z))(\mu(X,Z) - g_x(Z))\right]}{\sqrt{\mathbb{E}\left[(\mu(X,Z) - g_x(Z))^2\right]}} \\
&= \frac{\mathbb{E}\left[(Y - g_y(Z))(\mu(X,Z) - g(Z))\right]}{\sqrt{\mathbb{E}\left[(\mu(X,Z) - g_x(Z))^2\right]}} + \frac{\mathbb{E}\left[(Y - g_y(Z))(g(Z) - g_x(Z))\right]}{\sqrt{\mathbb{E}\left[(\mu(X,Z) - g_x(Z))^2\right]}} \\
&\leq \frac{\mathbb{E}\left[(Y - g_y(Z))(\mu(X,Z) - g(Z))\right]}{\sqrt{\mathbb{E}\left[h^2(W)\right]}} + \frac{\mathbb{E}\left[(Y - g_y(Z))(g(Z) - g_x(Z))\right]}{\sqrt{\mathbb{E}\left[(\mu(X,Z) - g_x(Z))^2\right]}} \\
&= \frac{\mathbb{E}\left[Yh(W))\right]}{\sqrt{\mathbb{E}\left[h^2(W)\right]}} + \frac{\mathbb{E}\left[(\epsilon(Y,W) + h^\star(W) + g^\star(Z) - g_y(Z))(g(Z) - g_x(Z))\right]}{\sqrt{\mathbb{E}\left[(\mu(X,Z) - g_x(Z))^2\right]}} \\
&= f(\mu) + \frac{\mathbb{E}\left[(g^\star(Z) - g_y(Z))(g(Z) - g_x(Z))\right]}{\sqrt{\mathbb{E}\left[(\mu(X,Z) - g_x(Z))^2\right]}}, \\
&\leq \mathcal{I} + \frac{\mathbb{E}\left[|g^\star(Z) - g_y(Z)| \cdot |g(Z) - g_x(Z)|\right]}{\sqrt{\mathbb{E}\left[h^2(W)\right]}} \qquad \text{(A.1.27)}
\end{aligned}
$$

where the first equality comes from (A.1.25), the second equality is by rearranging, the first inequality holds due to $\mathbb{E}\left[(Y - g_y(Z))(\mu(X,Z) - g_x(Z))\right] > 0$ and (A.1.26), the third equality holds since $\mathbb{E}\left[g_y(Z)(\mu(X,Z) - g(Z))\right] = \mathbb{E}\left[g_y(Z)h(W)\right] = 0$ by (A.1.4) and we expand Y as in (A.1.3), the last equality holds by (A.1.4), (A.1.5) and (A.1.7), and the last inequality holds due to Lemma 1.2.2, $\mathbb{E}\left[|g^\star(Z) - g_y(Z)| \cdot |g(Z) - g_x(Z)|\right] > 0$ and (A.1.26). In the following, we bound $\mathbb{E}\left[|g^\star(Z) - g_y(Z)| \cdot |g(Z) - g_x(Z)|\right]$. Since we denote $g_x(z) = \mathbb{E}_{Q_x}\left[\mu(X,Z) \mid Z = z\right]$ with Q_x being the estimate of the true conditional distribution of X given Z (i.e., $P_{X|Z}$, abbrevi-

ated as P_x), we can rewrite $|g(Z) - g_x(Z)|$ then bound it as:

$$\begin{aligned}
|g(Z) - g_x(Z)| &= |\mathbb{E}_{P_x}[\mu(X,Z) \mid Z] - \mathbb{E}_{Q_x}[\mu(X,Z) \mid Z]| \\
&= |\mathbb{E}_{P_x}[h(W) + g(Z) \mid Z] - \mathbb{E}_{Q_x}[h(W) + g(Z) \mid Z]| \\
&= |\mathbb{E}_{P_x}[h(W) \mid Z] - \mathbb{E}_{Q_x}[h(W) \mid Z]| \\
&= \left| \int h(x,Z)(1 - \delta(x,Z)) dP_{X|Z}(x \mid Z) \right| \\
&= |\mathbb{E}_{P_x}[h(W)(1 - \delta(W)) \mid Z]| \leq \sqrt{\mathbb{E}_{P_x}[h^2(W) \mid Z]} \sqrt{\chi^2\left(Q_x \| P_{X|Z}\right)},
\end{aligned} \tag{A.1.28}$$

where the second equality holds due to (A.1.2), the third equality holds since $g(Z) \in \mathscr{A}(Z)$, the fourth equality holds since Q_x is absolutely continuous with respect to $P_{X|Z}$ and we denote $\delta(x,Z) := \frac{dQ_x(x|Z)}{dP_{X|Z}(x|Z)}$ and rewrite the third line in the form of integral, and the last inequality holds by the Cauchy–Schwarz inequality and the definition of the χ^2 divergence. Hence replacing the term $|g(Z) - g_x(Z)|$ in (A.1.27) by its upper bound in (A.1.28) produces the following

$$f_{Q_y,Q_x}(\mu) \leq \mathcal{I} + \frac{\mathbb{E}\left[|g^\star(Z) - g_y(Z)| \sqrt{\mathbb{E}_{P_x}[h^2(W) \mid Z]} \sqrt{\chi^2\left(Q_x \| P_{X|Z}\right)}\right]}{\sqrt{\mathbb{E}[h^2(W)]}}. \tag{A.1.29}$$

Now we will bound $\text{III} := \mathbb{E}\left[|g^\star(Z) - g_y(Z)| \sqrt{\mathbb{E}_{P_x}[h^2(W) \mid Z]} \sqrt{\chi^2\left(Q_x \| P_{X|Z}\right)}\right]$ in three different versions.

Firstly, we apply the Cauchy–Schwarz inequality to $\sqrt{\mathbb{E}_{P_x}[h^2(W) \mid Z]} \sqrt{\chi^2\left(Q_x \| P_{X|Z}\right)}$ and

$|g^\star(Z) - g_y(Z)|$, producing

$$\begin{aligned}
\text{III} &= \mathbb{E}\left[|g^\star(Z) - g_y(Z)|\sqrt{\mathbb{E}_{P_x}\left[h^2(W) \mid Z\right]}\sqrt{\chi^2\left(Q_x \| P_{X|Z}\right)}\right] \\
&\leq \sqrt{\mathbb{E}\left[(g^\star(Z) - g_y(Z))^2\right]}\sqrt{\mathbb{E}\left[\mathbb{E}_{P_x}\left[h^2(W) \mid Z\right]\chi^2\left(Q_x \| P_{X|Z}\right)\right]} \\
&= \sqrt{\mathbb{E}\left[(g^\star(Z) - g_y(Z))^2\right]}\sqrt{\mathbb{E}\left[\mathbb{E}_{P_x}\left[h^2(W)\chi^2\left(Q_x \| P_{X|Z}\right) \mid Z\right]\right]} \\
&= \sqrt{\mathbb{E}\left[(g^\star(Z) - g_y(Z))^2\right]}\sqrt{\mathbb{E}\left[h^2(W)\chi^2\left(Q_x \| P_{X|Z}\right)\right]}, \qquad \text{(A.1.30)}
\end{aligned}$$

where the second equality holds since $\chi^2\left(Q_x \| P_{X|Z}\right) \in \mathscr{A}(Z)$, and the last equality holds due to the notation $P_x = P_{X|Z}$ and the law of total expectation. Noting the definition of III and combining (A.1.29) with (A.1.30) yields

$$f_{Q_y,Q_x}(\mu) \leq \mathcal{I} + \sqrt{\mathbb{E}\left[(g^\star(Z) - g_y(Z))^2\right]}\sqrt{\mathbb{E}\left[\left(\frac{h(W)}{\sqrt{\mathbb{E}\left[h^2(W)\right]}}\right)^2 \chi^2\left(Q_x \| P_{X|Z}\right)\right]}.$$

Recalling the notations:

$$g^\star(Z) = \mathbb{E}\left[Y \mid Z\right], \quad g_y(Z) = \mathbb{E}_{Q_y}\left[Y \mid Z\right], \quad h(W) = \mu(X,Z) - \mathbb{E}\left[\mu(X,Z) \mid Z\right], \text{ (A.1.31)}$$

and $w_\mu(X,Z) = \frac{(\mu(X,Z) - \mathbb{E}[\mu(X,Z) \mid Z])^2}{\mathbb{E}[(\mu(X,Z) - \mathbb{E}[\mu(X,Z) \mid Z])^2]}$, (1.2.9) is thus established.

Secondly, we apply the Cauchy–Schwarz inequality to the term $\sqrt{\chi^2\left(Q_x \| P_{X|Z}\right)}$ and the term

$|g^\star(Z) - g_y(Z)|\sqrt{\mathbb{E}_{P_x}\left[h^2(W) \mid Z\right]}$ in III, producing

$$\begin{aligned}
\text{III} &= \mathbb{E}\left[|g^\star(Z) - g_y(Z)|\sqrt{\mathbb{E}_{P_x}\left[h^2(W) \mid Z\right]}\sqrt{\chi^2\left(Q_x \| P_{X|Z}\right)}\right] \\
&\leq \sqrt{\mathbb{E}\left[\chi^2\left(Q_x \| P_{X|Z}\right)\right]}\sqrt{\mathbb{E}\left[\mathbb{E}_{P_x}\left[h^2(W) \mid Z\right](g^\star(Z) - g_y(Z))^2\right]} \\
&= \sqrt{\mathbb{E}\left[\chi^2\left(Q_x \| P_{X|Z}\right)\right]}\sqrt{\mathbb{E}\left[\mathbb{E}_{P_x}\left[h^2(W)(g^\star(Z) - g_y(Z))^2 \mid Z\right]\right]} \\
&= \sqrt{\mathbb{E}\left[\chi^2\left(Q_x \| P_{X|Z}\right)\right]}\sqrt{\mathbb{E}\left[h^2(W)(g^\star(Z) - g_y(Z))^2\right]}, \qquad \text{(A.1.32)}
\end{aligned}$$

where the second equality holds since $(g^\star(Z) - g_y(Z))^2 \in \mathscr{A}(Z)$, and the last equality holds due to the notation $P_x = P_{X|Z}$ and the law of total expectation. Combining (A.1.29) and (A.1.31) with (A.1.32) and recalling $w_\mu(X, Z) = \frac{(\mu(X,Z) - \mathbb{E}[\mu(X,Z) \mid Z])^2}{\mathbb{E}[(\mu(X,Z) - \mathbb{E}[\mu(X,Z) \mid Z])^2]} = \frac{h^2(W)}{\mathbb{E}[h^2(W)]}$ yields a different bound on $f_{Q_y,Q_x}(\mu)$, namely,

$$\begin{aligned}
&f_{Q_y,Q_x}(\mu) \leq f(\mu) + \Delta', \text{ where} \\
&\Delta' = \sqrt{\mathbb{E}\left[\chi^2\left(Q_x \| P_{X|Z}\right)\right]}\sqrt{\mathbb{E}\left[w_\mu(X, Z)(\mathbb{E}\left[Y \mid Z\right] - \mathbb{E}_{Q_y}\left[Y \mid Z\right])^2\right]}.
\end{aligned} \qquad \text{(A.1.33)}$$

Lastly, we apply the Cauchy–Schwarz inequality to the term $\sqrt{\mathbb{E}_{P_x}\left[h^2(W) \mid Z\right]}$ and the term $|g^\star(Z) - g_y(Z)|\sqrt{\chi^2\left(Q_x \| P_{X|Z}\right)}$ in III, producing

$$\begin{aligned}
\text{III} = &\ \mathbb{E}\left[|g^\star(Z) - g_y(Z)|\sqrt{\mathbb{E}_{P_x}\left[h^2(W) \mid Z\right]}\sqrt{\chi^2\left(Q_x \| P_{X|Z}\right)}\right] \\
\leq &\ \sqrt{\mathbb{E}\left[\mathbb{E}_{P_x}\left[h^2(W) \mid Z\right]\right]}\sqrt{\mathbb{E}\left[\chi^2\left(Q_x \| P_{X|Z}\right)(g^\star(Z) - g_y(Z))^2\right]} \\
= &\ \sqrt{\mathbb{E}\left[h^2(W)\right]}\sqrt{\mathbb{E}\left[\chi^2\left(Q_x \| P_{X|Z}\right)(g^\star(Z) - g_y(Z))^2\right]} \\
= &\ \sqrt{\mathbb{E}\left[h^2(W)\right]}\left(\mathbb{E}\left[(g^\star(Z) - g_y(Z))^4\right]\right)^{1/4}\left(\mathbb{E}\left[\left(\chi^2\left(Q_x \| P_{X|Z}\right)\right)^2\right]\right)^{1/4}, \qquad \text{(A.1.34)}
\end{aligned}$$

where the second equality holds due to the notation $P_x = P_{X|Z}$ and the law of total expectation, and the last inequality holds by applying the Cauchy–Schwarz inequality again. Combining

(A.1.29) and (A.1.31) with (A.1.34) yields a final different bound on $f_{Q_y,Q_x}(\mu)$, namely,

$$\begin{aligned} &f_{Q_y,Q_x}(\mu) \leq f(\mu) + \Delta'', \text{ where} \\ \Delta'' &= \left(\mathbb{E}\left[\left(\mathbb{E}\left[Y \mid Z\right] - \mathbb{E}_{Q_y}\left[Y \mid Z\right]\right)^4\right]\right)^{1/4} \left(\mathbb{E}\left[\left(\chi^2\left(Q_x \| P_{X|Z}\right)\right)^2\right]\right)^{1/4}. \end{aligned} \tag{A.1.35}$$

□

A.1.2 Proofs in Section 1.2.3

Proof of Theorem 1.2.4. We prove by contradiction. Suppose there exists an upper confidence bound procedure ensuring asymptotic coverage such that (1.2.10) holds, that is, there exists a joint law over (Y, X, Z), denoted by $F_\infty \in \mathcal{F}$ such that

$$\limsup_{n\to\infty} \mathbb{P}_\infty\left(U(D_n) - \mathcal{I}^2_{F_\infty} < \mathbb{E}_\infty\left[\mathrm{Var}_\infty\left(Y \mid X, Z\right)\right]\right) > \alpha. \tag{A.1.36}$$

where $\mathbb{P}_\infty,\ \mathbb{E}_\infty,\ \mathrm{Var}_\infty$ denote that the data generating distribution for i.i.d. sample D_n is F_∞. Note that $\mathbb{P}_\infty\left(U(D_n) - \mathcal{I}^2_{F_\infty} < \mathbb{E}_\infty\left[\mathrm{Var}_\infty\left(Y \mid X, Z\right)\right]\right) = \mathbb{P}_\infty\left(U(D_n) < \mathbb{E}_\infty\left[\mathrm{Var}_\infty\left(Y \mid Z\right)\right]\right)$ by the definition of $\mathcal{I}^2_{F_\infty}$. Let $\lambda_1 = \mathbb{E}_\infty\left[\mathrm{Var}_\infty\left(Y \mid Z\right)\right]$. When $\lambda_1 = 0$, we have $\mathbb{E}_\infty\left[\mathrm{Var}_\infty\left(Y \mid Z\right)\right] = \mathbb{E}_\infty\left[\mathrm{Var}_\infty\left(Y \mid X, Z\right)\right] = \mathcal{I}^2_{F_\infty} = 0$ and immediately show

$$\limsup_{n\to\infty} \mathbb{P}_\infty\left(U(D_n) - \mathcal{I}^2_{F_\infty} < \mathbb{E}_\infty\left[\mathrm{Var}_\infty\left(Y \mid X, Z\right)\right]\right) = \limsup_{n\to\infty} \mathbb{P}_\infty\left(U(D_n) < \mathcal{I}^2_{F_\infty}\right) \leq \alpha,$$

which contradicts (A.1.36). In the following we consider the case where $\lambda_1 > 0$. Now we construct a sequence of joint laws over (Y, X, Z), denoted by $\{F_k\}_{k=1}^\infty$, $F_k \in \mathcal{F}$, such that the conditional

distribution of $\epsilon \mid X, Z$ is the same as that under F_∞, where $\epsilon = Y - \mathbb{E}[Y \mid X, Z]$, that is,

$$\mathbb{P}_k(\epsilon \mid X, Z) = \mathbb{P}_\infty(\epsilon \mid X, Z), \qquad \forall\, k \geq 1 \tag{A.1.37}$$

and there exist Borel sets $A_k \in \mathbb{R}^{p-1}$ satisfying the following:

(a) $\mathbb{P}_k(Z \in A_k) = 1/k$;

(b) $\mathbb{P}_k(Y \mid X, Z) = \mathbb{P}_\infty(Y \mid X, Z)$ when $Z \notin A_k$;

(c) $\mathbb{E}_k[\mu_k^\star(X, Z) \mid Z] = \mathbb{E}_\infty[\mu_\infty^\star(X, Z) \mid Z]$ when $Z \in A_k$;

(d) $\mathrm{Var}_k(\mu_k^\star(X, Z) \mid Z) = \mathrm{Var}_\infty(\mu_\infty^\star(X, Z) \mid Z) + k\left(2\lambda_1 - \mathcal{I}_{F_\infty}^2\right)$ when $Z \in A_k$;

where $\mathbb{P}_k, \mathbb{E}_k, \mathrm{Var}_k$ denote that the data generating distribution for i.i.d. sample D_n is F_k, and $\mu_k^\star(X, Z) := \mathbb{E}_k[Y \mid X, Z], \mu_\infty^\star(X, Z) := \mathbb{E}_\infty[Y \mid X, Z]$. According to the statement of Theorem 1.2.4, the covariate distribution $P_{X,Z}$ is continuous and fixed. Therefore we have (a) is possible and immediately know

$$\mathbb{P}_k(X, Z) = \mathbb{P}_\infty(X, Z), \qquad \forall\, k \geq 1. \tag{A.1.38}$$

Note here $\mathbb{E}_k[\cdot \mid Z], \mathrm{Var}_k(\cdot \mid Z)$ are the same as $\mathbb{E}_\infty[\cdot \mid Z], \mathrm{Var}_\infty(\cdot \mid Z)$ due to (A.1.38). Hence we can calculate $\mathcal{I}_{F_k}$ through the following

$$\begin{aligned}
\mathcal{I}_{F_k}^2 - \mathcal{I}_{F_\infty}^2 &= \mathbb{E}_\infty\left[\mathbb{1}_{\{A_k\}}\left(\mathrm{Var}_\infty(\mu_k^\star(X, Z) \mid Z) - \mathrm{Var}_\infty(\mu_\infty^\star(X, Z) \mid Z)\right)\right] \\
&= \mathbb{E}_\infty\left[\mathbb{1}_{\{A_k\}} k\left(2\lambda_1 - \mathcal{I}_{F_\infty}^2\right)\right] \\
&= 2\lambda_1 - \mathcal{I}_{F_\infty}^2 =: \lambda_2,
\end{aligned} \tag{A.1.39}$$

where the first equality comes from the definition of $\mathcal{I}^2_F$, (A.1.38) and (b), the second equality holds due to (d) and the third equality holds due to (a). Therefore $\mathcal{I}^2_{F_k} = 2\lambda_1$. We should also check whether F_k belongs to $\mathcal{F}$. Indeed, we consider the following

$$\begin{aligned}
\mathrm{Var}_k(Y) &= \mathbb{E}_k[\mathrm{Var}_k(Y \mid X, Z)] + \mathrm{Var}_k(\mathbb{E}_k[Y \mid X, Z]) \\
&= \mathbb{E}_k[\mathrm{Var}_k(\epsilon \mid X, Z)] + \mathrm{Var}_k(\mathbb{E}_k[Y \mid Z]) + \mathcal{I}^2_{F_k} \\
&= \mathbb{E}_\infty[\mathrm{Var}_k(\epsilon \mid X, Z)] + \mathrm{Var}_\infty(\mathbb{E}_k[Y \mid Z]) + \mathcal{I}^2_{F_k} \\
&= \mathbb{E}_\infty[\mathrm{Var}_\infty(\epsilon \mid X, Z)] + \mathrm{Var}_\infty(\mathbb{E}_\infty[Y \mid Z]) + \mathcal{I}^2_{F_k} \\
&= \mathbb{E}_\infty[\mathrm{Var}_\infty(\epsilon \mid X, Z)] + \mathrm{Var}_\infty(\mathbb{E}_\infty[Y \mid Z]) + \mathcal{I}^2_{F_\infty} + \lambda_2 \\
&= \mathrm{Var}_\infty(Y) + \lambda_2 < \infty,
\end{aligned}$$

where the first equality comes from the law of total variance, the second equality holds as a result of the decomposition $Y = \mu^\star(X, Z) + \epsilon$ and the equivalent expression of the mMSE gap (1.2.2), the third equality holds due to (A.1.38), the fourth equality holds due to (A.1.37), (b) and (c), the fifth equality comes from (A.1.39). Thus we verify $F_k \in \mathcal{F}$, $\forall\, k \geq 1$. As the upper confidence bound procedure U ensures asymptotic coverage validity and $\mathcal{I}^2_{F_k} = 2\lambda_1$, we have

$$\mathbb{P}_k(U(D_n) \geq 2\lambda_1) \geq 1 - \alpha + o_k(1) \tag{A.1.40}$$

where the subscript in $o_k(1)$ emphasizes that the convergence is with respect to data generating function F_k. Remark we only require for fixed k, $o_k(1) \to 0$ as $n \to \infty$. Also notice the following

$$|\mathbb{P}_\infty(U(D_n) \geq 2\lambda_1) - \mathbb{P}_k(U(D_n) \geq 2\lambda_1)| \leq d_{TV}(F_k, F_\infty) \leq \frac{1}{k}, \quad \forall\, k \geq 1, \tag{A.1.41}$$

where the first inequality comes from the property of total variation distance and the second equality holds as a result of (a), according to the construction of F_k. Combining (A.1.40) and (A.1.41) yields the following

$$\mathbb{P}_\infty \left(U(D_n) \geq 2\lambda_1 \right) \geq 1 - \alpha - 1/k + o_k(1), \quad \forall\, k \geq 1.$$

First let $n \to \infty$ then send k to infinity, we obtain

$$\liminf_{n\to\infty} \mathbb{P}_\infty \left(U(D_n) \geq 2\lambda_1 \right) \geq 1 - \alpha,$$

which contradicts

$$\limsup_{n\to\infty} \mathbb{P}_\infty \left(U(D_n) < \mathbb{E}_\infty \left[\mathrm{Var}_\infty \left(Y \mid Z \right) \right] = \lambda_1 \right) > \alpha.$$

□

A.1.3 Proofs in Section 1.2.4

Proof of Theorem 1.2.5. As in the proof of Theorem 1.2.3, we immediately have coverage validity when $\mu(X, Z) \in \mathscr{A}(Z)$. Otherwise, it suffices to show

$$\mathbb{P}\left(\frac{\bar{R}}{\sqrt{\bar{V}}} - \frac{z_\alpha s}{\sqrt{n}} \leq f(\mu) \right) \geq 1 - \alpha - o(1). \tag{A.1.42}$$

for any given $K > 1$, where the sample mean $(\bar{R}, \bar{V})$ and sample covariance matrix $\widehat{\Sigma}$ are defined the same way as in Algorithm 1 except that R_i, V_i are replaced by their Monte Carlo estimators

R_i^K, V_i^K as defined below.

$$
\begin{aligned}
R_i^K &:= Y_i \left(\mu(X_i, Z_i) - \frac{1}{K} \sum_{k=1}^{K} \mu(X_i^{(k)}, Z_i) \right), \\
V_i^K &:= \frac{1}{K-1} \sum_{k=1}^{K} \left(\mu(X_i^{(k)}, Z_i) - \frac{1}{K} \sum_{k=1}^{K} \mu(X_i^{(k)}, Z_i) \right)^2,
\end{aligned}
\tag{A.1.43}
$$

for any fixed $K > 1$.

First we verify

$$
\mathbb{E}\left[R_i^K\right] = \mathbb{E}\left[Yh(W)\right], \quad \mathbb{E}\left[V_i^K\right] = \mathbb{E}\left[h^2(W)\right]. \tag{A.1.44}
$$

By the construction of the null samples, $X_i^{(k)}$ satisfy the following properties:

$$
\{X_i^{(k)}\}_{k=1}^K \perp\!\!\!\perp (X_i, Y_i) \mid Z_i, \tag{A.1.45}
$$

$$
\{X_i^{(k)}\}_{k=1}^K \mid Z_i \overset{i.i.d.}{\sim} X_i \mid Z_i, \tag{A.1.46}
$$

thus we have

$$
\mathbb{E}\left[\frac{1}{K} \sum_{k=1}^{K} \mu(\widetilde{X}_i^{(k)}, Z_i) \,\middle|\, Z_i \right] = \mathbb{E}\left[\mu(X_i, Z_i) \mid Z_i\right], \tag{A.1.47}
$$

$$
\mathbb{E}\left[\frac{1}{K-1} \sum_{k=1}^{K} \left(\mu(\widetilde{X}_i^{(k)}, Z_i) - \frac{1}{K} \sum_{k=1}^{K} \mu(\widetilde{X}_i^{(k)}, Z_i) \right)^2 \,\middle|\, Z_i \right] = \mathrm{Var}\left(\mu(X_i, Z_i) \mid Z_i\right), \tag{A.1.48}
$$

and further obtain

$$\begin{aligned}
\mathbb{E}\left[R_i^K\right] &= \mathbb{E}\left[Y_i\left(\mu(X_i,Z_i)-\frac{1}{K}\sum_{k=1}^{K}\mu(\widetilde{X}_i^{(k)},Z_i)\right)\right] \\
&= \mathbb{E}\left[Y_i\mu(W_i)\right]-\mathbb{E}\left[\mathbb{E}\left[Y_i\mid Z_i\right]\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\mu(\widetilde{X}_i^{(k)},Z_i)\,\middle|\, Z_i\right]\right] \\
&= \mathbb{E}\left[Y_i\mu(W_i)\right]-\mathbb{E}\left[\mathbb{E}\left[Y_i\mid Z_i\right]\mathbb{E}\left[\mu(X_i,Z_i)\mid Z_i\right]\right] \\
&= \mathbb{E}\left[Y_i\mu(W_i)\right]-\mathbb{E}\left[Y_i\mathbb{E}\left[\mu(X_i,Z_i)\mid Z_i\right]\right]=\mathbb{E}\left[Yh(W)\right],
\end{aligned}$$

where the first equality holds due to (A.1.43), the second equality holds due to (A.1.45), the third equality holds due to (A.1.47), the fourth equality comes from the tower property of total expectation and the last one is by the definition of $h(W)$. Regarding the term $\mathbb{E}\left[V_i^K\right]$, (A.1.48) and (A.1.5) immediately imply $\mathbb{E}\left[V_i^K\right]=\mathbb{E}\left[h^2(W)\right]$.

To prove (A.1.42), we can follow a similar strategy as in the proof of Theorem 1.2.3. Note Appendix A.1.1 considers 4 different cases then deals with them separately. Essentially we can conduct similar analysis, but to avoid lengthy proof, we focus on the most complicated case where $\mathrm{Var}\left(Yh(W)\right)>0$ and $\mathrm{Var}\left(\mathrm{Var}\left(h(X)\mid Z\right)\right)>0$ and omit the derivations for the other three cases. Under the moment conditions $\mathbb{E}\left[Y^4\right],\mathbb{E}\left[h^4(W)\right]<\infty$, we have $\mathbb{E}\left[R_i^K\right]=\mathbb{E}\left[Yh(W)\right]<\infty$ and $\mathbb{E}\left[V_i^K\right]=\mathbb{E}\left[h^2(W)\right]<\infty$.

By applying the multivariate central limit theorem and the delta method, we obtain the following asymptotic normality result as in the proof of Theorem 1.2.3: as $n\to\infty$,

$$\sqrt{n}\left(\frac{\frac{1}{n}\sum_{i=1}^{n}R_i^K}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}V_i^K}}-f(\mu)\right)\xrightarrow{d}\mathcal{N}\left(0,\widetilde{\sigma}_0^2\right), \tag{A.1.49}$$

where $\widetilde{\sigma}_0^2$ is similarly defined as in (A.1.15) and its positiveness will be proved over the course of derivations from (A.1.59) toward the end of this proof. Due to the law of large numbers and the

continuous mapping theorem, we can prove $s \xrightarrow{p} \widetilde{\sigma}_0$ as in Appendix A.1.1. The asymptotic normality and the consistency result only require us to verify the finiteness of $\Sigma_{11} = \mathrm{Var}\left(R_i^K\right), \Sigma_{12} = \mathrm{Cov}\left(R_i^K, V_i^K\right), \Sigma_{22} = \mathrm{Var}\left(V_i^K\right)$. Since $\mathbb{E}\left[R_i^K\right], \mathbb{E}\left[V_i^K\right] < \infty$ under the stated moment conditions and $\mathrm{Cov}\left(R_i^K, V_i^K\right) \leq \sqrt{\mathrm{Var}\left(R_i^K\right)\mathrm{Var}\left(V_i^K\right)}$ by the Cauchy–Schwarz inequality, it suffices to prove

$$\mathbb{E}\left[|R_i^K|^2\right] < \infty, \mathbb{E}\left[|V_i^K|^2\right] < \infty. \tag{A.1.50}$$

Denote $\bar{h}_i^K = \frac{1}{K}\sum_{k=1}^K h(\widetilde{X}_i^{(k)}, Z_i)$ and we rewrite R_i^K and V_i^K.

$$\begin{aligned}
R_i^K &= Y_i\left(\mu(X_i, Z_i) - \frac{1}{K}\sum_{k=1}^K \mu(\widetilde{X}_i^{(k)}, Z_i)\right) \\
&= Y_i\left(\mu(X_i, Z_i) - \mathbb{E}\left[\mu(X_i, Z_i) \mid Z_i\right] - \frac{1}{K}\sum_{k=1}^K (\mu(\widetilde{X}_i^{(k)}, Z_i) - \mathbb{E}\left[\mu(\widetilde{X}_i^{(k)}, Z_i) \mid Z_i\right])\right) \\
&= Y_i\left(h(X_i, Z_i) - \frac{1}{K}\sum_{k=1}^K h(\widetilde{X}_i^{(k)}, Z_i)\right) \qquad \text{(A.1.51)} \\
&= Y_i(h(X_i, Z_i) - \bar{h}_i^K) \qquad \text{(A.1.52)}
\end{aligned}$$

where the first equality holds by (A.1.43), the second equality holds by (A.1.47) and the third equality holds by the definition of $h(w)$.

$$\begin{aligned}
V_i^K &= \frac{1}{K-1}\sum_{k=1}^K \left(\mu(X_i^{(k)}, Z_i) - \frac{1}{K}\sum_{k=1}^K \mu(X_i^{(k)}, Z_i)\right)^2 \\
&= \frac{1}{K-1}\sum_{k=1}^K \left(h(X_i^{(k)}, Z_i) - \frac{1}{K}\sum_{k=1}^K h(X_i^{(k)}, Z_i)\right)^2 \\
&= \frac{1}{K-1}\sum_{k=1}^K h^2(X_i^{(k)}, Z_i) - \frac{K}{K-1}\left(\frac{1}{K}\sum_{k=1}^K h(X_i^{(k)}, Z_i)\right)^2 \qquad \text{(A.1.53)} \\
&= \frac{K}{K-1}\left(\frac{1}{K}\sum_{k=1}^K h^2(X_i^{(k)}, Z_i) - (\bar{h}_i^K)^2\right) \qquad \text{(A.1.54)}
\end{aligned}$$

where the first equality holds by (A.1.43), the second equality holds due to similar derivations as (A.1.51) and the last two equalities are simply by expanding and rearranging. Now we bound

$$\begin{aligned}\left(\mathbb{E}\left[|R_i^K|^2\right]\right)^2 &= \left(\mathbb{E}\left[Y_i^2(h(X_i,Z_i)-\bar{h}_i^K)^2\right]\right)^2 \\ &\leq \mathbb{E}\left[Y^4\right]\mathbb{E}\left[(h(X_i,Z_i)-\bar{h}_i^K)^4\right] \\ &\leq \mathbb{E}\left[Y^4\right]\cdot 2^{4-1}\left(\mathbb{E}\left[h^4(X_i,Z_i)+\mathbb{E}\left[\left(\bar{h}_i^K\right)^4\right]\right]\right) \end{aligned} \tag{A.1.55}$$

where the first equality holds due to (A.1.52), the first inequality holds by the Cauchy–Schwarz inequality, the second inequality comes from the C_r inequality. Regarding $\mathbb{E}\left[|V_i^K|^2\right]$, we have

$$\begin{aligned}\mathbb{E}\left[|V_i^K|^2\right] &= \mathbb{E}\left[\left|\frac{K}{K-1}\left(\frac{1}{K}\sum_{k=1}^K h^2(X_i^{(k)},Z_i)-(\bar{h}_i^K)^2\right)\right|^2\right] \\ &\leq \frac{2^{2-1}K^2}{(K-1)^2}\mathbb{E}\left[\left(\frac{1}{K}\sum_{k=1}^K h^2(X_i^{(k)},Z_i)\right)^2\right]+\frac{2^{2-1}K^2}{(K-1)^2}\mathbb{E}\left[(\bar{h}_i^K)^4\right] \\ &\leq 2^3\left(\mathbb{E}\left[\left(\frac{1}{K}\sum_{k=1}^K h^2(X_i^{(k)},Z_i)\right)^2\right]+\mathbb{E}\left[(\bar{h}_i^K)^4\right]\right) := 2^3(\mathrm{II}+\mathbb{E}\left[(\bar{h}_i^K)^4\right]),\end{aligned} \tag{A.1.56}$$

where the first equality holds by (A.1.54), the first inequality holds due to the C_r inequality, and the second inequality comes from rearranging and the fact that $K \leq 2(K-1)$ (since $K > 1$). The term II and $\mathbb{E}\left[(\bar{h}_i^K)^4\right]$ can be bounded using the same strategy. Below we give the bounding details

of $\mathbb{E}\left[(\bar{h}_i^K)^4\right]$ and omit that of II. By the tower property of conditional expectation, we have

$$\begin{aligned}\mathbb{E}\left[(\bar{h}_i^K)^4\right] &= \mathbb{E}\left[\left(\frac{1}{K}\sum_{k=1}^{K} h(\widetilde{X}_i^{(k)}, Z_i)\right)^4\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\left(\frac{1}{K}\sum_{k=1}^{K} h(\widetilde{X}_i^{(k)}, Z_i)\right)^4 \middle| Z_i\right]\right]. \end{aligned} \tag{A.1.57}$$

To bound $\mathbb{E}\left[\left(\frac{1}{K}\sum_{k=1}^{K} h(\widetilde{X}_i^{(k)}, Z_i)\right)^4 \middle| Z_i\right]$, we notice that, conditional on Z_i, $\{h(\widetilde{X}_i^{(k)}, Z_i)\}_{k=1}^K$ are i.i.d. mean zero random variables, hence we can apply the extension of the Bahr–Esseen inequality in Dharmadhikari et al. (1969) to obtain

$$\mathbb{E}\left[\left(\sum_{k=1}^{K} h(\widetilde{X}_i^{(k)}, Z_i)\right)^4 \middle| Z_i\right] \leq c_{4,K} \sum_{k=1}^{K} \mathbb{E}\left[h^4(\widetilde{X}_i^{(k)}, Z_i) \middle| Z_i\right], \tag{A.1.58}$$

Note for generic $d \geq 2$ and n, the term $c_{d,n}$ is defined as

$$c_{d,n} = n^{d/2-1}\frac{d(d-1)}{2}\max\{1, 2^{d-3}\}\left[1 + 2d^{-1}D_{2m}^{(d-2)/2m}\right]$$

where the integer m satisfies $2m \leq d < 2m+2$, and

$$D_{2m} = \sum_{t=1}^{m} \frac{t^{2m-1}}{(t-1)!}.$$

We then can simply bound $c_{4,K}$ by $C_4 K$ for some universal constant C_4 which do not depend on K. Therefore, combining (A.1.57) and (A.1.58) gives us

$$\begin{aligned}\mathbb{E}\left[(\bar{h}_i^K)^4\right] &\leq \mathbb{E}\left[\frac{C_4 K}{K^4}\sum_{k=1}^{K}\mathbb{E}\left[h^4(\widetilde{X}_i^{(k)}, Z_i)\middle| Z_i\right]\right] \\ &= \frac{C_4}{K^2}\mathbb{E}\left[\mathbb{E}\left[h^4(X_i, Z_i)\middle| Z_i\right]\right] = \frac{C_4}{K^2}\mathbb{E}\left[h^4(W)\right]\end{aligned}$$

where the equality holds by (A.1.46) and the second equality holds by the tower property of conditional expectation. Since $\mathbb{E}\left[h^4(W)\right] < \infty$, we have $\mathbb{E}\left[(\bar{h}_i^K)^4\right] < \infty$. The finiteness of II is similarly proved. Due to (A.1.55) and (A.1.56), we thus establish (A.1.50) under the stated moment conditions $\mathbb{E}\left[Y^4\right], \mathbb{E}\left[h^4(W)\right] < \infty$. Applying Slutsky's theorem to (A.1.49) and the consistency result that $s \xrightarrow{p} \widetilde{\sigma}_0$, we have

$$\frac{\sqrt{n}}{s}\left(\frac{\frac{1}{n}\sum_{i=1}^n R_i^K}{\sqrt{\frac{1}{n}\sum_{i=1}^n V_i^K}} - f(\mu)\right) \xrightarrow{d} \mathcal{N}(0,1),$$

which establishes (A.1.42).

Now we will verify the positiveness of $\widetilde{\sigma}_0$ as promised. Recall in the proof of Theorem 1.2.3, the variance term in the asymptotic normality result is also denoted as $\widetilde{\sigma}_0^2$ and admits the following expression

$$\mathbb{E}\left[(R_i - \mathbb{E}\left[Yh(W)\right] - 0.5\,\mathbb{E}\left[Yh(W)\right](V_i - 1))^2\right] = \mathbb{E}\left[(A+B)^2\right] > 0 \tag{A.1.59}$$

according to (A.1.21), where A and B are defined in (A.1.22) and (A.1.23) and $\mathbb{E}\left[(A+B)^2\right] > 0$ as proved over the course of derivations from (A.1.21) to the end of the proof of Theorem 1.2.3. In this proof, it is not hard to see $\widetilde{\sigma}_0^2$ has a similar form except that R_i, V_i in the above expression are replaced by their Monte Carlo estimators R_i^K, V_i^K, thus giving

$$\begin{aligned}
\widetilde{\sigma}_0^2 &= \mathbb{E}\left[(R_i^K - \mathbb{E}\left[Yh(W)\right] - 0.5\,\mathbb{E}\left[Yh(W)\right](V_i^K - 1))^2\right] \\
&= \mathbb{E}\left[(Y_i(h(X_i,Z_i) - \bar{h}_i^K) - \mathbb{E}\left[Yh(W)\right] - 0.5\,\mathbb{E}\left[Yh(W)\right](V_i^K - 1))^2\right] \\
&= \mathbb{E}\left[(\mathrm{III}_1 - \mathrm{III}_2)^2\right], \qquad \text{(A.1.60)}
\end{aligned}$$

where the second equality holds by (A.1.52) and rearranging, the terms $\mathrm{III}_1, \mathrm{III}_2$ in the last equality

are defined as:

$$\begin{aligned}
\mathrm{III}_1 &:= Y_i h(W_i) - \mathbb{E}\left[Y h(W)\right] - 0.5\,\mathbb{E}\left[Y h(W)\right]\left(\mathrm{Var}\left(h(W_i) \mid Z_i\right) - 1\right) \\
\mathrm{III}_2 &:= Y_i \bar{h}_i^K + 0.5\,\mathbb{E}\left[Y h(W)\right]\left(V_i^K - \mathrm{Var}\left(h(W_i) \mid Z_i\right)\right).
\end{aligned}$$

To bound $\mathbb{E}\left[\left(\mathrm{III}_1 - \mathrm{III}_2\right)^2\right]$, we will show $\mathbb{E}\left[\mathrm{III}_2 \mid Y_i, W_i\right] = 0$. Recall the definition that $\bar{h}_i^K = \frac{1}{K}\sum_{k=1}^K h(\widetilde{X}_i^{(k)}, Z_i)$, we obtain

$$\mathbb{E}\left[\bar{h}_i^K \mid Y_i, W_i\right] = \mathbb{E}\left[\bar{h}_i^K \mid Z_i\right] = \mathbb{E}\left[h(W_i) \mid Z_i\right] = 0,$$

where the first equality holds due to $W_i = (X_1, Z_i)$ and (A.1.45), the second equality holds by (A.1.46), and the last equality holds due to (A.1.4). Similarly we have

$$\mathbb{E}\left[V_i^K \mid Y_i, W_i\right] = \mathbb{E}\left[V_i^K \mid Z_i\right] = \mathrm{Var}\left(h(W_i) \mid Z_i\right),$$

due to (A.1.43), (A.1.45), and (A.1.47). Thus we have shown

$$\mathbb{E}\left[\mathrm{III}_2 \mid Y_i, W_i\right] = 0. \tag{A.1.61}$$

Applying the tower property of conditional expectation to (A.1.60) then expanding yields the following expression:

$$\begin{aligned}
\widetilde{\sigma}_0^2 &= \mathbb{E}\left[\mathbb{E}\left[\left(\mathrm{III}_1^2 + \mathrm{III}_2^2 - 2\mathrm{III}_1\mathrm{III}_2\right) \mid Y_i, W_i\right]\right] \\
&= \mathbb{E}\left[\mathrm{III}_1^2 + \mathbb{E}\left[\mathrm{III}_2^2 \mid Y_i, W_i\right] - 2\mathrm{III}_1\mathbb{E}\left[\mathrm{III}_2 \mid Y_i, W_i\right]\right] \\
&= \mathbb{E}\left[\mathrm{III}_1^2 + \mathbb{E}\left[\mathrm{III}_2^2 \mid Y_i, W_i\right]\right] \\
&\geq \mathbb{E}\left[\mathrm{III}_1^2\right] = \mathbb{E}\left[(A+B)^2\right],
\end{aligned} \tag{A.1.62}$$

where the second equality holds since $\mathrm{III}_1 \in \mathscr{A}(Y_i, W_i)$, and the third equality comes from (A.1.61). Note in the last line we have $\mathrm{III}_1 = A + B$ due to the definitions of A, B in (A.1.22) and (A.1.23) and $\mathbb{E}\left[(A+B)^2\right] > 0$ due to (A.1.59). Note $\mathbb{E}\left[(A+B)^2\right]$ does not depend on K, therefore we establish the positiveness of $\widetilde{\sigma}_0$ for any $K > 1$. □

A.1.4 Proofs in Section 1.2.5

Proof of Theorem 1.2.6. First we write

$$\mathcal{I} - L_\alpha^n(\mu_n) = \mathcal{I} - f(\mu_n) + f(\mu_n) - L_\alpha^n(\mu_n),$$

where $f(\mu_n)$ is defined as

$$f(\mu_n) =: \frac{\mathbb{E}\left[\mathrm{Cov}(\mu^\star(X,Z), \mu_n(X,Z) \mid Z)\right]}{\sqrt{\mathbb{E}\left[\mathrm{Var}(\mu_n(X,Z) \mid Z)\right]}}.$$

Then it suffices to separately show

$$\mathcal{I} - f(\mu_n) = O_p\left(\inf_{\mu' \in S_{\mu_n}} \mathbb{E}\left[(\mu_n'(X,Z) - \mu^\star(X,Z))^2\right]\right), \tag{A.1.63}$$

$$f(\mu_n) - L_\alpha^n(\mu_n) = O_p\left(n^{-1/2}\right). \tag{A.1.64}$$

In the following, we first show (A.1.64). Recall the definitions in Algorithm 1, when $\mu(X,Z) \in \mathscr{A}(Z)$, we have $f(\mu_n) = L_\alpha^n(\mu_n) = 0$, hence in the following we focus on the case where $\mu(X,Z) \notin \mathscr{A}(Z)$. Note we have

$$L_\alpha^n(\mu_n) \geq \frac{\bar{R}}{\sqrt{V}} - \frac{z_\alpha s}{\sqrt{n}},$$

then since $f(\mu_n) - L_\alpha^n(\mu_n) \leq s\left(\left|\left(\frac{\bar{R}}{\sqrt{V}} - f(\mu_n)\right)/s\right| + \frac{z_\alpha}{\sqrt{n}}\right)$, it suffices to show

$$T := \frac{\bar{R}/\sqrt{V} - f(\mu_n)}{s} = O_p\left(n^{-1/2}\right), \quad s = O_p(1).$$

For given μ_n, showing the above is quite straightforward: in the proof of Theorem 1.2.3, we establish the asymptotic normality of T; we also show s converges in probability to $\widetilde{\sigma}_0$ (which is the variance of the asymptotic normal distribution, as defined in (A.1.15)). For a sequence of working regression functions μ_n, we need more work and the stated uniform moment conditions. The proof proceeds through verifying the following: note that by definition of bounded in probability, $T = O_p\left(n^{-1/2}\right)$ says for any $\epsilon > 0$, there exists M for which

$$\sup_n P(\sqrt{n}|T| > M) \leq \epsilon.$$

The case that $\mu(X, Z) \in \mathscr{A}(Z)$, i.e., $\mathbb{E}\left[\mathrm{Var}\left(\mu_n(X, Z) \mid Z\right)\right] = 0$, was dealt with in the first sentence after (A.1.64). Now it suffices to show for any μ_n in the function class $\mathcal{U} := \{\mu : \mathbb{E}\left[\mu^{12}(X, Z)\right]/(\mathbb{E}\left[\mathrm{Var}\left(\mu(X, Z) \mid Z\right)\right])^6 \leq C\}$,

$$\sup_n \mathbb{P}\left(\sqrt{n}|T| > M\right) \leq \epsilon, \tag{A.1.65}$$

and the choice of M (when fixing ϵ) is uniform over $\mu_n \in \mathcal{U}$. Define the standard Gaussian random variable by G. Then we have

$$\mathbb{P}\left(\sqrt{n}|T| > M\right) \leq \mathbb{P}\left(|G| > M\right) + \Delta, \tag{A.1.66}$$

where Δ is defined as

$$\Delta := \sup_{\mu_n \in \mathcal{U}} \sup_{M>0} \left| \mathbb{P}\left(\sqrt{n}|T| > M\right) - \mathbb{P}\left(|G| > M\right) \right|. \tag{A.1.67}$$

Due to (A.1.9), $\mathbb{E}\left[\mu^{12}(X,Z)\right] < \infty$ implies $\mathbb{E}\left[h^{12}(W)\right] < \infty$, where h is defined in (A.1.2). In the following proof, we will only assume weaker moment conditions, i.e., $\mathbb{E}\left[\mathrm{Var}\left(\mu_n(X,Z) \mid Z\right)\right] = 0$ or $\frac{\mathbb{E}\left[\mu_n^{12}(X,Z)\right]}{\mathbb{E}[\mathrm{Var}(\mu_n(X,Z) \mid Z)]^6} \leq C$ stated in Theorem 1.2.6 is replaced by $\mathbb{E}\left[\mathrm{Var}\left(\mu_n(X,Z) \mid Z\right)\right] = 0$ or $\frac{\mathbb{E}\left[h_n^{12}(X,Z)\right]}{\mathbb{E}[\mathrm{Var}(\mu_n(X,Z) \mid Z)]^6} \leq C$, where h_n is defined accordingly.

In the proof of Theorem A.3.1, we assume $\mathbb{E}\left[h^2(W)\right] = 1$ without loss of generality. This is because we can always scale h by dividing by $\sqrt{\mathbb{E}\left[h^2(W)\right]}$ when the given working regression function satisfies $\mu(X,Z) \notin \mathscr{A}(Z)$. The floodgate inference procedure and results are the same with the corresponding scaled version $\widetilde{h}(W)$. And the scaled version still satisfies the finite moment condition $\mathbb{E}\left[\widetilde{h}^{12}(W)\right] < \infty$. Now we are dealing with a sequence of working regression functions μ_n. If we scale h_n analogously by dividing it by $\sqrt{\mathbb{E}\left[h_n^2(W)\right]}$, the corresponding function sequence $\{\widetilde{h}_n\}$ does not necessarily satisfy the uniform moment condition, i.e., for all n, $\mathbb{E}\left[\widetilde{h}_n^{12}(W)\right] < C$ for some constant C. But the moment conditions $\mathbb{E}[Y^{12}] < \infty$ and $\mathbb{E}\left[h_n^{12}(X,Z)\right] / (\mathbb{E}\left[\mathrm{Var}\left(\mu_n(X,Z) \mid Z\right)\right])^6 = \mathbb{E}\left[h_n^{12}(W)\right] / (\sqrt{\mathbb{E}\left[h_n^2(W)\right]})^{12} \leq C$ for all n ensure the uniform moment bound after scaling, hence for the following we can assume $\mathbb{E}\left[h_n^2(W)\right] = 1$.

According to the proof of Theorem A.3.1, we have the following Berry–Esseen bound

$$\sup_{M>0} \left| \mathbb{P}\left(\sqrt{n}|T| > M\right) - \mathbb{P}\left(|G| > M\right) \right| = O\left(\frac{1}{\sqrt{n}}\right),$$

which relies on verifying the following:

(i) $\mathbb{E}\left[|U_{01}|^3\right], \mathbb{E}\left[|U_{02}|^3\right], \mathbb{E}\left[|U_{03}|^3\right], \mathbb{E}\left[|U_{04}|^3\right], \mathbb{E}\left[|U_{05}|^3\right] < \infty$,

(ii) $\widetilde{\sigma}_0^2(\mu_n) = H_2(\mathbf{0}) > 0$,

(iii) $\widetilde{\sigma}^2(\mu_n) = \|L(U_0)\|_2 > 0$.

Note the above terms are defined similarly as in the proof of Theorem A.3.1 except the dependence on μ_n (but we abbreviate the notation dependence on μ_n for the random variables). We have $\widetilde{\sigma}^2(\mu_n) = 1$ due to the derivations after (A.3.21) in the proof of Theorem A.3.1. To show the constant in the above rate of $\frac{1}{\sqrt{n}}$ is uniformly bounded, we need to prove $\inf_{\mu_n \in \mathcal{U}} \widetilde{\sigma}^2(\mu_n) > 0$ and uniformly control the the 3rd moments in the condition (i). First notice that

$$\begin{aligned}
\inf_{\mu_n \in \mathcal{U}} \widetilde{\sigma}^2(\mu_n) &\geq \inf_{\mu_n \in \mathcal{U}} \mathbb{E}\left[\mathrm{Var}\left(Y h_n(W) \mid Z\right)\right] \\
&\geq \inf_{\mu_n \in \mathcal{U}} \mathbb{E}\left[\mathrm{Var}\left(Y h_n(W) \mid X, Z\right)\right] \\
&= \inf_{\mu_n \in \mathcal{U}} \mathbb{E}\left[h_n^2(W) \mathrm{Var}\left(Y \mid X, Z\right)\right] \\
&\geq \tau > 0
\end{aligned}$$

where the first inequality holds due to (A.1.24), the second inequality holds as a result of the law of total conditional variance, the last equality holds by the assumption that $\mathbb{E}\left[h_n^2(W)\right] = 1$ and the moment lower bound condition $\mathrm{Var}\left(Y|X,Z\right) \geq \tau > 0$. Assuming $\mathbb{E}[Y^{12}] < \infty$ and $\mathbb{E}\left[\mu_n^{12}(X,Z)\right] / (\mathbb{E}\left[\mathrm{Var}\left(\mu_n(X,Z) \mid Z\right)\right])^6 \leq C$, we can uniformly control the moments $\mathbb{E}\left[|U_{01}|^3\right], \mathbb{E}\left[|U_{02}|^3\right], \mathbb{E}\left[|U_{03}|^3\right], \mathbb{E}\left[|U_{04}|^3\right], \mathbb{E}\left[|U_{05}|^3\right]$, therefore establish the rate of $\frac{1}{\sqrt{n}}$ in (A.1.67):

$$\Delta = O\left(\frac{1}{\sqrt{n}}\right).$$

Combining this with (A.1.66), we have

$$\sup_{\mu_n \in \mathcal{U}} \mathbb{P}\left(\sqrt{n}|T| > M\right) \leq \mathbb{P}\left(|G| > M\right) + \frac{C'}{\sqrt{n}}$$

for some constant C' depending on C, τ and $\mathbb{E}\left[Y^{12}\right]$. Therefore we obtain (A.1.65) and the choice of M can be universally chosen over $\mu_n \in \mathcal{U}$, which finally establishes $T = O_p\left(n^{-1/2}\right)$. Using similar strategies, we can prove $s = O_p(1)$. Hence we have shown (A.1.64).

Now we proceed to prove (A.1.63), first it can be simplified into the following form due to (A.1.6) and (A.1.8),

$$\mathcal{I} - f(\mu_n) = \sqrt{\mathbb{E}[(h^\star)^2(W)]} - \frac{\mathbb{E}\left[h_n(W)h^\star(W)\right]}{\sqrt{\mathbb{E}\left[h_n^2(W)\right]}} \tag{A.1.68}$$

where $h_n(W) = \mu_n(W) - \mathbb{E}\left[\mu_n(W) \mid Z\right]$ and $h^\star$ are defined the same way. Remark we have $0/0 = 0$ by convention for (A.1.68). We also find it is more convenient to work with $f(\bar{\mu}_n)$ (note $f(\mu_n) = f(\bar{\mu}_n)$), recall that the definition of $\bar{\mu}_n$:

$$\bar{\mu}_n(x,z) := \sqrt{\frac{\mathcal{I}}{\mathbb{E}\left[h_n^2(W)\right]}}\left(\mu_n(x,z) - \mathbb{E}\left[\mu_n(X,Z) \mid Z = z\right]\right) + \mathbb{E}\left[\mu^\star(X,Z) \mid Z = z\right],$$

and similarly denote $\bar{h}_n(w) = \bar{\mu}_n(x,z) - \mathbb{E}\left[\bar{\mu}_n(X,Z) \mid Z = z\right]$. When $\mu(X,Z) \in \mathscr{A}(Z)$, we have $\bar{\mu}_n(x,z) = \mathbb{E}\left[\mu^\star(X,Z) \mid Z = z\right], \bar{h}_n(w) = 0$, thus

$$\mathcal{I} - f(\mu_n) = \mathcal{I} = \frac{\mathbb{E}\left[(\bar{h}_n(W) - h^\star(W))^2\right]}{\sqrt{\mathbb{E}[(h^\star)^2(W)]}} \tag{A.1.69}$$

Otherwise when $\mathbb{E}\left[h_n^2(W)\right] > 0$, we have $\sqrt{\mathbb{E}\left[\bar{\mu}_n^2(W)\right]} = \mathcal{I}$. In this case, we rewrite the right hand side of (A.1.68) in terms of $\bar{\mu}_n$ and further simplify it as below,

$$\frac{\mathbb{E}\left[(\bar{h}_n(W) - h^\star(W))^2\right] - \left(\sqrt{\mathbb{E}\left[\bar{h}_n^2(W)\right]} - \sqrt{\mathbb{E}[(h^\star)^2(W)]}\right)^2}{2\sqrt{\mathbb{E}\left[\bar{h}_n^2(W)\right]}} = \frac{\mathbb{E}\left[(\bar{h}_n(W) - h^\star(W))^2\right]}{2\sqrt{\mathbb{E}[(h^\star)^2(W)]}}$$

which says that

$$\mathcal{I} - f(\mu_n) = \frac{\mathbb{E}\left[(\bar{h}_n(W) - h^\star(W))^2\right]}{2\sqrt{\mathbb{E}[(h^\star)^2(W)]}} \tag{A.1.70}$$

Note that $\sqrt{\mathbb{E}[(h^\star)^2(W)]} = \mathcal{I}$ which does not depend on μ, hence it suffices to show

$$\mathbb{E}\left[(\bar{h}_n(W) - h^\star(W))^2\right] = O_p\left(\inf_{\mu' \in S_{\mu_n}} \mathbb{E}\left[(\mu'(X,Z) - \mu^\star(X,Z))^2\right]\right). \tag{A.1.71}$$

We prove it by considering two cases:

(a) $\mathbb{E}\left[h_n(W)h^\star(W)\right] \leq 0$,

(b) $\mathbb{E}\left[h_n(W)h^\star(W)\right] > 0$.

Regarding case (a), we have

$$\begin{aligned}
\inf_{\mu' \in S_{\mu_n}} \mathbb{E}\left[(\mu'(X,Z) - \mu^\star(X,Z))^2\right] &= \inf_{c>0, \forall g(z)} \left(\mathbb{E}\left[(ch_n(W) - h^\star(W))^2\right] + \mathbb{E}\left[(g(Z) - \mathbb{E}\left[\mu^\star(W) \mid Z\right])^2\right]\right) \\
&= \inf_{c>0} \mathbb{E}\left[(ch_n(W) - h^\star(W))^2\right] \\
&= \mathbb{E}\left[(h^\star)^2(W)\right] + \inf_{c>0} c^2\mathbb{E}\left[h_n^2(W)\right] - 2c\mathbb{E}\left[h_n(W)h^\star(W)\right] \\
&= \mathbb{E}\left[(h^\star)^2(W)\right]
\end{aligned}$$

where the first equality holds by the definition of S_{μ_n} and the fact that, for any $g(Z)$,

$$\mathbb{E}\left[h^\star(W)g(Z)\right] = \mathbb{E}\left[g(Z)\mathbb{E}\left[h^\star(W) \mid Z\right]\right] = 0$$

and similarly $\mathbb{E}\left[h_n(W)g(Z)\right] = 0$. The second equality holds by choosing $g(z)$ to be $\mathbb{E}\left[h^\star(W) \mid Z = z\right]$. The third equality is simply from expanding and the last equality holds in case (a). Noticing

$$\mathbb{E}\left[(\bar{h}_n(W) - h^\star(W))^2\right] \leq 2\left(\mathbb{E}\left[\bar{h}_n^2(W)\right] + \mathbb{E}\left[(h^\star)^2(W)\right]\right) = 4\mathbb{E}\left[(h^\star)^2(W)\right]$$

we thus establish (A.1.71). Regarding case (b), we have

$$\begin{aligned}
\inf_{\mu' \in S_{\mu_n}} \mathbb{E}\left[(\mu'(X,Z) - \mu^\star(X,Z))^2\right] &= \inf_{c>0} \mathbb{E}\left[(ch_n(W) - h^\star(W))^2\right] \\
&= \inf_{c>0} \mathbb{E}\left[(ch_n(W) - h_0(W) + h_0(W) - h^\star(W))^2\right] \\
&= \mathbb{E}\left[(h_0(W) - h^\star(W))^2\right] + \inf_{c>0} \mathbb{E}\left[(ch_n(W) - h_0(W))^2\right] \\
&= \mathbb{E}\left[(h_0(W) - h^\star(W))^2\right] \\
&= \mathbb{E}\left[(h^\star)^2(W)\right] - \mathbb{E}\left[(h_0(W))^2\right] \qquad \text{(A.1.72)}
\end{aligned}$$

where in the second equality, h_0 is defined to be

$$h_0(w) := \frac{\mathbb{E}\left[h_n(W)h^\star(W)\right]}{\mathbb{E}\left[h_n^2(W)\right]} h_n(w).$$

It satisfies the property $\mathbb{E}\left[h_n(W)\left(h^\star(W) - h_0(W)\right)\right] = 0$ thus the third equality holds. The fourth equality comes from choosing c to be $\frac{\mathbb{E}[h_n(W)h^\star(W)]}{\mathbb{E}[h_n^2(W)]}$, which is positive in case (b). The last equality holds again due to $\mathbb{E}\left[h_n(W)\left(h^\star(W) - h_0(W)\right)\right] = 0$. And we have

$$\begin{aligned}
\mathbb{E}\left[(\bar{h}_n(W) - h^\star(W))^2\right] &= 2\mathbb{E}\left[(h^\star)^2(W)\right] - 2\mathbb{E}\left[\bar{h}_n(W)h^\star(W)\right] \\
&= 2\mathbb{E}\left[(h^\star)^2(W)\right] - 2\mathbb{E}\left[(h_0(W))^2\right]\rho \qquad \text{(A.1.73)}
\end{aligned}$$

where ρ denotes the following term and can be further simplified based on the definition of $\bar{h}_n(W)$ and $h_0(W)$.

$$\begin{aligned}
\rho &:= \frac{\mathbb{E}\left[\bar{h}_n(W)h^\star(W)\right]}{\mathbb{E}\left[(h_0(W))^2\right]} \\
&= \frac{\mathcal{I}\sqrt{\mathbb{E}\left[h_n^2(W)\right]}}{\mathbb{E}\left[h_n(W)h^\star(W)\right]}
\end{aligned}$$

thus we have $\rho > 0$ in case (b) and $\rho \geq 1$ by the Cauchy–Schwarz inequality. Combining this with (A.1.72) and (A.1.73) yields (A.1.71). Finally we establish the bound in (1.2.11).

□

A.1.5 Proofs in Section 1.3.1

Proof of Lemma 1.3.2. We prove this lemma by a small trick, taking advantage of the idea of symmetry. Remember as in (A.1.45), X's null copy $\widetilde{X}$ is constructed such that

$$\widetilde{X} \perp\!\!\!\perp (X, Y) \mid Z, \quad \text{and} \quad \widetilde{X} \mid Z \stackrel{d}{=} X \mid Z. \tag{A.1.74}$$

We can define the null copy of $\widetilde{Y}$ by drawing from the conditional distribution of of Y given Z, without looking at (X, Y). Remark that introducing $\widetilde{Y}$ is just for the convenience of proof and does not necessarily mean we need to be able to sample it. Formally it satisfy

$$\widetilde{Y} \perp\!\!\!\perp (X, Y) \mid Z, \ \ \widetilde{Y} \mid Z \stackrel{d}{=} Y \mid Z \tag{A.1.75}$$

More specifically, we "generate" $\widetilde{Y}$ conditioning on $(\widetilde{X}, Z)$, following the same conditional distribution as $Y|X, Z$ (It can be verified this will satisfy (A.1.75)). Now by the symmetry argument, we have

$$\mathbb{E}\left[\mathbb{1}_{\{Y\cdot[\mu(\widetilde{X},Z)-\mathbb{E}[\mu(X,Z)\,|\,Z]]<0\}}\right] = \mathbb{E}\left[\mathbb{1}_{\{\widetilde{Y}\cdot[\mu(X,Z)-\mathbb{E}[\mu(X,Z)\,|\,Z]]<0\}}\right]. \tag{A.1.76}$$

Let $W = (X, Z)$ and define $g(Z) := \mathbb{E}\left[\mu(W) \,|\, Z\right],\ h(W) := \mu(W) - g(Z)$ with the associated functions denoted by $g(z),\ h(w)$, we can rewrite $f_{\ell_1}(\mu)/2$ as

$$\begin{aligned} f_{\ell_1}(\mu)/2 &= \mathbb{P}\big(Y(\mu(\widetilde{X}, Z) - \mathbb{E}\left[\mu(X, Z) \,|\, Z\right]) < 0\big) - \mathbb{P}\big(Y(\mu(X, Z) - \mathbb{E}\left[\mu(X, Z) \,|\, Z\right]) < 0\big) \\ &= \mathbb{E}\left[\mathbb{1}_{\{\widetilde{Y}\cdot[\mu(W) - \mathbb{E}[\mu(W) \,|\, Z]] < 0\}}\right] - \mathbb{E}\left[\mathbb{1}_{\{Y\cdot[\mu(W) - \mathbb{E}[\mu(W) \,|\, Z]] < 0\}}\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\left(\mathbb{1}_{\{\widetilde{Y}\cdot[\mu(W) - \mathbb{E}[\mu(W) \,|\, Z]] < 0\}} - \mathbb{1}_{\{Y\cdot[\mu(W) - \mathbb{E}[\mu(W) \,|\, Z]] < 0\}}\right) \,\middle|\, W\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\left(\mathbb{1}_{\{\widetilde{Y}\cdot h(W) < 0\}} - \mathbb{1}_{\{Y\cdot h(W) < 0\}}\right) \,\middle|\, W\right]\right] \end{aligned}$$

where the second equality is by (A.1.76), the third one comes from the law of total expectation and the fourth one is by the definition of $h(W)$. Now it suffices to consider maximizing the following quantity

$$\mathbb{E}\left[\left(\mathbb{1}_{\{\widetilde{Y}\cdot h(W) < 0\}} - \mathbb{1}_{\{Y\cdot h(W) < 0\}}\right) \,\middle|\, W = w\right] \tag{A.1.77}$$

for each $w = (x, z)$. Due to the property (A.1.75), we have

$$\mathbb{P}\left(\widetilde{Y} = y \,|\, W\right) = \mathbb{P}\left(\widetilde{Y} = y \,|\, Z\right) = \mathbb{P}\left(Y = y \,|\, Z\right) \quad y \in \{-1, 1\},$$

hence we can simplify the conditional expectation of the first indicator function in (A.1.77) into the following

$$\begin{aligned} \mathbb{E}\left[\mathbb{1}_{\{\widetilde{Y}\cdot h(W) < 0\}} \,|\, W = w\right] &= \mathbb{P}\left(\widetilde{Y} = 1,\ h(W) < 0 \,|\, W = w\right) + \mathbb{P}\left(\widetilde{Y} = -1,\ h(W) > 0 \,|\, W = w\right) \\ &= \mathbb{P}\left(Y = 1 \,|\, Z = z\right) \mathbb{1}_{\{h(w) < 0\}} + \mathbb{P}\left(Y = -1 \,|\, Z = z\right) \mathbb{1}_{\{h(w) > 0\}}. \end{aligned} \tag{A.1.78}$$

Similarly we have

$$\mathbb{E}\left[\mathbb{1}_{\{Y\cdot h(W)<0\}} \,|\, W=w\right] = \mathbb{P}\left(Y=1 \,|\, W=w\right)\mathbb{1}_{\{h(w)<0\}} + \mathbb{P}\left(Y=-1 \,|\, W=w\right)\mathbb{1}_{\{h(w)>0\}}. \tag{A.1.79}$$

When $\mathbb{E}\left[Y \,|\, W=w\right] > \mathbb{E}\left[Y \,|\, Z=z\right]$, we have

$$\mathbb{P}\left(Y=1 \,|\, W=w\right) > \mathbb{P}\left(Y=1 \,|\, Z=z\right),\ \mathbb{P}\left(Y=-1 \,|\, W=w\right) < \mathbb{P}\left(Y=-1 \,|\, Z=z\right),$$

hence in this case, by comparing (A.1.78) and (A.1.79) we know $h(w) > 0$ will maximize (A.1.77) with maximum value

$$\begin{aligned}\mathbb{P}\left(Y=-1 \,|\, Z=z\right) - \mathbb{P}\left(Y=-1 \,|\, W=w\right) &= (1-\mathbb{E}\left[Y \,|\, Z=z\right])/2 - (1-\mathbb{E}\left[Y \,|\, W=w\right])/2 \\ &= (\mathbb{E}\left[Y \,|\, W=w\right] - \mathbb{E}\left[Y \,|\, Z=z\right])/2. \end{aligned} \tag{A.1.80}$$

Similarly we can figure out the maximizer of $h(w)$ when $\mathbb{E}\left[Y \,|\, W=w\right] < \mathbb{E}\left[Y \,|\, Z=z\right]$. Finally we have

$$h(w)\begin{cases} > 0, \ \text{ when } \mathbb{E}\left[Y \,|\, W=w\right] > \mathbb{E}\left[Y \,|\, Z=z\right] \\ < 0, \ \text{ when } \mathbb{E}\left[Y \,|\, W=w\right] < \mathbb{E}\left[Y \,|\, Z=z\right] \\ \text{can be any choice, } \ \text{ when } \mathbb{E}\left[Y \,|\, W=w\right] = \mathbb{E}\left[Y \,|\, Z=z\right] \end{cases} \tag{A.1.81}$$

will maximize (A.1.77) with the maximum value $|\mathbb{E}\left[Y \,|\, W=w\right] - \mathbb{E}\left[Y \,|\, Z=z\right]|/2$. Remark the definition of $h(w) = \mu(w) - g(z)$, we can restate (A.1.81) as

$$\begin{cases} \mu(x,z) = \mu(w) > g(z), \ \text{ when } \mathbb{E}\left[Y \,|\, W=w\right] > \mathbb{E}\left[Y \,|\, Z=z\right] \\ \mu(x,z) = \mu(w) < g(z), \ \text{ when } \mathbb{E}\left[Y \,|\, W=w\right] < \mathbb{E}\left[Y \,|\, Z=z\right] \\ \text{can be any choice, } \ \text{ when } \mathbb{E}\left[Y \,|\, W=w\right] = \mathbb{E}\left[Y \,|\, Z=z\right] \end{cases} \tag{A.1.82}$$

where again $g(z) = \mathbb{E}[\mu(X, Z) \mid Z = z]$. Apparently, choosing $\mu(x, z)$ to be the true regression function $\mu^\star(x, z)$ will satisfy (A.1.82). Hence we show $f_{\ell_1}(\mu)$ is maximized at $\mu^\star$ with maximum value

$$\mathbb{E}\left|\mathbb{E}[Y \mid Z] - \mathbb{E}[Y \mid X, Z]\right|$$

which equals $\mathcal{I}_{\ell_1}$. Clearly from (A.1.82), $\mu^\star(x, z)$ is not the unique maximizer and any function in the set described in the following set can attain the maximum.

$$\{\mu : \mathbb{R}^p \to \mathbb{R} \mid \text{sign}\left(\mu(x, z) - \mathbb{E}[\mu(X, Z) \mid Z = z]\right) = \text{sign}\left(\mathbb{E}[Y \mid X = x] - \mathbb{E}[Y \mid Z = z]\right)\}. \tag{A.1.83}$$

□

Proof of Theorem 1.3.3. According to Algorithm 10, we first denote

$$\begin{aligned} &U := \mu(X, Z), \;\; g(z) := \mathbb{E}[\mu(X, Z) \mid Z = z], \\ &G_z(u) := \mathbb{P}\left(U < u \mid Z = z\right), \;\; F_z(u) := \mathbb{P}\left(U \le u \mid Z = z\right). \end{aligned} \tag{A.1.84}$$

thus have the following expression of R_i:

$$R_i = G_{Z_i}(g(Z_i))\mathbb{1}_{\{Y_i=1\}} + (1 - F_{Z_i}(g(Z_i)))\mathbb{1}_{\{Y_i=-1\}} - \mathbb{1}_{\{Y_i(\mu(W_i)-g(Z_i))<0\}}$$

First we prove that $\mathbb{E}[R_i] = f_{\ell_1}(\mu)/2$. Recall the definition of $f_{\ell_1}(\mu)$ in (1.3.2),

$$f_{\ell_1}(\mu)/2 = \mathbb{E}\left[\mathbb{1}_{\{Y\cdot[\mu(\widetilde{X},Z)-\mathbb{E}[\mu(X,Z)\mid Z]]<0\}}\right] - \mathbb{E}\left[\mathbb{1}_{\{Y\cdot[\mu(X,Z)-\mathbb{E}[\mu(X,Z)\mid Z]]<0\}}\right],$$

let $W = (X, Z)$, then it suffices to show the following

$$\mathbb{E}\left[G_Z(g(Z))\mathbb{1}_{\{Y=1\}} + (1 - F_Z(g(Z)))\mathbb{1}_{\{Y=-1\}}\right] = \mathbb{E}\left[\mathbb{1}_{\{Y\cdot[\mu(\widetilde{X},Z)-\mathbb{E}[\mu(X,Z)\,|\,Z]]<0\}}\right]. \tag{A.1.85}$$

By the law of total expectation we can rewrite the right hand side as

$$\mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{\{Y\cdot[\mu(\widetilde{X},Z)-\mathbb{E}[\mu(X,Z)\,|\,Z]]<0\}} \,|\, Z, Y\right]\right].$$

Due to the property (A.1.74), we have $\widetilde{X} \perp\!\!\!\perp (Y, Z) \mid Z$ and $\widetilde{X} \mid Z \sim X \mid Z$, which yields

$$\mathbb{E}\left[\mathbb{1}_{\{Y\cdot[\mu(\widetilde{X},Z)-\mathbb{E}[\mu(X,Z)\,|\,Z]]<0\}} \,|\, Z = z, Y = 1\right] = G_Z(g(Z))\mathbb{1}_{\{Y=1\}}.$$

And we can do similar derivations when $Y = -1$. Thus we can prove $\mathbb{E}[R_i] = f_{\ell_1}(\mu)/2$ by showing (A.1.85). In light of the deterministic relationship in Lemma 1.3.2, we have $\{L_n^\alpha(\mu) \leq f_{\ell_1}(\mu)\} \subset \{L_n^\alpha(\mu) \leq \mathcal{I}_{\ell_1}\}$, hence it suffices to prove

$$\mathbb{P}\left(L_n^\alpha(\mu) \leq f_{\ell_1}(\mu)\right) \geq 1 - \alpha - O(n^{-1/2}). \tag{A.1.86}$$

Note that $\mathrm{Var}\,(R_i)$ always exist due to the boundedness. When $\mathrm{Var}\,(R_i) = 0$, we have $R_i = f_{\ell_1}(\mu)/2 = \bar{R}$ and $s = 0$, thus $L_n^\alpha(\mu) = f_{\ell_1}(\mu)$, hence (A.1.86) trivially holds. Remark this includes the case when $\mu(X, Z) \in \mathscr{A}(Z)$. Otherwise, applying Lemma A.3.4 to i.i.d. bounded random variables R_i will yield (A.1.86), where the constant will depend on $\mathrm{Var}\,(R_i)$. □

A.1.6 Proofs in Section 1.3.2

Proof of Theorem 1.3.4. When $\mathcal{T}$ is degenerate or $\mu(X) \in \mathscr{A}(Z)$, we immediately have $L_n^{\alpha,\mathcal{T}}(\mu) = 0$ according to Algorithm 11, which implies the coverage validity. Below we focus on the non-trivial

case. Due to the deterministic relationship

$$f_n^{\mathcal{T}}(\mu) \leq f_n^{\mathcal{T}}(\mu^\star) \leq f(\mu^\star) = \mathcal{I},$$

it suffices to prove

$$\mathbb{P}_P\left(L_n^{\alpha,\mathcal{T}}(\mu) \leq f_n^{\mathcal{T}}(\mu)\right) \geq 1 - \alpha - o(1). \tag{A.1.87}$$

which can be reduced to establishing certain asymptotic normality based on i.i.d. random variables $R_m, V_m, m \in [n_1]$ whenever the variance of the asymptotic distribution is nonzero. First, we verify that under the stated conditions, all the involving moments are finite, which can be reduced to show

$$\operatorname{Var}(R_m), \operatorname{Var}(V_m) < \infty.$$

For a given n_2, it can be further reduced to the following

$$\operatorname{Var}\left(Y_i\left(\mu(X_i, Z_i) - \mathbb{E}\left[\mu(X_i, Z_i) \mid \boldsymbol{Z}_m, \boldsymbol{T}_m\right]\right)\right)$$
$$\operatorname{Var}\left(\operatorname{Var}\left(\mu(X_i, Z_i) \mid \boldsymbol{Z}_m, \boldsymbol{T}_m\right)\right) < \infty.$$

Using similar strategies in the proof of Theorem 1.2.3, we can show the above holds under the moment conditions $\mathbb{E}\left[Y^4\right], \mathbb{E}\left[\mu^4(X)\right] < \infty$ by the Cauchy–Schwarz inequality and the tower property of conditional expectation.

Note that in the proof of the main result, i.e. Theorem 1.2.3, we consider four different cases based on whether some variances are zero or not. Here we only pursue the asymptotic coverage validity, then the discussion on those four different cases becomes very straighforward. When both the variances of R_m, V_m are zero, we have $\bar{R}/\bar{V} = f_n^{\mathcal{T}}(\mu),\ s^2 = 0$, then (A.1.87) holds im-

mediately. When $\mathrm{Var}\,(V_m) = 0$, we can simply establish the asymptotic normality by the central limit theorem. Otherwise, delta method can be applied. Here we give the derivation for the most non-trivial case where $\mathrm{Var}\,(R_m)\,, \mathrm{Var}\,(V_m) > 0$. Denote random vectors $\{U_m\}_{m=1}^{n_1} = \{(U_{m1}, U_{m2})\}_{m=1}^{n_1} \overset{i.i.d.}{\sim} U = (U_1, U_2)$ to be

$$
\begin{aligned}
U_{m1} &= R_m - \mathbb{E}\left[Y_i \left(\mu(X_i, Z_i) - \mathbb{E}\left[\mu(X_i, Z_i) \mid \boldsymbol{Z}_m, \boldsymbol{T}_m\right]\right], \qquad \text{(A.1.88)} \\
U_{m2} &= V_m - \mathbb{E}\left[\mathrm{Var}\left(\mu(X_i, Z_i) \mid \boldsymbol{Z}_m, \boldsymbol{T}_m\right)\right] \qquad \text{(A.1.89)}
\end{aligned}
$$

hence we have $\mathbb{E}\left[U\right] = 0$. Denote $h^{\mathcal{T}}(W_i) = \mu(X_i, Z_i) - \mathbb{E}\left[\mu(X_i, Z_i) \mid \boldsymbol{Z}_m, \boldsymbol{T}_m\right]$, we have the following holds

$$
\begin{aligned}
f_n^{\mathcal{T}}(\mu) &= \frac{\mathbb{E}\left[\mathrm{Cov}(\mu^\star(X_i, Z_i), \mu(X_i, Z_i) \mid \boldsymbol{Z}, \boldsymbol{T})\right]}{\sqrt{\mathbb{E}\left[\mathrm{Var}(\mu(X_i, Z_i) \mid \boldsymbol{Z}, \boldsymbol{T})\right]}} \\
&= \frac{\mathbb{E}\left[\mathrm{Cov}(\mu^\star(X_i, Z_i), h^{\mathcal{T}}(W_i) \mid \boldsymbol{Z}, \boldsymbol{T})\right]}{\sqrt{\mathbb{E}\left[\mathbb{E}\left[(h^{\mathcal{T}}(W_i)^2)\right]\right]}} \\
&= \frac{\mathbb{E}\left[\mu^\star(X_i, Z_i) h^{\mathcal{T}}(W_i)\right]}{\sqrt{\mathbb{E}\left[\mathbb{E}\left[(h^{\mathcal{T}}(W_i)^2)\right]\right]}} \\
&= \frac{\mathbb{E}\left[Y_i h^{\mathcal{T}}(W_i)\right]}{\sqrt{\mathbb{E}\left[(h^{\mathcal{T}}(W_i)^2)\right]}},
\end{aligned}
$$

where the first equality holds by the definition of $f_n^{\mathcal{T}}(\mu)$, the second inequality holds by the definition of $h^{\mathcal{T}}(W_i)$. Regarding the third equality, we make use of the fact $\mathbb{E}\left[h^{\mathcal{T}}(W_i) \mid \boldsymbol{Z}_m, \boldsymbol{T}_m\right] = 0$ and the tower property of conditional expectation. The last inequality holds by the tower property of conditional expectation and the fact that $h^{\mathcal{T}}(W_i) \in \mathscr{A}(\boldsymbol{X}_m, \boldsymbol{Z}_m)$. Let $T = \bar{R}/\bar{V}$, then $T - f_n^{\mathcal{T}}(\mu)$ can be rewritten as

$$
T - f_n^{\mathcal{T}}(\mu) = \frac{\bar{U}_1 + \mathbb{E}\left[Y_i h^{\mathcal{T}}(W_i)\right]}{\sqrt{\bar{U}_2 + \mathbb{E}\left[(h^{\mathcal{T}}(W_i)^2)\right]}} - \frac{\mathbb{E}\left[Y_i h^{\mathcal{T}}(W_i)\right]}{\sqrt{\mathbb{E}\left[(h^{\mathcal{T}}(W_i)^2)\right]}} := H(\bar{U})
$$

where $\bar{U} = (\bar{U}_1, \bar{U}_2) = \frac{1}{n_1}\sum_{i=1}^{n} U_m$ and $H : \mathbb{R}^2 \to \mathbb{R}$ is defined through the following:

$$H(x) = H(x_1, x_2) := \frac{x_1 + \mathbb{E}\left[Y_i h^{\mathcal{T}}(W_i)\right]}{\sqrt{x_2 + \mathbb{E}\left[(h^{\mathcal{T}}(W_i)^2)\right]}} - \frac{\mathbb{E}\left[Y_i h^{\mathcal{T}}(W_i)\right]}{\sqrt{\mathbb{E}\left[(h^{\mathcal{T}}(W_i)^2)\right]}} := H(\bar{U})$$

when $x_2 > -\mathbb{E}\left[(h^{\mathcal{T}}(W_i)^2)\right]$ and is set to be $\frac{\mathbb{E}\left[Y_i h^{\mathcal{T}}(W_i)\right]}{\sqrt{\mathbb{E}\left[(h^{\mathcal{T}}(W_i)^2)\right]}}$ otherwise. Note that the first order derivatives of $H(x)$ exists, by applying the multivariate Delta method to mean zero random vectors $\{(U_{m1}, U_{m2})\}_{m=1}^{n_1}$ with the nonlinear function chosen as H, we have

$$\sqrt{n_1}(T - f_n^{\mathcal{T}}(\mu)) \xrightarrow{d} \mathcal{N}\left(0, \widetilde{\sigma}^2\right)$$

whenever the variance term $\widetilde{\sigma}^2$ is nonzero. Exactly following the strategy in the proof of Theorem 1.2.3, we have $\widetilde{\sigma}^2 > 0$ under the case where $\mathrm{Var}\left(R_m\right), \mathrm{Var}\left(V_m\right) > 0$. Also notice s^2 is a consistent estimator of $\widetilde{\sigma}^2$, then by the argument of Slutsky's Theorem, (A.1.87) is established. □

A.2 An example for projection methods

Consider covariates $W = (W_1, W_2)$ distributed as $W_1 \sim \mathcal{N}(0,1)$ and $W_2 = W_1^2 + \mathcal{N}(0,1)$. Let $Y = W_1^2 + \mathcal{N}(0,1)$, with all the Gaussian random variables independent. Then W_1 is the only important variable; formally: $W_1 \not\perp\!\!\!\perp Y \mid W_2$ and $W_2 \perp\!\!\!\perp Y \mid W_1$. But the projection parameters are $(\mathbb{E}\left[W^\top W\right])^{-1}\mathbb{E}\left[WY\right] = (0, \frac{3}{4})^\top$, i.e., zero for the non-null covariate and non-zero for the null covariate.

A.3 Rate results

Theorem A.3.1 (Floodgate validity). *For any given working regression function $\mu : \mathbb{R}^p \to \mathbb{R}$ and i.i.d. data $\{(Y_i, X_i, Z_i)\}_{i=1}^{n}$, if $\mathbb{E}[Y^{12}]$, $\mathbb{E}[\mu^{12}(X,Z)] < \infty$, then $L_n^{\alpha}(\mu)$ from Algorithm 1*

satisfies

$$\mathbb{P}\left(L_n^{\alpha}(\mu) \leq \mathcal{I}\right) \geq 1 - \alpha - Cn^{-1/2}$$

for some constant C depending only on the moments of Y and $\mu(X, Z)$.

The proof can be found in Appendix A.3.1. Establishing the $n^{-1/2}$ rate requires relatively recent Berry–Esseen-type results for the delta method (Pinelis et al., 2016) and also necessitates the existence of 12th moments.

Theorem A.3.2. *Under the conditions of Theorem A.3.1 and the additional moment condition that $\mathbb{E}\left[\mathrm{Var}\left(Y(\mu(X, Z) - \mathbb{E}\left[\mu(X, Z) \mid Z\right]) \mid Z\right)\right] > 0$, $L_{n,K}^{\alpha}(\mu)$ computed by replacing R_i and V_i with R_i^K and V_i^K, respectively, in Algorithm 1 satisfies*

$$\inf_{K>1} \mathbb{P}\left(L_{n,K}^{\alpha}(\mu) \leq \mathcal{I}\right) \geq 1 - \alpha - Cn^{-1/2}$$

for some constant C depending only on the moments of Y and $\mu(X, Z)$.

The proof can be found in Appendix A.3.1. Note that the additional assumption beyond Theorem A.3.1 of $\mathbb{E}\left[\mathrm{Var}\left(Y(\mu(X, Z) - \mathbb{E}\left[\mu(X, Z) \mid Z\right]) \mid Z\right)\right] > 0$ is only needed for $n^{-1/2}$-rate coverage validity *uniformly* over $K > 1$, and could be removed for the same result for any fixed $K > 1$.

A.3.1 Proofs in Appendix A.3

Theorem A.3.1

Proof of Theorem A.3.1. Recall in Algorithm 1, we denote $R_i = Y_i\big(\mu(X_i, Z_i) - \mathbb{E}\left[\mu(X_i, Z_i) \mid Z_i\right]\big)$ and $V_i = \mathrm{Var}\left(\mu(X_i, Z_i) \mid Z_i\right)$ for each $i \in [n]$, and compute their sample mean $(\bar{R}, \bar{V})$ and sam-

ple covariance matrix $\widehat{\Sigma}$. The LCB is constructed as

$$L_n^{\alpha}(\mu) = \max\left\{\frac{\bar{R}}{\sqrt{\bar{V}}} - \frac{z_\alpha s}{\sqrt{n}}, 0\right\}, \text{ where } s^2 = \frac{1}{\bar{V}}\left[\left(\frac{\bar{R}}{2\bar{V}}\right)^2 \widehat{\Sigma}_{22} + \widehat{\Sigma}_{11} - \frac{\bar{R}}{\bar{V}}\widehat{\Sigma}_{12}\right].$$

Following exactly the same discussions as those from the beginning to (A.1.10) in the proof of Theorem 1.2.3, we have

- Theorem A.3.1 can be proved under the weaker moment conditions that $\mathbb{E}\left[Y^{12}\right], \mathbb{E}\left[h^{12}(W)\right] < \infty$, which is assumed for the following proof;
- it suffices to prove

$$\mathbb{P}\left(\frac{\bar{R}}{\sqrt{\bar{V}}} - \frac{z_\alpha s}{\sqrt{n}} \leq f(\mu)\right) \geq 1 - \alpha - C/\sqrt{n} \tag{A.3.1}$$

  for some constant C when $\mathbb{E}\left[\mathrm{Var}(\mu(X, Z) \mid Z)\right] \neq 0$;
- we can assume $\mathbb{E}\left[h^2(W)\right] = 1$ without loss of generality.

We will utilize Berry–Esseen-type bounds to prove (A.3.1). Now we still consider the following four cases.

(I) $\mathrm{Var}\left(Yh(W)\right) = 0$ and $\mathrm{Var}\left(\mathrm{Var}\left(h(W) \mid Z\right)\right) = 0$.

(II) $\mathrm{Var}\left(Yh(W)\right) > 0$ and $\mathrm{Var}\left(\mathrm{Var}\left(h(W) \mid Z\right)\right) = 0$.

(III) $\mathrm{Var}\left(Yh(W)\right) = 0$ and $\mathrm{Var}\left(\mathrm{Var}\left(h(W) \mid Z\right)\right) > 0$.

(IV) $\mathrm{Var}\left(Yh(W)\right) > 0$ and $\mathrm{Var}\left(\mathrm{Var}\left(h(X) \mid Z\right)\right) > 0$.

Note that assuming $\mathbb{E}\left[Y^{12}\right]$ and $\mathbb{E}\left[h^{12}(W)\right] < \infty$ ensures all the above variances exist due to the same bounding strategy as (A.1.9).

Case (I): (A.3.1) holds by the discussion for Case (I) in the proof of Theorem 1.2.3.

Case (II): due to the derivations for Case (II) in the proof of Theorem 1.2.3, the problem is reduced to showing

$$\mathbb{P}\left(\bar{R} - \frac{z_\alpha(\widehat{\Sigma}_{11})^{1/2}}{\sqrt{n}} \leq \mathbb{E}\left[Yh(W)\right]\right) \geq 1 - \alpha - C/\sqrt{n}. \tag{A.3.2}$$

As mentioned in the proof of Theorem 1.2.3, $\bar{R}$ is simply the sample mean estimator of the quantity $\mathbb{E}\left[Yh(W)\right]$ and $\widehat{\Sigma}_{11}$ is the corresponding sample variance. Therefore, the CLT and Slutsky's theorem immediately establish the asymptotic coverage validity. To prove the $1/\sqrt{n}$ rate in (A.3.2), stronger results are needed. The classical Berry–Esseen bound serves as the main ingredient, which states that

Lemma A.3.3 (Berry–Esseen bound)**.** *There exists a positive constant C, such that for $i.i.d.$ mean zero random variables $X_1, \ldots, X_n$ satisfying*

(1) $\mathbb{E}[X_1^2] = \sigma^2 > 0$

(2) $\mathbb{E}[|X_1|^3] = \rho < \infty$

if we define $F_n(x)$ to be the cumulative distribution function (CDF) of the scaled average $\sqrt{n}\bar{X}/\sigma$ and denote the CDF of the standard normal distribution by $\Phi(x)$, then we have

$$\sup_{x\in\mathbb{R}} |F_n(x) - \Phi(x)| \leq \frac{C\rho}{\sigma^3\sqrt{n}}. \tag{A.3.3}$$

Since σ in the above result is generally unknown and usually replaced by the sample variance $s_\sigma^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$, we need the following lemma, which is proved in Bentkus et al. (1996).

Lemma A.3.4 (Berry–Esseen bound for Student's statistic)**.** *Under the same conditions as in Lemma A.3.3, if we redefine $F_n(x)$ to be the cumulative distribution function (CDF) of the Student t-statistic*

$\sqrt{n}\bar{X}/s_\sigma$, *then we have the following Berry–Esseen bound*

$$\sup_{x\in\mathbb{R}} |F_n(x) - \Phi(x)| \leq \frac{C'\rho}{\sigma^3\sqrt{n}}. \tag{A.3.4}$$

To apply Lemma A.3.4, since we are in Case (II) where $\text{Var}\left(\text{Var}\left(h(W) \mid Z\right)\right) = 0$ and $\text{Var}\left(Yh(W)\right) > 0$, it suffices to verify the finiteness of the term "ρ" in our context:

$$\begin{aligned} \rho &= \mathbb{E}\left[\left|Yh(W) - \mathbb{E}\left[Yh(W)\right]\right|^3\right] \\ &\leq 2^{3-1}\left(\mathbb{E}\left[Y^3h^3(W)\right] + \left|\mathbb{E}\left[Yh(W)\right]\right|^3\right) < \infty \end{aligned}$$

where the equality holds since we assume $\mathbb{E}\left[h^2(W)\right] = 1$ and the inequality comes from the C_r inequality. For the last inequality, using the Cauchy–Schwarz inequality and the fact that higher moments dominate lower moments, we obtain the finiteness when assuming $\mathbb{E}\left[Y^6\right], \mathbb{E}\left[h^6(W)\right] < \infty$, which holds under the assumed moment conditions. Now by applying the Berry–Esseen bound in Lemma A.3.4 with $\bar{X} = \bar{R} - \mathbb{E}\left[Yh(W)\right]$ and $s_\sigma^2 = \widehat{\Sigma}_{11}$, we obtain (A.3.2).

Case (III): due to (A.1.11), we have

$$\frac{\bar{R}}{\sqrt{\bar{V}}} - \frac{z_\alpha s}{\sqrt{n}} = \frac{\mathbb{E}\left[Yh(W)\right]}{\sqrt{\bar{V}}} - \frac{z_\alpha s}{\sqrt{n}}, \text{ where } s^2 = \frac{1}{\bar{V}}\left(\frac{\mathbb{E}\left[Yh(W)\right]}{2\bar{V}}\right)^2 \widehat{\Sigma}_{22}.$$

Note $\frac{\mathbb{E}[Yh(W)]}{\sqrt{\bar{V}}}$ is a nonlinear function of the moment estimators, so the following asymptotic normality result is a direct consequence of the multivariate delta method,

$$\sqrt{n}\left(\frac{\mathbb{E}\left[Yh(W)\right]}{\sqrt{\bar{V}}} - f(\mu)\right) \xrightarrow{d} \mathcal{N}\left(0, \widetilde{\sigma}_0^2\right),$$

where $\widetilde{\sigma}_0^2 = H_2(\mathbf{0})$ will be specified later (see the definition of $H_2(x)$ in (A.3.10)) and s^2 in $L_n^\alpha(\mu)$ is a consistent estimator of it. To establish the rate $1/\sqrt{n}$, the classical Berry–Esseen result needs to

be extended for nonlinear statistics. Note that Case (IV) involves a nonlinear statistic too, and is a bit more complicated. Hence we focus on Case (IV) and omit the very similar proof for Case (III).

Case (IV): Denote $T := \left(\frac{\bar{R}}{\sqrt{\bar{V}}} - f(\mu)\right)/s$. Under specific moment conditions, we will establish the Berry–Esseen-type bound below:

$$\sup_{t\in\mathbb{R}} \left|\mathbb{P}\left(\sqrt{n}T \le t\right) - \Phi(t)\right| = O\left(\frac{1}{\sqrt{n}}\right) \tag{A.3.5}$$

where $\Phi(t)$ denotes the CDF of the standard normal distribution.

The proof relies on a careful analysis of nonlinear statistics. We take advantage of the results in a recent paper (Pinelis et al., 2016) that establishes Berry–Esseen bounds with rate $1/\sqrt{n}$ for the multivariate delta method when the function applied to the sample mean estimator satisfies certain smoothness conditions. And the constants in the rate depend on the distribution only through several moments. Specifically, consider $U, U_1, \ldots, U_n$ to be i.i.d. random vectors on a set $\mathcal{X}$ and a functional $H : \mathcal{X} \to \mathbb{R}$ which satisfies the following smoothness condition:

Condition A.3.5. *There exists $\varepsilon, M_\varepsilon > 0$ and a continuous linear functional $L : \mathcal{X} \to \mathbb{R}$ such that*

$$|H(x) - L(x)| \le M_\varepsilon \|x\|^2 \quad \textit{for all } x \in \mathcal{X} \textit{ with } \|x\| \le \varepsilon \tag{A.3.6}$$

We can think of L as the first-order Taylor expansion of H. This smoothness condition basically requires H to be nearly linear around the origin and can be satisfied if its second derivatives are bounded in the small neighbourhood $\{x : \|x\| \le \varepsilon\}$. Before stating Pinelis et al. (2016)'s result (we change their notation to avoid conflicts with the notation in the main text of this paper), define $\bar{U} := \frac{1}{n}\sum_{i=1}^n U_i$ and

$$\widetilde{\sigma} := \|L(U)\|_2, \;\; \nu_p := \|U\|_p, \;\; \varsigma_p := \frac{\|L(U)\|_p}{\widetilde{\sigma}},$$

where for a given random vector $U = (U_1, \cdots, U_d) \in \mathbb{R}^d$, $\|U\|_p$ is defined as $\|U\|_p = (\mathbb{E}[\|U\|^p])^{1/p}$ with $\|u\|^p := \sum_{j=1}^d |u_j|^p$.

