Data-Rich Personalized Causal Inference

by

Abhin Shah

B.Tech., Indian Institute of Technology, Bombay (2018)
S.M., Massachusetts Institute of Technology (2021)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2025

Authored by:	Abhin Shah Department of Electrical Engineering and Computer Science January 10, 2025
Certified by:	Devavrat Shah Andrew (1956) and Erna Viterbi Professor Thesis Supervisor
Certified by:	Gregory W. Wornell Sumitomo Professor of Engineering Thesis Co-Supervisor
Accepted by:	Leslie A. Kolodziejski Professor of Electrical Engineering and Computer Science Chair, Department Committee on Graduate Studies

Data-Rich Personalized Causal Inference

by

Abhin Shah

Submitted to the Department of Electrical Engineering and Computer Science
on January 10, 2025 in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Abstract

There is a growing interest in individual-level causal questions to enable personalized decision-making. For example, what happens to a particular patient's health if we prescribe a drug to them, or what happens to a particular consumer's behavior if we recommend a product to them? Conducting large-scale randomized experiments to answer such questions is impractical—if not infeasible—due to cost, the level of personalization, or ethical concerns. Observational data offer a valuable alternative, but their lack of explicit randomization makes statistical analysis particularly challenging.

In this thesis, we exploit the richness of modern observational data to develop methods for personalized causal inference. In the first part, we introduce a framework for causal inference using exponential family modeling. In particular, we reduce answering causal questions to learning exponential family from one sample. En route, we introduce a computationally tractable alternative to maximum likelihood estimation for learning exponential family. In the second part, we leverage ideas from doubly robust estimation to enable causal inference with black-box matrix completion under a latent factor model.

Thesis supervisors:

Devavrat Shah Andrew (1956) and Erna Viterbi Professor	Gregory W. Wornell Sumitomo Professor of Engineering

Alberto Abadie
Professor of Economics

Acknowledgments

This thesis has been made possible through the support, guidance and encouragement of many individuals, to whom I am deeply grateful. First, I would like to express my profound gratitude to my advisors Devavrat Shah, Greg Wornell, and Alberto Abadie. All three have been a constant source of wisdom, inspiration, and support. Their distinct advising styles have challenged and nurtured me in ways that I could not have anticipated, and I am deeply thankful for the privilege of learning from them. They have been extremely patient with me despite the circuitous route taken by my PhD research. It has truly been an honor to work with all of them and I could not have asked for a better team to guide me through my PhD journey.

I also thank Raaz Dwivedi for his invaluable mentorship as well as friendship throughout my PhD journey. Some of my most cherished research moments were shared with him and I am grateful for his guidance, encouragement, and support. I am also deeply grateful to my mentors from my internships at Google and IBM, as well as my undergraduate advisors, whose guidance and encouragement played a pivotal role in shaping my academic path and inspiring me to embark on this journey.

My heartfelt gratitude goes to my labmates in Devavrat's SPPIN group and Greg's SIA group, both past and present. The countless discussions, collaborations, and shared experiences during group outings have been an integral part of my PhD journey. I have learned and grown immensely, both intellectually and personally, thanks to the camaraderie and insights we have exchanged.

To all my friends who have been by my side throughout this journey—thank you for making these years so much more meaningful and fulfilling. Being away from home is never easy, but your presence has brightened my path and helped me navigate the challenges. The memories we have created together will forever hold a cherished place in my heart.

My deepest gratitude goes to my family for all their love and care, even from halfway across the world. Since my childhood, my parents and my brother have instilled in me the courage to dream big and follow my ambitions. Their countless sacrifices and unwavering support over the years are something I will forever be grateful for. Finally, to Neha, thank you for always being there for me and for being my constant source of strength and encouragement. Your support has meant everything to me.

Grants and fundings: My Ph.D. research was generously supported by the Jacobs Family Presidential Fellowship. In addition, my research was also supported by the following grants: NSF CNS-1523546, NSF DMS-2022448, NSF CCF-1717610, NSF CCF-1816209, ONR N000014-23-1-2803, DARPA HR0011-16-C-0030, MIT-IBM Watson Lab under Agreement No. W1771646, and MIT-KACST project.

Contents

List of Figures

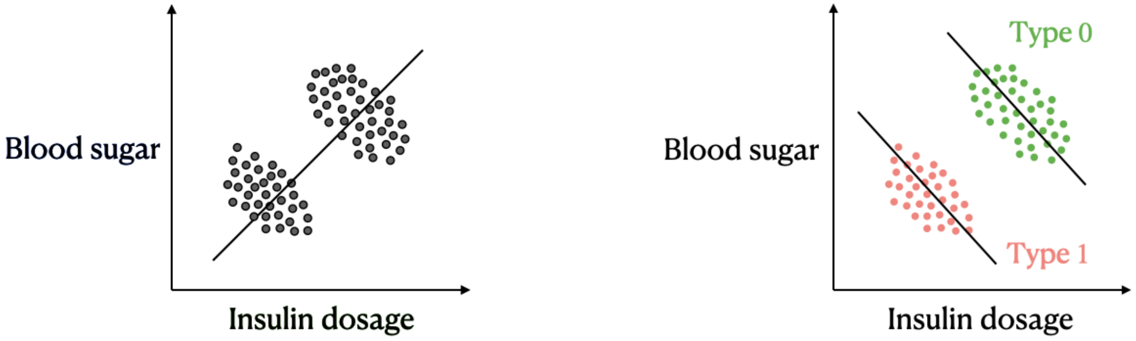
Chapter 1

Introduction

This thesis focuses on questions in causal inference arising in applications such as healthcare, e-commerce, and finance. In particular, we are interested in individual-level what-if questions. For example, *what* happens to a patient's health *if* we prescribe them a drug, or *what* happens to a consumer's behavior *if* we expose them to a product? These personalized inferential tasks, i.e., determining *what happens to an individual's outcome if we do an intervention/action*, are foundational for personalized data-driven decision-making, and complement conventional causal inference methods that focus on population-level inference, e.g., *what happens to the outcome averaged across the population if we do an intervention*?

Experimental data: Experimental data, collected through controlled trials where interventions are systematically assigned to subjects, are the gold standard for causal inference. As the interventions are systematically assigned, often through randomization, the differences between groups can typically be attributed directly due to the intervention. However, while controlled trials are fundamental in the natural and social sciences, they are becoming increasingly costly in engineering and operational contexts. For example, when treatments are continuous and/or high-dimensional, the sheer number of feasible choices limits experimentation. Likewise, as the level of personalization now required grows, conducting experiments is becoming increasingly impractical. For example, the diversity of individuals and their unique contexts would require an enormous number of trials to cover all possible combinations of individual characteristics and interventions. Moreover, assigning actions at random may be ethically concerning in certain domains. For example, it may be unethical to expose individuals to interventions with uncertain risks merely for the sake of experimentation. These considerations make it necessary to explore alternative approaches, such as using observational data, i.e., data acquired without researcher manipulation.

Observational data: Observational data presents its own challenges. The key challenge in deriving causal insights from such data is what is known as confounding. To illustrate this, consider the classic example of Simpson's Paradox. Focusing on the context of diabetes, we are interested in the relationship between insulin dosage and blood sugar levels, as in Figure 1.1. Doctors anticipate that a higher insulin level should lead to lower blood sugar levels. In reality, if you collect raw data and try to fit a



(a) Relationship without grouping data by the type of diabetes.

(b) Relationship after grouping data by the type of diabetes.

Figure 1.1: Illustration of Simpson's Paradox in insulin dosage and blood sugar levels. (a) Relationship without grouping data by the type of diabetes, showing a positive association. (b) Relationship after grouping data by the type of diabetes (Type 0 and Type 1), revealing the expected negative association within each group.

standard ordinary least squares, you would get a plot as shown in Figure 1.1a, which is entirely counter-intuitive. If you examine the data more carefully, you realize that there are two different clusters representing the two types of diabetic patients. Fitting ordinary least squares separately for these two types, indeed recovers the expected behavior, as shown in Figure 1.1b. This is such a paradox because we reach totally opposite conclusions by simply dividing the data into groups.

In short, there was 'confounding' due to the type of diabetes. That is, the type of diabetes introduces spurious associations, as it leads doctors to prescribe different insulin levels and also leads to different responses from patients. If we always know such confounders, then causal inference is simply a prediction problem after appropriately dividing the data into groups. However, in most real-life applications, there are hidden or unobserved confounders, as the reasons for intervention assignment are unknown. In other words, unobserved factors can create spurious associations between interventions and outcomes. This necessitates an emphasis on the principle that *correlation is not causation*, as we can never be certain of recording all the relevant factors. Consequently, we need principled methods to answer causal questions from observational data.

Data-Rich Environments: This thesis builds methods for personalized causal inference, overcoming challenges of observational data by harnessing modern data-rich environments. We define modern data-rich environments as those featuring many outcome measurements across a wide range of units. Our interest in data-rich environments stems from the emergence of digital platforms (e.g., internet retailers, social media companies, and ride-sharing companies), electronic medical records systems, IoT devices, and other real-time digitized data systems, which gather economic and social behavior data with unprecedented scope and granularity.

Consider the following example from mobile health, a personalized healthcare technology, that is gaining prominence lately. Companies like Apple and Fitbit use smart watches to collect massive amounts of observational data. For example, these companies record the exercise routine of individuals over a period of days. Every day, the smartwatch records the exercise performed and the amount of calories burned by an individual. One of the goals of these companies is to provide accessible personalized care to users. For example, there is a growing demand to build systems that recommend personalized workout routines. Considering the exercise sequence as the intervention and the corresponding calories burned as the outcome, a key question that needs to be addressed to build such systems is as follows: what happens to an individual's calorie count if we recommend them a different sequence of exercises? This is an individual-level what-if question as before, but the interventions and the outcomes have naturally become high-dimensional.

The individual's exercise choices and the calories burned are influenced by many factors. For example, stress levels, diet, sleep quality, heart-rate. It turns out that some of these factors, such as stress levels and diet are not recorded by default, and hence act as unobserved confounders. To develop intuition on how such repeated measurements could help perform personalized causal inference in the presence of unobserved confounding, suppose only the (binarized) stress level acts as the unobserved confounder. Consider two simple scenarios. In the first, stress level is fixed over days, say always high. In the second, stress varies with day, say high with probability 1/2 and low with probability 1/2. If we only have one measurement of (exercise, calories) from each of these scenarios, then statistically speaking, it is not easy to distinguish between them. But, if we have repeated measurements of (exercise, calories) from each of these scenarios, then we can actually distinguish the two scenarios. In fact, the more measurements we have, the easier it becomes. In other words, even if the confounder is unobserved, but has some structure to it, it can be exploited with repeated measurements, or in high-dimensions.

In this thesis, we consider scenarios where the amount of variation in the unobserved confounders is appropriately controlled and leverage the availability of many outcome measurements in data-rich environments. Repeated measurements represent just one aspect of modern data's richness. For example, companies like Apple and Fitbit collect extensive smartwatch data across many individuals. Our hope is to extract insights from such data by observing that oftentimes individuals behave in a similar fashion. Finally, we also want to draw power from auxiliary information recorded by the smartwatch, such as sleep-quality and heart-rate. In summary, the depth and breadth of modern observational data offer a timely opportunity to develop methods for personalized causal inference.

Below, we provide an overview of the contributions of this thesis. Primarily, we exploit two different modeling structures to account for the unobserved confounding, namely exponential family modeling and latent factor modeling. In the next chapter (i.e., Chapter 2) of this thesis, we consider the classical problem of learning exponential family distributions in a computationally efficient manner. The resulting estimator lies at the heart of the methodology developed in Chapter 3, where we use exponential family modeling to perform personalized causal inference in scenarios with sequential dependence between interventions and outcomes. In the last chapter (i.e., Chapter 4) of this thesis, we use latent factor modeling to perform doubly robust personalized causal inference.

1.1 Thesis Overview

1.1.1 Computationally Efficient Learning of Exponential Family: Chapter 2

An exponential family is a set of parametric probability distributions, first introduced by Fisher (1934), and later generalized by Darmois (1935), Koopman (1936), and Pitman (1936). Exponential families play an important role in statistical inference and arise in many diverse applications for a variety of reasons. Indeed, they are analytically tractable, arise as the solution to several natural optimization problems on the space of probability distributions, and have robust generalization properties; see, e.g., Barndorff-Nielsen (2014); Brown (1986).

Consider the classical problem of learning the natural parameters of a k-parameter exponential family in a computationally and statistically efficient manner. We focus on the setting where the support as well as the natural parameters and statistics are appropriately bounded. The obvious approach for learning these parameters from independent, identically distributed samples is to use the maximum likelihood estimator (MLE). While MLE has many attractive asymptotic properties such as consistency, asymptotically normality, and asymptotically efficiency, it is not directly applicable in high dimensions due to the computational intractability of calculating the partition function (i.e., the normalization constant) (Jerrum and Sinclair, 1989; Valiant, 1979). In fact, even approximating the partition function, up to a multiplicative error, is NP-hard in general (Sly and Sun, 2012).

Via a novel loss function we develop a computationally and statistically efficient estimator that is consistent as well as asymptotically normal under mild conditions. At the population level, we show that the methodology can be viewed as the maximum likelihood estimation of a re-parameterized distribution belonging to the same class of exponential families. We show further that the estimator can be interpreted as a solution to minimizing a particular Bregman score as well as an instance of minimizing the *surrogate* likelihood of Jeon and Lin (2006). We provide finite sample guarantees to achieve an ℓ_2 error of α in the parameter estimates with sample complexity $O(\mathsf{poly}(k)/\alpha^2)$. Moreover, the method achieves the order-optimal sample complexity $O(\mathsf{log}(k)/\alpha^2)$ when tailored for node-wise sparse Markov random fields (Shah et al., 2021d; Vuffray et al., 2016a, 2022a). A preliminary version of this work appeared in Shah et al. (2021a) and the full version appeared in Shah et al. (2024).

1.1.2 Causal Inference via Exponential Family Modeling: Chapter 3

Given an action-outcome pair, counterfactuals reveal the potential outcome, i.e., the outcome if an intervention (with a different action) had been implemented. Consider a movie streaming platform interacting with a customer, over many days, who watches a movie on the platform daily based on observed and unobserved factors. Given historical data of many customers, the platforms seeks to maximize every customer's viewing time and asks: *what would have happened to each customer's viewing time if they were exposed to a different sequence of movies?* In addition to the spurious associations caused by the unobserved factors, this task is challenging as each customer's viewing time could sequentially depended on prior interactions in addition to the ongoing interaction. Further, each customer provides only a single interaction trajectory and the customers could be heterogeneous in that they may have different responses to same sequence of movies.

The econometrics literature on panel data, where one observes multiple outcomes for each unit, investigates such questions, representing potential outcomes as a tensor with units (individuals), measurements (days), and interventions (movies) as different axes. For linear panel data settings, a common approach is factor modeling, where potential outcomes and interventions (binary or multi-ary) are assumed to be independent conditional on some latent factors. See, e.g., difference-in-difference methods (Angrist and Pischke, 2009; Bertrand et al., 2004), synthetic control (Abadie et al., 2010a; Abadie and Gardeazabal, 2003a), its variants (Arkhangelsky et al., 2021; Dwivedi et al., 2022b), and extensions to multi-ary interventions in synthetic interventions (Agarwal et al., 2020). For non-linear panel data settings, the most commonly used models include probit, logit, Poisson, negative binomial, proportional hazard, and tobit models (see Fernández-Val and Weidner (2018) for an overview) where some parametric model characterises the distribution of the outcomes conditional on the unobserved covariates, the observed covariates, and the interventions. However, these works do not allow outcomes and interventions to explicitly depend on past outcomes and interventions.

We use exponential family to estimate the potential outcome tensor and accommodate (a) sequential dependence of outcomes and interventions on past outcomes and interventions, and (b) unseen interventions from a compact set. We model the conditional distribution of outcomes as an exponential family and reduce learning the potential outcome tensor (with n units and p measurements) to learning parameters of n different distributions from the same exponential family, each with only one p-dimensional sample. Our convex estimator jointly learns all n parameter vectors and results in finite sample recovery rate of $O(p^{-1/2})$ for individual-level mean of outcomes. Our framework extends some of the widely used panel data models from econometrics. In particular, we allow for dynamics in the outcomes, the interventions, and the observed covariates for the linear and logistic unit fixed effect models as well as the linear and logistic time fixed effect models. Further, we allow the causal effect to vary with unit and time for the unit fixed effect models, and the effect to vary with time for the time fixed effect models. Our framework also enables imputing sparsely missing unobserved factors and denoising data with sparse measurement errors. En route, we derive sufficient conditions for compactly supported distributions to satisfy the logarithmic Sobolev inequality. Methodologically, our work generalizes prior work of (a) Dagan et al. (2021); Kandiros et al. (2021) on learning Ising models (and their extensions to discrete, continuous, or mixed variables) from a single sample, where we learn the dependencies between variables and (b) Shah et al. (2021d); Vuffray et al.

(2016a, 2022a) on learning Markov random fields (a sub-class of exponential family) from multiple independent samples, where we allow the samples to be non-identically distributed. This work is currently under review and a preprint version can be found at Shah et al. (2022).

1.1.3 Causal Inference via Latent Factor Modeling: Chapter 4

In causal inference, model-based and design-based are two complementary identification strategies. The former employs restrictions on the process that determines how observed/unobserved factors affect potential outcomes, while the latter employs restrictions on the process that determines how observed/unobserved factors affect intervention assignments. Thus, the two primary approaches to estimating treatment effects are methods based on modeling outcomes and those based on modeling assignments. Consider the example of an internet-retail platform where customers interact with various product categories. For each consumer-category pair, the platform makes decisions to either offer a discount or not, and records whether the consumer purchased a product in the category. Outcome-based methods operate by imputing the missing potential outcomes for each consumer-product category pair. This process involves predicting whether a consumer, who received a discount, would have made the purchase without the discount (i.e., the potential outcome without discount), and conversely, if a consumer who did not receive the discount would have purchased the product had they received the discount (i.e., the potential outcome with discount). In contrast, assignment-based methods estimate the probabilities of consumers receiving discounts in each product category and adjust for missing potential outcomes by weighting observed outcomes inversely to the probability of missingness.

A substantial body of literature has explored outcome-based methods, particularly in settings where all confounding factors are measured (see, e.g., Abadie and Imbens, 2006; Angrist, 1998; Cochran, 1968; Rosenbaum and Rubin, 1983b, among many others). Imputing potential outcomes in the presence of unobserved confounders poses a more complex challenge. In this context, a commonly adopted framework is the synthetic control method and its variants (see, e.g., Abadie et al., 2010a; Abadie and Gardeazabal, 2003a; Arkhangelsky et al., 2021; Cattaneo et al., 2021). An alternative but related approach to outcome imputation under unobserved confounding is the latent factor framework (Bai, 2009; Bai and Ng, 2002; Xiong and Pelger, 2023), wherein each element of the large-dimensional outcome vector is influenced by the same low-dimensional vector of unobserved confounders. Matrix completion methods (see, e.g., Agarwal et al., 2023; Athey et al., 2021; Bai and Ng, 2021; Chatterjee, 2015; Dwivedi et al., 2022a) which have found widespread applications in recommendation systems and panel data models, are closely related to latent factor models. Similarly, existing assignment-based procedures to estimate treatment effects rely on the assumption of no unmeasured confounding (see, e.g., Hirano et al., 2003; Robins et al., 2000; Wooldridge, 2007), common trends restrictions (Abadie, 2005), or the availability of an instrumental variable (Abadie, 2003; Sloczynski et al., 2024). Doubly robust estimators (see Bang and Robins, 2005; Chernozhukov et al., 2018; Robins et al., 1994) combine model-based and design-based strategies to provide estimators that remain consistent as long as either of the two sets of restrictions is correct. In settings with no unobserved factors, these relationships are often estimated using machine learning, a technique known as double machine learning. However, despite their popularity, doubly robust estimators are unavailable for settings with unobserved factors, such as the panel data setting described earlier.

We propose a doubly-robust estimator of treatment effects in the presence of unobserved confounding by leveraging information on both the outcome process and the intervention assignment mechanism under a latent factor framework. The core identification concept is that if each element of a high-dimensional outcome vector is influenced by a common low-dimensional vector of unobserved confounders, it becomes possible to remove the influence of the confounders and identify treatment effects. Our method combines outcome imputation and inverse probability weighting with a new cross-fitting approach for matrix completion. We show that the proposed doubly-robust estimator has better finite-sample guarantees than alternative outcome-based and assignment-based estimators. Furthermore, the doubly-robust estimator is approximately Gaussian, asymptotically unbiased, and converges at a parametric rate, under provably valid error rates for matrix completion, irrespective of other properties of the matrix completion algorithm used for estimation. This work is currently under review and a preprint version can be found at Abadie et al. (2023).

Chapter 2

Computationally Efficient Learning of Exponential Family

2.1 Introduction

Consider a random vector $\mathsf{x} = (x_1, \cdots, x_p)$ with support $\mathcal{X} \subset \mathbb{R}^p$. An exponential family is a set of parametric probability distributions with probability densities of the following canonical form

$$f_{\mathsf{x}}(\boldsymbol{x}; \boldsymbol{\theta}) \propto \exp\bigl(\boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{\phi}(\boldsymbol{x}) + \beta(\boldsymbol{x})\bigr), \tag{2.1}$$

where $\boldsymbol{x} \in \mathcal{X}$ is a realization of the underlying random variable x, $\boldsymbol{\theta} \in \mathbb{R}^k$ is the natural parameter, $\boldsymbol{\phi}\colon \mathcal{X} \to \mathbb{R}^k$ is the natural statistic, k denotes the number of parameters, and β is the log base function. The family is specified by fixing $\mathcal{X}$, $\boldsymbol{\phi}$, and β, and $\boldsymbol{\theta}$ is varied to obtain different distributions within the family.

We focus on exponential families with bounded support as introduced by Hogg and Craig (1956). These so-called "truncated" exponential families share the same parametric form with their non-truncated counterparts up to a normalizing constant, and naturally arise in applications due to limitations of data acquisition.

As we will develop, of interest are families in which there are constraints on the (convex) parameter set $\mathit{\Theta}$ that can, in general, be expressed in terms of a matrix norm bound. Accordingly, for $k = k_1 k_2$ we express (2.1) in the convenient form[1]

$$f_{\mathsf{x}}(\boldsymbol{x}; \boldsymbol{\Theta}) \propto \exp\bigl(\langle \boldsymbol{\Theta}, \boldsymbol{\Phi}(\boldsymbol{x})\rangle\bigr), \qquad \boldsymbol{\Theta} \in \mathit{\Theta},$$

where $\boldsymbol{\Theta} = [\Theta_{ij}] \in \mathbb{R}^{k_1 \times k_2}$ is the natural parameter, $\boldsymbol{\Phi} = [\Phi_{ij}]\colon \mathcal{X} \to \mathbb{R}^{k_1 \times k_2}$ is the natural statistic, and $\bigl\langle \boldsymbol{\Theta}, \boldsymbol{\Phi}(\boldsymbol{x})\bigr\rangle$ denotes the matrix inner product, i.e.,

$$\bigl\langle \boldsymbol{\Theta}, \boldsymbol{\Phi}(\boldsymbol{x})\bigr\rangle = \sum_{i \in [k_1], j \in [k_2]} \Theta_{ij}\, \Phi_{ij}(\boldsymbol{x})$$

[1] For brevity, we absorb the log base function β into the natural statistics and let the corresponding entry of the natural parameter be 1.

We restrict attention to exponential families that are minimal so that the parameters are identifiable. This means there does not exist a nonzero matrix $\mathbf{U} \in \mathbb{R}^{k_1 \times k_2}$ such that $\big\langle \mathbf{U}, \mathbf{\Phi}(\boldsymbol{x}) \big\rangle$ is equal to a constant for all $\boldsymbol{x} \in \mathcal{X}$. Any non-minimal family can be reduced to a minimal one by eliminating components of the natural statistic.

If the natural statistic $\mathbf{\Phi}$ and the support $\mathcal{X}$ are known, then learning a distribution in the exponential family is equivalent to learning the corresponding natural parameter $\mathbf{\Theta}$.

The obvious approach for learning these parameters from independent, identically distributed (i.i.d.) samples is to use the maximum likelihood estimator (MLE) for $f_{\mathsf{x}}(\cdot; \mathbf{\Theta})$. Given i.i.d. data $\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(n)}$, the MLE of $f_{\mathsf{x}}(\cdot; \mathbf{\Theta})$ minimizes the loss function

$$-\frac{1}{n} \sum_{t=1}^{n} \big\langle \mathbf{\Theta}, \mathbf{\Phi}(\boldsymbol{x}^{(t)}) \big\rangle + \ln \int_{\mathcal{X}} \exp\big(\big\langle \mathbf{\Theta}, \mathbf{\Phi}(\boldsymbol{x}) \big\rangle \big) \, \mathrm{d}\boldsymbol{x}. \tag{2.3}$$

The MLE has many attractive asymptotic properties: a) consistency (Ferguson, 2017, Theorem 17), i.e., as the sample size goes to infinity, the bias in the estimated parameters goes to zero; b) asymptotic normality (Ferguson, 2017, Theorem 18), i.e., as the sample size goes to infinity, normalized estimation error coverges to a Gaussian distribution; and c) asymptotic efficiency (Ferguson, 2017, Theorem 20), i.e., as the sample size goes to infinity, the variance in the estimation error attains the minimum possible value among all consistent estimators.

However, the MLE computation is computationally hard (Jerrum and Sinclair, 1989; Valiant, 1979). Indeed, evaluating the objective function (2.3) is generally infeasible due to the presence of the log partition function. In fact, even approximating the partition function, up to a multiplicative error, is NP-hard in general (Sly and Sun, 2012).

As such, the MLE is typically unsuitable for practical use in high-dimensional settings. Accordingly, the focus of this work is on establishing—by construction—the existence of computationally and statistically efficient methods for learning the parameters in such settings.

2.1.1 Some Basic Notation

We use the set notation $[t] \triangleq \{1, \dots, t\}$ for any natural t. Random variables are denoted using sans-serif fonts; e.g., v, and deterministic quantities are denoted using serifed fonts; e.g., v. Vectors are denoted using bold face; e.g., $\boldsymbol{v}$. We use v_i to denote the ith element of $\boldsymbol{v}$, so, e.g., $\boldsymbol{v} = (v_1, \dots, v_t)$. We use

$$\|\boldsymbol{v}\|_q \triangleq \left(\sum_{i=1}^{t} |v_i|^q \right)^{1/q}$$

to denote the ℓ_q norm of $\boldsymbol{v} \in \mathbb{R}^t$ for $q \geq 1$, and

$$\|\boldsymbol{v}\|_\infty \triangleq \max_{i \in [t]} |v_i|$$

to denote the ℓ_∞ norm.

Matrices are denoted using upper-case bold face; e.g., $\mathbf{M}$. For a matrix $\mathbf{M} \in \mathbb{R}^{u \times v}$, we denote the element in ith row and jth column by M_{ij}, and the singular values of the matrix by $\sigma_i(\mathbf{M})$ for $i \in [\min\{u, v\}]$. We denote the matrix maximum norm by

$$\|\mathbf{M}\|_{\max} \triangleq \max_{i \in [u], j \in [v]} |M_{ij}|,$$

the entry-wise $L_{p,q}$ norm by

$$\|\mathbf{M}\|_{p,q} \triangleq \bigg(\sum_{j \in [v]} \bigg(\sum_{i \in [u]} |M_{ij}|^p \bigg)^{q/p} \bigg)^{1/q},$$

the Schatten p-norm by

$$\|\mathbf{M}\|_p^\star \triangleq \bigg(\sum_{i \in [\min\{u,v\}]} \sigma_i^p(\mathbf{M}) \bigg)^{1/p},$$

and the operator norm by

$$\|\mathbf{M}\|_p \triangleq \max_{\boldsymbol{y} \colon \|\boldsymbol{y}\|_p = 1} \|\mathbf{M}\boldsymbol{y}\|_p.$$

For convenience, we denote the Frobenius norm by $\|\mathbf{M}\|_{\mathrm{F}} \triangleq \|\mathbf{M}\|_{2,2}$, the nuclear norm by $\|\mathbf{M}\|^\star \triangleq \|\mathbf{M}\|_1^\star$, and the spectral norm by $\|\mathbf{M}\| \triangleq \|\mathbf{M}\|_2$.

We denote the Frobenius or trace inner product of matrices $\mathbf{M}, \mathbf{N} \in \mathbb{R}^{u \times v}$ by

$$\langle \mathbf{M}, \mathbf{N} \rangle \triangleq \sum_{i \in [u], j \in [v]} M_{ij} N_{ij}.$$

More generally, we use $\mathcal{R}(\mathbf{M})$ to denote a generic matrix norm, and $\mathcal{R}^*(\mathbf{M}) \triangleq \sup_{\mathbf{N}} \{\langle \mathbf{M}, \mathbf{N} \rangle \colon \mathcal{R}(\mathbf{N}) \leq 1\}$ to denote the corresponding dual norm.

We denote the vectorization of a matrix $\mathbf{M} \in \mathbb{R}^{u \times v}$ by $\mathrm{vec}(\mathbf{M}) \in \mathbb{R}^{uv \times 1}$, the ordering of the elements within which is not important as long as it is consistent. We use $\mathbf{0} \in \mathbb{R}^{k_1 \times k_2}$ to denote the matrix whose entries are all zero. Finally, we use $\mathcal{B}_q(b)$ to denote a p-dimensional ℓ_q ball of radius b centered at $\mathbf{0} \in \mathbb{R}^p$, for $q \in \{1, 2\}$.

2.1.2 Outline

The remainder of this chapter is organized as follows. Section 2.2 provides, for convenience, a concise summary of the chapter's contributions, and Section 2.3 contains a summary of related work as context for the present development. In Section 2.4, we formulate the problem of interest and provide examples. In Section 2.5, we describe the proposed learning methodology, and in Section 2.6, we provide our analysis and key results including the connections to the MLE of of a re-parameterized distribution, and our development of consistency, asymptotic normality, and finite sample guarantees. In Section 2.7, we provide our empirical findings. Finally, Section 2.8 contains some concluding remarks.

2.2 Summary of Contributions

The contributions of this chapter include the following.

2.2.1 Estimator

Given samples $\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(n)}$ from $f_{\mathbf{x}}(\cdot; \boldsymbol{\Theta}^*)$ for $\boldsymbol{\Theta}^* \in \varTheta$, we develop and analyze learning a member of the exponential family (2.2) with parameter set $\varTheta$ via the estimator

$$\hat{\boldsymbol{\Theta}}_n \in \arg\min_{\boldsymbol{\Theta} \in \varTheta} \mathcal{L}_n(\boldsymbol{\Theta}) \tag{2.4a}$$

with the convex loss function

$$\mathcal{L}_n(\boldsymbol{\Theta}) = \frac{1}{n} \sum_{t=1}^{n} \exp\bigl(-\bigl\langle \boldsymbol{\Theta}, \boldsymbol{\varPhi}(\boldsymbol{x}^{(t)}) \bigr\rangle\bigr), \tag{2.4b}$$

where

$$\boldsymbol{\varPhi}(\cdot) \triangleq \boldsymbol{\Phi}(\cdot) - \mathbb{E}_{\mathcal{U}_{\mathcal{X}}}[\boldsymbol{\Phi}(\cdot)] \tag{2.4c}$$

are the centered statistics,[2] with $\mathcal{U}_{\mathcal{X}}$ denoting the uniform distribution on $\mathcal{X}$. We establish that $\hat{\boldsymbol{\Theta}}_n$ is consistent and (under mild further restrictions) asymptotically normal (see Theorem 2.2 in Section 2.6.4). We show that the loss function in (2.4b) is a smooth function of $\boldsymbol{\Theta}$ (see Proposition 2.1 in Section 2.5) and satisfies a restricted strong convexity property (see Proposition 2.2 in Section 2.5). This implies, using Agarwal et al. (2010), that an ϵ-optimal solution $\hat{\boldsymbol{\Theta}}_{\epsilon,n}$ can be obtained in $O(\log(1/\epsilon))$ iterations. In addition, we provide rigorous finite sample guarantees for $\hat{\boldsymbol{\Theta}}_n$ to achieve an error of α (in the Frobenius norm) with respect to the true natural parameter $\boldsymbol{\Theta}^*$ with $O(\mathrm{poly}(k_1 k_2)/\alpha^2)$ samples (see Theorem 2.3 in Section 2.6.5). We note the loss function in (2.4b) is a generalization of the loss functions proposed in Shah et al. (2021d); Vuffray et al. (2022b, 2016b) for learning node-wise-sparse Markov random fields (MRFs). In particular, Theorem 2.3 can be specialized to recover the natural parameters of sparse MRFs with $O(\log(k_1 k_2)/\alpha^2)$ samples (see Remark 2.3 in Section 2.6.5). Beyond local structure on $\boldsymbol{\Theta}$ (of node-wise-sparsity) as in MRFs, our framework can capture various forms of global structure on $\boldsymbol{\Theta}$, e.g., bounded maximum norm, bounded Frobenius norm, bounded nuclear norm (see Section 2.4.1).

2.2.2 Connections

We establish relationships between our method and various existing methods in the literature. First, we show that the estimator that minimizes the population version of the loss function in (2.4b), viz.,

$$\mathcal{L}(\boldsymbol{\Theta}) = \mathbb{E}\bigl[\exp\bigl(-\bigl\langle \boldsymbol{\Theta}, \boldsymbol{\varPhi}(\mathbf{x}) \bigr\rangle\bigr)\bigr], \tag{2.5}$$

[2] This centering plays a key role in ensuring that our loss function is proper.

is equivalent to the estimator that minimizes the Kullback-Leibler (KL) divergence

$$D\big(\mathcal{U}_{\mathcal{X}}(\cdot)\big\|f_{\mathbf{x}}(\cdot;\boldsymbol{\Theta}^* - \boldsymbol{\Theta})\big),$$

(see Theorem 2.1 in Section 2.6.1). Accordingly, at the population level, the method can be viewed as the MLE of the parametric family $f_{\mathbf{x}}(\cdot;\boldsymbol{\Theta}^* - \boldsymbol{\Theta})$. It follows from this divergence relation that $\mathcal{L}(\boldsymbol{\Theta})$ is minimized if and only if $\boldsymbol{\Theta} = \boldsymbol{\Theta}^*$, which in turn implies that $\mathcal{L}(\boldsymbol{\Theta})$ is a proper loss function. This connection provides intuitive justification for the estimator (2.4). Second, we demonstrate that the estimator in (2.4) can be interpreted as a solution to minimizing a particular Bregman score, and thus connect our method with score-based methods (see Proposition 2.4 in Section 2.6.2). Phrased differently, we show that optimizing a specific separable Bregman score is equivalent to the learning task of interest. And as a result, our work establishes the computational tractability of this score. Finally, we show that our estimator can be viewed as an instance of the surrogate likelihood estimator proposed by Jeon and Lin Jeon and Lin (2006), and thus draw connections with nonparametric density estimation (see Proposition 2.5 in Section 2.6.3).

2.3 Background and Related Work

First, there are two broad lines of approaches to overcome the computational hardness of the MLE: approximating the MLE; and selecting a surrogate objective. In the sequel, we provide a summary of representative examples; further discussion is deferred to Appendix 2.A.

2.3.1 Approximating the MLE

Techniques in this category typically approximate the MLE by approximating the log-partition function. Examples include: approximating the gradient of log-likelihood with a stochastic estimator by minimizing the contrastive divergence (Hinton, 2002); upper bounding the log-partition function by an iterative tree-reweighted belief propagation algorithm (Wainwright et al., 2003); and using Monte Carlo methods like importance sampling for estimating the partition function (Robert and Casella, 2013). Since these methods approximate the partition function, they come at the cost of an approximation error or result in a biased estimator.

2.3.2 Selecting a surrogate objective

This approach avoids the partition function computation by selecting a surrogate objective that is easier to compute. One class of examples are pseudo-likelihood estimators, which approximate the joint distribution with the product of conditional distributions, each of which only represents the distribution of a single variable conditioned on the remaining variables (Besag, 1975a). Another class are score-matching estimators, which minimizes the Fisher divergence between the true log density and the model log density (Hyvärinen, 2007; Hyvärinen and Dayan, 2005). Even though score-matching does not require evaluating the partition function, it is computationally expensive as it requires computing third order derivatives for optimization. And yet another class are estimators based on kernel Stein discrepancy, which measures the kernel mean discrepancy between a data distribution and a model density using Stein's identity (Chwialkowski et al., 2016; Liu et al., 2016). This measure is directly characterized by the choice of the kernel; see Appendix 2.A.2 for further discussion.

Next, there are a variety of methods for learning classes of exponential families with specific structure. Representative examples are as follows.

2.3.3 Learning the Gaussian distribution

The literature on learning exponential family distributions with unbounded support has been largely restricted to Gaussian distributions. Moreover, learning such distributions has commonly focused on those described by sparse graphical models. Examples include the neighborhood selection scheme (Meinshausen et al., 2006), graphical lasso (Friedman et al., 2008), and constrained ℓ_1-minimization for inverse matrix estimation (CLIME) (Cai et al., 2011). However, finite sample analysis of these methods requires various conditions that are hard to verify in practice, such as the incoherence assumption (see Jalali et al. (2011); Wainwright et al. (2006)) and the precision matrix having bounded condition number. In a recent work, Misra et al. (2020) provides the first polynomial-time algorithm whose sample complexity matches the information-theoretic lower bound of Wang et al. (2010) without the aforementioned conditions. A faster alternative to Misra et al. (2020), for a specific subclass of Gaussian graphical models, is proposed in Kelner et al. (2020).

There has also been a similarly long history of learning truncated Gaussian distributions dating back to Galton (1898), Pearson (1902); Pearson and Lee (1908), Lee (1914), and Fisher (1931). More recently, Daskalakis et al. (2018) shows that it is possible to learn, in polynomial time, the mean vector and the covariance matrix of a p-dimensional truncated Gaussian distribution, up to an ℓ_2 error of α with $O(p^2/\alpha^2)$ samples—i.e., with a sample complexity of the same order as when there is no truncation.

2.3.4 Learning sparse MRFs

MRFs are an important class of exponential family distributions, and often arise out of maximum entropy formulations; see, e.g., Wainwright and Jordan (2008). A popular method for learning node-wise sparse MRFs is estimating node-neighborhoods (i.e., fitting conditional distributions of each node conditioned on the rest of the nodes). More recent work considers a subclass of node-wise sparse pairwise continuous MRFs in which the node-conditional distribution of $\mathsf{x}_i \in \mathcal{X}_i$ for every $i \in [p]$ arises from the exponential family

$$f_{\mathsf{x}_i|\mathsf{x}_{-i}}(x_i|x_{-i}) \propto \exp\bigg(\Big[\Theta_i + \sum_{j\in[p], j\neq i} \Theta_{ij}\, \phi(x_j)\Big]\phi(x_i)\bigg), \tag{2.6}$$

where $\phi(x_i)$ is the natural statistics and

$$\Theta_i + \sum_{j \in [p], j \neq i} \Theta_{ij}\, \phi(x_j)$$

is the natural parameter.[3] Yang et al. (2015) show that only the joint distribution

$$f_{\mathbf{x}}(\boldsymbol{x}) \propto \exp\biggl(\sum_{i \in [p]} \Theta_i\, \phi(x_i) + \sum_{j \neq i} \Theta_{ij}\, \phi(x_i)\, \phi(x_j)\biggr). \tag{2.7}$$

is consistent with the node-conditional distributions (2.6). In turn, to learn the distribution (2.7) for linear $\phi(\cdot)$, Yang et al. (2015) proposes an ℓ_1 regularized node-conditional log-likelihood. However, the associated finite sample analysis requires the following conditions: incoherence, dependency (see Jalali et al. (2011); Wainwright et al. (2006)), bounded moments of the variables, and local smoothness of the log-partition function.

Tansey et al. (2015) extend the approach in Yang et al. (2015) to vector-space MRFs (i.e., vector natural parameters and natural statistics) and nonlinear $\phi(\cdot)$. In partiular, they propose a sparse group lasso (Simon et al., 2013), regularized node-conditional log-likelihood, and an alternating direction method of multipliers (ADMM) based approach for solving the associated optimization problem. However, the analysis continues to require the conditions of Yang et al. (2015).

While node-conditional log-likelihood has been a natural choice for learning exponential family MRFs of the form (2.7), M-estimation (Shah et al., 2021d; Vuffray et al., 2022b, 2016b) and maximum pseudo-likelihood estimation (Dagan et al., 2021; Ning et al., 2017; Yang et al., 2018) have recently gained popularity. The objective function in M-estimation is a sample average and the estimator is generally consistent and asymptotically normal. Shah et al. (2021d) propose the following M-estimation (inspired by Vuffray et al. (2022b, 2016b)) for vector-space MRFs and nonlinear $\phi(\cdot)$:

$$\min \frac{1}{n} \sum_{t=1}^{n} \exp\biggl(-\Theta_i\, \varphi(x_i^{(t)}) - \sum_{j \in [p], j \neq i} \Theta_{ij}\, \varphi(x_i^{(t)})\, \varphi(x_j^{(t)})\biggr),$$

with

$$\varphi(x_i) \triangleq \phi(x_i) - \int \phi(x_i')\, \mathcal{U}_{\mathcal{X}_i}(x_i')\, \mathrm{d}x_i'.$$

An entropic descent algorithm (borrowed from Vuffray et al. (2022b)) is used to solve the optimization of (2.8), and the associated finite-sample bounds rely on the bounded domain of the variables and a variance lower condition (which is naturally satisfied by linear $\phi(\cdot)$).

Yuan et al. (2016) consider sparse pairwise exponential family MRFs of the form

$$f_{\mathbf{x}}(\boldsymbol{x}) \propto \exp\biggl(\sum_{i \in [p]} \Theta_i\, \phi(x_i) + \sum_{j \neq i} \Theta_{ij}\, \psi(x_i, x_j)\biggr), \tag{2.9}$$

[3] Under node-wise sparsity, $\sum_{j \in [p], j \neq i} \mathbb{1}(\Theta_{ij} \neq 0)$ is bounded by a constant for every $i \in [p]$.

which is a broader class than those described by (2.7). For this class, they propose an $\ell_{2,1}$ regularized joint likelihood and an $\ell_{2,1}$ regularized node-conditional likelihood. They further develop Monte-Carlo approximations via proximal gradient descent. The correpsonding finite-sample analysis requires restricted strong convexity (of the Hessian of the negative log-likelihood of the joint density) and bounded moment-generating function of the variables.

Building upon Shah et al. (2021d); Vuffray et al. (2022b, 2016b), Ren et al. (2021) study the learning of continuous exponential family distributions through a series of numerical experiments. They consider unbounded distributions and allow for terms corresponding to multi-wise interactions in the joint density. They assume local structure on the parameters as in MRFs and their estimator is defined via a series of node-wise optimization problems. Notably, Ren et al. (2021) reports that among the methods described above, those based on M-estimation have superior numerical performance compared to the ones based on pseudo-likelihood estimation.

Likewise, motivated by causal inference applications and building upon Shah et al. (2021d) and Dagan et al. (2021), Shah et al. (2022) consider learning the node-conditional distribution corresponding to the joint distribution in (2.7) when certain variables remain unobserved, and provide finite sample guarantees for learning the counterfactual means.

We emphasize that the above developments are limited to scenarios in which the natural parameters are node-wise sparse, and so none apply to setting in which the natural parameters have, e.g., a bounded nuclear norm (corresponding to a convex relaxation of a low-rank constraint). Such constraints are among those of interest in the present work.

2.3.5 Score-based method

Score-based methods are also applicable to learning exponential families. A scoring rule $S(\boldsymbol{x}, Q)$ is a numerical score assigned to a realization $\boldsymbol{x}$ of a random variable $\mathbf{x}$ and it measures the quality of a predictive distribution Q (for which the probability density is $q(\cdot)$). If P is the true distribution of $\mathbf{x}$, the divergence associated with a scoring rule is defined as

$$D_S(P\|Q) \triangleq \mathbb{E}_P[S(\mathbf{x}, Q) - S(\mathbf{x}, P)].$$

Within this framework, the MLE is an example of a scoring rule with the choice $S(\cdot, Q) = -\log q(\cdot)$ and the resulting divergence is the KL divergence.

To avoid the intractability of MLE, Hyvärinen (2007); Hyvärinen and Dayan (2005) propose the scoring rule

$$S(\cdot, Q) = \Delta \log q(\cdot) + \frac{1}{2}\big\|\nabla \log q(\cdot)\big\|_2^2,$$

where Δ is the Laplacian operator and ∇ is the gradient. This method is called *score-matching* and the resulting divergence is the Fisher divergence. Score-matching is widely used for estimating unnormalized probability distributions because computing the scoring rule $S(\cdot, Q)$ does not require knowing the partition function. Despite the flexibility of this approach, it is computationally expensive in high dimensions since it requires computing the trace of the unnormalized density's Hessian (and its derivatives for optimization). Additionally, it breaks down for models in which the second derivative grows very rapidly.

Truncated exponential family distributions are learned using the principle of score-matching in Liu et al. (2022). This work builds on the framework of generalized score-matching (Hyvärinen, 2007) and proposes a novel estimator that minimizes a weighted Fisher divergence. It is shown that the estimator is a special case of the one minimizing a Stein discrepancy. However, the associated finite sample analysis also relies on assumptions that are hard to verify—e.g., the assumption that the optimal parameter is well-separated from other neighboring parameters in terms of the population objective.

2.3.6 Nonparametric density estimation

Finally, exponential families can also be learned via approaches based on nonparametric density estimation. Goodd and Gaskins (1971) introduce the idea of penalized log-likelihood for nonparametric density estimation. The logistic density transform methodology, commonly used today for nonparametric density estimation, was introduced by Leonard (1978) to incorporate the positivity ($f_{\mathbf{x}}(\cdot) \geq 0$) and normalization ($\int_{\mathcal{X}} f_{\mathbf{x}}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} = 1$) constraints. They considered densities of the form

$$f_{\mathbf{x}}(\boldsymbol{x}) \propto \exp(\eta(\boldsymbol{x})), \tag{2.10}$$

with some constraints on $\eta(\cdot)$ for identifiability, and propose to estimate $\eta(\cdot)$ by minimizing the penalized log likelihood

$$\frac{1}{n}\sum_{i=1}^{n} \eta(\boldsymbol{x}_i) + \log \int_{\mathcal{X}} \exp(\eta(\boldsymbol{x})) \, \mathrm{d}\boldsymbol{x} + \lambda J(\eta), \tag{2.11}$$

where $\lambda \geq 0$ is a smoothing parameter and $J(\eta)$ is a penalty functional.

While this method has been successful in low dimensions, it scales poorly in higher dimensions. In high-dimensional problems, the main difficulty is in computing the requisite multidimensional integral in (2.11), which does not decompose in general. To circumvent this computational limitation, Jeon and Lin (2006) propose a penalized M-estimation (surrogate likelihood) method that minimizes (over η)

$$\frac{1}{n}\sum_{i=1}^{n} \exp(-\eta(\boldsymbol{x}_i)) + \int_{\mathcal{X}} \rho(\boldsymbol{x}) \, \eta(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} + \lambda \, J(\eta), \tag{2.12}$$

where $\eta(\cdot)$ lies in a reproducing kernel Hilbert space (RKHS) and $\rho(\cdot)$ is some fixed known density with the same support as the unknown density (2.10). The resulting density estimate in this formulation is $\hat{f}_{\mathbf{x}}(\boldsymbol{x}) \propto \rho(\boldsymbol{x}) \exp(\hat{\eta}(\boldsymbol{x}))$. It is shown that with appropriate choices of $\rho(\cdot)$, the integral $\int_{\mathcal{X}} \rho(\boldsymbol{x}) \, \eta(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}$ can be decomposed into sums of products of one-dimensional integrals, allowing faster computation. However, the selection of λ that delivers reasonable performance requires the evaluation of the normalization $\int_{\mathcal{X}} \rho(\boldsymbol{x}) \exp(\eta(\boldsymbol{x})) \, \mathrm{d}\boldsymbol{x}$. Additionally, the properties of the estimator in (2.12) are generally not yet known.

2.4 Problem Formulation

To start, we let $\mathsf{x} = (\mathsf{x}_1, \dots, \mathsf{x}_p)$ represent a p-dimensional vector of continuous random variables.[4] For any $i \in [p]$, we use $\mathcal{X}_i$ to denote the support of x_i, which is a bounded measurable subset of $\mathbb{R}$. And $\boldsymbol{x} = (x_1, \dots, x_p) \in \mathcal{X} \triangleq \mathcal{X}_1 \times \dots \times \mathcal{X}_p$ denotes a realization of x.

With $\boldsymbol{\Theta}^*$ denoting the true natural parameter, the learning task is as follows. Given n independent samples $\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(n)}$ from $f_{\mathsf{x}}(\boldsymbol{x}; \boldsymbol{\Theta}^*)$ of the form (2.2), compute an estimate $\hat{\boldsymbol{\Theta}}$ of $\boldsymbol{\Theta}^*$ in polynomial time such that $\|\boldsymbol{\Theta}^* - \hat{\boldsymbol{\Theta}}\|_{\mathrm{F}}$ is arbitrarily small.

In this formulation, the family (2.2) is minimal and truncated, and both the natural statistic $\boldsymbol{\Phi}$ and the support $\mathcal{X}$ are known. In the sequel, we describe additional constraints we impose on the natural parameter and statistic.

2.4.1 Natural parameter $\boldsymbol{\Theta}$

We restrict our attention to convex parameter sets $\varTheta$ such that a suitable norm of the natural parameter $\boldsymbol{\Theta} \in \mathbb{R}^{k_1 \times k_2}$ is bounded. Examples include bounded maximum norm $\|\boldsymbol{\Theta}\|_{\max}$, bounded Frobenius norm $\|\boldsymbol{\Theta}\|_{\mathrm{F}}$, and bounded nuclear norm $\|\boldsymbol{\Theta}\|^{\star}$.

Assumption 2.1 (Bounded norm of $\boldsymbol{\Theta}$)**.**

$$\varTheta \triangleq \big\{\boldsymbol{\Theta} \colon \mathcal{R}(\boldsymbol{\Theta}) \le r\big\},$$

where $\mathcal{R} \colon \mathbb{R}^{k_1 \times k_2} \to \mathbb{R}_+$ is a norm and r is a known constant.

Assumption 2.1 provides flexibility in the problem specification; a practitioner has the option to choose from a variety of constraints on the natural parameters (that can be handled by our framework). For example, in some applications every entry of the parameter is bounded while in some other case the sum of singular values of the parameter matrix is bounded (convex relaxation of low-rank parameter matrix).

Remark 2.1. *We note that p is not assumed to be a constant. Instead, we think of k_1 and k_2 as implicit functions of p. Typically, for an exponential family, the quantity of interest is the number of parameters, i.e., $k = k_1 k_2$, and this quantity scales polynomially in p, e.g., $k = O(p^t)$ for t-wise MRFs over binary alphabets (see Section 2.7).*

2.4.2 Natural statistic $\boldsymbol{\Phi}$

We further restrict our attention to suitably bounded natural statistic $\boldsymbol{\Phi}(\boldsymbol{x}) \colon \mathcal{X} \to \mathbb{R}^{k_1 \times k_2}$. We combine two notions of boundedness. First, we assume that the dual norm (defined with respect to $\mathcal{R}$ used in the definition of the parameter set in Assumption 2.1) of the natural statistic is bounded:[5]

[4] Even though we focus on continuous variables, our framework applies equally to discrete or mixed variables.

[5] This enables us to bound the matrix inner product between the natural parameter $\boldsymbol{\Theta}$ and the natural statistic $\boldsymbol{\Phi}(\cdot)$.

Assumption 2.2 (Bounded dual norm of $\mathbf{\Phi}$)**.** *The dual norm $\mathcal{R}^*$ of the natural statistic $\mathbf{\Phi}$ is bounded by a constant $\bar{r}$. Formally,*

$$\mathcal{R}^*(\mathbf{\Phi}(\boldsymbol{x})) \le \bar{r},$$

for any $\boldsymbol{x} \in \mathcal{X}$.

Examples of dual norms include the $L_{1,1}$ norm $\|\mathbf{\Phi}(\boldsymbol{x})\|_{1,1}$, the Frobenius norm $\|\mathbf{\Phi}(\boldsymbol{x})\|_{\mathrm{F}}$, and the spectral norm $\|\mathbf{\Phi}(\boldsymbol{x})\|$, when the underlying norm $\mathcal{R}$ is the maximum norm $\|\mathbf{\Theta}\|_{\max}$, the Frobenius norm $\|\mathbf{\Theta}\|_{\mathrm{F}}$, and the nuclear norm $\|\mathbf{\Theta}\|^{\star}$, respectively. While we require this assumption for our analysis, our empirical findings (later in Section 2.7) suggest that the assumption may not be a strict requirement in practice.

Second, we require that the maximum norm of the natural statistic $\mathbf{\Phi}(\cdot)$ also be bounded:

Assumption 2.3 (Bounded maximum norm of $\mathbf{\Phi}$)**.** *For any $\boldsymbol{x} \in \mathcal{X}$,*

$$\|\mathbf{\Phi}(\boldsymbol{x})\|_{\max} \le \phi_{\max},$$

where $\phi_{\max}$ is a constant.

Examples of natural statistics and their support satisfying Assumptions 2.2 and 2.3 include both polynomial and trigonometric statistics; see Appendix 2.B for further discussion.

Finally, we further restrict attention to the case in which the autocorrelation matrix of $\mathrm{vec}(\boldsymbol{\Phi}(\mathbf{x}))$ has positive eigenvalues.

Assumption 2.4 (Positive definite $\boldsymbol{\Phi}$ autocorrelation)**.** *The minimum eigenvalue $\lambda_{\min}$ of $\mathbb{E}_{\mathbf{x}}[\mathrm{vec}(\boldsymbol{\Phi}(\mathbf{x}))\,\mathrm{vec}(\boldsymbol{\Phi}(\mathbf{x}))^{\mathrm{T}}]$ is positive.*

2.5 Learning Algorithm

Our learning algorithm (2.4) draws inspiration from the recent advancements in learning sparse MRFs Shah et al. (2021d); Vuffray et al. (2022b, 2016b). The loss function $\mathcal{L}_n(\mathbf{\Theta})$, defined in (2.4b), is an empirical average of the inverse of the functional of $\mathbf{x}$ to which the probability density $f_{\mathbf{x}}(\boldsymbol{x};\mathbf{\Theta})$ in (2.2) is proportional (with centered natural statistic (2.4c)). As defined in (2.4a), the associated estimator $\hat{\mathbf{\Theta}}_n$, minimizes $\mathcal{L}_n(\mathbf{\Theta})$ over all $\mathbf{\Theta}$ satisfying Assumption 2.1.

Note that (2.4) is a convex optimization problem, i.e., we are minimizing a convex function $\mathcal{L}_n$ over a convex set $\mathit{\Theta}$. Moreover, the loss function $\mathcal{L}_n(\mathbf{\Theta})$ has key additional structure. First, $\mathcal{L}_n(\mathbf{\Theta})$ obeys the following smoothness property, which we verify in Appendix 2.C.

Proposition 2.1 (Smoothness of $\mathcal{L}_n$)**.** *Consider any $\mathbf{\Theta} \in \mathit{\Theta}$. Under Assumptions 2.1, 2.2, and 2.3, $\mathcal{L}_n(\mathbf{\Theta})$ is a $4k_1k_2\phi_{\max}^2\mathrm{e}^{2r\bar{r}}$ smooth function of $\mathbf{\Theta}$.*

Second, $\mathcal{L}_n(\mathbf{\Theta})$ obeys the following restricted strong convexity property with high probability, which we verify in Appendix 2.D.

Proposition 2.2 (Restricted strong convexity of $\mathcal{L}_n$)**.** *Suppose Assumptions 2.1, 2.2, 2.3, and 2.4 are satisfied. Define*

$$\gamma(k_1, k_2) = \max_{\mathbf{M} \in 4\varTheta \setminus \{\mathbf{0}\}} \frac{\|\mathbf{M}\|_{1,1}}{\|\mathbf{M}\|_{\mathrm{F}}}. \tag{2.13}$$

Consider any $\boldsymbol{\Theta} \in \mathbb{R}^{k_1 \times k_2}$ such that $\Delta \triangleq \boldsymbol{\Theta} - \boldsymbol{\Theta}^*$ satisfies $\Delta \in 4\varTheta$. Then, for any fixed $\delta \in (0,1)$ and provided

$$n > \frac{128\phi_{\max}^4 \gamma(k_1, k_2)^4}{\lambda_{\min}^2} \log\Big(\frac{2k_1^2 k_2^2}{\delta}\Big),$$

the residual loss

$$\delta\mathcal{L}_n(\Delta, \boldsymbol{\Theta}^*) \triangleq \mathcal{L}_n(\boldsymbol{\Theta}^* + \Delta) - \mathcal{L}_n(\boldsymbol{\Theta}^*) - \langle \nabla\mathcal{L}_n(\boldsymbol{\Theta}^*), \Delta \rangle \tag{2.14}$$

satisfies

$$\delta\mathcal{L}_n(\Delta, \boldsymbol{\Theta}^*) \geq \frac{\lambda_{\min} \mathrm{e}^{-2r\overline{r}}}{4 + 16r\overline{r}} \|\Delta\|_{\mathrm{F}}^2,$$

with probability at least $1 - \delta$.

As a result of the properties established by Propositions 2.1 and 2.2, there exist efficient implementations for finding an ϵ-optimal solution of $\hat{\boldsymbol{\Theta}}_n$.[6] In particular, from (Agarwal et al., 2010, Theorem 1), the run-time of such an implementation scales as $O(\log(1/\epsilon))$ under these properties. Furthermore, from Slater's condition (which holds because $\mathrm{int}(\varTheta) \neq \emptyset$), we can express $\hat{\boldsymbol{\Theta}}_n$ as a solution to the following unconstrained optimization:

$$\hat{\boldsymbol{\Theta}}_n \in \operatorname*{arg\,min}_{\boldsymbol{\Theta} \in \mathbb{R}^{k_1 \times k_2}} \mathcal{L}_n(\boldsymbol{\Theta}) + \lambda_n \mathcal{R}(\boldsymbol{\Theta}), \tag{2.15}$$

where λ_n is a regularization penalty. As we develop in Section 2.6, $\hat{\boldsymbol{\Theta}}_n$ is close to $\boldsymbol{\Theta}^*$ (in Frobenius norm) when λ_n is appropriately chosen. In particular, from (Negahban et al., 2012, Corollary 1), this is possible whenever the loss function satisfies restricted strong convexity and the regularization penalty is such that $\lambda_n \geq 2\mathcal{R}^*(\nabla\mathcal{L}_n(\boldsymbol{\Theta}^*))$. We note that, as required, the addition of the regularization preserves restricted strong convexity of the optimization in (2.15) due to convexity of norms. In turn, to choose an appropriate λ_n, we define

$$g(k_1, k_2) = \max_{\mathbf{M} \in \mathbb{R}^{k_1 \times k_2} \setminus \{\mathbf{0}\}} \frac{\mathcal{R}^*(\mathbf{M})}{\|\mathbf{M}\|_{\max}}. \tag{2.16}$$

Then, it suffices to bound $g(\cdot, \cdot)$ and the maximum norm of the gradient of $\mathcal{L}_n(\boldsymbol{\Theta})$, evaluated at the true natural parameter. We bound the former in Section 2.6 and the latter below, with a proof in Appendix 2.E.

[6] Recall that $\hat{\boldsymbol{\Theta}}_{\epsilon,n}$ is an ϵ-optimal solution of $\hat{\boldsymbol{\Theta}}_n$ if $\mathcal{L}_n(\hat{\boldsymbol{\Theta}}_{\epsilon,n}) \leq \mathcal{L}_n(\hat{\boldsymbol{\Theta}}_n) + \epsilon$ for any $\epsilon > 0$.

Proposition 2.3 (Bounded $\|\nabla \mathcal{L}_n(\boldsymbol{\Theta}^*)\|_{\max}$)**.** *Suppose Assumptions 2.1, 2.2, and 2.3 are satisfied. Fix any $\epsilon > 0$ and $\delta \in (0,1)$. Then provided*

$$n > \frac{8\phi_{\max}^2 \exp(4r\bar{r})}{\epsilon^2} \log\Big(\frac{2k_1k_2}{\delta}\Big),$$

we have[7]

$$\|\nabla \mathcal{L}_n(\boldsymbol{\Theta}^*)\|_{\max} \leq \epsilon,$$

with probability at least $1-\delta$.

In our implementation, we obtain the ϵ-optimal solution $\hat{\boldsymbol{\Theta}}_{\epsilon,n}$ using projected gradient descent under the assumption that r is known, i.e., we perform the optimization in (2.4). Alternatively, one could obtain $\hat{\boldsymbol{\Theta}}_{\epsilon,n}$ by performing the optimization in (2.15), where the appropriate choice of λ_n needs the knowledge of a lower bound on $r\bar{r}$; see Section 2.6 for details.

Remark 2.2. *While (2.15) is a convex optimization problem, computing the loss function as well as its gradient requires centering of the natural statistics. If the natural statistics are polynomials or trigonometric, centering them is relatively straightforward since the expectation (2.4c) can be evaluated in closed-form in these cases. Such statistics are prevalent in a broad segment of the literature, including studies on Ising models and MRFs more generally. For other statistics, centering may be more difficult. Examples include functions featuring exponentials of random variables which do not yield a tractable integral, or functions with discontinuities or other behaviors that preclude straightforward analytical integration and necessitate specialized numerical techniques. Alternatively, one might require assuming the existence of computationally efficient sampling or that obtaining approximately random samples of $\mathbf{x}$ is computationally efficient, as in Diakonikolas et al. (2021).*

2.6 Analysis and Main Results

In this section, we develop the properties of the learning algorithm of Section 2.5 and interpret its structure.

2.6.1 Connection with MLE of $f_{\mathbf{x}}(\cdot; \boldsymbol{\Theta}^* - \boldsymbol{\Theta})$

First we establish the following interpretation of the population version (2.5) of the loss function (2.4b).

Theorem 2.1 (Minimizing population loss function $\iff$ minimizing KL divergence)**.** *We have*

$$\arg\min_{\boldsymbol{\Theta} \in \varTheta} \mathcal{L}(\boldsymbol{\Theta}) = \arg\min_{\boldsymbol{\Theta} \in \varTheta} D(\mathcal{U}_{\mathcal{X}}(\cdot) \| f_{\mathbf{x}}(\cdot; \boldsymbol{\Theta}^* - \boldsymbol{\Theta})).$$

Moreover, $\boldsymbol{\Theta}^*$ is the unique minimizer of $\mathcal{L}(\boldsymbol{\Theta})$.

[7] Ideally, one would consider the gradient of $\mathcal{L}_n(\text{vec}(\boldsymbol{\Theta}))$. However, for the ease of the exposition we abuse the terminology.

Proof of Theorem 2.1. We have

$$f_{\mathsf{x}}(\boldsymbol{x};\boldsymbol{\Theta}^*-\boldsymbol{\Theta}) = \frac{\exp\big(\langle\boldsymbol{\Theta}^*-\boldsymbol{\Theta},\mathbf{\Phi}(\boldsymbol{x})\rangle\big)}{\int_{\mathcal{X}}\exp\big(\langle\boldsymbol{\Theta}^*-\boldsymbol{\Theta},\mathbf{\Phi}(\boldsymbol{y})\rangle\big)\,\mathrm{d}\boldsymbol{y}} \overset{(a)}{=} \frac{\exp\big(\langle\boldsymbol{\Theta}^*-\boldsymbol{\Theta},\boldsymbol{\varPhi}(\boldsymbol{x})\rangle\big)}{\int_{\mathcal{X}}\exp\big(\langle\boldsymbol{\Theta}^*-\boldsymbol{\Theta},\boldsymbol{\varPhi}(\boldsymbol{y})\rangle\big)\,\mathrm{d}\boldsymbol{y}}$$
$$\overset{(b)}{=} \frac{f_{\mathsf{x}}(\boldsymbol{x};\boldsymbol{\Theta}^*)\exp\big(-\langle\boldsymbol{\Theta},\boldsymbol{\varPhi}(\boldsymbol{x})\rangle\big)}{\int_{\mathcal{X}}f_{\mathsf{x}}(\boldsymbol{x};\boldsymbol{\Theta}^*)\exp\big(-\langle\boldsymbol{\Theta},\boldsymbol{\varPhi}(\boldsymbol{y})\rangle\big)\,\mathrm{d}\boldsymbol{y}}$$
$$\overset{(c)}{=} \frac{f_{\mathsf{x}}(\boldsymbol{x};\boldsymbol{\Theta}^*)\exp\big(-\langle\boldsymbol{\Theta},\boldsymbol{\varPhi}(\boldsymbol{x})\rangle\big)}{\mathcal{L}(\boldsymbol{\Theta})}, \tag{2.17}$$

where (a) follows because $\mathbb{E}_{\mathcal{U}_{\mathcal{X}}}[\mathbf{\Phi}(\mathbf{x})]$ is a constant, (b) follows by dividing the numerator and the denominator by the constant $\int_{\mathcal{X}}\exp\big(\langle\boldsymbol{\Theta}^*,\boldsymbol{\varPhi}(\boldsymbol{y})\rangle\big)\,\mathrm{d}\boldsymbol{y}$ and using the definition of $f_{\mathsf{x}}(\boldsymbol{x};\boldsymbol{\Theta}^*)$, and (c) follows from definition of $\mathcal{L}(\boldsymbol{\Theta})$. In turn, we have

$$D(\mathcal{U}_{\mathcal{X}}(\cdot)\|f_{\mathsf{x}}(\cdot;\boldsymbol{\Theta}^*-\boldsymbol{\Theta})) \overset{(a)}{=} \mathbb{E}_{\mathcal{U}_{\mathcal{X}}}\left[\log\left(\frac{\mathcal{U}_{\mathcal{X}}(\cdot)\mathcal{L}(\boldsymbol{\Theta})}{f_{\mathsf{x}}(\cdot;\boldsymbol{\Theta}^*)\exp\big(-\langle\boldsymbol{\Theta},\boldsymbol{\varPhi}(\cdot)\rangle\big)}\right)\right]$$
$$\overset{(b)}{=} \mathbb{E}_{\mathcal{U}_{\mathcal{X}}}\left[\log\left(\frac{\mathcal{U}_{\mathcal{X}}(\cdot)}{f_{\mathsf{x}}(\cdot;\boldsymbol{\Theta}^*)}\right)\right] + \mathbb{E}_{\mathcal{U}_{\mathcal{X}}}\Big[\langle\boldsymbol{\Theta},\boldsymbol{\varPhi}(\cdot)\rangle\Big] + \log\mathcal{L}(\boldsymbol{\Theta})$$
$$\overset{(c)}{=} \mathbb{E}_{\mathcal{U}_{\mathcal{X}}}\left[\log\left(\frac{\mathcal{U}_{\mathcal{X}}(\cdot)}{f_{\mathsf{x}}(\cdot;\boldsymbol{\Theta}^*)}\right)\right] + \big\langle\boldsymbol{\Theta},\mathbb{E}_{\mathcal{U}_{\mathcal{X}}}[\boldsymbol{\varPhi}(\cdot)]\big\rangle + \log\mathcal{L}(\boldsymbol{\Theta})$$
$$\overset{(d)}{=} \mathbb{E}_{\mathcal{U}_{\mathcal{X}}}\left[\log\left(\frac{\mathcal{U}_{\mathcal{X}}(\cdot)}{f_{\mathsf{x}}(\cdot;\boldsymbol{\Theta}^*)}\right)\right] + \log\mathcal{L}(\boldsymbol{\Theta}), \tag{2.18}$$

where (a) follows from (2.17) and the definition of KL divergence, (b) follows because $\log(abc) = \log a + \log b + \log c$ and $\mathcal{L}(\boldsymbol{\Theta})$ is a constant, (c) follows from the linearity of the expectation, and (d) follows because $\mathbb{E}_{\mathcal{U}_{\mathcal{X}}}[\boldsymbol{\varPhi}(\mathbf{x})] = 0$ for centered natural statistics. Now the first term in (2.18) does not depend on $\boldsymbol{\Theta}$, so

$$\arg\min_{\boldsymbol{\Theta}\in\varTheta} D(\mathcal{U}_{\mathcal{X}}(\cdot)\|f_{\mathsf{x}}(\cdot;\boldsymbol{\Theta}^*-\boldsymbol{\Theta})) = \arg\min_{\boldsymbol{\Theta}\in\varTheta}\log\mathcal{L}(\boldsymbol{\Theta}) \overset{(a)}{=} \arg\min_{\boldsymbol{\Theta}\in\varTheta}\mathcal{L}(\boldsymbol{\Theta}),$$

where (a) follows because log is a monotonic function. Further, the KL divergence between $\mathcal{U}_{\mathcal{X}}(\cdot)$ and $f_{\mathsf{x}}(\cdot;\boldsymbol{\Theta}^*-\boldsymbol{\Theta})$ is uniquely minimized when $\mathcal{U}_{\mathcal{X}}(\cdot) = f_{\mathsf{x}}(\cdot;\boldsymbol{\Theta}^*-\boldsymbol{\Theta})$. Recall that the natural statistics are such that the exponential family is minimal. Therefore, $\mathcal{U}_{\mathcal{X}}(\cdot) = f_{\mathsf{x}}(\cdot;\boldsymbol{\Theta}^*-\boldsymbol{\Theta})$ if and only if $\boldsymbol{\Theta} = \boldsymbol{\Theta}^*$. Thus, $\boldsymbol{\Theta}^* \in \arg\min_{\boldsymbol{\Theta}\in\varTheta}\mathcal{L}(\boldsymbol{\Theta})$, and it is the unique minimizer of $\mathcal{L}(\boldsymbol{\Theta})$. □

2.6.2 Connections to Bregman score

We now interpet (2.4) as optimizing a particular Bregman score. Specifically, we show that the term $\exp\big(-\langle\boldsymbol{\Theta},\boldsymbol{\varPhi}(\boldsymbol{x})\rangle\big)$ in the loss function is a Bregman scoring rule.

First, score-based estimators are defined as follows. Let $q(\cdot;\boldsymbol{\Theta})$ be a measurable function parameterized by $\boldsymbol{\Theta}\in\varTheta$. Let $\mathbf{x}$ be a random variable whose distribution is proportional to $q(\cdot;\boldsymbol{\Theta}^*)$ for $\boldsymbol{\Theta}^*\in\varTheta$. A scoring rule $S(\boldsymbol{x}, q(\boldsymbol{x};\boldsymbol{\Theta}))$ (Gneiting and Raftery, 2007) is a numerical score assigned to a realization $\boldsymbol{x}$ of $\mathbf{x}$ and it measures the quality of the predictive function $q(\cdot;\boldsymbol{\Theta})$. Given samples $\boldsymbol{x}^{(1)},\dots,\boldsymbol{x}^{(n)}$ of $\mathbf{x}$, an optimal score estimator $\hat{\boldsymbol{\Theta}}_{n,S}$ for $\boldsymbol{\Theta}^*$ with scoring rule S is

$$\hat{\boldsymbol{\Theta}}_{n,S} \in \arg\min_{\boldsymbol{\Theta}\in\mathit{\Theta}} \frac{1}{n}\sum_{t=1}^{n} S(\boldsymbol{x}^{(t)}, q(\boldsymbol{x}^{(t)};\boldsymbol{\Theta})). \tag{2.19}$$

Furthermore, a scoring rule is *proper* when $\mathbb{E}[S(\mathbf{x}, q(\mathbf{x};\boldsymbol{\Theta}))]$ is uniquely minimized at $\boldsymbol{\Theta}=\boldsymbol{\Theta}^*$. Finally, the (separable) Bregman scoring rule (Grünwald and Dawid, 2004) associated with a convex and differentiable function $\psi\colon \mathbb{R}^+ \to \mathbb{R}$ and a baseline measure ρ is given by

$$S_{\psi,\rho}(\boldsymbol{x}, q(\boldsymbol{x};\boldsymbol{\Theta})) = -\psi'(q(\boldsymbol{x};\boldsymbol{\Theta})) - \mathbb{E}_\rho[\psi(q(\boldsymbol{x};\boldsymbol{\Theta})) - q(\boldsymbol{x};\boldsymbol{\Theta})\,\psi'(q(\boldsymbol{x};\boldsymbol{\Theta}))]. \tag{2.20}$$

With this framework, we have the following relationship.

Proposition 2.4 (Loss function is equivalent of a Bregman score)**.** *Let* $\psi(\cdot) = -\log(\cdot)$, $\rho(\cdot) = \mathcal{U}_{\mathcal{X}}(\cdot)$, *and* $q(\cdot;\boldsymbol{\Theta}) = \exp\big(\langle\boldsymbol{\Theta}, \boldsymbol{\Phi}(\cdot)\rangle\big)$. *Then*

$$S_{\psi,\rho}(\cdot, q(\cdot;\boldsymbol{\Theta})) = \exp\big(-\langle\boldsymbol{\Theta}, \boldsymbol{\Phi}(\cdot)\rangle\big) - 1$$

and

$$\hat{\boldsymbol{\Theta}}_{n,S_{\psi,\rho}} = \hat{\boldsymbol{\Theta}}_n.$$

Proof of Proposition 2.4. With $\psi(\cdot) = -\log(\cdot)$, the Bregman scoring rule in (2.20) simplifies to

$$S_{\psi,\rho}(\boldsymbol{x}, q(\boldsymbol{x};\boldsymbol{\Theta})) = 1/q(\boldsymbol{x};\boldsymbol{\Theta}) + \mathbb{E}_\rho[\log(q(\boldsymbol{x};\boldsymbol{\Theta})) - 1].$$

In turn, with $q(\cdot;\boldsymbol{\Theta}) = \exp\big(\langle\boldsymbol{\Theta}, \boldsymbol{\Phi}(\cdot)\rangle\big)$ and $\rho(\cdot) = \mathcal{U}_{\mathcal{X}}(\cdot)$, we have

$$\begin{aligned} S_{\psi,\rho}(\boldsymbol{x}, q(\boldsymbol{x};\boldsymbol{\Theta})) &= \exp\big(-\langle\boldsymbol{\Theta}, \boldsymbol{\Phi}(\boldsymbol{x})\rangle\big) + \mathbb{E}_{\mathcal{U}_{\mathcal{X}}}\Big[\big\langle\boldsymbol{\Theta}, \boldsymbol{\Phi}(\mathbf{x})\big\rangle\Big] - 1 \\ &\overset{(a)}{=} \exp\big(-\langle\boldsymbol{\Theta}, \boldsymbol{\Phi}(\boldsymbol{x})\rangle\big) + \big\langle\boldsymbol{\Theta}, \mathbb{E}_{\mathcal{U}_{\mathcal{X}}}[\boldsymbol{\Phi}(\mathbf{x})]\big\rangle - 1 \\ &\overset{(b)}{=} \exp\big(-\langle\boldsymbol{\Theta}, \boldsymbol{\Phi}(\boldsymbol{x})\rangle\big) - 1, \end{aligned} \tag{2.21}$$

where (a) follows from the linearity of the expectation and (b) follows because $\mathbb{E}_{\mathcal{U}_{\mathcal{X}}}[\boldsymbol{\Phi}(\mathbf{x})] = 0$ for centered natural statistics. The equivalence between $\hat{\boldsymbol{\Theta}}_{n,S_{\psi,\rho}}$ and $\hat{\boldsymbol{\Theta}}_n$ follows by plugging (2.21) in (2.19). □

We note that by the choice $q(\cdot;\boldsymbol{\Theta}) = \exp\big(\langle\boldsymbol{\Theta}, \boldsymbol{\Phi}(\cdot)\rangle\big)$, we inherently make use of the extension of Bregman scoring rule beyond the probability simplex (Painsky and Wornell, 2019). Further, having established that $\exp\big(-\langle\boldsymbol{\Theta}, \boldsymbol{\Phi}(\boldsymbol{x})\rangle\big)$ is a Bregman scoring rule, it immediately follows that the loss function is proper, since this is a property of all Bregman rules (Gneiting and Raftery, 2007).

2.6.3 Connections to nonparametric density estimation

We now interpret the loss function in (2.4) as an instance of the surrogate likelihood proposed by Jeon and Lin (2006). As described in Section 2.3, to bypass the computational hardness of the MLE, Jeon and Lin (2006) propose using the following surrogate likelihood [cf. (2.12)] for learning nonparametric densities of the form $f_{\mathbf{x}}(\boldsymbol{x}) \propto \exp(\eta(\boldsymbol{x}))$:

$$\mathcal{J}_n(\eta) = \frac{1}{n}\sum_{t=1}^{n} \exp\bigl(-\eta(\boldsymbol{x}^{(t)})\bigr) + \int_{\mathcal{X}} \rho(\boldsymbol{x})\, \eta(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}, \tag{2.22}$$

where $\rho(\cdot)$ is some fixed known density with the same support $\mathcal{X}$ as the unknown density $f_{\mathbf{x}}(\boldsymbol{x})$. The following proposition establishes that the loss function in (2.4) is a surrogate likelihood for a specific choice of $\rho(\cdot)$.

Proposition 2.5 (Loss function is an instance of the surrogate likelihood)**.** *Let* $\rho(\cdot) = \mathcal{U}_{\mathcal{X}}(\cdot)$ *and* $\eta(\cdot) = \bigl\langle \boldsymbol{\Theta}, \boldsymbol{\Phi}(\cdot)\bigr\rangle$*. Then* $\mathcal{J}_n(\eta) = \mathcal{L}_n(\boldsymbol{\Theta})$*.*

Proof of Proposition 2.5. Using $\rho(\cdot) = \mathcal{U}_{\mathcal{X}}(\cdot)$ and $\eta(\cdot) = \bigl\langle \boldsymbol{\Theta}, \boldsymbol{\Phi}(\cdot)\bigr\rangle$ in (2.22), we have

$$\begin{aligned}\mathcal{J}_n(\eta) &= \frac{1}{n}\sum_{t=1}^{n} \exp\bigl(-\bigl\langle \boldsymbol{\Theta}, \boldsymbol{\Phi}(\boldsymbol{x}^{(t)})\bigr\rangle\bigr) + \int_{\mathcal{X}} \mathcal{U}_{\mathcal{X}}(\boldsymbol{x}) \bigl\langle \boldsymbol{\Theta}, \boldsymbol{\Phi}(\boldsymbol{x})\bigr\rangle\, \mathrm{d}\boldsymbol{x} \\ &\overset{(a)}{=} \mathcal{L}_n(\boldsymbol{\Theta}) + \Bigl\langle \boldsymbol{\Theta}, \int_{\mathcal{X}} \mathcal{U}_{\mathcal{X}}(\boldsymbol{x})\, \boldsymbol{\Phi}(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}\Bigr\rangle \overset{(b)}{=} \mathcal{L}_n(\boldsymbol{\Theta}),\end{aligned}$$

where (a) follows from (2.4b) and because the integral of a sum is equal to the sum of the integrals and (b) follows because $\mathbb{E}_{\mathcal{U}_{\mathcal{X}}}[\boldsymbol{\Phi}(\mathbf{x})] = 0$ for centered natural statistics. □

2.6.4 Consistency and normality

To start, we note that the asymptotic theory of MLE cannot be invoked to establish consistency and asymptotic normality of $\hat{\boldsymbol{\Theta}}_n$. Indeed, from Theorem 2.1 we have the population version of $\hat{\boldsymbol{\Theta}}_n$ is equivalent to the maximum likelihood estimate of $f_{\mathbf{x}}(\cdot; \boldsymbol{\Theta}^* - \boldsymbol{\Theta})$, not $f_{\mathbf{x}}(\cdot; \boldsymbol{\Theta})$. And, as such, there is no direct connection between $\hat{\boldsymbol{\Theta}}_n$ and the finite sample maximum likelihood estimate of $f_{\mathbf{x}}(\cdot; \boldsymbol{\Theta})$ or $f_{\mathbf{x}}(\cdot; \boldsymbol{\Theta}^* - \boldsymbol{\Theta})$. Instead, we establish consistency and asymptotic normality of the proposed estimator $\hat{\boldsymbol{\Theta}}_n$ by invoking the asymptotic theory of M-estimation.

Let $\mathbf{A}(\boldsymbol{\Theta}^*)$ denote the covariance matrix of $\mathrm{vec}\bigl(\boldsymbol{\Phi}(\mathbf{x}) \exp\bigl(-\bigl\langle \boldsymbol{\Theta}^*, \boldsymbol{\Phi}(\mathbf{x})\bigr\rangle\bigr)\bigr)$. Let $\mathbf{B}(\boldsymbol{\Theta}^*)$ denote the cross-covariance matrix of $\mathrm{vec}(\boldsymbol{\Phi}(\mathbf{x}))$ and $\mathrm{vec}(\boldsymbol{\Phi}(\mathbf{x}) \exp\bigl(-\bigl\langle \boldsymbol{\Theta}^*, \boldsymbol{\Phi}(\mathbf{x})\bigr\rangle\bigr))$. Let $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ represent the multi-variate Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

Theorem 2.2 (Consistency and asymptotic normality)**.** *Let Assumptions 2.1, 2.2, and 2.3 be satisfied. Let* $\hat{\boldsymbol{\Theta}}_n$ *be a solution of* (2.4)*. Then* $\hat{\boldsymbol{\Theta}}_n \xrightarrow{\mathrm{p}} \boldsymbol{\Theta}^*$ *as* $n \to \infty$*. Further, assuming* $\boldsymbol{\Theta}^* \in \mathrm{int}(\mathit{\Theta})$ *and* $\mathbf{B}(\boldsymbol{\Theta}^*)$ *is invertible, we have* $\sqrt{n}\,\mathrm{vec}(\hat{\boldsymbol{\Theta}}_n - \boldsymbol{\Theta}^*) \xrightarrow{\mathrm{d}} \mathcal{N}(\mathrm{vec}(\mathbf{0}), \mathbf{B}(\boldsymbol{\Theta}^*)^{-1}\mathbf{A}(\boldsymbol{\Theta}^*)\, \mathbf{B}(\boldsymbol{\Theta}^*)^{-1})$*.*

The proof of Theorem 2.2 is provided in Appendix 2.F. The proof is based on two key observations: (a) $\hat{\boldsymbol{\Theta}}_n$ is an M-estimator (which follows from (2.4b)) and (b) $\mathcal{L}(\boldsymbol{\Theta})$ is uniquely minimized at $\boldsymbol{\Theta}^*$ (which follows from Theorem 2.1).

2.6.5 Finite sample guarantees

Theorem 2.3 below shows that, given sufficiently many samples, $\hat{\mathbf{\Theta}}_n$ is close to the true natural parameter $\mathbf{\Theta}^*$ in the Frobenius norm with high probability.

Theorem 2.3 (Finite sample guarantees)**.** *Let Assumptions 2.1, 2.2, 2.3, and 2.4 be satisfied. Define*

$$\Psi(k_1, k_2) = \max_{\mathbf{M} \in \varTheta \setminus \{\mathbf{0}\}} \frac{\mathcal{R}(\mathbf{M})}{\|\mathbf{M}\|_{\mathrm{F}}}.$$

Let $\hat{\mathbf{\Theta}}_n$ be a minimizer of the optimization in (2.15). Then for any $\alpha > 0$ and $\delta \in (0,1)$, we have $\|\hat{\mathbf{\Theta}}_n - \mathbf{\Theta}^*\|_{\mathrm{F}} \leq \alpha$ with probability at least $1 - \delta$ as long as

$$n = \Omega\biggl(\frac{\max\bigl\{\gamma(k_1,k_2)^4, g(k_1,k_2)^2\Psi(k_1,k_2)^2\bigr\}}{\alpha^2\lambda_{\min}^2} \cdot \log\biggl(\frac{4k_1^2k_2^2}{\delta}\biggr)\biggr), \tag{2.23}$$

where $\gamma(\cdot,\cdot)$ and $g(\cdot,\cdot)$ are defined in (2.13) and (2.16), respectively.

The proof of Theorem 2.3, provided in Appendix 2.G, builds on techniques in Negahban et al. (2012); Shah et al. (2021d); Vuffray et al. (2022b, 2016b) and is based on two key properties of the loss function $\mathcal{L}_n(\mathbf{\Theta})$: with sufficiently many samples, a) the loss function $\mathcal{L}_n(\mathbf{\Theta})$ naturally obeys the restricted strong convexity with high probability (which we established in Proposition 2.2); and b) $\|\nabla \mathcal{L}_n(\mathbf{\Theta}^*)\|_{\max}$ is bounded with high probability (which we established in Proposition 2.3). The proof also reveals the dependence of the sample complexity on $r, \overline{r}$ and $\phi_{\max}$, where the regularization penalty λ_n is chosen as

$$\lambda_n = \frac{\alpha\lambda_{\min}}{(12 + 48r\overline{r})\,\mathrm{e}^{2r\overline{r}}\Psi(k_1,k_2)}.$$

The sample complexity in (2.23) depends on the functions $\gamma(k_1,k_2)$, $g(k_1,k_2)$, and $\Psi(k_1,k_2)$, which in turn depend on the choice of the norm $\mathcal{R}$. For the entry-wise $L_{p,q}$ norms, the Schatten p-norms, and the operator norms, the sample complexity is bounded as follows, with a proof in Appendix 2.H.