Theorem A.3.6. *(Pinelis et al., 2016, Theorem 2.11) Let $\mathcal{X}$ be a Hilbert space, let H satisfy Condition A.3.5 for some $\epsilon > 0$, and assume $\mathbb{E}[U] = 0$, $\widetilde{\sigma} > 0$ and $\nu_3 < \infty$, then*

$$\sup_{t\in\mathbb{R}} \left| \mathbb{P}\left(\frac{\sqrt{n} H(\bar{U})}{\widetilde{\sigma}} \leq t \right) - \Phi(t) \right| \leq \frac{C}{\sqrt{n}} \tag{A.3.7}$$

where the constant C depends on the distribution of U only through $\widetilde{\sigma}, \nu_2, \nu_3, \varsigma_3$ (it also depends on the smoothness of the functional H through ϵ, M_ϵ).

Note that the above result is a generalization of the standard Berry–Esseen bound. $\widetilde{\sigma}^2$ is the variance term of the asymptotic normal distribution. ς_3 is closely related to the term ρ/σ^2 in (A.3.3). The quantities $\widetilde{\sigma}, \nu_2, \nu_3, \varsigma_3$ involved in the constant C only involve up to third moments, which is in accordance with the standard Berry–Esseen bound in Lemmas A.3.3 and A.3.4. Note the existence of $\widetilde{\sigma}, \nu_2, \varsigma_3$ is implied by $\nu_3 < \infty$ due to the fact that lower moments can be controlled by higher moments, together with the linearity of the functional L. To apply Theorem A.3.6 to our problem, we first let $\mathcal{X} = \mathbb{R}^5$ and random vectors $\{U_i\}_{i=1}^n = \{(U_{i1}, U_{i2}, U_{i3}, U_{i4}, U_{i5})\}_{i=1}^n \overset{i.i.d.}{\sim} U_0 = (U_{01}, U_{02}, U_{03}, U_{04}, U_{05})$ to be

$$\begin{aligned} U_{i1} = R_i - \mathbb{E}[Yh(W)], \; U_{i2} = V_i - \mathbb{E}\left[h^2(W)\right], \\ U_{i3} = Y_i^2 h^2(W_i) - \mathbb{E}\left[Y^2 h^2(W)\right], \; U_{i4} = (\mathrm{Var}(h(W_i) \mid Z_i))^2 - \mathbb{E}\left[(\mathrm{Var}(h(W) \mid Z))^2\right], \\ U_{i5} = R_i \mathrm{Var}(h(W_i) \mid Z_i) - \mathbb{E}[Yh(W)\mathrm{Var}(h(W) \mid Z)]. \end{aligned} \tag{A.3.8}$$

Recall the definition $R_i = Y_i\big(\mu(X_i, Z_i) - \mathbb{E}[\mu(X, Z_i) \mid Z_i]\big)$ and $V_i = \mathrm{Var}(\mu(X_i, Z_i) \mid Z_i)$, hence we have $\mathbb{E}[U_i] = \mathbb{E}[U_0] = \mathbf{0}$. Let $\bar{U} = (\bar{U}_1, \bar{U}_2, \bar{U}_3, \bar{U}_4, \bar{U}_5) = \frac{1}{n}\sum_{i=1}^n U_i \in \mathbb{R}^5$, recall

the definition $T = \left(\frac{\bar{R}}{\sqrt{\bar{V}}} - f(\mu)\right)/s$ where $s^2 = \frac{1}{\bar{V}}\left[\left(\frac{\bar{R}}{2\bar{V}}\right)^2 \widehat{\Sigma}_{22} + \widehat{\Sigma}_{11} - \frac{\bar{R}}{\bar{V}}\widehat{\Sigma}_{12}\right]$, then T can be rewritten as

$$T = H(\bar{U}) := \frac{H_1(\bar{U}_1, \bar{U}_2)}{\sqrt{H_2(\bar{U})}},$$

where $H_1(\bar{U}_1, \bar{U}_2)$ and $H_2(\bar{U})$ are defined as

$$H_1(\bar{U}_1, \bar{U}_2) := \frac{\bar{U}_1 + \mathbb{E}\left[Yh(W)\right]}{\sqrt{\bar{U}_2 + \mathbb{E}\left[h^2(W)\right]}} - \frac{\mathbb{E}\left[Yh(W)\right]}{\sqrt{\mathbb{E}\left[h^2(W)\right]}}, \tag{A.3.9}$$

$$\begin{aligned} H_2(\bar{U}) := \frac{1}{\bar{U}_2 + \mathbb{E}\left[h^2(W)\right]}\Bigg[& \left(\frac{\bar{U}_1 + \mathbb{E}\left[Yh(W)\right]}{2(\bar{U}_2 + \mathbb{E}\left[h^2(W)\right])}\right)^2 \left(\bar{U}_4 + \mathbb{E}\left[(\mathrm{Var}\left(h(W) \mid Z\right))^2\right] - (\bar{U}_2 + \mathbb{E}\left[h^2(W)\right])^2\right) \\ & + \bar{U}_3 + \mathbb{E}\left[Y^2 h^2(W)\right] - (\bar{U}_1 + \mathbb{E}\left[Yh(W)\right])^2 \\ & - \frac{\bar{U}_1 + \mathbb{E}\left[Yh(W)\right]}{\bar{U}_2 + \mathbb{E}\left[h^2(W)\right]}\left(\bar{U}_5 + \mathbb{E}\left[Yh(W)\mathrm{Var}\left(h(W) \mid Z\right)\right] - (\bar{U}_1 + \mathbb{E}\left[Yh(W)\right])(\bar{U}_2 + \mathbb{E}\left[h^2(W)\right])\right)\Bigg]. \end{aligned} \tag{A.3.10}$$

Note $H(x) = H(x_1, x_2, x_3, x_4, x_5) : \mathbb{R}^5 \to \mathbb{R}$ is defined by replacing the above $\bar{U} = (\bar{U}_1, \bar{U}_2, \bar{U}_3, \bar{U}_4, \bar{U}_5)$ by $x := (x_1, x_2, x_3, x_4, x_5)$ respectively. When $x_2 > -\mathbb{E}\left[h^2(W)\right]$ or $H_2(x) = 0$, $H(x)$ is set to be 0. If we can verify the conditions for $T = H(\bar{U})$, Theorem A.3.6 implies

$$\sup_{t\in\mathbb{R}} \left|\mathbb{P}\left(\sqrt{n}T \le t\widetilde{\sigma}\right) - \Phi(t)\right| \le \frac{C}{\sqrt{n}},$$

for some constant C, where $\widetilde{\sigma} = \|L(U_0)\|_2 > 0$ (we will define $L(x)$ shortly and subsequently show $\widetilde{\sigma} = 1$). Theorem A.3.6 says that the constant C above only depends on some universal constants and $\widetilde{\sigma}, \nu_2, \nu_3, \varsigma_3$, which are the moments of U (i.e., the moments of U_i). Since U_i (defined in the three lines around (A.3.8)) is a function of Y_i and $h(W_i)$, we will apply the Cauchy–Schwarz inequality to further bound the moments of U by the moments of Y and $h(W) = \mu(X, Z) - \mathbb{E}\left[\mu(X, Z) \mid Z\right]$. First we need to verify Condition A.3.5, i.e., there exists $\varepsilon, M_\varepsilon > 0$ and a continu-

ous linear functional $L : \mathbb{R}^5 \to \mathbb{R}$ such that

$$|H(x) - L(x)| \leq M_\varepsilon \|x\|^2 \text{ for all } x \in \mathbb{R}^5 \text{ with } \|x\| \leq \varepsilon. \tag{A.3.11}$$

Second, we will show $\widetilde{\sigma}$, ν_3, and ς_3 are finite under the stated moment conditions. Regarding the smoothness condition, consider the first order Taylor expansion of H at zero,

$$H(\mathbf{0}) + \frac{\partial H}{\partial x_1}(\mathbf{0})x_1 + \frac{\partial H}{\partial x_2}(\mathbf{0})x_2 + \frac{\partial H}{\partial x_3}(\mathbf{0})x_3 + \frac{\partial H}{\partial x_4}(\mathbf{0})x_4 + \frac{\partial H}{\partial x_5}(\mathbf{0})x_5.$$

Note that for $H(\mathbf{0}) = H_1(\mathbf{0})/\sqrt{H_2(\mathbf{0})}$, we have $H_1(\mathbf{0}) = 0$ and $H_2(\mathbf{0}) > 0$ (denote $\widetilde{\sigma}_0^2 := H_2(\mathbf{0})$ and we will show it is positive over the course of derivations from (A.3.17) to (A.3.21). After simplifying the expression of $H_2(\mathbf{0})$, we give the explicit form of $\widetilde{\sigma}_0^2$ below:

$$\begin{aligned}\widetilde{\sigma}_0^2 = \frac{1}{\mathbb{E}\left[h^2(W)\right]}\Bigg[&\left(\frac{\mathbb{E}\left[Yh(W)\right]}{2(\mathbb{E}\left[h^2(W)\right])}\right)^2 \mathrm{Var}\left(\mathrm{Var}\left(h(W) \mid Z\right)\right) + \mathrm{Var}\left(Yh(W)\right) \\ &- \frac{\mathbb{E}\left[Yh(W)\right]}{\mathbb{E}\left[h^2(W)\right]}\mathrm{Cov}\left(Yh(W), \mathrm{Var}\left(h(W) \mid Z\right)\right)\Bigg].\end{aligned} \tag{A.3.12}$$

Using the chain rule of derivatives, we have for $m \in [5]$,

$$\frac{\partial H}{\partial x_m}(\mathbf{0}) = \frac{\partial H_1}{\partial x_m}(\mathbf{0})/\sqrt{H_2(\mathbf{0})} - \frac{H_1(\mathbf{0})}{2H_2(\mathbf{0})^{3/2}} \cdot \frac{\partial H_2}{\partial x_m}(\mathbf{0}) = \frac{\partial H_1}{\partial x_m}(\mathbf{0})/\widetilde{\sigma}_0.$$

Since $H_1(x_1, x_2)$ only depends on x_1, x_2, we only need to evaluate two partial derivatives to compute the first order Taylor expansion of H at zero, yielding the following linear function

$$\frac{1}{\widetilde{\sigma}_0}\left(\frac{1}{\sqrt{\mathbb{E}\left[h^2(W)\right]}}x_1 - \frac{\mathbb{E}\left[Yh(W)\right]}{2(\sqrt{\mathbb{E}\left[h^2(W)\right]})^3}x_2\right) := L(x), \tag{A.3.13}$$

which is denoted by $L(x) = L(x_1, x_2)$ and satisfies $L(\mathbf{0}) = 0$. Note that when $\epsilon = \mathbb{E}\left[h^2(W)\right]/2$,

we have

$$\min_{\|x\|\leq\epsilon}(x_2+\mathbb{E}\left[h^2(W)\right])=\mathbb{E}\left[h^2(W)\right]-\epsilon>0.$$

Since $H_2(x)$ is continuous around zero and $H_2(\mathbf{0})>0$ (which will be shown in the following proof), we can similarly choose ϵ sufficiently small such that $\min_{\|x\|\leq\epsilon}H_2(x)>0$. Recall $H(x)=H_1(x)/\sqrt{H_2(x)}$, where H_1, H_2 are defined in (A.3.9) and (A.3.10), so $H(x)$ is continuous on $\{x:\|x\|\leq\varepsilon\}$. Furthermore, its second partial derivatives exist and are continuous over the compact set $\{x:\|x\|\leq\varepsilon\}$, thus are also bounded, which implies that there exists $M_\epsilon>0$ such that (A.3.11) holds.

As for $\widetilde{\sigma}$, ν_3, and ς_3, we will now establish the following moment bounds:

$$0<\widetilde{\sigma}:=\|L(U_0)\|_2<\infty, \tag{A.3.14}$$

$$\nu_2:=\|U_0\|_2,\qquad \nu_3:=\|U_0\|_3<\infty,$$

$$\varsigma_3:=\frac{\|L(U_0)\|_3}{\widetilde{\sigma}}<\infty. \tag{A.3.15}$$

Note that $\nu_3^3=\|U_0\|_3^3=\mathbb{E}\left[|U_{01}|^3\right]+\mathbb{E}\left[|U_{02}|^3\right]+\mathbb{E}\left[|U_{03}|^3\right]+\mathbb{E}\left[|U_{04}|^3\right]+\mathbb{E}\left[|U_{05}|^3\right]$ and

$$\begin{aligned}(\varsigma_3\widetilde{\sigma})^3=\mathbb{E}\left[|L(U_0)|^3\right]&=\frac{1}{\widetilde{\sigma}_0^3}\mathbb{E}\left[\left|\frac{1}{\sqrt{\mathbb{E}\left[h^2(W)\right]}}U_{01}-\frac{\mathbb{E}\left[Yh(W)\right]}{2(\sqrt{\mathbb{E}\left[h^2(W)\right]})^3}U_{02}\right|^3\right]\\&\leq\frac{2^{3-1}}{\widetilde{\sigma}_0^3}\left(\frac{1}{(\sqrt{\mathbb{E}\left[h^2(W)\right]})^3}\mathbb{E}\left[|U_{01}|^3\right]+\frac{(\mathbb{E}\left[Yh(W)\right])^3}{8(\sqrt{\mathbb{E}\left[h^2(W)\right]})^9}\mathbb{E}\left[|U_{02}|^3\right]\right)\end{aligned} \tag{A.3.16}$$

where the equalities hold due to the definitions of L and ς_3 in (A.3.13), (A.3.15), and the inequality holds as a result of the C_r inequality. Due to the fact that the finiteness of higher moments implies that of lower moments and (A.3.16), we only need to show

(i) $\mathbb{E}\left[|U_{01}|^3\right], \mathbb{E}\left[|U_{02}|^3\right], \mathbb{E}\left[|U_{03}|^3\right], \mathbb{E}\left[|U_{04}|^3\right], \mathbb{E}\left[|U_{05}|^3\right] < \infty$,

(ii) $\widetilde{\sigma}_0^2 = H_2(\mathbf{0}) > 0$,

(iii) $\widetilde{\sigma}^2 = \|L(U_0)\|_2 > 0$,

under the stated moment conditions. For (iii), actually we will show $\widetilde{\sigma}^2 = 1$.

Starting with (i), we have

$$\begin{aligned}
\mathbb{E}\left[|U_{02}|^3\right] = \mathbb{E}\left[|U_{i2}|^3\right] &= \mathbb{E}\left[\left|V_i - \mathbb{E}\left[h^2(W)\right]\right|^3\right] \\
&\leq 2^{3-1}\left(\mathbb{E}\left[\left|\mathrm{Var}\left(\mu(W_i) \mid Z_i\right)\right|^3\right] + \left(\mathbb{E}\left[h^2(W)\right]\right)^3\right) \\
&\leq 2^{3-1}\left(\mathbb{E}\left[\mathbb{E}\left[h^6(W_i) \mid Z_i\right]\right] + \left(\mathbb{E}\left[h^2(W)\right]\right)^3\right) < \infty,
\end{aligned}$$

where the first inequality comes from the C_r inequality, the second holds by the definition of h and Jensen's inequality, and the third inequality holds due to the tower property of conditional expectation and $\mathbb{E}\left[h^6(W)\right] < \infty$ under the assumed moment conditions. For the term $\mathbb{E}\left[|U_{01}|^3\right]$, we have

$$\begin{aligned}
\mathbb{E}\left[|U_{01}|^3\right] = \mathbb{E}\left[|U_{i1}|^3\right] &= \mathbb{E}\left[\left|R_i - \mathbb{E}\left[Yh(W)\right]\right|^3\right] \\
&\leq 2^{3-1}\left(\mathbb{E}\left[\left|Y_i(\mu(W_i) - \mathbb{E}\left[\mu(W_i) \mid Z_i\right])\right|^3\right] + \left(\mathbb{E}\left[Yh(W)\right]\right)^3\right) \\
&= 2^{3-1}\left(\mathbb{E}[|Y^3h^3(W)|] + \left(\mathbb{E}\left[Yh(W)\right]\right)^3\right) < \infty,
\end{aligned}$$

where the first inequality holds due to the C_r inequality and the second inequality holds due to the Cauchy–Schwarz inequality and the assumed moment conditions. The same approach and inequalities can be used for the other three terms, i.e., we have $\mathbb{E}\left[|U_{03}|^3\right], \mathbb{E}\left[|U_{04}|^3\right], \mathbb{E}\left[|U_{05}|^3\right] < \infty$. Note that U_{03}, U_{04}, and U_{05} involve higher-order polynomials of $Y_ih(W_i)$ and $\mathrm{Var}\left(h(W_i) \mid Z_i\right)$ than U_{01}, U_{02}, and thus require assuming bounded 12th moments to ensure the boundedness of

their third absolute moments, hence the assumptions in Theorem A.3.1 that $\mathbb{E}\left[Y^{12}\right] < \infty$ and $\mathbb{E}\left[h^{12}(W)\right] < \infty$.

Regarding (ii) and (iii): recalling the definitions of $\widetilde{\sigma}^2$ and L in (A.3.13), (A.3.14), we have

$$\begin{aligned}
\widetilde{\sigma}_0^2\widetilde{\sigma}^2 = \widetilde{\sigma}_0^2\|L(U_0)\|_2 &= \frac{1}{\mathbb{E}\left[h^2(W)\right]}\mathbb{E}\left[\left(U_{i1} - \frac{\mathbb{E}\left[Yh(W)\right]}{2\mathbb{E}\left[h^2(W)\right]}U_{i2}\right)^2\right] \\
&= \mathbb{E}\left[\left(U_{i1} - \frac{\mathbb{E}\left[Yh(W)\right]}{2}U_{i2}\right)^2\right] \qquad \text{(A.3.17)} \\
&= \mathbb{E}\left[\left(R_i - \mathbb{E}\left[Yh(W)\right] - \frac{\mathbb{E}\left[Yh(W)\right]}{2}(\mathrm{Var}\left(h(W_i) \mid Z_i\right) - 1)\right)^2\right] \\
&= \mathbb{E}\left[(A+B)^2\right], \qquad \text{(A.3.18)}
\end{aligned}$$

where the third equality holds since $\mathbb{E}\left[h^2(W)\right] = 1$ as assumed without loss of generality, the fourth one comes from (A.3.8), and the last one is by rearranging with A, B defined as:

$$\begin{aligned}
A &:= Y_i h(W_i) - \mathbb{E}\left[Y_i h(W_i) \mid Z_i\right], \qquad \text{(A.3.19)} \\
B &:= \mathbb{E}\left[Y_i h(W_i) \mid Z_i\right] - \mathbb{E}\left[Yh(W)\right] - \frac{\mathbb{E}\left[Yh(W)\right]}{2}(\mathrm{Var}\left(h(W_i) \mid Z_i\right) - 1). \qquad \text{(A.3.20)}
\end{aligned}$$

The above terms A, B have equivalent expressions as the terms A, B defined in the proof of Theorem 1.2.3 (see (A.1.22), (A.1.23)). Note $\mathbb{E}\left[(A+B)^2\right] > 0$, as proved over the course of derivations from (A.1.20) to the end of the proof of Theorem 1.2.3. Due to (A.3.18), we then have $\widetilde{\sigma}_0^2\widetilde{\sigma}^2$ in this proof is nonzero, thus finish showing (ii).

Now we will verify $\widetilde{\sigma} = 1$. According to (A.3.17), we equivalently write down

$$\begin{aligned}
\widetilde{\sigma}_0^2\widetilde{\sigma}^2 &= \mathbb{E}\left[\left(U_{i1} - \frac{\mathbb{E}\left[Yh(W)\right]}{2}U_{i2}\right)^2\right] \\
&= \mathbb{E}\left[\left(\left(1, -\frac{\mathbb{E}\left[Yh(W)\right]}{2}\right)(U_{i1}, U_{i2})^\top\right)^2\right] \\
&= \boldsymbol{a}^\top \Sigma_U \boldsymbol{a}, \qquad \text{(A.3.21)}
\end{aligned}$$

where $\boldsymbol{a}^\top := \left(1, -\frac{\mathbb{E}[Yh(W]}{2}\right)$ and Σ_U is the covariance matrix for the random vector U_i, which can be explicitly written as

$$\Sigma_U = \begin{pmatrix} \mathrm{Var}\left(Yh(W)\right) & \mathrm{Cov}\left(Yh(W), \mathrm{Var}\left(h(W) \mid Z\right)\right) \\ \mathrm{Cov}\left(Yh(W), \mathrm{Var}\left(h(W) \mid Z\right)\right) & \mathrm{Var}\left(\mathrm{Var}\left(h(W) \mid Z\right)\right) \end{pmatrix}.$$

We immediately have $\widetilde{\sigma}_0^2\widetilde{\sigma}^2 = \boldsymbol{a}^\top \Sigma_U \boldsymbol{a} = H_2(\mathbf{0}) = \widetilde{\sigma}_0^2$ due to the expression of $\widetilde{\sigma}_0^2$ in (A.3.12), (A.3.21) and $\mathbb{E}\left[h^2(W)\right] = 1$ as assumed; hence, $\widetilde{\sigma} = 1$.

Having verified (i), (ii) and (iii), we thus prove the Berry–Esseen-type bound in (A.3.5), which completes the proof for case (IV). Therefore, the asymptotic coverage validity with a rate of $1/\sqrt{n}$ for the lower confidence bounds produced by Algorithm 1 has been established. □

Theorem A.3.2

Proof of Theorem A.3.2. Similarly as in the proofs of Theorem 1.2.3 and Theorem A.3.1, we immediately have coverage validity when $\mu(X, Z) \in \mathscr{A}(Z)$. Otherwise, it suffices to show

$$\inf_{K>1} \mathbb{P}\left(\frac{\bar{R}}{\sqrt{\bar{V}}} - \frac{z_\alpha s}{\sqrt{n}} \leq f(\mu)\right) \geq 1 - \alpha - C/\sqrt{n} \qquad \text{(A.3.22)}$$

for some constant C, where the sample mean $(\bar{R}, \bar{V})$ and sample covariance matrix $\widehat{\Sigma}$ are defined the same way as in Algorithm 1 except that R_i, V_i are replaced by their Monte Carlo estimators R_i^K, V_i^K as defined below:

$$R_i^K = Y_i \left(\mu(X_i, Z_i) - \frac{1}{K} \sum_{k=1}^{K} \mu(X_i^{(k)}, Z_i) \right),$$
$$V_i^K = \frac{1}{K-1} \sum_{k=1}^{K} \left(\mu(X_i^{(k)}, Z_i) - \frac{1}{K} \sum_{k=1}^{K} \mu(X_i^{(k)}, Z_i) \right)^2,$$

Recall that the proof in Appendix A.1.1 considers 4 cases then deals with them separately. Essentially we can conduct similar analysis, but to avoid lengthy derivations, we focus on Case IV. Note we also make the extra assumption $\mathbb{E}\left[\mathrm{Var}\left(Y(\mu(X,Z) - \mathbb{E}\left[\mu(X,Z) \mid Z\right]) \mid Z\right)\right] > 0$ to simplify the proof.

In the proof of Theorem 1.2.5, we have the following asymptotic normality result:

$$\sqrt{n} \left(\frac{\frac{1}{n}\sum_{i=1}^{n} R_i^K}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} V_i^K}} - f(\mu) \right) \xrightarrow{d} \mathcal{N}\left(0, \widetilde{\sigma}_0^2\right).$$

To establish (A.3.22), we follow the proof strategy of Theorem A.3.1. Specifically, we apply the Berry–Esseen bound for nonlinear statistics (see Theorem A.3.6 in Appendix A.3.1).

Again we first introduce some notations for the following proof: let random vectors $\{U_i\}_{i=1}^n = \{(U_{i1}, U_{i2}, U_{i3}, U_{i4}, U_{i5})\}_{i=1}^n \overset{i.i.d.}{\sim} U_0 = (U_{01}, U_{02}, U_{03}, U_{04}, U_{05})$ to be

$$U_{i1} = R_i^K - \mathbb{E}\left[Yh(W)\right], \; U_{i2} = V_i^K - \mathbb{E}\left[h^2(W)\right], \tag{A.3.23}$$

$$U_{i3} = (R_i^K)^2 - \mathbb{E}\left[(R_i^K)^2\right], \; U_{i4} = (V_i^K)^2 - \mathbb{E}\left[(V_i^K)^2\right], \; U_{i5} = R_i^K V_i^K - \mathbb{E}\left[R_i^K V_i^K\right].$$

Note by the construction of the null samples, $X_i^{(k)}$ satisfy the two properties in (A.1.45) and (A.1.46) and we have (A.1.47), (A.1.48) hold. Recall (A.1.44) in the proof of Theorem 1.2.5 states $\mathbb{E}\left[R_i^K\right] = \mathbb{E}\left[Yh(W)\right], \mathbb{E}\left[V_i^K\right] = \mathbb{E}\left[h^2(W)\right]$, hence $\mathbb{E}\left[U_{i1}\right] = \mathbb{E}\left[U_{i2}\right] = 0$. Straightforwardly, $\mathbb{E}\left[U_{i3}\right] = \mathbb{E}\left[U_{i4}\right] = \mathbb{E}\left[U_{i5}\right] = 0$. Thus we have $\mathbb{E}\left[U_0\right] = \mathbf{0}$. Now we denote $\bar{U} = (\bar{U}_1, \bar{U}_2, \bar{U}_3, \bar{U}_4, \bar{U}_5) = \frac{1}{n}\sum_{i=1}^n U_i$ and rewrite the following expression,

$$\frac{1}{s}\left(\frac{\frac{1}{n}\sum_{i=1}^n R_i^K}{\sqrt{\frac{1}{n}\sum_{i=1}^n V_i^K}} - f(\mu)\right) := H(\bar{U}) := \frac{H_1(\bar{U}_1, \bar{U}_2)}{\sqrt{H_2(\bar{U})}},$$

where s is similarly defined as in Algorithm 1 except that R_i, V_i are replaced by R_i^K, V_i^K. Here $H(x) = H(x_1, x_2, x_3, x_4, x_5) : \mathbb{R}^5 \to \mathbb{R}$ is the same as in the proof of Theorem A.3.1. Therefore the smoothness condition, i.e., Condition (A.3.5), holds by the same argument as in Appendix A.3.1. The continuous linear functional L is also defined the same way. To apply Theorem A.3.6, it remains to verify the following moment bound conditions on U_0 and $L(U_0)$,

$$\begin{aligned} 0 < \widetilde{\sigma} := \|L(U_0)\|_2 < \infty, \\ \nu_2 := \|U_0\|_2,\ \nu_3 := \|U_0\|_3 < \infty, \\ \varsigma_3 := \frac{\|L(U_0)\|_3}{\widetilde{\sigma}} < \infty. \end{aligned}$$

Note that $\nu_3^3 = \|U_0\|_3^3 = \mathbb{E}\left[|U_{01}|^3\right] + \mathbb{E}\left[|U_{02}|^3\right] + \mathbb{E}\left[|U_{03}|^3\right] + \mathbb{E}\left[|U_{04}|^3\right] + \mathbb{E}\left[|U_{05}|^3\right]$ and we can bound $(\varsigma_3\widetilde{\sigma})^3$ similarly as in the proof of Theorem A.3.1:

$$\begin{aligned} (\varsigma_3\widetilde{\sigma})^3 &= \mathbb{E}\left[|L(U_0)|^3\right] = \mathbb{E}\left[\left|\frac{1}{\sqrt{\mathbb{E}\left[h^2(W)\right]}}U_{01} - \frac{\mathbb{E}\left[Yh(W)\right]}{2(\sqrt{\mathbb{E}\left[h^2(W)\right]})^3}U_{02}\right|^3\right] \\ &\leq 2^{3-1}\left(A\frac{1}{(\sqrt{\mathbb{E}\left[h^2(W)\right]})^3}\mathbb{E}\left[|U_{01}|^3\right] + \frac{(\mathbb{E}\left[Yh(W)\right])^3}{8(\sqrt{\mathbb{E}\left[h^2(W)\right]})^9}\mathbb{E}\left[|U_{02}|^3\right]\right), \quad \text{(A.3.24)} \end{aligned}$$

Due to the fact that the finiteness of higher moments implies that of lower moments and (A.3.24), we only need to show

(i) $\mathbb{E}\left[|U_{01}|^3\right], \mathbb{E}\left[|U_{02}|^3\right], \mathbb{E}\left[|U_{03}|^3\right], \mathbb{E}\left[|U_{04}|^3\right], \mathbb{E}\left[|U_{05}|^3\right] < \infty$

(ii) $\widetilde{\sigma}_0^2 = H_2(\mathbf{0}) > 0$

(iii) $\widetilde{\sigma}^2 = \|L(U_0)\|_2 > 0$

under the stated moment conditions. For (iii), we have $\widetilde{\sigma}^2 = 1$, due to the derivations in the proof of Theorem A.3.1. Hence we will focus on the first two conditions in the following. Appendix A.3.1 verifies (i) and (ii) for any given $K > 1$. In this proof, we will actually show

$$\sup_{K>1} \mathbb{E}\left[|U_{0j}|^3\right] < \infty, \ \forall j \in [5], \quad \inf_{K>1} \widetilde{\sigma}_0^2 > 0.$$

Note the definitions of $U_0 = (U_{01}, U_{02}, U_{03}, U_{04}, U_{05})$ and $\widetilde{\sigma}_0^2$ depend on K. To simplify notations, we do not make this dependence explicit. By the definitions in (A.3.23), we bound U_{01}, U_{02} as below:

$$\begin{aligned}
\mathbb{E}\left[|U_{01}|^3\right] = \mathbb{E}\left[|U_{i1}|^3\right] &= \mathbb{E}\left[|R_i^K - \mathbb{E}\left[Yh(W)\right]|^3\right] \\
&\leq 2^{3-1}\left(\mathbb{E}\left[|R_i^K|^3\right] + \left(\mathbb{E}\left[Yh(W)\right]\right)^3\right), \\
\mathbb{E}\left[|U_{02}|^3\right] = \mathbb{E}\left[|U_{i2}|^3\right] &= \mathbb{E}\left[|V_i^K - \mathbb{E}\left[h^2(W)\right]|^3\right] \\
&\leq 2^{3-1}\left(\mathbb{E}\left[|V_i^K|^3\right] + \left(\mathbb{E}\left[h^2(W)\right]\right)^3\right),
\end{aligned}$$

where the inequalities hold due to the C_r inequality. Recalling in the proof of Theorem 1.2.5, we show $\mathbb{E}\left[|R_i^K|^2\right] < \infty, \mathbb{E}\left[|V_i^K|^2\right] < \infty$ under the condition $\mathbb{E}\left[Y^4\right], \mathbb{E}\left[h^4(W)\right] < \infty$ over the course of derivations from (A.1.50) to the end of that proof. The derivations are mainly based on the C_r inequality and the Bahr–Esseen inequality in Dharmadhikari et al. (1969). Us-

ing the same bounding strategy, we can show $\mathbb{E}\left[|R_i^K|^3\right], \mathbb{E}\left[|V_i^K|^3\right] < \infty$ when assuming $\mathbb{E}\left[Y^6\right], \mathbb{E}\left[h^6(W)\right] < \infty$. Hence we obtain $\sup_{K>1}\mathbb{E}\left[|U_{01}|^3\right], \sup_{K>1}\mathbb{E}\left[|U_{02}|^3\right] < \infty$ under the above moment conditions. And nearly identical derivations as in bounding $\mathbb{E}\left[|U_{01}|^3\right]$ and $\mathbb{E}\left[|U_{02}|^3\right]$ suffice to show $\sup_{K>1}\mathbb{E}\left[|U_{03}|^3\right], \sup_{K>1}\mathbb{E}\left[|U_{04}|^3\right], \sup_{K>1}\mathbb{E}\left[|U_{05}|^3\right] < \infty$ under the stronger moment boundedness conditions $\mathbb{E}\left[Y^{12}\right] < \infty, \mathbb{E}\left[h^{12}(W)\right] < \infty$ stated in Theorem 1.2.5.

Regarding (ii), we notice that

$$\widetilde{\sigma}_0^2 \geq \mathbb{E}\left[(A+B)^2\right] \geq \mathbb{E}\left[\mathrm{Var}\left(Yh(W) \mid Z\right)\right], \tag{A.3.25}$$

where the first inequality holds due to (A.1.62), A, B are defined as (A.1.22) and (A.1.23) in the proof of Theorem 1.2.3, and the second inequality holds by (A.1.24). The above lower bound for $\widetilde{\sigma}_0^2$ does not depend on K and implies the positiveness of $\inf_{K>1}\widetilde{\sigma}_0$ under the assumed condition $\mathbb{E}\left[\mathrm{Var}\left(Yh(W) \mid Z\right)\right] = \mathbb{E}\left[\mathrm{Var}\left(Y(\mu(X,Z) - \mathbb{E}\left[\mu(X,Z) \mid Z\right]) \mid Z\right)\right] > 0$.

Therefore, we obtain the Berry–Esseen bound for nonlinear statistics by applying Theorem A.3.6. Finally we conclude the asymptotic coverage with a rate of $n^{-1/2}$, i.e.,

$$\inf_{K>1}\mathbb{P}\left(L_{n,K}^{\alpha}(\mu) \leq \mathcal{I}\right) \geq 1 - \alpha - Cn^{-1/2},$$

where the constant C only depends on the moments of Y and $h(X,Z) = \mu(X,Z) - \mathbb{E}\left[\mu(X,Z) \mid Z\right]$.

□

A.4 Applicability of the Model-X assumption

Model-X floodgate assumes knowing the distribution of $P_{X|Z}$. This may not always hold in practice, but in some important instances, $P_{X|Z}$ may be (A) known due to experimental randomization,

(B) well-modeled a priori due to domain expertise, or (C) accurately estimated from a large unlabeled data set. For example, (A) holds in the high-dimensional experiments of conjoint analysis (Luce & Tukey, 1964; Hainmueller & Hopkins, 2014), (B) holds in the study of the microbiome where accurate covariate simulators exist (Ren et al., 2016), and a combination of (B) and (C) hold in genomics, where the model-X framework has been repeatedly and successfully applied for controlled variable selection (Sesia et al., 2019; Katsevich & Sabatti, 2019; Sesia et al., 2020b; Bates et al., 2020; Sesia et al., 2020a).

We also quantify the robustness of our inferences to this assumption in Appendix A.5 and show it can be relaxed to parametric models (Section 1.3.2), and indeed model-X approaches have shown promising empirical performance in a number of applications in which it is unclear whether any of (A), (B), or (C) hold, such as bacterial classification from spectroscopic data (Chia et al., 2020) and single cell regulatory screening (Katsevich & Roeder, 2020).

A.5 Robustness

To explain how the floodgate idea is not tied to the model-X assumption, a double-robustness type result (Lemma 2.3) is presented in Remark 1.2.3.1. It involves an approximated floodgate functional (1.2.8) and says that the inferential statements are valid as long as either of the models of $X \mid Z$ or $Y \mid Z$ is correctly specified. For ease of exposition, Algorithm 1.2.3 and Theorem 1.2.3 focus on a particular floodgate procedure which requires knowing $P_{X|Z}$. However, it is still of interest to study the robustness of floodgate (in Algorithm 1.2.3) to misspecification of $P_{X|Z}$. Specifically, we consider the case when the true distribution $P_{X|Z}$ used in floodgate is replaced by an approximation $Q_{X|Z}$.

Notationally, let $Q = P_{Y|X,Z} \times Q_{X|Z} \times P_Z$ (we need not consider misspecification in the distributions of Z or $Y \mid X, Z$ since these are not inputs to floodgate), and let f^Q be an analogue

of f with certain expectations replaced by expectations over Q (we will denote such expectations by $\mathbb{E}_Q[\cdot]$); see Equation (A.5.5) for a formal definition. It is not hard to see that floodgate with input $Q_{X|Z}$ produces an asymptotically-valid LCB for $f^Q(\mu)$, from which we immediately draw the following conclusions.

First, if μ does not actually depend on X, i.e., $\mathrm{Var}_Q(\mu(X,Z) \mid Z) \overset{a.s.}{=} 0$, then $f^Q(\mu) = 0$ regardless of Q and floodgate is trivially asymptotically-valid. Second, when μ does depend on X, floodgate's inference will still be approximately valid as long as $f^Q(\mu) - f(\mu) \approx 0$, and this difference can be bounded by, for instance, the χ^2 divergence between $P_{X|Z}$ and $Q_{X|Z}$. The third, and perhaps most interesting, conclusion is that the gap between $\mathcal{I}$ and $f(\mu)$ grants floodgate an *extra* layer of robustness as long as $\mathcal{I} - f(\mu)$ is large compared to $f^Q(\mu) - f(\mu)$. Thus even if $Q_{X|Z}$ is a bad approximation of $P_{X|Z}$, floodgate's inference may be saved if $f(\mu)$ is an *even worse* approximation of $\mathcal{I}$, and this latter approximation is related to that of μ for $\mu^\star$. To make this last relation precise, we quantify μ's approximation of $\mu^\star$ by focusing on a particular representative of S_μ: for any $\mu : \mathbb{R}^p \to \mathbb{R}$,

$$\bar{\mu}(x,z) = \sqrt{\frac{\mathbb{E}\left[\mathrm{Var}(\mu^\star(X,Z) \mid Z)\right]}{\mathbb{E}\left[\mathrm{Var}(\mu(X,Z) \mid Z)\right]}}\Big(\mu(x,z) - \mathbb{E}\left[\mu(X,Z) \mid Z = z\right]\Big) + \mathbb{E}\left[\mu^\star(X,Z) \mid Z = z\right], \tag{A.5.1}$$

where $0/0 = 0$. We can think of $\bar{\mu}$ as a generally accurate representative from S_μ, in that it takes μ and corrects its conditional mean and expected conditional variance to match $\mu^\star$. Note that $\bar{\mu} = \mu^\star$ whenever $\mu^\star \in S_\mu$, which includes anytime $\mathcal{I} = 0$. Since the LCB from floodgate with input $Q_{X|Z}$ is asymptotically-valid for $f^Q(\mu)$ under certain moment conditions and the proof can be done similarly as Theorem 1.2.3, we will focus on quantifying the difference between $f^Q(\mu)$ and $\mathcal{I}$ in the following robustness result.

Theorem A.5.1 (Floodgate robustness). *For data $\{(Y_i, X_i, Z_i)\}_{i=1}^n$ i.i.d. draws from P satisfying $\mathbb{E}[Y^4] < \infty$, a sequence of working regression functions $\mu_n : \mathbb{R}^p \to \mathbb{R}$ such that for some C and all*

n either $\mathrm{Var}_{Q^{(n)}}\left(\mu_n(X,Z) \mid Z\right) \overset{a.s.}{=} 0$ *or* $\frac{\max\left\{\mathbb{E}\left[\mu_n^4(X,Z)\right], \mathbb{E}_{Q^{(n)}}\left[\mu_n^4(X,Z)\right]\right\}}{\mathbb{E}\left[\mathrm{Var}_{Q^{(n)}}\left(\mu_n(X,Z) \mid Z\right)\right]^2} \leq C$*, and a sequence of conditional distributions* $Q_{X|Z}^{(n)}$*, the difference between* $f^{Q^{(n)}}(\mu)$ *and* $\mathcal{I}$ *can be controlled as*

$$\Delta_n = f^{Q^{(n)}}(\mu_n) - \mathcal{I} \leq c_1 \sqrt{\mathbb{E}\left[\chi^2\left(P_{X|Z} \,\|\, Q_{X|Z}^{(n)}\right)\right]} - c_2\, \mathbb{E}\left[(\bar{\mu}_n(X,Z) - \mu^\star(X,Z))^2\right] \tag{A.5.2}$$

for some positive c_1 *and* c_2 *that depend on* P*, where* $\chi^2(\cdot \,\|\, \cdot)$ *denotes the* χ^2 *divergence.*

The proof of Theorem A.5.1 can be found in Appendix A.5.1. Equation (A.5.2) formalizes that larger MSE of $\bar{\mu}_n$ actually *improves* robustness, although we remind the reader once again that when $\mathcal{I} = 0$, the MSE of $\bar{\mu}_n$ is always zero by construction in Equation (A.5.1). Given the $n^{-1/2}$-rate half-width lower-bound for floodgate, a sufficient condition for asymptotically-exact coverage is

$$\sqrt{\mathbb{E}\left[\chi^2\left(P_{X|Z} \,\|\, Q_{X|Z}^{(n)}\right)\right]} = o\left(n^{-1/2} + \mathbb{E}\left[(\bar{\mu}_n(X,Z) - \mu^\star(X,Z))^2\right]\right). \tag{A.5.3}$$

When $Q_{X|Z}^{(n)}$ is a standard parametric estimator based on N_n independent samples, the left-hand side has a $O(N_n^{-1/2})$ rate. Thus if $N_n \gg \min\{n, \mathbb{E}\left[(\bar{\mu}_n(X,Z) - \mu^\star(X,Z))^2\right]^{-2}\}$, then floodgate's coverage will be asymptotically-exact. For certain parametric models for $X \mid Z$, Section 1.3.2 shows how to modify floodgate to attain asymptotically-exact inference without the need for estimation at all.

Theorem A.5.1 treats the sequence $Q_{X|Z}^{(n)}$ as fixed, which of course means $Q_{X|Z}^{(n)}$ can be estimated from any data that is independent of the data floodgate is applied to. This means the same data can be used to estimate μ_n and $Q_{X|Z}^{(n)}$. For $Q_{X|Z}^{(n)}$ however, this strict separation may not be necessary in practice, and in our simulations we found floodgate to be quite robust to estimating $Q_{X|Z}^{(n)}$ on samples that included those used as input to floodgate; see Section 1.4.5.

Another layer of robustness beyond that addressed in this section can be injected by replacing

$P_{X|Z}$ in floodgate with $P_{X|Z,T}$ for some random variable T. For instance, floodgate's model-X assumption can be formally relaxed to only needing to know a fixed-dimensional model for $P_{X|Z}$ by conditioning on T that is a sufficient statistic for that model; see Section 1.3.2 for details. More generally, conditioning on T that is a function of $\{(X, Z)\}_{i=1}^n$ may induce some degree of robustness, as conditioning on the order statistics of the X_i can in conditional independence testing (Berrett et al., 2020).

A.5.1 Proofs in Appendix A.5

In the case where the conditional distribution of X given Z is specified as $Q_{X|Z}$ (in the following, we often denote the true conditional distribution by $P := P_{X|Z}$ and the specified conditional distribution by $Q := Q_{X|Z}$ without causing confusion), the floodgate functional with input $Q_{X|Z}$ is denoted by $f^Q(\mu)$. Note that $f(\mu)$ can be rewritten with explicit subscripts as below (here we use the equivalent expression of $f(\mu)$ in (A.1.7) and expand $h(W)$).

$$f(\mu) = \frac{\mathbb{E}_P\left[Y\left(\mu(X,Z) - \mathbb{E}_P\left[\mu(X,Z) \mid Z\right]\right)\right]}{\sqrt{\mathbb{E}_{P_Z}\left[\mathrm{Var}_P\left(\mu(X,Z) \mid Z\right)\right]}} \tag{A.5.4}$$

Therefore, $f^Q(\mu)$ admits the following expression:

$$f^Q(\mu) := \frac{\mathbb{E}_P\left[Y\left(\mu(X,Z) - \mathbb{E}_Q\left[\mu(X,Z) \mid Z\right]\right)\right]}{\sqrt{\mathbb{E}_{P_Z}\left[\mathrm{Var}_Q\left(\mu(X,Z) \mid Z\right)\right]}}. \tag{A.5.5}$$

Denote $\omega(x,z) := \frac{dP_{X|Z}(x|z)}{dQ_{X|Z}(x|z)}$. Note that $\omega(x,z)$ is the ratio of conditional densities if we are in the continuous case; $\omega(x,z)$ is the ratio of conditional probability mass function in discrete case. Then we can quantify the difference between $f(\mu)$ and $f^Q(\mu)$ as in Lemma A.5.2.

Lemma A.5.2. *Assuming $\mathbb{E}\left[Y^4\right] < \infty$, consider two joint distributions P, Q over (X, Z), defined as $P(x,z) = P_{X|Z}(x|z)P_Z(z), Q(x,z) = Q_{X|Z}(x|z)P_Z(z)$. If we denote $\mathcal{U}$ to be the class of*

functions $\mu : \mathbb{R}^p \to \mathbb{R}$ *satisfying one of the following conditions:*

- $\mu(X, Z) \in \mathscr{A}(Z)$*;*
- $\max\{\mathbb{E}_P\left[\mu^4(X,Z)\right], \mathbb{E}_Q\left[\mu^4(X,Z)\right]\}/(\mathbb{E}_{P_Z}\left[\mathrm{Var}_Q\left(\mu(X,Z) \mid Z\right)\right])^2 \leq c_0$.

for some constants c_0*, then we have the following bounds*

$$\Delta(P, Q) := \sup_{\mu \in \mathcal{U}} |\theta^Q(\mu) - f(\mu)| \leq C\sqrt{\mathbb{E}_{P_Z}\left[\chi^2\left(P_{X|Z} \| Q_{X|Z}\right)\right]} \tag{A.5.6}$$

for some constant C *only depending on* $\mathbb{E}\left[Y^4\right]$ *and* c_0*, where the* χ^2 *divergence between two distributions* P, Q *on the probability space* Ω *is defined as* $\chi^2\left(P \| Q\right) := \int_\Omega (\frac{dP}{dQ} - 1)^2 dQ$.

When the $X \mid Z$ model is misspecified, the inferential validity will not hold in general, without adjustment on the lower confidence bound. Lemma A.5.2 gives a quantitative characterization about how much we need to adjust.

Proof of Lemma A.5.2. When the support of Q does not contain the support of P, the χ^2 divergence between P and Q is infinite, which immediately proves (A.5.6). From now, we work with the case where the support of Q contains the support of P. When $\mu(X, Z) \in \mathscr{A}(Z)$, $f(\mu) = f^Q(\mu) = 0$, thus the statement holds. Now we deal with the nontrivial case where $\mathbb{E}_{P_Z}\left[\mathrm{Var}_Q\left(\mu(X,Z) \mid Z\right)\right] > 0$. Without loss of generality, we assume $\mathbb{E}_{P_Z}\left[\mathrm{Var}_Q\left(\mu(X,Z) \mid Z\right)\right] = 1$ for the following proof (since floodgate is invariate to positive scaling of μ). Then the stated moment conditions on μ imply

$$\mathbb{E}_P\left[\mu^4(X,Z)\right], \mathbb{E}_Q\left[\mu^4(X,Z)\right] \leq c_0. \tag{A.5.7}$$

First we simplify $f(\mu)$ and $f^Q(\mu)$ into

$$\begin{aligned} f(\mu) &= \frac{\mathbb{E}_P\left[\mu^\star(X,Z)\left(\mu(X,Z)-\mathbb{E}_{P_{X|Z}}\left[\mu(X,Z)\,|\,Z\right]\right)\right]}{\sqrt{\mathbb{E}_{P_Z}\left[\mathrm{Var}_{P_{X|Z}}\left(\mu(X,Z)\,|\,Z\right)\right]}} = \frac{\mathbb{E}_P\left[\mu^\star(W)\left(\mu(W)-\mathbb{E}_P\left[\mu(W)\,|\,Z\right]\right)\right]}{\sqrt{\mathbb{E}_{P_Z}\left[\mathrm{Var}_P\left(\mu(W)\,|\,Z\right)\right]}} \\ f^Q(\mu) &= \frac{\mathbb{E}_P\left[\mu^\star(X,Z)\left(\mu(X,Z)-\mathbb{E}_{Q_{X|Z}}\left[\mu(X,Z)\,|\,Z\right]\right)\right]}{\sqrt{\mathbb{E}_{P_Z}\left[\mathrm{Var}_{Q_{X|Z}}\left(\mu(X,Z)\,|\,Z\right)\right]}} = \frac{\mathbb{E}_P\left[\mu^\star(W)\left(\mu(W)-\mathbb{E}_Q\left[\mu(W)\,|\,Z\right]\right)\right]}{\sqrt{\mathbb{E}_{P_Z}\left[\mathrm{Var}_Q\left(\mu(W)\,|\,Z\right)\right]}} \end{aligned}$$

due to (A.1.4), where we denote $W=(X,Z)$ (thus $w=(x,z)$). Noticing the following facts

$$\left|\frac{a}{\sqrt{b}}-\frac{c}{\sqrt{d}}\right| = \left|\frac{a\sqrt{d}-c\sqrt{b}}{\sqrt{bd}}\right| \le \frac{a}{\sqrt{bd}}\left|\sqrt{b}-\sqrt{d}\right| + \frac{1}{\sqrt{d}}\left|a-c\right| \le \frac{a}{\sqrt{b}}\cdot\frac{1}{d}\left|b-d\right| + \frac{1}{\sqrt{d}}\left|a-c\right|,$$

we let a, c to be the numerators of $f(\mu)$ and $f^Q(\mu)$ respectively and $\sqrt{b}$, $\sqrt{d}$ to be their denominators. Before dealing with $|b-d|$ and $|c-d|$, we have the following bounds on the terms $a/\sqrt{b}$ and $1/d$.

$$a/\sqrt{b} = f(\mu) \le \mathcal{I} \le \left(\mathbb{E}_P\left[Y^4\right]\right)^{1/4},\ \ 1/d = 1/\mathbb{E}_{P_Z}\left[\mathrm{Var}_Q\left(\mu(X,Z)\,|\,Z\right)\right] = 1, \qquad \text{(A.5.8)}$$

where the first equality is by Lemma 1.2.2 and the second one is by applying Jensen's inequality ($\mathbb{E}_{P_Z}\left[\mathrm{Var}_P\left(\mathbb{E}\left[Y\,|\,X,Z\right]|\,Z\right)\right] \le \mathbb{E}_{P_Z}\left[\mathbb{E}_P\left[\left(\mathbb{E}\left[Y\,|\,X,Z\right]\right)^2|\,Z\right]\right] \le \mathbb{E}\left[Y^2\right] \le \sqrt{\mathbb{E}\left[Y^4\right]}$). The equality holds by assumption. Now it suffices to consider bounding $|b-d|$ and $|c-d|$ in terms of the expected χ^2 divergence between $P_{X|Z}$ and $Q_{X|Z}$. We have the following equations for $|a-c|$:

$$\begin{aligned} |a-c| &= \left|\mathbb{E}_P\left[\mu^\star(W)\left(\mu(W)-\mathbb{E}_P\left[\mu(W)\,|\,Z\right]\right)\right]-\mathbb{E}_P\left[\mu^\star(W)\left(\mu(W)-\mathbb{E}_Q\left[\mu(W)\,|\,Z\right]\right)\right]\right| \\ &= \left|\mathbb{E}_P\left[\mu^\star(W)\left(\mathbb{E}_P\left[\mu(W)\,|\,Z\right]-\mathbb{E}_Q\left[\mu(W)\,|\,Z\right]\right)\right]\right| \\ &= \left|\mathbb{E}_{P_Z}\left[\mathbb{E}_P\left[\mu^\star(W)\,|\,Z\right]\left(\mathbb{E}_P\left[\mu(W)\,|\,Z\right]-\mathbb{E}_Q\left[\mu(W)\,|\,Z\right]\right)\right]\right|. \qquad \text{(A.5.9)} \end{aligned}$$

Now we rewrite $|\mathbb{E}_P[\mu(W) \mid Z] - \mathbb{E}_Q[\mu(W) \mid Z]|$ in the form of integral then bound it as

$$\begin{aligned}
|\mathbb{E}_P[\mu(W) \mid Z] - \mathbb{E}_Q[\mu(W) \mid Z]| &= \left| \int \mu(x, Z)(1 - \omega(x, Z)) dQ_{X|Z}(x \mid Z) \right| \\
&\leq \sqrt{\mathbb{E}_{Q_{X|Z}}[\mu^2(X, Z) \mid Z]} \sqrt{\int (1 - w(x, Z))^2 dQ_{X|Z}(x \mid Z)} \\
&= \sqrt{\mathbb{E}_{Q_{X|Z}}[\mu^2(W) \mid Z]} \sqrt{\chi^2\left(P_{X|Z} \| Q_{X|Z}\right)}, \qquad \text{(A.5.10)}
\end{aligned}$$

where $\omega(x, Z) = \frac{dP_{X|Z}(x|Z)}{dQ_{X|Z}(x|Z)}$ and the above inequality is from the Cauchy–Schwarz inequality. Hence we can plug (A.5.10) into (A.5.9) and further bound $|a - c|$ by

$$\begin{aligned}
|a - c| &\leq \mathbb{E}_{P_Z}\left[\mathbb{E}_{P_{X|Z}}[\mu^\star(W) \mid Z] \sqrt{\mathbb{E}_{Q_{X|Z}}[\mu^2(W) \mid Z]} \sqrt{\chi^2\left(P_{X|Z} \| Q_{X|Z}\right)}\right] \\
&\leq \sqrt{\mathbb{E}_{P_Z}\left[(\mathbb{E}_{P_{X|Z}}[\mu^\star(W) \mid Z])^2 \mathbb{E}_{Q_{X|Z}}[\mu^2(W) \mid Z]\right]} \cdot \sqrt{\mathbb{E}_{P_Z}\left[\chi^2\left(P_{X|Z} \| Q_{X|Z}\right)\right]}.
\end{aligned} \qquad \text{(A.5.11)}$$

For the first part of the product in (A.5.11), we can apply the Cauchy–Schwarz inequality and Jensen's inequality and bound it by $(\mathbb{E}_P\left[(\mu^\star)^4(W)\right] \mathbb{E}_Q\left[\mu^4(W)\right])^{1/4}$, which is upper bounded by some constant under the stated condition $\mathbb{E}\left[Y^4\right] < \infty$ and $\mathbb{E}_Q\left[\mu^4(X, Z)\right] \leq c_0$ (from (A.5.7)). Regarding $|b - d|$, we have

$$\begin{aligned}
|b - d| &= |\mathbb{E}_{P_Z}[\mathrm{Var}_P(\mu(W) \mid Z)] - \mathbb{E}_{P_Z}[\mathrm{Var}_Q(\mu(X, Z) \mid Z)]| \\
&\leq \left|\mathbb{E}_{P_Z}\left[(\mathbb{E}_P[\mu(W) \mid Z])^2 - (\mathbb{E}_Q[\mu(W) \mid Z])^2\right]\right| \\
&\quad + \left|\mathbb{E}_{P_Z}\left[\mathbb{E}_P\left[\mu^2(W) \mid Z\right] - \mathbb{E}_Q\left[\mu^2(W) \mid Z\right]\right]\right|. \qquad \text{(A.5.12)}
\end{aligned}$$

Similarly as (A.5.10), we obtain

$$\left|\mathbb{E}_P\left[\mu^2(W) \mid Z\right] - \mathbb{E}_Q\left[\mu^2(W) \mid Z\right]\right| \leq \sqrt{\mathbb{E}_{Q_{X|Z}}[\mu^4(W) \mid Z]} \sqrt{\chi^2\left(P_{X|Z} \| Q_{X|Z}\right)}.$$

Then under the moment bounds $\mathbb{E}_Q\left[\mu^4(X,Z)\right] \leq c_0$ in (A.5.7), we show the second term in (A.5.12) is upper bounded by $\sqrt{c_0 \mathbb{E}_{P_Z}\left[\chi^2\left(P_{X|Z} \| Q_{X|Z}\right)\right]}$. Regarding the first term in (A.5.12), we can write

$$\left(\mathbb{E}_P\left[\mu(W) \mid Z\right]\right)^2 - \left(\mathbb{E}_Q\left[\mu(W) \mid Z\right]\right)^2 = \left(\mathbb{E}_P\left[\mu(W) \mid Z\right] - \mathbb{E}_Q\left[\mu(W) \mid Z\right]\right)\left(\mathbb{E}_P\left[\mu(W) \mid Z\right] + \mathbb{E}_Q\left[\mu(W) \mid Z\right]\right)$$

then apply similar strategies in deriving (A.5.9) and (A.5.11) to control the above term under a bound $C\sqrt{\mathbb{E}_{P_Z}\left[\chi^2\left(P_{X|Z} \| Q_{X|Z}\right)\right]}$ for some constant C. And this will make use of the moment bound conditions $\mathbb{E}_P\left[\mu^4(X,Z)\right]$,$\mathbb{E}_Q\left[\mu^4(X,Z)\right] \leq c_0$ in (A.5.7). Finally we establish (A.5.6). □

Proof of Theorem A.5.1. First notice that Δ_n can be decomposed into two parts:

$$\Delta_n = f^{Q^{(n)}}(\mu_n) - \mathcal{I} = (f^{Q^{(n)}}(\mu_n) - f(\mu_n)) - (\mathcal{I} - f(\mu_n)). \tag{A.5.13}$$

In the following, we will deal with $f^{Q^{(n)}}(\mu_n) - f(\mu_n)$ and $\mathcal{I} - f(\mu_n)$ separately. Applying Lemma A.5.2 to P, $Q^{(n)}$ and μ_n under the stated conditions gives

$$(f^{Q^{(n)}}(\mu_n) - f(\mu_n)) \leq c_1 \sqrt{\mathbb{E}\left[\chi^2\left(P_{X|Z} \,\|\, Q^{(n)}_{X|Z}\right)\right]} \tag{A.5.14}$$

for some constant c_1 only depending on $\mathbb{E}\left[Y^4\right]$ and c_0. Regarding the term $\mathcal{I} - f(\mu_n)$, we recall the derivations in the proof of Theorem 1.2.6, specifically (A.1.69) and (A.1.70), then obtain

$$\mathcal{I} - f(\mu_n) \geq \frac{\mathbb{E}\left[(\bar{h}_n(W) - h^\star(W))^2\right]}{2\mathcal{I}} = \frac{\mathbb{E}\left[(\bar{\mu}_n(W) - \mu^\star(W))^2\right]}{2\mathcal{I}}, \tag{A.5.15}$$

where the equality holds by the definition of $h^\star$, $\bar{\mu}_n$ and $\bar{h}_n$. Combining (A.5.13), (A.5.14) and (A.5.15) yields (A.5.2). □

A.6 Details of extending the mMSE gap

A.6.1 Taking the supremum over transformations

Drawing inspiration from the maximum correlation coefficient (Hirschfeld, 1935), taking the supremum of the mMSE gap over transformations of Y leads to other desirable properties. For a set $\mathcal{G}$ of functions g mapping Y to its sample space, let $\mathcal{I}_{\mathcal{G}} = \sup_{g \in \mathcal{G}} \mathcal{I}_{\text{sf}}(g(Y))$, where $\mathcal{I}_{\text{sf}}(g(Y))$ denotes the scale-free version of the mMSE gap when Y is replaced by $g(Y)$. Then for any fixed function $g \in \mathcal{G}$, floodgate's LCB for $\mathcal{I}_{\text{sf}}(g(Y))$ is also an asymptotically valid LCB for $\mathcal{I}_{\mathcal{G}}$. And like μ, g can be chosen based on an independent split of the data to make the gap between $\mathcal{I}_{\text{sf}}(g(Y))$ and $\mathcal{I}_{\mathcal{G}}$ as small as possible. If $\mathcal{G}$ forms a group, then it is immediate that $\mathcal{I}_{\mathcal{G}}$ takes the same value when $g(Y)$ is used as the response, for any $g \in \mathcal{G}$, i.e., $\mathcal{I}_{\mathcal{G}}$ is invariant to any transformation $g \in \mathcal{G}$ of Y. For instance, we might choose $\mathcal{G}$ to be the group of all strictly monotone functions, or of all bijections. Regardless of whether $\mathcal{G}$ is a group or not, if it is large enough that it contains all bounded continuous functions then, by the Portmanteau Theorem, $\mathcal{I}_{\mathcal{G}}$ will be zero if *and only if* $Y \perp\!\!\!\perp X \mid Z$. That is, for sufficiently large $\mathcal{G}$, $\mathcal{I}_{\mathcal{G}}$ satisfies the key property of the MOVI in Azadkia & Chatterjee (2019) and floodgate provides asymptotically valid inference for it. A natural choice* of $\mathcal{G}$ satisfying such property is $\{\mathbb{1}_{\{y \le t\}} : t \in \mathbb{R}\}$ as

$$\mathcal{I}_{\mathcal{G}} = \sup_{t \in \mathbb{R}} \frac{\mathbb{E}\left[\text{Var}\left(\mathbb{E}\left[\mathbb{1}_{\{Y \le t\}} \mid X, Z\right] \mid Z\right)\right]}{\text{Var}\left(\mathbb{1}_{\{Y \le t\}}\right)}.$$

The above quantity is related to the measure of conditional dependence in Azadkia & Chatterjee (2019) as both involve $\mathbb{E}\left[\text{Var}\left(\mathbb{E}\left[\mathbb{1}_{\{Y \le t\}} \mid X, Z\right] \mid Z\right)\right]$.

*We are grateful to an anonymous reviewer for suggesting this choice.

A.6.2 Extending via the RKHS framework

A reviewer pointed out a very interesting work (Huang et al., 2020b) which came out after our arXiv preprint. To handle X, Y, Z from general topological spaces, Huang et al. (2020b) proposes the kernel partial correlation coefficient (KPC) to measure conditional dependence and provides consistent estimation methods. Huang et al. (2020b) mentioned the numerator of KPC with a linear kernel equals to the mMSE gap considered in our paper. In this section, we discuss how to extend the floodgate inferential approach via reproducing kernel Hilbert spaces (RKHS) to apply to the KPC. For ease of exposition, we focus on the numerator of KPC and call it the average kernel maximum mean discrepancy (AKMMD). Note that the AKMMD with a characteristic kernel will be zero if *and only if* $Y \perp\!\!\!\perp X \mid Z$.

Recall the equivalent expression of the mMSE gap in (1.2.4)

$$\mathcal{I}^2 = \mathbb{E}\left[\left(\mathbb{E}\left[Y \mid X, Z\right] - \mathbb{E}\left[Y \mid Z\right]\right)^2\right],$$

where $\mathbb{E}\left[Y \mid X, Z\right]$ can be viewed as the kernel embedding of $P_{Y|X,Z}$ under a special linear kernel. Then $\mathcal{I}^2$ essentially quantifies the distance between $P_{Y|X,Z}$ and $P_{Y|Z}$ via the maximum mean discrepancy (MMD). To extend this idea using a general kernel, we introduce some new notations and preliminary concepts about RKHS. Suppose (Y, X, Z) take values in some topological space $\mathcal{Y} \times \mathcal{X} \times \mathcal{Z}$ and let P be the joint distribution over (Y, X, Z). The marginal distribution of Y is denoted by P_Y. Sometimes this subscript is dropped when doing so does not cause confusion. We use the bold $\boldsymbol{\mu}$ notation for kernel mean embeddings, which should be differentiated from the working regression function in the main text. Denote by $\mathcal{H}_\mathcal{Y}$ an RKHS with kernel $\mathcal{K}(\cdot, \cdot)$ on the space $\mathcal{Y}$, where $\mathcal{K} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is a symmetric and positive semidefinite function such that $\mathcal{K}(\cdot, y)$ is measurable function on $\mathcal{Y}, \forall\, y \in \mathcal{Y}$. The inner product and norm on the RKHS $\mathcal{H}_\mathcal{Y}$ are denoted by $\langle \cdot, \cdot \rangle_{\mathcal{H}_\mathcal{Y}}$ and $|| \cdot ||_{\mathcal{H}_\mathcal{Y}}$, with the subscripts often dropped for simplicity. The kernel reproducing

property implies that $h(y) = \langle \mathcal{K}(\cdot, y), h \rangle_{\mathcal{H}_\mathcal{Y}}$. First we introduce the definitions of the kernel mean embedding and the MMD (Deb et al., 2020; Huang et al., 2020b).

Definition A.6.1 (Kernel mean embedding). *Suppose* $Y \sim P_Y$ *and* $\mathbb{E}_P\left[\sqrt{\mathcal{K}(Y,Y)}\right] < \infty$. *There exists (Deb et al., 2020; Huang et al., 2020b) a unique* $\boldsymbol{\mu}_P \in \mathcal{H}_\mathcal{Y}$ *satisfying*

$$\langle \boldsymbol{\mu}_P, h \rangle_{\mathcal{H}_\mathcal{Y}} = \mathbb{E}_P\left[h(Y)\right], \quad \text{for all } h \in \mathcal{H}_\mathcal{Y},$$

which is called the kernel mean embedding of P_Y *into* $\mathcal{H}_\mathcal{Y}$.

Definition A.6.2 (Maximum mean discrepancy). *We measure the distance between two distributions* P_1, P_2 *via the MMD (with respect to the kernel* $\mathcal{K}(\cdot,\cdot)$*), defined as*

$$\mathrm{MMD}(P_1, P_2) := ||\boldsymbol{\mu}_{P_1} - \boldsymbol{\mu}_{P_2}||_{\mathcal{H}_\mathcal{Y}}.$$

It also has the following equivalent representation (Deb et al., 2020; Huang et al., 2020b):

$$\mathrm{MMD}^2(P_1, P_2) := \mathbb{E}\left[\mathcal{K}(U, U')\right] + \mathbb{E}\left[\mathcal{K}(V, V')\right] - 2\mathbb{E}\left[\mathcal{K}(U, V)\right],$$

where $U, U' \overset{\text{i.i.d.}}{\sim} P_1$, $V, V' \overset{\text{i.i.d.}}{\sim} P_2$ *and* $U \perp\!\!\!\perp V$.

Now we are ready to define the AKMMD.

Definition A.6.3 (average kernel maximum mean discrepancy). *The* average kernel maximum mean discrepancy *for variable* X *is defined as*

$$\mathcal{I}^2_\mathcal{K} = \mathbb{E}\left[\mathrm{MMD}^2(P_{Y|X,Z}, P_{Y|Z})\right] \tag{A.6.1}$$

whenever all the above expectations exist.

We also present its alternative expression in terms of the kernel:

$$\mathcal{I}^2_{\mathcal{K}} = \mathbb{E}\left[\mathcal{K}(Y_2, \widetilde{Y}_2)\right] - \mathbb{E}\left[\mathcal{K}(Y_1, \widetilde{Y}_1)\right] = \mathbb{E}\left[\mathbb{E}\left[\mathcal{K}(Y_2, \widetilde{Y}_2) \,|\, X, Z\right]\right] - \mathbb{E}\left[\mathbb{E}\left[\mathcal{K}(Y_1, \widetilde{Y}_1) \,|\, Z\right]\right],$$

where $Y_1, \widetilde{Y}_1, Y_2, \widetilde{Y}_2$ are defined as below

$$Y_1 \mid X \sim P_{Y|Z}, \quad \widetilde{Y}_1 \mid X \sim P_{Y|Z}, \quad \text{and } Y_1 \perp\!\!\!\perp \widetilde{Y}_1 \mid X,$$
$$(X, Z) \sim P_{X,Z}, \quad Y_2 \mid X, Z \sim P_{Y|X,Z}, \quad \widetilde{Y}_2 \mid X, Z \sim P_{Y|X,Z}, \quad \text{and } Y_2 \perp\!\!\!\perp \widetilde{Y}_2 \mid X, Z.$$

The floodgate functional constitutes a deterministic lower bound for the mMSE gap for any working regression function μ. As we are now dealing with mean embeddings with a general kernel, we will replace the role of μ with $Q_{Y|X,Z}$, an estimate of the full conditional distribution of $Y \mid X, Z$ (as opposed to just its conditional mean). Let $Q = Q_{Y|X,Z} \times P_{X,Z}$ and the associated conditional distribution of Y given Z by $Q_{Y|Z}$. For notational simplicity, $Q_{Y|X,Z}$ and $Q_{Y|Z}$ are both sometimes abbreviated simply as Q. Given any non-random conditional distribution $Q_{Y|X,Z}$, we consider the kernel floodgate functional

$$f_{\mathcal{K}}(Q) := \frac{\mathbb{E}\left[\mathcal{K}(Y, Y_2^Q)\right] - \mathbb{E}\left[\mathcal{K}(Y, Y_1^Q)\right]}{\sqrt{\mathbb{E}\left[\mathcal{K}(Y_2^Q, \widetilde{Y}_2^Q)\right] - \mathbb{E}\left[\mathcal{K}(Y_2^Q, Y_1^Q)\right]}}, \tag{A.6.2}$$

where the involved random variables are defined through

$$\begin{aligned} (X, Z) \sim P_{X,Z}, \quad Y \mid X, Z \sim P_{Y|X,Z}, \quad Y \mid Z \sim P_{Y|Z} \\ Y_2^Q, \widetilde{Y}_2^Q \mid X, Z \overset{i.i.d.}{\sim} Q_{Y|X,Z}, \quad Y \perp\!\!\!\perp (Y_2^Q, \widetilde{Y}_2^Q) \mid X, Z, \\ Y_1^Q \mid Z \overset{i.i.d.}{\sim} Q_{Y|Z}, \quad Y_1^Q \perp\!\!\!\perp (X, Y, Y_2^Q, \widetilde{Y}_2^Q) \mid Z. \end{aligned} \tag{A.6.3}$$

Lemma A.6.4 shows $f_{\mathcal{K}}$ tightly satisfies the lower-bounding property, as f does in Lemma 1.2.2. The proof can be found in Appendix A.6.3.

Lemma A.6.4. *For any Q such that $f_{\mathcal{K}}(Q)$ exists, we have $f_{\mathcal{K}}(Q) \leq \mathcal{I}_{\mathcal{K}}$, with equality when $Q = P_{Y|X,Z}$.*

Therefore, we can provide an LCB for $\mathcal{I}_{\mathcal{K}}$ via a LCB for $f_{\mathcal{K}}(Q)$ with some choice of Q. Since the definition of $f_{\mathcal{K}}(Q)$ involves null Y samples such as $Y_2^Q, \widetilde{Y}_2^Q, Y_1^Q$, we will follow (A.6.3) to generate null samples of Y then construct i.i.d. unbiased estimates of the numerator and the denominator of $f_{\mathcal{K}}(Q)$ respectively. Based on the CLT and the delta method, we can derive asymptotically valid LCBs for $f_{\mathcal{K}}(Q)$. This idea is spelled out in Algorithm 9.

A.6.3 Proofs in Appendix A.6.2

Proof of Lemma A.6.4. Recall the form of the kernel floodgate functional in (A.6.2)

$$f_{\mathcal{K}}(Q) = \frac{\mathbb{E}\left[\mathcal{K}(Y, Y_2^Q)\right] - \mathbb{E}\left[\mathcal{K}(Y, Y_1^Q)\right]}{\sqrt{\mathbb{E}\left[\mathcal{K}(Y_2^Q, \widetilde{Y}_2^Q)\right] - \mathbb{E}\left[\mathcal{K}(Y_2^Q, Y_1^Q)\right]}} := \frac{\mathrm{II}_1}{\sqrt{\mathrm{II}_2}},$$

where $X, Z, Y, Y_2^Q, \widetilde{Y}_2^Q, Y_1^Q$ are defined as

$$(X, Z) \sim P_{X,Z}, \quad Y \mid X, Z \sim P_{Y|X,Z}, \quad Y \mid Z \sim P_{Y|Z} \tag{A.6.4}$$

$$Y_2^Q, \widetilde{Y}_2^Q \mid X, Z \overset{i.i.d.}{\sim} Q_{Y|X,Z}, \quad Y \perp\!\!\!\perp (Y_2^Q, \widetilde{Y}_2^Q) \mid X, Z, \tag{A.6.5}$$

$$Y_1^Q \mid Z \overset{i.i.d.}{\sim} Q_{Y|Z}, \quad Y_1^Q \perp\!\!\!\perp (X, Y, Y_2^Q, \widetilde{Y}_2^Q) \mid Z. \tag{A.6.6}$$

Denote the true conditional distributions $P_{Y|X,Z}, P_{Y|Z}$ by F, G respectively, the estimated conditional distributions $Q_{Y|X,Z}, Q_{Y|Z}$ by $F_{\mathfrak{q}}, G_{\mathfrak{q}}$ respectively, and the kernel mean embeddings of

Algorithm 9 Kernel floodgate

Input: Data $\{(Y_i, W_i)\}_{i=1}^n$, a chosen kernel $\mathcal{K}(\cdot, \cdot)$, a estimated conditional distribution of $P_{Y|X,Z}$, denoted by $Q_{Y|X,Z}$, resampling number M, $P_{X|Z}$, number of null replicates K, and a confidence level $\alpha \in (0, 1)$.

1: For each $i \in [n]$, draw $\{Y_{2,i}^{(m)}\}_{m=1}^M$ from $Q_{Y|X,Z}$ given (X_i, Z_i); given Z_i, draw i.i.d. null samples $\{\widetilde{X}_i^{(k)}\}_{k=1}^K$ from $P_{X|Z}$, then draw $\{Y_{1,i}^{(k,m)}\}_{m=1}^M$ from $Q_{Y|X,Z}$ given $(X_i, \widetilde{Z}_i^{(k)})$ for each $k \in [K]$. Denote $Y_{2,i}^{(m)} = Y_{1,i}^{(0,m)}$ for each $m \in [M]$.

2: Compute

$$\begin{aligned} R_i &= \frac{1}{M}\sum_{m=1}^{M} \mathcal{K}\left(Y_i, Y_{2,i}^{(m)}\right) - \frac{1}{KM}\sum_{k=1}^{K}\sum_{m=1}^{M} \mathcal{K}\left(Y_i, Y_{1,i}^{(k,m)}\right) \\ V_i &= \frac{2}{(K+1)M(M-1)}\sum_{k=0}^{K} \sum_{1\le m_1<m_2\le M} \mathcal{K}\left(Y_{1,i}^{(k,m_1)}, Y_{1,i}^{(k,m_2)}\right) \\ &\quad - \frac{2}{K(K+1)M^2}\sum_{m_1,m_2=1}^{M} \sum_{0\le k_1<k_2\le K} \mathcal{K}\left(Y_{1,i}^{(k_1,m_1)}, Y_{1,i}^{(k_2,m_2)}\right) \end{aligned}$$

for each $i \in [n]$, and their sample mean $(\bar{R}, \bar{V})$ and sample covariance matrix $\widehat{\Sigma}$, and compute $s^2 = \frac{1}{\bar{V}}\left[\left(\frac{\bar{R}}{2\bar{V}}\right)^2 \widehat{\Sigma}_{22} + \widehat{\Sigma}_{11} - \frac{\bar{R}}{\bar{V}}\widehat{\Sigma}_{12}\right]$.

Output: Lower confidence bound $L_n^\alpha(\mu) = \max\left\{\frac{\bar{R}}{\sqrt{\bar{V}}} - \frac{z_\alpha s}{\sqrt{n}},\ 0\right\}$, with the convention that $0/0 = 0$.

those conditional distributions by $\boldsymbol{\mu}_F, \boldsymbol{\mu}_G, \boldsymbol{\mu}_{F_{\mathfrak{q}}}, \boldsymbol{\mu}_{G_{\mathfrak{q}}}$. First notice

$$\left\langle \boldsymbol{\mu}_F, \boldsymbol{\mu}_{F_{\mathfrak{q}}} \right\rangle_{\mathcal{H}_{\mathcal{Y}}} = \left\langle \mathbb{E}\left[\mathcal{K}(\cdot, Y) \mid X, Z\right], \mathbb{E}\left[\mathcal{K}(\cdot, Y_2^Q) \mid X, Z\right] \right\rangle_{\mathcal{H}_{\mathcal{Y}}} = \mathbb{E}\left[\mathcal{K}(Y, Y_2^Q) \mid X, Z\right]$$

by (A.6.4), (A.6.5) and the definition of the kernel embedding. Similarly, we have the following equalities,

$$\mathbb{E}\left[\mathcal{K}(Y, Y_2^Q)\right] = \mathbb{E}\left[\mathbb{E}\left[\mathcal{K}(Y, Y_2^Q) \mid X, Z\right]\right] = \mathbb{E}\left[\left\langle \boldsymbol{\mu}_F, \boldsymbol{\mu}_{F_{\mathfrak{q}}} \right\rangle_{\mathcal{H}_{\mathcal{Y}}}\right], \tag{A.6.7}$$

$$\mathbb{E}\left[\mathcal{K}(Y, Y_1^Q)\right] = \mathbb{E}\left[\mathbb{E}\left[\mathcal{K}(Y, Y_1^Q) \mid Z\right]\right] = \mathbb{E}\left[\left\langle \boldsymbol{\mu}_G, \boldsymbol{\mu}_{G_{\mathfrak{q}}} \right\rangle_{\mathcal{H}_{\mathcal{Y}}}\right], \tag{A.6.8}$$

$$\mathbb{E}\left[\mathcal{K}(Y_2^Q, \widetilde{Y}_2^Q)\right] = \mathbb{E}\left[\mathbb{E}\left[\mathcal{K}(Y_2^Q, \widetilde{Y}_2^Q) \mid X, Z\right]\right] = \mathbb{E}\left[\left\langle \boldsymbol{\mu}_{F_{\mathfrak{q}}}, \boldsymbol{\mu}_{F_{\mathfrak{q}}} \right\rangle_{\mathcal{H}_{\mathcal{Y}}}\right], \tag{A.6.9}$$

$$\mathbb{E}\left[\mathcal{K}(Y_2^Q, Y_1^Q)\right] = \mathbb{E}\left[\mathbb{E}\left[\mathcal{K}(Y_2^Q, Y_1^Q) \mid X, Z\right]\right] = \mathbb{E}\left[\left\langle \boldsymbol{\mu}_{F_{\mathfrak{q}}}, \boldsymbol{\mu}_{G_{\mathfrak{q}}} \right\rangle_{\mathcal{H}_{\mathcal{Y}}}\right], \tag{A.6.10}$$

where we also apply the law of total expectation. Note that the subscripts for the expectation in the above equations are abbreviated. In addition to the these equalities, our derivation also relies on a key result $\mathbb{E}\left[\left\langle \boldsymbol{\mu}_G, \boldsymbol{\mu}_{F_{\mathfrak{q}}} \right\rangle\right] = \mathbb{E}\left[\left\langle \boldsymbol{\mu}_G, \boldsymbol{\mu}_{G_{\mathfrak{q}}} \right\rangle\right]$. Consider $\widetilde{Y}$ satisfying $\widetilde{Y} \mid X, Z \sim P_{Y|Z}$, $\widetilde{Y} \perp\!\!\!\perp Y_2^Q \mid X, Z, \widetilde{Y} \perp\!\!\!\perp Y_1^Q \mid Z$, then we prove the key result as below,

$$\begin{aligned}
\mathbb{E}\left[\left\langle \boldsymbol{\mu}_G, \boldsymbol{\mu}_{F_{\mathfrak{q}}} \right\rangle_{\mathcal{H}_{\mathcal{Y}}}\right] &= \mathbb{E}_{X,Z}\left[\mathbb{E}\left[\mathcal{K}(\widetilde{Y}, Y_2^Q) \mid X, Z\right]\right] \\
&= \mathbb{E}\left[\mathcal{K}(\widetilde{Y}, Y_2^Q)\right] \\
&= \mathbb{E}_Z\left[\mathbb{E}\left[\mathcal{K}(\widetilde{Y}, Y_2^Q) \mid Z\right]\right] \\
&= \mathbb{E}_Z\left[\mathbb{E}\left[\mathcal{K}(\widetilde{Y}, Y_1^Q) \mid Z\right]\right] = \mathbb{E}\left[\left\langle \boldsymbol{\mu}_G, \boldsymbol{\mu}_{G_{\mathfrak{q}}} \right\rangle_{\mathcal{H}_{\mathcal{Y}}}\right],
\end{aligned} \tag{A.6.11}$$

where the first and the last equalities hold by the definition of the kernel mean embedding, the second and the third equalities hold by the law of total expectation, and the fourth equality holds by the definitions of $Y_1^Q, Y_2^Q, \widetilde{Y}$.

Therefore we can rewrite the numerator of $f_{\mathcal{K}}(Q)$ as

$$\begin{aligned}
\mathrm{II}_1 &= \mathbb{E}\left[\mathcal{K}(Y, Y_2^Q)\right] - \mathbb{E}\left[\mathcal{K}(Y, Y_1^Q)\right] \\
&= \mathbb{E}\left[\left\langle \boldsymbol{\mu}_F, \boldsymbol{\mu}_{F_{\mathfrak{q}}} \right\rangle_{\mathcal{H}_{\mathcal{Y}}}\right] - \mathbb{E}\left[\left\langle \boldsymbol{\mu}_F, \boldsymbol{\mu}_{G_{\mathfrak{q}}} \right\rangle_{\mathcal{H}_{\mathcal{Y}}}\right] \\
&= \mathbb{E}\left[\left\langle \boldsymbol{\mu}_F, \boldsymbol{\mu}_{F_{\mathfrak{q}}} - \boldsymbol{\mu}_{G_{\mathfrak{q}}} \right\rangle_{\mathcal{H}_{\mathcal{Y}}}\right] \\
&= \mathbb{E}\left[\left\langle \boldsymbol{\mu}_F - \boldsymbol{\mu}_G, \boldsymbol{\mu}_{F_{\mathfrak{q}}} - \boldsymbol{\mu}_{G_{\mathfrak{q}}} \right\rangle_{\mathcal{H}_{\mathcal{Y}}}\right] + \mathbb{E}\left[\left\langle \boldsymbol{\mu}_G, \boldsymbol{\mu}_{F_{\mathfrak{q}}} - \boldsymbol{\mu}_{G_{\mathfrak{q}}} \right\rangle_{\mathcal{H}_{\mathcal{Y}}}\right] \\
&= \mathbb{E}\left[\left\langle \boldsymbol{\mu}_F - \boldsymbol{\mu}_G, \boldsymbol{\mu}_{F_{\mathfrak{q}}} - \boldsymbol{\mu}_{G_{\mathfrak{q}}} \right\rangle_{\mathcal{H}_{\mathcal{Y}}}\right] + \mathbb{E}\left[\left\langle \boldsymbol{\mu}_G, \boldsymbol{\mu}_{F_{\mathfrak{q}}} \right\rangle_{\mathcal{H}_{\mathcal{Y}}}\right] - \mathbb{E}\left[\left\langle \boldsymbol{\mu}_G, \boldsymbol{\mu}_{G_{\mathfrak{q}}} \right\rangle_{\mathcal{H}_{\mathcal{Y}}}\right] \\
&= \mathbb{E}\left[\left\langle \boldsymbol{\mu}_F - \boldsymbol{\mu}_G, \boldsymbol{\mu}_{F_{\mathfrak{q}}} - \boldsymbol{\mu}_{G_{\mathfrak{q}}} \right\rangle_{\mathcal{H}_{\mathcal{Y}}}\right] \\
&\leq \mathbb{E}\left[||\boldsymbol{\mu}_F - \boldsymbol{\mu}_G||_{\mathcal{H}_{\mathcal{Y}}} ||\boldsymbol{\mu}_{F_{\mathfrak{q}}} - \boldsymbol{\mu}_{G_{\mathfrak{q}}}||_{\mathcal{H}_{\mathcal{Y}}}\right] \\
&\leq \sqrt{\mathbb{E}\left[||\boldsymbol{\mu}_F - \boldsymbol{\mu}_G||^2_{\mathcal{H}_{\mathcal{Y}}}\right]} \sqrt{\mathbb{E}\left[||\boldsymbol{\mu}_{F_{\mathfrak{q}}} - \boldsymbol{\mu}_{G_{\mathfrak{q}}}||^2_{\mathcal{H}_{\mathcal{Y}}}\right]}, \qquad \text{(A.6.12)}
\end{aligned}$$

where the first line holds due to (A.6.7) and (A.6.8), the second to the fourth equalities hold by rearranging, the fifth equality holds due to (A.6.11), the last two inequalities hold by the Cauchy–Schwarz inequality. Regarding the denominator of $f_{\mathcal{K}}(Q)$, we rewrite II_2 in terms of the kernel embedding

$$\begin{aligned}
\mathrm{II}_2 &= \mathbb{E}\left[\mathcal{K}(Y_2^Q, \widetilde{Y}_2^Q)\right] - \mathbb{E}\left[\mathcal{K}(Y_2^Q, Y_1^Q)\right] \\
&= \mathbb{E}\left[\left\langle \boldsymbol{\mu}_{F_{\mathfrak{q}}}, \boldsymbol{\mu}_{F_{\mathfrak{q}}} \right\rangle_{\mathcal{H}_{\mathcal{Y}}}\right] - \mathbb{E}\left[\left\langle \boldsymbol{\mu}_{G_{\mathfrak{q}}}, \boldsymbol{\mu}_{F_{\mathfrak{q}}} \right\rangle_{\mathcal{H}_{\mathcal{Y}}}\right] \\
&= \mathbb{E}\left[\left\langle \boldsymbol{\mu}_{F_{\mathfrak{q}}}, \boldsymbol{\mu}_{F_{\mathfrak{q}}} \right\rangle_{\mathcal{H}_{\mathcal{Y}}}\right] + \mathbb{E}\left[\left\langle \boldsymbol{\mu}_{G_{\mathfrak{q}}}, \boldsymbol{\mu}_{G_{\mathfrak{q}}} \right\rangle_{\mathcal{H}_{\mathcal{Y}}}\right] - 2\mathbb{E}\left[\left\langle \boldsymbol{\mu}_{G_{\mathfrak{q}}}, \boldsymbol{\mu}_{F_{\mathfrak{q}}} \right\rangle_{\mathcal{H}_{\mathcal{Y}}}\right] \\
&= \mathbb{E}\left[||\boldsymbol{\mu}_{F_{\mathfrak{q}}} - \boldsymbol{\mu}_{G_{\mathfrak{q}}}||^2_{\mathcal{H}_{\mathcal{Y}}}\right], \qquad \text{(A.6.13)}
\end{aligned}$$

where the second equality holds due to (A.6.9) and (A.6.10) and the third equality holds since $\mathbb{E}\left[\left\langle \boldsymbol{\mu}_{G_{\mathfrak{q}}}, \boldsymbol{\mu}_{F_{\mathfrak{q}}} \right\rangle_{\mathcal{H}_{\mathcal{Y}}}\right] = \mathbb{E}\left[\left\langle \boldsymbol{\mu}_{G_{\mathfrak{q}}}, \boldsymbol{\mu}_{G_{\mathfrak{q}}} \right\rangle_{\mathcal{H}_{\mathcal{Y}}}\right]$ can be similarly derived as (A.6.11). As $\mathcal{I}^2_{\mathcal{K}}$ has equivalent expressions $\mathcal{I}^2_{\mathcal{K}} = \mathbb{E}\left[\mathrm{MMD}^2(P_{Y|X,Z}, P_{Y|Z})\right] = \mathbb{E}\left[||\boldsymbol{\mu}_F - \boldsymbol{\mu}_G||^2_{\mathcal{H}_{\mathcal{Y}}}\right]$, we have $f_{\mathcal{K}}(Q) \leq \mathcal{I}_{\mathcal{K}}$

by combining (A.6.2), (A.6.12), and (A.6.13). □

A.7 Transporting inference to other covariate distributions

To present how to perform inference on a target population whose covariate distribution differs from the distribution the study samples are drawn from, let Q denote the target distribution for all the random variables (Y, X, Z), but assume that $Q_{Y|X,Z} = P_{Y|X,Z}$ and that $Q_{X|Z}$ and the likelihood ratio Q_Z/P_Z are known (note this last requirement is trivially satisfied if only $X \mid Z$ changes between the study and target distributions, i.e., we know $Q_Z = P_Z$). Overloading notation slightly, let Q and P also denote the real-valued densities of random variables under their respective distributions (so, e.g., $P(Y = y \,|\, Z = z)$ denotes the density of $Y \mid Z = z$ under P evaluated at the value y), which we assume to exist. We can now define a weighted analogue of the floodgate functional (1.2.6):

$$f^w(\mu) = \frac{\mathbb{E}_P[(Y - \mu(\widetilde{X}, Z))^2 w(X, Z) w_1(\widetilde{X}, Z) - (Y - \mu(X, Z))^2 w(X, Z)]}{\sqrt{2\mathbb{E}_P[(\mu(X, Z) - \mu(\widetilde{X}, Z))^2 w(X, Z) w_1(\widetilde{X}, Z)]}}, \qquad \text{(A.7.1)}$$

where $w(x, z) = w_0(z)w_1(x, z)$, $w_0(z) = \frac{Q(Z=z)}{P(Z=z)}$, $w_1(x, z) = \frac{Q(X=x \,|\, Z=z)}{P(X=x \,|\, Z=z)}$, and $\widetilde{X} \sim P_{X|Z}$ conditionally independently of Y and X. The following Lemma certifies that f^w satisfies property (a) of a floodgate functional for $\mathcal{I}_Q^2 = \mathbb{E}_Q\left[\mathrm{Var}_Q\left(\mathbb{E}_Q[Y \mid X, Z] \mid Z\right)\right]$, the mMSE gap with respect to Q.

Lemma A.7.1. *If $Q_{Y|X,Z} = P_{Y|X,Z}$, then for any μ such that $f^w(\mu)$ exists, $f^w(\mu) \leq \mathcal{I}_Q$, with equality when $\mu = \mu^\star$.*

The proof is immediate from Lemma 1.2.2 if we notice that the ratio of the joint distribution of

$(Y, X, \widetilde{X}, Z)$ under the two populations equals

$$\frac{Q(Y,X,Z)Q(\widetilde{X} \mid Z)}{P(Y,X,Z)P(\widetilde{X} \mid Z)} = \frac{Q(Y \mid X,Z)}{P(Y \mid X,Z)} \frac{Q(X,Z)}{P(X,Z)} \frac{Q(\widetilde{X} \mid Z)}{P(\widetilde{X} \mid Z)} = w_1(\widetilde{X}, Z) w(X,Z), \quad \text{(A.7.2)}$$

where the last equality follows from $P_{Y|X,Z} = Q_{Y|X,Z}$. Floodgate property (b) of f^w can be established in the same way as for f by computing weighted versions of R_i and V_i from Algorithm 1 according to the weights in Equation (A.7.1), applying the central limit theorem, and combining them with the delta method.

A.8 Algorithm details for inference on the MACM gap

Recall the construction of the floodgate functional ((1.3.2) in Section 1.3.1):

$$f_{\ell_1}(\mu) = 2\mathbb{P}\big(Y(\mu(\widetilde{X}, Z) - \mathbb{E}\left[\mu(X,Z) \mid Z\right]) < 0\big) - 2\mathbb{P}\big(Y(\mu(X,Z) - \mathbb{E}\left[\mu(X,Z) \mid Z\right]) < 0\big).$$

We can define random variables which are i.i.d. and unbiased for $f_{\ell_1}(\mu)$ then construct CLT-based confidence bounds, as formalized in Algorithm 10. Algorithm 10 involves computing the terms

Algorithm 10 Floodgate for the MACM gap

Input: Data $\{(Y_i, X_i, Z_i)\}_{i=1}^n$, $P_{X|Z}$, a working regression function $\mu : \mathbb{R}^p \to \mathbb{R}$, and a confidence level $\alpha \in (0,1)$.

Let $U_i = \mu(X_i, Z_i) - \mathbb{E}[\mu(X_i, Z_i) \mid Z_i]$ and compute

$$R_i = \begin{cases} \mathbb{P}\left(U_i < 0 \mid Z_i\right) - \mathbb{1}_{\{U_i<0\}} & \text{if } Y_i = 1 \\ \mathbb{P}\left(U_i > 0 \mid Z_i\right) - \mathbb{1}_{\{U_i>0\}} & \text{if } Y_i = -1 \end{cases}$$

for $i \in [n]$, and compute its sample mean $\bar{R}$ and sample variance s^2.

return Lower confidence bound $L_n^\alpha(\mu) = 2\max\left\{\bar{R} - \frac{z_\alpha s}{\sqrt{n}}, 0\right\}$.

$\mathbb{E}[\mu(X_i, Z_i) \mid Z_i]$ and evaluating the CDF of the conditional distribution $\mu(X,Z) \mid Z = z$ at the value $\mathbb{E}[\mu(X_i, Z_i) \mid Z_i]$, which is not analytically possible in general. Unlike in Section 1.2.4,

where users can replace $\mathbb{E}\left[\mu(X,Z) \mid Z\right]$ and $\mathrm{Var}\left(\mu(X,Z) \mid Z\right)$ by their Monte Carlo estimators without it impacting asymptotic normality, we need slightly more assumptions when inferring the MACM gap due to the discontinuous indicator functions in the definition of $f_{\ell_1}(\mu)$. Before stating the required assumptions, we introduce some notation, all of which is specific to a given working regression function μ.

$$
\begin{aligned}
&U := \mu(X,Z), \;\; g(z) := \mathbb{E}[\mu(X,Z) \mid Z = z], \\
&G_z(u) := \mathbb{P}\left(U < u \mid Z = z\right), \;\; F_z(u) := \mathbb{P}\left(U \leq u \mid Z = z\right). \\
&\varsigma(z) := \sqrt{\mathrm{Var}\left(\mu(X,Z) \mid Z = z\right)}, \\
&C_{u,z,y} := \frac{\max\{|G_{z,y}(u) - G_{z,y}(g(z))|, |F_{z,y}(u) - F_{z,y}(g(z))|\}}{|u - g(z)|}
\end{aligned} \tag{A.8.1}
$$

where $F_{z,y}(u)$ is the CDF of $\mu(X,Z) \mid Z = z, Y = y$ evaluated at u, $G_{z,y}(u)$ is the limit from the left of the same CDF at u, and with the convention for $C_{u,z,y}$ that $0/0 = 0$ (so it is well-defined when $u = g(z)$). Now we are ready to state Assumption A.8.1.

Assumption A.8.1. *Assume the joint distribution over (Y,X,Z) and the nonrandom function $\mu : \mathbb{R}^p \to \mathbb{R}$ satisfy the following on a set of values of $Y = y, Z = z$ of probability 1:*

(a) There exists a $\delta_{z,y} > 0$ and finite $C_{z,y}$ such that

$$C_{u,z,y} \leq C_{z,y} \;\; \textit{when} \; |u - g(z)| \leq \varsigma(z)\delta_{z,y}.$$

(b) The above $C_{z,y}$ and $\delta_{z,y}$ satisfy

$$\mathbb{E}\left[C_{Z,Y}^2\right] < \infty, \; \mathbb{E}\left[\frac{1}{\delta_{Z,Y}}\right] < \infty.$$

(c) $\mathbb{E}\left[\varsigma^2(Z)\right] < \infty$, $\mathbb{E}\left[\frac{\mathbb{E}\left[|\mu(X,Z)-\mathbb{E}[\mu(X,Z)\,|\,Z]|^3\,|\,Z\right]}{\varsigma^3(Z)}\right] < \infty$.