Corollary 2.1. *For any $p, q \geq 1$, if $\mathcal{R}(\mathbf{\Theta}) = \|\mathbf{\Theta}\|_{p,q}$, $\mathcal{R}(\mathbf{\Theta}) = \|\mathbf{\Theta}\|_p^\star$, or $\mathcal{R}(\mathbf{\Theta}) = \|\mathbf{\Theta}\|_p$, the sample complexity in (2.23) simplifies to*

$$n = \Omega\biggl(\frac{\mathrm{poly}(k_1k_2)}{\alpha^2}\log\biggl(\frac{4k_1^2k_2^2}{\delta}\biggr)\biggr).$$

For certain norms, tighter bounds can be obtained on the sample complexity. The following corollary (whose straightforward proof we omit) provides a formal version of our finite sample guarantees for the examples in Section 2.4, viz., when the underlying norm $\mathcal{R}$ is either the maximum norm, the Frobenius norm, or the nuclear norm. We note that Theorem 2.3 can be specialized for other norms as well.

Corollary 2.2. *If $\mathcal{R}(\mathbf{\Theta}) = \|\mathbf{\Theta}\|_{\max}$, $\mathcal{R}(\mathbf{\Theta}) = \|\mathbf{\Theta}\|_{\mathrm{F}}$, or $\mathcal{R}(\mathbf{\Theta}) = \|\mathbf{\Theta}\|^{\star}$, the sample complexity in* (2.23) *simplifies to*

$$n = \Omega\bigg(\frac{k_1^2 k_2^2}{\alpha^2}\log\bigg(\frac{4k_1^2k_2^2}{\delta}\bigg)\bigg).$$

Remark 2.3. *The result in Theorem 2.3 can also be specialized for learning node-wise-sparse pairwise MRFs. Under this setting, as is typical, the machinery developed could be applied to the node-conditional distribution of x_i, i.e., the conditional distribution in* (2.6)*, for every $i \in [p]$, one at a time. Then with $k_1 = p$ and $k_2 = 1$, the parameter set $\varTheta$ is defined as the set of all r-sparse p-dimensional vectors where r is assumed to be a constant. To enforce the sparsity, $\mathcal{R}$ is chosen to be the ℓ_1 norm resulting in $\mathcal{R}^*$ being equal to the maximum norm. Then it is easy to see that $\gamma(p)$ and $\Psi(p)$ are $O(\sqrt{r})$, and $g(p) = 1$. As a result, the sample complexity in* (2.23) *can be simplified to*

$$n = \Omega\bigg(\frac{1}{\alpha^2}\log\bigg(\frac{p}{\sqrt{\delta}}\bigg)\bigg).$$

The logarithmic dependence on p is consistent with the literature on binary, discrete, Gaussian as well as continuous MRFs Daskalakis et al. (2018); Shah et al. (2021d); Vuffray et al. (2022b, 2016b). The $1/\alpha^2$ dependence on the error tolerance is consistent with the literature on binary and Gaussian MRFs Daskalakis et al. (2018); Vuffray et al. (2016b) and is an improvement over the literature on discrete and continuous MRFs Shah et al. (2021d); Vuffray et al. (2022b).

2.7 Simulations

In this section, we demonstrate our experimental findings on the three examples from Section 2.4 using synthetic data. Specifically, we consider the Frobenius norm constraint on the parameters in the first example, the maximum norm constraint in the second example, and the nuclear norm constraint (which is a relaxation of the low-rank constraint) in the third example. For each of these examples, we make certain choices of the natural statistics and the natural support, with details in Appendix 2.B on how Assumptions 2.2 and 2.3 hold.

2.7.1 Frobenius norm constraint

We consider the random vector $\mathbf{x}$ belonging to $\mathcal{X}$ for two different choices of $\mathcal{X}$: (a) $\mathcal{X} = \mathcal{B}_1(b)$ and (b) $\mathcal{X} = [-b, b]^p$ for some $b \in \mathbb{R}_+$. We let $k_1 = k_2 = p$ and let the natural statistics be polynomials of degree two i.e., $\Phi_{ij} = x_i x_j$ for all $i \in [p], j \in [p]$. Summarizing, the family of distributions considered is as follows:

$$f_{\mathbf{x}}(\boldsymbol{x}; \mathbf{\Theta}) \propto \exp\Big(\sum_{i,j\in[p]} \Theta_{ij} x_i x_j\Big), \tag{2.24}$$

where $\mathbf{x} \in \mathcal{X}$ and $\|\mathbf{\Theta}\|_{\mathrm{F}} \leq r$ for some constant r. As in Section 2.4, let $f_{\mathbf{x}}(\boldsymbol{x}; \mathbf{\Theta}^*)$ denote the true distribution of $\mathbf{x}$ and $\mathbf{\Theta}^*$ denote the true natural parameter of interest such that $\|\mathbf{\Theta}^*\|_{\mathrm{F}} \leq r$. Further, we have $\mathit{\Theta} = \{\mathbf{\Theta} \in \mathbb{R}^{p \times p} \colon \|\mathbf{\Theta}\|_{\mathrm{F}} \leq r\}$.

For our first choice of $\mathcal{X}$, i.e., $\mathcal{X} = \mathcal{B}_1(b)$, the family of distributions in (2.24) satisfies Assumption 2.2 with $\bar{r} = (1+b)^2$ and Assumption 2.3 with $\phi_{\max} = \max\{1, b^2\}$. For our second choice of $\mathcal{X}$, i.e., $\mathcal{X} = [-b, b]^p$, the family of distributions in (2.24) satisfies Assumption 2.3 with $\phi_{\max} = \max\{1, b^2\}$. In contrast, the constant $\bar{r}$ in Assumption 2.2 scales quadratically in p. As a result, the analytical bound on the sample complexity from Corollary 2.2 suggests an exponential dependence on p (see equation (2.43) in the proof of Theorem 2.3 for the dependence on $\bar{r}$). However, we see that the empirical bound on the sample complexity scales only polynomially in p, i.e., it is in agreement with Corollary 2.2, suggesting that Assumption 2.2 may not be a strict requirement in practice. For brevity, we only provide results with $\mathcal{X} = \mathcal{B}_1(b)$. The results with $\mathcal{X} = [-b, b]^p$ are analogous.

We choose $b = 1$ and let the true natural parameter $\mathbf{\Theta}^*$ be as follows:

$$\Theta^*_{ij} = \begin{cases} 1/\sqrt{p} & \text{if } i = 1 \text{ or } j = 1 \text{ or } i = j, \\ 0 & \text{otherwise.} \end{cases} \tag{2.25}$$

This choice ensures that $\|\mathbf{\Theta}^*\|_{\mathrm{F}} \leq r$, i.e., $r = 1$. Further, the choice also ensures that the maximum node-degree in the underlying undirected graphical model is p and the total number of edges scale linearly with p. This is easy to see as the undirected graph is a star graph with x_1 as the center of the star. We note that this is in contrast with the literature on node-wise-sparse pairwise MRFs (see Section 2.3) where the total number of edges scale linearly with p but the maximum node-degree does not depend on p. Therefore, the techniques developed to learn the parameters of such MRFs are not useful here. We also note that the $1/\sqrt{p}$ scaling in (2.25) is consistent with the Sherington-Kirkpatrick model (Sherrington and Kirkpatrick, 1975). Finally, to draw high-quality samples from (2.24), we employ brute-force sampling using fine discretization with 100 bins per dimension.

In Fig. 2.1a, we plot the scaling of errors in our estimates for $\mathbf{\Theta}^*$, i.e., $\|\hat{\mathbf{\Theta}}_n - \mathbf{\Theta}^*\|_{\mathrm{F}}$ as a function of the number of samples n for various p. Likewise, we present how the error scales as the dimension p grows for various n in Fig. 2.1b. We plot the averaged error across 100 independent trials along with ±1 standard error (the standard error is too small to be visible in our results). To help see the error scaling, we display the best linear fit (fitted on the log-log scale) and mention an empirical decay rate in the legend based on the slope of that fit, e.g., for a slope of -0.47 for estimating $\mathbf{\Theta}^*$ when $p = 7$, we report an empirical rate of $n^{-0.47}$ for the averaged error.

2.7.2 Maximum norm constraint

We consider the same model as in Section 2.7.1 except for the choices of $\mathbf{\Theta}^*$ and $\mathcal{X}$. We let the true natural parameter $\mathbf{\Theta}^*$ be as follows:

$$\Theta^*_{ij} = -0.1 - 0.4 \cdot \mathbb{1}(i = j) - 0.2 \cdot \mathbb{1}(|i - j| = 1) - 0.1 \cdot \mathbb{1}(|i - j| = 2).$$

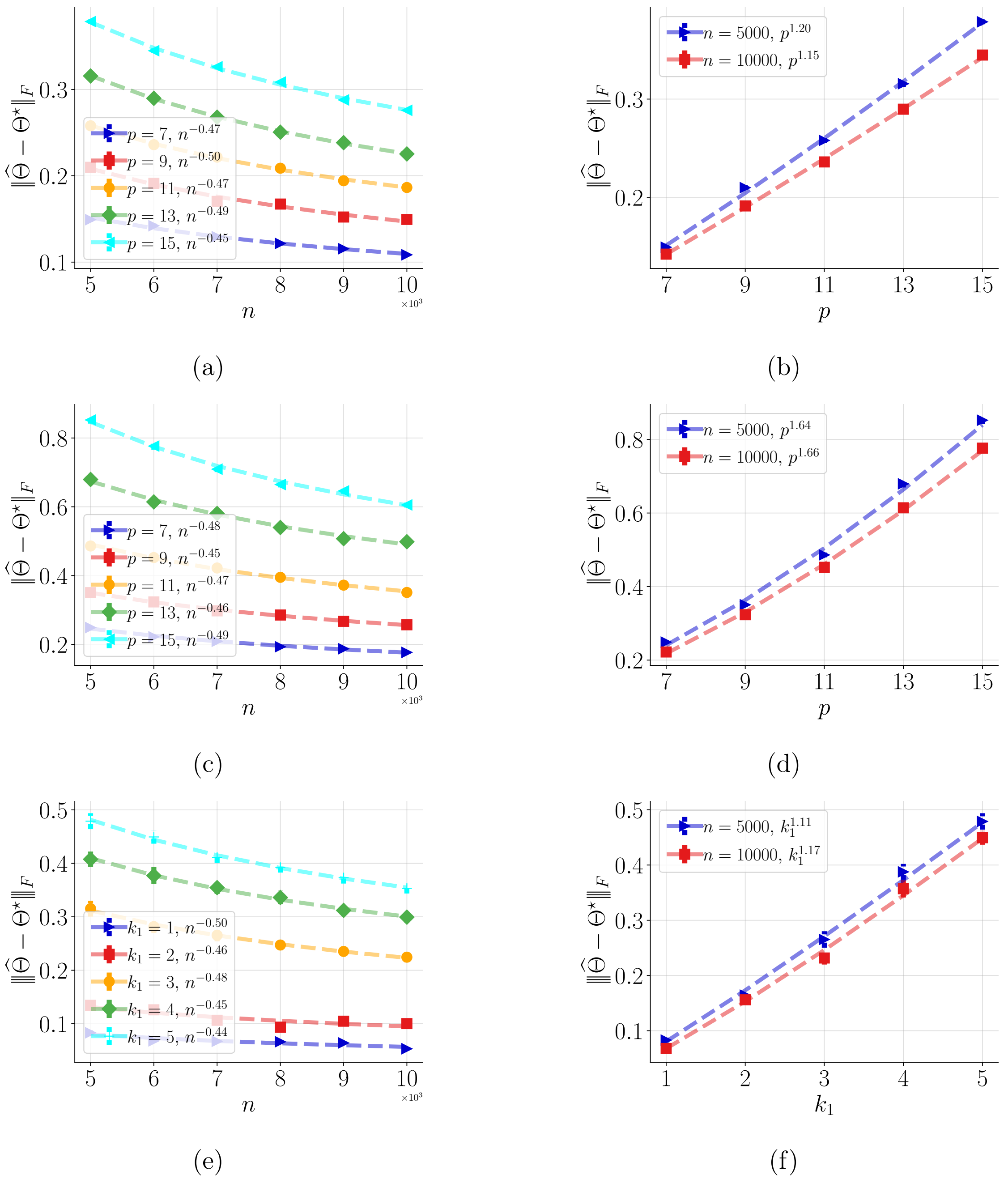


Figure 2.1: Error scaling for Frobenius norm constraint (in a and b), maximum norm constraint (in c and d), and the nuclear norm constraint (in e and f) with number of samples n for various p or k_1 (in a, c, e) and with number of parameters p or k_1 for various n (in b, d, f).

This choice ensures that $\|\mathbf{\Theta}^*\|_{\max} \leq r$, i.e., $r = 0.5$. Further, the choice also ensures that the maximum node-degree in the underlying undirected graphical model is p and the total number of edges scale quadratically with p. This is easy to see as the undirected graph is a complete graph as every entry of $\mathbf{\Theta}^*$ is non-zero. Further, we also note the choice of $\mathbf{\Theta}^*$ ensures that the inverse of $\mathbf{\Theta}^*$ is positive semi-definite. Therefore, the distribution of $\mathbf{x}$ is equivalent to a Gaussian with mean equal to zero and inverse covariance equal to $\mathbf{\Theta}^*$ but the support truncated to $\mathcal{X}$. Then we use the `tmvtnorm` package (Wilhelm and Manjunath, 2010a) to generate samples from (2.24) via rejection sampling, and choose $\mathcal{X} = [-b, b]^p$ (with $b = 1$) for a higher acceptance probability.

In Fig. 2.1c, we plot the scaling of errors in our estimates for $\mathbf{\Theta}^*$, i.e., $\|\hat{\mathbf{\Theta}}_n - \mathbf{\Theta}^*\|_{\mathrm{F}}$ as a function of the number of samples n for various p. Likewise, we present how the error scales as the dimension p grows for various n in Fig. 2.1d.

2.7.3 Nuclear norm constraint

For this constraint, we let $\mathcal{X} = \mathcal{B}_2(b)$. We consider the dimension $p = 2$ and vary the number of natural parameters $k = k_1 k_2$. We let the natural statistics be polynomials of varying degree, i.e., $\Phi_{ij} = x_1^i x_2^j$ for all $i \in [k_1], j \in [k_2]$. Summarizing, the family of distribution considered is as follows:

$$f_{\mathbf{x}}(\boldsymbol{x}; \mathbf{\Theta}) \propto \exp\Big(\sum_{i \in [k_1], j \in [k_2]} \Theta_{i,j} x_1^i x_2^j \Big), \tag{2.26}$$

where $\mathbf{x} \in \mathcal{X}$ and $\|\mathbf{\Theta}^{(1)}\|^{\star} = r$ for some constant r. As in Section 2.4, let $f_{\mathbf{x}}(\boldsymbol{x}; \mathbf{\Theta}^*)$ denote the true distribution of $\mathbf{x}$ and $\mathbf{\Theta}^*$ denote the true natural parameter of interest such that $\|\mathbf{\Theta}^*\|^{\star} \leq r$. Further, we have $\mathit{\Theta} = \big\{ \mathbf{\Theta} \in \mathbb{R}^{k_1 \times k_2} : \|\mathbf{\Theta}\|^{\star} \leq r \big\}$.

In our simulations, we fix $k_2 = 2$ and vary k_1 from 1 to 5. For $k_1 = 1$, we let the true natural parameter $\mathbf{\Theta}^*$ be as follows:

$$\mathbf{\Theta}^* = \begin{bmatrix} 1 & 0.8 \end{bmatrix}.$$

This choice ensures that $\|\mathbf{\Theta}^*\|^{\star} \leq r$, i.e., $r = 1$. To ensure that $r = 1$ for $k_1 > 1$, we let $\Theta^*_{i,j} = \Theta^*_{i-1,j}/2$ i.e., we let every row of $\mathbf{\Theta}^*$ to be a multiple of its first row. To draw high-quality samples from (2.26), we employ brute-force sampling using fine discretization with 100 bins per dimension.

In Fig. 2.1e, we plot the scaling of errors in our estimates for $\mathbf{\Theta}^*$, i.e., $\|\hat{\mathbf{\Theta}}_n - \mathbf{\Theta}^*\|_{\mathrm{F}}$ as a function of the number of samples n for various k_1. Likewise, we present how the error scales as the dimension k_1 grows for various n in Fig. 2.1f.

2.7.4 Results

We observe that the error $\|\hat{\mathbf{\Theta}}_n - \mathbf{\Theta}^*\|_{\mathrm{F}}$, for all three constraints, admits a scaling of between $n^{-0.50}$ and $n^{-0.44}$ for various p. These empirical rates indicate a parametric error rate of $\sqrt{1/n}$, which is consistent with the theoretical rate. For the Frobenius norm constraint, the error $\|\hat{\mathbf{\Theta}}_n - \mathbf{\Theta}^*\|_{\mathrm{F}}$ admits a scaling of between $p^{1.15}$ and $p^{1.20}$ for different n. These empirical rates indicate a parametric error rate of $p^2 \log p$, which is an improvement over the theoretical rate of $p^4 \log p$ suggested by Corollary 2.2. For the maximum norm constraint, the error $\|\hat{\mathbf{\Theta}}_n - \mathbf{\Theta}^*\|_{\mathrm{F}}$ admits a scaling of between $p^{1.64}$ and $p^{1.66}$ for different n. These empirical rates indicate a parametric error rate of $p^3 \log p$, which is an improvement over the theoretical rate of $p^4 \log p$ suggested by

Corollary 2.2. Lastly, for the nuclear norm constraint, the error $\|\hat{\mathbf{\Theta}}_n - \mathbf{\Theta}^*\|_{\mathrm{F}}$ admits a scaling of between $k_1^{1.11}$ and $k_1^{1.17}$ for different n. These empirical rates indicate a parametric error rate of $k_1^2 \log k_1$, which matches the theoretical rate from Corollary 2.2 (when k_2 is treated as a constant).

2.8 Concluding Remarks

In this work, we develop a computationally efficient alternative to the MLE for learning distributions in a k-parameter exponential family from i.i.d. samples. While our estimator is consistent and asymptotically normal, it is not asymptotically efficient. Focusing on node-wise-sparse pairwise exponential family MRFs, which is a special case of the setting considered in our work, (Shah et al., 2021d, Appendix U.2) provides one such example where the asymptotic covariance matrix of $\hat{\mathbf{\Theta}}_n$ from Theorem 2.2 does not coincide with the inverse of the Fisher information matrix. Investigating the possibility of a single estimator that achieves computational and asymptotic efficiency for the class of exponential family in our work could be an interesting future direction.

We emphasize that the focus of our work includes but is not limited to exponential families associated with node-wise-sparse MRFs, i.e., undirected graphical models, and towards general exponential families. The former focuses on local assumptions on the parameters such as node-wise-sparsity, and the sample complexity depends logarithmically on the parameter dimension i.e., $O(\log(k))$. In contrast, our work can handle local as well as global structures on the parameters, e.g., a maximum norm constraint, a Frobenius norm constraint, or a nuclear norm constraint (see Section 2.7), and our loss function in (2.4b) is a generalization of the interaction screening objective (Vuffray et al., 2016b) and generalized interaction screening objective (Shah et al., 2021d; Vuffray et al., 2022b). Similarly, for node-wise-sparse MRFs there has been a lot of work to relax the assumptions required for learning (see the discussion on Assumption 2.4 below). Since our work focuses on global structures associated with the parameters, we leave the question of relaxing the assumptions required for learning as an open question.

It is also worth commenting on Assumption 2.4. For node-wise-sparse pairwise exponential family MRFs (e.g., Ising models), which is a special case of the setting considered in our work, Assumption 2.4 is proven (e.g., Shah et al. (2021d, Appendix T.1) provides one such analysis for a condition that is equivalent to Assumption 2.4 for sparse continuous graphical model). However, such analysis typically requires a bound on the ℓ_1 norm of the parameters associated with each node as in MRFs. Since the focus of our work is beyond the exponential families associated with node-wise-sparse MRFs, we view Assumption 2.4 as an adequate condition to rule out certain singular distributions (as evident in the proof of Proposition 2.2 where this condition is used to effectively lower bounds the variance of a non-constant random variable). Therefore, we expect this assumption to hold for most real-world applications. Further, we highlight that the MLE in (2.3) remains computationally intractable even under Assumption 2.4. To see this, one could again focus on node-wise-sparse pairwise exponential family MRFs where Assumption 2.4 is proven and the MLE is still known to be computationally intractable.

Finally, while truncated exponential families are important classes of distributions, it requires boundedness of the support and does not capture a few widely used non-compact distributions, i.e., distributions with infinite support (e.g., Gaussian distribution, Laplace distribution). While, conceptually, most non-compact distributions could be truncated by introducing a controlled amount of error, we believe this assumption could be lifted as for exponential families: $\mathbb{P}(|\mathsf{x}_i| \geq \delta \log \gamma) \leq c\gamma^{-\delta}$ where $c > 0$ is a constant and $\gamma > 0$. Alternatively, the notion of multiplicative regularizing distribution from Ren et al. (2021) could also be used. We believe extending our work to the non-compact setup could be a valuable direction for future work.

Appendix

2.A Further Related Work

In this section, we expand on Section 2.3, summarizing additional examples of work on learning sparse MRFs, score-based methods, and nonparametric density estimation, including the related literature on Stein discrepancy.

2.A.1 Learning sparse MRFs

Additional investigations into learning sparse exponential family MRFs beyond those discussed in Section 2.3.4 include the following.

Following Yang et al. (2015), Suggala et al. (2017) propose an ℓ_1-regularized node-conditional log-likelihood to learn the node-conditional density in (2.6) for nonlinear $\phi(\cdot)$. They use an alternating minimization technique with proximal gradient descent to solve the resulting optimization problem. However, the analysis requires restricted strong convexity, bounded domain of the variables, nonnegative node parameters, and some assumptions on gradient of the population loss that are hard to verify.

Yang et al. (2018) introduce a nonparametric component to the node-conditional density in (2.6) while focusing on linear $\phi(\cdot)$; specifically,

$$f_{\mathbf{x}}(\boldsymbol{x}) \propto \exp\Big(\sum_{i\in[p]} \eta_i(x_i) + \sum_{j\neq i} \Theta_{ij} x_i x_j\Big),$$

where $\eta_i(\cdot)$ is the nonparametric node-wise term. They propose a node-conditional pseudo-likelihood (introduced in Ning et al. (2017)) regularized by a nonconvex penalty and use an adaptive multi-stage convex relaxation method to solve the resulting optimization problem. However, their finite-sample bounds require bounded moments of the variables, a sparse eigenvalue condition on their loss function, and local smoothness of the log-partition function.

Sun et al. (2015) investigate an infinite dimensional sparse pairwise exponential family MRFs where they assume that the node and the edge potentials lie in an RKHS. They use a penalized version of the score-matching objective of Hyvärinen and Dayan (2005). However, the associated finite-sample analysis requires incoherence and dependency conditions (see Jalali et al. (2011); Wainwright et al. (2006)).

Lin et al. (2016) consider the joint distribution in (2.9), while restricting the variables to be nonnegative. They propose a group lasso regularized score-matching objective

(Hyvärinen, 2007), with a focus on nonnegative data. However, the associated finite-sample analysis requires the incoherence condition.

2.A.2 Score-based and Stein discrepancy methods

Additional developments of score-based methods beyond those discussed in Section 2.3.5 include those based on Stein discrepency.

A Stein discrepancy is a quantitative measure of how well a predictive distribution, Q, matches a distribution of interest, P, based on a generalization of the classical Stein's identity to multivariate distributions (Gorham and Mackey, 2015). Stein's identity defines an infinite number of identities indexed by a critic function f and, like the score-matching method, does not require evaluation of the partition function. By focusing on Stein discrepancies constructed from a RKHS, Liu et al. (2016) and Chwialkowski et al. (2016) independently propose the kernel Stein discrepancy as a test statistic to access the goodness-of-fit for unnormalized densities.

In contemporaneous work, Gorham and Mackey (2017) develop kernel Stein discrepancies as tools for explicitly measuring and comparing sample quality (e.g., for judging which sample approximation offered a better fit to P). In particular, Gorham and Mackey (2017) provides a criteria for selecting a suitable kernel and recommends the rational quadratic kernel (a particular form of inverse multiquadric kernel) as a simple default choice.

In addition to the kernel Stein discrepancies, one can also use non-kernel Stein discrepancies as surrogate objective functions for estimation, as the Stein discrepancy objectives are convex in the natural parameters of an exponential family (Barp et al., 2019). For instance, the computable spanner graph Stein discrepancy (Gorham and Mackey, 2015), which has no tuning parameters (such as a kernel), can be used if kernel selection is a concern. In Barp et al. (2019) it is demonstrated that the Fisher divergence, the minimization criterion used by the score-matching method, can be viewed as a special case of a non-kernel Stein discrepancy with a non-kernel Stein set of test functions. It is also showed that other methods, including contrastive divergence (Hinton, 2002), can be viewed as Stein discrepancies with respect to a different class of critic functions.

Dai et al. (2019) leverage the primal-dual formulation of the maximum likelihood estimator (MLE) to avoid estimating the normalizing constant, at the cost of introducing dual variables that must be jointly estimated. They demonstrate that many other methods, including contrastive divergence Hinton (2002), pseudo-likelihood Besag (1975a), score-matching (Hyvärinen and Dayan, 2005), and the minimum Stein discrepancy estimator (Barp et al., 2019), are special cases of their estimator. However, this approach leads to expensive optimization problems, as it relies on adversarial optimization (see Rhodes et al. (2020) for details). Liu et al. (2019) propose an inference method for unnormalized models known as the discriminative likelihood estimator. This estimator follows the KL divergence minimization criterion and is implemented via density ratio estimation and a Stein operator. However, this method requires certain hard-to-verify conditions. Ryu et al. (2024) provide a unified framework encompassing various estimators for learning exponential family, based on noise-contrastive estimation

(Gutmann and Hyvärinen, 2010, 2012).

2.A.3 Nonparametric density estimation

Additional investigations into approaches based on nonparametric density estimation beyond those in Section 2.3.6 include the following.

Silverman (1982) proposes to estimate the log density, $\eta(\cdot) = \log f_{\mathbf{x}}(\cdot)$, which eliminates the positivity constraint, and augmented Leonard's formulation in (2.11) by introducing a functional $\int_{\mathcal{X}} \exp(\eta(\boldsymbol{x}))\,\mathrm{d}\boldsymbol{x}$, effectively enforcing the unity constraint. The properties of the following penalized estimator are studied,

$$-\min \frac{1}{n}\sum_{i=1}^{n} \eta(\boldsymbol{x}_i) + \int_{\mathcal{X}} \exp(\eta(\boldsymbol{x}))\,\mathrm{d}\boldsymbol{x} + \lambda J(\eta), \tag{2.27}$$

in the setting where $\eta(\cdot)$ lies in a RKHS. The formulation in (2.27) is further analyzed by O'Sullivan (1988), who provides a practical algorithm with cross-validated λ. However, similar to the formulation of Leonard (1978), the formulation in Silverman (1982) scales poorly in high-dimensional settings.

More generally, the formulation in Leonard (1978) in (2.11) evolved through a series of works. Gu and Qiu (1993) investigate the properties of the estimator in (2.11) over a RKHS, although they use a finite-dimensional function space approximation (composed of the linear span of kernel functions) to the RKHS. Gu (1993) provides a practical algorithm with cross-validated λ for the estimator analyzed in Gu and Qiu (1993). Gu and Wang (2003) further improve upon the algorithm in Gu (1993) by offering a direct strategy for cross-validation. However, this function-space approximation lacks strong statistical guarantees.

Gu et al. (2013) provide a practical method for selecting a cross-validated λ for the formulation of Jeon and Lin (2006) in (2.12), using the function-space approximation in Gu and Qiu (1993). However, their algorithm struggles in high-dimensional settings, where accurate estimation becomes challenging.

2.B Examples of Natural Statistics

Examples of natural statistics and their support satisfying Assumptions 2.2 and 2.3 including the following.

2.B.1 Polynomial statistics

Consider natural statistics that are polynomials in $\mathbf{x}$ of maximum degree l, i.e.,

$$\prod_{i\in[p]} x_i^{l_i} \tag{2.28}$$

such that $l_i \in [l] \cup \{0\}$ for all $i \in [p]$ and $\sum_{i\in[p]} l_i \leq l$ for some $l < p$. Furthermore, suppose $\mathcal{X} \subset \mathcal{X}_+ \triangleq [-b, b]^p$ for some $b \in \mathbb{R}_+$. Then, Assumptions 2.2 and 2.3 are verified as follows for some representative matrix norms.

First, we have

$$\|\mathbf{\Phi}(\boldsymbol{x})\|_{\max} = \max_{u\in[k_1],v\in[k_2]} |\Phi_{uv}(\boldsymbol{x})| \leq \max\{1, b^l\}.$$

so Assumption 2.3 holds with $\phi_{\max} = \max\{1, b^l\}$.

Next, we verify Assumption 2.2 under each of maximum, Frobenius, and nuclear norms.

2.B.1.1 Maximum norm

For maximum norm, the natural support is $\mathcal{X} = \mathcal{B}_1(b) \subset \mathcal{X}_+$. The dual norm $\mathcal{R}^*$ is the matrix $L_{1,1}$, so Assumption 2.2 holds since

$$\mathcal{R}^*(\mathbf{\Phi}(\boldsymbol{x})) = \|\mathbf{\Phi}(\boldsymbol{x})\|_{1,1} \leq \sum_{\substack{l_i\in[l]\cup\{0\}:\\ \sum_i l_i \leq l}} \prod_{i\in[p]} |x_i|^{l_i} \leq (1 + \|\boldsymbol{x}\|_1)^l \overset{(a)}{\leq} (1+b)^l, \tag{2.29}$$

where (a) follows because $\boldsymbol{x} \in \mathcal{B}_1(b)$.

2.B.1.2 Frobenius norm

For the Frobenius norm, the natural support is $\mathcal{X} = \mathcal{B}_1(b) \subset \mathcal{X}_+$. The dual norm $\mathcal{R}^*$ is the Frobenius norm itself, so Assumption 2.2 holds since

$$\mathcal{R}^*(\mathbf{\Phi}(\boldsymbol{x})) = \|\mathbf{\Phi}(\boldsymbol{x})\|_{\mathrm{F}} \overset{(a)}{\leq} \|\mathbf{\Phi}(\boldsymbol{x})\|_{1,1} \overset{(b)}{\leq} (1+b)^l,$$

where (a) follows because Frobenius norm is bounded by matrix $L_{1,1}$ norm and (b) follows from (2.29).

2.B.1.3 Nuclear norm

Finally, for the nuclear norm, the natural support is $\mathcal{X} = \mathcal{B}_2(b) \subset \mathcal{X}_+$. Consider $l = 2$. The dual norm $\mathcal{R}^*$ is the matrix spectral norm, and we have $\mathbf{\Phi}(\boldsymbol{x}) = \tilde{\mathsf{x}}\tilde{\mathsf{x}}^{\mathrm{T}}$ where $\tilde{\mathsf{x}} = (1, \mathsf{x}_1, \ldots, \mathsf{x}_p)$. Hence, $\mathbf{\Phi}(\boldsymbol{x})$ is a rank-1 matrix and the spectral norm is equal to the sum of the diagonal entries. Accordingly, Assumption 2.2 holds since

$$\mathcal{R}^*(\mathbf{\Phi}(\boldsymbol{x})) = \|\mathbf{\Phi}(\boldsymbol{x})\| \leq (1 + \|\boldsymbol{x}\|_2^2) \overset{(a)}{\leq} (1+b^2),$$

where (a) follows because $\boldsymbol{x} \in \mathcal{B}_2(b)$.

2.B.2 Trigonometric statistics

Now consider natural statistics that are sines and cosines of $\mathbf{x}$ with l different frequencies, i.e.,

$$\sin\Big(\sum_{i\in[p]} l_i x_i\Big) \qquad \text{and} \qquad \cos\Big(\sum_{i\in[p]} l_i x_i\Big) \tag{2.30}$$

such that $l_i \in [l] \cup \{0\}$ for all $i \in [p]$. Furthermore, let $\mathcal{X} \subset \mathbb{R}^p$ be arbitary. Assumptions 2.2 and 2.3 are verified as follows for a representative matrix norm.

First, for any $\boldsymbol{x} \in \mathcal{X}$, we have

$$\|\boldsymbol{\Phi}(\boldsymbol{x})\|_{\max} = \max_{u\in[k_1],v\in[k_2]} |\Phi_{uv}(\boldsymbol{x})| \leq 1,$$

so Assumption 2.3 holds with $\phi_{\max} = 1$.

Second, for the $L_{1,1}$ norm, the dual norm $\mathcal{R}^*$ is the matrix maximum norm. Accordingly, for any $\boldsymbol{x} \in \mathcal{X}$, Assumption 2.2 holds since

$$\mathcal{R}^*(\boldsymbol{\Phi}(\boldsymbol{x})) = \|\boldsymbol{\Phi}(\boldsymbol{x})\|_{\max} \leq \phi_{\max} = 1.$$

2.B.3 Combinations of polynomial and trigonometric statistics

Now consider natural statistics that are combinations of polynomials of $\mathbf{x}$ with maximum degree l—as in (2.28)—and sines and cosines of $\mathbf{x}$ of $\tilde{l}$ different frequencies—as in (2.30). Furthermore, consider the support $\mathcal{X} = \mathcal{X}_+ = [-b, b]^p$ for $b \in \mathbb{R}_+$. Assumptions 2.2 and 2.3 are verified as follows for the case of a representative norm.

First, from Appendices 2.B.1 and 2.B.2, it is straightforward to verify that Assumption 2.3 holds with $\phi_{\max} = \max\{1, b^l\}$.

Second, for the case of the $L_{1,1}$ norm, since $\mathcal{R}^*$ is the matrix maximum norm it follows that

$$\mathcal{R}^*(\boldsymbol{\Phi}(\boldsymbol{x})) = \|\boldsymbol{\Phi}(\boldsymbol{x})\|_{\max} \leq \phi_{\max} = \max\{1, b^l\},$$

so Assumption 2.2 holds.

2.C Proof of Proposition 2.1: Smoothness of $\mathcal{L}_n$

To show the desired smoothness of $\mathcal{L}_n(\boldsymbol{\Theta})$, we show that the largest eigenvalue of the Hessian[8] of $\mathcal{L}_n(\boldsymbol{\Theta})$ is upper bounded by $4k_1k_2\phi_{\max}^2 e^{2r\bar{r}}$. First, we simplify the Hessian of $\mathcal{L}_n(\boldsymbol{\Theta})$, i.e., $\nabla^2\mathcal{L}_n(\boldsymbol{\Theta})$. The component of the Hessian of $\mathcal{L}_n(\boldsymbol{\Theta})$ corresponding to $\Theta_{u_1v_1}$ and $\Theta_{u_2v_2}$ for some $u_1, u_2 \in [k_1]$ and $v_1, v_2 \in [k_2]$ is given by

$$\frac{\partial^2 \mathcal{L}_n(\boldsymbol{\Theta})}{\partial\Theta_{u_1v_1}\partial\Theta_{u_2v_2}} = \frac{1}{n}\sum_{t=1}^{n} \mathit{\Phi}_{u_1v_1}(\boldsymbol{x}^{(t)})\, \mathit{\Phi}_{u_2v_2}(\boldsymbol{x}^{(t)}) \cdot \exp\Big(-\big\langle \boldsymbol{\Theta}, \boldsymbol{\mathit{\Phi}}(\boldsymbol{x}^{(t)})\big\rangle\Big).$$

From the Gershgorin circle theorem, we know that the largest eigenvalue of any matrix is upper bounded by the largest absolute row sum or column sum. Let $\lambda_{\max}(\nabla^2\mathcal{L}_n(\boldsymbol{\Theta}))$ denote the largest eigenvalue of $\nabla^2\mathcal{L}_n(\boldsymbol{\Theta})$. We have the following

$$\lambda_{\max}(\nabla^2\mathcal{L}_n(\boldsymbol{\Theta})) \leq \max_{u_2,v_2} \sum_{u_1,v_1} \left| \frac{\partial^2 \mathcal{L}_n(\boldsymbol{\Theta})}{\partial\Theta_{u_1v_1}\partial\Theta_{u_2v_2}} \right|$$

[8] Ideally, one would consider the Hessian of $\mathcal{L}_n(\text{vec}(\boldsymbol{\Theta}))$. However, we abuse the terminology for the sake of clarity and simplicity in the exposition.

$$= \max_{u_2, v_2} \sum_{u_1, v_1} \left| \frac{1}{n} \sum_{t=1}^{n} \varPhi_{u_1 v_1}(\boldsymbol{x}^{(t)}) \varPhi_{u_2 v_2}(\boldsymbol{x}^{(t)}) \cdot \exp\Big(- \big\langle \boldsymbol{\Theta}, \boldsymbol{\varPhi}(\boldsymbol{x}^{(t)}) \big\rangle \Big) \right|. \tag{2.31}$$

To bound (2.31), we bound the absolute inner product between $\boldsymbol{\Theta}$ and $\boldsymbol{\varPhi}$, i.e., $\big|\big\langle \boldsymbol{\Theta}, \boldsymbol{\varPhi}(\boldsymbol{x}) \big\rangle\big|$ for any $\boldsymbol{x} \in \mathcal{X}$. We have

$$\begin{aligned} \big|\big\langle \boldsymbol{\Theta}, \boldsymbol{\varPhi}(\boldsymbol{x}) \big\rangle\big| &\overset{(a)}{\leq} \mathcal{R}(\boldsymbol{\Theta})\, \mathcal{R}^*(\boldsymbol{\varPhi}(\boldsymbol{x})) \overset{(b)}{\leq} r\Big(\mathcal{R}^*(\boldsymbol{\Phi}(\boldsymbol{x})) + \mathcal{R}^*(\mathbb{E}_{\mathcal{U}_{\mathcal{X}}}[\boldsymbol{\Phi}(\mathbf{x})]) \Big) \\ &\overset{(c)}{\leq} r\Big(\mathcal{R}^*(\boldsymbol{\Phi}(\boldsymbol{x})) + \mathbb{E}_{\mathcal{U}_{\mathcal{X}}}[\mathcal{R}^*(\boldsymbol{\Phi}(\mathbf{x}))] \Big) \overset{(d)}{\leq} 2r\bar{r}, \end{aligned} \tag{2.32}$$

where (a) follows from the definition of a dual norm, (b) follows from the definition of centered natural statistics, the triangle inequality, and because $\mathcal{R}(\boldsymbol{\Theta}) \leq r$, (c) follows from convexity of norms, and (d) follows from Assumption 2.2. Likewise, using Assumption 2.3, we can bound $\|\boldsymbol{\varPhi}(\boldsymbol{x})\|_{\max}$ by $2\phi_{\max}$ for any $\boldsymbol{x} \in \mathcal{X}$. Using these bounds in (2.31), we have

$$\lambda_{\max}(\nabla^2 \mathcal{L}_n(\boldsymbol{\Theta})) \leq \max_{u_2, v_2} \sum_{u_1, v_1} 4\phi_{\max}^2 \mathrm{e}^{2r\bar{r}} = 4k_1 k_2 \phi_{\max}^2 \mathrm{e}^{2r\bar{r}}.$$

Therefore, $\mathcal{L}_n(\boldsymbol{\Theta})$ is a $4k_1 k_2 \phi_{\max}^2 \mathrm{e}^{2r\bar{r}}$ smooth function of $\boldsymbol{\Theta}$.

2.D Proof of Proposition 2.2: Restricted strong convexity of $\mathcal{L}_n$

The following lemma is useful in our analysis.

Lemma 2.1. *For $u_1, u_2 \in [k_1]$ and $v_1, v_2 \in [k_2]$, let*

$$H_{u_1 v_1 u_2 v_2} \triangleq \mathbb{E}\big[\varPhi_{u_1 v_1}(\mathbf{x})\, \varPhi_{u_2 v_2}(\mathbf{x})\big], \tag{2.33}$$

and

$$\hat{H}_{u_1 v_1 u_2 v_2} \triangleq \frac{1}{n} \sum_{t=1}^{n} \varPhi_{u_1 v_1}(\boldsymbol{x}^{(t)})\, \varPhi_{u_2 v_2}(\boldsymbol{x}^{(t)}). \tag{2.34}$$

Fix any $\epsilon > 0$ and $\delta \in (0,1)$. Then if Assumption 2.3 is satisfied and

$$n > \frac{32\phi_{\max}^4}{\epsilon^2} \log\Big(\frac{2k_1^2 k_2^2}{\delta} \Big),$$

we have

$$|\hat{H}_{u_1 v_1 u_2 v_2} - H_{u_1 v_1 u_2 v_2}| < \epsilon,$$

with probability at least $1 - \delta$.

Proof of Lemma 2.1. Fix $u_1, u_2 \in [k_1]$ and $v_1, v_2 \in [k_2]$. The random variable defined as $Y_{u_1v_1u_2v_2} \triangleq \varPhi_{u_1v_1}(\mathbf{x})\,\varPhi_{u_2v_2}(\mathbf{x})$ satisfies $|Y_{u_1v_1u_2v_2}| \leq 4\phi_{\max}^2$ (from the definition of centered natural statistics, triangle inequality, convexity of norms, and Assumption 2.3). Using the Hoeffding inequality we get

$$\mathbb{P}\big(|\hat{H}_{u_1v_1u_2v_2} - H_{u_1v_1u_2v_2}| > \epsilon\big) < 2\exp\bigg(-\frac{n\epsilon^2}{32\phi_{\max}^4}\bigg).$$

The proof follows by using the union bound over all $u_1, u_2 \in [k_1]$ and $v_1, v_2 \in [k_2]$. □

We now proceed to establishing Proposition 2.2. First, we simplify the gradient of $\mathcal{L}_n(\mathbf{\Theta})$[9] evaluated at $\mathbf{\Theta}^*$. For any $u \in [k_1]$ and $v \in [k_2]$, the component of the gradient of $\mathcal{L}_n(\mathbf{\Theta})$ corresponding to Θ_{uv} evaluated at $\mathbf{\Theta}^*$ is given by

$$\frac{\partial \mathcal{L}_n(\mathbf{\Theta}^*)}{\partial \Theta_{uv}} = \frac{-1}{n}\sum_{t=1}^{n} \varPhi_{uv}(\boldsymbol{x}^{(t)})\exp\big(-\big\langle \mathbf{\Theta}^*, \boldsymbol{\varPhi}(\boldsymbol{x}^{(t)})\big\rangle\big). \tag{2.35}$$

We now provide the desired lower bound on the residual. Substituting (2.4b) and (2.35) in (2.14), we have

$$\begin{aligned}
\delta\mathcal{L}_n(\Delta, \mathbf{\Theta}^*) &= \frac{1}{n}\sum_{t=1}^{n} \exp\big(-\big\langle \mathbf{\Theta}^*, \boldsymbol{\varPhi}(\boldsymbol{x}^{(t)})\big\rangle\big)\cdot\Big[\exp\big(-\big\langle \Delta, \boldsymbol{\varPhi}(\boldsymbol{x}^{(t)})\big\rangle\big) - 1 + \big\langle \Delta, \boldsymbol{\varPhi}(\boldsymbol{x}^{(t)})\big\rangle\Big]\\
&\overset{(a)}{\geq} \frac{\mathrm{e}^{-2r\bar{r}}}{n}\sum_{t=1}^{n}\Big[\exp\big(-\big\langle \Delta, \boldsymbol{\varPhi}(\boldsymbol{x}^{(t)})\big\rangle\big) - 1 + \big\langle \Delta, \boldsymbol{\varPhi}(\boldsymbol{x}^{(t)})\big\rangle\Big]\\
&\overset{(b)}{\geq} \frac{\mathrm{e}^{-2r\bar{r}}}{n}\sum_{t=1}^{n}\frac{\big|\big\langle \Delta, \boldsymbol{\varPhi}(\boldsymbol{x}^{(t)})\big\rangle\big|^2}{2 + \big|\big\langle \Delta, \boldsymbol{\varPhi}(\boldsymbol{x}^{(t)})\big\rangle\big|}\\
&\overset{(c)}{\geq} \frac{\mathrm{e}^{-2r\bar{r}}}{2 + 8r\bar{r}}\cdot\frac{1}{n}\sum_{t=1}^{n}\big|\big\langle \Delta, \boldsymbol{\varPhi}(\boldsymbol{x}^{(t)})\big\rangle\big|^2\\
&\overset{(d)}{=} \frac{\mathrm{e}^{-2r\bar{r}}}{2 + 8r\bar{r}}\sum_{u_1=1}^{k_1}\sum_{v_1=1}^{k_2}\sum_{u_2=1}^{k_1}\sum_{v_2=1}^{k_2}\Delta_{u_1v_1}\hat{H}_{u_1v_1u_2v_2}\Delta_{u_2v_2}\\
&= \frac{\mathrm{e}^{-2r\bar{r}}}{2 + 8r\bar{r}}\sum_{u_1=1}^{k_1}\sum_{v_1=1}^{k_2}\sum_{u_2=1}^{k_1}\sum_{v_2=1}^{k_2}\Delta_{u_1v_1}\cdot[H_{u_1v_1u_2v_2} + \hat{H}_{u_1v_1u_2v_2} - H_{u_1v_1u_2v_2}]\Delta_{u_2v_2},
\end{aligned}$$

where (a) follows because $-\big\langle \mathbf{\Theta}^*, \boldsymbol{\varPhi}(\boldsymbol{x})\big\rangle \geq -2r\bar{r}$ for every $\boldsymbol{x} \in \mathcal{X}$ from (2.32), (b) follows because

$$\mathrm{e}^{-z} - 1 + z \geq \frac{z^2}{2 + |z|}, \qquad z \in \mathbb{R},$$

(c) follows because $-\big\langle \Delta, \boldsymbol{\varPhi}(\boldsymbol{x})\big\rangle \geq -8r\bar{r}$ for every $\boldsymbol{x} \in \mathcal{X}$ and $\Delta \in 4\varTheta$ from arguments similar to (2.32), and (d) follows from (2.34).

[9] Ideally, one would consider the gradient of $\mathcal{L}_n(\mathrm{vec}(\mathbf{\Theta}))$. However, we abuse the terminology for the sake of clarity and simplicity in the exposition.

Let the number of samples satisfy

$$n > \frac{128\phi_{\max}^4 \gamma(k_1,k_2)^4}{\lambda_{\min}^2} \log\Big(\frac{2k_1^2k_2^2}{\delta}\Big).$$

Using Lemma 2.1 with $\epsilon \hookleftarrow \lambda_{\min}/[2\gamma^2(k_1,k_2)]$, $\delta \hookleftarrow \delta$, and the triangle inequality, we have the following with probability at least $1-\delta$

$$\begin{aligned}
\delta\mathcal{L}_n(\Delta, \boldsymbol{\Theta}^*) &\geq \frac{\mathrm{e}^{-2r\bar{r}}}{2+8r\bar{r}} \sum_{u_1=1}^{k_1}\sum_{v_1=1}^{k_2}\sum_{u_2=1}^{k_1}\sum_{v_2=1}^{k_2} \Big[\Delta_{u_1v_1} H_{u_1v_1u_2v_2}\Delta_{u_2v_2} - \frac{\lambda_{\min}}{2\gamma^2(k_1,k_2)}\|\Delta\|_{1,1}^2\Big] \\
&\overset{(a)}{\geq} \frac{\mathrm{e}^{-2r\bar{r}}}{2+8r\bar{r}} \sum_{u_1=1}^{k_1}\sum_{v_1=1}^{k_2}\sum_{u_2=1}^{k_1}\sum_{v_2=1}^{k_2} \Big[\Delta_{u_1v_1} H_{u_1v_1u_2v_2}\Delta_{u_2v_2} - \frac{\lambda_{\min}}{2}\|\Delta\|_{\mathrm{F}}^2\Big] \\
&\overset{(b)}{=} \frac{\mathrm{e}^{-2r\bar{r}}}{2+8r\bar{r}} \Big[\mathrm{vec}(\Delta)\mathbb{E}[\mathrm{vec}(\boldsymbol{\Phi}(\mathbf{x}))\,\mathrm{vec}(\boldsymbol{\Phi}(\mathbf{x}))^{\mathrm{T}}]\,\mathrm{vec}(\Delta)^{\mathrm{T}} - \frac{\lambda_{\min}}{2}\|\Delta\|_{\mathrm{F}}^2\Big] \\
&\overset{(c)}{\geq} \frac{\mathrm{e}^{-2r\bar{r}}}{2+8r\bar{r}} \Big[\lambda_{\min}\|\,\mathrm{vec}(\Delta)\|_2^2 - \frac{\lambda_{\min}}{2}\|\Delta\|_{\mathrm{F}}^2\Big] \\
&\overset{(d)}{=} \frac{\mathrm{e}^{-2r\bar{r}}}{2+8r\bar{r}} \cdot \frac{\lambda_{\min}}{2}\|\Delta\|_{\mathrm{F}}^2,
\end{aligned}$$

where (a) follows because $\|\Delta\|_{1,1} \leq \gamma(k_1,k_2)\|\Delta\|_{\mathrm{F}}$, (b) follows from (2.33), (c) follows from the Courant-Fischer theorem (because $\mathbb{E}[\mathrm{vec}(\boldsymbol{\Phi}(\mathbf{x}))\,\mathrm{vec}(\boldsymbol{\Phi}(\mathbf{x}))^{\mathrm{T}}]$ is a symmetric matrix) and Assumption 2.4, and (d) follows because $\|\,\mathrm{vec}(\Delta)\|_2 = \|\Delta\|_{\mathrm{F}}$.

2.E Proof of Proposition 2.3: Bounded $\|\nabla\mathcal{L}_n(\boldsymbol{\Theta}^*)\|_{\max}$

The following lemma is useful in our analysis.

Lemma 2.2. *For any $u \in [k_1]$ and $v \in [k_2]$, define the random variable*

$$\mathsf{x}_{uv} \triangleq -\Phi_{uv}(\boldsymbol{x})\exp\big(-\langle\boldsymbol{\Theta}^*, \boldsymbol{\Phi}(\boldsymbol{x})\rangle\big). \tag{2.36}$$

Then, we have $\mathbb{E}[\mathsf{x}_{uv}] = 0$, where the expectation is with respect to $f_{\mathbf{x}}(\boldsymbol{x};\boldsymbol{\Theta}^*)$.

Proof of Lemma 2.2. Fix any $u \in [k_1]$ and $v \in [k_2]$. Using (2.36), we have

$$\begin{aligned}
\mathbb{E}[\mathsf{x}_{uv}] &= -\int_{\mathcal{X}} f_{\mathbf{x}}(\boldsymbol{x};\boldsymbol{\Theta}^*)\,\Phi_{uv}(\boldsymbol{x})\exp\big(-\langle\boldsymbol{\Theta}^*, \boldsymbol{\Phi}(\boldsymbol{x})\rangle\big)\,\mathrm{d}\boldsymbol{x} \\
&\overset{(a)}{=} \frac{-\int_{\mathcal{X}}\Phi_{uv}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}}{\int_{\mathcal{X}}\exp\big(\langle\boldsymbol{\Theta}^*, \boldsymbol{\Phi}(\boldsymbol{y})\rangle\big)\,\mathrm{d}\boldsymbol{y}} \overset{(b)}{=} 0,
\end{aligned}$$

where (a) follows from the definition of $f_{\mathbf{x}}(\boldsymbol{x};\boldsymbol{\Theta}^*)$ and because $\mathbb{E}_{\mathcal{U}_{\mathcal{X}}}[\boldsymbol{\Phi}(\mathbf{x})]$ is a constant, and (b) follows because $\int_{\mathcal{X}}\boldsymbol{\Phi}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = 0$ for centered natural statistics. □

We now proceed to establishing Proposition 2.3.

Proof of Proposition 2.3. Fix $u \in [k_1]$ and $v \in [k_2]$. We start by simplifying the gradient of the $\mathcal{L}_n(\boldsymbol{\Theta})$ evaluated at $\boldsymbol{\Theta}^*$. The component of the gradient of $\mathcal{L}_n(\boldsymbol{\Theta})$ corresponding to Θ_{uv}, evaluated at $\boldsymbol{\Theta}^*$, is given by

$$\frac{\partial \mathcal{L}_n(\boldsymbol{\Theta}^*)}{\partial \Theta_{uv}} = -\frac{1}{n}\sum_{t=1}^{n} \varPhi_{uv}(\boldsymbol{x}^{(t)}) \exp\Big(-\big\langle \boldsymbol{\Theta}^*, \boldsymbol{\varPhi}(\boldsymbol{x}^{(t)})\big\rangle\Big).$$

Each term in the above summation is distributed as per the random variable x_{uv} (see (2.36)). The random variable x_{uv} has zero mean (see Lemma 2.2) and satisfies $|\mathsf{x}_{uv}| \leq 2\phi_{\max} \exp(2r\bar{r})$ (from Assumption 2.3 and arguments similar to (2.32)). Using the Hoeffding's inequality, we have

$$\mathbb{P}\bigg(\bigg|\frac{\partial \mathcal{L}_n(\boldsymbol{\Theta}^*)}{\partial \Theta_{uv}}\bigg| > \epsilon\bigg) < 2\exp\bigg(\frac{-n\epsilon^2}{8\phi_{\max}^2 \exp(4r\bar{r})}\bigg). \tag{2.37}$$

The proof follows by using (2.37) and the union bound over all $u \in [k_1]$ and $v \in [k_2]$. □

2.F Proof of Theorem 2.2: Consistency and asymptotic normality

We prove Theorem 2.2 using the theory developed for M-estimation. In particular, observe that $\hat{\boldsymbol{\Theta}}_n$ is an M-estimator i.e., $\hat{\boldsymbol{\Theta}}_n$ is a sample average. Then, we invoke Theorem 4.1.1 and Theorem 4.1.3 of Amemiya (1985) to prove the consistency and the asymptotic normality of $\hat{\boldsymbol{\Theta}}_n$, respectively.

We divide the proof in two parts.

We first show that $\hat{\boldsymbol{\Theta}}_n$ is asymptotically consistent by applying Amemiya (1985, Theorem 4.1.1), viz.,

Theorem 2.4. *Let $z_1, \ldots, z_n$ be i.i.d. samples of a random variable z. Let $q(\mathsf{z};\theta)$ be some function of z parameterized by $\theta \in \Upsilon$. Let θ^* be the true underlying parameter. Define*

$$Q_n(\theta) \triangleq \frac{1}{n}\sum_{t=1}^{n} q(z_t;\theta) \qquad \textit{and} \qquad \hat{\theta}_n \in \operatorname*{arg\,min}_{\theta\in\Upsilon} Q_n(\theta).$$

Suppose the following conditions hold:

(a) Υ is compact,

(b) $Q_n(\theta) \xrightarrow{\text{p}} Q(\theta)$ uniformly for some nonstochastic $Q(\cdot)$,

(c) $Q(\theta)$ is continuous, and

(d) $Q(\theta)$ is uniquely minimized at θ^*.

Then, $\hat{\theta}_n$ is consistent for θ^* i.e., $\hat{\theta}_n \xrightarrow{\text{p}} \theta^*$ as $n \to \infty$.

To apply Theorem 2.4, we let $\mathsf{z} \triangleq \mathbf{x}$, $\theta \triangleq \boldsymbol{\Theta}$, $\hat{\theta}_n \triangleq \hat{\boldsymbol{\Theta}}_n$, $\theta^* \triangleq \boldsymbol{\Theta}^*$, $\Upsilon = \varTheta$, $q(z;\theta) \triangleq \exp\big(-\big\langle \boldsymbol{\Theta}, \boldsymbol{\varPhi}(\boldsymbol{x})\big\rangle\big)$, and $Q_n(\theta) \triangleq \mathcal{L}_n(\boldsymbol{\Theta})$. With these choices, we verify the theorem's conditions as follows.

Condition (a) We have $\mathit{\Theta} = \{\mathbf{\Theta} \colon \mathcal{R}(\mathbf{\Theta}) \leq r\}$ which is bounded and closed. Therefore, $\mathit{\Theta}$ is compact.

Condition (b) To verify that $\mathcal{L}_n(\mathbf{\Theta}) \xrightarrow{\mathrm{p}} \mathcal{L}(\mathbf{\Theta})$ uniformly for $\mathcal{L}(\mathbf{\Theta})$ in (2.5), we apply (Jennrich, 1969, Theorem 2), viz.,

Theorem 2.5. *Let $z_1, \ldots, z_n$ be i.i.d. samples of a random variable z. Let $g(z;\theta)$ be a function of z parameterized by $\theta \in \Upsilon$. Then, $n^{-1} \sum_{t\in[n]} g(z_t, \theta) \xrightarrow{\mathrm{p}} \mathbb{E}[g(\mathsf{z}, \theta)]$ uniformly if*

(i) Υ is compact,

(ii) $g(\mathsf{z}, \theta)$ is continuous at each $\theta \in \Upsilon$ with probability one,

(iii) $g(\mathsf{z}, \theta)$ is dominated by a function $G(\mathsf{z})$ i.e., $|g(\mathsf{z}, \theta)| \leq G(\mathsf{z})$, and

(iv) $\mathbb{E}[G(\mathsf{z})] < \infty$.

Using Theorem 2.5 with $\mathsf{z} \triangleq \mathbf{x}$, $\theta \triangleq \mathbf{\Theta}$, $\Upsilon \triangleq \mathit{\Theta}$, $g(\mathsf{z}, \theta) \triangleq \exp\bigl(-\bigl\langle \mathbf{\Theta}, \boldsymbol{\Phi}(\boldsymbol{x})\bigr\rangle\bigr)$, and $G(\mathsf{z}) \triangleq \exp(2r\overline{r})$, along with (2.32), we conclude that $\mathcal{L}_n(\mathbf{\Theta}) \xrightarrow{\mathrm{p}} \mathcal{L}(\mathbf{\Theta})$ uniformly.

Condition (c) Note that $\exp\bigl(-\bigl\langle \mathbf{\Theta}, \boldsymbol{\Phi}(\boldsymbol{x})\bigr\rangle\bigr)$ is a continuous function of $\mathbf{\Theta} \in \mathit{\Theta}$ and that $f_{\mathbf{x}}(\boldsymbol{x}; \mathbf{\Theta}^*)$ is not a function of $\mathbf{\Theta}$. Therefore, $\mathcal{L}(\mathbf{\Theta})$ is continuous for all $\mathbf{\Theta} \in \mathit{\Theta}$.

Condition (d) From Theorem 2.1, we have that $\mathcal{L}(\mathbf{\Theta})$ is uniquely minimized at $\mathbf{\Theta}^*$.

Since conditions (a)–(d) are satisfied, the consistency of $\hat{\mathbf{\Theta}}_n$ follows.

We next show that $\hat{\mathbf{\Theta}}_n$ is asymptotically normal by applying Amemiya (1985, Theorem 4.1.3), viz.:

Theorem 2.6. *Let $z_1, \ldots, z_n$ be i.i.d. samples of a random variable z. Let $q(\mathsf{z};\theta)$ be some function of z parameterized by $\theta \in \Upsilon$. Let θ^* be the true underlying parameter. Define*

$$Q_n(\theta) \triangleq \frac{1}{n} \sum_{t=1}^{n} q(z_t; \theta) \qquad \textit{and} \qquad \hat{\theta}_n \in \operatorname*{arg\,min}_{\theta\in\Upsilon} Q_n(\theta).$$

Suppose the following conditions hold:

(a) $\hat{\theta}_n$ is consistent for θ^*.

(b) $\theta^* \in \mathrm{int}(\Upsilon)$.

(c) Q_n is twice continuously differentiable in an open and convex neighbourhood of θ^*.

(d) $\sqrt{n}\nabla Q_n(\theta)|_{\theta=\theta^*} \xrightarrow{\mathrm{d}} \mathcal{N}(\mathbf{0}, \mathbf{A}(\theta^*))$.

(e) $\nabla^2 Q_n(\theta)|_{\theta=\hat{\theta}_n} \xrightarrow{\mathrm{p}} \mathbf{B}(\theta^*)$ with $\mathbf{B}(\theta)$ finite, nonsingular, and continuous at θ^*.

Then, $\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{\mathrm{d}} \mathcal{N}(\mathbf{0}, \mathbf{B}^{-1}(\theta^*)\mathbf{A}(\theta^*)\mathbf{B}^{-1}(\theta^*))$*.*

To apply Theorem 2.6, we let $z \triangleq \mathbf{x}$, $\theta \triangleq \mathbf{\Theta}$, $\hat{\theta}_n \triangleq \hat{\mathbf{\Theta}}_n$, $\theta^* \triangleq \mathbf{\Theta}^*$, $\Upsilon = \mathit{\Theta}$, $q(z;\theta) \triangleq \exp\big(-\langle \mathbf{\Theta}, \boldsymbol{\Phi}(\boldsymbol{x})\rangle\big)$, and $Q_n(\theta) \triangleq \mathcal{L}_n(\mathbf{\Theta})$. With these choices, we verify the theorem's conditions as follows.

Condition (a) We established $\hat{\mathbf{\Theta}}_n$ is consistent for $\mathbf{\Theta}^*$ in the first part of the proof.

Condition (b) We have assumed $\mathbf{\Theta}^* \in \mathrm{int}(\mathit{\Theta})$.

Condition (c) Fix $u_1, u_2 \in [k_1]$ and $v_1, v_2 \in [k_2]$. We have

$$\frac{\partial^2 \mathcal{L}_n(\mathbf{\Theta})}{\partial \Theta_{u_1v_1} \partial \Theta_{u_2v_2}} = \frac{1}{n}\sum_{t=1}^{n} \mathit{\Phi}_{u_1v_1}(\boldsymbol{x}^{(t)})\mathit{\Phi}_{u_2v_2}(\boldsymbol{x}^{(t)}) \cdot \exp\big(-\langle \mathbf{\Theta}, \boldsymbol{\Phi}(\boldsymbol{x}^{(t)})\rangle\big).$$

Hence, $\partial^2 \mathcal{L}_n(\mathbf{\Theta})/\partial \Theta_{u_1v_1}\partial\Theta_{u_2v_2}$ exists. Using the continuity of $\boldsymbol{\Phi}(\cdot)$ and $\exp\big(-\langle \mathbf{\Theta}, \boldsymbol{\Phi}(\cdot)\rangle\big)$, we see that $\partial^2 \mathcal{L}_n(\mathbf{\Theta})/\partial \Theta_{u_1v_1}\partial\Theta_{u_2v_2}$ is continuous in an open and convex neighborhood of $\mathbf{\Theta}^*$.

Condition (d) For any $u \in [k_1]$ and $v \in [k_2]$, define the random variable

$$\mathsf{x}_{uv} \triangleq -\mathit{\Phi}_{uv}(\boldsymbol{x}) \exp\big(-\langle \mathbf{\Theta}^*, \boldsymbol{\Phi}(\boldsymbol{x})\rangle\big).$$

The component of the gradient of $\mathcal{L}_n(\mathrm{vec}(\mathbf{\Theta}))$ corresponding to Θ_{uv}, evaluated at $\mathbf{\Theta}^*$, is given by

$$\frac{\partial \mathcal{L}_n(\mathbf{\Theta}^*)}{\partial \Theta_{uv}} = -\frac{1}{n}\sum_{t=1}^{n} \mathit{\Phi}_{uv}(\boldsymbol{x}^{(t)}) \exp\big(-\langle \mathbf{\Theta}^*, \boldsymbol{\Phi}(\boldsymbol{x}^{(t)})\rangle\big).$$

Each term in the above summation is distributed as per the random variable x_{uv}. The random variable x_{uv} has zero mean (see Lemma 2.2). Using this, and the multivariate central limit theorem (Van der Vaart, 2000, Example 2.18), we have

$$\sqrt{n}\nabla\mathcal{L}_n(\mathrm{vec}(\mathbf{\Theta}))|_{\mathbf{\Theta}=\mathbf{\Theta}^*} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{A}(\mathbf{\Theta}^*)),$$

where $\mathbf{A}(\mathbf{\Theta}^*)$ is the covariance matrix of $\mathrm{vec}\big(\boldsymbol{\Phi}(\mathbf{x})\exp\big(-\langle \mathbf{\Theta}^*, \boldsymbol{\Phi}(\mathbf{x})\rangle\big)\big)$.

Condition (e) Finally, we verify that

$$\nabla^2\mathcal{L}_n(\mathrm{vec}(\mathbf{\Theta}))|_{\mathbf{\Theta}=\hat{\mathbf{\Theta}}_n} \xrightarrow{\mathrm{p}} \mathbf{B}(\mathbf{\Theta}^*)$$

with $\mathbf{B}(\mathbf{\Theta})$ finite, nonsingular, and continuous at $\mathbf{\Theta}^*$. We start by showing that

$$\nabla^2\mathcal{L}_n(\mathrm{vec}(\mathbf{\Theta}))|_{\mathbf{\Theta}=\hat{\mathbf{\Theta}}_n} \xrightarrow{p} \nabla^2\mathcal{L}(\mathrm{vec}(\mathbf{\Theta}))|_{\mathbf{\Theta}=\mathbf{\Theta}^*}. \tag{2.38}$$

Using the uniform law of large numbers (Jennrich, 1969, Theorem 2) for any $\boldsymbol{\Theta} \in \mathit{\Theta}$ results in

$$\nabla^2 \mathcal{L}_n(\mathrm{vec}(\boldsymbol{\Theta})) \xrightarrow{p} \nabla^2 \mathcal{L}(\mathrm{vec}(\boldsymbol{\Theta})). \tag{2.39}$$

Using the consistency of $\hat{\boldsymbol{\Theta}}_n$, and the continuous mapping theorem, we have

$$\nabla^2 \mathcal{L}(\mathrm{vec}(\boldsymbol{\Theta}))|_{\boldsymbol{\Theta}=\hat{\boldsymbol{\Theta}}_n} \xrightarrow{p} \nabla^2 \mathcal{L}(\mathrm{vec}(\boldsymbol{\Theta}))|_{\boldsymbol{\Theta}=\boldsymbol{\Theta}^*}. \tag{2.40}$$

Let $u_1, u_2 \in [k_1]$ and $v_1, v_2 \in [k_2]$. From (2.39) and (2.40), for any $\epsilon > 0$ and $\delta > 0$, there exists integers n_1, n_2 such that for $n \geq \max\{n_1, n_2\}$, we have

$$\mathbb{P}\left(\left|\frac{\partial^2 \mathcal{L}_n(\hat{\boldsymbol{\Theta}}_n)}{\partial \Theta_{u_1 v_1} \partial \Theta_{u_2 v_2}} - \frac{\partial^2 \mathcal{L}(\hat{\boldsymbol{\Theta}}_n)}{\partial \Theta_{u_1 v_1} \partial \Theta_{u_2 v_2}}\right| > \frac{\epsilon}{2}\right) \leq \frac{\delta}{2}$$

and

$$\mathbb{P}\left(\left|\frac{\partial^2 \mathcal{L}(\hat{\boldsymbol{\Theta}}_n)}{\partial \Theta_{u_1 v_1} \partial \Theta_{u_2 v_2}} - \frac{\partial^2 \mathcal{L}(\boldsymbol{\Theta}^*)}{\partial \Theta_{u_1 v_1} \partial \Theta_{u_2 v_2}}\right| > \frac{\epsilon}{2}\right) \leq \frac{\delta}{2}.$$

Now for $n \geq \max\{n_1, n_2\}$, we have, using the triangle inequality,

$$\mathbb{P}\left(\left|\frac{\partial^2 \mathcal{L}_n(\hat{\boldsymbol{\Theta}}_n)}{\partial \Theta_{u_1 v_1} \partial \Theta_{u_2 v_2}} - \frac{\partial^2 \mathcal{L}(\boldsymbol{\Theta}^*)}{\partial \Theta_{u_1 v_1} \partial \Theta_{u_2 v_2}}\right| > \epsilon\right) \leq \frac{\delta}{2} + \frac{\delta}{2} = \delta.$$

Hence, we have (2.38). Using the definition of $\mathcal{L}(\boldsymbol{\Theta})$, we have

$$\begin{aligned}
\partial^2 \mathcal{L}(\boldsymbol{\Theta}^*)/\partial \Theta_{u_1 v_1} \partial \Theta_{u_2 v_2} &= \mathbb{E}\big[\mathit{\Phi}_{u_1 v_1}(\mathbf{x}) \mathit{\Phi}_{u_2 v_2}(\mathbf{x}) \exp\big(-\big\langle \boldsymbol{\Theta}^*, \boldsymbol{\mathit{\Phi}}(\mathbf{x})\big\rangle\big)\big] \\
&\overset{(a)}{=} \mathbb{E}\big[\mathit{\Phi}_{u_1 v_1}(\mathbf{x}) \mathit{\Phi}_{u_2 v_2}(\mathbf{x}) \exp\big(-\big\langle \boldsymbol{\Theta}^*, \boldsymbol{\mathit{\Phi}}(\mathbf{x})\big\rangle\big)\big] \\
&\quad - \mathbb{E}\big[\mathit{\Phi}_{u_1 v_1}(\mathbf{x})\big] \mathbb{E}\Big[\mathit{\Phi}_{u_2 v_2}(\mathbf{x}) \exp\big(-\big\langle \boldsymbol{\Theta}^*, \boldsymbol{\mathit{\Phi}}(\mathbf{x})\big\rangle\big)\Big] \\
&= \mathrm{cov}\big(\mathit{\Phi}_{u_1 v_1}(\mathbf{x}), \mathit{\Phi}_{u_2 v_2}(\mathbf{x}) \exp\big(-\big\langle \boldsymbol{\Theta}^*, \boldsymbol{\mathit{\Phi}}(\mathbf{x})\big\rangle\big)\big),
\end{aligned}$$

where to obtain (a) we have used that, from Lemma 2.2,

$$\mathbb{E}\big[\mathit{\Phi}_{u_2 v_2}(\mathbf{x}) \exp\big(-\big\langle \boldsymbol{\Theta}^*, \boldsymbol{\mathit{\Phi}}(\mathbf{x})\big\rangle\big)\big] = 0$$

for $u_2 \in [k_1]$ and $v_2 \in [k_2]$, whence

$$\nabla^2 \mathcal{L}_n(\mathrm{vec}(\boldsymbol{\Theta}))|_{\boldsymbol{\Theta}=\hat{\boldsymbol{\Theta}}_n} \xrightarrow{p} \mathbf{B}(\boldsymbol{\Theta}^*),$$

where $\mathbf{B}(\boldsymbol{\Theta}^*)$ is the cross-covariance matrix of $\mathrm{vec}(\boldsymbol{\mathit{\Phi}}(\mathbf{x}))$ and $\mathrm{vec}\big(\boldsymbol{\mathit{\Phi}}(\mathbf{x}) \exp\big(-\big\langle \boldsymbol{\Theta}^*, \boldsymbol{\mathit{\Phi}}(\mathbf{x})\big\rangle\big)\big)$. Finiteness and continuity of $\boldsymbol{\mathit{\Phi}}(\mathbf{x})$ and $\boldsymbol{\mathit{\Phi}}(\mathbf{x}) \exp\big(-\big\langle \boldsymbol{\Theta}^*, \boldsymbol{\mathit{\Phi}}(\mathbf{x})\big\rangle\big)$ implies the finiteness and continuity of $\mathbf{B}(\boldsymbol{\Theta}^*)$. By assumption, the cross-covariance matrix of $\mathrm{vec}(\boldsymbol{\mathit{\Phi}}(\mathbf{x}))$ and $\mathrm{vec}\big(\boldsymbol{\mathit{\Phi}}(\mathbf{x}) \exp\big(-\big\langle \boldsymbol{\Theta}^*, \boldsymbol{\mathit{\Phi}}(\mathbf{x})\big\rangle\big)\big)$ is invertible.

Since conditions (a)–(e) are satisfied, the asymptotic normality of $\hat{\boldsymbol{\Theta}}_n$ follows.

2.G Proof of Theorem 2.3: Finite sample guarantees

We establish our result by applying Negahban et al. (2012, Corollary 1), viz.:

Theorem 2.7. *Let $z_1, \ldots, z_n$ be i.i.d. samples of a random variable z. Let $q(\mathsf{z};\theta)$ be some convex and differentiable function of z parameterized by $\theta \in \Upsilon$. Define*

$$\hat{\theta}_n \in \arg\min_{\theta} \frac{1}{n}\sum_{i=1}^{n} q(z_i;\theta) + \lambda_n \mathcal{R}(\theta),$$

where λ_n is a regularization penalty and $\mathcal{R}$ is a norm. Let θ^* be the true underlying parameter, i.e., $\theta^* \in \arg\min_{\theta\in\Upsilon} \mathbb{E}[q(z;\theta)]$. Suppose the following conditions hold:

(a) The regularization penalty is such that $\lambda_n \geq 2\mathcal{R}^*(\nabla\mathcal{L}_n(\theta^*))$ where $\mathcal{R}^*$ is the dual norm of $\mathcal{R}$.

(b) The loss function satisfies a restricted strong convexity condition with curvature $\kappa > 0$, i.e., $\delta\mathcal{L}_n(\Delta,\theta^*) \geq \kappa\|\Delta\|_2^2$, where $\delta\mathcal{L}_n(\Delta,\theta^*)$ is the residual of the first-order Taylor expansion.

Then $\hat{\theta}_n$ is such that

$$\|\hat{\theta}_n - \theta^*\|_2 \leq 3\frac{\lambda_n}{\kappa}\Psi(\Upsilon) \text{ where } \Psi(\Upsilon) = \max_{\boldsymbol{v}\in\Upsilon\setminus\{\mathbf{0}\}} \frac{\mathcal{R}(\boldsymbol{v})}{\|\boldsymbol{v}\|_2}. \tag{2.41}$$

To apply Theorem 2.7, we let $\mathsf{z} \triangleq \mathbf{x}$, $\theta \triangleq \boldsymbol{\Theta}$, $\hat{\theta}_n \triangleq \hat{\boldsymbol{\Theta}}_n$, $\theta^* \triangleq \boldsymbol{\Theta}^*$, $\Upsilon = \mathit{\Theta}$, $q(z;\theta) \triangleq \exp\bigl(-\bigl\langle \boldsymbol{\Theta}, \boldsymbol{\Phi}(\boldsymbol{x})\bigr\rangle\bigr)$,

$$\lambda_n = \frac{2\alpha\lambda_{\min}}{(24+96r\bar{r})\,\mathrm{e}^{2r\bar{r}}\Psi(k_1,k_2)}, \tag{2.42a}$$

and

$$\kappa = \frac{\lambda_{\min}\,\mathrm{e}^{-2r\bar{r}}}{4+16r\bar{r}}. \tag{2.42b}$$

While Theorem 2.7 is a deterministic result, we use probabilistic analysis to show that the necessary conditions hold, resulting in a high probability bound. Towards that, let the number of samples satisfy

$$n \geq \max\{n_1, n_2\},$$

where

$$n_1 \triangleq \frac{128\phi_{\max}^4\gamma(k_1,k_2)^4}{\lambda_{\min}^2}\log\Bigl(\frac{4k_1^2k_2^2}{\delta}\Bigr),$$

and

$$n_2 \triangleq \frac{8\phi_{\max}^2(24+96r\bar{r})^2\,\mathrm{e}^{8r\bar{r}}\,g(k_1,k_2)^2\,\Psi(k_1,k_2)^2}{\alpha^2\lambda_{\min}^2}\cdot\log\Bigl(\frac{4k_1k_2}{\delta}\Bigr). \tag{2.43}$$

In other words, we have (2.23). With these choices, we verify the theorem's conditions as follows.

Condition (a) To establish this, we need to show that the choice of λ_n in (2.42a) upper bounds $\mathcal{R}^*(\nabla\mathcal{L}_n(\boldsymbol{\Theta}^*))$. From (2.16) in Section 2.6, we have $\mathcal{R}^*(\nabla\mathcal{L}_n(\boldsymbol{\Theta}^*)) \leq g(k_1,k_2)\|\nabla\mathcal{L}_n(\boldsymbol{\Theta}^*)\|_{\max}$. Then, using Proposition 2.3 with $\delta \hookleftarrow \delta/2$ and

$$\epsilon \hookleftarrow \frac{\alpha\lambda_{\min}}{(24+96r\overline{r})\,\mathrm{e}^{2r\overline{r}}g(k_1,k_2)\Psi(k_1,k_2)},$$

as long as $n \geq n_2$, we have $2\mathcal{R}^*(\nabla\mathcal{L}_n(\boldsymbol{\Theta}^*)) \leq 2g(k_1,k_2)\,\epsilon = \lambda_n$ with probability at least $1-\delta/2$.

Condition (b) Let $\Delta = \hat{\boldsymbol{\Theta}}_n - \boldsymbol{\Theta}^*$. Then, Lemma 1 of Negahban et al. (2012) implies that Δ is such that $\mathcal{R}(\Delta) \leq 4\mathcal{R}(\boldsymbol{\Theta}^*)$, i.e., $\Delta \in 4\varTheta$. Then, using Proposition 2.2 with $\delta \hookleftarrow \delta/2$ and κ as in (2.42b), as long as $n \geq n_1$, we have $\delta\mathcal{L}_n(\Delta,\boldsymbol{\Theta}^*) \geq \kappa\|\Delta\|_{\mathrm{F}}^2$ with probability at least $1-\delta/2$.

Putting everything together in (2.41), we have $\|\hat{\boldsymbol{\Theta}}_n - \boldsymbol{\Theta}^*\|_{\mathrm{F}} \leq \alpha$ with probability at least $1-\delta$.

2.H Proof of Corollary 2.1

To provide a poly(k_1k_2) dependence on the sample complexity, we show that each of the functions $\gamma(k_1,k_2)$, $g(k_1,k_2)$, and $\Psi(k_1,k_2)$ can be bounded by a poly(k_1k_2) term. First, it is easy to see

$$\gamma(k_1,k_2) \leq \sqrt{k_1k_2},$$

for any $\varTheta$. Next, we note that any bound on $g(k_1,k_2)$ is a bound on $\Psi(k_1,k_2)$ because a) $\|\mathbf{M}\|_{\max} \leq \|\mathbf{M}\|_{\mathrm{F}}$ for any $\mathbf{M}$ and b) each of the entry-wise $L_{p,q}$ norms, the Schatten p-norms, and the operator p-norms are closed under the dual operation. Below, we provide a bound on $g(k_1,k_2)$ for each of these family of norms.

2.H.1 Entry-wise $L_{p,q}$ norms

For any matrix $\mathbf{M} \in \mathbb{R}^{k_1\times k_2}$, we have

$$\begin{aligned}\|\mathbf{M}\|_{p,q} = \Bigg(\sum_{j\in[k_2]}\Bigg(\sum_{i\in[k_1]}|M_{ij}|^p\Bigg)^{q/p}\Bigg)^{1/q} &\leq \Bigg(\sum_{j\in[k_2]}\Bigg(\sum_{i\in[k_1]}\|\mathbf{M}\|_{\max}^p\Bigg)^{q/p}\Bigg)^{1/q}\\ &= k_1^{1/p}k_2^{1/q}\|\mathbf{M}\|_{\max}.\end{aligned}$$

Therefore, $g(k_1,k_2) \leq k_1^{1/p}k_2^{1/q}$.

2.H.2 Schatten p-norms

For any matrix $\mathbf{M} \in \mathbb{R}^{k_1\times k_2}$, let r denote its rank. We have

$$\|\mathbf{M}\|_p^\star = \Bigg(\sum_{i\in[r]}\sigma_i^p(\mathbf{M})\Bigg)^{1/p} \overset{(a)}{\leq} \sum_{i\in[r]}\sigma_i(\mathbf{M}) \overset{(b)}{\leq} \sqrt{rk_1k_2}\|\mathbf{M}\|_{\max}$$

$$\overset{(c)}{\leq} \sqrt{\min\{k_1, k_2\} k_1 k_2} \|\mathbf{M}\|_{\max},$$

where (a) follows because of the monotonicity of the Schatten p-norms, (b) follows because $\|\mathbf{M}\|^\star \leq \sqrt{r k_1 k_2} \|\mathbf{M}\|_{\max}$, and (c) follows because $r \leq \min\{k_1, k_2\}$. Therefore, $g(k_1, k_2) \leq \sqrt{\min\{k_1, k_2\} k_1 k_2}$.

2.H.3 Operator p-norms

Define $q \triangleq p/(p-1)$. For any matrix $\mathbf{M} \in \mathbb{R}^{k_1 \times k_2}$, let $[\mathbf{M}]_i$ denote the ith row of $\mathbf{M}$ for $i \in k_1$. We have

$$\begin{aligned}
\|\mathbf{M}\|_p = \max_{\boldsymbol{y}\colon \|\boldsymbol{y}\|_p = 1} \|\mathbf{M}\boldsymbol{y}\|_p &\overset{(a)}{\leq} k_1^{1/p} \max_{\boldsymbol{y}\colon \|\boldsymbol{y}\|_p = 1} \|\mathbf{M}\boldsymbol{y}\|_\infty \\
&\overset{(b)}{\leq} k_1^{1/p} \max_{\boldsymbol{y}\colon \|\boldsymbol{y}\|_p = 1} \max_{i \in [k_1]} \|[\mathbf{M}]_i\|_q \|\boldsymbol{y}\|_p \\
&\leq k_1^{1/p} \max_{i \in [k_1]} \|[\mathbf{M}]_i\|_q \\
&\overset{(c)}{\leq} k_1^{1/p} k_2^{1/q} \max_{i \in [k_1]} \|[\mathbf{M}]_i\|_\infty = k_1^{1/p} k_2^{1-1/p} \|\mathbf{M}\|_{\max},
\end{aligned}$$

where (a) follows because $\|\boldsymbol{v}\|_p \leq m^{1/p} \|\boldsymbol{v}\|_\infty$ for any vector $\boldsymbol{v} \in \mathbb{R}^m$ and $p \geq 1$, (b) follows from the definition of the infinity norm of a vector and using the Hölder's inequality, and (c) follows because $\|\boldsymbol{v}\|_q \leq m^{1/q} \|\boldsymbol{v}\|_\infty$ for any vector $\boldsymbol{v} \in \mathbb{R}^m$ and $q \geq 1$. Therefore, $g(k_1, k_2) \leq k_1^{1/p} k_2^{1-1/p}$.

Chapter 3

Causal Inference via Exponential Family Modeling

3.1 Introduction

We are interested in the problem of unit-level counterfactual inference owing to the increasing importance of personalized decision-making in many domains. As a motivating example, consider an observational dataset corresponding to an interaction between a recommender system and a user over time. At each time, the user was exposed to a product based on observed demographic factors as well as factors that are not observed in the dataset, e.g., user's energy level (i.e., whether they're feeling energetic or tired). Additionally, at each time, the user's engagement level, which could have sequentially depended on the prior interaction in addition to the ongoing interaction, was also recorded. Also, the system could have sequentially adapted its recommendation. Given such data of many heterogeneous users (e.g., a movie recommender system for a streaming media platform), we want to infer each user's average engagement level if it were exposed to a different sequence of products while the observed and the unobserved factors remain unchanged. This task is challenging since: (a) the unobserved factors could give rise to spurious associations, (b) the users could be heterogeneous in that they may have different responses to same sequence of products, and (c) each user provides a single interaction trajectory.

More generally, to address counterfactual problems of this kind, we consider an observational setting where a unit undergoes multiple interventions (or treatments) denoted by $\mathbf{a}$. We denote the outcomes of interest by $\mathbf{y}$, and allow the interventions $\mathbf{a}$ and the outcomes $\mathbf{y}$ to be confounded by observed covariates $\mathbf{v}$ as well as unobserved covariates $\mathbf{z}$. The graphical structure shown in Figure 3.1.1 captures these interactions for a unit and is at the heart of our problem. In the recommender system example above, a unit corresponds to a user, $\mathbf{a}$ corresponds to the products recommended, $\mathbf{y}$ corresponds to the engagement levels, $\mathbf{v}$ corresponds to the observed demographic factors, and $\mathbf{z}$ corresponds to the unobserved energy levels (see Figure 3.1.2). More generally, a unit may comprise of one or more users/individuals. We consider n heterogeneous and independent units indexed by $i \in [n] \triangleq \{1, \cdots, n\}$, and assume access to one

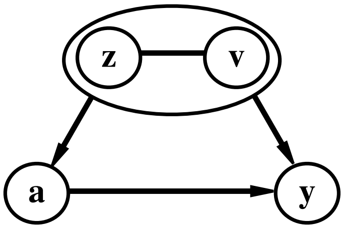


Figure 3.1.1: A generic model covered by our methodology. Directed arrows denote causation and undirected arrows denote association. All directed arrows denote high-level causal links, i.e., aggregated low-level causal links. For example, the high-level causal link between **a** and **y** captures all low-level causal links between any element of **a** and any element of **y**. Our methodology does not assume knowledge of any low-level causal link and is applicable to any graphical model with high-level causal links between variables as in this model.