These assumptions are placed because we have to construct the Monte Carlo estimator of $\mathbb{E}\left[\mu(X,Z)\,|\,Z\right]$ then plug it into the discontinuous indicator functions in $f_{\ell_1}(\mu)$. Assumptions A.8.1(a) and A.8.1(b) are smoothness requirements on the the CDF of $\mu(X,Z)\mid Z, Y$ around $\mathbb{E}\left[\mu(X,Z)\,|\,Z\right]$. Assumption A.8.1$(c)$ specifies mild moment bound conditions on $\mu(X,Z)$. To see that they are actually sensible, we consider the example of logistic regression and walk through those assumptions in Appendix A.8.1.

Assume that we can sample $(M+K)$ copies of X_i from $P_{X_i|Z_i}$ conditionally independently of X_i and Y_i, which are denoted by $\{\widetilde{X}_i^{(m)}\}_{m=1}^M$, $\{\widetilde{X}_i^{(k)}\}_{k=1}^K$, and thus replace $g(Z_i)$ (i.e. $\mathbb{E}[\mu(X_i,Z_i)\,|\,Z_i]$) and R_i, respectively, by the sample estimators

$$g^M(Z_i) = \frac{1}{M}\sum_{m=1}^{M}\mu(\widetilde{X}_i^{(m)}, Z_i),\ R_i^{M,K} = \frac{1}{K}\sum_{k=1}^{K}\left(\mathbb{1}_{\left\{Y_i(\mu(\widetilde{X}_i^{(k)},Z_i)-g^M(Z_i))<0\right\}}\right) - \mathbb{1}_{\{Y_i(\mu(X_i,Z_i))-g^M(Z_i))<0\}}$$

Theorem A.8.2. *Under the same setting as in Theorem 1.3.3, if either (i)* $\mathbb{E}\left[\mathrm{Var}(\mu(X,Z)\,|\,Z)\right] = 0$ *or (ii)* $\mathbb{E}\left[\mathrm{Var}\left(\mathbb{1}_{\{Y\cdot[\mu(X,Z)-\mathbb{E}[\mu(X,Z)\,|\,Z]]<0\}}\,|\,Z,Y\right)\right] > 0$ *holds together with Assumption A.8.1 and* $n/M = o(1)$*, then* $L^{\alpha}_{n,M,K}(\mu)$ *computed by replacing* $g(Z_i)$ *and* R_i *with* $g^M(Z_i)$ *and* $R_i^{M,K}$*, respectively, in Algorithm 10 satisfies*

$$\mathbb{P}\left(L^{\alpha}_{n,M,K}(\mu) \le \mathcal{I}_{\ell_1}\right) \ge 1-\alpha+o(1).$$

The proof can be found in Appendix A.8.2. Intuitively when we construct a lot more null samples to estimate the term $g(Z_i)$, our inferential validity improves. Formally, when $n^2/M = O(1)$, we can improve the asymptotic miscoverage to $O(n^{-1/2})$. Note that we only place a rate assumption on M (but put no requirement on K).

A.8.1 Illustration of assumption A.8.1

We consider the joint distribution over W to be p-dimensional multivariate Gaussian with $X = W_j, Z = W_{\text{-}j}$ for some $1 \leq j \leq p$, and Y follows a generalized linear model with logistic link. That is,

$$W \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}), \;\; \mu^\star(W) = 2\mathbb{P}(Y = 1 \,|\, W) - 1, \quad \text{where } \mathbb{P}(Y = 1 \,|\, W) = \frac{\exp(W\beta^\star)}{1 + \exp(W\beta^\star)}, \; \beta^\star \in \mathbb{R}^p.$$

Choosing logistic regression as the fitting algorithm, we have $U := \mu(X, Z)$ takes the following form

$$U := \mu(W) = \frac{2\exp(W\beta)}{1 + \exp(W\beta)} - 1$$

where $\beta \in \mathbb{R}^p$ is the fitted regression coefficient vector and $\beta_j \neq 0$ whenever $\mathbb{E}[\mathrm{Var}(\mu(X, Z) \,|\, Z)] > 0$. Conditional on Z, U follows a logit-normal distribution (defined as the logistic function transformation of normal random variable) up to constant shift and scaling. Note that the probability density function (PDF) of logit-normal distribution with parameters a, σ is

$$h_{\text{logit}}(u) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(\text{logit}(u) - a)^2}{2\sigma^2}\right) \frac{1}{u(1-u)}, \;\; u \in (0, 1) \tag{A.8.2}$$

where $\text{logit}(u) = \log(u/(1-u))$ is the logit function. Note $h_{\text{logit}}(u)$ is bounded over its support. Regarding the PDF of $U \mid Z = z, Y = 1$, which is denoted as $h_{z,1}(u)$, we first notice the following expression

$$h(x \mid Z = z, Y = 1) = \frac{h(x \mid Z = z)\mathbb{P}(Y = 1 \,|\, W = w)}{\int h(x \mid Z = z)\mathbb{P}(Y = 1 \,|\, W = w)\, dx} \tag{A.8.3}$$

where $w_j = x, w_{\text{-}j} = z, h(x \mid Z = z, Y = 1)$ and $h(x \mid Z = z, Y = 1)$ denote the density functions of $X \mid Z = z, Y = 1$ and $X \mid Z = z$. Since $\text{logit}(z)$ is one-to-one mapping, we have

$f_{z,1}(z)$ (up to constant shift and scaling) takes the form similar to (A.8.3)

$$h_{z,1}(u) = \frac{h_{\text{logit}}(u)\mathbb{P}\left(Y = 1 \mid W = w\right)}{\int h_{\text{logit}}(u)\mathbb{P}\left(Y = 1 \mid W = w\right) dx} \tag{A.8.4}$$

where $w = (x, z) = \mu^{-1}(u)$, and we denote the PDF of $U \mid Z = z$ as $h_{\text{logit}}(u)$ without causing confusion (the parameters of $h_{\text{logit}}(u)$ depend on z, β). Therefore we can show $h_{z,1}(z)$ is bounded (similarly for $h_{z,-1}(z)$).

The boundedness of $h_{z,y}(u)$ implies that the corresponding CDF $F_{z,y}$ ($F_{z,y} = G_{z,y}$ in this case) satisfies a Lipschitz condition over its support. Hence $\delta_{z,y}$ can be chosen to be greater than some positive constant uniformly, so that $\mathbb{E}\left[\frac{1}{\delta_{Z,Y}}\right] < \infty$ holds. Though the Lipschitz constant does depend on z, β, it is easy to verify $\mathbb{E}\left[C_{Z,Y}^2\right] < \infty$, thus assumption (b) holds. And assumption (c) is just a regular moment condition.

A.8.2 Proofs in Appendix A.8

Proof of Theorem A.8.2. Similar to the proof of Theorem 1.3.3, it suffices to deal with the case where $\mu(X, Z) \notin \mathscr{A}(Z)$ and prove

$$\mathbb{P}\left(L_{n,M,K}^{\alpha}(\mu) \leq f_{\ell_1}(\mu)\right) \geq 1 - \alpha + o(1). \tag{A.8.5}$$

Note that in Algorithm 10, $\mathbb{E}\left[R_i\right] = f_{\ell_1}(\mu)/2$. But when $g(Z_i)$ (i.e., $\mathbb{E}[\mu(X_i, Z_i) \mid Z_i]$) and R_i are replaced by $g^M(Z_i)$ and $R_i^{M,K}$, respectively, in Algorithm 10, we do not have $\mathbb{E}\left[R_i^{M,K}\right]$ equal to $f_{\ell_1}(\mu)/2$ anymore. Note that $f_{\ell_1}(\mu)/2$ equals the following

$$f_{\ell_1}(\mu)/2 = \mathbb{E}\left[\mathbb{1}_{\{Y\cdot[\mu(\widetilde{X},Z)-\mathbb{E}[\mu(X,Z)\mid Z]]<0\}}\right] - \mathbb{E}\left[\mathbb{1}_{\{Y\cdot[\mu(X,Z)-\mathbb{E}[\mu(X,Z)\mid Z]]<0\}}\right], \tag{A.8.6}$$

and $R_i^{M,K}$ is defined as

$$R_i^{M,K} = \frac{1}{K}\sum_{k=1}^{K}\left(\mathbb{1}_{\left\{Y_i(\mu(\widetilde{X}_i^{(k)},Z_i)-g^M(Z_i))<0\right\}}\right) - \mathbb{1}_{\{Y_i(\mu(X_i,Z_i))-g^M(Z_i))<0\}} \tag{A.8.7}$$

Remark the value of $\mathbb{E}\left[R_i^{M,K}\right]$ does not depend on K, hence we simplify the notation into R_i^M without causing confusion. Actually we can show as $M \to \infty, \mathbb{E}\left[R_i^M\right] \to f_{\ell_1}(\mu)/2$. Indeed, we need to show $\sqrt{n}|\mathbb{E}\left[R_i^M\right] - f_{\ell_1}(\mu)/2| = o(1)$ in order to prove (A.8.5). Also remark that in Section 1.3.1, it is mentioned that under a stronger condition $n^2/M = O(1)$ (which will imply $\sqrt{n}|\mathbb{E}\left[R_i^M\right] - f_{\ell_1}(\mu)/2| = O(1/\sqrt{n})$), we can additionally establish a rate for $n^{-1/2}$ for the asymptotic coverage validity in Theorem A.8.2. In either cases, it is reduced to prove

$$\left|\mathbb{E}\left[R_i^M\right] - \frac{f_{\ell_1}(\mu)}{2}\right| = O\left(\frac{1}{\sqrt{M}}\right) \tag{A.8.8}$$

First we ignore the i subscripts and get rid of the average over K null samples in the definition of $R_i^{M,K}$, then $\mathbb{E}\left[R_i^M\right]$ can be simplified into

$$\mathbb{E}\left[\mathbb{1}_{\left\{Y(\mu(\widetilde{X},Z)-g^M(Z))<0\right\}} - \mathbb{1}_{\{Y(\mu(X,Z)-g^M(Z))<0\}}\right] \tag{A.8.9}$$

where $g^M(Z) = \frac{1}{M}\sum_{m=1}^{M}\mu(\widetilde{X}^{(m)}, Z)$. To bound $\left|\mathbb{E}\left[R_i^M\right] - f_{\ell_1}(\mu)/2\right|$, we consider the two terms in (A.8.6) and separately bound

$$\begin{aligned}
\mathrm{II}_1 &:= \left|\mathbb{E}\left[\mathbb{1}_{\left\{Y(\mu(\widetilde{X},Z)-g^M(Z))<0\right\}} - \mathbb{1}_{\{Y\cdot[\mu(\widetilde{X},Z)-\mathbb{E}[\mu(X,Z)\,|\,Z]]<0}\right]\right|, \\
\mathrm{II}_2 &:= \left|\mathbb{E}\left[\mathbb{1}_{\{Y(\mu(X,Z)-g^M(Z))<0\}} - \mathbb{1}_{\{Y\cdot[\mu(X,Z)-\mathbb{E}[\mu(X,Z)\,|\,Z]]<0\}}\right]\right|.
\end{aligned}$$

Starting from the second term above, we rewrite it as

$$\begin{aligned}
\mathrm{II}_2 = & \left|\mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{\{Y(\mu(X,Z)-g^M(Z))<0\}} - \mathbb{1}_{\{Y\cdot[\mu(X,Z)-\mathbb{E}[\mu(X,Z)\,|\,Z]]<0\}} \,|\, Z, Y, \{\widetilde{X}^{(m)}\}_{m=1}^M\right]\right]\right| \\
\leq & \left|\mathbb{E}\left[\mathbb{1}_{\{Y=1\}}\mathbb{E}\left[\mathbb{1}_{\{\mu(X,Z)<g^M(Z)\}} - \mathbb{1}_{\{\mu(X,Z)<\mathbb{E}[\mu(X,Z)\,|\,Z]\}} \,|\, Z, Y, \{\widetilde{X}^{(m)}\}_{m=1}^M\right]\right]\right| \\
& + \left|\mathbb{E}\left[\mathbb{1}_{\{Y=-1\}}\mathbb{E}\left[\mathbb{1}_{\{\mu(X,Z)>g^M(Z)\}} - \mathbb{1}_{\{\mu(X,Z)>\mathbb{E}[\mu(X,Z)\,|\,Z]\}} \,|\, Z, Y, \{\widetilde{X}^{(m)}\}_{m=1}^M\right]\right]\right| \\
\leq & \ \mathbb{E}\left[\max\{\left|G_{Z,Y}(g^M(Z)) - G_{Z,Y}(g(Z))\right|, \left|F_{Z,Y}(g^M(Z)) - F_{Z,Y}(g(Z))\right|\}\right] \\
:= & \ \mathbb{E}\left[A\right] \qquad \text{(A.8.10)}
\end{aligned}$$

where the first equality is by the law of total expectation, the first and the second inequality are simply expanding and rearranging. By construction, $\mu(\widetilde{X}^{(m)}, Z), m \in [M]$ are i.i.d. random variables conditioning on Z, Y, then by central limit theorem we have

$$\frac{\sqrt{M}(g^M(Z) - g(Z))}{\varsigma(Z)} \xrightarrow{d} \mathcal{N}(0,1)$$

conditioning on Z, Y. Further we obtain the following from the Berry–Esseen bound i.e. Lemma A.3.3:

$$\left|\mathbb{P}\left(\left|\frac{\sqrt{M}|g^M(Z) - g(Z)|}{\varsigma(Z)}\right| > \sqrt{M}\delta_{Z,Y} \,\middle|\, Z, Y\right) - \overline{\Phi}(\left|\sqrt{M}\delta_{Z,Y}\right|)\right| \leq \frac{C}{\sqrt{M}} \cdot \frac{\mathbb{E}\left[|\mu^3(X,Z)|\,|\,Z\right]}{\varsigma^3(Z)} \qquad \text{(A.8.11)}$$

for any $\delta_{Z,Y}$ when conditioning on Z, Y, where $\overline{\Phi}(x) = 1 - \Phi(x)$ and C is some constant which does not depend on the distribution of (Y, X, Z). Regarding (A.8.10), by considering the event $B := \{|g^M(Z) - g(Z)|/\varsigma(Z) \leq \delta_{Z,Y}\}$, we can decompose (A.8.10) into

$$\mathbb{E}\left[A\right] = \mathbb{E}\left[A\mathbb{1}_{\{B\}}\right] + \mathbb{E}\left[A\mathbb{1}_{\{B^c\}}\right] \qquad \text{(A.8.12)}$$

For the first term, we have

$$\begin{aligned}
\mathbb{E}\left[A\mathbb{1}_{\{B\}}\right] &\leq \mathbb{E}\left[C_{g^M(Z),Z,Y}\left|g^M(Z)-g(Z)\right|\mathbb{1}_{\{B\}}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[C_{g^M(Z),Z,Y}\left|g^M(Z)-g(Z)\right|\mathbb{1}_{\{B\}}\,\middle|\,Z,Y\right]\right] \\
&\leq \mathbb{E}\left[C_{Z,Y}\mathbb{E}\left[\left|g^M(Z)-g(Z)\right|\,\middle|\,Z,Y\right]\right] \\
&\leq \mathbb{E}\left[C_{Z,Y}\sqrt{\mathbb{E}\left[\left|g^M(Z)-g(Z)\right|^2\,\middle|\,Z,Y\right]}\right] \qquad \text{(A.8.13)}
\end{aligned}$$

where the first inequality is by the definition of $C_{u,z,y}$, the first equality is from the law of total expectation, the second inequality holds by (a) in Assumption A.8.1 and the last inequality holds due to the Cauchy–Schwarz inequality. Remember we have $g^M(Z) = \frac{1}{M}\sum_{m=1}^{M}\mu(\widetilde{X}^{(m)},Z)$ where $\mu(\widetilde{X}^{(m)},Z), m \in [M]$ are i.i.d. random variables with mean $g(Z)$ when conditioning on Z,Y, hence (A.8.13) equals

$$\mathbb{E}\left[C_{Z,Y}\sqrt{\frac{\varsigma^2(Z)}{M}}\right] \leq \frac{1}{\sqrt{M}}\sqrt{\mathbb{E}\left[C_{Z,Y}^2\right]}\sqrt{\mathbb{E}\left[\varsigma^2(Z)\right]} = O\left(\frac{1}{\sqrt{M}}\right)$$

where the first inequality is from the Cauchy–Schwarz inequality and the second one holds by (b) and (c) in Assumption A.8.1. Now we have showed

$$\mathbb{E}\left[A\mathbb{1}_{\{B\}}\right] = O\left(\frac{1}{\sqrt{M}}\right), \qquad \text{(A.8.14)}$$

it suffices to prove the same rate for $\mathbb{E}\left[A\mathbb{1}_{\{B^c\}}\right]$:

$$\begin{aligned}
\mathbb{E}\left[A\mathbb{1}_{\{B^c\}}\right] &\leq 2\,\mathbb{P}\left(B^c\right) \\
&= 2\,\mathbb{E}\left[\mathbb{P}\left(B^c \mid Z\right)\right] \\
&= 2\,\mathbb{E}\left[\mathbb{P}\left(\sqrt{M}|g^M(Z) - g(Z)|/\varsigma(Z) > \sqrt{M}\delta_{Z,Y} \mid Z\right)\right] \\
&\leq 2\mathbb{E}\left[\bar{\Phi}(\left|\sqrt{M}\delta_{Z,Y}\right|) + \frac{C}{\sqrt{M}} \cdot \frac{\mathbb{E}\left[|\mu^3(X,Z)| \mid Z\right]}{\varsigma^3(Z)}\right] \\
&\leq 2\mathbb{E}\left[\frac{2}{\sqrt{2\pi}}\frac{\exp\{-M\delta_{Z,Y}^2\}}{\sqrt{M}\delta_{Z,Y}} + \frac{C}{\sqrt{M}} \cdot \frac{\mathbb{E}\left[|\mu^3(X,Z)| \mid Z\right]}{\varsigma^3(Z)}\right]
\end{aligned}$$

where the first inequality holds since $F_{z,y}(u), G_{z,y}(u)$ are bounded between 0 and 1, the first equality is due to the law of total expectation, the second equality is from the definition of the event B, the second inequality holds due to (A.8.11) and the last inequality is a result of Mill's Ratio, see Proposition 2.1.2 in Vershynin (2018). Under (b) and (c) in Assumption A.8.1, the following holds

$$\mathbb{E}\left[A\mathbb{1}_{\{B^c\}}\right] = O\left(\frac{1}{\sqrt{M}}\right). \tag{A.8.15}$$

Finally we prove

$$\left|\mathbb{E}\left[\mathbb{1}_{\{Y(\mu(X,Z)-g^M(Z))<0\}} - \mathbb{1}_{\{Y\cdot[\mu(X,Z)-\mathbb{E}[\mu(X,Z)\mid Z]]<0\}}\right]\right| = O\left(\frac{1}{\sqrt{M}}\right).$$

Regarding the term

$$\mathrm{II}_1 = \left|\mathbb{E}\left[\mathbb{1}_{\{Y(\mu(\widetilde{X},Z)-g^M(Z))<0\}} - \mathbb{1}_{\{Y\cdot[\mu(\widetilde{X},Z)-\mathbb{E}[\mu(X,Z)\mid Z]]<0}\right]\right|$$

All of the steps are the same except that the CDF (and its limit) of the conditional distribution $X \mid Z, Y$ are replaced by those of $X \mid Z$, i.e. $F_z(u)$ and $G_z(u)$ as defined in (A.8.1). Hence it suffices to

notice the following derivations for $F_z(u)$:

$$\begin{aligned} F_z(u) = \mathbb{P}\left(U \le u \,|\, Z = z\right) &= \mathbb{E}_{Y|Z=z}\left[\mathbb{P}\left(U \le u \,|\, Z = z, Y\right) | \, Z = z\right] \\ &= \mathbb{E}_{Y|Z=z}\left[F_{z,Y}(u) \,|\, Z = z\right], \end{aligned}$$

and similarly for $G_z(u)$. Together with the definition of $C_{u,z,y}$ and (a) in Assumption A.8.1, the above equations yield

$$\max\{|F_z(u) - F_z(g(z))|, |G_z(u) - G_z(g(z))|\} \le C_{z,y}|u - g(z)|$$

over the region $|u - g(z)| \le \varsigma(z)\delta_{z,y}$. Then the other steps follow as those of proving the term II_2. Finally, we obtain a rate of $O\left(\frac{1}{\sqrt{M}}\right)$ for $\left|\mathbb{E}\left[R_i^M\right] - f_{\ell_1}(\mu)/2.\right|$.

In the following, we prove the stronger version of (A.8.5), i.e.,

$$\mathbb{P}\left(L^{\alpha}_{n,M,K}(\mu) \le f_{\ell_1}(\mu)\right) \ge 1 - \alpha - O\left(\frac{1}{\sqrt{n}}\right), \tag{A.8.16}$$

when assuming $n^2/M = O(1)$. For this it suffices to establish the following Berry–Esseen bound:

$$\Delta := \sup_{t\in\mathbb{R}}\left|\mathbb{P}\left(\sqrt{n}\left(\frac{\bar{R} - f_{\ell_1}(\mu)/2}{s}\right) \le t\right) - \Phi(t)\right| = O\left(\frac{1}{\sqrt{n}}\right),$$

where $\bar{R}$ and s are defined similarly as in Algorithm 10 except that $g(Z_i)$ and R_i are replaced with $g^M(Z_i)$ and $R_i^{M,K}$, respectively. Notice that

$$\begin{aligned} \Delta &= \sup_{t\in\mathbb{R}}\left|\mathbb{P}\left(\sqrt{n}\left(\frac{\bar{R} - \mathbb{E}\left[R_i^M\right]}{s}\right) \le t + \sqrt{n}\frac{(\mathbb{E}\left[R_i^M\right] - f_{\ell_1}(\mu)/2)}{s}\right) - \Phi(t)\right| \\ &\le \sup_{t\in\mathbb{R}}\left|\mathbb{P}\left(\sqrt{n}\left(\frac{\bar{R} - \mathbb{E}\left[R_i^M\right]}{s}\right) \le t\right) - \Phi(t)\right| + \sup_{t\in\mathbb{R}}\left|\Phi\left(t + \sqrt{n}\frac{(\mathbb{E}\left[R_i^M\right] - f_{\ell_1}(\mu)/2)}{s}\right) - \Phi(t)\right| \\ &:= \Delta_1 + \Delta_2 \end{aligned}$$

Since the first derivative of $\Phi(t)$ is bounded by $1/\sqrt{2\pi}$ over $\mathbb{R}$, we have

$$\Delta_2 \quad \leq \quad \frac{\sqrt{n}}{\sqrt{2\pi}} \frac{|f_{\ell_1}(\mu)/2 - \mathbb{E}\left[R_i^M\right]|}{\sqrt{\mathrm{Var}\left(R_i^M\right)}} \cdot (\sqrt{\mathrm{Var}\left(R_i^M\right)}/s)$$

by Taylor expansion. Note that as a result of (A.8.8), we have

$$\sqrt{n}|\mathbb{E}\left[R_i^M\right] - f_{\ell_1}(\mu)/2| = O(1/\sqrt{n}). \tag{A.8.17}$$

Then it suffices to prove $\Delta_1 = O(1/\sqrt{n})$ and $\mathrm{Var}\left(R_i^M\right) > 0$ (since s is simply the sample mean estimator of $\mathrm{Var}\left(R_i^M\right)$ thus consistent). $\Delta_1 = O(1/\sqrt{n})$ holds when applying the triangular array version of the Berry–Esseen bound in Lemma A.3.4 (note that the result is stated in a way such that the bound clearly applies to the triangular array with i.i.d. rows $\{R_i^{M,K}\}_{i=1}^n$ for each M). The only thing we need to deal with is to verify the following uniform moment conditions:

(i) $\sup_{M,K} \mathbb{E}\left[\left|R_i^{M,K} - \mathbb{E}\left[R_i^{M,K}\right]\right|^3\right] < \infty$,

(ii) $\inf_{M,K} \mathrm{Var}\left(R_i^{M,K}\right) > 0$.

where we go back to the original notation $R_i^{M,K}$ from the simplified one R_i^M since the above moments do depend on both M and K. Since $R_i^{M,K}$ is always bounded, (i) holds. Regarding (ii),

notice that we have the following

$$
\begin{aligned}
&\mathrm{Var}\left(R_i^{M,K}\right) \\
&= \mathbb{E}\left[\mathrm{Var}\left(R_i^{M,K} \mid Z_i, Y_i, \{\widetilde{X}_i^{(m)}\}_{m=1}^M\right)\right] + \mathrm{Var}\left(\mathbb{E}\left[R_i^{M,K} \mid Z_i, Y_i, \{\widetilde{X}_i^{(m)}\}_{m=1}^M\right]\right) \\
&\geq \mathbb{E}\left[\mathrm{Var}\left(R_i^{M,K} \mid Z_i, Y_i, \{\widetilde{X}_i^{(m)}\}_{m=1}^M\right)\right] \\
&= \mathbb{E}\left[\mathrm{Var}\left(\frac{1}{K}\sum_{k=1}^K \left(\mathbb{1}_{\left\{Y_i(\mu(\widetilde{X}_i^{(k)}, Z_i) - g^M(Z_i)) < 0\right\}}\right) - \mathbb{1}_{\{Y_i(\mu(X_i, Z_i)) - g^M(Z_i)) < 0\}} \,\middle|\, Z_i, Y_i, \{\widetilde{X}_i^{(m)}\}_{m=1}^M\right)\right] \\
&\geq \mathbb{E}\left[\mathrm{Var}\left(\mathbb{1}_{\{Y_i(\mu(X_i, Z_i)) - g^M(Z_i)) < 0\}} \,\middle|\, Z_i, Y_i, \{\widetilde{X}_i^{(m)}\}_{m=1}^M\right)\right] := \sigma_M^2 \qquad \text{(A.8.18)}
\end{aligned}
$$

where the first equality is due to the law of total expectation, the second equality is by the definition of $R_i^{M,K}$, the second inequality holds since $\{\widetilde{X}_i^{(k)}\}_{k=1}^K \perp\!\!\!\perp X_i \mid Z_i, Y_i, \{\widetilde{X}_i^{(m)}\}_{m=1}^M$ due to the construction of $\{\widetilde{X}_i^{(k)}\}_{k=1}^K$ and the variance of first term is non-negative. Before dealing with (A.8.18), notice the stated condition

$$
\sigma_0^2 := \mathbb{E}\left[\mathrm{Var}\left(\mathbb{1}_{\{Y_i(\mu(X_i, Z_i)) - g(Z_i)) < 0\}} \mid Z_i, Y_i\right)\right] > 0
$$

Thus to establish (ii), it suffices to show $\sigma_M^2 \to \sigma_0^2$ as $M \to \infty$. Recall the derivations in (A.8.10) for bounding the term II_2, we can similarly bound $|\sigma_M^2 - \sigma_0^2|$ by the following quantity:

$$
\begin{aligned}
|\sigma_M^2 - \sigma_0^2| &\leq \mathbb{E}\left[3\max\{\left|G_{Z,Y}(g^M(Z)) - G_{Z,Y}(g(Z))\right|, \left|F_{Z,Y}(g^M(Z)) - F_{Z,Y}(g(Z))\right|\}\right] \\
&= 3\mathbb{E}\left[A\right] = 3(\mathbb{E}\left[A\mathbb{1}_{\{B\}}\right] + \mathbb{E}\left[A\mathbb{1}_{\{B^c\}}\right]) = O\left(\frac{1}{\sqrt{M}}\right).
\end{aligned}
$$

where the last equality holds due to the results (A.8.14) and (A.8.15) from previous derivations for the term II_2. Finally we conclude (A.8.16), which immediately implies a weaker version of the result, i.e.the statement of Theorem A.8.2. □

A.9 Co-sufficient floodgate details

The strategy described in Section 1.3.2 is formalized in Algorithm 11 (under the simplifying assumption that the number of batches, n_2, evenly divides the sample size n).

Algorithm 11 Co-sufficient floodgate

Input: The inputs of Algorithm 1, a sufficient statistic functional $\mathcal{T}$, and a batch size n_2.

1: Let $n_1 = n/n_2$ and for $m \in [n_1]$, denote $(\boldsymbol{X}_m, \boldsymbol{Z}_m) = \{X_i, Z_i\}_{i=(m-1)n_2+1}^{mn_2}$, and let $\boldsymbol{T}_m = \mathcal{T}(\boldsymbol{X}_m, \boldsymbol{Z}_m)$.

2: For $m \in [n_1]$, compute

$$R_m = \frac{1}{n_2} \sum_{i=(m-1)n_2+1}^{mn_2} Y_i \left(\mu(X_i, Z_i) - \mathbb{E}\left[\mu(X_i, Z_i) \mid \boldsymbol{Z}_m, \boldsymbol{T}_m\right]\right),$$

$$V_m = \frac{1}{n_2} \sum_{i=(m-1)n_2+1}^{mn_2} \mathrm{Var}\left(\mu(X_i, Z_i) \mid \boldsymbol{Z}_m, \boldsymbol{T}_m\right),$$

their sample mean $(\bar{R}, \bar{V})$, their sample covariance matrix $\widehat{\Sigma}$, and $s^2 = \frac{1}{\bar{V}}\left[\left(\frac{\bar{R}}{2\bar{V}}\right)^2 \widehat{\Sigma}_{22} + \widehat{\Sigma}_{11} - \frac{\bar{R}}{\bar{V}}\widehat{\Sigma}_{12}\right]$.

3: **return** Lower confidence bound $L_n^{\alpha,\mathcal{T}}(\mu) = \max\left\{\frac{\bar{R}}{\sqrt{\bar{V}}} - \frac{z_\alpha s}{\sqrt{n_1}},\ 0\right\}$, with the convention that $0/0 = 0$.

A.9.1 Monte Carlo analogue of co-sufficient floodgate

Similarly as in Section 1.2, when the conditional expectations in Algorithm 11 do not have closed-form expressions, Monte Carlo provides a general approach: within each batch, we can sample K copies $\widetilde{\boldsymbol{X}}_m^{(k)}$ of $\boldsymbol{X}_m$ from the conditional distribution $\boldsymbol{X}_m \mid \boldsymbol{Z}_m, \boldsymbol{T}_m$, conditionally independently

of $\boldsymbol{X}_m$ and $\boldsymbol{y}$ and thus replace R_m and V_m, respectively, by the sample estimators

$$(R_m^K, V_m^K) = \frac{1}{n_2}\left(\sum_{i=(m-1)n_2+1}^{mn_2} Y_i\left(\mu(X_i, Z_i) - \frac{1}{K}\sum_{k=1}^{K}\mu(\widetilde{X}_i^{(k)}, Z_i)\right),\right.$$
$$\left.\sum_{i=(m-1)n_2+1}^{mn_2} \frac{1}{K-1}\sum_{k=1}^{K}\left(\mu(X_i^{(k)}, Z_i) - \frac{1}{K}\sum_{k=1}^{K}\mu(\widetilde{X}_i^{(k)}, Z_i)\right)^2\right)$$

We defer to future work a proof of validity of the Monte Carlo analogue of co-sufficient floodgate following similar techniques as Theorem 1.2.5.

A.9.2 Proofs in Appendix A.9

Lemma A.9.1. *Under the moment conditions* $\mathbb{E}\left[\mu^2(X,Z)\right], \mathbb{E}\left[(\mu^\star)^2(X,Z)\right] < \infty$*, we can quantify the gap between* $f(\mu)$ *and* $f_n^{\mathcal{T}}(\mu)$ *as below.*

$$f(\mu) - f_n^{\mathcal{T}}(\mu) = O\left(\max\{\mathrm{II}(\mu), \mathrm{II}(\mu^\star)\}\right) \tag{A.9.1}$$

where $\mathrm{II}(\mu) = \mathbb{E}_{\boldsymbol{Z}}\left[\mathrm{Var}_{\boldsymbol{T}|\boldsymbol{Z}}\left(\mathbb{E}\left[\mu(X_i, Z_i) \mid \boldsymbol{Z}, \boldsymbol{T}\right]\right)\right]$.

When this lemma is used in the proof of Proposition 1.3.5 and 1.3.6, the natural sufficient statistic and $f_n^{\mathcal{T}}(\mu)$ are actually defined based on the batch $\mathcal{B}_m$ whose sample size is n_2. We do not carry these in the above notation, but use generic $(\boldsymbol{X}, \boldsymbol{Z})$ instead, where $(\boldsymbol{X}, \boldsymbol{Z}) = \{(X_i, Z_i)\}_{i=1}^n$.

Proof of Lemma A.9.1. Recall the definition of $f(\mu)$ and $f_n^{\mathcal{T}}(\mu)$,

$$f(\mu) = \frac{\mathbb{E}\left[\mathrm{Cov}(\mu^\star(X,Z), \mu(X,Z) \mid Z)\right]}{\sqrt{\mathbb{E}\left[\mathrm{Var}(\mu(X,Z) \mid Z)\right]}}, \tag{A.9.2}$$

$$f_n^{\mathcal{T}}(\mu) = \frac{\mathbb{E}\left[\mathrm{Cov}(\mu^\star(X_i,Z_i), \mu(X_i,Z_i) \mid \boldsymbol{Z}, \boldsymbol{T})\right]}{\sqrt{\mathbb{E}\left[\mathrm{Var}(\mu(X_i,Z_i) \mid \boldsymbol{Z}, \boldsymbol{T})\right]}}, \tag{A.9.3}$$

then denote $W_i = (X_i, Z_i)$, $h(W_i) := \mu(W_i) - \mathbb{E}\left[\mu(W_i) \mid Z_i\right], h^{\mathcal{T}}(W_i) := \mu^\star(W_i) - \mathbb{E}\left[\mu^\star(W_i) \mid \boldsymbol{Z}, \boldsymbol{T}\right]$ and assume $\mathbb{E}\left[h^2(W_i)\right] = 1$ without loss of generality. First notice a simple fact $|\frac{a}{b} - \frac{c}{d}| = \frac{|ad-bc|}{bd} = \frac{|ad-cd+cd-bc|}{bd} \leq \frac{|a-c|}{b} + \frac{c|b-d|}{bd}$ for $a, b, c, d > 0$, then let the numerator and denominator of $f(\mu)$ in (A.9.2) to be a, b respectively (similarly denote c, d for $f_n^{\mathcal{T}}(\mu)$ in (A.9.3)). And we have

$$\max\{\frac{1}{b}, \frac{c}{bd}\} \leq 1 + f_n^{\mathcal{T}}(\mu) \leq 1 + f_n^{\mathcal{T}}(\mu^\star) \leq 1 + f(\mu^\star) \leq 1 + \mathbb{E}\left[(\mu^\star)^2(X, Z)\right] < \infty,$$

hence it suffices to bound $|a - c|$ and $|b - d|$. First we have the following

$$\begin{aligned} a - c &= \mathbb{E}\left[\mathrm{Cov}\left(\mu^\star(W_i), \mu(W_i) \mid \boldsymbol{Z}\right)\right] - \mathbb{E}\left[\mathrm{Cov}\left(\mu^\star(W_i), \mu(W_i) \mid \boldsymbol{Z}, \boldsymbol{T}\right)\right] \quad \text{(A.9.4)} \\ &= \mathbb{E}\left[\mathrm{Cov}\left(\mathbb{E}\left[\mu^\star(W_i) \mid \boldsymbol{Z}, \boldsymbol{T}\right], \mathbb{E}\left[\mu(W_i) \mid \boldsymbol{Z}, \boldsymbol{T}\right] \mid \boldsymbol{Z}\right)\right] \\ &= \mathbb{E}_{\boldsymbol{Z}}\left[\mathrm{Cov}_{\boldsymbol{T}|\boldsymbol{Z}}\left(\mathbb{E}\left[\mu^\star(W_i) \mid \boldsymbol{Z}, \boldsymbol{T}\right], \mathbb{E}\left[\mu(W_i) \mid \boldsymbol{Z}, \boldsymbol{T}\right]\right)\right]. \end{aligned}$$

where the first equality holds due to the independence among $i.i.d.$ samples $(\boldsymbol{X}, \boldsymbol{Z}) = \{(X_i, Z_i)\}_{i=1}^n$. For the second equality, we apply the law of total covariance to the covariance term $\mathrm{Cov}\left(\mu^\star(W_i), \mu(W_i) \mid \boldsymbol{Z}\right)$ then cancel out the second term of the first line, leading to the term in the second line. Finally we spell out the randomness of the expectation and covariance through explicit subscripts in the last inequality. They by applying Cauchy–Schwarz inequality, we obtain

$$|a - c| \leq \sqrt{\mathbb{E}_{\boldsymbol{Z}}\left[\mathrm{Var}_{\boldsymbol{T}|\boldsymbol{Z}}\left(\mathbb{E}\left[\mu^\star(W_i) \mid \boldsymbol{Z}, \boldsymbol{T}\right]\right)\right]} \sqrt{\mathbb{E}_{\boldsymbol{Z}}\left[\mathrm{Var}_{\boldsymbol{T}|\boldsymbol{Z}}\left(\mathbb{E}\left[\mu(W_i) \mid \boldsymbol{Z}, \boldsymbol{T}\right]\right)\right]} \quad \text{(A.9.5)}$$

Regarding the term $|b-d|$, we have

$$\begin{aligned}
|b-d| &= \left|\sqrt{\mathbb{E}\left[h^2(W_i)\right]}-\sqrt{\mathbb{E}\left[(h^{\mathcal{T}})^2(W_i)\right]}\right| \\
&= \frac{\left|\mathbb{E}\left[h^2(W_i)\right]-\mathbb{E}\left[(h^{\mathcal{T}})^2(W_i)\right]\right|}{\sqrt{\mathbb{E}\left[h^2(W_i)\right]}+\sqrt{\mathbb{E}\left[(h^{\mathcal{T}})^2(W_i)\right]}} \\
&\leq \frac{\left|\mathbb{E}\left[h^2(W_i)\right]-\mathbb{E}\left[(h^{\mathcal{T}})^2(W_i)\right]\right|}{\sqrt{\mathbb{E}\left[h^2(W_i)\right]}} \\
&\leq \mathbb{E}\left[\mathrm{Var}\left(\mu(W_i)\mid \boldsymbol{Z}\right)\right]-\mathbb{E}\left[\mathrm{Var}\left(\mu(W_i)\mid \boldsymbol{Z},\boldsymbol{T}\right)\right] \\
&= \mathbb{E}_{\boldsymbol{Z}}\left[\mathrm{Var}_{\boldsymbol{T}|\boldsymbol{Z}}\left(\mathbb{E}\left[\mu(W_i)\mid \boldsymbol{Z},\boldsymbol{T}\right]\right)\right] \qquad \text{(A.9.6)}
\end{aligned}$$

where we use the assumption $\mathbb{E}\left[h^2(W_i)\right]=1$ and the definition of $h, h^{\mathcal{T}}$ in the second inequality. The last equality holds as a result of applying the law of total variance to the variance term $\mathrm{Var}\left(\mu(W_i)\mid \boldsymbol{Z}\right)$ then getting the second term of line 4 cancelled out. Finally, combining (A.9.5) and (A.9.6) establishes the bound in (A.9.1). □

Proposition 1.3.5

Proof of Proposition 1.3.5. Throughout the proof, the natural sufficient statistic and $f_n^{\mathcal{T}}(\mu)$ are defined based on the batch $\mathcal{B}_m$ whose sample size is n_2. But we will abbreviate the notation dependence on it for simplicity and use a generic n instead of n_2 to avoid carrying too many subscripts, without causing any confusion. Now we present a roadmap of this proof.

(i) due to Lemma A.9.1, it suffices to bound the term $\mathrm{II}(\mu)$, $\mathrm{II}(\mu^\star)$ in (A.9.1).

(ii) we bound $\mathrm{II}(\mu)$, $\mathrm{II}(\mu^\star)$ with the same strategy. Specifically, we will show

$$\mathrm{II}(\mu)=O\left(\mathbb{E}_{Z_i}\left[\mathbb{E}_F\left[\mu^2(W_i)\right]\mathbb{E}\left[h_{ii}\mid Z_i\right]\right]\right)$$

and similarly for $\mathrm{II}(\mu^\star)$ under the stated model, where F denotes the conditional distribution

of $X_i|\boldsymbol{Z}$, and h_{ii} is the ith diagonal term of the hat matrix $\boldsymbol{H}$, which is defined later. This terminology comes from the fact that we can treat X_j as response variable, $(1, Z)$ as predictors, the natural sufficient statistic for this low dimensional multivariate Gaussian distribution is equivalent to the OLS estimator.

(iii) Regarding the term $\mathbb{E}\left[h_{ii} \mid Z_i\right]$ above, we can carefully bound it by $1/(n-1) + \mathbb{E}\left[\boldsymbol{\Xi} \mid Z_i\right]$, where $\boldsymbol{\Xi}$ is defined in (A.9.16).

(iv) Simply expanding $\mathbb{E}\left[\boldsymbol{\Xi} \mid Z_i\right]$ into three terms: $\mathrm{III}_1, \mathrm{III}_2, \mathrm{III}_3$, which are defined in (A.9.17), (A.9.18) and (A.9.18), we will show $\mathrm{III}_2 = 0$ and figure out the stochastic representation of $\mathrm{III}_1, \mathrm{III}_3$, which turns out to be related to chi-squared, Wishart and inverse-Wishart random variables.

(v) Cauchy–Schwarz inequalities together with some properties of those random variables (chi-squared, Wishart and inverse-Wishart) and the stated moment conditions finally gives us the result in (1.3.4).

Having proved Lemma A.9.1, now we directly start with step (ii). Notice the following

$$\begin{aligned}
\mathrm{II}(\mu) &= \mathbb{E}_{\boldsymbol{Z}}\left[\mathrm{Var}_{\boldsymbol{T}|\boldsymbol{Z}}\left(\mathbb{E}\left[\mu(W_i) \mid \boldsymbol{Z}, \boldsymbol{T}\right]\right)\right] \\
&= \mathbb{E}_{\boldsymbol{Z}}\left[\mathbb{E}_{\boldsymbol{T}|\boldsymbol{Z}}\left[\left(\mathbb{E}_F\left[\mu(W_i)\right] - \mathbb{E}_{F_{\boldsymbol{T}}}\left[\mu(W_i)\right]\right)^2\right]\right] \\
&= \mathbb{E}_{\boldsymbol{Z}}\left[\mathrm{Var}_F\left(\mu(W_i)\right)\mathbb{E}_{\boldsymbol{T}|\boldsymbol{Z}}\left[\frac{\left(\mathbb{E}_F\left[\mu(W_i)\right] - \mathbb{E}_{F_{\boldsymbol{T}}}\left[\mu(W_i)\right]\right)^2}{\mathrm{Var}_F\left(\mu(W_i)\right)}\right]\right] \\
&\leq \mathbb{E}_{\boldsymbol{Z}}\left[\mathrm{Var}_F\left(\mu(W_i)\right)\min\left\{\mathbb{E}_{\boldsymbol{T}|\boldsymbol{Z}}\left[\chi^2(F_{\boldsymbol{T}} \| F)\right], 2\right\}\right] \qquad \text{(A.9.7)}
\end{aligned}$$

where the second equality is just rewriting the conditional variance, with F denoting the conditional distribution $X_i|\boldsymbol{Z}$ and $F_{\boldsymbol{T}}$ denoting the conditional distribution $X_i|\boldsymbol{Z}, \boldsymbol{T}$. Here we abbreviate the subscript dependence on i for notation simplicity. The third equality holds since

$\mathrm{Var}_F\left(\mu(W_i)\right) \in \mathscr{A}(\boldsymbol{Z})$. Regarding the last inequality, we make use of the variational representation of χ^2-divergence:

$$\chi^2(P\|Q) = \sup_{\mu} \frac{(\mathbb{E}_P(\mu) - \mathbb{E}_Q(\mu))^2}{\mathrm{Var}_Q(\mu)}$$

and the fact that

$$\begin{aligned}
& \mathbb{E}_{\boldsymbol{T}|\boldsymbol{Z}}\left[\frac{\left(\mathbb{E}_F\left[\mu(W_i)\right] - \mathbb{E}_{F_{\boldsymbol{T}}}\left[\mu(W_i)\right]\right)^2}{\mathrm{Var}_F\left(\mu(W_i)\right)}\right] \\
\leq & \frac{\mathbb{E}_{\boldsymbol{T}|\boldsymbol{Z}}\left[\mathbb{E}_F\left[\mu^2(W_i)\right]\right] + \mathbb{E}_{\boldsymbol{T}|\boldsymbol{Z}}\left[\mathbb{E}_{F_{\boldsymbol{T}}}\left[\mu^2(W_i)\right]\right] - 2\mathbb{E}_{\boldsymbol{T}|\boldsymbol{Z}}\left[\mathbb{E}_{F_{\boldsymbol{T}}}\left[\mu(W_i)\right]\mathbb{E}_F\left[\mu(W_i)\right]\right]}{\mathrm{Var}_F\left(\mu(W_i)\right)} \\
= & \frac{\mathbb{E}_F\left[\mu^2(W_i)\right] + \mathbb{E}_F\left[\mu^2(W_i)\right] - 2(\mathbb{E}_F\left[\mu(W_i)\right])^2}{\mathrm{Var}_F\left(\mu(W_i)\right)} \\
= & \frac{2\mathrm{Var}_F\left(\mu(W_i)\right)}{\mathrm{Var}_F\left(\mu(W_i)\right)} = 2
\end{aligned}$$

where the first inequality is from expanding the quadratic term and the fact $\left(\mathbb{E}_F\left[\mu(W_i)\right]\right)^2 \leq \mathbb{E}_F\left[\mu^2(W_i)\right]$, $\left(\mathbb{E}_{F_{\boldsymbol{T}}}\left[\mu(W_i)\right]\right)^2 \leq \mathbb{E}_{F_{\boldsymbol{T}}}\left[\mu^2(W_i)\right]$, the first equality holds as a result of the tower property of conditional expectation and $\mathbb{E}_F\left[\mu(W_i)\right] \in \mathscr{A}(\boldsymbol{Z})$. Denote $u_i = (1, Z_i)^\top$ and the following n by p matrix by $\boldsymbol{U}$:

$$\boldsymbol{U} = \begin{pmatrix} u_1^\top \\ \vdots \\ u_n^\top \end{pmatrix} = (\mathbf{1}, \boldsymbol{Z}) \tag{A.9.8}$$

Recall that the sufficient statistic (here we ignore the batching index)

$$\boldsymbol{T} = (\sum_{i\in[n]} X_i, \sum_{i\in[n]} X_i Z_i) = \boldsymbol{U}^\top \boldsymbol{X},$$

under the stated multivariate Gaussian model, we know $\boldsymbol{X} \mid \boldsymbol{Z} \sim \mathcal{N}\left(\boldsymbol{U}\gamma, \sigma^2 \boldsymbol{I}_n\right)$, then the conditional distribution of $(X_i, \boldsymbol{T}) \mid \boldsymbol{Z}$ can be specified as below

$$\begin{pmatrix} X_i \\ \boldsymbol{T} \end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix} (1, Z_i)\gamma \\ \boldsymbol{U}^\top \boldsymbol{U}\gamma \end{bmatrix}, \sigma^2 \begin{bmatrix} 1 & e_i^\top \boldsymbol{U} \\ \boldsymbol{U}^\top e_i^\top & \boldsymbol{U}^\top \boldsymbol{U} \end{bmatrix} \right) \tag{A.9.9}$$

where $e_i \in \mathbb{R}^n$, $(e_1, \cdots, e_n)$ forms the standard orthogonal basis. Noticing the above joint distribution is multivariate Gaussian, we can immediately derive the conditional distribution as below,

$$X_i \mid \boldsymbol{Z}, \boldsymbol{T} \sim \mathcal{N}\left(e_i^\top \boldsymbol{U}(\boldsymbol{U}^\top \boldsymbol{U})^{-1}\boldsymbol{U}^\top \boldsymbol{X}, \sigma^2(1 - e_i^\top \boldsymbol{U}(\boldsymbol{U}^\top \boldsymbol{U})^{-1}\boldsymbol{U}^\top e_i)\right).$$

Denote $\boldsymbol{H} = \boldsymbol{U}(\boldsymbol{U}^\top \boldsymbol{U})^{-1}\boldsymbol{U}^\top$, which is the "hat" matrix. Now we compactly write down the following two conditional distributions:

$$\begin{aligned} F_{\boldsymbol{T}} &: X_i \mid \boldsymbol{Z}, \boldsymbol{T} \sim \mathcal{N}\left(e_i^\top \boldsymbol{H}\boldsymbol{X}, \sigma^2(1 - h_{ii})\right) \\ F &: X_i \mid \boldsymbol{Z} \sim \mathcal{N}\left((1, Z_i)\gamma, \sigma^2\right) \end{aligned}$$

Note the sufficient statistic $\boldsymbol{T}$ is equivalent to

$$\widehat{\gamma}^{OLS} = (\boldsymbol{U}^\top \boldsymbol{U})^{-1}\boldsymbol{U}^\top \boldsymbol{X}$$

whenever $\boldsymbol{U}^\top \boldsymbol{U}$ is nonsingular. Here $\widehat{\gamma}^{OLS}$ is the OLS estimator for γ (when treating X as response variable, $(1, Z)$ as predictors). Simply, we have

$$\widehat{\gamma}^{OLS} \sim \mathcal{N}\left(\gamma, \sigma^2(\boldsymbol{U}^\top \boldsymbol{U})^{-1}\right)$$

Now we are ready to calculate $\chi^2(F_{\boldsymbol{T}}\|F)$. First,

$$\begin{aligned} e_i^\top \boldsymbol{H}\boldsymbol{X} - (1, Z_i)\gamma &= e_i^\top \boldsymbol{U}\widehat{\gamma}^{OLS} - (1, Z_i)\gamma \\ &= e_i^\top \boldsymbol{U}(\widehat{\gamma}^{OLS} - \gamma) \sim \mathcal{N}(0, \sigma^2 h_{ii}) \end{aligned} \tag{A.9.10}$$

Since $2\sigma^2 > \sigma^2(1 - h_{ii})$, applying Lemma A.9.2 yields the following

$$\begin{aligned} \chi^2(F_{\boldsymbol{T}}\|F) &= \frac{1}{2}\left[\frac{1}{\sqrt{1-h_{ii}^2}}\exp\left\{\frac{(e_i^\top \boldsymbol{H}\boldsymbol{X} - (1, Z_i)\gamma)^2}{\sigma^2(1+h_{ii})}\right\} - 1\right] \\ &\leq \frac{1}{\sqrt{1-h_{ii}}}\exp\left\{\frac{(e_i^\top \boldsymbol{H}\boldsymbol{X} - (1, Z_i)\gamma)^2}{\sigma^2(1+h_{ii})}\right\} - 1 \\ &= \frac{1}{\sqrt{1-h_{ii}}}\exp\left\{\frac{h_{ii}G^2}{1+h_{ii}}\right\} - 1 \end{aligned} \tag{A.9.11}$$

where $G \sim \mathcal{N}(0,1)$ is independent from $\boldsymbol{X}$ and the last equality holds due to (A.9.10). Plugin (A.9.11) back to (A.9.7), we have

$$\begin{aligned} \mathrm{II}(\mu) &\leq \mathbb{E}_{\boldsymbol{Z}}\left[\mathrm{Var}_F\left(\mu(W_i)\right)\min\left\{\mathbb{E}_{\boldsymbol{T}|\boldsymbol{Z}}\left[\chi^2(F_{\boldsymbol{T}}\|F)\right], 2\right\}\right] \\ &\leq \mathbb{E}_{\boldsymbol{Z}}\left[\mathrm{Var}_F\left(\mu(W_i)\right)\min\left\{\mathbb{E}_{\boldsymbol{T}|\boldsymbol{Z}}\left[\frac{1}{\sqrt{1-h_{ii}}}\exp\left\{\frac{h_{ii}G^2}{1+h_{ii}}\right\} - 1\right], 2\right\}\right] \end{aligned}$$

Note the moment generating function for χ_1^2 random variable is $\frac{1}{\sqrt{1-2t}}$ when $t < 1/2$. Since the expectation of $\exp\left\{\frac{h_{ii}G^2}{1+h_{ii}}\right\}$ does not always exist, we consider two events E and E^c such that conditional on the event E, the expectation exists and the probability of event E^c is small. More

specifically, define the event $E = \{h_{ii} < \frac{1}{2}\}$, which implies

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{T}|\boldsymbol{Z}}\left[\frac{1}{\sqrt{1-h_{ii}}}\exp\left\{\frac{h_{ii}G^2}{1+h_{ii}}\right\}\right] - 1 &= \frac{1}{\sqrt{1-h_{ii}}\sqrt{1-2h_{ii}/(1+h_{ii})}} - 1 \\
&= \frac{\sqrt{1+h_{ii}}}{1-h_{ii}} - 1 \\
&\leq \frac{1+h_{ii}}{1-h_{ii}} - 1 \\
&\leq 4h_{ii}
\end{aligned}$$

hence we can bound $\mathrm{II}(\mu)$ by the summation of the following two terms:

$$\mathrm{II}_1 := \mathbb{E}_{\boldsymbol{Z}}\left[\mathrm{Var}_F\left(\mu(W_i)\right)\mathbb{1}_{\{E\}}\cdot 4h_{ii}\right],\ \mathrm{II}_2 := \mathbb{E}_{\boldsymbol{Z}}\left[\mathrm{Var}_F\left(\mu(W_i)\right)\mathbb{1}_{\{E^c\}}\cdot 2\right]$$

Regarding II_1, the following holds:

$$\mathrm{II}_1 \leq 4\,\mathbb{E}_{Z_i}\left[\mathbb{E}_F\left[\mu^2(W_i)\right]\mathbb{E}\left[h_{ii}\,|\,Z_i\right]\right],$$

where we apply the tower property of conditional expectation and $\mathrm{Var}_F\left(\mu(W_i)\right) \leq \mathbb{E}_F\left[\mu^2(W_i)\right] \in \mathscr{A}(Z_i)$ Regarding II_2, we have

$$\begin{aligned}
\mathrm{II}_2 &= 2\,\mathbb{E}_{\boldsymbol{Z}}\left[\mathrm{Var}_F\left(\mu(W_i)\right)\mathbb{1}_{\{E^c\}}\right] \\
&= 2\,\mathbb{E}_{\boldsymbol{Z}}\left[\mathrm{Var}_F\left(\mu(W_i)\right)\mathbb{E}\left[\mathbb{1}_{\{E^c\}}\,|\,Z_i\right]\right] \\
&\leq 2\,\mathbb{E}_{Z_i}\left[\mathbb{E}_F\left[\mu^2(W_i)\right]\mathbb{P}\left(h_{ii}\geq\frac{1}{2}\,|\,Z_i\right)\right] \\
&\leq 4\,\mathbb{E}_{Z_i}\left[\mathbb{E}_F\left[\mu^2(W_i)\right]\mathbb{E}\left[h_{ii}\,|\,Z_i\right]\right]
\end{aligned}$$

where the second equality comes from the tower property of conditional expectation and $\mathrm{Var}_F\left(\mu(W_i)\right) \in \mathscr{A}(Z_i)$ and the last inequality holds due to Markov's inequality. Now we can compactly write

down the following bound for $\mathrm{II}(\mu)$,

$$\mathrm{II}(\mu) \leq \mathrm{II}_1 + \mathrm{II}_2 \leq 8\,\mathbb{E}_{Z_i}\left[\mathbb{E}_F\left[\mu^2(W_i)\right]\mathbb{E}\left[h_{ii} \mid Z_i\right]\right], \tag{A.9.12}$$

Similarly we obtain $\mathrm{II}(\mu^\star) = O\left(\mathbb{E}_{Z_i}\left[\mathbb{E}_F\left[(\mu^\star)^2(W_i)\right]\mathbb{E}\left[h_{ii} \mid Z_i\right]\right)\right)$. Now we proceed step (iii), i.e. calculating $\mathbb{E}\left[h_{ii} \mid Z_i\right]$. Notice h_{ii} is the ith diagonal term of the "hat" matrix, which involves $\{w_i\}_{i=1}^n$. In order to bound the conditional expectation of h_{ii} given Z_i in a sharp way, we carefully expand h_{ii} and try to get w_i separated from $\{w_m\}_{m\neq i}$. Recall the definition of $\boldsymbol{U} = (\mathbf{1}, \boldsymbol{Z})$ in (A.9.8), we can rewrite

$$\boldsymbol{U}^\top\boldsymbol{U} = \sum_{m\neq i} u_m u_m^\top + u_i u_i^\top, \quad \boldsymbol{A} := \sum_{m\neq i} u_m u_m^\top$$

Note that $h_{ii} = u_i^\top(\boldsymbol{U}^\top\boldsymbol{U})^{-1}u_i$ since $\boldsymbol{H} = \boldsymbol{U}(\boldsymbol{U}^\top\boldsymbol{U})^{-1}\boldsymbol{U}^\top$, hence we have

$$h_{ii} = u_i^\top(\boldsymbol{A} + u_i u_i^\top)^{-1}u_i$$

As $n > p$, $\boldsymbol{A}$ is almost surely positive definite thus invertible, then applying Sherman–Morrison formula to $\boldsymbol{A}$ and $u_i u_i^\top$ yields the following

$$h_{ii} = u_i^\top\boldsymbol{A}^{-1}u_i - \frac{(u_i^\top\boldsymbol{A}^{-1}u_i)^2}{1 + u_i^\top\boldsymbol{A}^{-1}u_i} \leq u_i^\top\boldsymbol{A}^{-1}u_i. \tag{A.9.13}$$

Since $\boldsymbol{A}$ also involves the unit vector $\mathbf{1}_{n-1}$, it is easier when we first project $\boldsymbol{Z}_{\cdot i}$ on $\mathbf{1}_{n-1}$ then work with the orthogonal complement. Bearing this idea in mind, we denote $\boldsymbol{\Omega} = (\mathbf{1}_{n-1}, \boldsymbol{Z}_{\cdot i})$ which is a

$n-1$ by p matrix, then rewrite $\boldsymbol{A}$ as

$$\boldsymbol{A} = \boldsymbol{\Omega}^\top \boldsymbol{\Omega} = \begin{pmatrix} \mathbf{1}_{n-1}^\top \mathbf{1}_{n-1} & \mathbf{1}_{n-1}^\top \boldsymbol{Z}_{\cdot i} \\ \boldsymbol{Z}_{\cdot i}^\top \mathbf{1}_{n-1} & \boldsymbol{Z}_{\cdot i}^\top \boldsymbol{Z}_{\cdot i} \end{pmatrix}$$

where $\mathbf{I}_{n-1}$ is the $(n-1)$ dimensional identity matrix. Denote

$$\overline{\boldsymbol{Z}_{\cdot i}} := \frac{1}{n-1} \sum_{m \neq i} Z_m = \frac{1}{n-1} \mathbf{1}_{n-1}^\top \boldsymbol{Z}_{\cdot i} \quad \boldsymbol{\Gamma} := \begin{pmatrix} 1 & -\overline{\boldsymbol{Z}_{\cdot i}} \\ \mathbf{0} & \mathbf{I}_{n-1} \end{pmatrix}, \tag{A.9.14}$$

we have

$$\begin{aligned} \boldsymbol{\Omega}\boldsymbol{\Gamma} = (\mathbf{1}_{n-1}, \boldsymbol{Z}_{\cdot i})\boldsymbol{\Gamma} &= (\mathbf{1}_{n-1}, \boldsymbol{Z}_{\cdot i} - \mathbf{1}_{n-1}\overline{\boldsymbol{Z}_{\cdot i}}) \\ &= (\mathbf{1}_{n-1}, (\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1})\boldsymbol{Z}_{\cdot i}). \end{aligned}$$

where $\boldsymbol{P}_{n-1} = \mathbf{1}_{n-1}\mathbf{1}_{n-1}^\top/(n-1)$ is the projection matrix onto $\mathbf{1}_{n-1}$. Then we immediately have

$$(\boldsymbol{\Omega}\boldsymbol{\Gamma})^\top \boldsymbol{\Omega}\boldsymbol{\Gamma} = \begin{pmatrix} n-1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{Z}_{\cdot i}^\top (\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1}) \boldsymbol{Z}_{\cdot i} \end{pmatrix}$$

since $\boldsymbol{P}_{n-1}\mathbf{1}_{n-1} = \mathbf{1}_{n-1}$, $(\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1})\mathbf{1}_{n-1} = \mathbf{0}$ and

$$u_i^\top \boldsymbol{\Gamma} = (1, Z_i)\boldsymbol{\Gamma} = (1, Z_i - \overline{\boldsymbol{Z}_{\cdot i}}). \tag{A.9.15}$$

Combining (A.9.14) with (A.9.15) yields the following

$$\begin{aligned} u_i^\top \boldsymbol{A}^{-1} u_i &= u_i^\top (\boldsymbol{\Omega}^\top \boldsymbol{\Omega})^{-1} u_i \\ &= u_i^\top \boldsymbol{\Gamma} ((\boldsymbol{\Omega}\boldsymbol{\Gamma})^\top \boldsymbol{\Omega}\boldsymbol{\Gamma})^{-1} \boldsymbol{\Gamma}^\top w_i \\ &= \frac{1}{n-1} + (Z_i - \overline{\boldsymbol{Z}_{\cdot i}})(\boldsymbol{Z}_{\cdot i}^\top (\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1}) \boldsymbol{Z}_{\cdot i})^{-1} (Z_i - \overline{\boldsymbol{Z}_{\cdot i}})^\top, \end{aligned}$$

which together with (A.9.13) implies $\mathbb{E}\left[h_{ii} \mid Z_i\right] \leq \mathbb{E}\left[u_i^\top \boldsymbol{A}^{-1} u_i \mid Z_i\right] = 1/(n-1) + \mathbb{E}\left[\boldsymbol{\Xi} \mid Z_i\right]$, where

$$\boldsymbol{\Xi} = (Z_i - \overline{\boldsymbol{Z}_{\cdot i}})(\boldsymbol{Z}_{\cdot i}^\top (\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1}) \boldsymbol{Z}_{\cdot i})^{-1} (Z_i - \overline{\boldsymbol{Z}_{\cdot i}})^\top. \tag{A.9.16}$$

As the problem has been reduced to calculating $\mathbb{E}\left[\boldsymbol{\Xi} \mid Z_i\right]$, we arrive at the step (iv) now. Write $(\boldsymbol{Z}_i - \overline{\boldsymbol{Z}_{\cdot i}}) = (\boldsymbol{Z}_i - \boldsymbol{v}_0) - (\overline{\boldsymbol{Z}_{\cdot i}} - \boldsymbol{v}_0)$, where $\boldsymbol{v}_0$ is the mean of Gaussian random variable Z, we can expand $\mathbb{E}\left[\boldsymbol{\Xi} \mid Z_i\right] = \mathrm{III}_1 + \mathrm{III}_2 + \mathrm{III}_3$, where

$$\mathrm{III}_1 = (\boldsymbol{Z}_i - \boldsymbol{v}_0)\mathbb{E}\left[(\boldsymbol{Z}_{\cdot i}^\top (\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1}) \boldsymbol{Z}_{\cdot i})^{-1} \mid \boldsymbol{Z}_i\right] (\boldsymbol{Z}_i - \boldsymbol{v}_0)^\top \tag{A.9.17}$$

$$\mathrm{III}_2 = -2(\boldsymbol{Z}_i - \boldsymbol{v}_0)\mathbb{E}\left[(\boldsymbol{Z}_{\cdot i}^\top (\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1}) \boldsymbol{Z}_{\cdot i})^{-1} (\overline{\boldsymbol{Z}_{\cdot i}} - \boldsymbol{v}_0)^\top \mid \boldsymbol{Z}_i\right] \tag{A.9.18}$$

$$\mathrm{III}_3 = \mathbb{E}\left[(\overline{\boldsymbol{Z}_{\cdot i}} - \boldsymbol{v}_0)(\boldsymbol{Z}_{\cdot i}^\top (\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1}) \boldsymbol{Z}_{\cdot i})^{-1} (\overline{\boldsymbol{Z}_{\cdot i}} - \boldsymbol{v}_0)^\top \mid \boldsymbol{Z}_i\right] \tag{A.9.19}$$

Below we are going to show $\mathrm{III}_2 = 0$ and derive $\mathrm{III}_1, \mathrm{III}_3$ carefully. Regarding the term III_1, we exactly write down its stochastic representation. Under the state Gaussian model, we have $\boldsymbol{Z}_{\cdot i}^\top \sim \mathcal{N}\left(\boldsymbol{v}_0 \mathbf{1}_{n-1}^\top, \mathbf{I}_{n-1} \otimes \boldsymbol{\Sigma}_0\right)$, then $(\boldsymbol{Z}_{\cdot i}^\top (\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1}) \boldsymbol{Z}_{\cdot i})^{-1}$ follows an inverse Wishart distribution i.e.

$$(\boldsymbol{Z}_{\cdot i}^\top (\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1}) \boldsymbol{Z}_{\cdot i})^{-1} \sim \mathcal{W}_{p-1}^{-1}(\boldsymbol{\Sigma}_0^{-1}, n-2)$$

and $\boldsymbol{Z}_{\cdot i} \perp\!\!\!\perp \boldsymbol{Z}_i$, hence we can calculate

$$\mathbb{E}\left[(\boldsymbol{Z}_{\cdot i}^\top(\mathbf{I}_{n-1}-\boldsymbol{P}_{n-1})\boldsymbol{Z}_{\cdot i})^{-1} \mid \boldsymbol{Z}_i\right] = \frac{\boldsymbol{\Sigma}_0^{-1}}{n-p-2}.$$

Plug in the above equation into (A.9.17), we have

$$\mathrm{III}_1 = (\boldsymbol{Z}_i-\boldsymbol{v}_0)\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{Z}_i-\boldsymbol{v}_0)^\top = \frac{\boldsymbol{\Phi}}{n-p-2}, \quad \text{where } \boldsymbol{\Phi}\sim\chi^2_{p-1},\ \boldsymbol{\Phi}\perp\!\!\!\perp \boldsymbol{Z}_{\cdot i}. \tag{A.9.20}$$

Regarding the term III_2 in (A.9.18), we first denote $\boldsymbol{Z} = \boldsymbol{Z}_{\cdot i} - \mathbf{1}_{n-1}\boldsymbol{v}_0$ and notice

$$\boldsymbol{Z}\sim\mathcal{N}\left(\mathbf{0}, \mathbf{I}_{n-1}\otimes\boldsymbol{\Sigma}_0\right), \quad \mathbf{1}_{n-1}^\top\boldsymbol{Z} = (n-1)(\overline{\boldsymbol{Z}_{\cdot i}}-\boldsymbol{v}_0), \tag{A.9.21}$$

then rewrite III_2 as below

$$\mathrm{III}_2 = -2(\boldsymbol{Z}_i-\boldsymbol{v}_0)\mathbb{E}\left[((\boldsymbol{Z}+\mathbf{1}_{n-1}\boldsymbol{v}_0)^\top(\mathbf{I}_{n-1}-\boldsymbol{P}_{n-1})(\boldsymbol{Z}+\mathbf{1}_{n-1}\boldsymbol{v}_0))^{-1}\frac{(\mathbf{1}_{n-1}^\top\boldsymbol{Z})^\top}{n-1}\right]$$

where we also makes use of the fact that

$$(\boldsymbol{Z}_{\cdot i}^\top(\mathbf{I}_{n-1}-\boldsymbol{P}_{n-1})\boldsymbol{Z}_{\cdot i})^{-1}(\overline{\boldsymbol{Z}_{\cdot i}}-\boldsymbol{v}_0)^\top \perp\!\!\!\perp \boldsymbol{Z}_i$$

Noticing that $(\mathbf{1}_{n-1}\boldsymbol{v}_0)^\top(\mathbf{I}_{n-1}-\boldsymbol{P}_{n-1}) = \mathbf{0}$, we can simplify further

$$\mathrm{III}_2 = -\frac{2}{n-1}(\boldsymbol{Z}_i-\boldsymbol{v}_0)\mathbb{E}\left[(\boldsymbol{Z}^\top(\mathbf{I}_{n-1}-\boldsymbol{P}_{n-1})\boldsymbol{Z})^{-1}(\mathbf{1}_{n-1}^\top\boldsymbol{Z})^\top\right] \tag{A.9.22}$$

Notice in the above equation, $\boldsymbol{Z}^\top(\mathbf{I}_{n-1}-\boldsymbol{P}_{n-1})$ is the orthogonal complement of $\boldsymbol{Z}^\top\mathbf{1}_{n-1}$, which implies independence under the Gaussian distribution assumption, which we will now use to prove the expectation in (A.9.22) equals zero. Formally, we first have $(\boldsymbol{Z}^\top(\mathbf{I}_{n-1}-\boldsymbol{P}_{n-1}), \boldsymbol{Z}^\top\mathbf{1}_{n-1})$ are

multivariate Gaussian. Introducing the vectorization of matrix and the Kronecker product, we can express in the following way:

$$\text{vec}(\boldsymbol{Z}^\top(\mathbf{I}_{n-1}-\boldsymbol{P}_{n-1}))=(\mathbf{I}_{n-1}-\boldsymbol{P}_{n-1})\otimes\mathbf{I}_{p-1}\text{vec}(\boldsymbol{Z}^\top),\ \ \text{vec}(\boldsymbol{Z}^\top)=\mathbf{1}_{n-1}\otimes\mathbf{I}_{p-1}\text{vec}(\boldsymbol{Z}^\top).$$

Now we are ready to calculate the covariance

$$\begin{aligned}
&\text{Cov}\left(\text{vec}(\boldsymbol{Z}^\top(\mathbf{I}_{n-1}-\boldsymbol{P}_{n-1})),\text{vec}(\boldsymbol{Z}^\top\mathbf{1}_{n-1})\right)\\
=\ &((\mathbf{I}_{n-1}-\boldsymbol{P}_{n-1})\otimes\mathbf{I}_{p-1})(\mathbf{I}_{n-1}\otimes\boldsymbol{\Sigma}_0)(\mathbf{1}_{n-1}\otimes\mathbf{I}_{p-1})^\top\\
=\ &((\mathbf{I}_{n-1}-\boldsymbol{P}_{n-1})\mathbf{I}_{n-1}\mathbf{1}_{n-1})\otimes(\mathbf{I}_{p-1}\boldsymbol{\Sigma}_0\mathbf{I}_{p-1})=\mathbf{0}
\end{aligned}$$

where in above equalities we use the fact $\text{Var}\left(\text{vec}(\boldsymbol{Z}^\top)\right)=\mathbf{I}_{n-1}\otimes\boldsymbol{\Sigma}_0$ in (A.9.21) and the mixed-product property of the Kronecker product. Therefore

$$\boldsymbol{Z}^\top(\mathbf{I}_{n-1}-\boldsymbol{P}_{n-1})\perp\!\!\!\perp\boldsymbol{Z}^\top\mathbf{1}_{n-1}\implies \text{III}_2=0 \tag{A.9.23}$$

Regarding the term III_3, first denote $\boldsymbol{\Psi}_1=\boldsymbol{Z}^\top\boldsymbol{P}_{n-1}\boldsymbol{Z}$ and $\boldsymbol{\Psi}_2=\boldsymbol{Z}^\top(\mathbf{I}_{n-1}-\boldsymbol{P}_{n-1})\boldsymbol{Z}$, we obtain two independent Wishart random variables i.e.

$$\boldsymbol{\Psi}_1\sim\mathcal{W}_{p-1}(\boldsymbol{\Sigma}_0,1),\ \ \boldsymbol{\Psi}_2\sim\mathcal{W}_{p-1}(\boldsymbol{\Sigma}_0,n-2),\ \ \boldsymbol{\Psi}_1\perp\!\!\!\perp\boldsymbol{\Psi}_2.$$

Then III_3 can be calculated as below

$$\begin{aligned}
\mathrm{III}_3 &= \mathbb{E}\left[(\overline{\boldsymbol{Z}_{\cdot i}} - \boldsymbol{v}_0)(\boldsymbol{Z}_{\cdot i}^\top(\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1})\boldsymbol{Z}_{\cdot i})^{-1}(\overline{\boldsymbol{Z}_{\cdot i}} - \boldsymbol{v}_0)^\top \mid \boldsymbol{Z}_i\right] \\
&= \mathbb{E}\left[\mathbf{1}_{n-1}^\top \boldsymbol{Z}(\boldsymbol{Z}^\top(\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1})\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top \mathbf{1}_{n-1}\right]/(n-1)^2 \\
&= \mathbb{E}\left[\mathrm{Tr}\left(\mathbf{1}_{n-1}^\top \boldsymbol{Z}(\boldsymbol{Z}^\top(\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1})\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top \mathbf{1}_{n-1}\right)\right]/(n-1)^2 \\
&= \mathbb{E}\left[\mathrm{Tr}(\boldsymbol{\Psi}_1\boldsymbol{\Psi}_2^{-1})\right]/(n-1) \\
&= \mathrm{Tr}\,\mathbb{E}\left[\boldsymbol{\Psi}_1\boldsymbol{\Psi}_2^{-1}\right]/(n-1) \\
&= \mathrm{Tr}(\mathbb{E}\left[\boldsymbol{\Psi}_1\right]\mathbb{E}\left[\boldsymbol{\Psi}_2^{-1}\right])/(n-1) \\
&= \mathrm{Tr}(\boldsymbol{\Sigma}_0\frac{\boldsymbol{\Sigma}_0^{-1}}{n-p-2})/(n-1) \\
&= \frac{p}{(n-1)(n-p-2)}
\end{aligned} \tag{A.9.24}$$

where the first equality is from (A.9.19), the second equality is similarly obtained as (A.9.22), the fourth equality holds by the fact $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$ and the definition of $\boldsymbol{\Psi}_1$ and $\boldsymbol{\Psi}_2$, the sixth equality holds due to $\boldsymbol{\Psi}_1 \perp\!\!\!\perp \boldsymbol{\Psi}_2$. So far we have shown $\mathrm{III}_2 = 0$ and figured out the stochastic representation of $\mathrm{III}_2, \mathrm{III}_3$, which are also further simplified using the properties of Wishart and inverse-Wishart random variables. These bring us to the final stage i.e. step (v). Combining (A.9.13), (A.9.20), (A.9.23) and (A.9.24), we finally obtain

$$\begin{aligned}
\mathbb{E}\left[h_{ii} \mid Z_i\right] &\leq \mathbb{E}\left[u_i^\top \boldsymbol{A}^{-1} u_i \mid Z_i\right] \\
&\leq \frac{1}{n-1} + \mathbb{E}\left[\boldsymbol{\Xi} \mid Z_i\right] \\
&= \frac{1}{n-1} + \mathrm{III}_1 + \mathrm{III}_2 + \mathrm{III}_3 \\
&\leq \frac{1}{n-1} \cdot \frac{n-2}{n-p-2} + \frac{\boldsymbol{\Phi}}{n-p-2}
\end{aligned} \tag{A.9.25}$$

Recall the bound for $\mathrm{II}(\mu)$ in (A.9.12), then we apply the Cauchy–Schwarz inequality to $\mathbb{E}\left[\mu^2(W_i) \mid \boldsymbol{Z}_i\right]$ and $\mathbb{E}\left[h_{ii} \mid \boldsymbol{Z}_i\right]$, which yields

$$\begin{aligned}
\mathrm{II}(\mu) &\leq 8\,\mathbb{E}_{Z_i}\left[\mathbb{E}_F\left[\mu^2(W_i)\right]\mathbb{E}\left[h_{ii} \mid Z_i\right]\right] \\
&\leq \frac{8(n-2)\mathbb{E}\left[\mu^2(W_i)\right]}{(n-1)(n-p-2)} + \frac{8\sqrt{\mathbb{E}\left[\boldsymbol{\Phi}^2\right]}}{n-p-2}\sqrt{\mathbb{E}_{\boldsymbol{Z}_i}\left[\mathbb{E}\left[\mu^4(W_i) \mid \boldsymbol{Z}_i\right]\right]} \\
&\leq \frac{8\sqrt{\mathbb{E}\left[\mu^4(X,Z)\right]}}{n-p-2}\left(1+\sqrt{\mathbb{E}\left[\boldsymbol{\Phi}^2\right]}\right) \qquad \text{(A.9.26)}
\end{aligned}$$

where in the above equality, $\boldsymbol{\Phi} \sim \chi^2_{p-1}$ and is independent from $\boldsymbol{Z}_{\text{-}i}$. Since $\mathbb{E}\left[\boldsymbol{\Phi}^2\right] \leq p^2$, under the assumption $\mathbb{E}\left[\mu^4(X,Z)\right] < \infty$, we obtain the following bound on $\mathrm{II}(\mu)$,

$$\mathrm{II}(\mu) = O\left(\frac{p}{n-p-2}\right). \qquad \text{(A.9.27)}$$

Replacing the μ function by $\mu^\star$ and applying the assumption $\mathbb{E}\left[(\mu^\star)^4(X,Z)\right] < \infty$, we can establish the same rate for $\mathrm{II}(\mu^\star)$. Shifting back to the n_2 notation, we finally establish (1.3.4), i.e.

$$f(\mu) - f_n^{\mathcal{T}}(\mu) = O\left(\frac{p}{n_2-p-2}\right).$$

□

Proposition 1.3.6

Proof of Proposition 1.3.6. From the proposition statement, we know the sufficient statistic $\boldsymbol{T}_m$ and $f_n^{\mathcal{T}}(\mu)$ are defined based on the batch $\mathcal{B}_m$ whose sample size is n_2. Again, we will abbreviate the notation dependence for simplicity, i.e. use a generic n instead of n_2, use $\boldsymbol{T}$ and $\boldsymbol{Z}$ instead of $\boldsymbol{T}_m$ and $\boldsymbol{Z}_m$, as we did in the proof of Proposition 1.3.5. Following the derivations up to (A.9.7) in the

proof of Proposition 1.3.5, it suffices to deal with the following term:

$$\Pi(\mu) := \mathbb{E}_{\boldsymbol{Z}}\left[\mathrm{Var}_F\left(\mu(W_i)\right)\mathbb{E}_{\boldsymbol{T}|\boldsymbol{Z}}\left[\chi^2(F^{\boldsymbol{T}}\|F)\right]\right].$$

where F denotes the conditional distribution $X_i|\boldsymbol{Z}$ and $F_{\boldsymbol{T}}$ denotes the conditional distribution $X_i|\boldsymbol{Z},\boldsymbol{T}$. Below we will consider quantifying the χ^2 divergence between $F_{\boldsymbol{T}}$ and F, Let k_1, k_2 be $W_{i,j-1}, W_{i,j+1}$ respectively, we can write down the probability mass function of $F_{\boldsymbol{T}}$ and F:

$$F : \mathbb{P}\left(X_i \mid \boldsymbol{Z}\right) = \prod_{k=1}^{K}\left(q(k,k_1,k_2)\right)^{\mathbb{1}\left\{X_i=k, W_{i,j-1}=k_1, W_{i,j+1}=k_1\right\}} \tag{A.9.28}$$

$$F_{\boldsymbol{T}} : \mathbb{P}\left(X_i \mid \boldsymbol{Z},\boldsymbol{T}\right) = \prod_{k=1}^{K}\left(\widehat{q}(k,k_1,k_2)\right)^{\mathbb{1}\left\{X_i=k, W_{i,j-1}=k_1, W_{i,j+1}=k_1\right\}} \tag{A.9.29}$$

where $\widehat{q}(k,k_1,k_2) = N(k,k_1,k_2)/N(:,k_1,k_2)$ and $N(:,k_1,k_2) = \sum_{i=1}^{n}\mathbb{1}_{\{W_{i,j-1}=k_1, W_{i,j+1}=k_2\}}$. Recall the definition of χ^2 divergence between two discrete distributions, we have

$$\chi^2(F_{\boldsymbol{T}}\|F) = \sum_{k=1}^{K}\frac{\left(\widehat{q}(k,k_1,k_2) - q(k,k_1,k_2)\right)^2}{q(k,k_1,k_2)}$$

Notice that

$$\mathbb{E}_{\boldsymbol{T}|\boldsymbol{Z}}\left[\widehat{q}(k,k_1,k_2)\right] = q(k,k_1,k_2), \quad \mathrm{Var}_{\boldsymbol{T}|\boldsymbol{Z}}\left(\widehat{q}(k,k_1,k_2)\right) = \frac{q(k,k_1,k_2)(1-q(k,k_1,k_2))}{N(:,k_1,k_2)}$$

hence we can calculate the following conditional expectation,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{T}|\boldsymbol{Z}}\left[\chi^2(F_{\boldsymbol{T}}\|F)\right] &= \sum_{k=1}^{K}\mathbb{E}_{\boldsymbol{T}|\boldsymbol{Z}}\left[\frac{(\widehat{q}(k,k_1,k_2)-q(k,k_1,k_2))^2}{q(k,k_1,k_2)}\right] \\
&= \sum_{k=1}^{K}\frac{q(k,k_1,k_2)(1-q(k,k_1,k_2))}{N(:,k_1,k_2)q(k,k_1,k_2)} \\
&= \sum_{k=1}^{K}\frac{K-1}{N(:,k_1,k_2)} \qquad \text{(A.9.30)}
\end{aligned}$$

where we use the fact $\sum_{k=1}^{K} q(k,k_1,k_2) = 1$ in the last equality. Now $\Pi(\mu)$ can be calculated as below.

$$\begin{aligned}
\Pi(\mu) &= \mathbb{E}_{\boldsymbol{Z}}\left[\mathrm{Var}_F\left(\mu(W_i)\right)\mathbb{E}_{\boldsymbol{T}|\boldsymbol{Z}}\left[\chi^2(F_{\boldsymbol{T}}\|F)\right]\right] \\
&= \mathbb{E}_{Z_i}\left[\mathrm{Var}_F\left(\mu(W_i)\right)\mathbb{E}\left[\mathbb{E}_{\boldsymbol{T}|\boldsymbol{Z}}\left[\chi^2(F_{\boldsymbol{T}}\|F)\right]\mid Z_i\right]\right] \\
&= \mathbb{E}_{Z_i}\left[\mathrm{Var}_F\left(\mu(W_i)\right)\mathbb{E}\left[\frac{K-1}{N(:,W_{i,j-1},W_{i,j+1})}\mid Z_i\right]\right] \\
&= \mathbb{E}_{Z_i}\left[\mathrm{Var}_F\left(\mu(W_i)\right)\mathbb{E}\left[\frac{K-1}{1+N_{n-1}(W_{i,j-1},W_{i,j+1})}\mid Z_i\right]\right] \qquad \text{(A.9.31)}
\end{aligned}$$

where the second equality comes from the tower property of conditional expectation, the third equality holds due to (A.9.30) and $k_1 = W_{i,j-1}, k_2 = W_{i,j+1}$. In term of the fourth equality, we simply use the new notation that $N_{n-1}(W_{i,j-1},W_{i,j+1}) = \sum_{m\neq i}^{n}\mathbb{1}_{\{W_{m,j-1}=W_{i,j-1},W_{m,j+1}=W_{i,j+1}\}}$. Due to the independence among $i.i.d.$ samples $\{W_i\}_{i=1}^n$, we have, when conditioning on $Z_i = W_{i,\text{-}j}$

$$\mathbb{1}_{\{W_{m,j-1}=W_{i,j-1},\, W_{m,j+1}=W_{i,j+1}\}} \overset{i.i.d.}{\sim} \text{Bern}(q(W_{i,j-1},W_{i,j+1})), \quad m\in[n],\, m\neq i.$$

where $q(W_{i,j-1}, W_{i,j+1}) = \mathbb{P}\left(W_{j-1} = W_{i,j-1}, W_{j+1} = W_{i,j+1} \mid Z_i\right)$. Given a binomial random variable $B \sim \text{Bin}(n, q)$, we have the following fact by elementary calculus,

$$\mathbb{E}\left[\frac{1}{1+B}\right] = \frac{1}{(n+1)q} \cdot (1 - (1-q)^{n+1}). \tag{A.9.32}$$

hence we can bound the term $\Pi(\mu)$ as below

$$\begin{aligned}
\Pi(\mu) &= \frac{K-1}{n}\mathbb{E}_{Z_i}\left[\text{Var}_F\left(\mu(W_i)\right)\frac{1-(1-q(W_{i,j-1},W_{i,j+1}))^n}{q(W_{i,j-1},W_{i,j+1})}\right] && \text{(A.9.33)}\\
&\leq \frac{K-1}{n}\mathbb{E}_{Z_i}\left[\text{Var}_F\left(\mu(W_i)\right)\right]\frac{K^2}{K^2\min\{q(k_1,k_2)\}} && \text{(A.9.34)}\\
&\leq \frac{K^3}{n}\frac{\mathbb{E}\left[\mu^2(X,Z)\right]}{q_0} && \text{(A.9.35)}
\end{aligned}$$

where the equality holds due to (A.9.31) and (A.9.32). And in the second line, we lower bound $q(W_{i,j-1}, W_{i,j+1})$ by $\min\{q(k_1,k_2)\}$. Assuming $K^2\min\{\mathbb{P}\left(W_{j-1}=k_1, W_{j+1}=k_2\right)\}_{k_1,k_2\in[K]}\} \geq q_0 > 0$ gives us the third line. Then we can establish $\Pi(\mu) = O\left(\frac{K^3}{n}\right)$ (and similarly for $\Pi(\mu^\star)$) under the stated moment condition $\mathbb{E}\left[(\mu)^2(X,Z)\right], \mathbb{E}\left[(\mu^\star)^2(X,Z)\right] < \infty$. Finally, making use of the rate result about $\Pi(\mu), \Pi(\mu^\star)$ and following the same derivation as in Proposition 1.3.5, we have $f(\mu) - f_n^{\mathcal{T}}(\mu) = O\left(\frac{K^3}{n_2}\right)$, where we shift back to the n_2 notation. □

Ancillary lemmas

Lemma A.9.2 can be similarly derived as the expression for the Rényi divergence between two multivariate Gaussian distributions in Section 2.2.4 of Gil (2011). For completeness, we still present our proof below.

Lemma A.9.2. *The χ^2-divergence between $P : \mathcal{N}(\boldsymbol{a}_1, \Sigma_1)$ and $Q : \mathcal{N}(\boldsymbol{a}_2, \Sigma_2)$ equals the follow-*

ing whenever $2\Sigma_2 - \Sigma_1 \succ 0$*:*

$$\frac{|\Sigma_2|}{|\Sigma_1|^{\frac{1}{2}}|2\Sigma_2 - \Sigma_1|^{\frac{1}{2}}} \exp\left\{(\boldsymbol{a}_1 - \boldsymbol{a}_2)^\top (2\Sigma_2 - \Sigma_1)^{-1}(\boldsymbol{a}_1 - \boldsymbol{a}_2)\right\} - 1.$$

where $\boldsymbol{a}_1, \boldsymbol{a}_2 \in \mathbb{R}^d$*,* $\Sigma_1, \Sigma_2 \in \mathbb{R}^{d\times d}$*,* $\Sigma \succ 0$ *means a matrix* Σ *is positive definite and* $|\Sigma|$ *denotes its determinant.*

Proof of Lemma A.9.2. According to the definition of the χ^2-divergence, we have

$$\chi^2(P\|Q) := \int \left(\frac{dP}{dQ}\right)^2 dQ - 1 = \int \frac{p^2(x)}{q(x)} dx - 1, \tag{A.9.36}$$

where $p(x), q(x)$ are the Gaussian density functions. For multivariate Gaussian random variable with mean $\boldsymbol{a} \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$, the density function equals the following

$$f(x) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(x - \boldsymbol{a})^\top \Sigma^{-1}(x - \boldsymbol{a})\right\}, \quad x \in \mathbb{R}^d. \tag{A.9.37}$$

Hence we can calculate the χ^2-divergence as below,

$$\begin{aligned}\chi^2(P\|Q) &= \frac{|\Sigma_2|^{\frac{1}{2}}}{|\Sigma_1|} \int_{\mathbb{R}^d} \frac{1}{(2\pi)^{\frac{d}{2}}} \exp\left\{-\frac{1}{2}(x - \boldsymbol{a}_1)^\top (2\Sigma_1^{-1})(x - \boldsymbol{a}_1) + \frac{1}{2}(x - \boldsymbol{a}_2)^\top \Sigma_2^{-1}(x - \boldsymbol{a}_2)\right\} dx - 1 \\ &:= \frac{|\Sigma_2|^{\frac{1}{2}}}{|\Sigma_1|} \int_{\mathbb{R}^d} \frac{1}{(2\pi)^{\frac{d}{2}}} \exp\left\{\mathrm{II}_1 + \mathrm{II}_2 + \mathrm{II}_3\right\} dx - 1, \end{aligned} \tag{A.9.38}$$

where the first equality holds following the definition in (A.9.36) and the second equality comes from expanding the term in the exponent and combining, together with the following new nota-

tions:

$$\mathrm{II}_1 := -\frac{1}{2}x^\top(2\Sigma_1^{-1} - \Sigma_2^{-1})x \tag{A.9.39}$$

$$\mathrm{II}_2 := -\frac{1}{2}\cdot(-2x^\top)(2\Sigma_1^{-1}\boldsymbol{a}_1 - \Sigma_2^{-1}\boldsymbol{a}_2) \tag{A.9.40}$$

$$\mathrm{II}_3 := -\frac{1}{2}(2\boldsymbol{a}_1^\top\Sigma_1^{-1}\boldsymbol{a}_1 - \boldsymbol{a}_2\Sigma_2^{-1}\boldsymbol{a}_2) \tag{A.9.41}$$

Let $\Sigma_\star^{-1} = 2\Sigma_1^{-1} - \Sigma_2^{-1}$, $\Sigma_\star^{-1}\boldsymbol{a}_\star = 2\Sigma_1^{-1}\boldsymbol{a}_1 - \Sigma_2^{-1}\boldsymbol{a}_2$ (since we assume the positive definiteness of $2\Sigma_2 - \Sigma_1$, which implies $2\Sigma_1^{-1} - \Sigma_2^{-1} \succ 0$, hence $\Sigma_\star$ and $\boldsymbol{a}_\star$ are well-defined), then we have

$$(\Sigma_1^{-1}\Sigma_\star\Sigma_2^{-1})^{-1} = \Sigma_2\Sigma_\star^{-1}\Sigma_1 = 2\Sigma_2 - \Sigma_1 \tag{A.9.42}$$

$$2\Sigma_\star\Sigma_1^{-1} - \mathrm{I}_d = \Sigma_\star(2\Sigma_1^{-1} - \Sigma_\star^{-1}) = \Sigma_\star\Sigma_2^{-1} \tag{A.9.43}$$

$$\begin{aligned}\frac{1}{2}\boldsymbol{a}_\star^\top\Sigma_\star^{-1}\boldsymbol{a}_\star &= \frac{1}{2}(2\Sigma_1^{-1}\boldsymbol{a}_1 - \Sigma_2^{-1}\boldsymbol{a}_2)^\top\Sigma_\star(2\Sigma_1^{-1}\boldsymbol{a}_1 - \Sigma_2^{-1}\boldsymbol{a}_2)\\ &= 2\boldsymbol{a}_1^\top\Sigma_1^{-1}\Sigma_\star\Sigma_1^{-1}\boldsymbol{a}_1 - 2\boldsymbol{a}_1^\top\Sigma_1^{-1}\Sigma_\star\Sigma_2^{-1}\boldsymbol{a}_2 + \frac{1}{2}\boldsymbol{a}_2^\top\Sigma_2^{-1}\Sigma_\star\Sigma_2^{-1}\boldsymbol{a}_2\\ &= 2\boldsymbol{a}_1^\top\Sigma_1^{-1}\Sigma_\star\Sigma_1^{-1}\boldsymbol{a}_1 - 2\boldsymbol{a}_1^\top(2\Sigma_2 - \Sigma_1)^{-1}\boldsymbol{a}_2 + \frac{1}{2}\boldsymbol{a}_2^\top\Sigma_2^{-1}\Sigma_\star\Sigma_2^{-1}\boldsymbol{a}_2\end{aligned} \tag{A.9.44}$$

where the first and the second line hold by the definition of $\Sigma_\star$, the second equality holds since $\Sigma_\star^{-1} = \Sigma_\star^{-1}\Sigma_\star\Sigma_\star^{-1}$, the third line is simply from expanding and the last equality comes from (A.9.42). The above equations will be used a lot for the incoming derivations. Now the term in the

exponent can be written as

$$
\begin{aligned}
&\mathrm{II}_1 + \mathrm{II}_2 + \mathrm{II}_3 \\
&= -\frac{1}{2}(x^\top \Sigma_\star^{-1} x - 2x^\top \Sigma_\star^{-1} \boldsymbol{a}_\star) + \mathrm{II}_3 \\
&= -\frac{1}{2}(x - \boldsymbol{a}_\star)^\top \Sigma_\star^{-1}(x - \boldsymbol{a}_\star) + \frac{1}{2}\boldsymbol{a}_\star^\top \Sigma_\star^{-1} \boldsymbol{a}_\star - \frac{1}{2}(2\boldsymbol{a}_1^\top \Sigma_1^{-1} \boldsymbol{a}_1 - \boldsymbol{a}_2 \Sigma_2^{-1} \boldsymbol{a}_2) \\
&= \lambda(x) + \boldsymbol{a}_1^\top \Sigma_1^{-1}(2\Sigma_\star \Sigma_1^{-1} - \mathrm{I}_d)\boldsymbol{a}_1 - 2\boldsymbol{a}_1^\top (2\Sigma_2 - \Sigma_1)^{-1}\boldsymbol{a}_2 + \frac{1}{2}\boldsymbol{a}_2^\top \Sigma_2^{-1}(\Sigma_\star \Sigma_2^{-1} + \mathrm{I}_d)\boldsymbol{a}_2 \\
&= \lambda(x) + \boldsymbol{a}_1^\top \Sigma_1^{-1} \Sigma_\star \Sigma_2^{-1} \boldsymbol{a}_1 - 2\boldsymbol{a}_1^\top (2\Sigma_2 - \Sigma_1)^{-1}\boldsymbol{a}_2 + \boldsymbol{a}_2^\top \Sigma_2^{-1} \Sigma_\star \Sigma_1^{-1} \boldsymbol{a}_2 \\
&= \lambda(x) + \boldsymbol{a}_1^\top (2\Sigma_2 - \Sigma_1)^{-1}\boldsymbol{a}_1 - 2\boldsymbol{a}_1^\top (2\Sigma_2 - \Sigma_1)^{-1}\boldsymbol{a}_2 + \boldsymbol{a}_2^\top (2\Sigma_2 - \Sigma_1)^{-1}\boldsymbol{a}_2 \\
&= \lambda(x) + (\boldsymbol{a}_1 - \boldsymbol{a}_2)^\top (2\Sigma_2 - \Sigma_1)^{-1}(\boldsymbol{a}_1 - \boldsymbol{a}_2) := \lambda(x) + Q(\boldsymbol{a}_1, \boldsymbol{a}_2, \Sigma_1, \Sigma_2) \qquad \text{(A.9.45)}
\end{aligned}
$$

where the first equality holds by the definition of $\Sigma_\star$, $\boldsymbol{a}_\star$ and (A.9.39), (A.9.40), and the second equality holds due to (A.9.41). Regarding the third equality, we denote the term which depends on x by $\lambda(x) := -\frac{1}{2}(x - \boldsymbol{a}_\star)^\top \Sigma_\star^{-1}(x - \boldsymbol{a}_\star)$. As for the other constant terms in the third line, we simply combine (A.9.44) with the expansion of the term II_3 and rearrange them into three terms: $\boldsymbol{a}_1^\top(\cdot)\boldsymbol{a}_1$, $\boldsymbol{a}_1^\top(\cdot)\boldsymbol{a}_2$ and $\boldsymbol{a}_2^\top(\cdot)\boldsymbol{a}_2$. The fourth equality holds as a result of applying (A.9.43) twice and the last equality is simply from rearranging. Since only the term $\lambda(x)$ depends on x, we can

simplify the χ^2-divergence into the following

$$\begin{aligned}
\chi^2(P\|Q) &= \frac{|\Sigma_2|^{\frac{1}{2}}}{|\Sigma_1|}\exp\{Q(\boldsymbol{a}_1,\boldsymbol{a}_2,\Sigma_1,\Sigma_2)\}\int_{\mathbb{R}^d}\frac{1}{(2\pi)^{\frac{d}{2}}}\exp\{\lambda(x)\}\,dx-1\\
&= \frac{|\Sigma_2|^{\frac{1}{2}}}{|\Sigma_1|}\exp\{Q(\boldsymbol{a}_1,\boldsymbol{a}_2,\Sigma_1,\Sigma_2)\}\int_{\mathbb{R}^d}\frac{|\Sigma_\star|^{\frac{1}{2}}}{(2\pi)^{\frac{d}{2}}|\Sigma_\star|^{\frac{1}{2}}}\exp\{\lambda(x)\}\,dx-1\\
&= \frac{|\Sigma_2|^{\frac{1}{2}}}{|\Sigma_1|}|\Sigma_\star|^{\frac{1}{2}}\exp\{Q(\boldsymbol{a}_1,\boldsymbol{a}_2,\Sigma_1,\Sigma_2)\}-1\\
&= \frac{|\Sigma_2|}{|\Sigma_1|^{\frac{1}{2}}}|\Sigma_1^{-1}\Sigma_\star\Sigma_2^{-1}|^{\frac{1}{2}}\exp\{Q(\boldsymbol{a}_1,\boldsymbol{a}_2,\Sigma_1,\Sigma_2)\}-1\\
&= \frac{|\Sigma_2|}{|\Sigma_1|^{\frac{1}{2}}|2\Sigma_2-\Sigma_1|^{\frac{1}{2}}}\exp\left\{(\boldsymbol{a}_1-\boldsymbol{a}_2)^\top(2\Sigma_2-\Sigma_1)^{-1}(\boldsymbol{a}_1-\boldsymbol{a}_2)\right\}-1
\end{aligned}$$

where the first equality comes from (A.9.38) and (A.9.45), the third equality holds due to the definition of $\lambda(x)$ and the fact that $\int f(x)dx = 1$, where $f(x)$ is the Gaussian density function with the mean $\boldsymbol{a}_\star$ and covariance matrix $\Sigma_\star$), the fourth equality holds by making use of the properties of determinant and the last equality holds as a result of (A.9.42). □

A.10 Further simulation details

Source code for conducting floodgate in our simulation studies can be found at https://github.com/LuZhangH/floodgate.