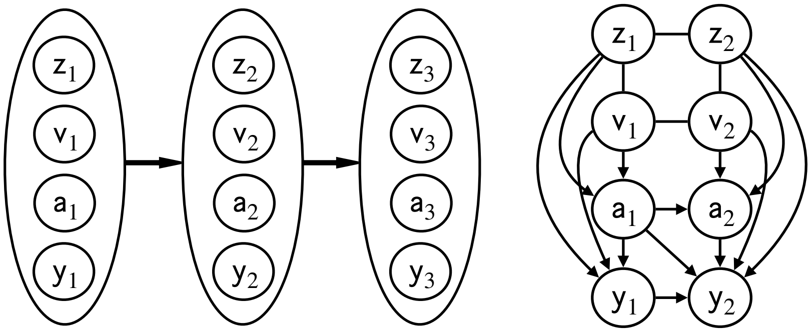


Figure 3.1.2: An example graphical model for a sequential recommender system (consistent with the model in Figure 3.1.1) interacting with a user at 3 time points where z_t, v_t, a_t, and y_t denote the user's unobserved energy levels, observed demographic factors, the product exposed to the user, and the user's engagement level, respectively, at time t. The left subplot illustrates the high-level dependency between the variables (with thick arrows) while the right subplot expands on it for time 1 and 2 (with thin arrows).

observation per unit with $(\boldsymbol{v}^{(i)}, \boldsymbol{a}^{(i)}, \boldsymbol{y}^{(i)})$ denoting the realizations of $(\mathbf{v}, \mathbf{a}, \mathbf{y})$ for unit i.

We operate within the Neyman-Rubin potential outcomes framework (Neyman, 1923; Rubin, 1974) and denote the potential outcome of unit $i \in [n]$ under interventions $\boldsymbol{a}$ by $\boldsymbol{y}^{(i)}(\boldsymbol{a})$. Given the realizations $\{(\boldsymbol{v}^{(i)}, \boldsymbol{a}^{(i)}, \boldsymbol{y}^{(i)})\}_{i=1}^{n}$, our goal is to answer counterfactual questions for these n units. For example, what would the potential outcomes $\boldsymbol{y}^{(i)}(\widetilde{\boldsymbol{a}}^{(i)})$ for interventions $\widetilde{\boldsymbol{a}}^{(i)} \neq \boldsymbol{a}^{(i)}$ be, while the observed and unobserved covariates remain unchanged? Under the graphical model in Figure 3.1.1 and the stable unit treatment value assumption (SUTVA), i.e., the potential outcomes of unit i are not affected by the interventions at other units[1], learning unit-level counterfactual distributions is equivalent to learning unit-level conditional distributions

$$\left\{f_{\mathbf{y}|\mathbf{a},\mathbf{z},\mathbf{v}}(\mathbf{y} = \cdot | \mathbf{a} = \cdot, \boldsymbol{z}^{(i)}, \boldsymbol{v}^{(i)})\right\}_{i=1}^{n}. \tag{3.1}$$

[1] We note that the potential outcomes of a user can be affected by other users in the same unit (if a unit comprises of multiple users) but not by users in a different unit (see Figure 3.3.1).

Here, the i-th distribution represents the conditional distribution for the outcomes $\mathbf{y}$ as a function of the interventions $\mathbf{a}$, while keeping the observed covariates $\mathbf{v}$ and the unobserved covariates $\mathbf{z}$ fixed at the corresponding realizations for unit i, i.e., $\boldsymbol{v}^{(i)}$ and $\boldsymbol{z}^{(i)}$, respectively.

Such questions cannot be answered without structural assumptions due to two key challenges: (a) unobserved confounding and (b) single observation per unit. First, the unobserved covariates $\mathbf{z}$ introduce spurious statistical dependence between interventions and outcomes, termed unobserved confounding, which results in biased estimates. Second, we only observe one realization, namely the outcomes $\boldsymbol{y}^{(i)}(\boldsymbol{a}^{(i)})$ under the interventions $\boldsymbol{a}^{(i)}$, that is consistent with the unit-level conditional distribution $f_{\mathbf{y}|\mathbf{a},\mathbf{z},\mathbf{v}}(\boldsymbol{y}|\boldsymbol{a}, \boldsymbol{z}^{(i)}, \boldsymbol{v}^{(i)})$. As a result, we need to learn n heterogeneous conditional distributions while having access to only one sample from each of them.

In this work, we model the conditional distribution of the outcomes of interest conditioned on the unobserved covariates, the observed covariates, the intervention as an exponential family distribution motivated by the principle of maximum entropy.[2] With this model structure, we show that both the aforementioned challenges can be tackled. In particular, we show that the n unit-level conditional distributions in (3.1) lead to n distributions from the same exponential family, albeit with parameters that vary across units. The parameter corresponding to the i^{th} unit, for brevity in terminology denoted by $\gamma^{(i)}$ (defined later), captures the effect of $\boldsymbol{z}^{(i)}$ and helps tackle the challenge of unobserved confounding. However, the challenge still remains to learn n heterogeneous exponential family distributions with one sample per distribution.

This challenge has been addressed in two specific scenarios in the literature: (a) if the unobserved confounding is identical across units, i.e., the parameters $\{\gamma^{(i)}\}_{i=1}^{n}$ were all equal, then the challenge boils down to learning parameters of a single exponential family distribution from n samples, which has been well-studied (cf. Shah et al. (2021b) for an overview); (b) if $\mathbf{v}$, $\mathbf{a}$, and $\mathbf{y}$ take binary values and have pairwise interactions, then the challenge boils down to learning parameters of an Ising model (a special sub-class of exponential family defined later) with one sample. This specific challenge has been studied under restricted settings: (i) where the dependencies between the variables are known (e.g., Kandiros et al. (2021); Mukherjee et al. (2021)) and (ii) where a specific subset of the parameters are known (Dagan et al., 2021). In this work, we consider a generalized setting where $\mathbf{v}$, $\mathbf{a}$, and $\mathbf{y}$ can be either discrete, continuous, or both, and do not assume that the underlying dependencies or a specific subset of parameters are known.

3.1.1 Summary of contributions

This work introduces a method to learn unit-level counterfactual distributions from observational studies, in the presence of unobserved confounding, with one sample per unit, using exponential family modeling. For every unit $i \in [n]$, we reduce learning its counterfactual distribution to learning the unit-specific parameter $\gamma^{(i)}$ with access to one

[2] Exponential family distributions are the maximum entropy distributions given linear constraints on distributions such as specifying the moments (see Jaynes (1957)).

sample $(\boldsymbol{v}^{(i)}, \boldsymbol{a}^{(i)}, \boldsymbol{y}^{(i)})$ from unit i. Here, $\{\gamma^{(1)}, \cdots, \gamma^{(n)}\}$ are parameters of n different distributions from the same exponential family. The specific technical contributions are as follows:

1. We introduce a convex (and strictly proper) loss function (Definition 3.1) that pools the data $\big\{(\boldsymbol{v}^{(i)}, \boldsymbol{a}^{(i)}, \boldsymbol{y}^{(i)})\big\}_{i=1}^{n}$ across all n samples to jointly learn all n parameters $\{\gamma^{(i)}\}_{i=1}^{n}$.
2. For every unit i, we prove that the mean squared errors of our estimates of (a) $\gamma^{(i)}$ (Theorem 3.1) and (b) the expected potential outcomes under alternate interventions (Theorem 3.2) scale linearly with the metric entropy of the underlying parameter space. For instance, when $\gamma^{(i)}$ is s-sparse linear combination of k known vectors (Corollary 3.1), the error—just with one sample—decays as $O(s \log k/p_y)$, where p_y is the dimension of the outcome $\mathbf{y}$.
3. Our framework extends some of the widely used panel data models from econometrics. In particular, we allow for dynamics in the outcomes, the actions, and the observed covariates for the linear and logistic unit fixed effect models as well as the linear and logistic time fixed effect models. Further, we allow parameters to vary with unit and time for the unit fixed effect models, and the parameters to vary with time for the time fixed effect models.
4. We apply our method to impute missing covariates when they are sparse. Formally, we consider a setup (with no systematically unobserved covariates) where the observed covariates are entirely missing for some fixed fraction of the units. Specifically, for unit i with missing covariates, only $(\boldsymbol{a}^{(i)}, \boldsymbol{y}^{(i)})$ is observed. For every such unit, we show that our method can recover the missing covariates with the mean squared error decaying as $O(p_v/p_y)$, where p_v and p_y are the dimensions of $\mathbf{v}$ and $\mathbf{y}$, respectively (Proposition 3.4).
5. Methodologically, our work advances two threads: (a) learning Ising models (and their extensions to discrete, continuous, or mixed variables) from a single sample, where we learn the dependencies between variables, generalizing prior work Dagan et al. (2021); Kandiros et al. (2021) and (b) learning Markov random fields (a sub-class of exponential family) from multiple independent but non-identical samples, generalizing prior work Shah et al. (2021c); Vuffray et al. (2016a, 2022a).
6. In our analysis, we (a) derive sufficient conditions for a continuous random vector supported on a compact set to satisfy the logarithmic Sobolev inequality (Proposition 3.5) and (b) provide new concentration bounds for arbitrary functions of a continuous random vector that satisfies the logarithmic Sobolev inequality (Proposition 3.6). These results may be of independent interest.

Outline. Section 3.2 discusses background and related work. We discuss our formulation and algorithm in Section 3.3 and present their analysis in Section 3.4. We provide some extensions of our model in Section 3.5 and provide some connections to panel data models in Section 3.6. We develop an application of our methodology to impute missing covariates in Section 3.7. We sketch the proof of our main result in Section 3.8 with detailed proofs deferred to the appendices. We conclude with a discussion in Section 3.9.

Notation. For any positive integer n, let $[n] \coloneqq \{1, \cdots, n\}$. For a deterministic sequence $u_1, \cdots, u_n$, we let $\boldsymbol{u} \coloneqq (u_1, \cdots, u_n)$. For a random sequence $\mathsf{u}_1, \cdots, \mathsf{u}_n$, we let $\mathbf{u} \coloneqq (\mathsf{u}_1, \cdots, \mathsf{u}_n)$. For a vector $\boldsymbol{u} \in \mathbb{R}^p$, we use u_t to denote its t^{th} coordinate and $\boldsymbol{u}_{-t} \in \mathbb{R}^{p-1}$ to denote the vector after deleting the t^{th} coordinate. We denote the ℓ_0, ℓ_q ($q \geq 1$), and ℓ_∞ norms of a vector $\boldsymbol{v}$ by $\|\boldsymbol{v}\|_0$, $\|\boldsymbol{v}\|_p$, and $\|\boldsymbol{v}\|_\infty$, respectively. For a matrix $\mathbf{M} \in \mathbb{R}^{p\times p}$, we denote the element in t^{th} row and u^{th} column by $\mathbf{M}_{tu}$, the t^{th} row by $\mathbf{M}_t$, and the vector obtained after deleting $\mathbf{M}_{tt}$ from $\mathbf{M}_t$ by $\mathbf{M}_{t,-t}$. Further, we denote the matrix maximum norm by $\|\mathbf{M}\|_{\max}$, the Frobenius norm by $\|\mathbf{M}\|_{\mathrm{F}}$, the spectral norm (operator 2-norm) by $\|\mathbf{M}\|_{\mathrm{op}}$, the induced $1-$norm (operator 1-norm) by $\|\mathbf{M}\|_1$, the induced ∞-norm (operator ∞-norm) by $\|\mathbf{M}\|_\infty$, and the $(2,\infty)$-norm by $\|\mathbf{M}\|_{2,\infty}$. For any matrix $\mathbf{M}$, let $\lambda_{\max}(\mathbf{M})$ and $\lambda_{\min}(\mathbf{M})$ denote the largest and the smallest eigenvalues of $\mathbf{M}$, respectively. Finally, for vectors $\widehat{\boldsymbol{u}} \in \mathbb{R}^p$ and $\widetilde{\boldsymbol{u}} \in \mathbb{R}^p$, the mean squared error between $\widehat{\boldsymbol{u}}$ and $\widetilde{\boldsymbol{u}}$ is defined as $\mathrm{MSE}(\widehat{\boldsymbol{u}}, \widetilde{\boldsymbol{u}}) \triangleq p^{-1}\sum_{t\in[p]}(\widehat{u}_t - \widetilde{u}_t)^2$.

3.2 Background and Related Work

This work builds on two vast bodies of literature: exponential family learning and unit-level counterfactual inference with unobserved confounding. For a detailed literature overview of the former, we refer the readers to Bresler (2015); Klivans and Meka (2017); Shah et al. (2021c); Vuffray et al. (2022a) (for a special sub-class, Markov random fields (MRFs)[3]) and Shah et al. (2021b) for general exponential families. For an introduction to counterfactual inference, see the books Hernán and Robins (2020); Imbens and Rubin (2015b) for settings with no unobserved confounding and Pearl (2009); Pearl et al. (2016) for settings with known causal mechanism (in the form of a causal graph).

3.2.1 Exponential family learning

There is a series of works for learning Ising models, a special MRF with binary variables and an instance of a pair-wise exponential family, from a single sample. Such a model has two distinct sets of parameters capturing the contribution of nodes and edges in the underlying undirected graph, referred to as the external field and the interaction matrix.[4] Many strategies exist for learning such a model when the interaction matrix is known up to a constant and under varying assumptions on the external field; see, e.g., Bhattacharya and Mukherjee (2018); Chatterjee (2007); Daskalakis et al. (2019); Ghosal and Mukherjee (2020); Kandiros et al. (2021); Mukherjee et al. (2021). More recently, Dagan et al. (2021) provide guarantees for learning the interaction matrix from a single sample when the external field is known. Kandiros et al. (2021) and Mukherjee et al. (2021) extend the tools in Dagan et al. (2021) to learn the external field for an Ising model with a known interaction matrix (up to a scalar multiple). Notably, all of these works are based on the pseudo-likelihood estimation (Besag, 1975b). Our work

[3] MRFs can be naturally represented as exponential family distributions with certain sparsity constraints on the parameters via the principle of maximum entropy (Wainwright et al., 2008).

[4] E.g., in our model (defined later in Eq. (3.2)), $\phi^{(y)}$ and $\Phi^{(y,y)}$ correspond to the external field and the interaction matrix, respectively.

extends the techniques and results from Dagan et al. (2021) to learn the external field from one sample of continuous variables with an estimated interaction matrix.

Vuffray et al. (2016a) introduced a novel M-estimation-based loss function for learning Ising models from many independent and identically distributed samples. Vuffray et al. (2022a) and Shah et al. (2021c) generalize it to learn general MRFs with multi-ary discrete and continuous variables, respectively. Ren et al. (2021) showed that this loss function has superior numerical performance compared to the ones based on pseudo-likelihood. We contribute to this line of work by generalizing that loss function further to learn MRFs with discrete, continuous, and mixed variables with independent but not identically distributed samples.

For settings closer to our work, namely, exponential families with unobserved variables, the two common modeling approaches include restricted Boltzmann machines (Bresler and Buhai, 2020; Bresler et al., 2019; Goel, 2020) and latent variable Gaussian graphical models; see, e.g., Chandrasekaran et al. (2012); Ma et al. (2013); Vinyes and Obozinski (2018); Wang et al. (2023). While the former assumes a bipartite structure with edges only across observed and unobserved variables, the latter imposes a Gaussian generative model. In this thread, most related to our set-up is the work by Taeb et al. (2020) as they model the conditional distribution of the observed variables conditioned on the unobserved variables as an exponential family similar to us. They provide empirically promising results for recovering the underlying graph and the number of unobserved variables (assumed to be small), albeit with limited theoretical guarantees. In contrast, here we provide parameter estimation error in the presence of unobserved variables (notably, we cover all the models they considered).

3.2.2 Unit-level counterfactual inference

Recent years have seen an active interest in developing different strategies for unit-level inference with unobserved confounding.

For the settings with univariate outcomes for each unit, a common approach to deal with unobserved confounding is the instrumental variable (IV) method (Imbens and Angrist, 1994) when one has access to a variable—the IV—that induces changes in intervention assignment but has no independent effect on outcomes allowing causal effect estimation. Recent works for IV methods with unit-level inference include Athey et al. (2019); Hartford et al. (2017); Semenova and Chernozhukov (2021); Singh et al. (2019); Syrgkanis et al. (2019); Wang et al. (2022); Xu et al. (2020). Another approach for univariate outcomes, called causal sensitivity analysis (Rosenbaum and Rubin, 1983a), estimates the worst-case effect on the causal estimand as a function of the extent of unobserved confounding in a given dataset under varying assumptions on the generative model. For such analysis with unit-level guarantees, see, e.g., Jesson et al. (2021); Jin et al. (2023); Kallus et al. (2019); Yadlowsky et al. (2022); Yin et al. (2022). In another related thread, Arkhangelsky and Imbens (2018) use an exponential family to model the unit-wise distribution of the observed covariates and interventions conditioned on the unobserved covariates. They connect this model to the commonly used unit fixed effects model for the outcomes (Angrist and Pischke, 2009), and provide estimates for the average treatment effect given multiple units with the same set of unobserved covariates. By contrast, our work uses an exponential family to model the unit-wise distribution of the outcomes conditioned on the interventions, the observed covariates, and the unobserved covariates. Further, we allow each unit to have a different set of unobserved covariates while the intervention and the outcome can be high-dimensional, and provide the first unit-level counterfactual inference guarantee with an exponential family model.

Closer to our work are those on panel or longitudinal data settings, where one observes multiple outcomes for each unit. For linear panel data settings, a common approach is factor modeling, where potential outcomes and interventions (binary or multi-ary) are assumed to be independent conditional on some latent factors. See, e.g., difference-in-difference methods (Angrist and Pischke, 2009; Bertrand et al., 2004), synthetic control (Abadie et al., 2010b; Abadie and Gardeazabal, 2003b), its variants Arkhangelsky et al. (2021); Dwivedi et al. (2022b), and extensions to multi-ary interventions in synthetic interventions (Agarwal et al., 2020) and sequential experiments (Dwivedi et al., 2022a). For non-linear panel data settings, the most commonly used models include probit, logit, Poisson, negative binomial, proportional hazard, and tobit models (see Fernández-Val and Weidner (2018) for an overview) where some parametric model characterises the distribution of the outcomes conditional on the unobserved covariates, the observed covariates, and the interventions. Notably, these works on linear and non-linear panel data directly estimate effects (averaged over all observed and unobserved covariates or unit-level for given observed and unobserved covariates) for finitely many interventions when the intervention assignment has special structure, while we focus on learning the counterfactual distributions while allowing for multi-ary discrete and continuous interventions without any special structure. Our work also generalizes some of these models by allowing for dynamics in the outcomes, the actions, and the observed covariates.

3.3 Problem Formulation and Algorithm

This section formalizes the problem, specifies our model, and defines the inference tasks of interest.

3.3.1 Underlying causal mechanism and counterfactual distributions

We consider a counterfactual inference task where units go through $p_a \geq 1$ interventions. For every unit, we observe $p_y \geq 1$ outcomes of interest. The interventions and the outcomes could be confounded by $p_v \geq 0$ observed covariates as well as $p_z \geq 0$ unobserved covariates. Additionally, the observed covariates and the unobserved covariates could be arbitrarily associated. We denote the random vector associated with the interventions, the outcomes, the observed covariates, and the unobserved covariates by $\mathbf{a} \triangleq (a_1, \cdots, a_{p_a}) \in \mathcal{A}^{p_a}$, $\mathbf{y} = (y_1, \cdots, y_{p_y}) \in \mathcal{Y}^{p_y}$, $\mathbf{v} \triangleq (v_1, \cdots, v_{p_v}) \in \mathcal{V}^{p_v}$, and $\mathbf{z} \triangleq (z_1, \cdots, z_{p_z}) \in \mathcal{Z}^{p_z}$, respectively, where $\mathcal{A}, \mathcal{Y}, \mathcal{V}$, and $\mathcal{Z}$ denote the support of

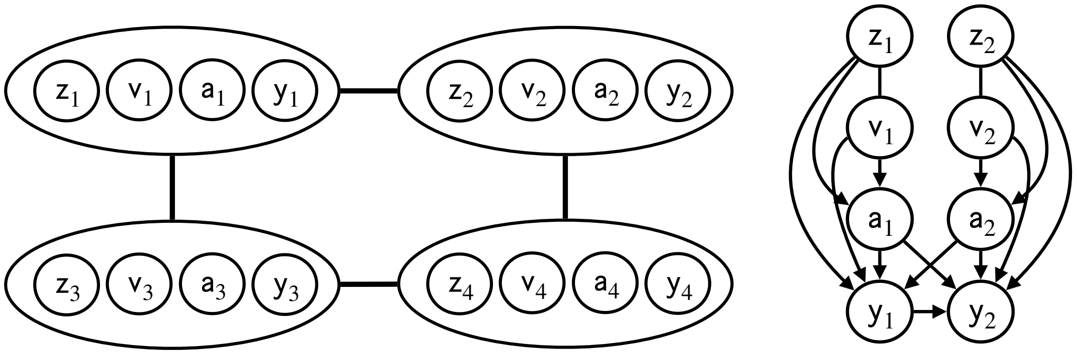


Figure 3.3.1: A graphical model for a single unit in the network setting with 4 users; arrows have same meaning as in Figures 3.1.1 and 3.1.2. Here v_t, z_t, a_t, and y_t denote user t's observed factors, unobserved factors, exposed product, and engagement level, respectively. The left plot illustrates the high-level dependency between the variables of different users in the network, and the right plot expands on it for (user 1, user 2) pair. Analogous dependencies exist for (user 1, user 3), (user 2, user 4), and (user 3, user 4) pairs.

interventions, outcomes, observed covariates, and unobserved covariates, respectively. We allow these sets to contain discrete, continuous, or mixed values.

Causal mechanism. We summarize the causal relationship between the random vectors $\mathbf{z}$, $\mathbf{v}$, $\mathbf{a}$, and $\mathbf{y}$ in Figure 3.1.1 where we denote the arbitrary association between $\mathbf{z}$ and $\mathbf{v}$ by an undirected arrow, and the causal association between (i) $(\mathbf{z}, \mathbf{v})$ and $\mathbf{a}$, (ii) $(\mathbf{z}, \mathbf{v})$ and $\mathbf{y}$, and (iii) $\mathbf{a}$ and $\mathbf{y}$ by directed arrows. More generally, we are interested in any setup where the high-level causal links are consistent with the graphical model in Figure 3.1.1 . We assume access to n independent realizations indexed by $i \in [n]$: $\boldsymbol{v}^{(i)}$, $\boldsymbol{a}^{(i)}$, and $\boldsymbol{y}^{(i)}$ denote the realizations of $\mathbf{v}$, $\mathbf{a}$, and $\mathbf{y}$ for unit i, respectively. For every realized tuple $(\boldsymbol{v}^{(i)}, \boldsymbol{a}^{(i)}, \boldsymbol{y}^{(i)})$, there is a corresponding realization $\boldsymbol{z}^{(i)}$ of the unobserved covariates $\mathbf{z}$ that is unobserved. Next, we discuss some examples covered by our framework.

Examples: sequential and network settings. While Figure 3.1.1 exhibits the high-level causal links between $\mathbf{z}$, $\mathbf{v}$, $\mathbf{a}$, and $\mathbf{y}$, there could be complex low-level causal links between elements of these vectors. We do not assume any knowledge of such low-level causal links. In Figure 3.1.2 , we provide an instance of a sequential setting covered by our work where every unit's (i) a_{t+1} depends on a_t in addition to v_{t+1} and $\mathbf{z}$, and (ii) y_{t+1} depends on a_t and y_t in addition to a_{t+1}, v_{t+1} and $\mathbf{z}$. Another classical example covered by our framework includes the network setting where a single unit represents a social network where multiple users are linked to each other by interpersonal relationships as shown in Figure 3.3.1. Similar to the sequential recommender system, every user was exposed to a product based on observed demographic factors as well as certain unobserved factors, and the user's engagement level was recorded. The engagement level of user t, i.e., y_t, depended its observed demographic factors v_t, its unobserved factors z_t, its exposed product a_t as well as on the product exposed to its neighbor u, i.e., $\mathbf{a}_u$. Further, $\mathbf{y}_t$ could have been associated with $\mathbf{y}_u$. The users in the same unit affect each other but the users in different units do not affect each other.

Unit-level counterfactual distributions. We denote the Neyman-Rubin potential outcomes of unit $i \in [n]$ under interventions $\boldsymbol{a} \in \mathcal{A}^{p_a}$ by $\boldsymbol{y}^{(i)}(\boldsymbol{a})$. We make the stable unit treatment value assumption (SUTVA) (Rubin, 1980) for the observed outcome, i.e., $\boldsymbol{y}^{(i)} = \boldsymbol{y}^{(i)}(\boldsymbol{a}^{(i)})$ for all units $i \in [n]$. For independent units with the causal mechanism and SUTVA assumed here, the unit-level counterfactual distributions are equivalent to certain unit-level conditional distributions as we now argue. Consider unit $i \in [n]$ and fix the observed covariates and the unobserved covariates at $\boldsymbol{v}^{(i)}$ and $\boldsymbol{z}^{(i)}$, respectively. Then, let $\widetilde{\boldsymbol{y}}^{(i)}$ be a realization of $\mathbf{y}$ when $\mathbf{a} = \widetilde{\boldsymbol{a}}^{(i)}$. We are interested in the distribution of the potential outcomes of unit i for interventions $\widetilde{\boldsymbol{a}}^{(i)}$, i.e., the distribution of $\boldsymbol{y}^{(i)}(\widetilde{\boldsymbol{a}}^{(i)})$ given $\mathbf{v} = \boldsymbol{v}^{(i)}, \mathbf{z} = \boldsymbol{z}^{(i)}$. Under the causal framework considered here (see Figure 3.1.1), it is equivalent to the distribution of $\boldsymbol{y}^{(i)}(\widetilde{\boldsymbol{a}}^{(i)})$ given $\mathbf{a} = \widetilde{\boldsymbol{a}}^{(i)}, \mathbf{v} = \boldsymbol{v}^{(i)}, \mathbf{z} = \boldsymbol{z}^{(i)}$ since $(\mathbf{z}, \mathbf{v})$ satisfy ignorability (Imbens and Rubin, 2015b; Pearl, 2009), i.e., the potential outcomes are independent of the interventions given $(\mathbf{z}, \mathbf{v})$. Further, under SUTVA, it is equivalent to the distribution of $\widetilde{\boldsymbol{y}}^{(i)}$ given $\mathbf{a} = \widetilde{\boldsymbol{a}}^{(i)}, \mathbf{v} = \boldsymbol{v}^{(i)}, \mathbf{z} = \boldsymbol{z}^{(i)}$, i.e., $f_{\mathbf{y}|\mathbf{a},\mathbf{v},\mathbf{z}}(\mathbf{y} = \cdot | \mathbf{a} = \widetilde{\boldsymbol{a}}^{(i)}, \boldsymbol{z}^{(i)}, \boldsymbol{v}^{(i)})$. Therefore, our goal is to learn the n unit-level conditional distributions in Eq. (3.1). Now, we proceed to the modeling details.

3.3.2 Exponential family modeling and its consequences

We start by parameterizing the conditional distribution $f_{\mathbf{y}|\mathbf{a},\mathbf{v},\mathbf{z}}$ with parameters $\phi^{(y)} \in \mathbb{R}^{p_y \times 1}$ and $\Phi^{(u,y)} \in \mathbb{R}^{p_u \times p_y}$ for all $\mathbf{u} \in \{\mathbf{z}, \mathbf{v}, \mathbf{a}, \mathbf{y}\}$, and natural statistics $\mathbf{y}$ and $\mathbf{y}\mathbf{y}^\top$ so that

$$f_{\mathbf{y}|\mathbf{a},\mathbf{v},\mathbf{z}}(\boldsymbol{y}|\boldsymbol{a}, \boldsymbol{v}, \boldsymbol{z}) \propto \exp\Big(\big[\phi^{(y)^\top} + 2\boldsymbol{z}^\top \Phi^{(z,y)} + 2\boldsymbol{v}^\top \Phi^{(v,y)} + 2\boldsymbol{a}^\top \Phi^{(a,y)}\big]\boldsymbol{y} + \boldsymbol{y}^\top \Phi^{(y,y)} \boldsymbol{y}\Big), \quad (3.2)$$

where $\boldsymbol{z} \triangleq (z_1, \cdots, z_{p_z})$, $\boldsymbol{v} \triangleq (v_1, \cdots, v_{p_v})$, $\boldsymbol{a} \triangleq (a_1, \cdots, a_{p_a})$, and $\boldsymbol{y} \triangleq (y_1, \cdots, y_{p_y})$ denote realizations of $\mathbf{z}$, $\mathbf{v}$, $\mathbf{a}$, and $\mathbf{y}$, respectively. Here, the parameter $\Phi^{(u,y)}$ captures the interaction between $\mathbf{u}$ and $\mathbf{y}$, for all $\mathbf{u} \in \{\mathbf{z}, \mathbf{v}, \mathbf{a}, \mathbf{y}\}$. [5] Without loss of generality, we can assume $\Phi^{(y,y)}$ to be a symmetric matrix. We note that the conditional distribution $f_{\mathbf{y}|\mathbf{a},\mathbf{v},\mathbf{z}}$ being an exponential family puts no restrictions on the marginal distribution $f_{\mathbf{z},\mathbf{v},\mathbf{a}}$ of the unobserved covariates, the observed covariates, and the interventions as is the case with semi-parametric causal models (Kennedy, 2016). We provide various examples of panel data models consistent with (3.2) in Section 3.6.

We make two key observations: (a) the term $\Phi^{(z,y)\top}\boldsymbol{z}$ captures the effect of unobserved covariates $\boldsymbol{z}$ on $f_{\mathbf{y}|\mathbf{a},\mathbf{v},\mathbf{z}}(\mathbf{y} = \cdot | \mathbf{a} = \cdot, \boldsymbol{v}, \boldsymbol{z})$ and (b) the task of learning $f_{\mathbf{y}|\mathbf{a},\mathbf{v},\mathbf{z}}(\mathbf{y} = \cdot | \mathbf{a} = \cdot, \boldsymbol{v}, \boldsymbol{z})$ in Eq. (3.2) as a function of $\mathbf{a}$ reduces to learning

$$\text{(i) } \phi^{(y)} + 2\Phi^{(z,y)\top}\boldsymbol{z} + 2\Phi^{(v,y)\top}\boldsymbol{v}, \qquad \text{(ii) } \Phi^{(a,y)}, \quad \text{and} \quad \text{(iii) } \Phi^{(y,y)}. \qquad (3.3)$$

[5] The exponential family in Eq. (3.2) is same as the one considered in Taeb et al. (2020, Equation 1.3).

That is, learning the unit-level conditional distribution for unit i is equivalent to learning

$$\gamma^{(i)} = \big\{\phi^{(y)} + 2\Phi^{(z,y)\top}\boldsymbol{z}^{(i)} + 2\Phi^{(v,y)\top}\boldsymbol{v}^{(i)}, \Phi^{(a,y)}, \Phi^{(y,y)}\big\},$$

where the notation $\gamma^{(i)}$ is the same as in Section 3.1.

Next, we note that learning the three quantities in Eq. (3.3) is subsumed in learning the parameters $\theta(\boldsymbol{z}) \in \mathbb{R}^{p_y \times 1}$ and $\Theta \in \mathbb{R}^{p_y \times \widetilde{p}}$ where $\widetilde{p} \triangleq p_y + p_a + p_v$,

$$\theta(\boldsymbol{z}) \triangleq \phi^{(y)} + 2\Phi^{(z,y)\top}\boldsymbol{z} \quad \text{and} \quad \Theta \triangleq \big[\Phi^{(y,y)}, \Phi^{(y,a)}, \Phi^{(y,v)}\big],$$

with $\Phi^{(y,a)} = \Phi^{(a,y)^\top} \in \mathbb{R}^{p_y \times p_a}$ and $\Phi^{(y,v)} = \Phi^{(v,y)^\top} \in \mathbb{R}^{p_y \times p_v}$. To exploit this, we observe that the (unit-level) conditional distribution $f_{\mathbf{y}|\mathbf{a},\mathbf{v},\mathbf{z}}$ in Eq. (3.2) can be reparameterized as follows:

$$f_{\mathbf{y}|\mathbf{a},\mathbf{v},\mathbf{z}}\big(\boldsymbol{y}|\boldsymbol{a}, \boldsymbol{v}, \boldsymbol{z}; \theta(\boldsymbol{z}), \Theta\big) \propto \exp\Big([\theta(\boldsymbol{z})]^\top \boldsymbol{y} + 2\boldsymbol{v}^\top \Phi^{(v,y)}\boldsymbol{y} + 2\boldsymbol{a}^\top \Phi^{(a,y)}\boldsymbol{y} + \boldsymbol{y}^\top \Phi^{(y,y)}\boldsymbol{y}\Big). \tag{3.4}$$

Given some estimates for $\theta(\boldsymbol{z})$ and Θ, using their appropriate components also yields an estimate of the three quantities in Eq. (3.3) for any $\mathbf{v} = \boldsymbol{v}$. To summarize, the spurious associations or unobserved confounding between $\boldsymbol{a}$ and $\boldsymbol{y}$ introduced due to unobserved $\mathbf{z}$ are fully captured by $\Phi^{(z,y)\top}\boldsymbol{z}$ or equivalently by $\theta(\boldsymbol{z})$; thereby, learning unit-level counterfactual distributions require us to learn these unit-level parameters.

3.3.2.1 Reduced inference task and modeling constraints

Let $f_{\mathbf{y}|\mathbf{a},\mathbf{v},\mathbf{z}}\big(\cdot\,|\boldsymbol{a}, \boldsymbol{v}, \boldsymbol{z}; \theta^\star(\boldsymbol{z}), \Theta^\star\big)$ denote the true distribution of $\mathbf{y}$ conditioned on $\mathbf{a} = \boldsymbol{a}$, $\mathbf{v} = \boldsymbol{v}$, and $\mathbf{z} = \boldsymbol{z}$ as in Eq. (3.4). Then, for all $i \in [n]$, the realization $(\boldsymbol{y}^{(i)}, \boldsymbol{a}^{(i)}, \boldsymbol{v}^{(i)})$ is consistent with the conditional distribution $f_{\mathbf{y}|\mathbf{a},\mathbf{v},\mathbf{z}}\big(\cdot\,|\boldsymbol{a}^{(i)}, \boldsymbol{v}^{(i)}, \boldsymbol{z}^{(i)}; \theta^\star(\boldsymbol{z}^{(i)}), \Theta^\star\big)$ where we do not observe $\boldsymbol{z}^{(i)}$. Our primary goal is to learn the n unit-level counterfactual distributions, which as noted above simplifies to estimating the following parameters:

$$\text{(i) Unit-level } \theta^{\star(i)} \triangleq \theta^\star(\boldsymbol{z}^{(i)}) \text{ for } i \in [n], \quad \text{and (ii) Population-level} \quad \Theta^\star. \tag{3.5}$$

Our secondary goal is to estimate the expected potential outcomes for any given unit i (with $\mathbf{v} = \boldsymbol{v}^{(i)}, \mathbf{z} = \boldsymbol{z}^{(i)}$) and an alternate intervention $\widetilde{\boldsymbol{a}}^{(i)}$:

$$\mu^{(i)}(\widetilde{\boldsymbol{a}}^{(i)}) \triangleq \mathbb{E}[\boldsymbol{y}^{(i)}(\widetilde{\boldsymbol{a}}^{(i)})|\mathbf{v} = \boldsymbol{v}^{(i)}, \mathbf{z} = \boldsymbol{z}^{(i)}], \tag{3.6}$$

where $\boldsymbol{y}^{(i)}(\widetilde{\boldsymbol{a}}^{(i)})$ denotes the potential outcomes for unit $i \in [n]$ under interventions $\widetilde{\boldsymbol{a}}^{(i)} \in \mathcal{A}^{p_a}$.

For ease of exposition, we consider compact continuous sets $\mathcal{V}$, $\mathcal{A}$, and $\mathcal{Y}$ with $\mathcal{V} = \mathcal{A} = \mathcal{Y} \triangleq \mathcal{X} = [-x_{\max}, x_{\max}]$ for a given $x_{\max}$. It is straightforward to extend the analysis when $\mathcal{V} \neq \mathcal{A} \neq \mathcal{Y}$. In Section 3.5.2, we consider compact discrete and mixed sets. Throughout this paper, it is convenient to further constrain the model as follows:

Assumption 3.1 (Bounded and sparse parameters)**.** *The true model parameters Eq.* (3.5) *satisfy*

$$\theta^{\star(i)} \in \Lambda_\theta \triangleq \left\{\theta \in \mathbb{R}^{p_y \times 1} : \left\|\theta\right\|_\infty \leq \alpha\right\} \text{ for all } i \in [n], \tag{3.7}$$

$$\Theta^\star \in \Lambda_\Theta \triangleq \left\{\Theta = [\Phi^{(y,y)}, \Phi^{(y,a)}, \Phi^{(y,v)}] \in \mathbb{R}^{p_y \times \widetilde{p}} : \Phi^{(y,y)} = \Phi^{(y,y)^\top}, \|\Theta\|_{\max} \leq \alpha, \ \|\Theta\|_\infty \leq \beta\right\}. \tag{3.8}$$

While Eq. (3.7) bounds the unit-level parameters (a necessary condition for model identifiability (Santhanam and Wainwright, 2012)), Eq. (3.8) bounds the ℓ_1 norm of the interaction of each $\mathsf{y}_t \in \mathbf{y}$ with the vector $\mathbf{x} \triangleq (\mathbf{y}, \mathbf{a}, \mathbf{v})$ in Eq. (3.4). As a result, Assumption 3.1 implies that the exponential family in Eq. (3.4) corresponds to MRFs (see Section 3.2), also known as undirected graphical models (defined in Section 3.H). We note that Assumption 3.1 is standard in the literature on learning MRFs (Bresler, 2015; Klivans and Meka, 2017; Shah et al., 2021c; Vuffray et al., 2016a, 2022a). We are now ready to state our algorithm.

3.3.3 An efficient algorithm via a convex objective

We first describe our strategy to estimate the parameters in Eq. (3.5). Then, we use the estimated parameters to estimate the expected potential outcomes in Eq. (3.6). We remark that for exponential families considered here, maximum likelihood for parameter estimation is not computationally tractable (Shah et al., 2021b; Wainwright et al., 2008). As a result, we resort to an alternative objective function inspired by the convex loss functions used in Shah et al. (2021c); Vuffray et al. (2016a, 2022a) as they do not depend on the partition function of the distribution. These loss functions are designed in a specific way (see below for details): (i) the sufficient statistics of the conditional distribution of a variable given all other variables are *centered* by adding appropriate constants, (ii) the loss function is an empirical average of the sum of the inverses of all of these conditional distributions (without the partition function) with *centered* sufficient statistics.

3.3.3.1 Parameter estimation

Our convex objective function jointly learns all the parameters of interest by pooling the observations across all n units and exploiting the exponential family structure of $\mathbf{y}$ conditioned on $\mathbf{a} = \boldsymbol{a}$, $\mathbf{v} = \boldsymbol{v}$, and $\mathbf{z} = \boldsymbol{z}$ in Eq. (3.4), i.e., the objective explicitly utilizes the fact that the population-level parameter $\Theta^\star$ is shared across units. In particular, we use the following two steps.

Centering sufficient statistics of the conditional distribution of a variable. With $\mathbf{x} = (\mathbf{y}, \mathbf{a}, \mathbf{v})$, we have $\mathsf{x}_t = \mathsf{y}_t$ and $\mathbf{x}_{-t} = (\mathbf{y}_{-t}, \mathbf{a}, \mathbf{v})$ for every $t \in [p_y]$. Then, consider the conditional distribution $f_{\mathsf{x}_t | \mathbf{x}_{-t}, \mathbf{z}}$ of the random variable x_t conditioned on $\mathbf{x}_{-t} = \boldsymbol{x}_{-t}$ and $\mathbf{z} = \boldsymbol{z}$ for any $t \in [p_y]$:

$$f_{\mathsf{x}_t | \mathbf{x}_{-t}, \mathbf{z}}(x_t | \boldsymbol{x}_{-t}, \boldsymbol{z}; \theta_t(\boldsymbol{z}), \Theta_t) \propto \exp\left(\left[\theta_t(\boldsymbol{z}) + 2\Theta_{t,-t}^\top \boldsymbol{x}_{-t}\right] x_t + \Theta_{tt} x_t^2\right), \tag{3.9}$$

where $\theta_t(\boldsymbol{z})$ is the t^{th} element of $\theta(\boldsymbol{z})$, Θ_t is the t^{th} row of Θ, Θ_{tt} is the t^{th} element of Θ_t, and $\Theta_{t,-t} \triangleq \Theta_t \setminus \Theta_{tt} \in \mathbb{R}^{\widetilde{p}-1}$ is the vector obtained after deleting Θ_{tt} from Θ_t. Then, the sufficient statistics in Eq. (3.9), namely x_t and x_t^2, are centered by subtracting their expected value with respect to the uniform distribution on $\mathcal{X}$ resulting in

$$f_{\mathsf{x}_t|\mathbf{x}_{-t},\mathbf{z}}\big(x_t|\boldsymbol{x}_{-t},\boldsymbol{z};\theta_t(\boldsymbol{z}),\Theta_t\big) \propto \exp\left(\big[\theta_t(\boldsymbol{z}) + 2\Theta_{t,-t}^\top \boldsymbol{x}_{-t}\big]x_t + \Theta_{tt}\Big(x_t^2 - \frac{x_{\max}^2}{3}\Big)\right), \quad (3.10)$$

as the integral of x_t and x_t^2 with respect to the uniform distribution on $\mathcal{X}$ is 0 and $x_{\max}^2/3$, respectively. As we see later (in Proposition 3.1), this centering ensures that our loss function is a proper loss function as well as leads to connections with the surrogate likelihood (Shah et al., 2021c, Proposition. 4.1). We emphasize that the term $x_{\max}^2/3$ inside the exponent in Eq. (3.10) is vacuous (as it is a constant) and the distribution in Eq. (3.10) is equivalent to the one in Eq. (3.9).

Constructing the loss function. Next, the loss function (defined below) is designed to be an empirical average of the sum over $t \in [p_y]$ of the inverse of the term in the right hand side of Eq. (3.10).

Definition 3.1 (Loss function)**.** *Given the samples $\{\boldsymbol{x}^{(i)}\}_{i\in[n]}$, the loss $\mathcal{L} : \mathbb{R}^{p_y\times(n+\widetilde{p})} \to \mathbb{R}$ is given by*

$$\mathcal{L}\big(\underline{\Theta}\big) = \frac{1}{n}\sum_{t\in[p_y]}\sum_{i\in[n]}\exp\left(-\big[\theta_t^{(i)} + 2\Theta_{t,-t}^\top \boldsymbol{x}_{-t}^{(i)}\big]x_t^{(i)} - \Theta_{tt}\Big([x_t^{(i)}]^2 - \frac{x_{\max}^2}{3}\Big)\right) \textit{ where}$$

$$\underline{\Theta} \triangleq \begin{bmatrix} \underline{\Theta}_1^\top \\ \vdots \\ \underline{\Theta}_p^\top \end{bmatrix} \in \mathbb{R}^{p_y\times(n+\widetilde{p})}, \textit{ with } \underline{\Theta}_t \triangleq \big\{\theta_t^{(1)},\cdots,\theta_t^{(n)},\Theta_t\big\} \in \mathbb{R}^{n+p}. \quad (3.11)$$

Our estimate of $\underline{\Theta}^\star$ (defined analogous to $\underline{\Theta}$) is given by

$$\widehat{\underline{\Theta}} \in \underset{\underline{\Theta}\in\Lambda_\theta^n\times\Lambda_\Theta}{\arg\min}\ \mathcal{L}\big(\underline{\Theta}\big). \quad (3.12)$$

We note Eq. (3.12) is a convex optimization problem, and a projected gradient descent algorithm (see Section 3.A.2) returns an ϵ-optimal estimate where $\widehat{\underline{\Theta}}_\epsilon$ is said to be an ϵ-optimal estimate if $\mathcal{L}\big(\widehat{\underline{\Theta}}_\epsilon\big) \le \mathcal{L}\big(\widehat{\underline{\Theta}}\big) + \epsilon$ for any $\epsilon > 0$. The loss function $\mathcal{L}$ admits a notable property (see Section 3.A.1 for the proof).

Proposition 3.1 (Proper loss function)**.** *The loss function $\mathcal{L}$ is strictly proper, i.e., $\underline{\Theta}^\star = \arg\min_{\underline{\Theta}\in\Lambda_\theta^n\times\Lambda_\Theta} \mathbb{E}_{\mathbf{y}|\mathbf{a},\mathbf{v},\mathbf{z}}\big[\mathcal{L}\big(\underline{\Theta}\big)\big]$.*

Proposition 3.1 shows that the solution of the idealized convex program $\min_{\underline{\Theta}\in\Lambda_\theta^n\times\Lambda_\Theta} \mathbb{E}_{\mathbf{y}|\mathbf{a},\mathbf{v},\mathbf{z}}\big[\mathcal{L}\big(\underline{\Theta}\big)\big]$ is unique and equal to $\underline{\Theta}^\star$. In this idealized convex program, conditioned on the realized values of the unobserved covariates of the n units $\boldsymbol{z}^{(1)},\cdots,\boldsymbol{z}^{(n)}$, the observed covariates of the n units $\boldsymbol{v}^{(1)},\cdots,\boldsymbol{v}^{(n)}$, and the interventions of the n units $\boldsymbol{a}^{(1)},\cdots,\boldsymbol{a}^{(n)}$, the loss function is averaged over all the randomness in the outcomes. In other words, for every $i \in [n]$, the idealized convex program has infinite samples from $f_{\mathsf{y}|\mathsf{a},\mathsf{v},\mathsf{z}}$ with unobserved covariates $\mathbf{z}$, observed covariates $\mathbf{v}$, and interventions $\mathbf{a}$ conditioned to be $\boldsymbol{z}^{(i)}$, $\boldsymbol{v}^{(i)}$, and $\boldsymbol{a}^{(i)}$, respectively. Thus, the convex program in Eq. (3.12) can be seen as a single sample version of this idealized program, thereby providing an intuitive justification of our loss function (instead of a maximum likelihood objective, which is not tractable here). As we show later in our proofs (see Section 3.8 for an overview), different partial averages on the RHS of Eq. (3.11) also admit useful properties and are critical to our analyses.

We note that loss function in Eq. (3.11) is a generalization of the loss functions used in Shah et al. (2021c); Vuffray et al. (2016a, 2022a). In particular, if the unobserved confounding is identical across units, i.e., $\theta^{\star(1)} = \cdots = \theta^{\star(n)}$, then $\mathcal{L}\big(\underline{\Theta}\big)$ in Eq. (3.11) can be decomposed into p_y independent loss functions, one for every $t \in [p_y]$. These decomposed loss functions are identical to the ones used in these prior works.

3.3.3.2 Causal estimate

Given the estimate $\widehat{\underline{\Theta}}$, our estimate of the expected potential outcome $\mu^{(i)}(\widetilde{\boldsymbol{a}}^{(i)})$ under an alternate intervention $\widetilde{\boldsymbol{a}}^{(i)} \in \mathcal{A}^{p_a}$ Eq. (3.6) is derived as follows: First, we identify $\widehat{\Phi}^{(u,y)} \in \mathbb{R}^{p_u \times p_y}$ to be the component of $\widehat{\Theta}$ corresponding to $\mathbf{u}$ and $\mathbf{y}$ for all $\mathbf{u} \in \{\mathbf{v}, \mathbf{a}, \mathbf{y}\}$. Next, we estimate the conditional distribution of $\mathbf{y}$ for unit i as a function of the interventions $\mathbf{a}$, while keeping $\mathbf{v} = \boldsymbol{v}^{(i)}$ and $\mathbf{z} = \boldsymbol{z}^{(i)}$ fixed as

$$\widehat{f}^{(i)}_{\mathsf{y}|\mathsf{a}}(\boldsymbol{y}|\boldsymbol{a}) \propto \exp\Big(\big[\widehat{\theta}^{(i)} + 2\boldsymbol{v}^{(i)\top}\widehat{\Phi}^{(v,y)} + 2\boldsymbol{a}^\top\widehat{\Phi}^{(a,y)}\big]\boldsymbol{y} + \boldsymbol{y}^\top\widehat{\Phi}^{(y,y)}\boldsymbol{y}\Big). \tag{3.13}$$

Finally, we estimate $\mu^{(i)}(\widetilde{\boldsymbol{a}}^{(i)})$ as the mean under the above conditional distribution, given by

$$\widehat{\mu}^{(i)}(\widetilde{\boldsymbol{a}}^{(i)}) \triangleq \mathbb{E}_{\widehat{f}^{(i)}_{\mathsf{y}|\mathsf{a}}}[\mathbf{y}|\mathbf{a} = \widetilde{\boldsymbol{a}}^{(i)}], \tag{3.14}$$

which can be computed by standard algorithms for estimating marginals of graphical models, e.g., via the junction tree algorithm (Wainwright et al., 2008) or message-passing algorithms.

In general, estimating the marginals exactly is computationally hard for undirected graphical models. While the junction tree algorithm works well for graphical models with small treewidth (Wainwright et al., 2008, Section. 2.5), e.g., for trees or chains as in hidden Markov models or state-space models, message-passing algorithms are the default choice for computing approximate marginals for complex graphs, especially with cycles. However, message-passing algorithms may induce additional approximations, which we do not discuss here. For the linear panel data models, estimating the expected potential outcome $\mu^{(i)}(\widetilde{\boldsymbol{a}}^{(i)})$ is equivalent to parameter estimation as we will see in Section 3.6.

3.4 Analysis and Main Results

In this section, we analyze our estimates. First, we provide our guarantee on estimating the unit-level and the population-level parameters in Section 3.4.1. Next, we provide our guarantee on estimating the causal estimand of interest in Section 3.4.2. Before stating our main results, we define a standard notion of complexity of the sets Λ_θ and Λ_Θ, namely metric entropy (defined below) that our guarantees rely on.

Definition 3.2 (ε-covering number and metric entropy)**.** *Given a set $\mathcal{S} \subset \mathbb{R}^{p_1 \times p_2}$ and a scalar $\varepsilon > 0$, we use $\mathcal{C}(\mathcal{S}, \varepsilon)$ to denote the ε-covering number of $\mathcal{S}$ with respect to $\|\cdot\|_{1,1}$, i.e., $\mathcal{C}(\mathcal{S}, \varepsilon)$ denotes the minimum cardinality over all possible subsets $\mathcal{T} \subset \mathcal{S}$ that satisfy $\mathcal{S} \subset \cup_{\mathbf{T} \in \mathcal{T}} \mathcal{B}(\mathbf{T}; \varepsilon)$, where $\mathcal{B}(\mathbf{T}; \varepsilon) \triangleq \{\mathbf{S} \in \mathbb{R}^{p_1 \times p_2} : \|\mathbf{T} - \mathbf{S}\|_{1,1} \leq \varepsilon\}$. We let $\mathcal{M}_\Theta(\varepsilon) \triangleq \log \mathcal{C}(\Lambda_\Theta, \varepsilon)$ denote the metric entropy of $\Lambda_\Theta \subset \mathbb{R}^{p_y \times \widetilde{p}}$, $\mathcal{M}_\theta(\varepsilon) \triangleq \log \mathcal{C}(\Lambda_\theta, \varepsilon)$ denote the metric entropy of $\Lambda_\theta \subset \mathbb{R}^{p_y \times 1}$, and $\mathcal{M}_{\theta,n}(\varepsilon) \triangleq n\mathcal{M}_\theta(n\varepsilon)$ denote a scaled version of the latter.*

Next, we state two settings with upper bounds on the metric entropy of Λ_θ, and we use them as running examples to unpack our general results throughout this paper.

Example 3.1 (Linear combination)**.** *Consider a set Λ_θ containing vectors with bounded entries that are also a linear combination of k known vectors in $\mathbb{R}^{p_y}$ collected as $\mathbf{D} \in \mathbb{R}^{p_y \times k}$, i.e., $\Lambda_\theta = \{\mathbf{D}\mathbf{c} : \mathbf{c} \in \mathbb{R}^k, \|\mathbf{D}\mathbf{c}\|_\infty \leq \alpha\}$. Then, Dagan et al. (2021, Lemma. 11) implies that $\mathcal{M}_\theta(\eta) = O\big(k \log\big(1 + \frac{\alpha}{\eta}\big)\big)$. Further, $\mathcal{M}_{\theta,n}(\eta) = O\big(\frac{\alpha k}{\eta}\big)$.*

Example 3.2 (Sparse linear combination)**.** *Consider a set Λ_θ containing vectors with bounded entries that are also a* s-sparse *linear combination of k known vectors in $\mathbb{R}^{p_y}$ collected as $\mathbf{D} \in \mathbb{R}^{p_y \times k}$, i.e., $\Lambda_\theta = \{\mathbf{D}\mathbf{c} : \mathbf{c} \in \mathbb{R}^k, \big\|a\big\|_0 \leq s, \|\mathbf{D}\mathbf{c}\|_\infty \leq \alpha\}$. Then Dagan et al. (2021, Corollary. 4) implies that $\mathcal{M}_\theta(\eta) = O\big(s \log k \log\big(1 + \frac{\alpha}{\eta}\big)\big)$. Further, $\mathcal{M}_{\theta,n}(\eta) = O\big(\frac{\alpha s \log k}{\eta}\big)$.*

3.4.1 Guarantee on quality of parameter estimate

Our non-asymptotic guarantees use an assumption of a lower bound on the smallest eigenvalue of a suitable set of autocorrelation matrices.

Assumption 3.2. *For any $\boldsymbol{z} \in \mathcal{Z}^{p_z}$, $\boldsymbol{v} \in \mathcal{V}^{p_v}$, $\boldsymbol{a} \in \mathcal{A}^{p_a}$, and $t \in [p_y]$, let $\lambda_{\min}(\boldsymbol{z}, \boldsymbol{v}, \boldsymbol{a}, t)$ denote the smallest eigenvalue of the matrix $\mathbb{E}_{\mathsf{y}|\mathsf{z},\mathsf{v},\mathsf{a}}\big[\widetilde{\mathsf{x}}\ \widetilde{\mathsf{x}}^\top | \mathsf{z} = \boldsymbol{z}, \mathsf{v} = \boldsymbol{v}, \mathsf{a} = \boldsymbol{a}\big]$ where $\widetilde{\mathsf{x}} \triangleq \big(\mathsf{x}_t, 2\mathsf{x}_{-t}\mathsf{x}_t, \mathsf{x}_t^2 - x_{\max}^2/3\big) \in \mathbb{R}^{\widetilde{p}+1}$ with $\mathsf{x} = (\mathsf{y}, \mathsf{a}, \mathsf{v})$. We assume $\lambda_{\min} \triangleq \min_{\boldsymbol{z} \in \mathcal{Z}^{p_z}, \boldsymbol{v} \in \mathcal{V}^{p_v}, \boldsymbol{a} \in \mathcal{A}^{p_a}, t \in [p_y]} \lambda_{\min}(\boldsymbol{z}, \boldsymbol{v}, \boldsymbol{a}, t)$ is strictly positive.*

We note that all eigenvalues of any autocorrelation matrix are non-negative implying $\lambda_{\min}(\boldsymbol{z}, \boldsymbol{v}, \boldsymbol{a}, t) \geq 0$ for all $\boldsymbol{z} \in \mathcal{Z}^{p_z}$, $\boldsymbol{v} \in \mathcal{V}^{p_v}$, $\boldsymbol{a} \in \mathcal{A}^{p_a}$, and $t \in [p_y]$. Assumption 3.2 requires $\lambda_{\min}(\boldsymbol{z}, \boldsymbol{v}, \boldsymbol{a}, t) > 0$ for all $\boldsymbol{z} \in \mathcal{Z}^{p_z}$, $\boldsymbol{v} \in \mathcal{V}^{p_v}$, $\boldsymbol{a} \in \mathcal{A}^{p_a}$, and $t \in [p_y]$, and serves as a sufficient condition to rule out certain singular distributions (Shah et al., 2021b, Section. 5). Essentially, we use this assumption to lower bound the variance of a non-constant random variable (Section 3.B.1).

We are now ready to state our main result that characterizes a high probability bound on the estimation error for the estimate $\widehat{\Theta}$ computed via Eq. (3.12). To simplify the presentation, we use c and c' to denote universal constants or constants that depend on the parameters $\alpha, x_{\max}$, and $\lambda_{\min}$ and can take a different value in each appearance.

Theorem 3.1 (Guarantee on quality of parameter estimate)**.** *Suppose Assumptions 3.1 and 3.2 hold. Fix an $\varepsilon > 0$ and $\delta \in (0,1)$, and define*

$$R(\varepsilon,\delta) \triangleq \max\{ce^{c'\beta}\sqrt{\log(\log p_y/\delta) + \mathcal{M}_{\theta}(ce^{-c'\beta})}, \varepsilon\gamma\} \ \textit{with} \ \gamma \triangleq \max_{\theta,\overline{\theta}\in\Lambda_\theta} \frac{\|\theta-\overline{\theta}\|_1}{\|\theta-\overline{\theta}\|_2} \quad (3.15)$$

and

$$\widetilde{\mathcal{M}}_{\theta,n}(\varepsilon,\delta) \triangleq \mathcal{M}_{\theta,n}\Big(\frac{\varepsilon^2}{\widetilde{p}}\Big) + \mathcal{M}_{\theta}\big(R^2(\varepsilon,\delta)\big). \quad (3.16)$$

Then, with probability at least $1-\delta$, the estimates $\widehat{\Theta}, \widehat{\theta}^{(1)}, \cdots, \widehat{\theta}^{(n)}$ defined in Eq. (3.12) satisfy

$$\|\widehat{\Theta} - \Theta^{\star}\|_{2,\infty} \leq \varepsilon \qquad \textit{when} \quad n \geq \frac{ce^{c'\beta}p_y^2\Big(\log\frac{p_y}{\delta} + \mathcal{M}_{\Theta}(\varepsilon^2) + \mathcal{M}_{\theta,n}(\varepsilon^2)\Big)}{\varepsilon^4}, \quad (3.17)$$

and

$$\max_{i\in[n]} \|\widehat{\theta}^{(i)} - \theta^{\star(i)}\|_2 \leq R\Big(\varepsilon, \frac{\delta}{n}\Big) \quad \textit{when} \quad n \geq \frac{ce^{c'\beta}p_y^2\widetilde{p}^2\Big(\log\frac{np_y}{\delta} + \mathcal{M}_{\Theta}(\frac{\varepsilon^2}{\widetilde{p}}) + \widetilde{\mathcal{M}}_{\theta,n}(\varepsilon, \frac{\delta}{n})\Big)}{\varepsilon^4}. \quad (3.18)$$

We split the proof into two parts: First, we establish the bound Eq. (3.17) in Section 3.B, which we then use to establish the bound Eq. (3.18) in Section 3.C.

We note that $\mathcal{M}_{\Theta}(\varepsilon^2) = O\big(\beta^2 p_y \log\frac{1}{\varepsilon^2}\big)$. Therefore, our guarantee in Eq. (3.17) provides a non-asymptotic error bound of order

$$\frac{p_y^2(p_y \log\sqrt{n}p_y + \mathcal{M}_{\theta,n}(n^{-1/2}))}{n^{1/4}},$$

(where we treat β as a constant) for estimating $\Theta^{\star}$ although the n samples have different unit-level parameters $\{\theta^{\star(i)}\}_{i=1}^{n}$. On the other hand, after squaring both sides and dividing by p_y, the guarantee Eq. (3.18) for the unit-level parameters can be simplified as follows:[6] whenever $n \geq c'\varepsilon^{-4}p_y^2\widetilde{p}^2(p_y\log\frac{p_y\widetilde{p}}{\delta\varepsilon^2} + \mathcal{M}_{\theta,n}(\varepsilon^2/\widetilde{p}) + \mathcal{M}_{\theta}(c))$, we have

$$\mathrm{MSE}(\widehat{\theta}^{(i)}, \theta^{\star(i)}) \leq \max\Big\{\varepsilon^2, \frac{\mathcal{M}_{\theta}(c) + \log(\log\frac{p_y}{\delta})}{p_y}\Big\}, \quad (3.19)$$

where we use $\gamma \leq \sqrt{p}$ in Eq. (3.15) and treat β as a constant. For large n so that ε is small, this error scales linearly with the metric entropy $\mathcal{M}_{\theta}$—the error becomes worse as the unit-level parameter set Λ_{θ} becomes more complex.

The next corollary (stated without proof) provides a formal version of the population-level guarantee in Eq. (3.17) and the unit-level guarantee in Eq. (3.19) for the two examples discussed earlier. We treat β as a constant and note that the dependence is exponential as in Theorem 3.1.

[6] We replace δ/n in Eq. (3.18) by δ as we do not require a union bound over $i \in [n]$ for unit-wise guarantees.

Corollary 3.1 (Consequences for examples)**.** *Suppose Assumptions 3.1 and 3.2 hold. Then, for any fixed $\varepsilon > 0$ and $\delta \in (0, 1)$, the following results hold with probability at least $1 - \delta$.*

(a) Linear combination: *If Λ_θ is as in Example 3.1, then for all $i \in [n]$,*

$$\|\widehat{\Theta} - \Theta^\star\|_{2,\infty} \leq \varepsilon \qquad \textit{for} \quad n \geq \frac{cp_y^2\big(p_y \log \frac{p_y}{\delta\varepsilon^2} + \frac{k}{\varepsilon^2}\big)}{\varepsilon^4}$$

$$\mathrm{MSE}(\widehat{\theta}^{(i)}, \theta^{\star(i)}) \leq \max\Big\{\varepsilon^2, \frac{c\big(k + \log(\log \frac{p_y}{\delta})\big)}{p_y}\Big\} \quad \textit{for} \quad n \geq \frac{cp_y^2\widetilde{p}^2\big(p_y \log \frac{p_y\widetilde{p}}{\delta\varepsilon^2} + \frac{\widetilde{p}k}{\varepsilon^2}\big)}{\varepsilon^4}.$$

(b) Sparse linear combination: *If Λ_θ is as in Example 3.2, then for all $i \in [n]$,*

$$\|\widehat{\Theta} - \Theta^\star\|_{2,\infty} \leq \varepsilon \qquad \textit{for} \quad n \geq \frac{cp_y^2\big(p_y \log \frac{p_y}{\delta\varepsilon^2} + \frac{s \log k}{\varepsilon^2}\big)}{\varepsilon^4}$$

$$\mathrm{MSE}(\widehat{\theta}^{(i)}, \theta^{\star(i)}) \leq \max\Big\{\varepsilon^2, \frac{c\big(s \log k + \log(\log \frac{p_y}{\delta})\big)}{p_y}\Big\} \quad \textit{for} \quad n \geq \frac{cp_y^2\widetilde{p}^2\big(p_y \log \frac{p_y\widetilde{p}}{\delta\varepsilon^2} + \frac{s\widetilde{p} \log k}{\varepsilon^2}\big)}{\varepsilon^4}.$$

Corollary 3.1 states that, as long as n is polynomially large in $(p_y, \widetilde{p})$, our strategy learns the unit-level parameters (on average in terms of mean square error across coordinates) for each user if p_y is large compared to either the number of vectors k (Example 3.1) or the sparsity parameter s (Example 3.2).

Sharpness of guarantees and generalization of prior results. The exponential dependence on β in Theorem 3.1 is unavoidable given the lower bounds for learning exponential families even with i.i.d. samples (Santhanam and Wainwright, 2012). Regarding the dependence on error tolerance ε, prior works with suitable analogs of our loss function provide two different error scaling: (i) $1/\varepsilon^4$ in Shah et al. (2021b,c); Vuffray et al. (2022a) and (ii) $1/\varepsilon^2$ in Vuffray et al. (2016a) and Shah et al. (2023). The works in category (ii) use techniques from Negahban et al. (2012), and it remains an interesting future direction to see whether similar ideas could be used to sharpen the error scaling of $1/\varepsilon^4$ to the parametric rate of $1/\varepsilon^2$ in Theorem 3.1. We note that improving the dependence on ε in Eq. (3.17) improves the dependence on ε as well as p_y in Eq. (3.18) (see the proof in Section 3.C for details). In the special case of equal unit-level parameters ($\theta^{\star(1)} = \cdots = \theta^{\star(n)}$), the analysis in Section 3.B to establish the bound Eq. (3.17) can be modified to recover (up to constants) prior guarantee (Shah et al., 2021c, Lemma. 9.1) on learning exponential family from n i.i.d. samples. Further, the guarantee Eq. (3.18) recovers the prior guarantee (Kandiros et al., 2021, Theorem. 6) as a special case where the authors consider learning an Ising model from one sample when the population-level parameter is known up to a scaling factor.

3.4.2 Guarantee on quality of outcome estimate

Our non-asymptotic guarantee on outcome estimate assumes that the following matrices are suitably stable under small perturbation in the parameters: (i) the covariance matrix of $\mathbf{y}$ conditioned on $\mathbf{a}$, $\mathbf{v}$, and $\mathbf{z}$ and (ii) the cross-covariance matrix of $\mathbf{y}$ and $y_t\mathbf{y}$ conditioned on $\mathbf{a}$, $\mathbf{v}$, and $\mathbf{z}$ for all $t \in [p_y]$.

Assumption 3.3. *For any set $\mathbb{B}$ containing θ, Θ, there exists a constant $C(\mathbb{B})$ such that*

$$\sup_{\theta,\Theta\in\mathbb{B}} \max\Big\{\|\mathbb{C}\mathrm{ov}_{\theta,\Theta}(\mathbf{y},\mathbf{y}|\boldsymbol{a},\boldsymbol{v},\boldsymbol{z})\|_{\mathrm{op}}, \max_{t\in[p_y]}\|\mathbb{C}\mathrm{ov}_{\theta,\Theta}(\mathbf{y},y_t\mathbf{y}|\boldsymbol{a},\boldsymbol{z},\boldsymbol{v})\|_{\mathrm{op}}\Big\} \leq C(\mathbb{B}), \tag{3.20}$$

almost surely. The expectation in Eq. (3.20) is with respect to the distribution of $\mathbf{y}$ conditioned on $\mathbf{a} = \boldsymbol{a}$, $\mathbf{v} = \boldsymbol{v}$, and $\mathbf{z} = \boldsymbol{z}$ which is fully parameterized by θ and Θ, and can be obtained from Eq. (3.4) after replacing $\theta(\boldsymbol{z})$ by θ.

In Section 3.D.2, we show that $C(\mathbb{B})$ is a constant for a class of distributions. We note that this assumption is common in the literature on learning Gaussian graphical models to rule out singular distributions (Ma and Michailidis, 2016; Won and Kim, 2006; Zhou et al., 2011).

We are now ready to state our guarantee for the estimate $\widehat{\mu}^{(i)}(\widetilde{\boldsymbol{a}}^{(i)})$ (see Eq. (3.14)) of the expected potential outcomes for any unit $i \in [n]$ under an alternate intervention $\widetilde{\boldsymbol{a}}^{(i)} \in \mathcal{A}^{p_a}$. We assume $p_v = p_a = p_y = p$ for brevity. See the proof in Section 3.D where we also state a more general result.

Theorem 3.2 (Guarantee on quality of outcome estimate)**.** *Suppose Assumptions 3.1 to 3.3 hold. Then for any fixed $\varepsilon > 0$ and $\delta \in (0,1)$, the estimates $\{\widehat{\mu}^{(i)}(\widetilde{\boldsymbol{a}}^{(i)})\}_{i=1}^n$ defined in Eq. (3.14) for any $\{\widetilde{\boldsymbol{a}}^{(i)} \in \mathcal{A}^{p_a}\}_{i=1}^n$ satisfy*

$$\max_{i\in[n]}\frac{\|\mu^{(i)}(\widetilde{\boldsymbol{a}}^{(i)})-\widehat{\mu}^{(i)}(\widetilde{\boldsymbol{a}}^{(i)})\|_2}{C(\mathbb{B}_i)} \leq R\Big(\varepsilon,\frac{\delta}{n}\Big)+p\varepsilon \;\; \textit{for} \;\; n \geq \frac{ce^{c'\beta}p^4\Big(\log\frac{np}{\delta}+\mathcal{M}_\Theta(\frac{\varepsilon^2}{p})+\widetilde{\mathcal{M}}_{\theta,n}(\varepsilon,\frac{\delta}{n})\Big)}{\varepsilon^4}, \tag{3.21}$$

with probability at least $1-\delta$, where $R(\varepsilon,\delta)$ was defined in Eq. (3.15), $\widetilde{\mathcal{M}}_{\theta,n}(\varepsilon,\delta)$ was defined in Eq. (3.16), $C(\mathbb{B})$ was defined in Eq. (3.20), and

$$\mathbb{B}_i \triangleq \Big\{\theta \in \Lambda_\theta : \|\theta-\theta^{\star(i)}\|_2 \leq R\Big(\varepsilon,\frac{\delta}{n}\Big)\Big\} \times \Big\{\Theta \in \Lambda_\Theta : \max_{t\in[p_y]}\|\Theta_t-\Theta_t^\star\|_2 \leq \varepsilon\Big\}.$$

Repeating the algebra as in Eq. (3.19) and treating $C(\mathbb{B}_i)$ as a constant, the bound Eq. (3.21) yields the following simplified bound for the MSE of our mean outcome estimate $\mu^{(i)}(\widetilde{\boldsymbol{a}}^{(i)})$ for unit $i \in [n]$ under treatment $\widetilde{\boldsymbol{a}}^{(i)} \in \mathcal{A}^{p_a}$: whenever $n \geq c'\varepsilon^{-4}p^4(p\log\frac{p^2}{\delta\varepsilon^2} + \mathcal{M}_{\theta,n}(\varepsilon^2/p) + \mathcal{M}_\theta(c))$, we have

$$\mathrm{MSE}(\mu^{(i)}(\widetilde{\boldsymbol{a}}^{(i)}),\widehat{\mu}^{(i)}(\widetilde{\boldsymbol{a}}^{(i)})) \leq \varepsilon^2 + \frac{\mathcal{M}_\theta(c)+\log(\log\frac{p}{\delta})}{p}.$$

This bound is of the same order as in Eq. (3.19) and can be formalized for the two examples (Examples 3.1 and 3.2) by deriving a suitable analog of Corollary 3.1. In a nutshell, in both settings, the unit-level expected potential outcomes can be estimated well when the total number of units n is large and the observations for each unit are high dimensional compared to the number of vectors k in Example 3.1 or the sparsity parameter s in Example 3.2. We omit a formal statement for brevity.

Finally, we also note that as in Theorem 3.1, the exponential dependence on β is expected to be unavoidable due to the principle of conjugate duality (Wainwright et al., 2008), i.e., the existence of a unique mapping from the parameters to the means and vice versa for the exponential family. Moreover, as in the discussion after Corollary 3.1, the sharpness of the rate of $1/\varepsilon^4$ is left for future work. Improving the dependency on ε in Eq. (3.21) would also improve the dependency on p.

3.5 Possible Extensions

We now discuss how to extend our theoretical results with various relaxations of the exponential family modeling.

3.5.1 Higher order terms in the conditional exponential family

In Section 3.3.2, we described how our framework and results apply when the conditional distribution $f_{\mathbf{y}|\mathbf{a},\mathbf{v},\mathbf{z}}$ is modeled as the exponential family distribution in Eq. (3.2) where the term inside the exponent is linear in $(\mathbf{z}, \mathbf{v}, \mathbf{a})$ and quadratic in $\mathbf{y}$. We now describe how our framework and results are applicable when the conditional distribution $f_{\mathbf{y}|\mathbf{a},\mathbf{v},\mathbf{z}}$ is modeled as the following exponential family distribution

$$f_{\mathbf{y}|\mathbf{a},\mathbf{v},\mathbf{z}}(\boldsymbol{y}|\boldsymbol{a}, \boldsymbol{z}, \boldsymbol{v}) \propto \exp\big(q_{\Phi}(\boldsymbol{v}, \boldsymbol{a}, \boldsymbol{y})\big)\exp\big(2\boldsymbol{z}^{\top}\Phi^{(z,y)}\boldsymbol{y}\big), \tag{3.22}$$

where $q_{\Phi}(\boldsymbol{v}, \boldsymbol{a}, \boldsymbol{y})$ is some bounded degree polynomial in $(\boldsymbol{v}, \boldsymbol{a}, \boldsymbol{y})$ parameterized by Φ, i.e., the term inside the exponent is linear in $\mathbf{z}$ and arbitrary bounded degree polynomial in $(\mathbf{v}, \mathbf{a}, \mathbf{y})$. We note that every term in $q_{\Phi}(\boldsymbol{v}, \boldsymbol{a}, \boldsymbol{y})$ needs to depend on $\mathbf{y}$ for it to contribute to $f_{\mathbf{y}|\mathbf{a},\mathbf{v},\mathbf{z}}$ in Eq. (3.22). For convenience, hereon, we ignore any dependence on $\mathbf{v}$, and abuse notation to let $q_{\Phi}(\boldsymbol{a}, \boldsymbol{y}) = q_{\Phi}(\boldsymbol{v}, \boldsymbol{a}, \boldsymbol{y})$. Then, in Eq. (3.2), $q_{\Phi}(\boldsymbol{a}, \boldsymbol{y})$ was a polynomial of degree 2 , i.e.,

$$q_{\Phi}(\boldsymbol{a}, \boldsymbol{y}) = q_{\Phi}^{(2)}(\boldsymbol{a}, \boldsymbol{y}) \triangleq \mathtt{Sum}\Big(\phi^{(y)} \odot \boldsymbol{y} + 2\Phi^{(a,y)} \odot \big(\boldsymbol{a} \otimes \boldsymbol{y}\big) + \Phi^{(y,y)} \odot \big(\boldsymbol{y} \otimes \boldsymbol{y}\big)\Big),$$

where $\odot$ denotes the Hadamard product, $\otimes$ denotes the Kronecker product, $\Phi = (\phi^{(y)}, \Phi^{(a,y)}, \Phi^{(y,y)})$ with $\Phi^{(y,y)}$ being symmetric, and $\mathtt{Sum}(s_1 + \cdots + s_h) \in \mathbb{R}$ sums, over all $i \in [h]$, all the entries of s_i which could be a real number/vector/matrix/tensor. To explain how the loss function in Eq. (3.11) needs to be modified for general $q_{\Phi}(\boldsymbol{a}, \boldsymbol{y})$, we consider a polynomial of degree 3:

$$q_{\Phi}(\boldsymbol{a}, \boldsymbol{y}) = q_{\Phi}^{(2)}(\boldsymbol{a}, \boldsymbol{y}) + \mathtt{Sum}\Big(\sum_{(u_1,u_2)\in\{(a,a),(a,y),(y,y)\}} c_{u_1,u_2} \cdot \Phi^{(u_1,u_2,y)} \odot \big(\boldsymbol{u}_1 \otimes \boldsymbol{u}_2 \otimes \boldsymbol{y}\big)\Big),$$

where $c_{a,a} = c_{a,y} = 3$, $c_{y,y} = 1$ are constants chosen for consistency, and $\Phi^{(u_1,u_2,y)} \in \mathbb{R}^{p_{u_1} \times p_{u_2} \times p_y}$ is symmetric with respect to indices that are repeated for every $(u_1, u_2) \in \{(a, a), (a, y), (y, y)\}$. We illustrate the two steps from Section 3.3.3.1 below.

Centering sufficient statistics of the conditional distribution of a variable The conditional distribution $f_{y_t|\mathbf{y}_{-t},\mathbf{a},\mathbf{z}}$ of the random variable y_t conditioned on $\mathbf{y}_{-t} = \boldsymbol{y}_{-t}$,

$\mathbf{a} = \boldsymbol{a}$, and $\mathbf{z} = \boldsymbol{z}$ for every $t \in [p_y]$ is given by

$$\begin{aligned}
&f_{\mathsf{y}_t|\mathbf{y}_{-t},\mathbf{a},\mathbf{z}}\big(y_t|\boldsymbol{y}_{-t},\boldsymbol{a},\boldsymbol{z}\big) \propto \\
&\quad\exp\bigg(\texttt{Sum}\bigg(\Big[\phi_t(\boldsymbol{z}) + \sum_{u\in\{y_{-t},a\}} 2\Phi^{(u,y_t)} \odot \boldsymbol{u} + \sum_{(u_1,u_2)\in\{(a,a),(a,y_{-t}),(y_{-t},y_{-t})\}} c_{u_1,u_2}\Phi^{(u_1,u_2,y_t)} \odot \big(\boldsymbol{u}_1 \otimes \boldsymbol{u}_2\big)\Big]y_t \\
&\qquad + \Big[\Phi^{(y_t,y_t)} + \sum_{u\in\{y_{-t},a\}} 3\Phi^{(u,y_t,y_t)} \odot \boldsymbol{u}\Big]\Big(y_t^2 - \frac{x_{\max}^2}{3}\Big) + \Phi^{(y_t,y_t,y_t)}y_t^3\bigg)\bigg),
\end{aligned} \tag{3.23}$$

where $\phi_t(\boldsymbol{z}) \triangleq \phi^{(y_t)} + 2\Phi^{(z,y_t)} \odot \boldsymbol{z}$, $c_{y_{-t},y_{-t}} = 3$, and $c_{a,y_{-t}} = 6$. Let Φ_t denote the concatenation of all the remaining parameters in Eq. (3.23). As in Eq. (3.10), the term $x_{\max}^2/3$ inside the exponent is vacuous and centers the sufficient statistics y_t^2. The other sufficient statistics, i.e., y_t and y_t^3, are naturally centered as their integrals with respect to the uniform distribution on $\mathcal{X}$ are both zeros.

Constructing the loss function Now, it is easy to see that the corresponding loss $\mathcal{L}$ is given by

$$\begin{aligned}
&\mathcal{L} = \frac{1}{n}\sum_{t\in[p_y]}\sum_{i\in[n]} \\
&\quad\exp\bigg(-\texttt{Sum}\bigg(\Big[\phi_t^{(i)} + \sum_{u\in\{y_{-t},a\}} 2\Phi^{(u,y_t)} \odot \boldsymbol{u}^{(i)} + \sum_{(u_1,u_2)\in\{(a,a),(a,y_{-t}),(y_{-t},y_{-t})\}} c_{u_1,u_2}\Phi^{(u_1,u_2,y_t)} \odot \big(\boldsymbol{u}_1^{(i)} \otimes \boldsymbol{u}_2^{(i)}\big)\Big]y_t^{(i)} \\
&\qquad + \Big[\Phi^{(y_t,y_t)} + \sum_{u\in\{y_{-t},a\}} 3\Phi^{(u,y_t,y_t)} \odot \boldsymbol{u}^{(i)}\Big]\Big(\big[y_t^{(i)}\big]^2 - \frac{x_{\max}^2}{3}\Big) + \Phi^{(y_t,y_t,y_t)}\big[y_t^{(i)}\big]^3\bigg)\bigg),
\end{aligned}$$

and minimizing this convex loss results in the estimates of $\{\phi_t^{(i)}\}_{i\in[n]}$ and $\{\Phi_t\}_{t\in p_y}$. Consequently, the guarantees in Section 3.4 continue to hold as long as Assumptions 3.1 to 3.3 are appropriately generalized.

Tilting the base distribution. We note that the exponential family in Eq. (3.2) can be rewritten as

$$\begin{aligned}
f_{\mathbf{y}|\mathbf{a},\mathbf{v},\mathbf{z}}(\boldsymbol{y}|\boldsymbol{a},\boldsymbol{z},\boldsymbol{v}) \propto \exp\big(2\boldsymbol{z}^\top\Phi^{(z,y)}\boldsymbol{y}\big)\exp\big(2\boldsymbol{v}^\top\Phi^{(v,y)}\boldsymbol{y}\big)\exp\big(2\boldsymbol{a}^\top\Phi^{(a,y)}\boldsymbol{y}\big) \\
\exp\big({\phi^{(y)}}^\top\boldsymbol{y} + \boldsymbol{y}^\top\Phi^{(y,y)}\boldsymbol{y}\big),
\end{aligned}$$

where $\exp\big({\phi^{(y)}}^\top\boldsymbol{y} + \boldsymbol{y}^\top\Phi^{(y,y)}\boldsymbol{y}\big)$ stands for a base distribution on $\mathbf{y}$ which is exponentially tilted by $\mathbf{z}$, $\mathbf{v}$, and $\mathbf{a}$, i.e., by $\exp\big(2\boldsymbol{z}^\top\Phi^{(z,y)}\boldsymbol{y}\big)$, $\exp\big(2\boldsymbol{v}^\top\Phi^{(v,y)}\boldsymbol{y}\big)$, and $\exp\big(2\boldsymbol{a}^\top\Phi^{(a,y)}\boldsymbol{y}\big)$, respectively. Then, generalizing the exponential family in Eq. (3.2) to the one in Eq. (3.22) is equivalent to saying that our approach and results continue to apply when (a) the base distribution on $\mathbf{y}$ is an exponential family distribution where the term inside the exponent is arbitrary bounded degree polynomial (instead of quadratic) and (b) the exponent of the exponential tilting of this base distribution by $(\mathbf{v}, \mathbf{a})$ is arbitrary bounded degree polynomial (instead of linear).

3.5.2 Discrete and mixed variables

In Section 3.3.2, we described how our framework and results are applicable when the support of $\mathbf{v}$, $\mathbf{a}$, and $\mathbf{y}$ are bounded continuous sets, i.e., $\mathcal{V} = \mathcal{A} = \mathcal{Y} = [-x_{\max}, x_{\max}]$. Since we only model the conditional distribution $f_{\mathbf{y}|\mathbf{a},\mathbf{v},\mathbf{z}}$ as an exponential family distribution, we only need boundedness as a restriction on the support of $\mathbf{v}$ and $\mathbf{a}$. Now, we describe how to adapt our loss function when $\mathbf{y} = (y_1, \cdots, y_{p_y}) \in \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_{p_y}$ where $\mathcal{Y}_t$ is either a discrete compact set or a continuous compact set for $t \in [p_y]$.

We note that the conditional distribution $f_{y_t|\mathbf{y}_{-t},\mathbf{v},\mathbf{a},\mathbf{z}}$ of the random variable y_t conditioned on $\mathbf{y}_{-t} = \boldsymbol{y}_{-t}$, $\mathbf{v} = \boldsymbol{v}$, $\mathbf{a} = \boldsymbol{a}$, and $\mathbf{z} = \boldsymbol{z}$ for every $t \in [p_y]$ is still consistent with the conditional distribution $f_{x_t|\mathbf{x}_{-t},\mathbf{z}}$ in Eq. (3.9). However, the constants used to center the sufficient statistics in Eq. (3.10) may change. More precisely, for any $t \in [p_y]$, the sufficient statistics y_t and y_t^2 are centered by subtracting $\mathbb{E}_{\mathcal{U}_t}[y_t]$ and $\mathbb{E}_{\mathcal{U}_t}[y_t^2]$, respectively where $\mathcal{U}_t$ denotes the uniform distribution supported over $\mathcal{Y}_t$. Consequently, the loss function in Eq. (3.11) as well as Assumption 3.2 can be adapted, and the guarantees in Section 3.4 continue to hold.

3.6 Connections to Panel Data Models

We now describe how the framework of exponential family can be applied to extend some of the common panel data models from econometrics. For simplicity, we let $p_y = p_a = p_v = p$. Consider the following generic model: for all $t \in [p]$ and $i \in [n]$

$$y_t^{(i)} = g(\boldsymbol{z}_{1:t}^{(i)}, \boldsymbol{v}_{1:t}^{(i)}, \boldsymbol{a}_{1:t}^{(i)}, \boldsymbol{y}_{1:t-1}^{(i)}, \eta_t^{(i)}), \tag{3.24}$$

where $\boldsymbol{u}_{1:t}^{(i)} \triangleq (u_1, \cdots, u_t)$ for all $\boldsymbol{u} \in \{\boldsymbol{z}, \boldsymbol{v}, \boldsymbol{a}, \boldsymbol{y}\}$, $\{\eta_t^{(i)}\}_{i\in[n],t\in[p]}$ is the idiosyncratic error, and g is a link function. The most commonly used panel data models are special cases of the model in Eq. (3.24) as illustrated below.

Example 3.3 (The linear unit fixed effects model)**.** *The linear unit fixed effects model is a non-dynamic model with a time-constant unobserved component and a linear link function as below: for all $t \in [p]$ and $i \in [n]$*

$$y_t^{(i)} = \beta^{(a)} a_t^{(i)} + \beta^{(v)} v_t^{(i)} + z^{(i)} + \eta_t^{(i)}.$$

Here, the unobserved covariate is constant across times but introduces heterogeneity across units, i.e., $z_t^{(i)} = z^{(i)}$, where $z^{(1)}, \cdots, z^{(n)}$ are known as unit effects.

Example 3.4 (The linear time fixed effects model)**.** *The linear time fixed effects model is a non-dynamic model with a unit-constant unobserved component and a linear link function as below: for all $t \in [p]$ and $i \in [n]$*

$$y_t^{(i)} = \beta^{(a)} a_t^{(i)} + \beta^{(v)} v_t^{(i)} + z_t + \eta_t^{(i)}.$$

Here, the unobserved covariate is constant across units but introduces heterogeneity across times, i.e., $z_t^{(i)} = z_t$, where $z_1, \cdots, z_p$ are known as time effects.