A.10.1 Nonlinear model setup

Consider W which follows a Gaussian copula distribution with $X = W_{j_0}, Z = W_{\text{-}j_0}$ for some j_0 ($1 \le j_0 \le p$), i.e.,

$$W^{\text{latent}} \sim AR(1),\ W_j = 2\varphi(X_j^{\text{latent}}) - 1,\ \forall\, 1 \le j \le p. \tag{A.10.1}$$

Hence the marginal distribution for W_j is Unif$[-1, 1]$ (in fact, these are the inputs to the fitting methods we use in floodgate, not the AR(1) latent variables W^{latent}). We consider the following conditional model for Y given W, with standard Gaussian noise,

$$\mu^\star(x, z) = \mu^\star(w) := \sum_{j \in S^1} g_j(w_j) + \sum_{(j,l) \in S^2} g_j(w_j) g_l(w_l) + \sum_{(j,l,m) \in S^3} g_j(w_j) g_l(w_l) g_m(w_m) \tag{A.10.2}$$

where each function $g_j(x)$ is randomly chosen from the following:

$$\sin(\pi x), \cos(\pi x), \sin(\pi x/2), \cos(\pi x) I(x > 0), x \sin(\pi x), x, |x|, x^2, x^3, \exp(x) - 1. \tag{A.10.3}$$

S^1 basically contains the main effect terms, while S^2 contain the pairs of variables with first order interactions. Tuples of variables involving second order interaction are denoted by S^3. For a given amplitude, (A.10.2) is scaled by the amplitude value divided by $\sqrt{n}$.

Now we describe the construction of S^1, S^2, S^3. First we randomly pick 30 variables into $S_\star$ and initialize $S_{\text{wl}} = S_\star$. 15 of them will be randomly assigned into S^1 and removed from S_{wl}. Among these 15 variables in S^1, we further choose 10 variables into 5 pairs randomly, which will be included in S^2. Regarding the other pairs in S^2, each time we randomly pick 2 variables from $S_\star$ with the unscaled weight being $2|S_{\text{wl}}|/|S_\star|$ for variables in S_{wl}, $|S_\star \setminus S_{\text{wl}}|/|S_\star|$ for the others, then add them as a pair into S^2. Once picked, the variables will be removed from S_{wl}. This process iterates until $|S_{\text{wl}}| \leq 5$. Regarding the construction of S^3, each time we randomly pick 3 variables from $S_\star$ with the unscaled weight being $1.5|S_{\text{wl}}|/|S_\star|$ for variables in S_{wl}, $|S_\star \setminus S_{\text{wl}}|/|S_\star|$ for the others, then add them as a tuple into S^3. Once picked, the variables will be removed from S_{wl}. This process iterates until $|S_{\text{wl}}| = 0$.

A.10.2 Implementation details of fitting algorithms

Regarding how to obtain the working regression function, there will be four different fitting algorithms for non-binary responses:

- *LASSO*: We fit a linear model by 10-fold cross-validated LASSO and output a working regression function. The subsequent inference step will be quite fast. First, as implied by Algorithm 1, $L_n^\alpha(\mu)$ will be set to zero for unselected variables, without any computation. Second, as alluded to in Section 1.2.4, we can analytically compute the conditional quantities in Algorithm 1.

- *Ridge*: We again use 10-fold cross-validation to choose the penalty parameter for Ridge regression. It is also fast to perform floodgate on, due to the second point mentioned above.

- *SAM*: We consider additive modelling, for example the sparse additive models (SAM) proposed in Ravikumar et al. (2009). As suggested by the name, it carries out sparse penalization and our method will assign $L_n^\alpha(\mu) = 0$ to unselected variables, as in *lasso*.

- *Random Forest*: Random forest (Breiman, 2001) is included as a purely nonlinear machine learning algorithm. While random forest do not generally conduct variable selection, we rank variables based on the heuristic importance measure and use the top 50 variables to run Algorithm 1 and set $L_n^\alpha(\mu) = 0$ for the remaining ones. Remark this is only for the concern of speed and does not have any negative impact on the inferential validity.

There are two additional fitting algorithms for binary responses: logistic regression with L1 regularization and L2 regularization, denoted by *Binom_LASSO* and *Binom_Ridge* respectively. Both use 10-fold cross-validation to choose the penalty parameter.

A.10.3 Implementation details of ordinary least squares

When the conditional model of $Y \mid X, Z$ is linear, i.e., $\mathbb{E}[Y \mid X, Z] = X\beta + Z\theta$ with $(\beta, \theta) \in \mathbb{R}^p$ the coefficients, the mMSE gap for X is closely related to its linear coefficient, formally

$$\mathcal{I} = |\beta| \sqrt{\mathbb{E}[\mathrm{Var}(X \mid Z)]}.$$

When the sample size n is greater than the number of variables p, ordinary least squares (OLS) can provide valid confidence intervals for β. However, there does not seem to exist a non-conservative way to transform the OLS confidence interval for β into a confidence bound for $|\beta|$. So instead, we provide OLS with further oracle information: the sign of β (we only compare half-widths of non-null covariates, and hence never construct OLS LCBs when $\beta = 0$). In particular, if [LCI, UCI] denotes a standard OLS 2-sided, equal-tailed $1 - 2\alpha$ confidence interval for β, then the OLS LCB for $\mathcal{I}$ we use is

$$\mathrm{LCB}_{\mathrm{OLS}} = \begin{cases} \mathrm{LCI}\sqrt{\mathbb{E}[\mathrm{Var}(X \mid Z)]} & \text{if } \beta > 0 \\ -\mathrm{UCI}\sqrt{\mathbb{E}[\mathrm{Var}(X \mid Z)]} & \text{if } \beta < 0 \end{cases} \tag{A.10.4}$$

which guarantees exact $1 - \alpha$ coverage of $\mathcal{I}$ for any nonzero value of β. We again emphasize that, in order to construct this interval, OLS uses the oracle information of the sign of β (this information is not available to floodgate in our simulations).

A.10.4 Plots deferred from the main paper

Effect of sample splitting proportion

The corresponding coverage plots of Figure 1.1 are given in Figure A.1. Figures A.2 and A.3 are additional plots with different simulation parameters specified in the captions. Figures A.1 and A.3 show that in the simulations in Section 1.4.2, the coverage of floodgate is consistently at or above the

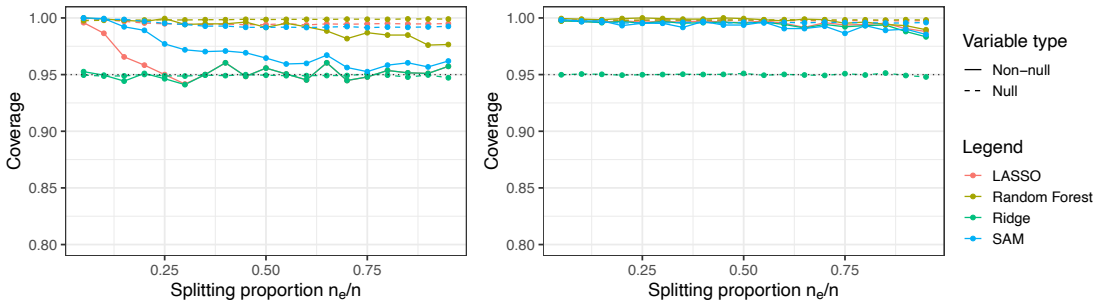


Figure A.1: Coverage for the the linear-$\mu^\star$ (left) and nonlinear-$\mu^\star$ (right) simulations of Section 1.4.2. The coefficient amplitude is chosen to be 10 for the left panel and the sample size n equals 3000 in the right panel; see Section 1.4.1 for remaining details. Standard errors are below 0.007 (left) and 0.003 (right).

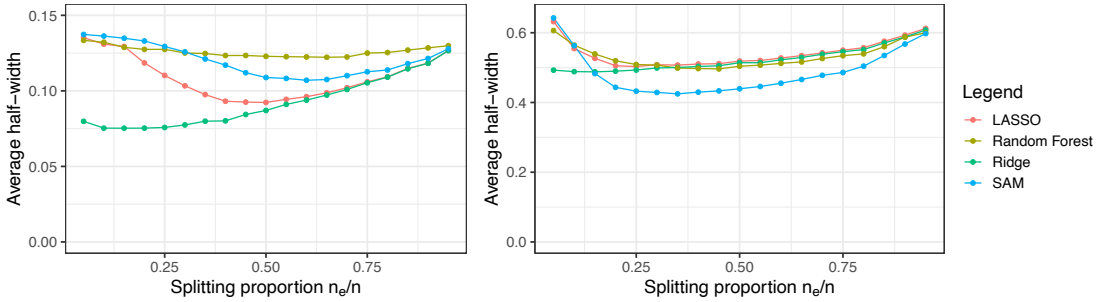


Figure A.2: Average half-widths for the linear-$\mu^\star$ (left) and nonlinear-$\mu^\star$ (right) simulations of Section 1.4.2. The coefficient amplitude is chosen to be 5 for the left panel and the sample size n equals 1000 in the right panel; see Section 1.4.1 for remaining details. Standard errors are below 0.002 (left) and 0.01 (right).

nominal 95% level.

Effect of covariate dimension

The corresponding coverage plots of Figure 1.2 are given in Figure A.4. Figures A.5 and A.6 are additional plots with different simulation parameters specified in the captions. Figures A.4 and A.6 show that in these simulations, the coverage of floodgate is consistently at or above the nominal 95% level.

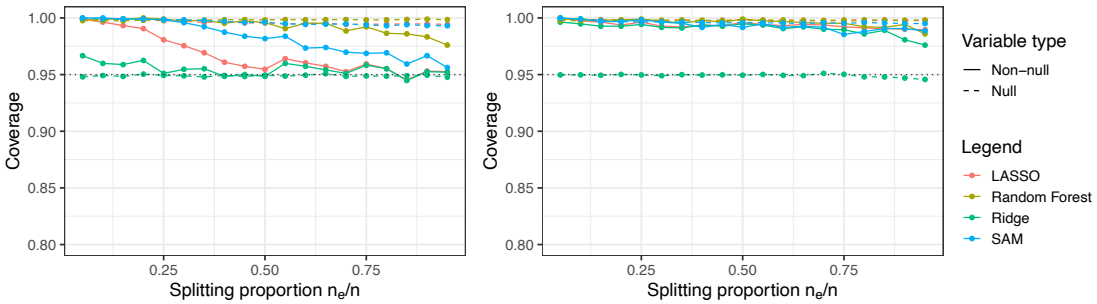


Figure A.3: Coverage for the the linear-$\mu^\star$ (left) and nonlinear-$\mu^\star$ (right) simulations of Section 1.4.2. The coefficient amplitude is chosen to be 5 for the left panel and the sample size n equals 1000 in the right panel; see Section 1.4.1 for remaining details. Standard errors are below 0.006 (left) and 0.004 (right).

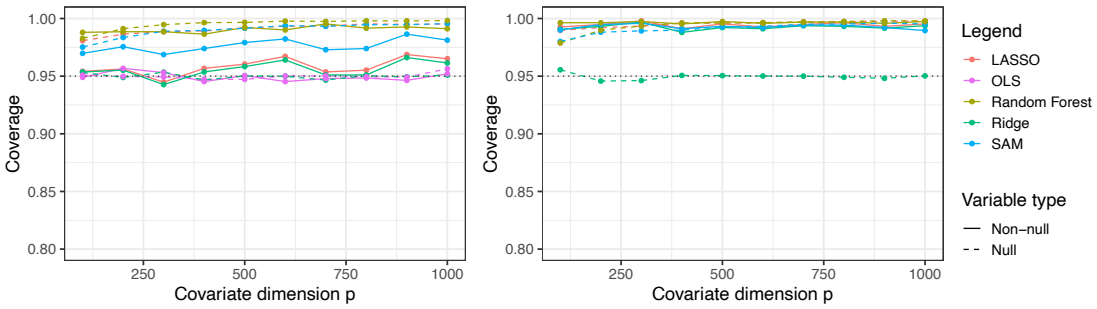


Figure A.4: Coverage for the linear-$\mu^\star$ (left) and nonlinear-$\mu^\star$ (right) simulations of Section 1.4.3. OLS is run on the full sample. p is varied on the x-axis; see Section 1.4.1 for remaining details. Standard errors are below 0.006 (left) and 0.004 (right).

Comparison with Williamson et al. (2020)

The corresponding coverage plot of Figure 1.3 is given in Figure A.7, where we see both methods have coverages above the nominal level. In addition to the example in Section 1.4.4, we also compare floodgate with W20b in the higher-dimensional setting of the left panel of Figure 1.2. Due to the computational challenge of running Williamson et al. (2020)'s method, we only consider the two most efficient algorithms (LASSO and Ridge) among the four described in Appendix A.10.2. Figure A.8 shows W20b to have slightly less consistent coverage than floodgate, but also reinforces the

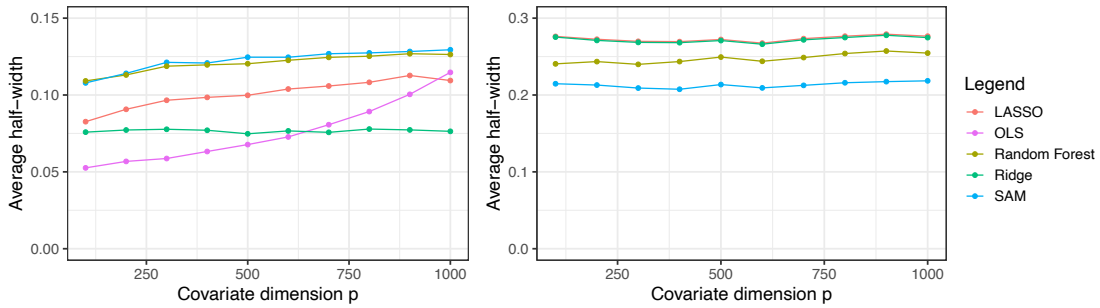


Figure A.5: Average half-widths for the linear-$\mu^\star$ (left) and nonlinear-$\mu^\star$ (right) simulations of Section 1.4.3. The splitting proportion is chosen to be 0.25 for the left panel and the sample size n equals 3000 in the right panel. p is varied on the x-axis; see Section 1.4.1 for remaining details. Standard errors are below 0.002 (left) and 0.005 (right).

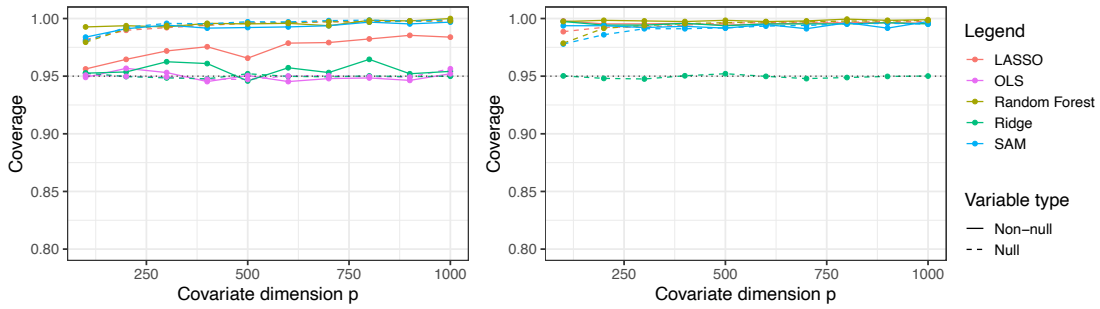


Figure A.6: Coverage for the linear-$\mu^\star$ (left) and nonlinear-$\mu^\star$ (right) simulations of Section 1.4.3. The splitting proportion is chosen to be 0.25 for the left panel and the sample size n equals 3000 in the right panel. p is varied on the x-axis; see Section 1.4.1 for remaining details. Standard errors are below 0.006 (left) and 0.004 (right).

general picture from the lower-dimensional simulation in Section 1.4.4 that W20b's LCBs are quite close to zero compared with floodgate's.

Robustness

Figure A.9 studies the robustness of floodgate for a nonlinear $\mu^\star$. We see the coverage being rather conservative for the non-null variables, reflecting the coverage-protective gap between $f(\mu)$ and $f(\mu^\star) = \mathcal{I}$. Figure A.10 shows that in the simulations of linear models and nonlinear models, the

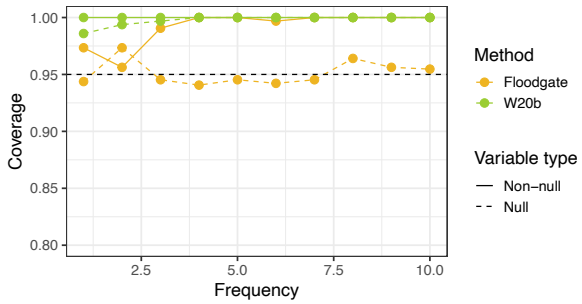


Figure A.7: Coverage for floodgate and W20b in the sine function simulation of Section 1.4.4. The frequency λ is varied on the x-axis, and the dotted black line in the plot shows the nominal coverage level $1 - \alpha$. The results are averaged over 640 independent replicates, and the standard errors are below 0.006.

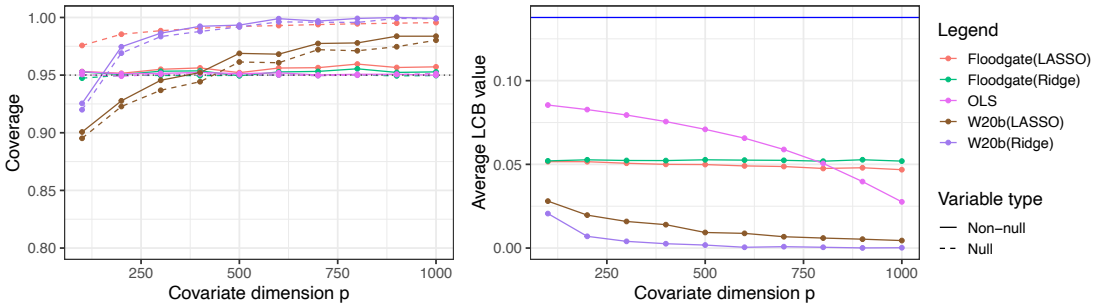


Figure A.8: Coverage (left) and average LCB values (right) for floodgate, W20b, and OLS (run on the full sample) in the linear-$\mu^\star$ simulation of Section 1.4.4. p is varied on the x-axis, and the solid blue line in the right-hand plot shows the value of $\mathcal{I}$; see Section 1.4.1 for remaining details. The results are averaged over 640 independent replicates, and the standard errors are below 0.012 (left) and 0.004 (right).

average half-width of floodgate is robust to estimation error in $P_{X|Z}$.

Co-sufficient floodgate

In this section, we demonstrate the performance of co-sufficient floodgate in a linear setting. Figure A.11 tells a similar story as Figure 1.6 in Section 1.4.7. Note that despite the linearity of the true model in Figure A.11, the LASSO performs poorly because the true model is quite dense (30 of the 50 covariates are non-null), which also explains why ridge regression performs so well.

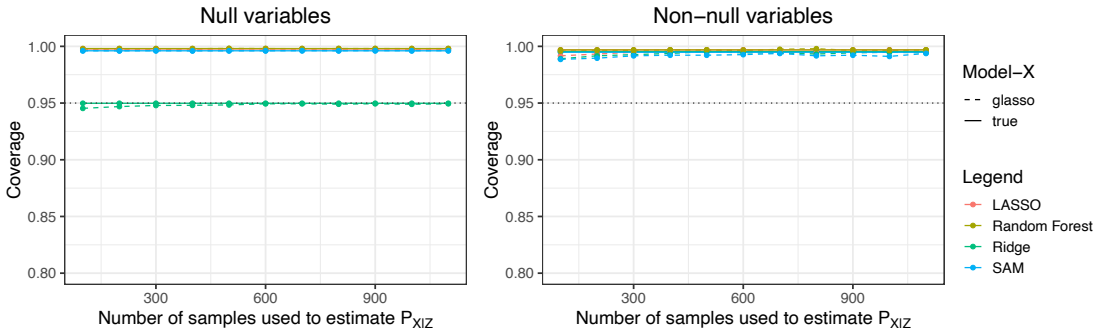


Figure A.9: Coverage of null (left) and non-null (right) covariates when the covariate distribution is estimated in-sample for the nonlinear-$\mu^\star$ simulations of Section 1.4.5. See Section 1.4.1 for remaining details. Standard errors are below 0.001 (left) and 0.003 (right).

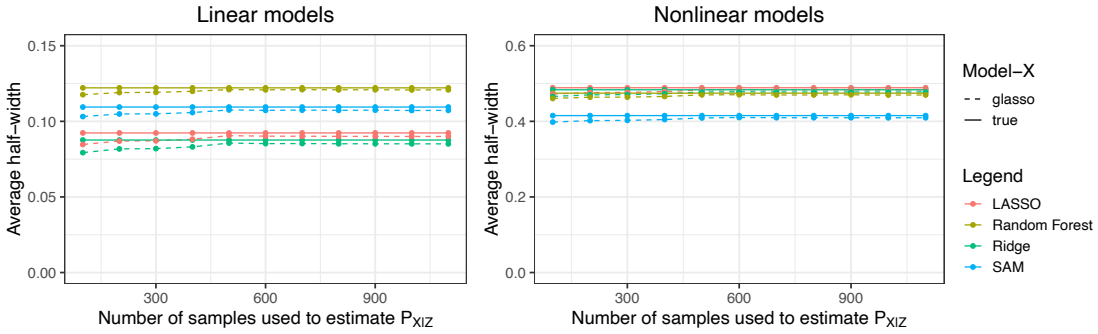


Figure A.10: Half-width plot of non-null covariates when the covariate distribution is estimated in-sample for the linear-$\mu^\star$ (left) and nonlinear-$\mu^\star$ (right) simulations of Section 1.4.5. See Section 1.4.1 for remaining details. Standard errors are below 0.002 (left) and 0.007 (right).

Effect of covariate dependence

In Figure A.12, we vary the covariate autocorrelation coefficient and plot the average half-widths of floodgate LCBs of non-null covariates under distributions with the linear (left panel) and the nonlinear (right panel) $\mu^\star$ described in Section 1.4.1, respectively. The left panel of Figure A.12 also includes a curve for OLS. Since $\mathcal{I}$ in a linear model is proportional to $\sqrt{\mathbb{E}\left[\mathrm{Var}\left(X \mid Z\right)\right]}$ which varies with the autocorrelation coefficient, we divided the half-widths in Figure A.12 by this quantity to make it easier to compare values across the x-axis. The main takeaway is that the effect of covariate

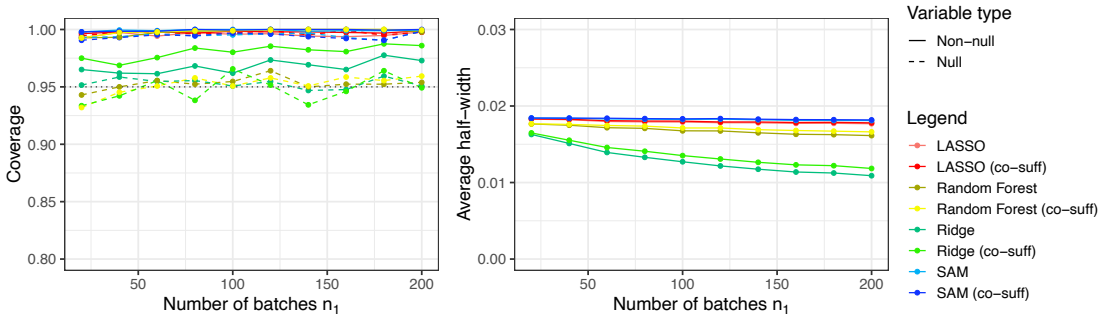


Figure A.11: Coverage (left) and average half-widths (right) for co-sufficient floodgate and original floodgate in the linear-$\mu^\star$ simulations. The number of batches n_1 is varied over the x-axis. See Section 1.4.1 and 1.4.7 for remaining details. Standard errors are below 0.008 (left) and 0.001 (right).

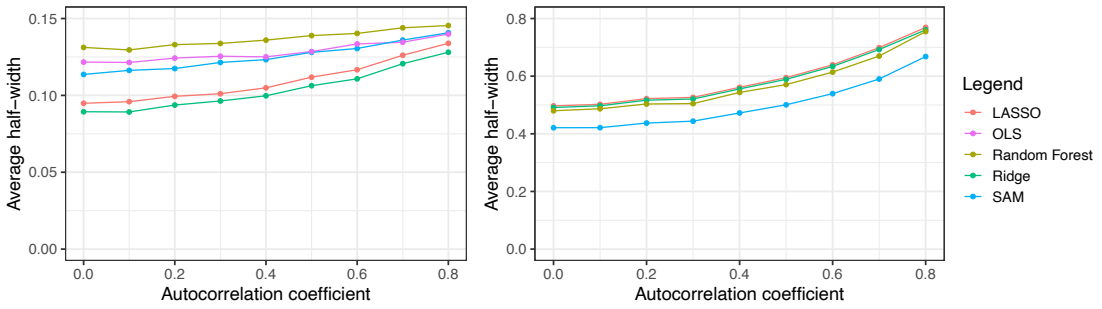


Figure A.12: Average half-widths for the linear-$\mu^\star$ (left) and nonlinear-$\mu^\star$ (right) simulations of Section A.10.4. The covariate dimension $p = 1000$ and the covariate autocorrelation coefficient is varied on the x-axis; see Section 1.4.1 for remaining details. Standard errors are below 0.002 (left) and 0.009 (right).

dependence on floodgate is somewhat mild until the dependence gets very large (> 0.5 correlation). This behavior is intuitive, and indeed we see a parallel trend in the curves for OLS inference in Figure A.12. The corresponding coverage plots of Figure A.12 are given in Figure A.13. Figures A.14 and A.15 are additional plots with a different covariate dimension specified in the captions. Figures A.13 and A.15 show that the coverage of floodgate is consistently at or above the nominal 95% level.

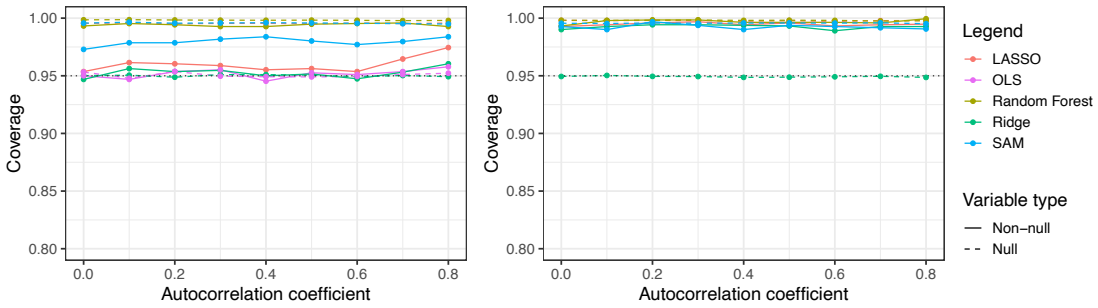


Figure A.13: Coverage for the linear-$\mu^\star$ (left) and nonlinear-$\mu^\star$ (right) simulations of Section A.10.4. The covariate dimension $p = 1000$ and the covariate autocorrelation coefficient is varied on the x-axis; see Section 1.4.1 for remaining details. Standard errors are below 0.006 (left) and 0.003 (right).

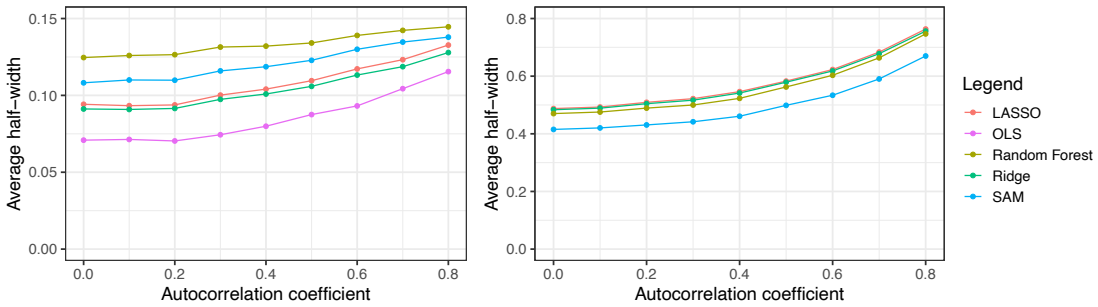


Figure A.14: Average half-widths for the linear-$\mu^\star$ (left) and nonlinear-$\mu^\star$ (right) simulations of Section A.10.4. The covariate dimension $p = 500$ and the covariate autocorrelation coefficient is varied on the x-axis; see Section 1.4.1 for remaining details. Standard errors are below 0.002 (left) and 0.01(right).

Effect of sample size

In Figures A.16 and A.17, we vary the sample size and plot the coverages and average half-widths of floodgate LCBs of non-null covariates under distributions with the linear and the nonlinear $\mu^\star$ described in Section 1.4.1, respectively. The main takeaway is that the accuracy of floodgate depends heavily on sample size. Note that in these plots, the signal size is scaled down by the square root of the sample size, so the *selection* problem is roughly getting no easier as the sample size increases, but we still see that floodgate can achieve much more accurate inference for larger sample sizes.

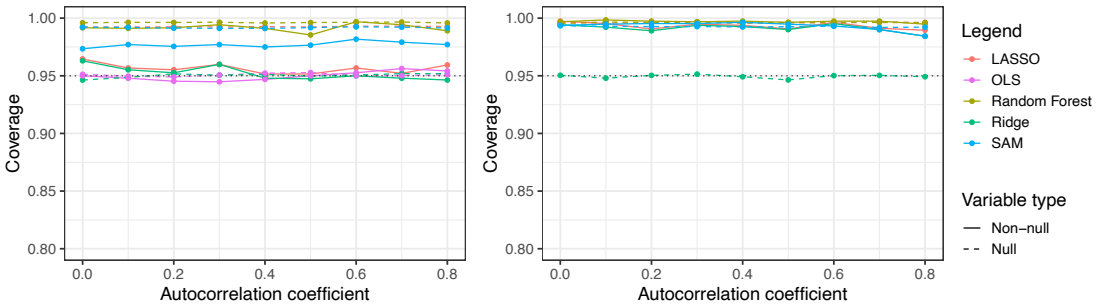


Figure A.15: Coverage for the linear-$\mu^\star$ (left) and nonlinear-$\mu^\star$ (right) simulations of Section A.10.4. The covariate dimension $p = 500$ and the covariate autocorrelation coefficient is varied on the x-axis; see Section 1.4.1 for remaining details. Standard errors are below 0.007 (left) and 0.004 (right).

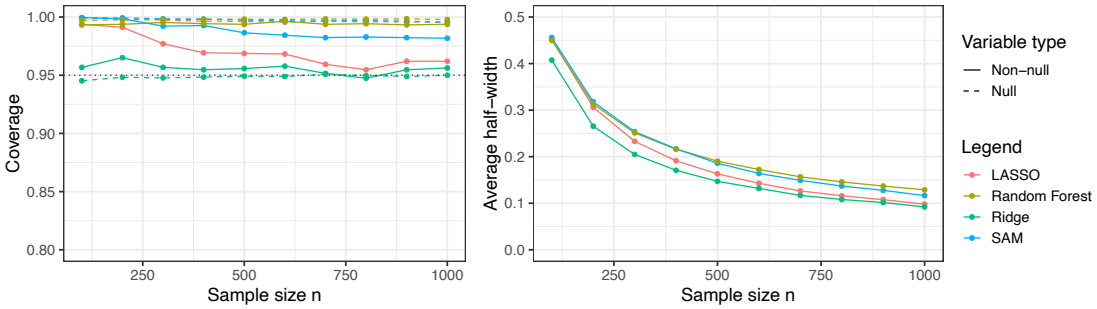


Figure A.16: Coverage (left) and average half-widths (right) for the linear-$\mu^\star$ simulations of Section A.10.4. The sample size n is varied on the x-axis; see Section 1.4.1 for remaining details. Standard errors are below 0.007 (left) and 0.003 (right).

A.11 Implementation details of genomics application

As mentioned in Section 1.2.6, the floodgate approach can be immediately generalized to conduct inference on the importance of a group of variables. This is practically useful in our application to the genomic data, where we group nearby SNPs whose effects are usually found challenging to be distinguished. Specifically, we use the exact same grouping at the same seven resolutions as Sesia et al. (2020b).

Regarding the genotype modelling, we consider the hidden Markov models (HMM) (Scheet &

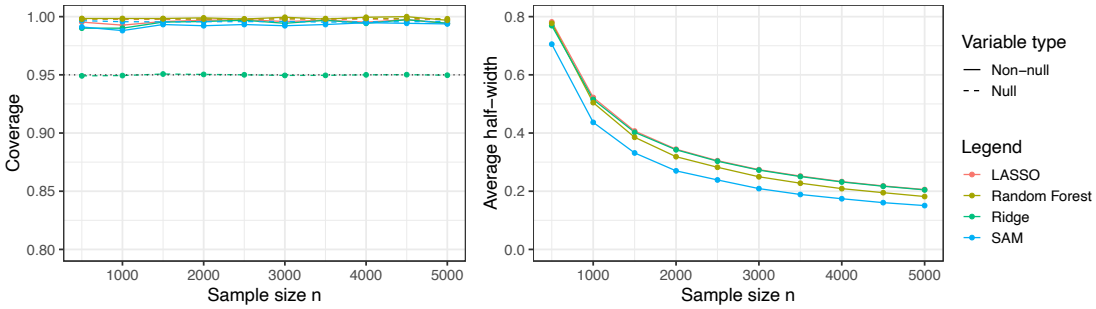


Figure A.17: Coverage (left) and average half-widths (right) for the nonlinear-$\mu^\star$ simulations of Section A.10.4. The sample size n is varied on the x-axis; see Section 1.4.1 for remaining details. Standard errors are below 0.004 (left) and 0.011 (right).

Stephens, 2006), as used in Sesia et al. (2019, 2020b), which provides a good description of the linkage disequilibrium (LD) structure. We obtain the fitted HMM parameters from Sesia et al. (2020b) on the UK Biobank data. Since HMM does not offer simple closed form expressions of the conditional quantities in Algorithm 1, we generate null copies of the genotypes and use them for the Monte Carlo analogue of floodgate. Below we simply describe the generating procedure. Under the HMM, we denote the covariates by W (genotypes or haplotypes) and the unobserved hidden states (local ancestries) by A, with the joint distribution over W denoted by P_W, the joint distribution over A denoted by P_A, which is the latent Markov chain model. For a given contiguous group of variables g_j, we can sample the null copy of W_{g_j} as follows:

(1) Marginalize out W_{g_j} and recompute the parameters of the new HMM $P_{\text{-}g_j}$ over $W_{\text{-}g_j}$.

(2) Sample the hidden states $A_{\text{-}g_j}$ by applying the forward-backward algorithm to $W_{\text{-}g_j}$, with the new HMM $P_{\text{-}g_j}$.

(3) Given $A_{\text{-}g_j}$, sample A_{g_j} according to the latent Markov chain model P_A.

(4) Sample $\widetilde{W}_{g_j}$ given A_{g_j} according to the emission distribution of the group g_j in the model of P_W.

To see why the above procedure produces a valid null copy of W_{g_j}, consider the following joint distribution, conditioning on $W_{\text{-}g_j}$

$$P_{\text{joint}} : (W_{g_j}, A_{g_j}, A_{\text{-}g_j}) \mid W_{\text{-}g_j}$$

If we sample $(\widetilde{W}_{g_j}, A_{g_j}, A_{\text{-}g_j})$ from the above joint conditional distribution, without looking at W_{g_j} or Y, then $\widetilde{W}_{g_j}$ has the same conditional distribution as W_{g_j}, given $W_{\text{-}g_j}$ and is conditionally independent from (W_{g_j}, Y), and thus is a valid null copy of W_{g_j}. Regarding how to sample from P_{joint}, we take advantage of the HMM structure and sample $A_{\text{-}g_j}, A_{g_j}, \widetilde{W}_{g_j}$ sequentially since

$$A_{g_j} \mid A_{\text{-}g_j}, W_{\text{-}g_j} \stackrel{d}{=} A_{g_j} \mid A_{\text{-}g_j}, \tag{A.11.1}$$

$$W_{g_j} \mid A_{g_j}, A_{\text{-}g_j}, W_{\text{-}g_j} \stackrel{d}{=} W_{g_j} \mid A_{g_j}. \tag{A.11.2}$$

Sampling from $A_{\text{-}g_j} \mid W_{\text{-}g_j}$ is feasible since $P_{\text{-}g_j}$ is still a HMM whenever the group g_j is contiguous. Under the HMM with particular parameterization in Scheet & Stephens (2006), the cost of the forward-backward algorithm can be reduced, see Sesia et al. (2020b) for more details. We remark that marginalizing out W_{g_j} only changes the transition structure around the group g_j and the special parameterization over other variables is still beneficial in terms of the computation cost. Sampling of A_{g_j} and $\widetilde{W}_{g_j}$ is computationally cheap due to (A.11.1) and (A.11.2). For a given number of null copies K, we will repeat the steps (2)-(4) for K times. But we remark the involving sampling probabilities only have to be computed once.

Regarding the quality control and data prepossessing of the UK Biobank data, we follow the Neale Lab GWAS with application 31063; details can be found on `http://www.nealelab.is/uk-biobank`. A few subjects withdrew consent and are removed from the analysis. Our final data set consisted of $361,128$ unrelated subjects and $591,513$ SNPs along 22 chromosomes.

For the platelet count phenotype, the analysis by Sesia et al. (2020b) makes several selections over the whole genome at seven different resolution levels. We focus on chromosome 12 and look at 248 selected groups from their analysis. For a given group of variables, we generate $K = 5$ null copies following the null copy generation procedure described above.

We applied floodgate with a 50-50 data split and fitted μ to the first half using the cross-validated LASSO as in Sesia et al. (2020b) and included both genotypes (SNPs from chromosomes 1–22) and the non-genetic variables sex, age and squared age. We centered Y by its sample mean from the first half of the data (the half used to fit μ) before applying floodgate. Although this changes nothing in theory, it does improve robustness as small biases in $\mu(X_i, Z_i) - \mathbb{E}\left[\mu(X_i, Z_i) \mid Z_i\right]$ would otherwise get multiplied by Y_i's mean in the computation of R_i in Algorithm 1.

Although our fitting of a linear model in no way changes the validity of floodgate's inference of the completely model-free mMSE gap, it does desensitize the LCB itself to the nonlinearities and interactions that partially motivated $\mathcal{I}$ as an object of inference in the first place. Our reasoning is purely pragmatic: as the universe of nonlinearities/interactions is exponentially larger than that of linear models, fitting such models requires either very strong nonlinear/interaction effects or prior knowledge of a curated set of likely nonlinearities/interactions. It is our understanding that nearly all genetic effects, linear and nonlinear/interaction alike, tend to be relatively weak, and the authors are not geneticists by training and thus lack the domain knowledge necessary to leverage the full flexibility of floodgate. Although we were already able to find substantial heritability for many blocks of SNPs with our default choice of the LASSO, it is our sincere hope and expectation that geneticists who specialize in the study of platelet count or similar traits would be able to find even more heritability using floodgate.

We report LCBs for all blocks simultaneously, although computationally we only actually run floodgate on those selected by Sesia et al. (2020b). Although their selection used all of the data (including the data we used for floodgate), it does not affect the marginal validity of the LCBs we re-

port, as explained in the last paragraph of Section 1.2.6.

B

Appendix of Chapter 2

B.1 Proofs for main text

Throughout the proofs, we denote $\epsilon(Y, X) := Y - \mathbb{E}[Y \mid X] = Y - \mu^\star(X)$ and simply have $\mathbb{E}[\epsilon(Y, X) \mid X] = 0$. By the law of total expectation, we have

$$\langle \epsilon(Y, X), g(X) \rangle = \mathbb{E}[g(X)\epsilon(Y, X)] = \mathbb{E}[g(X)\mathbb{E}[\epsilon(Y, X) \mid X]] = 0. \tag{B.1.1}$$

Due to the definition of $\mathcal{P}_{\mathcal{S}}$, we have, for random variables U and V.

$$\langle \mathcal{P}_{\mathcal{S}} U, V - \mathcal{P}_{\mathcal{S}} V \rangle = 0. \tag{B.1.2}$$

Let $\mathcal{S}^{\perp}$ be the orthogonal complement of $\mathcal{S}$. Due to the orthogonal decomposition, we have $U = \mathcal{P}_{\mathcal{S}} U + \mathcal{P}_{\mathcal{S}}^{\perp} U$ hence write $\mathcal{P}_{\mathcal{S}}^{\perp} = \mathbb{1} - \mathcal{P}_{\mathcal{S}}$ with $\mathbb{1}$ being the identity operator.

B.1.1 Proofs in Section 2.2

Proof of Lemma 2.2.2. Recall that $\epsilon(Y, X) = Y - \mathbb{E}[Y \mid X] = Y - \mu^{\star}(X)$. We simply have $\inf_{\mu(X) \in L_2(X)} \mathbb{E}\left[(Y - \mu(X))^2\right] = \mathbb{E}\left[(Y - \mu^{\star}(X))^2\right] = \mathbb{E}\left[\epsilon^2(Y, X)\right]$. Then $\mathcal{I}_{\mathcal{S}}^2$ can be rewritten as

$$\begin{aligned}
\mathcal{I}_{\mathcal{S}}^2 = & \inf_{\mu(X) \in \mathcal{S}} \mathbb{E}\left[(Y - \mu(X))^2\right] - \inf_{\mu(X) \in L_2(X)} \mathbb{E}\left[(Y - \mu(X))^2\right] \\
= & \inf_{\mu(X) \in \mathcal{S}} \mathbb{E}\left[(\epsilon(Y, X) + \mu^{\star}(X) - \mu(X))^2\right] - \mathbb{E}\left[\epsilon^2(Y, X)\right] \\
= & \inf_{\mu(X) \in \mathcal{S}} \{\mathbb{E}\left[(\mu^{\star}(X) - \mu(X))^2\right] + 2\mathbb{E}\left[\epsilon(Y, X)(\mu^{\star}(X) - \mu(X))\right] + \mathbb{E}\left[\epsilon^2(Y, X)\right]\} - \mathbb{E}\left[\epsilon^2(Y, X)\right] \\
= & \inf_{\mu(X) \in \mathcal{S}} \mathbb{E}\left[(\mu^{\star}(X) - \mu(X))^2\right] \\
= & ||\mathcal{P}_{\mathcal{S}}^{\perp} \mu^{\star}(X)||^2,
\end{aligned}$$

where the third equality holds by expansion, the fourth equality holds due to (B.1.1) and the cancellation of $\mathbb{E}\left[\epsilon^2(Y, X)\right]$, and the last equality is by the definition of projections. Therefore, we prove the concise expression $\mathcal{I}_{\mathcal{S}} = ||\mathcal{P}_{\mathcal{S}}^{\perp} \mu^{\star}(X)||$. □

Proof of Lemma 2.2.3. Recall that $\mathcal{P}_{\mathcal{S}}^{\perp} = \mathbb{1} - \mathcal{P}_{\mathcal{S}}$ with $\mathbb{1}$ being the identity operator. We bound

$f(\mu)$ as below

$$\begin{aligned} f(\mu) &= \left\langle 2Y - \mu(X), \mathcal{P}_{\mathcal{S}}^{\perp}\mu(X) \right\rangle \\ &= \left\langle 2\epsilon(Y,X) + 2\mathcal{P}_{\mathcal{S}}^{\perp}\mu^{\star}(X) - \mathcal{P}_{\mathcal{S}}^{\perp}\mu(X) + 2\mathcal{P}_{\mathcal{S}}\mu^{\star}(X) - \mathcal{P}_{\mathcal{S}}\mu(X), \mathcal{P}_{\mathcal{S}}^{\perp}\mu(X) \right\rangle \\ &= \left\langle 2\mathcal{P}_{\mathcal{S}}^{\perp}\mu^{\star}(X) - \mathcal{P}_{\mathcal{S}}^{\perp}\mu(X), \mathcal{P}_{\mathcal{S}}^{\perp}\mu(X) \right\rangle \\ &= -\left\langle \mathcal{P}_{\mathcal{S}}^{\perp}\mu^{\star}(X) - \mathcal{P}_{\mathcal{S}}^{\perp}\mu(X), \mathcal{P}_{\mathcal{S}}^{\perp}\mu^{\star}(X) - \mathcal{P}_{\mathcal{S}}^{\perp}\mu(X) \right\rangle + \left\langle \mathcal{P}_{\mathcal{S}}^{\perp}\mu^{\star}(X), \mathcal{P}_{\mathcal{S}}^{\perp}\mu^{\star}(X) \right\rangle \\ &= -||\mathcal{P}_{\mathcal{S}}^{\perp}\mu^{\star}(X) - \mathcal{P}_{\mathcal{S}}^{\perp}\mu(X)||^2 + \mathcal{I}_{\mathcal{S}}^2 \leq \mathcal{I}_{\mathcal{S}}^2, \end{aligned}$$

where the first equality holds by the definition of $f(\mu)$ in (2.2.3), the second equality holds by the definition of $\epsilon(Y,X)$ and the orthogonal decomposition $\mathcal{P}_{\mathcal{S}} + \mathcal{P}_{\mathcal{S}}^{\perp} = \mathbb{I}$ with $\mathbb{I}$ being the identity operator, the third equality holds due to (B.1.1) and (B.1.2), the fourth equality is by rearranging, the fifth equality holds due to Lemma 2.2.2, and the last equality holds by the non-negativeness of norms.

□

Proof of Theorem 2.2.4. Recall the expression of $L_n^{\alpha}(\mu)$. We notice

$$\{L_n^{\alpha}(\mu) \leq \mathcal{I}_{\mathcal{S}}^2\} = \left\{ \max\left\{ \bar{R} - \frac{z_{\alpha} s}{\sqrt{n}}, 0 \right\} \leq \mathcal{I}_{\mathcal{S}}^2 \right\} \supset \left\{ \bar{R} - \frac{z_{\alpha} s}{\sqrt{n}} \leq \mathcal{I}_{\mathcal{S}}^2 \right\}$$

due to the non-negativeness of $\mathcal{I}_{\mathcal{S}}^2$. Lemma 2.2.3 says $f(\mu) \leq \mathcal{I}_{\mathcal{S}}^2$. Thus it suffices to prove

$$1 - \alpha \leq \liminf_{n\to\infty} \mathbb{P}\left(\bar{R} - \frac{z_{\alpha} s}{\sqrt{n}} \leq f(\mu) \right) = \liminf_{n\to\infty} \mathbb{P}\left(\frac{\sqrt{n}(\bar{R} - f(\mu))}{s} \leq z_{\alpha} \right). \tag{B.1.3}$$

By construction in Algorithm 2, $\bar{R}$ and s are respectively the sample mean and sample standard deviation of i.i.d. random variables $R_i = \langle 2Y_i - \mu(X_i), \mu(X_i) - \mathcal{P}_{\mathcal{S}}\mu(X_i) \rangle$ whose expectation equals $f(\mu)$. Also note z_{α} is the $(1-\alpha)$th quantile of the standard normal distribution. Then

(B.1.3) immediately holds as a result of applying CLT to i.i.d. random variables $\{R_i\}_{i=1}^n$. □

B.1.2 Proofs in Section 2.3

Proof of Lemma 2.3.1. Consider the following minimization problem

$$\begin{aligned}
& \arg\min_{\nu(X)\in\mathcal{S}} ||\mu(X) - \nu(X)||^2 \\
= \; & \arg\min_{\lambda} \mathbb{E}\left[(\mu(X) - \lambda(t(X)))^2\right] \\
= \; & \arg\min_{\lambda} \mathbb{E}\left[(\mu(X) - \mathbb{E}\left[\mu(X) \,|\, t(X)\right] + \mathbb{E}\left[\mu(X) \,|\, t(X)\right] - \lambda(t(X)))^2\right] \\
= \; & \arg\min_{\lambda} \mathbb{E}\left[(\mu(X) - \mathbb{E}\left[\mu(X) \,|\, t(X)\right])^2\right] + \mathbb{E}\left[(\mathbb{E}\left[\mu(X) \,|\, t(X)\right] - \lambda(t(X)))^2\right] \\
= \; & \mathbb{E}\left[(\mu(X) - \mathbb{E}\left[\mu(X) \,|\, t(X)\right])^2\right] + \arg\min_{\lambda} \mathbb{E}\left[(\mathbb{E}\left[\mu(X) \,|\, t(X)\right] - \lambda(t(X)))^2\right],
\end{aligned}$$

where $\nu(X) = \lambda(t(X))$. For the above equalities, the third equality holds since

$$\begin{aligned}
& \mathbb{E}\left[(\mu(X) - \mathbb{E}\left[\mu(X) \,|\, t(X)\right])(\mathbb{E}\left[\mu(X) \,|\, t(X)\right] - \lambda(t(X)))\right] \\
= \; & \mathbb{E}\left[(\mathbb{E}\left[\mu(X) \,|\, t(X)\right] - \lambda(t(X)))\mathbb{E}\left[(\mu(X) - \mathbb{E}\left[\mu(X) \,|\, t(X)\right]) \,|\, t(X)\right]\right] = 0.
\end{aligned}$$

due to the law of total expectation. By the definition of $\mathcal{P}_{\mathcal{S}}$, we have

$$\begin{aligned}
\mathcal{P}_{\mathcal{S}}\mu(X) = \arg\min_{\nu(X)\in\mathcal{S}} ||\mu(X) - \nu(X)||^2 \quad &= \quad \arg\min_{\lambda} \mathbb{E}\left[(\mathbb{E}\left[\mu(X) \,|\, t(X)\right] - \lambda(t(X)))^2\right] \\
&= \quad \mathbb{E}\left[\mu(X) \,|\, t(X)\right].
\end{aligned}$$

□

Proof of Lemma 2.3.2. We have

$$\begin{aligned}
\mathbb{E}\left[(2Y-\mu(X))\mu(\widetilde{X})\right] &= \mathbb{E}\left[\mathbb{E}\left[(2Y-\mu(X))\mu(\widetilde{X})\,|\,X\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[(2Y-\mu(X))\,|\,t(X)\right]\mathbb{E}\left[\mu(\widetilde{X})\,|\,t(X)\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[(2Y-\mu(X))\,|\,t(X)\right]\mathbb{E}\left[\mu(X)\,|\,t(X)\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[(2Y-\mu(X))\mathbb{E}\left[\mu(X)\,|\,t(X)\right]\,|\,t(X)\right]\right] \\
&= \mathbb{E}\left[(2Y-\mu(X))\mathbb{E}\left[\mu(X)\,|\,t(X)\right]\right] = \langle 2Y-\mu(X), \mathcal{P}_{\mathcal{S}}\mu(X)\rangle\,,
\end{aligned}$$

where the first and fifth equalities hold by the law of total expectation, the second and third equalities hold due to (2.3.2), the fourth equality holds since $\mathbb{E}\left[\mu(X)\,|\,t(X)\right] \in \sigma(t(X))$, and the last equality holds by Lemma 2.3.1. Therefore we establish $\mathbb{E}\left[(2Y-\mu(X))(\mu(X)-\mu(\widetilde{X}))\right] = \langle 2Y-\mu(X), \mu(X)-\mathcal{P}_{\mathcal{S}}\mu(X)\rangle = f(\mu)$. □

B.1.3 Proofs in Section 2.3.5

Proof of Theorem 2.3.8. Recall the definition $f(\mu) = \langle 2Y-\mu(X), \mu(X)-\mathcal{P}_{\mathcal{S}}\mu(X)\rangle = \langle 2Y-\mu(X), \mathcal{P}_{\mathcal{S}^\perp}\mu(X)\rangle$. Due to the derivations in the proof of Theorem 2.2.4, it suffices to quantify the difference between the expectation of $R_i = \left\langle 2Y_i-\mu(X_i), \mathcal{P}_{12}^N\mu(X_i)\right\rangle$ and $f(\mu)$, i.e.,

$$|\mathbb{E}\left[R_i\right] - f(\mu)| = |\left\langle 2Y-\mu(X), \mathcal{P}_{12}^N\mu(X) - \mathcal{P}_{\mathcal{S}^\perp}\mu(X)\right\rangle|. \tag{B.1.4}$$

To bound $\mathcal{P}_{12}^N\mu(X) - \mathcal{P}_{\mathcal{S}^\perp}\mu(X)$, we apply the convergence result in Theorem 2.3.7 and (2.3.6) with $\mathcal{M} = \mathcal{S}^\perp, \mathcal{M}_1 = \mathcal{S}_1^\perp, \mathcal{M}_2 = \mathcal{S}_2^\perp$ and obtain

$$||\mathcal{P}_{12}^N\mu(X) - \mathcal{P}_{\mathcal{S}^\perp}\mu(X)|| \leq \rho^{2N-1}||\mu(X)||, \tag{B.1.5}$$

where $\rho = \sup\{\langle v_1, v_2\rangle : v_j \in \mathcal{M}_j \cap (\mathcal{M})^{\perp}, ||v_j|| \leq 1\}$. Applying the Cauchy–Schwarz inequality then (B.1.5) to (B.1.4), we have

$$|\mathbb{E}[R_i] - f(\mu)| \leq ||2Y - \mu(X)|| \cdot ||\mu(X)|| \rho^{2N-1} \leq 3c_0\rho^{2N-1}, \tag{B.1.6}$$

where the last inequality holds since $c_0 = \max\{\mathbb{E}\left[Y^2\right], \mathbb{E}\left[\mu^2(X)\right]\}$. Choosing the number of alternating steps N such that $2N - 1 \geq \frac{\log(\epsilon/3c_0)}{\log(\rho)}$, we have $3c_0\rho^{2N-1} \leq \epsilon$ since $\rho \leq 1$. Then following similar derivations in the proof of Theorem 2.2.4, we have

$$1 - \alpha \leq \liminf_{n\to\infty} \mathbb{P}\left(L_n^{\alpha}(\mu, N) \leq \mathcal{I}_{\mathcal{S}}^2 + |\mathbb{E}[R_i] - f(\mu)|\right) \leq \liminf_{n\to\infty} \mathbb{P}\left(L_n^{\alpha}(\mu, N) \leq \mathcal{I}_{\mathcal{S}}^2 + \epsilon\right)$$

thus establish (2.3.7) in Theorem 2.3.8. □

Proof of Lemma 2.3.9. For $N = 1$, we have

$$\begin{aligned} \text{RHS} &= \mathbb{1} - A(1, \mathcal{P}_{\mathcal{S}_1}, \mathcal{P}_{\mathcal{S}_1}) - A(1, \mathcal{P}_{\mathcal{S}_2}, \mathcal{P}_{\mathcal{S}_2}) + \mathcal{P}_{\mathcal{S}_2}\mathcal{P}_{\mathcal{S}_1} \\ &= \mathbb{1} - \mathcal{P}_{\mathcal{S}_1} - \mathcal{P}_{\mathcal{S}_2} + \mathcal{P}_{\mathcal{S}_2}\mathcal{P}_{\mathcal{S}_1} = (\mathbb{1} - \mathcal{P}_{\mathcal{S}_2})(\mathbb{1} - \mathcal{P}_{\mathcal{S}_1}) = \text{LHS}. \end{aligned}$$

We prove by induction. Assume (2.3.8) holds for N, and consider the case of $N + 1$:

$$\begin{aligned} \text{LHS} &= \mathcal{P}_{21}\mathcal{P}_{21}^N = \mathcal{P}_{21} + \sum_{s=1}^{N-1} \mathcal{P}_{21}\left(A(2s, \mathcal{P}_{\mathcal{S}_1}, \mathcal{P}_{\mathcal{S}_2}) + A(2s, \mathcal{P}_{\mathcal{S}_2}, \mathcal{P}_{\mathcal{S}_1})\right) \\ &\quad - \sum_{s=1}^{N} \mathcal{P}_{21}\left(A(2s-1, \mathcal{P}_{\mathcal{S}_1}, \mathcal{P}_{\mathcal{S}_1}) + A(2s-1, \mathcal{P}_{\mathcal{S}_2}, \mathcal{P}_{\mathcal{S}_2})\right) + \mathcal{P}_{21}(\mathcal{P}_{\mathcal{S}_2}\mathcal{P}_{\mathcal{S}_1})^N \\ &:= \text{II}_1 + \text{II}_2 + \text{II}_3 + \text{II}_4. \end{aligned}$$

First we have $\mathrm{II}_1 = \mathbb{1} - A(1, \mathcal{P}_{\mathcal{S}_1}, \mathcal{P}_{\mathcal{S}_1}) - A(1, \mathcal{P}_{\mathcal{S}_2}, \mathcal{P}_{\mathcal{S}_2}) + \mathcal{P}_{\mathcal{S}_2}\mathcal{P}_{\mathcal{S}_1}$. As for II_2, we have

$$
\begin{aligned}
\mathrm{II}_2 = & \sum_{s=1}^{N-1} (\mathbb{1} - \mathcal{P}_{\mathcal{S}_1} - \mathcal{P}_{\mathcal{S}_2} + \mathcal{P}_{\mathcal{S}_2}\mathcal{P}_{\mathcal{S}_1}) \left(A(2s, \mathcal{P}_{\mathcal{S}_1}, \mathcal{P}_{\mathcal{S}_2}) + A(2s, \mathcal{P}_{\mathcal{S}_2}, \mathcal{P}_{\mathcal{S}_1})\right) \\
= & \sum_{s=1}^{N-1} \left(A(2s, \mathcal{P}_{\mathcal{S}_1}, \mathcal{P}_{\mathcal{S}_2}) + A(2s, \mathcal{P}_{\mathcal{S}_2}, \mathcal{P}_{\mathcal{S}_1}) + A(2s+1, \mathcal{P}_{\mathcal{S}_2}, \mathcal{P}_{\mathcal{S}_2}) + A(2s+2, \mathcal{P}_{\mathcal{S}_2}, \mathcal{P}_{\mathcal{S}_1})\right) \\
& - \sum_{s=1}^{N-1} \left(A(2s, \mathcal{P}_{\mathcal{S}_1}, \mathcal{P}_{\mathcal{S}_2}) + A(2s+1, \mathcal{P}_{\mathcal{S}_1}, \mathcal{P}_{\mathcal{S}_1}) + A(2s+1, \mathcal{P}_{\mathcal{S}_2}, \mathcal{P}_{\mathcal{S}_2}) + A(2s, \mathcal{P}_{\mathcal{S}_2}, \mathcal{P}_{\mathcal{S}_1})\right) \\
= & \sum_{s=1}^{N-1} \left(A(2s+2, \mathcal{P}_{\mathcal{S}_2}, \mathcal{P}_{\mathcal{S}_1}) - A(2s+1, \mathcal{P}_{\mathcal{S}_1}, \mathcal{P}_{\mathcal{S}_1})\right) \\
= & \sum_{s=2}^{N} \left(A(2s, \mathcal{P}_{\mathcal{S}_2}, \mathcal{P}_{\mathcal{S}_1}) - A(2s-1, \mathcal{P}_{\mathcal{S}_1}, \mathcal{P}_{\mathcal{S}_1})\right),
\end{aligned}
$$

where the second equality holds since $\mathcal{P}_1 A(s, \mathcal{P}_1, \mathcal{P}_0) = A(s, \mathcal{P}_1, \mathcal{P}_0)$ and $\mathcal{P}_0 A(s, \mathcal{P}_1, \mathcal{P}_0) = A(s+1, \mathcal{P}_0, \mathcal{P}_0)$, the third equality comes from term cancelling and the last equality holds due to a change of summation index. And similarly for II_3, we obtain the following equations,

$$
\begin{aligned}
\mathrm{II}_3 = & -\sum_{s=1}^{N} (\mathbb{1} - \mathcal{P}_{\mathcal{S}_1} - \mathcal{P}_{\mathcal{S}_2} + \mathcal{P}_{\mathcal{S}_2}\mathcal{P}_{\mathcal{S}_1}) \left(A(2s-1, \mathcal{P}_{\mathcal{S}_1}, \mathcal{P}_{\mathcal{S}_1}) + A(2s-1, \mathcal{P}_{\mathcal{S}_2}, \mathcal{P}_{\mathcal{S}_2})\right) \\
= & -\sum_{s=1}^{N} \left(A(2s-1, \mathcal{P}_{\mathcal{S}_1}, \mathcal{P}_{\mathcal{S}_1}) + A(2s-1, \mathcal{P}_{\mathcal{S}_2}, \mathcal{P}_{\mathcal{S}_2}) + A(2s, \mathcal{P}_{\mathcal{S}_2}, \mathcal{P}_{\mathcal{S}_1}) + A(2s+1, \mathcal{P}_{\mathcal{S}_2}, \mathcal{P}_{\mathcal{S}_2})\right) \\
& + \sum_{s=1}^{N} \left(A(2s-1, \mathcal{P}_{\mathcal{S}_1}, \mathcal{P}_{\mathcal{S}_1}) + A(2s, \mathcal{P}_{\mathcal{S}_1}, \mathcal{P}_{\mathcal{S}_2}) + A(2s, \mathcal{P}_{\mathcal{S}_2}, \mathcal{P}_{\mathcal{S}_1}) + A(2s-1, \mathcal{P}_{\mathcal{S}_2}, \mathcal{P}_{\mathcal{S}_2})\right) \\
= & -\sum_{s=1}^{N} \left(A(2s+1, \mathcal{P}_{\mathcal{S}_2}, \mathcal{P}_{\mathcal{S}_2}) - A(2s, \mathcal{P}_{\mathcal{S}_1}, \mathcal{P}_{\mathcal{S}_2})\right).
\end{aligned}
$$

Simply for II_4, we have

$$\begin{aligned}\mathrm{II}_4 &= (\mathcal{P}_{\mathcal{S}_2}\mathcal{P}_{\mathcal{S}_1})^N - A(2N+1, \mathcal{P}_{\mathcal{S}_1}, \mathcal{P}_{\mathcal{S}_1}) - (\mathcal{P}_{\mathcal{S}_2}\mathcal{P}_{\mathcal{S}_1})^N + (\mathcal{P}_{\mathcal{S}_2}\mathcal{P}_{\mathcal{S}_1})^{N+1} \\ &= -A(2N+1, \mathcal{P}_{\mathcal{S}_1}, \mathcal{P}_{\mathcal{S}_1}) + (\mathcal{P}_{\mathcal{S}_2}\mathcal{P}_{\mathcal{S}_1})^{N+1}.\end{aligned}$$

Combining the expressions of $\mathrm{II}_1, \mathrm{II}_2, \mathrm{II}_3, \mathrm{II}_4$ yields the following thus finally establish (2.3.8):

$$\begin{aligned}\mathrm{LHS} = \mathrm{II}_1 + \mathrm{II}_2 + \mathrm{II}_3 + \mathrm{II}_4 &= \mathbb{1} - A(1, \mathcal{P}_{\mathcal{S}_1}, \mathcal{P}_{\mathcal{S}_1}) - A(1, \mathcal{P}_{\mathcal{S}_2}, \mathcal{P}_{\mathcal{S}_2}) + \mathcal{P}_{\mathcal{S}_2}\mathcal{P}_{\mathcal{S}_1} \\ &+ \sum_{s=2}^{N} (A(2s, \mathcal{P}_{\mathcal{S}_2}, \mathcal{P}_{\mathcal{S}_1}) - A(2s-1, \mathcal{P}_{\mathcal{S}_1}, \mathcal{P}_{\mathcal{S}_1})) - \sum_{s=1}^{N} (A(2s+1, \mathcal{P}_{\mathcal{S}_2}, \mathcal{P}_{\mathcal{S}_2}) - A(2s, \mathcal{P}_{\mathcal{S}_1}, \mathcal{P}_{\mathcal{S}_2})) \\ &- A(2N+1, \mathcal{P}_{\mathcal{S}_1}, \mathcal{P}_{\mathcal{S}_1}) + (\mathcal{P}_{\mathcal{S}_2}\mathcal{P}_{\mathcal{S}_1})^{N+1} \\ &= \mathbb{1} + \left(\mathcal{P}_{\mathcal{S}_2}\mathcal{P}_{\mathcal{S}_1} + \sum_{s=2}^{N} A(2s, \mathcal{P}_{\mathcal{S}_2}, \mathcal{P}_{\mathcal{S}_1}) + \sum_{s=1}^{N} A(2s, \mathcal{P}_{\mathcal{S}_1}, \mathcal{P}_{\mathcal{S}_2})\right) \\ &- \left(A(1, \mathcal{P}_{\mathcal{S}_1}, \mathcal{P}_{\mathcal{S}_1}) + \sum_{s=2}^{N} A(2s-1, \mathcal{P}_{\mathcal{S}_1}, \mathcal{P}_{\mathcal{S}_1}) + A(2N+1, \mathcal{P}_{\mathcal{S}_1}, \mathcal{P}_{\mathcal{S}_1})\right) \\ &- \left(A(1, \mathcal{P}_{\mathcal{S}_2}, \mathcal{P}_{\mathcal{S}_2}) + \sum_{s=1}^{N} A(2s+1, \mathcal{P}_{\mathcal{S}_2}, \mathcal{P}_{\mathcal{S}_2})\right) + (\mathcal{P}_{\mathcal{S}_2}\mathcal{P}_{\mathcal{S}_1})^{N+1} \\ &= \mathbb{1} + \sum_{s=1}^{N} (A(2s, \mathcal{P}_{\mathcal{S}_1}, \mathcal{P}_{\mathcal{S}_2}) + A(2s, \mathcal{P}_{\mathcal{S}_2}, \mathcal{P}_{\mathcal{S}_1})) - \sum_{s=1}^{N+1} A(2s-1, \mathcal{P}_{\mathcal{S}_1}, \mathcal{P}_{\mathcal{S}_1}) \\ &- \sum_{s=1}^{N+1} A(2s-1, \mathcal{P}_{\mathcal{S}_2}, \mathcal{P}_{\mathcal{S}_2}) + (\mathcal{P}_{\mathcal{S}_2}\mathcal{P}_{\mathcal{S}_1})^{N+1} = \mathrm{RHS}.\end{aligned}$$

□

Proof of Lemma 2.3.10. The proof is based on induction. First, when $t = 1$, we have

$$\begin{aligned}
\mathbb{E}\left[(2Y - \mu(X))\mu(\widetilde{X}_1^{(1,1)}, \widetilde{X}_2^{(1,0)}, Z)\right] &= \mathbb{E}\left[(2Y - \mu(X))\mu(\widetilde{X}_1^{(1,1)}, X_2, Z)\right] \\
&= \mathbb{E}\left[(2\mu^\star(X) - \mu(X))\mu(\widetilde{X}_1^{(1,1)}, X_2, Z)\right] \\
&= \mathbb{E}\left[(2\mu^\star(X) - \mu(X))\mathbb{E}\left[\mu(\widetilde{X}_1^{(1,1)}, X_2, Z) \,|\, X\right]\right] \\
&= \mathbb{E}\left[(2\mu^\star(X) - \mu(X))\mathbb{E}\left[\mu(\widetilde{X}_1^{(1,1)}, X_2, Z) \,|\, X_2, Z\right]\right] \\
&= \mathbb{E}\left[(2\mu^\star(X) - \mu(X))\mathbb{E}\left[\mu(X_1, X_2, Z) \,|\, X_2, Z\right]\right] \\
&= \mathbb{E}\left[(2\mu^\star(X) - \mu(X))A(1, \mathcal{P}_{\mathcal{S}_2}, \mathcal{P}_{\mathcal{S}_2})\mu(X)\right] \\
&= \mathbb{E}\left[(2Y - \mu(X))A(1, \mathcal{P}_{\mathcal{S}_2}, \mathcal{P}_{\mathcal{S}_2})\mu(X)\right],
\end{aligned}$$

where the first and sixth equalities hold by definition, the second and seventh equalities holds due to (B.1.1), the third equality holds by the law of total expectation, and the fourth and fifth equalities hold due to the construction of $\widetilde{X}_1^{(1,1)}$. Thus the second equation in (2.3.9) holds for $t = 1$. Similarly we also have

$$\begin{aligned}
&\mathbb{E}\left[(2Y - \mu(X))\mu(\widetilde{X}_1^{(1,1)}, \widetilde{X}_2^{(1,1)}, Z)\right] \\
= \;& \mathbb{E}\left[(2\mu^\star(X) - \mu(X))\mathbb{E}\left[\mu(\widetilde{X}_1^{(1,1)}, \widetilde{X}_2^{(1,1)}, Z) \,|\, X\right]\right] \\
= \;& \mathbb{E}\left[(2\mu^\star(X) - \mu(X))\mathbb{E}\left[\mathbb{E}\left[\mu(\widetilde{X}_1^{(1,1)}, \widetilde{X}_2^{(1,1)}, Z) \,|\, X, \widetilde{X}_1^{(1,1)}\right] | \, X\right]\right] \\
= \;& \mathbb{E}\left[(2\mu^\star(X) - \mu(X))\mathbb{E}\left[\mathbb{E}\left[\mu(\widetilde{X}_1^{(1,1)}, \widetilde{X}_2^{(1,1)}, Z) \,|\, Z, \widetilde{X}_1^{(1,1)}\right] | \, X\right]\right] \\
= \;& \mathbb{E}\left[(2\mu^\star(X) - \mu(X))\mathbb{E}\left[g(\widetilde{X}_1^{(1,1)}, Z) \,|\, X\right]\right] \\
= \;& \mathbb{E}\left[(2\mu^\star(X) - \mu(X))A(1, \mathcal{P}_{\mathcal{S}_2}, \mathcal{P}_{\mathcal{S}_2})g(X_1, Z)\right] \\
= \;& \mathbb{E}\left[(2\mu^\star(X) - \mu(X))A(1, \mathcal{P}_{\mathcal{S}_2}, \mathcal{P}_{\mathcal{S}_2})A(1, \mathcal{P}_{\mathcal{S}_1}, \mathcal{P}_{\mathcal{S}_1})\mu(X)\right] \\
= \;& \mathbb{E}\left[(2Y - \mu(X))A(2, \mathcal{P}_{\mathcal{S}_2}, \mathcal{P}_{\mathcal{S}_1})\mu(X)\right],
\end{aligned}$$

where $g(\widetilde{X}_1^{(1,1)}, Z)$ in the fourth line denotes $\mathbb{E}\left[\mu(\widetilde{X}_1^{(1,1)}, \widetilde{X}_2^{(1,1)}, Z) \mid Z, \widetilde{X}_1^{(1,1)}\right]$. Thus the equation line in (2.3.9) holds for $t = 1$. Similarly, we can prove the last two lines in (2.3.9) hold for the case $t = 1$. Now assume the four equations in (2.3.9) hold for t and consider the case of $t+1$, we have

$$
\begin{aligned}
& \mathbb{E}\left[(2Y - \mu(X))\mu(\widetilde{X}_1^{(1,t+1)}, \widetilde{X}_2^{(1,t+1)}, Z)\right] \\
= \; & \mathbb{E}\left[(2\mu^\star(X) - \mu(X))\mathbb{E}\left[\mu(\widetilde{X}_1^{(1,t+1)}, \widetilde{X}_2^{(1,t+1)}, Z) \mid X\right]\right] \\
= \; & \mathbb{E}\left[(2\mu^\star(X) - \mu(X))\mathbb{E}\left[\mathbb{E}\left[\mu(\widetilde{X}_1^{(1,t+1)}, \widetilde{X}_2^{(1,t+1)}, Z) \mid X, \widetilde{X}_1^{(1,1)}\right] \mid X\right]\right] \\
= \; & \mathbb{E}\left[(2\mu^\star(X) - \mu(X))\mathbb{E}\left[\mathbb{E}\left[\mu(\widetilde{X}_1^{(1,t+1)}, \widetilde{X}_2^{(1,t)}, Z) \mid Z, \widetilde{X}_1^{(1,1)}\right] \mid X\right]\right] \\
= \; & \mathbb{E}\left[(2\mu^\star(X) - \mu(X))\mathbb{E}\left[A(2t-1, \mathcal{P}_{\mathcal{S}_1}, \mathcal{P}_{\mathcal{S}_1})\mu(\widetilde{X}_1^{(1,1)}, X_2, Z) \mid X\right]\right] \\
= \; & \mathbb{E}\left[(2\mu^\star(X) - \mu(X))\mathbb{E}\left[\mu^{11}(\widetilde{X}_1^{(1,1)}, X_2, Z) \mid X_1, X_2, Z\right]\right] \\
= \; & \mathbb{E}\left[(2\mu^\star(X) - \mu(X))\mathbb{E}\left[\mu^{11}(X_1, X_2, Z) \mid X_2, Z\right]\right] \\
= \; & \mathbb{E}\left[(2\mu^\star(X) - \mu(X))A(1, \mathcal{P}_{\mathcal{S}_2}, \mathcal{P}_{\mathcal{S}_2})\mu^{11}(X_1, X_2, Z)\right] \\
= \; & \mathbb{E}\left[(2\mu^\star(X) - \mu(X))A(1, \mathcal{P}_{\mathcal{S}_2}, \mathcal{P}_{\mathcal{S}_2})A(2t-1, \mathcal{P}_{\mathcal{S}_1}, \mathcal{P}_{\mathcal{S}_1})\mu(X_1, X_2, Z)\right] \\
= \; & \mathbb{E}\left[(2Y - \mu(X))A(2t, \mathcal{P}_{\mathcal{S}_2}, \mathcal{P}_{\mathcal{S}_1})\mu(X)\right],
\end{aligned}
$$

where the fourth equality holds by treating $\widetilde{X}_1^{(1,1)}$ as $\widetilde{X}_1^{(2,0)}$ in Algorithm 5 and applying the fourth equation in (2.3.9) under t, and $\mu^{11}(\widetilde{X}_1^{(1,1)}, X_2, Z)$ in the fifth line denotes the conditional expectation $\mathbb{E}\left[A(2t-1, \mathcal{P}_{\mathcal{S}_1}, \mathcal{P}_{\mathcal{S}_1})\mu(\widetilde{X}_1^{(1,1)}, X_2, Z) \mid X\right]$. Thus the second equation in (2.3.9) holds for $t+1$. Similarly, we can show prove the other three equations for the case of $t+1$. Therefore, we are done by induction. □

Proof of Lemma 2.3.11. Start from simplifying $\mathcal{P}_{12}\mu(X)$ as below

$$
\begin{aligned}
& \mathcal{P}_{12}\mu(X) \\
= \; & \left(\mathbb{1} - A(1, \mathcal{P}_{\mathcal{S}_1}, \mathcal{P}_{\mathcal{S}_1}) - A(1, \mathcal{P}_{\mathcal{S}_2}, \mathcal{P}_{\mathcal{S}_2}) + \mathcal{P}_{\mathcal{S}_2}\mathcal{P}_{\mathcal{S}_1}\right)\mu(X) \\
= \; & \mu(X) - \mathbb{E}\left[\mu(X) \,|\, X_1, Z\right] - \mathbb{E}\left[\mu(X) \,|\, X_2, Z\right] + \mathcal{P}_{\mathcal{S}_2}\mathcal{P}_{\mathcal{S}_1}\mu(X) \\
= \; & \mu(X) - \mu_1(X_1, Z)\mathbb{E}\left[\mu_2(X_2, Z) \,|\, X_1, Z\right] - \mu_1(X_2, Z)\mathbb{E}\left[\mu_1(X_1, Z) \,|\, X_2, Z\right] + \mathcal{P}_{\mathcal{S}_2}\mathcal{P}_{\mathcal{S}_1}\mu(X) \\
= \; & \mu(X) - \mu_1(X_1, Z)\mathbb{E}\left[\mu_2(X_2, Z) \,|\, Z\right] - \mu_1(X_2, Z)\mathbb{E}\left[\mu_1(X_1, Z) \,|\, Z\right] + \mathbb{E}\left[\mu_1(X_1, Z) \,|\, Z\right]\mathbb{E}\left[\mu_2(X_2, Z) \,|\, Z\right] \\
= \; & \left(\mu_1(X_1, Z) - \mathbb{E}\left[\mu_1(X_1, Z) \,|\, Z\right]\right)\left(\mu_2(X_2, Z) - \mathbb{E}\left[\mu_2(X_2, Z) \,|\, Z\right]\right).
\end{aligned}
$$

Then we notice the following key result

$$
\begin{aligned}
\mathcal{P}_{\mathcal{S}_1}\mathcal{P}_{12}\mu(X) &= \mathcal{P}_{\mathcal{S}_1}\left(\mu_1(X_1, Z) - \mathbb{E}\left[\mu_1(X_1, Z) \,|\, Z\right]\right)\left(\mu_2(X_2, Z) - \mathbb{E}\left[\mu_2(X_2, Z) \,|\, Z\right]\right) \\
&= \left(\mu_1(X_1, Z) - \mathbb{E}\left[\mu_1(X_1, Z) \,|\, Z\right]\right)\mathcal{P}_{\mathcal{S}_1}\left(\mu_2(X_2, Z) - \mathbb{E}\left[\mu_2(X_2, Z) \,|\, Z\right]\right) \\
&= \left(\mu_1(X_1, Z) - \mathbb{E}\left[\mu_1(X_1, Z) \,|\, Z\right]\right)\left(\mathbb{E}\left[\mu_2(X_2, Z) \,|\, X_1, Z\right] - \mathbb{E}\left[\mu_2(X_2, Z) \,|\, Z\right]\right) \\
&= \left(\mu_1(X_1, Z) - \mathbb{E}\left[\mu_1(X_1, Z) \,|\, Z\right]\right)\left(\mathbb{E}\left[\mu_2(X_2, Z) \,|\, Z\right] - \mathbb{E}\left[\mu_2(X_2, Z) \,|\, Z\right]\right) = 0.
\end{aligned}
$$

Hence $(\mathbb{1} - \mathcal{P}_{\mathcal{S}_1})\mathcal{P}_{21}\mu(X) = \mathcal{P}_{12}\mu(X)$ and similarly $(\mathbb{1} - \mathcal{P}_{\mathcal{S}_2})\mathcal{P}_{12}\mu(X) = \mathcal{P}_{12}\mu(X)$ holds. Therefore $\mathcal{P}_{12}^N\mu(X) = \mathcal{P}_{12}\mu(X)$ holds for any $N \geq 1$, and we obtain $\mathcal{P}_{\mathcal{S}^\perp}\mu(X) = \left(\mu_1(X_1, Z) - \mathbb{E}\left[\mu_1(X_1, Z) \,|\, Z\right]\right)\left(\mu_2(X_2, Z) - \mathbb{E}\left[\mu_2(X_2, Z) \,|\, Z\right]\right)$. When $\mu_1(X_2, Z) =$

$X_1, \mu_2(X_2, Z) = X_2$, we have

$$\begin{aligned}
||\mathcal{P}_{\mathcal{S}^\perp}\mu(X)||^2 &= \mathbb{E}\left[(X_1 - \mathbb{E}[X_1 \mid Z])^2 (X_2 - \mathbb{E}[X_2 \mid Z])^2\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[(X_1 - \mathbb{E}[X_1 \mid Z])^2 (X_2 - \mathbb{E}[X_2 \mid Z])^2 \mid Z\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[(X_1 - \mathbb{E}[X_1 \mid Z])^2 \mid Z\right] \mathbb{E}\left[(X_2 - \mathbb{E}[X_2 \mid Z])^2 \mid Z\right]\right] \\
&= \mathbb{E}\left[\mathrm{Var}(X_1 \mid Z)\, \mathrm{Var}(X_2 \mid Z)\right],
\end{aligned}$$

thus establish Lemma 2.3.11. □

B.1.4 Proofs in Section 2.4

Proof of Lemma 2.4.1. Following similar derivations as in the proof of Lemma 2.2.3, we can reduce $f(\mu) = \langle 2Y - \mu(X), \mu(X) - \mathcal{P}_{\mathcal{G}}\mu(X)\rangle$ to $\langle 2\mu^\star(X) - \mu(X), \mu(X) - \mathcal{P}_{\mathcal{G}}\mu(X)\rangle$. Then we have

$$\begin{aligned}
f(\mu) &= \langle 2\mu^\star(X) - \mu(X), \mu(X) - \mathcal{P}_{\mathcal{G}}\mu(X)\rangle \\
&= \langle 2(\mu^\star(X) - \mathcal{P}_{\mathcal{G}}\mu^\star(X)) - (\mu(X) - \mathcal{P}_{\mathcal{G}}\mu(X)), \mu(X) - \mathcal{P}_{\mathcal{G}}\mu(X)\rangle \\
&\quad + \langle 2\mathcal{P}_{\mathcal{G}}\mu^\star(X) - \mathcal{P}_{\mathcal{G}}\mu(X), \mu(X) - \mathcal{P}_{\mathcal{G}}\mu(X)\rangle \\
&\leq \langle 2(\mu^\star(X) - \mathcal{P}_{\mathcal{G}}\mu^\star(X)) - (\mu(X) - \mathcal{P}_{\mathcal{G}}\mu(X)), \mu(X) - \mathcal{P}_{\mathcal{G}}\mu(X)\rangle := \mathrm{II}, \qquad \text{(B.1.7)}
\end{aligned}$$

where the second equality is by rearranging and the inequality holds due to the non-positiveness of $\langle 2\mathcal{P}_{\mathcal{G}}\mu^\star(X) - \mathcal{P}_{\mathcal{G}}\mu(X), \mu(X) - \mathcal{P}_{\mathcal{G}}\mu(X)\rangle$ (which is a result of the property of a convex cone (Ingram & Marsh, 1991)). When $\mu = \mu^\star$, we have $\langle 2\mathcal{P}_{\mathcal{G}}\mu^\star(X) - \mathcal{P}_{\mathcal{G}}\mu(X), \mu(X) - \mathcal{P}_{\mathcal{G}}\mu(X)\rangle = \langle \mathcal{P}_{\mathcal{G}}\mu^\star(X), \mu^\star(X) - \mathcal{P}_{\mathcal{G}}\mu^\star(X)\rangle = 0$ (Ingram & Marsh, 1991). Regarding the term II in (B.1.7),

we can rewrite it as

$$\begin{aligned}\text{II} &= -||(\mu^\star(X) - \mathcal{P}_\mathcal{G}\mu^\star(X)) - (\mu(X) - \mathcal{P}_\mathcal{G}\mu(X))||^2 + ||\mu^\star(X) - \mathcal{P}_\mathcal{G}\mu^\star(X)||^2 \\ &= -||(\mu^\star(X) - \mathcal{P}_\mathcal{G}\mu^\star(X)) - (\mu(X) - \mathcal{P}_\mathcal{G}\mu(X))||^2 + \mathcal{I}_\mathcal{G}^2 \leq \mathcal{I}_\mathcal{G}^2, \end{aligned} \tag{B.1.8}$$

where the first equality is by rearranging similarly as in the proof of Lemma 2.2.3, and the inequality holds due to the derivations in the proof of Lemma 2.2.2 and attains equality at $\mu^\star$. Combining (B.1.7) and (B.1.8), we prove $f(\mu) \leq \mathcal{I}_\mathcal{G}^2$. Since both the two inequalities in (B.1.7) and (B.1.8) are tight at $\mu^\star$, we have $f(\mu^\star) = \mathcal{I}_\mathcal{G}^2$. □

Proof of Theorem 2.4.6. Recall that $f(\mu) = \langle 2Y - \mu(X), \mu(X) - \mathcal{P}_\mathcal{G}\mu(X)\rangle$. Due to the derivations in the proof of Theorem 2.2.4, it suffices to quantify the difference between the expectation of $R_i = (2Y_i - \mu(X_i))\big(\mu(X_i) - g_N(X_i)\big)$ and $f(\mu)$. We bound it as below:

$$\begin{aligned} |\mathbb{E}\left[R_i\right] - f(\mu)| &= |\left\langle 2Y - \mu(X), g_N(X) - \mathcal{P}_\mathcal{G}\mu(X)\right\rangle| \\ &\leq ||2Y - \mu(X)||||g_N(X) - \mathcal{P}_\mathcal{G}\mu(X)|| \\ &\leq 3\sqrt{c_0}(\mathbb{E}\left[(g_N(X) - \mathcal{P}_\mathcal{G}\mu(X))^2\right])^{1/2}, \end{aligned}$$

where the first inequality is by the Cauchy–Schwarz inequality and the second inequality holds due to the Minkowski inequality and the definition of c_0. Then following similar derivations in the proof of Theorem 2.2.4, we have

$$1 - \alpha \leq \liminf_{n\to\infty} \mathbb{P}\left(L_n^\alpha(\mu, N) \leq \mathcal{I}_\mathcal{G}^2 + |\mathbb{E}\left[R_i\right] - f(\mu)|\right) \leq \liminf_{n\to\infty} \mathbb{P}\left(L_n^\alpha(\mu, N) \leq \mathcal{I}_\mathcal{G}^2 + \epsilon_N\right)$$

with $\epsilon_N = 3\sqrt{c_0}(\mathbb{E}\left[(g_N(X) - \mathcal{P}_\mathcal{G}\mu(X))^2\right])^{1/2}$. Hence Theorem 2.4.6 is proved. □

B.2 Methodological details deferred

B.2.1 A simple example in Section 2.3.1

Here we present a simple example with Gaussian covariates and linear transformation and show how to generate null samples.

Example B.2.1. Suppose $X \in \mathbb{R}^p$ and $X \sim \mathcal{N}(\mathbf{0}, \Sigma)$. The transformation function $t : \mathbb{R}^p \to \mathbb{R}^r$ is a linear function $t(x) = B^\top x$ where $B^\top = (\beta_1, \cdots, \beta_r) \in \mathbb{R}^{r \times p}$.

First denote $L = \Sigma^{1/2} B \in \mathbb{R}^{p \times r}$ and $\Pi_{\mathcal{S}} = L(L^\top L)^{-1} L^\top \in \mathbb{R}^{p \times p}$. Then we generate the null sample $\widetilde{X}$ as below:

Lemma B.2.2. *$\widetilde{X}$ satisfies the properties in* (2.3.2) *if it is constructed through the following procedure:*

1. *sample $\widetilde{Z}^0$ independently from $\mathcal{N}(\mathbf{0}, \mathrm{I}_p)$;*
2. *let $Z = \Sigma^{-1/2} X$ and $\widetilde{Z} = \Pi_{\mathcal{S}} Z + (\mathrm{I}_p - \Pi_{\mathcal{S}})\widetilde{Z}^0$;*
3. *set $\widetilde{X} = \Sigma^{1/2} \widetilde{Z}$.*

In the above procedure, we note

$$\Pi_{\mathcal{S}} Z = L(L^\top L)^{-1} L^\top \Sigma^{-1/2} X = L(L^\top L)^{-1} (\Sigma^{1/2} B)^\top \Sigma^{-1/2} X = L(L^\top L)^{-1} t(X),$$

hence the sampling of $\widetilde{X}$ is independent from (X, Y) conditioning on $t(X)$.

Proof of Lemma B.2.2. First note that

$$\begin{aligned} \widetilde{Z} &= \Pi_{\mathcal{S}} Z + (\mathrm{I}_p - \Pi_{\mathcal{S}})\widetilde{Z}^0 \\ &= L(L^\top L)^{-1} B^\top \Sigma^{1/2} \Sigma^{-1/2} X + (\mathrm{I}_p - \Pi_{\mathcal{S}})\widetilde{Z}^0 \\ &= L(L^\top L)^{-1} t(X) + (\mathrm{I}_p - \Pi_{\mathcal{S}})\widetilde{Z}^0. \end{aligned} \tag{B.2.1}$$

By construction, we immediate have $\widetilde{X} = \Sigma^{1/2}\widetilde{Z}$ satisfies $\widetilde{X} \perp\!\!\!\perp (X, Y) \mid t(X)$. Note we also have

$$\begin{aligned} Z = (\Pi_{\mathcal{S}} + \mathrm{I}_p - \Pi_{\mathcal{S}})Z &= \Pi_{\mathcal{S}}Z + (\mathrm{I}_p - \Pi_{\mathcal{S}})Z \\ &= L(L^\top L)^{-1}t(X) + (\mathrm{I}_p - \Pi_{\mathcal{S}})Z. \end{aligned} \tag{B.2.2}$$

Since both Z and $\widetilde{Z}$ follow the Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathrm{I}_p)$, we have

$$(\mathrm{I}_p - \Pi_{\mathcal{S}})\widetilde{Z}^0 \overset{d}{\sim} (\mathrm{I}_p - \Pi_{\mathcal{S}})Z. \tag{B.2.3}$$

and $(\mathrm{I}_p - \Pi_{\mathcal{S}})Z$ still follow a multivariate Gaussian distribution. We also have $t(X) = B^\top X$ are Gaussian random vectors and

$$\begin{aligned} \mathrm{Cov}\left(t(X), (\mathrm{I}_p - \Pi_{\mathcal{S}})Z\right) &= \mathrm{Cov}\left(L^\top Z, (\mathrm{I}_p - \Pi_{\mathcal{S}})Z\right) \\ &= L^\top \mathrm{I}_p(\mathrm{I}_p - \Pi_{\mathcal{S}}) = L^\top - L^\top L(L^\top L)^{-1}L^\top = O. \end{aligned} \tag{B.2.4}$$

Hence we have $t(X) \perp\!\!\!\perp (\mathrm{I}_p - \Pi_{\mathcal{S}})Z$ and

$$(\mathrm{I}_p - \Pi_{\mathcal{S}})Z \overset{d}{\sim} (\mathrm{I}_p - \Pi_{\mathcal{S}})Z \mid t(X). \tag{B.2.5}$$

By the construction of $\widetilde{Z}^0$, we have

$$(\mathrm{I}_p - \Pi_{\mathcal{S}})\widetilde{Z}^0 \overset{d}{\sim} (\mathrm{I}_p - \Pi_{\mathcal{S}})\widetilde{Z}^0 \mid t(X). \tag{B.2.6}$$

Therefore we finally have

$$(\mathrm{I}_p - \Pi_{\mathcal{S}})\widetilde{Z}^0 \mid t(X) \overset{d}{\sim} (\mathrm{I}_p - \Pi_{\mathcal{S}})Z \mid t(X) \tag{B.2.7}$$

and establish the property that $\widetilde{X} \mid t(X) \overset{d}{\sim} X \mid t(X)$. □

B.2.2 Details of MCMC sampling for Section 2.3.1

In this section, we present the computation details of generating null sample $\widetilde{X}^{(1)}$ of $X^{(1)}$ given $U^{(2)}$, more specifically, sampling from the conditional distribution of $X^{(1)}$ given $U^{(2)}$. Recall that the linear transformation between $X^{(1)}$ and $U^{(2)}$ defines a connected manifold

$$\mathcal{M} = \{x \in \mathbb{R}^{r_1} : q_k(x) = \sum_{j=1}^{r_1} \mathrm{W}^{(2)}_{kj} x_j - U^{(2)}_k = 0, k \in [r_2]\}$$

which is a linear subspace. Zappa et al. (2018) provides generic sampling algorithms for running MCMC on manifolds. We will follow the procedures to draw from the conditional distribution of $X^{(1)}$ given $U^{(2)}$ i.e., sample from the distribution of $X^{(1)}$ on the manifold $\mathcal{M}$. Specifically, let $Q_x \in \mathbb{R}^{r_1 \times r_2}$ be the transpose of the Jacobian of the overall constraint function $q : \mathbb{R}^{r_1} \to \mathbb{R}^{r_2}$. The entries of Q_x are

$$(Q_x)_{jk} = \frac{\partial q_k(x)}{\partial x_j} = \mathrm{W}^{(2)}_{kj},$$

which does not depend on x. Simply we have $Q_x = (\mathrm{W}^{(2)})^\top$. Obviously, Q_x has full rank r_2 everywhere on $\mathcal{M}$. By the implicit function theorem, we have the dimension of $\mathcal{M}$ is $r_1 - r_2$ and the tangent space $T_x := T_x\mathcal{M}$ at a point $x \in \mathcal{M}$ is well-defined. And the gradients $\{\nabla q_k(x)\}_{k=1}^{r_2}$ form a basis of the orthogonal space $T_x^\perp := T_x\mathcal{M}^\perp$. $\mathcal{M}$ inherits the metric from the ambient space $\mathbb{R}^{r_1}$ by restriction. The corresponding volume element is r_1-dimensional Hausdorff measure, which is denoted by $\sigma(dx)$. Denote the density function of $X^{(1)}$ as ρ with $\rho(dx) \propto p(x)\sigma(dx)$. We leverage the transformation of random variable density function formula to compute the expression of $p(x)$. With the above $U^{(2)}$, $\rho(dx)$ and $\mathcal{M}$, applying the MCMC surface sampling algorithm

produces a sequence of samples $\{\widetilde{X}_t^{(1)}\} \in \mathcal{M}$ satisfying the property that

$$\widetilde{X}_t^{(1)} \sim \rho \quad \Longrightarrow \quad \widetilde{X}_{t+1}^{(1)} \sim \rho.$$

According to Zappa et al. (2018), the relevant MCMC theory shows that if $\mathcal{M}$ is connected, compact and smooth, then the algorithm is geometrically ergodic. Here in our example, $\mathcal{M}$ is a linear subspace thus satisfies the conditions. The proposal process begins with a tangential move $x \to x + v$ with $v \in T_x$. We follow Zappa et al. (2018)'s choice and use an isotropic r_1-dimensional Gaussian distribution with some width s centered at x,

$$\varphi(v|x) = \frac{1}{(2\pi)^{r_1/2} s^{r_1}} \exp\left\{-\frac{||v||^2}{2s^2}\right\}.$$

And we generate v using an orthonormal basis for T_x. The orthonormal basis is found as the last $(r_2 - r_1)$ columns of the $r_1 \times r_1$ matrix in the QR decomposition of Q_x. Given x and v, the projection step is nothing for our example since the tangent space T_x is the same as the constraint manifold $\mathcal{M}$, i.e., $y = x + v$. To make sure the detailed balance condition, we also need to make the reverse proposal. That is, we have to choose $v' \in T_y$ so that $x = y + v' + w'$ with $w' \perp T_y$. In our example where $\mathcal{M}$ is a linear subspace, this is also quite straightforward since we can choose $v' = -v$ and $w' = 0$. Obviously, we have $v' \in T_y$ and $w' \perp T_y$ (here $T_x = T_y = \mathcal{M}$). Then, we compute an acceptance probability $a(y|x)$ using the Metropolis Hastings formula,

$$a(y \mid x) = \min\left\{1, \frac{p(y)\varphi(v' \mid x)}{p(x)\varphi(v \mid x)}\right\}.$$

Now we can summarize the above MCMC sampling procedures as Algorithm 12.