Example 3.5 (The non-linear unit fixed effects model)**.** *The non-linear unit fixed effects model is a non-dynamic model with a time-constant unobserved component and a non-linear link function as below: for all $t \in [p]$ and $i \in [n]$*

$$y_t^{(i)} = \mathbb{1}\Big(\beta^{(a)} a_t^{(i)} + \beta^{(v)} v_t^{(i)} + z^{(i)} + \eta_t^{(i)}\Big).$$

Here, as in Example 3.3, the unobserved covariate is constant across times but introduces heterogeneity across units, i.e., $z_t^{(i)} = z^{(i)}$, where $z^{(1)}, \cdots, z^{(n)}$ are known as unit effects.

Example 3.6 (The non-linear time fixed effects model)**.** *The non-linear time fixed effects model is a non-dynamic model with a unit-constant unobserved component and a non-linear link function as below: for all $t \in [p]$ and $i \in [n]$*

$$y_t^{(i)} = \mathbb{1}\Big(\beta^{(a)} a_t^{(i)} + \beta^{(v)} v_t^{(i)} + z_t + \eta_t^{(i)}\Big).$$

Here, as in Example 3.4, the unobserved covariate is constant across units but introduces heterogeneity across times, i.e., $z_t^{(i)} = z_t$, where $z_1, \cdots, z_p$ are known as time effects.

In Examples 3.3 to 3.6, the outcome at time t depends only on the unobserved covariate, the observed covariate, and the intervention at time t. Further, the slopes corresponding to observed covariates and interventions, $\beta^{(v)}$ and $\beta^{(a)}$, are constant across units and times. Therefore, for the linear models in Examples 3.3 and 3.4, the causal effect is equal to $\beta^{(a)}$ for all units and all times. Likewise, for the non-linear models in Examples 3.5 and 3.6, the heterogeneity in the causal effect (across units or times) is driven only by the unobserved covariate.

In this section, we show how to incorporate dynamics in Examples 3.3 to 3.6, i.e., we allow the current outcome to also depend on the previous observed covariates, interventions, and outcomes. Further, for the unit fixed effects models in Examples 3.3 and 3.5, we show how to incorporate a slope varying across units and times in an additive manner (i.e., $\beta^{(a)} = \beta_i^{(a)} + \beta_t^{(a)}$ and $\beta^{(v)} = \beta_i^{(v)} + \beta_t^{(v)}$ for unit i and time t), and for the time fixed effects models in Examples 3.3 and 3.5, we show how to incorporate a slope varying across times. Towards that, we consider the following extensions of Examples 3.3 to 3.6.

Example 3.7 (The dynamic linear unit fixed effects model with additive slopes)**.** *The dynamic linear unit fixed effects model with additive slopes is a model with a time-constant unobserved component and a linear link function as below: for all $t \in [p]$ and $i \in [n]$*

$$y_t^{(i)} = \beta_i^{(a)} a_t^{(i)} + \sum_{j=t-d}^{t} \beta_{t,j}^{(a)} a_j^{(i)} + \beta_i^{(v)} v_t^{(i)} + \sum_{j=t-d}^{t} \beta_{t,j}^{(v)} v_j^{(i)} + \sum_{j=t-d}^{t-1} \beta_{t,j}^{(y)} y_j^{(i)} + z^{(i)} + \eta_t^{(i)}. \tag{3.25}$$

This model recovers the model in Example 3.3 when $d = 0$, $\beta_i^{(a)} + \beta_{t,t}^{(a)} = \beta^{(a)}$, and $\beta_i^{(v)} + \beta_{t,t}^{(v)} = \beta^{(v)}$.

Example 3.8 (The dynamic linear time fixed effects model)**.** *The dynamic linear time fixed effects model is a model with a unit-constant unobserved component and a linear link function as below: for all $t \in [p]$ and $i \in [n]$*

$$y_t^{(i)} = \sum_{j=t-d}^{t} \beta_{t,j}^{(a)} a_j^{(i)} + \sum_{j=t-d}^{t} \beta_{t,j}^{(v)} v_j^{(i)} + \sum_{j=t-d}^{t-1} \beta_{t,j}^{(y)} y_j^{(i)} + z_t + \eta_t^{(i)}.$$

This model recovers the model in Example 3.4 when $d = 0$, $\beta_{t,t}^{(a)} = \beta^{(a)}$, and $\beta_{t,t}^{(v)} = \beta^{(v)}$.

Example 3.9 (The dynamic non-linear unit fixed effects model with additive slopes)**.** *The dynamic non-linear unit fixed effects model with additive slopes is a model with a time-constant unobserved component and a non-linear link function as below: for all $t \in [p]$ and $i \in [n]$*

$$y_t^{(i)} = \mathbb{1}\Big(\beta_i^{(a)} a_t^{(i)} + \sum_{j=t-d}^{t} \beta_{t,j}^{(a)} a_j^{(i)} + \beta_i^{(v)} v_t^{(i)} + \sum_{j=t-d}^{t} \beta_{t,j}^{(v)} v_j^{(i)} + \sum_{j=t-d}^{t-1} \beta_{t,j}^{(y)} y_j^{(i)} + z^{(i)} + \eta_t^{(i)}\Big).$$

This model recovers the model in Example 3.5 when $d = 0$, $\beta_i^{(a)} + \beta_{t,t}^{(a)} = \beta^{(a)}$, and $\beta_i^{(v)} + \beta_{t,t}^{(v)} = \beta^{(v)}$.

Example 3.10 (The dynamic non-linear time fixed effects model)**.** *The dynamic non-linear time fixed effects model is a model with a unit-constant unobserved component and a non-linear link function as below: for all $t \in [p]$ and $i \in [n]$*

$$y_t^{(i)} = \mathbb{1}\Big(\sum_{j=t-d}^{t} \beta_{t,j}^{(a)} a_j^{(i)} + \sum_{j=t-d}^{t} \beta_{t,j}^{(v)} v_j^{(i)} + \sum_{j=t-d}^{t-1} \beta_{t,j}^{(y)} y_j^{(i)} + z_t + \eta_t^{(i)}\Big).$$

This model recovers the model in Example 3.6 when $d = 0$, $\beta_i^{(a)} + \beta_{t,t}^{(a)} = \beta^{(a)}$, and $\beta_i^{(v)} + \beta_{t,t}^{(v)} = \beta^{(v)}$.

To represent the slopes varying across times in Examples 3.7 to 3.10, we define the following upper-triangular matrices

$$\mathbf{B}^{(v)} \in \mathbb{R}^{p \times p} \quad \text{such that} \quad \mathbf{B}^{(v)}_{t_1,t_2} = \begin{cases} 0 \quad \text{if} \quad t_2 < t_1 \quad \text{or} \quad t_2 - t_1 > d \\ \beta^{(v)}_{t_2,t_1} \quad \text{otherwise} \end{cases} \tag{3.26}$$

$$\mathbf{B}^{(a)} \in \mathbb{R}^{p \times p} \quad \text{such that} \quad \mathbf{B}^{(a)}_{t_1,t_2} = \begin{cases} 0 \quad \text{if} \quad t_2 < t_1 \quad \text{or} \quad t_2 - t_1 > d \\ \beta^{(a)}_{t_2,t_1} \quad \text{otherwise} \end{cases} \tag{3.27}$$

$$\mathbf{B}^{(y)} \in \mathbb{R}^{p \times p} \quad \text{such that} \quad \mathbf{B}^{(y)}_{t_1,t_2} = \begin{cases} 0 \quad \text{if} \quad t_2 < t_1 \quad \text{or} \quad t_2 - t_1 > d \\ -1 \quad \text{if} \quad t_2 = t_1 \\ \beta^{(y)}_{t_2,t_1} \quad \text{otherwise} \end{cases} . \tag{3.28}$$

To be able to recover the unknown time varying slope matrices $\mathbf{B}^{(v)}$, $\mathbf{B}^{(a)}$, and $\mathbf{B}^{(y)}$, as well as the unknown unit varying slopes $\{\beta_i^{(v)}, \beta_i^{(a)}\}_{i \in [n]}$, we make the following assumptions. First, we assume that these slopes and the unobserved covariates are bounded.

Assumption 3.4 (Bounded slopes and unobserved covariates)**.**

(a) The unobserved covariates in Examples 3.7 and 3.9 is such that $|z^{(i)}| \leq z_{\max}$ for all $i \in [n]$,

(b) The slopes varying across units and times are such that $\max\{|\beta_i^{(u)}|, \|\mathbf{B}^{(u)}\|_{\max}\} \leq \beta_{\max}$ for all $i \in [n]$ and $u \in \{v, a\}$. Further, $\|\mathbf{B}^{(y)}\|_{\max} \leq \overline{\beta}_{\max} \triangleq \max\{1, \beta_{\max}\}$.

Next, we impose certain distributional assumptions on the idiosyncratic errors. In particular, for the linear models in Examples 3.7 and 3.8, we assume the these errors are coming from zero-mean truncated Gaussian distribution and for the non-linear models in Examples 3.9 and 3.10, we assume the these errors are coming from zero-mean logistic distribution.

Assumption 3.5 (Different error models)**.**

(a) The idiosyncratic errors $\{\boldsymbol{\eta}^{(i)} \triangleq (\eta_1^{(i)}, \cdots, \eta_p^{(i)})\}_{i\in[n]}$ are independently distributed as per $f(\boldsymbol{\eta}) \propto \exp\big(\boldsymbol{\eta}^\top \mathbf{E} \boldsymbol{\eta}\big)$ where $\mathbf{E} \in \mathbb{R}^{p\times p}$ is a symmetric matrix such that $\|\mathbf{E}\|_\infty \leq \beta$ akin to Assumption 3.1.

(b) The idiosyncratic errors $\{\eta_t^{(i)}\}_{i\in[n], t\in[p]}$ are independently distributed as per a logistic distribution with location parameter 0 and scale parameter 1.

Finally, we assume bounded eigenvalues for certain matrices.

Assumption 3.6 (Bounded eigenvalues)**.** *Let $\Phi^{(u,y)} \triangleq \mathbf{B}^{(u)}\mathbf{E}\boldsymbol{B}^{(y)\top} \in \mathbb{R}^{p\times p}$ for every $u \in \{v, a, y\}$ where $\mathbf{E}$ is as in Assumption 3.5(a). Let $\boldsymbol{o}^{(i)} \triangleq \big(1, \boldsymbol{v}^{(i)}, \boldsymbol{a}^{(i)}\big) \in \mathbb{R}^{p\times 3}$ for every $i \in [n]$.*

(a) The eigenvalues of $\Phi^{(y,y)\top}\Phi^{(y,y)}$ are lower bounded by κ_1 for some $\kappa_1 > 0$.

(b) For every $i \in [n]$ and $u \in \{v, a, y\}$, the eigenvalues of $\boldsymbol{o}^{(i)\top}\Phi^{(u,y)}\Phi^{(u,y)\top}\boldsymbol{o}^{(i)}$ are lower bounded by $\kappa_2 p$ for some $\kappa_2 > 0$.

(c) For every $i \in [n]$, the eigenvalues of $\boldsymbol{o}^{(i)\top}\boldsymbol{o}^{(i)}$ are lower bounded by $\kappa_3 p$ for some $\kappa_3 > 0$.

The following result, proven in Section 3.E.1, provides guarantees for recovering the unknown slopes for the linear models in Examples 3.7 and 3.8.

Proposition 3.2 (Guarantees for Examples 3.7 and 3.8)**.** *Suppose the idiosyncratic errors are as in Assumption 3.5(a). Suppose Assumption 3.2, Assumption 3.4, Assumption 3.6(a), and Assumption 3.6(b) hold. Fix any $\varepsilon > 0$ and $\delta \in (0,1)$. Define $\beta^{(y)} = 1$, $x \triangleq \{y, a, v\}$, $\beta_i^{(z)} \triangleq z^{(i)}$ for $i \in [n]$ for Example 3.9 and $\boldsymbol{z} = (z_1, \cdots, z_p)$ for Example 3.10. For any $u \in x$, if $\mathbf{B}^{(u)} = -\beta^{(u)}\mathbf{I}$ with $|\beta^{(u)}| \geq \beta_{\min}$, then there exists estimates $\{\widehat{\mathbf{B}}^{(w)}\}_{w\in x\setminus\{u\}}$, $\big\{(\beta_i^{(v)}, \beta_i^{(a)}, \beta_i^{(y)})\big\}_{i=1}^n$, and $\widehat{\boldsymbol{z}}$ such that, the following results hold with probability at least $1-\delta$.*

(a) Linear Unit Fixed Effects Model: *For the model in Example 3.7,*

$$\max_{w\in x\setminus\{u\}} \|\widehat{\mathbf{B}}^{(w)} - \mathbf{B}^{(w)}\|_{2,\infty} \leq \varepsilon \quad \textit{for} \quad n \geq \frac{cp^4\big(\log\frac{p}{\delta} + p\log\frac{p}{\varepsilon^4\kappa_1^2}\big)}{\varepsilon^8\kappa_1^4}, \tag{3.29}$$

$$\max_{w\in\{z,v,a\},i\in[n]} \big|\beta_i^{(w)}\big|^2 \leq \frac{c}{\kappa_2(1-\varepsilon)}\left(\varepsilon^2\kappa_2^2 + \frac{\log(\log\frac{np}{\delta})}{p}\right) \quad \textit{for} \quad n \geq \frac{cp^6\big(\log\frac{np}{\delta} + \frac{p^2}{\varepsilon^2\kappa_2^2}\big)}{\varepsilon^4\kappa_2^4}. \tag{3.30}$$

(b) Linear Time Fixed Effects Model: *For the model in Example 3.8,*

$$\max_{w\in x\backslash\{u\}} \|\widehat{\mathbf{B}}^{(w)} - \mathbf{B}^{(w)}\|_{2,\infty} \leq \varepsilon \quad \textit{for} \quad n \geq \frac{cp\log\frac{p}{\sqrt{\delta}}}{\varepsilon_1^4\kappa_1^2} \tag{3.31}$$

$$\max\big\{\|\widehat{\boldsymbol{z}} - \boldsymbol{z}\|_2, \max_{w\in\{v,a,y\}} \|\widehat{\mathbf{B}}^{(w)} - \mathbf{B}^{(w)}\|_{2,\infty}\big\} \leq \varepsilon \quad \textit{for} \quad n \geq \frac{cp^3\log\frac{p}{\sqrt{\delta}}}{\varepsilon_1^4\kappa_1^2}. \tag{3.32}$$

Proposition 3.2 shows that the unknown parameters can be recovered as long as there is no dynamics in either the outcomes, the interventions, or the observed covariates.

The following result, proven in Section 3.E.2, provides guarantees for recovering the unknown slopes for the linear models in Examples 3.9 and 3.10.

Proposition 3.3 (Guarantees for Examples 3.9 and 3.10)**.** *Suppose the idiosyncratic errors are as in Assumption 3.5(b). Suppose Assumption 3.2, Assumption 3.4, Assumption 3.6(a), and Assumption 3.6(c) hold. Fix any* $\varepsilon > 0$ *and* $\delta \in (0,1)$*. Define* $\beta_i^{(z)} \triangleq z^{(i)}$ *for* $i \in [n]$ *for Example 3.9 and* $\boldsymbol{z} = (z_1, \cdots, z_p)$ *for Example 3.10. Then, there exists estimates* $\{\widehat{\mathbf{B}}^{(u)}\}_{u\in\{v,a,y\}}$, $\big\{(\beta_i^{(v)}, \beta_i^{(a)}, \beta_i^{(y)})\big\}_{i=1}^n$, *and* $\widehat{\boldsymbol{z}}$ *such that, the following results hold with probability at least* $1-\delta$.

(a) Non-linear Unit Fixed Effects Model: *For the model in Example 3.9,*

$$\max_{w\in\{v,a,y\}} \|\widehat{\mathbf{B}}^{(w)} - \mathbf{B}^{(w)}\|_{2,\infty} \leq \varepsilon \quad \textit{for} \quad n \geq \frac{cp^2\big(\log\frac{p}{\delta} + p\log\frac{1}{\varepsilon^2} + \frac{1}{\varepsilon^2}\big)}{\varepsilon^4}$$

$$\max_{w\in\{z,v,a\},i\in[n]} \big|\beta_i^{(w)}\big|^2 \leq \frac{1}{\kappa_3}\max\left\{\varepsilon^2, \frac{c\log(\log\frac{p}{\delta})}{p}\right\} \quad \textit{for} \quad n \geq \frac{cp^4\big(\log\frac{p}{\delta} + \frac{p}{\varepsilon^2}\big)}{\varepsilon^4}.$$

(b) Non-linear Time Fixed Effects Model: *For the model in Example 3.10,*

$$\max\big\{\|\widehat{\boldsymbol{z}} - \boldsymbol{z}\|_2, \max_{w\in\{v,a,y\}} \|\widehat{\mathbf{B}}^{(w)} - \mathbf{B}^{(w)}\|_{2,\infty}\big\} \leq \varepsilon \quad \textit{for} \quad n \geq \frac{c\log\frac{p}{\sqrt{\delta}}}{\varepsilon^2}.$$

We emphasize that our methodology also recovers the unobserved covariates $\{z^{(i)}\}_{i\in[n]}$ for Examples 3.7 and 3.9 and $\boldsymbol{z}$ for Examples 3.8 and 3.10.

3.7 Application: Imputing Missing Covariates

Consider a setting with no systematically unobserved covariates $\mathbf{z}$; instead, elements of $(\mathbf{v}, \mathbf{a}, \mathbf{y})$ are missing or have measurement error for some fraction of the units. Our goal is to impute these missing values or denoise the measurement error in the observed values.

For the ease of exposition, we assume the observed covariates $\mathbf{v}$ can have measurement error but the interventions and the outcomes do not have any measurement error. Our analysis remains the same (i) when observed covariates $\mathbf{v}$ are missing instead of having measurement error or (ii) when interventions / outcomes have measurement error / are missing. We note that our analysis also applies to the scenario where the unobserved covariates $\mathbf{z}$ are observed for some fraction of the units and need to be imputed for the remaining fraction of the units.

Problem setup. For every unit $i \in [n]$, along with the interventions $\boldsymbol{a}^{(i)}$ and the outcomes $\boldsymbol{y}^{(i)}$, we observe $\overline{\boldsymbol{v}}^{(i)} = \boldsymbol{v}^{(i)} + \Delta\boldsymbol{v}^{(i)}$ instead of true covariates $\boldsymbol{v}^{(i)}$ where $\Delta\boldsymbol{v}^{(i)}$ denotes (unobserved) bounded measurement error. We assume that a certain number of units (known to us) have no measurement error: say, $\Delta\boldsymbol{v}^{(i)} = 0$ for all $i \in \{n/2+1, \cdots, n\}$.

Questions of interest. Besides counterfactual estimates, our goal is to estimate $\Delta\boldsymbol{v}^{(i)}$ for units with measurement error.

3.7.1 A theoretical guarantee

Our methodology can be applied to estimate these measurement errors when the conditional distribution of the observed outcomes $\mathbf{y} \in \mathcal{X}^{p_y}$ given the interventions $\mathbf{a} \in \mathcal{X}^{p_a}$ and the true covariates $\mathbf{v} \in \mathcal{X}^{p_v}$ can be modeled as an exponential family, parameterized by a vector $\phi^{(y)} \in \mathbb{R}^{p_y \times 1}$ and matrices $\Phi^{(u,y)} \in \mathbb{R}^{p_u \times p_y}$ for all $\mathbf{u} \in \{\mathbf{v}, \mathbf{a}, \mathbf{y}\}$

$$f_{\mathbf{y}|\mathbf{a},\mathbf{v}}(\boldsymbol{y}|\boldsymbol{a},\boldsymbol{v}) \propto \exp\Big(\big[\phi^{(y)^\top} + 2\boldsymbol{v}^\top\Phi^{(v,y)} + 2\boldsymbol{a}^\top\Phi^{(a,y)}\big]\boldsymbol{y} + \boldsymbol{y}^\top\Phi^{(y,y)}\boldsymbol{y}\Big) \tag{3.33}$$

where $\boldsymbol{v} \triangleq (v_1, \cdots, v_{p_v})$, $\boldsymbol{a} \triangleq (a_1, \cdots, a_{p_a})$, and $\boldsymbol{y} \triangleq (y_1, \cdots, y_{p_y})$ denote realizations of $\mathbf{v}$, $\mathbf{a}$, and $\mathbf{y}$, respectively. To estimate the counterfactual distribution, we decompose $\mathbf{v}$ into $\overline{\mathbf{v}}$ and $\Delta\mathbf{v}$, and obtain the distribution of the outcome $\mathbf{y}$ conditioned on $(\mathbf{a}, \overline{\mathbf{v}}, \Delta\mathbf{v}) = (\boldsymbol{a}, \overline{\boldsymbol{v}}, \Delta\boldsymbol{v})$ as follows

$$f_{\mathbf{y}|\mathbf{a},\overline{\mathbf{v}},\Delta\mathbf{v}}(\boldsymbol{y}|\boldsymbol{a},\overline{\boldsymbol{v}},\Delta\boldsymbol{v}) \propto \exp\Big(\big[\phi^{(y)^\top} + 2\Delta\boldsymbol{v}^\top\Phi^{(v,y)} + 2\overline{\boldsymbol{v}}^\top\Phi^{(v,y)} + 2\boldsymbol{a}^\top\Phi^{(a,y)}\big]\boldsymbol{y} + \boldsymbol{y}^\top\Phi^{(y,y)}\boldsymbol{y}\Big) \tag{3.34}$$

As in Section 3.3.2, to estimate the counterfactual distribution, it suffices to learn

$$\theta(\Delta\boldsymbol{v}) \triangleq \phi^{(y)} + 2\Phi^{(v,y)\top}\Delta\boldsymbol{v} \quad \text{and} \quad \Theta \triangleq \big[\Phi^{(y,y)}, \Phi^{(y,a)}, \Phi^{(y,v)}\big], \tag{3.35}$$

with $\Phi^{(y,a)} = \Phi^{(a,y)^\top} \in \mathbb{R}^{p_y \times p_a}$ and $\Phi^{(y,v)} = \Phi^{(v,y)^\top} \in \mathbb{R}^{p_y \times p_v}$. Here, we also aim to learn $\Delta\boldsymbol{v}$.

Let $f_{\mathbf{y}|\mathbf{a},\mathbf{v}}(\cdot; \boldsymbol{a}, \boldsymbol{v}; \phi^{(y)^\star}, \Theta^\star)$ denote the true data generating distribution of $\mathbf{y}$ conditioned on $(\mathbf{a}, \mathbf{v}) = (\boldsymbol{a}, \boldsymbol{v})$ in Eq. (3.33) and let $f_{\mathbf{y}|\mathbf{a},\overline{\mathbf{v}},\Delta\mathbf{v}}\big(\cdot\,|\boldsymbol{a}, \overline{\boldsymbol{v}}, \Delta\boldsymbol{v}; \theta^\star(\Delta\boldsymbol{v}), \Theta^\star\big)$ denote the true distribution of $\mathbf{y}$ conditioned on $(\mathbf{a}, \overline{\mathbf{v}}, \Delta\mathbf{v}) = (\boldsymbol{a}, \overline{\boldsymbol{v}}, \Delta\boldsymbol{v})$ in Eq. (3.34). We assume (a) $\max\big\{\|\Delta\boldsymbol{v}\|_\infty, \|\phi^{(y)^\star}\|_\infty, \|\Theta^\star\|_{\max}\big\} \le \alpha$ and (b) $\|\Theta^\star\|_\infty \le \beta$ analogous to Assumption 3.1 where the row-wise ℓ_1 sparsity in (b) is assumed to be induced by row-wise ℓ_0 sparsity, i.e., $\|\Theta_t^\star\|_0 \le \beta/\alpha$ for all $t \in [p]$. Then, given realizations $\{\boldsymbol{y}^{(i)}, \boldsymbol{a}^{(i)}, \overline{\boldsymbol{v}}^{(i)}\}_{i=1}^n$

consistent with $\big\{f_{\mathsf{y}|\mathsf{a},\bar{\mathsf{v}},\Delta\mathsf{v}}\big(\boldsymbol{y}^{(i)}|\boldsymbol{a}^{(i)},\overline{\boldsymbol{v}}^{(i)},\Delta\boldsymbol{v}^{(i)};\theta^{\star}(\Delta\boldsymbol{v}^{(i)}),\Theta^{\star}\big)\big\}_{i=1}^{n}$ first, we estimate the parameters $\phi^{(y)\star}$ and $\Theta^{\star}$ using the realizations for units $\{n/2+1,\cdots,n\}$. Next, we exploit the structure in the problem to show that $\theta^{\star(i)} \triangleq \theta^{\star}(\Delta\boldsymbol{v}^{(i)})$ can be written as a linear combination of known vectors with some error, for every unit $i \in \{1,\cdots,n/2\}$. Then, we use the loss function in Eq. (3.12) to estimate $\{\theta^{\star(i)}\}_{i=1}^{n}$ and obtain estimates of $\{\Delta\boldsymbol{v}^{(i)}\}_{i=1}^{n}$ as by-products. In particular, the estimate of the coefficients associated with the aforementioned linear combination for $\theta^{\star(i)}$ turn out to be our estimate of the measurement error $\Delta\boldsymbol{v}^{(i)}$ for every $i \in \{1,\cdots,n/2\}$. For $i \in \{n/2+1,\cdots,n\}$, estimating $\theta^{\star(i)}$ and $\Delta\boldsymbol{v}^{(i)}$ is straightforward since $\theta^{\star(i)} = \phi^{(y)\star}$ and $\Delta\boldsymbol{v}^{(i)} = 0$. We provide our guarantee on estimating $\Theta^{\star}$, $\theta^{\star(i)}$ for $i \in [n]$, and $\Delta\boldsymbol{v}^{(i)}$ for $i \in [n]$ below with a proof in Section 3.F.

Proposition 3.4 (Impute missing covariates)**.** *Suppose the eigenvalues of $\boldsymbol{D}^{\top}\boldsymbol{D}$ are lower bounded by κp for some $\kappa > 0$ where $\boldsymbol{D} \triangleq \big[\phi^{(y)\star}, 2\Phi^{(y,v)\star}\big] \in \mathbb{R}^{p\times(p_v+1)}$. Then, for any fixed $\varepsilon_1 > 0$ and $\delta \in (0,1)$, there exists estimates $\widehat{\Theta}$ and $\big\{\widehat{\theta}^{(i)}\big\}_{i=1}^{n}$ such that, with probability at least $1-\delta$,*

$$\|\widehat{\Theta}-\Theta^{\star}\|_{2,\infty} \le \varepsilon_1 \quad \textit{for} \quad n \ge \frac{ce^{c'\beta}\log\frac{p_y}{\sqrt{\delta}}}{\varepsilon_1^2},$$

and

$$\max_{i\in[n]} \mathrm{MSE}(\widehat{\theta}^{(i)},\theta^{\star(i)}) \le \max\bigg\{\varepsilon_1^2, \frac{ce^{c'\beta}\big(p_v+\log(\log\frac{np_y}{\delta})\big)}{p_y}\bigg\} \textit{ for } n \ge \frac{ce^{c'\beta}\widetilde{p}^2\big(\log\frac{\sqrt{n}p_y}{\sqrt{\delta}}+p_v\big)}{\varepsilon_1^2}.$$

Further, for any fixed $\varepsilon_2 > 0$, if $\varepsilon_2 \le \frac{1}{8}\sqrt{\frac{p_y}{p_v+1}}$, there exist estimates $\big\{\widehat{\Delta\boldsymbol{v}}^{(i)}\big\}_{i=1}^{n}$ such that,

$$\max_{i\in[n]} \|\widehat{\Delta\boldsymbol{v}}^{(i)} - \Delta\boldsymbol{v}^{(i)}\|_2^2 \le \frac{ce^{c'\beta}\big(p_v+\log(\log\frac{np_y}{\delta})\big)}{p_y\kappa} + \frac{4\varepsilon_2^2\kappa}{p_y},$$

with probability at least $1-\delta$, whenever $n \ge ce^{c'\beta}\kappa^{-2}\varepsilon_2^{-2}(p_v+1)p_y\widetilde{p}^2\big(\log\frac{\sqrt{n}p_y}{\sqrt{\delta}}+p_v\big)$.

The above guarantees can be simplified as follows by treating β and κ as constants as well as ignoring the constants, and the logarithmic factors in n and δ (denoted by $\precsim$ and $\succsim$): for any $\varepsilon_1 > 0$ and $\frac{1}{8}\sqrt{\frac{p_y}{p_v+1}} \ge \varepsilon_2 > 0$

$$\|\widehat{\Theta}-\Theta^{\star}\|_{2,\infty} \le \varepsilon_1 \qquad \text{when} \quad n \succsim \frac{\log p_y}{\varepsilon_1^2}, \tag{3.36}$$

$$\max_{i\in[n]} \mathrm{MSE}(\widehat{\theta}^{(i)},\theta^{\star(i)}) \precsim \max\Big\{\varepsilon_1^2, \frac{p_v}{p_y}\Big\} \qquad \text{when} \quad n \succsim \frac{\widetilde{p}^2(\log p_y + p_v)}{\varepsilon_1^2}, \tag{3.37}$$

and

$$\max_{i\in[n]} \|\widehat{\Delta\boldsymbol{v}}^{(i)} - \Delta\boldsymbol{v}^{(i)}\|_2^2 \precsim \frac{p_v}{p_y} + \frac{\varepsilon_2^2}{p_y} \qquad \text{when} \quad n \succsim \frac{p_v p_y \widetilde{p}^2(\log p_y + p_v)}{\varepsilon_2^2}. \tag{3.38}$$

For large n, whenever, $\max\left\{\varepsilon_1^2, \frac{p_v}{p_y}\right\} = \frac{p_v}{p_y}$ and $\max\left\{\frac{\varepsilon_2^2}{p_y}, \frac{p_v}{p_y}\right\} = \frac{p_v}{p_y}$, the guarantees in Eqs. (3.37) and (3.38) can be written as

$$\max_{i\in[n]} \mathrm{MSE}(\widehat{\theta}^{(i)}, \theta^{\star(i)}) \precsim \frac{p_v}{p_y} \qquad \text{when} \quad n \succsim \frac{\widetilde{p}^2 p_y(\log p_y + p_v)}{p_v}, \tag{3.39}$$

and

$$\max_{i\in[n]} \|\widehat{\Delta \boldsymbol{v}}^{(i)} - \Delta\boldsymbol{v}^{(i)}\|_2^2 \precsim \frac{p_v}{p_y} \qquad \text{when} \quad n \succsim p_y \widetilde{p}^2(\log p_y + p_v). \tag{3.40}$$

Remark. The measurement errors can be recovered well as long as enough units with no measurement error are observed (i.e., $n/2$ is large) and the observation per unit is high dimensional (i.e., p_y is large compared to p_v).

3.7.2 Simulations

We now present some simulation results to empirically evaluate the error scaling of our parameter estimates with three key aspects of the application above: number of units n, total dimension $\widetilde{p}$, and dimension p_v of covariates with measurement error.

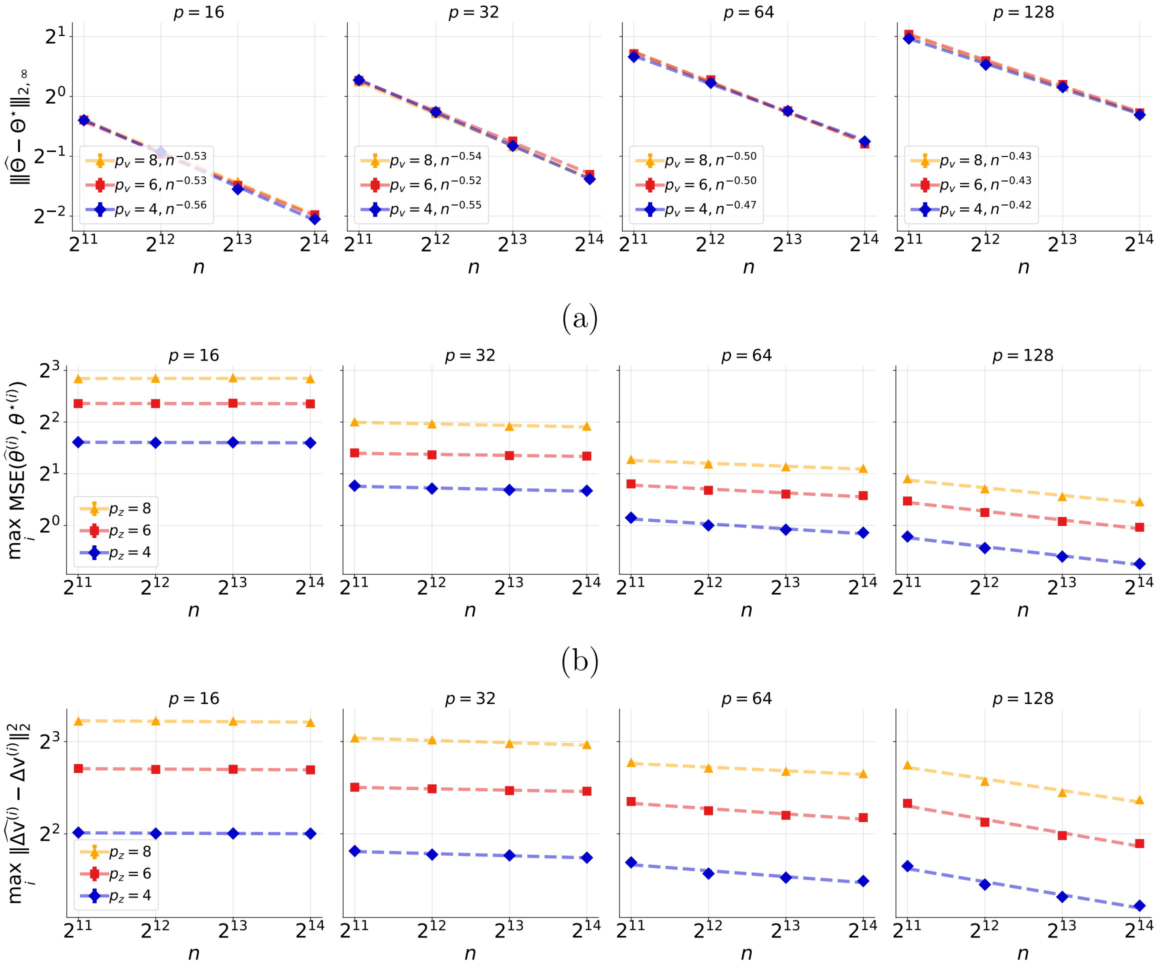
3.7.2.1 Data generation

We choose $\mathcal{X} = [-1, 1]$ and $p_a = p_y = (\widetilde{p} - p_v)/2$. The true joint distribution of $(\mathbf{v}, \mathbf{a}, \mathbf{y})$ is set as a truncated Gaussian distribution with the parameters $\phi^\star = \mathbf{1} \in \mathbb{R}^p$ and a positive definite $\Phi^\star \in \mathbb{R}^{p\times p}$ generated using *sklearn* package (Pedregosa et al., 2011) such that $\alpha = 6$, $\beta = 4$, and $\kappa = 0.15$. We draw n i.i.d. samples $\{\boldsymbol{y}^{(i)}, \boldsymbol{a}^{(i)}, \boldsymbol{v}^{(i)}\}_{i=1}^n$ from this true distrbution using *tmvtnorm* package (Wilhelm and Manjunath, 2010b). Next, we generate $\Delta\boldsymbol{v}^{(i)}$ uniformly from $[0.9, 1]^{p_v}$ for units $i \in \{1, \cdots, n/2\}$ while setting $\Delta\boldsymbol{v}^{(i)} = \mathbf{0}$ for other units. Combining $\{\boldsymbol{y}^{(i)}, \boldsymbol{a}^{(i)}, \boldsymbol{v}^{(i)}\}_{i=1}^n$ and $\{\Delta\boldsymbol{v}^{(i)}\}_{i=1}^n$ yields $\{\boldsymbol{y}^{(i)}, \boldsymbol{a}^{(i)}, \overline{\boldsymbol{v}}^{(i)}\}_{i=1}^n$ (see Eq. (3.34)).

3.7.2.2 Plot details

In Figure 3.7.1, we plot the scaling of errors in our estimates for $\Theta^\star$ in the top row, $\{\theta^{\star(i)}\}_{i=1}^n$ in the middle row, and $\{\Delta\boldsymbol{v}^{(i)}\}_{i=1}^n$ in the bottom row. In particular, we present how the error scales as the number of units n grows for various $\widetilde{p}$ and p_v. We plot the averaged error across 50 independent trials along with ± 1 standard error (the standard error is too small to be visible in our results).

To help see the error scaling, we provide the least squares fit on the log-log scale (log error vs log x-axis). We display the best linear fit and mention an empirical decay rate in the legend based on the slope of that fit, e.g., for a slope of -0.56 for estimating $\Theta^\star$ when $\widetilde{p} = 16$ and $p_v = 4$, we report an empirical rate of $n^{-0.56}$ for the averaged error. In the middle row and the bottom row of Figure 3.7.1, the rates vary from $n^{0.00}$ to $n^{-0.17}$, and we omit these weak dependencies in the legend to reduce clutter.



(a)

(b)

(c)

Figure 3.7.1: Error scaling with number of units n, for various $\widetilde{p}$ and p_v, for our estimates of $\Theta^\star$ (top row), $\{\theta^{\star(i)}\}_{i=1}^n$ (middle row), and $\{\Delta \boldsymbol{v}^{(i)}\}_{i=1}^n$ (bottom row).

3.7.2.3 Error scaling for $\widehat{\Theta}$

From the first row of Figure 3.7.1, we observe that the error $\|\widehat{\Theta} - \Theta^\star\|_{2,\infty}$ admits a scaling of between $n^{-0.56}$ and $n^{-0.42}$ for various $\widetilde{p}$ and p_v. These empirical rates indicate a parametric error rate of $n^{-0.5}$ for $\|\widehat{\Theta} - \Theta^\star\|_{2,\infty}$, consistent with the scaling of ε^{-2} in Eq. (3.36). Further, as expected, the error $\|\widehat{\Theta} - \Theta^\star\|_{2,\infty}$ does not depend on p_v but increases with an increase in $\widetilde{p}$.

3.7.2.4 Error scaling for $\widehat{\boldsymbol{\theta}}^{(i)}$

In the middle row of Figure 3.7.1, we see the error $\max_{i \in [n]} \mathrm{MSE}(\widehat{\theta}^{(i)}, \theta^{\star(i)})$ has a weak dependence on n for a fixed $\widetilde{p}$ and p_v, decreases with an increase in $\widetilde{p}$ for any fixed n and p_v, and increases with an increase in p_v for any fixed n and $\widetilde{p}$. This is consistent with Eq. (3.37) when $\max\left\{\varepsilon_1^2, \frac{p_v}{p}\right\} = \frac{p_v}{p_y} = \frac{2p_v}{\widetilde{p} - p_v}$ (see Eq. (3.39)). Further, we note that the decay of the error with $\widetilde{p}$ is slower for smaller n (cf. $n = 2^{11}$ vs $n = 2^{14}$). This is expected from Eq. (3.37) where the n required to ensure $\max\big\{\varepsilon_1^2, \frac{p_v}{p_y}\big\} = \frac{p_v}{p_y}$ increases with an increase in p_y, and therefore $\widetilde{p}$. As a result, for larger $\widetilde{p}$, ε_1^2 comes into the picture explaining the increased dependence of the error on n (cf. $\widetilde{p} = 16$ vs $\widetilde{p} = 128$).

3.7.2.5 Error scaling for $\widehat{\Delta \boldsymbol{v}}^{(i)}$

The trends in $\max_{i\in[n]} \|\widehat{\Delta \boldsymbol{v}}^{(i)} - \Delta \boldsymbol{v}^{(i)}\|_2^2$ are similar to $\max_{i\in[n]} \mathrm{MSE}(\widehat{\theta}^{(i)}, \theta^{\star(i)})$. In the bottom row of Figure 3.7.1, we see $\max_{i\in[n]}\|\widehat{\Delta \boldsymbol{v}}^{(i)} - \Delta \boldsymbol{v}^{(i)}\|_2^2$ has a weak dependence on n for a fixed $\widetilde{p}$ and p_v, decreases with an increase in $\widetilde{p}$ for any fixed n and p_v, and increases with an increase in p_v for any fixed n and $\widetilde{p}$. This is consistent with Eq. (3.38) when $\max\big\{\frac{\varepsilon_2^2}{p_y}, \frac{p_v}{p_y}\big\} = \frac{p_v}{p_y} = \frac{2p_v}{\widetilde{p} - p_v}$ (see Eq. (3.40)). For the same reason mentioned in the previous paragraph, we see a slower decay in the error with $\widetilde{p}$ for smaller n (cf. $n = 2^{11}$ vs $n = 2^{14}$), and a higher dependence of the error on n for larger $\widetilde{p}$ (cf. $\widetilde{p} = 16$ vs $\widetilde{p} = 128$).

3.8 Proof Sketch for Theorem 3.1

Our proof of Theorem 3.1 proceeds in two stages (see Figure 3.8.1 for an overview). First, we establish Eq. (3.17) for estimating $\Theta^\star$. Next, we use this guarantee to establish the unit-level guarantee Eq. (3.18) for each of $\big\{\theta^{\star(1)}, \cdots, \theta^{\star(n)}\big\}$ by substituting $\Theta = \widehat{\Theta}$ in Eq. (3.12), i.e., analyzing the following convex optimization problem:

$$\{\widehat{\theta}^{(1)}, \cdots, \widehat{\theta}^{(n)}\} \in \underset{\{\theta^{(1)},\cdots,\theta^{(n)}\}\in\Lambda_\theta^n}{\arg\min} \mathcal{L}\big(\widehat{\Theta}, \theta^{(1)}, \cdots, \theta^{(n)}\big). \tag{3.41}$$

3.8.1 Estimating the population-level parameter

In the first part, we show that all points $\underline{\Theta} \in \Lambda_\Theta \times \Lambda_\theta^n$, such that $\|\Theta_t - \Theta_t^\star\|_2 \geq \varepsilon$ for at least one $t \in [p_y]$, uniformly satisfy

$$\mathcal{L}(\underline{\Theta}) \geq \mathcal{L}(\underline{\Theta}^\star) + \Omega(\varepsilon^2) \text{ for } n \geq \frac{ce^{c'\beta}p_y^2}{\varepsilon^4} \cdot \Big(\log\frac{p_y}{\delta\varepsilon^2} + \mathcal{M}_\Theta\big(\varepsilon^2\big) + \mathcal{M}_{\theta,n}\big(\varepsilon^2\big)\Big), \tag{3.42}$$

with probability at least $1 - \delta$. Then, we conclude the proof using contraposition.

To prove Eq. (3.42), we first decompose the convex (and positive) objective $\mathcal{L}(\underline{\Theta})$ in Eq. (3.11) as a sum of p_y convex (and positive) auxiliary objectives $\mathcal{L}_t$, namely, $\mathcal{L}(\underline{\Theta}) = \sum_{t\in[p_y]} \mathcal{L}_t\big(\underline{\Theta}_t\big)$ where

$$\mathcal{L}_t\big(\underline{\Theta}_t\big) \triangleq \frac{1}{n}\sum_{i\in[n]} \exp\Big(-[\theta_t^{(i)} + 2\Theta_{t,-t}^\top \boldsymbol{x}_{-t}^{(i)}]x_t^{(i)} - \Theta_{tt}\Big[[x_t^{(i)}]^2 - \frac{x_{\max}^2}{3}\Big]\Big). \tag{3.43}$$

Next, for any fixed $t \in [p_y]$, $\varepsilon > 0$, and $\underline{\Theta} \in \Lambda_\theta^n \times \Lambda_\Theta$ with $\|\Theta_t - \Theta_t^\star\|_2 \geq \varepsilon$, we show (see Lemma 3.1)

$$\mathcal{L}_t(\underline{\Theta}_t) \geq \mathcal{L}_t(\underline{\Theta}_t^\star) + \Omega(\varepsilon^2) - \varepsilon_1 \quad \text{whenever} \quad n \geq \frac{ce^{c'\beta} \log \frac{p_y}{\delta}}{\varepsilon_1^2}, \tag{3.44}$$

and then establish the same bound uniformly for all $t \in [p_y]$ with probability $1 - \delta$. Taking a sum over t on both sides of Eq. (3.44), we conclude that for any fixed $\underline{\Theta}$ with $\|\Theta_t - \Theta_t^\star\|_2 \geq \varepsilon$ for some $t \in [p_y]$,

$$\mathcal{L}(\underline{\Theta}) \geq \mathcal{L}(\underline{\Theta}^\star) + \Omega(\varepsilon^2) \quad \text{whenever} \quad n \geq \frac{ce^{c'\beta} p_y^2 \log \frac{p_y}{\delta}}{\varepsilon^4}, \tag{3.45}$$

with probability at least $1-\delta$ where we substituted $\varepsilon_1 = c\varepsilon^2/p_y$. Finally, we conclude Eq. (3.42) by using Eq. (3.45), the Lipschitzness of $\mathcal{L}$ (see Lemma 3.2), and a covering number argument (see Section 3.B).

We establish Eq. (3.44) (Lemma 3.1) via Lemma 3.3, which provides suitable concentration and anti-concentration results for the first-order and second-order derivatives, respectively, for the auxiliary objective $\mathcal{L}_t$ in Eq. (3.43). We prove Lemma 3.3 by extending the results from Shah et al. (2021c) to the setting with non-identical but independent samples $\{\boldsymbol{y}^{(i)} \sim f_{\mathbf{y}|\mathbf{a},\mathbf{v},\mathbf{z}}\big(\cdot\,|\boldsymbol{a}^{(i)}, \boldsymbol{v}^{(i)}, \boldsymbol{z}^{(i)}; \theta^\star(\boldsymbol{z}^{(i)}), \Theta^\star\big)\}_{i=1}^n$.

3.8.2 Estimating the unit-level parameters

In the second part, we decompose the convex optimization problem in Eq. (3.41) into n convex optimization problems:

$$\mathcal{L}^{(i)}\big(\theta^{(i)}\big) \triangleq \sum_{t \in [p_y]} \exp\Big(- \big[\theta_t^{(i)} + 2\widehat{\Theta}_{t,-t}^\top \boldsymbol{x}_{-t}^{(i)}\big] x_t^{(i)} - \widehat{\Theta}_{tt}\big([x_t^{(i)}]^2 - \frac{x_{\max}^2}{3}\big)\Big) \text{ for } i \in [n]. \tag{3.46}$$

Noting that the set Λ_θ^n places independent constraints on the n unit-level parameters, namely $\theta^{(i)} \in \Lambda_\theta$, independently for all $i \in [n]$ and combining Eqs. (3.11) and (3.41), we find that

$$\min_{\{\theta^{(1)},\cdots,\theta^{(n)}\} \in \Lambda_\theta^n} \mathcal{L}\big(\widehat{\Theta}, \theta^{(1)}, \cdots, \theta^{(n)}\big) \overset{Eq.\ (3.46)}{=} \frac{1}{n} \sum_{i \in [n]} \min_{\theta^{(i)} \in \Lambda_\theta} \mathcal{L}^{(i)}\big(\theta^{(i)}\big) \implies \widehat{\theta}^{(i)} \in \arg\min_{\theta^{(i)} \in \Lambda_\theta} \mathcal{L}^{(i)}\big(\theta^{(i)}\big),$$

for each $i \in [n]$. Next, we establish that with probability at least $1 - \delta$,

$$\mathcal{L}^{(i)}\big(\theta^{(i)}\big) \geq \mathcal{L}^{(i)}\big(\theta^{\star(i)}\big) + R^2(\varepsilon, \delta) \text{ when } n \geq \frac{ce^{c'\beta} p_y^2 \widetilde{p}^2 \Big(\log \frac{p_y \widetilde{p}}{\delta \varepsilon^2} + \mathcal{M}_\Theta\big(\frac{\varepsilon^2}{\widetilde{p}}\big) + \widetilde{\mathcal{M}}_{\theta,n}(\varepsilon, \delta)\Big)}{\varepsilon^4}, \tag{3.47}$$

uniformly for all points $\theta^{(i)} \in \Lambda_\theta$ with $\|\theta^{(i)} - \theta^{\star(i)}\|_2 \geq R(\varepsilon, \delta)$ (see Eq. (3.15)). We conclude the proof by contraposition with the basic inequality $\mathcal{L}^{(i)}(\widehat{\theta}^{(i)}) \leq \mathcal{L}^{(i)}(\theta^{\star(i)})$ and a standard union bound over all $i \in [n]$.

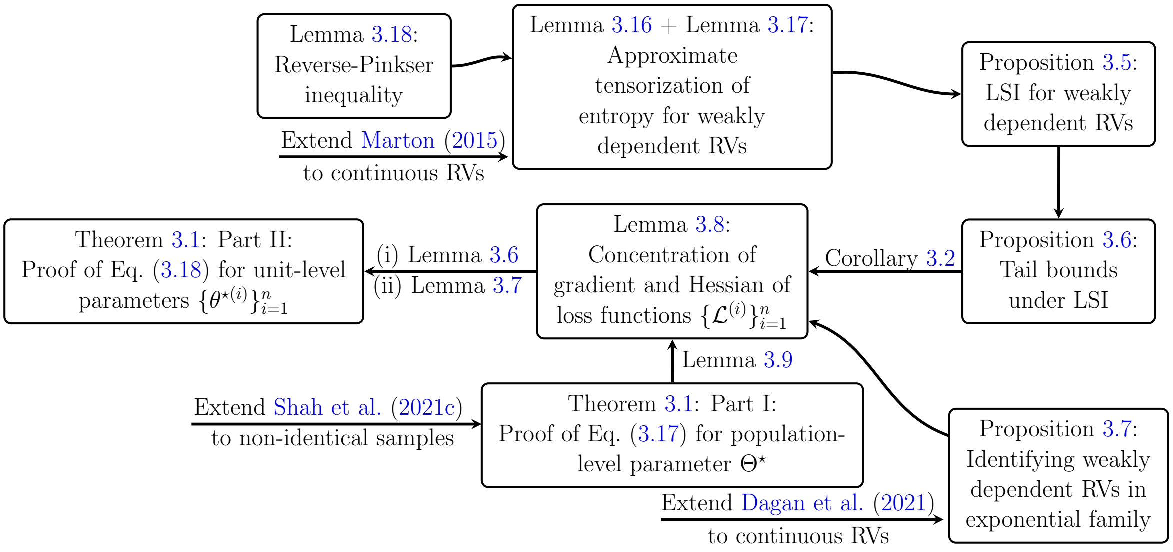

Lemma 3.18: Reverse-Pinkser inequality
Lemma 3.16 + Lemma 3.17: Approximate tensorization of entropy for weakly dependent RVs
Proposition 3.5: LSI for weakly dependent RVs
Extend Marton (2015) to continuous RVs
Theorem 3.1: Part II: Proof of Eq. (3.18) for unit-level parameters $\{\theta^{\star(i)}\}_{i=1}^n$
(i) Lemma 3.6
(ii) Lemma 3.7
Lemma 3.8: Concentration of gradient and Hessian of loss functions $\{\mathcal{L}^{(i)}\}_{i=1}^n$
Corollary 3.2
Proposition 3.6: Tail bounds under LSI
Lemma 3.9
Extend Shah et al. (2021c) to non-identical samples
Theorem 3.1: Part I: Proof of Eq. (3.17) for population-level parameter $\Theta^\star$
Proposition 3.7: Identifying weakly dependent RVs in exponential family
Extend Dagan et al. (2021) to continuous RVs

Figure 3.8.1: Sketch diagram of the results and the proof techniques for Theorem 3.1. First, we establish Eq. (3.17) for estimating $\Theta^\star$ by extending Shah et al. (2021c, Proposition I.1, Proposition I.2) for i.i.d. data to non-identical samples. Next, we use Eq. (3.17) to establish Eq. (3.18) for the unit-level parameters $\{\theta^{\star(i)}\}_{i=1}^n$ via suitable concentration results for derivatives of the auxiliary loss functions in kEq. (3.46). En route, we establish three results of independent interest: (i) Proposition 3.5 that shows that weakly dependent and bounded random variables satisfy logarithmic Sobolev inequality (LSI) by both extending Marton (2015, Theorem. 1, Theorem. 2) and establishing a reverse-Pinkser inequality to continuous random vectors; (ii) Proposition 3.6 that extends the tail bounds Dagan et al. (2021, Theorem. 6) to continuous distributions satisfying LSI; and (iii) Proposition 3.7 that extends the conditioning trick Dagan et al. (2021, Lemma. 2) for identifying a weakly dependent subset to continuous random vectors.

The proof of Eq. (3.47) mimics the same road map as that for Eq. (3.42). Lemma 3.6 shows that for any fixed $\theta^{(i)} \in \Lambda_\theta$, if $\theta^{(i)}$ is far from $\theta^{\star(i)}$, then with high probability $\mathcal{L}^{(i)}\big(\theta^{(i)}\big)$ is significantly larger than $\mathcal{L}^{(i)}\big(\theta^{\star(i)}\big)$. We prove Lemma 3.6 via concentration of derivatives of $\mathcal{L}^{(i)}$ Eq. (3.46) in Lemma 3.8, this objective's Lipschitznes in Lemma 3.7, and a covering number argument (see Section 3.C).

The proof of Lemma 3.8 involves several novel arguments: First, for a τ-Sparse Graphical Model (Definition 3.8), i.e., a generalization of the random vector $\mathbf{y}$ in Eq. (3.4), Proposition 3.7 identifies a subset that satisfies Dobrushin's uniqueness condition (Definition 3.4) after conditioning on the complementary subset. Second, Proposition 3.5 shows that a bounded and weakly dependent continuous random vector (defined using Dobrushin's uniqueness condition) satisfies the logarithmic Sobolev inequality (LSI). Third, Proposition 3.6 establishes tail bounds for arbitrary functions of a continuous random vector that satisfies LSI. Putting together these results and a robustness result (Lemma 3.9) while invoking concentration results to account for the estimation error for $\Theta^\star$, yields Lemma 3.8.

3.9 Concluding Remarks

We introduce an exponential family approach to learn unit-level counterfactual distributions from a single sample per unit even when there is unobserved confounding. By conditioning on the latent confounders and using a novel convex loss function, we estimate the parameters of unit-level counterfactual distributions given the information about what actually happened. The resulting estimates of unit-level counterfactual distributions enable us to estimate any functional of each unit's potential outcomes under alternate interventions. We analyze each unit's expected potential outcomes under alternate interventions, thereby providing a guarantee on unit-level counterfactual effects, i.e., individual treatment effects. We note that our approach makes only macro-level assumptions about the underlying causal graph and does not assume the knowledge of the micro-level causal graph.

A side product of our results is a strategy for answering interventional questions, e.g., to estimate average treatment effects. These questions are equivalent to estimating distributions of the form $f_{\mathbf{y}|\mathrm{do}(\mathbf{a})}(\boldsymbol{y}|\mathrm{do}(\mathbf{a}=\boldsymbol{a}))$ where the do-operator (Pearl, 2009) forces $\mathbf{a}$ to be $\boldsymbol{a}$. Under the causal framework considered (Figure 3.1.2, we have $f_{\mathbf{y}|\mathrm{do}(\mathbf{a})}(\boldsymbol{y}|\mathrm{do}(\mathbf{a}=\boldsymbol{a})) = \mathbb{E}_{\mathbf{v},\mathbf{z}}[f_{\mathbf{y}|\mathbf{a},\mathbf{z},\mathbf{v}}(\boldsymbol{y}|\boldsymbol{a},\boldsymbol{v},\boldsymbol{z})]$. Consequently, the mixture distribution $n^{-1}\sum_{i\in[n]}\widehat{f}^{(i)}_{\mathbf{y}|\mathbf{a}}(\boldsymbol{y}|\boldsymbol{a})$ with $\widehat{f}^{(i)}_{\mathbf{y}|\mathbf{a}}(\boldsymbol{y}|\boldsymbol{a})$ defined in Eq. (3.13), serves as a natural estimate via our strategy. Investigating the efficacy of this estimator is an interesting future direction.

In this work, the conditional exponential family distribution of $\mathbf{y}$ in Section 3.3.2 or in Section 3.5.1 was such that the effect of unobserved covariates $\mathbf{z}$—after conditioning on them—was captured by a first-order interaction term varying with the realized value of $\mathbf{z}$ for each unit, e.g., $\{\theta(\boldsymbol{z}^{(i)})\}_{i=1}^{n}$for the conditional distribution in Section 3.3.2. Focusing on Section 3.3.2, the conditional distributions could also have higher-order interaction terms that vary with $\mathbf{z}$. Focusing on Section 3.5.1, the exponent of the exponential tilting of the base distribution of the outcomes by the unobserved covariates could have higher-order terms. For such cases, while our analysis for population-level parameters (Theorem 3.1 Part I's proof in Section 3.B) is likely to extend easily, new arguments for analyzing quadratic (or higher-order) interaction terms that vary for each unit seem necessary. Developing these results, e.g., suitable analogs of Dobrushin's condition for higher-order exponential family, present an exciting future venue for research.

Our methodology can be useful for a class of multi-task learning problems (Caruana, 1997), e.g., when we have multiple logistic regression tasks with some commonalities. For a logistic regression task, the exponential family model Eq. (3.4) has been used by Dagan et al. (2021) to allow dependencies between the labels via the parameter Θ (instead of assuming independence between the labels), e.g., for spatio-temporal data. They consider a single regression task and assume that the dependency matrix Θ is known up to a constant and learn a task-specific parameter $\theta(\boldsymbol{z})$ (where $\boldsymbol{z}$ denotes a task). Our model and methodology apply to the case of fully unknown Θ given multiple datasets that share the same dependency parameter Θ but have varying task-specific parameters $\theta(\boldsymbol{z})$; and provide a tractable way to estimate all these parameters together. In fact, our framework and results also apply beyond the quadratic dependencies captured by

Θ as described in Section 3.5.1. Analyzing whether our methodology can be extended beyond logistic regression models for multi-task learning is a question worthy of further investigation.

Appendix

3.A Proper loss function and projected gradient descent

In this section, we prove Proposition 3.1 showing that the loss function in Eq. (3.11) is a proper loss function. We also provide an algorithm to obtain an ϵ-optimal estimate of $\underline{\widehat{\Theta}}$.

3.A.1 Proof of Proposition 3.1

Fix any $\boldsymbol{z} \in \mathcal{Z}^{p_z}$, $\boldsymbol{v} \in \mathcal{V}^{p_v}$, and $\boldsymbol{a} \in \mathcal{A}^{p_a}$, and recall $\mathbf{x} = (\mathbf{y}, \mathbf{a}, \mathbf{v})$. Fix any $t \in [p_y]$ and define the following parametric distribution

$$u_{\mathsf{y}|\mathsf{a},\mathsf{v},\mathsf{z}}\big(\boldsymbol{y}|\boldsymbol{a},\boldsymbol{v},\boldsymbol{z};\theta_t(\boldsymbol{z}),\Theta_t\big) \propto \frac{f_{\mathsf{y}|\mathsf{a},\mathsf{v},\mathsf{z}}\big(\boldsymbol{y}|\boldsymbol{a},\boldsymbol{v},\boldsymbol{z};\theta^\star(\boldsymbol{z}),\Theta^\star\big)}{f_{\mathsf{x}_t|\mathsf{x}_{-t},\mathsf{z}}(x_t|\boldsymbol{x}_{-t},\boldsymbol{z};\theta_t(\boldsymbol{z}),\Theta_t)}, \tag{3.48}$$

where $f_{\mathsf{y}|\mathsf{a},\mathsf{v},\mathsf{z}}\big(\boldsymbol{y}|\boldsymbol{a},\boldsymbol{v},\boldsymbol{z};\theta^\star(\boldsymbol{z}),\Theta^\star\big)$ is as defined in Eq. (3.4) and $f_{\mathsf{x}_t|\mathsf{x}_{-t},\mathsf{z}}(x_t|\boldsymbol{x}_{-t},\boldsymbol{z};\theta_t(\boldsymbol{z}),\Theta_t)$ is as defined in Eq. (3.10). Letting $\overline{x}_t \triangleq x_t^2 - x_{\max}^2/3$ and using Eq. (3.10), we can write $u_{\mathsf{y}|\mathsf{a},\mathsf{v},\mathsf{z}}\big(\boldsymbol{y}|\boldsymbol{a},\boldsymbol{v},\boldsymbol{z};\theta_t(\boldsymbol{z}),\Theta_t\big)$ in Eq. (3.48) as

$$\begin{aligned} u_{\mathsf{y}|\mathsf{a},\mathsf{v},\mathsf{z}}\big(\boldsymbol{y}|\boldsymbol{a},\boldsymbol{v},\boldsymbol{z};\theta_t(\boldsymbol{z}),\Theta_t\big) \propto f_{\mathsf{y}|\mathsf{a},\mathsf{v},\mathsf{z}}&\big(\boldsymbol{y}|\boldsymbol{a},\boldsymbol{v},\boldsymbol{z};\theta^\star(\boldsymbol{z}),\Theta^\star\big)\cdot \\ &\exp\big(-[\theta_t(\boldsymbol{z}) + 2\Theta_{t,-t}^\top \boldsymbol{x}_{-t}]x_t - \Theta_{tt}\overline{x}_t\big). \end{aligned}$$

Then, we have

$$\begin{aligned} &u_{\mathsf{y}|\mathsf{a},\mathsf{v},\mathsf{z}}\big(\boldsymbol{y}|\boldsymbol{a},\boldsymbol{v},\boldsymbol{z};\theta_t(\boldsymbol{z}),\Theta_t\big) \\ &= \frac{f_{\mathsf{y}|\mathsf{a},\mathsf{v},\mathsf{z}}\big(\boldsymbol{y}|\boldsymbol{a},\boldsymbol{v},\boldsymbol{z};\theta^\star(\boldsymbol{z}),\Theta^\star\big)\exp\big(-[\theta_t(\boldsymbol{z})+2\Theta_{t,-t}^\top\boldsymbol{x}_{-t}]x_t-\Theta_{tt}\overline{x}_t\big)}{\int_{\boldsymbol{y}\in\mathcal{Y}^{p_y}} f_{\mathsf{y}|\mathsf{a},\mathsf{v},\mathsf{z}}\big(\boldsymbol{y}|\boldsymbol{a},\boldsymbol{v},\boldsymbol{z};\theta^\star(\boldsymbol{z}),\Theta^\star\big)\exp\big(-[\theta_t(\boldsymbol{z})+2\Theta_{t,-t}^\top\boldsymbol{x}_{-t}]x_t-\Theta_{tt}\overline{x}_t\big)d\boldsymbol{y}} \\ &= \frac{f_{\mathsf{y}|\mathsf{a},\mathsf{v},\mathsf{z}}\big(\boldsymbol{y}|\boldsymbol{a},\boldsymbol{v},\boldsymbol{z};\theta^\star(\boldsymbol{z}),\Theta^\star\big)\exp\big(-[\theta_t(\boldsymbol{z}) + 2\Theta_{t,-t}^\top\boldsymbol{x}_{-t}]x_t - \Theta_{tt}\overline{x}_t\big)}{\mathbb{E}_{\mathsf{y}|\mathsf{a},\mathsf{v},\mathsf{z}}\Big[\exp\big(-[\theta_t(\boldsymbol{z}) + 2\Theta_{t,-t}^\top\boldsymbol{x}_{-t}]x_t - \Theta_{tt}\overline{x}_t\big)\Big]}. \end{aligned} \tag{3.49}$$

For $\theta_t(\boldsymbol{z}) = \theta_t^\star(\boldsymbol{z})$, and $\Theta_t = \Theta_t^\star$, we can write an expression for $u_{\mathsf{y}|\mathsf{a},\mathsf{v},\mathsf{z}}\big(\boldsymbol{y}|\boldsymbol{a},\boldsymbol{v},\boldsymbol{z};\theta_t^\star(\boldsymbol{z}),\Theta_t^\star\big)$ which does not depend on y_t functionally. From Eqs. (3.10) and (3.48), we have

$$u_{\mathsf{y}|\mathsf{a},\mathsf{v},\mathsf{z}}\big(\boldsymbol{y}|\boldsymbol{a},\boldsymbol{v},\boldsymbol{z};\theta_t^\star(\boldsymbol{z}),\Theta_t^\star\big) \propto f_{\mathsf{y}_{-t}|\mathsf{a},\mathsf{v},\mathsf{z}}\big(\boldsymbol{y}_{-t}|\boldsymbol{a},\boldsymbol{v},\boldsymbol{z};\theta^\star(\boldsymbol{z}),\Theta^\star\big). \tag{3.50}$$

Now, consider the difference between the following KL divergences:

$$
\begin{aligned}
&\mathsf{KL}\left(u_{\mathsf{y|a,v,z}}\big(\,\cdot\,|\boldsymbol{a},\boldsymbol{v},\boldsymbol{z};\theta_t^\star(\boldsymbol{z}),\Theta_t^\star\big)\,\big\|u_{\mathsf{y|a,v,z}}\big(\,\cdot\,|\boldsymbol{a},\boldsymbol{v},\boldsymbol{z};\theta_t(\boldsymbol{z}),\Theta_t\big)\right)\\
&\qquad\qquad-\mathsf{KL}\left(u_{\mathsf{y|a,v,z}}\big(\,\cdot\,|\boldsymbol{a},\boldsymbol{v},\boldsymbol{z};\theta_t^\star(\boldsymbol{z}),\Theta_t^\star\big)\,\big\|f_{\mathsf{y|a,v,z}}\big(\,\cdot\,|\boldsymbol{a},\boldsymbol{v},\boldsymbol{z};\theta^\star(\boldsymbol{z}),\Theta^\star\big)\right)\\
&\overset{(a)}{=}\int_{\boldsymbol{y}\in\mathcal{Y}^{p_y}}u_{\mathsf{y|a,v,z}}\big(\boldsymbol{y}|\boldsymbol{a},\boldsymbol{v},\boldsymbol{z};\theta_t^\star(\boldsymbol{z}),\Theta_t^\star\big)\log\frac{f_{\mathsf{y|a,v,z}}\big(\boldsymbol{y}|\boldsymbol{a},\boldsymbol{v},\boldsymbol{z};\theta^\star(\boldsymbol{z}),\Theta^\star\big)}{u_{\mathsf{y|a,v,z}}\big(\boldsymbol{y}|\boldsymbol{a},\boldsymbol{v},\boldsymbol{z};\theta_t(\boldsymbol{z}),\Theta_t\big)}d\boldsymbol{y}\\
&\overset{(b)}{=}\int_{\boldsymbol{y}\in\mathcal{Y}^{p_y}}u_{\mathsf{y|a,v,z}}\big(\boldsymbol{y}|\boldsymbol{a},\boldsymbol{v},\boldsymbol{z};\theta_t^\star(\boldsymbol{z}),\Theta_t^\star\big)\log\frac{\mathbb{E}_{\mathsf{y|a,v,z}}\big[\exp\big(-[\theta_t(\boldsymbol{z})+2\Theta_{t,-t}^\top\boldsymbol{x}_{-t}]x_t-\Theta_{tt}\overline{x}_t\big)\big]}{\exp\big(-[\theta_t(\boldsymbol{z})+2\Theta_{t,-t}^\top\boldsymbol{x}_{-t}]x_t-\Theta_{tt}\overline{x}_t\big)}d\boldsymbol{y}\\
&=\log\mathbb{E}_{\mathsf{y|a,v,z}}\Big[\exp\big(-[\theta_t(\boldsymbol{z})+2\Theta_{t,-t}^\top\boldsymbol{x}_{-t}]x_t-\Theta_{tt}\overline{x}_t\big)\Big]\\
&\qquad\qquad-\int_{\boldsymbol{y}\in\mathcal{Y}^{p_y}}u_{\mathsf{y|a,v,z}}\big(\boldsymbol{y}|\boldsymbol{a},\boldsymbol{v},\boldsymbol{z};\theta_t^\star(\boldsymbol{z}),\Theta_t^\star\big)\big([\theta_t(\boldsymbol{z})+2\Theta_{t,-t}^\top\boldsymbol{x}_{-t}]x_t+\Theta_{tt}\overline{x}_t\big)d\boldsymbol{y}\\
&\overset{(c)}{=}\log\mathbb{E}_{\mathsf{y|a,v,z}}\Big[\exp\big(-[\theta_t(\boldsymbol{z})+2\Theta_{t,-t}^\top\boldsymbol{x}_{-t}]x_t-\Theta_{tt}\overline{x}_t\big)\Big],
\end{aligned}
\tag{3.51}
$$

where (a) follows from the definition of KL-divergence, (b) follows from Eq. (3.49), and (c) follows because integral is zero since $u_{\mathsf{y|a,v,z}}\big(\boldsymbol{y}|\boldsymbol{a},\boldsymbol{v},\boldsymbol{z};\theta_t^\star(\boldsymbol{z}),\Theta_t^\star\big)$ does not functionally depend on $y_t = x_t$ as in Eq. (3.50), and $\int_{x_t\in\mathcal{X}} x_t dx_t = 0$ and $\int_{x_t\in\mathcal{X}} \overline{x}_t dx_t = 0$. Now, from Eqs. (3.11) and (3.51) we can write

$$
\begin{aligned}
&\mathbb{E}_{\mathsf{y|a,v,z}}\big[\mathcal{L}(\underline{\Theta})\big]=\frac{1}{n}\sum_{t\in[p_y]}\sum_{i\in[n]}\exp\Bigg(\\
&\quad\mathsf{KL}\left(u_{\mathsf{y|a,v,z}}\big(\,\cdot\,|\boldsymbol{a}^{(i)},\boldsymbol{v}^{(i)},\boldsymbol{z}^{(i)};\theta_t^\star(\boldsymbol{z}^{(i)}),\Theta_t^\star\big)\,\big\|u_{\mathsf{y|a,v,z}}\big(\,\cdot\,|\boldsymbol{a}^{(i)},\boldsymbol{v}^{(i)},\boldsymbol{z}^{(i)};\theta_t(\boldsymbol{z}^{(i)}),\Theta_t\big)\right)\\
&\quad-\mathsf{KL}\left(u_{\mathsf{y|a,v,z}}\big(\,\cdot\,|\boldsymbol{a}^{(i)},\boldsymbol{v}^{(i)},\boldsymbol{z}^{(i)};\theta_t^\star(\boldsymbol{z}^{(i)}),\Theta_t^\star\big)\,\big\|f_{\mathsf{y|a,v,z}}\big(\,\cdot\,|\boldsymbol{a}^{(i)},\boldsymbol{v}^{(i)},\boldsymbol{z}^{(i)};\theta^\star(\boldsymbol{z}^{(i)}),\Theta^\star\big)\right)\Bigg).
\end{aligned}
\tag{3.52}
$$

We note that the parameters only show up in the first KL-divergence term in the right-hand-side of Eq. (3.52). Therefore, it is easy to see that $\mathbb{E}_{\mathsf{y|a,v,z}}\big[\mathcal{L}(\underline{\Theta})\big]$ is minimized uniquely when $\theta_t(\boldsymbol{z}^{(i)}) = \theta_t^\star(\boldsymbol{z}^{(i)})$ and $\Theta_t = \Theta_t^\star$ for all $t \in [p_y]$ and all $i \in [n]$, i.e., when $\underline{\Theta} = \underline{\Theta}^\star$.

3.A.2 Algorithm

In this section, we provide a projected gradient descent algorithm to return an ϵ-optimal estimate of the convex optimization in Eq. (3.12). We note that alternative algorithms (including Frank-Wolfe) can also be used.

We note that, in general, projecting onto the space $\Lambda_\theta^n \times \Lambda_\Theta$ may not be easy depending on the specific form of Λ_θ. For Examples 3.1 and 3.2, projecting on Λ_θ is equivalent to projecting onto the k-dimensional vector $\mathbf{a}$. For Example 3.2, the ℓ_0-sparsity is relaxed to ℓ_1 sparsity. We also do not focus on any issues that may arise due to the choice of the step size η.

```
Algorithm 1: Projected Gradient Descent
  Input: number of iterations $\tau$, step size $\eta$, $\epsilon$, parameter sets $\Lambda_\theta$ and $\Lambda_\Theta$
  Output: $\epsilon$-optimal estimate $\widehat{\underline{\Theta}}_\epsilon$
  Initialization: $\underline{\Theta}^{(0)} = \mathbf{0}$
  for $j = 0, \cdots, \tau$ do
      $\underline{\Theta}^{(j+1)} \leftarrow \arg\min_{\underline{\Theta} \in \Lambda_\theta^n \times \Lambda_\Theta} \|\underline{\Theta}^{(j)} - \eta \nabla \mathcal{L}(\underline{\Theta}^{(j)}) - \underline{\Theta}\|_2$
  $\widehat{\underline{\Theta}}_\epsilon \leftarrow \underline{\Theta}^{(\tau+1)}$
```

3.B Proof of Theorem 3.1 Part I: Recovering population-level parameter

To prove this part, it is sufficient to show that all points $\underline{\Theta} \in \Lambda_\Theta \times \Lambda_\theta^n$, such that $\|\Theta_t - \Theta_t^\star\|_2 \geq \varepsilon$ for at least one $t \in [p_y]$, uniformly satisfy

$$\mathcal{L}(\underline{\Theta}) \geq \mathcal{L}(\underline{\Theta}^\star) + \Omega(\varepsilon^2) \text{ for } n \geq \frac{ce^{c'\beta} p_y^2}{\varepsilon^4} \cdot \bigg(\log \frac{p_y}{\delta} + \mathcal{M}_\Theta\big(\varepsilon^2\big) + \mathcal{M}_{\theta,n}\big(\varepsilon^2\big) \bigg), \tag{3.53}$$

with probability at least $1-\delta$. Then, the guarantee in Theorem 3.1 follows from Eq. (3.12) by contraposition.

To that end, we decompose $\mathcal{L}(\underline{\Theta})$ in Eq. (3.11) as a sum of p_y convex (and positive) auxiliary objectives $\mathcal{L}_t\big(\underline{\Theta}_t\big)$, i.e., $\mathcal{L}(\underline{\Theta}) = \sum_{t \in [p_y]} \mathcal{L}_t\big(\underline{\Theta}_t\big)$ where

$$\mathcal{L}_t\big(\underline{\Theta}_t\big) \triangleq \frac{1}{n} \sum_{i \in [n]} \exp\Big(-[\theta_t^{(i)} + 2\Theta_{t,-t}^\top \boldsymbol{x}_{-t}^{(i)}] x_t^{(i)} - \Theta_{tt} \overline{x}_t^{(i)} \Big), \tag{3.54}$$

with $\overline{x}_t^{(i)} = \big[x_t^{(i)}\big]^2 - x_{\max}^2/3$ and $\underline{\Theta}_t = \big\{\theta_t^{(1)}, \cdots, \theta_t^{(n)}, \Theta_t\big\}$ as defined in Eq. (3.11). The lemma below, proven in Section 3.B.1, shows that for any fixed and feasible $\underline{\Theta}_t$, if Θ_t is far from $\Theta_t^\star$, then with high probability $\mathcal{L}_t\big(\underline{\Theta}_t\big)$ is significantly larger than $\mathcal{L}_t\big(\underline{\Theta}_t^\star\big)$. The lemma uses the following constants that depend on parameters $\tau \triangleq (\alpha, \beta, x_{\max})$:

$$C_{1,\tau} \triangleq \alpha + 2\beta x_{\max} \quad \text{and} \quad C_{2,\tau} \triangleq \exp\left(x_{\max}(\alpha + 2\beta x_{\max})\right). \tag{3.55}$$

Lemma 3.1 (Gap between the loss function for a fixed parameter)**.** *Consider any $\underline{\Theta} \in \Lambda_\theta^n \times \Lambda_\Theta$. Fix any $\delta \in (0,1)$. Then, we have uniformly for all $t \in [p_y]$*

$$\mathcal{L}_t\big(\underline{\Theta}_t\big) \geq \mathcal{L}_t\big(\underline{\Theta}_t^\star\big) + \frac{\lambda_{\min}\big\|\Theta_t - \Theta_t^\star\big\|_2^2}{2C_{2,\tau}} - \varepsilon \quad \textit{for} \quad n \geq \frac{ce^{c'\beta}\log(p_y/\delta)}{\varepsilon^2},$$

with probability at least $1-\delta$, where $C_{2,\tau}$ was defined in Eq. (3.55).

Next, we show that the loss function $\mathcal{L}$ is Lipschitz (see Section 3.B.2 for the proof).

Lemma 3.2 (Lipschitzness of the loss function)**.** *Consider any $\underline{\Theta}, \underline{\widetilde{\Theta}} \in \Lambda_{\underline{\Theta}}$. Then, the loss function $\mathcal{L}$ is $2x_{\max}^2 C_{2,\tau}$-Lipschitz in a suitably-adjusted ℓ_1 norm:*

$$\left|\mathcal{L}(\underline{\widetilde{\Theta}}) - \mathcal{L}(\underline{\Theta})\right| \leq 2x_{\max}^2 C_{2,\tau}\Big(\sum_{t\in[p_y]} \|\widetilde{\Theta}_t - \Theta_t\|_1 + \frac{1}{n}\sum_{i\in[n]} \|\widetilde{\theta}^{(i)} - \theta^{(i)}\|_1\Big), \tag{3.56}$$

where the constant $C_{2,\tau}$ was defined in Eq. (3.55).

Given these lemmas, we now proceed with the proof.

Proof strategy. We want to show that all points $\underline{\Theta} \in \Lambda_\Theta \times \Lambda_\theta^n$, such that $\|\Theta_t - \Theta_t^\star\|_2 \geq \varepsilon$ for at least one $t \in [p_y]$, uniformly satisfy Eq. (3.53) with probability at least $1-\delta$. To do so, we consider the set of feasible $\underline{\Theta}$ such that the distance of Θ_t from $\Theta_t^\star$ is at least $\varepsilon > 0$ in ℓ_2 norm for some $t \in [p_y]$, and denote the set by $\Lambda_\Theta^\varepsilon \times \Lambda_\theta^n$ (see Eq. (3.57) and Eq. (3.7)). Then, using an appropriate covering set of $\Lambda_\Theta^\varepsilon \times \Lambda_\theta^n$ and the Lipschitzness of $\mathcal{L}$, we show that the value of $\mathcal{L}$ at all points in $\Lambda_\Theta^\varepsilon \times \Lambda_\theta^n$ is uniformly $\Omega(\varepsilon^2)$ larger than the value of $\mathcal{L}$ at $\underline{\Theta}^\star$ with high probability.

Arguments for points in the covering set. Define the set

$$\Lambda_\Theta^\varepsilon \triangleq \Big\{\Theta = [\Phi^{(y,y)}, \Phi^{(y,a)}, \Phi^{(y,v)}] \in \mathbb{R}^{p_y \times \widetilde{p}} : \Phi^{(y,y)} = \Phi^{(y,y)^\top},$$
$$\|\Theta\|_{\max} \leq \alpha, \|\Theta\|_\infty \leq \beta, \max_{t\in[p_y]} \|\Theta_t^\star - \Theta_t\|_2 \geq \varepsilon\Big\}. \tag{3.57}$$

Let $\mathcal{U}(\Lambda_\Theta^\varepsilon, \varepsilon')$ be the ε'-cover of smallest size for the set $\Lambda_\Theta^\varepsilon$ with respect to $\|\cdot\|_1$ (see Definition 3.2) and let $\mathcal{C}(\Lambda_\Theta^\varepsilon, \varepsilon') = |\mathcal{U}(\Lambda_\Theta^\varepsilon, \varepsilon')|$ be the ε'-covering number. Similarly, let $\mathcal{U}(\Lambda_\theta^\varepsilon, \varepsilon'')$ be the ε''-cover of the smallest size for the set Λ_θ^n with respect to $\|\cdot\|_1$ and let $\mathcal{C}(\Lambda_\theta^n, \varepsilon'') = |\mathcal{U}(\Lambda_\theta^\varepsilon, \varepsilon'')|$ be the ε''-covering number. We choose

$$\varepsilon' \triangleq \frac{\lambda_{\min}\varepsilon^2}{32x_{\max}^2 C_{2,\tau}^2} \quad \text{and} \quad \varepsilon'' \triangleq \frac{\lambda_{\min}\varepsilon^2 n}{32x_{\max}^2 C_{2,\tau}^2}. \tag{3.58}$$

Now, we argue by a union bound that the value of $\mathcal{L}$ at all points in $\mathcal{U}(\Lambda_\Theta^\varepsilon, \varepsilon') \times \mathcal{U}(\Lambda_\theta^n, \varepsilon'')$ is uniformly $\Omega(\varepsilon^2)$ larger than $\mathcal{L}(\underline{\Theta}^\star)$ with high probability. For any $\underline{\Theta} \in \mathcal{U}(\Lambda_\Theta^\varepsilon, \varepsilon') \times \mathcal{U}(\Lambda_\theta^n, \varepsilon'')$, we have

$$\sum_{t\in[p_y]} \|\Theta_t^\star - \Theta_t\|_2^2 \overset{(a)}{\geq} \varepsilon^2, \tag{3.59}$$

where (a) follows because $\mathcal{U}(\Lambda_\Theta^\varepsilon, \varepsilon') \subseteq \Lambda_\Theta^\varepsilon$. Now, applying Lemma 3.1 with $\varepsilon \hookleftarrow \lambda_{\min}\varepsilon^2/4C_{2,\tau}p_y$ and $\delta \hookleftarrow \delta/(\mathcal{C}(\Lambda_\Theta^\varepsilon, \varepsilon') + \mathcal{C}(\Lambda_\theta^n, \varepsilon''))$ and summing over $t \in [p_y]$, we find that

$$\sum_{t\in[p_y]} \mathcal{L}_t(\underline{\Theta}_t) \geq \sum_{t\in[p_y]} \left(\mathcal{L}_t(\underline{\Theta}_t^\star) + \frac{\lambda_{\min}\|\Theta_t - \Theta_t^\star\|_2^2}{2C_{2,\tau}} - \frac{\lambda_{\min}\varepsilon^2}{4C_{2,\tau}p_y}\right)$$

$$\begin{aligned}\implies \qquad \mathcal{L}(\underline{\Theta}) &\ge \mathcal{L}(\underline{\Theta}^\star) + \frac{\lambda_{\min}}{2C_{2,\tau}} \sum_{t\in[p_y]} \|\Theta_t^\star - \Theta_t\|_2^2 - \frac{\lambda_{\min}\varepsilon^2}{4C_{2,\tau}} \\ &\overset{Eq.\ (3.59)}{\ge} \mathcal{L}(\underline{\Theta}^\star) + \frac{\lambda_{\min}\varepsilon^2}{4C_{2,\tau}},\end{aligned}$$

with probability at least $1 - \delta/(\mathcal{C}(\Lambda_\Theta^\varepsilon, \varepsilon') + \mathcal{C}(\Lambda_\theta^n, \varepsilon''))$ whenever

$$n \ge \frac{ce^{c'\beta} p_y^2 \log\left((\mathcal{C}(\Lambda_\Theta^\varepsilon, \varepsilon') \times \mathcal{C}(\Lambda_\theta^n, \varepsilon'')) \cdot p_y/\delta\right)}{\lambda_{\min}^2 \varepsilon^4}. \tag{3.60}$$

By applying the union bound over $\mathcal{U}(\Lambda_\Theta^\varepsilon, \varepsilon') \times \mathcal{U}(\Lambda_\theta^n, \varepsilon'')$, as long as n satisfies Eq. (3.60), we have

$$\mathcal{L}(\underline{\Theta}) \ge \mathcal{L}(\underline{\Theta}^\star) + \frac{\lambda_{\min}\varepsilon^2}{4C_{2,\tau}} \text{ uniformly for every } \underline{\Theta} \in \mathcal{U}(\Lambda_\Theta^\varepsilon, \varepsilon') \times \mathcal{U}(\Lambda_\theta^n, \varepsilon''), \tag{3.61}$$

with probability at least $1 - \delta$.

Arguments for points outside the covering set. Now, we establish the claim Eq. (3.53) for an arbitrary $\underline{\widetilde{\Theta}} \in \Lambda_\Theta^\varepsilon \times \Lambda_\theta^n$ conditional on the event that Eq. (3.61) holds. Given a fixed $\underline{\widetilde{\Theta}} \in \Lambda_\Theta^\varepsilon \times \Lambda_\theta^n$, let $\underline{\Theta}$ be (one of) the point(s) in the cover $\mathcal{U}(\Lambda_\Theta^\varepsilon, \varepsilon') \times \mathcal{U}(\Lambda_\theta^n, \varepsilon'')$ that satisfies $\sum_{t\in[p_y]} \|\widetilde{\Theta}_t - \Theta_t\|_1 \le \varepsilon'$ and $\sum_{i\in[n]} \|\widetilde{\theta}^{(i)} - \theta^{(i)}\|_1 \le \varepsilon''$ (there exists such a point by Definition 3.2). Then, the choices Eq. (3.58) and Lemma 3.2 put together imply that

$$\begin{aligned}\mathcal{L}(\underline{\widetilde{\Theta}}) &\ge \mathcal{L}(\underline{\Theta}) - 2x_{\max}^2 C_{2,\tau}\Big(\sum_{t\in[p_y]} \|\widetilde{\Theta}_t - \Theta_t\|_1 + \frac{1}{n}\sum_{i\in[n]} \|\widetilde{\theta}^{(i)} - \theta^{(i)}\|_1\Big) \\ &\ge \mathcal{L}(\underline{\Theta}) - 2x_{\max}^2 C_{2,\tau}\Big(\varepsilon' + \frac{\varepsilon''}{n}\Big) \overset{Eq.\ (3.58)}{\ge} \mathcal{L}(\underline{\Theta}) - \frac{\lambda_{\min}\varepsilon^2}{8C_{2,\tau}} \overset{Eq.\ (3.61)}{\ge} \mathcal{L}(\underline{\Theta}^\star) + \frac{\lambda_{\min}\varepsilon^2}{8C_{2,\tau}}.\end{aligned}$$

Bounding n. Using $\Lambda_\Theta^\varepsilon \subseteq \Lambda_\Theta$ and the outer product definition of θ^n, we find that

$$\mathcal{C}(\Lambda_\Theta^\varepsilon, \varepsilon') \le \mathcal{C}(\Lambda_\Theta, \varepsilon') \quad \text{and} \quad \mathcal{C}(\Lambda_\theta^n, \varepsilon'') = (\mathcal{C}(\Lambda_\theta, \varepsilon''))^n. \tag{3.62}$$

Putting together Eqs. (3.58) and (3.62), the lower bound Eq. (3.60) can be replaced by

$$n \ge \frac{ce^{c'\beta} p_y^2}{\lambda_{\min}^2 \varepsilon^4} \cdot \left(\log\frac{p_y}{\delta} + \log\mathcal{C}\Big(\Lambda_\Theta, \frac{\lambda_{\min}\varepsilon^2}{ce^{c'\beta}}\Big) + n\log\mathcal{C}\Big(\Lambda_\theta, \frac{\lambda_{\min} n\varepsilon^2}{ce^{c'\beta}}\Big)\right),$$

which yields the claim immediately after noting that

$$\log\mathcal{C}\Big(\Lambda_\Theta, \frac{\lambda_{\min}\varepsilon^2}{ce^{c'\beta}}\Big) = \mathcal{M}_\Theta\Big(\frac{\lambda_{\min} n\varepsilon^2}{ce^{c'\beta}}\Big) \text{ and } \log\mathcal{C}\Big(\Lambda_\theta, \frac{\lambda_{\min} n\varepsilon^2}{ce^{c'\beta}}\Big) = \mathcal{M}_\theta\Big(\frac{\lambda_{\min} n\varepsilon^2}{ce^{c'\beta}}\Big).$$

3.B.1 Proof of Lemma 3.1: Gap between the loss function for a fixed parameter

Fix any $\varepsilon > 0$, any $\delta \in (0,1)$, and $t \in [p_y]$. Consider any direction $\underline{\Omega}_t \triangleq \big\{\omega_t^{(1)}, \cdots, \omega_t^{(n)}, \Omega_t\big\} \in \mathbb{R}^{n+\widetilde{p}}$ along the parameter $\underline{\Theta}_t$, i.e.,

$$\underline{\Omega}_t = \underline{\Theta}_t - \underline{\Theta}_t^\star, \quad \text{and} \quad \Omega_t = \Theta_t - \Theta_t^\star. \tag{3.63}$$

We denote the first-order and the second-order directional derivatives of the loss function $\mathcal{L}_t$ in Eq. (3.54) along the direction $\underline{\Omega}_t$ evaluated at $\underline{\Theta}_t$ by $\partial_{\underline{\Omega}_t}\mathcal{L}_t(\underline{\Theta}_t)$ and $\partial^2_{\underline{\Omega}_t^2}\mathcal{L}_t(\underline{\Theta}_t)$, respectively. Below, we state a lemma (with proof divided across Section 3.B.1.1 and Section 3.B.1.2) that provides us a control on $\partial_{\underline{\Omega}_t}\mathcal{L}_t(\underline{\Theta}_t)$ and $\partial^2_{\underline{\Omega}_t^2}\mathcal{L}_t(\underline{\Theta}_t)$. The assumptions of Lemma 3.1 remain in force.

Lemma 3.3 (Control on first and second directional derivatives)**.** *For any fixed $\varepsilon_1, \varepsilon_2 > 0$, $\delta_1, \delta_2 \in (0,1)$, $t \in [p_y]$, $\underline{\Theta} \in \Lambda_\theta^n \times \Lambda_\Theta$ defined in Eq. (3.11) and Ω_t defined in Eq. (3.63), we have the following:*

(a) Concentration of first directional derivative*: with probability at least $1-\delta_1$,*

$$\big|\partial_{\underline{\Omega}_t}\mathcal{L}_t(\underline{\Theta}_t^\star)\big| \leq \varepsilon_1 \textit{ for } n \geq \frac{8C_{1,\tau}^2 C_{2,\tau}^2 x_{\max}^2 \log\frac{2p_y}{\delta_1}}{\varepsilon_1^2} \textit{ and uniformly for all } t \in [p_y].$$

(b) Anti-concentration of second directional derivative*: with probability at least $1-\delta_2$,*

$$\partial^2_{\underline{\Omega}_t^2}\mathcal{L}_t(\underline{\Theta}_t) \geq \frac{\lambda_{\min}\big\|\Omega_t\big\|_2^2}{C_{2,\tau}} - \varepsilon_2 \textit{ for } n \geq \frac{32C_{1,\tau}^4 x_{\max}^4 \log\frac{2p_y}{\delta_2}}{\varepsilon_2^2 C_{2,\tau}^2} \textit{ and uniformly for all } t \in [p_y].$$

Given this lemma, we now proceed with the proof. Define a function $g : [0,1] \to \mathbb{R}^{n+\widetilde{p}}$

$$g(a) \triangleq \underline{\Theta}_t^\star + a(\underline{\Theta}_t - \underline{\Theta}_t^\star).$$

Notice that $g(0) = \underline{\Theta}_t^\star$ and $g(1) = \underline{\Theta}_t$ as well as

$$\frac{d\mathcal{L}_t(g(a))}{da} = \partial_{\underline{\Omega}_t}\mathcal{L}_t(\widetilde{\underline{\Theta}}_t)\big|_{\widetilde{\underline{\Theta}}_t = g(a)} \quad \text{and} \quad \frac{d^2\mathcal{L}_t(g(a))}{da^2} = \partial^2_{\underline{\Omega}_t^2}\mathcal{L}_t(\widetilde{\underline{\Theta}}_t)\big|_{\widetilde{\underline{\Theta}}_t = g(a)}. \tag{3.64}$$

By the fundamental theorem of calculus, we have

$$\frac{d\mathcal{L}_t(g(a))}{da} \geq \frac{d\mathcal{L}_t(g(a))}{da}\Big|_{a=0} + a \min_{a \in (0,1)} \frac{d^2\mathcal{L}_t(g(a))}{da^2}. \tag{3.65}$$

Integrating both sides of Eq. (3.65) with respect to a, we obtain

$$\begin{aligned}\mathcal{L}_t(g(a)) - \mathcal{L}_t(g(0)) &\geq a\frac{d\mathcal{L}_t(g(a))}{da}\Big|_{a=0} + \frac{a^2}{2}\min_{a \in (0,1)} \frac{d^2\mathcal{L}_t(g(a))}{da^2} \\ &\overset{Eq.\ (3.64)}{=} a\partial_{\underline{\Omega}_t}\mathcal{L}_t(\widetilde{\underline{\Theta}}_t)\big|_{\widetilde{\underline{\Theta}}_t = g(0)} + \frac{a^2}{2}\min_{a \in (0,1)} \partial^2_{\underline{\Omega}_t^2}\mathcal{L}_t(\widetilde{\underline{\Theta}}_t)\big|_{\widetilde{\underline{\Theta}}_t = g(a)}\end{aligned}$$

$$\begin{aligned}&\overset{(a)}{=} a\partial_{\underline{\Omega}_t}\mathcal{L}_t(\underline{\Theta}_t^\star) + \frac{a^2}{2}\min_{a\in(0,1)} \partial^2_{\underline{\Omega}_t^2}\mathcal{L}_t(\widetilde{\underline{\Theta}}_t)\big|_{\widetilde{\underline{\Theta}}_t=g(a)}\\ &\overset{(b)}{\geq} -a\big|\partial_{\underline{\Omega}_t}\mathcal{L}_t(\underline{\Theta}_t^\star)\big| + \frac{a^2}{2}\min_{a\in(0,1)} \partial^2_{\underline{\Omega}_t^2}\mathcal{L}_t(\widetilde{\underline{\Theta}}_t)\big|_{\widetilde{\underline{\Theta}}_t=g(a)},\end{aligned} \tag{3.66}$$

where (a) follows because $g(0)=\underline{\Theta}_t^\star$ and (b) follows by the triangle inequality. Plugging in $a=1$ in Eq. (3.66) as well as using $g(0)=\underline{\Theta}_t^\star$ and $g(1)=\underline{\Theta}_t$, we find that

$$\mathcal{L}_t(\underline{\Theta}_t) - \mathcal{L}_t(\underline{\Theta}_t^\star) \geq -\big|\partial_{\underline{\Omega}_t}\mathcal{L}_t(\underline{\Theta}_t^\star)\big| + \frac{1}{2}\min_{a\in(0,1)} \partial^2_{\underline{\Omega}_t^2}\mathcal{L}_t(\widetilde{\underline{\Theta}}_t)\big|_{\widetilde{\underline{\Theta}}_t=g(a)}.$$

Now, we use Lemma 3.3 with

$$\varepsilon_1 \mapsfrom \frac{\varepsilon}{2}, \quad \delta_1 \mapsfrom \frac{\delta}{2}, \quad \varepsilon_2 \mapsfrom \varepsilon, \quad \text{and} \quad \delta_2 \mapsfrom \frac{\delta}{2}.$$

Thus for $n \geq \dfrac{ce^{c'\beta}\log(p_y/\delta)}{\varepsilon^2}$, we have

$$\mathcal{L}_t(\underline{\Theta}_t) - \mathcal{L}_t(\underline{\Theta}_t^\star) \geq -\frac{\varepsilon}{2} + \frac{1}{2}\bigg(\frac{\lambda_{\min}\big\|\underline{\Omega}_t\big\|_2^2}{C_{2,\tau}} - \varepsilon\bigg) = \frac{\lambda_{\min}\big\|\underline{\Omega}_t\big\|_2^2}{2C_{2,\tau}} - \varepsilon,$$

uniformly for all $t\in[p_y]$, with probability at least $1-\delta$.

3.B.1.1 Proof of Lemma 3.3(a): Concentration of first directional derivative

For every $t\in[p_y]$ with $\underline{\Omega}_t$ defined in Eq. (3.63), we claim that the first-order directional derivative of the loss function defined in Eq. (3.54) is given by

$$\partial_{\underline{\Omega}_t}\mathcal{L}_t(\underline{\Theta}_t) = -\frac{1}{n}\sum_{i\in[n]}\Big([\Delta_t^{(i)}]^\top\widetilde{\boldsymbol{x}}^{(i)}\Big)\exp\Big(-[\theta_t^{(i)} + 2\Theta_{t,-t}^\top\boldsymbol{x}_{-t}^{(i)}]x_t^{(i)} - \Theta_{tt}\overline{x}_t^{(i)}\Big), \tag{3.67}$$

where $\Delta_t^{(i)} \triangleq \begin{bmatrix}\omega_t^{(i)}\\ \Omega_{t,-t}^\top\\ \Omega_{tt}\end{bmatrix} \in \mathbb{R}^{\widetilde{p}+1}$ and $\widetilde{\boldsymbol{x}}^{(i)} \triangleq \begin{bmatrix}x_t^{(i)}\\ 2\boldsymbol{x}_{-t}^{(i)}x_t^{(i)}\\ \overline{x}_t^{(i)}\end{bmatrix} \in \mathbb{R}^{\widetilde{p}+1}$ for all $i\in[n]$ with $\overline{x}_t^{(i)} = \big[x_t^{(i)}\big]^2 - x_{\max}^2/3$. We provide a proof at the end.

Next, we claim that the mean of the first-order directional derivative evaluated at the true parameter is zero. We provide a proof at the end.

Lemma 3.4 (Zero-meanness of first directional derivative). *For every $t\in[p_y]$ with $\underline{\Omega}_t$ defined in Eq. (3.63), we have $\mathbb{E}\big[\partial_{\underline{\Omega}_t}\mathcal{L}_t(\underline{\Theta}_t^\star)\big]=0$.*

Given these, we proceed to show the concentration of the first-order directional derivative evaluated at the true parameter. Fix any $t\in[p_y]$. From Eq. (3.67), we have

$$\partial_{\underline{\Omega}_t}\mathcal{L}_t(\underline{\Theta}_t^\star) \overset{Eq.\ (3.67)}{=} -\frac{1}{n}\sum_{i\in[n]}\Big([\Delta_t^{(i)}]^\top\widetilde{\boldsymbol{x}}^{(i)}\Big)\exp\Big(-[\theta_t^{\star(i)} + 2\Theta_{t,-t}^{\star\top}\boldsymbol{x}_{-t}^{(i)}]x_t^{(i)} - \Theta_{tt}^\star\overline{x}_t^{(i)}\Big).$$

Each term in the above summation is an independent random variable and is bounded as follows

$$\begin{aligned}
&\Big|\Big([\Delta_t^{(i)}]^\top \widetilde{\boldsymbol{x}}^{(i)}\Big) \times \exp\Big(- [\theta_t^{\star(i)} + 2\Theta_{t,-t}^{\star\top}\boldsymbol{x}_{-t}^{(i)}]x_t^{(i)} - \Theta_{tt}^{\star}\overline{x}_t^{(i)}\Big)\Big| \\
&\overset{(a)}{=} \Big|\Big(\omega_t^{(i)}x_t^{(i)} + 2\Omega_{t,-t}^\top\boldsymbol{x}_{-t}^{(i)}x_t^{(i)} + \Omega_{tt}\overline{x}_t^{(i)}\Big) \times \exp\Big(- [\theta_t^{\star(i)} + 2\Theta_{t,-t}^{\star\top}\boldsymbol{x}_{-t}^{(i)}]x_t^{(i)} - \Theta_{tt}^{\star}\overline{x}_t^{(i)}\Big)\Big| \\
&\overset{(b)}{\leq} \big||\omega_t^{(i)}| + 2\|\Omega_t\|_1\|\boldsymbol{x}^{(i)}\|_\infty\big| \times x_{\max} \times \exp\Big(\big(|\theta_t^{\star(i)}| + 2\|\Theta_t^\star\|_1\|\boldsymbol{x}^{(i)}\|_\infty\big)x_{\max}\Big) \\
&\overset{(c)}{\leq} \big(2\alpha + 4\beta x_{\max}\big) \times x_{\max} \times \exp\Big(\big(\alpha + 2\beta x_{\max}\big)x_{\max}\Big) \overset{Eq.\ (3.55)}{=} 2C_{1,\tau}C_{2,\tau}x_{\max},
\end{aligned}$$

where (a) follows by plugging in $\Delta_t^{(i)}$ and $\widetilde{\boldsymbol{x}}^{(i)}$, (b) follows from triangle inequality, Cauchy–Schwarz inequality, and because $\|\boldsymbol{x}^{(i)}\|_\infty \leq x_{\max}$ for all $i \in [n]$, and (c) follows because $\theta^{\star(i)} \in \Lambda_\theta$ for all $i \in [n]$, $\Theta^\star \in \Lambda_\Theta$, $\omega^{(i)} \in 2\Lambda_\theta$ for all $i \in [n]$, $\Omega \in 2\Lambda_\Theta$, and $\|\boldsymbol{x}^{(i)}\|_\infty \leq x_{\max}$ for all $i \in [n]$.

Further, from Lemma 3.4, we have $\mathbb{E}\big[\partial_{\underline{\Omega}_t}\mathcal{L}_t(\underline{\Theta}_t^\star)\big] = 0$. Therefore, using the Hoeffding's inequality results in

$$\mathbb{P}\Big(\big|\partial_{\underline{\Omega}_t}\mathcal{L}_t(\underline{\Theta}_t^\star)\big| > \varepsilon_1\Big) < 2\exp\Big(- \frac{n\varepsilon_1^2}{8C_{1,\tau}^2C_{2,\tau}^2x_{\max}^2}\Big).$$

The proof follows by using the union bound over all $t \in [p_y]$.

3.B.1.1.1 Proof of Eq. (3.67): Expression for first directional derivative

Fix any $t \in [p_y]$. The first-order partial derivatives of $\mathcal{L}_t$ with respect to entries of $\underline{\Theta}_t$ defined in Eq. (3.54) are given by

$$\frac{\partial \mathcal{L}_t(\underline{\Theta}_t)}{\partial \theta_t^{(i)}} = \frac{-1}{n}x_t^{(i)}\exp\Big(-[\theta_t^{(i)} + 2\Theta_{t,-t}^\top\boldsymbol{x}_{-t}^{(i)}]x_t^{(i)} - \Theta_{tt}\overline{x}_t^{(i)}\Big) \text{ for all } i \in [n], \quad \text{and}$$

$$\frac{\partial \mathcal{L}_t(\underline{\Theta}_t)}{\partial \Theta_{tu}} = \begin{cases} \frac{-2}{n}\sum_{i\in[n]} x_t^{(i)}x_u^{(i)}\exp\Big(-[\theta_t^{(i)} + 2\Theta_{t,-t}^\top\boldsymbol{x}_{-t}^{(i)}]x_t^{(i)} - \Theta_{tt}\overline{x}_t^{(i)}\Big) \text{ for all } u \in [\widetilde{p}]\backslash\{t\}. \\ \frac{-1}{n}\sum_{i\in[n]} \overline{x}_t^{(i)}\exp\Big(-[\theta_t^{(i)} + 2\Theta_{t,-t}^\top\boldsymbol{x}_{-t}^{(i)}]x_t^{(i)} - \Theta_{tt}\overline{x}_t^{(i)}\Big) \text{ for } u = t. \end{cases}$$

Now, we can write the first-order directional derivative of $\mathcal{L}_t$ as

$$\begin{aligned}
\partial_{\underline{\Omega}_t}\mathcal{L}_t(\underline{\Theta}_t) &\triangleq \lim_{h\to 0}\frac{\mathcal{L}_t(\underline{\Theta}_t + h\underline{\Omega}_t) - \mathcal{L}_t(\underline{\Theta}_t)}{h} = \sum_{i\in[n]}\omega_t^{(i)}\frac{\partial \mathcal{L}_t(\underline{\Theta}_t)}{\partial \theta_t^{(i)}} + \sum_{u\in[\widetilde{p}]}\Omega_{tu}\frac{\partial \mathcal{L}_t(\underline{\Theta}_t)}{\partial \Theta_{tu}} \\
&= \frac{-1}{n}\sum_{i\in[n]}\Big(\omega_t^{(i)}x_t^{(i)} + 2\sum_{u\in[\widetilde{p}]\backslash\{t\}}\Omega_{tu}x_t^{(i)}x_u^{(i)} + \Omega_{tt}\overline{x}_t^{(i)}\Big)\exp\Big(-[\theta_t^{(i)} + 2\Theta_{t,-t}^\top\boldsymbol{x}_{-t}^{(i)}]x_t^{(i)} - \Theta_{tt}\overline{x}_t^{(i)}\Big) \\
&= \frac{-1}{n}\sum_{i\in[n]}\Big(\omega_t^{(i)}x_t^{(i)} + 2\Omega_{t,-t}^\top\boldsymbol{x}_{-t}^{(i)}x_t^{(i)} + \Omega_{tt}\overline{x}_t^{(i)}\Big)\exp\Big(-[\theta_t^{(i)} + 2\Theta_{t,-t}^\top\boldsymbol{x}_{-t}^{(i)}]x_t^{(i)} - \Theta_{tt}\overline{x}_t^{(i)}\Big) \\
&\overset{(a)}{=} \frac{-1}{n}\sum_{i\in[n]}\Big([\Delta_t^{(i)}]^\top\widetilde{\boldsymbol{x}}^{(i)}\Big)\exp\Big(-[\theta_t^{(i)} + 2\Theta_{t,-t}^\top\boldsymbol{x}_{-t}^{(i)}]x_t^{(i)} - \Theta_{tt}\overline{x}_t^{(i)}\Big),
\end{aligned}$$

where (a) follows from the definitions of $\Delta_t^{(i)}$ and $\widetilde{\boldsymbol{x}}^{(i)}$.

3.B.1.1.2 Proof of Lemma 3.4: Zero-meanness of first directional derivative

Fix any $t \in [p_y]$. From Eq. (3.67), we have

$$
\begin{aligned}
&\mathbb{E}\big[\partial_{\underline{\Omega}_t} \mathcal{L}_t(\underline{\Theta}_t^\star)\big] \\
&\overset{Eq.~(3.67)}{=} -\frac{1}{n} \sum_{i \in [n]} \mathbb{E}_{\boldsymbol{x}^{(i)}, \boldsymbol{z}^{(i)}} \bigg[\Big([\Delta_t^{(i)}]^\top \widetilde{\boldsymbol{x}}^{(i)} \Big) \exp\Big(- [\theta_t^{\star(i)} + 2\Theta_{t,-t}^{\star\top} \boldsymbol{x}_{-t}^{(i)}] x_t^{(i)} - \Theta_{tt}^\star \overline{x}_t^{(i)} \Big) \bigg] \\
&\overset{(a)}{=} -\frac{1}{n} \sum_{i \in [n]} \sum_{u \in [\widetilde{p}+1]} \mathbb{E}_{\boldsymbol{a}^{(i)}, \boldsymbol{v}^{(i)}, \boldsymbol{z}^{(i)}} \bigg[\Delta_{tu}^{(i)} \cdot \\
&\qquad\qquad \mathbb{E}_{\boldsymbol{y}^{(i)} | \boldsymbol{a}^{(i)}, \boldsymbol{v}^{(i)}, \boldsymbol{z}^{(i)}} \Big[\widetilde{\boldsymbol{x}}_u^{(i)} \exp\Big(- [\theta_t^\star(\boldsymbol{z}^{(i)}) + 2\Theta_{t,-t}^{\star\top} \boldsymbol{x}_{-t}^{(i)}] x_t^{(i)} - \Theta_{tt}^\star \overline{x}_t^{(i)} \Big) \Big] \bigg],
\end{aligned}
$$

where (a) follows by linearity of expectation and by plugging in $\theta_t^{\star(i)} = \theta_t^\star(\boldsymbol{z}^{(i)})$. Now to complete the proof, we show that for any $i \in [n], u \in [\widetilde{p}+1]$, $\boldsymbol{a}^{(i)} \in \mathcal{A}^{p_z}$, $\boldsymbol{v}^{(i)} \in \mathcal{V}^{p_z}$, and $\boldsymbol{z}^{(i)} \in \mathcal{Z}^{p_z}$, we have

$$
\mathbb{E}_{\boldsymbol{y}^{(i)} | \boldsymbol{a}^{(i)}, \boldsymbol{v}^{(i)}, \boldsymbol{z}^{(i)}} \Big[\widetilde{\boldsymbol{x}}_u^{(i)} \exp\Big(- [\theta_t^\star(\boldsymbol{z}^{(i)}) + 2\Theta_{t,-t}^{\star\top} \boldsymbol{x}_{-t}^{(i)}] x_t^{(i)} - \Theta_{tt}^\star \overline{x}_t^{(i)} \Big) \Big] = 0.
$$

Fix any $i \in [n], u \in [\widetilde{p}+1]$, $\boldsymbol{a}^{(i)} \in \mathcal{A}^{p_z}$, $\boldsymbol{v}^{(i)} \in \mathcal{V}^{p_z}$, and $\boldsymbol{z}^{(i)} \in \mathcal{Z}^{p_z}$. We have

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{y}^{(i)} | \boldsymbol{a}^{(i)}, \boldsymbol{v}^{(i)}, \boldsymbol{z}^{(i)}} \Big[\widetilde{\boldsymbol{x}}_u^{(i)} \exp\Big(- [\theta_t^\star(\boldsymbol{z}^{(i)}) + 2\Theta_{t,-t}^{\star\top} \boldsymbol{x}_{-t}^{(i)}] x_t^{(i)} - \Theta_{tt}^\star \overline{x}_t^{(i)} \Big) \Big] \\
&= \int_{\mathcal{X}^p} \widetilde{x}_u^{(i)} \exp\Big(- [\theta_t^\star(\boldsymbol{z}^{(i)}) + 2\Theta_{t,-t}^{\star\top} \boldsymbol{x}_{-t}^{(i)}] x_t^{(i)} - \Theta_{tt}^\star \overline{x}_t^{(i)} \Big) f_{\mathbf{y}|\mathbf{a},\mathbf{v},\mathbf{z}}\big(\boldsymbol{y}^{(i)} | \boldsymbol{a}^{(i)}, \boldsymbol{v}^{(i)}, \boldsymbol{z}^{(i)}\big) d\boldsymbol{y}^{(i)} \\
&= \int_{\mathcal{X}^p} \widetilde{x}_u^{(i)} \exp\Big(-[\theta_t^\star(\boldsymbol{z}^{(i)}) + 2\Theta_{t,-t}^{\star\top} \boldsymbol{x}_{-t}^{(i)}] x_t^{(i)} - \Theta_{tt}^\star \overline{x}_t^{(i)} \Big) f_{\mathbf{y}_{-t}|\mathbf{a},\mathbf{v},\mathbf{z}}\big(\boldsymbol{y}_{-t}^{(i)} | \boldsymbol{a}^{(i)}, \boldsymbol{v}^{(i)}, \boldsymbol{z}^{(i)}\big) \times \\
&\qquad\qquad f_{\mathsf{x}_t|\mathbf{x}_{-t},\mathbf{z}}\big(x_t^{(i)} | \boldsymbol{x}_{-t}^{(i)}, \boldsymbol{z}^{(i)}; \theta_t^\star(\boldsymbol{z}^{(i)}), \Theta_t^\star\big) d\boldsymbol{y}^{(i)} \\
&\overset{(a)}{=} \int_{\mathcal{X}^p} \frac{\widetilde{x}_u^{(i)} f_{\mathbf{y}_{-t}|\mathbf{a},\mathbf{v},\mathbf{z}}\big(\boldsymbol{y}_{-t}^{(i)} | \boldsymbol{a}^{(i)}, \boldsymbol{v}^{(i)}, \boldsymbol{z}^{(i)}\big) d\boldsymbol{y}^{(i)}}{\int_{\mathcal{X}} \exp\Big([\theta_t^\star(\boldsymbol{z}^{(i)}) + 2\Theta_{t,-t}^{\star\top} \boldsymbol{x}_{-t}^{(i)}] x_t^{(i)} + \Theta_{tt}^\star \overline{x}_t^{(i)} \Big) dx_t^{(i)}} \\
&= \int_{\mathcal{X}^{p-1}} \bigg[\int_{\mathcal{X}} \widetilde{x}_u^{(i)} dx_t^{(i)} \bigg] \frac{f_{\mathbf{y}_{-t}|\mathbf{a},\mathbf{v},\mathbf{z}}\big(\boldsymbol{y}_{-t}^{(i)} | \boldsymbol{a}^{(i)}, \boldsymbol{v}^{(i)}, \boldsymbol{z}^{(i)}\big) d\boldsymbol{y}_{-t}^{(i)}}{\int_{\mathcal{X}} \exp\Big([\theta_t^\star(\boldsymbol{z}^{(i)}) + 2\Theta_{t,-t}^{\star\top} \boldsymbol{x}_{-t}^{(i)}] x_t^{(i)} + \Theta_{tt}^\star \overline{x}_t^{(i)} \Big) dx_t^{(i)}} \\
&\overset{(b)}{=} 0,
\end{aligned}
$$

where (a) follows by plugging in $f_{\mathsf{x}_t|\mathbf{x}_{-t},\mathbf{z}}\big(x_t^{(i)} | \boldsymbol{x}_{-t}^{(i)}, \boldsymbol{z}^{(i)}; \theta_t^\star(\boldsymbol{z}^{(i)}), \Theta_t^\star\big)$ from Eq. (3.10) and (b) follows because $\int_{\mathcal{X}} x_t^{(i)} dx_t^{(i)} = 0$ and $\int_{\mathcal{X}} \overline{x}_t^{(i)} dx_t^{(i)} = 0$.

3.B.1.2 Proof of Lemma 3.3(b): Anti-concentration of second directional derivative

We start by claiming that the second-order directional derivative can be lower bounded by a quadratic form. We provide a proof at the end.

Lemma 3.5 (Lower bound on the second directional derivative)**.** *For every $t \in [p_y]$ with $\underline{\Omega}_t$ defined in Eq.* (3.63)*, we have*

$$\partial^2_{\underline{\Omega}_t^2} \mathcal{L}_t(\underline{\Theta}_t) \geq \frac{1}{nC_{2,\tau}} \sum_{i \in [n]} \Big([\Delta_t^{(i)}]^\top \widetilde{\boldsymbol{x}}^{(i)} \Big)^2,$$

where $\Delta_t^{(i)} \triangleq \begin{bmatrix} \omega_t^{(i)} \\ \Omega_{t,-t}^\top \\ \Omega_{tt} \end{bmatrix} \in \mathbb{R}^{\widetilde{p}+1}$ *and* $\widetilde{\boldsymbol{x}}^{(i)} \triangleq \begin{bmatrix} x_t^{(i)} \\ 2\boldsymbol{x}_{-t}^{(i)} x_t^{(i)} \\ \overline{x}_t^{(i)} \end{bmatrix} \in \mathbb{R}^{\widetilde{p}+1}$ *for all $i \in [n]$ with $\overline{x}_t^{(i)} = \big[x_t^{(i)}\big]^2 - x_{\max}^2/3$ and the constant $C_{2,\tau}$ was defined in Eq.* (3.55)*.*

Given this, we proceed to show the anti-concentration of the second-order directional derivative. Fix any $t \in [p_y]$ and any $\underline{\Theta} \in \Lambda_\theta^n \times \Lambda_\Theta$. From Lemma 3.5, we have

$$\partial^2_{\underline{\Omega}_t^2} \mathcal{L}_t(\underline{\Theta}_t) \geq \frac{1}{nC_{2,\tau}} \sum_{i \in [n]} \Big([\Delta_t^{(i)}]^\top \widetilde{\boldsymbol{x}}^{(i)} \Big)^2. \tag{3.68}$$

First, using the Hoeffding's inequality, let us show concentration of $\frac{1}{n}\sum_{i \in [n]} \Big([\Delta_t^{(i)}]^\top \widetilde{\boldsymbol{x}}^{(i)} \Big)^2$ around its mean. We observe that each term in the summation is an independent random variable and is bounded as follows

$$\begin{aligned} \Big([\Delta_t^{(i)}]^\top \widetilde{\boldsymbol{x}}^{(i)} \Big)^2 &\overset{(a)}{=} \big(\omega_t^{(i)} x_t^{(i)} + 2\Omega_{t,-t}^\top \boldsymbol{x}_{-t}^{(i)} x_t^{(i)} + \Omega_{tt} \overline{x}_t^{(i)} \big)^2 \\ &\overset{(b)}{\leq} \big(|\omega_t^{(i)}| + 2\|\Omega_t\|_1 \|\boldsymbol{x}^{(i)}\|_\infty \big)^2 x_{\max}^2 \overset{(c)}{\leq} \big(2\alpha + 4\beta x_{\max} \big)^2 x_{\max}^2 \overset{Eq.\ (3.55)}{=} 4C_{1,\tau}^2 x_{\max}^2, \end{aligned}$$

where (a) follows by plugging in $\Delta_t^{(i)}$ and $\widetilde{\boldsymbol{x}}^{(i)}$, (b) follows from triangle inequality, Cauchy–Schwarz inequality and because $\|\boldsymbol{x}^{(i)}\|_\infty \leq x_{\max}$ for all $i \in [n]$, and (c) follows because $\Omega \in 2\Lambda_\Theta$, $\omega^{(i)} \in 2\Lambda_\theta$, and $\|\boldsymbol{x}^{(i)}\|_\infty \leq x_{\max}$ for all $i \in [n]$. Then, from the Hoeffding's inequality, for any $\varepsilon > 0$ we have

$$\mathbb{P}\bigg(\bigg| \frac{1}{n} \sum_{i \in [n]} \Big([\Delta_t^{(i)}]^\top \widetilde{\boldsymbol{x}}^{(i)} \Big)^2 - \frac{1}{n} \sum_{i \in [n]} \mathbb{E}\Big[\Big([\Delta_t^{(i)}]^\top \widetilde{\boldsymbol{x}}^{(i)} \Big)^2 \Big] \bigg| > \varepsilon \bigg) < 2 \exp\bigg(- \frac{n\varepsilon^2}{32 C_{1,\tau}^4 x_{\max}^4} \bigg).$$

Applying the union bound over all $t \in [p_y]$, for any $\delta \in (0,1)$ and uniformly for all $t \in [p_y]$, we have

$$\frac{1}{n} \sum_{i \in [n]} \Big([\Delta_t^{(i)}]^\top \widetilde{\boldsymbol{x}}^{(i)} \Big)^2 \geq \frac{1}{n} \sum_{i \in [n]} \mathbb{E}\Big[\Big([\Delta_t^{(i)}]^\top \widetilde{\boldsymbol{x}}^{(i)} \Big)^2 \Big] - \varepsilon, \tag{3.69}$$

with probability at least $1 - \delta$ as long as

$$n \geq \frac{32 C_{1,\tau}^4 x_{\max}^4}{\varepsilon^2} \log\bigg(\frac{2p_y}{\delta} \bigg).$$

Now, we lower bound $\mathbb{E}\Big[\Big([\Delta_t^{(i)}]^\top \widetilde{\boldsymbol{x}}^{(i)}\Big)^2\Big]$ for every $t \in [p_y]$ and every $i \in [n]$. Fix any $t \in [p_y]$ and $i \in [n]$. We have

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x}^{(i)},\boldsymbol{z}^{(i)}}\bigg[\Big([\Delta_t^{(i)}]^\top \widetilde{\boldsymbol{x}}^{(i)}\Big)^2\bigg] &= \mathbb{E}_{\boldsymbol{a}^{(i)},\boldsymbol{v}^{(i)},\boldsymbol{z}^{(i)}}\bigg[[\Delta_t^{(i)}]^\top \mathbb{E}_{\boldsymbol{y}^{(i)}|\boldsymbol{a}^{(i)},\boldsymbol{v}^{(i)},\boldsymbol{z}^{(i)}}\Big[\widetilde{\boldsymbol{x}}^{(i)}\widetilde{\boldsymbol{x}}^{(i)\top}|\boldsymbol{z}^{(i)}\Big]\Delta_t^{(i)}\bigg] \\ &\overset{(a)}{\geq} \lambda_{\min}\mathbb{E}_{\boldsymbol{a}^{(i)},\boldsymbol{v}^{(i)},\boldsymbol{z}^{(i)}}\Big[\|\Delta_t^{(i)}\|_2^2\Big] \overset{(b)}{\geq} \lambda_{\min}\|\Omega_t\|_2^2, \qquad (3.70)\end{aligned}$$

where (a) follows from Assumption 3.2 and (b) follows from the definition of $\Delta_t^{(i)}$. Combining Eqs. (3.68) to (3.70), for any $\delta \in (0,1)$ and uniformly for all $t \in [p_y]$, we have

$$\partial^2_{\underline{\Omega}_t^2}\mathcal{L}_t(\underline{\Theta}_t) \geq \frac{1}{C_{2,\tau}}\bigg(\lambda_{\min}\|\Omega_t\|_2^2 - \varepsilon\bigg),$$

with probability at least $1-\delta$ as long as

$$n \geq \frac{32C_{1,\tau}^4 x_{\max}^4}{\varepsilon^2}\log\bigg(\frac{2p_y}{\delta}\bigg).$$

Choosing $\varepsilon = \varepsilon_2 C_{2,\tau}$ and $\delta = \delta_2$ yields the claim.

3.B.1.2.1 Proof of Lemma 3.5: Lower bound on the second directional derivative

For every $t \in [p_y]$ with $\underline{\Omega}_t$ defined in Eq. (3.63), we claim that the second-order directional derivative of the loss function defined in Eq. (3.54) is given by

$$\partial^2_{\underline{\Omega}_t^2}\mathcal{L}_t(\underline{\Theta}_t) = \frac{1}{n}\sum_{i\in[n]}\Big([\Delta_t^{(i)}]^\top\widetilde{\boldsymbol{x}}^{(i)}\Big)^2 \exp\Big(-[\theta_t^{(i)} + 2\Theta_{t,-t}^\top\boldsymbol{x}_{-t}^{(i)}]x_t^{(i)} - \Theta_{tt}\overline{x}_t^{(i)}\Big), \qquad (3.71)$$

where $\Delta_t^{(i)} \triangleq \begin{bmatrix}\omega_t^{(i)} \\ \Omega_{t,-t}^\top \\ \Omega_{tt}\end{bmatrix} \in \mathbb{R}^{\widetilde{p}+1}$ and $\widetilde{\boldsymbol{x}}^{(i)} \triangleq \begin{bmatrix}x_t^{(i)} \\ 2\boldsymbol{x}_{-t}^{(i)}x_t^{(i)} \\ \overline{x}_t^{(i)}\end{bmatrix} \in \mathbb{R}^{\widetilde{p}+1}$ for all $i \in [n]$ with $\overline{x}_t^{(i)} = \big[x_t^{(i)}\big]^2 - x_{\max}^2/3$. We provide a proof at the end.

Given this claim, we proceed to prove the lower bound on the second directional derivative. Fix any $t \in [p_y]$. From Eq. (3.71), we have

$$\begin{aligned}\partial^2_{\underline{\Omega}_t^2}\mathcal{L}_t(\underline{\Theta}_t) &= \frac{1}{n}\sum_{i\in[n]}\Big([\Delta_t^{(i)}]^\top\widetilde{\boldsymbol{x}}^{(i)}\Big)^2 \times \exp\Big(-[\theta_t^{(i)} + 2\Theta_{t,-t}^\top\boldsymbol{x}_{-t}^{(i)}]x_t^{(i)} - \Theta_{tt}\overline{x}_t^{(i)}\Big) \\ &\overset{(a)}{\geq} \frac{1}{n}\sum_{i\in[n]}\Big([\Delta_t^{(i)}]^\top\widetilde{\boldsymbol{x}}^{(i)}\Big)^2 \times \exp\Big(-\big(|\theta_t^{(i)}| + 2\|\Theta_t\|_1\|\boldsymbol{x}^{(i)}\|_\infty\big)x_{\max}\Big) \\ &\overset{(b)}{\geq} \frac{1}{n}\sum_{i\in[n]}\Big([\Delta_t^{(i)}]^\top\widetilde{\boldsymbol{x}}^{(i)}\Big)^2 \times \exp\Big(-(\alpha + 2\beta x_{\max})x_{\max}\Big)\end{aligned}$$

$$\overset{Eq.~(3.55)}{=} \frac{1}{C_{2,\tau} n} \sum_{i\in[n]} \Big([\Delta_t^{(i)}]^\top \widetilde{\boldsymbol{x}}^{(i)}\Big)^2,$$

where (a) follows from triangle inequality, Cauchy–Schwarz inequality and because $\|\boldsymbol{x}^{(i)}\|_\infty \leq x_{\max}$ for all $i \in [n]$, and (b) follows because $\theta^{(i)} \in \Lambda_\theta$ for all $i \in [n]$, $\Theta \in \Lambda_\Theta$, and $\|\boldsymbol{x}^{(i)}\|_\infty \leq x_{\max}$ for all $i \in [n]$.

3.B.1.2.2 Proof of Eq. (3.71): Expression for second directional derivative

Fix any $t \in [p_y]$. The second-order partial derivatives of $\mathcal{L}_t$ with respect to entries of $\underline{\Theta}_t$ defined in Eq. (3.11) are given by

$$\frac{\partial^2 \mathcal{L}_t(\underline{\Theta}_t)}{\partial \big[\theta_t^{(i)}\big]^2} = \frac{1}{n}\big[x_t^{(i)}\big]^2 \exp\Big(- [\theta_t^{(i)} + 2\Theta_{t,-t}^\top \boldsymbol{x}_{-t}^{(i)}]x_t^{(i)} - \Theta_{tt}\overline{x}_t^{(i)}\Big) \quad \text{for all} \quad i \in [n],$$

$$\frac{\partial^2 \mathcal{L}_t(\underline{\Theta}_t)}{\partial \Theta_{tu}\Theta_{tv}} = \begin{cases} \frac{4}{n}\sum_{i\in[n]} \big[x_t^{(i)}\big]^2 x_u^{(i)} x_v^{(i)} \exp\Big(- [\theta_t^{(i)} + 2\Theta_{t,-t}^\top \boldsymbol{x}_{-t}^{(i)}]x_t^{(i)} - \Theta_{tt}\overline{x}_t^{(i)}\Big) \\ \qquad\qquad\qquad\qquad \text{for all } u, v \in [\widetilde{p}]\setminus\{t\}\,. \\ \frac{2}{n}\sum_{i\in[n]} \overline{x}_t^{(i)} x_t^{(i)} x_u^{(i)} \exp\Big(- [\theta_t^{(i)} + 2\Theta_{t,-t}^\top \boldsymbol{x}_{-t}^{(i)}]x_t^{(i)} - \Theta_{tt}\overline{x}_t^{(i)}\Big) \\ \qquad\qquad\qquad\qquad \text{for all } u \in [\widetilde{p}]\setminus\{t\} \text{ and } v = t. \\ \frac{2}{n}\sum_{i\in[n]} \overline{x}_t^{(i)} x_t^{(i)} x_v^{(i)} \exp\Big(- [\theta_t^{(i)} + 2\Theta_{t,-t}^\top \boldsymbol{x}_{-t}^{(i)}]x_t^{(i)} - \Theta_{tt}\overline{x}_t^{(i)}\Big) \\ \qquad\qquad\qquad\qquad \text{for all } v \in [\widetilde{p}]\setminus\{t\} \text{ and } u = t. \\ \frac{1}{n}\sum_{i\in[n]} \big[\overline{x}_t^{(i)}\big]^2 \exp\Big(- [\theta_t^{(i)} + 2\Theta_{t,-t}^\top \boldsymbol{x}_{-t}^{(i)}]x_t^{(i)} - \Theta_{tt}\overline{x}_t^{(i)}\Big) \\ \qquad\qquad\qquad\qquad \text{for } v = t \text{ and } u = t. \end{cases}$$

$$\frac{\partial^2 \mathcal{L}_t(\underline{\Theta}_t)}{\partial \Theta_{tu}\theta_t^{(i)}} = \frac{\partial^2 \mathcal{L}_t(\underline{\Theta}_t)}{\partial \theta_t^{(i)}\Theta_{tu}} = \begin{cases} \frac{2}{n}\big[x_t^{(i)}\big]^2 x_u^{(i)} \exp\Big(-[\theta_t^{(i)} + 2\Theta_{t,-t}^\top \boldsymbol{x}_{-t}^{(i)}]x_t^{(i)} - \Theta_{tt}\overline{x}_t^{(i)}\Big) \\ \qquad\qquad\qquad\qquad \text{for all } i \in [n], u \in [\widetilde{p}]\setminus\{t\}\,. \\ \frac{1}{n} x_t^{(i)} \overline{x}_t^{(i)} \exp\Big(-[\theta_t^{(i)} + 2\Theta_{t,-t}^\top \boldsymbol{x}_{-t}^{(i)}]x_t^{(i)} - \Theta_{tt}\overline{x}_t^{(i)}\Big) \\ \qquad\qquad\qquad\qquad \text{for all } i \in [n], u = t. \end{cases}$$

Now, we can write the second-order directional derivative of $\mathcal{L}_t$ as

$$\begin{aligned} \partial^2_{\underline{\Omega}_t^2} \mathcal{L}_t(\underline{\Theta}_t) &\triangleq \lim_{h\to 0} \frac{\partial_{\underline{\Omega}_t}\mathcal{L}_t(\underline{\Theta}_t + h\underline{\Omega}_t) - \partial_{\underline{\Omega}_t}\mathcal{L}_t(\underline{\Theta}_t)}{h} \\ &= \sum_{i\in[n]} \big[\omega_t^{(i)}\big]^2 \frac{\partial^2 \mathcal{L}_t(\underline{\Theta}_t)}{\partial \big[\theta_t^{(i)}\big]^2} + \sum_{u\in[\widetilde{p}]}\sum_{v\in[\widetilde{p}]} \Omega_{tu}\Omega_{tv} \frac{\partial^2 \mathcal{L}_t(\underline{\Theta}_t)}{\partial \Theta_{tu}\Theta_{tv}} + 2\sum_{i\in[n]}\sum_{u\in[\widetilde{p}]} \omega_t^{(i)}\Omega_{tu} \frac{\partial^2 \mathcal{L}_t(\underline{\Theta}_t)}{\partial \Theta_{tu}\theta_t^{(i)}} \\ &= \frac{1}{n}\sum_{i\in[n]} \Big(\big[\omega_t^{(i)} x_t^{(i)}\big]^2 + 4\sum_{u\in[\widetilde{p}]\setminus\{t\}} \Omega_{tu} x_t^{(i)} x_u^{(i)} \sum_{v\in[\widetilde{p}]\setminus\{t\}} \Omega_{tv} x_t^{(i)} x_v^{(i)} + 4\Omega_{tt}\overline{x}_t^{(i)} \sum_{u\in[\widetilde{p}]\setminus\{t\}} \Omega_{tu} x_t^{(i)} x_u^{(i)} + \big[\Omega_{tt}\overline{x}_t^{(i)}\big]^2 \\ &\qquad + 4\omega_t^{(i)} x_t^{(i)} \sum_{u\in[\widetilde{p}]\setminus\{t\}} \Omega_{tu} x_t^{(i)} x_u^{(i)} + 2\omega_t^{(i)} x_t^{(i)} \big[\Omega_{tt}\overline{x}_t^{(i)}\big]\Big) \times \exp\Big(-[\theta_t^{(i)} + 2\Theta_{t,-t}^\top \boldsymbol{x}_{-t}^{(i)}]x_t^{(i)} - \Theta_{tt}\overline{x}_t^{(i)}\Big) \\ &= \frac{1}{n}\sum_{i\in[n]} \big(\omega_t^{(i)} x_t^{(i)} + 2\Omega_{t,-t}^\top \boldsymbol{x}_{-t}^{(i)} x_t^{(i)} + \Omega_{tt}\overline{x}_t^{(i)}\big)^2 \exp\Big(- [\theta_t^{(i)} + 2\Theta_{t,-t}^\top \boldsymbol{x}_{-t}^{(i)}]x_t^{(i)} - \Theta_{tt}\overline{x}_t^{(i)}\Big) \end{aligned}$$

$$\overset{(a)}{=} \frac{1}{n} \sum_{i \in [n]} \Big([\Delta_t^{(i)}]^\top \widetilde{\boldsymbol{x}}^{(i)} \Big)^2 \exp\Big(-[\theta_t^{(i)} + 2\Theta_{t,-t}^\top \boldsymbol{x}_{-t}^{(i)}] x_t^{(i)} - \Theta_{tt} \overline{x}_t^{(i)} \Big),$$

where (a) follows from the definitions of $\Delta_t^{(i)}$ and $\widetilde{\boldsymbol{x}}^{(i)}$.

3.B.2 Proof of Lemma 3.2: Lipschitzness of the loss function

Consider any direction $\underline{\Omega} = \underline{\widetilde{\Theta}} - \underline{\Theta}$. Now, define the function $q : [0,1] \to \mathbb{R}$ as follows

$$q(a) = \mathcal{L}\big(\underline{\Theta} + a(\underline{\widetilde{\Theta}} - \underline{\Theta})\big). \tag{3.72}$$

Then, the desired inequality in Eq. (3.56) is equivalent to

$$|q(1) - q(0)| \leq 2x_{\max}^2 C_{2,\tau} \Big(\sum_{t \in [p]} \|\Omega_t\|_1 + \frac{1}{n} \sum_{i \in [n]} \|\omega^{(i)}\|_1 \Big).$$

From the mean value theorem, there exists $a' \in (0,1)$ such that

$$|q(1) - q(0)| = \left| \frac{dq(a')}{da} \right| \overset{Eq.\ (3.72)}{=} \left| \frac{d\mathcal{L}\big(\underline{\Theta} + a(\underline{\widetilde{\Theta}} - \underline{\Theta})\big)}{da} \right| \overset{Eq.\ (3.64)}{=} \left| \partial_{\underline{\Omega}} \mathcal{L}(\underline{\Theta}) \big|_{\underline{\Theta} = \underline{\Theta} + a(\underline{\widetilde{\Theta}} - \underline{\Theta})} \right|. \tag{3.73}$$

Using Eq. (3.67) in Eq. (3.73), we can write

$$\begin{aligned}
&\big|q(1) - q(0)\big| \\
&= \frac{1}{n} \bigg| \sum_{t \in [p]} \sum_{i \in [n]} \Big([\Delta_t^{(i)}]^\top \widetilde{\boldsymbol{x}}^{(i)} \Big) \times \exp\Big(- \Big[\big(\theta_t^{(i)} + a'(\widetilde{\theta}_t^{(i)} - \theta_t^{(i)}) \big) + \\
&\qquad\qquad 2\big(\Theta_{t,-t} + a'(\widetilde{\Theta}_{t,-t} - \Theta_{t,-t}) \big)^\top \boldsymbol{x}_{-t}^{(i)} \Big] x_t^{(i)} - \big(\Theta_{tt} + a'(\widetilde{\Theta}_{tt} - \Theta_{tt}) \big) \overline{x}_t^{(i)} \Big) \bigg| \\
&\overset{(a)}{\leq} \exp\Big(\big([(1-a')\alpha + a'\alpha] + 2[(1-a')\beta + a'\beta] x_{\max} \big) x_{\max} \Big) \frac{1}{n} \bigg| \sum_{t \in [p]} \sum_{i \in [n]} \Big([\Delta_t^{(i)}]^\top \widetilde{\boldsymbol{x}}^{(i)} \Big) \bigg| \\
&\overset{(b)}{\leq} \frac{2x_{\max}^2 C_{2,\tau}}{n} \sum_{t \in [p]} \sum_{i \in [n]} \|\Delta_t^{(i)}\|_1 \overset{(c)}{=} 2x_{\max}^2 C_{2,\tau} \Big(\sum_{t \in [p]} \|\Omega_t\|_1 + \frac{1}{n} \sum_{i \in [n]} \|\omega^{(i)}\|_1 \Big),
\end{aligned}$$

where (a) follows from triangle inequality, Cauchy–Schwarz inequality, $\theta^{(i)}, \widetilde{\theta}^{(i)} \in \Lambda_\theta$, $\Theta, \widetilde{\Theta} \in \Lambda_\Theta$, and $\|\boldsymbol{x}^{(i)}\|_\infty \leq x_{\max}$ for all $i \in [n]$, (b) follows from Eq. (3.55), the triangle inequality, and because $\|\boldsymbol{x}^{(i)}\|_\infty \leq x_{\max}$ for all $i \in [n]$, and (c) follows from the definition of $\Delta_t^{(i)}$.

3.C Proof of Theorem 3.1 Part II: Recovering unit-level parameters

To analyze our estimate of the unit-level parameters, we use the estimate $\widehat{\Theta}$ of the population-level parameter $\Theta^\star$ along with the associated guarantee provided in Theorem 3.1 Part I. We note that the constraints on the unit-level parameters in Eq. (3.12) are independent across units, i.e., $\theta^{(i)} \in \Lambda_\theta$ independently for all $i \in [n]$. Therefore, we look at n independent convex optimization problems by decomposing the loss function $\mathcal{L}$ in Eq. (3.11) and the estimate $\underline{\widehat{\Theta}}$ in Eq. (3.12) as follows: For $i \in [n]$, we define

$$\mathcal{L}^{(i)}\big(\theta^{(i)}\big) \triangleq \sum_{t\in[p_y]} \exp\Big(-[\theta_t^{(i)} + 2\widehat{\Theta}_{t,-t}^\top \boldsymbol{x}_{-t}^{(i)}]x_t^{(i)} - \widehat{\Theta}_{tt}\overline{x}_t^{(i)}\Big)$$
$$\text{and} \quad \widehat{\theta}^{(i)} \triangleq \operatorname*{arg\,min}_{\theta^{(i)}\in\Lambda_\theta} \mathcal{L}^{(i)}\big(\theta^{(i)}\big). \tag{3.74}$$

Now, fix any $i \in [n]$. From Eq. (3.74), we have $\mathcal{L}^{(i)}\big(\widehat{\theta}^{(i)}\big) \leq \mathcal{L}^{(i)}\big(\theta^{\star(i)}\big)$. Using contraposition, to prove this part, it is sufficient to show that all points $\theta^{(i)} \in \Lambda_\theta$ that satisfy $\|\theta^{(i)} - \theta^{\star(i)}\|_2 \geq R(\varepsilon, \delta)$ also uniformly satisfy

$$\mathcal{L}^{(i)}\big(\theta^{(i)}\big) \geq \mathcal{L}^{(i)}\big(\theta^{\star(i)}\big) + R^2(\varepsilon,\delta) \text{ when } n \geq \frac{ce^{c'\beta}p_y^2\widetilde{p}^2\Big(\log\frac{p_y}{\delta} + \mathcal{M}_\Theta(\frac{\varepsilon^2}{\widetilde{p}}) + \widetilde{\mathcal{M}}_{\theta,n}(\varepsilon,\delta)\Big)}{\varepsilon^4}, \tag{3.75}$$

with probability at least $1-\delta$ where $R(\varepsilon,\delta)$ was defined in Eq. (3.15) and $\widetilde{\mathcal{M}}_{\theta,n}(\varepsilon,\delta)$ was defined in Eq. (3.16). Then, the guarantee in Theorem 3.1 follows by applying a union bound over all $i \in [n]$.

To that end, the lemma below, proven in Section 3.C.1, shows that for any fixed $\theta^{(i)} \in \Lambda_\theta$, if $\theta^{(i)}$ is far from $\theta^{\star(i)}$, then with high probability $\mathcal{L}^{(i)}\big(\theta^{(i)}\big)$ is significantly larger than $\mathcal{L}^{(i)}\big(\theta^{\star(i)}\big)$.

Lemma 3.6 (Gap between the loss function for a fixed parameter)**.** *Fix any $\varepsilon > 0$, $\delta \in (0,1)$, and $i \in [n]$. Then, for any $\theta^{(i)} \in \Lambda_\theta$ such that $\|\theta^{(i)} - \theta^{\star(i)}\|_2 \geq \varepsilon\gamma$ (see Eq. (3.15)), we have*

$$\mathcal{L}^{(i)}\big(\theta^{(i)}\big) \geq \mathcal{L}^{(i)}\big(\theta^{\star(i)}\big) + \frac{2^{2.5}\beta x_{\max}^4}{\pi e C_{2,\tau}^5}\|\theta^{(i)} - \theta^{\star(i)}\|_2^2$$
$$\textit{for} \quad n \geq \frac{ce^{c'\beta}p_y^2\widetilde{p}^2\Big(\log\frac{p_y}{\delta} + \mathcal{M}_\Theta(\frac{\varepsilon^2}{\widetilde{p}}) + \mathcal{M}_{\theta,n}(\frac{\varepsilon^2}{\widetilde{p}})\Big)}{\varepsilon^4},$$

with probability at least $1 - \delta - c\beta^2\log p_y \cdot \exp(-e^{-c'\beta}\|\theta^{(i)} - \theta^{\star(i)}\|_2^2)$ where $C_{2,\tau}$ was defined in Eq. (3.55).

Note. When we invoke Lemma 3.6, we ensure that $c\beta^2\log p_y \cdot \exp(-e^{-c'\beta}\|\theta^{(i)} - \theta^{\star(i)}\|_2^2)$ is of the same order as δ.

Next, we show that the loss function $\mathcal{L}^{(i)}$ is Lipschitz (see Section 3.C.2 for the proof).

Lemma 3.7 (Lipschitzness of the loss function)**.** *Consider any $i \in [n]$. Then, the loss function $\mathcal{L}^{(i)}$ is Lipschitz with respect to the ℓ_1 norm $\|\cdot\|_1$ and with Lipschitz constant $x_{\max}C_{2,\tau}$, i.e.,*

$$\big|\mathcal{L}^{(i)}\big(\widetilde{\theta}^{(i)}\big) - \mathcal{L}^{(i)}\big(\theta^{(i)}\big)\big| \leq x_{\max}C_{2,\tau}\|\widetilde{\theta}^{(i)} - \theta^{(i)}\|_1 \quad \textit{for all} \quad \theta^{(i)}, \widetilde{\theta}^{(i)} \in \Lambda_\theta, \tag{3.76}$$

where the constant $C_{2,\tau}$ was defined in Eq. (3.55).

Given these lemmas, we now proceed with the proof.

Proof strategy. We want to show that all points $\theta^{(i)} \in \Lambda_\theta$, that satisfy $\|\theta^{(i)} - \theta^{\star(i)}\|_2 \geq R(\varepsilon, \delta)$, uniformly satisfy Eq. (3.75) with probability at least $1-\delta$. To do so, we consider the set of points $\Lambda_\theta^r \subset \Lambda_\theta$ whose distance from $\theta^{\star(i)}$ is at least $r > 0$ in ℓ_2 norm. Then, using an appropriate covering set of Λ_θ^r and the Lipschitzness of $\mathcal{L}^{(i)}$, we show that the value of $\mathcal{L}^{(i)}$ at all points in Λ_θ^r is uniformly $\Omega(r^2)$ larger than the value of $\mathcal{L}^{(i)}$ at $\theta^{\star(i)}$ with high probability. Finally, we choose r small enough to make the failure probability smaller than δ.

Arguments for points in the covering set. Consider any $r \geq \varepsilon\gamma$ (where γ is defined in Eq. (3.15)) and the set of elements $\Lambda_\theta^r \triangleq \big\{\theta^{(i)} \in \Lambda_\theta : \|\theta^{\star(i)} - \theta^{(i)}\|_2 \geq r\big\}$. Let $\mathcal{U}(\Lambda_\theta^r, \varepsilon')$ be the ε'-cover of the smallest size for the set Λ_θ^r with respect to $\|\cdot\|_1$ (see Definition 3.2) and let $\mathcal{C}(\Lambda_\theta^r, \varepsilon')$ be the ε'-covering number where

$$\varepsilon' \triangleq \frac{2\sqrt{2}\beta x_{\max}^3 r^2}{\pi e C_{2,\tau}^6}. \tag{3.77}$$

Now, we argue by a union bound that the value of $\mathcal{L}^{(i)}$ at all points in $\mathcal{U}(\Lambda_\theta^r, \varepsilon')$ is uniformly $\Omega(r^2)$ larger than $\mathcal{L}^{(i)}(\theta^{\star(i)})$ with high probability. For any $\theta^{(i)} \in \mathcal{U}(\Lambda_\theta^r, \varepsilon')$, we have

$$\|\theta^{\star(i)} - \theta^{(i)}\|_2 \overset{(a)}{\geq} r, \tag{3.78}$$

where (a) follows because $\mathcal{U}(\Lambda_\theta^r, \varepsilon') \subseteq \Lambda_\theta^r$. Now, applying Lemma 3.6 with $\varepsilon \hookleftarrow \varepsilon$ and $\delta \hookleftarrow \delta/2\mathcal{C}(\Lambda_\theta^r, \varepsilon')$, we have

$$\mathcal{L}^{(i)}\big(\theta^{(i)}\big) \geq \mathcal{L}^{(i)}\big(\theta^{\star(i)}\big) + \frac{4\sqrt{2}\beta x_{\max}^4}{\pi e C_{2,\tau}^5}\|\theta^{\star(i)} - \theta^{(i)}\|_2^2 \overset{Eq.~(3.78)}{\geq} \mathcal{L}^{(i)}\big(\theta^{\star(i)}\big) + \frac{4\sqrt{2}\beta x_{\max}^4 r^2}{\pi e C_{2,\tau}^5},$$

with probability at least $1-\delta/2\mathcal{C}(\Lambda_\theta^r, \varepsilon') - c\beta^2 \log p_y \cdot \exp(-e^{-c'\beta}\|\theta^{(i)} - \theta^{\star(i)}\|_2^2)$ whenever

$$n \geq \frac{c e^{c'\beta} p_y^2 \widetilde{p}^2 \Big(\log \frac{\mathcal{C}(\Lambda_\theta^r, \varepsilon') \cdot p_y}{\delta} + \mathcal{M}_\Theta(\frac{\varepsilon^2}{\widetilde{p}}) + \mathcal{M}_{\theta,n}(\frac{\varepsilon^2}{\widetilde{p}})\Big)}{\varepsilon^4}. \tag{3.79}$$

By applying the union bound over $\mathcal{U}(\Lambda_\theta^r, \varepsilon')$, as long as n satisfies Eq. (3.79), we have

$$\mathcal{L}^{(i)}\big(\theta^{(i)}\big) \geq \mathcal{L}^{(i)}\big(\theta^{\star(i)}\big) + \frac{4\sqrt{2}\beta x_{\max}^4 r^2}{\pi e C_{2,\tau}^5} \text{ uniformly for every } \theta^{(i)} \in \mathcal{U}(\Lambda_\theta^r, \varepsilon'), \tag{3.80}$$

with probability at least $1 - \delta/2 - c\beta^2\mathcal{C}(\Lambda_\theta^r, \varepsilon') \log p_y \cdot \exp(-e^{-c'\beta}\|\theta^{(i)} - \theta^{\star(i)}\|_2^2)$ which can lower bounded by $1 - \delta/2 - c\beta^2\mathcal{C}(\Lambda_\theta^r, \varepsilon') \log p_y \cdot \exp(-e^{-c'\beta}r^2)$ using Eq. (3.78).

Arguments for points outside the covering set. Next, we establish the claim Eq. (3.75) for an arbitrary $\widetilde{\theta}^{(i)} \in \Lambda_\theta^r$ conditional on the event that Eq. (3.80) holds. Given a fixed $\widetilde{\theta}^{(i)} \in \Lambda_\theta^r$, let $\theta^{(i)}$ be (one of) the point(s) in the $\mathcal{U}(\Lambda_\theta^r, \varepsilon')$ that satisfies

$\|\theta^{(i)} - \widetilde{\theta}^{(i)}\|_1 \leq \varepsilon'$ (there exists such a point by Definition 3.2) Then, the choices Eq. (3.77) and Lemma 3.7 put together imply that

$$\begin{aligned}\mathcal{L}^{(i)}\big(\widetilde{\theta}^{(i)}\big) \geq \mathcal{L}^{(i)}\big(\theta^{(i)}\big) - x_{\max} C_{2,\tau} \|\theta^{(i)} - \widetilde{\theta}^{(i)}\|_1 &\geq \mathcal{L}^{(i)}\big(\theta^{(i)}\big) - x_{\max} C_{2,\tau}\varepsilon' \\ &\overset{Eq.\ (3.77)}{\geq} \mathcal{L}^{(i)}\big(\theta^{(i)}\big) - \frac{2\sqrt{2}\beta x_{\max}^4 r^2}{\pi e C_{2,\tau}^5} \\ &\overset{Eq.\ (3.80)}{\geq} \mathcal{L}^{(i)}\big(\theta^{\star(i)}\big) + \frac{2\sqrt{2}\beta x_{\max}^4 r^2}{\pi e C_{2,\tau}^5},\end{aligned}$$

It remains to bound sample size n and the failure probability δ.

Bounding n. Using $\Lambda_\theta^r \subseteq \Lambda_\theta$, we find that

$$\mathcal{C}(\Lambda_\theta^r, \varepsilon') \overset{(a)}{\leq} \mathcal{C}(\Lambda_\theta, \varepsilon'). \tag{3.81}$$

Putting together Eq. (3.77) and Eq. (3.81), the lower bound Eq. (3.79) can be replaced by

$$n \geq \frac{c e^{c'\beta} p_y^2 \widetilde{p}^2 \Big(\log \frac{p_y}{\delta} + \mathcal{M}_\Theta(\frac{\varepsilon^2}{\widetilde{p}}) + \mathcal{M}_\theta\big(r^2\big) + \mathcal{M}_{\theta,n}(\frac{\varepsilon^2}{\widetilde{p}}) \Big)}{\varepsilon^4}.$$

Bounding δ. To bound the failure probability by δ, it is sufficient to chose r such that

$$\delta \geq \delta/2 + c\beta^2 \mathcal{C}(\Lambda_\theta^r, \varepsilon') \log p_y \cdot \exp(-e^{-c'\beta} r^2). \tag{3.82}$$

From Eq. (3.81) and Eq. (3.82), it is sufficient to chose r such that

$$\delta \geq \delta/2 + c\beta^2 \mathcal{C}(\Lambda_\theta, \varepsilon') \log p_y \cdot \exp(-e^{-c'\beta} r^2). \tag{3.83}$$

Re-arranging and taking logarithm on both sides of Eq. (3.83) and using Eq. (3.77), we have

$$\log \delta \geq c \Bigg[\log\big(\beta^2 \log p_y\big) + \mathcal{M}_\theta\Big(\frac{r^2}{c e^{c'\beta}}\Big) - e^{-c'\beta} r^2 \Bigg]. \tag{3.84}$$

Finally, Eq. (3.84) holds whenever

$$r \geq c e^{c'\beta} \sqrt{\log \frac{\beta^2 \log p_y}{\delta} + \mathcal{M}_\theta(c e^{-c'\beta})}.$$

Recalling that the choice of r was such that $r \geq \varepsilon\gamma$ completes the proof.

3.C.1 Proof of Lemma 3.6: Gap between the loss function for a fixed parameter

Fix any $\varepsilon > 0$, any $\delta \in (0,1)$, and any $i \in [n]$. Consider any direction $\omega^{(i)} \in \mathbb{R}^{p_y}$ along the parameter $\theta^{(i)}$, i.e.,

$$\omega^{(i)} = \theta^{(i)} - \theta^{\star(i)}. \tag{3.85}$$

We denote the first-order and the second-order directional derivatives of the loss function $\mathcal{L}^{(i)}$ in Eq. (3.74) along the direction $\omega^{(i)}$ evaluated at $\theta^{(i)}$ by $\partial_{\omega^{(i)}}(\mathcal{L}^{(i)}(\theta^{(i)}))$ and $\partial^2_{[\omega^{(i)}]^2}\mathcal{L}^{(i)}(\theta^{(i)})$, respectively. Below, we state a lemma (with proof divided across Section 3.C.1.1 and Section 3.C.1.2) that provides us a control on $\partial_{\omega^{(i)}}(\mathcal{L}^{(i)}(\theta^{\star(i)}))$ and $\partial^2_{[\omega^{(i)}]^2}\mathcal{L}^{(i)}(\theta^{(i)})$. The assumptions of Lemma 3.6 remain in force.

Lemma 3.8 (Control on first and second directional derivatives)**.** *For any fixed $\varepsilon_1, \varepsilon_2 > 0$, $\delta_1 \in (0,1)$, $i \in [n]$, $\theta^{(i)} \in \Lambda_\theta$ with $\omega^{(i)}$ defined in Eq. (3.85), we have the following:*

(a) Concentration of first directional derivative*: We have*

$$\left|\partial_{\omega^{(i)}}(\mathcal{L}^{(i)}(\theta^{\star(i)}))\right| \leq \varepsilon_1\|\omega^{(i)}\|_1 + \varepsilon_2\|\omega^{(i)}\|_2^2 \quad \textit{for}$$

$$n \geq \frac{ce^{c'\beta}p_y^2\widetilde{p}^2\Big(\log\frac{p_y}{\delta_1} + \mathcal{M}_\Theta(\frac{\varepsilon_1^2}{\widetilde{p}}) + \mathcal{M}_{\theta,n}(\frac{\varepsilon_1^2}{\widetilde{p}})\Big)}{\varepsilon_1^4},$$

with probability at least $1 - \delta_1 - O\Bigg(\beta^2\log p_y \exp\left(\frac{-\varepsilon_2^2\|\omega^{(i)}\|_2^2}{e^{c'\beta}}\right)\Bigg)$.

(b) Anti-concentration of second directional derivative*: We have*

$$\partial^2_{[\omega^{(i)}]^2}\mathcal{L}^{(i)}(\theta^{(i)}) \geq \frac{32\sqrt{2}\beta x_{\max}^4}{\pi e C_{2,\tau}^5}\|\omega^{(i)}\|_2^2,$$

with probability at least $1 - O\Bigg(\beta^2\log p_y \exp\left(\frac{-\|\omega^{(i)}\|_2^2}{e^{c'\beta}}\right)\Bigg)$ *where $C_{2,\tau}$ was defined in Eq. (3.55).*

Given this lemma, we now proceed with the proof. Define a function $g : [0,1] \to \mathbb{R}^{p_y}$ as follows

$$g(a) = \theta^{\star(i)} + a(\theta^{(i)} - \theta^{\star(i)}).$$

Notice that $g(0) = \theta^{\star(i)}$ and $g(1) = \theta^{(i)}$ as well as

$$\frac{d\mathcal{L}^{(i)}(g(a))}{da} = \partial_{\omega^{(i)}}(\mathcal{L}^{(i)}(\widetilde{\theta}^{(i)}))\big|_{\widetilde{\theta}^{(i)}=g(a)} \quad \text{and} \quad \frac{d^2\mathcal{L}^{(i)}(g(a))}{da^2} = \partial^2_{[\omega^{(i)}]^2}\mathcal{L}^{(i)}(\widetilde{\theta}^{(i)})\big|_{\widetilde{\theta}^{(i)}=g(a)}. \tag{3.86}$$

By the fundamental theorem of calculus, we have

$$\frac{d\mathcal{L}^{(i)}(g(a))}{da} \geq \frac{d\mathcal{L}^{(i)}(g(a))}{da}\Big|_{a=0} + a\min_{a\in(0,1)}\frac{d^2\mathcal{L}^{(i)}(g(a))}{da^2}. \tag{3.87}$$

Integrating both sides of Eq. (3.87) with respect to a, we obtain

$$\mathcal{L}^{(i)}(g(a)) - \mathcal{L}^{(i)}(g(0)) \geq a\frac{d\mathcal{L}^{(i)}(g(a))}{da}\Big|_{a=0} + \frac{a^2}{2}\min_{a\in(0,1)}\frac{d^2\mathcal{L}^{(i)}(g(a))}{da^2}$$

$$\begin{aligned}
&\overset{Eq.~(3.86)}{=} a\partial_{\omega^{(i)}}(\mathcal{L}^{(i)}(\widetilde{\theta}^{(i)}))\big|_{\widetilde{\theta}^{(i)}=g(0)} + \frac{a^2}{2}\min_{a\in(0,1)} \partial^2_{[\omega^{(i)}]^2}\mathcal{L}^{(i)}(\widetilde{\theta}^{(i)})\big|_{\widetilde{\theta}^{(i)}=g(a)} \\
&\overset{(a)}{=} a\partial_{\omega^{(i)}}(\mathcal{L}^{(i)}(\theta^{\star(i)})) + \frac{a^2}{2}\min_{a\in(0,1)} \partial^2_{[\omega^{(i)}]^2}\mathcal{L}^{(i)}(\widetilde{\theta}^{(i)})\big|_{\widetilde{\theta}^{(i)}=g(a)} \\
&\overset{(b)}{\geq} -a\big|\partial_{\omega^{(i)}}(\mathcal{L}^{(i)}(\theta^{\star(i)}))\big| + \frac{a^2}{2}\min_{a\in(0,1)} \partial^2_{[\omega^{(i)}]^2}\mathcal{L}^{(i)}(\widetilde{\theta}^{(i)})\big|_{\widetilde{\theta}^{(i)}=g(a)},
\end{aligned} \tag{3.88}$$

where (a) follows because $g(0) = \theta^{\star(i)}$, and (b) follows by the triangle inequality. Plugging in $a = 1$ in Eq. (3.88) as well as using $g(0) = \theta^{\star(i)}$ and $g(1) = \theta^{(i)}$, we find that

$$\mathcal{L}^{(i)}(\theta^{(i)}) - \mathcal{L}^{(i)}(\theta^{\star(i)}) \geq -\big|\partial_{\omega^{(i)}}(\mathcal{L}^{(i)}(\theta^{\star(i)}))\big| + \frac{1}{2}\min_{a\in(0,1)} \partial^2_{[\omega^{(i)}]^2}\mathcal{L}^{(i)}(\widetilde{\theta}^{(i)})\big|_{\widetilde{\theta}^{(i)}=g(a)}.$$

Now, we use Lemma 3.8 with $\varepsilon_1 \mapsfrom 4\sqrt{2}\beta x^4_{\max}\varepsilon/\pi e C^5_{2,\tau}$, $\varepsilon_2 \mapsfrom 8\sqrt{2}\beta x^4_{\max}/\pi e C^5_{2,\tau}$, and $\delta_1 \mapsfrom \delta$. Therefore, with probability at least $1 - \delta - O\Big(\beta^2 \log p_y \exp\Big(\frac{-\|\omega^{(i)}\|_2^2}{e^{c'\beta}}\Big)\Big)$ and as long as

$$n \geq \frac{c e^{c'\beta} p_y^2 \widetilde{p}^2\Big(\log\frac{p_y}{\delta} + \mathcal{M}_{\Theta}(\frac{\varepsilon^2}{\widetilde{p}}) + \mathcal{M}_{\theta,n}(\frac{\varepsilon^2}{\widetilde{p}})\Big)}{\varepsilon^4},$$

we have

$$\begin{aligned}
\mathcal{L}^{(i)}(\theta^{(i)}) - \mathcal{L}^{(i)}(\theta^{\star(i)}) &\geq -\frac{2^{2.5}\beta x^4_{\max}\varepsilon}{\pi e C^5_{2,\tau}}\|\omega^{(i)}\|_1 - \frac{2^{3.5}\beta x^4_{\max}}{\pi e C^5_{2,\tau}}\|\omega^{(i)}\|_2^2 + \frac{2^{4.5}\beta x^4_{\max}}{\pi e C^5_{2,\tau}}\|\omega^{(i)}\|_2^2 \\
&= -\frac{2^{2.5}\beta x^4_{\max}\varepsilon}{\pi e C^5_{2,\tau}}\|\omega^{(i)}\|_1 + \frac{2^{3.5}\beta x^4_{\max}}{\pi e C^5_{2,\tau}}\|\omega^{(i)}\|_2^2 \\
&\overset{Eq.~(3.15)}{\geq} -\frac{2^{2.5}\beta x^4_{\max}\varepsilon\gamma}{\pi e C^5_{2,\tau}}\|\omega^{(i)}\|_2 + \frac{2^{3.5}\beta x^4_{\max}}{\pi e C^5_{2,\tau}}\|\omega^{(i)}\|_2^2 \\
&\overset{(a)}{\geq} -\frac{2^{2.5}\beta x^4_{\max}}{\pi e C^5_{2,\tau}}\|\omega^{(i)}\|_2^2 + \frac{2^{3.5}\beta x^4_{\max}}{\pi e C^5_{2,\tau}}\|\omega^{(i)}\|_2^2 = \frac{2^{2.5}\beta x^4_{\max}}{\pi e C^5_{2,\tau}}\|\omega^{(i)}\|_2^2,
\end{aligned}$$

where (a) follows because $\|\omega^{(i)}\|_2 = \|\theta^{(i)} - \theta^{\star(i)}\|_2 \geq \varepsilon\gamma$ according to the lemma statement.

3.C.1.1 Proof of Lemma 3.8(a): Concentration of first directional derivative

Fix some $i \in [n]$ and some $\theta^{(i)} \in \Lambda_\theta$. Let $\omega^{(i)}$ be as defined in Eq. (3.85). We claim that the first-order directional derivative of $\mathcal{L}^{(i)}$ defined in Eq. (3.74) is given by

$$\partial_{\omega^{(i)}}(\mathcal{L}^{(i)}(\theta^{(i)})) = -\sum_{t\in[p_y]} \omega_t^{(i)} x_t^{(i)} \exp\Big(-[\theta_t^{(i)} + 2\widehat{\Theta}_{t,-t}^{\top}\boldsymbol{x}_{-t}^{(i)}]x_t^{(i)} - \widehat{\Theta}_{tt}\overline{x}_t^{(i)}\Big). \tag{3.89}$$

We provide a proof at the end. For now, we assume the claim and proceed.

We note that the tuple of random vectors $(\mathsf{y}, \mathsf{a}, \mathsf{v}, \mathsf{z})$ corresponds to a τ-SGM (see Definition 3.8) with $\tau \triangleq (\alpha, \beta, \beta, x_{\max}, \Theta)$. To show the concentration, we use Proposition 3.7 (see Section 3.H) with $\lambda = \frac{1}{4\sqrt{2}x_{\max}^2}$, decompose $\partial_{\omega^{(i)}}(\mathcal{L}^{(i)}(\theta^{\star(i)}))$ as a sum of $L = 1024\beta^2 x_{\max}^4 \log 4p_y$, and focus on these L terms. Consider the L subsets $S_1, \cdots, S_L \in [p_y]$ obtained from Proposition 3.7 with $\lambda = \frac{1}{4\sqrt{2}x_{\max}^2}$ and define

$$\psi_u(\boldsymbol{x}^{(i)}; \omega^{(i)}) \triangleq \sum_{t \in S_u} \omega_t^{(i)} x_t^{(i)} \exp\Big(-[\theta_t^{\star(i)} + 2\widehat{\Theta}_{t,-t}^\top \boldsymbol{x}_{-t}^{(i)}] x_t^{(i)} - \widehat{\Theta}_{tt} \overline{x}_t^{(i)}\Big) \text{ for every } u \in L. \tag{3.90}$$

Now, we decompose $\partial_{\omega^{(i)}}(\mathcal{L}^{(i)}(\theta^{\star(i)}))$ as a sum of the L terms defined above. More precisely, we have

$$\begin{aligned}\partial_{\omega^{(i)}}(\mathcal{L}^{(i)}(\theta^{\star(i)})) &\overset{Eq.\ (3.89)}{=} - \sum_{t \in [p_y]} \omega_t^{(i)} x_t^{(i)} \exp\Big(- [\theta_t^{\star(i)} + 2\widehat{\Theta}_{t,-t}^\top \boldsymbol{x}_{-t}^{(i)}] x_t^{(i)} - \widehat{\Theta}_{tt} \overline{x}_t^{(i)}\Big) \\ &\overset{(a)}{=} -\frac{1}{L'} \sum_{u \in [L]} \sum_{t \in S_u} \omega_t^{(i)} x_t^{(i)} \exp\Big(- [\theta_t^{\star(i)} + 2\widehat{\Theta}_{t,-t}^\top \boldsymbol{x}_{-t}^{(i)}] x_t^{(i)} - \widehat{\Theta}_{tt} \overline{x}_t^{(i)}\Big) \\ &\overset{Eq.\ (3.90)}{=} -\frac{1}{L'} \sum_{u \in [L]} \psi_u(\boldsymbol{x}^{(i)}; \omega^{(i)}),\end{aligned} \tag{3.91}$$

where (a) follows because each $t \in [p_y]$ appears in exactly $L' = \lceil L/32\sqrt{2}\beta x_{\max}^2 \rceil$ of the sets $S_1, \cdots, S_L$ according to Proposition 3.7(a) (with $\lambda = \frac{1}{4\sqrt{2}x_{\max}^2}$). Now, we focus on the L terms in Eq. (3.91).

Consider any $u \in [L]$. We claim that conditioned on $\boldsymbol{x}_{-S_u}^{(i)}$ and $\boldsymbol{z}^{(i)}$, the expected value of $\psi_u(\boldsymbol{x}^{(i)}; \omega^{(i)})$ can be upper bounded uniformly across all $u \in [L]$. We provide a proof at the end.

Lemma 3.9 (Upper bound on expected ψ_u). *Fix $\varepsilon > 0$, $\delta \in (0, 1)$, $i \in [n]$ and $\theta^{(i)} \in \Lambda_\theta$. Then, with $\omega^{(i)}$ defined in Eq. (3.85) and given $\boldsymbol{z}^{(i)}$ and $\boldsymbol{x}_{-S_u}^{(i)}$ for all $u \in [L]$, we have*

$$\max_{u \in [L]} \mathbb{E}\Big[\psi_u(\boldsymbol{x}^{(i)}; \omega^{(i)}) \mid \boldsymbol{x}_{-S_u}^{(i)}, \boldsymbol{z}^{(i)}\Big] \le \varepsilon \|\omega^{(i)}\|_1 \quad \textit{for}$$

$$n \ge \frac{c e^{c'\beta} p_y^2 \widetilde{p}^2 \Big(\log \frac{p_y}{\delta} + \mathcal{M}_\Theta(\frac{\varepsilon^2}{\widetilde{p}}) + \mathcal{M}_{\theta,n}(\frac{\varepsilon^2}{\widetilde{p}})\Big)}{\varepsilon^4},$$

with probability at least $1 - \delta$.

Consider again any $u \in [L]$. Now, we claim that conditioned on $\boldsymbol{x}_{-S_u}^{(i)}$ and $\boldsymbol{z}^{(i)}$, $\psi_u(\boldsymbol{x}^{(i)}; \omega^{(i)})$ concentrates around its conditional expected value. We provide a proof at the end.

Lemma 3.10 (Concentration of ψ_u). *Fix $\varepsilon > 0$, $i \in [n]$, $u \in [L]$, and $\theta^{(i)} \in \Lambda_\theta$. Then, with $\omega^{(i)}$ defined in Eq. (3.85) and given $\boldsymbol{z}^{(i)}$ and $\boldsymbol{x}_{-S_u}^{(i)}$, we have*

$$\Big|\psi_u(\boldsymbol{x}^{(i)}; \omega^{(i)}) - \mathbb{E}\big[\psi_u(\boldsymbol{x}^{(i)}; \omega^{(i)}) \mid \boldsymbol{x}_{-S_u}^{(i)}, \boldsymbol{z}^{(i)}\big]\Big| \le \varepsilon,$$

with probability at least $1 - \exp\left(\dfrac{-\varepsilon^2}{e^{c'\beta}\|\omega^{(i)}\|_2^2}\right)$.

Given these lemmas, we proceed to show the concentration of $\partial_{\omega^{(i)}}(\mathcal{L}^{(i)}(\theta^{\star(i)}))$. To that end, for any $u \in [L]$, given $\boldsymbol{x}^{(i)}_{-S_u}$ and $\boldsymbol{z}^{(i)}$, let E_u denote the event that

$$\psi_u(\boldsymbol{x}^{(i)};\omega^{(i)}) \leq \mathbb{E}\big[\psi_u(\boldsymbol{x}^{(i)};\omega^{(i)})|\boldsymbol{x}^{(i)}_{-S_u},\boldsymbol{z}^{(i)}\big] + \frac{1}{32\sqrt{2}\beta x_{\max}^2}\varepsilon_2\|\omega^{(i)}\|_2^2. \tag{3.92}$$

Since E_u in an indicator event, using the law of total expectation results in

$$\mathbb{P}(E_u) = \mathbb{E}\Big[\mathbb{P}(E_u|\boldsymbol{x}^{(i)}_{-S_u},\boldsymbol{z}^{(i)})\Big] \overset{(a)}{\geq} 1 - \exp\left(\frac{-\varepsilon_2^2\|\omega^{(i)}\|_2^2}{e^{c'\beta}}\right).$$

where (a) follows from Lemma 3.10 with $\varepsilon \hookleftarrow \dfrac{\varepsilon_2\|\omega^{(i)}\|_2^2}{32\sqrt{2}\beta x_{\max}^2}$. Now, by applying the union bound over all $u \in [L]$ where $L = 1024\beta^2 x_{\max}^4 \log 4p_y$, we have

$$\mathbb{P}\Big(\bigcap_{u\in L} E_u\Big) \geq 1 - O\bigg(\beta^2 \log p_y \exp\left(\frac{-\varepsilon_2^2\|\omega^{(i)}\|_2^2}{e^{c'\beta}}\right)\bigg).$$

Now, assume the event $\cap_{u\in L}E_u$ holds. Whenever this holds, we also have

$$\begin{aligned}\big|\partial_{\omega^{(i)}}(\mathcal{L}^{(i)}(\theta^{\star(i)}))\big| &\overset{Eq.\ (3.91)}{\leq} \frac{1}{L'}\sum_{u\in[L]}\big|\psi_u(\boldsymbol{x}^{(i)};\omega^{(i)})\big| \\ &\overset{Eq.\ (3.92)}{\leq} \frac{1}{L'}\sum_{u\in[L]}\Big|\mathbb{E}\big[\psi_u(\boldsymbol{x}^{(i)};\omega^{(i)})|\boldsymbol{x}^{(i)}_{-S_u},\boldsymbol{z}^{(i)}\big] + \frac{1}{32\sqrt{2}\beta x_{\max}^2}\varepsilon_2\|\omega^{(i)}\|_2^2\Big|,\end{aligned} \tag{3.93}$$

where $L' = \lceil L/32\sqrt{2}\beta x_{\max}^2\rceil$. Further, using Lemma 3.9 in Eq. (3.93) with $\varepsilon \hookleftarrow \dfrac{\varepsilon_1}{32\sqrt{2}\beta x_{\max}^2}$ and $\delta \hookleftarrow \delta_1$, whenever

$$n \geq \frac{ce^{c'\beta}p_y^2\widetilde{p}^2\Big(\log\frac{p_y}{\delta_1} + \mathcal{M}_\Theta(\frac{\varepsilon_1^2}{\widetilde{p}}) + \mathcal{M}_{\theta,n}(\frac{\varepsilon_1^2}{\widetilde{p}})\Big)}{\varepsilon_1^4},$$

with probability at least $1-\delta_1$, we have,

$$\begin{aligned}\big|\partial_{\omega^{(i)}}(\mathcal{L}^{(i)}(\theta^{\star(i)}))\big| &\leq \frac{1}{L'}\sum_{u\in[L]}\Big(\frac{1}{32\sqrt{2}\beta x_{\max}^2}\varepsilon_1\|\omega^{(i)}\|_1 + \frac{1}{32\sqrt{2}\beta x_{\max}^2}\varepsilon_2\|\omega^{(i)}\|_2^2\Big) \\ &= \frac{L}{32\sqrt{2}\beta x_{\max}^2 L'}\Big(\varepsilon_1\|\omega^{(i)}\|_1 + \varepsilon_2\|\omega^{(i)}\|_2^2\Big) \overset{(a)}{\leq} \varepsilon_1\|\omega^{(i)}\|_1 + \varepsilon_2\|\omega^{(i)}\|_2^2,\end{aligned}$$

where (a) follows because $L' = \lceil L/32\sqrt{2}\beta x_{\max}^2\rceil$.