In addition to the algorithm above, we also provide details on the initialization. There are two choices:

Algorithm 12 MCMC surface sampling

Input: $x = \widetilde{X}_t^{(1)}$, the (unnormalized) density function $p(x)$.

1: Set $Q_x = (\mathrm{W}^{(2)})^\top$ and find orthonormal bases for T_x and $T_x^\perp$ using the QR decomposition of Q_x.

2: Generate $v \in T_x$ with $v \in \varphi(v \mid x)$ using the orthonormal basis of T_x and set the proposal $y = x + v$.

3: Accept the proposal with probability

$$a(y \mid x) = \min\left\{1, \frac{p(y)\varphi(-v \mid x)}{p(x)\varphi(v \mid x)}\right\}.$$

.

Output: Set $\widetilde{X}_{t+1}^{(1)} = y$ upon acceptance; otherwise set $\widetilde{X}_{t+1}^{(1)} = \widetilde{X}_t^{(1)}$.

(a) Set the original sample $\widetilde{X}^{(1)}$ to be the initialization point. Then we immediately have

$$\widetilde{X}_0^{(1)} \sim \rho \quad \Longrightarrow \quad \widetilde{X}_t^{(1)} \sim \rho, \quad \forall t.$$

But the multiple null samples $\{\widetilde{X}_t^{(1k)}\}_{k=1}^K$ will not be conditionally independent anymore.

(b) Randomly choose one initialization point from the linear subspace $\mathcal{M}$ using the the implementation in Van den Meersche et al. (2009). By construction, the multiple null samples $\{\widetilde{X}_t^{(1k)}\}_{k=1}^K$ will be conditionally independent. But $\widetilde{X}_t^{(1k)}$ will not follow the target distribution exactly.

Either choice can be used in practice since we expect the conditional dependence in (a) or the approximation error in (b) will vanish thus we obtain valid null samples, as the number of iterations goes to infinity.

B.2.3 Details in Section 2.3.2

Suppose $X \sim \mathbb{R}^p$ follows a mean zero multivariate Gaussian distribution with covariance matrix Σ_X and the perturbation random variable $\delta \sim \mathcal{N}(0, \Sigma_\delta)$. Following the notations in Example

B.2.1, we have

$$(X, \delta) \in \mathbb{R}^{2p} \overset{\text{Linear transformaiton}}{\longrightarrow} t(X, \delta) = B^\top (X, \delta) \in \mathbb{R}^p \tag{B.2.8}$$

where $B^\top = (\mathrm{I}_p, \mathrm{I}_p) \in \mathbb{R}^{p \times 2p}$ and the covariance matrix of (X, δ) is given by

$$\Sigma := \begin{bmatrix} \Sigma_X & O \\ O & \Sigma_\delta \end{bmatrix}. \tag{B.2.9}$$

Then the projection matrix $\Pi_{\mathcal{S}}$ equals

$$\Pi_{\mathcal{S}} = \begin{bmatrix} \Sigma_X^{1/2} \\ \Sigma_\delta^{1/2} \end{bmatrix} (\Sigma_X + \Sigma_\delta)^{-1} \begin{bmatrix} \Sigma_X^{1/2} & \Sigma_\delta^{1/2} \end{bmatrix}. \tag{B.2.10}$$

With such Σ and $\Pi_{\mathcal{S}}$, we can utilize the procedures in Appendix B.2.1 to generate null samples of (X, δ).

B.2.4 Details in Section 2.3.5

For completeness, we present the multiple chain versions of Algorithms 4 and 5 as Algorithms 13 and 14.

Algorithm 13 Sampling with multiple null replicates

Input: (X_1, X_2, Z), number of alternating steps N and number of null replicates K. For $k \in [K]$, let $\widetilde{X}_1^{(2,0,k)} = X_1, \widetilde{X}_2^{(2,0,k)} = X_2$.
  for t from 1 to N **do**
    Sample $\widetilde{X}_1^{(1,t,k)}$ conditional on $(\widetilde{X}_2^{(1,t-1,k)}, Z)$, independently for $k \in [K]$.
    Sample $\widetilde{X}_2^{(1,t,k)}$ conditional on $(\widetilde{X}_1^{(1,t,k)}, Z)$, independently for $k \in [K]$.
  end for
Output: $\{\{\widetilde{X}_1^{(1,t,k)}, \widetilde{X}_2^{(1,t,k)}\}_{t=1}^N\}_{k=1}^K$.

Algorithm 14 Sampling with multiple null replicates

Input: (X_1, X_2, Z), number of alternating steps N and number of null replicates K. For $k \in [K]$, let $\widetilde{X}_1^{(2,0,k)} = X_1, \widetilde{X}_2^{(2,0,k)} = X_2$.

for t from 1 to N **do**

Sample $\widetilde{X}_2^{(2,t)}$ conditional on $(\widetilde{X}_1^{(2,t-1)}, Z)$, independently for $k \in [K]$.

Sample $\widetilde{X}_1^{(2,t)}$ conditional on $(\widetilde{X}_2^{(2,t)}, Z)$, independently for $k \in [K]$.

end for

Output: $\{\{\widetilde{X}_1^{(2,t,k)}, \widetilde{X}_2^{(2,t,k)}\}_{t=1}^{N}\}_{k=1}^{K}$.

C

Appendix of Chapter 3

Supplementary material to

StarTrek: Combinatorial Variable Selection with False Discovery Rate Control

This document contains the supplementary material to the paper "StarTrek: Combinatorial Variable Selection with False Discovery Rate Control". Appendix C.1 presents the proofs of the FDR

control results. In Appendix C.2, we provide the proofs of two types of Cramér-type comparison bounds for Gaussian maxima. Appendix C.3 proves the Cramér-type deviation bounds for the Gaussian multiplier bootstrap. In Appendix C.4, we establish the validity and a power result of our test on the degree of a single node. Appendix C.5 contains some plots and tables deferred from the main paper.

C.1 Proofs for FDR control

In this section, we aim to prove Theorem 3.5.2. In order to prove the theorem, we need Lemma C.1.1 which is about the test of single node degree. Remark that this lemma proves the asymptotic validity of the test in Algorithm 7 and provides a power analysis. The signal strength condition is only required for the power analysis part. To see why Lemma C.1.1 is useful for establishing FDR control for our StarTrek procedure in Algorithm 8, we notice the following equivalence:

$$\{\psi_{j,\alpha} = 1\} = \{\alpha_j \leq \alpha\}, \tag{C.1.1}$$

where α is a given type-I error level, $\psi_{j,\alpha}$ is the test described in Algorithm 7, and α_j is defined in Algorithm 8. First, we show $\{\alpha_j \leq \alpha\} \subset \{\psi_{j,\alpha} = 1\}$. Note

$$\begin{aligned} \{\alpha_j \leq \alpha\} &= \bigcap_{1\leq s\leq k_\tau} \{\widehat{c}^{-1}(\sqrt{n}|\widetilde{\boldsymbol{\Theta}}_{j,(s)}|, E_j^{(s)}) \leq \alpha\} \\ &= \bigcap_{1\leq s\leq k_\tau} \{\sqrt{n}|\widetilde{\boldsymbol{\Theta}}_{j,(s)}| \geq \widehat{c}(\alpha, E_j^{(s)})\}, \end{aligned} \tag{C.1.2}$$

where $E_j^{(s)} := \{(j,\ell) : \ell \neq j, |\widetilde{\boldsymbol{\Theta}}_{j\ell}| \leq |\widetilde{\boldsymbol{\Theta}}_{j,(s)}|\}$. The first equality is due to the definition of α_j and the second equality holds by the definition of $\widehat{c}^{-1}$. Examining (C.1.2), we immediately know $\sqrt{n}|\widetilde{\boldsymbol{\Theta}}_{j,(1)}| \geq \widehat{c}(\alpha, E_j^{(1)})$ (here $E_j^{(1)} = E_0 = \{(k,j) : k \in [d], k \neq j\}$),

thus the edge corresponding to $\widetilde{\mathbf{\Theta}}_{j,(1)}$ will be rejected in the first iteration of Algorithm 7. Regarding the edge corresponding to $\widetilde{\mathbf{\Theta}}_{j,(2)}$, if $\sqrt{n}|\widetilde{\mathbf{\Theta}}_{j,(2)}| \geq \widehat{c}(\alpha, E_j^{(1)})$, then it will be rejected in the first iteration, too. Otherwise, Algorithm 7 enters the second iteration. Since (C.1.2) implies $\sqrt{n}|\widetilde{\mathbf{\Theta}}_{j,(2)}| \geq \widehat{c}(\alpha, E_j^{(2)})$, we know the edge corresponding to $\widetilde{\mathbf{\Theta}}_{j,(2)}$ must be rejected in the second iteration of Algorithm 7. Following this kind of argument, we are able to show that (C.1.2) implies that all those edges corresponding to $\{\widetilde{\mathbf{\Theta}}_{j,(s)}, 1 \leq s \leq k_\tau\}$ will be rejected according to Algorithm 7. Since the number of rejected edges is at least k_τ, we have $\psi_{j,\alpha} = 1$. Second, we show $\{\psi_{j,\alpha} = 1\} \subset \{\alpha_j \leq \alpha\}$. If $\psi_{j,\alpha} = 1\}$, we know the edges corresponding to $\{\widetilde{\mathbf{\Theta}}_{j,(s)}, 1 \leq s \leq k_\tau\}$ will be rejected, which immediately imply $\sqrt{n}|\widetilde{\mathbf{\Theta}}_{j,(1)}| \geq \widehat{c}(\alpha, E_j^{(1)})$. Regarding the edge corresponding to $\widetilde{\mathbf{\Theta}}_{j,(2)}$, it must get rejected in the first two iterations of Algorithm 7. In either cases, we always have $\sqrt{n}|\widetilde{\mathbf{\Theta}}_{j,(2)}| \geq \widehat{c}(\alpha, E_j^{(2)})$ due to $E_j^{(2)} \subset E_j^{(1)}$ and the fact that $\widehat{c}(\alpha, E) \leq \widehat{c}(\alpha, E')$ when $E \subset E'$. Finally, we establish (C.1.1).

Lemma C.1.1. *Under the same conditions as Lemma 3.2.1, given some $1 \leq j \leq d$, we have the following results.*

(i) Additionally, suppose for any $|\mathbf{\Theta}_{jk}| > 0$, we also have $|\mathbf{\Theta}_{jk}| \geq c\sqrt{\log d/n}$ for some constant $c > 0$. Under the alternative hypothesis $H_{1j} : \|\mathbf{\Theta}_{j,-j}\|_0 \geq k_\tau$, we then have for any $\alpha \in (0, 1)$,

$$\lim_{(n,d)\to\infty} \mathbb{P}(\psi_{j,\alpha} = 1) = 1.$$

(ii) Under the null hypothesis $H_{0j} : \|\mathbf{\Theta}_{j,-j}\|_0 < k_\tau$, we have for any $u \in (0, 1)$,

$$\lim_{(n,d)\to\infty} \mathbb{P}\left(\psi_{j,\alpha} = 1\right) \leq \alpha.$$

The proof of the above lemma is deferred to Appendix C.4.1. The maximum statistic used in our

testing procedure takes the form of $T_E = \max_{(j,k)\in E} \sqrt{n}|\widetilde{\mathbf{\Theta}}^{\mathrm{d}}_{jk}|$. In our key proof procedure, we deal with the case where $E = \{(j,k) : \mathbf{\Theta}_{jk} = 0\}$. Since some of the results hold for general E, we will work with the general notations. Specifically, through out Appendices C.1.1 and C.1.2, we introduce the following notations: in order to approximate

$$T_E := \max_{(j,k)\in E} \sqrt{n} \left| (\widehat{\mathbf{\Theta}}^{\mathrm{d}}_{jk} / \sqrt{\widehat{\mathbf{\Theta}}^{\mathrm{d}}_{jj} \widehat{\mathbf{\Theta}}^{\mathrm{d}}_{kk}} - \mathbf{\Theta}_{jk} / \sqrt{\mathbf{\Theta}_{jj}\mathbf{\Theta}_{jk}}) \right| \tag{C.1.3}$$

by the multiplier bootstrap process

$$T^{\mathcal{B}}_E := \max_{(j,k)\in E} \frac{1}{\sqrt{n\,\widehat{\mathbf{\Theta}}_{jj}\widehat{\mathbf{\Theta}}_{kk}}} \left| \sum_{i=1}^n \widehat{\mathbf{\Theta}}_j^\top (\boldsymbol{X}_i \boldsymbol{X}_i^\top \widehat{\mathbf{\Theta}}_k - \mathbf{e}_k)\xi_i \right|, \tag{C.1.4}$$

we define two intermediate processes:

$$\breve{T}_E := \max_{(j,k)\in E} \left| \frac{1}{\sqrt{n\,\mathbf{\Theta}_{jj}\mathbf{\Theta}_{kk}}} \sum_{i=1}^n \mathbf{\Theta}_j^\top (\boldsymbol{X}_i \boldsymbol{X}_i^\top \mathbf{\Theta}_k - \mathbf{e}_k) \right|, \tag{C.1.5}$$

$$\breve{T}^{\mathcal{B}}_E := \max_{(j,k)\in E} \left| \frac{1}{\sqrt{n\,\mathbf{\Theta}_{jj}\mathbf{\Theta}_{kk}}} \sum_{i=1}^n \mathbf{\Theta}_j^\top (\boldsymbol{X}_i \boldsymbol{X}_i^\top \mathbf{\Theta}_k - \mathbf{e}_k)\xi_i \right|. \tag{C.1.6}$$

C.1.1 Proof of Theorem 3.5.2

Proof of Theorem 3.5.2. Given some $j \in \mathcal{H}_0$, denote $N_{0j} = \{(j,k) : \mathbf{\Theta}_{jk} = 0\}$. By the first part of Lemma C.1.1, we have $\forall j \in \mathcal{B}$,

$$\mathbb{P}\left(\psi_{j,\alpha} = 1\right) > 1 - 3/d^2, \tag{C.1.7}$$

when $\alpha = \Omega(1/d)$, where $\mathcal{B} := \{j \in \mathcal{H}_0^c : \forall k \in \text{supp}(\mathbf{\Theta}_j), |\mathbf{\Theta}_{jk}| > c\sqrt{\log d/n}\}$. Note that we have

$$\mathbb{P}\left(\frac{q|\mathcal{B}|}{d} \le \widehat{\alpha} \le 1\right) \ge \mathbb{P}\left(\frac{q|\mathcal{B}|/d \cdot d}{\sum_{j\in[d]} \psi_{j,q|\mathcal{B}|/d}} \le q\right) = \mathbb{P}\left(\frac{|\mathcal{B}|}{\sum_{j\in[d]} \psi_{j,q|\mathcal{B}|/d}} \le 1\right) \ge 1 - 3/d, \tag{C.1.8}$$

where the first inequality is by (3.2.2) and the last inequality is due to $q\frac{|\mathcal{B}|}{d} = \Omega(1/d)$, (C.1.7) and the the union bound. Rewrite the FDP (with $\widehat{\alpha}$) as

$$\text{FDP}(\widehat{\alpha}) := \frac{\sum_{j\in\mathcal{H}_0} \psi_{j,\widehat{\alpha}}}{\max\left\{1, \sum_{j\in[d]} \psi_{j,\widehat{\alpha}}\right\}} = \frac{\widehat{\alpha} d}{\max\left\{1, \sum_{j\in[d]} \psi_{j,\widehat{\alpha}}\right\}} \cdot \frac{\sum_{j\in\mathcal{H}_0} \psi_{j,\widehat{\alpha}}}{d_0\widehat{\alpha}} \cdot \frac{d_0}{d},$$

and notice that

$$\frac{\widehat{\alpha} d}{\max\left\{1, \sum_{j\in[d]} \psi_{j,\widehat{\alpha}}\right\}} \cdot \frac{d_0}{d} \le \frac{q d_0}{d} \le q.$$

Then it suffices to control the $\text{FDP}(\widehat{\alpha})$ by dealing with $\left(\sum_{j\in\mathcal{H}_0} \psi_{j,\widehat{\alpha}}\right)/d_0\widehat{\alpha}$. By (C.1.8), the FDP control problem is now reduced to showing

$$\sup_{\alpha\in[\alpha_L,1]} \frac{\sum_{j\in\mathcal{H}_0} \psi_{j,\alpha}}{d_0\alpha} \le 1 + o_{\mathbb{P}}(1),$$

where $\alpha_L = q|\mathcal{B}|/d$, By (C.4.4) in the proof of the second part of Lemma C.1.1, $\psi_{j,\alpha} = 1$ implies that $\max_{e\in N_{0j}} \sqrt{n}|\widetilde{\mathbf{\Theta}}_e^{\text{d}} - \mathbf{\Theta}_e^\star| \ge \widehat{c}(\alpha, N_{0j})$, where $N_{0j} = \{(j,k) : \mathbf{\Theta}_{jk} = 0\} = \{(j,k) : \mathbf{\Theta}_{jk}^\star = 0\}$. Therefore, we have

$$\frac{\sum_{j\in\mathcal{H}_0} \psi_{j,\alpha}}{d_0\alpha} \le \frac{\sum_{j\in\mathcal{H}_0} \mathbb{1}(\max_{e\in N_{0j}} \sqrt{n}|\widetilde{\mathbf{\Theta}}_e^{\text{d}} - \mathbf{\Theta}_e^\star| \ge \widehat{c}(\alpha, N_{0j}))}{d_0\alpha}.$$

Hence it suffices to prove that

$$\sup_{\alpha\in[\alpha_L,1]}\left|\frac{\sum_{j\in\mathcal{H}_0}\mathbb{1}(\max_{e\in N_{0j}}\sqrt{n}|\widetilde{\boldsymbol{\Theta}}^{\mathrm{d}}_e-\boldsymbol{\Theta}^\star_e|\geq\widehat{c}(\alpha,N_{0j}))}{d_0\alpha}-1\right|\to 0 \quad \text{in probability.} \quad \text{(C.1.9)}$$

In order to prove (C.1.9), we construct a discrete grid of the interval $[\alpha_L, 1]$. The number of grid points is denoted by λ_d and will be decided later. First, we let $t_1 := \widehat{c}(1, N_{0j}) = 0, t_{\lambda_d} := \widehat{c}(\alpha_L, N_{0j})$. Here $\widehat{c}(\alpha_L, N_{0j}) = \inf\left\{t\in\mathbb{R}:\mathbb{P}_\xi\left(T^{\mathcal{B}}_{N_{0j}}\leq t\right)\geq 1-\alpha\right\}$ is the quantile based on the Gaussian multiplier bootstrap process and depends on the data $\boldsymbol{X}$. Note that the involving random vectors in the Gaussian multiplier bootstrap process are Gaussian conditioning on the data $\boldsymbol{X}$ and have bounded variances with probability growing to 1. Since $\alpha_L = \Omega(1/d)$, then by the maximal inequalities for sub-Gaussian random variables (Lemma 5.2 in van Handel (2014)), we have $t_{\lambda_d} = O(\sqrt{\log d})$ with probability growing to 1. Second, note there exists h_d such that $h_d t_{\lambda_d} = o(1)$ and $t_{\lambda_d}/h_d = O(\log d)$. Based on such h_d, we construct equally spaced sequences $\{t_m\}_{m=1}^{\lambda_d}$ over the range $[t_1, t_{\lambda_d}] = [0, t_{\lambda_d}]$ with $t_m - t_{m-1} = h_d$. Then by setting α_m such that $t_m = \widehat{c}(\alpha_m, N_{0j})$, we obtain a discrete grid $\{\alpha_m\}_{m=1}^{\lambda_d}$ of the interval $[\alpha_L, 1]$. For such $\alpha_m, 1\leq m\leq\lambda_d$, we have

$$\begin{aligned}\max_{1\leq m\leq\lambda_d}\left|\frac{\alpha_{m-1}}{\alpha_m}-1\right| &= \max_{1\leq m\leq\lambda_d}\left|\frac{\mathbb{P}\left(T^{\mathcal{B}}_{N_{0j}}>t_{m-1}\right)}{\mathbb{P}\left(T^{\mathcal{B}}_{N_{0j}}>t_m\right)}-1\right| \\ &\leq \max_{1\leq m\leq\lambda_d} C''(t_m-t_{m-1})(t_m+1)\exp(C'(t_m-t_{m-1})(t_m+1)) = o(1)\end{aligned} \quad \text{(C.1.10)}$$

with probability growing to 1, where the first equality holds by the definition of α_m, the first inequality holds due to part 2 and 3 of Theorem 2.1 in Kuchibhotla et al. (2021) (by first choosing $r-\epsilon, r+\epsilon$ in part 3 to be t_{m-1}, t_m respectively then letting $r-\epsilon, r$ in part 2 to be t_{m-1}, t_m respectively). And the right hand side of the inequality is $o(1)$ since $(t_m - t_{m-1})t_m \leq h_d t_{\lambda_d} = o(1)$

with probability growing to 1.

Denote $I_j(\alpha) = \mathbb{1}(\max_{e\in N_{0j}} \sqrt{n}|\widetilde{\mathbf{\Theta}}_e^{\mathrm{d}} - \mathbf{\Theta}_e^{\star}| \geq \widehat{c}(\alpha, N_{0j}))$. Then given $\alpha_m \leq \alpha \leq \alpha_{m-1}$, for $m = 1, \cdots, \lambda_d$, we have

$$\frac{\sum_{j\in\mathcal{H}_0} I_j(\alpha_m)}{d_0\alpha_m} \cdot \frac{\alpha_m}{\alpha_{m-1}} \leq \frac{\sum_{j\in\mathcal{H}_0} I_j(\alpha)}{d_0\alpha} \leq \frac{\sum_{j\in\mathcal{H}_0} I_j(\alpha_{m-1})}{d_0\alpha_{m-1}} \cdot \frac{\alpha_{m-1}}{\alpha_m}. \tag{C.1.11}$$

Hence by (C.1.10) and (C.1.11), showing (C.1.9) is reduced to proving

$$\max_{1\leq m\leq\lambda_d} \left| \frac{\sum_{j\in\mathcal{H}_0} I_j(\alpha_m)}{d_0\alpha_m} - 1 \right| \to 0, \quad \text{in probability.} \tag{C.1.12}$$

Then it suffices to show that, for any $\epsilon > 0$,

$$\mathbb{P}\left(\max_{1\leq m\leq\lambda_d} \left| \frac{\sum_{j\in\mathcal{H}_0} I_j(\alpha_m)}{d_0\alpha_m} - 1 \right| \geq \epsilon \right) \to 0.$$

By the union bound argument and Chebyshev's inequality, we have

$$\begin{aligned}
&\mathbb{P}\left(\max_{1\leq m\leq\lambda_d} \left| \frac{\sum_{j\in\mathcal{H}_0} I_j(\alpha_m)}{d_0\alpha_m} - 1 \right| \geq \epsilon \right) \\
\leq\ & \sum_{m=1}^{\lambda_d} \mathbb{P}\left(\left| \frac{\sum_{j\in\mathcal{H}_0} I_j(\alpha_m)}{d_0\alpha_m} - 1 \right| \geq \epsilon \right) \\
\leq\ & \sum_{m=1}^{\lambda_d} \frac{\mathbb{E}\left[\sum_{j\in\mathcal{H}_0} I_j(\alpha_m) - d_0\alpha_m \right]^2}{\epsilon^2 d_0^2 \alpha_m^2} \qquad \text{(C.1.13)} \\
=\ & \underbrace{\sum_{m=1}^{\lambda_d} \frac{\sum_{j\in\mathcal{H}_0} \mathrm{Var}\left(I_j(\alpha_m) - d_0\alpha_m \right)}{\epsilon^2 d_0^2 \alpha_m^2}}_{\mathrm{III}_1} + \underbrace{\sum_{m=1}^{\lambda_d} \frac{\left(\mathbb{E}\left[\sum_{j\in\mathcal{H}_0} I_j(\alpha_m) - d_0\alpha_m \right] \right)^2}{\epsilon^2 d_0^2 \alpha_m^2}}_{\mathrm{III}_2} \qquad \text{(C.1.14)} \\
&+ \underbrace{\sum_{m=1}^{\lambda_d} \frac{\sum_{j_1,j_2\in\mathcal{H}_0, j_1\neq j_2} \mathrm{Cov}\left(I_{j_1}(\alpha_m), I_{j_2}(\alpha_m) \right)}{\epsilon^2 d_0^2 \alpha_m^2}}_{\mathrm{III}_3}.
\end{aligned}$$

By Lemma C.1.2 and Lemma C.1.3, we have

$$\begin{aligned}\mathrm{III}_1+\mathrm{III}_2+\mathrm{III}_3 &\leq \frac{C' t_{\lambda_d}}{\epsilon^2 h_d}\left(\frac{d}{d_0|\mathcal{B}|}+\eta^2(d,n)\right)+\frac{C''' d}{\epsilon^2|\mathcal{B}|d_0}\cdot\frac{t_{\lambda_d}}{h_d}\cdot\left(1+\eta(d,n)d_0+\frac{|S|\log d}{d_0 p}\right)\\ &\leq \frac{C_1 t_{\lambda_d}\eta^2(d,n)}{\epsilon^2 h_d}+\frac{C_2}{\epsilon^2\rho d_0}\cdot\frac{t_{\lambda_d}}{h_d}\cdot\left(1+\eta(d,n)d_0+\frac{|S|\log d}{d_0 p}\right), \qquad \text{(C.1.15)}\end{aligned}$$

where we substitute $\zeta_1 = s(\log d)^2/\sqrt{n}$, $\zeta_2 = 1/d^2$ and $\alpha_L = q|\mathcal{B}|/d = \Omega(\rho)$ in $\eta(d,n,\zeta_1,\zeta_2,\alpha_L)$ of Lemma C.1.2 and note $|\mathcal{B}|>0$ then obtain the concise form $\eta(d,n)$ below,

$$\eta(d,n)=\frac{(\log d)^{19/6}}{n^{1/6}}+\frac{(\log d)^{11/6}}{\rho^{1/3}n^{1/6}}+\frac{s(\log d)^3}{n^{1/2}}+\frac{1}{d}.$$

Recall that $t_{\lambda_d} = q(\alpha_L; T^B_{N_{0j}}) = O\left(\sqrt{\log d}\right)$ with probability growing to 1 and $t_{\lambda_d}/h_d = O(\log d)$. Under Assumption 3.5.1, we have

$$\frac{\log d}{\rho}\left(\frac{(\log d)^{19/6}}{n^{1/6}}+\frac{(\log d)^{11/6}}{\rho^{1/3}n^{1/6}}+\frac{s(\log d)^3}{n^{1/2}}\right)=o(1), \quad \frac{\log d}{\rho d_0}+\frac{(\log d)^2|S|}{\rho d_0^2 p}=o(1),$$

and thus $\mathrm{III}_1+\mathrm{III}_2+\mathrm{III}_3 = o(1)$ with probability growing to 1. Therefore, we have proved (C.1.9), and finally establish the FDP control result below,

$$\mathrm{FDP}(\widehat{\alpha})\leq q\frac{d_0}{d}+o_{\mathbb{P}}(1).$$

In order to establish FDR control, it remains to check the uniformly integrability of the random variable sequence in (C.1.12). We pause to note the following result: for a sequence of random variable $R_1, R_2, \cdots$, we $\sup_n \mathbb{E}\left[|R_n|\mathbb{1}(|R_n|>x)\right] \leq x^{-1}\sup_n \mathbb{E}\left[R_n^2\right]$ due to Markov's inequality. Then to show the uniform integrability of the random variable sequence $\{R_n\}_{n=1}^{\infty}$, where

$R_n = \max_{1\le m\le\lambda_d}\left|\frac{\sum_{j\in\mathcal{H}_0} I_j(\alpha_m)}{d_0\alpha_m} - 1\right|$, it suffices to show $\sup_n \mathbb{E}\left[R_n^2\right] < \infty$. Indeed, we have

$$\begin{aligned} & \sup_n \mathbb{E}\left[\left(\max_{1\le m\le\lambda_d}\left|\frac{\sum_{j\in\mathcal{H}_0} I_j(\alpha_m)}{d_0\alpha_m} - 1\right|\right)^2\right] \\ \le\ & \sup_n \sum_{m=1}^{\lambda_d} \frac{\mathbb{E}\left[\sum_{j\in\mathcal{H}_0} I_j(\alpha_m) - d_0\alpha_m\right]^2}{d_0^2\alpha_m^2} \\ =\ & \sup_n \epsilon^2(\mathrm{III}_1 + \mathrm{III}_2 + \mathrm{III}_3). \end{aligned}$$

Since $\mathrm{III}_1+\mathrm{III}_2+\mathrm{III}_3 = o(1)$ with probability growing to 1, we immediately have $\sup_n \mathbb{E}\left[R_n^2\right] < \infty$, thus finally establish the FDR control result:

$$\lim_{(n,d)\to\infty} \mathrm{FDR} \le q\frac{d_0}{d}.$$

□

C.1.2 Ancillary lemmas for Theorem 3.5.2

Lemma C.1.2. *Recalling the definitions of* $\mathrm{III}_1, \mathrm{III}_2$ *in* (C.1.14)*, we have*

$$\mathrm{III}_1 + \mathrm{III}_2 \le \frac{C' t_{\lambda_d}}{\epsilon^2 h_d}\left(\frac{1}{\rho d_0} + \eta^2(d, n, \zeta_1, \zeta_2, \alpha_L)\right),$$

where $\eta(d, n, \zeta_1, \zeta_2, \alpha_L) = O\big(\frac{(\log d)^{19/6}}{n^{1/6}} + \frac{(\log d)^{11/6}}{n^{1/6}\alpha_L^{1/3}} + \zeta_1\log d + \frac{\zeta_2}{\alpha_L}\big)$ *with* $\zeta_1 = s(\log d)^2/\sqrt{n}$, $\zeta_2 = 1/d^2$.

Proof of Lemma C.1.2. First note the definitions of $T_E, \breve{T}_E, T_E^{\mathcal{B}}$ and $\breve{T}_E^{\mathcal{B}}$ in (C.1.3), (C.1.5), (C.1.4) and (C.1.6) respectively, then we apply Proposition C.3.2 to $T = T_E, T_{\boldsymbol{Y}} = \breve{T}_E, T^{\mathcal{B}} = T_E^{\mathcal{B}}, T_{\mathcal{W}} = \breve{T}_E^{\mathcal{B}}$ with $E = N_{0j}$. And we can find the terms ζ_1, ζ_2 in (C.3.4), (C.3.5) to be $s(\log d)^2/\sqrt{n}, 1/d^2$ respectively, due to (C.4.29) and (C.4.34) (i.e., the bound on the differences $T_E - T_0, T_E^{\mathcal{B}} - T_0^{\mathcal{B}}$) in

the proof of Lemma 3.2.1. Thus we have

$$\left| \frac{\mathbb{P}(\max_{e\in N_{0j}} \sqrt{n}|\widetilde{\mathbf{\Theta}}^{\mathrm{d}}_e - \mathbf{\Theta}^{\star}_e| \geq \widehat{c}(\alpha, N_{0j}))}{\alpha} - 1 \right| = \eta(d, n, \zeta_1, \zeta_2, \alpha_L), \tag{C.1.16}$$

where $\mathbf{\Theta}^{\star}_e = 0, e \in N_{0j}$ and $\eta(d, n, \zeta_1, \zeta_2, \alpha_L) = O\left(\frac{(\log d)^{19/6}}{n^{1/6}} + \zeta_1 \log d + \frac{\zeta_2}{\alpha_L}\right)$ with $\zeta_1 = s(\log d)^2/\sqrt{n}, \zeta_2 = 1/d^2$. Recalling the definition of III_2 in (C.1.14), we have

$$\mathrm{III}_2 = \sum_{m=1}^{\lambda_d} \frac{\left(\mathbb{E}\left[\sum_{j\in\mathcal{H}_0} I_j(\alpha_m) - d_0\alpha_m\right]\right)^2}{\epsilon^2 d_0^2 \alpha_m^2},$$

where $I_j(\alpha) = \mathbb{1}(\max_{e\in N_{0j}} \sqrt{n}|\widetilde{\mathbf{\Theta}}^{\mathrm{d}}_e - \mathbf{\Theta}^{\star}_e| \geq \widehat{c}(\alpha, N_{0j}))$. Note that $\alpha_m \in [\alpha_L, 1], \forall\, 1 \leq m \leq \lambda_d$, then we arrive at the following bound

$$\mathrm{III}_2 \leq \frac{\lambda_d}{\epsilon^2} \cdot \eta^2(d, n, \zeta_1, \zeta_2, \alpha_L) \leq \frac{t_{\lambda_d}}{\epsilon^2 h_d} \cdot \eta^2(d, n, \zeta_1, \zeta_2, \alpha_L) \tag{C.1.17}$$

up to some constant, where the first inequality holds by (C.1.16). As for the second inequality, we recall the construction of $\{t_m\}_{m=1}^{\lambda_d}$ (over the course of derivations from (C.1.9) to (C.1.10)) in the proof of Theorem 3.5.2 thus note $\alpha_1 = 1, t_1 = 0$ and $t_{\lambda_d} - t_1 = \sum_{m=2}^{\lambda_d}(t_m - t_{m-1}) = (\lambda_d - 1)h_d$. Regarding the term III_1, we have

$$\begin{aligned} \mathrm{III}_1 &= \sum_{m=1}^{\lambda_d} \frac{\sum_{j\in\mathcal{H}_0} \mathrm{Var}\left(I_j(\alpha_m) - d_0\alpha_m\right)}{\epsilon^2 d_0^2 \alpha_m^2} \\ &= \sum_{m=1}^{\lambda_d} \frac{\sum_{j\in\mathcal{H}_0} \mathbb{E}(I_j(\alpha_m))(1 - \mathbb{E}(I_j(\alpha_m))}{\epsilon^2 d_0^2 \alpha_m^2} \leq \frac{1}{\epsilon^2 d_0} \sum_{m=1}^{\lambda_d} \frac{C}{\alpha_m} \leq \frac{C}{\epsilon^2 d_0 \alpha_L} \cdot \frac{t_{\lambda_d}}{h_d}, \end{aligned} \tag{C.1.18}$$

where the first inequality holds due to (C.1.16) and the second inequality holds since $\alpha_m \geq \alpha_L$

$\forall\, 1 \le m \le \lambda_d$ and $t_{\lambda_d} = (\lambda_d - 1)h_d$. Therefore, combining (C.1.17) with (C.1.18), we obtain

$$\mathrm{III}_1 + \mathrm{III}_2 \le \frac{1}{\epsilon^2} \cdot \frac{t_{\lambda_d}}{h_d} \left(\frac{C}{d_0 \alpha_L} + \eta^2(d, n, \zeta_1, \zeta_2, \alpha_L) \right) \le \frac{C' t_{\lambda_d}}{\epsilon^2 h_d} \left(\frac{1}{\rho d_0} + \eta^2(d, n, \zeta_1, \zeta_2, \alpha_L) \right)$$

for some constant C', where the second inequality holds by the definition $\alpha_L = q|\mathcal{B}|/d$ in the proof of Theorem 3.5.2 and the definition $\rho = |\mathcal{B}|/d$ in Section 3.5. □

Lemma C.1.3. *Recalling the definition of* III_3 *in* (C.1.14)*, we have*

$$\mathrm{III}_3 \le \frac{C''' t_{\lambda_d}}{\rho \epsilon^2 d_0 h_d} \left(1 + \eta(d, n, \zeta_1, \zeta_2, \alpha_L) d_0 + \frac{|S| \log d}{d_0 p} \right),$$

where $\eta(d, n, \zeta_1, \zeta_2, \alpha_L) = O\big(\frac{(\log d)^{19/6}}{n^{1/6}} + \frac{(\log d)^{11/6}}{n^{1/6} \alpha_L^{1/3}} + \zeta_1 \log d + \frac{\zeta_2}{\alpha_L} \big)$ *with* $\zeta_1 = s(\log d)^2/\sqrt{n}$*,* $\zeta_2 = 1/d^2$*.*

Proof of Lemma C.1.3. Note that III_3 in (C.1.14) equals

$$\begin{aligned} \mathrm{III}_3 = \sum_{m=1}^{\lambda_d} \frac{\sum_{j_1, j_2 \in \mathcal{H}_0, j_1 \neq j_2} \mathrm{Cov}\left(I_{j_1}(\alpha_m), I_{j_2}(\alpha_m) \right)}{\epsilon^2 d_0^2 \alpha_m^2}, \\ \text{where } I_j(\alpha) = \mathbb{1}\big(\max_{e \in N_{0j}} \sqrt{n} |\widetilde{\boldsymbol{\Theta}}_e^{\mathrm{d}} - \boldsymbol{\Theta}_e^{\star}| \ge \widehat{c}(\alpha, N_{0j}) \big) \end{aligned} \tag{C.1.19}$$

for $j \in \{j_1, j_2\}$. To quantify the covariance between $I_{j_1}(\alpha_m)$ and $I_{j_2}(\alpha_m)$ for $j_1, j_2 \in \mathcal{H}_0, j_1 \neq j_2$, we define

$$W_j(\alpha) = \mathbb{1}\big(\max_{e \in N_{0j}} |Z_e| \ge c(\alpha, N_{0j}) \big), \tag{C.1.20}$$

where $(Z_e)_{e \in E}$ (with $E = N_{0j}$) is a Gaussian random vector and shares the same mean vector and covariance matrix as the term $\big(\frac{1}{\sqrt{n\, \boldsymbol{\Theta}_{jj} \boldsymbol{\Theta}_{kk}}} \sum_{i=1}^n \boldsymbol{\Theta}_j^\top (\boldsymbol{X}_i \boldsymbol{X}_i^\top \boldsymbol{\Theta}_k - \mathbf{e}_k) \big)_{(j,k) \in E}$ in $\breve{T}_E$. Here $\breve{T}_E$

(with $E = N_{0j}$) has the explicit form below

$$\breve{T}_E = \max_{(j,k)\in E} \frac{1}{\sqrt{n\,\boldsymbol{\Theta}_{jj}\boldsymbol{\Theta}_{kk}}} \left| \sum_{i=1}^{n} \boldsymbol{\Theta}_j^\top (\boldsymbol{X}_i \boldsymbol{X}_i^\top \boldsymbol{\Theta}_k - \mathbf{e}_k) \right|.$$

Remark here $\breve{T}_E$ corresponds to the term $T_{\boldsymbol{Y}}$ in Proposition C.3.2 and $\max_{e\in E} |Z_e|$ corresponds to the term $T_{\boldsymbol{Z}}$ in Proposition C.3.1. And $c(\alpha, N_{0j})$ is the corresponding Gaussian maxima quantile $q(\alpha; T_{\boldsymbol{Z}})$ (which does not need to be computed). Since $\mathbb{P}(T_{\boldsymbol{Z}} > q(\alpha; T_{\boldsymbol{Z}})) = \alpha$, we immediately have have $\mathbb{E}[W_j(\alpha)] = \mathbb{P}\left(\max_{e\in N_{0j}} \sqrt{n}|Z_e| \geq c(\alpha, N_{0j})\right) = \alpha$.

Now we replace $I_{j_1}(\alpha), I_{j_2}(\alpha)$ in III_3 by $W_{j_1}(\alpha), W_{j_2}(\alpha)$ and define III_3' as

$$\mathrm{III}_3' := \sum_{m=1}^{\lambda_d} \frac{\sum_{j_1,j_2\in\mathcal{H}_0, j_1\neq j_2} \mathrm{Cov}\left(W_{j_1}(\alpha_m), W_{j_2}(\alpha_m)\right)}{\epsilon^2 d_0^2 \alpha_m^2}. \tag{C.1.21}$$

To bound the difference between III_3 and III_3', we first note $\mathrm{Cov}(I_{j_1}(\alpha), I_{j_2}(\alpha)) = \mathbb{E}[I_{j_1}(\alpha)I_{j_2}(\alpha)] - \mathbb{E}[I_{j_1}(\alpha)]\mathbb{E}[I_{j_2}(\alpha)]$ then separately deal with the term $|\mathbb{E}\left[I_{j_1}(\alpha)I_{j_2}(\alpha)\right] - \mathbb{E}\left[W_{j_1}(\alpha)W_{j_2}(\alpha)\right]|$ and the term $|\mathbb{E}\left[I_{j_1}(\alpha)\right]\mathbb{E}\left[I_{j_2}(\alpha)\right] - \mathbb{E}\left[W_{j_1}(\alpha)\right]\mathbb{E}\left[G_{j_2}(\alpha)\right]|$.

By Lemma C.1.5, we have up to some constant factor,

$$\frac{|\mathbb{E}\left[I_{j_1}(\alpha)I_{j_2}(\alpha)\right] - \mathbb{E}\left[W_{j_1}(\alpha)W_{j_2}(\alpha)\right]|}{\alpha^2} \leq \frac{\eta(d, n, \zeta_1, \zeta_2, \alpha_L)}{\alpha}.$$

Applying the same strategy to the term $\mathbb{E}\left[I_{j_1}(\alpha)\right]\mathbb{E}\left[I_{j_2}(\alpha)\right]$, we obtain

$$\frac{|\mathbb{E}\left[I_{j_1}(\alpha)\right]\mathbb{E}\left[I_{j_2}(\alpha)\right] - \mathbb{E}\left[W_{j_1}(\alpha)\right]\mathbb{E}\left[G_{j_2}(\alpha)\right]|}{\alpha^2} \leq \frac{\eta(d, n, \zeta_1, \zeta_2, \alpha_L)}{\alpha}.$$

Combining the above two inequalities, and noting the definition of III_3' in (C.1.21), we derive the

following bound on the difference between III_3 and III_3',

$$\left|\mathrm{III}_3 - \mathrm{III}_3'\right| \leq \frac{1}{\epsilon^2} \sum_{m=1}^{\lambda_d} \frac{\eta(d, n, \zeta_1, \zeta_2, \alpha_L)}{\alpha_m} \leq \frac{C' t_{\lambda_d}}{\rho \epsilon^2 h_d} \cdot \eta(d, n, \zeta_1, \zeta_2, \alpha_L).$$

where the second inequality holds due to the fact $\alpha_m \geq \alpha_L \, \forall \, 1 \leq m \leq \lambda_d$ and $t_{\lambda_d} = (\lambda_d - 1)h_d$, the definition $\alpha_L = q|\mathcal{B}|/d$ in the proof of Theorem 3.5.2, and the definition $\rho = |\mathcal{B}|/d$ in Section 3.5.

The above bound on $|\mathrm{III}_3 - \mathrm{III}_3'|$, when combined with Lemma C.1.4, immediately establishes

$$\begin{aligned} \mathrm{III}_3 &\leq \frac{C' t_{\lambda_d}}{\rho \epsilon^2 h_d} \cdot \eta(d, n, \zeta_1, \zeta_2, \alpha_L) + \frac{C'' t_{\lambda_d}}{\rho \epsilon^2 d_0 h_d} \left(1 + C_{\boldsymbol{\Theta}} \frac{|S| \log d}{d_0 p}\right) \\ &\leq \frac{C''' t_{\lambda_d}}{\rho \epsilon^2 d_0 h_d} \left(1 + \eta(d, n, \zeta_1, \zeta_2, \alpha_L) d_0 + \frac{|S| \log d}{d_0 p}\right), \end{aligned}$$

for some constant C'''. □

Lemma C.1.4. *Recalling the term* III_3' *from* (C.1.21) *in the proof of Lemma C.1.3, we have*

$$\mathrm{III}_3' = \sum_{m=1}^{\lambda_d} \frac{\sum_{j_1, j_2 \in \mathcal{H}_0, j_1 \neq j_2} \mathrm{Cov}(W_{j_1}(\alpha_m), W_{j_2}(\alpha_m))}{\epsilon^2 d_0^2 \alpha_m^2} \leq \frac{C'' t_{\lambda_d}}{\rho \epsilon^2 d_0 h_d} \left(1 + C_{\boldsymbol{\Theta}} \frac{|S| \log d}{d_0 p}\right).$$

Proof of Lemma C.1.4. Similarly as in the proof of Lemma C.1.3, we define $(Z_e)_{e \in N_{0j_1} \cup N_{0j_2}}$ to be jointly Gaussian such that this $(|N_{0j_1}| + |N_{0j_2}|)$-dimensional Gaussian random vector shares the same mean vector and covariance matrix as the term $(\frac{1}{\sqrt{n \, \boldsymbol{\Theta}_{jj} \boldsymbol{\Theta}_{kk}}} \sum_{i=1}^n \boldsymbol{\Theta}_j^\top (\boldsymbol{X}_i \boldsymbol{X}_i^\top \boldsymbol{\Theta}_k - \mathbf{e}_k))_{(j,k) \in N_{0j_1} \cup N_{0j_2}}$. Note that the two sub-vectors $(Z_e)_{e \in N_{0j_1}}$ and $(Z_e)_{e \in N_{0j_1}}$ are generally dependent. Then we define $(Z_e')_{e \in N_{0j_1}}, (Z_e')_{e \in N_{0j_2}}$ to be two Gaussian random vectors such that

$$(Z'_e)_{e\in N_{0j_1}} \overset{d}{=} (Z_e)_{e\in N_{0j_1}},\ (Z'_e)_{e\in N_{0j_2}} \overset{d}{=} (Z_e)_{e\in N_{0j_2}} \text{ and } (Z'_e)_{e\in N_{0j_1}} \perp\!\!\!\perp (Z'_e)_{e\in N_{0j_2}}. \tag{C.1.22}$$

Recalling the definition of $W_j(\alpha)$ in (C.1.20): $W_j(\alpha) = \mathbb{1}(\max_{e\in N_{0j}} |Z_e| \geq c(\alpha, N_{0j}))$, we thus have the following,

$$\begin{aligned}
\mathrm{IV}_{j_1j_2}(\alpha) &:= \frac{|\mathrm{Cov}(W_{j_1}(\alpha_m), W_{j_2}(\alpha_m))|}{\alpha^2} && \text{(C.1.23)}\\
&= \frac{|\mathbb{E}[W_{j_1}(\alpha)W_{j_2}(\alpha)] - \mathbb{E}[W_{j_1}(\alpha)]\,\mathbb{E}[W_{j_2}(\alpha)]|}{\alpha^2}\\
&= \frac{1}{\alpha^2}\Big|\mathbb{P}(\max_{e\in N_{0j_1}} |Z_e| \geq c(\alpha, N_{0j_1}), \max_{e\in N_{0j_2}} |Z_e| \geq c(\alpha, N_{0j_2})) -\\
&\quad \mathbb{P}(\max_{e\in N_{0j_1}} |Z'_e| \geq c(\alpha, N_{0j_1}), \max_{e\in N_{0j_2}} |Z'_e| \geq c(\alpha, N_{0j_2}))\Big|\\
&= \frac{1}{\alpha^2}\Big|\mathbb{P}(\max_{e\in N_{0j_1}} |Z_e| \geq t, \max_{e\in N_{0j_2}} |Z_e| \geq t) - \mathbb{P}(\max_{e\in N_{0j_1}} |Z'_e| \geq t, \max_{e\in N_{0j_2}} |Z'_e| \geq t)\Big|\\
&= \frac{1}{\alpha^2}\Big|\mathbb{P}(\max_{e\in N_{0j_1}\cup N_{0j_2}} |Z_e| \geq t) - \mathbb{P}(\max_{e\in N_{0j_1}\cup N_{0j_2}} |Z'_e| \geq t)\Big|, && \text{(C.1.24)}
\end{aligned}$$

where the third equality follows due to the construction of $(Z_e)_{e\in N_{0j_1}\cup N_{0j_2}}, (Z'_e)_{e\in N_{0j_1}\cup N_{0j_2}}$. Note that in the fourth equality, we assume $c(\alpha, N_{0j_1}) = c(\alpha, N_{0j_2}) := t$ without loss of generality, since we can rescale one of the maximum statistic by rescaling the Gaussian random vectors. Remark that the scaling will not break down the application of Theorem 3.3.2, which will be explained in detail later in this proof. The last inequality holds by (C.1.22) and the fact that $\mathbb{P}(A \cap B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cup B)$.

Notice that we can apply the Cramér-type Gaussian comparison bound with ℓ_0 norm to control (C.1.24). Specifically, we first figure out the difference between the covariance matrices of $(Z_e)_{e\in N_{0j_1}\cup N_{0j_2}}$ and $(Z'_e)_{e\in N_{0j_1}\cup N_{0j_2}}$. Denote the covariance matrices by $\mathbf{\Sigma}^Z$ and $\mathbf{\Sigma}^{Z'}$ respectively. As these two Gaussian random vectors have two sub-vectors, we write their covariance matri-

ces in a block form

$$\mathbf{\Sigma}^{Z} = \begin{pmatrix} \mathbf{\Sigma}^{Z}_{11} & \mathbf{\Sigma}^{Z}_{12} \\ \mathbf{\Sigma}^{Z}_{21} & \mathbf{\Sigma}^{Z}_{22} \end{pmatrix}, \ \mathbf{\Sigma}^{Z'} = \begin{pmatrix} \mathbf{\Sigma}^{Z'}_{11} & \boldsymbol{O} \\ \boldsymbol{O} & \mathbf{\Sigma}^{Z'}_{22} \end{pmatrix}.$$

where $\mathbf{\Sigma}^{Z'}$ is block diagonal due to (C.1.22). Note that we also have $\mathbf{\Sigma}^{Z}_{11} = \mathbf{\Sigma}^{Z'}_{11}$ and $\mathbf{\Sigma}^{Z}_{22} = \mathbf{\Sigma}^{Z'}_{22}$. Then we have

$$\mathbf{\Sigma}^{Z} - \mathbf{\Sigma}^{Z'} = \begin{pmatrix} \boldsymbol{O} & \mathbf{\Sigma}^{Z}_{12} \\ \mathbf{\Sigma}^{Z}_{21} & \boldsymbol{O} \end{pmatrix}. \tag{C.1.25}$$

Throughout the following proof, we assume $\mathbf{\Theta}_{jj} = 1, j \in [d]$ without loss of generality, since the standardized version is considered in $\breve{T}_E$ (C.1.5). Recall that $(Z_e)_{e \in N_{0j_1} \cup N_{0j_2}}$ shares the same covariance structure as $(Y_e)_{e \in N_{0j_1} \cup N_{0j_2}}$ where Y_e (with $e = (j, k)$) is defined as

$$Y_e := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{\Theta}_j^{\top} (\boldsymbol{X}_i \boldsymbol{X}_i^{\top} \mathbf{\Theta}_k - \mathbf{e}_k).$$

Then we are ready to calculate the covariance matrix $\mathbf{\Sigma}^{Z}$. Specifically, we compute the entries in each block. Regarding the block $\mathbf{\Sigma}^{Z}_{11}$, for any $k, k' \in N_{0j_1}$ where $N_{0j_1} = \{k : \mathbf{\Theta}_{j_1 k} = 0\}$, we have the corresponding (k, k') entry in $\mathbf{\Sigma}^{Z}_{11}$ equals

$$\mathrm{Cov}(\mathbf{\Theta}_{j_1}^{\top} (\boldsymbol{X}_i \boldsymbol{X}_i^{\top} \mathbf{\Theta}_k - \mathbf{e}_k), \mathbf{\Theta}_{j_1}^{\top} (\boldsymbol{X}_i \boldsymbol{X}_i^{\top} \mathbf{\Theta}_{k'} - \mathbf{e}_{k'})) = \mathbf{\Theta}_{j_1 j_1} \mathbf{\Theta}_{k k'} + \mathbf{\Theta}_{j_1 k} \mathbf{\Theta}_{j_1 k'} = \mathbf{\Theta}_{k k'}, \tag{C.1.26}$$

by applying Isserlis' theorem (Isserlis, 1918) and noting $\mathbf{\Theta}_{j_1 k} = \mathbf{\Theta}_{j_1 k'} = 0$. Similar results hold for the block $\mathbf{\Sigma}^{Z}_{22}$. Regarding the block $\mathbf{\Sigma}^{Z}_{12}$, consider $k_1 \in N_{0j_1}, k_2 \in N_{0j_2}$, then we have the corresponding (k_1, k_2) entry in the block equals

$$\mathrm{Cov}(\mathbf{\Theta}_{j_1}^{\top} (\boldsymbol{X}_i \boldsymbol{X}_i^{\top} \mathbf{\Theta}_{k_1} - \mathbf{e}_{k_1}), \mathbf{\Theta}_{j_2}^{\top} (\boldsymbol{X}_i \boldsymbol{X}_i^{\top} \mathbf{\Theta}_{k_2} - \mathbf{e}_{k_2})) = \mathbf{\Theta}_{j_1 j_2} \mathbf{\Theta}_{k_1 k_2} + \mathbf{\Theta}_{j_1 k_2} \mathbf{\Theta}_{j_2 k_1}. \tag{C.1.27}$$

Now we have fully characterized the covariance matrix $\mathbf{\Sigma}^Z$ and the covariance matrix difference in (C.1.25) for any $j_1, j_2 \in \mathcal{H}_0, j_1 \neq j_2$. Specifically, we have $||\mathbf{\Sigma}^Z - \mathbf{\Sigma}^{Z'}||_0 = ||\mathbf{\Sigma}^Z_{12}||_0 = \sum_{k_1 \in N_{0j_1}, k_2 \in N_{0j_2}} \mathbb{1}(\mathbf{\Theta}_{j_1 j_2}\mathbf{\Theta}_{k_1 k_2} + \mathbf{\Theta}_{j_1 k_2}\mathbf{\Theta}_{j_2 k_1} \neq 0)$. Based on whether $\mathbf{\Theta}_{j_1 j_2}$ is zero or not, we consider the following two cases then handle them separately:

- Case 1: $\mathbf{\Theta}_{j_1 j_2} = 0$. If $k_1 = k_2$, then we have the covariance matrix entry (C.1.27) equal zero; If $k_1 \neq k_2$, then (C.1.27) is nonzero only if $\mathbf{\Theta}_{j_1 k_2} \neq 0, \mathbf{\Theta}_{j_2 k_1} \neq 0$ (i.e., $k_2 \notin N_{0j_1}, k_1 \notin N_{0j_2}$). By the fact $j_1, j_2 \in \mathcal{H}_0, j_1 \neq j_2$ and the definition of $\mathcal{H}_0 = \{j : \|\mathbf{\Theta}_{j,-j}\|_0 < k_\tau\}$, we have $\#\{(k_1, k_2) : k_1 \neq k_2, \mathbf{\Theta}_{j_1 k_2} \neq 0, \mathbf{\Theta}_{j_2 k_1} \neq 0\} \leq k_\tau^2$. Hence $||\mathbf{\Sigma}^Z - \mathbf{\Sigma}^{Z'}||_0 \leq k_\tau^2$.

- Case 2: $\mathbf{\Theta}_{j_1 j_2} \neq 0$. The covariance matrix entry (C.1.27) is nonzero only if $\mathbf{\Theta}_{j_1 k_2} \neq 0, \mathbf{\Theta}_{j_2 k_1} \neq 0$ (i.e., $k_2 \notin N_{0j_1}, k_1 \notin N_{0j_2}$) or $\mathbf{\Theta}_{k_1 k_2} \neq 0$.

We start from the simpler case, i.e., Case 2 where $\mathbf{\Theta}_{j_1 j_2} \neq 0$. Simply, we obtain

$$\mathrm{IV}_{j_1 j_2}(\alpha) = \frac{|\operatorname{Cov}(W_{j_1}(\alpha), W_{j_2}(\alpha))|}{\alpha^2} \leq \frac{\operatorname{Var}(W_{j_1}(\alpha))}{\alpha^2} + \frac{\operatorname{Var}(W_{j_2}(\alpha))}{\alpha^2} \leq \frac{C}{\alpha},$$

for some constant C since $\operatorname{Var}(W_j(\alpha)) = \mathbb{E}[W_j(\alpha)](1 - \mathbb{E}[W_j(\alpha)]) = \alpha(1-\alpha)$ for $j = j_1, j_2$. For a fixed j_1, we also know that $|\{j_2 \in \mathcal{H}_0 : j_2 \neq j_1, \mathbf{\Theta}_{j_1 j_2} \neq 0\}| < k_\tau$. Then we have

$$\sum_{m=1}^{\lambda_d} \sum_{\mathbf{\Theta}_{j_1 j_2} \neq 0} \frac{\mathrm{IV}_{j_1 j_2}(\alpha_m)}{\epsilon^2 d_0^2} \leq \sum_{m=1}^{\lambda_d} \frac{d_0 k_\tau}{\epsilon^2 d_0^2} \cdot \frac{C}{\alpha_m} \leq \frac{1}{\epsilon^2 d_0} \sum_{m=1}^{\lambda_d} \frac{C'}{\alpha_m}, \tag{C.1.28}$$

where the last inequality holds due to the same derivations for III_1 in the proof of Lemma C.1.2.

Regarding Case 1 where $\mathbf{\Theta}_{j_1 j_2} = 0$, we will give a more careful treatment to $\mathrm{IV}_{j_1 j_2}(\alpha)$ in (C.1.23). Due to the discussion about Case 1, we have $||\mathbf{\Sigma}^Z - \mathbf{\Sigma}^{Z'}||_0 \leq k_\tau^2$. This fact will be utilized to derive a nice bound on III_3'. Indeed, we can apply Theorem 3.3.2 to (C.1.24) (with U

and V chosen to be $Z_e)_{e\in N_{0j_1}\cup N_{0j_2}}$ and $(Z'_e)_{e\in N_{0j_1}\cup N_{0j_2}}$ respectively) and obtain

$$\mathrm{IV}_{j_1j_2}(\alpha) \le \frac{\log d}{\alpha p}\left(\sum_{k_1\in N_{0j_1},k_2\in N_{0j_2},k_1\neq k_2} \mathbb{1}(\boldsymbol{\Theta}_{j_1k_2}\boldsymbol{\Theta}_{j_2k_1}\neq 0)\right). \tag{C.1.29}$$

when $\boldsymbol{\Theta}_{j_1j_2} = 0$ (i.e., under Case 1). Recall Theorem 3.3.2 assumes for Gaussian random vectors U and V, there exists a disjoint $\mathfrak{p}$-partition of nodes $\cup_{\ell=1}^{\mathfrak{p}}\mathcal{C}_\ell = [d]$ such that $\sigma^U_{jk} = \sigma^V_{jk} = 0$ when $j \in \mathcal{C}_\ell$ and $k \in \mathcal{C}_{\ell'}$ for some $\ell \neq \ell'$. This is the connectivity assumption. Theorem 3.3.2 also assumes that U and V have unit variances i.e., $\sigma^U_{jj} = \sigma^V_{jj} = 1, j \in [d]$ and there exists some $\sigma_0 < 1$ such that $|\sigma^V_{jk}| \le \sigma_0$ for any $j \neq k$ and $|\{(j,k) : j \neq k, |\sigma^U_{jk}| > \sigma_0\}| \le b_0$ for some constant b_0. Under its general version (which is actually proved in Appendix C.2.2), we only need to assume $a_0 \le \sigma^U_{jj} = \sigma^V_{jj} \le a_1, \forall j \in [d]$, and given any $j \in \mathcal{C}_\ell$ with some ℓ, there exists at least one $m \in \mathcal{C}_{\ell'}$ such that $\sigma^U_{jj} = \sigma^V_{jj} = \sigma^U_{mm} = \sigma^V_{mm}$ for any $\ell' \neq \ell$. From now, we will call it the general variance condition. Accordingly, we assume there exists some $\sigma_0 < 1$ such that $|\sigma^V_{jk}/\sqrt{\sigma^V_{jj}\sigma^V_{kk}}| \le \sigma_0$ for any $j \neq k$ and $|\{(j,k) : j \neq k, |\sigma^U_{jk}|\sqrt{\sigma^U_{jj}\sigma^U_{kk}} > \sigma_0\}| \le b_0$ for some constant b_0. Such condition is referred as the general covariance assumption. Below we give the details of applying Theorem 3.3.2 (with a general version of the variance assumption) by checking those three conditions.

We start from the connectivity assumption and the general variance condition. Notice that in Section 3.5, p denotes the number of connected components in the associated graph $\mathcal{G}$ of $\boldsymbol{X}$. Then we know there exist disjoint partitions of nodes $\cup_{\ell=1}^{p}\mathcal{C}^X_\ell = [d]$ such that $\boldsymbol{\Theta}_{jk} = 0$ when $j \in \mathcal{C}^X_\ell, k \in \mathcal{C}^X_{\ell'}$ for some $\ell \neq \ell'$. We will utilize this fact to examine the covariance matrices of $U := (Z_e)_{e\in N_{0j_1}\cup N_{0j_2}}$ and $V := (Z'_e)_{e\in N_{0j_1}\cup N_{0j_2}}$ and show the connectivity assumption holds. Note that for given $j_1, j_2 \in \mathcal{H}_0, j_1 \neq j_2$, there exist at least $p-2$ components $\cup_{\ell=1}^{p-2}\mathcal{C}^X_\ell$ such that j_1 and j_2 do not belong to them. Without loss of generality, we write $j_1, j_2 \notin \cup_{\ell=1}^{p-2}\mathcal{C}^X_\ell$. Thus we have $\cup_{l=1}^{p-2}\mathcal{C}^X_\ell \subset N_{0j_1} \cap N_{0j_2}$ by definition.

In the following, we will show the number of connected components on the associated graph of the Gaussian random vector $U := (Z_e)_{e\in N_{0j_1}\cup N_{0j_2}}$ is at least $2(p-2)$ by examining its covariance matrix $\mathbf{\Sigma}_Z$. First we focus on the covariance entries in the block $\mathbf{\Sigma}_{11}^Z$. When $\ell_1, \ell_2 \in [p-2]$ and $\ell_1 \neq \ell_2$, we have for any $k \in \mathcal{C}_{\ell_1}^X, k' \in \mathcal{C}_{\ell_2}^X$ (thus $k, k' \in N_{0j_1} \cap N_{0j_2}$), the (k, k') covariance entry (C.1.26) in the block $\mathbf{\Sigma}_{11}^Z$ equals

$$\mathbf{\Theta}_{j_1j_1}\mathbf{\Theta}_{kk'} + \mathbf{\Theta}_{j_1k}\mathbf{\Theta}_{j_1k'} = \mathbf{\Theta}_{j_1j_1}\mathbf{\Theta}_{kk'} = 0, \tag{C.1.30}$$

where the first equality holds since $k, k' \in N_{0j_1}$, and the second equality holds since $\ell_1 \neq \ell_2$. Similarly, we have the (k, k') covariance entry in the block $\mathbf{\Sigma}_{22}^Z$ also equals to zero. Next we compute the covariance entries in the block $\mathbf{\Sigma}_{12}^Z$. For the same (k, k'), we know that $k \in N_{0j_1}, k' \in N_{0j_2}$. Thus the corresponding covariance entry (C.1.27) equals

$$\mathbf{\Theta}_{j_1j_2}\mathbf{\Theta}_{kk'} + \mathbf{\Theta}_{j_1k'}\mathbf{\Theta}_{j_2k} = 0, \tag{C.1.31}$$

since we also have $k \in N_{0j_2}, k' \in N_{0j_1}$ and $k \in \mathcal{C}_{\ell_1}^X, k' \in \mathcal{C}_{\ell_2}^X$ for some $\ell_1 \neq \ell_2$. Denote the nodes in the associated graph of $\mathbf{\Sigma}^Z$ by $\mathcal{V}_Z := \{(j, k) : k \in N_{0j}, j = j_1, j_2\}$. Remark here we use a pair (j, k) to represent a node since there exists some $k \in N_{0j_1} \cap N_{0j_2}$ and we have to distinguish the covariance entries (j_1, k) and (j_2, k). Based on previous calculations, we immediately find $\cup_{\ell=1}^{2(p-2)} \mathcal{C}_\ell^Z \subset \mathcal{V}_Z$, where $\mathcal{C}_\ell^Z$ is chosen to be

$$\mathcal{C}_\ell^Z = \begin{cases} \{(j_1, k) : k \in \mathcal{C}_\ell^X\} & \text{when } 1 \leq \ell \leq p-2, \\ \{(j_2, k) : k \in \mathcal{C}_\ell^X\} & \text{when } p-1 \leq \ell \leq 2(p-2). \end{cases} \tag{C.1.32}$$

Further, we know they form different components on the associated graph of $\mathbf{\Sigma}^Z$. This is due to (C.1.30) and (C.1.31). The above results also apply to the Gaussian random vector $V := (Z'_e)_{e\in N_{0j_1}\cup N_{0j_2}}$

by construction of Z'_e, i.e., we have the same subset of nodes $\cup_{\ell=1}^{2(p-2)}\mathcal{C}_\ell^Z \subset \mathcal{V}_Z$ from different components on the associated graph of $\mathbf{\Sigma}^{Z'}$.

When $k \in \mathcal{C}_\ell$ for some $\ell \in [p-2]$, the corresponding diagonal entries of the covariance matrices $\mathbf{\Sigma}^Z, \mathbf{\Sigma}^{Z'}$ equal

$$\mathbf{\Theta}_{j_1j_1}\mathbf{\Theta}_{kk} + \mathbf{\Theta}_{j_1k}\mathbf{\Theta}_{j_1k} = \mathbf{\Theta}_{j_1j_1}\mathbf{\Theta}_{kk} = 1 = \mathbf{\Theta}_{j_2j_2}\mathbf{\Theta}_{kk},$$

where the first equality holds since $\mathbf{\Theta}_{j_1k} = 0$ when $k \in \mathcal{C}_\ell \subset N_{0j_1}$. As for the second equality, we use the fact that $\mathbf{\Theta}_{jj} = 1, j \in [d]$. This is because $\breve{T}_E$ in (C.1.5) considers the standardized version $\mathbf{\Theta}_{jk}/\sqrt{\mathbf{\Theta}_{jj}\mathbf{\Theta}_{kk}}$. Remark that the rescaling in Lemma C.1.4 is performed on one of the two random vectors $(Z'_e)_{e\in N_{0j_1}}, (Z'_e)_{e\in N_{0j_2}}$. Then we have the variances across the $p-2$ components $\cup_{\ell=1}^{p-2}\mathcal{C}_\ell^Z$ are the same. The variances across the other $p-2$ components $\cup_{\ell=p-1}^{2(p-2)}\mathcal{C}_\ell^Z$ are also the same. Finally, we show there exist at least $p-2$ components $\cup_{\ell=1}^{p-2}\mathcal{C}_\ell^Z$ (or $\cup_{\ell=p-1}^{2(p-2)}\mathcal{C}_\ell^Z$) satisfying the requirement in the connectivity assumption and the general variance condition.

Regarding the general covariance condition, we first note that $\mathbf{\Theta} \in \mathcal{U}(M, s, r_0)$ which says that $\lambda_{\min}(\mathbf{\Theta}) \geq 1/r_0, \lambda_{\max}(\mathbf{\Theta}) \leq r_0$. Thus we have $\max_{j,k\in[d],j\neq k}|\mathbf{\Theta}_{jk}| \leq \sigma_0$ for some $\sigma_0 < 1$. Below we will examine all the off-diagonal entries of $\mathbf{\Sigma}^Z$ and $\mathbf{\Sigma}^{Z'}$. Regarding the block $\mathbf{\Sigma}^Z_{11}$, for any $k, k' \in N_{0j_1}, k \neq k'$ where $N_{0j_1} = \{k : \mathbf{\Theta}_{j_1k} = 0\}$, (C.1.26) says that the corresponding (k, k') entry in $\mathbf{\Sigma}^Z_{11}$ equals $\mathbf{\Theta}_{kk'}$ (here we have $|\mathbf{\Theta}_{kk'}| \leq \sigma_0$). Similar results hold for the block $\mathbf{\Sigma}^Z_{22}$. Regarding the block $\mathbf{\Sigma}^Z_{12}$, consider $k_1 \in N_{0j_1}, k_2 \in N_{0j_2}$, then we have the corresponding (k_1, k_2) entry in the block equals $\mathbf{\Theta}_{j_1k_2}\mathbf{\Theta}_{j_2k_1}$. This is due to (C.1.27) and the fact that $\mathbf{\Theta}_{j_1j_2} = 0$ under Case 1. Only when $k_2 = j_1, k_1 = j_2$, we have $\mathbf{\Theta}_{j_1k_2}\mathbf{\Theta}_{j_2k_1} = 1$. Otherwise, $|\mathbf{\Theta}_{j_1k_2}\mathbf{\Theta}_{j_2k_1}| \leq \sigma_0^2 < \sigma_0$ always holds. As for the $\mathbf{\Sigma}^{Z'}$, since its block $\mathbf{\Sigma}^{Z'}_{12} = \boldsymbol{O}$, we immediately have the absolute values of all its off-diagonal entries is bounded by σ_0. In summary, we verify the covariance condition of Theorem 3.3.2 (here U and V are chosen to be $Z_e)_{e\in N_{0j_1}\cup N_{0j_2}}$ and

$(Z'_e)_{e\in N_{0j_1}\cup N_{0j_2}}$ respectively).

Having checked all the three conditions, we now obtain

$$\begin{aligned}
&\sum_{m=1}^{\lambda_d}\sum_{\boldsymbol{\Theta}_{j_1j_2}=0}\frac{\mathrm{IV}_{j_1j_2}(\alpha_m)}{\epsilon^2 d_0^2}\\
\leq\ &\sum_{m=1}^{\lambda_d}\left\{\frac{1}{\epsilon^2 d_0^2}\cdot\frac{\log d}{\alpha_m p}\left(\sum_{k_1\in N_{0j_1},k_2\in N_{0j_2},k_1\neq k_2}\mathbb{1}(\boldsymbol{\Theta}_{j_1k_2}\boldsymbol{\Theta}_{j_2k_1}\neq 0)\right)\right\}\\
\leq\ &\frac{C_{\boldsymbol{\Theta}}|S|\log d}{\epsilon^2 d_0 p}\left(\frac{1}{d_0}\sum_{m=1}^{\lambda_d}\frac{C'}{\alpha_m}\right), \qquad \text{(C.1.33)}
\end{aligned}$$

where S represents the set

$$S=\{(j_1,j_2,k_1,k_2): j_1,j_2\in\mathcal{H}_0, j_1\neq j_2, k_1\neq k_2, \boldsymbol{\Theta}_{j_1j_2}=\boldsymbol{\Theta}_{j_1k_1}=\boldsymbol{\Theta}_{j_2k_2}=0, \boldsymbol{\Theta}_{j_1k_2}\neq 0, \boldsymbol{\Theta}_{j_2k_1}\neq 0\}$$

as defined in Section 3.5, and $C_{\boldsymbol{\Theta}}$ is some universal constant over $\boldsymbol{\Theta}\in\mathcal{U}(M,s,r_0)$. Finally, combining (C.1.33) with (C.1.28), we obtain the following bound on III'_3,

$$\begin{aligned}
\mathrm{III}'_3\ &\leq\ \frac{C_{\boldsymbol{\Theta}}|S|\log d}{\epsilon^2 d_0 p}\left(\frac{1}{d_0}\sum_{m=1}^{\lambda_d}\frac{C'}{\alpha_m}\right)+\frac{1}{\epsilon^2 d_0}\sum_{m=1}^{\lambda_d}\frac{C'}{\alpha_m}\\
&=\ \left(1+\frac{C_{\boldsymbol{\Theta}}|S|\log d}{d_0 p}\right)\cdot\frac{1}{\epsilon^2 d_0}\sum_{m=1}^{\lambda_d}\frac{C'}{\alpha_m}\\
&\leq\ \frac{C''t_{\lambda_d}}{\rho\epsilon^2 d_0 h_d}\left(1+\frac{C_{\boldsymbol{\Theta}}|S|\log d}{d_0 p}\right),
\end{aligned}$$

where the last inequality holds due to the same derivations for III_1 in the proof of Lemma C.1.2.

□

Lemma C.1.5. *Recall the definitions of $I_j(\alpha)$ and $W_j(\alpha)$ in* (C.1.19) *and* (C.1.20)*, for $j_1, j_2\in$*

$\mathcal{H}_0, j_1 \neq j_2$, *when* $\alpha \in [\alpha_L, 1]$, *we have*

$$\left|\mathbb{E}\left[I_{j_1}(\alpha)I_{j_2}(\alpha)\right] - \mathbb{E}\left[W_{j_1}(\alpha)W_{j_2}(\alpha)\right]\right| \leq \eta(d, n, \zeta_1, \zeta_2, \alpha_L)\alpha. \tag{C.1.34}$$

Proof of Lemma C.1.5. First express $\left|\mathbb{E}\left[I_{j_1}(\alpha)I_{j_2}(\alpha)\right] - \mathbb{E}\left[W_{j_1}(\alpha)W_{j_2}(\alpha)\right]\right|$ as

$$\begin{aligned}
&\left|\mathbb{E}\left[I_{j_1}(\alpha)I_{j_2}(\alpha)\right] - \mathbb{E}\left[W_{j_1}(\alpha)W_{j_2}(\alpha)\right]\right| \\
=\ &\Big|\mathbb{P}\Big(\max_{e\in N_{0j_1}} \sqrt{n}|\widetilde{\mathbf{\Theta}}^{\mathrm{d}}_e| \geq \widehat{c}(\alpha, N_{0j_1}), \max_{e\in N_{0j_2}} \sqrt{n}|\widetilde{\mathbf{\Theta}}^{\mathrm{d}}_e| \geq \widehat{c}(\alpha, N_{0j_2})\Big) \\
&\quad - \mathbb{P}\Big(\max_{e\in N_{0j_1}} |Z_e| \geq c(\alpha, N_{0j_1}), \max_{e\in N_{0j_1}} |Z_e| \geq c(\alpha, N_{0j_2})\Big)\Big| \\
=\ &\Big|\mathbb{P}\Big(T_{N_{0j_1}} \geq \widehat{c}(\alpha, N_{0j_1}), T_{N_{0j_2}} \geq \widehat{c}(\alpha, N_{0j_2})\Big) \\
&\quad - \mathbb{P}\Big(\max_{e\in N_{0j_1}} |Z_e| \geq c(\alpha, N_{0j_1}), \max_{e\in N_{0j_1}} |Z_e| \geq c(\alpha, N_{0j_2})\Big)\Big|,
\end{aligned} \tag{C.1.35}$$

where the second equality holds by the definition of T_E in (C.1.3) and the definitions of N_{0j_1}, N_{0j_2}. Now proving the bound in (C.1.34) is reduced to showing

$$\begin{aligned}
&\Big|\mathbb{P}\Big(T_{N_{0j_1}} \geq \widehat{c}(\alpha, N_{0j_1}), T_{N_{0j_2}} \geq \widehat{c}(\alpha, N_{0j_2})\Big) - \mathbb{P}\Big(\max_{e\in N_{0j_1}} |Z_e| \geq c(\alpha, N_{0j_1}), \max_{e\in N_{0j_1}} |Z_e| \geq c(\alpha, N_{0j_2})\Big)\Big| \\
&\leq \eta(d, n, \zeta_1, \zeta_2, \alpha_L)\alpha.
\end{aligned} \tag{C.1.36}$$

We first relate the notations in the above expression to the notations in Appendix C.3: $T_{N_{0j_1}}, T_{N_{0j_2}}$ correspond to T; $\widehat{c}(\alpha, N_{0j_1}), \widehat{c}(\alpha, N_{0j_2})$ correspond to $q_\xi(\alpha, T^{\mathcal{B}})$; $\max_{e\in N_{0j_1}} |Z_e|, \max_{e\in N_{0j_2}} |Z_e|$ correspond to $T_{\boldsymbol{Z}}$; $c(\alpha, N_{0j_1}), c(\alpha, N_{0j_2})$ correspond to $q(\alpha; T_{\boldsymbol{Z}})$. In Appendix C.3, we prove Propositions C.3.1 and C.3.2. And the strategy can be used to derive the bound on (C.1.35). First, we note that $T_{N_{0j_1}}, T_{N_{0j_2}}$ satisfy the conditions of Proposition C.3.2, i.e., (C.3.4) and (C.3.5). This is due to the same derivations as the first parapraph of the proof of Lemma C.1.2. Since the proving

strategy is quite similar, we omit the proof of (C.1.36) for simplicity. Instead, we prove (C.1.37), i.e., when $\alpha \in [\alpha_L, 1]$,

$$\begin{aligned} \mathrm{D} :=& \left| \mathbb{P}\left(T_{\boldsymbol{Y}_1} \geq q_\xi(\alpha; T_{\boldsymbol{W}_1}), T_{\boldsymbol{Y}_2} \geq q_\xi(\alpha; T_{\boldsymbol{W}_2})\right) - \mathbb{P}\left(T_{\boldsymbol{Z}_1} \geq q(\alpha; T_{\boldsymbol{Z}_1}), T_{\boldsymbol{Z}_2} \geq q(\alpha; T_{\boldsymbol{Z}_2})\right) \right| \\ \leq& C\alpha \left(\frac{(\log d)^{11/6}}{n^{1/6}\alpha_L^{1/3}} + \frac{(\log d)^{19/6}}{n^{1/6}} \right), \end{aligned} \tag{C.1.37}$$

where $T_{\boldsymbol{Y}_1}, T_{\boldsymbol{Y}_2}$ correspond to $\breve{T}_E$ with $E = N_{0j_1}, N_{0j_2}$ respectively, $T_{\boldsymbol{W}_1}, T_{\boldsymbol{W}_2}$ correspond to $\breve{T}_E^{\mathcal{B}}$ with $E = N_{0j_1}, N_{0j_2}$ respectively, and $T_{\boldsymbol{Z}_1} = \max_{e \in N_{0j_1}} |Z_e|, T_{\boldsymbol{Z}_2} = \max_{e \in N_{0j_2}} |Z_e|$. As for the quantiles, $q_\xi(\alpha; T_{\boldsymbol{W}_1}), q_\xi(\alpha; T_{\boldsymbol{W}_2})$ are the Gaussian multiplier bootstrap quantiles based on $T_{\boldsymbol{W}_1}, T_{\boldsymbol{W}_2}$. $q(\alpha; T_{\boldsymbol{Z}_1}), q(\alpha; T_{\boldsymbol{Z}_2})$ are the quantiles of the Gaussian maxima $T_{\boldsymbol{Z}_1}, T_{\boldsymbol{Z}_2}$. Denote $A_1 = \{T_{\boldsymbol{Y}_1} \geq q_\xi(\alpha; T_{\boldsymbol{W}_1})\}, A_2 = \{T_{\boldsymbol{Y}_2} \geq q_\xi(\alpha; T_{\boldsymbol{W}_2})\}, B_1 = \{T_{\boldsymbol{Y}_1} \geq q(\alpha; T_{\boldsymbol{Z}_1})\}$, $B_2 = \{T_{\boldsymbol{Y}_2} \geq q(\alpha; T_{\boldsymbol{Z}_2})\}$, we have

$$\begin{aligned} \mathrm{D}_{12} :=& \left| \mathbb{P}\left(T_{\boldsymbol{Y}_1} \geq q_\xi(\alpha; T_{\boldsymbol{W}_1}), T_{\boldsymbol{Y}_2} \geq q_\xi(\alpha; T_{\boldsymbol{W}_2})\right) - \mathbb{P}\left(T_{\boldsymbol{Y}_1} \geq q(\alpha; T_{\boldsymbol{Z}_1}), T_{\boldsymbol{Y}_2} \geq q(\alpha; T_{\boldsymbol{Z}_2})\right) \right| \\ \leq& \mathbb{P}\left((A_1 \cap A_2) \ominus (B_1 \cap B_2)\right) \\ =& \mathbb{P}\left((A_1 \cap A_2) \cap (B_1^c \cup B_2^c)\right) + \mathbb{P}\left((B_1 \cap B_2) \cap (A_1^c \cup A_2^c)\right) \\ \leq& \mathbb{P}\left(A_1 \cap B_1^c\right) + \mathbb{P}\left(A_2 \cap B_2^c\right) + \mathbb{P}\left(B_1 \cap A_1^c\right) + \mathbb{P}\left(B_2 \cap A_2^c\right) \\ =& \mathbb{P}\left((A_1 \cap B_1^c) \cup (B_1 \cap A_1^c)\right) + \mathbb{P}\left((A_2 \cap B_2^c) \cup (B_2 \cap A_2^c)\right) \\ =& \mathbb{P}\left(A_1 \ominus B_1\right) + \mathbb{P}\left(A_2 \ominus B_2\right). \end{aligned} \tag{C.1.38}$$

By (C.3.14) and (C.3.15), we can bound (C.1.38) as

$$\mathrm{D}_{12} \leq \mathbb{P}\left(A_1 \ominus B_1\right) + \mathbb{P}\left(A_2 \ominus B_2\right) \leq 2C'\alpha \left(\frac{(\log d)^{11/6}}{n^{1/6}\alpha_L^{1/3}} + \frac{(\log d)^{19/6}}{n^{1/6}} \right). \tag{C.1.39}$$

By the triangle inequality, we have the following bound on D,

$$
\begin{aligned}
\mathrm{D} = & \left| \mathbb{P}\left(T_{\boldsymbol{Y}_1} \geq q_\xi(\alpha; T_{\boldsymbol{W}_1}), T_{\boldsymbol{Y}_2} \geq q_\xi(\alpha; T_{\boldsymbol{W}_2})\right) - \mathbb{P}\left(T_{\boldsymbol{Z}_1} \geq q(\alpha; T_{\boldsymbol{Z}_1}), T_{\boldsymbol{Z}_2} \geq q(\alpha; T_{\boldsymbol{Z}_2})\right) \right| \\
\leq & \ \mathrm{D}_{12} + \left| \mathbb{P}\left(T_{\boldsymbol{Y}_1} \geq q(\alpha; T_{\boldsymbol{Z}_1}), T_{\boldsymbol{Y}_2} \geq q(\alpha; T_{\boldsymbol{Z}_2})\right) - \mathbb{P}\left(T_{\boldsymbol{Z}_1} \geq q(\alpha; T_{\boldsymbol{Z}_1}), T_{\boldsymbol{Z}_2} \geq q(\alpha; T_{\boldsymbol{Z}_2})\right) \right|. \\
\leq & \ \mathrm{D}_{12} + \left| \mathbb{P}\left(T_{\boldsymbol{Y}_1} \geq q(\alpha; T_{\boldsymbol{Z}_1})\right) - \mathbb{P}\left(T_{\boldsymbol{Z}_1} \geq q(\alpha; T_{\boldsymbol{Z}_1})\right) \right| + \left| \mathbb{P}\left(T_{\boldsymbol{Y}_2} \geq q(\alpha; T_{\boldsymbol{Z}_2})\right) - \mathbb{P}\left(T_{\boldsymbol{Z}_2} \geq q(\alpha; T_{\boldsymbol{Z}_2})\right) \right| \\
& + \underbrace{\left| \mathbb{P}\left(\{T_{\boldsymbol{Y}_1} \geq q(\alpha; T_{\boldsymbol{Z}_1})\} \cup \{T_{\boldsymbol{Y}_2} \geq q(\alpha; T_{\boldsymbol{Z}_2})\}\right) - \mathbb{P}\left(\{T_{\boldsymbol{Z}_1} \geq q(\alpha; T_{\boldsymbol{Z}_1})\} \cup \{T_{\boldsymbol{Z}_2} \geq q(\alpha; T_{\boldsymbol{Z}_2})\}\right) \right|}_{\mathrm{D}'_{12}},
\end{aligned}
\tag{C.1.40}
$$

where the last inequality holds since $\mathbb{P}(A \cap B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cup B)$. For the second term and the third term in (C.1.40), we can directly apply the results (C.3.10) in Proposition C.3.1 and bound them as

$$
\begin{aligned}
& \left| \mathbb{P}\left(T_{\boldsymbol{Y}_1} \geq q(\alpha; T_{\boldsymbol{Z}_1})\right) - \mathbb{P}\left(T_{\boldsymbol{Z}_1} \geq q(\alpha; T_{\boldsymbol{Z}_1})\right) \right| + \left| \mathbb{P}\left(T_{\boldsymbol{Y}_2} \geq q(\alpha; T_{\boldsymbol{Z}_2})\right) - \mathbb{P}\left(T_{\boldsymbol{Z}_2} \geq q(\alpha; T_{\boldsymbol{Z}_2})\right) \right| \\
& \leq C\alpha \cdot \frac{(\log d)^{19/6}}{n^{1/6}}
\end{aligned}
\tag{C.1.41}
$$

for some constant C. Regarding the term D'_{12}, we assume $q(\alpha; T_{\boldsymbol{Z}_2}) = q(\alpha; T_{\boldsymbol{Z}_2}) := t$ without loss of generality. This is because $q(\alpha; T_{\boldsymbol{Z}_1}), q(\alpha; T_{\boldsymbol{Z}_2})$ are all deterministic values and we can rescale the random vector inside one of the maximum statistics $T_{\boldsymbol{Z}_1}, T_{\boldsymbol{Z}_2}$. Now we rewrite D'_{12}

based on $q(\alpha; T_{\boldsymbol{Z}_2}) = q(\alpha; T_{\boldsymbol{Z}_2}) = t$ and derive the following bound:

$$\begin{aligned} \mathrm{D}'_{12} &= \left| \mathbb{P}\left(\max\{T_{\boldsymbol{Y}_1}, T_{\boldsymbol{Y}_2}\} \geq t\right) - \mathbb{P}\left(\max\{T_{\boldsymbol{Z}_1}, T_{\boldsymbol{Z}_2}\} \geq t\right) \right| && \text{(C.1.42)} \\ &\leq \frac{C''(\log d)^{19/6}}{n^{1/6}} \cdot \mathbb{P}\left(\max\{T_{\boldsymbol{Z}_1}, T_{\boldsymbol{Z}_2}\} \geq t\right) \\ &\leq \frac{C''(\log d)^{19/6}}{n^{1/6}} \cdot \left(\mathbb{P}\left(T_{\boldsymbol{Z}_1} \geq q(\alpha; T_{\boldsymbol{Z}_1})\right) + \mathbb{P}\left(T_{\boldsymbol{Z}_2} \geq q(\alpha; T_{\boldsymbol{Z}_2})\right)\right) \\ &= 2C''\alpha \cdot \frac{(\log d)^{19/6}}{n^{1/6}}, && \text{(C.1.43)} \end{aligned}$$

where the first inequality holds by applying Corollary 5.1 of Kuchibhotla et al. (2021) similarly as in the derivation of (C.3.10). Here we briefly explain why Corollary 5.1 of Kuchibhotla et al. (2021) is applicable to (C.1.42). Note that $\max\{T_{\boldsymbol{Y}_1}, T_{\boldsymbol{Y}_2}\} = T_{\boldsymbol{Y}_{12}}$ is the maximum statistic with respect to the random vectors which concatenate the random vectors involved in $T_{\boldsymbol{Y}_1}, T_{\boldsymbol{Y}_2}$. Write $T_{\boldsymbol{Y}_1}, T_{\boldsymbol{Y}_2}$ explicitly as

$$T_{\boldsymbol{Y}_1} := \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{Y}_i^{(1)} \right\|_\infty, \quad T_{\boldsymbol{Y}_2} := \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{Y}_i^{(2)} \right\|_\infty,$$

and denote $\boldsymbol{Y}_i^{(12)} = (\boldsymbol{Y}_i^{(1)}, \boldsymbol{Y}_i^{(2)})$, then $T_{\boldsymbol{Y}_{12}}$ is defined as

$$T_{\boldsymbol{Y}_{12}} := \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{Y}_i^{(12)} \right\|_\infty.$$

By the definition of $\boldsymbol{Z}_1, \boldsymbol{Z}_2$, we have $\mathrm{Cov}((\boldsymbol{Z}_1^\top, \boldsymbol{Z}_2^\top)^\top) = \mathrm{Cov}((\boldsymbol{Y}_1^\top, \boldsymbol{Y}_2^\top)^\top)$. Hence we can apply Corollary 5.1 of Kuchibhotla et al. (2021) to (C.1.42). Now we combine (C.1.39), (C.1.40), (C.1.41) with (C.1.43) and obtain the following bound

$$\mathrm{D} \leq C\alpha \left(\frac{(\log d)^{11/6}}{n^{1/6} \alpha_L^{1/3}} + \frac{(\log d)^{19/6}}{n^{1/6}} \right),$$

for some constant C, thus (C.1.37) is established. The above strategy of obtaining (C.1.37) can be

similarly applied to the term in (C.1.35), then establishes the bound in (C.1.34).

□

C.1.3 Proof of Theorem 3.4.2

Proof of Theorem 3.4.2. Throughout the proof, we condition on the design matrix $\boldsymbol{X}$, but without explicitly writing it out in order to simplify the notation. In the context of selecting hub response variables, we recall $\mathcal{H}_0 = \{j \in [d_1] : ||\boldsymbol{\Theta}_j||_0 \geq k_\tau\}$ and $d_0 = |\mathcal{H}_0|$. For a non-hub response variable $j \in \mathcal{H}_0$, let N_{0j} be the set of its null covariates, i.e., $N_{0j} = \{(j,k) : \boldsymbol{\Theta}_{jk} = 0\}$.

To establish FDR control, we follow the same derivations as in the proof of Theorem 3.5.2. Specifically, it suffices to bound

$$\sum_{m=1}^{\lambda_d} \frac{\mathrm{Var}[\sum_{j\in\mathcal{H}_0} I_j(\alpha_m) - d_0\alpha_m]}{\epsilon^2 d_0^2 \alpha_m^2} + \sum_{m=1}^{\lambda_d} \frac{(\mathbb{E}[\sum_{j\in\mathcal{H}_0} I_j(\alpha_m) - d_0\alpha_m])^2}{\epsilon^2 d_0^2 \alpha_m^2}$$
$$+ \sum_{m=1}^{\lambda_d} \frac{\sum_{j_1,j_2\in\mathcal{H}_0, j_1\neq j_2} \mathrm{Cov}(I_{j_1}(\alpha_m), I_{j_2}(\alpha_m))}{\epsilon^2 d_0^2 \alpha_m^2} := \mathrm{III}_1 + \mathrm{III}_2 + 0 \qquad \text{(C.1.44)}$$

for any $\epsilon > 0$. In the above terms, the sequence $\{\alpha_m\}_{m=1}^{\lambda_d}$ is chosen similarly as in the proof of Theorem 3.5.2 and $I_j(\alpha)$ is defined as

$$I_j(\alpha) = \mathbb{1}(\max_{e\in N_{0j}} \sqrt{n}|\widetilde{\boldsymbol{\Theta}}_e^{\mathrm{d}}| \geq \widehat{c}(\alpha, N_{0j})),$$

where $\widetilde{\boldsymbol{\Theta}}_j^{\mathrm{d}}$ is the debiased Lasso estimator defined in (3.4.2). Note that the cross term in (C.1.44) equals zero as $\mathrm{Cov}(I_{j_1}(\alpha_m), I_{j_2}(\alpha_m)) = 0$. This is because $\boldsymbol{Y}^{(j)}, j \in [d_1]$ are conditionally independent given $\mathbf{X}$. Therefore it suffices to bound III_1 and III_2. By applying Lemma C.1.2 with the term $\eta(d, n, \zeta_1, \zeta_2, \alpha_L)$ replaced by $\eta_0(d_1, d_2, n, \zeta_1, \zeta_2, \alpha_L)$ in Lemma C.1.8, (C.1.44) can be

controlled by

$$\mathrm{III}_2 + \mathrm{III}_2 \leq \frac{C' t_{\lambda_{d_2}}}{\epsilon^2 h_{d_2}} \left(\frac{d_1}{d_0 |\mathcal{B}|} + \eta_0(d_1, d_2, n, \zeta_1, \zeta_2, \alpha_L) \right),$$

where $\alpha_L = q|\mathcal{B}|/d_1$ and $t_{\lambda_{d_2}}, h_{d_2}$ are similarly defined as in the proof of Theorem 3.5.2. According to Lemma C.1.8, we have the explicit form of $\eta_0(d_1, d_2, n, \zeta_1, \zeta_2, \delta, \alpha_L)$:

$$\eta_0(d_1, d_2, n, \zeta_1, \zeta_2, \delta, \alpha_L) = \zeta_1 \log d_2 + (\log d_2)^{5/2} \delta^{1/2} + \frac{\eta + \zeta_2}{\alpha_L},$$

where $\zeta_1 = O(s \log d_2 / \sqrt{n})$, $\zeta_2 = O(e^{-c_1 n} + d_2^{-\widetilde{c}_0 \wedge c_2})$, δ satisfies $\frac{1}{\delta} \sqrt{\frac{s \log d_2}{n}} = O(1)$ and $\eta = e^{-c_1 n} + \frac{1}{d_2} + \frac{1}{n\delta^2}$. By rearranging, we obtain the following bound on $\mathrm{III}_2 + \mathrm{III}_2$:

$$\frac{\log d_2}{\epsilon^2} \left(\frac{1}{d_0 \rho} + \frac{s(\log d_2)^2}{n^{1/2}} + (\log d_2)^{5/2} \delta^{1/2} + \frac{1}{n\delta^2 \rho} + \frac{1}{\rho} \big(\frac{1}{d_2} + e^{-c_1 n} + d_2^{-\widetilde{c}_0 \wedge c_2} \big) \right).$$

where $\rho = \mathcal{B}/d_1$. We choose δ to be $\frac{1}{(n\rho)^{2/5} \log d_2}$ and have $\delta > \frac{1}{n^{2/5} \log d_2}$ (since $\rho < 1$). Thus this choice of δ satisfies the requirement in Lemma C.1.8. Finally we have (C.1.44) is bounded as

$$\frac{\log d_2}{\epsilon^2} \left(\frac{1}{d_0 \rho} + \frac{s(\log d_2)^2}{n^{1/2}} + \frac{(\log d_2)^2}{(n\rho)^{1/5}} + \frac{1}{\rho d_2} \right).$$

Under the stated assumption in Theorem 3.4.2, the above term is $o(1)$. Thus the FDP control result is established. Due to similar derivations as in Theorem 3.5.2, the FDR control result follows. □

C.1.4 Ancillary lemmas for Theorem 3.4.2

To prove FDR control, we will establish a key result, i.e., Lemma C.1.8 in this section. Recall that in Section 3.4, we utilize the following result

$$\sqrt{n}(\widetilde{\boldsymbol{\Theta}}_j^{\mathrm{d}} - \boldsymbol{\Theta}_j) = Z_j + \Xi, \quad Z_j | \boldsymbol{X} \sim \mathcal{N}(0, \sigma_j^2 M \widehat{\boldsymbol{\Sigma}} M^\top).$$

and approximate the quantile of the maximum statistics $T_E = \max_{(j,k)\in E} \sqrt{n}|\widetilde{\boldsymbol{\Theta}}^{\mathrm{d}}_{jk}|$ by $T^{\mathcal{N}}_E = \max_{(j,k)\in E} |Z_{jk}|$. Lemma C.1.8 basically establishes the Cramér deviation bounds for such quantile approximation. Note that this lemma can be seen as a special case of Proposition C.3.1 since the involving random vector $\sqrt{n}(\widetilde{\boldsymbol{\Theta}}^{\mathrm{d}}_j - \boldsymbol{\Theta}_j)$ can be decomposed into a Gaussian random vector plus some error term. Hence we do not need to use the results in Kuchibhotla et al. (2021) to handle the case of a general random vector (and quantify Gaussian approximation errors).

In this section, we will define some notations similar to the theoretical results in Appendix C.3. First, we will drop the j-th subscript for simplicity. Without loss of generality, we prove relevant results for $E = \{(j,k) : k \in [d_2]\}$ and drop the subscript E. Note the results hold for any $j \in [d_1]$ and any subset of $\{(j,k) : k \in [d_2]\}$. Now we rewrite (3.4.4) using new notations, i.e.,

$$\sqrt{n}(\widetilde{\boldsymbol{\Theta}}^{\mathrm{d}}_j - \boldsymbol{\Theta}_j) = \boldsymbol{Z} + \Xi, \quad \boldsymbol{Z}|\boldsymbol{X} \sim \mathcal{N}(0, \sigma_j^2 M\widehat{\boldsymbol{\Sigma}}M^\top), \tag{C.1.45}$$

and denote its maximum by $T_{\boldsymbol{Z}} = ||\boldsymbol{Z}||_\infty$. Intuitively, we can use the quantile of $T_{\boldsymbol{Z}}$ to approximate the quantile of $T := \sqrt{n}||\widetilde{\boldsymbol{\Theta}}^{\mathrm{d}}_j - \boldsymbol{\Theta}_j||_\infty$. Since the covariance matrix $\sigma_j^2 M\widehat{\boldsymbol{\Sigma}}M^\top$ of the Gaussian random vector $\boldsymbol{Z}$ is not completely known, we can not directly compute its quantile (denoted by $q(\alpha; \boldsymbol{Z})$). Instead, we first estimate the unknown parameter σ_j by $\widehat{\sigma}_j$, which is constructed according to (3.4.5). Then we define $\boldsymbol{W} \sim \mathcal{N}(0, \widehat{\sigma}_j^2 M\widehat{\boldsymbol{\Sigma}}M^\top)$ (given the data $\boldsymbol{X}, \boldsymbol{Y}^{(j)}$), and denote its maximum by $T_{\boldsymbol{W}} = ||W||_\infty$. We will approximate the unknown quantile of T by the conditional quantile $q_\xi(\alpha; T_{\boldsymbol{W}})$. Here we use the ξ subscript to emphasize that we are conditioning on the data when defining such quantiles.

Due to the existence of the term Ξ in (C.1.45), there also exist additional estimation errors when we approximate the quantiles of T by the conditional quantiles $q_\xi(\alpha; T_{\boldsymbol{W}})$. Lemma C.1.7 characterizes such approximation errors. As for the difference between the distributions of the two Gaussian random vectors $\boldsymbol{W}$ and $\boldsymbol{Z}$, Lemma C.1.7 provides a bound on the maximal difference of their

covariance matrices, which is denoted by Δ_∞. Finally, Lemma C.1.8 builds on these results and establishes the Cramér-type deviation bounds for the quantile approximation of T.

Lemma C.1.6. *In the context of multiple linear models, we have*

$$\mathbb{P}(|T - T_{\boldsymbol{Z}}| > \zeta_1) < \zeta_2,$$

where $\zeta_1 = O(s \log d_2/\sqrt{n})$ *and* $\zeta_2 = O(e^{-c_1 n} + d_2^{-\widetilde{c}_0 \wedge c_2})$.

Proof of Lemma C.1.6. By Theorem 2.5 in Javanmard & Montanari (2014a), we have

$$\sqrt{n}(\widetilde{\boldsymbol{\Theta}}_j^{\mathrm{d}} - \boldsymbol{\Theta}_j) = \boldsymbol{Z} + \Xi, \quad \boldsymbol{Z}|\boldsymbol{X} \sim \mathcal{N}(0, \sigma_j^2 M \widehat{\boldsymbol{\Sigma}} M^\top),$$

and

$$\mathbb{P}\left(\|\Xi\|_\infty \geq \left(\frac{16 a c \sigma}{C_{\min}} \right) \frac{s \log d_2}{\sqrt{n}} \right) \leq 4\, e^{-c_1 n} + 4\, d_2^{-\widetilde{c}_0 \wedge c_2}\,.$$

Thus we immediately obtain the following bound on the difference between T and $T_{\boldsymbol{Z}}$:

$$\mathbb{P}(|T - T_{\boldsymbol{Z}}| > \zeta_1) < \zeta_2$$

where $\zeta_1 = O(s \log d_2/\sqrt{n})$ and $\zeta_2 = O(e^{-c_1 n} + d_2^{-\widetilde{c}_0 \wedge c_2})$. □

Lemma C.1.7. *For the the maximal difference term* $\Delta_\infty = ||\widehat{\sigma}^2 M \widehat{\boldsymbol{\Sigma}} M^\top - \sigma^2 M \widehat{\boldsymbol{\Sigma}} M^\top||_{\max}$, *we have*

$$\mathbb{P}\left(\Delta_\infty \geq \delta\right) \leq \eta, \tag{C.1.46}$$

where δ *satisfies* $\frac{1}{\delta}\sqrt{\frac{s \log d_2}{n}} = O(1)$ *and* $\eta = O\left(e^{-c_1 n} + \frac{1}{d_2} + \frac{1}{n\delta^2}\right)$.

Proof of Lemma C.1.7. To bound Δ_∞, we start with the term $|\widehat{\sigma}/\sigma - 1|$. First we denote

$$\mathcal{E}_n = \mathcal{E}_n(\phi_0, s_0, K) := \Big\{ \boldsymbol{X} \in \mathbb{R}^{n \times d_1} : \min_{S:|S| \le s_0} \phi(\widehat{\boldsymbol{\Sigma}}, S) \ge \phi_0, \max_{j \in [d_1]} \boldsymbol{\Sigma}_{jj} \le K, \boldsymbol{\Sigma} = (\boldsymbol{X}^\top \boldsymbol{X})/n \Big\}$$

similarly as in Theorem 7.(a) of Javanmard & Montanari (2014a), where $\phi(\widehat{\boldsymbol{\Sigma}}, S)$ is the compatibility constant as defined in Definition 1 of Javanmard & Montanari (2014a). Following the proof of Lemma 14 in Javanmard & Montanari (2014a), we have

$$\begin{aligned} \mathbb{P}\left(\left| \frac{\widehat{\sigma}}{\sigma} - 1 \right| \ge \epsilon \right) &\le \mathbb{P}\left(\boldsymbol{X} \notin \mathcal{E}_n \right) + \sup_{\boldsymbol{X} \in \mathcal{E}_n} \mathbb{P}\Big(\Big| \frac{\widehat{\sigma}}{\sigma} - 1 \Big| \ge \epsilon \,\Big|\, \boldsymbol{X} \Big) \\ &\le 4e^{-c_1 n} + \sup_{\boldsymbol{X} \in \mathcal{E}_n} \mathbb{P}\Big(\frac{||\boldsymbol{X}^\top \boldsymbol{E}||_\infty}{n\sigma^\star} \ge \widetilde{\lambda}/4 \,\Big|\, \boldsymbol{X} \Big) + \sup_{\boldsymbol{X} \in \mathcal{E}_n} \mathbb{P}\Big(\Big| \frac{\sigma^\star}{\sigma} - 1 \Big| \ge \frac{\epsilon}{10} \,\Big|\, \boldsymbol{X} \Big) \end{aligned} \tag{C.1.47}$$

where $\widetilde{\lambda} = 10\sqrt{(2 \log d_2)/n}$, $\sigma^\star$ is the oracle estimator of σ introduced in Sun & Zhang (2012) and ϵ satisfies $\frac{2\sqrt{s}\widetilde{\lambda}}{\sigma^\star \phi_0} \le \frac{\epsilon}{2} < a_0$.

Now we separately bound the last two terms in (C.1.47). The second term in (C.1.47) can be bounded by the derivation in the proof of Theorem 2 (ii) (Sun & Zhang, 2012), i.e.,

$$\begin{aligned} \sup_{\boldsymbol{X} \in \mathcal{E}_n} \mathbb{P}\Big(\frac{||\boldsymbol{X}^\top \boldsymbol{E}||_\infty}{n\sigma^\star} \ge \widetilde{\lambda}/4 \,\Big|\, \boldsymbol{X} \Big) &\le d_2 \mathbb{P}\Big(|L_k| \ge \sqrt{2 \log(d_2^{25/4})/n} \,\Big|\, \boldsymbol{X} \Big) \\ &\le d_2 \cdot \frac{C}{d_2^{25/4}\sqrt{\log d_2}} \le \frac{C}{d_2}, \end{aligned} \tag{C.1.48}$$

where L_k is the k-th element of $\frac{\boldsymbol{X}^\top \boldsymbol{E}}{n\sigma^\star}$ and $\frac{\sqrt{n-1}L_k}{\sqrt{1-L_k^2}}$ follows the Student's t-distribution with $n-1$ degrees of freedom. Then (C.1.48) holds due to equation (A7) in Sun & Zhang (2012) together with the union bound. As for the last term in (C.1.47), we note $n(\sigma^\star/\sigma)^2$ follows the χ^2_n distribu-

tion according to Sun & Zhang (2012). Thus by Markov's inequality, we have

$$\sup_{\boldsymbol{X}\in\mathcal{E}_n} \mathbb{P}\Big(\Big|\frac{\sigma^\star}{\sigma}-1\Big|\geq\frac{\epsilon}{10}\,\Big|\,\boldsymbol{X}\Big) \;\leq\; \frac{C'\mathbb{E}\left[(n(\sigma^\star/\sigma)^2-n)^2\right]}{n^2\epsilon^2}\leq\frac{2C'}{n\epsilon^2}. \tag{C.1.49}$$

Now we arrive at the following bound on Δ_∞:

$$\begin{aligned}
&\mathbb{P}\Big(\Delta_\infty\geq(\epsilon^2+2\epsilon)\cdot\sigma^2||M\widehat{\boldsymbol{\Sigma}}M^\top||_{\max}\Big)\\
=\;&\mathbb{P}\Big(||\widehat{\sigma}^2M\widehat{\boldsymbol{\Sigma}}M^\top-\sigma^2M\widehat{\boldsymbol{\Sigma}}M^\top||_{\max}\geq(\epsilon^2+2\epsilon)\cdot\sigma^2||M\widehat{\boldsymbol{\Sigma}}M^\top||_{\max}\Big)\\
\leq\;&\mathbb{P}\Big(\Big|\frac{\widehat{\sigma}}{\sigma}-1\Big|\geq\epsilon\Big)\\
\leq\;&4e^{-c_1n}+\frac{C}{d_2}+\frac{2C'}{n\epsilon^2},
\end{aligned}$$

where the last inequality comes from combining (C.1.47), (C.1.48) with (C.1.49). Note that the proof of Theorem 16 in Javanmard & Montanari (2014a) shows that $||M\widehat{\boldsymbol{\Sigma}}M^\top||_{\max}=O(1)$. Hence, we finally establish (C.1.46) with

$$\delta=\sigma^2(\epsilon^2+2\epsilon)||M\widehat{\boldsymbol{\Sigma}}M^\top||_{\max}=C_\sigma\epsilon,\quad \eta=O\left(e^{-c_1n}+\frac{1}{d_2}+\frac{1}{n\delta^2}\right).$$

where C_σ is some constant and $\delta=C_\sigma\epsilon$ satisfies $\frac{1}{\delta}\sqrt{\frac{s\log d_2}{n}}=O(1)$ due to the choice of ϵ. □

Lemma C.1.8. *Based on the result about the approximation error between T and $T_{\boldsymbol{Z}}$ (Lemma C.1.6) and the bound on $||\Delta||_\infty$ in Lemma C.1.7, we have*

$$\sup_{\alpha\in[\alpha_L,1]}\left|\frac{\mathbb{P}(T>q(\alpha;T_{\boldsymbol{W}}))}{\mathbb{P}(T_{\boldsymbol{Z}}>q(\alpha;T_{\boldsymbol{Z}}))}-1\right|=O\left(\eta_0(d_1,d_2,n,\zeta_1,\zeta_2,\delta,\alpha_L)\right), \tag{C.1.50}$$

where $\eta_0(d_1,d_2,n,\zeta_1,\zeta_2,\delta,\alpha_L):=\zeta_1\log d_2+(\log d_2)^{5/2}\delta^{1/2}+\frac{\eta+\zeta_2}{\alpha_L}$ with $\zeta_1=O(s\log d_2/\sqrt{n})$, $\zeta_2=O(e^{-c_1n}+d_2^{-\widetilde{c}_0\wedge c_2})$. Here δ is a term to be determined and we requite $\frac{1}{\delta}\sqrt{\frac{s\log d_2}{n}}=O(1)$. η

depends on δ, i.e., $\eta = e^{-c_1 n} + \frac{1}{d_2} + \frac{1}{n\delta^2}$.

Proof of Lemma C.1.8. First we have $\mathbb{P}(|T - T_{\boldsymbol{Z}}| > \zeta_1) < \zeta_2$ by Lemma C.1.6, thus we obtain

$$\left| \frac{\mathbb{P}(T > q(\alpha; T_{\boldsymbol{W}}))}{\mathbb{P}(T_{\boldsymbol{Z}} > q(\alpha; T_{\boldsymbol{Z}}))} - 1 \right| \leq \max\{\mathrm{II}_1, \mathrm{II}_2\} + \frac{2\zeta_2}{\alpha}$$

for $\alpha \in [\alpha_L, 1]$, where II_1 and II_2 are defined as:

$$\mathrm{II}_1 := \left| \frac{\mathbb{P}(T_{\boldsymbol{Z}} > q(\alpha; T_{\boldsymbol{W}}) + \zeta_1)}{\mathbb{P}(T_{\boldsymbol{Z}} > q(\alpha; T_{\boldsymbol{Z}}))} - 1 \right|, \quad \mathrm{II}_2 := \left| \frac{\mathbb{P}(T_{\boldsymbol{Z}} > q(\alpha; T_{\boldsymbol{W}}) - \zeta_1)}{\mathbb{P}(T_{\boldsymbol{Z}} > q(\alpha; T_{\boldsymbol{Z}}))} - 1 \right|.$$

The above two terms can be bounded similarly. Take II_1 as an example, we use similar strategy as in Proposition C.3.1. Consider the event $S := \{\Delta_\infty \leq \delta\}$ where δ satisfies $\frac{1}{\delta}\sqrt{\frac{s \log d_2}{n}} = O(1)$, we apply Lemma C.3.3 and bound II_1 by

$$\frac{1}{1 - \pi(\delta)} \cdot \mathrm{II}_{11} + \mathrm{II}_{12} + \frac{\mathbb{P}\left(\Delta_\infty > \delta\right)}{\alpha},$$

where II_{11} and II_{11} are defined as

$$\begin{aligned} \mathrm{II}_{11} &:= \frac{1 - \pi(\delta)}{\alpha} \left| \mathbb{P}\Big(T_{\boldsymbol{Z}} > q(\frac{\alpha}{1 - \pi(\delta)}; T_{\boldsymbol{Z}}) + \zeta_1\Big) - \mathbb{P}(T_{\boldsymbol{Z}} > q(\frac{\alpha}{1 - \pi(\delta)}; T_{\boldsymbol{Z}})) \right|, \\ \mathrm{II}_{12} &:= \frac{1}{\alpha} \left| \mathbb{P}\Big(T_{\boldsymbol{Z}} > q(\frac{\alpha}{1 - \pi(\delta)}; T_{\boldsymbol{Z}})\Big) - \mathbb{P}(T_{\boldsymbol{Z}} > q(\alpha; T_{\boldsymbol{Z}})) \right|, \end{aligned} \tag{C.1.51}$$

where $\pi(\Delta_\infty) = [A(\Delta_\infty) + 1]e^{M_1 (\log d)^{3/2} A(\Delta_\infty)} - 1$. By applying the part 3 of Theorem 2.1 in Kuchibhotla et al. (2021) (with $r + \epsilon = q(\frac{\alpha}{1-\pi(\delta)}; T_{\boldsymbol{Z}}) + \zeta_1, r - \epsilon = q(\frac{\alpha}{1-\pi(\delta)}; T_{\boldsymbol{Z}})$) to the Gaussian random vector $\boldsymbol{Z}$, we have

$$\mathrm{II}_{11} \leq K_4 \zeta_1 \big(q(\frac{\alpha}{1 - \pi(\delta)}; T_{\boldsymbol{Z}}) + \zeta_1/2\big) \leq C \zeta_1 \log d_2. \tag{C.1.52}$$

where the second inequality holds due to the similar reason stated in the proof of Proposition C.3.2. And the term II_{11} can be simply derived as

$$\mathrm{II}_{12} = \frac{1}{\alpha}\left|\frac{\alpha}{1-\pi(\delta)} - \alpha\right| = \frac{\pi(\delta)}{1-\pi(\delta)}. \tag{C.1.53}$$

Combing the results above, we have

$$\left|\frac{\mathbb{P}(T > q(\alpha; T_{\boldsymbol{W}}))}{\mathbb{P}(T_{\boldsymbol{Z}} > q(\alpha; T_{\boldsymbol{Z}}))} - 1\right| \leq C'\zeta_1 \log d_2 + \frac{\pi(\delta)}{1-\pi(\delta)} + \frac{\pi(\delta)}{1+\pi(\delta)} + \frac{2\mathbb{P}\left(\Delta_\infty > \delta\right)}{\alpha} + \frac{2\zeta_2}{\alpha}.$$

Applying the bound in Lemma C.1.7, we finally establish (C.1.50) i.e., $\eta_0(d_1, d_2, n, \zeta_1, \zeta_2, \alpha_L) := \zeta_1 \log d_2 + (\log d_2)^{5/2}\delta^{1/2} + \frac{\eta+\zeta_2}{\alpha_L}$ up to some constant factor, where $\eta = e^{-c_1 n} + \frac{1}{d_2} + \frac{1}{n\delta^2}$. □

C.2 Proofs of Cramér-type comparison bounds

In this section, we will prove two types of Cramér-type comparison bounds: Theorems 3.3.1 and 3.3.2. One of the challenges to derive the comparison bounds for Gaussian maxima is that the maximum function is non-smooth. In order to show the Cramér-type comparison bound, we first consider smooth approximation of the maximum. The following lemma from Bentkus (1990) show the existence of such smooth approximation.

Lemma C.2.1 (Theorem 1, Bentkus (1990)). *Consider the Euclidean space $\mathbb{R}^d$ with ℓ_∞-norm, for any $t, \epsilon \geq 0$, there exists a smooth approximating function $\varphi_{r,\epsilon}$ satisfying the following:*

(a) $\varphi_{r,\epsilon} : \mathbb{R}^d \to [0,1], \varphi_{r,\epsilon} \in \mathbb{C}^\infty$, *where $\mathbb{C}^\infty$ is the smooth function class with functions differentiable for all degrees of differentiation.*

(b) $\varphi_{r,\epsilon}(x) = 1$ *if* $||x||_\infty \leq r$, $\varphi_{r,\epsilon}(x) = 0$ *if* $||x||_\infty \geq r + \epsilon$,

(c) $\sup_{x\in\mathbb{R}^d} ||D^j \varphi_{r,\epsilon}(x)||_1 \leq c(j)\epsilon^{-j}\log^{j-1}(d+1)$,

where $||D^j\varphi_{r,\epsilon}(x)||_1 = \sum_{i_1=1}^d \cdots \sum_{i_j=1}^d \left|\frac{\partial^j \varphi_{r,\epsilon}(x)}{\partial x_{i_1}\cdots\partial x_{i_j}}\right|$ *and the constants* $c(j)$ *only depends on* j.

Remark C.2.1.1. *Kuchibhotla et al. (2021) gives a concrete example of* $\varphi_{r,\epsilon}(x)$ *satisfying the three properties in Lemma C.2.1:*

$$\varphi_{r,\epsilon}(x) = g_0\left(\frac{2(F_\beta(z_x - r\mathbf{1}_{2d}) - \epsilon/2)}{\epsilon}\right), \tag{C.2.1}$$

where $\beta = 2\log(2d)/\epsilon$, $g_0(t) := 30\mathbb{1}(0 \le t \le 1)\int_t^1 s^2(1-s)^2 ds + \mathbb{1}(t \le 0)$, $F_\beta(\cdot)$ *is the "softmax" function*

$$F_\beta(z) := \frac{1}{\beta}\log\Big(\sum_{m=1}^{2d}\exp(\beta z_m)\Big) \quad \textit{for } z \in \mathbb{R}^{2d},$$

$z_x = (x^\top, -x^\top)^\top$, *and* $\mathbf{1}_{2d}$ *is the vector of* 1*'s of dimension* $2d$.

In fact, in the proof of Theorem 3.3.1, we do not need a specific form of $\varphi_{r,\epsilon}(x)$ *and any function satisfying Lemma C.2.1 will work. While in the proof of Theorem 3.3.2, we need to utilize the specific form in* (C.2.1).

C.2.1 Proof of Theorem 3.3.1

As mentioned in Remark 3.3.1.1, we can prove the Cramér-type comparison bound with max norm difference as $M_3(\log d)^{3/2}A(\Delta_\infty)e^{M_3(\log d)^{3/2}A(\Delta_\infty)}$, without the assumption on Δ_∞. Therefore we state the more general form of Theorem 3.3.1 below and give its proof. Note that under the assumption $(\log d)^5\Delta_\infty = O(1)$ and the discussions in Remark 3.3.1.1, the bound (3.3.1) in Theorem 3.3.1 immediately follows from Theorem C.2.2.

Theorem C.2.2 (CCB with max norm difference)**.** *Let* U *and* V *be two Gaussian random vectors*

and we have

$$\sup_{0\le t\le C_0\sqrt{\log d}}\left|\frac{\mathbb{P}(||U||_\infty > t)}{\mathbb{P}(||V||_\infty > t)} - 1\right| \le M_3(\log d)^{3/2}A(\Delta_\infty)e^{M_1(\log d)^{3/2}A(\Delta_\infty)}, \qquad \text{(C.2.2)}$$

where $C_0 > 0$ *is some constant,* $A(\Delta_\infty) = M_1 \log d\Delta_\infty^{1/2}\exp\left(M_2\log^2 d\Delta_\infty^{1/2}\right)$*, the constants* M_1, M_2 *only depend on* $\min_{1\le j\le d}\{\sigma_{jj}^U, \sigma_{jj}^V\}, \max_{1\le j\le d}\{\sigma_{jj}^U, \sigma_{jj}^V\}$*, and* M_3 *is a universal constant.*

Proof of Theorem C.2.2. Using the smooth approximation in Lemma C.2.1, we can bound the difference between the distribution functions of Gaussian maxima as

$$\begin{aligned} & \left|\mathbb{P}(||U||_\infty > t) - \mathbb{P}(||V||_\infty > t)\right| \\ = \; & \left|\mathbb{E}[\mathbb{1}(||U||_\infty \le t) - \mathbb{1}(||V||_\infty \le t)]\right| \\ \le \; & \mathbb{P}(t-\epsilon \le ||V||_\infty \le t+\epsilon) + \max_{j=1,2}\left|\mathbb{E}\varphi_j(U) - \mathbb{E}\varphi_j(V)\right|, \qquad \text{(C.2.3)} \end{aligned}$$

where $\varphi_1(x) := \varphi_{t,\epsilon}(x), \varphi_2(x) := \varphi_{t-\epsilon,\epsilon}(x)$. Regarding the inequality in (C.2.3), we first notice that

$$\mathbb{1}(||x||_\infty \le t) = \varphi_{t,\epsilon}(x) - \mathbb{1}(t < ||x||_\infty < t+\epsilon)\cdot\varphi_{t,\epsilon}(x) = \varphi_{t-\epsilon,\epsilon}(x) - \mathbb{1}(t-\epsilon < ||x||_\infty < t)\cdot\varphi_{t-\epsilon,\epsilon}(x),$$

where the first equality is due to property (b) in Lemma C.2.1. Hence we have

$$\begin{aligned} \mathbb{1}(||U||_\infty \le t) &\le \varphi_j(U), \quad j = 1,2 \\ \mathbb{1}(||V||_\infty \le t) &\ge \varphi_1(V) - \mathbb{1}(t < \|V\| < t+\epsilon), \\ \mathbb{1}(||V||_\infty \le t) &\ge \varphi_2(V) - \mathbb{1}(t-\epsilon < \|V\| < t), \end{aligned}$$

then (C.2.3) immediately follows by combining the above three inequalities.

The first term in (C.2.3) is related to the anti-concentration inequalities for the Gaussian maxima. By applying Theorem 2.1 in Kuchibhotla et al. (2021), we have

$$\mathbb{P}(t-\epsilon \leq ||V||_\infty \leq t+\epsilon) \leq K_1(t+1)\epsilon \exp(K_2(t+1)\epsilon)\mathbb{P}(||V||_\infty > t). \tag{C.2.4}$$

The explicit forms of K_1, K_2 can be found in Theorem 2.1 of Kuchibhotla et al. (2021). They only depend on $\min_{1\leq j\leq d}\{\sigma_{jj}^U, \sigma_{jj}^V\}$, $\max_{1\leq j\leq d}\{\sigma_{jj}^U, \sigma_{jj}^V\}$ and the median of Gaussian maxima. Remark that the median of $||V||_\infty$ is bounded by $O(\sqrt{\log d})$ by the maximal inequalities for sub-Gaussian random variables (Lemma 5.2 in van Handel (2014)). Plugging this into the explicit form of K_1, K_2 in Theorem 2.1 of Kuchibhotla et al. (2021), we have $K_1 = O(\log d), K_2 = O(\log^2 d)$. Then (C.2.4) can be written as

$$\mathbb{P}(t-\epsilon \leq ||V||_\infty \leq t+\epsilon) \leq M_1 \log d(t+1)\epsilon \exp(M_2 \log^2 d\,(t+1)\epsilon)\mathbb{P}(||V||_\infty > t),$$

for some constants M_1, M_2 only depending on $\min_{1\leq j\leq d}\{\sigma_{jj}^U, \sigma_{jj}^V\}, \max_{1\leq j\leq d}\{\sigma_{jj}^U, \sigma_{jj}^V\}$.

Overall the above bound has only a logarithmic dependence on the dimension d, similar to the anti-concentration bounds from Chernozhukov et al. (2014). But it quantifies the deviation with respect to the tail probability of the Gaussian maxima, thus offers a more refined characterization, which is crucial to our proof.

Now we deal with the second term in (C.2.3). It is not hard to check that the following proof works for both φ_1 and φ_2. Therefore, without loss of generality, we use a unified notation φ to represent either functions. We consider the Slepian interpolation between U and V: $W(s) :=$

$\sqrt{s}U + \sqrt{1-s}V,\ \ s \in [0,1]$. Let $\Psi_t(s) = \mathbb{E}[\varphi(W(s))]$, then we have

$$|\mathbb{E}\varphi(U) - \mathbb{E}\varphi(V)| = |\Psi_t(1) - \Psi_t(0)| = \left| \int_0^1 \Psi_t'(s) ds \right|, \tag{C.2.5}$$

where $\Psi_t'(s) = \frac{1}{2}\sum_{j=1}^d \mathbb{E}[\partial_j \varphi(W(s))(s^{-1/2}U_j - (1-s)^{-1/2}V_j)]$. Applying Stein's identity (Lemma 2 of Chernozhukov et al. (2015)) to $(s^{-1/2}U_j - (1-s)^{-1/2}V_j, W(s)^\top)^\top$ and $\partial_j \varphi(W(s))$, we have

$$\Psi_t'(s) = \frac{1}{2}\sum_{j,k=1}^d (\sigma_{jk}^U - \sigma_{jk}^V)\mathbb{E}[\partial_j \partial_k \varphi(W(s))]. \tag{C.2.6}$$

Hence we obtain the following bound on (C.2.5),

$$\begin{aligned}
\left| \int_0^1 \Psi_t'(s) ds \right| &\leq \frac{1}{2}\sum_{j,k=1}^d |\sigma_{jk}^U - \sigma_{jk}^V| \cdot \left| \int_0^1 \mathbb{E}[\partial_j \partial_k \varphi(W(s))] ds \right| \\
&\leq \frac{\Delta_\infty}{2} \int_0^1 \sum_{j,k=1}^d \mathbb{E}[|\partial_j \partial_k \varphi(W(s))|] ds \\
&\leq \frac{\Delta_\infty}{2} \int_0^1 \sum_{j,k=1}^d \mathbb{E}[|\partial_j \partial_k \varphi(W(s))| \cdot \mathbb{1}(t-\epsilon \leq ||W(s)||_\infty \leq t+\epsilon)] ds \\
&\leq \frac{\Delta_\infty}{2} \int_0^1 \sup_{x \in \mathbb{R}^d} ||D^2 \varphi(x)||_1 \cdot \mathbb{E}[\mathbb{1}(t-\epsilon \leq ||W(s)||_\infty \leq t+\epsilon))] ds \\
&\leq \frac{c(2)\Delta_\infty \log(d+1)}{2\epsilon^2} \int_0^1 \mathbb{P}(t-\epsilon \leq ||W(s)||_\infty \leq t+\epsilon) ds
\end{aligned} \tag{C.2.7}$$

where the second inequality is by the definition of Δ_∞ and the third one comes from the property (b) in Lemma C.2.1 for $\varphi_j(x), j = 1, 2$ (recalling $\varphi_1(x) = \varphi_{t,\epsilon}(x)$ and $\varphi_2(x) = \varphi_{t-\epsilon,\epsilon}(x)$). Note that property (c) gives a upper bound for the partial derivative terms. Thus the fourth inequality holds.

By the definition of Slepian interpolation, we have, for any $s \in [0,1]$, $W(s)$ is a Gaussian ran-

dom vector and the variances can be controlled between $\min_{1\le j\le d}\{\sigma_{jj}^U, \sigma_{jj}^V\}$ and $\max_{1\le j\le d}\{\sigma_{jj}^U, \sigma_{jj}^V\}$. The median of $||W(s)||_\infty$ can also be similarly bounded by $O(\sqrt{\log d})$ as $||V||_\infty$. Applying the anti-concentration inequalities again to $W(s)$ in (C.2.7), we thus obtain

$$\left|\int_0^1 \Psi_t'(s)ds\right| \le \frac{c(2)\Delta_\infty \log(d+1)}{2\epsilon^2}\cdot M_1 \log d(t+1)\epsilon \exp(M_2 \log^2 d(t+1)\epsilon)\cdot \int_0^1 \mathbb{P}(||W(s)||_\infty > t)ds. \tag{C.2.8}$$

Let $Q_t(u) = \mathbb{P}(||W(u)||_\infty > t)$ and $R_t(u) = Q_t(u)/Q_t(0) - 1$. Combining (C.2.3), (C.2.4), (C.2.5) and (C.2.8), we have

$$\begin{aligned} |Q_t(1) - Q_t(0)| = &\left|\mathbb{P}(||U||_\infty > t) - \mathbb{P}(||V||_\infty > t)\right| \\ \le &\ M_1 \log d(t+1)\epsilon \exp(M_2 \log^2 d(t+1)\epsilon) Q_t(0) \\ &+ \frac{c(2)\Delta_\infty \log(d+1)}{2\epsilon^2} M_1 \log d(t+1)\epsilon \exp(M_2 \log^2 d(t+1)\epsilon)\int_0^1 Q_t(s)ds. \end{aligned} \tag{C.2.9}$$

If starting with the interpolation between $W(s)$ and V instead of that between U and V, we can similarly obtain the bound on $|Q_t(s) - Q_t(0)|$. And the integral $\int_0^1 Q_t(s)ds$ in (C.2.9) can be directly replaced by $\int_0^u Q_t(s)ds$. Namely, we have

$$\frac{|Q_t(u) - Q_t(0)|}{|Q_t(0)|} = |R_t(u)| \le A(t,\epsilon)B(\Delta_\infty,\epsilon)\int_0^u |R_t(s)|ds + A(t,\epsilon)B(\Delta_\infty,\epsilon)\cdot u + A(t,\epsilon), \tag{C.2.10}$$

where we denote $A(t,\epsilon) = M_1 \log d(t+1)\epsilon \exp(M_2 \log^2 d(t+1)\epsilon)$ and $B(\Delta_\infty,\epsilon) = \frac{c(2)\Delta_\infty \log(d+1)}{2\epsilon^2}$.

Notice that (C.2.10) is an integral inequality and we can thus bound $R_t(s)$ by Grönwall's in-

equality (Grönwall, 1919)

$$|R_t(u)| \leq (A(t,\epsilon)B(\Delta_\infty,\epsilon)u + A(t,\epsilon))e^{A(t,\epsilon)B(\Delta_\infty,\epsilon)u}.$$

In particular, we have $|R_t(1)| \leq (A(t,\epsilon)B(\Delta_\infty,\epsilon) + A(t,\epsilon))e^{A(t,\epsilon)B(\Delta_\infty,\epsilon)}$. Remember that ϵ is the smoothing parameter that controls the level of approximation. Choosing $\epsilon = \Delta_\infty^{1/2}/(t+1)$, we then have $A(\Delta_\infty) := A(t,\epsilon) = M_1 \log d \Delta_\infty^{1/2} \exp\left(M_2 \log^2 d \Delta_\infty^{1/2}\right)$ for some constants M_1, M_2 only depending on $\min_{1\leq j\leq d}\{\sigma_{jj}^U, \sigma_{jj}^V\}$, $\max_{1\leq j\leq d}\{\sigma_{jj}^U, \sigma_{jj}^V\}$ and $B(t) := B(\Delta_\infty,\epsilon) = \frac{c(2)\log(d+1)(t+1)^2}{2}$. When $0 \leq t \leq C_0\sqrt{\log d}$, we have $B(t) \leq M_1(\log d)^{3/2}$ for some universal constant M_3. Therefore the bound in (C.2.2) is established, i.e.,

$$\sup_{0\leq t\leq C_0\sqrt{\log d}} |R_t(1)| \leq M_3(\log d)^{3/2} A(\Delta_\infty) e^{M_3(\log d)^{3/2}A(\Delta_\infty)}.$$

□

C.2.2 Proof of Theorem 3.3.2

Before proving Theorem 3.3.2, we note its assumption about the connectivity can be relaxed. Therefore, we first present Theorem C.2.4 with a weaker connectivity assumption, which is stated below.

Assumption C.2.3 ($\mathfrak{p}$-connectivity property)**.** We say two Gaussian random vectors U and V satisfy the $\mathfrak{p}$-connectivity property if for any j such that $\sigma_{jk}^U \neq \sigma_{jk}^V$ for some k, there exists a subset $\mathcal{E}_0 \subset [d]$ satisfying the following three requirements:

(a) $j \in \mathcal{E}_0, |\mathcal{E}_0| = \mathfrak{p} + 1$;

(b) When $m, m' \in \mathcal{E}_0$ and $m \neq m'$, $\sigma_{mm}^U = \sigma_{m'm'}^U$ and $\sigma_{mm'}^U = \sigma_{mm'}^V = 0$ hold;

(c) $\forall\, k \in [d], |\{m \in \mathcal{E}_0 : |\sigma_{km}^U| + |\sigma_{km}^V| \neq 0\}| \leq c_0$ for some constant c_0.

This assumption gives a characterization of the connectivity of the associated graphs of the Gaussian random vectors U and V. Below we give a few sufficient conditions (SC) for it.

SC1 U and V have unit variances. There exists a disjoint $(\mathfrak{p}+2)$-partition of nodes $\cup_{\ell=1}^{\mathfrak{p}+2}\mathcal{C}_\ell = [d]$ such that $\sigma_{jk}^U = \sigma_{jk}^V = 0$ when $j \in \mathcal{C}_\ell$ and $k \in \mathcal{C}_{\ell'}$ for some $\ell \neq \ell'$.

SC2 U and V have unit variances. There exist disjoint partitions of nodes $\cup_{\ell=1}^{\mathfrak{p}+2}\mathcal{C}_\ell^U = \cup_{\ell=1}^{\mathfrak{p}+2}\mathcal{C}_\ell^V = [d]$, such that σ_{jk}^U (σ_{jk}^V) equals 0 when j, k belong to different elements $\mathcal{C}_\ell^U$ ($\mathcal{C}_\ell^V$), and $\forall \ell \in [\mathfrak{p}+2], \mathcal{C}_\ell^U \cap \mathcal{C}_\ell^V \neq \varnothing$.

SC3 $\forall\, s \in [0,1]$, the Gaussian random vector $W(s) := \sqrt{s}U + \sqrt{1-s}V$ always has the same variances σ_s^2 across different components. The associated graph of $W(s)$ has at least $\mathfrak{p}+2$ components, i.e., there exists a disjoint partition of nodes $\cup_{\ell=1}^{\mathfrak{p}+2}\mathcal{C}_\ell^W = [d]$, such that each $\mathcal{C}_\ell^W$ comes from a different component. And the partition $\cup_{\ell=1}^{\mathfrak{p}+2}\mathcal{C}_\ell^W = [d]$ works any $s \in [0,1]$.

Remark C.2.3.1. *Note that the above first condition SC1 is the main assumption of Theorem 3.3.2 (except that $\mathfrak{p}+2$ is replaced by $\mathfrak{p}$). It is immediate that the condition SC1 implies SC2. We will verify SC2 is indeed a sufficient condition of Assumption C.2.3 in the following paragraph. Regarding SC3, its sufficiency can be verified similarly, thus we omit the details.*

Simply, we have $\sigma_{jj}^U = \sigma_{jj}^V = 1, j \in [d]$ by the unit variance assumption. For any j such that $\sigma_{jk}^U \neq \sigma_{jk}^V$ for some k, we will construct a subset $\mathcal{E}_0$ and show it satisfies the three requirements (a), (b) and (c). Note that the condition SC1 assumes the existence of disjoint partitions of nodes $\cup_{\ell=1}^{\mathfrak{p}+2}\mathcal{C}_\ell^U = \cup_{\ell=1}^{\mathfrak{p}+2}\mathcal{C}_\ell^V = [d]$. We suppose $j \in \mathcal{C}_{\ell_1}^U \cap \mathcal{C}_{\ell_2}^V$ for some ℓ_1, ℓ_2, then $\mathcal{E}_0$ is constructed by including j and picking one element m_ℓ from $\mathcal{C}_\ell^U \cap \mathcal{C}_\ell^V$ for each $\ell \in [\mathfrak{p}+2] \setminus \{\ell_1, \ell_2\}$. As $\mathcal{C}_\ell^U \cap \mathcal{C}_\ell^V \neq \varnothing, \forall \ell \in [\mathfrak{p}+2]$, we have $|\mathcal{E}_0| \geq 1+\mathfrak{p}$, hence the requirement (a) is satisfied. Regarding the requirement (b), when $m, m' \in \mathcal{E}_0, m \neq m'$, we immediately have $\sigma_{mm}^U = \sigma_{m'm'}^V = 1$ by the unit variance assumption. Since every element in $\mathcal{E}_0$ comes from a different component $\mathcal{C}_\ell^U$ ($\mathcal{C}_\ell^V$), we

also have $\sigma^U_{mm'} = \sigma^V_{mm'} = 0$ *when* $m, m' \in \mathcal{E}_0, m \neq m'$. *Lastly, due to the same reason, we have* $\forall k \in [d], |\{m \in \mathcal{E}_0 : |\sigma^U_{km}| + |\sigma^V_{km}| \neq 0\}| \leq 2$. *Hence the requirement (c) is also satisfied.*

Now we prove Theorem C.2.4, which is stated below. Note that it requires weaker connectivity assumption compared with Theorem 3.3.2 but needs to assume minimal eigenvalue conditions.

Theorem C.2.4 (CCB with elementwise ℓ_0 norm difference)**.** *Consider the two Gaussian random vectors* U *and* V *to have equal variances* $\sigma^U_{jj} = \sigma^V_{jj} = O(1)$, *for* $j \in [d]$ *and we assume* $\lambda_{\min}(\mathbf{\Sigma}^U) \geq 1/b_0 > 0, \lambda_{\min}(\mathbf{\Sigma}^V) \geq 1/b_0 > 0$ *for some constant* $b_0 > 0$. *Suppose* U *and* V *also satisfy Assumption C.2.3, we then have*

$$\sup_{0 \leq t \leq C_0 \sqrt{\log d}} \left| \frac{\mathbb{P}(||U||_\infty > t)}{\mathbb{P}(||V||_\infty > t)} - 1 \right| = O\left(\frac{\Delta_0 \log d}{\mathfrak{p}} \right). \tag{C.2.11}$$

for some constant $C_0 > 0$.

Proof of Theorem C.2.4. Following the same derivations as in Theorem C.2.2, we have

$$\begin{aligned} & |\mathbb{P}(||U||_\infty > t) - \mathbb{P}(||V||_\infty > t)| \\ \leq \; & M_1 \log d(t+1)\epsilon \exp(M_2 \log^2 d(t+1)\epsilon)\mathbb{P}(||V||_\infty > t) + \max_{j=1,2} |\mathbb{E}[\varphi_j(U)] - \mathbb{E}[\varphi_j(V)]| \\ \leq \; & M_1 \log d(t+1)\epsilon \exp(M_2 \log^2 d(t+1)\epsilon)\mathbb{P}(||V||_\infty > t) + \left| \int_0^1 \Psi'_t(s) ds \right|, \end{aligned} \tag{C.2.12}$$

where $\epsilon = c/\max\{(\log d)^{3/2}, \mathfrak{p} \log d\}$ for some small enough constant $c > 0$, and the constants M_1, M_2 only depend on $\min_{1 \leq j \leq d}\{\sigma^U_{jj}\}, \max_{1 \leq j \leq d}\{\sigma^U_{jj}\}$. The above two inequalities hold by

(C.2.3), (C.2.4) and (C.2.5). We further bound $\left|\int_0^1 \Psi_t'(s)ds\right|$ as below,

$$
\begin{aligned}
&\left|\int_0^1 \Psi_t'(s)ds\right| \\
\leq\ & \frac{1}{2}\sum_{j,k=1}^{d} |\sigma_{jk}^U - \sigma_{jk}^V| \left|\int_0^1 \mathbb{E}[\partial_j\partial_k\varphi(W(s))]ds\right| \\
\leq\ & \frac{M}{2}\sum_{j\neq k, \sigma_{jk}^U \neq \sigma_{jk}^V} \int_0^1 \mathbb{E}[|\partial_j\partial_k\varphi(W(s))|]ds \\
\leq\ & \frac{M}{2}\sum_{j\neq k, \sigma_{jk}^U \neq \sigma_{jk}^V} \int_0^1 \mathbb{E}[|\partial_j\partial_k\varphi(W(s))| \cdot \mathbb{1}(t-\epsilon \leq ||W(s)||_\infty \leq t+\epsilon)]ds \\
\leq\ & \frac{M\Delta_0}{2}\max_{j\neq k, \sigma_{jk}^U \neq \sigma_{jk}^V} \int_0^1 \mathbb{E}[|\partial_j\partial_k\varphi(W(s))| \cdot \mathbb{1}(t-\epsilon \leq ||W(s)||_\infty \leq t+\epsilon)]ds, \quad \text{(C.2.13)}
\end{aligned}
$$

where the first inequality holds due to (C.2.6), the second inequality is because $\sigma_{jk}^U = O(1), \sigma_{jk}^V = O(1)$ for all j, k and the constant M only depends on the maximal variances of the elements of U, V, the third inequality holds by the property (b) in Lemma C.2.1 for $\varphi_j(x), j = 1, 2$, and the last inequality holds by the definition of Δ_0. Note that $\varphi_1(x) := \varphi_{t,\epsilon}(x), \varphi_2(x) := \varphi_{t-\epsilon,\epsilon}(x)$ as defined in the proof of Theorem C.2.2. We use the same strategy to deal with $\varphi_1(x)$ and $\varphi_2(x)$. Below we give the derivations when $\varphi = \varphi_1(x)$ and it is not hard to check these derivations work for $\varphi_2(x)$ as well. Recall the explicit construction of $\varphi : \mathbb{R}^d \to \mathbb{R}$ introduced in Remark C.2.1.1,

$$
\varphi(x) = \varphi_{r,\epsilon}(x) = g_0\left(\frac{2(F_\beta(z_x - r\mathbf{1}_{2d}) - \epsilon/2)}{\epsilon}\right),
$$

where $\beta = 2\log(2d)/\epsilon$, $g_0(t) := 30\mathbb{1}(0 \leq t \leq 1)\int_t^1 s^2(1-s)^2 ds + \mathbb{1}(t \leq 0)$, F_β is the "softmax" function

$$
F_\beta(z) := \frac{1}{\beta}\log\Big(\sum_{m=1}^{2d}\exp(\beta z_m)\Big) \quad \text{for } z \in \mathbb{R}^{2d},
$$

$z_x = (x^\top, -x^\top)^\top$ and $\mathbf{1}_{2d}$ is the vector of 1's of dimension $2d$.

To bound (C.2.13), we consider the case where $j \neq k$ and $\sigma^U_{jk} \neq \sigma^V_{jk}$. Note that

$$|\partial_j \partial_k \varphi(W(s))| \leq ||g''||_\infty |\widetilde{\pi}_j(Z)\widetilde{\pi}_k(Z)| + \beta ||g'||_\infty |\widetilde{\pi}_j(Z)\widetilde{\pi}_k(Z)|, \tag{C.2.14}$$

where $g(t) := g_0(\frac{2(t-\epsilon/2)}{\epsilon})$, $Z := W(s)$ and

$$\widetilde{\pi}_j(z) := \frac{e^{\beta z_j} - e^{-\beta z_j}}{\sum_{m=1}^d e^{\beta z_m} + \sum_{m=1}^d e^{-\beta z_m}}.$$

The above result follows from a direct calculation. Due to the boundedness of $||g_0'||_\infty$, $||g_0''||_\infty$ and $\beta = 2\log(2d)/\epsilon$, we obtain the following bound on (C.2.14),

$$\begin{aligned}
|\partial_j \partial_k \varphi(W(s))| &\leq (||g''||_\infty + \beta ||g'||_\infty)|\widetilde{\pi}_j(Z)\widetilde{\pi}_k(Z)| \\
&\leq \Big(\frac{4}{\epsilon^2}||g_0''||_\infty + \frac{2\beta}{\epsilon}||g_0'||_\infty\Big)|\widetilde{\pi}_j(Z)\widetilde{\pi}_k(Z)| \\
&\leq \frac{C_1 \log(2d)}{\epsilon^2}|\widetilde{\pi}_j(Z)\widetilde{\pi}_k(Z)| \leq \frac{C_1 \log(2d)}{\epsilon^2}|\pi_j(Z)\pi_k(Z)|,
\end{aligned}$$

for some constant C_1, where $\pi_j(z) = e^{\beta|z_j|}/\sum_{m=1}^d e^{\beta|z_m|}$. Recalling $Z = W(s)$, we have

$$\begin{aligned}
&\int_0^1 \mathbb{E}\left[|\partial_j \partial_k \varphi(W(s))| \cdot \mathbb{1}(t-\epsilon \leq ||W(s)||_\infty \leq t+\epsilon)\right] ds \\
\leq\ & \frac{C_1 \log(2d)}{\epsilon^2} \int_0^1 \mathbb{E}\left[\pi_j(Z)\pi_k(Z) \cdot \mathbb{1}(t-\epsilon \leq ||Z||_\infty \leq t+\epsilon)\right] ds \\
=\ & \frac{C_1 \log(2d)}{\epsilon^2} \mathbb{P}(||V||_\infty > t) \int_0^1 \underbrace{\frac{\mathbb{E}\left[\pi_j(Z)\pi_k(Z) \cdot \mathbb{1}(t-\epsilon \leq ||Z||_\infty \leq t+\epsilon)\right]}{\mathbb{P}(||V||_\infty > t)}}_{\text{II(s)}} ds.
\end{aligned} \tag{C.2.15}$$

Below we focus on bounding the term II(s) for any $s \in [0,1]$. First we rewrite $\pi_j(Z)\pi_k(Z)$ and

simply derive the following inequality,

$$\begin{aligned}\pi_j(Z)\pi_k(Z) &= \frac{e^{\beta|Z_j|}}{\sum_{m=1}^d e^{\beta|Z_m|}} \cdot \frac{e^{\beta|Z_k|}}{\sum_{m=1}^d e^{\beta|Z_m|}} \\ &= \frac{e^{-\beta(||Z||_\infty - |Z_j|)} \cdot e^{-\beta(||Z||_\infty - |Z_k|)}}{(1 + \sum_{|Z_m| \neq ||Z||_\infty} e^{-\beta(||Z||_\infty - |Z_m|)})^2} \\ &\leq e^{-\beta(||Z||_\infty - |Z_j|)} \cdot e^{-\beta(||Z||_\infty - |Z_k|)}, \end{aligned} \tag{C.2.16}$$

where the second equality comes from dividing both the numerator and denominator by $e^{2\beta||Z||_\infty}$ in the first line. Note that $\mathbb{P}(|Z_j| = |Z_k|) = 0$ since the random vector Z follows a non-degenerate d-dimensional multivariate Gaussian distribution. Hence we have

$$1 = \mathbb{1}(|Z_j| = ||Z||_\infty, |Z_k| < ||Z||_\infty) + \mathbb{1}(|Z_j| < ||Z||_\infty), \text{ almost surely.} \tag{C.2.17}$$

Plugging the equality (C.2.17) into (C.2.16), we can further bound $\pi_j(Z)\pi_k(Z)$ as

$$\pi_j(Z)\pi_k(Z) \leq e^{-\beta(||Z||_\infty - |Z_k|)} \cdot \mathbb{1}(|Z_k| < ||Z||_\infty) + e^{-\beta(||Z||_\infty - |Z_j|)} \cdot \mathbb{1}(|Z_j| < ||Z||_\infty), \text{ almost surely.}$$

Then we can bound $\mathrm{II}(s)$ by

$$\begin{aligned}\mathrm{II}(s) &= \frac{\mathbb{E}\left[\pi_j(Z)\pi_k(Z) \cdot \mathbb{1}(t - \epsilon \leq ||Z||_\infty \leq t + \epsilon)\right]}{\mathbb{P}(||V||_\infty > t)} \\ &\leq \frac{\mathbb{E}[e^{-\beta(||Z||_\infty - |Z_k|)} \cdot \mathbb{1}(|Z_k| < ||Z||_\infty)\mathbb{1}(t - \epsilon \leq ||Z||_\infty \leq t + \epsilon)]}{\mathbb{P}(||V||_\infty > t)} \qquad \text{(C.2.18)} \\ &\quad + \frac{\mathbb{E}[e^{-\beta(||Z||_\infty - |Z_j|)} \cdot \mathbb{1}(|Z_j| < ||Z||_\infty)\mathbb{1}(t - \epsilon \leq ||Z||_\infty \leq t + \epsilon)]}{\mathbb{P}(||V||_\infty > t)}. \qquad \text{(C.2.19)}\end{aligned}$$

We use the same strategy to bound (C.2.18) and (C.2.19). Below we give the derivations for bounding (C.2.19) and note these also work for (C.2.18).

For any $j \neq k$ and $\sigma_{jk}^U \neq \sigma_{jk}^V$, Assumption C.2.3 says that there exists a subset $\mathcal{E}_0 \subset [d]$

satisfying $j \in \mathcal{E}_0$, $|\mathcal{E}_0| = \mathfrak{p}+1$, and $\sigma^U_{mm} = \sigma^U_{m'm'}, \sigma^U_{mm'} = \sigma^V_{mm'} = 0$ when $m, m' \in \mathcal{E}_0, m \neq m'$. This implies the following: when $s = 0$ or 1 (i.e., $Z = U$ or V), we can find a $\mathfrak{p}-$dimensional random vector G such that (Z_j, G) are all independent and $\mathrm{Var}(G_\ell) = \mathrm{Var}(Z_j) = \sigma_j^2$ for $\ell \in [\mathfrak{p}]$. Note that G is constructed as $(Z_m)_{m\in\mathcal{E}_0, m\neq j}$ with $\mathcal{E}_0$ being the same for $Z = U$ and V. Therefore, for any $s \in (0,1), Z = W(s) = \sqrt{s}U + \sqrt{1-s}V$, we can construct $G = (Z_m)_{m\in\mathcal{E}_0, m\neq j}$ such that (Z_j, G) are all independent and $\mathrm{Var}(G_\ell) = \mathrm{Var}(Z_j) = \sigma_j^2$ for $\ell \in [\mathfrak{p}]$. Throughout the following proof and the lemmas in Appendix C.2.3, we will use the notation Z, G without making the dependence on s explicitly. And we denote the indices of the random variables in G (among Z) by $\mathcal{E}_G$, i.e., $\mathcal{E}_G = \mathcal{E}_0 \setminus \{j\} = \{m \in [d] : Z_m = G_\ell \text{ for some } \ell \in [\mathfrak{p}]\}$.

We will consider two separate cases based on whether $||G||_\infty = ||Z||_\infty$ holds. Formally, we write $\mathbb{1}(|Z_j| < ||Z||_\infty) \leq \mathbb{1}(E_1) + \mathbb{1}(E_2)$ with E_1 and E_2 defined as

$$E_1 := \{||Z||_\infty > ||G||_\infty, ||Z||_\infty > |Z_j|\}, \tag{C.2.20}$$

$$E_2 := \{||G||_\infty = ||Z||_\infty > |Z_j|\}. \tag{C.2.21}$$

Then the numerator of the fraction in (C.2.19) can be bounded by the summation of the following two terms:

$$\begin{aligned} \mathrm{II}_1 &:= \mathbb{E}\left[e^{-\beta(||Z||_\infty - |Z_j|)} \cdot \mathbb{1}(E_1) \cdot \mathbb{1}(t-\epsilon \leq ||Z||_\infty \leq t+\epsilon)\right], \\ \mathrm{II}_2 &:= \mathbb{E}\left[e^{-\beta(||Z||_\infty - |Z_j|)} \cdot \mathbb{1}(E_2) \cdot \mathbb{1}(t-\epsilon \leq ||Z||_\infty \leq t+\epsilon)\right]. \end{aligned} \tag{C.2.22}$$

Combining (C.2.19) with (C.2.22) and applying Lemmas C.2.5 and C.2.6, we have

$$\mathrm{II}(s) \leq \frac{2(\mathrm{II}_1 + \mathrm{II}_2)}{\mathbb{P}(||V||_\infty > t)} \leq \frac{C'\epsilon \log d}{\beta \mathfrak{p}}, \quad \forall s \in [0,1], \tag{C.2.23}$$

for some constant C'. By (C.2.12), (C.2.13), (C.2.15) and (C.2.23), we thus obtain the following

inequality

$$\begin{aligned} & |\mathbb{P}(||U||_\infty > t) - \mathbb{P}(||V||_\infty > t)| \\ \leq \; & A(t,\epsilon)\mathbb{P}(||V||_\infty > t) + \frac{C_1 M \Delta_0 \log(2d)}{2\epsilon^2}\mathbb{P}(||V||_\infty > t) \cdot \frac{C'\epsilon \log d}{\beta \mathfrak{p}} \\ = \; & \mathbb{P}(||V||_\infty > t)\left(A(t,\epsilon) + B(\Delta_0, \mathfrak{p})\right), \end{aligned} \tag{C.2.24}$$

where $A(t,\epsilon) := M_1 \log d(t+1)\epsilon \exp(M_2 \log^2 d(t+1)\epsilon)$, $B(\Delta_0, \mathfrak{p}) := C''(\log d/\mathfrak{p})\Delta_0$ for some constants M_1, M_2, C''. In the last line, we also subsitute $\beta = \frac{2\log(2d)}{\epsilon}$. By re-arranging (C.2.24), we finally have

$$\left|\frac{\mathbb{P}(||U||_\infty > t)}{\mathbb{P}(||V||_\infty > t)} - 1\right| \leq A(t,\epsilon) + B(\Delta_0, \mathfrak{p}).$$

Since $0 \leq t \leq C_0\sqrt{\log d}$ and $\epsilon = c/\max\{(\log d)^{3/2}, \mathfrak{p}\log d\}$ for some small enough constant $c > 0$, we have $A(t,\epsilon) = O(B(\Delta_0, \mathfrak{p}))$. Then (C.2.11) can be established, i.e.,

$$\sup_{0\leq t\leq C_0\sqrt{\log d}} \left|\frac{\mathbb{P}(||U||_\infty > t)}{\mathbb{P}(||V||_\infty > t)} - 1\right| \leq C''' B(\Delta_0, \mathfrak{p}) = O\left(\frac{\Delta_0 \log d}{\mathfrak{p}}\right).$$

□

Now we prove Theorem 3.3.2 using similar strategies as in Theorem C.2.4. Recall that the connectivity assumption in Theorem 3.3.2 assumes that there exists a disjoint $\mathfrak{p}$-partition of nodes $\cup_{\ell=1}^{\mathfrak{p}} \mathcal{C}_\ell = [d]$ such that $\sigma^U_{jk} = \sigma^V_{jk} = 0$ when $j \in \mathcal{C}_\ell$ and $k \in \mathcal{C}_{\ell'}$ for some $\ell \neq \ell'$. Since this connectivity assumption is stronger than that in Theorem C.2.4, we are able to do slightly more careful analysis in Lemma C.2.5. As a result, the minimal eigenvalue condition is no longer needed. Also note that Theorem 3.3.2 assumes the unit variance condition and there exists some $\sigma_0 < 1$ such that $|\sigma^V_{jk}| \leq \sigma_0, |\sigma^U_{jk}| \leq \sigma_0$ for any $j \neq k$. Both the variance condition and the covariance

condition can be relaxed. In the following proof, we establish the Cramér-type comparison bound under a general variance condition. This general version is actually used in the proof of Theorem 3.5.2. Specifically, the general variance condition says that $a_0 \leq \sigma_{jj}^U = \sigma_{jj}^V \leq a_1,\ \forall j \in [d]$. After relaxing the unit variance assumption, some balanced variance assumption on the above components $\mathcal{C}_\ell$ is required. It says that given any $j \in \mathcal{C}_\ell$ with some ℓ, there exists at least one $m \in \mathcal{C}_{\ell'}$ such that $\sigma_{jj}^U = \sigma_{jj}^V = \sigma_{mm}^U = \sigma_{mm}^V$ for any $\ell' \neq \ell$. Remark this condition is mainly needed for Lemma C.2.12. We will call all these assumptions about variances as general variance condition. Denote $\widetilde{\sigma}_{jk}^U = \sigma_{jk}^U / \sqrt{\sigma_{jj}^U \sigma_{kk}^U}$. Accordingly, the covariance condition on σ_{jk} in Theorem 3.3.2 can also be relaxed into the following: there exists some $\sigma_0 < 1$ such that $|\widetilde{\sigma}_{jk}^V| = |\sigma_{jk}^V| / \sqrt{\sigma_{jj}^V \sigma_{kk}^V} \leq \sigma_0$ for any $j \neq k$ and $|\{(j,k) : j \neq k, |\widetilde{\sigma}_{jk}^U| = |\sigma_{jk}^U| / \sqrt{\sigma_{jj}^U \sigma_{kk}^U} > \sigma_0\}| \leq b_0$ for some constant b_0. We will call this condition as general covariance condition.

Proof of Theorem 3.3.2. Following exactly the same derivations in Theorem C.2.4 (up to (C.2.22)), we arrive at the following

$$\mathrm{II}(s) \leq \frac{2(\mathrm{II}_1 + \mathrm{II}_2)}{\mathbb{P}(||V||_\infty > t)},$$

where $\mathrm{II}(s), \mathrm{II}_1, \mathrm{II}_2$ are defined in (C.2.15) and (C.2.22), except that the random vector G can be constructed to satisfy more properties. Assuming the connectivity assumption of Theorem 3.3.2 and the general variance condition, we construct G by choosing one random variable Z_m from each component (except the one to which Z_j belongs) satisfying $\mathrm{Var}\,(Z_m) = \sigma_{mm}^U = \sigma_{mm}^V = \sigma_{jj}^U = \sigma_{jj}^V = \mathrm{Var}\,(Z_j)$. Such construction still satisfies the mentioned properties in Theorem C.2.4. Specifically, (G, Z_j) consists of $(\mathfrak{p}+1)$ i.i.d. Gaussian random variables. Moreover, for any $k \neq j, k \notin \mathcal{E}_G = \{m \in [d] : Z_m = G_\ell \text{ for some } \ell \in [\mathfrak{p}]\}$, there exists at most one $m \in \{j\} \cup \mathcal{E}_G$, such that Z_k and Z_m belong to the same component. Based on this property, we prove Lemma C.2.12 and Lemma C.2.13, which do not require minimal eigenvalue conditions compared with Lemma C.2.5 and Lemma C.2.7. We still apply Lemma C.2.13 to bound the term II_2. Regarding

the term II_1, we control it by using Lemma C.2.6. Therefore, we obtain the following

$$\mathrm{II}(s) \leq \frac{C'\epsilon \log d}{\beta \mathfrak{p}} \left(1 + \frac{b_0}{\sqrt{1-(s+(1-s)\sigma_0)^2}}\right). \tag{C.2.25}$$

Note a simple calculus result:

$$\int_0^1 \frac{b_0}{\sqrt{1-(s+(1-s)\sigma_0)^2}} \leq \frac{0.5\pi b_0}{1-\sigma_0} < C''$$

for some constant C'' when $\sigma_0 < 1$. Combining the above bound with (C.2.25), (C.2.12), (C.2.13), (C.2.15) and (C.2.23), we establish the bound (3.3.2) thus prove Theorem 3.3.2. □

C.2.3 Ancillary lemmas for Theorem C.2.4

Throughout the lemmas in this section, we will use Z and G without making the dependence on s explicitly, as mentioned in the proof of Theorem C.2.4.

Lemma C.2.5. *Suppose* $\lambda_{\min}(\mathbf{\Sigma}^U) \geq 1/b_0 > 0, \lambda_{\min}(\mathbf{\Sigma}^V) \geq 1/b_0 > 0$ *for some constant* $b_0 > 0$. *For the term* $\mathrm{II}_1 = \mathbb{E}\left[e^{-\beta(||Z||_\infty - |Z_j|)} \cdot \mathbb{1}(E_1) \cdot \mathbb{1}(t-\epsilon \leq ||Z||_\infty \leq t+\epsilon)\right]$ *with* E_1 *defined in* (C.2.20) *and* $\epsilon = c/\max\{(\log d)^{3/2}, \mathfrak{p}\log d\}$ *for some small enough constant* $c > 0$, *whenever* t *satisfies* $0 \leq t \leq C_0\sqrt{\log d}$ *for some constant* $C_0 > 0$, *we have*

$$\frac{\mathrm{II}_1}{\mathbb{P}(||V||_\infty > t)} \leq \frac{C'\epsilon \log d}{\beta \mathfrak{p}}. \tag{C.2.26}$$

Proof of Lemma C.2.5. We will bound II_1 by the law of total expectation. Specifically, we first calculate the conditional expectation given (G, Z_j) then take expectation with respect to (G, Z_j). Denoting the conditional density function of $||Z||_\infty \mid Z_j = z_j, G = g$ by $f_{g,z_j}(u)$, we write out the integral form of II_1 as

$$
\begin{aligned}
\mathrm{II}_1 =\ & \mathbb{E}\left[e^{-\beta(||Z||_\infty - |Z_j|)} \cdot \mathbb{1}(||Z||_\infty > ||G||_\infty, ||Z||_\infty > Z_j) \cdot \mathbb{1}(t-\epsilon \le ||Z||_\infty \le t+\epsilon)\right] \\
=\ & \mathbb{E}\left[e^{\beta|Z_j|} \cdot \mathbb{1}(||G||_\infty \le t+\epsilon, |Z_j| \le t+\epsilon)\left(\int_{t-\epsilon}^{t+\epsilon} f_{G,Z_j}(u) e^{-\beta u} \mathbb{1}(u > ||G||_\infty, u > |Z_j|) du\right)\right] \\
\le\ & \mathbb{E}\left[e^{\beta|Z_j|} \cdot \mathbb{1}(||G||_\infty \le t+\epsilon, |Z_j| \le t+\epsilon)\left(\int_{t-\epsilon}^{t+\epsilon} C\sqrt{\log d} \cdot e^{-\beta u} \mathbb{1}(u > |Z_j|) du\right)\right] \\
\le\ & C\sqrt{\log d}\,\mathbb{P}(||G||_\infty \le t+\epsilon)\mathbb{E}\left[\int_{|z_j|\le t+\epsilon} \phi\left(\frac{z_j}{\sigma_j}\right) e^{\beta|z_j|}\left(\int_{t-\epsilon}^{t+\epsilon} e^{-\beta u}\mathbb{1}(u > |z_j|) du\right) dz_j\right] \\
\le\ & C\sqrt{\log d}\,\mathbb{P}(||G||_\infty \le t+\epsilon) \underbrace{\int_{|z_j|\le t+\epsilon} \phi\left(\frac{z_j}{\sigma_j}\right) e^{\beta|z_j|}\left(\int_{t-\epsilon}^{t+\epsilon} e^{-\beta u}\mathbb{1}(u > |z_j|) du\right) dz_j}_{\mathrm{III}},
\end{aligned}
\tag{C.2.27}
$$

where the first inequality holds since $\mathbb{1}(u > ||G||_\infty, |Z_j|) \le \mathbb{1}(u > |Z_j|)$ and the conditional density function $f_{g,z_j}(u)$ is bounded by $C\sqrt{\log d}$ when $||g||_\infty, |z_j| < u \le t+\epsilon$ and $0 \le t \le C_0\sqrt{\log d}$, as a result of Lemma C.2.7. Recall that $\phi(\cdot)$ denotes the standard Gaussian PDF. We use the fact that $Z_j \perp\!\!\!\perp G, Z_j \sim \mathcal{N}(0, \sigma_j^2)$ and write out the integral form of the expectation with

respect to Z_j, thus the second inequality follows. Then the integral III can be further rewritten as

$$\begin{aligned}
\text{III} &= 2\int_{t-\epsilon}^{t+\epsilon} e^{-\beta u}\left(\int_0^u \phi\left(\frac{x}{\sigma_j}\right)e^{\beta x}dx\right)du \\
&= 2\int_{t-\epsilon}^{t+\epsilon} e^{-\beta u}\left(e^{\frac{\beta^2\sigma_j^2}{2}}\int_0^u \phi\left(\frac{x}{\sigma_j}-\beta\sigma_j\right)dx\right)du \\
&= 2\int_{t-\epsilon}^{t+\epsilon} e^{-\beta u}\left(e^{\frac{\beta^2\sigma_j^2}{2}}\int_{-\beta\sigma_j}^{u/\sigma_j-\beta\sigma_j} \phi\left(x\right)dx\right)du \\
&\leq 2\int_{t-\epsilon}^{t+\epsilon} e^{-\beta u}\left(e^{\frac{\beta^2\sigma_j^2}{2}}\bar{\Phi}\big(\beta\sigma_j-u/\sigma_j\big)\right)du \\
&\leq 2\int_{t-\epsilon}^{t+\epsilon} e^{-\beta u}\left(e^{\frac{\beta^2\sigma_j^2}{2}}\frac{e^{-\frac{(\beta\sigma_j-u/\sigma_j)^2}{2}}}{\beta\sigma_j-u/\sigma_j}\right)du \\
&\leq \frac{4}{\beta\sigma_j}\int_{t-\epsilon}^{t+\epsilon} e^{-\beta u}\left(e^{\beta u}e^{-\frac{u}{2\sigma_j}}\right)du \leq \frac{8\epsilon}{\beta\sigma_j}\exp\left(-\frac{(t-\epsilon)^2}{2\sigma_j^2}\right), \qquad \text{(C.2.28)}
\end{aligned}$$

where the first equality holds by Fubini's theorem, and the second equality holds by the definition of $\phi(\cdot)$. Regarding the first inequality, we use the fact that $u/\sigma_j-\beta\sigma_j < 2u/\sigma_j-\beta\sigma_j < 0$ for $u \leq t+\epsilon$ and $t \leq C_0\sqrt{\log d}$. This is because $\beta = \frac{2\log(2d)}{\epsilon}$ and $\epsilon = c/\max\{(\log d)^{3/2}, \mathfrak{p}\log d\}$ for some small enough constant $c > 0$. Then $\int_{-\beta\sigma_j}^{u/\sigma_j-\beta\sigma_j}\phi\left(x\right)dx \leq \bar{\Phi}(\beta\sigma_j-u/\sigma_j)$, recalling $\bar{\Phi} = 1-\Phi$, where Φ is the standard Gaussian CDF. The second inequality holds as a result of Lemma C.2.8. The third inequality holds due to $\beta\sigma_j > 2u/\sigma_j$ for $u \leq t+\epsilon$.

By (C.2.27) and (C.2.28), we arrive at the following bound

$$\begin{aligned}
\frac{\text{II}_1}{\mathbb{P}(||V||_\infty > t)} &\leq C\sqrt{\log d}\cdot\frac{\mathbb{P}(||G||_\infty \leq t+\epsilon)}{\mathbb{P}(||V||_\infty > t)}\cdot\frac{8\epsilon}{\beta\sigma_j}\exp\left(-\frac{(t-\epsilon)^2}{2\sigma_j^2}\right) \\
&\leq C\sqrt{\log d}\cdot\frac{C_1\epsilon}{\beta}\cdot\frac{\mathbb{P}(||G||_\infty \leq t+\epsilon)}{\mathbb{P}(||G||_\infty > t)}\cdot\phi\big(\frac{t-\epsilon}{\sigma_j}\big)/\sigma_j \\
&= C\sqrt{\log d}\cdot\frac{C_1\epsilon}{\beta}\cdot\underbrace{\frac{(1-2\bar{\Phi}(\frac{t+\epsilon}{\sigma_j}))^{\mathfrak{p}}}{1-(1-2\bar{\Phi}\big(\frac{t}{\sigma_j}\big))^{\mathfrak{p}}}\cdot\phi\big(\frac{t-\epsilon}{\sigma_j}\big)/\sigma_j}_{\Lambda(t,\epsilon,\mathfrak{p})},
\end{aligned}$$

for some constants C, C_1, where the second inequality holds due to the definition of $\phi(z)$ and $\mathbb{P}(||V||_\infty > t) \geq \mathbb{P}(||G||_\infty > t)$. This is because

$$\mathbb{P}(||V||_\infty > t) \geq \mathbb{P}(\max_{k \in \mathcal{E}_G} |V_k| > t) = \mathbb{P}(||G_V||_\infty > t) = \mathbb{P}(||G||_\infty > t), \tag{C.2.29}$$

where $G_V = (Z_m)_{m \in \mathcal{E}_0, m \neq j}$ with $Z = V$ has the same distribution as G. Regarding the last line, by the construction of $G = (G_\ell)_{\ell \in [\mathfrak{p}]} = (Z_m)_{m \in \mathcal{E}_0, m \neq j}$ in the proof of Theorem C.2.4, we have $\{G_\ell\}_{\ell \in [\mathfrak{p}]}$ are $\mathfrak{p}$ i.i.d. Gaussian random variables with $\mathrm{Var}\,(G_\ell) = \mathrm{Var}\,(Z_j) = \sigma_j^2$. By applying Lemma C.2.9 to the term $\Lambda(t, \epsilon, \mathfrak{p})$ in the last line, we further obtain,

$$\frac{\mathrm{II}_1}{\mathbb{P}(||V||_\infty > t)} \quad \leq \quad C' \sqrt{\log d} \cdot \frac{\epsilon}{\beta} \frac{\sqrt{\log d}}{\mathfrak{p}} = \frac{C' \epsilon \log d}{\beta \mathfrak{p}},$$

for some constant C', therefore (C.2.26) is established. □

Lemma C.2.6. *For the term* $\mathrm{II}_2 = \mathbb{E}\left[e^{-\beta(||Z||_\infty - |Z_j|)} \cdot \mathbb{1}(E_2) \cdot \mathbb{1}(t - \epsilon \leq ||Z||_\infty \leq t + \epsilon)\right]$ *with* E_2 *defined in* (C.2.21) *and* $\epsilon = c/\max\{(\log d)^{3/2}, \mathfrak{p} \log d\}$ *for some small enough constant* $c > 0$*, whenever* t *satisfies* $0 \leq t \leq C_0 \sqrt{\log d}$ *for some constant* $C_0 > 0$*, we have*

$$\frac{\mathrm{II}_2}{\mathbb{P}(||V||_\infty > t)} \leq \frac{C'' \epsilon \sqrt{\log d}}{\beta \mathfrak{p}}. \tag{C.2.30}$$

Proof of Lemma C.2.6. By the definition of E_2 in (C.2.21) and the tower property, we have

$$\begin{aligned}
\mathrm{II}_2 &= \mathbb{E}\left[e^{-\beta(||Z||_\infty - |Z_j|)} \cdot \mathbb{1}(||G||_\infty = ||Z||_\infty > |Z_j|) \cdot \mathbb{1}(t - \epsilon \leq ||Z||_\infty \leq t + \epsilon)\right] \\
&= \mathbb{E}\left[e^{-\beta(||G||_\infty - |Z_j|)} \cdot \mathbb{1}(||G||_\infty = ||Z||_\infty, ||G||_\infty > |Z_j|) \cdot \mathbb{1}(t - \epsilon \leq ||G||_\infty \leq t + \epsilon)\right] \\
&\leq \mathbb{E}\left[e^{-\beta(||G||_\infty - |Z_j|)} \cdot \mathbb{1}(||G||_\infty > |Z_j|) \cdot \mathbb{1}(t - \epsilon \leq ||G||_\infty \leq t + \epsilon)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[e^{\beta |Z_j|} \mathbb{1}(|Z_j| < ||G||_\infty) \,|\, G\right] e^{-\beta ||G||_\infty} \cdot \mathbb{1}(t - \epsilon \leq ||G||_\infty \leq t + \epsilon)\right]. \quad \text{(C.2.31)}
\end{aligned}$$

First we bound $\mathrm{III}(g) := \mathbb{E}\left[e^{\beta|Z_j|}\mathbb{1}(|Z_j| < ||G||_\infty) \,|\, G = g\right]$ when $||g||_\infty \in [t-\epsilon, t+\epsilon]$. Specifically,

$$\begin{aligned}
\mathrm{III}(g) = \frac{2}{\sigma_j}\int_0^{||g||_\infty} e^{\beta x}\phi\Big(\frac{x}{\sigma_j}\Big)dx &= \frac{2e^{\beta^2\sigma_j^2/2}}{\sigma_j}\int_0^{||g||_\infty}\phi\Big(\frac{x-\beta\sigma_j^2}{\sigma_j}\Big)dx \\
&\le 2e^{\beta^2\sigma_j^2/2}\int_{-\infty}^{||g||_\infty/\sigma_j-\beta\sigma_j}\phi(y)dy \\
&= 2e^{\beta^2\sigma_j^2/2}\,\bar{\Phi}(\beta\sigma_j - ||g||_\infty/\sigma_j) \\
&\le 2e^{\beta^2\sigma_j^2/2}\frac{\phi\left(\beta\sigma_j - ||g||_\infty/\sigma_j\right)}{\beta\sigma_j - ||g||_\infty/\sigma_j} \\
&\le \frac{4}{\beta\sigma_j}\phi\left(\frac{||g||_\infty}{\sigma_j}\right)e^{\beta||g||_\infty}, \qquad \text{(C.2.32)}
\end{aligned}$$

where the first equality holds due to $Z_j \perp\!\!\!\perp G$, and the second equality comes from rearranging. The first inequality holds by the change of variable $y = (x-\beta\sigma_j^2)/\sigma_j$ and setting the lower limit of the integral as $-\infty$. Because $\beta = \frac{2\log(2d)}{\epsilon}$ and $\epsilon = c/\max\{(\log d)^{3/2}, \mathfrak{p}\log d\}$ for some small enough constant $c > 0$, we have $||g||_\infty/\sigma_j < \beta\sigma_j$ for $||g||_\infty \le t+\epsilon$ and $t \le C_0\sqrt{\log d}$. Then the second inequality holds as a result of Lemma C.2.8 and the fact that $\beta\sigma_j - ||g||_\infty/\sigma_j > 0$. The last inequality comes from rearranging and the fact that $\beta\sigma_j > 2||g||_\infty/\sigma_j$ for $||g||_\infty \le t+\epsilon$ and $t \le C_0\sqrt{\log d}$. Combining (C.2.32) with (C.2.31), we have

$$\begin{aligned}
\mathrm{II}_2 &\le \mathbb{E}\left[\mathrm{III}(G)\cdot e^{-\beta||G||_\infty}\cdot\mathbb{1}(t-\epsilon \le ||G||_\infty \le t+\epsilon)\right] \\
&\le \frac{4}{\beta\sigma_j}\mathbb{E}\left[\phi\Big(\frac{||G||_\infty}{\sigma_j}\Big)e^{\beta(||G||_\infty)}\cdot e^{-\beta||G||_\infty}\cdot\mathbb{1}(t-\epsilon \le ||G||_\infty \le t+\epsilon)\right] \\
&\le \frac{4}{\beta\sigma_j}\int_{t-\epsilon}^{t+\epsilon}\phi(\frac{y}{\sigma_j})f(y)dy, \qquad \text{(C.2.33)}
\end{aligned}$$

where $f(y)$ denotes the PDF of $||G||_\infty$. As $\{G_\ell\}_{\ell\in[\mathfrak{p}]}$ are i.i.d. Gaussian random variables satisfying

$\forall\, \ell \in [\mathfrak{p}], \mathbb{E}\left[G_\ell\right] = 0$ and $\mathrm{Var}\left(G_\ell\right) = \sigma_j^2$, we have for $y > 0$,

$$\mathbb{P}(||G||_\infty \le y) = \mathbb{P}(\bigcup_{\ell\in[\mathfrak{p}]} |G_\ell| \le y) = (1 - 2\mathbb{P}(G_\ell/\sigma_j > y/\sigma_j))^{\mathfrak{p}} = (1 - 2\bar{\Phi}(y/\sigma_j))^{\mathfrak{p}}. \tag{C.2.34}$$

Thus we have the PDF of $||G||_\infty$ equals $f(y) = \frac{2\mathfrak{p}}{\sigma_j}\left(1 - 2\bar{\Phi}(\frac{y}{\sigma_j}))^{\mathfrak{p}}\right)^{\frac{\mathfrak{p}-1}{\mathfrak{p}}} \phi(\frac{y}{\sigma_j})$. Plugging the expression of $f(y)$ into (C.2.33), we further derive the following bound

$$\begin{aligned}
\frac{\mathrm{II}_2}{\mathbb{P}(||V||_\infty > t)} &\le \frac{8\mathfrak{p}}{\beta\sigma_j^2}\int_{t-\epsilon}^{t+\epsilon} \frac{\left(1 - 2\bar{\Phi}(\frac{y}{\sigma_j}))^{\mathfrak{p}}\right)^{\frac{\mathfrak{p}-1}{\mathfrak{p}}} \phi^2(\frac{y}{\sigma_j})}{\mathbb{P}(||V||_\infty > t)} dy \\
&\le \frac{8\mathfrak{p}}{\beta\sigma_j^2}\int_{t-\epsilon}^{t+\epsilon} \frac{\left(1 - 2\bar{\Phi}(\frac{y}{\sigma_j}))^{\mathfrak{p}}\right)^{\frac{\mathfrak{p}-1}{\mathfrak{p}}} \phi^2(\frac{y}{\sigma_j})}{1 - \mathbb{P}(||G||_\infty \le t)} dy \\
&= \frac{8\mathfrak{p}}{\beta\sigma_j^2}\int_{t-\epsilon}^{t+\epsilon} \frac{\left((1 - 2\bar{\Phi}(\frac{y}{\sigma_j}))^{\mathfrak{p}}\right)^{\frac{\mathfrak{p}-1}{\mathfrak{p}}} \phi^2(\frac{y}{\sigma_j})}{1 - (1 - 2\bar{\Phi}(\frac{t}{\sigma_j}))^{\mathfrak{p}}} dy \\
&\le \frac{16\epsilon}{\beta\sigma_j^2\mathfrak{p}} \frac{\left((1 - 2\bar{\Phi}(\frac{t+\epsilon}{\sigma_j}))^{\mathfrak{p}}\right)^{\frac{\mathfrak{p}-1}{\mathfrak{p}}} (\mathfrak{p}\phi(\frac{t-\epsilon}{\sigma_j}))^2}{1 - (1 - 2\bar{\Phi}(\frac{t+\epsilon}{\sigma_j}))^{\mathfrak{p}}} \le \frac{C''\epsilon\sqrt{\log d}}{\beta\mathfrak{p}},
\end{aligned}$$

for some constant C', where the second inequality holds due to (C.2.29), as mentioned in the proof of Lemma C.2.5. The equality holds as a result of substituting the expression of $\mathbb{P}(||G||_\infty \le t)$ by (C.2.34). The third inequality holds since $1 - 2\bar{\Phi}(z)$ is monotonically increasing and $\phi(z)$ is monotonically decreasing when $z \ge 0$. As for the last line, we apply Lemma C.2.10. Finally, (C.2.30) is established. □

Lemma C.2.7. *Suppose* $\lambda_{\min}(\mathbf{\Sigma}^U) \ge 1/b_0 > 0, \lambda_{\min}(\mathbf{\Sigma}^V) \ge 1/b_0 > 0$ *for some constant* $b_0 > 0$. *Recall that the density function of the conditional distribution of* $||Z||_\infty \mid \{Z_j = z_j, G = g\}$ *is denoted by* $f_{g,z_j}(z)$. *Suppose* $\epsilon > 0$, *when* $0 \le t \le C_0\sqrt{\log d}$ *for some constant* $C_0 > 0$ *and*

$|z_j|, ||g||_\infty \le t + \epsilon$, *we have*

$$f_{g,z_j}(z) \le C\sqrt{\log d}, \quad \forall\, z \in (\max\{|z_j|, ||g||_\infty\}, t+\epsilon].$$

Proof of Lemma C.2.7. First we introduce some new notations. Let $(\sigma_{jk})_{1\le j,k\le d} \in \mathbb{R}^{d\times d}$ be the covariance matrix of Z. For given j, we denote

$$\sigma_{kk\cdot j} := \sigma_{kk} - \sigma_{kj}^2\sigma_{jj}^{-1} - \sum_{m\in\mathcal{E}_G} \sigma_{km}^2\sigma_{mm}^{-1}. \tag{C.2.35}$$

As $z \in (\max\{|z_j|, ||g||_\infty\}, t+\epsilon]$, we can choose δ such that $0 < \delta < z - \max\{|z_j|, ||g||_\infty\}$. Throughout the following proof, we will work with such δ. Since $\max\{|z_j|, ||g||_\infty\} - z < -\delta$, we have

$$\mathbb{P}\left(\big|||Z||_\infty - z\big| \le \delta \,|\, Z_j = z_j, G = g\right) = \mathbb{P}\left(\big|||X||_\infty - z\big| \le \delta \,|\, Z_j = z_j, G = g\right), \tag{C.2.36}$$

where X denotes the $(d - \mathfrak{p} - 1)$-dimensional random vector by excluding Z_j, G from Z and therefore $||Z||_\infty = \max\{||X||_\infty, |Z_j|, ||G||_\infty\}$.

Recalling $G = (G_\ell)_{\ell\in[\mathfrak{p}]} = (Z_m)_{m\in\mathcal{E}_G}$, where $\mathcal{E}_G$ denotes the indices of the random variables in G (among Z), i.e., $\mathcal{E}_G = \{m \in [d] : Z_m = G_\ell \text{ for some } \ell \in [\mathfrak{p}]\}$, we have

$$||X||_\infty = \max_{k\in[d],\, k\notin\{j,\mathcal{E}_G\}} \{\max\{Z_k, -Z_k\}\}.$$

Given j and the choice of G, we also denote

$$\underline{\sigma_{\cdot j}} := \min_{k\in[d], k\notin\{j,\mathcal{E}_G\}} \sqrt{\sigma_{kk\cdot j}}, \quad \bar{\rho}_j := \max_{k\in[d], k\notin\{j,\mathcal{E}_G\}} \frac{|\sigma_{jk}|}{\sigma_{jj}}. \tag{C.2.37}$$

For each $k \in [d], k \notin \{j, \mathcal{E}_G\}$, the conditional expectation $\mathbb{E}[Z_k \mid Z_j, G]$ has the following expression,

$$\mathbb{E}[Z_k \mid Z_j, G] = \sigma_{kj}\sigma_{jj}^{-1} Z_j + \sum_{m \in \mathcal{E}_G} (\sigma_{km}\sigma_{mm}^{-1} Z_m), \tag{C.2.38}$$

since (Z_j, G) are all independent. Note that the requirement (c) in Assumption C.2.3 says $\forall\, k \in [d], |\{m \in \mathcal{E}_0 : |\sigma_{km}^U| + |\sigma_{km}^V| \neq 0\}| \leq c_0$, we thus have

$$\begin{aligned} \sum_{m \in \mathcal{E}_G} \mathbb{1}(\sigma_{km} \neq 0) &= \sum_{m \in \mathcal{E}_G} \mathbb{1}((s\sigma_{km}^U + (1-s)\sigma_{km}^V) \neq 0) \\ &\leq \sum_{m \in \mathcal{E}_G} \mathbb{1}(\sigma_{km}^U \neq 0 \text{ or } \sigma_{km}^V \neq 0) \leq c_0, \end{aligned} \tag{C.2.39}$$

where the first equality holds by the definition of σ_{km} and $Z = W(s) = \sqrt{s}U + \sqrt{1-s}V$. Combining (C.2.39) with (C.2.38), it yields the following bound on $|\mathbb{E}[Z_k \mid Z_j = z_j, G = g]|$,

$$\begin{aligned} |\mathbb{E}[Z_k \mid Z_j = z_j, G = g]| &= \Big|\sigma_{kj}\sigma_{jj}^{-1} z_j + \sum_{m \in \mathcal{E}_G} (\sigma_{km}\sigma_{mm}^{-1} z_m)\Big| \\ &\leq \bar{\rho}_j(|z_j| + c_0||g||_\infty), \end{aligned} \tag{C.2.40}$$

where $\bar{\rho}_j = \max_{k \in \mathcal{E}_X} \frac{|\sigma_{jk}|}{\sigma_{jj}}$ as defined. Denoting $\mathcal{E}_X := \{k : k \in [d],\ k \notin \{j, \mathcal{E}_G\}\}$, we define the following random variables,

$$\widetilde{W}_{2k-1} = \frac{Z_k - z}{\sqrt{\sigma_{kk\cdot j}}} + \frac{\widetilde{z}}{\underline{\sigma_{\cdot j}}}, \quad \widetilde{W}_{2k} = \frac{-Z_k - z}{\sqrt{\sigma_{kk\cdot j}}} + \frac{\widetilde{z}}{\underline{\sigma_{\cdot j}}}, \quad k \in \mathcal{E}_X, \tag{C.2.41}$$

where $\widetilde{z} = z + \bar{\rho}_j(|z_j| + c_0||g||_\infty)$. Then by the definitions of $\sigma_{kk\cdot j}$, $\underline{\sigma_{\cdot j}}$ and $\bar{\rho}_j$ in (C.2.35) and

(C.2.37), we have the above random variables satisfy the following properties,

$$\mathbb{E}\left[\widetilde{W}_m \,|\, Z_j = z_j, G = g\right] \geq 0, \quad \text{Var}\left(\widetilde{W}_m \,|\, Z_j = z_j, G = g\right) = 1,$$

where $m = 2k - 1$ or $2k$ and $k \in \mathcal{E}_X$. Denote those random variables defined in (C.2.41) by $\{\widetilde{W}_m\}$ for notation simplicity. We let $q_{z_j,g}(w)$ be the PDF of the conditional distribution of $\max_m\{\widetilde{W}_m\} \mid Z_j = z_j, G = g$. Then we will apply the derivation of Step 2 in Theorem 3 of Chernozhukov et al. (2015) to bound $q_{z_j,g}(w)$. Note that for the following derivations, we always conditional on the event $Z_j = z_j, G = g$. First, we verify the condition on $\{\widetilde{W}_m\}$. Since $|\text{Corr}(U_j, U_k)| \neq 1, |\text{Corr}(V_j, V_k)| \neq 1$ for distinct $j, k \in [d]$, we then have the correlation between $\widetilde{W}_{m_1}$ and $\widetilde{W}_{m_2}$ for $m_1 \neq m_2$ is less than 1. Therefore, by applying the derivation of Step 2 in Theorem 3 of Chernozhukov et al. (2015) to $\{\widetilde{W}_m\}$, we have

$$q_{z_j,g}(w) \leq h(w) := 2(w \vee 1) \exp\left\{-\frac{(w - \bar{w} - a_d)_+^2}{2}\right\},$$

where $\bar{w} = \max_m \mathbb{E}\left[\widetilde{W}_m \,|\, Z_j = z_j, G = g\right]$ and

$$a_d = \max_m \mathbb{E}\left[\left(\widetilde{W}_m - \mathbb{E}\left[\widetilde{W}_m \,|\, Z_j = z_j, G = g\right]\right) | \, Z_j = z_j, G = g\right].$$

When $w \leq \bar{w} + a_d$, we have $h(w) \leq 2(\bar{w} + a_d)$. To deal with the case where $w > \bar{w} + a_d$, we consider

$$\begin{aligned}
\log(h(w)) &= \log(2w) - \frac{(w - \bar{w} - a_d)^2}{2}, \\
\frac{d\log(h(w))}{dw} &= \frac{1}{w} - (w - \bar{w} - a_d), \\
\frac{d^2\log(h(w))}{dw^2} &= -\frac{1}{w^2} - 1 < 0.
\end{aligned}$$

Solving $\frac{d}{dw}\log(h(w)) = 0$ yields $w^\star = \frac{\bar{w}+a_d+\sqrt{(\bar{w}+a_d)^2+4}}{2}$. Therefore, the PDF of the conditional distribution of $\max_m\{\widetilde{W}_m\} \mid Z_j = z_j, G = g$ can be bounded by

$$h(w) \le h(w^\star) \le 3(\bar{w} + a_d). \tag{C.2.42}$$

Now we have

$$\begin{aligned}
& \mathbb{P}\left(\left|||Z||_\infty - z\right| \le \delta \mid Z_j = z_j, G = g\right) \\
= \;& \mathbb{P}\left(|||X||_\infty - z| \le \delta \mid Z_j = z_j, G = g\right) \\
= \;& \mathbb{P}\left(\left|\max_{k\in\mathcal{E}_X}\{Z_k, -Z_k\} - z\right| \le \delta \mid Z_j = z_j, G = g\right) \\
\le \;& \mathbb{P}\left(\left|\max_{k\in\mathcal{E}_X}\left\{\frac{Z_k - z}{\sqrt{\sigma_{kk\cdot j}}}, \frac{-Z_k - z}{\sqrt{\sigma_{kk\cdot j}}}\right\}\right| \le \frac{\delta}{\underline{\sigma_{\cdot j}}} \mid Z_j = z_j, G = g\right) \\
\le \;& \sup_{y\in\mathbb{R}} \mathbb{P}\left(\left|\max_{k\in\mathcal{E}_X}\left\{\frac{Z_k - z}{\sqrt{\sigma_{kk\cdot j}}} + \frac{\widetilde{z}}{\underline{\sigma_{\cdot j}}}, \frac{-Z_k - z}{\sqrt{\sigma_{kk\cdot j}}} + \frac{\widetilde{z}}{\underline{\sigma_{\cdot j}}}\right\} - y\right| \le \frac{\delta}{\underline{\sigma_{\cdot j}}} \mid Z_j = z_j, G = g\right) \\
= \;& \sup_{y\in\mathbb{R}} \mathbb{P}\left(\left|\max_m\{\widetilde{W}_m\} - y\right| \le \frac{\delta}{\underline{\sigma_{\cdot j}}} \mid Z_j = z_j, G = g\right) \le \frac{6\delta}{\underline{\sigma_{\cdot j}}}(\bar{w} + a_d),
\end{aligned} \tag{C.2.43}$$

where the first equality holds by (C.2.36), the second equality holds by the definition of X and $\mathcal{E}_X$, the first inequality holds since $\underline{\sigma_{\cdot j}} = \min_{k\neq j}\sqrt{\sigma_{kk\cdot j}}$, the third equality holds by the definition of $\{\widetilde{W}_m\}$ in (C.2.41), and the last inequality holds by the bound on $h(w)$ in (C.2.42). Regarding the

quantity $\bar{w} = \max_m \mathbb{E}\left[\widetilde{W}_m \mid Z_j = z_j, G = g\right]$, we have

$$\begin{aligned}
\bar{w} &= \max_{k\in\mathcal{E}_X}\left\{\frac{\pm\mathbb{E}\left[Z_k \mid Z_j = z_j, G = g\right] - z}{\sqrt{\sigma_{kk\cdot j}}} + \frac{\widetilde{z}}{\underline{\sigma_{\cdot j}}}\right\} \\
&\leq \max_{k\in\mathcal{E}_X}\left\{\frac{\pm\mathbb{E}\left[Z_k \mid Z_j = z_j, G = g\right]}{\sqrt{\sigma_{kk\cdot j}}}\right\} + \max_{k\in\mathcal{E}_X}\left\{\frac{1}{\underline{\sigma_{\cdot j}}} - \frac{1}{\sqrt{\sigma_{kk\cdot j}}}\right\} z + \frac{\bar{\rho}_j(|z_j| + c_0||g||_\infty)}{\underline{\sigma_{\cdot j}}} \\
&\leq \max_{k\in\mathcal{E}_X}\left\{\frac{\pm\left(\sigma_{kj}\sigma_{jj}^{-1}z_j + \sum_{m\in\mathcal{E}_G}(\sigma_{km}\sigma_{mm}^{-1}z_m)\right)}{\sqrt{\sigma_{kk\cdot j}}}\right\} + \frac{z}{\underline{\sigma_{\cdot j}}} + \frac{\bar{\rho}_j(|z_j| + c_0||g||_\infty)}{\underline{\sigma_{\cdot j}}} \\
&\leq \frac{2\bar{\rho}_j(|z_j| + c_0||g||_\infty)}{\underline{\sigma_{\cdot j}}} + \frac{z}{\underline{\sigma_{\cdot j}}} \leq \frac{2\bar{\rho}_j(1 + c_0) + 1}{\underline{\sigma_{\cdot j}}}(t + \epsilon), \qquad \text{(C.2.44)}
\end{aligned}$$

where $\max\{\pm A\} := \max\{A, -A\}$, the first inequality holds by the definition of $\widetilde{z}$, the second inequality holds by (C.2.38), and the last inequality holds by the definitions of $\bar{\rho}_j$ and $\underline{\sigma_{\cdot j}}$ and the fact $\sum_{m\in\mathcal{E}_G}\mathbb{1}(\sigma_{km} \neq 0) \leq c_0$.