3.C.1.1.1 Proof of Eq. (3.89): Expression for first directional derivative

Fix any $i \in [n]$. The first-order partial derivatives of $\mathcal{L}^{(i)}$ (defined in Eq. (3.74)) with respect to the entries of the parameter vector $\theta^{(i)}$ are given by

$$\frac{\partial \mathcal{L}^{(i)}(\theta^{(i)})}{\partial \theta_t^{(i)}} = -x_t^{(i)} \exp\Big(- [\theta_t^{(i)} + 2\widehat{\Theta}_{t,-t}^\top \boldsymbol{x}_{-t}^{(i)}] x_t^{(i)} - \widehat{\Theta}_{tt} \overline{x}_t^{(i)} \Big) \quad \text{for all} \quad t \in [p_y].$$

Now, we can write the first-order directional derivative of $\mathcal{L}^{(i)}$ as

$$\begin{aligned} \partial_{\omega^{(i)}}(\mathcal{L}^{(i)}(\theta^{(i)})) \triangleq \lim_{h \to 0} \frac{\mathcal{L}^{(i)}(\theta^{(i)} + h\omega^{(i)}) - \mathcal{L}^{(i)}(\theta^{(i)})}{h} &= \sum_{t \in [p_y]} \omega_t^{(i)} \frac{\partial \mathcal{L}^{(i)}(\theta^{(i)})}{\partial \theta_t^{(i)}} \\ &= - \sum_{t \in [p_y]} \omega_t^{(i)} x_t^{(i)} \exp\Big(- [\theta_t^{(i)} + 2\widehat{\Theta}_{t,-t}^\top \boldsymbol{x}_{-t}^{(i)}] x_t^{(i)} - \widehat{\Theta}_{tt} \overline{x}_t^{(i)} \Big). \end{aligned}$$

3.C.1.1.2 Proof of Lemma 3.9: Upper bound on expected ψ_u

Fix any $i \in [n]$, $u \in [L]$, and $\theta^{(i)} \in \Lambda_\theta$. Then, given $\boldsymbol{x}_{-S_u}^{(i)}$ and $\boldsymbol{z}^{(i)}$, we have

$$\begin{aligned} &\mathbb{E}\Big[\psi_u(\boldsymbol{x}^{(i)}; \omega^{(i)}) \mid \boldsymbol{x}_{-S_u}^{(i)}, \boldsymbol{z}^{(i)}\Big] \\ &\overset{(a)}{=} \mathbb{E}\Big[\sum_{t \in S_u} \omega_t^{(i)} x_t^{(i)} \exp\Big(- [\theta_t^{\star(i)} + 2\widehat{\Theta}_{t,-t}^\top \boldsymbol{x}_{-t}^{(i)}] x_t^{(i)} - \widehat{\Theta}_{tt} \overline{x}_t^{(i)} \Big) \Big| \boldsymbol{x}_{-S_u}^{(i)}, \boldsymbol{z}^{(i)} \Big] \\ &\overset{(b)}{=} \sum_{t \in S_u} \omega_t^{(i)} \mathbb{E}\Big[x_t^{(i)} \exp\Big(- [\theta_t^{\star(i)} + 2\widehat{\Theta}_{t,-t}^\top \boldsymbol{x}_{-t}^{(i)}] x_t^{(i)} - \widehat{\Theta}_{tt} \overline{x}_t^{(i)} \Big) \Big| \boldsymbol{x}_{-S_u}^{(i)}, \boldsymbol{z}^{(i)} \Big] \\ &\overset{(c)}{=} \sum_{t \in S_u} \omega_t^{(i)} \mathbb{E}\Big[\mathbb{E}\Big[x_t^{(i)} \exp\big(- [\theta_t^{\star(i)} + 2\widehat{\Theta}_{t,-t}^\top \boldsymbol{x}_{-t}^{(i)}] x_t^{(i)} - \widehat{\Theta}_{tt} \overline{x}_t^{(i)} \big) \Big| \boldsymbol{x}_{-t}^{(i)}, \boldsymbol{z}^{(i)} \Big] \Big| \boldsymbol{x}_{-S_u}^{(i)}, \boldsymbol{z}^{(i)} \Big], \end{aligned} \tag{3.94}$$

where (a) follows from the definition of $\psi_u(\boldsymbol{x}^{(i)}; \omega^{(i)})$ in Eq. (3.90), (b) follows from linearity of expectation, and (c) follows from the law of total expectation, i.e., $\mathbb{E}[\mathbb{E}[Y|X, Z]|Z] = \mathbb{E}[Y|Z]$ since $\boldsymbol{x}_{-S_u}^{(i)} \subseteq \boldsymbol{x}_{-t}^{(i)}$. Now, we bound $\mathbb{E}\Big[x_t^{(i)} \exp\big(- [\theta_t^{\star(i)} + 2\widehat{\Theta}_{t,-t}^\top \boldsymbol{x}_{-t}^{(i)}] x_t^{(i)} - \widehat{\Theta}_{tt} \overline{x}_t^{(i)} \big) \mid \boldsymbol{x}_{-t}^{(i)}, \boldsymbol{z}^{(i)}\Big]$ for every $t \in S_u$. We have

$$\begin{aligned} &\mathbb{E}\Big[x_t^{(i)} \exp\big(- [\theta_t^{\star(i)} + 2\widehat{\Theta}_{t,-t}^\top \boldsymbol{x}_{-t}^{(i)}] x_t^{(i)} - \widehat{\Theta}_{tt} \overline{x}_t^{(i)} \big) \Big| \boldsymbol{x}_{-t}^{(i)}, \boldsymbol{z}^{(i)} \Big] \\ &= \int_{\mathcal{X}} x_t^{(i)} \exp\big(- [\theta_t^{\star(i)} + 2\widehat{\Theta}_{t,-t}^\top \boldsymbol{x}_{-t}^{(i)}] x_t^{(i)} - \widehat{\Theta}_{tt} \overline{x}_t^{(i)} \big) f_{\mathsf{x}_t | \boldsymbol{\mathsf{x}}_{-t}, \boldsymbol{\mathsf{z}}}\big(x_t^{(i)} | \boldsymbol{x}_{-t}^{(i)}, \boldsymbol{z}^{(i)}; \theta_t^\star(\boldsymbol{z}^{(i)}), \Theta_t^\star \big) dx_t^{(i)} \\ &\overset{(a)}{=} \frac{\int_{\mathcal{X}} x_t^{(i)} \exp\big(2[\Theta_{t,-t}^\star - \widehat{\Theta}_{t,-t}]^\top \boldsymbol{x}_{-t}^{(i)} x_t^{(i)} + [\Theta_{tt}^\star - \widehat{\Theta}_{tt}] \overline{x}_t^{(i)} \big) dx_t^{(i)}}{\int_{\mathcal{X}} \exp\Big([\theta_t^\star(\boldsymbol{z}^{(i)}) + 2\Theta_{t,-t}^{\star\top} \boldsymbol{x}_{-t}^{(i)}] x_t^{(i)} + \Theta_{tt}^\star \overline{x}_t^{(i)} \Big) dx_t^{(i)}} \\ &\overset{(b)}{=} \frac{\int_{\mathcal{X}} x_t^{(i)} \Big[1 + 2[\Theta_{t,-t}^\star - \widehat{\Theta}_{t,-t}]^\top \boldsymbol{x}_{-t}^{(i)} x_t^{(i)} + [\Theta_{tt}^\star - \widehat{\Theta}_{tt}] \overline{x}_t^{(i)} \Big] dx_t^{(i)}}{\int_{\mathcal{X}} \exp\Big([\theta_t^\star(\boldsymbol{z}^{(i)}) + 2\Theta_{t,-t}^{\star\top} \boldsymbol{x}_{-t}^{(i)}] x_t^{(i)} + \Theta_{tt}^\star \overline{x}_t^{(i)} \Big) dx_t^{(i)}} \end{aligned}$$

$$+\frac{\int_{\mathcal{X}} x_t^{(i)}\Big[o\Big([\Theta^{\star}_{t,-t}-\widehat{\Theta}_{t,-t}]^{\top}\boldsymbol{x}_{-t}^{(i)}x_t^{(i)}+[\Theta^{\star}_{tt}-\widehat{\Theta}_{tt}]\overline{x}_t^{(i)}\Big)^2\Big]dx_t^{(i)}}{\int_{\mathcal{X}}\exp\Big([\theta_t^{\star}(\boldsymbol{z}^{(i)})+2\Theta^{\star\top}_{t,-t}\boldsymbol{x}_{-t}^{(i)}]x_t^{(i)}+\Theta^{\star}_{tt}\overline{x}_t^{(i)}\Big)dx_t^{(i)}}$$

$$\overset{(c)}{=}\frac{4x_{\max}^3[\Theta^{\star}_{t,-t}-\widehat{\Theta}_{t,-t}]^{\top}\boldsymbol{x}_{-t}^{(i)}}{3\int_{\mathcal{X}}\exp\Big([\theta_t^{\star}(\boldsymbol{z}^{(i)})+2\Theta^{\star\top}_{t,-t}\boldsymbol{x}_{-t}^{(i)}]x_t^{(i)}+\Theta^{\star}_{tt}\overline{x}_t^{(i)}\Big)dx_t^{(i)}}$$

$$+\frac{x_{\max}^5\big([\Theta^{\star}_{t,-t}-\widehat{\Theta}_{t,-t}]^{\top}\boldsymbol{x}_{-t}^{(i)}\big)\big(\Theta^{\star}_{tt}-\widehat{\Theta}_{tt}\big)o(1)}{\int_{\mathcal{X}}\exp\Big([\theta_t^{\star}(\boldsymbol{z}^{(i)})+2\Theta^{\star\top}_{t,-t}\boldsymbol{x}_{-t}^{(i)}]x_t^{(i)}+\Theta^{\star}_{tt}\overline{x}_t^{(i)}\Big)dx_t^{(i)}}, \tag{3.95}$$

where (a) follows from Eq. (3.10) and $\theta^{\star(i)}=\theta^{\star}(\boldsymbol{z}^{(i)})\ \forall i\in[n]$, (b) follows by using the Taylor series expansion $\exp(y)=1+y+o(y^2)$ around zero, (c) follows because $\int_{\mathcal{X}} x_t^{(i)}dx_t^{(i)}=\int_{\mathcal{X}} x_t^{(i)}\overline{x}_t^{(i)}dx_t^{(i)}=\int_{\mathcal{X}}\big(x_t^{(i)}\big)^3dx_t^{(i)}=\int_{\mathcal{X}} x_t^{(i)}\big(\overline{x}_t^{(i)}\big)^2dx_t^{(i)}=0$, $\int_{\mathcal{X}}\big(x_t^{(i)}\big)^2dx_t^{(i)}=2x_{\max}^3/3$, and $\int_{\mathcal{X}}\big(x_t^{(i)}\big)^2\overline{x}_t^{(i)}dx_t^{(i)}=8x_{\max}^5/45$.

Now, we bound the numerators in Eq. (3.95) by using $\|\Theta_t^{\star}-\widehat{\Theta}_t\|_1\leq\sqrt{\widetilde{p}}\|\Theta_t^{\star}-\widehat{\Theta}_t\|_2$. Then, we invoke Theorem 3.1 to bound $\|\Theta_t^{\star}-\widehat{\Theta}_t\|_2$ by $\varepsilon\mapsfrom\frac{3\varepsilon}{2C_{2,\tau}x_{\max}^3\sqrt{\widetilde{p}}}$. Therefore, we subsume the second term by the first term resulting in the following bound:

$$\mathbb{E}\Big[x_t^{(i)}\exp\big(-[\theta_t^{\star(i)}+2\widehat{\Theta}_{t,-t}^{\top}\boldsymbol{x}_{-t}^{(i)}]x_t^{(i)}-\widehat{\Theta}_{tt}\overline{x}_t^{(i)}\big)\,\big|\,\boldsymbol{x}_{-t}^{(i)},\boldsymbol{z}^{(i)}\Big]\leq\frac{2C_{2,\tau}x_{\max}^3\sqrt{\widetilde{p}}\|\Theta_t^{\star}-\widehat{\Theta}_t\|_2}{3}, \tag{3.96}$$

where we have used the triangle inequality, $\|\boldsymbol{x}^{(i)}\|_{\infty}\leq x_{\max}$ for all $i\in[n]$ as well as $\|\Theta_t^{\star}-\widehat{\Theta}_t\|_1\leq\sqrt{\widetilde{p}}\|\Theta_t^{\star}-\widehat{\Theta}_t\|_2$ to upper bound the numerator, and the arguments used in the proof of Lemma 3.13 as well as $\int_{\mathcal{X}} dx_t^{(i)}=2x_{\max}$ to lower bound the denominator. Using Theorem 3.1 in Eq. (3.96) with $\varepsilon\mapsfrom\dfrac{3\varepsilon}{2C_{2,\tau}x_{\max}^3\sqrt{\widetilde{p}}}$ and $\delta\mapsfrom\delta$, we have

$$\mathbb{E}\Big[x_t^{(i)}\exp\Big(-[\theta_t^{\star(i)}+2\widehat{\Theta}_{t,-t}^{\top}\boldsymbol{x}_{-t}^{(i)}]x_t^{(i)}-\widehat{\Theta}_{tt}\overline{x}_t^{(i)}\Big)\,\Big|\,\boldsymbol{x}_{-t}^{(i)},\boldsymbol{z}^{(i)}\Big]\leq\varepsilon, \tag{3.97}$$

with probability at least $1-\delta$ as long as

$$n\geq\frac{ce^{c'\beta}p_y^2\widetilde{p}^2\Big(\log\frac{p_y}{\delta}+\mathcal{M}_{\Theta}(\frac{\varepsilon^2}{\widetilde{p}})+\mathcal{M}_{\theta,n}(\frac{\varepsilon^2}{\widetilde{p}})\Big)}{\varepsilon^4}. \tag{3.98}$$

Using Eq. (3.97) and triangle inequality in Eq. (3.94), we have

$$\mathbb{E}\Big[\psi_u(\boldsymbol{x}^{(i)};\omega^{(i)})\,\Big|\,\boldsymbol{x}_{-S_u}^{(i)},\boldsymbol{z}^{(i)}\Big]\leq\varepsilon\sum_{t\in S_u}\big|\omega_t^{(i)}\big|\leq\varepsilon\|\omega^{(i)}\|_1,$$

with probability at least $1-\delta$ as long as n satisfies Eq. (3.98).

3.C.1.1.3 Proof of Lemma 3.10: Concentration of ψ_u To show this concentration result, we use Corollary 3.2 Eq. (3.223) for the function q_2. To that end, we note that the tuple of random vectors $(\mathbf{y},\mathbf{a},\mathbf{v},\mathbf{z})$ corresponds to a τ-Sgm (see Definition 3.8) with $\tau \triangleq (\alpha, \beta, \beta, x_{\max}, \Theta)$. However, the random vector $\mathbf{y}$ conditioned on $(\mathbf{a}, \mathbf{v}, \mathbf{z})$ need not satisfy the Dobrushin's uniqueness condition (Definition 3.4). Therefore, we cannot apply Corollary 3.2 Eq. (3.223) as is. To resolve this, we resort to Proposition 3.7 with $\lambda = \frac{1}{4\sqrt{2}x_{\max}^2}$ to reduce the random vector $\mathbf{y}$ conditioned on $(\mathbf{a}, \mathbf{v}, \mathbf{z})$ to Dobrushin's regime.

Fix any $u \in [L]$. Then, from Proposition 3.7(b), (i) the tuple of random vectors $\{\mathbf{y}_{S_u}, \mathbf{a}, \mathbf{v}, (\mathbf{y}_{-S_u}, \mathbf{z})\}$ corresponds to a τ_1-SGM with $\tau_1 \triangleq (\alpha{+}2\beta x_{\max}, \beta, \frac{1}{4\sqrt{2}x_{\max}^2}, x_{\max}, \Theta_{\backslash S_u})$, and (ii) the random vector $\mathbf{y}_{S_u}$ conditioned on $(\mathbf{y}_{-S_u}, \mathbf{a}, \mathbf{v}, \mathbf{z})$ satisfies the Dobrushin's uniqueness condition (Definition 3.4) with coupling matrix $2\sqrt{2}x_{\max}^2|\Theta_{\backslash S_u}|$ such that $2\sqrt{2}x_{\max}^2\||\Theta_{\backslash S_u}|\|_{\mathrm{op}} \leq 2\sqrt{2}x_{\max}^2\lambda \leq 1/2$. Now, for any fixed $i \in [n]$, we apply Corollary 3.2 Eq. (3.223) for the function q_2 with $\varepsilon \hookleftarrow \varepsilon$ for a given $\boldsymbol{x}_{-S_u}^{(i)}$ and $\boldsymbol{z}^{(i)}$, to obtain

$$\mathbb{P}\bigg(\Big|\psi_u(\boldsymbol{x}^{(i)}; \omega^{(i)}) - \mathbb{E}\Big[\psi_u(\boldsymbol{x}^{(i)}; \omega^{(i)}) \;\Big|\; \boldsymbol{x}_{-S_u}^{(i)}, \mathbf{z}\Big]\Big| \geq \varepsilon \;\Big|\; \boldsymbol{x}_{-S_u}^{(i)}, \mathbf{z}\bigg) \leq \exp\bigg(\frac{-\varepsilon^2}{e^{c'\beta}\|\omega^{(i)}\|_2^2}\bigg).$$

3.C.1.2 Proof of Lemma 3.8(b): Anti-concentration of second directional derivative

Fix some $i \in [n]$ and some $\theta^{(i)} \in \Lambda_\theta$. Let $\omega^{(i)}$ be as defined in Eq. (3.85). We claim that the second-order directional derivative of $\mathcal{L}^{(i)}$ defined in Eq. (3.74) is given by

$$\partial^2_{[\omega^{(i)}]^2}\mathcal{L}^{(i)}(\theta^{(i)}) = \sum_{t \in [p_y]} \big(\omega_t^{(i)} x_t^{(i)}\big)^2 \exp\Big(-[\theta_t^{(i)} + 2\widehat{\Theta}_{t,-t}^\top \boldsymbol{x}_{-t}^{(i)}]x_t^{(i)} - \widehat{\Theta}_{tt}\overline{x}_t^{(i)}\Big). \tag{3.99}$$

We provide a proof at the end. For now, we assume the claim and proceed. Now, we lower bound $\partial^2_{[\omega^{(i)}]^2}\mathcal{L}^{(i)}(\theta^{(i)})$ by a quadratic form as follows

$$\begin{aligned}\partial^2_{[\omega^{(i)}]^2}\mathcal{L}^{(i)}(\theta^{(i)}) &\overset{(a)}{\geq} \sum_{t \in [p_y]} \big(\omega_t^{(i)} x_t^{(i)}\big)^2 \times \exp\Big(-\big(|\theta_t^{(i)}|{+}2\|\widehat{\Theta}_t\|_1\|\boldsymbol{x}^{(i)}\|_\infty\big)x_{\max}\Big)\\ &\overset{(b)}{\geq} \sum_{t \in [p_y]} \big(\omega_t^{(i)} x_t^{(i)}\big)^2 \times \exp\Big(-(\alpha{+}2\beta x_{\max})x_{\max}\Big) \overset{Eq.\ (3.55)}{=} \frac{1}{C_{2,\tau}}\sum_{t \in [p_y]} \big(\omega_t^{(i)} x_t^{(i)}\big)^2,\end{aligned} \tag{3.100}$$

where (a) follows from Eq. (3.99) by triangle inequality, Cauchy–Schwarz inequality, and because $\|\boldsymbol{x}^{(i)}\|_\infty \leq x_{\max}$ for all $i \in [n]$, and (b) follows because $\widehat{\Theta} \in \Lambda_\Theta$, $\theta^{(i)} \in \Lambda_\theta$, and $\|\boldsymbol{x}^{(i)}\|_\infty \leq x_{\max}$ for all $i \in [n]$.

Now, to show the anti-concentration of $\partial^2_{[\omega^{(i)}]^2}\mathcal{L}^{(i)}(\theta^{(i)})$, we show the anti-concentration of the quadratic form in Eq. (3.100). To that end, we note that the tuple of random vectors $(\mathbf{y}, \mathbf{a}, \mathbf{v}, \mathbf{z})$ corresponds to a τ-SGM (see Definition 3.8) with $\tau \triangleq (\alpha, \beta, \beta, x_{\max}, \Theta)$. Then, we decompose the quadratic form in Eq. (3.100) as a sum of $L = 1024\beta^2 x_{\max}^4 \log 4p_y$ terms using Proposition 3.7 (see Section 3.H) with $\lambda = \frac{1}{4\sqrt{2}x_{\max}^2}$ and focus on these L

terms. Consider the L subsets $S_1, \cdots, S_L \in [p_y]$ obtained from Proposition 3.7 and define

$$\overline{\psi}_u(\boldsymbol{x}^{(i)}; \omega^{(i)}) \triangleq \sum_{t \in S_u} \big(\omega_t^{(i)} x_t^{(i)}\big)^2 \quad \text{for every} \quad u \in L. \tag{3.101}$$

Then, we have

$$\sum_{t \in [p_y]} \big(\omega_t^{(i)} x_t^{(i)}\big)^2 \overset{(a)}{=} \frac{1}{L'} \sum_{u \in [L]} \sum_{t \in S_u} \big(\omega_t^{(i)} x_t^{(i)}\big)^2 \overset{Eq.~(3.101)}{=} \frac{1}{L'} \sum_{u \in [L]} \overline{\psi}_u(\boldsymbol{x}^{(i)}; \omega^{(i)}), \tag{3.102}$$

where (a) follows because each $t \in [p_y]$ appears in exactly $L' = \lceil L/32\sqrt{2}\beta x_{\max}^2 \rceil$ of the sets $S_1, \cdots, S_L$ according to Proposition 3.7(a) (with $\lambda = \frac{1}{4\sqrt{2}x_{\max}^2}$). Now, we focus on the L terms in Eq. (3.102).

Consider any $u \in [L]$. We claim that conditioned on $\boldsymbol{x}_{-S_u}^{(i)}$ and $\boldsymbol{z}^{(i)}$, the expected value of $\overline{\psi}_u(\boldsymbol{x}^{(i)}; \omega^{(i)})$ can be upper bounded uniformly across all $u \in [L]$. We provide a proof at the end.

Lemma 3.11 (Lower bound on expected $\overline{\psi}_u$)**.** *Fix $i \in [n]$ and $\theta^{(i)} \in \Lambda_\theta$. Then, with $\omega^{(i)}$ defined in Eq. (3.85) and given $\boldsymbol{z}^{(i)}$ and $\boldsymbol{x}_{-S_u}^{(i)}$, we have*

$$\min_{u \in [L]} \mathbb{E}\Big[\overline{\psi}_u(\boldsymbol{x}^{(i)}; \omega^{(i)}) \;\big|\; \boldsymbol{x}_{-S_u}^{(i)}, \boldsymbol{z}^{(i)}\Big] \geq \frac{2x_{\max}^2}{\pi e C_{2,\tau}^4} \|\omega^{(i)}\|_2^2,$$

where the constant $C_{2,\tau}$ was defined in Eq. (3.55).

Consider again any $u \in [L]$. Now, we claim that conditioned on $\boldsymbol{x}_{-S_u}^{(i)}$ and $\boldsymbol{z}^{(i)}$, $\overline{\psi}_u(\boldsymbol{x}^{(i)}; \omega^{(i)})$ concentrates around its conditional expected value. We provide a proof at the end.

Lemma 3.12 (Concentration of $\overline{\psi}_u$)**.** *Fix $\varepsilon > 0$, $i \in [n]$, $u \in [L]$, and $\theta^{(i)} \in \Lambda_\theta$. Then, with $\omega^{(i)}$ defined in Eq. (3.85) and given $\boldsymbol{z}^{(i)}$ and $\boldsymbol{x}_{-S_u}^{(i)}$, we have*

$$\Big|\overline{\psi}_u(\boldsymbol{x}^{(i)}; \omega^{(i)}) - \mathbb{E}\big[\overline{\psi}_u(\boldsymbol{x}^{(i)}; \omega^{(i)}) \;\big|\; \boldsymbol{x}_{-S_u}^{(i)}, \boldsymbol{z}^{(i)}\big]\Big| \leq \varepsilon,$$

with probability at least $1 - \exp\left(\dfrac{-\varepsilon^2}{e^{c'\beta}\|\omega^{(i)}\|_2^2}\right)$.

Given these lemmas, we proceed to show the anti-concentration of the quadratic form in Eq. (3.100) implying the anti-concentration of $\partial^2_{[\omega^{(i)}]^2}\mathcal{L}^{(i)}(\theta^{(i)})$. To that end, for any $u \in [L]$, given $\boldsymbol{x}_{-S_u}^{(i)}$ and $\boldsymbol{z}^{(i)}$, let E_u denote the event that

$$\overline{\psi}_u(\boldsymbol{x}^{(i)}; \omega^{(i)}) \geq \mathbb{E}\big[\overline{\psi}_u(\boldsymbol{x}^{(i)}; \omega^{(i)}) | \boldsymbol{x}_{-S_u}^{(i)}, \boldsymbol{z}^{(i)}\big] - \frac{x_{\max}^2}{\pi e C_{2,\tau}^4} \|\omega^{(i)}\|_2^2. \tag{3.103}$$

Since E_u in an indicator event, using the law of total expectation results in

$$\mathbb{P}(E_u) = \mathbb{E}\Big[\mathbb{P}(E_u|\boldsymbol{x}_{-S_u}^{(i)}, \boldsymbol{z}^{(i)})\Big] \overset{(a)}{\geq} 1 - \exp\left(\frac{\|\omega^{(i)}\|_2^2}{e^{c'\beta}}\right),$$

where (a) follows from Lemma 3.12 with $\varepsilon \hookleftarrow \dfrac{x_{\max}^2}{\pi e C_{2,\tau}^4}\|\omega^{(i)}\|_2^2$. Now, by applying the union bound over all $u \in [L]$ where $L = 1024\beta^2 x_{\max}^4 \log 4p_y$, we have

$$\mathbb{P}\Big(\bigcap_{u\in L} E_u\Big) \geq 1 - O\Big(\beta^2 \log p_y \exp\left(\frac{\|\omega^{(i)}\|_2^2}{e^{c'\beta}}\right)\Big).$$

Now, assume the event $\cap_{u\in L} E_u$ holds. Whenever this holds, we also have

$$\begin{aligned}\sum_{t\in[p_y]} \big(\omega_t^{(i)} x_t^{(i)}\big)^2 &\overset{Eq.\ (3.102)}{=} \frac{1}{L'}\sum_{u\in[L]} \overline{\psi}_u(\boldsymbol{x}^{(i)}; \omega^{(i)}) \\ &\overset{Eq.\ (3.103)}{\geq} \frac{1}{L'}\sum_{u\in[L]}\Big(\mathbb{E}\big[\overline{\psi}_u(\boldsymbol{x}^{(i)}; \omega^{(i)})|\boldsymbol{x}_{-S_u}^{(i)}, \boldsymbol{z}^{(i)}\big] - \frac{x_{\max}^2}{\pi e C_{2,\tau}^4}\|\omega^{(i)}\|_2^2\Big) \\ &\overset{(a)}{\geq} \frac{1}{L'}\sum_{u\in[L]} \frac{x_{\max}^2}{\pi e C_{2,\tau}^4}\|\omega^{(i)}\|_2^2 = \frac{x_{\max}^2 L}{\pi e L' C_{2,\tau}^4}\|\omega^{(i)}\|_2^2, \end{aligned} \tag{3.104}$$

where $L' = \lceil L/32\sqrt{2}\beta x_{\max}^2 \rceil$ and (a) follows from Lemma 3.11. Finally, approximating $L' = L/32\sqrt{2}\beta x_{\max}^2$ and using Eq. (3.100), we have

$$\partial^2_{[\omega^{(i)}]^2}\mathcal{L}^{(i)}(\theta^{(i)}) \geq \frac{1}{C_{2,\tau}}\sum_{t\in[p_y]} \big(\omega_t^{(i)} x_t^{(i)}\big)^2 \overset{Eq.\ (3.104)}{\geq} \frac{32\sqrt{2}\beta x_{\max}^4}{\pi e C_{2,\tau}^5}\|\omega^{(i)}\|_2^2,$$

which completes the proof.

3.C.1.2.1 Proof of Eq. (3.99): Expression for second directional derivative

Fix any $i \in [n]$. The second-order partial derivatives of $\mathcal{L}^{(i)}$ (defined in Eq. (3.74)) with respect to the entries of the parameter vector $\theta^{(i)}$ are given by

$$\frac{\partial^2 \mathcal{L}^{(i)}(\theta^{(i)})}{\partial\big[\theta_t^{(i)}\big]^2} = \big[x_t^{(i)}\big]^2 \exp\Big(-[\theta_t^{(i)} + 2\widehat{\Theta}_{t,-t}^\top \boldsymbol{x}_{-t}^{(i)}]x_t^{(i)} - \widehat{\Theta}_{tt}\overline{x}_t^{(i)}\Big) \quad \text{for all} \quad t \in [p_y].$$

Now, we can write the second-order directional derivative of $\mathcal{L}^{(i)}$ as

$$\begin{aligned}\partial^2_{[\omega^{(i)}]^2}\mathcal{L}^{(i)}(\theta^{(i)}) &\triangleq \lim_{h\to 0} \frac{\partial_{\omega^{(i)}}\mathcal{L}^{(i)}(\theta^{(i)} + h\omega^{(i)}) - \partial_{\omega^{(i)}}\mathcal{L}^{(i)}(\theta^{(i)})}{h} = \sum_{t\in[p_y]} \big[\omega_t^{(i)}\big]^2 \frac{\partial^2\mathcal{L}^{(i)}(\theta^{(i)})}{\partial\big[\theta_t^{(i)}\big]^2} \\ &= \sum_{t\in[p_y]} \big(\omega_t^{(i)} x_t^{(i)}\big)^2 \exp\Big(-[\theta_t^{(i)} + 2\widehat{\Theta}_{t,-t}^\top \boldsymbol{x}_{-t}^{(i)}]x_t^{(i)} - \widehat{\Theta}_{tt}\overline{x}_t^{(i)}\Big).\end{aligned}$$

3.C.1.2.2 Proof of Lemma 3.11: Lower bound on expected $\overline{\psi}_u$

First, we claim that the conditional variance of $x_t^{(i)}$ conditioned on $\mathbf{x}_{-t} = \boldsymbol{x}_{-t}^{(i)}$ and $\mathbf{z} = \boldsymbol{z}^{(i)}$ is lower bounded by a constant for every $t \in [p_y]$ and $i \in [n]$. We provide a proof at this end of this section.

Lemma 3.13 (Lower bound on the conditional variance)**.** *We have*

$$\mathbb{V}ar\big(x_t^{(i)} \big| \boldsymbol{x}_{-t}^{(i)}, \boldsymbol{z}^{(i)}\big) \geq \frac{2x_{\max}^2}{\pi e C_{2,\tau}^4} \qquad \textit{for all } t \in [p_y] \textit{ and } i \in [n],$$

where the constant $C_{2,\tau}$ was defined in Eq. (3.55).

Given this lemma, we proceed. Fix any $i \in [n]$, $u \in [L]$, and $\theta^{(i)} \in \Lambda_\theta$. Then, given $\boldsymbol{x}_{-S_u}^{(i)}$ and $\boldsymbol{z}^{(i)}$, we have

$$\begin{aligned}
\mathbb{E}\Big[\overline{\psi}_u(\boldsymbol{x}^{(i)}; \omega^{(i)}) \mid \boldsymbol{x}_{-S_u}^{(i)}, \boldsymbol{z}^{(i)}\Big] &\overset{Eq.\ (3.101)}{=} \mathbb{E}\Big[\sum_{t \in S_u} \big(\omega_t^{(i)} x_t^{(i)}\big)^2 \mid \boldsymbol{x}_{-S_u}^{(i)}, \boldsymbol{z}^{(i)}\Big] \\
&\overset{(a)}{=} \sum_{t \in S_u} \mathbb{E}\Big[\big(\omega_t^{(i)} x_t^{(i)}\big)^2 \mid \boldsymbol{x}_{-S_u}^{(i)}, \boldsymbol{z}^{(i)}\Big] \\
&\overset{(b)}{=} \sum_{t \in S_u} \mathbb{E}\Big[\mathbb{E}\Big[\big(\omega_t^{(i)} x_t^{(i)}\big)^2 \Big| \boldsymbol{x}_{-t}^{(i)}, \boldsymbol{z}^{(i)}\Big] \mid \boldsymbol{x}_{-S_u}^{(i)}, \boldsymbol{z}^{(i)}\Big] \\
&\overset{(c)}{\geq} \sum_{t \in S_u} \mathbb{E}\Big[\mathbb{V}\mathrm{ar}\Big(\omega_t^{(i)} x_t^{(i)} \Big| \boldsymbol{x}_{-t}^{(i)}, \boldsymbol{z}^{(i)}\Big) \mid \boldsymbol{x}_{-S_u}^{(i)}, \boldsymbol{z}^{(i)}\Big] \\
&\overset{(d)}{\geq} \frac{2x_{\max}^2}{\pi e C_{2,\tau}^4} \|\omega^{(i)}\|_2^2,
\end{aligned}$$

where (a) follows from linearity of expectation, (b) follows from the law of total expectation i.e., $\mathbb{E}[\mathbb{E}[Y|X,Z]|Z] = \mathbb{E}[Y|Z]$ since $\boldsymbol{x}_{-S_u}^{(i)} \subseteq \boldsymbol{x}_{-t}^{(i)}$, (c) follows follows from the fact that for any random variable a, $\mathbb{E}[a^2] \geq \mathbb{V}\mathrm{ar}[a]$, and (d) follows from Lemma 3.13.

3.C.1.2.3 Proof of Lemma 3.12: Concentration of $\overline{\psi}_u$

To show this concentration result, we use Corollary 3.2 Eq. (3.223) for the function q_1. To that end, we note that the tuple of random vectors $(\mathbf{y}, \mathbf{a}, \mathbf{v}, \mathbf{z})$ corresponds to a τ-SGM (see Definition 3.8) with $\tau \triangleq (\alpha, \beta, \beta, x_{\max}, \Theta)$. However, the random vector $\mathbf{y}$ conditioned on $(\mathbf{a}, \mathbf{v}, \mathbf{z})$ need not satisfy the Dobrushin's uniqueness condition (Definition 3.4). Therefore, we cannot apply Corollary 3.2 Eq. (3.223) as is. To resolve this, we resort to Proposition 3.7 with $\lambda = \frac{1}{4\sqrt{2}x_{\max}^2}$ to reduce the random vector $\mathbf{y}$ conditioned on $(\mathbf{a}, \mathbf{v}, \mathbf{z})$ to Dobrushin's regime.

Fix any $u \in [L]$. Then, from Proposition 3.7(b), (i) the tuple of random vectors $\{\mathbf{y}_{S_u}, \mathbf{a}, \mathbf{v}, (\mathbf{y}_{-S_u}, \mathbf{z})\}$ corresponds to a τ_1-SGM with $\tau_1 \triangleq (\alpha{+}2\beta x_{\max}, \beta, \frac{1}{4\sqrt{2}x_{\max}^2}, x_{\max}, \Theta_{\backslash S_u})$, and (ii) the random vector $\mathbf{y}_{S_u}$ conditioned on $(\mathbf{y}_{-S_u}, \mathbf{a}, \mathbf{v}, \mathbf{z})$ satisfies the Dobrushin's uniqueness condition (Definition 3.4) with coupling matrix $2\sqrt{2}x_{\max}^2|\Theta_{\backslash S_u}|$ such that

$2\sqrt{2}x_{\max}^2 |||\Theta_{\setminus S_u}|||_{\mathrm{op}} \leq 2\sqrt{2}x_{\max}^2\lambda \leq 1/2$. Now, for any fixed $i \in [n]$, we apply Corollary 3.2 Eq. (3.223) for the function q_1 with $\varepsilon = \varepsilon$ for a given $\boldsymbol{x}_{-S_u}^{(i)}$ and $\boldsymbol{z}^{(i)}$, to obtain

$$\mathbb{P}\bigg(\Big|\overline{\psi}_u(\boldsymbol{x}^{(i)};\omega^{(i)}) - \mathbb{E}\Big[\overline{\psi}_u(\boldsymbol{x}^{(i)};\omega^{(i)}) \;\Big|\; \boldsymbol{x}_{-S_u}^{(i)}, \boldsymbol{z}^{(i)}\Big]\Big| \geq \varepsilon \;\Big|\; \boldsymbol{x}_{-S_u}^{(i)}, \boldsymbol{z}^{(i)}\bigg) \leq \exp\bigg(\frac{-\varepsilon^2}{e^{c'\beta}\|\omega^{(i)}\|_2^2}\bigg).$$

3.C.1.2.4 Proof of Lemma 3.13: Lower bound on the conditional variance

For any random variable x, let $h(\mathsf{x})$ denote the differential entropy of x. Fix any $t \in [p_y]$ and $i \in [n]$. Then, from Shannon's entropy inequality $(2h(\cdot) \leq \log\sqrt{2\pi e\mathbb{V}\mathrm{ar}(\cdot)})$, we have

$$2\pi e\mathbb{V}\mathrm{ar}\big(x_t^{(i)}\big|\boldsymbol{x}_{-t}^{(i)}, \boldsymbol{z}^{(i)}\big) \geq \exp\Big(2h\big(x_t^{(i)}\big|\boldsymbol{x}_{-t}^{(i)}, \boldsymbol{z}^{(i)}\big)\Big). \tag{3.105}$$

Therefore, to bound the variance, it suffices to bound the differential entropy. We have

$$\begin{aligned}
&-h\big(x_t^{(i)}\big|\boldsymbol{x}_{-t}^{(i)}, \boldsymbol{z}^{(i)}\big)\\
&= \int\limits_{\mathcal{X}^{\widetilde{p}}\times\mathcal{Z}^{p_z}} f_{\mathbf{x},\mathbf{z}}(\boldsymbol{x}^{(i)}, \boldsymbol{z}^{(i)}) \log\Big(f_{\mathsf{x}_t|\mathbf{x}_{-t},\mathbf{z}}\big(x_t^{(i)}|\boldsymbol{x}_{-t}^{(i)}, \boldsymbol{z}^{(i)}; \theta_t^\star(\boldsymbol{z}^{(i)}), \Theta_t^\star\big)\Big) d\boldsymbol{x}^{(i)} d\boldsymbol{z}^{(i)}\\
&= \int\limits_{\mathcal{X}^{\widetilde{p}}\times\mathcal{Z}^{p_z}} f_{\mathbf{x},\mathbf{z}}(\boldsymbol{x}^{(i)}, \boldsymbol{z}^{(i)}) \log\bigg(\frac{\exp\big([\theta_t^\star(\boldsymbol{z}^{(i)}) + 2\Theta_{t,-t}^{\star\top}\boldsymbol{x}_{-t}^{(i)}]x_t^{(i)} + \Theta_{tt}^\star\overline{x}_t^{(i)}\big)}{\int_{\mathcal{X}}\exp\big([\theta_t^\star(\boldsymbol{z}^{(i)}) + 2\Theta_{t,-t}^{\star\top}\boldsymbol{x}_{-t}^{(i)}]x_t^{(i)} + \Theta_{tt}^\star\overline{x}_t^{(i)}\big)dx_t^{(i)}}\bigg) d\boldsymbol{x}^{(i)} d\boldsymbol{z}^{(i)}\\
&\overset{(a)}{\geq} \int\limits_{\mathcal{X}^{\widetilde{p}}\times\mathcal{Z}^{p_z}} f_{\mathbf{x},\mathbf{z}}(\boldsymbol{x}^{(i)}, \boldsymbol{z}^{(i)}) \log\bigg(\frac{\exp\big((|\theta_t^\star(\boldsymbol{z}^{(i)})| + 2\|\Theta_t^\star\|_1\|\boldsymbol{x}^{(i)}\|_\infty)x_{\max}\big)}{\int_{\mathcal{X}}\exp\big(-(|\theta_t^\star(\boldsymbol{z}^{(i)})| + 2\|\Theta_t^\star\|_1\|\boldsymbol{x}^{(i)}\|_\infty)x_{\max}\big)dx_t^{(i)}}\bigg) d\boldsymbol{x}^{(i)} d\boldsymbol{z}^{(i)}\\
&\overset{(b)}{\geq} \int\limits_{\mathcal{X}^{\widetilde{p}}\times\mathcal{Z}^{p_z}} f_{\mathbf{x},\mathbf{z}}(\boldsymbol{x}^{(i)}, \boldsymbol{z}^{(i)}) \log\bigg(\frac{\exp\big((\alpha + 2\beta x_{\max})x_{\max}\big)}{\int_{\mathcal{X}}\exp\big(-(\alpha + 2\beta x_{\max})x_{\max}\big)dx_t^{(i)}}\bigg) d\boldsymbol{x}^{(i)} d\boldsymbol{z}^{(i)}\\
&\overset{(c)}{=} \int\limits_{\mathcal{X}^{\widetilde{p}}\times\mathcal{Z}^{p_z}} f_{\mathbf{x},\mathbf{z}}(\boldsymbol{x}^{(i)}, \boldsymbol{z}^{(i)}) \log\bigg(\frac{C_{2,\tau}^2}{2x_{\max}}\bigg) d\boldsymbol{x}^{(i)} d\boldsymbol{z}^{(i)} = \log\bigg(\frac{C_{2,\tau}^2}{2x_{\max}}\bigg),
\end{aligned} \tag{3.106}$$

where (a) follows from triangle inequality and Cauchy–Schwarz inequality and because $\|\boldsymbol{x}^{(i)}\|_\infty \leq x_{\max}$ for all $i \in [n]$, (b) follows because $\theta^\star(\boldsymbol{z}^{(i)}) \in \Lambda_\theta$ for all $i \in [n]$, $\Theta^\star \in \Lambda_\Theta$, $\|\boldsymbol{x}^{(i)}\|_\infty \leq x_{\max}$ for all $i \in [n]$, and (c) follows because $\int_{\mathcal{X}} dx_t^{(i)} = 2x_{\max}$. Combining Eqs. (3.105) and (3.106) completes the proof.

3.C.2 Proof of Lemma 3.7: Lipschitzness of the loss function

Fix any $i \in [n]$, any $\theta^{(i)}, \widetilde{\theta}^{(i)} \in \Lambda_\theta$. Consider the direction $\omega^{(i)} = \widetilde{\theta}^{(i)} - \theta^{(i)}$, and define the function $q : [0,1] \to \mathbb{R}$ as follows

$$q(a) = \mathcal{L}^{(i)}\big(\theta^{(i)} + a(\widetilde{\theta}^{(i)} - \theta^{(i)})\big). \tag{3.107}$$

Then, the desired inequality in Eq. (3.76) is equivalent to

$$|q(1) - q(0)| \leq x_{\max} C_{2,\tau} \|\omega^{(i)}\|_1.$$

From the mean value theorem, there exists $a' \in (0, 1)$ such that

$$|q(1) - q(0)| = \left|\frac{dq(a')}{da}\right|. \tag{3.108}$$

Therefore, we have

$$\begin{aligned} \big|q(1) - q(0)\big| &\overset{Eq.~(3.108)}{=} \left|\frac{dq(a')}{da}\right| \overset{Eq.~(3.107)}{=} \left|\frac{d\mathcal{L}^{(i)}\big(\theta^{(i)} + a'(\widetilde{\theta}^{(i)} - \theta^{(i)})\big)}{da}\right| \\ &\overset{Eq.~(3.86)}{=} \left|\partial_{\omega^{(i)}}(\mathcal{L}^{(i)}(\theta^{(i)}))\big|_{\theta^{(i)} = \theta^{(i)} + a'(\widetilde{\theta}^{(i)} - \theta^{(i)})}\right|. \end{aligned} \tag{3.109}$$

Using Eq. (3.89) in Eq. (3.109), we have

$$\begin{aligned} \big|q(1) - q(0)\big| &= \Big|\sum_{t \in [p_y]} \omega_t^{(i)} x_t^{(i)} \exp\Big(-[\theta_t^{(i)} + a'(\widetilde{\theta}_t^{(i)} - \theta_t^{(i)}) + 2\widehat{\Theta}_{t,-t}^\top \boldsymbol{x}_t^{(i)}] x_t^{(i)} - \widehat{\Theta}_{tt} \overline{x}_t^{(i)}\Big)\Big| \\ &\overset{(a)}{\leq} x_{\max} \sum_{t \in [p_y]} \big|\omega_t^{(i)}\big| \exp\Big(\Big[\big|(1 - a')\theta_t^{(i)}\big| + \big|a'\widetilde{\theta}_t^{(i)}\big| + 2\|\widehat{\Theta}_t\|_1 \|\boldsymbol{x}^{(i)}\|_\infty\Big] x_{\max}\Big) \\ &\overset{(b)}{\leq} x_{\max} \exp\Big(\big((1 - a')\alpha + a'\alpha + 2\beta x_{\max}\big) x_{\max}\Big) \sum_{t \in [p_y]} \big|\omega_t^{(i)}\big| \\ &\overset{Eq.~(3.55)}{=} x_{\max} C_{2,\tau} \|\omega^{(i)}\|_1, \end{aligned}$$

where (a) follows from triangle inequality, Cauchy–Schwarz inequality, and because $\|\boldsymbol{x}^{(i)}\|_\infty \leq x_{\max}$ for all $i \in [n]$ and (b) follows because $\theta^{(i)}, \widetilde{\theta}^{(i)} \in \Lambda_\theta$, $\widehat{\Theta} \in \Lambda_\Theta$, and $\|\boldsymbol{x}^{(i)}\|_\infty \leq x_{\max}$ for all $i \in [n]$.

3.D Proof of Theorem 3.2: Guarantee on quality of outcome estimate

Fix any unit $i \in [n]$ and an alternate intervention $\widetilde{\boldsymbol{a}}^{(i)} \in \mathcal{A}^{p_a}$. Then, we have

$$\mu^{(i)}(\widetilde{\boldsymbol{a}}^{(i)}) \overset{Eq.~(3.6)}{=} \mathbb{E}[\boldsymbol{y}^{(i)}(\widetilde{\boldsymbol{a}}^{(i)}) | \mathbf{v} = \boldsymbol{v}^{(i)}, \mathbf{z} = \boldsymbol{z}^{(i)}] \overset{(a)}{=} \mathbb{E}[\mathbf{y} | \mathbf{a} = \widetilde{\boldsymbol{a}}^{(i)}, \mathbf{v} = \boldsymbol{v}^{(i)}, \mathbf{z} = \boldsymbol{z}^{(i)}],$$

where (a) follows because the unit-level counterfactual distribution is equivalent to unit-level conditional distribution under the causal framework considered as described in Section 3.3.1. To obtain a convenient expression for $\mathbb{E}[\mathbf{y} | \mathbf{a} = \widetilde{\boldsymbol{a}}^{(i)}, \mathbf{v} = \boldsymbol{v}^{(i)}, \mathbf{z} = \boldsymbol{z}^{(i)}]$, we identify $\Phi^{\star(u,y)} \in \mathbb{R}^{p_u \times p_y}$ to be the component of $\Theta^\star$ corresponding to $\mathbf{u}$ and $\mathbf{y}$ for all $\mathbf{u} \in \{\mathbf{v}, \mathbf{a}, \mathbf{y}\}$. Then, the conditional distribution of $\mathbf{y}$ as a function of the interventions

$\mathbf{a}$, while keeping $\mathbf{v}$ and $\mathbf{z}$ fixed at the corresponding realizations for unit i, i.e., $\boldsymbol{v}^{(i)}$ and $\boldsymbol{z}^{(i)}$, respectively, can be written as

$$f^{(i)}_{\mathbf{y}|\mathbf{a}}(\boldsymbol{y}|\boldsymbol{a}) \propto \exp\Big(\big[\theta^{\star(i)} + 2\boldsymbol{v}^{(i)\top}\Phi^{\star(v,y)} + 2\boldsymbol{a}^{\top}\Phi^{\star(a,y)}\big]\boldsymbol{y} + \boldsymbol{y}^{\top}\Phi^{\star(y,y)}\boldsymbol{y}\Big). \tag{3.110}$$

Therefore, we have

$$\mathbb{E}[\mathbf{y}|\mathbf{a} = \widetilde{\boldsymbol{a}}^{(i)}, \mathbf{v} = \boldsymbol{v}^{(i)}, \mathbf{z} = \boldsymbol{z}^{(i)}] = \mathbb{E}_{f^{(i)}_{\mathbf{y}|\mathbf{a}}}[\mathbf{y}|\mathbf{a} = \widetilde{\boldsymbol{a}}^{(i)}].$$

Now, consider the p_w dimensional random vector $\mathbf{w}$ supported on $\mathcal{X}^{p_w}$ with distribution $f_{\mathbf{w}}$ parameterized by $\psi \in \mathbb{R}^{p_w}$ and $\Psi \in \mathbb{R}^{p_w \times p_w}$ as follows

$$f_{\mathbf{w}}(\boldsymbol{w}|\psi, \Psi) \propto \exp(\psi^{\top}\boldsymbol{w} + \boldsymbol{w}^{\top}\Psi\boldsymbol{w}). \tag{3.111}$$

Then, note that $\widehat{f}^{(i)}_{\mathbf{y}|\mathbf{a}}(\boldsymbol{y}|\boldsymbol{a})$ in Eq. (3.13) and $f^{(i)}_{\mathbf{y}|\mathbf{a}}(\boldsymbol{y}|\boldsymbol{a})$ in Eq. (3.110) belong to the set $\{f_{\mathbf{w}}(\cdot|\psi, \Psi) : \psi \in \mathbb{R}^{p_w}, \Psi \in \mathbb{R}^{p_w \times p_w}\}$ for some ψ and Ψ. Now, we consider any two distributions in this set, namely $f_{\mathbf{w}}(\boldsymbol{w}|\widehat{\psi}, \widehat{\Psi})$ and $f_{\mathbf{w}}(\boldsymbol{w}|\psi^{\star}, \Psi^{\star})$. Then, we claim that the two norm of the difference of the mean vectors of these distributions is bounded as below. We provide a proof at the end.

Lemma 3.14 (Perturbation in the mean vector)**.** *For any $\psi \in \mathbb{R}^{p_w}$ and $\Psi \in \mathbb{R}^{p_w \times p_w}$, let $\mu_{\psi,\Psi}(\mathbf{w}) \in \mathbb{R}^{p_w}$ and $\mathbb{C}\mathrm{ov}_{\psi,\Psi}(\mathbf{w}, \mathbf{w}) \in \mathbb{R}^{p_w \times p_w}$ denote the mean vector and the covariance matrix of $\mathbf{w}$, respectively, with respect to $f_{\mathbf{w}}$ in Eq. (3.111). Then, for any $\widehat{\psi}, \psi^{\star} \in \mathbb{R}^{p_w}$ and $\widehat{\Psi}, \Psi^{\star} \in \mathbb{R}^{p_w \times p_w}$, there exists some $t \in (0, 1)$, $\widetilde{\psi} \triangleq t\widehat{\psi} + (1-t)\psi^{\star}$ and $\widetilde{\Psi} \triangleq t\widetilde{\psi} + (1-t)\widetilde{\psi}$ such that*

$$\begin{aligned}\|\mu_{\widehat{\psi},\widehat{\Psi}}(\mathbf{w}) - \mu_{\psi^{\star},\Psi^{\star}}(\mathbf{w})\|_2 \leq \|\mathbb{C}\mathrm{ov}_{\widetilde{\psi},\widetilde{\Psi}}(\mathbf{w}, \mathbf{w})\|_{\mathrm{op}}\|(\widehat{\psi} - \psi^{\star})\|_2 \\ + \sum_{t_3 \in [p_w]} \|\mathbb{C}\mathrm{ov}_{\widetilde{\psi},\widetilde{\Psi}}(\mathbf{w}, w_{t_3}\mathbf{w})\|_{\mathrm{op}}\|(\widehat{\Psi}_{t_3} - \Psi^{\star}_{t_3})\|_2.\end{aligned}$$

Given this lemma, we proceed with the proof. By applying this lemma to $\widehat{f}^{(i)}_{\mathbf{y}|\mathbf{a}}(\boldsymbol{y}|\boldsymbol{a})$ in Eq. (3.13) and $f^{(i)}_{\mathbf{y}|\mathbf{a}}(\boldsymbol{y}|\boldsymbol{a})$ in Eq. (3.110), we see that it is sufficient to show the following bound

$$\begin{aligned}\|(\theta^{\star(i)} - \widehat{\theta}^{(i)}) + 2\boldsymbol{v}^{(i)\top}(\Phi^{\star(v,y)} - \widehat{\Phi}^{(v,y)}) + 2\widetilde{\boldsymbol{a}}^{(i)\top}(\Phi^{\star(a,y)} - \widehat{\Phi}^{(a,y)})\|_2 \\ + \sum_{t \in [p_y]} \|\Phi^{\star(y,y)}_t - \widehat{\Phi}^{(y,y)}_t\|_2 \leq R(\varepsilon, \delta/n) + p\varepsilon.\end{aligned}$$

To that end, we have

$$\sum_{t \in [p_y]} \|\Phi^{\star(y,y)}_t - \widehat{\Phi}^{(y,y)}_t\|_2 \overset{(a)}{\leq} \sum_{t \in [p_y]} \|\Theta^{\star}_t - \widehat{\Theta}_t\|_2, \tag{3.112}$$

where (a) follows because ℓ_2 norm of any sub-vector is no more than ℓ_2 norm of the vector. Similarly, we have

$$\|(\theta^{\star(i)} - \widehat{\theta}^{(i)}) + 2\boldsymbol{v}^{(i)\top}(\Phi^{\star(v,y)} - \widehat{\Phi}^{(v,y)}) + 2\widetilde{\boldsymbol{a}}^{(i)\top}(\Phi^{\star(a,y)} - \widehat{\Phi}^{(a,y)})\|_2$$

$$\begin{aligned}
&\overset{(a)}{\le} \|\theta^{\star(i)} - \widehat{\theta}^{(i)}\|_2 + 2\|\boldsymbol{v}^{(i)\top}(\Phi^{\star(v,y)} - \widehat{\Phi}^{(v,y)})\|_2 + 2\|\widetilde{\boldsymbol{a}}^{(i)\top}(\Phi^{\star(a,y)} - \widehat{\Phi}^{(a,y)})\|_2 \\
&\overset{(b)}{\le} \|\theta^{\star(i)} - \widehat{\theta}^{(i)}\|_2 + 2\|\boldsymbol{v}^{(i)}\|_2 \|\Phi^{\star(v,y)} - \widehat{\Phi}^{(v,y)}\|_{\mathrm{op}} + 2\|\widetilde{\boldsymbol{a}}^{(i)}\|_2 \|(\Phi^{\star(a,y)} - \widehat{\Phi}^{(a,y)})\|_{\mathrm{op}} \\
&\overset{(c)}{\le} \|\theta^{\star(i)} - \widehat{\theta}^{(i)}\|_2 + 2\Big(\|\boldsymbol{v}^{(i)}\|_2 + \|\widetilde{\boldsymbol{a}}^{(i)}\|_2\Big)\|\Theta^\star - \widehat{\Theta}\|_{\mathrm{op}} \\
&\overset{(d)}{\le} \|\theta^{\star(i)} - \widehat{\theta}^{(i)}\|_2 + 2\Big(\|\boldsymbol{v}^{(i)}\|_2 + \|\widetilde{\boldsymbol{a}}^{(i)}\|_2\Big)\|\Theta^\star - \widehat{\Theta}\|_{\mathrm{F}} \\
&\overset{(e)}{\le} \|\theta^{\star(i)} - \widehat{\theta}^{(i)}\|_2 + 2x_{\max}\big(\sqrt{p_v} + \sqrt{p_a}\big)\sqrt{p_y}\|\Theta^\star - \widehat{\Theta}\|_{2,\infty}, \qquad (3.113)
\end{aligned}$$

where (a) follows from triangle inequality, (b) follows because induced matrix norms are submultiplicative, (c) follows because operator norm of any sub-matrix is no more than operator norm of the matrix and ℓ_2 norm of any sub-vector is no more than ℓ_2 norm of the vector, (d) follows because the operator norm is no more than the Frobenius norm, and (e) follows from the relationship between the matrix norms and because $\max\{\|\boldsymbol{v}^{(i)}\|_\infty, \|\boldsymbol{a}^{(i)}\|_\infty\} \le x_{\max}$ for all $i \in [n]$.

Now, combining Eqs. (3.112) and (3.113), we have

$$\begin{aligned}
&\|(\theta^{\star(i)} - \widehat{\theta}^{(i)}) + 2\boldsymbol{v}^{(i)\top}(\Phi^{\star(v,y)} - \widehat{\Phi}^{(v,y)}) + 2\widetilde{\boldsymbol{a}}^{(i)\top}(\Phi^{\star(a,y)} - \widehat{\Phi}^{(a,y)})\|_2 + \sum_{t \in [p_y]} \|\Phi_t^{\star(y,y)} - \widehat{\Phi}_t^{(y,y)}\|_2 \\
&\le \|\theta^{\star(i)} - \widehat{\theta}^{(i)}\|_2 + 2x_{\max}\big(\sqrt{p_v} + \sqrt{p_a}\big)\sqrt{p_y}\|\Theta^\star - \widehat{\Theta}\|_{2,\infty} + \sum_{t \in [p_y]} \|\Theta_t^\star - \widehat{\Theta}_t\|_2 \\
&\overset{(a)}{\le} R(\varepsilon, \delta/n) + 2x_{\max}\big(\sqrt{p_v} + \sqrt{p_a}\big)\sqrt{p_y}\varepsilon + p_y\varepsilon,
\end{aligned}$$

and (a) follows from Theorem 3.1. The proof is complete by rescaling ε and absorbing the constants in c.

3.D.1 Proof of Lemma 3.14: Perturbation in the mean vector

Let $Z(\psi, \Psi) \in \mathbb{R}_+$ denote the log-partition function of $f_{\mathsf{w}}(\cdot|\psi, \Psi)$ in Eq. (3.111). Then, from (Busa-Fekete et al., 2019, Theorem 1), we have

$$\|\mu_{\widehat{\psi},\widehat{\Psi}}(\mathsf{w}) - \mu_{\psi^\star,\Psi^\star}(\mathsf{w})\|_2 = \|\nabla_{\widehat{\psi}} Z(\widehat{\psi}, \widehat{\Psi}) - \nabla_{\psi^\star} Z(\psi^\star, \Psi^\star)\|_2. \qquad (3.114)$$

For $t_1, t_2, t_3 \in [p_w]$, consider $\frac{\partial^2 Z(\psi,\Psi)}{\partial\psi_{t_1}\partial\psi_{t_2}}$ and $\frac{\partial^2 Z(\psi,\Psi)}{\partial\psi_{t_1}\partial\Psi_{t_2,t_3}}$. Using the fact that the Hessian of the log partition function of any regular exponential family is the covariance matrix of the associated sufficient statistic, we have

$$\frac{\partial^2 Z(\psi, \Psi)}{\partial\psi_{t_1}\partial\psi_{t_2}} = \mathbb{C}\mathrm{ov}_{\psi,\Psi}(\mathsf{w}_{t_1}, \mathsf{w}_{t_2}) \quad \text{and} \quad \frac{\partial^2 Z(\psi, \Psi)}{\partial\psi_{t_1}\partial\Psi_{t_2,t_3}} = \mathbb{C}\mathrm{ov}_{\psi,\Psi}(\mathsf{w}_{t_1}, \mathsf{w}_{t_2}\mathsf{w}_{t_3}). \qquad (3.115)$$

Now, for some $c \in (0, 1)$, $\widetilde{\psi} \triangleq c\widehat{\psi} + (1-c)\psi^\star$ and $\widetilde{\Psi} \triangleq c\widetilde{\psi} + (1-c)\widetilde{\psi}$, we have the following from the mean value theorem

$$\frac{\partial Z(\widehat{\psi}, \widehat{\Psi})}{\partial\widehat{\psi}_{t_1}} - \frac{\partial Z(\psi^\star, \Psi^\star)}{\partial\psi^\star_{t_1}}$$

$$= \sum_{t_2\in[p_w]} \frac{\partial^2 Z(\widetilde{\psi},\widetilde{\Psi})}{\partial\widetilde{\psi}_{t_2}\partial\widetilde{\psi}_{t_1}} \cdot (\widehat{\psi}_{t_2} - \psi^\star_{t_2}) + \sum_{t_2\in[p_w]}\sum_{t_3\in[p_w]} \frac{\partial^2 Z(\widetilde{\psi},\widetilde{\Psi})}{\partial\widetilde{\Psi}_{t_2,t_3}\partial\widetilde{\psi}_{t_1}} \cdot (\widehat{\Psi}_{t_2,t_3} - \Psi^\star_{t_2,t_3})$$

$$\overset{Eq.\ (3.115)}{=} \sum_{t_2\in[p_w]} \mathbb{C}\mathrm{ov}_{\widetilde{\psi},\widetilde{\Psi}}(\mathsf{w}_{t_1}, \mathsf{w}_{t_2})\cdot(\widehat{\psi}_{t_2} - \psi^\star_{t_2}) + \sum_{t_3\in[p_w]}\sum_{t_2\in[p_w]} \mathbb{C}\mathrm{ov}_{\widetilde{\psi},\widetilde{\Psi}}(\mathsf{w}_{t_1}, \mathsf{w}_{t_3}\mathsf{w}_{t_2})\cdot(\widehat{\Psi}_{t_3,t_2} - \Psi^\star_{t_3,t_2}).$$

Now, using the triangle inequality and sub-multiplicativity of induced matrix norms, we have

$$\|\nabla_{\widehat{\psi}} Z(\widehat{\psi},\widehat{\Psi}) - \nabla_{\psi^\star} Z(\psi^\star,\Psi^\star)\|_2 \leq \|\mathbb{C}\mathrm{ov}_{\widetilde{\psi},\widetilde{\Psi}}(\mathbf{w},\mathbf{w})\|_{\mathrm{op}}\|(\widehat{\psi} - \psi^\star)\|_2 + \sum_{t_3\in[p_w]} \|\mathbb{C}\mathrm{ov}_{\widetilde{\psi},\widetilde{\Psi}}(\mathbf{w}, \mathsf{w}_{t_3}\mathbf{w})\|_{\mathrm{op}}\|(\widehat{\Psi}_{t_3} - \Psi^\star_{t_3})\|_2. \tag{3.116}$$

Combining Eqs. (3.114) and (3.116) completes the proof.

3.D.2 Bounded operator norms for perturbations in the parameters

In Section 3.4.2, we assumed the operator norms of (i) the covariance matrix of $\mathbf{y}$ conditioned on $\mathbf{a}$, $\mathbf{z}$, and $\mathbf{v}$ and (ii) the cross-covariance matrix of $\mathbf{y}$ and $\mathsf{y}_t\mathbf{y}$ conditioned on $\mathbf{a}$, $\mathbf{z}$, and $\mathbf{v}$ for all $t \in [p_y]$ to remain bounded for small perturbation in the parameters. In this section, we provide examples where these hold.

Suppose the distribution of $\mathbf{y}$ conditioned on $\mathbf{a}$, $\mathbf{z}$, and $\mathbf{v}$ is a Gaussian distribution. For simplicity, let the mean of this distribution be zero. Then, for any $t, u, v \in [p_y]$,

$$\mathbb{C}\mathrm{ov}_{\theta,\Theta}(\mathsf{y}_u, \mathsf{y}_t\mathsf{y}_v | \boldsymbol{a}, \boldsymbol{z}, \boldsymbol{v}) = \mathbb{E}_{\theta,\Theta}(\mathsf{y}_u\mathsf{y}_t\mathsf{y}_v | \boldsymbol{a}, \boldsymbol{z}, \boldsymbol{v}) \overset{(a)}{=} 0.$$

where (a) follows because $\mathbb{E}_{\theta,\Theta}(\mathsf{y}_u\mathsf{y}_t\mathsf{y}_v | \boldsymbol{a}, \boldsymbol{z}, \boldsymbol{v})$ is the third cumulant of $\mathsf{y}_u\mathsf{y}_t\mathsf{y}_v | \mathbf{a}, \mathbf{z}, \mathbf{v}$ and the third cumulant for any Gaussian distribution is zero (Holmquist, 1988). Then,

$$\max_{t\in[p_y]} \|\mathbb{C}\mathrm{ov}_{\theta,\Theta}(\mathbf{y}, \mathsf{y}_t\mathbf{y} | \boldsymbol{a}, \boldsymbol{z}, \boldsymbol{v})\|_{\mathrm{op}} = 0. \tag{3.117}$$

Further, Eq. (3.117) also holds for small perturbations in θ and Θ as the distribution of $\mathbf{y}$ conditioned on $\mathbf{a}$, $\mathbf{z}$, and $\mathbf{v}$ would still be a Gaussian distribution.

Now, we bound $\|\mathbb{C}\mathrm{ov}_{\theta,\Theta}(\mathbf{y}, \mathbf{y} | \boldsymbol{a}, \boldsymbol{z}, \boldsymbol{v})\|_{\mathrm{op}}$ under additional conditions. For simplicity, suppose $\mathbb{V}\mathrm{ar}_{\theta,\Theta}(\mathsf{y}_t | \boldsymbol{a}, \boldsymbol{z}, \boldsymbol{v}) = 1$ for all $t \in [p_y]$. Further, suppose the (undirected) graphical structure associated with elements of $\mathbf{y}$, i.e., $\mathsf{y}_1, \cdots, \mathsf{y}_{p_y}$, is a chain (This would be true for the motivating example in Figure 3.1.1). If the correlation between any two elements of $\mathbf{y}$ connected by an edge in the tree is equal to $\rho \in [0, 1]$ (This is equivalent to all the off-diagonal non-zero entries of Θ being the same), then for any $u, v \in [p_y]$,

$$\mathbb{C}\mathrm{ov}_{\theta,\Theta}(\mathsf{y}_u, \mathsf{y}_v | \boldsymbol{a}, \boldsymbol{z}, \boldsymbol{v}) \overset{(a)}{=} \rho^{|u-v|},$$

where (a) follows by the correlation decay property for Gaussian tree models (Tan et al., 2010, Equation. 18). Then, for any $0 \leq \rho < 1$

$$\|\mathbb{C}\mathrm{ov}_{\theta,\Theta}(\mathbf{y}, \mathbf{y}|\boldsymbol{a}, \boldsymbol{z}, \boldsymbol{v})\|_{\mathrm{op}} \overset{(a)}{\leq} \frac{1+\rho}{1-\rho}, \tag{3.118}$$

where (a) follows from Trench (1999). Further, Eq. (3.118) holds for small perturbations in θ and Θ as long as $\rho < 1$. Therefore, $C(\mathbb{B})$ in Eq. (3.20) is a constant (with respect to p) for small perturbations in θ and Θ.

While we showed that $C(\mathbb{B})$ is a constant for a class of Gaussian distributions, we except similar results for truncated Gaussian distributions and exponential family distributions in Eq. (3.2).

3.E Proofs of Propositions 3.2 and 3.3

3.E.1 Proof of Proposition 3.2

3.E.1.1 Guarantees for Example 3.7

We divide the proof into multiple parts for convenience. First, we express the outcome generating process in Eq. (3.25) in a vector form. Then, we obtain the conditional distribution of the outcome vector given the intervention vector, the observed covariates, and the unobserved covariate. Then, we provide guarantees on recovering the parameters that vary with time. Finally, we provide guarantees on recovering the parameters that vary with unit.

3.E.1.1.1 Expressing Eq. (3.25) in a vector form We define

$$\begin{aligned}
&\boldsymbol{\beta}^{(y,j)} \in \mathbb{R}^p \quad \text{such that} \quad \boldsymbol{\beta}_t^{(y,j)} = \beta_{t,t-j}^{(y)} \quad \text{for every} \quad t \in [p] \quad \text{and} \quad j \in [d]\\
&\boldsymbol{\beta}^{(a,j)} \in \mathbb{R}^p \quad \text{such that} \quad \boldsymbol{\beta}_t^{(a,j)} = \beta_{t,t-j}^{(a)} \quad \text{for every} \quad t \in [p] \quad \text{and} \quad j \in \{0\} \cup [d]\\
&\boldsymbol{\beta}^{(v,j)} \in \mathbb{R}^p \quad \text{such that} \quad \boldsymbol{\beta}_t^{(v,j)} = \beta_{t,t-j}^{(v)} \quad \text{for every} \quad t \in [p] \quad \text{and} \quad j \in \{0\} \cup [d],
\end{aligned}$$

where $\beta_{t,t-j}^{(u)} = 0$ if $t - j \leq 0$ for every $u \in \{y, a, v\}$. Further, we define

$$\begin{aligned}
&\boldsymbol{y}^{(i,j)} \in \mathbb{R}^p \quad \text{such that} \quad \boldsymbol{y}_t^{(i,j)} = y_{t-j}^{(i)} \quad \text{for every} \quad i \in [n], \quad t \in [p] \quad \text{and} \quad j \in [d]\\
&\boldsymbol{a}^{(i,j)} \in \mathbb{R}^p \quad \text{such that} \quad \boldsymbol{a}_t^{(i,j)} = a_{t-j}^{(i)} \quad \text{for every} \quad i \in [n], \quad t \in [p] \quad \text{and} \quad j \in \{0\} \cup [d]\\
&\boldsymbol{v}^{(i,j)} \in \mathbb{R}^p \quad \text{such that} \quad \boldsymbol{v}_t^{(i,j)} = v_{t-j}^{(i)} \quad \text{for every} \quad i \in [n], \quad t \in [p] \quad \text{and} \quad j \in \{0\} \cup [d],
\end{aligned}$$

where $y_{t-j}^{(i)} = a_{t-j}^{(i)} = v_{t-j}^{(i)} = 0$ for $t - j \leq 0$ for all $i \in [n]$. Then, Eq. (3.25) can be written as

$$\boldsymbol{\eta}^{(i)} = \boldsymbol{y}^{(i)} - \sum_{j=1}^{d} \boldsymbol{\beta}^{(y,j)} \odot \boldsymbol{y}^{(i,j)} - \beta_i^{(a)} \boldsymbol{a}^{(i)} - \sum_{j=0}^{d} \boldsymbol{\beta}^{(a,j)} \odot \boldsymbol{a}^{(i,j)} - \beta_i^{(v)} \boldsymbol{v}^{(i)} - \sum_{j=0}^{d} \boldsymbol{\beta}^{(v,j)} \odot \boldsymbol{v}^{(i,j)} - z^{(i)} \mathbf{1}, \tag{3.119}$$

where $\odot$ denotes the element-wise multiplication.

3.E.1.1.2 Obtaining the conditional distribution of the outcome For every unit $i \in [n]$, the distribution of $\mathbf{y} = \boldsymbol{y}^{(i)}$ given $\mathbf{a} = \boldsymbol{a}^{(i)}$, $\mathbf{v} = \boldsymbol{v}^{(i)}$, and $\mathbf{z} = z^{(i)}\mathbf{1}$ is given by $f_{\mathbf{y}|\mathbf{a},\mathbf{v},\mathbf{z}}(\boldsymbol{y}^{(i)}|\boldsymbol{a}^{(i)}, \boldsymbol{v}^{(i)}, z^{(i)}\mathbf{1}) \propto \exp\big(\boldsymbol{\eta}^{(i)\top}\mathbf{E}\boldsymbol{\eta}^{(i)}\big)$. We claim that this can expressed in a form akin to Eq. (3.4). To see this, plugging in $\boldsymbol{\eta}^{(i)}$ from (3.119), we have

$$
\begin{aligned}
&f_{\mathbf{y}|\mathbf{a},\mathbf{v},\mathbf{z}}(\boldsymbol{y}^{(i)}|\boldsymbol{a}^{(i)}, \boldsymbol{v}^{(i)}, z^{(i)}\mathbf{1})\\
&\qquad \propto \exp\bigg(-2\Big[z^{(i)}\mathbf{1} + \beta_i^{(v)}\boldsymbol{v}^{(i)} + \beta_i^{(a)}\boldsymbol{a}^{(i)}\Big]^\top \cdot \Big[\mathbf{E}\boldsymbol{y}^{(i)} - \sum_{j=1}^{d}\mathbf{E}\big(\boldsymbol{\beta}^{(y,j)} \odot \boldsymbol{y}^{(i,j)}\big)\Big]\\
&\qquad -2\sum_{j_1=0}^{d}\Big[\boldsymbol{\beta}^{(v,j)} \odot \boldsymbol{v}^{(i,j)} + \boldsymbol{\beta}^{(a,j)} \odot \boldsymbol{a}^{(i,j)}\Big]^\top \cdot \Big[\mathbf{E}\boldsymbol{y}^{(i)} - \sum_{j=1}^{d}\mathbf{E}\big(\boldsymbol{\beta}^{(y,j)} \odot \boldsymbol{y}^{(i,j)}\big)\Big]\\
&\qquad + \Big[\boldsymbol{y}^{(i)} - \sum_{\bar{j}=1}^{d}\big(\boldsymbol{\beta}^{(y,\bar{j})} \odot \boldsymbol{y}^{(i,\bar{j})}\big)\Big]^\top \cdot \Big[\mathbf{E}\boldsymbol{y}^{(i)} - \sum_{j=1}^{d}\mathbf{E}\big(\boldsymbol{\beta}^{(y,j)} \odot \boldsymbol{y}^{(i,j)}\big)\Big]\bigg). \qquad (3.120)
\end{aligned}
$$

To convert Eq. (3.120) into the familiar form, we define the interaction matrices

$$
\begin{aligned}
&\Phi^{(z,y)} \in \mathbb{R}^{p\times p} \quad \text{such that} \quad \Phi^{(z,y)}_{t_1,t_2} = -\mathbf{E}_{t_1,t_2} + \sum_{j=1}^{d}\mathbf{E}_{t_1,t_2+j}\cdot\beta^{(y)}_{t_2+j,t_2} \quad \text{for every} \quad t_1, t_2 \in [p]\\
&\Phi^{(v,y)} \in \mathbb{R}^{p\times p} \quad \text{such that} \quad \Phi^{(v,y)}_{t_1,t_2} = \sum_{j=0}^{d}\Phi^{(z,y)}_{t_1+j,t_2}\cdot\beta^{(v)}_{t_1+j,t_1} \quad \text{for every} \quad t_1, t_2 \in [p]\\
&\Phi^{(a,y)} \in \mathbb{R}^{p\times p} \quad \text{such that} \quad \Phi^{(a,y)}_{t_1,t_2} = \sum_{j=0}^{d}\Phi^{(z,y)}_{t_1+j,t_2}\cdot\beta^{(a)}_{t_1+j,t_1} \quad \text{for every} \quad t_1, t_2 \in [p], \text{ and}\\
&\Phi^{(y,y)} \in \mathbb{R}^{p\times p} \quad \text{such that} \quad \Phi^{(y,y)}_{t_1,t_2} = -\Phi^{(z,y)}_{t_1,t_2} + \sum_{j=1}^{d}\Phi^{(z,y)}_{t_1+j,t_2}\cdot\beta^{(y)}_{t_1+j,t_1} \quad \text{for every} \quad t_1, t_2 \in [p],
\end{aligned}
$$

where $\mathbf{E}_{t_1,t_2+j} = 0$ for $t_2 + j > p$, $\Phi^{(z,y)}_{t_1+j,t_2} = 0$ for $t_1 + j > p$, and $\beta^{(u)}_{t+j,t} = 0$ for $t + j > p$ for every $u \in \{y, a, v\}$. We note that $\Phi^{(y,y)}$ is a symmetric matrix. Further, these interaction matrices are indeed such that:

$$
\Phi^{(z,y)} = \mathbf{E}\mathbf{B}^{(y)\top} \quad \text{and} \quad \Phi^{(u,y)} = \mathbf{B}^{(u)}\Phi^{(z,y)} \quad \text{for every} \quad u \in \{v, a, y\}. \qquad (3.121)
$$

Then, we define the unit-level and the population-level parameters

$$
\theta^{(i)} \triangleq 2\Phi^{(z,y)\top}\Big[z^{(i)}\mathbf{1} + \beta_i^{(v)}\boldsymbol{v}^{(i)} + \beta_i^{(a)}\boldsymbol{a}^{(i)}\Big] \quad \text{and} \quad \Theta \triangleq \Big[\Phi^{(y,y)}, \Phi^{(y,a)}, \Phi^{(y,v)}\Big], \qquad (3.122)
$$

with $\Phi^{(y,a)} = \Phi^{(a,y)^\top} \in \mathbb{R}^{p\times p_a}$ and $\Phi^{(y,v)} = \Phi^{(v,y)^\top} \in \mathbb{R}^{p\times p_v}$. Plugging these in (3.120), we have

$$
\begin{aligned}
&f_{\mathbf{y}|\mathbf{a},\mathbf{v},\mathbf{z}}(\boldsymbol{y}^{(i)}|\boldsymbol{a}^{(i)}, \boldsymbol{v}^{(i)}, z^{(i)}\mathbf{1}; \theta^{(i)}, \Theta)\\
&\qquad \propto \exp\Big(\theta^{(i)^\top}\boldsymbol{y}^{(i)} + 2\boldsymbol{v}^{(i)^\top}\Phi^{(v,y)}\boldsymbol{y}^{(i)} + 2\boldsymbol{a}^{(i)^\top}\Phi^{(a,y)}\boldsymbol{y}^{(i)} + \boldsymbol{y}^{(i)^\top}\Phi^{(y,y)}\boldsymbol{y}^{(i)}\Big), \qquad (3.123)
\end{aligned}
$$

which is akin to Eq. (3.4).

3.E.1.1.3 Assumption 3.1 holds To enable the application of the machinery developed in Section 3.3.3, we show that the unit-level and the population-level parameters defined in Eq. (3.122) satisfy Assumption 3.1. We claim

$$\|\theta^{(i)}\|_\infty \leq 2(z_{\max} + 2\beta_{\max}x_{\max})(1 + d\beta_{\max})\beta \tag{3.124}$$

$$\|\Theta\|_{\max} \leq \overline{\beta}_{\max}(1 + d\beta_{\max})\beta, \text{ and} \tag{3.125}$$

$$\|\Theta\|_\infty \leq (1 + 2\beta_{\max} + 3d\beta_{\max})(1 + d\beta_{\max})\beta. \tag{3.126}$$

Proof of Eq. (3.124). We have

$$\begin{aligned}
\|\theta^{(i)}\|_\infty &\overset{Eq.~(3.122)}{=} 2\|\Phi^{(z,y)\top}[z^{(i)}\mathbf{1} + \beta_i^{(v)}\boldsymbol{v}^{(i)} + \beta_i^{(a)}\boldsymbol{a}^{(i)}]\|_\infty \\
&\overset{(a)}{\leq} 2\|\Phi^{(z,y)}\|_1\|z^{(i)}\mathbf{1} + \beta_i^{(v)}\boldsymbol{v}^{(i)} + \beta_i^{(a)}\boldsymbol{a}^{(i)}\|_\infty \\
&\overset{(b)}{\leq} 2(z_{\max} + 2\beta_{\max}x_{\max})\|\Phi^{(z,y)}\|_1 \\
&\overset{(c)}{\leq} 2(z_{\max} + 2\beta_{\max}x_{\max})\|\mathbf{E}\|_1\|\mathbf{B}^{(y)}\|_\infty \\
&\overset{(d)}{\leq} 2(z_{\max} + 2\beta_{\max}x_{\max})\|\mathbf{E}\|_\infty\|\mathbf{B}^{(y)}\|_\infty, \\
&\overset{(e)}{\leq} 2(z_{\max} + 2\beta_{\max}x_{\max})(1 + d\beta_{\max})\beta,
\end{aligned} \tag{3.127}$$

where (a) follows from standard matrix norm inequalities, (b) follows from Assumption 3.4 and because $\max\{\boldsymbol{v}^{(i)}, \boldsymbol{a}^{(i)}\} \leq x_{\max}$ for all $i \in [n]$, (c) follows from sub-multiplicativity of induced matrix norms, (d) follows because $\mathbf{E}$ is symmetric, and (e) follows from Eq. (3.28), Assumption 3.4, and Assumption 3.5(a).

Proof of Eq. (3.125). For any $u \in \{v, a, y\}$, we have

$$\begin{aligned}
\|\Phi^{(u,y)}\|_{\max} \overset{Eq.~(3.121)}{=} \|\mathbf{B}^{(u)}\Phi^{(z,y)}\|_{\max} \overset{(a)}{\leq} \|\mathbf{B}^{(u)}\|_{\max}\|\Phi^{(z,y)}\|_1 &\overset{(b)}{\leq} \overline{\beta}_{\max}\|\Phi^{(z,y)}\|_1 \\
&\overset{(c)}{\leq} \overline{\beta}_{\max}(1 + d\beta_{\max})\beta,
\end{aligned}$$

where (a) follows from standard matrix norm inequalities, (b) follows from Assumption 3.4(b), and (c) follows as in Eq. (3.127). Then, Eq. (3.125) follows by noting that $\|\Theta\|_{\max} = \max_{u\in\{v,a,y\}} \|\Phi^{(u,y)}\|_{\max}$.

Proof of Eq. (3.126). For any $u \in \{v, a, y\}$, we have

$$\|\Phi^{(u,y)}\|_1 \overset{Eq.~(3.121)}{=} \|\mathbf{B}^{(u)}\Phi^{(z,y)}\|_1 \overset{(a)}{\leq} \|\mathbf{B}^{(u)}\|_1\|\Phi^{(z,y)}\|_1 \overset{(b)}{\leq} \|\mathbf{B}^{(u)}\|_1(1 + d\beta_{\max})\beta, \tag{3.128}$$

where (a) follows from sub-multiplicativity of induced matrix norms and (b) follows as in Eq. (3.127). Then, Eq. (3.126) follows by noting that

$$\|\Theta\|_\infty \leq \sum_{u\in\{y,a,v\}} \|\Phi^{(u,y)}\|_1 \overset{(a)}{\leq} (1 + d\beta_{\max} + 2(d+1)\beta_{\max})(1 + d\beta_{\max})\beta,$$

where (a) follows from Eqs. (3.26) to (3.28) and (3.128) and Assumption 3.4.

3.E.1.1.4 No dynamics in the outcomes In this scenario, we have $\beta^{(y)}_{t_2,t_1} = 0$ for all $t_1, t_2 \in [p]$, i.e., $\mathbf{B}^{(y)} = -\mathbf{I}$ and $\Phi^{(z,y)} = -\Phi^{(y,y)}$. Then, for every $i \in [n]$, $\theta^{(i)}$ can be expressed as a linear combination of known vectors. In particular,

$$\theta^{(i)} = \mathbf{D}^{(i)}\mathbf{c}^{(i)}, \tag{3.129}$$

where

$$\mathbf{D}^{(i)} \triangleq -2\Phi^{(y,y)\top}\boldsymbol{o}^{(i)} \in \mathbb{R}^{p\times 3} \quad \text{and} \quad \mathbf{c}^{(i)} \triangleq \begin{bmatrix} z^{(i)} \\ \beta_i^{(v)} \\ \beta_i^{(a)} \end{bmatrix} \in \mathbb{R}^{3\times 1}. \tag{3.130}$$

Therefore, using Example 3.1, the sum of the metric entropy $\mathcal{M}_{\Theta}(\varepsilon^2) + \mathcal{M}_{\theta,n}(\varepsilon^2)$ is $O\big(p\log\frac{1}{\varepsilon^2} + \frac{1}{\varepsilon^2}\big)$, which is driven by the unknowns Θ and $\mathbf{c}^{(i)}$.