Let δ in (C.2.43) go to 0, we get the following bound on the density function of the conditional distribution of $||Z||_\infty \mid \{Z_j = z_j, G = g\}$, i.e., when $0 \leq t \leq C_0\sqrt{\log d}$ and $|z_j|, ||g||_\infty \leq t + \epsilon$,

$$f_{g,z_j}(z) \leq \frac{6}{\underline{\sigma_{\cdot j}}}(\bar{w} + a_d) \leq \frac{6}{\underline{\sigma_{\cdot j}}}\left(\frac{2\bar{\rho}_j(1 + c_0) + 1}{\underline{\sigma_{\cdot j}}}C_1\sqrt{\log d} + C_2\sqrt{\log d}\right), \qquad \text{(C.2.45)}$$

for any $z \in (\max\{|z_j|, ||g||_\infty\}, t + \epsilon]$. The first inequality holds by (C.2.43). Regarding the second inequality, we apply the result in (C.2.44) and bound $(t + \epsilon)$ and a_d by $C_1\sqrt{\log d}$ for some constant C_1. Note $a_d \leq C_1\sqrt{\log d}$ is because of the maximal inequalities for sub-Gaussian random variables (Lemma 5.2 in van Handel (2014)). As for $\bar{\rho}_j = \max_{k\in\mathcal{E}_X}\frac{|\sigma_{jk}|}{\sigma_{jj}}$, we have

$$\bar{\rho}_j^2 \leq \max_{k\neq j}\frac{\sigma_{kk}}{\sigma_{jj}} \leq \frac{\max_j \sigma_{jj}^U}{\min_j \sigma_{jj}^U} \leq \frac{\max_j \sigma_{jj}^U}{\lambda_{\min}(\mathbf{\Sigma}^U)} = O(1),$$

where the first inequality holds by the Cauchy-Schwarz inequality, the second inequality holds

by the definition of Z and $\sigma_{jj}^U = \sigma_{jj}^V$, the third inequality holds by the fact that $\min_j \sigma_{jj}^U \geq \lambda_{\min}(\mathbf{\Sigma}^U)$, and the last step holds under the stated assumption of Theorem C.2.4. As for $\underline{\sigma_{\cdot j}} = \min_{k \in \mathcal{E}_X} \sqrt{\sigma_{kk \cdot j}}$ where $\sigma_{kk \cdot j} = \sigma_{kk} - \sigma_{kj}^2 \sigma_{jj}^{-1} - \sum_{m \in \mathcal{E}_G} \sigma_{km}^2 \sigma_{mm}^{-1} = \mathrm{Var}\left(Z_k \mid Z_j, G\right)$, we have

$$
\begin{aligned}
\frac{1}{\underline{\sigma_{\cdot j}^2}} &= \frac{1}{\min_{k \in \mathcal{E}_X} \mathrm{Var}\left(Z_k \mid Z_j, G\right)} \\
&\leq \frac{1}{\min_k \mathrm{Var}\left(Z_k \mid Z_{\cdot k}\right)} \\
&= \max_k ((\mathbf{\Sigma}^Z)^{-1})_{kk} \\
&\leq \lambda_{\max}((\mathbf{\Sigma}^Z)^{-1}) \\
&= 1/\lambda_{\min}(\mathbf{\Sigma}^Z) \\
&\leq (\min\{\lambda_{\min}(\mathbf{\Sigma}^U), \lambda_{\min}(\mathbf{\Sigma}^V\})^{-1} \leq b_0,
\end{aligned}
$$

under the stated assumption that $\lambda_{\min}(\mathbf{\Sigma}^U) \geq 1/b_0$, $\lambda_{\min}(\mathbf{\Sigma}^V) \geq 1/b_0$, where the first inequality holds since (Z_j, G) is a sub-vector of $Z_{\cdot k} := Z_{(1:d)\setminus k}$, the second equality holds by the relationship between the partial variances and the inverse covariance matrix, and the last three hold by the definitions of $\lambda_{\min}(\cdot)$, $\lambda_{\max}(\cdot)$. Thus we have $f_{g,z_j}(z) \leq C\sqrt{\log d}$ for some constant C, i.e., Lemma C.2.7 is proved.

□

Lemma C.2.8. *For $z > 0$, we have*

$$
\frac{\phi(z)}{2(z \vee 1)} \leq \bar{\Phi}(z) = 1 - \Phi(z) \leq \frac{\phi(z)}{z},
$$

where $\phi(z), \Phi(z)$ is the PDF and CDF of the standard Gaussian distribution respectively.

Proof of Lemma C.2.8. This is a simple fact derived from Mill's inequality; see the derivations in the proof of Theorem 3 in Chernozhukov et al. (2015). □

Lemma C.2.9. *Whenever* $0 \le t \le C_0\sqrt{\log d}$ *for some constant* $C_0 > 0$*, and* $\epsilon = c/\max\{(\log d)^{3/2}, \mathfrak{p}\log d\}$ *for some small enough constant* $c > 0$*, we have*

$$\Lambda(t,\epsilon,\mathfrak{p}) := \frac{(1-2\bar{\Phi}(\frac{t+\epsilon}{\sigma_j}))^{\mathfrak{p}}}{1-(1-2\bar{\Phi}(\frac{t}{\sigma_j}))^{\mathfrak{p}}}\cdot\phi(\frac{t-\epsilon}{\sigma_j}) = O\left(\frac{\sqrt{\log d}}{\mathfrak{p}}\right). \tag{C.2.46}$$

Proof of Lemma C.2.9. By Lemma C.2.8, we can simplify $\Lambda(t,\epsilon,\mathfrak{p})$ into the following

$$\Lambda(t,\epsilon,\mathfrak{p}) \le \frac{\left(1-\frac{\phi(\frac{t+\epsilon}{\sigma_j})}{\frac{t+\epsilon}{\sigma_j}\vee 1}\right)^{\mathfrak{p}}}{1-\left(1-\frac{\phi(\frac{t}{\sigma_j})}{\frac{t}{\sigma}\vee 1}\right)^{\mathfrak{p}}}\cdot\phi(\frac{t-\epsilon}{\sigma_j}).$$

When $\frac{t}{\sigma_j} \le 1$, we have $\frac{t+\epsilon}{\sigma_j} \le 2$ due to the choice of ϵ. Because $\frac{t}{\sigma_j} > 0$, $\frac{t+\epsilon}{\sigma_j} > 0$ and $\phi(z)$ is monotonically decreasing when $z > 0$, we then have the the bound below,

$$\Lambda(t,\epsilon,p) \le \frac{(1-\phi(2)/2)^{\mathfrak{p}}}{1-(1-\phi(1))^{\mathfrak{p}}} = O\left(\frac{\sqrt{\log d}}{\mathfrak{p}}\right),$$

where the second inequality holds due to $0 < \phi(2) < \phi(1) < 0.5$ and $\mathfrak{p} > 1$. Now it suffices to consider the case where $\frac{t+\epsilon}{\sigma_j} > \frac{t}{\sigma_j} > 1$ and deal with the following

$$\Lambda(t,\epsilon,\mathfrak{p}) \le \frac{\left(1-\frac{\phi(\frac{t+\epsilon}{\sigma_j})}{\frac{t+\epsilon}{\sigma_j}}\right)^{\mathfrak{p}}}{1-\left(1-\frac{\phi(\frac{t+\epsilon}{\sigma_j})}{\frac{t+\epsilon}{\sigma_j}}\right)^{\mathfrak{p}}}\cdot\phi\left(\frac{t-\epsilon}{\sigma_j}\right).$$

We further bound $\Lambda(t, \epsilon, \mathfrak{p})$ as

$$\begin{aligned}
\Lambda(t, \epsilon, \mathfrak{p}) &\leq \frac{\left(1 - \frac{\phi(\frac{t+\epsilon}{\sigma_j})}{\frac{t+\epsilon}{\sigma_j}}\right)^{\mathfrak{p}}}{1 - \left(1 - \frac{\phi(\frac{t+\epsilon}{\sigma_j})}{\frac{t+\epsilon}{\sigma_j}}\right)^{\mathfrak{p}}} \cdot \phi\Big(\frac{t+\epsilon}{\sigma_j}\Big) \cdot e^{\frac{t\epsilon}{2\sigma_j^2}} \\
&\leq 2\frac{\left(1 - \frac{\phi(\frac{t+\epsilon}{\sigma_j})}{\frac{t+\epsilon}{\sigma_j}}\right)^{\mathfrak{p}}}{1 - \left(1 - \frac{\phi(\frac{t+\epsilon}{\sigma_j})}{\frac{t+\epsilon}{\sigma_j}}\right)^{\mathfrak{p}}} \cdot \phi\Big(\frac{t+\epsilon}{\sigma_j}\Big) \\
&:= 2e^{H(\lambda)} \cdot \frac{t+\epsilon}{\mathfrak{p}\sigma_j}
\end{aligned}$$

where the first inequality comes from rearranging, the second inequality holds since $\exp\left(\frac{t\epsilon}{2\sigma_j^2}\right) < 2$ for $t \leq C_0\sqrt{\log d}$. This is because $\epsilon = c/\max\{(\log d)^{3/2}, \mathfrak{p}\log d\}$ for some small enough constant $c > 0$. The last line holds by rewriting using some new notations: $\lambda := \mathfrak{p}\,\frac{\phi(\frac{t+\epsilon}{\sigma_j})}{\frac{t+\epsilon}{\sigma_j}}$ and

$$H(\lambda) := \log\left(\frac{(1-\frac{\lambda}{\mathfrak{p}})^{\mathfrak{p}}}{1-(1-\frac{\lambda}{\mathfrak{p}})^{\mathfrak{p}}} \cdot \lambda\right) = \mathfrak{p}\log\left(1-\frac{\lambda}{\mathfrak{p}}\right) - \log\left(1-(1-\frac{\lambda}{\mathfrak{p}})^{\mathfrak{p}}\right) + \log\lambda \quad \text{(C.2.47)}$$

Since $\frac{t+\epsilon}{\sigma_j} > 1$, we have $0 < \lambda < \mathfrak{p}$. Below we will first deal with $H(\lambda)$ then obtain the bound on

$\Lambda(t, \epsilon, \mathfrak{p})$. To bound $H(\lambda)$, consider taking the derivative of $H(\lambda)$ with respect to λ, then we have

$$\begin{aligned} H'(\lambda) &= \frac{\mathfrak{p}}{\lambda - \mathfrak{p}} - \frac{(1 - \frac{\lambda}{\mathfrak{p}})^{(\mathfrak{p}-1)}}{1 - (1 - \frac{\lambda}{\mathfrak{p}})^{\mathfrak{p}}} + \frac{1}{\lambda} \\ &= \frac{\mathfrak{p}}{\lambda - \mathfrak{p}} - \frac{1}{1 - \frac{\lambda}{\mathfrak{p}}} \cdot \frac{(1 - \frac{\lambda}{\mathfrak{p}})^{\mathfrak{p}}}{1 - (1 - \frac{\lambda}{\mathfrak{p}})^{\mathfrak{p}}} + \frac{1}{\lambda} \\ &\leq \frac{\mathfrak{p}}{\lambda - \mathfrak{p}} - \frac{1}{1 - \frac{\lambda}{\mathfrak{p}}} \cdot \frac{1 - \lambda}{1 - (1 - \lambda)} + \frac{1}{\lambda} \\ &= \frac{\mathfrak{p}}{\lambda - \mathfrak{p}} + \frac{1}{1 - \frac{\lambda}{\mathfrak{p}}} - \frac{1}{\lambda} \cdot \frac{1}{1 - \frac{\lambda}{\mathfrak{p}}} + \frac{1}{\lambda} \\ &\leq \frac{1}{\lambda}\Big(1 - \frac{\mathfrak{p}}{\mathfrak{p} - \lambda}\Big) < 0, \end{aligned}$$

where the first inequality holds by the Bernoulli's inequality: $(1 + x)^r \geq 1 + rx$ when $r \in \mathbb{N}, 1 + x \geq 0$, and the last inequality holds since $0 < \lambda < \mathfrak{p}$. Now we have $H(\lambda)$ is monotone decreasing. When $0 \leq t \leq C_0\sqrt{\log d}$, we will first find the lower bound on $\lambda = \mathfrak{p}\, \frac{\phi(\frac{t+\epsilon}{\sigma_j})}{\frac{t+\epsilon}{\sigma_j}}$, denoted by $\underline{\lambda}$. Then we have $H(\lambda)$ is bounded by $H(\underline{\lambda})$ due to its monotonicity. Regarding $\underline{\lambda}$, we denote $\bar{x} := 2C_0\sqrt{\log d}/\sigma_j$ and note $\frac{\phi(x)}{x}$ is monotone decreasing when $x \geq 0$. Then we have, when $0 \leq t \leq C_0\sqrt{\log d}$,

$$\mathfrak{p}\, \frac{\phi(\frac{t+\epsilon}{\sigma_j})}{\frac{t+\epsilon}{\sigma_j}} \geq \mathfrak{p}\, \frac{\phi(\bar{x})}{\bar{x}} \geq \frac{\mathfrak{p}}{d^{a_1}} := \underline{\lambda},$$

where $a_1 > 2$. Therefore we obtain

$$H(\lambda) \leq H(\underline{\lambda}) = \log\left(\frac{(1 - \frac{\lambda}{\mathfrak{p}})^{\mathfrak{p}}}{1 - (1 - \frac{\lambda}{\mathfrak{p}})^{\mathfrak{p}}} \cdot \lambda\right)\Bigg|_{\lambda=\underline{\lambda}} \leq \log\left(\frac{\lambda}{1 - (1 - \mathfrak{p}\frac{\lambda}{\mathfrak{p}} + \frac{(\mathfrak{p}-1)\mathfrak{p}}{2}\frac{\lambda^2}{\mathfrak{p}^2})}\right)\Bigg|_{\lambda=\underline{\lambda}} \leq C', \tag{C.2.48}$$

where the second inequality holds due to the fact that $(1 - \frac{\lambda}{\mathfrak{p}})^{\mathfrak{p}} \leq 1$, $\frac{\lambda}{\mathfrak{p}} \in [0, 1]$ and Lemma C.2.11.

The third inequality holds since $\underline{\lambda} = \frac{\mathfrak{p}}{d^{a_1}} \le \frac{1}{d^{a_1-1}} < \frac{1}{d}$, then we have

$$\left(\frac{\lambda}{1-(1-\mathfrak{p}\frac{\lambda}{\mathfrak{p}}+\frac{(\mathfrak{p}-1)\mathfrak{p}}{2}\frac{\lambda^2}{\mathfrak{p}^2})}\right)\Bigg|_{\lambda=\underline{\lambda}} = \frac{\underline{\lambda}}{\underline{\lambda}-\frac{2(\mathfrak{p}-1)}{\mathfrak{p}}\underline{\lambda}^2} \le \frac{\underline{\lambda}}{\underline{\lambda}-2\underline{\lambda}^2} = \frac{1}{1-2\underline{\lambda}} \le C_1',$$

for some constant C_1'. Now we figure out the bound on $\Lambda(t,\epsilon,\mathfrak{p})$,

$$\begin{aligned}
\Lambda(t,\epsilon,\mathfrak{p}) &\le \frac{\left(1-\frac{\phi(\frac{t+\epsilon}{\sigma_j})}{\frac{t+\epsilon}{\sigma_j}}\right)^{\mathfrak{p}}}{1-\left(1-\frac{\phi(\frac{t+\epsilon}{\sigma_j})}{\frac{t+\epsilon}{\sigma_j}}\right)^{\mathfrak{p}}}\cdot\phi\Big(\frac{t-\epsilon}{\sigma_j}\Big) \\
&\le \frac{\left(1-\frac{\phi(\frac{t+\epsilon}{\sigma_j})}{\frac{t+\epsilon}{\sigma_j}}\right)^{\mathfrak{p}}}{1-\left(1-\frac{\phi(\frac{t+\epsilon}{\sigma_j})}{\frac{t+\epsilon}{\sigma_j}}\right)^{\mathfrak{p}}}\cdot\phi\Big(\frac{t+\epsilon}{\sigma_j}\Big)\cdot e^{\frac{t\epsilon}{2\sigma_j^2}} \\
&\le 2\,\frac{\left(1-\frac{\phi(\frac{t+\epsilon}{\sigma_j})}{\frac{t+\epsilon}{\sigma_j}}\right)^{\mathfrak{p}}}{1-\left(1-\frac{\phi(\frac{t+\epsilon}{\sigma_j})}{\frac{t+\epsilon}{\sigma_j}}\right)^{\mathfrak{p}}}\cdot\phi\Big(\frac{t+\epsilon}{\sigma_j}\Big) \\
&\le 2e^{H(\lambda)}\cdot\frac{t+\epsilon}{\mathfrak{p}\sigma_j} \le \frac{C\sqrt{\log d}}{\mathfrak{p}},
\end{aligned}$$

where the second inequality comes from rearranging, the third inequality holds since $\exp\left(\frac{t\epsilon}{2\sigma_j^2}\right) < 2$ for $t \le C_0\sqrt{\log d}$. This is because $\epsilon = c/\max\{(\log d)^{3/2}, \mathfrak{p}\log d\}$ for some small enough constant $c > 0$. And the last line holds by (C.2.47) and (C.2.48). Therefore Lemma C.2.9 is established. □

Lemma C.2.10. *Under the same conditions as Lemma C.2.9, we have*

$$\frac{(1-2\bar{\Phi}(\frac{t+\epsilon}{\sigma_j}))^{\mathfrak{p}-1}}{1-(1-2\bar{\Phi}(\frac{t}{\sigma_j}))^{\mathfrak{p}}}\cdot\Big(\mathfrak{p}\phi\big(\frac{t-\epsilon}{\sigma_j}\big)\Big)^2 = O\left(\sqrt{\log d}\right). \tag{C.2.49}$$

Proof. Note that the result (C.2.46) in Lemma C.2.9 can be rewritten as

$$\mathfrak{p}\,\Lambda(t,\epsilon,\mathfrak{p}) = \frac{(1-2\bar{\Phi}(\frac{t+\epsilon}{\sigma_j}))^{\mathfrak{p}}}{1-(1-2\bar{\Phi}(\frac{t}{\sigma_j}))^{\mathfrak{p}}} \cdot \Big(\mathfrak{p}\phi\big(\frac{t-\epsilon}{\sigma_j}\big)\Big) = O\left(\sqrt{\log d}\right).$$

By similar derivations as in the proof of Lemma C.2.9, we can establish

$$\frac{(1-2\bar{\Phi}(\frac{t+\epsilon}{\sigma_j}))^{\mathfrak{p}-1}}{1-(1-2\bar{\Phi}(\frac{t}{\sigma_j}))^{\mathfrak{p}}} \cdot \Big(\mathfrak{p}\phi\big(\frac{t-\epsilon}{\sigma_j}\big)\Big)^2 = O\left(\sqrt{\log d}\right).$$

□

Lemma C.2.11. *For $x \in [0,1]$, we have $(1-x)^{\mathfrak{p}} \leq 1 - \mathfrak{p}x + 0.5\mathfrak{p}(\mathfrak{p}-1)x^2$.*

Proof. When $\mathfrak{p} = 1$, the above simply holds. Now we consider the case where $\mathfrak{p} > 1$. Let $Q(x) = (1-x)^{\mathfrak{p}} - (1 - \mathfrak{p}x + 0.5\mathfrak{p}(\mathfrak{p}-1)x^2)$, we have $Q(0) = 0$ and

$$Q'(x) = -\mathfrak{p}(1-x)^{(\mathfrak{p}-1)} + \mathfrak{p} - \mathfrak{p}(\mathfrak{p}-1)x \leq -\mathfrak{p}(1-(\mathfrak{p}-1)x) + \mathfrak{p} - \mathfrak{p}(\mathfrak{p}-1)x = 0, \tag{C.2.50}$$

where the inequality holds by applying Bernoulli's inequality to $(1-x)^{(\mathfrak{p}-1)}$ for $\mathfrak{p} > 1, x \in [0,1]$. Therefore, $Q(x)$ is monotonically decreasing, and the statement is proved. □

C.2.4 Ancillary lemmas for Theorem 3.3.2

Remark C.2.11.1. *Recall that the connectivity assumption of Theorem 3.3.2 assumes that there exists a disjoint $\mathfrak{p}$-partition of nodes $\cup_{\ell=1}^{\mathfrak{p}} \mathcal{C}_\ell = [d]$ such that $\sigma^U_{jk} = \sigma^V_{jk} = 0$ when $j \in \mathcal{C}_\ell$ and $k \in \mathcal{C}_{\ell'}$ for some $\ell \neq \ell'$. The more general version of the variance condition assumes: $a_0 \leq \sigma^U_{jj} = \sigma^V_{jj} \leq a_1$, $\forall j \in [d]$; given any $j \in \mathcal{C}^U_\ell$ with some ℓ, there exists at least one $m \in \mathcal{C}^U_{\ell'}$ such that $\sigma^U_{jj} = \sigma^U_{mm}$ for any $\ell' \neq \ell$. Denote $\widetilde{\sigma}^U_{jk} = \sigma^U_{jk}/\sqrt{\sigma^U_{jj}\sigma^U_{kk}}$. And the general covariance condition says that there*

exists some $\sigma_0 < 1$ *such that* $|\widetilde{\sigma}^V_{jk}| = |\sigma^V_{jk}|/\sqrt{\sigma^V_{jj}\sigma^V_{kk}} \leq \sigma_0$ *for any* $j \neq k$ *and* $|\{(j,k) : j \neq k, |\widetilde{\sigma}^U_{jk}| = |\sigma^U_{jk}|/\sqrt{\sigma^U_{jj}\sigma^U_{kk}} > \sigma_0\}| \leq b_0$ *for some constant* b_0.

Lemma C.2.12. *For the term* $\mathrm{II}_1 = \mathbb{E}\left[e^{-\beta(||Z||_\infty - |Z_j|)} \cdot \mathbb{1}(E_1) \cdot \mathbb{1}(t - \epsilon \leq ||Z||_\infty \leq t + \epsilon)\right]$ *with* E_1 *defined in* (C.2.20) *and* $\epsilon = c/\max\{(\log d)^{3/2}, \mathfrak{p}\log d\}$ *for some small enough constant* $c > 0$, *whenever* t *satisfies* $0 \leq t \leq C_0\sqrt{\log d}$ *for some constant* $C_0 > 0$, *we have*

$$\frac{\mathrm{II}_1}{\mathbb{P}(||V||_\infty > t)} \leq \frac{C'\epsilon \log d}{\beta\mathfrak{p}}\left(1 + \frac{b_0}{\sqrt{1 - (s + (1-s)\sigma_0)^2}}\right). \tag{C.2.51}$$

for any $s \in (0,1)$, *where* $\sigma_0 < 1$ *and* b_0 *are the constants in the assumption of Theorem 3.3.2.*

Remark C.2.12.1. *Recall the definition of* $Z = W(s)$. *Hence the term* II_1 *depends on* s. *In Lemma C.2.5, we are able to derive a uniform upper bound when assuming the minimal eigenvalue condition as in Theorem C.2.4. Since Theorem 3.3.2 does not make assumptions about the minimal eigenvalue condition, we will bound the term* II_1 *differently and the upper bound depend on* s, *as showed in the following proof.*

Proof of Lemma C.2.12. We basically use the same proof strategy as Lemma but will separately deal with two cases. First recall that

$$\mathrm{II}_1 = \mathbb{E}\left[e^{-\beta(||Z||_\infty - |Z_j|)} \cdot \mathbb{1}(||Z||_\infty > ||G||_\infty, ||Z||_\infty > Z_j) \cdot \mathbb{1}(t - \epsilon \leq ||Z||_\infty \leq t + \epsilon)\right].$$

We define $Z^\dagger = (Z_k)_{k \in \mathcal{E}^\dagger}$ where

$$\mathcal{E}^\dagger = \{j\} \cup \mathcal{E}_G \cup \{k \in [d] : |\widetilde{\sigma}^U_{jk}| \leq \sigma_0, \max_{m \in \mathcal{E}_G}\{|\widetilde{\sigma}^U_{mk}|\} \leq \sigma_0\}. \tag{C.2.52}$$

Under the condition of Theorem 3.3.2, we have $|[d] \setminus \mathcal{E}^\dagger| \leq |\{(j,k) : j \neq k, |\widetilde{\sigma}^U_{jk}|\} > \sigma_0\}| \leq b_0$ for some constant b_0. Note we can write $1 = \mathbb{1}(||Z^\dagger||_\infty = ||Z||_\infty) + \sum_{k \in [d]\setminus\mathcal{E}^\dagger} \mathbb{1}(|Z_k| = ||Z||_\infty)$.

Then we have

$$\frac{\mathrm{II}_1}{\mathbb{P}(||V||_\infty > t)} \leq \frac{\mathrm{II}_1^\dagger}{\mathbb{P}(||V||_\infty > t)} + b_0 \cdot \max_{k\in[d]\setminus\mathcal{E}^\dagger} \frac{\mathrm{II}_1^{(k)}}{\mathbb{P}(||V||_\infty > t)}, \tag{C.2.53}$$

where $\mathrm{II}_1^\dagger$ and $\mathrm{II}_1^{(k)}$ are defined as

$$\begin{aligned}
\mathrm{II}_1^\dagger :=& \mathbb{E}\left[e^{-\beta(||Z^\dagger||_\infty - |Z_j|)} \cdot \mathbb{1}(||Z^\dagger||_\infty > ||G||_\infty, ||Z^\dagger||_\infty > Z_j) \cdot \mathbb{1}(t-\epsilon \leq ||Z^\dagger||_\infty \leq t+\epsilon)\right], \\
\mathrm{II}_1^{(k)} :=& \mathbb{E}\left[e^{-\beta(|Z_k| - |Z_j|)} \cdot \mathbb{1}(|Z_k| > ||G||_\infty, |Z_k| > Z_j) \cdot \mathbb{1}(t-\epsilon \leq |Z_k| \leq t+\epsilon)\right].
\end{aligned} \tag{C.2.54}$$

Denote the conditional density function of $||Z^\dagger||_\infty \mid Z_j = z_j, G = g$ by $f^\dagger_{g,z_j}(u)$. Then we apply exactly the same derivations as in Lemma C.2.5 (except that $f^\dagger_{g,z_j}(u)$ is bounded using Lemma C.2.13 instead of Lemma C.2.7) and obtain the following bound

$$\frac{\mathrm{II}_1^\dagger}{\mathbb{P}(||V||_\infty > t)} \leq \frac{C'\epsilon \log d}{\beta \mathfrak{p}}. \tag{C.2.55}$$

Regarding the term $\mathrm{II}_1^{(k)}$, we follow the same derivations as in the beginning of the proof of Lemma C.2.5. Specifically, we have

$$\begin{aligned}
\mathrm{II}_1^{(k)} =& \ \mathbb{E}\left[e^{-\beta(|Z_k| - |Z_j|)} \cdot \mathbb{1}(|Z_k| > ||G||_\infty, |Z_k| > Z_j) \cdot \mathbb{1}(t-\epsilon \leq |Z_k| \leq t+\epsilon)\right] \\
=& \ \mathbb{E}\left[e^{\beta|Z_j|} \cdot \mathbb{1}(||G||_\infty \leq t+\epsilon, |Z_j| \leq t+\epsilon) \left(\int_{t-\epsilon}^{t+\epsilon} f_{Z_j}(u) e^{-\beta u} \mathbb{1}(u > ||G||_\infty, u > |Z_j|) du\right)\right],
\end{aligned} \tag{C.2.56}$$

where $f_{Z_j,G}(u)$ denotes the conditional density of Z_k given Z_j, G. Recall the construction of G described in the proof of Theorem 3.3.2, we have for any $k \neq j, k \notin \mathcal{E}_G = \{m \in [d] : Z_m = G_\ell \text{ for some } \ell \in [\mathfrak{p}]\}$, there exists at most one $m \in \{j\} \cup \mathcal{E}_G$, such that Z_k and Z_m belong to

the same component. Denote that random variable by Z_{m_0}, then $f_{Z_j,G}(u)$ is just the conditional density function of Z_k given Z_{m_0}. Since Z follows a multivariate Gaussian distribution, we can immediately figure out the expression of the conditional density $f_{Z_{m_0}}(u)$ and simply derive a bound

$$\begin{aligned} f_{Z_j,G}(u) = f_{Z_{m_0}}(u) &\leq \frac{1}{\sqrt{2\pi \mathrm{Var}\left(Z_k \mid Z_{m_0}\right)}} \\ &= \frac{1}{\sqrt{2\pi(\sigma_{kk} - \sigma_{km_0}^2/\sigma_{m_0m_0}))}} \\ &= \frac{1}{\sqrt{2\pi\sigma_{kk}}} \cdot \frac{1}{1 - \sigma_{km_0}^2/(\sigma_{kk}\sigma_{m_0m_0})} \\ &\leq \frac{1}{\sqrt{2\pi a_0}} \cdot \frac{1}{1 - \sigma_{km_0}^2/(\sigma_{kk}\sigma_{m_0m_0})}, \end{aligned} \tag{C.2.57}$$

where $\sigma_{kk} = \mathrm{Var}\left(Z_k\right), \sigma_{m_0m_0} = \mathrm{Var}\left(Z_{m_0}\right), \sigma_{km_0} = \mathrm{Cov}\left(Z_k, Z_{m_0}\right)$ and we use the fact that $\sigma_{kk} = \mathrm{Var}\left(Z_k\right) = \sigma_{kk}^U \geq a_0$ (under the general variance assumption). Note $Z = \sqrt{s}U + \sqrt{1-s}V$, then we have $\sigma_{km_0}^2 = \left(\mathrm{Cov}\left(Z_k, Z_{m_0}\right)\right)^2 = (s\sigma_{km_0}^U + (1-s)\sigma_{km_0}^V)^2$ where $m_0 \in \{j\} \cup \mathcal{E}_G$. Since $|\widetilde{\sigma}_{km_0}^U| \leq 1$ by definition and $|\widetilde{\sigma}_{km_0}^V| \leq \sigma_0$ under the assumption of Theorem 3.3.2, we have

$$(s\sigma_{km_0}^U + (1-s)\sigma_{km_0}^V)^2/(\sigma_{kk}\sigma_{m_0m_0}) = (s\widetilde{\sigma}_{km_0}^U + (1-s)\widetilde{\sigma}_{km_0}^V)^2 \leq (s + (1-s)\sigma_0)^2. \tag{C.2.58}$$

Now we obtain a upper bound on the conditional density function $f_{Z_j,G}(u)$ based on (C.2.57) and (C.2.58). Combining this bound and following the same derivations as in Lemma C.2.5 to deal with the term in (C.2.56), we establish the upper bound on the term $\mathrm{II}_1^{(k)}/\mathbb{P}(||V||_\infty > t)$ for any $k \in [d] \setminus \mathcal{E}^\dagger$,

$$\frac{\mathrm{II}_1^{(k)}}{\mathbb{P}(||V||_\infty > t)} \leq \frac{C'\epsilon \log d}{\beta \mathfrak{p}} \cdot \frac{1}{\sqrt{1 - (s + (1-s)\sigma_0)^2}}. \tag{C.2.59}$$

Combining (C.2.53), (C.2.54),(C.2.55) with (C.2.59), we derive the bound in (C.2.51). □

Lemma C.2.13. *Recall that the density function of the conditional distribution of* $||Z^\dagger||_\infty \mid \{Z_j = z_j, G = g\}$ *is denoted by* $f^\dagger_{g,z_j}(z)$ *where* $Z^\dagger$ *is defined in .... Suppose* $\epsilon > 0$*, when* $0 \le t \le C_0\sqrt{\log d}$ *for some constant* $C_0 > 0$ *and* $|z_j|, ||g||_\infty \le t + \epsilon$*, we have*

$$f^\dagger_{g,z_j}(z) \le C\sqrt{\log d}, \quad \forall\, z \in (\max\{|z_j|, ||g||_\infty\}, t + \epsilon]. \tag{C.2.60}$$

where the finite constant C *depends on* a_0 *and* $\sigma_0 < 1$*.*

Proof of Lemma C.2.13. Following exactly the same derivations as in Lemma C.2.7 (up to (C.2.45)), we have

$$f^\dagger_{g,z_j}(z) \le \frac{6}{\underline{\sigma_{\cdot j}}}\left(\frac{2\bar\rho_j(1+c_0)+1}{\underline{\sigma_{\cdot j}}}C_1\sqrt{\log d} + C_2\sqrt{\log d}\right), \tag{C.2.61}$$

for any $z \in (\max\{|z_j|, ||g||_\infty\}, t + \epsilon]$, where C_1, C_2 are some constants. First, $\bar\rho_j$ is defined in (C.2.37). Simply, we have

$$\bar\rho_j \le \max_{k\neq j}\frac{|\sigma_{jk}|}{\sigma_{jj}} \le \frac{\max_j \sigma^U_{jj}}{\min_j \sigma^U_{jj}} \le \frac{a_1}{a_0}$$

under the general variance assumption. Recall the construction of G described in the proof of Theorem 3.3.2, we have for any $k \neq j, k \notin \mathcal{E}_G = \{m \in [d] : Z_m = G_\ell \text{ for some } \ell \in [\mathfrak{p}]\}$, there is at most one $m \in \mathcal{E}_G$, such that Z_k and Z_m belong to the same component. Then we have

$$\sum_{m\in\mathcal{E}_G} \mathbb{1}(\sigma_{km} \neq 0) \le 1,$$

hence $c_0 = 1$ by definition. Also note by the definition of $Z^\dagger$ and $\mathcal{E}^\dagger$ in (C.2.52), for any $k \in \mathcal{E}^\dagger, k \neq j, k \neq \mathcal{E}_G$, we have

$$\max\{|\widetilde\sigma^U_{jk}|, |\widetilde\sigma^V_{jk}|\}\} \le \sigma_0, \quad \max_{m\in\mathcal{E}_G}\{|\widetilde\sigma^U_{mk}|, |\widetilde\sigma^V_{mk}|\} \le \sigma_0 \tag{C.2.62}$$

under the assumption of Theorem 3.3.2. We will take advantage of this together with the above

property of G to derive a bound on $\underline{\sigma_{\cdot j}}$. Similarly as in (C.2.37), we have $\underline{\sigma^2_{\cdot j}} := \min_{k\in\mathcal{E}_X} \mathrm{Var}\,(Z_k \mid Z_j, G)$ with $\mathcal{E}_X := \{k \in \mathcal{E}^\dagger : k \neq j, k \notin \mathcal{E}_G\}$. For each $k \in \mathcal{E}_X$, we have it can at most belong to the same component as one of $\{j\} \cup \mathcal{E}_G$, due to the property of G. Then we have

$$\begin{aligned}
\mathrm{Var}\,(Z_k \mid Z_j, G) \geq\ & \min\{\mathrm{Var}\,(Z_k \mid Z_j)\,,\ \min_{m\in\mathcal{E}_G}\{\mathrm{Var}\,(Z_k \mid Z_m)\}\} \\
=\ & \sigma_{kk}\cdot\min\{1-\sigma^2_{jk}/(\sigma_{jj}\sigma_{kk}),\ \min_{m\in\mathcal{E}_G}\{1-\sigma^2_{mk}/(\sigma_{mm}\sigma_{kk})\}\} \\
\geq\ & a_0\cdot\min\{1-\sigma^2_{jk}/(\sigma_{jj}\sigma_{kk}),\ \min_{m\in\mathcal{E}_G}\{1-\sigma^2_{mk}/(\sigma_{mm}\sigma_{kk})\}\}. \qquad \text{(C.2.63)}
\end{aligned}$$

since (Z_j, G) are all independent and $\sigma_{kk} = \mathrm{Var}\,(Z_k) = \sigma^U_{kk} \geq a_0$ (under the general variance assumption). Recall the definition of $Z = \sqrt{s}U + \sqrt{1-s}V$, we have

$$|\sigma_{mk}|/\sqrt{\sigma_{mm}\sigma_{kk}} = |\mathrm{Cov}\,(Z_k, Z_m)\,|/\sqrt{\sigma_{mm}\sigma_{kk}} = |s\widetilde{\sigma}^U_{mk} + (1-s)\widetilde{\sigma}^V_{mk}| \leq \sigma_0, \quad \forall\, s \in [0,1], \qquad \text{(C.2.64)}$$

when $m \in \{j\} \cup \mathcal{E}_G$. This is due to (C.2.62). Then we can derive a bound on $1/\underline{\sigma^2_{\cdot j}}$, i.e.,

$$\begin{aligned}
\frac{1}{\underline{\sigma^2_{\cdot j}}} &= \frac{1}{\min_{k\in\mathcal{E}_X}\mathrm{Var}\,(Z_k \mid Z_j, G)} \\
&\leq \frac{1}{a_0\min_{k\in\mathcal{E}_X}\min\{1-\sigma^2_{jk}/(\sigma_{jj}\sigma_{kk}), \min_{m\in\mathcal{E}_G}\{1-\sigma^2_{mk}/(\sigma_{mm}\sigma_{kk})\}\}} \\
&\leq \frac{1}{a_0(1-\sigma_0^2)},
\end{aligned}$$

where the first inequality holds by (C.2.63) and the second equality holds by (C.2.64). Combining the above bound with (C.2.61), we finally establish (C.2.60) for some finite constant C. □

C.3 Ancillary propositions for FDR control

Throughout this section, we introduce some new notations. For a given mean zero random vector $\boldsymbol{Y} \in \mathbb{R}^d$ with positive semi-definite covariance matrix $\boldsymbol{\Sigma}^Y := \mathbb{E}\left[\boldsymbol{Y}\boldsymbol{Y}^\top\right] \in \mathbb{R}^{d\times d}$, we denote its Gaussian counterpart by $\boldsymbol{Z} \in \mathbb{R}^d$ (i.e., $\mathbb{E}\left[\boldsymbol{Z}\right] = \boldsymbol{0}$ and its covariance matrix $\mathbb{E}\left[\boldsymbol{Z}\boldsymbol{Z}^\top\right] := \boldsymbol{\Sigma}^Z$ equals $\boldsymbol{\Sigma}^Y = (\sigma^Y_{jk})_{1\le j,k\le d}$). Consider n i.i.d. copies of $\boldsymbol{Y}$, denoted by $\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_n \in \mathbb{R}^d$. We define the maximum $T_{\boldsymbol{Y}}$ and $T_{\boldsymbol{Z}}$ as below,

$$T_{\boldsymbol{Y}} := \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \boldsymbol{Y}_i \right\|_\infty, \quad T_{\boldsymbol{Z}} := ||\boldsymbol{Z}||_\infty, \tag{C.3.1}$$

where $q(\alpha; T_{\boldsymbol{Y}})$ and $q(\alpha; T_{\boldsymbol{Z}})$ ($\alpha \in [0,1]$) are the corresponding upper quantile functions. Define the Gaussian multiplier bootstrap counterpart as

$$T_{\boldsymbol{W}} := \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \boldsymbol{Y}_i \xi_i \right\|_\infty, \tag{C.3.2}$$

where $\xi_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ and are independent from $\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_n$. Let $q_\xi(\alpha; T_{\boldsymbol{W}})$ be the conditional quantile of $T_{\boldsymbol{W}}$, then we have $\mathbb{P}_\xi\left(T_{\boldsymbol{W}} \ge q_\xi(\alpha; T_{\boldsymbol{W}})\right) = \alpha$. Note that we use the ξ subscript to remind ourselves that the probability measure is induced by the multiplier random variables $\{\xi_i\}_{i=1}^n$ conditional on $\{\boldsymbol{Y}_i\}_{i=1}^n$. And we have the covariance matrix of $\frac{1}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{Y}_i\xi_i$ (conditional on $\{\boldsymbol{Y}_i\}_{i=1}^n$) equals $\boldsymbol{\Sigma}^W := \frac{1}{n}\sum_{i=1}^n \boldsymbol{Y}_i\boldsymbol{Y}_i^\top$. Denote $\Delta_\infty = ||\boldsymbol{\Sigma}^Z - \boldsymbol{\Sigma}^W||_\infty$, which measures the maximal differences between the true covariance matrix $\boldsymbol{\Sigma}^Z$ and the sample version $\boldsymbol{\Sigma}^W$.

C.3.1 Cramér-type deviation bounds for the Gaussian multiplier bootstrap

Based on the Cramér-type Gaussian comparison bound in Theorem 3.3.1, the Cramér-type approximation bound (Kuchibhotla et al., 2021), the maximal inequalities and a careful treatment to

the comparison of quantiles, we will establish the Cramér-type deviation bounds for the Gaussian multiplier bootstrap (CGMB) in this section.

Proposition C.3.1 (CGMB). *Assuming the covariance matrix $\mathbf{\Sigma}^Y$ satisfies $0 < c_1 \leq \sigma_{jj}^Y \leq c_2 < \infty$, for any $j \in [d]$ and $\boldsymbol{Y}$ satisfies the tail condition that $\max_{1\leq i\leq n} \max_{1\leq j\leq p} ||\boldsymbol{Y}_{ij}||_{\psi_1} \leq K_3$ for some constants c_1, c_2, K_3, under the scaling condition $(\log ed)^3(\log(ed+n))^{56/3}/n = o(1)$, we have the following bound,*

$$\sup_{\alpha\in[\alpha_L,1]} \left| \frac{\mathbb{P}(T_{\boldsymbol{Y}} > q_\xi(\alpha; T_{\boldsymbol{W}}))}{\mathbb{P}(T_{\boldsymbol{Z}} > q(\alpha; T_{\boldsymbol{Z}}))} - 1 \right| = O\left(\frac{(\log d)^{11/6}}{n^{1/6}\alpha_L^{1/3}} + \frac{(\log d)^{19/6}}{n^{1/6}} \right), \tag{C.3.3}$$

where α_L satisfies $q(\alpha_L; T_{\boldsymbol{Z}}) = O\left(\sqrt{\log d}\right)$ and $\frac{\log^{11} d}{n\alpha_L} = O(1)$.

The proof can be found in Appendix C.3.2. In practice, there are many situations where the relevant statistics come from the maxima of approximated averages. In particular, the test statistics in our node selection problem can not be directly expressed as maxima of scaled averages, but can be approximated by a $T_{\boldsymbol{Y}}$-like term with the approximation error suitably controlled. Therefore, we also prove an extended version of Proposition C.3.1. Suppose the statistics of interest and its Gaussian multiplier bootstrap counterpart, denoted by T and $T^{\mathcal{B}}$ respectively, can be approximated by $T_{\boldsymbol{Y}}$ (defined in (C.3.1)) and $T_{\boldsymbol{W}}$ (defined in (C.3.2)). The quantile functions $q(\alpha; T)$ and $q_\xi(\alpha; T^{\mathcal{B}})$ are defined correspondingly.

Proposition C.3.2 (CGMB with approximation). *Under the same conditions as in Proposition C.3.1 and the additional assumption about the differences between the maximum statistics:*

$$\mathbb{P}(|T - T_{\boldsymbol{Y}}| > \zeta_1) < \zeta_2, \tag{C.3.4}$$

$$\mathbb{P}(\mathbb{P}_\xi(|T^{\mathcal{B}} - T_{\boldsymbol{W}}| > \zeta_1) > \zeta_2) < \zeta_2, \tag{C.3.5}$$

where $\zeta_1, \zeta_2 \geq 0$ characterize the approximation error and satisfy $\zeta_1 \log d = O(1), \zeta_2 = O(\alpha_L)$, we

have the following Cramér-type deviation bound

$$\sup_{\alpha\in[\alpha_L,1]}\left|\frac{\mathbb{P}(T>q_\xi(\alpha;T^{\mathcal{B}}))}{\mathbb{P}(T_{\boldsymbol{Z}}>q(\alpha;T_{\boldsymbol{Z}}))}-1\right|=\eta(d,n,\zeta_1,\zeta_2,\alpha_L), \tag{C.3.6}$$

where $\eta(d,n,\zeta_1,\zeta_2,\alpha_L)=O\left(\frac{(\log d)^{19/6}}{n^{1/6}}+\frac{(\log d)^{11/6}}{n^{1/6}\alpha_L^{1/3}}+\zeta_1\log d+\frac{\zeta_2}{\alpha_L}\right)$.

C.3.2 Proof of Proposition C.3.1

Before proving Proposition C.3.1, we present Lemma C.3.3. It bounds the conditional quantile $q_\xi(\alpha;T_{\boldsymbol{W}})$ in terms of the quantile $q(\alpha;T_{\boldsymbol{Z}})$ of the Gaussian maxima $T_{\boldsymbol{Z}}$ when the maximal covariance matrix differences are controlled. In the proof of Lemma C.3.3, we apply the Cramér-type comparison bound (3.3.1), which is establised in Theorem 3.3.1. To simplify the notation, we denote the bound $C_1(\log d)^{5/2}\Delta_\infty^{1/2}$ in (3.3.1) by $\pi(\Delta_\infty)$, where the constant C_1 only depends on $\min_{1\le j\le d}\{\sigma_{jj}^U,\sigma_{jj}^V\}$, $\max_{1\le j\le d}\{\sigma_{jj}^U,\sigma_{jj}^V\}$.

Lemma C.3.3. *Suppose* δ *satisfies* $(\log d)^5\delta=O(1)$. *On the event* $\{\Delta_\infty\le\delta\}$, *we have*

$$q_\xi(\alpha;T_{\boldsymbol{W}})\ge q\Big(\frac{\alpha}{1-\pi(\delta)};T_{\boldsymbol{Z}}\Big), \tag{C.3.7}$$

$$q_\xi(\alpha;T_{\boldsymbol{W}})\le q\Big(\frac{\alpha}{1+\pi(\delta)};T_{\boldsymbol{Z}}\Big). \tag{C.3.8}$$

Proof of Lemma C.3.3. On the event $\{\Delta_\infty\le\delta\}$, we have $(\log d)^5\Delta_\infty\le(\log d)^5\delta=O(1)$, then by applying Theorem 3.3.1 to $\boldsymbol{Z}$ and $\boldsymbol{W}$, we obtain the following,

$$\sup_{0\le t\le C_0\sqrt{\log d}}\left|\frac{\mathbb{P}_\xi(T_{\boldsymbol{W}}>t)}{\mathbb{P}(T_{\boldsymbol{Z}}>t)}-1\right|\le\pi(\delta).$$

Therefore we have

$$\mathbb{P}_\xi\left(T_{\boldsymbol{W}} \geq q\Big(\frac{\alpha}{1-\pi(\delta)}; T_{\boldsymbol{Z}}\Big)\right) \geq \mathbb{P}\left(T_{\boldsymbol{Z}} \geq q\Big(\frac{\alpha}{1-\pi(\delta)}; T_{\boldsymbol{Z}}\Big)\right) \cdot (1-\pi(\delta)) = \alpha,$$

when t satisfies $0 \leq t \leq C_0\sqrt{\log d}$. Then $q_\xi(\alpha; T_{\boldsymbol{W}}) \geq q\Big(\frac{\alpha}{1-\pi(\delta)}; T_{\boldsymbol{Z}}\Big)$ immediately follows, i.e., (C.3.7) holds. Similarly, on the event $\{\Delta_\infty \leq \delta\}$, we have

$$\mathbb{P}_\xi\left(T_{\boldsymbol{W}} \geq q\Big(\frac{\alpha}{1+\pi(\delta)}; T_{\boldsymbol{Z}}\Big)\right) \leq \mathbb{P}\left(T_{\boldsymbol{Z}} \geq q\Big(\frac{\alpha}{1+\pi(\delta)}; T_{\boldsymbol{Z}}\Big)\right) \cdot (1+\pi(\delta)) = \alpha.$$

Thus $q_\xi(\alpha; T_{\boldsymbol{W}}) \leq q\Big(\frac{\alpha}{1+\pi(\delta)}; T_{\boldsymbol{Z}}\Big)$, i.e., (C.3.8) holds. □

Proof of Proposition C.3.1. By the triangle inequality, we have

$$\left|\frac{\mathbb{P}(T_{\boldsymbol{Y}} > q_\xi(\alpha; T_{\boldsymbol{W}}))}{\mathbb{P}(T_{\boldsymbol{Z}} > q(\alpha; T_{\boldsymbol{Z}}))} - 1\right| \leq \underbrace{\left|\frac{\mathbb{P}(T_{\boldsymbol{Y}} > q(\alpha; T_{\boldsymbol{Z}}))}{\mathbb{P}(T_{\boldsymbol{Z}} > q(\alpha; T_{\boldsymbol{Z}}))} - 1\right|}_{\mathrm{I}} + \underbrace{\frac{|\mathbb{P}(T_{\boldsymbol{Y}} > q_\xi(\alpha; T_{\boldsymbol{W}})) - \mathbb{P}(T_{\boldsymbol{Y}} > q(\alpha; T_{\boldsymbol{Z}}))|}{\mathbb{P}(T_{\boldsymbol{Z}} > q(\alpha; T_{\boldsymbol{Z}}))}}_{\mathrm{II}}. \tag{C.3.9}$$

Regarding the first term I, we will directly apply Corollary 5.1 in Kuchibhotla et al. (2021). Specifically, we verify the tail assumption on $\boldsymbol{Y}$ and the condition on the quantile that $q(\alpha; T_{\boldsymbol{Z}}) \leq q(\alpha_L; T_{\boldsymbol{Z}}) = O\left(\sqrt{\log d}\right)$ when $\alpha \in [\alpha_L, 1]$. Then we obtain the following bound

$$\mathrm{I} = \left|\frac{\mathbb{P}(T_{\boldsymbol{Y}} > q(\alpha; T_{\boldsymbol{Z}}))}{\mathbb{P}\left(T_{\boldsymbol{Z}} > q(\alpha; T_{\boldsymbol{Z}})\right)} - 1\right| = O\left(\frac{(\log d)^{19/6}}{n^{1/6}}\right). \tag{C.3.10}$$

Regarding the second term II, we write it as

$$\begin{aligned}
\mathrm{II} &= \frac{1}{\alpha}\left|\mathbb{P}(T_{\boldsymbol{Y}} > q_\xi(\alpha; T_{\boldsymbol{W}})) - \mathbb{P}(T_{\boldsymbol{Y}} > q(\alpha; T_{\boldsymbol{Z}}))\right| \\
&\leq \frac{1}{\alpha}\mathbb{P}(\{T_{\boldsymbol{Y}} > q_\xi(\alpha; T_{\boldsymbol{W}})\} \ominus \{T_{\boldsymbol{Y}} > q(\alpha; T_{\boldsymbol{Z}})\}) \\
&= \frac{1}{\alpha}\Big(\mathbb{P}(T_{\boldsymbol{Y}} > q_\xi(\alpha; T_{\boldsymbol{W}}), T_{\boldsymbol{Y}} \leq q(\alpha; T_{\boldsymbol{Z}})) + \mathbb{P}(T_{\boldsymbol{Y}} \leq q_\xi(\alpha; T_{\boldsymbol{W}}), T_{\boldsymbol{Y}} > q(\alpha; T_{\boldsymbol{Z}}))\Big) \\
&\leq \frac{1}{\alpha}\mathbb{P}(T_{\boldsymbol{Y}} > q_\xi(\alpha; T_{\boldsymbol{W}}), T_{\boldsymbol{Y}} \leq q(\alpha; T_{\boldsymbol{Z}}), \Delta_\infty \leq \delta) \\
&\quad + \frac{1}{\alpha}\mathbb{P}(T_{\boldsymbol{Y}} \leq q_\xi(\alpha; T_{\boldsymbol{W}}), T_{\boldsymbol{Y}} > q(\alpha; T_{\boldsymbol{Z}}), \Delta_\infty \leq \delta) + \frac{2\mathbb{P}\left(\Delta_\infty > \delta\right)}{\alpha},
\end{aligned}$$

where the first inequality holds by the definition of the symmetric difference; recall the symmetric difference between A and B is defined as $A \ominus B = (A \setminus B) \cup (B \setminus A)$). Remark that we will give the explicit choice of δ later in the proof. Now we apply Lemma C.3.3 (whose condition will be verified in (C.3.16)) and further bound II as,

$$\begin{aligned}
\mathrm{II} &\leq \frac{1}{\alpha}\Big(\mathbb{P}\big(T_{\boldsymbol{Y}} \geq q\big(\frac{\alpha}{1-\pi(\delta)}; T_{\boldsymbol{Z}}\big), T_{\boldsymbol{Y}} \leq q(\alpha; T_{\boldsymbol{Z}})\big) \\
&\qquad + \mathbb{P}(T_{\boldsymbol{Y}} \leq q\big(\frac{\alpha}{1+\pi(\delta)}; T_{\boldsymbol{Z}}\big), T_{\boldsymbol{Y}} > q(\alpha; T_{\boldsymbol{Z}}))\Big) + \frac{2\mathbb{P}\left(\Delta_\infty > \delta\right)}{\alpha} \\
&= \frac{1}{\alpha}\,\mathbb{P}\left(q\big(\frac{\alpha}{1-\pi(\delta)}; T_{\boldsymbol{Z}}\big) \leq T_{\boldsymbol{Y}} \leq q\big(\frac{\alpha}{1+\pi(\delta)}; T_{\boldsymbol{Z}}\big)\right) + \frac{2\mathbb{P}\left(\Delta_\infty > \delta\right)}{\alpha} \qquad \text{(C.3.11)} \\
&\leq \frac{1}{\alpha}\,\mathbb{P}\left(q\big(\frac{\alpha}{1-\pi(\delta)}; T_{\boldsymbol{Z}}\big) \leq T_{\boldsymbol{Z}} \leq q\big(\frac{\alpha}{1+\pi(\delta)}; T_{\boldsymbol{Z}}\big)\right) + \frac{2\mathbb{P}\left(\Delta_\infty > \delta\right)}{\alpha} + \mathrm{III} \\
&= \frac{2\pi(\delta)}{1-\pi^2(\delta)} + \frac{2\mathbb{P}\left(\Delta_\infty > \delta\right)}{\alpha} + \mathrm{III}, \qquad \text{(C.3.12)}
\end{aligned}$$

where the term III in the second inequality is defined as,

$$\mathrm{III} := \frac{1}{\alpha}\left|\mathbb{P}\left(q\big(\frac{\alpha}{1-\pi(\delta)}; T_{\boldsymbol{Z}}\big) \leq T_{\boldsymbol{Y}} \leq q\big(\frac{\alpha}{1+\pi(\delta)}; T_{\boldsymbol{Z}}\big)\right) - \mathbb{P}\left(q\big(\frac{\alpha}{1-\pi(\delta)}; T_{\boldsymbol{Z}}\big) \leq T_{\boldsymbol{Z}} \leq q\big(\frac{\alpha}{1+\pi(\delta)}; T_{\boldsymbol{Z}}\big)\right)\right|.$$

Below we further rewrite III as

$$\mathrm{III} = \frac{1}{\alpha}\left|\frac{\alpha}{1-\pi(\delta)}\cdot \mathrm{III}_1 - \frac{\alpha}{1+\pi(\delta)}\cdot \mathrm{III}_2\right|,$$

with $\mathrm{III}_1, \mathrm{III}_2$ defined as

$$\begin{aligned}
\mathrm{III}_1 &= \frac{\mathbb{P}\left(T_{\boldsymbol{Y}} > q\left(\frac{\alpha}{1-\pi(\delta)}; T_{\boldsymbol{Z}}\right)\right) - \mathbb{P}\left(T_{\boldsymbol{Z}} > q\left(\frac{\alpha}{1-\pi(\delta)}; T_{\boldsymbol{Z}}\right)\right)}{\mathbb{P}\left(T_{\boldsymbol{Z}} > q\left(\frac{\alpha}{1-\pi(\delta)}; T_{\boldsymbol{Z}}\right)\right)}, \\
\mathrm{III}_2 &= \frac{\mathbb{P}\left(T_{\boldsymbol{Y}} > q\left(\frac{\alpha}{1+\pi(\delta)}; T_{\boldsymbol{Z}}\right)\right) - \mathbb{P}\left(T_{\boldsymbol{Z}} > q\left(\frac{\alpha}{1+\pi(\delta)}; T_{\boldsymbol{Z}}\right)\right)}{\mathbb{P}\left(T_{\boldsymbol{Z}} > q\left(\frac{\alpha}{1+\pi(\delta)}; T_{\boldsymbol{Z}}\right)\right)}.
\end{aligned}$$

Thus by applying Corollary 5.1 of Kuchibhotla et al. (2021) to $\mathrm{III}_1, \mathrm{III}_2$ similarly as in (C.3.10), we have the following bound on III,

$$\mathrm{III} = O\left(\frac{(\log d)^{19/6}}{n^{1/6}}\right). \tag{C.3.13}$$

Combining (C.3.12) and (C.3.13) yields the following bound,

$$\begin{aligned}
\mathrm{II} &\leq \frac{1}{\alpha}\mathbb{P}(\{T_{\boldsymbol{Y}} > q_\xi(\alpha; T_{\boldsymbol{W}})\} \ominus \{T_{\boldsymbol{Y}} > q(\alpha; T_{\boldsymbol{Z}})\}) \\
&\leq \frac{C(\log d)^{19/6}}{n^{1/6}} + C_0'\pi(\delta) + \frac{C''\,\mathbb{P}\left(\Delta_\infty > \delta\right)}{\alpha} \\
&\leq \frac{C(\log d)^{19/6}}{n^{1/6}} + C'(\log d)^{5/2}\delta^{1/2} + \frac{C''\,\mathbb{E}\left[\Delta_\infty\right]}{\delta\alpha} \\
&= O\left(\left(\frac{\mathbb{E}\left[\Delta_\infty\right]\log^5 d}{\alpha}\right)^{1/3} + \frac{(\log d)^{19/6}}{n^{1/6}}\right),
\end{aligned} \tag{C.3.14}$$

where the second inequality holds due to the definition of $\pi(\delta)$ and Markov's inequality, the last line holds by choosing δ to be $(\mathbb{E}\left[\Delta_\infty\right])^{2/3}/(\alpha^{1/3}(\log d)^{5/3})$. We will bound the term $\mathbb{E}\left[\Delta_\infty\right]$ using Lemma C.1 in Chernozhukov et al. (2013). Specifically, under the stated tail assumption on

$\boldsymbol{Y}$, the condition (E.1) of Lemma C.1 in Chernozhukov et al. (2013) is satisfied; see Comment 2.2 in Chernozhukov et al. (2013). Thus we have

$$\mathbb{E}\left[\Delta_\infty\right] \leq \sqrt{\frac{B_n^2 \log d}{n}} \vee \frac{B_n^2 (\log(dn))^2 (\log d)}{n}, \tag{C.3.15}$$

where B_n equals some constant C which does not depend on n. As promised previously, we verify the assumption of Lemma C.3.3 for our choice of δ. Specifically, for $\delta = (\mathbb{E}\left[\Delta_\infty\right])^{2/3}/(\alpha^{1/3}(\log d)^{5/3})$, we have $(\log d)^5 \delta$ satisfies the following

$$(\log d)^5 \delta \leq \frac{(\log d)^5 (\mathbb{E}\left[\Delta_\infty\right])^{2/3}}{\alpha_L^{1/3} (\log d)^{5/3}} = \left(\frac{\log^{11} d}{n\alpha_L}\right)^{1/3} = O(1), \tag{C.3.16}$$

under the stated condition on α_L. Finally, when $\alpha \in [\alpha_L, 1]$, we combine (C.3.9), (C.3.10), (C.3.14) with (C.3.15), then establish (C.3.3), i.e.,

$$\sup_{\alpha \in [\alpha_L, 1]} \left| \frac{\mathbb{P}(T_{\boldsymbol{Y}} > q_\xi(\alpha; T_{\boldsymbol{W}}))}{\mathbb{P}(T_{\boldsymbol{Z}} > q(\alpha; T_{\boldsymbol{Z}}))} - 1 \right| = O\left(\frac{(\log d)^{11/6}}{n^{1/6} \alpha_L^{1/3}} + \frac{(\log d)^{19/6}}{n^{1/6}} \right).$$

□

C.3.3 Proof of Proposition C.3.2

Before proving Proposition C.3.2, we need to present a simple lemma. It translates the approximation error ζ_1, ζ_2 into the bounds on the quantiles. And its proof is quite straightforward thus omitted.

Lemma C.3.4. *Under the assumption in* (C.3.5)*, we have, for* $\alpha \in (0, 1)$*,*

$$\mathbb{P}(q_\xi(\alpha; T^{\mathcal{B}}) \leq q_\xi(\alpha + \zeta_2; T_{\boldsymbol{W}}) + \zeta_1) \geq 1 - \zeta_2,$$
$$\mathbb{P}(q_\xi(\alpha; T^{\mathcal{B}}) \geq q_\xi(\alpha - \zeta_2; T_{\boldsymbol{W}}) - \zeta_1) \geq 1 - \zeta_2.$$

Proof of Proposition C.3.2. By the triangle inequality, we have

$$\left|\frac{\mathbb{P}(T > q_\xi(\alpha; T^{\mathcal{B}}))}{\mathbb{P}(T_{\boldsymbol{Z}} > q(\alpha; T_{\boldsymbol{Z}}))} - 1\right| \le \underbrace{\left|\frac{\mathbb{P}(T_{\boldsymbol{Y}} > q(\alpha; T_{\boldsymbol{Z}}))}{\mathbb{P}(T_{\boldsymbol{Z}} > q(\alpha; T_{\boldsymbol{Z}}))} - 1\right|}_{\mathrm{I}} + \underbrace{\frac{|\mathbb{P}(T > q_\xi(\alpha; T^{\mathcal{B}})) - \mathbb{P}(T_{\boldsymbol{Y}} > q(\alpha; T_{\boldsymbol{Z}}))|}{\mathbb{P}(T_{\boldsymbol{Z}} > q(\alpha; T_{\boldsymbol{Z}}))}}_{\mathrm{II}}. \tag{C.3.17}$$

Note that (C.3.10) in the proof of Proposition C.3.1 immediately gives the bound on I, i.e.,

$$\mathrm{I} = O\Big(\frac{(\log d)^{19/6}}{n^{1/6}}\Big). \tag{C.3.18}$$

Regarding the term II, we have

$$\begin{aligned}
\mathrm{II} &= \frac{1}{\alpha}\left|\mathbb{P}(T > q_\xi(\alpha; T^{\mathcal{B}})) - \mathbb{P}(T_{\boldsymbol{Y}} > q(\alpha; T_{\boldsymbol{Z}}))\right| \\
&\le \frac{1}{\alpha}\left|\mathbb{P}(\{T > q_\xi(\alpha; T^{\mathcal{B}})\} \ominus \{T_{\boldsymbol{Y}} > q(\alpha; T_{\boldsymbol{Z}})\})\right| \\
&= \frac{1}{\alpha}\mathbb{P}(T > q_\xi(\alpha; T^{\mathcal{B}}), T_{\boldsymbol{Y}} \le q(\alpha; T_{\boldsymbol{Z}})) + \frac{1}{\alpha}\mathbb{P}(T \le q_\xi(\alpha; T^{\mathcal{B}}), T_{\boldsymbol{Y}} > q(\alpha; T_{\boldsymbol{Z}})).
\end{aligned} \tag{C.3.19}$$

To bound the two terms in (C.3.19), first notice that on the event $|T - T_{\boldsymbol{Y}}| > \zeta_1$, we have

$$\{T > q_\xi(\alpha; T^{\mathcal{B}}), T_{\boldsymbol{Y}} \le q(\alpha; T_{\boldsymbol{Z}})\} \subset \{T_{\boldsymbol{Y}} > q_\xi(\alpha; T^{\mathcal{B}}) - \zeta_1, T_{\boldsymbol{Y}} \le q(\alpha; T_{\boldsymbol{Z}})\}.$$

Then under the assumption in (C.3.4), i.e., $\mathbb{P}(|T - T_{\boldsymbol{Y}}| > \zeta_1) < \zeta_2$, we obtain

$$\mathbb{P}(T > q_\xi(\alpha; T^{\mathcal{B}}), T_{\boldsymbol{Y}} \le q(\alpha; T_{\boldsymbol{Z}})) \le \mathbb{P}(T_{\boldsymbol{Y}} > q_\xi(\alpha; T^{\mathcal{B}}) - \zeta_1, T_{\boldsymbol{Y}} \le q(\alpha; T_{\boldsymbol{Z}})) + \zeta_2.$$

Applying such strategies to the second term in (C.3.19) similarly, we get the following,

$$\begin{aligned} \mathrm{II} &\leq \mathrm{II}_1 + \mathrm{II}_2 + \frac{2\zeta_2}{\alpha}, \quad \text{where} \\ \mathrm{II}_1 &:= \frac{1}{\alpha}\mathbb{P}(T_{\boldsymbol{Y}} > q_\xi(\alpha; T^{\mathcal{B}}) - \zeta_1, T_{\boldsymbol{Y}} \leq q(\alpha; T_{\boldsymbol{Z}})), \\ \mathrm{II}_2 &:= \frac{1}{\alpha}\mathbb{P}(T_{\boldsymbol{Y}} \leq q_\xi(\alpha; T^{\mathcal{B}}) + \zeta_2, T_{\boldsymbol{Y}} > q(\alpha; T_{\boldsymbol{Z}})). \end{aligned} \tag{C.3.20}$$

Under the assumption (C.3.5), by Lemma C.3.4, we have

$$\begin{aligned} \mathbb{P}(q_\xi(\alpha; T^{\mathcal{B}}) \leq q_\xi(\alpha + \zeta_2; T_{\boldsymbol{W}}) + \zeta_1) &\geq 1 - \zeta_2, \\ \mathbb{P}(q_\xi(\alpha; T^{\mathcal{B}}) \geq q_\xi(\alpha - \zeta_2; T_{\boldsymbol{W}}) - \zeta_1) &\geq 1 - \zeta_2. \end{aligned}$$

Hence we can bound $\mathrm{II}_1, \mathrm{II}_2$ as below,

$$\begin{aligned} \mathrm{II}_1 &\leq \frac{1}{\alpha}\mathbb{P}\left(T_{\boldsymbol{Y}} > q_\xi(\alpha - \zeta_2; T_{\boldsymbol{W}}) - 2\zeta_1, T_{\boldsymbol{Y}} \leq q(\alpha; T_{\boldsymbol{Z}})\right) + \frac{\zeta_2}{\alpha}, \\ \mathrm{II}_2 &\leq \frac{1}{\alpha}\mathbb{P}\left(T_{\boldsymbol{Y}} \leq q_\xi(\alpha + \zeta_2; T_{\boldsymbol{W}}) + 2\zeta_1, T_{\boldsymbol{Y}} > q(\alpha; T_{\boldsymbol{Z}})\right) + \frac{\zeta_2}{\alpha}. \end{aligned}$$

Now we will use the strategy of deriving (C.3.11) in the proof of Proposition C.3.1, i.e., apply Lemma C.3.3, then we have,

$$\begin{aligned} \mathrm{II}_1 &\leq \frac{1}{\alpha}\mathbb{P}\left(T_{\boldsymbol{Y}} > q\Big(\frac{\alpha - \zeta_2}{1 - \pi(\delta)}; T_{\boldsymbol{Z}}\Big) - 2\zeta_1, T_{\boldsymbol{Y}} \leq q(\alpha; T_{\boldsymbol{Z}})\right) + \frac{\mathbb{P}\left(\Delta_\infty > \delta\right)}{\alpha} + \frac{\zeta_2}{\alpha}, \\ \mathrm{II}_2 &\leq \frac{1}{\alpha}\mathbb{P}\left(T_{\boldsymbol{Y}} \leq q\Big(\frac{\alpha + \zeta_2}{1 + \pi(\delta)}; T_{\boldsymbol{Z}}\Big) + 2\zeta_1, T_{\boldsymbol{Y}} > q(\alpha; T_{\boldsymbol{Z}})\right) + \frac{\mathbb{P}\left(\Delta_\infty > \delta\right)}{\alpha} + \frac{\zeta_2}{\alpha}. \end{aligned}$$

Combining the above two inequalities with (C.3.20), we have

$$\mathrm{II} \leq \mathrm{III} + \frac{2\mathbb{P}\left(\Delta_\infty > \delta\right)}{\alpha} + \frac{4\zeta_2}{\alpha}, \tag{C.3.21}$$

where III is defined as below,

$$\begin{aligned}
\text{III} :=\ & \frac{1}{\alpha}\Big|\mathbb{P}\left(T_{\boldsymbol{Y}} > q\Big(\frac{\alpha-\zeta_2}{1-\pi(\delta)}; T_{\boldsymbol{Z}}\Big) - 2\zeta_1\right) - \mathbb{P}\left(T_{\boldsymbol{Y}} > q\Big(\frac{\alpha+\zeta_2}{1+\pi(\delta)}; T_{\boldsymbol{Z}}\Big) + 2\zeta_1\right)\Big| \\
=\ & \frac{1}{\alpha}\Big|\mathbb{P}\left(T_{\boldsymbol{Y}} > q\Big(\frac{\alpha-\zeta_2}{1-\pi(\delta)}; T_{\boldsymbol{Z}}\Big) - 2\zeta_1\right) - \mathbb{P}\left(T_{\boldsymbol{Z}} > q\Big(\frac{\alpha-\zeta_2}{1-\pi(\delta)}; T_{\boldsymbol{Z}}\Big) - 2\zeta_1\right) \\
& - \mathbb{P}\left(T_{\boldsymbol{Y}} > q\Big(\frac{\alpha+\zeta_2}{1+\pi(\delta)}; T_{\boldsymbol{Z}}\Big) + 2\zeta_1\right) + \mathbb{P}\left(T_{\boldsymbol{Z}} > q\Big(\frac{\alpha+\zeta_2}{1+\pi(\delta)}; T_{\boldsymbol{Z}}\Big) + 2\zeta_1\right) \\
& + \mathbb{P}\left(T_{\boldsymbol{Z}} > q\Big(\frac{\alpha-\zeta_2}{1-\pi(\delta)}; T_{\boldsymbol{Z}}\Big) - 2\zeta_1\right) - \mathbb{P}\left(T_{\boldsymbol{Z}} > q\Big(\frac{\alpha+\zeta_2}{1+\pi(\delta)}; T_{\boldsymbol{Z}}\Big) + 2\zeta_1\right)\Big| \\
\leq\ & \text{III}_1 + \text{III}_2 + \text{III}_3.
\end{aligned}$$

The last line comes from the triangle inequality, with $\text{III}_1, \text{III}_2, \text{III}_3$ defined as,

$$\begin{aligned}
\text{III}_1 &:= \frac{1}{\alpha}\Big|\mathbb{P}\Big(T_{\boldsymbol{Y}} > q\Big(\frac{\alpha+\zeta_2}{1+\pi(\delta)}; T_{\boldsymbol{Z}}\Big) + 2\zeta_1\Big) - \mathbb{P}\Big(T_{\boldsymbol{Z}} > q\Big(\frac{\alpha+\zeta_2}{1+\pi(\delta)}; T_{\boldsymbol{Z}}\Big) + 2\zeta_1\Big)\Big|, \\
\text{III}_2 &:= \frac{1}{\alpha}\Big|\mathbb{P}\Big(T_{\boldsymbol{Y}} > q\Big(\frac{\alpha-\zeta_2}{1-\pi(\delta)}; T_{\boldsymbol{Z}}\Big) - 2\zeta_1\Big) - \mathbb{P}\Big(T_{\boldsymbol{Z}} > q\Big(\frac{\alpha-\zeta_2}{1-\pi(\delta)}; T_{\boldsymbol{Z}}\Big) - 2\zeta_1\Big)\Big|, \\
\text{III}_3 &:= \frac{1}{\alpha}\Big|\mathbb{P}\Big(T_{\boldsymbol{Z}} > q\Big(\frac{\alpha-\zeta_2}{1-\pi(\delta)}; T_{\boldsymbol{Z}}\Big) - 2\zeta_1\Big) - \mathbb{P}\Big(T_{\boldsymbol{Z}} > q\Big(\frac{\alpha+\zeta_2}{1+\pi(\delta)}; T_{\boldsymbol{Z}}\Big) + 2\zeta_1\Big)\Big|.
\end{aligned}$$

We first bound III_3 by the triangle inequality,

$$\begin{aligned}
\text{III}_3 &= \frac{1}{\alpha}\Big|\mathbb{P}\Big(T_{\boldsymbol{Z}} > q\Big(\frac{\alpha-\zeta_2}{1-\pi(\delta)}; T_{\boldsymbol{Z}}\Big) - 2\zeta_1\Big) - \mathbb{P}\Big(T_{\boldsymbol{Z}} > q\Big(\frac{\alpha+\zeta_2}{1+\pi(\delta)}; T_{\boldsymbol{Z}}\Big) + 2\zeta_1\Big)\Big| \\
&\leq \underbrace{\frac{1}{\alpha}\Big|\mathbb{P}\Big(T_{\boldsymbol{Z}} > q\Big(\frac{\alpha+\zeta_2}{1+\pi(\delta)}; T_{\boldsymbol{Z}}\Big) + 2\zeta_1\Big) - \frac{\alpha+\zeta_2}{1+\pi(\delta)}\Big|}_{\text{III}_{31}} \\
&+ \underbrace{\frac{1}{\alpha}\Big|\mathbb{P}\Big(T_{\boldsymbol{Z}} > q\Big(\frac{\alpha-\zeta_2}{1-\pi(\delta)}; T_{\boldsymbol{Z}}\Big) - 2\zeta_1\Big) - \frac{\alpha-\zeta_2}{1-\pi(\delta)}\Big|}_{\text{III}_{32}} + \underbrace{\frac{1}{\alpha}\Big|\frac{\alpha-\zeta_2}{1-\pi(\delta)} - \frac{\alpha+\zeta_2}{1+\pi(\delta)}\Big|}_{\text{III}_{33}}.
\end{aligned}$$

Note that III_{31} can be rewritten as

$$\mathrm{III}_{31} = \frac{\alpha+\zeta_2}{\alpha(1+\pi(\delta))} \cdot \frac{\left|\mathbb{P}\Big(T_{\boldsymbol{Z}} > q\big(\frac{\alpha+\zeta_2}{1+\pi(\delta)}; T_{\boldsymbol{Z}}\big) + 2\zeta_1\Big) - \mathbb{P}\left(T_{\boldsymbol{Z}} > q\big(\frac{\alpha+\zeta_2}{1+\pi(\delta)}; T_{\boldsymbol{Z}}\big)\right)\right|}{\mathbb{P}\left(T_{\boldsymbol{Z}} > q\big(\frac{\alpha+\zeta_2}{1+\pi(\delta)}; T_{\boldsymbol{Z}}\big)\right)} \tag{C.3.22}$$

$$\leq \frac{\alpha+\zeta_2}{\alpha(1+\pi(\delta))} \cdot K_4\zeta_1\Big(q\big(\frac{\alpha+\zeta_2}{1+\pi(\delta)}; T_{\boldsymbol{Z}}\big) + \zeta_1\Big) \leq C\zeta_1 \log d, \tag{C.3.23}$$

where the first inequality holds by applying a non-uniform anti-concentration bound. Specifically, we apply the part 3 of Theorem 2.1 in Kuchibhotla et al. (2021) (with $r - \epsilon = q\big(\frac{\alpha+\zeta_2}{1+\pi(\delta)}; T_{\boldsymbol{Z}}\big), r + \epsilon = q\big(\frac{\alpha+\zeta_2}{1+\pi(\delta)}; T_{\boldsymbol{Z}}\big) + 2\zeta_1$) to the Gaussian random vector $\boldsymbol{Z}$. Remark that the term K_3 is a constant only depending on $\min_{1\leq j\leq d}\{\sigma^Y_{jj}\}$, $\max_{1\leq j\leq d}\{\sigma^Y_{jj}\}$ and the median of Gaussian maxima (up to 2-nd power, hence at most of rate $O(\log d)$). As for the second inequality, under the assumption $\zeta_2 = O(\alpha_L)$, we have $\frac{\zeta_2}{\alpha} \leq \frac{\zeta_2}{\alpha_L} = O(1)$ when $\alpha \in [\alpha_L, 1]$; we also use the fact that $\zeta_1 = O(\sqrt{\log d})$ (which holds under the stated assumption), and $q\big(\frac{\alpha+\zeta_2}{1+\pi(\delta)}; T_{\boldsymbol{Z}}\big) = O(\sqrt{\log d})$ (which will be verified later in (C.3.27)). Thus we show $\mathrm{III}_{31} = O(\zeta_1 \log d)$. Similarly, III_{32} can be bounded as $O(\zeta_1 \log d)$. As for III_{33}, we have

$$\mathrm{III}_{33} = \frac{1}{\alpha}\Big|\frac{\alpha-\zeta_2}{1-\pi(\delta)} - \frac{\alpha+\zeta_2}{1+\pi(\delta)}\Big| \leq \frac{2\pi(\delta)}{1-\pi^2(\delta)} + \frac{2\zeta_2}{\alpha(1-\pi^2(\delta))}.$$

Thus by combining the bounds on $\mathrm{III}_{31}, \mathrm{III}_{32}, \mathrm{III}_{33}$, we obtain

$$\mathrm{III}_3 \leq \mathrm{III}_{31} + \mathrm{III}_{32} + \mathrm{III}_{33} \leq C'\zeta_1 \log d + \frac{2\pi(\delta)}{1-\pi^2(\delta)} + \frac{2\zeta_2}{\alpha(1-\pi^2(\delta))}. \tag{C.3.24}$$

Regarding the term III_1, we first consider the following,

$$\begin{aligned}
\mathrm{III}_{11} &:= \frac{1}{\alpha}\mathbb{P}\Big(T_{\boldsymbol{Z}} > q\big(\frac{\alpha+\zeta_2}{1+\pi(\delta)}; T_{\boldsymbol{Z}}\big) + 2\zeta_1\Big) \\
&\leq \frac{1}{\alpha}\mathbb{P}\Big(T_{\boldsymbol{Z}} > q\big(\frac{\alpha+\zeta_2}{1+\pi(\delta)}; T_{\boldsymbol{Z}}\big)\Big) \cdot \Big(1 + K_4\zeta_1\Big(q\big(\frac{\alpha+\zeta_2}{1+\pi(\delta)}; T_{\boldsymbol{Z}}\big) + \zeta_1\Big)\Big) \\
&= \frac{\alpha+\zeta_2}{\alpha(1+\pi(\delta))} \cdot \Big(1 + K_4\zeta_1\Big(q\big(\frac{\alpha+\zeta_2}{1+\pi(\delta)}; T_{\boldsymbol{Z}}\big) + \zeta_1\Big)\Big) \\
&\leq C'' + C\zeta_1 \log d = O(1),
\end{aligned}$$

where the first inequality holds due to the derivations from (C.3.22) to (C.3.23), the second inequality holds due to the last inequality in (C.3.23) and the stated assumption $\zeta_2 = O(\alpha_L)$. Then we bound III_1 in terms of III_{11} and write

$$\begin{aligned}
\mathrm{III}_1 &= \frac{1}{\alpha}\Big|\mathbb{P}\Big(T_{\boldsymbol{Y}} > q\Big(\frac{\alpha+\zeta_2}{1+\pi(\delta)}; T_{\boldsymbol{Z}}\Big) + 2\zeta_1\Big) - \mathbb{P}\Big(T_{\boldsymbol{Z}} > q\Big(\frac{\alpha+\zeta_2}{1+\pi(\delta)}; T_{\boldsymbol{Z}}\Big) + 2\zeta_1\Big)\Big| \\
&= \mathrm{III}_{11} \cdot \left|\frac{\mathbb{P}\Big(T_{\boldsymbol{Y}} > q\Big(\frac{\alpha+\zeta_2}{1+\pi(\delta)}; T_{\boldsymbol{Z}}\Big) + 2\zeta_1\Big) - \mathbb{P}\Big(T_{\boldsymbol{Z}} > q\Big(\frac{\alpha+\zeta_2}{1+\pi(\delta)}; T_{\boldsymbol{Z}}\Big) + 2\zeta_1\Big)}{\mathbb{P}\Big(T_{\boldsymbol{Z}} > q\Big(\frac{\alpha+\zeta_2}{1+\pi(\delta)}; T_{\boldsymbol{Z}}\Big) + 2\zeta_1\Big)}\right| \\
&\leq \mathrm{III}_{11} \cdot \frac{(\log d)^{19/6}}{n^{1/6}} = O\Big(\frac{(\log d)^{19/6}}{n^{1/6}}\Big),
\end{aligned}$$

where the inequality holds by applying Corollary 5.1 in Kuchibhotla et al. (2021) again to $T_{\boldsymbol{Y}}$ as the derivations of (C.3.10) in the proof of Proposition C.3.1. The term III_2 can be similarly bounded as III_1. Combining the above bounds on III_1, III_2 and (C.3.24) yields the following bound on III,

$$\mathrm{III} \leq \frac{C(\log d)^{19/6}}{n^{1/6}} + C'\zeta_1 \log d + \frac{2\pi(\delta)}{1-\pi^2(\delta)} + \frac{2\zeta_2}{\alpha(1-\pi^2(\delta))}. \tag{C.3.25}$$

By (C.3.17), (C.3.18), (C.3.21) and (C.3.25), we have, when $\alpha \in [\alpha_L, 1]$,

$$\begin{aligned}\left|\frac{\mathbb{P}(T > q_\xi(\alpha; T^{\mathcal{B}}))}{\mathbb{P}(T_{\boldsymbol{Z}} > q(\alpha; T_{\boldsymbol{Z}}))} - 1\right| &\leq \mathrm{I} + \mathrm{II} \leq \mathrm{I} + \mathrm{III} + \frac{2\mathbb{P}\left(\Delta_\infty > \delta\right)}{\alpha} + \frac{4\zeta_2}{\alpha} \\ &\leq \frac{C(\log d)^{19/6}}{n^{1/6}} + C'\zeta_1 \log d + \frac{C''\zeta_2}{\alpha} + \frac{2\pi(\delta)}{1-\pi^2(\delta)} + \frac{2\mathbb{P}\left(\Delta_\infty > \delta\right)}{\alpha} \\ &\leq \frac{C(\log d)^{19/6}}{n^{1/6}} + C'\zeta_1 \log d + \frac{C''\zeta_2}{\alpha} + \frac{C(\log d)^{11/6}}{n^{1/6}\alpha_L^{1/3}} \\ &= O\left(\frac{(\log d)^{19/6}}{n^{1/6}} + \frac{(\log d)^{11/6}}{n^{1/6}\alpha_L^{1/3}} + \zeta_1 \log d + \frac{\zeta_2}{\alpha_L}\right). \qquad \text{(C.3.26)}\end{aligned}$$

where the third line holds due to the derivations between (C.3.13) and (C.3.16) in the proof of Proposition C.3.1. Remark by the choice of δ and (C.3.16), we have $\pi(\delta) = O(1)$. Also note that $\zeta_2 = O(\alpha_L)$, hence we can show

$$q\big(\frac{\alpha + \zeta_2}{1 + \pi(\delta)}; T_{\boldsymbol{Z}}\big) = O(\sqrt{\log d}). \qquad \text{(C.3.27)}$$

when $\alpha \in [\alpha_L, 1]$. Hence we are able to verify $q\big(\frac{\alpha+\zeta_2}{1+\pi(\delta)}; T_{\boldsymbol{Z}}\big) = O(\sqrt{\log d})$, as promised when deriving (C.3.23). Denoting the bound in (C.3.26) by $\eta(d, n, \zeta_1, \zeta_2, \alpha_L)$, we finally establish (C.3.6), i.e.,

$$\sup_{\alpha \in [\alpha_L, 1]} \left|\frac{\mathbb{P}(T > q_\xi(\alpha; T^{\mathcal{B}}))}{\mathbb{P}(T_{\boldsymbol{Z}} > q(\alpha; T_{\boldsymbol{Z}}))} - 1\right| = \eta(d, n, \zeta_1, \zeta_2, \alpha_L).$$

□

C.4 Validity and power analysis of single node testing

In this section, we focus on Lemma C.1.1 and Lemma C.3.4. Note that these results are established using the same strategies as Theorem 4.1, Lemma S.1 and Theorem S.7 in Lu et al. (2017). We still present their proofs for completeness.

C.4.1 Proof of Lemma C.1.1

Proof. For given node j, we denote $N_{0j} = \{(j,k) : \mathbf{\Theta}_{jk} = 0\}$, then $N_{0j}^c = \{(j,k) : |\mathbf{\Theta}_{jk}| > 0\}$. First we consider the following event,

$$\mathcal{E} = \Big\{ \min_{e \in N_{0j}^c} \sqrt{n} |\widetilde{\mathbf{\Theta}}_e^{\mathrm{d}}| > \widehat{c}(\alpha, E_0) \Big\}, \quad \text{where } E_0 = \{(j,k) : k \neq j, k \in [d]\}.$$

By the definition of Algorithm 7, we immediately have the rejected edge set in the first iteration can be written as

$$E_1 = \{(j,k) \in E_0 : \sqrt{n} |\widetilde{\mathbf{\Theta}}_{jk}^{\mathrm{d}}| > \widehat{c}(\alpha, E_0)\}.$$

Regarding (i) i.e., under the alternative hypothesis $H_{1j} : \|\mathbf{\Theta}_{j,-j}\|_0 \geq k_\tau$, we first note $\psi_{j,\alpha} = 1$ on the event $\mathcal{E}$. Also notice that $N_{0j}^c \subseteq E_1$ given $\mathcal{E}$. Then the following bound immediately follows:

$$\mathbb{P}\left(\psi_{j,\alpha} = 1\right) \geq \mathbb{P}\left(\mathcal{E}\right). \tag{C.4.1}$$

We further derive a lower bound for $\mathbb{P}\left(\mathcal{E}\right)$ by the triangle inequality:

$$\mathbb{P}\left(\mathcal{E}\right) \geq \mathbb{P}\left(\min_{e \in N_{0j}^c} |\mathbf{\Theta}_e^\star| > \frac{\widehat{c}(\alpha, E_0)}{\sqrt{n}} + C_0 \sqrt{\frac{\log d}{n}} \text{ and } ||\widetilde{\mathbf{\Theta}}^{\mathrm{d}} - \mathbf{\Theta}^\star||_{\max} \leq C_0 \sqrt{\frac{\log d}{n}} \right). \tag{C.4.2}$$

For any fixed $\alpha \in (0,1)$, we consider sufficiently large d such that $1/d \leq \alpha$. By applying Lemma C.4.1, we have

$$\mathbb{P}\left(\mathbb{P}_\xi(T_{E_0}^{\mathcal{B}} \geq C_0 \sqrt{\log d} \mid \{\boldsymbol{X}_i\}_{i=1}^n) \leq 1/d \right) \geq 1 - 1/d^2, \tag{C.4.3}$$

where $E_0 = \{(j,k) : k \neq j, k \in [d]\}$. Recall the definition of $\widehat{c}(\alpha, E)$

$$\widehat{c}(\alpha, E) = \inf\left\{t \in \mathbb{R} : \mathbb{P}_\xi\left(T_E^{\mathcal{B}} \leq t\right) \geq 1 - \alpha\right\}.$$

We then have $\widehat{c}(\alpha, E_0) \leq C_0\sqrt{\log d}$ for some constant $C_0 > 0$, with probability greater than $1 - 1/d^2$. Choosing the constant in the signal strength condition of Lemma C.1.1 to be $2C_0$ (i.e., for any $(j,k) \in N_{0j}^c$, $|\mathbf{\Theta}_{jk}| \geq 2C_0\sqrt{\log d/n}$) and applying (C.4.37), we have with probability greater than $1 - 1/d^2$

$$\min_{e \in N_{0j}^c} |\mathbf{\Theta}_e^\star| \geq 2C_0\sqrt{\frac{\log d}{n}} \geq \frac{\widehat{c}(\alpha, E_0)}{\sqrt{n}} + C_0\sqrt{\frac{\log d}{n}} \text{ and}$$
$$\mathbb{P}\left(||\widetilde{\mathbf{\Theta}}^{\mathrm{d}} - \mathbf{\Theta}^\star||_{\max} \leq C_0\sqrt{\frac{\log d}{n}}\right) \geq 1 - 2/d^2.$$

Combining the above two inequalities with the earlier bounds (C.4.1) and (C.4.2), we obtain a lower bound for $\mathbb{P}(\psi_{j,\alpha} = 1)$ i.e., $\mathbb{P}(\psi_{j,\alpha} = 1) \geq \mathbb{P}(\mathcal{E}) > 1 - 3/d^2$. Therefore, we establish

$$\lim_{(n,d)\to\infty} \mathbb{P}(\psi_{j,\alpha} = 1) = 1.$$

Now we consider (ii), i.e., the case when $\|\mathbf{\Theta}_{j,-j}\|_0 < k_\tau$. Since $\|\mathbf{\Theta}_{j,-j}\|_0 \leq k_\tau - 1$, $\psi_{j,\alpha} = 1$ implies at least one edge in N_{0j} is rejected in Algorithm 7. Suppose the first rejected edge in N_{0j} is (j, k_*) and it is rejected at the t_*-th iteration. Then we have $N_{0j} \subseteq E_{t_*-1}$ and

$$\max_{e \in N_{0j}} \sqrt{n}|\widetilde{\mathbf{\Theta}}_e^{\mathrm{d}} - \mathbf{\Theta}_e^\star| \geq \sqrt{n}|\widetilde{\mathbf{\Theta}}_{jk_*}^{\mathrm{d}} - \mathbf{\Theta}_{jk_*}^\star| \geq \widehat{c}(\alpha, E_{t_*-1}) \geq \widehat{c}(\alpha, N_{0j}), \tag{C.4.4}$$

where the first inequality holds since $(j, k_*) \subset N_{0j}$, the second inequality holds since $\mathbf{\Theta}_{jk_*}^\star = 0$ and the edge (j, k_*) is rejected at the t_*-th iteration. The last inequality holds simply because $N_{0j} \subseteq$

E_{t_*-1}. Therefore by applying Lemma 3.2.1 with E chosen to be N_{0j}, we have

$$\lim_{(n,d)\to\infty} \mathbb{P}\left(\psi_{j,\alpha} = 1\right) \leq \alpha.$$

□

Lemma C.4.1. *Under the same conditions as Lemma 3.2.1, we have for any $j \in [d]$,*

$$\mathbb{P}\left(\mathbb{P}_\xi(T^{\mathcal{B}}_{E_{0j}} \geq C_0\sqrt{\log d} \mid \{\boldsymbol{X}_i\}_{i=1}^n) \leq 1/d\right) \geq 1 - 1/d^2 \tag{C.4.5}$$

holds for some constant $C_0 > 0$, where $E_{0j} := \{(j,k) : k \neq j, k \in [d]\}$.