Recovering population-level parameters as in Eq. (3.29) Applying Theorem 3.1, we obtain estimate $\widehat{\Theta}$ such that, with probability at least $1-\delta$, we have

$$\|\widehat{\Theta} - \Theta^\star\|_{2,\infty} \le \varepsilon_1 \quad \text{for} \quad n \ge \frac{cp^2\big(\log\frac{p}{\delta} + p\log\frac{1}{\varepsilon_1^2} + \frac{1}{\varepsilon_1^2}\big)}{\varepsilon_1^4}. \tag{3.131}$$

Let us condition on the event in Eq. (3.131). To estimate $\mathbf{B}^{(a)}$ and $\mathbf{B}^{(v)}$, consider Eq. (3.121). For every $w \in \{v, a\}$, we have

$$\Phi^{(w,y)} = -\mathbf{B}^{(w)}\Phi^{(y,y)}.$$

If we knew $\Phi^{(w,y)}$ and $\Phi^{(y,y)}$, estimating $\mathbf{B}^{(w)}$ would have been easy. While we do not know $\Phi^{(w,y)}$ and $\Phi^{(y,y)}$, we can produce estimates using $\widehat{\Theta}$. In particular, let $\widehat{\Phi}^{(w,y)}$ and $\widehat{\Phi}^{(y,y)}$ be the components of $\widehat{\Theta}$ that are estimates of $\Phi^{(w,y)}$ and $\Phi^{(y,y)}$, respectively. Then, we can estimate $\mathbf{B}^{(w)}$ by performing an error-in-variable regression. Specifically, defining $\Delta\Phi^{(w,y)} = \widehat{\Phi}^{(w,y)} - \Phi^{(w,y)}$, we have

$$\widehat{\Phi}^{(w,y)} = -\mathbf{B}^{(w)}\widehat{\Phi}^{(y,y)} + \mathbf{B}^{(w)}\Delta\Phi^{(y,y)} + \Delta\Phi^{(w,y)}$$

Then, we let $\widehat{\mathbf{B}}^{(w)}$ be such that

$$\widehat{\Phi}^{(w,y)} = -\widehat{\mathbf{B}}^{(w)}\widehat{\Phi}^{(y,y)}.$$

We claim that the choice of $\varepsilon_1 = c\varepsilon^2/\sqrt{p\log p}$ in Eq. (3.131) is such that

$$\|\widehat{\mathbf{B}}^{(w)} - \mathbf{B}^{(w)}\|_{2,\infty} \le c\varepsilon/\sqrt{\kappa_1}. \tag{3.132}$$

Then, the proof is complete by re-parameterizing ε.

Proof of Eq. (3.132). First, from the triangle inequality, we have

$$\|(\widehat{\mathbf{B}}^{(w)} - \mathbf{B}^{(w)})\Phi^{(y,y)}\|_{2,\infty} \le \|(\widehat{\mathbf{B}}^{(w)} - \mathbf{B}^{(w)})\widehat{\Phi}^{(y,y)}\|_{2,\infty} + \|(\widehat{\mathbf{B}}^{(w)} - \mathbf{B}^{(w)})\Delta\Phi^{(y,y)}\|_{2,\infty}. \tag{3.133}$$

Now, we bound both the terms on the right hand side of Eq. (3.133). To bound the first term in Eq. (3.133), we invoke (Shah et al., 2021c, Lemma N.1). We have

$$
\begin{aligned}
&\|(\widehat{\mathbf{B}}^{(w)} - \mathbf{B}^{(w)})\widehat{\Phi}^{(y,y)}\|_{2,\infty}^2 \\
&\leq p\bigg(c\|\mathbf{B}^{(w)}\Delta\Phi^{(y,y)}\|_{\max}^2 + c\|\widehat{\Phi}^{(y,y)}\|_{\max}\|\Delta\Phi^{(w)}\|_{\max}\|\mathbf{B}^{(w)}\|_{\infty}\sqrt{\frac{\log p}{p}}\bigg) \\
&\overset{(a)}{\leq} p\bigg(c\|\mathbf{B}^{(w)}\|_{2,\infty}^2\|\Delta\Phi^{(y,y)}\|_{1,2}^2 + c\overline{\beta}_{\max}(1+d\beta_{\max})\beta\|\Delta\Phi^{(w,y)}\|_{\max}\|\mathbf{B}^{(w)}\|_{\infty}\sqrt{\frac{\log p}{p}}\bigg) \\
&\overset{(b)}{\leq} p\bigg(c(1+d)\overline{\beta}_{\max}^2\|\Delta\Phi^{(y,y)}\|_{2,\infty}^2 + c\overline{\beta}_{\max}(1+d\beta_{\max})\beta(1+d)\overline{\beta}_{\max}\|\Delta\Phi^{(w,y)}\|_{\max}\sqrt{\frac{\log p}{p}}\bigg) \\
&\overset{(c)}{\leq} p\bigg(c(1+d)\overline{\beta}_{\max}^2\varepsilon_1^2 + c\overline{\beta}_{\max}(1+d\beta_{\max})\beta(1+d)\overline{\beta}_{\max}\varepsilon_1\sqrt{\frac{\log p}{p}}\bigg), \qquad (3.134)
\end{aligned}
$$

where (a) follows by Cauchy–Schwarz inequality and from Eq. (3.125), (b) follows from Eqs. (3.26) to (3.28) and Assumption 3.4(b), and (c) follows from Eq. (3.131) and because $\Delta\Phi^{(y,y)}$ is symmetric. To bound the second term in Eq. (3.133), we have

$$
\begin{aligned}
\|(\widehat{\mathbf{B}}^{(w)} - \mathbf{B}^{(w)})\Delta\Phi^{(y,y)}\|_{2,\infty}^2 &\overset{(a)}{\leq} p\|(\widehat{\mathbf{B}}^{(w)} - \mathbf{B}^{(w)})\Delta\Phi^{(y,y)}\|_{\max}^2 \\
&\overset{(b)}{\leq} p\|\widehat{\mathbf{B}}^{(w)} - \mathbf{B}^{(w)}\|_{2,\infty}^2\|\Delta\Phi^{(y,y)}\|_{1,2}^2 \\
&\overset{(c)}{\leq} p\big(\|\widehat{\mathbf{B}}^{(w)}\|_{2,\infty} + \|\mathbf{B}^{(w)}\|_{2,\infty}\big)^2\|\Delta\Phi^{(y,y)}\|_{2,\infty}^2 \\
&\overset{(d)}{\leq} cp(1+d)\ \overline{\beta}_{\max}^2\varepsilon_1^2, \qquad (3.135)
\end{aligned}
$$

where (a) follows from standard matrix norm inequalities, (b) follows by Cauchy–Schwarz inequality, (c) follows from the triangle inequality and because $\Delta\Phi^{(y,y)}$ is symmetric, and (d) follows similar to Eq. (3.134). Then, putting together Eqs. (3.133) to (3.135) and using the choice of $\varepsilon_1 = c\varepsilon/\sqrt{p}$, we have

$$
\|(\widehat{\mathbf{B}}^{(w)} - \mathbf{B}^{(w)})\Phi^{(y,y)}\|_{2,\infty}^2 \leq c\bigg(\frac{\varepsilon^4}{\log p} + \varepsilon^2\bigg) \overset{(a)}{\leq} c\varepsilon^2,
$$

where (a) holds as typically $\varepsilon \ll \sqrt{\log p}$. Finally, Eq. (3.132) follows after noting that Assumption 3.6(a) implies

$$
\|\widehat{\mathbf{B}}^{(w)} - \mathbf{B}^{(w)}\|_{2,\infty}^2 \leq \frac{1}{\kappa_1}\|(\widehat{\mathbf{B}}^{(w)} - \mathbf{B}^{(w)})\Phi^{(y,y)}\|_{2,\infty}^2.
$$

Recovering unit-level parameters as in Eq. (3.30) In Eq. (3.129), if we knew the matrix $\mathbf{D}^{(i)}$, estimating $\mathbf{c}^{(i)}$ would have been easy. While we do not know $\mathbf{D}^{(i)}$, we can produce an estimate using $\widehat{\Theta}$.

The following lemma, proven in Section 3.E.1.3, provides guarantees on recovering the coefficients in a linear combination when the basis are known with some error.

Lemma 3.15 (Coefficient recovery in a linear combination with noisy basis)**.** *Suppose $\theta \in \mathbb{R}^p$, $\mathbf{D} \in \mathbb{R}^{p\times k}$, and $\mathbf{c} \in \mathbb{R}^k$ are unknowns such that $\theta = \mathbf{D}\mathbf{c}$. Suppose we have estimates $\widehat{\mathbf{D}}$ and $\widehat{\theta} = \widehat{\mathbf{D}}\widehat{\mathbf{c}}$ such that*

$$\|\widehat{\mathbf{D}} - \mathbf{D}\|_{2,\infty} \leq \bar{\varepsilon}_1 \quad \text{and} \quad \mathrm{MSE}(\widehat{\theta}, \theta) \leq \bar{\varepsilon}_2. \tag{3.136}$$

Then,

$$\frac{1}{p}\|\widehat{\mathbf{c}} - \mathbf{c}\|_2^2 \leq \frac{2}{\lambda_{\min}(\widehat{\mathbf{D}}^\top\widehat{\mathbf{D}})}\Big(k\|\mathbf{c}\|_\infty^2\bar{\varepsilon}_1^2 + \bar{\varepsilon}_2\Big).$$

Given this lemma, we now proceed with the proof. Let us condition on the event in Eq. (3.131) and define $\widehat{\mathbf{D}}^{(i)} \triangleq -2\widehat{\Phi}^{(y,y)\top}\boldsymbol{o}^{(i)}$. Now, we write

$$\theta^{(i)} = \widehat{\mathbf{D}}^{(i)}\widetilde{\mathbf{c}}^{(i)} \quad \text{where} \quad \widetilde{\mathbf{c}}^{(i)} \triangleq \mathbf{c}^{(i)} + \zeta,$$

for some error term ζ. Applying Theorem 3.1, we obtain estimate $\widehat{\theta}^{(i)}$ such that, with probability at least $1-\delta$, we have

$$\mathrm{MSE}(\widehat{\theta}^{(i)}, \theta^{(i)}) \leq \max\Big\{\varepsilon_1^2, \frac{c\log(\log\frac{p}{\delta})}{p}\Big\} \quad \text{for} \quad n \geq \frac{cp^4\big(\log\frac{p}{\delta} + p\log\frac{p}{\varepsilon_1^2} + \frac{p}{\varepsilon_1^2}\big)}{\varepsilon_1^4}. \tag{3.137}$$

We note that the above estimate $\widehat{\theta}^{(i)}$ of the unit-level parameter $\theta^{(i)}$ is of the form $\widehat{\theta}^{(i)} = \widehat{\mathbf{D}}^{(i)}\widehat{\mathbf{c}}^{(i)}$. To prove the corresponding guarantee on $\widehat{\mathbf{c}}^{(i)}$, we invoke Lemma 3.15. Towards that, note that

$$\begin{aligned}\|\widehat{\mathbf{D}}^{(i)} - \mathbf{D}^{(i)}\|_{2,\infty} &\overset{(a)}{\leq} 2\sqrt{1+2x_{\max}^2}\|\widehat{\Phi}^{(y,y)\top}\mathbf{1} - \Phi^{(y,y)\top}\mathbf{1}\|_\infty \\ &\overset{(b)}{\leq} \sqrt{1+2x_{\max}^2}\sqrt{p}\|\widehat{\Phi}^{(y,y)} - \Phi^{(y,y)}\|_{2,\infty}, \\ &\overset{(c)}{\leq} \sqrt{1+2x_{\max}^2}\sqrt{p}\varepsilon_1,\end{aligned}$$

where (a) follows by noting that $\max\{\boldsymbol{v}^{(i)}, \boldsymbol{a}^{(i)}\} \leq x_{\max}$ for all $i \in [n]$, (b) follows from standard matrix norm inequalities, and (c) follows from Eq. (3.131).

Next, we claim that the eigenvalues of $\widehat{\mathbf{D}}^{(i)\top}\widehat{\mathbf{D}}^{(i)}$ are lower bounded by $\kappa_2 p(1-\varepsilon)$ with the choice $\varepsilon_1 = c\kappa\varepsilon/\sqrt{p}$. Then, conditioning on the event in Eq. (3.137) and invoking Lemma 3.15, we have

$$\frac{1}{p}\|\widehat{\mathbf{c}}^{(i)} - \mathbf{c}^{(i)}\|_2^2 \leq \frac{c}{\kappa_2 p(1-\varepsilon)}\bigg(\max\{z_{\max}^2, \beta_{\max}^2\}(1+2x_{\max}^2)\varepsilon^2\kappa_2^2 + \max\Big\{\frac{\varepsilon^2\kappa_2^2}{p}, \frac{\log(\log\frac{p}{\delta})}{p}\Big\}\bigg).$$

where we have used Assumption 3.4. The proof is complete by taking a union bound over $i \in [n]$.

It remains to show that the eigenvalues of $\widehat{\mathbf{D}}^{(i)\top}\widehat{\mathbf{D}}^{(i)}$ can be lower bounded by $\kappa_2 p(1-\varepsilon)$ with the choice $\varepsilon_1 = c\kappa_2\varepsilon/\sqrt{p}$. From Weyl's inequality (Bhatia, 2007, Theorem. 8.2), we have

$$\lambda_{\min}(\widehat{\mathbf{D}}^{(i)\top}\widehat{\mathbf{D}}^{(i)}) \geq \lambda_{\min}(\mathbf{D}^{(i)\top}\mathbf{D}^{(i)}) - \lambda_{\max}(\mathbf{D}^{(i)\top}\mathbf{D}^{(i)} - \widehat{\mathbf{D}}^{(i)\top}\widehat{\mathbf{D}}^{(i)})$$

$$\overset{(a)}{\geq} \kappa_2 p - \lambda_{\max}(\mathbf{D}^{(i)\top}\mathbf{D}^{(i)} - \widehat{\mathbf{D}}^{(i)\top}\widehat{\mathbf{D}}^{(i)}),$$

where (a) follows from Assumption 3.6(b). Now, it suffices to upper bound $\lambda_{\max}(\mathbf{D}^{(i)\top}\mathbf{D}^{(i)} - \widehat{\mathbf{D}}^{(i)\top}\widehat{\mathbf{D}}^{(i)})$ by $\kappa_2 p\varepsilon$. We have

$$\begin{aligned}
&\big|\lambda_{\max}(\mathbf{D}^{(i)\top}\mathbf{D}^{(i)} - \widehat{\mathbf{D}}^{(i)\top}\widehat{\mathbf{D}}^{(i)})\big| \\
&\overset{(a)}{=} \|\mathbf{D}^{(i)\top}\mathbf{D}^{(i)} - \widehat{\mathbf{D}}^{(i)\top}\widehat{\mathbf{D}}^{(i)}\|_{\mathrm{op}} \\
&\overset{(b)}{\leq} 3\|\mathbf{D}^{(i)\top}\mathbf{D}^{(i)} - \widehat{\mathbf{D}}^{(i)\top}\widehat{\mathbf{D}}^{(i)}\|_{\max} \\
&\overset{(c)}{\leq} 3\Big(\|\mathbf{D}^{(i)\top}\big(\mathbf{D}^{(i)} - \widehat{\mathbf{D}}^{(i)}\big)\|_{\max} + \|\big(\mathbf{D}^{(i)} - \widehat{\mathbf{D}}^{(i)\top}\big)\widehat{\mathbf{D}}^{(i)}\|_{\max}\Big) \\
&\overset{(d)}{\leq} 3\big(\|\mathbf{D}^{(i)\top}\|_{2,\infty} + \|\widehat{\mathbf{D}}^{(i)\top}\|_{2,\infty}\big)\|\mathbf{D}^{(i)} - \widehat{\mathbf{D}}^{(i)}\|_{1,2} \\
&\overset{(e)}{\leq} c\max\{1, x_{\max}^2\}\big(\|\Phi^{(y,y)\top}\mathbf{1}\|_2 + \|\widehat{\Phi}^{(y,y)\top}\mathbf{1}\|_2\big)\|\Delta\Phi^{(y,y)\top}\mathbf{1}\|_2 \\
&\overset{(f)}{\leq} c\max\{1, x_{\max}^2\}p\big(\|\Phi^{(y,y)}\|_\infty + \|\widehat{\Phi}^{(y,y)}\|_\infty\big)\|\Delta\Phi^{(y,y)}\|_\infty, \\
&\overset{(g)}{\leq} c\max\{1, x_{\max}^2\}p^{3/2}(1 + 2\beta_{\max} + 3d\beta_{\max})(1 + d\beta_{\max})\beta\varepsilon_1,
\end{aligned}$$

where (a) follows because $\mathbf{D}^{(i)\top}\mathbf{D}^{(i)} - \widehat{\mathbf{D}}^{(i)\top}\widehat{\mathbf{D}}^{(i)}$ is symmetric, (b) follows because $\|\mathbf{M}\|_{\mathrm{op}} \leq \|\mathbf{M}\|_{\mathrm{F}} \leq k\|\mathbf{M}\|_{\max}$ for any square matrix $\mathbf{M} \in \mathbb{R}^{k\times k}$, (c) follows from the triangle inequality, (d) follows by Cauchy–Schwarz inequality, (e) follows from Eq. (3.130) and because $\max\{\boldsymbol{v}^{(i)}, \boldsymbol{a}^{(i)}\} \leq x_{\max}$ for all $i \in [n]$, (f) follows from standard matrix norm inequalities, and (g) follows because $\|\Delta\Phi^{(y,y)}\|_\infty \leq \sqrt{p}\|\Delta\Phi^{(y,y)}\|_{2,\infty}$, and Eqs. (3.126) and (3.131). Then, the upper bound follows by the choice of ε_1.

3.E.1.1.5 No dynamics in the observed covariates

In this scenario, we have $\beta^{(v)}_{t_2,t_1} = 0$ if $t_1 \neq t_2$ and $\beta^{(v)}_{t_2,t_1} = \beta^{(v)}$ otherwise, i.e., $\mathbf{B}^{(v)} = -\beta^{(v)}\mathbf{I}$ and $\Phi^{(z,y)} = -\frac{1}{\beta^{(v)}}\Phi^{(v,y)}$. Then, for every $i \in [n]$, $\theta^{(i)}$ can be expressed as a linear combination of known vectors. In particular,

$$\theta^{(i)} = -\frac{1}{\beta^{(v)}}\mathbf{D}^{(i)}\mathbf{c}^{(i)},$$

where

$$\mathbf{D}^{(i)} \triangleq -2\Phi^{(v,y)\top}\boldsymbol{o}^{(i)} \in \mathbb{R}^{p\times 3} \quad \text{and} \quad \mathbf{c}^{(i)} \triangleq \begin{bmatrix} z^{(i)} \\ \beta^{(v)}_i \\ \beta^{(a)}_i \end{bmatrix} \in \mathbb{R}^{3\times 1}.$$

From Eq. (3.121), we have $\Phi^{(w,y)} = \mathbf{B}^{(w)}\Phi^{(z,y)} = -\frac{1}{\beta^{(v)}}\mathbf{B}^{(w)}\Phi^{(v,y)}$ for every $w \in \{a, y\}$. We claim that there exists $\widehat{\beta}^{(v)}$ such that

$$\big|\widehat{\beta}^{(v)} - \beta^{(v)}\big| \leq \varepsilon \quad \text{for} \quad n \geq \frac{cp\log\frac{1}{\delta\varepsilon^2}}{\varepsilon^4}. \tag{3.138}$$

Then, performing an analysis similar to the setting with no dynamics in the outcomes, results in estimates $\widehat{\mathbf{F}}^{(w)}$ of $\mathbf{F}^{(w)} \triangleq \mathbf{B}^{(w)}/\beta^{(v)}$ and $\widehat{\mathbf{m}}^{(i)}$ of $\mathbf{m}^{(i)} \triangleq \mathbf{c}^{(i)}/\beta^{(v)}$ such that

$$\begin{aligned} &\|\widehat{\mathbf{F}}^{(w)} - \mathbf{F}^{(w)}\|_{2,\infty} \le \varepsilon \quad \text{for} \quad n \ge \frac{cp^4\log^2 p\big(\log\frac{p}{\delta} + p\log\frac{p}{\varepsilon^4\kappa_1^2}\big)}{\varepsilon^8\kappa_1^4} \quad \text{and } u \in \{a, y\}, \text{ and} \\ &\|\widehat{\mathbf{m}}^{(i)} - \mathbf{m}^{(i)}\|_2^2 \le \frac{c}{\kappa_2(1-\varepsilon)}\bigg(\varepsilon^2\kappa_2^2 + \frac{\log(\log\frac{p}{\delta})}{p}\bigg) \quad \text{for} \quad n \ge \frac{cp^6\big(\log\frac{p}{\delta} + \frac{p^2}{\varepsilon^2\kappa_2^2}\big)}{\varepsilon^4\kappa_2^4}. \end{aligned} \tag{3.139}$$

Then, $\widehat{\mathbf{B}}^{(w)} \triangleq \widehat{\mathbf{F}}^{(w)}\widehat{\beta}^{(v)}$ is an estimate of $\mathbf{B}^{(w)}$ such that Eq. (3.29) holds as we have

$$\begin{aligned} \|\widehat{\mathbf{B}}^{(w)} - \mathbf{B}^{(w)}\|_{2,\infty} &= \|\widehat{\mathbf{F}}^{(w)}\widehat{\beta}^{(v)} - \mathbf{F}^{(w)}\beta^{(v)}\|_{2,\infty} \\ &= \|\big(\widehat{\mathbf{F}}^{(w)} - \mathbf{F}^{(w)}\big)\widehat{\beta}^{(v)} + \mathbf{F}^{(w)}\big(\widehat{\beta}^{(v)} - \beta^{(v)}\big)\|_{2,\infty} \\ &\overset{(a)}{\le} \|\big(\widehat{\mathbf{F}}^{(w)} - \mathbf{F}^{(w)}\big)\widehat{\beta}^{(v)}\|_{2,\infty} + \|\mathbf{F}^{(w)}\big(\widehat{\beta}^{(v)} - \beta^{(v)}\big)\|_{2,\infty} \\ &\overset{(b)}{\le} \beta_{\max}\|\widehat{\mathbf{F}}^{(w)} - \mathbf{F}^{(w)}\|_{2,\infty} + \|\mathbf{F}^{(w)}\|_{2,\infty}\big|\widehat{\beta}^{(v)} - \beta^{(v)}\big| \\ &\overset{(c)}{\le} \beta_{\max}\|\widehat{\mathbf{F}}^{(w)} - \mathbf{F}^{(w)}\|_{2,\infty} + \overline{\beta}_{\max}/\beta_i^{(v)}\sqrt{1+d}\big|\widehat{\beta}^{(v)} - \beta^{(v)}\big| \\ &\overset{(d)}{\le} \beta_{\max}\varepsilon + \overline{\beta}_{\max}/\beta_{\min}\sqrt{1+d}\varepsilon, \end{aligned}$$

where (a) follows from triangle inequality, (b) follows from Assumption 3.4(b), (c) follows from Eqs. (3.26) to (3.28) and Assumption 3.4(b), and (d) follows from Eqs. (3.138) and (3.139). The proof of Eq. (3.29) is complete by re-parameterizing ε.

Similarly, $\widehat{\mathbf{c}}^{(i)} \triangleq \widehat{\mathbf{m}}^{(i)}\widehat{\beta}^{(v)}$ is an estimate of $\mathbf{c}^{(i)}$ such that Eq. (3.30) holds. The proof is similar to above.

Proof of Eq. (3.138). Using $\Phi^{(y,y)} = -\frac{1}{\beta^{(v)}}\mathbf{B}^{(y)}\Phi^{(v,y)}$ and Eq. (3.28), we have

$$\beta^{(v)} = \frac{\Phi_{p,t}^{(v,y)}}{\Phi_{p,t}^{(y,y)}} \quad \text{for all} \quad t \in [p]. \tag{3.140}$$

Then, we define

$$\widehat{\beta}^{(v)} \triangleq \frac{1}{p}\mathbf{1}^\top\widetilde{\boldsymbol{\beta}}^{(v)} \quad \text{where} \quad \widetilde{\boldsymbol{\beta}}^{(v)} \in \mathbb{R}^p \quad \text{such that} \quad \widetilde{\boldsymbol{\beta}}_t^{(v)} \triangleq \frac{\widehat{\Phi}_{p,t}^{(v,y)}}{\widehat{\Phi}_{p,t}^{(y,y)}} \quad \text{for all} \quad t \in [p]. \tag{3.141}$$

We have

$$\big|\widehat{\beta}^{(v)} - \beta^{(v)}\big| = \frac{1}{p}\big|\mathbf{1}^\top\widetilde{\boldsymbol{\beta}}^{(v)} - \beta^{(v)}\mathbf{1}^\top\mathbf{1}\big| \overset{(a)}{\le} \frac{1}{\sqrt{p}}\|\widetilde{\boldsymbol{\beta}}^{(v)} - \beta^{(v)}\mathbf{1}\|_2, \tag{3.142}$$

where (a) follows from Cauchy–Schwarz inequality. It remains to bound the right hand side of Eq. (3.142) by $\varepsilon\sqrt{p}$.

Case 1: $\|\big(\widetilde{\boldsymbol{\beta}}^{(v)} - \beta^{(v)}\mathbf{1}\big) \odot \Phi_p^{(y,y)}\|_2 \geq \|\widetilde{\boldsymbol{\beta}}^{(v)} \odot \big(\widehat{\Phi}_p^{(y,y)} - \Phi_p^{(y,y)}\big)\|_2$. From Eqs. (3.140) and (3.141), we have

$$\begin{aligned}
\|\widehat{\Phi}_p^{(v,y)} - \Phi_p^{(v,y)}\|_2 &= \|\widetilde{\boldsymbol{\beta}}^{(v)} \odot \widehat{\Phi}_p^{(y,y)} - \beta^{(v)}\mathbf{1} \odot \Phi_p^{(y,y)}\|_2 \\
&= \|\widetilde{\boldsymbol{\beta}}^{(v)} \odot \widehat{\Phi}_p^{(y,y)} - \widetilde{\boldsymbol{\beta}}^{(v)} \odot \Phi_p^{(y,y)} + \widetilde{\boldsymbol{\beta}}^{(v)} \odot \Phi_p^{(y,y)} - \beta^{(v)}\mathbf{1} \odot \Phi_p^{(y,y)}\|_2 \\
&\overset{(a)}{\geq} \|\big(\widetilde{\boldsymbol{\beta}}^{(v)} - \beta^{(v)}\mathbf{1}\big) \odot \Phi_p^{(y,y)}\|_2 - \|\widetilde{\boldsymbol{\beta}}^{(v)} \odot \big(\widehat{\Phi}_p^{(y,y)} - \Phi_p^{(y,y)}\big)\|_2 \\
&\overset{(b)}{\geq} \sqrt{\kappa_1}\|\widetilde{\boldsymbol{\beta}}^{(v)} - \beta^{(v)}\mathbf{1}\|_2 - \beta_{\max}\|\widehat{\Phi}_p^{(y,y)} - \Phi_p^{(y,y)}\|_2, \qquad (3.143)
\end{aligned}$$

where (a) follows from triangle inequality and (b) follows from Assumption 3.6(a) and Assumption 3.4(b). Then, re-arranging Eq. (3.143), we have

$$\|\widetilde{\boldsymbol{\beta}}^{(v)} - \beta^{(v)}\mathbf{1}\|_2 \leq \frac{1}{\sqrt{\kappa_1}}\Big(\|\widehat{\Phi}_p^{(v,y)} - \Phi_p^{(v,y)}\|_2 + \beta_{\max}\|\widehat{\Phi}_p^{(y,y)} - \Phi_p^{(y,y)}\|_2\Big) \overset{(a)}{\leq} \frac{1}{\sqrt{\kappa_1}}\big(\varepsilon_1 + \beta_{\max}\varepsilon_1\big),$$

where (a) follows from a bound similar to Eq. (3.131). The proof is complete by choosing $\varepsilon_1 = c\varepsilon\sqrt{p}$.

Case 2: $\|\big(\widetilde{\boldsymbol{\beta}}^{(v)} - \beta^{(v)}\mathbf{1}\big) \odot \Phi_p^{(y,y)}\|_2 \leq \|\widetilde{\boldsymbol{\beta}}^{(v)} \odot \big(\widehat{\Phi}_p^{(y,y)} - \Phi_p^{(y,y)}\big)\|_2$. We have

$$\begin{aligned}
\sqrt{\kappa_1}\|\widetilde{\boldsymbol{\beta}}^{(v)} - \beta^{(v)}\mathbf{1}\|_2 \overset{(a)}{\leq} \|\big(\widetilde{\boldsymbol{\beta}}^{(v)} - \beta^{(v)}\mathbf{1}\big) \odot \Phi_p^{(y,y)}\|_2 &\leq \|\widetilde{\boldsymbol{\beta}}^{(v)} \odot \big(\widehat{\Phi}_p^{(y,y)} - \Phi_p^{(y,y)}\big)\|_2 \\
&\overset{(b)}{\leq} \beta_{\max}\|\widehat{\Phi}_p^{(y,y)} - \Phi_p^{(y,y)}\|_2 \overset{(c)}{\leq} \beta_{\max}\varepsilon_1,
\end{aligned}$$

where (a) follows from Assumption 3.6(a), (b) follows from Assumption 3.4(b), and (c) follows from a bound similar to Eq. (3.131). The proof is complete by choosing $\varepsilon_1 = c\varepsilon\sqrt{p}$.

3.E.1.1.6 No dynamics in the interventions In this scenario, we have $\beta_{t_2,t_1}^{(a)} = 0$ if $t_1 \neq t_2$ and $\beta_{t_2,t_1}^{(a)} = \beta^{(a)}$ otherwise, i.e., $\mathbf{B}^{(a)} = -\beta^{(a)}\mathbf{I}$ and $\Phi^{(z,y)} = -\frac{1}{\beta^{(a)}}\Phi^{(a,y)}$. Then, for every $i \in [n]$, $\theta^{(i)}$ can be expressed as a linear combination of known vectors. In particular,

$$\theta^{(i)} = -\frac{1}{\beta^{(a)}}\mathbf{D}^{(i)}\mathbf{c}^{(i)},$$

where

$$\mathbf{D}^{(i)} \triangleq -2\Phi^{(a,y)\top}\boldsymbol{o}^{(i)} \in \mathbb{R}^{p\times 3} \quad \text{and} \quad \mathbf{c}^{(i)} \triangleq \begin{bmatrix} z^{(i)} \\ \beta_i^{(v)} \\ \beta_i^{(a)} \end{bmatrix} \in \mathbb{R}^{3\times 1}.$$

Then, it is easy to see that the rest of the proof is similar to the setting with no dynamics in the observed covariates.

3.E.1.2 Guarantees for Example 3.8

In this setting, $\boldsymbol{z}^{(i)} = \boldsymbol{z}$ for all $i \in [n]$. Proceeding as in Section 3.E.1.1 to obtain Eq. (3.123), for every unit $i \in [n]$, we have

$$\begin{aligned} & f_{\mathbf{y}|\mathbf{a},\mathbf{v},\mathbf{z}}(\boldsymbol{y}^{(i)}|\boldsymbol{a}^{(i)}, \boldsymbol{v}^{(i)}, \boldsymbol{z}; \theta, \Theta) \\ & \quad \propto \exp\left(\theta^\top \boldsymbol{y}^{(i)} + 2\boldsymbol{v}^{(i)^\top} \Phi^{(v,y)} \boldsymbol{y}^{(i)} + 2\boldsymbol{a}^{(i)^\top} \Phi^{(a,y)} \boldsymbol{y}^{(i)} + \boldsymbol{y}^{(i)^\top} \Phi^{(y,y)} \boldsymbol{y}^{(i)}\right), \end{aligned}$$

where θ and Θ are as follows

$$\theta \triangleq 2\Phi^{(z,y)\top} \boldsymbol{z} \quad \text{and} \quad \Theta \triangleq \left[\Phi^{(y,y)}, \Phi^{(y,a)}, \Phi^{(y,v)}\right].$$

Then, using the methodology and analysis from Shah et al. (2023) (which is closely related to the one in this work) to obtain estimates $\widehat{\theta}$ and $\widehat{\Theta}$ such that with probability at least $1-\delta$, we have

$$\max\left\{\|\widehat{\theta} - \theta\|_2, \|\widehat{\Theta} - \Theta^\star\|_{2,\infty}\right\} \leq \varepsilon_1 \quad \text{whenever} \quad n \geq \frac{c \log \frac{p}{\sqrt{\delta}}}{\varepsilon_1^2}. \tag{3.144}$$

The rest of the proof is similar to Section 3.E.1.1 with the choice of $\varepsilon_1 = c\kappa_1 \varepsilon^2 / \sqrt{p \log p}$ in Eq. (3.144) for the guarantee in Eq. (3.31), and the choice of $\varepsilon_1 = c\kappa_1 \varepsilon^2 / p\sqrt{p \log p}$ in Eq. (3.144) for the guarantee in Eq. (3.32) (as $\|\boldsymbol{z}\|_2^2 = O(p)$).

3.E.1.3 Proof of Lemma 3.15: Coefficient recovery in a linear combination with noisy basis

We start by expressing θ in terms of $\widehat{\mathbf{D}}$, i.e.,

$$\theta = \widehat{\mathbf{D}}\widetilde{\mathbf{c}} \quad \text{where} \quad \widetilde{\mathbf{c}} \triangleq \mathbf{c} + \zeta, \tag{3.145}$$

for some error term ζ. Then, ζ can be controlled in following manner

$$\begin{aligned} \|\widehat{\mathbf{D}}\zeta\|_2 \overset{Eq.\ (3.145)}{=} \|\theta - \widehat{\mathbf{D}}\mathbf{c}\|_2 = \|\mathbf{D}\mathbf{c} - \widehat{\mathbf{D}}\mathbf{c}\|_2 & \overset{(a)}{\leq} \|\mathbf{D} - \widehat{\mathbf{D}}\|_{\text{op}} \|\mathbf{c}\|_2 \\ & \overset{(b)}{\leq} \left(\sqrt{p}\|\mathbf{D} - \widehat{\mathbf{D}}\|_{2,\infty}\right) \cdot \left(\sqrt{k}\|\mathbf{c}\|_\infty\right) \\ & \overset{(c)}{\leq} \sqrt{kp}\|\mathbf{c}\|_\infty \bar{\varepsilon}_1, \end{aligned} \tag{3.146}$$

where (a) follows from sub-multiplicativity of induced matrix norms, (b) follows from standard matrix norm inequalities, and (c) follows from Eq. (3.136).

Case 1: $\|\widehat{\mathbf{D}}\widehat{\mathbf{c}} - \widehat{\mathbf{D}}\mathbf{c}\|_2 \geq \|\widehat{\mathbf{D}}\zeta\|_2$. From Eq. (3.145) and triangle inequality, we find that

$$\|\widehat{\theta} - \theta\|_2 = \|\widehat{\mathbf{D}}\widehat{\mathbf{c}} - \widehat{\mathbf{D}}\mathbf{c} - \widehat{\mathbf{D}}\zeta\|_2 \geq \|\widehat{\mathbf{D}}\widehat{\mathbf{c}} - \widehat{\mathbf{D}}\mathbf{c}\|_2 - \|\widehat{\mathbf{D}}\zeta\|_2. \tag{3.147}$$

Then, doing standard algebra with Eq. (3.147) yields that

$$\text{MSE}(\widehat{\theta}, \theta) + \frac{\|\widehat{\mathbf{D}}\zeta\|_2^2}{p} \geq \frac{\|\widehat{\mathbf{D}}\widehat{\mathbf{c}} - \widehat{\mathbf{D}}\mathbf{c}\|_2^2}{2p} = \frac{(\widehat{\mathbf{c}} - \mathbf{c})^\top \widehat{\mathbf{D}}^\top \widehat{\mathbf{D}}(\widehat{\mathbf{c}} - \mathbf{c})}{2p}. \tag{3.148}$$

Then, by assumption on the eigenvalues of $\widehat{\mathbf{D}}^\top \widehat{\mathbf{D}}$, we have

$$\frac{\lambda_{\min}(\widehat{\mathbf{D}}^\top \widehat{\mathbf{D}})}{2p}\|\widehat{\mathbf{c}}-\mathbf{c}\|_2^2 \leq \frac{(\widehat{\mathbf{c}}-\mathbf{c})^\top \widehat{\mathbf{D}}^\top \widehat{\mathbf{D}}(\widehat{\mathbf{c}}-\mathbf{c})}{2p} \overset{(a)}{\leq} \bar{\varepsilon}_2 + k\|\mathbf{c}\|_\infty^2 \bar{\varepsilon}_1^2.$$

where (a) follows from Eqs. (3.136), (3.146), and (3.148).

Case 2: $\|\widehat{\mathbf{D}}\widehat{\mathbf{c}} - \widehat{\mathbf{D}}\mathbf{c}\|_2 \leq \|\widehat{\mathbf{D}}\zeta\|_2$**.** By assumption on the eigenvalues of $\widehat{\mathbf{D}}^\top \widehat{\mathbf{D}}$, we have

$$\frac{\lambda_{\min}(\widehat{\mathbf{D}}^\top \widehat{\mathbf{D}})}{2p}\|\widehat{\mathbf{c}}-\mathbf{c}\|_2^2 \leq \frac{(\widehat{\mathbf{c}}-\mathbf{c})^\top \widehat{\mathbf{D}}^\top \widehat{\mathbf{D}}(\widehat{\mathbf{c}}-\mathbf{c})}{2p} = \frac{\|\widehat{\mathbf{D}}\widehat{\mathbf{c}} - \widehat{\mathbf{D}}\mathbf{c}\|_2^2}{2p} \leq \frac{\|\widehat{\mathbf{D}}\zeta\|_2^2}{2p} \overset{(a)}{\leq} 2k\|\mathbf{c}\|_\infty^2 \bar{\varepsilon}_1^2,$$

where (a) follows from Eq. (3.146).

3.E.2 Proof of Proposition 3.3

3.E.2.1 Guarantees for Example 3.9

We divide the proof into two parts for convenience. First, we obtain the conditional distribution of the outcome vector given the intervention vector, the observed covariates, and the unobserved covariate. Then, we provide guarantees on recovering the parameters.

3.E.2.1.1 Obtaining the conditional distribution of the outcome For every unit $i \in [n]$ and time $t \in [p]$, the distribution of the outcome $\mathsf{y}_t = y_t^{(i)}$ given $\mathbf{y}_{1:t-1} = \boldsymbol{y}_{1:t-1}^{(i)}$, $\mathbf{a}_{1:t} = \boldsymbol{a}_{1:t}^{(i)}$, $\mathbf{v}_{1:t} = \boldsymbol{v}_{1:t}^{(i)}$, and $\boldsymbol{z}_{1:t} = z^{(i)}\mathbf{1}$ is given by

$$\begin{aligned}
&f_{\mathsf{y}_t|\mathbf{y}_{1:t-1},\mathbf{a}_{1:t},\mathbf{v}_{1:t},\mathbf{z}_{1:t}}(y_t^{(i)}|\boldsymbol{y}_{1:t-1}^{(i)}, \boldsymbol{a}_{1:t}^{(i)}, \boldsymbol{v}_{1:t}^{(i)}, z^{(i)}\mathbf{1}) \\
&\propto \exp\left(\left[z^{(i)} + \beta_i^{(v)} v_t^{(i)} + \beta_i^{(a)} a_t^{(i)} + \sum_{j=t-d}^{t} \beta_{t,j}^{(v)} v_j^{(i)} + \sum_{j=t-d}^{t} \beta_{t,j}^{(a)} a_j^{(i)} + \sum_{j=t-d}^{t-1} \beta_{t,j}^{(y)} y_j^{(i)}\right] y_t^{(i)}\right).
\end{aligned} \tag{3.149}$$

Then, the distribution of $\mathbf{y} = \boldsymbol{y}^{(i)}$ given $\mathbf{a} = \boldsymbol{a}^{(i)}$, $\mathbf{v} = \boldsymbol{v}^{(i)}$, and $\mathbf{z} = z^{(i)}\mathbf{1}$ is given by

$$\begin{aligned}
&f_{\mathbf{y}|\mathbf{a},\mathbf{v},\mathbf{z}}(\boldsymbol{y}^{(i)}|\boldsymbol{a}^{(i)}, \boldsymbol{v}^{(i)}, z^{(i)}\mathbf{1}) \\
&\overset{(a)}{=} \prod_{t\in[p]} f_{\mathsf{y}_t|\mathbf{y}_{1:t-1},\mathbf{a},\mathbf{z},\mathbf{v}}(y_t^{(i)}|\boldsymbol{y}_{1:t-1}^{(i)}, \boldsymbol{a}^{(i)}, \boldsymbol{z}^{(i)}, \boldsymbol{v}^{(i)}) \\
&\overset{(b)}{=} \prod_{t\in[p]} f_{\mathsf{y}_t|\mathbf{y}_{1:t-1},\mathbf{a}_{1:t},\mathbf{z}_{1:t},\mathbf{v}_{1:t}}(y_t^{(i)}|\boldsymbol{y}_{1:t-1}^{(i)}, \boldsymbol{a}_{1:t}^{(i)}, z^{(i)}\mathbf{1}, \boldsymbol{v}_{1:t}^{(i)}) \\
&\overset{(c)}{\propto} \prod_{t\in[p]} \exp\left(\left[z^{(i)} + \beta_i^{(v)} v_t^{(i)} + \beta_i^{(a)} a_t^{(i)} + \sum_{j=t-d}^{t} \beta_{t,j}^{(v)} v_j^{(i)} + \sum_{j=t-d}^{t} \beta_{t,j}^{(a)} a_j^{(i)} + \sum_{j=t-d}^{t-1} \beta_{t,j}^{(y)} y_j^{(i)}\right] y_t^{(i)}\right),
\end{aligned} \tag{3.150}$$

where (a) follows by the chain rule, (b) follows because $\mathsf{y}_t \perp\!\!\!\perp \mathbf{a}_{t+1:p}, \mathbf{z}_{t+1:p}, \mathbf{v}_{t+1:p} \mid \mathbf{y}_{1:t-1}, \mathbf{a}_{1:t}, \mathbf{z}_{1:t}, \mathbf{v}_{1:t}$ as per Figure 3.1.1, and (c) follows from Eq. (3.149). To convert Eq. (3.150) into the familiar form, we define the interaction matrices

$$\Phi^{(v,y)} \triangleq \mathbf{B}^{(v)}, \qquad \Phi^{(a,y)} \triangleq \mathbf{B}^{(a)}, \quad \text{and} \quad \Phi^{(y,y)} \triangleq (\mathbf{B}^{(y)} + \mathbf{B}^{(y)^\top})/2. \tag{3.151}$$

Then, we define the unit-level and the population-level parameters

$$\theta^{(i)} \triangleq \Big[z^{(i)}\mathbf{1} + \beta_i^{(v)}\boldsymbol{v}^{(i)} + \beta_i^{(a)}\boldsymbol{a}^{(i)}\Big] \quad \text{and} \quad \Theta \triangleq \big[\Phi^{(y,y)}, \Phi^{(y,a)}, \Phi^{(y,v)}\big], \tag{3.152}$$

with $\Phi^{(y,a)} = \Phi^{(a,y)^\top} \in \mathbb{R}^{p\times p_a}$ and $\Phi^{(y,v)} = \Phi^{(v,y)^\top} \in \mathbb{R}^{p\times p_v}$. Putting together Eqs. (3.26) to (3.28) and (3.150) to (3.152), we have

$$\begin{aligned} &f_{\mathbf{y}|\mathbf{a},\mathbf{v},\mathbf{z}}(\boldsymbol{y}^{(i)}|\boldsymbol{a}^{(i)}, \boldsymbol{v}^{(i)}, z^{(i)}\mathbf{1}; \theta^{(i)}, \Theta) \\ &\qquad \propto \exp\bigg(\theta^{(i)^\top}\boldsymbol{y}^{(i)} + \boldsymbol{v}^{(i)^\top}\Phi^{(v,y)}\boldsymbol{y}^{(i)} + \boldsymbol{a}^{(i)^\top}\Phi^{(a,y)}\boldsymbol{y}^{(i)} + \boldsymbol{y}^{(i)^\top}\Phi^{(y,y)}\boldsymbol{y}^{(i)}\bigg). \end{aligned} \tag{3.153}$$

3.E.2.1.2 Assumption 3.1 holds To enable the application of the machinery developed in Section 3.3.3, we show that the unit-level and the population-level parameters defined in Eq. (3.152) satisfy Assumption 3.1. We have

$$\begin{aligned} \|\theta^{(i)}\|_\infty &\overset{Eq.\ (3.152)}{=} \|z^{(i)}\mathbf{1} + \beta_i^{(v)}\boldsymbol{v}^{(i)} + \beta_i^{(a)}\boldsymbol{a}^{(i)}\|_\infty \overset{(a)}{\leq} z_{\max} + 2\beta_{\max}x_{\max} \\ \|\Theta\|_{\max} &\overset{Eq.\ (3.152)}{=} \max_{u\in\{v,a,y\}} \|\Phi^{(u,y)}\|_{\max} \overset{Eq.\ (3.151)}{=} \max_{u\in\{v,a,y\}} \|\mathbf{B}^{(u)}\|_{\max} \overset{(b)}{\leq} \overline{\beta}_{\max}, \text{ and} \\ \|\Theta\|_\infty &\overset{Eq.\ (3.152)}{\leq} \sum_{u\in\{y,a,v\}} \|\Phi^{(u,y)}\|_1 \overset{(c)}{\leq} 1 + d\beta_{\max} + 2(d+1)\beta_{\max}, \end{aligned}$$

where (a) follows from Assumption 3.4 and because $\max\{\boldsymbol{v}^{(i)}, \boldsymbol{a}^{(i)}\} \leq x_{\max}$ for all $i \in [n]$, (b) follows from Assumption 3.4(b) and (c) follows from Eqs. (3.26) to (3.28) and (3.151), and Assumption 3.4.

3.E.2.1.3 Recovering population-level and unit-level parameters For every $i \in [n]$, $\theta^{(i)}$ in Eq. (3.151) can be expressed as a linear combination of known vectors. In particular,

$$\theta^{(i)} = \boldsymbol{o}^{(i)}\mathbf{c}^{(i)} \quad \text{where} \quad \mathbf{c}^{(i)} \triangleq \begin{bmatrix} z^{(i)} \\ \beta_i^{(v)} \\ \beta_i^{(a)} \end{bmatrix} \in \mathbb{R}^{3\times 1}. \tag{3.154}$$

Applying Corollary 3.1, we obtain estimates $\widehat{\Theta}$ and $\widehat{\theta}^{(i)}$ such that, with probability at least $1-\delta$, we have

$$\|\widehat{\Theta} - \Theta^\star\|_{2,\infty} \leq \varepsilon \quad \text{for} \quad n \geq \frac{cp^2\big(\log\frac{p}{\delta} + p\log\frac{1}{\varepsilon^2} + \frac{1}{\varepsilon^2}\big)}{\varepsilon^4} \tag{3.155}$$

$$\mathrm{MSE}(\widehat{\theta}^{(i)}, \theta^{(i)}) \leq \max\bigg\{\varepsilon^2, \frac{c\log(\log\frac{p}{\delta})}{p}\bigg\} \quad \text{for} \quad n \geq \frac{cp^4\big(\log\frac{p}{\delta} + p\log\frac{p}{\varepsilon^2} + \frac{p}{\varepsilon^2}\big)}{\varepsilon^4}. \tag{3.156}$$

We note that the above estimate $\widehat{\theta}^{(i)}$ is of the form $\widehat{\theta}^{(i)} = \boldsymbol{o}^{(i)}\widehat{\mathbf{c}}^{(i)}$. Then, from Assumption 3.6(c)

$$\mathrm{MSE}(\widehat{\theta}^{(i)}, \theta^{(i)}) = \frac{\|\boldsymbol{o}^{(i)}\mathbf{c}^{(i)} - \boldsymbol{o}^{(i)}\widehat{\mathbf{c}}^{(i)}\|_2^2}{p} = \frac{(\mathbf{c}^{(i)} - \widehat{\mathbf{c}}^{(i)})^\top\boldsymbol{o}^{(i)^\top}\boldsymbol{o}^{(i)}(\mathbf{c}^{(i)} - \widehat{\mathbf{c}}^{(i)})}{p}$$

$$\geq \kappa_3 \|\mathbf{c}^{(i)} - \widehat{\mathbf{c}}^{(i)}\|_2^2. \tag{3.157}$$

Putting together Eqs. (3.151), (3.152), and (3.154) to (3.157) completes the proof.

3.E.2.2 Guarantees for Example 3.10

In this setting, $\boldsymbol{z}^{(i)} = \boldsymbol{z}$ for all $i \in [n]$. Proceeding as in Section 3.E.2.1 to obtain Eq. (3.153), for every unit $i \in [n]$, we have

$$f_{\mathbf{y}|\mathbf{a},\mathbf{v},\mathbf{z}}(\boldsymbol{y}^{(i)}|\boldsymbol{a}^{(i)}, \boldsymbol{v}^{(i)}, \boldsymbol{z}; \theta, \Theta)$$
$$\propto \exp\left(\theta^\top \boldsymbol{y}^{(i)} + \boldsymbol{v}^{(i)^\top} \Phi^{(v,y)} \boldsymbol{y}^{(i)} + \boldsymbol{a}^{(i)^\top} \Phi^{(a,y)} \boldsymbol{y}^{(i)} + \boldsymbol{y}^{(i)^\top} \Phi^{(y,y)} \boldsymbol{y}^{(i)}\right),$$

where θ and Θ are as follows

$$\theta \triangleq \boldsymbol{z} \quad \text{and} \quad \Theta \triangleq \left[\Phi^{(y,y)}, \Phi^{(y,a)}, \Phi^{(y,v)}\right]. \tag{3.158}$$

Then, using the methodology and analysis from Shah et al. (2023) (which is closely related to the one in this work) to obtain estimates $\widehat{\theta}$ and $\widehat{\Theta}$ such that with probability at least $1 - \delta$, we have

$$\max\left\{\|\widehat{\theta} - \theta\|_2, \|\widehat{\Theta} - \Theta^\star\|_{2,\infty}\right\} \leq \varepsilon \quad \text{whenever} \quad n \geq \frac{c \log \frac{p}{\sqrt{\delta}}}{\varepsilon^2}. \tag{3.159}$$

Putting together Eqs. (3.151), (3.158), and (3.159) completes the proof.

3.F Proof of Proposition 3.4: Impute missing covariates

Proof idea. First, we use units $i \in \{n/2 + 1, \cdots, n\}$ without any measurement error to estimate $\phi^{(y)^\star}$ and $\Theta^\star$, i.e., the parameters corresponding to the distribution of $\mathbf{y}$ conditioned on $(\mathbf{a}, \mathbf{v})$ (see Section 3.7.1). Next, for units $i \in \{1, \cdots, n/2\}$ with measurement error, we estimate $\theta^{\star(i)}$ by expressing it as a linear combination of the estimates of $\phi^{(y)^\star}$ and $\Theta^\star$ (enabling the use of Example 3.1). The coefficients of this linear combination turn out to be our estimates of the measurement error $\Delta \boldsymbol{v}^{(i)}$.

Estimate $\phi^{(y)^\star}$ and $\Theta^\star$. For units $i \in \{n/2+1, \cdots, n\}$, under our assumption $\Delta \boldsymbol{v}^{(i)} = 0$ implying $\theta^{\star(i)} = \phi^{(y)^\star}$. Therefore, in addition to the population-level parameter $\Theta^\star$, the unit-level parameter $\theta^{\star(i)} = \phi^{(y)^\star}$ is also shared for these units. Thus, learning $\phi^{(y)^\star}$ and $\Theta^\star$ boils down to learning parameters of a sparse graphical model (because of the assumptions in Section 3.7.1) from $n/2$ samples. We use the methodology and analysis from Shah et al. (2023) (which is closely related to the one in this work) to obtain estimates $\widehat{\phi}$ and $\widehat{\Theta}$ such that with probability at least $1 - \delta$, we have

$$\max\left\{\|\widehat{\phi} - \phi^{(y)^\star}\|_2, \|\widehat{\Theta} - \Theta^\star\|_{2,\infty}\right\} \leq \varepsilon_1 \quad \text{whenever} \quad n \geq \frac{ce^{c'\beta} \log \frac{p_y}{\sqrt{\delta}}}{\varepsilon_1^2}. \tag{3.160}$$

Recover the unit-level parameters. As the first step, for units $i \in \{1, \cdots, n/2\}$, we express the true unit-level parameters $\theta^{\star(i)}$ as a linear combination of known vectors. To that end, fix any $i \in [n/2]$. Then, using Eq. (3.35), we can write $\theta^{\star(i)}$ as a linear combination of $p_v + 1$ vectors, i.e.,

$$\theta^{\star(i)} = \mathbf{D}\mathbf{c}^{(i)},$$

where

$$\mathbf{D} \triangleq \left[\phi^{(y)\star}, 2\Phi^{(y,v)\star}\right] \in \mathbb{R}^{p_y \times (p_v+1)} \quad \text{and} \quad \mathbf{c}^{(i)} \triangleq \begin{bmatrix} 1 \\ \Delta\boldsymbol{v}^{(i)} \end{bmatrix} \in \mathbb{R}^{(p_v+1)\times 1}. \tag{3.161}$$

While we do not know the matrix $\mathbf{D}$, we can produce an estimate $\widehat{\mathbf{D}}$ using $\widehat{\phi}$ and $\widehat{\Theta}$. Let $\widehat{\Phi}^{(y,v)}$ be the component of $\widehat{\Theta}$ that is an estimate of $\Phi^{(y,v)\star}$, and define $\widehat{\mathbf{D}} \triangleq \left[\widehat{\phi}, 2\widehat{\Phi}^{(y,v)}\right]$. This estimate is such that, with probability at least $1-\delta$,

$$\|\widehat{\mathbf{D}} - \mathbf{D}\|_{2,\infty} \le \varepsilon_1 \quad \text{whenever} \quad n \ge \frac{ce^{c'\beta}\log\frac{p_y}{\sqrt{\delta}}}{\varepsilon_1^2}. \tag{3.162}$$

This guarantee follows directly from Eq. (3.160) and the definition of $\mathbf{D}$ in Eq. (3.161).

Now, we write

$$\theta^{\star(i)} = \widehat{\mathbf{D}}\widetilde{\mathbf{c}}^{(i)} \quad \text{where} \quad \widetilde{\mathbf{c}}^{(i)} \triangleq \mathbf{c}^{(i)} + \zeta,$$

for some error term ζ. Then, performing an analysis similar to one in Section 3.C while using the bound on n in Eq. (3.160) instead of the one in Eq. (3.17), and using Example 3.1, we obtain estimates $\widehat{\theta}^{(1)}, \cdots, \widehat{\theta}^{(n/2)}$ such that (see Corollary 3.1(a) for reference), with probability at least $1-\delta$, we have

$$\max_{i\in[n/2]} \mathrm{MSE}(\widehat{\theta}^{(i)}, \theta^{\star(i)}) \le \max\left\{\varepsilon_1^2, \frac{ce^{c'\beta}\big(p_v + \log(\log\frac{np_y}{\delta})\big)}{p_y}\right\}, \tag{3.163}$$

whenever $n \ge ce^{c'\beta}\varepsilon_1^{-2}\widetilde{p}^2\big(\log\frac{\sqrt{n}p_y}{\sqrt{\delta}} + p_v\big)$. We note that the above estimate $\widehat{\theta}^{(i)}$ of the unit-level parameter $\theta^{\star(i)}$ is of the form $\widehat{\theta}^{(i)} = \widehat{\mathbf{D}}\widehat{\mathbf{c}}^{(i)}$ for $i \in [n/2]$.

Recover the measurement error. We condition on the events Eqs. (3.162) and (3.163) happening. Then, we declare $\widehat{\mathbf{c}}^{(i)}$ as our estimate of the measurement error for unit $i \in [n/2]$ and prove the corresponding guarantee below by invoking Lemma 3.15.

We claim that the eigenvalues of $\widehat{\mathbf{D}}^\top\widehat{\mathbf{D}}$ can be lower bounded by $\kappa p_y/2$ with the choice $\varepsilon_1 = \kappa\varepsilon_2/\alpha\sqrt{p_y(p_v+1)}$, whenever $\varepsilon_2 \le \sqrt{p_y/(p_v+1)}/8$. Taking this claim as given at the moment, we continue our proof. From Lemma 3.15, with probability at least $1-\delta$, we have

$$\frac{1}{p_y}\max_{i\in[n/2]}\|\widehat{\mathbf{c}}^{(i)} - \mathbf{c}^{(i)}\|_2^2 \le \frac{4}{\kappa p_y}\left(p_v\left\|\mathbf{c}^{(i)}\right\|_\infty^2\varepsilon_1^2 + \max\left\{\varepsilon_1^2, \frac{ce^{c'\beta}\big(p_v + \log(\log\frac{np_y}{\delta})\big)}{p_y}\right\}\right)$$

$$\overset{(a)}{\leq} \frac{4}{\kappa p_y}\biggl(\frac{\kappa^2\varepsilon_2^2}{p_y} + \frac{c e^{c'\beta}\bigl(p_v + \log(\log\frac{np_y}{\delta})\bigr)}{p_y}\biggr), \tag{3.164}$$

where (a) follows from the choice of ε_1 and because $\bigl\|\mathbf{c}^{(i)}\bigr\|_\infty \leq \alpha$. Rearranging Eq. (3.164) completes the proof.

It remains to show that the eigenvalues of $\widehat{\mathbf{D}}^\top\widehat{\mathbf{D}}$ can be lower bounded by $\kappa p_y/2$ with the choice $\varepsilon_1 = \kappa\varepsilon_2/\alpha\sqrt{p_y(p_v+1)}$. From Weyl's inequality (Bhatia, 2007, Theorem. 8.2), we have

$$\lambda_{\min}(\widehat{\mathbf{D}}^\top\widehat{\mathbf{D}}) \geq \lambda_{\min}(\mathbf{D}^\top\mathbf{D}) - \lambda_{\max}(\mathbf{D}^\top\mathbf{D} - \widehat{\mathbf{D}}^\top\widehat{\mathbf{D}}) \overset{(a)}{\geq} \kappa p_y - \lambda_{\max}(\mathbf{D}^\top\mathbf{D} - \widehat{\mathbf{D}}^\top\widehat{\mathbf{D}}),$$

where (a) follows from the assumption on the eigenvalues of $\mathbf{D}^\top\mathbf{D}$. Now, it suffices to upper bound $\lambda_{\max}(\mathbf{D}^\top\mathbf{D} - \widehat{\mathbf{D}}^\top\widehat{\mathbf{D}})$ by $\kappa p_y/2$. We have

$$\begin{aligned}
\bigl|\lambda_{\max}(\mathbf{D}^\top\mathbf{D} - \widehat{\mathbf{D}}^\top\widehat{\mathbf{D}})\bigr| &\overset{(a)}{=} \|\mathbf{D}^\top\mathbf{D} - \widehat{\mathbf{D}}^\top\widehat{\mathbf{D}}\|_{\mathrm{op}} \\
&\overset{(b)}{\leq} (p_v+1)\|\mathbf{D}^\top\mathbf{D} - \widehat{\mathbf{D}}^\top\widehat{\mathbf{D}}\|_{\max} \\
&\overset{(c)}{\leq} (p_v+1)\Bigl(\|\mathbf{D}^\top(\mathbf{D}-\widehat{\mathbf{D}})\|_{\max} + \|(\mathbf{D}-\widehat{\mathbf{D}})^\top\widehat{\mathbf{D}}\|_{\max}\Bigr) \\
&\overset{(d)}{\leq} (p_v+1)\bigl(\|\mathbf{D}^\top\|_{2,\infty} + \|\widehat{\mathbf{D}}^\top\|_{2,\infty}\bigr)\|\mathbf{D}-\widehat{\mathbf{D}}\|_{1,2} \\
&\overset{(e)}{\leq} (p_v+1)(2\alpha\sqrt{p_y} + 2\alpha\sqrt{p_y})\cdot\sqrt{p_y}\varepsilon_1 \overset{(f)}{\leq} 4\kappa\varepsilon_2\sqrt{p_v+1}\sqrt{p_y} \overset{(g)}{\leq} \frac{\kappa p_y}{2},
\end{aligned}$$

where (a) follows because $\mathbf{D}^\top\mathbf{D} - \widehat{\mathbf{D}}^\top\widehat{\mathbf{D}}$ is symmetric, (b) follows from because $\|\mathbf{M}\|_{\mathrm{op}} \leq \|\mathbf{M}\|_{\mathrm{F}} \leq k\|\mathbf{M}\|_{\max}$ for any square matrix $\mathbf{M} \in \mathbb{R}^{k\times k}$, (c) follows from the triangle inequality, (d) follows by Cauchy–Schwarz inequality, (e) follows because $\|\widehat{\mathbf{D}}\|_{\max} \leq 2\alpha$, $\|\mathbf{D}\|_{\max} \leq 2\alpha$ (because of the assumptions in Section 3.7.1), and from Eqs. (3.160) and (3.161), (f) follows from the choice of ε_1, and (g) follows whenever $\varepsilon_2 \leq \frac{1}{8}\sqrt{\frac{p_y}{p_v+1}}$.

3.G Logarithmic Sobolev inequality and tail bounds

In this section, we present two results which may be of independent interest. First, we show that a random vector supported on a compact set satisfies the logarithmic Sobolev inequality (to be defined) if it satisfies the Dobrushin's uniqueness condition (to be defined). This result is a generalization of the result in Marton (2015) for discrete random vectors to continuous random vectors supported on a compact set. Next, we show that if a random vector satisfies the logarithmic Sobolev inequality, then any arbitrary function of the random vector concentrates around its mean. This result is a generalization of the result in Dagan et al. (2021) for discrete random vectors to continuous random vectors.

Throughout this section, we consider a p-dimensional random vector $\mathbf{u}$ supported on $\mathcal{U}^p$ with distribution $f_{\mathbf{u}}$ where $p \geq 1$. We start by defining the logarithmic Sobolev inequality (LSI). We use the convention $0\log 0 = 0$.

Definition 3.3 (Logarithmic Sobolev inequality)**.** *A random vector* $\mathbf{u}$ *satisfies the logarithmic Sobolev inequality with constant* $\sigma^2 > 0$ *(abbreviated as* $\mathrm{LSI}_{\mathbf{u}}(\sigma^2)$*) if*

$$\mathsf{Ent}_{\mathbf{u}}\left(q^2\right) \leq \sigma^2 \mathbb{E}_{\mathbf{u}}\Big[\big\|\nabla_{\mathbf{u}} q(\mathbf{u})\big\|_2^2\Big] \quad \textit{for all} \quad q : \mathcal{U}^p \to \mathbb{R}, \tag{3.165}$$

where $\mathsf{Ent}_{\mathbf{u}}(g) \triangleq \mathbb{E}_{\mathbf{u}}[g(\mathbf{u}) \log g(\mathbf{u})] - \mathbb{E}_{\mathbf{u}}[g(\mathbf{u})] \log \mathbb{E}_{\mathbf{u}}[g(\mathbf{u})]$ *denotes the entropy of the function* $g : \mathcal{U}^p \to \mathbb{R}_+$*.*

Next, we state the Dobrushin's uniqueness condition. For any distributions g and f, let $\|f - g\|_{\mathsf{TV}}$ denote the total variation distance between g and f.

Definition 3.4 (Dobrushin's uniqueness condition)**.** *A random vector* $\mathbf{u}$ *satisfies the Dobrushin's uniqueness condition with coupling matrix* $\Theta \in \mathbb{R}_+^{p \times p}$ *if* $\|\Theta\|_{\mathrm{op}} < 1$*, and for every* $t \in [p], s \in [p] \backslash \{t\}$*, and* $\boldsymbol{u}_{-t}, \widetilde{\boldsymbol{u}}_{-t} \in \mathcal{U}^{p-1}$ *differing only in the* s^{th} *coordinate,*

$$\|f_{\mathbf{u}_t | \mathbf{u}_{-t} = \boldsymbol{u}_{-t}} - f_{\mathbf{u}_t | \mathbf{u}_{-t} = \widetilde{\boldsymbol{u}}_{-t}}\|_{\mathsf{TV}} \leq \Theta_{ts}. \tag{3.166}$$

We note that the Dobrushin's uniqueness condition, as originally stated (see Marton (2015)) for Ising model, also requires $\Theta_{tt} = 0$ for all $t \in [p]$. This condition makes sense for Ising model where $\boldsymbol{u}_t^2 = 1$ for all $t \in [p]$. However, this is not true for continuous random vectors necessitating a need for modification in the condition.

From hereon, we let $\mathcal{U}^p$ be compact unless otherwise specified. Moreover, we define

$$f_{\min} \triangleq \min_{t \in [p], \boldsymbol{u} \in \mathcal{U}^p} f_{\mathbf{u}_t | \mathbf{u}_{-t}}(u_t | \boldsymbol{u}_{-t}). \tag{3.167}$$

Now, we provide the first main result of this section with a proof in Section 3.G.1.

Proposition 3.5 (Logarithmic Sobolev inequality)**.** *If a random vector* $\mathbf{u}$ *with* $f_{\min} > 0$ *(see Eq. (3.167)) satisfies (a) the Dobrushin's uniqueness condition (Definition 3.4) with coupling matrix* $\Theta \in \mathbb{R}_+^{p \times p}$*, and (b)* $\mathbf{u}_t | \mathbf{u}_{-t}$ *satisfies* $\mathrm{LSI}_{\mathbf{u}_t | \mathbf{u}_{-t} = \boldsymbol{u}_{-t}}(\sigma^2)$ *for all* $t \in [p]$ *and* $\boldsymbol{u}_{-t} \in \mathcal{U}^{p-1}$ *(see Definition 3.3), then it satisfies* $\mathrm{LSI}_{\mathbf{u}}(2\sigma^2 / (f_{\min}(1 - \|\Theta\|_{\mathrm{op}})^2))$*.*

Next, we define the notion of pseudo derivative and pseudo Hessian that come in handy in our proofs for providing upper bounds on the norm of the derivative and the Hessian.

Definition 3.5 (Pseudo derivative and Hessian)**.** *For a function* $q : \mathcal{U}^p \to \mathbb{R}$*, the functions* $\widetilde{\nabla} q : \mathcal{U}^p \to \mathbb{R}^{p_1}$ *and* $\widetilde{\nabla}^2 q : \mathcal{U}^p \to \mathbb{R}^{p_1 \times p_2}$ *(*$p_1, p_2 \geq 1$*) are, respectively, called a pseudo derivative and a pseudo Hessian for* q *if for all* $\boldsymbol{w} \in \mathcal{U}^p$ *and* $\rho \in \mathbb{R}^{p_1 \times 1}$*, we have*

$$\|\widetilde{\nabla} q(\boldsymbol{w})\|_2 \geq \|\nabla q(\boldsymbol{w})\|_2 \quad \textit{and} \quad \|\rho^\top \widetilde{\nabla}^2 q(\boldsymbol{w})\|_2 \geq \|\nabla\big[\rho^\top \widetilde{\nabla} q(\boldsymbol{w})\big]\|_2. \tag{3.168}$$

Finally, we provide the second main result of this section with a proof in Section 3.G.2.

Proposition 3.6 (Tail bounds for arbitrary functions under LSI)**.** *Given a random vector* $\mathbf{u}$ *satisfying* $\mathrm{LSI}_{\mathbf{u}}(\sigma^2)$*, any function* $q : \mathcal{U}^p \to \mathbb{R}$ *with a pseudo derivative* $\widetilde{\nabla} q$ *and pseudo Hessian* $\widetilde{\nabla}^2 q$ *(see Definition 3.5),* $\mathbf{u}$ *satisfies a tail bound, namely for any fixed* $\varepsilon > 0$*, we have*

$$\mathbb{P}\Big[\big|q_c(\mathbf{u})\big| \geq \varepsilon\Big] \leq \exp\left(\frac{-c}{\sigma^4} \min\left(\frac{\varepsilon^2}{\mathbb{E}\big[\|\widetilde{\nabla} q(\mathbf{u})\|_2\big]^2 + \max\limits_{\boldsymbol{u} \in \mathcal{U}^p} \|\widetilde{\nabla}^2 q(\boldsymbol{u})\|_{\mathrm{F}}^2}, \frac{\varepsilon}{\max\limits_{\boldsymbol{u} \in \mathcal{U}^p} \|\widetilde{\nabla}^2 q(\boldsymbol{u})\|_{\mathrm{op}}}\right)\right),$$

where $q_c(\mathbf{u}) = q(\mathbf{u}) - \mathbb{E}\big[q(\mathbf{u})\big]$ *and* c *is a universal constant.*

3.G.1 Proof of Proposition 3.5: Logarithmic Sobolev inequality

We start by defining the notion of W_2 distance (Marton, 2015) which is useful in the proof. We note that W_2 distance is a metric on the space of probability measures and satisfies triangle inequality.

Definition 3.6. *(Marton, 2015, W_2 distance) For random vectors $\mathbf{u}$ and $\mathbf{w}$ supported on $\mathcal{U}^p$ with distributions g and f, respectively, the W_2 distance is given by $W_2^2(g_{\mathbf{w}}, f_{\mathbf{u}}) \triangleq \inf_\pi \sum_{t\in[p]} \Big[\mathbb{P}_\pi(w_t \neq u_t)\Big]^2$, where the infimum is taken over all couplings $\pi(\mathbf{u}, \mathbf{w})$ such that $\pi(\mathbf{u}) = f(\mathbf{u})$ and $\pi(\mathbf{w}) = g(\mathbf{w})$.*

Given Definition 3.6, our next lemma states that if appropriate W_2 distances are bounded, then the KL divergence (denoted by $\mathsf{KL}\left(\cdot \,\|\cdot\right)$) and the entropy approximately tensorize. We provide a proof in Section 3.G.1.1.

Lemma 3.16 (Approximate tensorization of KL divergence and entropy)**.** *Given random vectors $\mathbf{u}$ and $\mathbf{w}$ supported on $\mathcal{U}^p$ with distributions g and f, respectively, such that $f_{\min} > 0$ (see Eq. (3.167)), if for all subsets $S \subseteq [p]$ (with $S^C \triangleq [p] \setminus S$) and all $\boldsymbol{w}_{S^C} \in \mathcal{U}^{p-|S|}$,*

$$W_2^2\big(g_{\mathbf{w}_S|\mathbf{w}_{S^C}=\boldsymbol{w}_{S^C}}, f_{\mathbf{u}_S|\mathbf{u}_{S^C}=\boldsymbol{w}_{S^C}}\big) \leq C \sum_{t\in S} \mathbb{E}\Big[\|g_{w_t|\mathbf{w}_{-t}=\boldsymbol{w}_{-t}} - f_{u_t|\mathbf{u}_{-t}=\boldsymbol{w}_{-t}}\|_{\mathsf{TV}}^2 \Big| \mathbf{w}_{S^C} = \boldsymbol{w}_{S^C}\Big], \tag{3.169}$$

almost surely for some constant $C \geq 1$, then

$$\mathsf{KL}\left(g_{\mathbf{w}} \,\|\, f_{\mathbf{u}}\right) \leq \frac{2C}{f_{\min}} \sum_{t\in[p]} \mathbb{E}\Big[\mathsf{KL}\left(g_{w_t|\mathbf{w}_{-t}=\boldsymbol{w}_{-t}} \,\big\|\, f_{u_t|\mathbf{u}_{-t}=\boldsymbol{w}_{-t}}\right)\Big], \quad \textit{and} \tag{3.170}$$

$$\mathsf{Ent}_{\mathbf{u}}\left(q\right) \leq \frac{2C}{f_{\min}} \sum_{t\in[p]} \mathbb{E}_{\mathbf{u}_{-t}}\Big[\mathsf{Ent}_{u_t|\mathbf{u}_{-t}}\left(q\right)\Big] \quad \textit{for any function } q : \mathcal{U}^p \to \mathbb{R}_+. \tag{3.171}$$

Next, we claim that if the random vector $\mathbf{u}$ satisfies Dobrushin's uniqueness condition, then the condition Eq. (3.169) of Lemma 3.16 is naturally satisfied. We provide a proof in Section 3.G.1.2.

Lemma 3.17 (Dobrushin's uniqueness implies approximate tensorization)**.** *Given random vectors $\mathbf{u}$ and $\mathbf{w}$ supported on $\mathcal{U}^p$ with distributions g and f, respectively, if $\mathbf{u}$ satisfies Dobrushin's uniqueness condition (see Definition 3.4) with coupling matrix $\Theta \in \mathbb{R}^{p\times p}$, then for all subsets $S \subseteq [p]$ (with $S^C \triangleq [p] \setminus S$) and all $\boldsymbol{w}_{S^C} \in \mathcal{U}^{p-|S|}$,*

$$W_2^2\big(g_{\mathbf{w}_S|\mathbf{w}_{S^C}=\boldsymbol{w}_{S^C}}, f_{\mathbf{u}_S|\mathbf{u}_{S^C}=\boldsymbol{w}_{S^C}}\big) \leq C \sum_{t\in S} \mathbb{E}\Big[\|g_{w_t|\mathbf{w}_{-t}=\boldsymbol{w}_{-t}} - f_{u_t|\mathbf{u}_{-t}=\boldsymbol{w}_{-t}}\|_{\mathsf{TV}}^2 \Big| \mathbf{w}_{S^C} = \boldsymbol{w}_{S^C}\Big], \tag{3.172}$$

almost surely where $C = \big(1 - \|\Theta\|_{\mathrm{op}}\big)^2$.

Now to prove Proposition 3.5, applying Lemmas 3.16 and 3.17 for an arbitrary function $f : \mathcal{U}^p \to \mathbb{R}$, we find that

$$\begin{aligned}\mathsf{Ent}_{\mathbf{u}}\left(q^2\right) &\leq \frac{2}{f_{\min}\left(1-\|\Theta\|_{\text{op}}\right)^2}\sum_{t\in[p]}\mathbb{E}_{\mathbf{u}_{-t}}\Big[\mathsf{Ent}_{u_t|\mathbf{u}_{-t}}\left(q^2\right)\Big] \\ &\overset{(a)}{\leq} \frac{2\sigma^2}{f_{\min}\left(1-\|\Theta\|_{\text{op}}\right)^2}\sum_{t\in[p]}\mathbb{E}_{\mathbf{u}_{-t}}\Big[\mathbb{E}_{u_t|\mathbf{u}_{-t}}\Big[\big\|\nabla_{u_t}q(u_t;\mathbf{u}_{-t})\big\|_2^2\Big]\Big] \\ &\overset{(b)}{=} \frac{2\sigma^2}{f_{\min}\left(1-\|\Theta\|_{\text{op}}\right)^2}\mathbb{E}_{\mathbf{u}_{-t}}\Big[\mathbb{E}_{u_t|\mathbf{u}_{-t}}\Big[\sum_{t\in[p]}\big\|\nabla_{u_t}q(u_t;\mathbf{u}_{-t})\big\|_2^2\Big]\Big] \\ &\overset{(c)}{=} \frac{2\sigma^2}{f_{\min}\left(1-\|\Theta\|_{\text{op}}\right)^2}\mathbb{E}_{\mathbf{u}}\Big[\big\|\nabla_{\mathbf{u}}q(\mathbf{u})\big\|_2^2\Big],\end{aligned}$$

where (a) follows because $u_t|\mathbf{u}_{-t}$ satisfies $\text{LSI}_{u_t|\mathbf{u}_{-t}=\boldsymbol{u}_{-t}}(\sigma^2)$ for all $t\in[p]$ and $\boldsymbol{u}_{-t}\in\mathcal{U}^{p-1}$, (b) follows by the linearity of expectation and (b) follows by the law of total expectation. The claim follows.

3.G.1.1 Proof of Lemma 3.16: Approximate tensorization of KL divergence and entropy

We start by establishing a reverse-Pinsker style inequality for distributions with compact support to bound their KL divergence by their total variation distance. We provide a proof at the end.

Lemma 3.18 (Reverse-Pinsker inequality)**.** *For any distributions g and f supported on $\mathcal{U}\subset\mathbb{R}$ such that $\min_{u\in\mathcal{U}} f(u) > 0$, we have* $\mathsf{KL}\left(g\,\|f\right) \leq \frac{4}{\min_{u\in\mathcal{U}} f(u)}\|g-f\|_{\mathsf{TV}}^2$.

Given Lemma 3.18, we proceed to prove Lemma 3.16.

3.G.1.1.1 Proof of bound Eq. (3.170) To prove Eq. (3.170), we show that the following inequality holds using the technique of mathematical induction on p:

$$\mathsf{KL}\left(g_{\mathbf{w}}\,\|f_{\mathbf{u}}\right) \leq \frac{4C}{f_{\min}}\sum_{t\in[p]}\mathbb{E}\Big[\|g_{w_t|\mathbf{w}_{-t}=\boldsymbol{w}_{-t}} - f_{u_t|\mathbf{u}_{-t}=\boldsymbol{w}_{-t}}\|_{\mathsf{TV}}^2\Big]. \tag{3.173}$$

Then, Eq. (3.170) follows by using Pinsker's inequality to bound the right hand side of Eq. (3.173).

3.G.1.1.2 Base case: $p = 1$ For the base case, we need to establish that the claim holds for all distributions supported on $\mathcal{U}$ that satisfy the required conditions. In other words, we need to show that

$$\mathsf{KL}\left(g_{w}\,\|f_{u}\right) \leq \frac{4C}{f_{\min}}\|g_{w}-f_{u}\|_{\mathsf{TV}}^2 \quad \text{for every} \quad t\in[p],$$

for all random variables w and u supported on $\mathcal{U}$ such that $f_{\min} = \min_{u\in\mathcal{U}} f_u(u) > 0$. This follows from Lemma 3.18 by observing that $C \geq 1$.

3.G.1.1.3 Inductive step Now, we assume that the claim holds for all distributions supported on $\mathcal{U}^{p-1}$ that satisfy the required conditions, and establish it for distributions supported on $\mathcal{U}^p$. From the chain rule of KL divergence, we have

$$\mathsf{KL}\left(g_{\mathbf{w}} \| f_{\mathbf{u}}\right) = \mathsf{KL}\left(g_{\mathsf{w}_t} \| f_{\mathsf{u}_t}\right) + \mathbb{E}\Big[\mathsf{KL}\left(g_{\mathbf{w}_{-t}|\mathsf{w}_t} \Big\| f_{\mathbf{u}_{-t}|\mathsf{u}_t}\right)\Big] \quad \text{for every} \quad t \in [p].$$

Taking an average over all $t \in [p]$, we have

$$\mathsf{KL}\left(g_{\mathbf{w}} \| f_{\mathbf{u}}\right) = \frac{1}{p}\sum_{t\in[p]} \mathsf{KL}\left(g_{\mathsf{w}_t} \| f_{\mathsf{u}_t}\right) + \frac{1}{p}\sum_{t\in[p]} \mathbb{E}\Big[\mathsf{KL}\left(g_{\mathbf{w}_{-t}|\mathsf{w}_t} \Big\| f_{\mathbf{u}_{-t}|\mathsf{u}_t}\right)\Big]. \tag{3.174}$$

Now, we bound the first term in Eq. (3.174). Let π^* be the coupling between $\mathbf{u}$ and $\mathbf{w}$ that achieves $W_2(g_{\mathbf{w}}, f_{\mathbf{u}})$, i.e.,[7]

$$\pi^* = \operatorname*{arg\,min}_{\pi:\pi(\mathbf{u})=f(\mathbf{u}),\pi(\mathbf{w})=g(\mathbf{w})} \sum_{t\in[p]} \Big[\mathbb{P}_{\pi}(\mathsf{w}_t \neq \mathsf{u}_t)\Big]^2. \tag{3.175}$$

Then, we have

$$\begin{aligned}
\frac{1}{p}\sum_{t\in[p]} \mathsf{KL}\left(g_{\mathsf{w}_t} \| f_{\mathsf{u}_t}\right) &\overset{(a)}{\leq} \frac{1}{p}\sum_{t\in[p]} \frac{4}{f_{\min}} \|g_{\mathsf{w}_t} - f_{\mathsf{u}_t}\|_{\mathsf{TV}}^2 \\
&\overset{(b)}{\leq} \frac{4}{pf_{\min}} \sum_{t\in[p]} \Big[\mathbb{P}_{\pi^*}(\mathsf{w}_t \neq \mathsf{u}_t)\Big]^2 \\
&\overset{(c)}{=} \frac{4}{pf_{\min}} W_2^2(g_{\mathbf{w}}, f_{\mathbf{u}}) \\
&\overset{Eq.\ (3.169)}{\leq} \frac{4C}{pf_{\min}} \sum_{t\in[p]} \mathbb{E}\Big[\|g_{\mathsf{w}_t|\mathbf{w}_{-t}=\boldsymbol{w}_{-t}} - f_{\mathsf{u}_t|\mathbf{u}_{-t}=\boldsymbol{w}_{-t}}\|_{\mathsf{TV}}^2\Big],
\end{aligned} \tag{3.176}$$

where (a) follows from Lemma 3.18 because lower bound on conditional implies lower bound on marginals, i.e., $\min_{t\in[p],u_t\in\mathcal{U}} f_{\mathsf{u}_t}(u_t) = \min_{t\in[p],u_t\in\mathcal{U}} \int_{\boldsymbol{u}_{-t}\in\mathcal{U}^{p-1}} f_{\mathsf{u}_t|\mathbf{u}_{-t}}(u_t|\boldsymbol{u}_{-t}) \cdot f_{\mathbf{u}_{-t}}(\boldsymbol{u}_{-t})d\boldsymbol{u}_{-t} > f_{\min}$, (b) follows from the connections of total variation distance to optimal transportation cost, i.e., $\|g_{\mathsf{w}} - f_{\mathsf{u}}\|_{\mathsf{TV}} = \inf_{\pi:\pi(\mathsf{u})=f(\mathsf{u}),\pi(\mathsf{w})=g(\mathsf{w})} \mathbb{P}_{\pi}(\mathsf{u} \neq \mathsf{w})$, and (c) follows from Definition 3.6 and Eq. (3.175).

Next, we bound the second term in Eq. (3.174). We have

$$\begin{aligned}
&\frac{1}{p}\sum_{t\in[p]} \mathbb{E}\Big[\mathsf{KL}\left(g_{\mathbf{w}_{-t}|\mathsf{w}_t} \Big\| f_{\mathbf{u}_{-t}|\mathsf{u}_t}\right)\Big] \\
&\overset{(a)}{\leq} \frac{1}{p}\sum_{t\in[p]} \mathbb{E}\bigg[\frac{4C}{f_{\min}} \sum_{s\in[p]\setminus\{t\}} \mathbb{E}\Big[\|g_{\mathsf{w}_s|\mathbf{w}_{-s}=\boldsymbol{w}_{-s}} - f_{\mathsf{u}_s|\mathbf{u}_{-s}=\boldsymbol{w}_{-s}}\|_{\mathsf{TV}}^2 \Big| \mathsf{w}_t = y_t\Big]\bigg]
\end{aligned}$$

[7] The minimum is achieved by using arguments similar to the ones used to show that the Wasserstein distance attains its minimum (Villani, 2009, Chapter 4).

$$\begin{aligned}&\overset{(b)}{=} \frac{4C}{p f_{\min}} \sum_{t\in[p]} \sum_{s\in[p]\setminus\{t\}} \mathbb{E}\Big[\|g_{w_s|\mathbf{w}_{-s}=\boldsymbol{w}_{-s}} - f_{u_s|\mathbf{u}_{-s}=\boldsymbol{w}_{-s}}\|_{\mathsf{TV}}^2\Big] \\ &= \frac{4C(p-1)}{p f_{\min}} \sum_{s\in[p]} \mathbb{E}\Big[\|g_{w_s|\mathbf{w}_{-s}=\boldsymbol{w}_{-s}} - f_{u_s|\mathbf{u}_{-s}=\boldsymbol{w}_{-s}}\|_{\mathsf{TV}}^2\Big], \qquad (3.177)\end{aligned}$$

where (a) follows from the inductive hypothesis and (b) follows from the law of total expectation. Then, Eq. (3.173) follows by putting Eqs. (3.174), (3.176), and (3.177) together.