Proof of Lemma C.4.1. Recall the definition of $T^{\mathcal{B}}_E$

$$T^{\mathcal{B}}_E := \max_{(j,k)\in E} \frac{1}{\sqrt{n\,\widehat{\boldsymbol{\Theta}}_{jj}\widehat{\boldsymbol{\Theta}}_{kk}}} \left| \sum_{i=1}^n \widehat{\boldsymbol{\Theta}}_j^\top \left(\boldsymbol{X}_i\boldsymbol{X}_i^\top \widehat{\boldsymbol{\Theta}}_k - \mathbf{e}_k\right)\xi_i \right|.$$

First we notice the following,

$$\begin{aligned}
&\mathbb{P}_\xi\big(T^{\mathcal{B}}_{E_{0j}} \geq C_0\sqrt{\log d} \mid \{\boldsymbol{X}_i\}_{i=1}^n\big) \\
=\ &\mathbb{P}_\xi\Big(\max_{(j,k)\in E_{0j}} \frac{1}{\sqrt{n\,\widehat{\boldsymbol{\Theta}}_{jj}\widehat{\boldsymbol{\Theta}}_{kk}}} \left| \sum_{i=1}^n \widehat{\boldsymbol{\Theta}}_j^\top \left(\boldsymbol{X}_i\boldsymbol{X}_i^\top \widehat{\boldsymbol{\Theta}}_k - \mathbf{e}_k\right)\xi_i \right| \geq C_0\sqrt{\log d} \mid \{\boldsymbol{X}_i\}_{i=1}^n\Big) \\
\leq\ &\sum_{(j,k)\in E_{0j}} \mathbb{P}_\xi\Big(\frac{1}{\sqrt{n\,\widehat{\boldsymbol{\Theta}}_{jj}\widehat{\boldsymbol{\Theta}}_{kk}}} \left| \sum_{i=1}^n \widehat{\boldsymbol{\Theta}}_j^\top \left(\boldsymbol{X}_i\boldsymbol{X}_i^\top \widehat{\boldsymbol{\Theta}}_k - \mathbf{e}_k\right)\xi_i \right| \geq C_0\sqrt{\log d} \mid \{\boldsymbol{X}_i\}_{i=1}^n\Big).
\end{aligned} \tag{C.4.6}$$

where the equality holds by the definition of $T^{\mathcal{B}}_{E_{0j}}$ and the inequality holds by the union bound. In the following, we will bound (C.4.6) for each $(j,k) \in E_{0j}$. Note that conditioning on $\{\boldsymbol{X}_i\}_{i=1}^n$,

the following random variable is a mean zero Gaussian random variable

$$G_{jk} := \frac{1}{\sqrt{n\,\widehat{\boldsymbol{\Theta}}_{jj}\widehat{\boldsymbol{\Theta}}_{kk}}}\Big|\sum_{i=1}^{n}\widehat{\boldsymbol{\Theta}}_j^\top\left(\boldsymbol{X}_i\boldsymbol{X}_i^\top\widehat{\boldsymbol{\Theta}}_k-\mathbf{e}_k\right)\xi_i.$$

Hence we will bound its conditional variance then apply the sub-Gaussian tail probability bound (in Section 2.1.2 of Wainwright (2019)). Specifically, we have with probability greater than $1-1/d^2$,

$$\begin{aligned}
\mathrm{Var}(G_{jk}\mid\{\boldsymbol{X}_i\}_{i=1}^n) &= \frac{\sum_{i=1}^n\left[\widehat{\boldsymbol{\Theta}}_j^\top(\boldsymbol{X}_i\boldsymbol{X}_i^\top\widehat{\boldsymbol{\Theta}}_k-\mathbf{e}_k)\right]^2}{n\,\widehat{\boldsymbol{\Theta}}_{jj}\widehat{\boldsymbol{\Theta}}_{kk}}\\
&= \frac{\sum_{i=1}^n\left[\widehat{\boldsymbol{\Theta}}_j^\top(\boldsymbol{X}_i\boldsymbol{X}_i^\top\widehat{\boldsymbol{\Theta}}_k-\mathbf{e}_k)\right]^2}{n\,\boldsymbol{\Theta}_{jj}\boldsymbol{\Theta}_{kk}}\left(\frac{\boldsymbol{\Theta}_{jj}\boldsymbol{\Theta}_{kk}-\widehat{\boldsymbol{\Theta}}_{jj}\widehat{\boldsymbol{\Theta}}_{kk}}{\widehat{\boldsymbol{\Theta}}_{jj}\widehat{\boldsymbol{\Theta}}_{kk}}+1\right)\\
&\le C'\frac{\sum_{i=1}^n\left[\widehat{\boldsymbol{\Theta}}_j^\top(\boldsymbol{X}_i\boldsymbol{X}_i^\top\widehat{\boldsymbol{\Theta}}_k-\mathbf{e}_k)\right]^2}{n\,\boldsymbol{\Theta}_{jj}\boldsymbol{\Theta}_{kk}}\\
&\le 2C'\sum_{i=1}^n\frac{\left[\boldsymbol{\Theta}_j^\top(\boldsymbol{X}_i\boldsymbol{X}_i^\top\boldsymbol{\Theta}_k-\mathbf{e}_k)\right]^2+\left[\widehat{\boldsymbol{\Theta}}_j^\top(\boldsymbol{X}_i\boldsymbol{X}_i^\top\widehat{\boldsymbol{\Theta}}_k-\mathbf{e}_k)-\boldsymbol{\Theta}_j^\top(\boldsymbol{X}_i\boldsymbol{X}_i^\top\boldsymbol{\Theta}_k-\mathbf{e}_k)\right]^2}{n\,\boldsymbol{\Theta}_{jj}\boldsymbol{\Theta}_{kk}}\\
&\le \frac{2C'}{\boldsymbol{\Theta}_{jj}\boldsymbol{\Theta}_{kk}}\Big(\left[M^2\max_i\|\boldsymbol{X}_i\boldsymbol{X}_i^\top-\boldsymbol{\Sigma}\|_{\max}\right]^2+\left[2M\|\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}\|_1\max_i\|\boldsymbol{X}_i\boldsymbol{X}_i^\top-\boldsymbol{\Sigma}\|_{\max}\right]^2\Big) \qquad \text{(C.4.7)}\\
&\le \frac{2C'}{\boldsymbol{\Theta}_{jj}\boldsymbol{\Theta}_{kk}}\left[\left(CM^2\sqrt{\log(dn)}\right)^2+\left(2MC\log(dn)\sqrt{\frac{s^2\log d}{n}}\right)^2\right] \qquad \text{(C.4.8)}\\
&\le 2C'r_0^2\left[\left(CM^2\sqrt{\log(dn)}\right)^2+\left(2MC\log(dn)\sqrt{\frac{s^2\log d}{n}}\right)^2\right]\\
&\le C''\log d
\end{aligned}$$

for some constant $C''>0$, where the first inequality holds by (C.4.30) and the fact $\min_{(j,k)\in E}\sqrt{\boldsymbol{\Theta}_{jj}\boldsymbol{\Theta}_{kk}}\ge\lambda_{\min}(\boldsymbol{\Theta})\ge 1/r_0$ (as $\boldsymbol{\Theta}\in\mathcal{U}(M,s,r_0)$), the second inequality holds by the triangle inequality, (C.4.7) and (C.4.8) holds due to (C.4.31), (C.4.32), (C.4.33) and (C.4.11), and the last two inequal-

ities hold due to the fact $\min_{(j,k)\in E}\sqrt{\boldsymbol{\Theta}_{jj}\boldsymbol{\Theta}_{kk}} \geq 1/r_0$ and the conditions of Lemma 3.2.1. Therefore, we have with probability greater than $1 - 1/d^2$,

$$\mathbb{P}_\xi\left(G_{jk} \geq C_0\sqrt{\log d} \mid \{\boldsymbol{X}_i\}_{i=1}^n\right) \leq \frac{1}{d^2}$$

for some constant $C_0 > 0$ by the sub-Gaussian tail probability bound (in Section 2.1.2 of Wainwright (2019)). Combining the above bound with (C.4.6), we have

$$\mathbb{P}\left(\mathbb{P}_\xi\big(T^{\mathcal{B}}_{E_{0j}} \geq C_0\sqrt{\log d} \mid \{\boldsymbol{X}_i\}_{i=1}^n\big) \leq d\cdot\frac{1}{d^2}\right) \geq 1 - 1/d^2$$

since $|E_{0j}| \leq d$. The above derivations hold for any $j \in [d]$, thus (C.4.5) is established. □

C.4.2 Proof of Lemma 3.2.1

We first recall the definition of $\mathcal{U}(M, s, r_0)$ and write down the statement of Lemma 3.2.1 below.

$$\mathcal{U}(M,s,r_0) = \Big\{\boldsymbol{\Theta} \in \mathbb{R}^{d\times d} \,\Big|\, \lambda_{\min}(\boldsymbol{\Theta}) \geq 1/r_0, \lambda_{\max}(\boldsymbol{\Theta}) \leq r_0, \max_{j\in[d]}\|\boldsymbol{\Theta}_j\|_0 \leq s, \|\boldsymbol{\Theta}\|_1 \leq M\Big\}. \tag{C.4.9}$$

Lemma C.4.2. *Suppose that $\boldsymbol{\Theta} \in \mathcal{U}(M, s, r_0)$. If $(\log(dn))^7/n + s^2(\log dn)^4/n = o(1)$, for any edge set $E \subseteq \mathcal{V}\times\mathcal{V}$, we have for any $\alpha \in [0,1]$,*

$$\lim_{(n,d)\to\infty}\sup_{\boldsymbol{\Theta}\in\mathcal{U}(M,s,r_0)}\sup_{\alpha\in(0,1)}\left|\mathbb{P}\left(\max_{e\in E}\sqrt{n}|\widetilde{\boldsymbol{\Theta}}^d_e - \boldsymbol{\Theta}^\star_e| > \widehat{c}(\alpha, E)\right) - \alpha\right| = 0. \tag{C.4.10}$$

Throughout the following parts, we will write the standardized one-step estimator explicitly:

$$\widehat{\boldsymbol{\Theta}}^{\mathrm{d}}_{jk}/\sqrt{\widehat{\boldsymbol{\Theta}}^{\mathrm{d}}_{jj}\widehat{\boldsymbol{\Theta}}^{\mathrm{d}}_{kk}}, \quad \text{where } \widehat{\boldsymbol{\Theta}}^{\mathrm{d}}_{jk} := \widehat{\boldsymbol{\Theta}}_{jk} - \frac{\widehat{\boldsymbol{\Theta}}_j^\top\left(\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\Theta}}_k - \mathbf{e}_k\right)}{\widehat{\boldsymbol{\Theta}}_j^\top\widehat{\boldsymbol{\Sigma}}_j}.$$

In order to prove (C.4.10), we need preliminary results on the estimation rates of CLIME estimator. Cai et al. (2011) gives the following theorem. We can also prove the same result for the GLasso estimator (Janková & van de Geer, 2018). Therefore, Lemma C.4.2 applies for both the CLIME estimator and the GLasso estimator. This also implies that the results in our paper apply to both the CLIME estimator and the GLasso estimator.

Lemma C.4.3. *Suppose* $\boldsymbol{\Theta} \in \mathcal{U}(M, s, r_0)$ *and we choose the tuning parameter* $\lambda \geq CM\sqrt{\log d/n}$ *in the CLIME estimator. With probability greater than* $1 - c/d^2$*, we have the following bounds:*

$$\begin{aligned} &||\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}||_{\max} \leq C\sqrt{\frac{\log d}{n}}, ||\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\Sigma}} - \mathbf{I}||_{\max} \leq CM\sqrt{\frac{\log d}{n}}, \text{ and} \\ &||\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}||_{\max} \leq CM\sqrt{\frac{\log d}{n}}, ||\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}||_1 \leq CM\sqrt{\frac{s^2 \log d}{n}}, \end{aligned} \tag{C.4.11}$$

where C *is a universal constant only depending on* r_0 *in* (C.4.9)*.*

Remark C.4.3.1. *Note the first inequality in* (C.4.11) *directly follows from Equation (26) in Cai et al. (2011), the second inequality follows from the constraint in the CLIME estimator and the third inequality holds due to Theorem 6 in Cai et al. (2011).*

Given a random variable Z, we define its ψ_ℓ-norm for $\ell \geq 1$ as $\|Z\|_{\psi_\ell} = \sup_{p \geq 1} p^{-1/\ell}(\mathbb{E}|Z|^p)^{1/p}$. The following lemma controls the ψ_ℓ-norm of $\boldsymbol{X}$ and gives the lower bound of the variance of the debiased estimator.

Lemma C.4.4. *There exist universal constants* c *and* C *only depending on* r_0 *in* (C.4.9) *such that*

$$\sup_{\|\mathbf{v}\|_2 = 1} \|\mathbf{v}^\top \boldsymbol{\Sigma}^{-1/2} \boldsymbol{X}\|_{\psi_2} \leq C \text{ and } \min_{j,k \in [d]} \mathbb{E}[(\boldsymbol{\Theta}_j^\top (\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{\Sigma})\boldsymbol{\Theta}_k)^2] \geq c. \tag{C.4.12}$$

Proof. The first inequality in (C.4.12) immediately follows since $\mathbf{v}^\top \boldsymbol{\Sigma}^{-1/2} \boldsymbol{X} \sim N(0, 1)$ for any $\|\mathbf{v}\|_2 = 1$. Regarding the second inequality, note that $\mathbb{E}[(\boldsymbol{\Theta}_j^\top (\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{\Sigma})\boldsymbol{\Theta}_k)^2] =$

$\mathrm{Var}\left(\boldsymbol{\Theta}_j^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{\Theta}_k\right)$. Below we calculate the expression of the general form $\mathrm{Var}\left(\mathbf{u}^\top \boldsymbol{X}\boldsymbol{X}^\top \mathbf{v}\right)$. Specifically, we apply Isserlis' theorem (Isserlis, 1918) to deal with the moments of Gaussian random variables. For any deterministic vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$, Isserlis' theorem says

$$\begin{aligned}
\mathrm{Var}(\mathbf{u}^\top \boldsymbol{X}\boldsymbol{X}^\top \mathbf{v}) &= \mathbb{E}[(\mathbf{u}^\top \boldsymbol{X})^2(\mathbf{v}^\top \boldsymbol{X})^2] - (\mathbb{E}[\mathbf{u}^\top \boldsymbol{X}\mathbf{v}^\top \boldsymbol{X}])^2 \\
&= \mathbb{E}[(\mathbf{u}^\top \boldsymbol{X})^2]\mathbb{E}[(\mathbf{v}^\top \boldsymbol{X})^2] + (\mathbb{E}[\mathbf{u}^\top \boldsymbol{X}\mathbf{v}^\top \boldsymbol{X}])^2 \\
&= (\mathbf{u}^\top \boldsymbol{\Sigma}\mathbf{u}^\top)(\mathbf{v}^\top \boldsymbol{\Sigma}\mathbf{v}^\top) + (\mathbf{u}^\top \boldsymbol{\Sigma}\mathbf{v}^\top)^2.
\end{aligned}$$

Therefore, we obtain the following,

$$\mathbb{E}[(\boldsymbol{\Theta}_j^\top(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{\Sigma})\boldsymbol{\Theta}_k)^2] = (\boldsymbol{\Theta}_j^\top \boldsymbol{\Sigma}\boldsymbol{\Theta}_j^\top)(\boldsymbol{\Theta}_k^\top \boldsymbol{\Sigma}\boldsymbol{\Theta}_k^\top) + (\boldsymbol{\Theta}_j^\top \boldsymbol{\Sigma}\boldsymbol{\Theta}_k^\top)^2 = \boldsymbol{\Theta}_{jj}\boldsymbol{\Theta}_{kk} + \boldsymbol{\Theta}_{jk}^2 \geq 1/r_0^2,$$

where the last inequality holds since $\lambda_{\min}(\boldsymbol{\Theta}) \geq 1/r_0$ when $\boldsymbol{\Theta} \in \mathcal{U}(M, s, r_0)$. □

Now we are ready to prove Lemma 3.2.1. Note the proof of this lemma follows a similar idea as the one used in Proposition 3.1 of Neykov et al. (2019). Since Lemma 3.2.1 involves the standardized version of the one-step estimator in Neykov et al. (2019), we still present the detailed proof for completeness.

Proof of Lemma 3.2.1. To approximate

$$T_E := \max_{(j,k)\in E} \sqrt{n}\left|(\widehat{\boldsymbol{\Theta}}_{jk}^{\mathrm{d}}/\sqrt{\widehat{\boldsymbol{\Theta}}_{jj}^{\mathrm{d}}\widehat{\boldsymbol{\Theta}}_{kk}^{\mathrm{d}}} - \boldsymbol{\Theta}_{jk}/\sqrt{\boldsymbol{\Theta}_{jj}\boldsymbol{\Theta}_{jk}})\right|, \tag{C.4.13}$$

by the multiplier bootstrap process

$$T_E^{\mathcal{B}} := \max_{(j,k)\in E} \frac{1}{\sqrt{n\,\widehat{\boldsymbol{\Theta}}_{jj}\widehat{\boldsymbol{\Theta}}_{kk}}}\left|\sum_{i=1}^n \widehat{\boldsymbol{\Theta}}_j^\top(\boldsymbol{X}_i\boldsymbol{X}_i^\top\widehat{\boldsymbol{\Theta}}_k - \mathbf{e}_k)\xi_i\right|, \tag{C.4.14}$$

we define two intermediate processes

$$\breve{T}_E := \max_{(j,k)\in E} \left| \frac{1}{\sqrt{n\,\boldsymbol{\Theta}_{jj}\boldsymbol{\Theta}_{kk}}} \sum_{i=1}^{n} \boldsymbol{\Theta}_j^\top (\boldsymbol{X}_i \boldsymbol{X}_i^\top \boldsymbol{\Theta}_k - \mathbf{e}_k) \right|, \tag{C.4.15}$$

$$\breve{T}_E^{\mathcal{B}} := \max_{(j,k)\in E} \left| \frac{1}{\sqrt{n\,\boldsymbol{\Theta}_{jj}\boldsymbol{\Theta}_{kk}}} \sum_{i=1}^{n} \boldsymbol{\Theta}_j^\top (\boldsymbol{X}_i \boldsymbol{X}_i^\top \boldsymbol{\Theta}_k - \mathbf{e}_k)\xi_i \right|. \tag{C.4.16}$$

The strategy of proving this lemma is to verify the three conditions in Corollary 3.1 of Chernozhukov et al. (2013):

(a) $\min_{j,k} \mathbb{E}[(\boldsymbol{\Theta}_j^\top (\boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{\Theta}_k - \mathbf{e}_k))^2] > c$ and $\max_{j,k\in[d]} ||\boldsymbol{\Theta}_j^\top (\boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{\Theta}_k - \mathbf{e}_k)||_{\psi_1} \le C$ for some positive constants c and C;

(b) $\mathbb{P}(|T_E - \breve{T}_E| > \zeta_1) < \zeta_2$ holds for some $\zeta_1, \zeta_2 > 0$;

(c) And $\mathbb{P}(\mathbb{P}_\xi(|T_E^{\mathcal{B}} - \breve{T}_E^{\mathcal{B}}| > \zeta_1 \mid \{\boldsymbol{X}_i\}_{i=1}^n) > \zeta_2) < \zeta_2$ holds for $\zeta_1\sqrt{\log d} + \zeta_2 = o(1)$.

Notice that in Chernozhukov et al. (2013), the original conditions require the last scaling to be $\zeta_1\sqrt{\log d} + \zeta_2 = o(n^{-c_1})$ for some c_1. This is because they pursue a stronger result that $|\mathbb{P}(T_E > \widehat{c}(\alpha, E)) - \alpha| = O(n^{-c_1})$. Since we do not emphasize on the polynomial decaying in our result, we only require $\zeta_1\sqrt{\log d} + \zeta_2 = o(1)$.

We start by checking the first condition (a). Lemma C.4.4 immediately implies the first part. By the second condition in (C.4.12), we have $\|\boldsymbol{X}_j\boldsymbol{X}_k - \mathbb{E}[\boldsymbol{X}_j\boldsymbol{X}_k]\|_{\psi_1} \le C$. By the definition of the ψ-norms, we have

$$\begin{aligned}\max_{j,k\in[d]} \|\boldsymbol{\Theta}_j^\top (\boldsymbol{X}_i \boldsymbol{X}_i^\top \boldsymbol{\Theta}_k - \mathbf{e}_k)\|_{\psi_1} &\le r_0^2 \|(\boldsymbol{X}_j\boldsymbol{X}_k - \mathbb{E}[\boldsymbol{X}_j\boldsymbol{X}_k])\|_{\psi_1} \\ &\le r_0^2 \sup_{\|\mathbf{v}\|_2=1} \|\mathbf{v}^\top \boldsymbol{X}\boldsymbol{X}^\top \mathbf{v} - \mathbb{E}[\mathbf{v}^\top \boldsymbol{X}\boldsymbol{X}^\top \mathbf{v}]\|_{\psi_1} = O(1).\end{aligned}$$

Regarding the condition (b), we check by bounding the difference $|T_E - \breve{T}_E|$. Recall the one-step

estimator

$$\widehat{\boldsymbol{\Theta}}^{\mathrm{d}}_{jk} = \widehat{\boldsymbol{\Theta}}_{jk} - \frac{\widehat{\boldsymbol{\Theta}}_j^\top \left(\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\Theta}}_k - \mathbf{e}_k\right)}{\widehat{\boldsymbol{\Theta}}_j^\top \widehat{\boldsymbol{\Sigma}}_j},$$

and plug it into T_E. Then we have the following bound,

$$\begin{aligned} |T_E - \breve{T}_E| &= \left| \max_{(j,k)\in E} \sqrt{n} \cdot \left| \frac{\widehat{\boldsymbol{\Theta}}^{\mathrm{d}}_{jk}}{\sqrt{\widehat{\boldsymbol{\Theta}}^{\mathrm{d}}_{jj}\widehat{\boldsymbol{\Theta}}^{\mathrm{d}}_{kk}}} - \frac{\boldsymbol{\Theta}_{jk}}{\sqrt{\boldsymbol{\Theta}_{jj}\boldsymbol{\Theta}_{kk}}} \right| - \max_{(j,k)\in E} \frac{\sqrt{n}}{\sqrt{\boldsymbol{\Theta}_{jj}\boldsymbol{\Theta}_{kk}}} \left| \boldsymbol{\Theta}_j^\top (\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Theta}_k - \mathbf{e}_k) \right| \right| \\ &\leq \frac{\mathrm{I}_1 \mathrm{I}_2}{\min_{(j,k)\in E} \sqrt{\boldsymbol{\Theta}_{jj}\boldsymbol{\Theta}_{kk}}} + \frac{\mathrm{I}_3}{\min_{(j,k)\in E} \sqrt{\widehat{\boldsymbol{\Theta}}^{\mathrm{d}}_{jj}\widehat{\boldsymbol{\Theta}}^{\mathrm{d}}_{kk}}}, \end{aligned} \tag{C.4.17}$$

where $\mathrm{I}_1 = \max_{(j,k)\in E} |\widehat{\boldsymbol{\Theta}}^{\mathrm{d}}_{jj}\widehat{\boldsymbol{\Theta}}^{\mathrm{d}}_{kk} - \boldsymbol{\Theta}_{jj}\boldsymbol{\Theta}_{kk}|$, $\mathrm{I}_2 = \max_{(j,k)\in E} |\sqrt{n} \cdot \boldsymbol{\Theta}_j^\top \left(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Theta}_k - \mathbf{e}_k\right)|$ and

$$\mathrm{I}_3 = \max_{(j,k)\in E} \left| \sqrt{n}(\widehat{\boldsymbol{\Theta}}^{\mathrm{d}}_{jk} - \boldsymbol{\Theta}_{jk}) - \sqrt{n} \cdot \boldsymbol{\Theta}_j^\top \left(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Theta}_k - \mathbf{e}_k\right) \right|.$$

Note I_1 can be bounded using Lemma C.4.5, i.e.,

$$\mathrm{I}_1 = \max_{(j,k)\in E} |\widehat{\boldsymbol{\Theta}}^{\mathrm{d}}_{jj}\widehat{\boldsymbol{\Theta}}^{\mathrm{d}}_{kk} - \boldsymbol{\Theta}_{jj}\boldsymbol{\Theta}_{kk}| \leq 2M \left\| \widehat{\boldsymbol{\Theta}}^{\mathrm{d}} - \boldsymbol{\Theta} \right\|_{\max} \leq CM^2 \sqrt{\frac{\log d}{n}}, \tag{C.4.18}$$

with probability $1 - 1/d^2$. As for the term I_2, we have

$$\begin{aligned} \mathrm{I}_2 = \max_{(j,k)\in E} \left| \sqrt{n} \boldsymbol{\Theta}_j^\top \left(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Theta}_k - \mathbf{e}_k \right| &= \max_{(j,k)\in E} \sqrt{n} \left| \boldsymbol{\Theta}_j^\top \left(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right) \boldsymbol{\Theta}_k \right| \\ &\leq \sqrt{n} M^2 \left\| \widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma} \right\|_{\max} \leq CM^2 \sqrt{\log d}. \end{aligned} \tag{C.4.19}$$

Denote $\breve{\boldsymbol{\Theta}}_k = (\widehat{\boldsymbol{\Theta}}_{k1}, \ldots, \widehat{\boldsymbol{\Theta}}_{k(j-1)}, \boldsymbol{\Theta}_{kj}, \widehat{\boldsymbol{\Theta}}_{k(j+1)}, \ldots, \widehat{\boldsymbol{\Theta}}_{kd})^\top \in \mathbb{R}^d$. To deal with the term I_3, we first rewrite the following

$$\sqrt{n}(\widehat{\boldsymbol{\Theta}}^{\mathrm{d}}_{jk} - \boldsymbol{\Theta}_{jk}) = -\sqrt{n} \cdot \frac{\widehat{\boldsymbol{\Theta}}_j^\top \left(\widehat{\boldsymbol{\Sigma}}\breve{\boldsymbol{\Theta}}_k - \mathbf{e}_k\right)}{\widehat{\boldsymbol{\Theta}}_j^\top \widehat{\boldsymbol{\Sigma}}_j}, \tag{C.4.20}$$

then quantify $\sqrt{n}\widehat{\boldsymbol{\Theta}}_j^\top\big(\widehat{\boldsymbol{\Sigma}}\check{\boldsymbol{\Theta}}_k - \mathbf{e}_k^\top\big)$. Notice that

$$\sqrt{n}\cdot\widehat{\boldsymbol{\Theta}}_j^\top\big(\widehat{\boldsymbol{\Sigma}}\check{\boldsymbol{\Theta}}_k - \mathbf{e}_k^\top\big) = \underbrace{\sqrt{n}\cdot\widehat{\boldsymbol{\Theta}}_j^\top\big(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Theta}_k - \mathbf{e}_k^\top\big)}_{\mathrm{II}_1} + \underbrace{\sqrt{n}\cdot\widehat{\boldsymbol{\Theta}}_j^\top\widehat{\boldsymbol{\Sigma}}\big(\check{\boldsymbol{\Theta}}_k - \boldsymbol{\Theta}_k\big)}_{\mathrm{II}_2}. \tag{C.4.21}$$

Further we expand II_1 as

$$\mathrm{II}_1 = \underbrace{\sqrt{n}\cdot\boldsymbol{\Theta}_j^\top\big(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Theta}_k - \mathbf{e}_k\big)}_{\mathrm{II}_{11}} + \underbrace{\sqrt{n}\cdot\big(\widehat{\boldsymbol{\Theta}}_j^\top - \boldsymbol{\Theta}_j^\top\big)\big(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Theta}_k - \mathbf{e}_k\big)}_{\mathrm{II}_{12}}, \tag{C.4.22}$$

where II_{11} can be rewritten as $\mathrm{II}_{11} = \frac{1}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{\Theta}_j^\top(\boldsymbol{X}_i\boldsymbol{X}_i^\top\boldsymbol{\Theta}_k - \mathbf{e}_k)$. We bound $|\mathrm{II}_{12}|$ as

$$|\mathrm{II}_{12}| = \sqrt{n}\cdot\big(\widehat{\boldsymbol{\Theta}}_j - \boldsymbol{\Theta}_j\big)^\top\big(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\big)\boldsymbol{\Theta}_k \le \sqrt{n}\cdot\big\|\widehat{\boldsymbol{\Theta}}_j - \boldsymbol{\Theta}_j\big\|_1\big\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\big\|_{\max}\|\boldsymbol{\Theta}_k\|_1. \tag{C.4.23}$$

According to Lemma C.4.3, (C.4.23) yields that

$$\max_{j,k\in[d]}|\mathrm{II}_{12}| \lesssim M^2\frac{s\log d}{\sqrt{n}}, \tag{C.4.24}$$

with probability $1 - 1/d^2$. By Hölder's inequality and Lemma C.4.3, we finally obtain the bound on II_2:

$$\max_{j,k\in[d]}|\mathrm{II}_2| \le \sqrt{n}\cdot\max_{j,k\in[d]}\|\widehat{\boldsymbol{\Theta}}_j^\top\widehat{\boldsymbol{\Sigma}}_{-j}\|_\infty\big\|\widehat{\boldsymbol{\Theta}}_k - \boldsymbol{\Theta}_k\big\|_1 \lesssim M^2\frac{s\log d}{\sqrt{n}}, \tag{C.4.25}$$

with probability $1 - 1/d^2$. Therefore, we conclude that by (C.4.24) and (C.4.25), with probability $1 - 1/d^2$, the following holds:

$$\max_{j,k\in[d]}\sqrt{n}\cdot\Big|\widehat{\boldsymbol{\Theta}}_j^\top\big(\widehat{\boldsymbol{\Sigma}}\check{\boldsymbol{\Theta}}_k - \mathbf{e}_k^\top\big) - \boldsymbol{\Theta}_j^\top\Big(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Theta}_k - \mathbf{e}_k^\top\Big)\Big| \lesssim M^2\frac{s\log d}{\sqrt{n}}. \tag{C.4.26}$$

Lemma C.4.3 also implies

$$\max_{j\in[d]} |\widehat{\boldsymbol{\Theta}}_j^\top \widehat{\boldsymbol{\Sigma}}_j - 1| \le \max_{j\in[d]} ||\widehat{\boldsymbol{\Theta}}_j^\top \widehat{\boldsymbol{\Sigma}} - \mathbf{e}_j||_\infty \lesssim M\sqrt{\frac{\log d}{n}}. \tag{C.4.27}$$

Combining (C.4.21), (C.4.22) with (C.4.26) and (C.4.27), for sufficiently large d, n, we have, with probability $1 - 1/d^2$, the following holds:

$$\begin{aligned}
\mathrm{I}_3 &\le \max_{(j,k)\in E} \sqrt{n}\Big|\frac{\widehat{\boldsymbol{\Theta}}_j^\top\big(\widehat{\boldsymbol{\Sigma}}\breve{\boldsymbol{\Theta}}_k - \mathbf{e}_k\big)}{\widehat{\boldsymbol{\Theta}}_j^\top\widehat{\boldsymbol{\Sigma}}_j} - \boldsymbol{\Theta}_j^\top\big(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Theta}_k - \mathbf{e}_k\big)\Big| \\
&\le \max_{(j,k)\in E} \big(2\sqrt{n}|\widehat{\boldsymbol{\Theta}}_j^\top\widehat{\boldsymbol{\Sigma}}_j - 1|\cdot|\boldsymbol{\Theta}_j^\top\big(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\big)\boldsymbol{\Theta}_k|\big) + 2\max_{(j,k)\in E}|\widehat{\boldsymbol{\Theta}}_j^\top\big(\widehat{\boldsymbol{\Sigma}}\breve{\boldsymbol{\Theta}}_k - \mathbf{e}_k\big) - \boldsymbol{\Theta}_j^\top\big(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Theta}_k - \mathbf{e}_k\big)| \\
&\le 2M\sqrt{n}\max_{j\in[d]}|\widehat{\boldsymbol{\Theta}}_j^\top\widehat{\boldsymbol{\Sigma}}_j - 1|\cdot||\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}||_{\max} + 2\max_{j,k\in[d]}(|I_{12}| + |I_2|) \lesssim M^2\frac{s\log d}{\sqrt{n}},
\end{aligned} \tag{C.4.28}$$

where the second inequality uses $|x/(1+\delta) - y| \le 2|y\delta| + 2|x - y|$ for any $|\delta| < 1/2$. Therefore, combining (C.4.17), (C.4.18),(C.4.19) with (C.4.28) and the fact $\min_{(j,k)\in E}\sqrt{\boldsymbol{\Theta}_{jj}\boldsymbol{\Theta}_{kk}} \ge \lambda_{\min}(\boldsymbol{\Theta}) \ge 1/r_0$ (as $\boldsymbol{\Theta} \in \mathcal{U}(M, s, r_0)$), we obtain the following:

$$\mathbb{P}(|T_E - \breve{T}_E| > \zeta_1) < \zeta_2, \tag{C.4.29}$$

where $\zeta_1 = s\log d/\sqrt{n}$ and $\zeta_2 = 1/d^2$; thus the condition (b) is verified. Also note that $\zeta_1\sqrt{\log d} + \zeta_2 = s(\log d)^{3/2}/\sqrt{n} + 1/d^2 = o(1)$ holds under the stated scaling condition of Lemma 3.2.1.

Regarding the third condition (c), we bound the difference between $T_E^{\mathcal{B}}$ and $\breve{T}_E^{\mathcal{B}}$ as

$$|T_E^{\mathcal{B}} - \breve{T}_E^{\mathcal{B}}| \le \max_{(j,k)\in E}\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n\Big(\frac{\widehat{\boldsymbol{\Theta}}_j^\top}{\sqrt{\widehat{\boldsymbol{\Theta}}_{jj}\widehat{\boldsymbol{\Theta}}_{kk}}}\big(\boldsymbol{X}_i\boldsymbol{X}_i^\top\widehat{\boldsymbol{\Theta}}_k - \mathbf{e}_k\big) - \frac{\boldsymbol{\Theta}_j^\top}{\sqrt{\boldsymbol{\Theta}_{jj}\boldsymbol{\Theta}_{kk}}}\big(\boldsymbol{X}_i\boldsymbol{X}_i^\top\boldsymbol{\Theta}_k - \mathbf{e}_k\big)\Big)\xi_i\Big|$$

Conditioning on the data $\{\boldsymbol{X}_i\}_{i=1}^n$, the right hand side of the above inequality is a suprema of a

Gaussian process. Therefore, we need to bound the following conditional variance

$$\max_{(j,k)\in E} \frac{1}{n}\sum_{i=1}^{n}\Big[\frac{\widehat{\boldsymbol{\Theta}}_j^\top}{\sqrt{\widehat{\boldsymbol{\Theta}}_{jj}\widehat{\boldsymbol{\Theta}}_{kk}}}\big(\boldsymbol{X}_i\boldsymbol{X}_i^\top\widehat{\boldsymbol{\Theta}}_k-\mathbf{e}_k\big)-\frac{\boldsymbol{\Theta}_j^\top}{\sqrt{\boldsymbol{\Theta}_{jj}\boldsymbol{\Theta}_{kk}}}\big(\boldsymbol{X}_i\boldsymbol{X}_i^\top\boldsymbol{\Theta}_k-\mathbf{e}_k\big)\Big]^2$$

Note the summand (for each i) can be bounded by

$$2\frac{\mathrm{III}_1\mathrm{III}_2}{\min\limits_{(j,k)\in E}\boldsymbol{\Theta}_{jj}\boldsymbol{\Theta}_{kk}}+2\frac{\mathrm{III}_3}{\min\limits_{(j,k)\in E}\widehat{\boldsymbol{\Theta}}_{jj}\widehat{\boldsymbol{\Theta}}_{kk}}$$

where III_1, III_2 and III_3 are defined and bounded as below:

$$\begin{aligned}
\mathrm{III}_1 &:= \max_{(j,k)\in E}|\widehat{\boldsymbol{\Theta}}_{jj}\widehat{\boldsymbol{\Theta}}_{kk}-\boldsymbol{\Theta}_{jj}\boldsymbol{\Theta}_{kk}|^2 \le \left(CM^2\sqrt{\frac{\log d}{n}}\right)^2 && \text{(C.4.30)}\\
\mathrm{III}_2 &:= \max_{(j,k)\in E}[\boldsymbol{\Theta}_j^\top\big(\boldsymbol{X}_i\boldsymbol{X}_i^\top\boldsymbol{\Theta}_k-\mathbf{e}_k\big)]^2=\max_{(j,k)\in E}[\boldsymbol{\Theta}_j^\top\big(\boldsymbol{X}_i\boldsymbol{X}_i^\top-\boldsymbol{\Sigma}\big)\boldsymbol{\Theta}_k]^2\\
&\le \Big[M^2\max_i\|\boldsymbol{X}_i\boldsymbol{X}_i^\top-\boldsymbol{\Sigma}\|_{\max}\Big]^2 && \text{(C.4.31)}\\
\mathrm{III}_3 &= \max_{(j,k)\in E}\Big|\widehat{\boldsymbol{\Theta}}_j^\top\big(\boldsymbol{X}_i\boldsymbol{X}_i^\top\widehat{\boldsymbol{\Theta}}_k-\mathbf{e}_k\big)-\boldsymbol{\Theta}_j^\top\big(\boldsymbol{X}_i\boldsymbol{X}_i^\top\boldsymbol{\Theta}_k-\mathbf{e}_k\big)\Big|^2\\
&\lesssim \Big[2M\|\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}\|_1\max_i\|\boldsymbol{X}_i\boldsymbol{X}_i^\top-\boldsymbol{\Sigma}\|_{\max}\Big]^2. && \text{(C.4.32)}
\end{aligned}$$

According to Lemma C.4.4, we have with probability $1-1/d^2$,

$$\max_i\|\boldsymbol{X}_i\boldsymbol{X}_i^\top-\boldsymbol{\Sigma}\|_{\max}\le C\sqrt{\log(dn)}. \qquad \text{(C.4.33)}$$

Therefore, the event

$$\mathcal{E}=\Big\{\max_{(j,k)\in E}\frac{1}{n}\sum_{i=1}^{n}\Big[\frac{\widehat{\boldsymbol{\Theta}}_j^\top}{\sqrt{\widehat{\boldsymbol{\Theta}}_{jj}\widehat{\boldsymbol{\Theta}}_{kk}}}\big(\boldsymbol{X}_i\boldsymbol{X}_i^\top\widehat{\boldsymbol{\Theta}}_k-\mathbf{e}_k\big)-\frac{\boldsymbol{\Theta}_j^\top}{\sqrt{\boldsymbol{\Theta}_{jj}\boldsymbol{\Theta}_{kk}}}\big(\boldsymbol{X}_i\boldsymbol{X}_i^\top\boldsymbol{\Theta}_k-\mathbf{e}_k\big)\Big]^2\le CM^2\frac{(s\log(dn))^2}{n}\Big\}$$

satisfies $\mathbb{P}(\mathcal{E}^c) < 1/d^2$. Therefore, by the maximal inequality, under the event $\mathcal{E}$, we have

$$\begin{aligned}
&\mathbb{E}\left[\max_{(j,k)\in E}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(\frac{\widehat{\boldsymbol{\Theta}}_j^\top}{\sqrt{\widehat{\boldsymbol{\Theta}}_{jj}\widehat{\boldsymbol{\Theta}}_{kk}}}\big(\boldsymbol{X}_i\boldsymbol{X}_i^\top\widehat{\boldsymbol{\Theta}}_k-\mathbf{e}_k\big)-\frac{\boldsymbol{\Theta}_j^\top}{\sqrt{\boldsymbol{\Theta}_{jj}\boldsymbol{\Theta}_{kk}}}\big(\boldsymbol{X}_i\boldsymbol{X}_i^\top\boldsymbol{\Theta}_k-\mathbf{e}_k\big)\Big)\xi_i \mid \{\boldsymbol{X}_i\}_{i=1}^n\right] \\
\lesssim\ & M^2\frac{(s\log dn)\sqrt{\log d}}{\sqrt{n}}.
\end{aligned}$$

Applying Borell's inequality, we have with probability $1-1/d^2$,

$$\mathbb{P}\left(\max_{(j,k)\in E}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{\widehat{\boldsymbol{\Theta}}_j^\top\big(\boldsymbol{X}_i\boldsymbol{X}_i^\top\widehat{\boldsymbol{\Theta}}_k-\mathbf{e}_k\big)}{\sqrt{\widehat{\boldsymbol{\Theta}}_{jj}\widehat{\boldsymbol{\Theta}}_{kk}}}-\frac{\boldsymbol{\Theta}_j^\top\big(\boldsymbol{X}_i\boldsymbol{X}_i^\top\boldsymbol{\Theta}_k-\mathbf{e}_k\big)}{\sqrt{\boldsymbol{\Theta}_{jj}\boldsymbol{\Theta}_{kk}}}\right)\xi_i > C\sqrt{\frac{s^2\log^4 dn}{n}} \mid \{\boldsymbol{X}_i\}_{i=1}^n\right)\le 1/d^2.$$

This implies that

$$\mathbb{P}\left(\mathbb{P}_\xi\big(|T_E^{\mathcal{B}}-\check{T}_E^{\mathcal{B}}|>\sqrt{(s^2\log^4 dn)/n}\mid\{\boldsymbol{X}_i\}_{i=1}^n\big)>1/d^2\right)<1/d^2.$$

Now we can verify the condition (c) by showing

$$\mathbb{P}(\mathbb{P}_\xi(|T_E^{\mathcal{B}}-\check{T}_E^{\mathcal{B}}|>\zeta_1\mid\{\boldsymbol{X}_i\}_{i=1}^n)>\zeta_2)<\zeta_2, \tag{C.4.34}$$

where $\zeta_1 = s(\log d)^2/\sqrt{n}$, $\zeta_2 = 1/d^2$ and the condition $\zeta_1\sqrt{\log d}+\zeta_2 = s(\log d)^{3/2}/\sqrt{n}+1/d^2 = o(1)$ holds under the stated scaling condition of Lemma 3.2.1. Therefore, by Corollary 3.1 of Chernozhukov et al. (2013), we have

$$\lim_{(n,d)\to\infty}|\mathbb{P}(T_E>\widehat{c}(\alpha,E))-\alpha|=0. \tag{C.4.35}$$

And it holds for any edge set E, thus the proof is complete. □

Lemma C.4.5. *Under the same conditions as Lemma 3.2.1, we have*

$$\mathbb{P}\Big(\max_{j,k\in[d]}|\widehat{\boldsymbol{\Theta}}^{d}_{jk}-\boldsymbol{\Theta}_{jk}|>C_0\sqrt{\frac{\log d}{n}}\Big)<\frac{2}{d^2}, \tag{C.4.36}$$

for some constant $C_0>0$.

Proof. By (C.4.20) and (C.4.28), we have with probability $1-1/d^2$,

$$\max_{j,k\in[d]}|\widehat{\boldsymbol{\Theta}}^{\mathrm{d}}_{jk}-\boldsymbol{\Theta}_{jk}+\boldsymbol{\Theta}_j^\top\big(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Theta}_k-\mathbf{e}_k\big)|\le C_1\frac{s\log d}{n}.$$

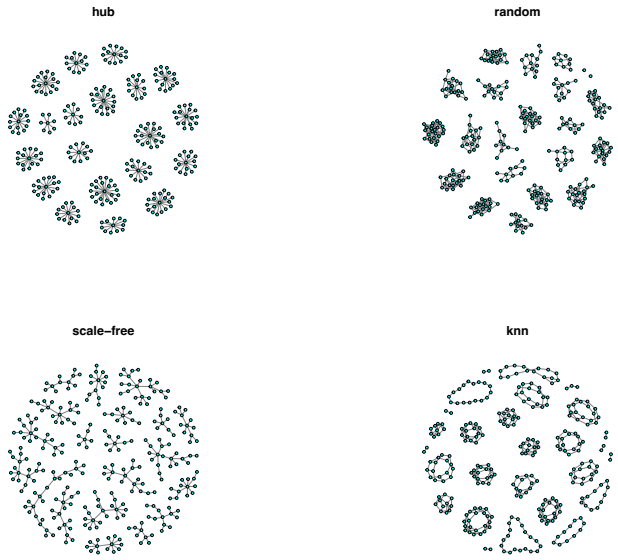
By Lemma C.4.4 and $\|\boldsymbol{\Theta}\|_2\le r_0$, we have $\|\boldsymbol{\Theta}_j^\top\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{\Theta}_k\|_{\psi_1}\le C_2r_0^2$. Applying the maximal inequality (Lemma 2.2.2 in Van Der Vaart & Wellner (1996)), we have for some constant $C_3>0$

$$\begin{aligned}&\mathbb{P}\Big(\max_{j,k\in[d]}|\boldsymbol{\Theta}_j^\top\big(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Theta}_k-\mathbf{e}_k\big)|>C_3r_0^2\sqrt{\frac{\log d}{n}}\Big)\\&\le\mathbb{P}\Big(\max_{j,k\in[d]}\Big|\frac{1}{n}\sum_{i=1}^n(\boldsymbol{\Theta}_j^\top\boldsymbol{X}_i\boldsymbol{X}_i^\top\boldsymbol{\Theta}_k-\mathbb{E}[\boldsymbol{\Theta}_j^\top\boldsymbol{X}_i\boldsymbol{X}_i^\top\boldsymbol{\Theta}_k)]\Big|>C_3r_0^2\sqrt{\frac{\log d}{n}}\Big)\le 1/d^2.\end{aligned}$$

With $C_0=C_1+C_3$, (C.4.36) is proved. And it is not hard to show a similar result for the standardized one-step estimator also holds, i.e.,

$$\mathbb{P}\Big(\max_{j,k\in[d]}|\widetilde{\boldsymbol{\Theta}}^{\mathrm{d}}_{jk}-\boldsymbol{\Theta}^{\star}_{jk}|>C_0'\sqrt{\frac{\log d}{n}}\Big)<\frac{2}{d^2} \tag{C.4.37}$$

for some constant $C_0'>0$. □



C.5 Tables and plots deferred from the main paper

C.5.1 Graph pattern demonstration

C.5.2 Tables of $q\frac{d_0}{d}$

C.5.3 Supplementary FDP and Power plots

Table C.1: $q\frac{d_0}{d}$

$d = 300$	$q = 0.1$			$q = 0.2$		
n	200	300	400	200	300	400
			$p = 20$			
hub	0.0930	0.0930	0.0930	0.1870	0.1870	0.1870
random	0.0620	0.0610	0.0600	0.1230	0.1220	0.1200
scale-free	0.0810	0.0810	0.0810	0.1620	0.1630	0.1620
knn	0.0680	0.0700	0.0690	0.1360	0.1390	0.1390
			$p = 30$			
hub	0.0900	0.0900	0.0900	0.1800	0.1800	0.1800
random	0.0810	0.0810	0.0810	0.1620	0.1620	0.1620
scale-free	0.0810	0.0810	0.0810	0.1620	0.1620	0.1610
knn	0.0730	0.0750	0.0740	0.1460	0.1510	0.1480

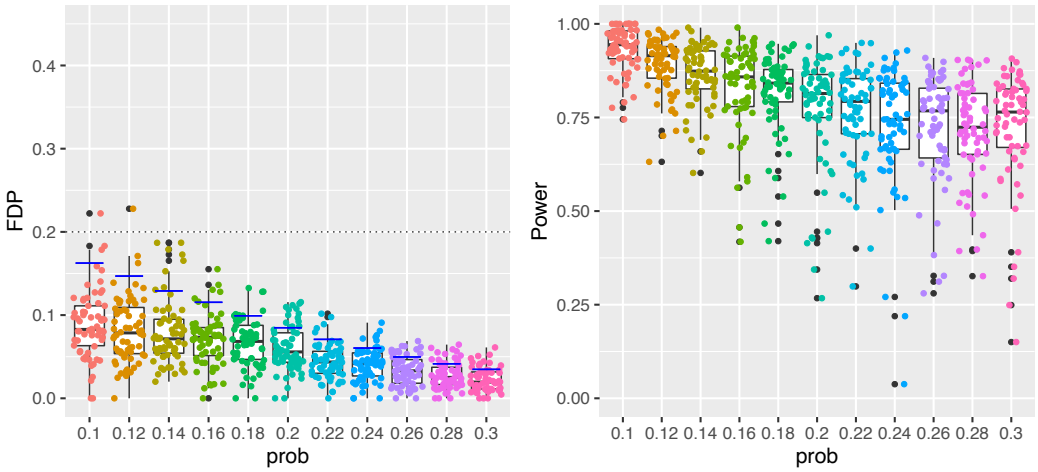


Figure C.1: FDP and power plots with $p = 20$ and the nominal FDR level $q = 0.2$. The other setups are the same as Figure 3.3.

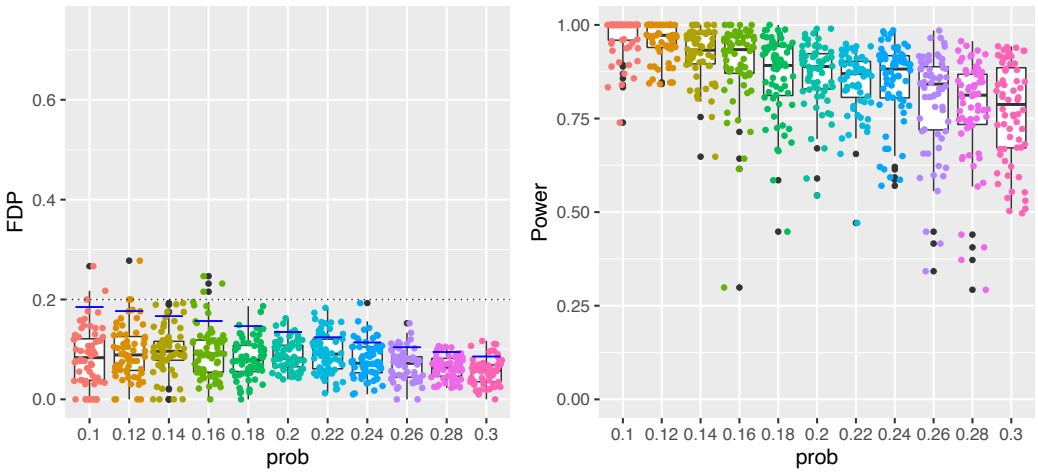


Figure C.2: FDP and power plots with $p = 30$ and the nominal FDR level $q = 0.2$. The other setups are the same as Figure 3.3.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security* (pp. 308–318).

Agrawal, R. & Srikant, R. (2000). Privacy-preserving data mining. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data* (pp. 439–450).

Ait-Sahalia, Y. & Duarte, J. (2003). Nonparametric option pricing under shape restrictions. *Journal of Econometrics*, 116(1-2), 9–47.

Alain, G. & Bengio, Y. (2016). Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*.

Allgöwer, F. (1995). Definition and computation of a nonlinearity measure. In *Nonlinear Control Systems Design 1995* (pp. 257–262). Elsevier.

Amit, Y. (1996). Convergence properties of the gibbs sampler for perturbations of gaussians. *The Annals of Statistics*, 24(1), 122–140.

Angrist, J. D. (2004). Treatment effect heterogeneity in theory and practice. *The economic journal*, 114(494), C52–C83.

Arnold, C. & Neunhoeffer, M. (2020). Really useful synthetic data–a framework to evaluate the quality of differentially private synthetic data. *arXiv preprint arXiv:2004.07740*.

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3), 337–404.

Azadkia, M. & Chatterjee, S. (2019). A simple measure of conditional dependence. *arXiv preprint arXiv:1910.12327*.

Bachman, P., Hjelm, R. D., & Buchwalter, W. (2019). Learning representations by maximizing mutual information across views. *Advances in neural information processing systems*, 32.

Bai, J., Zhang, X., Kang, X., Jin, L., Wang, P., & Wang, Z. (2019). Screening of core genes and pathways in breast cancer development via comprehensive analysis of multi gene expression datasets. *Oncology Letters*, 18(6), 5821–5830.

Banerjee, M., Durot, C., & Sen, B. (2019). Divide and conquer in nonstandard problems and the super-efficiency phenomenon. *The Annals of Statistics*, 47(2), 720–757.

Banerjee, O., El Ghaoui, L., & d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *The Journal of Machine Learning Research*, 9, 485–516.

Barber, R. F. & Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5), 2055 – 2085.

Barber, R. F. & Candès, E. J. (2019). A knockoff filter for high-dimensional selective inference. *The Annals of Statistics*, 47(5), 2504 – 2537.

Barber, R. F. & Janson, L. (2020). Testing goodness-of-fit and conditional independence with approximate co-sufficient sampling. *arXiv preprint arXiv:2007.09851*.

Barlow, R. E. & Brunk, H. D. (1972). The isotonic regression problem and its dual. *Journal of the American Statistical Association*, 67(337), 140–147.

Bates, D. M. & Watts, D. G. (1980). Relative curvature measures of nonlinearity. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(1), 1–16.

Bates, S., Sesia, M., Sabatti, C., & Candès, E. J. (2020). Causal inference in genetic trio studies. *Proceedings of the National Academy of Sciences*, 117(39), 24117–24126.

Beale, E. (1960). Confidence regions in non-linear estimation. *Journal of the Royal Statistical Society: Series B (Methodological)*, 22(1), 41–76.

Beaulieu-Jones, B. K., Wu, Z. S., Williams, C., Lee, R., Bhavnani, S. P., Byrd, J. B., & Greene, C. S. (2019). Privacy-preserving generative deep neural networks support clinical data sharing. *Circulation: Cardiovascular Quality and Outcomes*, 12(7), e005122.

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

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8), 1798–1828.

Benjamini, Y. (2010). Discovering the false discovery rate. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4), 405–416.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1), 289–300.

Benjamini, Y. & Yekutieli, D. (2001). The control of the false discovery rate in multiple testingunder dependency. *The Annals of Statistics*, 29(4), 1165 – 1188.

Benjamini, Y. & Yekutieli, D. (2005). False discovery rate–adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association*, 100(469), 71–81.

Bentkus, V. (1990). Smooth approximations of the norm and differentiable functions with bounded support in banach space l_∞^k. *Lithuanian Mathematical Journal*, 30(3), 223–230.

Bentkus, V., Götze, F., et al. (1996). The berry-esseen bound for student's statistic. *The Annals of Probability*, 24(1), 491–503.

Berk, R., Brown, L., Buja, A., Zhang, K., Zhao, L., et al. (2013). Valid post-selection inference. *The Annals of Statistics*, 41(2), 802–837.

Berrett, T. B., Wang, Y., Barber, R. F., & Samworth, R. J. (2020). The conditional permutation test for independence while controlling for confounders. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1), 175–197.

Blein, S., Barjhoux, L., investigators, G., Damiola, F., Dondon, M.-G., Eon-Marchais, S., Marcou, M., Caron, O., Lortholary, A., Buecher, B., Vennin, P., Berthet, P., Noguès, C., Lasset, C., Gauthier-Villars, M., Mazoyer, S., Stoppa-Lyonnet, D., Andrieu, N., Thomas, G., Sinilnikova, O. M., & Cox, D. G. (2015). Targeted sequencing of the mitochondrial genome of women at high risk of breast cancer without detectable mutations in brca1/2. *PLoS One*, 10(9), e0136192.

Blier, L. & Ollivier, Y. (2018). The description length of deep learning models. *Advances in Neural Information Processing Systems*, 31.

Bojinov, I., Simchi-Levi, D., & Zhao, J. (2020). Design and analysis of switchback experiments. *arXiv preprint arXiv:2009.00148*.

Bowman, A., Jones, M., & Gijbels, I. (1998). Testing monotonicity of regression. *Journal of computational and Graphical Statistics*, 7(4), 489–500.

Bowser, A., Wiggins, A., Shanley, L., Preece, J., & Henderson, S. (2014). Sharing data while protecting privacy in citizen science. *interactions*, 21(1), 70–73.

Boyd, S., Kim, S.-J., Vandenberghe, L., & Hassibi, A. (2007). A tutorial on geometric programming. *Optimization and engineering*, 8(1), 67–127.

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

Brunk, H., Barlow, R. E., Bartholomew, D. J., & Bremner, J. M. (1972). *Statistical inference under order restrictions.(the theory and application of isotonic regression)*. Technical report, Missouri Univ Columbia Dept of Statistics.

Bühlmann, P. et al. (2013). Statistical significance in high-dimensional linear models. *Bernoulli*, 19(4), 1212–1242.

Bühlmann, P., van de Geer, S., et al. (2015). High-dimensional inference in misspecified linear models. *Electronic Journal of Statistics*, 9(1), 1449–1473.

Buja, A., Berk, R. A., Brown, L. D., George, E. I., Pitkin, E., Traskin, M., Zhao, L., & Zhang, K. (2015). Models as Approximations–A conspiracy of random regressors and model deviations against classical inference in regression. *Statistical Science*, (pp. 1–44).

Buja, A. & Brown, L. (2014). Discussion:"a significance test for the lasso". *The Annals of Statistics*, 42(2), 509–517.

Buja, A., Brown, L., Berk, R., George, E., Pitkin, E., Traskin, M., Zhang, K., & Zhao, L. (2019a). Models as Approximations I: Consequences illustrated with linear regression. *Statistical Science*, 34(4), 523 – 544.

Buja, A., Brown, L., Kuchibhotla, A. K., Berk, R., George, E., Zhao, L., et al. (2019b). Models as Approximations II: A model-free theory of parametric regression. *Statistical Science*, 34(4), 545–565.

Bun, M. & Steinke, T. (2016). Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference* (pp. 635–658).: Springer.

Cai, T., Liu, W., & Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494), 594–607.

Cai, T., Liu, W., & Xia, Y. (2013). Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. *Journal of the American Statistical Association*, 108(501), 265–277.

Cai, T. T. & Ma, Z. (2013). Optimal hypothesis testing for high dimensional covariance matrices. *Bernoulli*, 19(5B), 2359 – 2388.

Cai, T. T. & Zhang, A. (2016). Inference for high-dimensional differential correlation matrices. *Journal of Multivariate Analysis*, 143, 107–126.

Candès, E. J., Fan, Y., Janson, L., & Lv, J. (2018a). Panning for gold: Model-X knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3), 551–577.

Candès, E. J., Fan, Y., Janson, L., & Lv, J. (2018b). Panning for gold: Model-X knockoffs for high-dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B*, 80(3), 551–577.

Castro, J., Gómez, D., & Tejada, J. (2009). Polynomial calculation of the shapley value based on sampling. *Computers & Operations Research*, 36(5), 1726–1730.

Chang, J., Shao, Q.-M., & Zhou, W.-X. (2016). Cramér-type moderate deviations for Studentized two-sample U-statistics with applications. *The Annals of Statistics*, 44(5), 1931 – 1956.

Charnes, A., Golany, B., Keane, M., & Rousseau, J. (1988). Extremal principle solutions of games in characteristic function form: core, Chebychev and Shapley value generalizations. In *Econometrics of Planning and Efficiency* (pp. 123–133). Springer.

Chatterjee, S. (2021). A new coefficient of correlation. *Journal of the American Statistical Association*, 116(536), 2009–2022.

Chatterjee, S., Guntuboyina, A., & Sen, B. (2015). On risk bounds in isotonic and other shape restricted regression problems. *The Annals of Statistics*, 43(4), 1774–1800.

Chen, Q., Xiang, C., Xue, M., Li, B., Borisov, N., Kaarfar, D., & Zhu, H. (2018). Differentially private data generative models. *arXiv preprint arXiv:1812.02274*.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International conference on machine learning* (pp. 1597–1607).: PMLR.

Chen, W.-C., Wang, C.-Y., Hung, Y.-H., Weng, T.-Y., Yen, M.-C., & Lai, M.-D. (2016). Systematic analysis of gene expression alterations and clinical outcomes for long-chain acyl-coenzyme a synthetase family in cancer. *PLoS One*, 11(5), e0155660.

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

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

Chernozhukov, V., Chetverikov, D., & Kato, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, 41(6), 2786–2819.

Chernozhukov, V., Chetverikov, D., & Kato, K. (2014). Anti-concentration and honest, adaptive confidence bands. *The Annals of Statistics*, 42(5), 1787–1818.

Chernozhukov, V., Chetverikov, D., & Kato, K. (2015). Comparison and anti-concentration bounds for maxima of gaussian random vectors. *Probability Theory and Related Fields*, 162(1), 47–70.

Chetverikov, D. (2019). Testing regression monotonicity in econometric models. *Econometric Theory*, 35(4), 729–776.

Chia, C., Sesia, M., Ho, C.-S., Jeffrey, S. S., Dionne, J., Candès, E. J., & Howe, R. T. (2020). Interpretable signal analysis with knockoffs enhances classification of bacterial raman spectra. *arXiv preprint arXiv:2006.04937*.

Choi, H. M. & Hobert, J. P. (2013). Analysis of mcmc algorithms for bayesian linear regression with laplace errors. *Journal of Multivariate Analysis*, 117, 32–40.

Covert, I., Lundberg, S. M., & Lee, S.-I. (2020). Understanding global feature contributions with additive importance measures. *Advances in Neural Information Processing Systems*, 33.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4), 303–314.

Dai, C., Lin, B., Xing, X., & Liu, J. S. (2020a). False discovery rate control via data splitting. *arXiv preprint arXiv:2002.08542*.

Dai, C., Lin, B., Xing, X., & Liu, J. S. (2020b). A scale-free approach for false discovery rate control in generalized linear models. *arXiv preprint arXiv:2007.01237*.

Dai, R. & Barber, R. (2016). The knockoff filter for FDR control in group-sparse and multitask regression. In *International Conference on Machine Learning* (pp. 1851–1859).: PMLR.

Dai, R., Song, H., Barber, R. F., & Raskutti, G. (2020c). The bias of isotonic regression. *Electronic journal of statistics*, 14(1), 801.

Deb, N., Ghosal, P., & Sen, B. (2020). Measuring association on topological spaces using kernels and geometric graphs. *arXiv preprint arXiv:2010.01768*.

Deb, N. & Sen, B. (2021). Multivariate rank-based distribution-free nonparametric testing using measure transportation. *Journal of the American Statistical Association*, (pp. 1–16).

Delgado, M. A. (1993). Testing the equality of nonparametric regression curves. *Statistics & probability letters*, 17(3), 199–204.

Dezeure, R., Bühlmann, P., & Zhang, C.-H. (2017). High-dimensional simultaneous inference with the bootstrap. *Test*, 26(4), 685–719.

Dharmadhikari, S., Jogdeo, K., et al. (1969). Bounds on moments of certain random variables. *The Annals of Mathematical Statistics*, 40(4), 1506–1509.

Diaconis, P., Khare, K., Saloff-Coste, L., et al. (2010). Stochastic alternating projections. *Illinois Journal of Mathematics*, 54(3), 963–979.

Ding, X. & Zhou, Z. (2020). Estimation and inference for precision matrices of nonstationary time series. *The Annals of Statistics*, 48(4), 2455 – 2477.

Dinur, I. & Nissim, K. (2003). Revealing information while preserving privacy. In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems* (pp. 202–210).

Dong, J., Roth, A., & Su, W. J. (2019). Gaussian differential privacy. *arXiv preprint arXiv:1905.02383*.

Drechsler, J. (2018). Some clarifications regarding fully synthetic data. In *International Conference on Privacy in Statistical Databases* (pp. 109–121).: Springer.

Dwork, C. (2008). Differential privacy: A survey of results. In *International conference on theory and applications of models of computation* (pp. 1–19).: Springer.

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference* (pp. 265–284).: Springer.

Dwork, C. & Rothblum, G. N. (2016). Concentrated differential privacy. *arXiv preprint arXiv:1603.01887*.

Dykstra, R. L. & Robertson, T. (1982). An algorithm for isotonic regression for two or more independent variables. *The Annals of Statistics*, 10(3), 708–716.

Egami, N. & Imai, K. (2018). Causal interaction in factorial experiments: Application to conjoint analysis. *Journal of the American Statistical Association*.

Eisenach, C., Bunea, F., Ning, Y., & Dinicu, C. (2020). High-dimensional inference for cluster-based graphical models. *Journal of Machine Learning Research*, 21(53).

Elliot, M. & Domingo-Ferrer, J. (2018). The future of statistical disclosure control. *arXiv preprint arXiv:1812.09204*.

Fan, J. (1992). Design-adaptive nonparametric regression. *Journal of the American statistical Association*, 87(420), 998–1004.

Fang, B. & Guntuboyina, A. (2019). On the risk of convex-constrained least squares estimators under misspecification. *Bernoulli*, 25(3), 2206–2244.

Feng, H. & Ning, Y. (2019). High-dimensional mixed graphical model with ordinal data: Parameter estimation and statistical inference. In *The 22nd International Conference on Artificial Intelligence and Statistics* (pp. 654–663).: PMLR.

Feng, J., Williamson, B., Simon, N., & Carone, M. (2018). Nonparametric variable importance using an augmented neural network with multi-task learning. In *International Conference on Machine Learning* (pp. 1496–1505).

Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177), 1–81.

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3), 432–441.

Fukumizu, K., Gretton, A., Sun, X., & Schölkopf, B. (2008). Kernel measures of conditional dependence. In *Advances in neural information processing systems* (pp. 489–496).

Gijbels, I., Hall, P., Jones, M., & Koch, I. (2000). Tests for monotonicity of a regression mean with guaranteed level. *Biometrika*, 87(3), 663–673.

Gil, M. (2011). *On Rényi divergence measures for continuous alphabet sources*. PhD thesis, Citeseer.

Gretton, A., Bousquet, O., Smola, A., & Schölkopf, B. (2005). Measuring statistical dependence with hilbert-schmidt norms. In *International Conference on Algorithmic Learning Theory* (pp. 63–77).: Springer.

Gretton, A., Fukumizu, K., Teo, C., Song, L., Schölkopf, B., & Smola, A. (2007). A kernel statistical test of independence. *Advances in Neural Information Processing Systems*, 20.

Grönwall, T. H. (1919). Note on the derivatives with respect to a parameter of the solutions of a system of differential equations. *Annals of Mathematics*, 20(4), 292–296.

Gu, Q., Cao, Y., Ning, Y., & Liu, H. (2015). Local and global inference for high dimensional nonparanormal graphical models. *arXiv preprint arXiv:1502.02347*.

Guay, M. et al. (1996). *Measurement of nonlinearity in chemical process control*. Queen's University.

Guntuboyina, A. & Sen, B. (2015). Global risk bounds and adaptation in univariate convex regression. *Probability Theory and Related Fields*, 163(1), 379–411.

Guntuboyina, A., Sen, B., et al. (2018). Nonparametric shape-restricted regression. *Statistical Science*, 33(4), 568–594.

Guttman, I. & Meeter, D. A. (1965). On beale's measures of non-linearity. *Technometrics*, 7(4), 623–637.

Hainmueller, J. & Hopkins, D. J. (2014). Public attitudes toward immigration. *Annual Review of Political Science*, 17(1), 225–249.

Hall, P. & Heckman, N. E. (2000). Testing for monotonicity of a regression mean by calibrating for linear functions. *Annals of Statistics*, (pp. 20–39).

Hamilton, D. C., Watts, D. G., & Bates, D. M. (1982). Accounting for intrinsic nonlinearity in nonlinear regression parameter inference regions. *The Annals of Statistics*, (pp. 386–393).

Han, Q., Wang, T., Chatterjee, S., & Samworth, R. J. (2019). Isotonic regression in general dimensions. *The Annals of Statistics*, 47(5), 2440–2471.

Hannah, L. A. & Dunson, D. B. (2013). Multivariate convex regression with adaptive partitioning. *The Journal of Machine Learning Research*, 14(1), 3261–3294.

Heller, R., Heller, Y., & Gorfine, M. (2013). A consistent multivariate test of association based on ranks of distances. *Biometrika*, 100(2), 503–510.

Hellwig, B., Madjar, K., Edlund, K., Marchan, R., Cadenas, C., Heimes, A.-S., Almstedt, K., Lebrecht, A., Sicking, I., Battista, M. J., Micke, P., Schmidt, M., Hengstler, J. G., & Rahnenführer, J. (2016). Epsin family member 3 and ribosome-related genes are associated with late metastasis in estrogen receptor-positive breast cancer and long-term survival in non-small cell lung cancer using a genome-wide identification and validation strategy. *PLoS One*, 11(12), 1–18.

Henaff, O. (2020). Data-efficient image recognition with contrastive predictive coding. In *International Conference on Machine Learning* (pp. 4182–4192).: PMLR.

Hirschfeld, H. O. (1935). A connection between correlation and contingency. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 31 (pp. 520–524).: Cambridge University Press.

Hoff, P. D. (2009). *A first course in Bayesian statistical methods*, volume 580. Springer.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2), 251–257.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5), 359–366.

Horvitz, E. & Mulligan, D. (2015). Data, privacy, and the greater good. *Science*, 349(6245), 253–255.

Huang, C., Kastelman, D., Bauman, J., & Tang, Y. (2020a). Control using predictions as covariates in switchback experiments.

Huang, D. & Janson, L. (2020). Relaxing the assumptions of knockoffs by conditioning. *The Annals of Statistics*, 48(5), 3021 – 3042.

Huang, Z., Deb, N., & Sen, B. (2020b). Kernel partial correlation coefficient–a measure of conditional dependence. *arXiv preprint arXiv:2012.14804*.

Ilyas, M. U., Shafiq, M. Z., Liu, A. X., & Radha, H. (2011). A distributed and privacy preserving algorithm for identifying information hubs in social networks. In *2011 Proceedings IEEE INFOCOM* (pp. 561–565).: IEEE.

Ingram, J. M. & Marsh, M. (1991). Projections onto convex cones in hilbert space. *Journal of approximation theory*, 64(3), 343–350.

Isaak, J. & Hanna, M. J. (2018). User data privacy: Facebook, cambridge analytica, and privacy protection. *Computer*, 51(8), 56–59.

Isserlis, L. (1918). On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika*, 12(1/2), 134–139.

Janková, J. & van de Geer, S. (2017). Honest confidence regions and optimality in high-dimensional precision matrix estimation. *Test*, 26(1), 143–162.

Janková, J. & van de Geer, S. (2018). Inference in high-dimensional graphical models. *arXiv preprint arXiv:1801.08512*.

Janson, L. (2017). *A Model-Free Approach to High-Dimensional Inference*. PhD thesis, Stanford University.

Javanmard, A. & Javadi, H. (2019). False discovery rate control via debiased lasso. *Electronic Journal of Statistics*, 13(1), 1212 – 1253.

Javanmard, A. & Montanari, A. (2013). Nearly optimal sample size in hypothesis testing for high-dimensional regression. In *2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)* (pp. 1427–1434).: IEEE.

Javanmard, A. & Montanari, A. (2014a). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15(1), 2869–2909.

Javanmard, A. & Montanari, A. (2014b). Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory. *IEEE Transactions on Information Theory*, 60(10), 6522–6554.

Jin, J., Ke, Z. T., Luo, S., & Wang, M. (2020). Estimating the number of communities by stepwise goodness-of-fit. *arXiv preprint arXiv:2009.09177*.

Johnstone, I. M. & Titterington, D. M. (2009). Statistical challenges of high-dimensional data.

Jung, Y. J. & Hobert, J. P. (2014). Spectral properties of mcmc algorithms for bayesian linear regression with generalized hyperbolic errors. *Statistics & Probability Letters*, 95, 92–100.

Karolczak, M. & Mickiewicz, A. (1995). Why to calculate, when to use, and how to understand curvature measures of nonlinearity. *Current Separations*, 14(1), 11.

Kastelman, D. & Ramesh, R. (2018). Switchback tests and randomized experimentation under network effects at doordash. *URL: https://medium. com/@DoorDash/switchback-tests-and-randomized-experimentationunder-network-effects-at-doordash-f1d938ab7c2a*.

Katsevich, E. & Ramdas, A. (2020). A theoretical treatment of conditional independence testing under model-X. *arXiv preprint arXiv:2005.05506*.

Katsevich, E. & Roeder, K. (2020). Conditional resampling improves sensitivity and specificity of single cell crispr regulatory screens. *bioRxiv*.

Katsevich, E. & Sabatti, C. (2019). Multilayer knockoff filter: Controlled variable selection at multiple resolutions. *The Annals of Applied Statistics*, 13(1), 1 – 33.

Kayalar, S. & Weinert, H. L. (1988). Error bounds for the method of alternating projections. *Mathematics of Control, Signals and Systems*, 1(1), 43–59.

Ke, Z. T., Ma, Y., & Lin, X. (2020). Estimation of the number of spiked eigenvalues in a covariance matrix by bulk eigenvalue matching analysis. *arXiv preprint arXiv:2006.00436*.

Kotchoni, R. (2018). Detecting and measuring nonlinearity. *Econometrics*, 6(3), 37.

Kruschke, J. K. (2010). Bayesian data analysis. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(5), 658–676.

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1), 1–27.

Kuchibhotla, A. K., Mukherjee, S., & Banerjee, D. (2021). High-dimensional CLT: Improvements, non-uniform extensions and large deviations. *Bernoulli*, 27(1), 192 – 217.

Kuhn, M. & Johnson, K. (2019). *Feature engineering and selection: A practical approach for predictive models*. CRC Press.

Kur, G., Gao, F., Guntuboyina, A., & Sen, B. (2020). Convex regression in multidimensions: Suboptimality of least squares estimators. *arXiv preprint arXiv:2006.02044*.

Lam, C. & Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *The Annals of Statistics*, 37(6B), 4254–4278.

Lee, J. D., Sun, D. L., Sun, Y., Taylor, J. E., et al. (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3), 907–927.

Lee, L.-F. & Pitt, M. M. (1986). Microeconometric demand system with binding nonnegativity constraints: the dual approach. *Econometrica: Journal of the Econometric Society*, (pp. 1237–1242).

Lee, R. K.-W., Hoang, T.-A., & Lim, E.-P. (2019). Discovering hidden topical hubs and authorities across multiple online social networks. *IEEE Transactions on Knowledge and Data Engineering*, 33(1), 70–84.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., & Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523), 1094–1111.

Leshno, M., Lin, V. Y., Pinkus, A., & Schocken, S. (1993). Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural networks*, 6(6), 861–867.

Li, J. & Maathuis, M. H. (2019). GGM knockoff filter: False discovery rate control for gaussian graphical models. *arXiv preprint arXiv:1908.11611*.

Li, L., Tchetgen Tchetgen, E., van der Vaart, A., & Robins, J. M. (2011). Higher order inference on a treatment effect under low regularity conditions. *Statistics & Probability Letters*, 81(7), 821–828. Statistics in Biological and Medical Sciences.

Li, N. & Stephens, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165(4), 2213–2233.

Li, X. R. (2012). Measure of nonlinearity for stochastic systems. In *2012 15th International Conference on Information Fusion* (pp. 1073–1080).: IEEE.

Li, Y., Giorgi, E. E., Beckman, K. B., Caberto, C., Kazma, R., Lum-Jones, A., Haiman, C. A., Marchand, L. L., Stram, D. O., Saxena, R., & Cheng, I. (2019). Association between mitochondrial genetic variation and breast cancer risk: The multiethnic cohort. *PLoS One*, 14(10), 1–14.

Lim, E. et al. (2020). The limiting behavior of isotonic and convex regression estimators when the model is misspecified. *Electronic Journal of Statistics*, 14(1), 2053–2097.

Lim, E. & Glynn, P. W. (2012). Consistency of multidimensional convex regression. *Operations Research*, 60(1), 196–208.

Lin, Y., Lee, D. D., & Saul, L. K. (2004). Nonnegative deconvolution for time of arrival estimation. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2 (pp. ii–377).: IEEE.

Liu, H. & Wang, L. (2017). Tiger: A tuning-insensitive approach for optimally estimating gaussian graphical models. *Electronic Journal of Statistics*, 11(1), 241–294.

Liu, L., Mukherjee, R., & Robins, J. M. (2019a). On assumption-free tests and confidence intervals for causal effects estimated by machine learning. *arXiv preprint arXiv:1904.04276*.

Liu, M., Katsevich, E., Janson, L., & Ramdas, A. (2021). Fast and powerful conditional randomization testing via distillation. *Biometrika*. asab039.

Liu, M., Xia, Y., Cai, T., & Cho, K. (2020). Integrative high dimensional multiple testing with heterogeneity under data sharing constraints. *arXiv preprint arXiv:2004.00816*.

Liu, W. (2013). Gaussian graphical model estimation with false discovery rate control. *The Annals of Statistics*, 41(6), 2948–2978.

Liu, W. & Luo, S. (2014). Hypothesis testing for high-dimensional regression models.

Liu, W. & Luo, X. (2015). Fast and adaptive sparse precision matrix estimation in high dimensions. *Journal of multivariate analysis*, 135, 153–162.

Liu, W. & Shao, Q.-M. (2010). Cramér-type moderate deviation for the maximum of the periodogram with application to simultaneous tests in gene expression time series. *The Annals of Statistics*, 38(3), 1913 – 1935.

Liu, W. & Shao, Q.-M. (2014). Phase transition and regularized bootstrap in large-scale t-tests with false discovery rate control. *The Annals of Statistics*, 42(5), 2003 – 2025.

Liu, Y., Gu, H.-Y., Zhu, J., Niu, Y.-M., Zhang, C., & Guo, G.-L. (2019b). Identification of hub genes and key pathways associated with bipolar disorder based on weighted gene co-expression network analysis. *Frontiers in Physiology*, 10, 1081.

Liu, Y., Yi, Y., Wu, W., Wu, K., & Zhang, W. (2019c). Bioinformatics prediction and analysis of hub genes and pathways of three types of gynecological cancer. *Oncology Letters*, 18(1), 617–628.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., et al. (2013). The genotype-tissue expression (GTEx) project. *Nature Genetics*, 45(6), 580–585.

Lou, W., Ding, B., Wang, S., & Fu, P. (2020). Overexpression of gpx3, a potential biomarker for diagnosis and prognosis of breast cancer, inhibits progression of breast cancer cells in vitro. *Cancer Cell International*, 20(1), 1–15.

Lu, J., Neykov, M., & Liu, H. (2017). Adaptive inferential method for monotone graph invariants. *arXiv preprint arXiv:1707.09114*.

Luce, R. & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology*, 1(1), 1–27.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, 2(1), 2522–5839.

Luscombe, N. M., Babu, M. M., Yu, H., Snyder, M., Teichmann, S. A., & Gerstein, M. (2004). Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, 431(7006), 308–312.

Luss, R., Rosset, S., & Shahar, M. (2012). Efficient regularized isotonic regression with application to gene–gene interaction search. *The Annals of Applied Statistics*, 6(1), 253–283.

Malvia, S., Bagadi, S. A. R., Pradhan, D., Chintamani, C., Bhatnagar, A., Arora, D., Sarin, R., & Saxena, S. (2019). Study of gene expression profiles of breast cancers in indian women. *Scientific Reports*, 9(1), 1–15.

Marino, N., German, R., Rao, X., Simpson, E., Liu, S., Wan, J., Liu, Y., Sandusky, G., Jacobsen, M., Stoval, M., Cao, S., & Storniolo, A. M. V. (2020). Upregulation of lipid metabolism genes in the breast prior to cancer diagnosis. *NPJ Breast Cancer*, 6(1), 1–13.

Mazumder, R., Choudhury, A., Iyengar, G., & Sen, B. (2019). A computational framework for multivariate convex regression and its variants. *Journal of the American Statistical Association*, 114(525), 318–331.

McCullagh, P. & Nelder, J. A. (2019). *Generalized linear models*. Routledge.

McSherry, F. & Talwar, K. (2007). Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)* (pp. 94–103).: IEEE.

Meinshausen, N. & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3), 1436–1462.

Meyer, M. C. (2003). A test for linear versus convex regression function using shape-restricted regression. *Biometrika*, (pp. 223–232).

Mohamed, M. M., Sabet, S., Peng, D.-F., Nouh, M. A., El-Shinawi, M., & El-Rifai, W. (2014). Promoter hypermethylation and suppression of glutathione peroxidase 3 are associated with inflammatory breast carcinogenesis. *Oxidative Medicine and Cellular Longevity*, 2014.

Neelon, B. & Dunson, D. B. (2004). Bayesian isotonic regression and trend analysis. *Biometrics*, 60(2), 398–406.

Nevo, D. & Ritov, Y. (2016). On bayesian robust regression with diverging number of predictors. *Electronic Journal of Statistics*, 10(2), 3045–3062.

Newey, W. K. & Robins, J. R. (2018). Cross-fitting and fast remainder rates for semiparametric estimation. *arXiv preprint arXiv:1801.09138*.

Newman, M. E., Watts, D. J., & Strogatz, S. H. (2002). Random graph models of social networks. *Proceedings of the National Academy of Sciences*, 99(suppl 1), 2566–2572.

Neykov, M., Lu, J., & Liu, H. (2019). Combinatorial inference for graphical models. *The Annals of Statistics*, 47(2), 795–827.

Nickl, R., Van De Geer, S., et al. (2013). Confidence sets in sparse regression. *The Annals of Statistics*, 41(6), 2852–2876.

Nissim, K., Raskhodnikova, S., & Smith, A. (2007). Smooth sensitivity and sampling in private data analysis. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing* (pp. 75–84).

Obozinski, G., Taskar, B., & Jordan, M. (2006). Multi-task feature selection. *Statistics Department, UC Berkeley, Tech. Rep*, 2(2.2), 2.

Owen, A. B. & Prieur, C. (2017). On shapley value for measuring importance of dependent inputs. *SIAM/ASA Journal on Uncertainty Quantification*, 5(1), 986–1002.

Park, J. & Muandet, K. (2020). A measure-theoretic approach to kernel conditional mean embeddings. *Advances in Neural Information Processing Systems*, 33, 21247–21259.

Peng, J., Wang, P., Zhou, N., & Zhu, J. (2009). Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, 104(486), 735–746.

Pinelis, I., Molzon, R., et al. (2016). Optimal-order bounds on the rate of convergence to normality in the multivariate delta method. *Electronic Journal of Statistics*, 10(1), 1001–1063.

Pinkus, A. (1999). Approximation theory of the mlp model in neural networks. *Acta numerica*, 8, 143–195.

Power, J. D., Schlaggar, B. L., Lessov-Schlaggar, C. N., & Petersen, S. E. (2013). Evidence for hubs in human functional brain networks. *Neuron*, 79(4), 798–813.

Raghunathan, T. E., Reiter, J. P., & Rubin, D. B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of official statistics*, 19(1), 1.

Ravikumar, P., Lafferty, J., Liu, H., & Wasserman, L. (2009). Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5), 1009–1030.

Ravikumar, P., Wainwright, M. J., Raskutti, G., & Yu, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5, 935–980.

Ren, B., Schwager, E., Tickle, T., & Huttenhower, C. (2016). sparsDOSSA sparse data observations for simulating synthetic abundance. *R package version 1.12. 0*.

Ren, Z., Sun, T., Zhang, C.-H., & Zhou, H. H. (2015). Asymptotic normality and optimalities in estimation of large Gaussian graphical models. *The Annals of Statistics*, 43(3), 991 – 1026.

Rinaldo, A., Wasserman, L., & G'Sell, M. (2019a). Bootstrapping and sample splitting for high-dimensional, assumption-lean inference. *The Annals of Statistics*, 47(6), 3438 – 3469.

Rinaldo, A., Wasserman, L., & G'Sell, M. (2019b). Bootstrapping and sample splitting for high-dimensional, assumption-lean inference. *The Annals of Statistics*, 47(6), 3438–3469.

Robins, J., Li, L., Tchetgen, E., van der Vaart, A., et al. (2008). Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and Statistics: Essays in Honor of David A. Freedman* (pp. 335–421). Institute of Mathematical Statistics.

Robins, J., Tchetgen, E. T., Li, L., & van der Vaart, A. (2009). Semiparametric minimax rates. *Electronic Journal of Statistics*, 3, 1305.

Robins, J. M., Li, L., Mukherjee, R., Tchetgen, E. T., van der Vaart, A., et al. (2017). Minimax estimation of a functional on a structured high-dimensional model. *The Annals of Statistics*, 45(5), 1951–1987.

Rothman, A. J., Bickel, P. J., Levina, E., & Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2, 494–515.

Rubinov, M. & Sporns, O. (2010). Complex network measures of brain connectivity: uses and interpretations. *Neuroimage*, 52(3), 1059–1069.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., & Tarantola, S. (2008). *Global sensitivity analysis: the primer*. John Wiley & Sons.

Scheet, P. & Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *The American Journal of Human Genetics*, 78(4), 629–644.

Seijo, E. & Sen, B. (2011). Nonparametric least squares estimation of a multivariate convex regression function. *The Annals of Statistics*, 39(3), 1633–1657.

Sen, B. & Meyer, M. (2017). Testing against a linear regression model using ideas from shape-restricted estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(2), 423–448.

Sesia, M., Bates, S., Candès, E. J., Marchini, J., & Sabatti, C. (2020a). Controlling the false discovery rate in gwas with population structure. *bioRxiv*.

Sesia, M., Katsevich, E., Bates, S., Candès, E. J., & Sabatti, C. (2020b). Multi-resolution localization of causal variants across the genome. *Nature communications*, 11(1), 1–10.

Sesia, M., Sabatti, C., & Candès, E. J. (2019). Gene hunting with hidden markov model knockoffs. *Biometrika*, 106(1), 1–18.

Shah, R. D. & Peters, J. (2020). The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3), 1514 – 1538.

Shao, X. & Zhang, J. (2014). Martingale difference correlation and its use in high-dimensional variable screening. *Journal of the American Statistical Association*, 109(507), 1302–1318.

Shapley, L. S. (1953). A value for n-person games. *Contributions to the Theory of Games*, 2(28), 307–317.

Shaw, P., Kabani, N. J., Lerch, J. P., Eckstrand, K., Lenroot, R., Gogtay, N., Greenstein, D., Clasen, L., Evans, A., Rapoport, J. L., Giedd, J. N., & Wise, S. P. (2008). Neurodevelopmental trajectories of the human cerebral cortex. *Journal of Neuroscience*, 28(14), 3586–3594.

Shen, S. & Lu, J. (2020). Combinatorial-probabilistic trade-off: Community properties test in the stochastic block models. *arXiv preprint arXiv:2010.15063*.

Shen, X., Pan, W., & Zhu, Y. (2012). Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association*, 107(497), 223–232.

Sirois, I., Aguilar-Mahecha, A., Lafleur, J., Fowler, E., Vu, V., Scriver, M., Buchanan, M., Chabot, C., Ramanathan, A., Balachandran, B., Légaré, S., Przybytkowski, E., Lan, C., Krzemien, U., Cavallone, L., Aleynikova, O., Ferrario, C., Guilbert, M.-C., Benlimame, N., Saad, A., Alaoui-Jamali, M., Saragovi, H. U., Josephy, S., O'Flanagan, C., Hursting, S. D., Richard, V. R., Zahedi, R. P., Borchers, C. H., Bareke, E., Nabavi, S., Tonellato, P., Roy, J.-A., Robidoux, A., Marcus, E. A., Mihalcioiu, C., Majewski, J., & Basik, M. (2019). A unique morphological phenotype in chemoresistant triple-negative breast cancer reveals metabolic reprogramming and PLIN4 expression as a molecular vulnerability. *Molecular Cancer Research*, 17(12), 2492–2507.

Slawski, M., Hein, M., & Campus, E. (2014). Sparse recovery for protein mass spectrometry data. *Practical Applications of Sparse Modeling*, 5, 79–98.

Slepian, D. (1962). The one-sided barrier problem for gaussian noise. *Bell System Technical Journal*, 41(2), 463–501.

Snoke, J., Raab, G. M., Nowok, B., Dibben, C., & Slavkovic, A. (2018). General and specific utility measures for synthetic data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181(3), 663–688.

Stephens, M., Smith, N. J., & Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *The American Journal of Human Genetics*, 68(4), 978–989.

Stephens, M. A. (2012). Goodness-of-fit and sufficiency: Exact and approximate tests. *Methodology and Computing in Applied Probability*, 14(3), 785–791.

Štrumbelj, E. & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41(3), 647–665.

Sun, T. & Zhang, C.-H. (2012). Scaled sparse linear regression. *Biometrika*, 99(4), 879–898.

Sur, P. & Candès, E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29), 14516–14525.

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International journal of uncertainty, fuzziness and knowledge-based systems*, 10(05), 557–570.

Székely, G. J. & Rizzo, M. L. (2013). Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8), 1249–1272.

Székely, G. J., Rizzo, M. L., & Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6), 2769–2794.

Tang, Y. & Huang, C. (2019). Analyzing switchback experiments by cluster robust standard error to prevent false positive results. *URL: https://doordash.engineering/2019/09/11/cluster-robust-standard-error-in-switchback-experiments/*.

Tansey, W., Veitch, V., Zhang, H., Rabadan, R., & Blei, D. M. (2022). The holdout randomization test for feature selection in black box models. *Journal of Computational and Graphical Statistics*, 31(1), 151–162.

Taylor, J., Lockhart, R., Tibshirani, R. J., & Tibshirani, R. (2014). Exact post-selection inference for forward stepwise and least angle regression. *arXiv preprint arXiv:1401.3889*, 7, 10–1.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267–288.

Van de Geer, S., Bühlmann, P., Ritov, Y., Dezeure, R., et al. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3), 1166–1202.

van den Heuvel, M. P. & Sporns, O. (2013). Network hubs in the human brain. *Trends in cognitive sciences*, 17(12), 683–696.

Van den Meersche, K., Soetaert, K., & Van Oevelen, D. (2009). xsample (): an r function for sampling linear inverse problems. *Journal of Statistical Software*, 30, 1–15.

Van Der Vaart, A. W. & Wellner, J. A. (1996). Weak convergence. In *Weak convergence and empirical processes* (pp. 16–28). Springer.

van Handel, R. (2014). *Probability in high dimension*. Technical report, Princeton University.

Varian, H. R. (1982). The nonparametric approach to demand analysis. *Econometrica: Journal of the Econometric Society*, (pp. 945–973).

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press.

Von Neumann, J. (1949). On rings of operators. reduction theory. *Annals of Mathematics*, (pp. 401–485).

Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Wang, C., Yuan, A., Cope, L., & Qin, J. (2021). A semiparametric isotonic regression model for skewed distributions with application to dna–rna–protein analysis. *Biometrics*.

Wang, X., Jiang, B., & Liu, J. S. (2017). Generalized r-squared for detecting dependence. *Biometrika*, 104(1), 129–139.

Wang, X., Pan, W., Hu, W., Tian, Y., & Zhang, H. (2015). Conditional distance correlation. *Journal of the American Statistical Association*, 110(512), 1726–1734.

Wasserman, L. (2006). *All of nonparametric statistics*. Springer Science & Business Media.

Watson, D. S. & Wright, M. N. (2019). Testing conditional predictive independence in supervised learning algorithms. *arXiv preprint arXiv:1901.09917*.

Wei, Y., Wainwright, M. J., Guntuboyina, A., et al. (2019). The geometry of hypothesis testing over convex cones: Generalized likelihood ratio tests and minimax radii. *Annals of Statistics*, 47(2), 994–1024.

Whitney, W. F., Song, M. J., Brandfonbrener, D., Altosaar, J., & Cho, K. (2020). Evaluating representations by the complexity of learning low-loss predictors. *arXiv preprint arXiv:2009.07368*.

Williams, E. (1962). Exact fiducial limits in non-linear estimation. *Journal of the Royal Statistical Society: Series B (Methodological)*, 24(1), 125–139.

Williamson, B. & Feng, J. (2020). Efficient nonparametric statistical inference on population feature importance using shapley values. In *International Conference on Machine Learning* (pp. 10282–10291).

Williamson, B. D., Gilbert, P. B., Carone, M., & Simon, N. (2019). Nonparametric variable importance assessment using machine learning techniques. *Biometrics*.

Williamson, B. D., Gilbert, P. B., Simon, N. R., & Carone, M. (2020). A unified approach for inference on algorithm-agnostic variable importance. *arXiv preprint arXiv:2004.03683*.

Xia, Y., Cai, T., & Cai, T. T. (2015). Testing differential networks with applications to the detection of gene-gene interactions. *Biometrika*, 102(2), 247–266.

Xia, Y., Cai, T., & Cai, T. T. (2018). Multiple testing of submatrices of a precision matrix with applications to identification of between pathway interactions. *Journal of the American Statistical Association*, 113(521), 328–339.

Xing, X., Zhao, Z., & Liu, J. S. (2019). Controlling false discovery rate using gaussian mirrors. *arXiv preprint arXiv:1911.09761*.

Yang, F. & Yuan, H. (2017). A non-iterative posterior sampling algorithm for laplace linear regression model. *Communications in Statistics-Simulation and Computation*, 46(3), 2488–2503.

Yang, Z., Ning, Y., & Liu, H. (2018). On semiparametric exponential family graphical models. *Journal of Machine Learning Research*, 19(1), 2314–2372.

Yatchew, A. (1998). Nonparametric regression techniques in economics. *Journal of Economic Literature*, 36(2), 669–721.

Yu, S. (2016). Big privacy: Challenges and opportunities of privacy study in the age of big data. *IEEE access*, 4, 2751–2763.

Yuan, L., Chen, L., Qian, K., Qian, G., Wu, C.-L., Wang, X., & Xiao, Y. (2017). Co-expression network analysis identified six hub genes in association with progression and prognosis in human clear cell renal cell carcinoma (ccrcc). *Genomics Data*, 14, 132–140.

Yuan, M. & Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1), 19–35.

Zappa, E., Holmes-Cerfon, M., & Goodman, J. (2018). Monte carlo on manifolds: sampling densities and integrating functions. *Communications on Pure and Applied Mathematics*, 71(12), 2609–2647.

Zhang, C.-H. (2002). Risk bounds in isotonic regression. *The Annals of Statistics*, 30(2), 528–555.

Zhang, C.-H. & Zhang, S. S. (2014). 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.

Zhang, K., Deng, M., Chen, T., Waterman, M. S., & Sun, F. (2002). A dynamic programming algorithm for haplotype block partitioning. *Proceedings of the National Academy of Sciences*, 99(11), 7335–7339.

Zhang, K., Peters, J., Janzing, D., & Schölkopf, B. (2011). Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence* (pp. 804–813).: AUAI Press.

Zhang, L. & Janson, L. (2020). Floodgate: Inference for model-free variable importance. *arXiv preprint arXiv:2007.01283*.

Zhang, L. & Lu, J. (2021). Startrek: Combinatorial variable selection with false discovery rate control. *arXiv preprint arXiv:2108.09904*.

Zhang, X. & Cheng, G. (2017). Simultaneous inference for high-dimensional linear models. *Journal of the American Statistical Association*, 112(518), 757–768.

Zhang, Z.-M., Chen, X.-Q., Lu, H.-M., Liang, Y.-Z., Fan, W., Xu, D., Zhou, J., Ye, F., & Yang, Z.-Y. (2014). Mixture analysis using reverse searching and non-negative least squares. *Chemometrics and Intelligent Laboratory Systems*, 137, 10–20.

Zhao, Q., Sur, P., & Candes, E. J. (2020). The asymptotic distribution of the mle in high-dimensional logistic models: Arbitrary covariance. *arXiv preprint arXiv:2001.09351*.

Zhao, T., Liu, H., Roeder, K., Lafferty, J., & Wasserman, L. (2012). The huge package for high-dimensional undirected graph estimation in R. *Journal of Machine Learning Research*, 13(1), 1059–1062.

Zhou, S., Rütimann, P., Xu, M., & Bühlmann, P. (2011). High-dimensional covariance estimation based on gaussian graphical models. *The Journal of Machine Learning Research*, 12, 2975–3026.

Zuk, O., Hechter, E., Sunyaev, S. R., & Lander, E. S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences*, 109(4), 1193–1198.