3.G.1.1.4 Proof of bound Eq. (3.171)

To prove Eq. (3.171), we note that Eq. (3.170) holds for any random vector $\mathbf{w}$ supported on $\mathcal{U}^p$. Consider $\mathbf{w}$ to be such that $q(\mathbf{u})/\mathbb{E}_{\mathbf{u}}[q(\mathbf{u})]$ is the Radon-Nikodym derivative of $g_{\mathbf{w}}$ with respect to $f_{\mathbf{u}}$. For any $\mathcal{A}^p \subseteq \mathcal{U}^p$, we have

$$\int_{\mathbf{w}\in\mathcal{A}^p} g_{\mathbf{w}} d\mathbf{w} = \int_{\mathbf{u}\in\mathcal{A}^p} \frac{q(\mathbf{u})}{\mathbb{E}_{\mathbf{u}}[q(\mathbf{u})]} f_{\mathbf{u}} d\mathbf{u}.$$

Integrating out w_t and u_t for $t \in [p]$, we have

$$\int_{\mathbf{w}_{-t}\in\mathcal{A}^{p-1}} g_{\mathbf{w}_{-t}} d\mathbf{w}_{-t} = \int_{\mathbf{u}_{-t}\in\mathcal{A}^{p-1}} \frac{\mathbb{E}_{u_t|\mathbf{u}_{-t}}\big[q(\mathbf{u})\big]}{\mathbb{E}_{\mathbf{u}}\big[q(\mathbf{u})\big]} f_{\mathbf{u}_{-t}} d\mathbf{u}_{-t},$$

implying

$$\frac{dg_{\mathbf{w}_{-t}}}{df_{\mathbf{u}_{-t}}} = \frac{\mathbb{E}_{u_t|\mathbf{u}_{-t}}\big[q(\mathbf{u})\big]}{\mathbb{E}_{\mathbf{u}}\big[q(\mathbf{u})\big]} \quad \text{and} \quad \frac{dg_{w_t|\mathbf{w}_{-t}}}{df_{u_t|\mathbf{u}_{-t}}} = \frac{q(\mathbf{u})}{\mathbb{E}_{u_t|\mathbf{u}_{-t}}\big[q(\mathbf{u})\big]} \quad \text{for all} \quad t\in[p]. \qquad (3.178)$$

We have

$$\begin{aligned}\mathsf{KL}\left(g_{\mathbf{w}} \,\|\, f_{\mathbf{u}}\right) &\overset{(a)}{=} \mathbb{E}_{\mathbf{u}}\left[\frac{dg_{\mathbf{w}}}{df_{\mathbf{u}}} \log \frac{dg_{\mathbf{w}}}{df_{\mathbf{u}}}\right] \\ &\overset{(b)}{=} \mathbb{E}_{\mathbf{u}}\left[\frac{q(\mathbf{u})}{\mathbb{E}_{\mathbf{u}}\big[q(\mathbf{u})\big]} \log \frac{q(\mathbf{u})}{\mathbb{E}_{\mathbf{u}}\big[q(\mathbf{u})\big]}\right] \\ &= \frac{1}{\mathbb{E}_{\mathbf{u}}\big[q(\mathbf{u})\big]}\Big(\mathbb{E}_{\mathbf{u}}\big[q(\mathbf{u})\log q(\mathbf{u})\big] - \mathbb{E}_{\mathbf{u}}\big[q(\mathbf{u})\big]\log \mathbb{E}_{\mathbf{u}}\big[q(\mathbf{u})\big]\Big) = \frac{\mathsf{Ent}_{\mathbf{u}}\left(q\right)}{\mathbb{E}_{\mathbf{u}}\big[q(\mathbf{u})\big]}, \\ &\qquad (3.179)\end{aligned}$$

where (a) follows from the definition of KL divergence and (b) follows from the choice of $\mathbf{w}$. Similarly, for every $t \in [p]$, we have

$$\begin{aligned}&\mathbb{E}_{\mathbf{w}_{-t}}\Big[\mathsf{KL}\left(g_{w_t|\mathbf{w}_{-t}=\boldsymbol{w}_{-t}} \,\|\, f_{u_t|\mathbf{u}_{-t}=\boldsymbol{w}_{-t}}\right)\Big] \\ &\overset{(a)}{=} \mathbb{E}_{\mathbf{w}_{-t}}\left[\mathbb{E}_{w_t|\mathbf{w}_{-t}}\left[\log \frac{dg_{w_t|\mathbf{w}_{-t}}}{df_{u_t|\mathbf{u}_{-t}}}\right]\right]\end{aligned}$$

$$
\begin{aligned}
&\overset{(b)}{=} \mathbb{E}_{\mathbf{w}}\left[\log \frac{dg_{w_t|\mathbf{w}_{-t}}}{df_{u_t|\mathbf{u}_{-t}}}\right]\\
&\overset{(c)}{=} \mathbb{E}_{\mathbf{u}}\left[\frac{dg_{\mathbf{w}}}{df_{\mathbf{u}}}\log \frac{dg_{w_t|\mathbf{w}_{-t}}}{df_{u_t|\mathbf{u}_{-t}}}\right]\\
&\overset{(d)}{=} \mathbb{E}_{\mathbf{u}}\left[\frac{q(\mathbf{u})}{\mathbb{E}_{\mathbf{u}}\big[q(\mathbf{u})\big]}\log \frac{q(\mathbf{u})}{\mathbb{E}_{u_t|\mathbf{u}_{-t}}\big[q(\mathbf{u})\big]}\right]\\
&\overset{(e)}{=} \frac{\mathbb{E}_{\mathbf{u}_{-t}}\Big[\mathbb{E}_{u_t|\mathbf{u}_{-t}}\big[q(\mathbf{u})\log q(\mathbf{u})\big] - \mathbb{E}_{u_t|\mathbf{u}_{-t}}\big[q(\mathbf{u})\log \mathbb{E}_{u_t|\mathbf{u}_{-t}}\big[q(\mathbf{u})\big]\big]\Big]}{\mathbb{E}_{\mathbf{u}}\big[q(\mathbf{u})\big]}\\
&\overset{(f)}{=} \frac{\mathbb{E}_{\mathbf{u}_{-t}}\big[\mathsf{Ent}_{u_t|\mathbf{u}_{-t}}(q)\,\big]}{\mathbb{E}\big[q(\mathbf{u})\big]},
\end{aligned}
\tag{3.180}
$$

where (a) follows from the definition of KL divergence, (b) follows from the law of total expectation, (c) follows from the definition of Radon-Nikodym derivative, (d) follows from the choice of $\mathbf{w}$ and Eq. (3.178), (e) follows from the law of total expectation, (f) follows from the definition of entropy. Then, Eq. (3.171) follows by putting Eqs. (3.170), (3.179), and (3.180) together.

3.G.1.1.5 Proof of Lemma 3.18: Reverse-Pinsker inequality

Using the facts (a) $\log a \geq 1 - \frac{1}{a}$ for all $a > 0$, and (b) $\min_{u\in\mathcal{U}} f(u) > 0$, we find that

$$
\log\frac{f(u)}{g(u)} \geq 1 - \frac{g(u)}{f(u)} \quad \text{for every} \quad u \in \mathcal{U}. \tag{3.181}
$$

Multiplying both sides of Eq. (3.181) by $g(u) \geq 0$ and rearranging terms yields that

$$
g(u)\log\frac{g(u)}{f(u)} \leq \frac{g^2(u)}{f(u)} - g(u) \quad \text{for every} \quad u \in \mathcal{U}. \tag{3.182}
$$

Now, we have

$$
\begin{aligned}
\mathsf{KL}\left(g\,\|f\right) = \int_{u\in\mathcal{U}} g(u)\log\frac{g(u)}{f(u)}dx &\overset{Eq.\ (3.182)}{\leq} \int_{u\in\mathcal{U}}\left(\frac{g^2(u)}{f(u)} - g(u)\right)dx\\
&\overset{(a)}{=} \int_{u\in\mathcal{U}} \frac{\big(g(u)-f(u)\big)^2}{f(u)}dx\\
&\leq \frac{1}{\min_{u\in\mathcal{U}} f(u)}\int_{u\in\mathcal{U}}\big(g(u)-f(u)\big)^2dx\\
&\overset{(b)}{\leq} \frac{1}{\min_{u\in\mathcal{U}} f(u)}\left(\int_{u\in\mathcal{U}}\big|g(u)-f(u)\big|dx\right)^2\\
&\overset{(c)}{=} \frac{1}{\min_{u\in\mathcal{U}} f(u)}\Big(2\|g-f\|_{\mathsf{TV}}\Big)^2\\
&= \frac{4}{\min_{u\in\mathcal{U}} f(u)}\|g-f\|_{\mathsf{TV}}^2,
\end{aligned}
$$

where (a) follows by simple manipulations, (b) follows by using the order of norms on Euclidean space, and (c) follows by the definition of the total variation distance.

3.G.1.2 Proof of Lemma 3.17: Dobrushin's uniqueness implies approximate tensorization

We start by defining the notion of Gibbs sampler which is useful in the proof.

Definition 3.7. *(Marton, 2015, Gibbs Sampler) For a random vector $\mathbf{u}$ with distribution f, define the Markov kernels and the Gibbs sampler as follows*

$$\Gamma_t(\boldsymbol{u}|\boldsymbol{u}') \triangleq \mathbb{1}(\boldsymbol{u}_{-t} = \boldsymbol{u}'_{-t}) f_{\mathrm{u}_t|\mathbf{u}_{-t}}(u_t|\boldsymbol{u}'_{-t}) \quad \textit{and} \quad \Gamma(\boldsymbol{u}|\boldsymbol{u}') \triangleq p^{-1} \sum_{t\in[p]} \Gamma_t(\boldsymbol{u}|\boldsymbol{u}'), \tag{3.183}$$

for all $t \in [p]$ and $\boldsymbol{u}, \boldsymbol{u}' \in \mathcal{U}^p$. That is, the kernel Γ_t leaves all but the t^{th} coordinate unchanged, and updates the t^{th} coordinate according to $f_{\mathrm{u}_t|\mathbf{u}_{-t}}$, and the sampler Γ selects an index $t \in [p]$ at random, and applies Γ_t. Further, for a random vector $\mathbf{w}$ with distribution g supported on $\mathcal{U}^p$, we also define

$$\begin{aligned} g_{\mathbf{w}}\Gamma_t(\boldsymbol{w}) &\triangleq \int g_{\mathbf{w}}(\boldsymbol{w}')\Gamma_t(\boldsymbol{w}|\boldsymbol{w}')d\boldsymbol{w}' \ \textit{for } t \in [p], \ \textit{and} \\ g_{\mathbf{w}}\Gamma(\boldsymbol{w}) &\triangleq \int g_{\mathbf{w}}(\boldsymbol{w}')\Gamma(\boldsymbol{w}|\boldsymbol{w}')d\boldsymbol{w}' \quad \textit{for all} \quad \boldsymbol{w} \in \mathcal{U}^p. \end{aligned} \tag{3.184}$$

We now proceed to prove Lemma 3.17 and split it in two cases: (i) $S = [p]$, and (ii) $S \subset [p]$.

3.G.1.2.1 Case (i) ($S = [p]$) Let Γ be the Gibbs sampler associated with the distribution f. Then,

$$W_2\big(g_{\mathbf{w}_S|\mathbf{w}_{S^C}}, f_{\mathbf{u}_S|\mathbf{u}_{S^C}}\big) = W_2(g_{\mathbf{w}}, f_{\mathbf{u}}) \overset{(a)}{\leq} W_2(g_{\mathbf{w}}, g_{\mathbf{w}}\Gamma) + W_2(g_{\mathbf{w}}\Gamma, f_{\mathbf{u}}), \tag{3.185}$$

where (a) follows from the triangle inequality. We claim that

$$W_2(g_{\mathbf{w}}, g_{\mathbf{w}}\Gamma) \leq \frac{1}{p}\sqrt{\sum_{t\in[p]} \mathbb{E}_{\mathbf{w}_{-t}}\Big[\|g_{\mathrm{w}_t|\mathbf{w}_{-t}=\boldsymbol{w}_{-t}} - f_{\mathrm{u}_t|\mathbf{u}_{-t}=\boldsymbol{w}_{-t}}\|_{\mathsf{TV}}^2\Big]}, \quad \text{and} \tag{3.186}$$

$$W_2(g_{\mathbf{w}}\Gamma, f_{\mathbf{u}}) \leq \bigg(1 - \frac{(1 - \|\Theta\|_{\mathrm{op}})}{p}\bigg) W_2(g_{\mathbf{w}}, f_{\mathbf{u}}). \tag{3.187}$$

Putting Eqs. (3.185) to (3.187) together, we have

$$\begin{aligned} W_2(g_{\mathbf{w}}, f_{\mathbf{u}}) \leq \frac{1}{p}\sqrt{\sum_{t\in[p]} \mathbb{E}_{\mathbf{w}_{-t}}\Big[\|g_{\mathrm{w}_t|\mathbf{w}_{-t}=\boldsymbol{w}_{-t}} - f_{\mathrm{u}_t|\mathbf{u}_{-t}=\boldsymbol{w}_{-t}}\|_{\mathsf{TV}}^2\Big]} \\ + \bigg(1 - \frac{(1 - \|\Theta\|_{\mathrm{op}})}{p}\bigg) W_2(g_{\mathbf{w}}, f_{\mathbf{u}}). \end{aligned} \tag{3.188}$$

Rearranging Eq. (3.188) results in Eq. (3.172) for $S = [p]$ as desired. It remains to prove our earlier claims Eqs. (3.186) and (3.187) which we now do one-by-one.

3.G.1.2.2 Proof of bound Eq. (3.186) on $W_2(g_{\mathbf{w}}, g_{\mathbf{w}}\Gamma)$ To bound $W_2(g_{\mathbf{w}}, g_{\mathbf{w}}\Gamma)$, we construct a random vector $\mathbf{w}^{\Gamma}$ such that it is coupled with the random vector $\mathbf{w}$. We select an index $b \in [p]$ at random, and define

$$y_v^{\Gamma} \triangleq y_v \quad \text{for all} \quad v \in [p] \setminus \{b\}.$$

Then, given b and $\mathbf{w}_{-b} = \boldsymbol{w}_{-b}$, we define the joint distribution of $(\mathsf{w}_b, \mathsf{w}_b^{\Gamma})$ to be the maximal coupling of $g_{\mathsf{w}_b|\mathbf{w}_{-b}=\boldsymbol{w}_{-b}}$ and $f_{\mathsf{u}_b|\mathbf{u}_{-b}=\boldsymbol{w}_{-b}}$ that achieves $\|g_{\mathsf{w}_b|\mathbf{w}_{-b}=\boldsymbol{w}_{-b}} - f_{\mathsf{u}_b|\mathbf{u}_{-b}=\boldsymbol{w}_{-b}}\|_{\mathsf{TV}}$. It is easy to see that the marginal distribution of $\mathbf{w}$ is $g_{\mathbf{w}}$ and the marginal distribution of $\mathbf{w}^{\Gamma}$ is $g_{\mathbf{w}}\Gamma$ (see Definition 3.7). Then, we have

$$\begin{aligned}
W_2^2(g_{\mathbf{w}}, g_{\mathbf{w}}\Gamma) &\overset{(a)}{\leq} \sum_{t\in[p]} \Big[\mathbb{P}(b=t)\mathbb{P}(\mathsf{w}_t \neq \mathsf{w}_t^{\Gamma}|b=t) + \mathbb{P}(b\neq t)\mathbb{P}(\mathsf{w}_t \neq \mathsf{w}_t^{\Gamma}|b\neq t)\Big]^2 \\
&\overset{(b)}{=} \sum_{t\in[p]} \Big[\frac{1}{p}\mathbb{P}(\mathsf{w}_t \neq \mathsf{w}_t^{\Gamma}|b=t)\Big]^2 \\
&\overset{(c)}{=} \frac{1}{p^2}\sum_{t\in[p]} \Big[\int_{\boldsymbol{w}_{-t}\in\mathcal{U}^{p-1}} \mathbb{P}(\mathsf{w}_t \neq \mathsf{w}_t^{\Gamma}|b=t, \mathbf{w}_{-t}=\boldsymbol{w}_{-t}) g_{\mathbf{w}_{-t}|b=t}(\boldsymbol{w}_{-t}|b=t) d\boldsymbol{w}_{-t}\Big]^2 \\
&\overset{(d)}{=} \frac{1}{p^2}\sum_{t\in[p]} \Big[\int_{\boldsymbol{w}_{-t}\in\mathcal{U}^{p-1}} \|g_{\mathsf{w}_t|\mathbf{w}_{-t}=\boldsymbol{w}_{-t}} - f_{\mathsf{u}_t|\mathbf{u}_{-t}=\boldsymbol{w}_{-t}}\|_{\mathsf{TV}} g_{\mathbf{w}_{-t}}(\boldsymbol{w}_{-t}) d\boldsymbol{w}_{-t}\Big]^2 \\
&= \frac{1}{p^2}\sum_{t\in[p]} \Big[\mathbb{E}_{\mathbf{w}_{-t}}\Big[\|g_{\mathsf{w}_t|\mathbf{w}_{-t}=\boldsymbol{w}_{-t}} - f_{\mathsf{u}_t|\mathbf{u}_{-t}=\boldsymbol{w}_{-t}}\|_{\mathsf{TV}}\Big]\Big]^2,
\end{aligned} \tag{3.189}$$

where (a) follows from Definition 3.6 and the Bayes rule, (b) follows because $\mathbb{P}(b=t) = \frac{1}{p}$ and $\mathbb{P}(\mathsf{w}_t \neq \mathsf{w}_t^{\Gamma}|b\neq t) = 0$, (c) follows by the law of total probability, and (d) follows because $g_{\mathbf{w}_{-t}|b=t}(\boldsymbol{w}_{-t}|b=t) = g_{\mathbf{w}_{-t}}(\boldsymbol{w}_{-t})$ and by the construction of the coupling between $\mathbf{w}$ and $\mathbf{w}^{\Gamma}$. Then, Eq. (3.186) follows by using Jensen's inequality in Eq. (3.189).

3.G.1.2.3 Proof of bound Eq. (3.187) on $W_2(g_{\mathbf{w}}\Gamma, f_{\mathbf{u}})$ We first show that $f_{\mathbf{u}}$ is an invariant measure for Γ, i.e., $f_{\mathbf{u}} = f_{\mathbf{u}}\Gamma$, implying $W_2(g_{\mathbf{w}}\Gamma, f_{\mathbf{u}}) = W_2(g_{\mathbf{w}}\Gamma, f_{\mathbf{u}}\Gamma)$, and then Γ is a contraction with respect to the W_2 distance with rate $1 - \frac{(1-\|\Theta\|_{\mathrm{op}})}{p}$, i.e., $W_2(g_{\mathbf{w}}\Gamma, f_{\mathbf{u}}\Gamma) \leq \Big(1 - \frac{(1-\|\Theta\|_{\mathrm{op}})}{p}\Big) W_2(g_{\mathbf{w}}, f_{\mathbf{u}})$, implying Eq. (3.187).

3.G.1.2.4 Proof of $f_{\mathbf{u}}$ being an invariant measure for Γ We have

$$\begin{aligned}
f_{\mathbf{u}}\Gamma(\boldsymbol{u}) &\overset{Eq.\ (3.184)}{=} \int_{\boldsymbol{u}'\in\mathcal{U}^p} f_{\mathbf{u}}(\boldsymbol{u}')\Gamma(\boldsymbol{u}|\boldsymbol{u}')d\boldsymbol{u}' \\
&\overset{Eq.\ (3.183)}{=} \int_{\boldsymbol{u}'\in\mathcal{U}^p} f_{\mathbf{u}}(\boldsymbol{u}')\Big(\frac{1}{p}\sum_{t\in[p]}\Gamma_t(\boldsymbol{u}|\boldsymbol{u}')\Big)d\boldsymbol{u}' \\
&\overset{Eq.\ (3.183)}{=} \frac{1}{p}\sum_{t\in[p]}\int_{\boldsymbol{u}'\in\mathcal{U}^p} f_{\mathbf{u}}(\boldsymbol{u}')\mathbb{1}(\boldsymbol{u}_{-t} = \boldsymbol{u}'_{-t}) f_{\mathsf{u}_t|\mathbf{u}_{-t}}(u_t|\boldsymbol{u}'_{-t})d\boldsymbol{u}'
\end{aligned}$$

$$= \frac{1}{p} \sum_{t \in [p]} f_{\mathsf{u}_t|\mathbf{u}_{-t}}(u_t|\boldsymbol{u}_{-t}) \int_{u'_t \in \mathcal{U}} f_{\mathbf{u}}(\boldsymbol{u}_{-t}, u'_t) du'_t$$

$$= \frac{1}{p} \sum_{t \in [p]} f_{\mathsf{u}_t|\mathbf{u}_{-t}}(u_t|\boldsymbol{u}_{-t}) f_{\mathbf{u}_{-t}}(\boldsymbol{u}_{-t}) = f_{\mathbf{u}}(\boldsymbol{u}).$$

3.G.1.2.5 Proof of Γ being a contraction w.r.t the W_2 distance

Let π^* be the coupling between $\mathbf{u}$ and $\mathbf{w}$ that achieves $W_2(g_{\mathbf{w}}, f_{\mathbf{u}})$ i.e.,[8]

$$\pi^* = \operatorname*{arg\,min}_{\pi : \pi(\mathbf{u}) = f(\mathbf{u}), \pi(\mathbf{w}) = g(\mathbf{w})} \sqrt{\sum_{t \in [p]} \Big[\mathbb{P}_\pi(\mathsf{w}_t \neq \mathsf{u}_t) \Big]^2}. \tag{3.190}$$

We construct random variables $\mathbf{u}'$ and $\mathbf{w}'$ as well as a coupling π' between them such that the marginal distribution of $\mathbf{u}'$ is $f_{\mathbf{u}}\Gamma$ and the marginal distribution of $\mathbf{w}'$ is $g_{\mathbf{w}}\Gamma$. We start by selecting an index $b \in [p]$ at random, and defining

$$w'_v \triangleq w_v \quad \text{and} \quad u'_v \triangleq u_v \quad \text{for all} \quad v \neq b. \tag{3.191}$$

Then, given b, $\mathbf{w}'_{-b} = \boldsymbol{w}_{-b}$, and $\mathbf{u}'_{-b} = \boldsymbol{u}_{-b}$, we define the joint distribution of $(\mathsf{w}'_b, \mathsf{u}'_b)$ to be the maximal coupling of $f_{\mathsf{u}_b|\mathbf{u}_{-b}}(\cdot|\boldsymbol{w}_{-b})$ and $f_{\mathsf{u}_b|\mathbf{u}_{-b}}(\cdot|\boldsymbol{u}_{-b})$ that achieves $\|f_{\mathsf{u}_b|\mathbf{u}_{-b}=\boldsymbol{w}_{-b}} - f_{\mathsf{u}_b|\mathbf{u}_{-b}=\boldsymbol{u}_{-b}}\|_{\mathsf{TV}}$.

Now, for every $t \in [p]$, we bound $\mathbb{P}_{\pi'}(\mathsf{w}'_t \neq \mathsf{u}'_t)$ in terms of $\mathbb{P}_{\pi^*}(\mathsf{w}_t \neq \mathsf{u}_t)$. To that end, we have

$$\begin{aligned} \mathbb{P}_{\pi'}(\mathsf{w}'_t \neq \mathsf{u}'_t) &\overset{(a)}{=} \mathbb{P}(b = t)\mathbb{P}_{\pi'}(\mathsf{w}'_t \neq \mathsf{u}'_t|b = t) + \mathbb{P}(b \neq t)\mathbb{P}_{\pi'}(\mathsf{w}'_t \neq \mathsf{u}'_t|b \neq t) \\ &\overset{(b)}{=} \frac{1}{p}\mathbb{P}_{\pi'}(\mathsf{w}'_t \neq \mathsf{u}'_t|b = t) + \Big(1 - \frac{1}{p}\Big)\mathbb{P}_{\pi^*}(\mathsf{w}_t \neq \mathsf{u}_t), \end{aligned} \tag{3.192}$$

where (a) follows from the Bayes rule and (b) follows because $\mathbb{P}(b = t) = \frac{1}{p}$ and Eq. (3.191). Focusing on $\mathbb{P}_{\pi'}(\mathsf{w}'_t \neq \mathsf{u}'_t|b = t)$ and using the law of total probability, we have

$$\begin{aligned} &\mathbb{P}_{\pi'}(\mathsf{w}'_t \neq \mathsf{u}'_t|b = t) \\ &= \int\limits_{\boldsymbol{w}_{-t}, \boldsymbol{u}_{-t} \in \mathcal{U}^{p-1}} \mathbb{P}_{\pi'}(\mathsf{w}'_t \neq \mathsf{u}'_t|b{=}t, \mathbf{w}'_{-t}{=}\boldsymbol{w}_{-t}, \mathbf{u}'_{-t}{=}\boldsymbol{u}_{-t})\pi'_{\mathbf{w}'_{-t},\mathbf{u}'_{-t}|b=t}(\boldsymbol{w}_{-t}, \boldsymbol{u}_{-t}|b{=}t)d\boldsymbol{w}_{-t}d\boldsymbol{u}_{-t} \\ &\overset{(a)}{=} \int\limits_{\boldsymbol{w}_{-t}, \boldsymbol{u}_{-t} \in \mathcal{U}^{p-1}} \|f_{\mathsf{u}_t|\mathbf{u}_{-t}=\boldsymbol{w}_{-t}} - f_{\mathsf{u}_t|\mathbf{u}_{-t}=\boldsymbol{u}_{-t}}\|_{\mathsf{TV}} \pi^*_{\mathbf{w}_{-t},\mathbf{u}_{-t}}(\boldsymbol{w}_{-t}, \boldsymbol{u}_{-t})d\boldsymbol{w}_{-t}d\boldsymbol{u}_{-t} \\ &= \mathbb{E}_{\pi^*_{\mathbf{w}_{-t},\mathbf{u}_{-t}}}\Big[\|f_{\mathsf{u}_t|\mathbf{u}_{-t}=\boldsymbol{w}_{-t}} - f_{\mathsf{u}_t|\mathbf{u}_{-t}=\boldsymbol{u}_{-t}}\|_{\mathsf{TV}}\Big] \end{aligned} \tag{3.193}$$

[8] The minimum is achieved by using arguments similar to the ones used to show that the Wasserstein distance attains its minimum (Villani, 2009, Chapter 4).

where (a) follows by the construction of the coupling between $\mathbf{w}'$ and $\mathbf{u}'$. Now, using the triangle inequality in Eq. (3.193), we have

$$\mathbb{P}_{\pi'}(w_t' \neq u_t' | b = t) \leq \mathbb{E}_{\pi^*_{\mathbf{w}_{-t}, \mathbf{u}_{-t}}} \Big[\sum_{s \in [p] \setminus \{t\}} \mathbb{1}(r_v = h_v = w_v \forall v < s) \mathbb{1}(r_v = h_v = u_v \forall v > s) \quad \times$$
$$\mathbb{1}(r_s = w_s, h_s = u_s) \| f_{u_t | \mathbf{u}_{-t} = \boldsymbol{r}_{-t}} - f_{u_t | \mathbf{u}_{-t} = \boldsymbol{h}_{-t}} \|_{\mathsf{TV}} \Big]$$
$$\overset{Eq. \ (3.166)}{\leq} \mathbb{E}_{\pi^*_{\mathbf{w}_{-t}, \mathbf{u}_{-t}}} \Big[\sum_{s \in [p] \setminus \{t\}} \Theta_{ts} \mathbb{1}(w_s \neq u_s) \Big] = \sum_{s \in [p] \setminus \{t\}} \Theta_{ts} \mathbb{P}_{\pi^*}(w_s \neq u_s). \tag{3.194}$$

Putting together Eqs. (3.192) and (3.194), we have

$$\mathbb{P}_{\pi'}(w_t' \neq u_t') \leq \frac{1}{p} \sum_{s \in [p] \setminus \{t\}} \Theta_{ts} \mathbb{P}_{\pi^*}(w_s \neq u_s) + \Big(1 - \frac{1}{p}\Big) \mathbb{P}_{\pi^*}(w_t \neq u_t). \tag{3.195}$$

Next, we use Eq. (3.195) to show contraction of Γ. To that end, we define $\mathrm{diag}(\Theta) \in \mathbb{R}^{p \times p}$ to be the matrix with diagonal same as Θ and all non-diagonal entries equal to zeros. Then, we have

$$\begin{aligned} W_2^2(g_{\mathbf{w}}\Gamma, f_{\mathbf{u}}\Gamma) &\overset{(a)}{\leq} \sum_{t \in [p]} \Big[\mathbb{P}_{\pi'}(w_t' \neq u_t') \Big]^2 \\ &\overset{Eq. \ (3.195)}{\leq} \sum_{t \in [p]} \Big[\frac{1}{p} \sum_{s \in [p] \setminus \{t\}} \Theta_{ts} \mathbb{P}_{\pi^*}(w_s \neq u_s) + \Big(1 - \frac{1}{p}\Big) \mathbb{P}_{\pi^*}(w_t \neq u_t) \Big]^2 \\ &\overset{(b)}{\leq} \Big\| \Big(1 - \frac{1}{p}\Big) I + \frac{1}{p} \Big(\Theta - \mathrm{diag}(\Theta) \Big) \Big\|_{\mathrm{op}}^2 \sum_{t \in [p]} \Big[\mathbb{P}_{\pi^*}(w_t \neq u_t) \Big]^2 \\ &\overset{(c)}{=} \Big\| \Big(1 - \frac{1}{p}\Big) I + \frac{1}{p} \Big(\Theta - \mathrm{diag}(\Theta) \Big) \Big\|_{\mathrm{op}}^2 W_2^2(g_{\mathbf{w}}, f_{\mathbf{u}}) \\ &\overset{(d)}{\leq} \Big(\Big(1 - \frac{1}{p}\Big) + \frac{1}{p} \| \Theta - \mathrm{diag}(\Theta) \|_{\mathrm{op}} \Big)^2 W_2^2(g_{\mathbf{w}}, f_{\mathbf{u}}) \\ &\overset{(e)}{\leq} \Big(\Big(1 - \frac{1}{p}\Big) + \frac{1}{p} \| \Theta \|_{\mathrm{op}} \Big)^2 W_2^2(g_{\mathbf{w}}, f_{\mathbf{u}}), \end{aligned} \tag{3.196}$$

where (a) follows from Definition 3.6, (b) follows by some linear algebraic manipulations, (c) follows from Definition 3.6 and Eq. (3.190), (d) follows from the triangle inequality, and (e) follows because $\|\mathbf{M}_1\|_{\mathrm{op}} \leq \|\mathbf{M}_2\|_{\mathrm{op}}$ for any matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ such that $0 \leq \mathbf{M}_1 \leq \mathbf{M}_2$ (component-wise). Then, contraction of Γ follows by taking square root on both sides of Eq. (3.196).

3.G.1.2.6 Case (ii) ($S \subset [p]$) We can directly verify that the matrix $\Theta_S \triangleq \{\Theta_{ts}\}_{t,u \in S}$ is such that $\|\Theta_S\|_{\mathrm{op}} \leq \|\Theta\|_{\mathrm{op}}$ This is true because the operator norm of any sub-matrix is no more than the operator norm of the matrix. Further, we note that for any $\boldsymbol{w}_{S^C} \in \mathcal{U}^{p-|S|}$, the random vector $\mathbf{u}_S|\mathbf{u}_{S^C} = \boldsymbol{w}_{S^C}$ with distribution $f_{\mathbf{u}_S|\mathbf{u}_{S^C}=\boldsymbol{w}_{S^C}}$ satisfies the Dobrushin's uniqueness condition (Definition 3.4) with coupling matrix Θ_S. Then, by performing an analysis similar to the one above, we have

$$W_2\big(g_{\mathbf{w}_S|\mathbf{w}_{S^C}}, f_{\mathbf{u}_S|\mathbf{u}_{S^C}}\big) \leq \frac{1}{(1-\|\Theta_S\|_{\mathrm{op}})}\sqrt{\sum_{t\in S}\mathbb{E}\Big[\|g_{w_t|\mathbf{w}_{-t}=\boldsymbol{w}_{-t}} - f_{u_t|\mathbf{u}_{-t}=\boldsymbol{w}_{-t}}\|_{\mathsf{TV}}^2\Big|\mathbf{w}_{S^C} = \boldsymbol{w}_{S^C}\Big]}$$
$$\overset{(a)}{\leq} \frac{1}{(1-\|\Theta\|_{\mathrm{op}})}\sqrt{\sum_{t\in S}\mathbb{E}\Big[\|g_{w_t|\mathbf{w}_{-t}=\boldsymbol{w}_{-t}} - f_{u_t|\mathbf{u}_{-t}=\boldsymbol{w}_{-t}}\|_{\mathsf{TV}}^2\Big|\mathbf{w}_{S^C} = \boldsymbol{w}_{S^C}\Big]},$$

where (a) follows because $\frac{1}{(1-\|\Theta_S\|_{\mathrm{op}})} \leq \frac{1}{(1-\|\Theta\|_{\mathrm{op}})}$. This completes the proof.

3.G.2 Proof of Proposition 3.6: Tail bounds for arbitrary functions under LSI

Fix a function $q : \mathcal{U}^p \to \mathbb{R}$. Fix any pseudo derivative $\widetilde{\nabla} q$ for q and any pseudo Hessian $\widetilde{\nabla}^2 q$ for q. To prove Proposition 3.6, we bound the p-th moment of $q(\mathbf{u}) - \mathbb{E}\big[q(\mathbf{u})\big]$ by certain norms of $\widetilde{\nabla}^2 q$ and $\mathbb{E}_{\mathbf{u}}\big[\widetilde{\nabla} q(\mathbf{u})\big]$. To that end, first, we claim that in order to control the p-th moment of $q(\mathbf{u}) - \mathbb{E}\big[q(\mathbf{u})\big]$, it is sufficient to control the p-th moment of $\big\|\nabla q(\mathbf{u})\big\|_2$. Then, using Eq. (3.168), we note that the p-th moment of $\big\|\nabla q(\mathbf{u})\big\|_2$ is bounded by the p-th moment of $\|\widetilde{\nabla} q(\mathbf{u})\|_2$. Next, we claim that the p-th moment of $\|\widetilde{\nabla} q(\mathbf{u})\|_2$ is bounded by a linear combination of appropriate norms of $\widetilde{\nabla}^2 q$ and $\mathbb{E}_{\mathbf{u}}\big[\widetilde{\nabla} q(\mathbf{u})\big]$. We formalize the claims below and divide the proof across Section 3.G.2.1 and Section 3.G.2.2.

Lemma 3.19 (Bounded p-th moments of $q(\mathbf{u}) - \mathbb{E}\big[q(\mathbf{u})\big]$ and $\|\widetilde{\nabla} q(\mathbf{u})\|_2$)**.** *If a random vector* $\mathbf{u}$ *satisfies* $\mathrm{LSI}_{\mathbf{u}}(\sigma^2)$*, then for any arbitrary function* $q : \mathcal{U}^p \to \mathbb{R}$*,*

$$\big\|q(\mathbf{u}) - \mathbb{E}\big[q(\mathbf{u})\big]\big\|_{L_p} \leq \sigma\sqrt{2p}\big\|\big\|\nabla q(\mathbf{u})\big\|_2\big\|_{L_p} \quad \textit{for any } p \geq 2. \tag{3.197}$$

Further, for any pseudo derivative $\widetilde{\nabla} q(\boldsymbol{u})$ *and any pseudo Hessian* $\widetilde{\nabla}^2 q(\boldsymbol{u})$ *for* q*, and even* $p \geq 2$*,*

$$\|\|\widetilde{\nabla} q(\mathbf{u})\|_2\|_{L_p} \leq 2c\sigma\big(\max_{\boldsymbol{u}\in\mathcal{U}^p}\|\widetilde{\nabla}^2 q(\boldsymbol{u})\|_{\mathrm{F}} + \sqrt{p}\max_{\boldsymbol{u}\in\mathcal{U}^p}\|\widetilde{\nabla}^2 q(\boldsymbol{u})\|_{\mathrm{op}}\big) + 4\|\mathbb{E}_{\mathbf{u}}\big[\widetilde{\nabla} q(\mathbf{u})\big]\|_2, \tag{3.198}$$

where $c \geq 0$ *is a universal constant.*

Given these lemmas, we proceed to prove Proposition 3.6. We let $q_c(\mathbf{u}) = q(\mathbf{u}) - \mathbb{E}\big[q(\mathbf{u})\big]$. Combining Eqs. (3.197) and (3.198) for any even $p \geq 2$, there exists a universal constant c' such that

$$\big\|q_c(\mathbf{u})\big\|_{L_p} \leq c'\sigma^2\Big(\sqrt{p}\max_{\boldsymbol{u}\in\mathcal{U}^p}\|\widetilde{\nabla}^2 q(\boldsymbol{u})\|_{\mathrm{F}} + p\max_{\boldsymbol{u}\in\mathcal{U}^p}\|\widetilde{\nabla}^2 q(\boldsymbol{u})\|_{\mathrm{op}} + \sqrt{p}\|\mathbb{E}_{\mathbf{u}}\big[\widetilde{\nabla} q(\mathbf{u})\big]\|_2\Big). \tag{3.199}$$

Now, we complete the proof by using Eq. (3.199) along with Markov's inequality for a specific choice of p. For any even $p \geq 2$, we have

$$\mathbb{P}\Big[\big|q_c(\mathbf{u})\big| > ec'\sigma^2\Big(\sqrt{p}\max_{\boldsymbol{u}\in\mathcal{U}^p}\|\widetilde{\nabla}^2 q(\boldsymbol{u})\|_{\mathrm{F}} + p\max_{\boldsymbol{u}\in\mathcal{U}^p}\|\widetilde{\nabla}^2 q(\boldsymbol{u})\|_{\mathrm{op}} + \sqrt{p}\|\mathbb{E}_{\mathbf{u}}\big[\widetilde{\nabla} q(\mathbf{u})\big]\|_2\Big)\Big]$$

$$
\begin{aligned}
&= \mathbb{P}\Big[\big|q_c(\mathbf{u})\big|^p > \big(ec'\sigma^2\big)^p\big(\sqrt{p}\max_{\boldsymbol{u}\in\mathcal{U}^p}\|\widetilde{\nabla}^2 q(\boldsymbol{u})\|_{\mathrm{F}} + p\max_{\boldsymbol{u}\in\mathcal{U}^p}\|\widetilde{\nabla}^2 q(\boldsymbol{u})\|_{\mathrm{op}} + \sqrt{p}\|\mathbb{E}_{\mathbf{u}}\big[\widetilde{\nabla} q(\mathbf{u})\big]\|_2\big)^p\Big] \\
&\overset{(a)}{\leq} \frac{\mathbb{E}\big[\big|q_c(\mathbf{u})\big|^p\big]}{\big(ec'\sigma^2\big)^p\big(\sqrt{p}\max_{\boldsymbol{u}\in\mathcal{U}^p}\|\widetilde{\nabla}^2 q(\boldsymbol{u})\|_{\mathrm{F}} + p\max_{\boldsymbol{u}\in\mathcal{U}^p}\|\widetilde{\nabla}^2 q(\boldsymbol{u})\|_{\mathrm{op}} + \sqrt{p}\|\mathbb{E}_{\mathbf{u}}\big[\widetilde{\nabla} q(\mathbf{u})\big]\|_2\big)^p} \\
&\overset{Eq.\ (3.199)}{\leq} e^{-p},
\end{aligned}
$$

where (a) follows from Markov's inequality. The proof is complete by choosing an appropriate universal constant c'', and and performing basic algebraic manipulations after letting

$$
p = \frac{1}{c''\sigma^2}\min\Bigg(\frac{\varepsilon^2}{\mathbb{E}\big[\|\widetilde{\nabla} q(\mathbf{u})\|_2\big]^2 + \max\limits_{\boldsymbol{u}\in\mathcal{U}^p}\|\widetilde{\nabla}^2 q(\boldsymbol{u})\|_{\mathrm{F}}^2}, \frac{\varepsilon}{\max\limits_{\boldsymbol{u}\in\mathcal{U}^p}\|\widetilde{\nabla}^2 q(\boldsymbol{u})\|_{\mathrm{op}}}\Bigg).
$$

We note that an even $p \geq 2$ can be ensured by choosing appropriate c''.

3.G.2.1 Proof of Lemma 3.19 Eq. (3.197): Bounded p-th moment of $q(\mathbf{u}) - \mathbb{E}\big[q(\mathbf{u})\big]$

Fix any $p \geq 2$. We start by using the following result from (Aida and Stroock, 1994, Theorem 3.4) since $\mathbf{u}$ satisfies $\mathrm{LSI}_{\mathbf{u}}(\sigma^2)$:

$$
\big\|q(\mathbf{u}) - \mathbb{E}\big[q(\mathbf{u})\big]\big\|_{L_p}^2 \leq \big\|q(\mathbf{u}) - \mathbb{E}\big[q(\mathbf{u})\big]\big\|_{L_2}^2 + 2\sigma^2(p-2)\big\|\big\|\nabla q(\mathbf{u})\big\|_2\big\|_{L_p}^2. \tag{3.200}
$$

Then, we bound the first term in Eq. (3.200) by using the fact that logarithmic Sobolev inequality implies Poincare inequality with the same constant:

$$
\big\|q(\mathbf{u}) - \mathbb{E}\big[q(\mathbf{u})\big]\big\|_{L_2}^2 = \mathbb{V}\mathrm{ar}(q(\mathbf{u})) \leq \sigma^2\mathbb{E}_{\mathbf{u}}\Big[\big\|\nabla q(\mathbf{u})\big\|_2^2\Big]. \tag{3.201}
$$

Putting together Eqs. (3.200) and (3.201), we have

$$
\begin{aligned}
\big\|q(\mathbf{u}) - \mathbb{E}\big[q(\mathbf{u})\big]\big\|_{L_p}^2 &\leq \sigma^2\mathbb{E}_{\mathbf{u}}\Big[\big\|\nabla q(\mathbf{u})\big\|_2^2\Big] + 2\sigma^2(p-2)\big\|\big\|\nabla q(\mathbf{u})\big\|_2\big\|_{L_p}^2 \\
&\overset{(a)}{\leq} \sigma^2\Big(\mathbb{E}_{\mathbf{u}}\Big[\big\|\nabla q(\mathbf{u})\big\|_2^p\Big]\Big)^{2/p} + 2\sigma^2(p-2)\big\|\big\|\nabla q(\mathbf{u})\big\|_2\big\|_{L_p}^2 \\
&\overset{(b)}{=} \sigma^2\big\|\big\|\nabla q(\mathbf{u})\big\|_2\big\|_{L_p}^2 + 2\sigma^2(p-2)\big\|\big\|\nabla q(\mathbf{u})\big\|_2\big\|_{L_p}^2 \\
&\leq 2\sigma^2 p\big\|\big\|\nabla q(\mathbf{u})\big\|_2\big\|_{L_p}^2,
\end{aligned} \tag{3.202}
$$

where (a) follows by Jensen's inequality and (b) follows by the definition of p-th moment. Taking square root on both sides of Eq. (3.202) completes the proof.

3.G.2.2 Proof of Lemma 3.19 Eq. (3.198): Bounded p-th moment of $\|\widetilde{\nabla} q(\mathbf{u})\|_2$

Fix any even $p \geq 2$. Fix any pseudo derivative $\widetilde{\nabla} q$ and any pseudo Hessian $\widetilde{\nabla}^2 q$. We start by obtaining a convenient bound on $\|\widetilde{\nabla} q(\boldsymbol{u})\|_2$ for every $\boldsymbol{u} \in \mathcal{U}^p$ and then proceed to bound the p-th moment of $\|\widetilde{\nabla} q(\mathbf{u})\|_2$.

Consider a p-dimensional standard normal random vector $\mathbf{g}$ independent of $\mathbf{u}$. For a given $\mathbf{u} = \boldsymbol{u} \in \mathcal{U}^p$, the random variable $\frac{\widetilde{\nabla} q(\boldsymbol{u})^\top \mathbf{g}}{\|\widetilde{\nabla} q(\boldsymbol{u})\|_2}$ is a standard normal random variable. Then, for every $\boldsymbol{u} \in \mathcal{U}^p$, we have

$$\Big\|\frac{\widetilde{\nabla} q(\boldsymbol{u})^\top \mathbf{g}}{\|\widetilde{\nabla} q(\boldsymbol{u})\|_2}\Big\|_{L_p} \overset{(a)}{=} \bigg(\mathbb{E}_{\mathbf{g}|\mathbf{u}=\boldsymbol{u}}\bigg[\bigg(\frac{\widetilde{\nabla} q(\boldsymbol{u})^\top \mathbf{g}}{\|\widetilde{\nabla} q(\boldsymbol{u})\|_2}\bigg)^p\bigg]\bigg)^{1/p} \overset{(b)}{\geq} \frac{\sqrt{p}}{2}, \tag{3.203}$$

where (a) follows from the definition of p-th moment, and (b) follows since $\big\|\boldsymbol{g}\big\|_{L_p} \geq \frac{\sqrt{p}}{2}$ for any standard normal random variable $\boldsymbol{g}$ and even $p \geq 2$. Rearranging Eq. (3.203), we have

$$\|\widetilde{\nabla} q(\boldsymbol{u})\|_2 \leq \frac{2}{\sqrt{p}}\Big(\mathbb{E}_{\mathbf{g}|\mathbf{u}=\boldsymbol{u}}\Big[\big(\widetilde{\nabla} q(\boldsymbol{u})^\top \mathbf{g}\big)^p\Big]\Big)^{1/p}. \tag{3.204}$$

Now, we proceed to bound the p-th moment of $\|\widetilde{\nabla} q(\mathbf{u})\|_2$ as follows

$$\begin{aligned}
\|\|\widetilde{\nabla} q(\mathbf{u})\|_2\|_{L_p} &\overset{(a)}{=} \Big(\mathbb{E}_{\mathbf{u}}\big[\|\widetilde{\nabla} q(\mathbf{u})\|_2^p\big]\Big)^{1/p} \\
&\overset{Eq.\ (3.204)}{\leq} \frac{2}{\sqrt{p}}\Big(\mathbb{E}_{\mathbf{u},\mathbf{g}}\Big[\big(\widetilde{\nabla} q(\mathbf{u})^\top \mathbf{g}\big)^p\Big]\Big)^{1/p} \\
&\overset{(b)}{=} \frac{2}{\sqrt{p}}\big\|\widetilde{\nabla} q(\mathbf{u})^\top \mathbf{g}\big\|_{L_p} \\
&\overset{(c)}{\leq} \frac{2}{\sqrt{p}}\Big(\big\|\widetilde{\nabla} q(\mathbf{u})^\top \mathbf{g} - \mathbb{E}_{\mathbf{u}}\big[\widetilde{\nabla} q(\mathbf{u})^\top \mathbf{g}\big]\big\|_{L_p} + \big\|\mathbb{E}_{\mathbf{u}}\big[\widetilde{\nabla} q(\mathbf{u})^\top \mathbf{g}\big]\big\|_{L_p}\Big),
\end{aligned} \tag{3.205}$$

where (a) and (b) follow from the definition of p-th moment and (c) follows by Minkowski's inequality. We claim that

$$\big\|\widetilde{\nabla} q(\mathbf{u})^\top \mathbf{g} - \mathbb{E}_{\mathbf{u}}\big[\widetilde{\nabla} q(\mathbf{u})^\top \mathbf{g}\big]\big\|_{L_p} \leq c\sigma\Big(\sqrt{p}\max_{\boldsymbol{u}\in\mathcal{U}^p}\|\widetilde{\nabla}^2 q(\boldsymbol{u})\|_{\mathrm{F}} + p\max_{\boldsymbol{u}\in\mathcal{U}^p}\|\widetilde{\nabla}^2 q(\boldsymbol{u})\|_{\mathrm{op}}\Big), \quad \&(3.206)$$

$$\big\|\mathbb{E}_{\mathbf{u}}\big[\widetilde{\nabla} q(\mathbf{u})^\top \mathbf{g}\big]\big\|_{L_p} \leq 2\sqrt{p}\big\|\mathbb{E}_{\mathbf{u}}\big[\widetilde{\nabla} q(\mathbf{u})\big]\big\|_2, \tag{3.207}$$

where $c \geq 0$ is a universal constant. Putting together Eqs. (3.205) to (3.207) completes the proof. It remains to prove our claims Eqs. (3.206) and (3.207) which we now do one-by-one.

3.G.2.2.1 Proof of bound Eq. (3.206) To start, we bound $\big(\mathbb{E}_{\mathbf{u}|\mathbf{g}=\boldsymbol{g}}\big[\big(\widetilde{\nabla} q(\mathbf{u})^\top \boldsymbol{g} - \mathbb{E}_{\mathbf{u}|\mathbf{g}=\boldsymbol{g}}\big[\widetilde{\nabla} q(\mathbf{u})^\top \boldsymbol{g}\big]\big)^p\big]\big)^{1/p}$ for every $\mathbf{g} = \boldsymbol{g}$. Then, we bound $\|\widetilde{\nabla} q(\mathbf{u})^\top \mathbf{g} - \mathbb{E}_{\mathbf{u}}\big[\widetilde{\nabla} q(\mathbf{u})^\top \mathbf{g}\big]\|_{L_p}$.

To that end, we define $h_{\boldsymbol{g}}(\mathbf{u}) \triangleq \widetilde{\nabla} q(\mathbf{u})^\top \boldsymbol{g} - \mathbb{E}_{\mathbf{u}|\mathbf{g}=\boldsymbol{g}}\big[\widetilde{\nabla} q(\mathbf{u})^\top \boldsymbol{g}\big]$ and observe that $\mathbb{E}_{\mathbf{u}|\mathbf{g}=\boldsymbol{g}}\big[h_{\boldsymbol{g}}(\mathbf{u})\big] = 0$. Now, applying Lemma 3.19 Eq. (3.197) to $h_{\boldsymbol{g}}(\cdot)$, we have

$$\big\|h_{\boldsymbol{g}}(\mathbf{u})\big\|_{L_p} \leq \sigma\sqrt{2p}\Big(\mathbb{E}_{\mathbf{u}|\mathbf{g}=\boldsymbol{g}}\Big[\big\|\nabla h_{\boldsymbol{g}}(\mathbf{u})\big\|_2^p\Big]\Big)^{1/p} \overset{(a)}{\leq} \sigma\sqrt{2p}\Big(\mathbb{E}_{\mathbf{u}|\mathbf{g}=\boldsymbol{g}}\Big[\big\|\nabla\big[\boldsymbol{g}^\top \widetilde{\nabla} q(\mathbf{u})\big]\big\|_2^p\Big]\Big)^{1/p}$$

$$\overset{Eq.~(3.168)}{\leq} \sigma\sqrt{2p}\Big(\mathbb{E}_{\mathbf{u}|\mathbf{g}=\boldsymbol{g}}\Big[\big\|\boldsymbol{g}^\top\widetilde{\nabla}^2 q(\mathbf{u})\big\|_2^p\Big]\Big)^{1/p}, \tag{3.208}$$

where (a) follows from the definition of $h_{\boldsymbol{g}}(\mathbf{u})$. Now, to obtain a bound on the RHS of Eq. (3.208), we further fix $\mathbf{u} = \boldsymbol{u}$. Then, we let $\mathbf{g'}$ be another p-dimensional standard normal vector and apply an inequality similar to Eq. (3.204) to $\boldsymbol{g}^\top\widetilde{\nabla}^2 q(\boldsymbol{u})$ obtaining

$$\big\|\boldsymbol{g}^\top\widetilde{\nabla}^2 q(\boldsymbol{u})\big\|_2 \leq \frac{2}{\sqrt{p}}\Big(\mathbb{E}_{\mathbf{g'}|\mathbf{u}=\boldsymbol{u},\mathbf{g}=\boldsymbol{g}}\Big[\Big(\boldsymbol{g}^\top\widetilde{\nabla}^2 q(\boldsymbol{u})\mathbf{g'}\Big)^p\Big]\Big)^{1/p},$$

which implies

$$\Big(\mathbb{E}_{\mathbf{u}|\mathbf{g}=\boldsymbol{g}}\Big[\big\|\boldsymbol{g}^\top\widetilde{\nabla}^2 q(\mathbf{u})\big\|_2^p\Big]\Big)^{1/p} \leq \frac{2}{\sqrt{p}}\Big(\mathbb{E}_{\mathbf{u},\mathbf{g'}|\mathbf{g}=\boldsymbol{g}}\Big[\Big(\nabla\boldsymbol{g}^\top\widetilde{\nabla}^2 q(\mathbf{u})\mathbf{g'}\Big)^p\Big]\Big)^{1/p}. \tag{3.209}$$

Putting together Eqs. (3.208) and (3.209), and using the definition of $h_{\boldsymbol{g}}(\mathbf{u})$, we have

$$\mathbb{E}_{\mathbf{u}|\mathbf{g}=\boldsymbol{g}}\Big[\Big(\widetilde{\nabla} q(\mathbf{u})^\top\boldsymbol{g} - \mathbb{E}_{\mathbf{u}|\mathbf{g}=\boldsymbol{g}}\Big[\widetilde{\nabla} q(\mathbf{u})^\top\boldsymbol{g}\Big]\Big)^p\Big] \leq (2\sqrt{2}\sigma)^p\mathbb{E}_{\mathbf{u},\mathbf{g'}|\mathbf{g}=\boldsymbol{g}}\Big[\Big(\boldsymbol{g}^\top\widetilde{\nabla}^2 q(\mathbf{u})\mathbf{g'}\Big)^p\Big]. \tag{3.210}$$

Now, we proceed to bound $\|\widetilde{\nabla} q(\mathbf{u})^\top\mathbf{g} - \mathbb{E}_{\mathbf{u}}\big[\widetilde{\nabla} q(\mathbf{u})^\top\mathbf{g}\big]\|_{L_p}$ as follows

$$\begin{aligned}\big\|\widetilde{\nabla} q(\mathbf{u})^\top\mathbf{g} - \mathbb{E}_{\mathbf{u}}\Big[\widetilde{\nabla} q(\mathbf{u})^\top\mathbf{g}\Big]\big\|_{L_p} &\overset{(a)}{=} \Big(\mathbb{E}_{\mathbf{u},\mathbf{g}}\Big[\Big(\widetilde{\nabla} q(\mathbf{u})^\top\mathbf{g} - \mathbb{E}_{\mathbf{u}}\Big[\widetilde{\nabla} q(\mathbf{u})^\top\mathbf{g}\Big]\Big)^p\Big]\Big)^{1/p}\\ &\overset{Eq.~(3.210)}{\leq} 2\sqrt{2}\sigma\Big(\mathbb{E}_{\mathbf{g},\mathbf{u},\mathbf{g'}}\Big[\Big(\mathbf{g}^\top\widetilde{\nabla}^2 q(\mathbf{u})\mathbf{g'}\Big)^p\Big]\Big)^{1/p},\end{aligned} \tag{3.211}$$

where (a) follows from the definition of p-th moment. Finally, to bound the RHS of Eq. (3.211), we fix $\mathbf{u} = \boldsymbol{u}$ and bound the p-th norm of the quadratic form $\mathbf{g}^\top\widetilde{\nabla}^2 q(\boldsymbol{u})\mathbf{g'}$ by the Hanson-Wright inequality resulting in

$$\begin{aligned}\Big(\mathbb{E}_{\boldsymbol{g},\mathbf{g'}|\mathbf{u}=\boldsymbol{u}}\Big[\Big(\mathbf{g}^\top\widetilde{\nabla}^2 q(\boldsymbol{u})\mathbf{g'}\Big)^p\Big]\Big)^{1/p} &\leq c\Big(\sqrt{p}\|\widetilde{\nabla}^2 q(\boldsymbol{u})\|_{\mathrm{F}} + p\|\widetilde{\nabla}^2 q(\boldsymbol{u})\|_{\mathrm{op}}\Big)\\ &\leq c\Big(\sqrt{p}\max_{\boldsymbol{u}\in\mathcal{U}^p}\|\widetilde{\nabla}^2 q(\boldsymbol{u})\|_{\mathrm{F}} + p\max_{\boldsymbol{u}\in\mathcal{U}^p}\|\widetilde{\nabla}^2 q(\boldsymbol{u})\|_{\mathrm{op}}\Big),\end{aligned} \tag{3.212}$$

where $c \geq 0$ is a universal constant. Then, Eq. (3.206) follows by putting together Eqs. (3.211) and (3.212).

3.G.2.2.2 Proof of bound Eq. (3.207)

By linearity of expectation, we have

$$\|\mathbb{E}_{\mathbf{u}}\big[\widetilde{\nabla} q(\mathbf{u})^\top\mathbf{g}\big]\|_{L_p} = \|\big(\mathbb{E}_{\mathbf{u}}\big[\widetilde{\nabla} q(\mathbf{u})\big]\big)^\top\mathbf{g}\|_{L_p}. \tag{3.213}$$

We note that the random variable $\dfrac{(\mathbb{E}_{\mathbf{u}}[\widetilde{\nabla} q(\mathbf{u})])^\top\mathbf{g}}{\|\mathbb{E}_{\mathbf{u}}[\widetilde{\nabla} q(\mathbf{u})]\|_2}$ is a standard normal random variable. Therefore,

$$\|\frac{\big(\mathbb{E}_{\mathbf{u}}\big[\widetilde{\nabla} q(\mathbf{u})\big]\big)^\top\mathbf{g}}{\|\mathbb{E}_{\mathbf{u}}\big[\widetilde{\nabla} q(\mathbf{u})\big]\|_2}\|_{L_p} \overset{(a)}{=} \Big(\mathbb{E}_{\mathbf{g}}\Big[\Big(\frac{\big(\mathbb{E}_{\mathbf{u}}\big[\widetilde{\nabla} q(\mathbf{u})\big]\big)^\top\mathbf{g}}{\|\mathbb{E}_{\mathbf{u}}\big[\widetilde{\nabla} q(\mathbf{u})\big]\|_2}\Big)^p\Big]\Big)^{1/p} \overset{(b)}{\leq} 2\sqrt{p}, \tag{3.214}$$

where (a) follows from the definition of p-th moment, and (b) follows since $\left\|\boldsymbol{g}\right\|_{L_p} \leq 2\sqrt{p}$ for any standard normal variable $\boldsymbol{g}$. Then, Eq. (3.207) follows by using Eq. (3.214) in Eq. (3.213).

3.H Identifying weakly dependent random variables

In Section 3.G, we derived (in Proposition 3.5) that a random vector (supported on a compact set) satisfies the logarithmic Sobolev inequality if it satisfies the Dobrushin's uniqueness condition (in Definition 3.4). Further, we also derived (Proposition 3.6) tail bounds for a random vector satisfying the logarithmic Sobolev inequality. Combining the two, we see that in order to use the tail bound, the random vector needs to satisfy the Dobrushin's uniqueness condition, i.e, the elements of the random vector should be weakly dependent. In this section, we show that any random vector (outside Dobrushin's regime) that is a τ-Sparse Graphical Model (to be defined) can be reduced to satisfy the Dobrushin's uniqueness condition. In particular, we show that by conditioning on a subset of the random vector, the unconditioned subset of the random vector (in the conditional distribution) are only weakly dependent. We exploit this trick in Lemma 3.10 and Lemma 3.12 to enable application of the tail bound in Section 3.G. The result below is a generalization of the result in Dagan et al. (2021) for discrete random vectors to continuous random vectors.

We start by defining the notion of τ-Sparse Graphical Model.

Definition 3.8 (τ-Sparse Graphical Model)**.** *A tuple of random vectors* $(\mathbf{y}, \mathbf{a}, \mathbf{v}, \mathbf{z})$ *supported on* $\mathcal{Y}^{p_y} \times \mathcal{A}^{p_a} \times \mathcal{V}^{p_v} \times \mathcal{Z}^{p_z}$ *is a* τ*-Sparse Graphical Model for model-parameters* $\tau \triangleq (\alpha, \xi, \zeta, x_{\max}, \Theta)$ *and denoted by* τ*-SGM if*

1. $\mathcal{U} \subseteq \mathcal{X} \triangleq [-x_{\max}, x_{\max}]$ *for every* $\mathcal{U} \in \{\mathcal{Y}, \mathcal{A}, \mathcal{V}\}$*,*

2. *for any realizations* $\boldsymbol{a} \in \mathcal{A}^{p_a}$*,* $\boldsymbol{v} \in \mathcal{V}^{p_v}$*, and* $\boldsymbol{z} \in \mathcal{Z}^{p_a}$*, the conditional probability distribution of* $\mathbf{y}$ *given* $\mathbf{a} = \boldsymbol{a}$*,* $\mathbf{v} = \boldsymbol{v}$*, and* $\mathbf{z} = \boldsymbol{z}$*, i.e.,* $f_{\mathbf{y}|\mathbf{a},\mathbf{v},\mathbf{z}}\big(\boldsymbol{y}|\boldsymbol{a}, \boldsymbol{v}, \boldsymbol{z}; \theta(\boldsymbol{z}), \Theta\big)$*, is as specified in Eq.* (3.4)*, such that* $\theta(\boldsymbol{z}) \in \mathbb{R}^{p_y}$ *depends on* $\boldsymbol{z}$ *and* $\Theta = [\Phi^{(y,y)}, \Phi^{(y,a)}, \Phi^{(y,v)}] \in \mathbb{R}^{p_y \times \widetilde{p}}$ *is independent of* $\boldsymbol{z}$ *where* $\widetilde{p} = p_y + p_a + p_v$ *and* $\Phi^{(y,y)} \in \mathbb{R}^{p_y \times p_y}$ *is symmetric.*

3. $\max\left\{\max_{\boldsymbol{z} \in \mathcal{Z}^{p_z}} \left\|\theta(\boldsymbol{z})\right\|_{\infty}, \|\Phi^{(y,y)}\|_{\max}\right\} \leq \alpha$*, and*

4. $\max\left\{\|\Phi^{(y,v)}\|_{\infty}, \|\Phi^{(y,a)}\|_{\infty}\right\} \leq \xi$*, and* $\|\Phi^{(y,y)}\|_{\infty} \leq \zeta$*.*

Now, we provide the main result of this section.

Proposition 3.7 (Identifying weakly dependent random variables)**.** *Given a tuple of random vectors* $(\mathbf{y}, \mathbf{a}, \mathbf{v}, \mathbf{z})$ *supported on* $\mathcal{Y}^{p_y} \times \mathcal{A}^{p_a} \times \mathcal{V}^{p_v} \times \mathcal{Z}^{p_z}$ *that is a* τ*-SGM (Definition 3.8) with* $\tau \triangleq (\alpha, \xi, \zeta, x_{\max}, \Theta)$*, and a scalar* $\lambda \in (0, \zeta]$*, there exists* $L \triangleq 32\zeta^2 \log 4p_y/\lambda^2$ *subsets* $S_1, \cdots, S_L \subseteq [p_y]$ *that satisfy the following properties:*

(a) For any $t \in [p_y]$*, we have* $\sum_{u=1}^{L} \mathbb{1}(t \in S_u) = \lceil \lambda L/(8\zeta) \rceil$*.*

(b) For any $u \in [L]$,

(i) the tuple of random vectors $(\mathbf{y}_{S_u}, \mathbf{a}, \mathbf{v}, (\mathbf{z}, \mathbf{y}_{-S_u}))$ correspond to a τ_1-SGM with $\tau_1 \triangleq (\alpha + 2x_{\max}\zeta, \xi, \lambda, x_{\max}, \Theta_{\setminus S_u})$ where $\Theta_{\setminus S_u}$ is obtained from Θ by removing the rows and columns corresponding to $[p_y] \setminus S_u$, and

(ii) the random vector $\mathbf{y}_{S_u}$ conditioned on $(\mathbf{a}, \mathbf{v}, (\mathbf{z}, \mathbf{y}_{-S_u}))$ satisfies the Dobrushin's uniqueness condition (Definition 3.4) with coupling matrix $2\sqrt{2}x_{\max}^2 |\Phi_{S_u}^{(y,y)}|$ whenever $\lambda \in \Big(0, \frac{1}{2\sqrt{2}x_{\max}^2}\Big)$, where $\Phi_{S_u}^{(y,y)} \triangleq \{\Phi_{rt}^{(y,y)}\}_{r,t \in S_u}$ such that $\||\Phi_{S_u}^{(y,y)}\||_{\mathrm{op}} \leq \lambda$.

Proof. Proof of Proposition 3.7: Identifying weakly dependent random variables We prove each part one-by-one using a generalization of Dagan et al. (2021, Lemma. 12).

Recall Dagan et al. (2021, Lemma. 12): Let $A \in \mathbb{R}^{p \times p}$ be a matrix with zeros on the diagonal and $\|A\|_\infty \leq 1$. Let $0 < \eta < 1$. Then, there exists subsets $\overline{S}_1, \cdots, \overline{S}_{\overline{L}} \subseteq [p]$ with $\overline{L} \triangleq 32 \log 4p/\eta^2$ such that

(a) For any $t \in [p]$, we have $\sum_{u=1}^{\overline{L}} \mathbb{1}(t \in \overline{S}_u) = \lceil \eta \overline{L}/8 \rceil$, and

(b) For any $u \in [\overline{L}]$ and $t \in \overline{S}_u$, $\sum_{r \in \overline{S}_u} |A_{tr}| \leq \eta$.

We claim that Dagan et al. (2021, Lemma. 12) holds even when A does not have zeros on the diagonal. The proof is exactly the same as the proof of Dagan et al. (2021, Lemma. 12).

Proof of part (a). From Definition 3.8, for any realizations $\boldsymbol{a} \in \mathcal{A}^{p_a}$, $\boldsymbol{v} \in \mathcal{V}^{p_v}$, and $\boldsymbol{z} \in \mathcal{Z}^{p_a}$, the conditional probability distribution of $\mathbf{y}$, given $\mathbf{a} = \boldsymbol{a}$, $\mathbf{v} = \boldsymbol{v}$, and $\mathbf{z} = \boldsymbol{z}$, is given by Eq. (3.4). Consider the matrix $\Phi^{(y,y)} \in \mathbb{R}^{p_y \times p_y}$ and define $A \triangleq \frac{1}{\zeta}\Phi^{(y,y)}$, Then, note that $\|A\|_\infty \leq 1$ and we can apply the generalization of Dagan et al. (2021, Lemma. 12) on A with $\eta = \frac{\lambda}{\zeta}$. Then part (a) follows directly from Dagan et al. (2021, Lemma. 12.1).

Proof of part (b)(i). To prove this part, consider the distribution of $\mathbf{y}_{S_u}$ conditioned on $\mathbf{y}_{-S_u} = \boldsymbol{y}_{-S_u}$, $\mathbf{a} = \boldsymbol{a}$, $\mathbf{v} = \boldsymbol{v}$, and $\mathbf{z} = \boldsymbol{z}$ for any $u \in [L]$, i.e., consider $f_{\mathbf{y}_{S_u}|\mathbf{y}_{-S_u},\mathbf{a},\mathbf{v},\mathbf{z}}(\boldsymbol{y}_{S_u}|\boldsymbol{y}_{-S_u}, \boldsymbol{a}, \boldsymbol{v}, \boldsymbol{z}) = f(\boldsymbol{y}_{S_u}|\boldsymbol{y}_{-S_u}, \boldsymbol{a}, \boldsymbol{v}, \boldsymbol{z})$ as follows

$$f(\boldsymbol{y}_{S_u}|\boldsymbol{y}_{-S_u}, \boldsymbol{a}, \boldsymbol{v}, \boldsymbol{z}) \propto \exp\bigg([\upsilon(\boldsymbol{z}, \boldsymbol{y}_{-S_u})]^\top \boldsymbol{y}_{S_u} + 2\big[\boldsymbol{v}^\top \Phi_{\cdot,S_u}^{(v,y)} + \boldsymbol{a}^\top \Phi_{\cdot,S_u}^{(a,y)}\big]\boldsymbol{y}_{S_u} + \boldsymbol{y}_{S_u}^\top \Phi_{S_u}^{(y,y)} \boldsymbol{y}_{S_u} \bigg),$$

where $\Phi_{\cdot,S_u}^{(w,y)} \triangleq \{\Phi_{rt}^{(w,y)}\}_{r \in [p_w], t \in S_u}$ for $w \in \{v, a\}$, $\Phi_{S_u}^{(y,y)} \triangleq \{\Phi_{rt}^{(y,y)}\}_{r,t \in S_u}$, and $\upsilon(\boldsymbol{z}, \boldsymbol{y}_{-S_u}) \in \mathbb{R}^{|S_u| \times 1}$ such that

$$\upsilon_t(\boldsymbol{z}, \boldsymbol{y}_{-S_u}) \triangleq \theta_t(\boldsymbol{z}) + 2 \sum_{k \in [p_y] \setminus S_u} \Phi_{tk}^{(y,y)} y_k \text{ for every } t \in S_u. \tag{3.215}$$

Now, to show that the tuple of random vectors $(\mathbf{y}_{S_u}, \mathbf{a}, \mathbf{v}, (\mathbf{z}, \mathbf{y}_{-S_u}))$ corresponds to an τ_1-SGM with $\tau_1 \triangleq (\alpha + 2x_{\max}\zeta, \xi, \lambda, x_{\max}, \Phi_{\cdot,S_u}^{(v,y)}, \Phi_{\cdot,S_u}^{(a,y)}, \Phi_{S_u}^{(y,y)})$, it suffices to establish that

$$\max\bigg\{ \max_{\substack{\boldsymbol{z} \in \mathcal{Z}^{p_z} \\ \boldsymbol{y}_{-S_u} \in \mathbb{R}^{p_y - |S_u|}}} \big\|\upsilon(\boldsymbol{z}, \boldsymbol{y}_{-S_u})\big\|_\infty, \|\Phi_{S_u}^{(y,y)}\|_{\max} \bigg\} \overset{(i)}{\leq} \alpha + 2x_{\max}\zeta \quad \text{and} \quad \|\Phi_{S_u}^{(y,y)}\|_\infty \overset{(ii)}{\leq} \lambda. \tag{3.216}$$

To establish (i) in Eq. (3.216), we note that

$$\|\Phi_{S_u}^{(y,y)}\|_{\max} \le \|\Phi^{(y,y)}\|_{\max} \overset{(a)}{\le} \alpha \quad \text{and} \tag{3.217}$$

$$\begin{aligned}\big\|\upsilon(\boldsymbol{z},\boldsymbol{y}_{-S_u})\big\|_\infty &\overset{(b)}{\le} \big\|\theta(\boldsymbol{z})\big\|_\infty + 2\max_{t\in S_u}\|\Phi_t^{(y,y)}\|_1\big\|\boldsymbol{y}\big\|_\infty \\ &\overset{(c)}{\le} \big\|\theta(\boldsymbol{z})\big\|_\infty + 2x_{\max}\|\Phi^{(y,y)}\|_\infty \overset{(d)}{\le} \alpha + 2x_{\max}\zeta,\end{aligned} \tag{3.218}$$

where (a) and (d) follow from Definition 3.8, (b) follows from Eq. (3.215) and the triangle inequality, and (c) follows from the definition of $\|\cdot\|_\infty$ and Definition 3.8. Then, (i) in Eq. (3.216) follows from Eq. (3.217) and Eq. (3.218). Next, to establish (ii) in Eq. (3.216), we again apply the generalization of Dagan et al. (2021, Lemma. 12) on the matrix $A = \frac{1}{\zeta}\Phi^{(y,y)}$ with $\eta = \frac{\lambda}{\zeta}$. Then, we have

$$\sum_{r\in S_u}\left|\frac{\Phi_{tr}^{(y,y)}}{\zeta}\right| \le \frac{\lambda}{\zeta} \quad \text{for all } t\in S_u,\ u\in[L]. \tag{3.219}$$

Therefore, we have

$$\|\Phi_{S_u}^{(y,y)}\|_\infty = \max_{t\in S_u}\Big(\sum_{r\in S_u}\big|\Phi_{tr}^{(y,y)}\big|\Big) \overset{Eq.\ (3.219)}{\le} \lambda, \tag{3.220}$$

as desired. The proof for this part is now complete.

Proof of part (b)(ii). We start by noting that the operator norm of a symmetric matrix is bounded by the infinity norm of the matrix. Then, from the analysis in part (b) (i), for any $u \in S_u$, we have

$$\||\Phi_{S_u}^{(y,y)}|\|_{\mathrm{op}} \le \||\Phi_{S_u}^{(y,y)}|\|_\infty \overset{Eq.\ (3.220)}{\le} \lambda.$$

Therefore, $\|2\sqrt{2}x_{\max}^2|\Phi_{S_u}^{(y,y)}|\|_{\mathrm{op}} < 1$ whenever $\lambda < 1/2\sqrt{2}x_{\max}^2$. It remains to show that for every $u \in [L]$, $t \in S_u, r \in S_u\backslash\{t\}$, $\mathbf{z} = \boldsymbol{z}$, $\mathbf{v} = \boldsymbol{v}$, $\mathbf{a} = \boldsymbol{a}$, and $\boldsymbol{y}_{-t}, \widetilde{\boldsymbol{y}}_{-t} \in \mathcal{Y}^{p_y-1}$ differing only in the r^{th} coordinate,

$$\|f_{\mathsf{y}_t|\mathbf{y}_{-t}=\boldsymbol{y}_{-t},\mathbf{a}=\boldsymbol{a},\mathbf{v}=\boldsymbol{v},\mathbf{z}=\boldsymbol{z}} - f_{\mathsf{y}_t|\mathbf{y}_{-t}=\widetilde{\boldsymbol{y}}_{-t},\mathbf{a}=\boldsymbol{a},\mathbf{v}=\boldsymbol{v},\mathbf{z}=\boldsymbol{z}}\|_{\mathsf{TV}} \le 2\sqrt{2}x_{\max}^2|\Phi_{tr}^{(y,y)}|.$$

To that end, fix any $u \in [L]$, any $t \in S_u$, any $r \in S_u\backslash\{t\}$, any $\mathbf{z} = \boldsymbol{z}$, $\mathbf{v} = \boldsymbol{v}$, $\mathbf{a} = \boldsymbol{a}$, and any $\boldsymbol{y}_{-t}, \widetilde{\boldsymbol{y}}_{-t} \in \mathcal{Y}^{p_y-1}$ differing only in the r^{th} coordinate. We have

$$\begin{aligned}&\|f_{\mathsf{y}_t|\mathbf{y}_{-t}=\boldsymbol{y}_{-t},\mathbf{a}=\boldsymbol{a},\mathbf{v}=\boldsymbol{v},\mathbf{z}=\boldsymbol{z}} - f_{\mathsf{y}_t|\mathbf{y}_{-t}=\widetilde{\boldsymbol{y}}_{-t},\mathbf{a}=\boldsymbol{a},\mathbf{v}=\boldsymbol{v},\mathbf{z}=\boldsymbol{z}}\|_{\mathsf{TV}}^2 \\ &\overset{(a)}{\le} \frac{1}{2}\mathsf{KL}\left(f_{\mathsf{y}_t|\mathbf{y}_{-t}=\boldsymbol{y}_{-t},\mathbf{a}=\boldsymbol{a},\mathbf{v}=\boldsymbol{v},\mathbf{z}=\boldsymbol{z}}\,\big\|\,f_{\mathsf{y}_t|\mathbf{y}_{-t}=\widetilde{\boldsymbol{y}}_{-t},\mathbf{a}=\boldsymbol{a},\mathbf{v}=\boldsymbol{v},\mathbf{z}=\boldsymbol{z}}\right) \\ &\overset{(b)}{=} \frac{1}{2}(2\Phi_{tr}^{(y,y)}y_r - 2\Phi_{tr}^{(y,y)}\widetilde{y}_r)^2x_{\max}^2 \overset{(c)}{\le} 8x_{\max}^4|\Phi_{tr}^{(y,y)}|^2,\end{aligned}$$

where (a) follows from Pinsker's inequality, (b) follows by (i) applying (Busa-Fekete et al., 2019, Theorem 1) to the exponential family parameterized as per $f_{\mathsf{y}_t|\mathbf{y}_{-t},\mathbf{a}=\boldsymbol{a},\mathbf{v}=\boldsymbol{v},\mathbf{z}=\boldsymbol{z}}$ in Eq. (3.10) and (ii) noting that the Hessian of the log partition function for any regular exponential family is the covariance matrix of the associated sufficient statistic which is bounded by $x_{\max}^2$ when $\mathcal{Y} \subseteq \mathcal{X} = [-x_{\max}, x_{\max}]$, and (c) follows because $y_r, \widetilde{y}_r \in \mathcal{Y}$. This completes the proof. □

3.I Supporting concentration results

In this section, we provide a corollary of Proposition 3.6 that is used to prove the concentration results in Lemma 3.10 and Lemma 3.12. To show any concentration result for the random vector $\mathbf{y}$ conditioned on $(\mathbf{a}, \mathbf{v}, \mathbf{z})$ via Proposition 3.6, we need $\mathbf{y}|\mathbf{a}, \mathbf{v}, \mathbf{z}$ to satisfy the logarithmic Sobolev inequality (defined in Eq. (3.165)). From Proposition 3.5, for this to be true, we need the random vector y_t conditioned on $(\mathbf{y}_{-t}, \mathbf{a}, \mathbf{v}, \mathbf{z})$ to satisfy the logarithmic Sobolev inequality for all $t \in [p_y]$. In the result below, we show this holds with a proof in Section 3.I.1. We define a $\tau \triangleq (\alpha, \xi, \zeta, x_{\max}, \Theta)$-dependent constant:

$$C_{3,\tau} \triangleq \exp\left(x_{\max}(\alpha + 2(2\xi + \zeta)x_{\max})\right). \tag{3.221}$$

Lemma 3.20 (Logarithmic Sobolev inequality for $\mathsf{y}_t|\mathbf{y}_{-t}, \mathbf{a}, \mathbf{v}, \mathbf{z}$)**.** *Given a tuple of random vectors $(\mathbf{y}, \mathbf{a}, \mathbf{v}, \mathbf{z})$ supported on $\mathcal{Y}^{p_y} \times \mathcal{A}^{p_a} \times \mathcal{V}^{p_v} \times \mathcal{Z}^{p_z}$ that is a τ-SGM (Definition 3.8) with $\tau \triangleq (\alpha, \xi, \zeta, x_{\max}, \Theta)$, $\mathsf{y}_t|\mathbf{y}_{-t}, \mathbf{a}, \mathbf{v}, \mathbf{z}$ satisfies $\mathrm{LSI}_{\mathsf{y}_t|\mathbf{y}_{-t}=\boldsymbol{y}_{-t},\mathbf{a}=\boldsymbol{a},\mathbf{v}=\boldsymbol{v},\mathbf{z}=\boldsymbol{z}}\Big(\frac{8x_{\max}^2}{\pi^2}C_{3,\tau}^2\Big)$ for all $t \in [p_y]$, $\boldsymbol{y}_{-t} \in \mathcal{Y}^{p_y-1}$, $\boldsymbol{a} \in \mathcal{A}^{p_a}$, $\boldsymbol{v} \in \mathcal{V}^{p_v}$, and $\boldsymbol{z} \in \mathcal{Z}^{p_z}$.*

Now, we state the desired corollary of Proposition 3.6 with a proof in Section 3.I.2. The corollary makes use of some $\tau \triangleq (\alpha, \xi, \zeta, x_{\max}, \Theta)$-dependent constants:

$$C_{4,\tau} \triangleq 1 + \alpha x_{\max} + 4x_{\max}^2(\xi + \zeta) \quad \text{and} \quad C_{5,\tau} \triangleq \frac{32x_{\max}^3 C_{3,\tau}^4}{\pi^2}. \tag{3.222}$$

Corollary 3.2 (Supporting concentration bounds)**.** *Suppose a tuple of random vectors $(\mathbf{y}, \mathbf{a}, \mathbf{v}, \mathbf{z})$ supported on $\mathcal{Y}^{p_y} \times \mathcal{A}^{p_a} \times \mathcal{V}^{p_v} \times \mathcal{Z}^{p_z}$ corresponds to a τ-SGM (Definition 3.8) with $\tau \triangleq (\alpha, \xi, \zeta, x_{\max}, \Theta)$, and $\mathbf{y}$ conditioned on $(\mathbf{a}, \mathbf{v}, \mathbf{z})$ satisfies the Dobrushin's uniqueness condition (Definition 3.4) with coupling matrix $\overline{\Phi}^{(y,y)}$. For any $\theta, \overline{\theta} \in \Lambda_\theta$ and $\Theta \in \Lambda_\Theta$, define the functions q_1 and q_2 as*

$$q_1(\mathbf{x}) \triangleq \sum_{t \in [p_y]} (\omega_t \mathsf{x}_t)^2 \quad \textit{and} \quad q_2(\mathbf{x}) \triangleq \sum_{t \in [p_y]} \omega_t \mathsf{x}_t \exp\Big(- [\theta_t + 2\Theta_{t,-t}^\top \mathbf{x}_{-t}]\mathsf{x}_t - \Theta_{tt}\overline{\mathsf{x}}_t\Big),$$

where $\omega = \overline{\theta} - \theta$ and $\overline{\mathsf{x}}_t \triangleq \mathsf{x}_t^2 - x_{\max}^2/3$. Then, for any $\varepsilon > 0$

$$\mathbb{P}\Big[\big|q_i(\mathbf{x}) - \mathbb{E}\big[q_i(\mathbf{x})\big|\mathbf{a}, \mathbf{v}, \mathbf{z}\big]\big| \geq \varepsilon \Big| \mathbf{a}, \mathbf{v}, \mathbf{z}\Big] \leq \exp\left(\frac{-c\big(1 - \|\overline{\Phi}^{(y,y)}\|_{\mathrm{op}}\big)^4 \varepsilon^2}{c_i \|\omega\|_2^2}\right) \quad \textit{for} \quad i = 1, 2, \tag{3.223}$$

where c is a universal constant, $c_1 \triangleq 16\alpha^2 x_{\max}^2 C_{5,\tau}^2$, and $c_2 \triangleq C_{3,\tau}^2 C_{4,\tau}^2 C_{5,\tau}^2$ with $C_{3,\tau}$ defined in Eq. (3.221) and $C_{4,\tau}$ and $C_{5,\tau}$ defined in Eq. (3.222).

3.I.1 Proof of Lemma 3.20: Logarithmic Sobolev inequality for $\mathsf{y}_t|\mathbf{y}_{-t}, \mathbf{a}, \mathbf{v}, \mathbf{z}$

Let u be the uniform distribution on $\mathcal{X}$. Then, u satisfies $\mathrm{LSI}_u\Big(\frac{8x_{\max}^2}{\pi^2}\Big)$ (see Ghang et al. (2014, Corollary. 2.4)). Then, using the Holley-Stroock perturbation principle (see

Holley and Stroock (1987, Page. 31), Ledoux (2001, Lemma. 1.2)), for every $t \in [p_y]$, $\boldsymbol{y}_{-t} \in \mathcal{Y}^{p_y-1}$, $\boldsymbol{a} \in \mathcal{A}^{p_a}$, $\boldsymbol{v} \in \mathcal{V}^{p_v}$, and $\boldsymbol{z} \in \mathcal{Z}^{p_z}$, $\mathsf{y}_t|\mathbf{y}_{-t} = \boldsymbol{y}_{-t}, \mathbf{a} = \boldsymbol{a}, \mathbf{v} = \boldsymbol{v}, \mathbf{z} = \boldsymbol{z}$ satisfies the logarithmic Sobolev inequality with a constant bounded by

$$\frac{8x_{\max}^2 \exp(\sup_{x_t \in \mathcal{X}} \psi(x_t; \boldsymbol{x}_{-t}, \boldsymbol{z}) - \inf_{x_t \in \mathcal{X}} \psi(x_t; \boldsymbol{x}_{-t}, \boldsymbol{z}))}{\pi^2},$$

where $\boldsymbol{x} = (\boldsymbol{y}, \boldsymbol{a}, \boldsymbol{v})$ and $\psi(x_t; \boldsymbol{x}_{-t}, \boldsymbol{z}) \triangleq -[\theta_t(\boldsymbol{z}) + 2\Theta_{t,-t}^\top \boldsymbol{x}_{-t}]x_t - \Theta_{tt}\overline{x}_t$ with $\overline{x}_t = x_t^2 - x_{\max}^2/3$. We have

$$\begin{aligned}\exp(\sup_{x_t \in \mathcal{X}} \psi(x_t; \boldsymbol{x}_{-t}, \boldsymbol{z}) - \inf_{x_t \in \mathcal{X}} \psi(x_t; \boldsymbol{x}_{-t}, \boldsymbol{z})) &\overset{(a)}{\leq} \exp\left(2\big|\theta_t(\boldsymbol{z}) + 2\Theta_{t,-t}^\top \boldsymbol{x}_{-t}\big| x_{\max} + 2|\Theta_{tt}| x_{\max}^2\right) \\ &\overset{(b)}{\leq} \exp\left((2\alpha + 4(2\xi + \zeta)x_{\max})x_{\max}\right) \\ &\overset{Eq.~(3.221)}{=} C_{3,\tau}^2,\end{aligned}$$

where (a) follows from Definition 3.8 and (b) follows by using Definition 3.8 along with triangle inequality and Cauchy–Schwarz inequality.

3.I.2 Proof of Corollary 3.2: Supporting concentration bounds

Let $\mathbf{x} = (\mathbf{y}, \mathbf{a}, \mathbf{v})$ and $\boldsymbol{x} = (\boldsymbol{y}, \boldsymbol{a}, \boldsymbol{v})$. To apply Proposition 3.6 to the random vector $\mathbf{y}$ conditioned on $(\mathbf{a}, \mathbf{v}, \mathbf{z})$, we need $\mathbf{y}|\mathbf{a}, \mathbf{v}, \mathbf{z}$ to satisfy the logarithmic Sobolev inequality. From Proposition 3.5, this is true if (i) $f_{\min} = \min_{t \in [p_y], \boldsymbol{x} \in \mathcal{X}^{\widetilde{p}}, \boldsymbol{z} \in \mathcal{X}^{p_z}} f_{\mathsf{x}_t|\mathbf{x}_{-t}, \mathbf{z}}(x_t|\boldsymbol{x}_{-t}, \boldsymbol{z}) > 0$ (see Eq. (3.167)), (ii) $\mathbf{y}|\mathbf{a}, \mathbf{v}, \mathbf{z}$ satisfies the Dobrushin's uniqueness condition, and (iii) $\mathsf{x}_t|\mathbf{x}_{-t}, \mathbf{z}$ satisfies the logarithmic Sobolev inequality for all $t \in [p_y]$. By assumption, $\mathbf{y}|\mathbf{a}, \mathbf{v}, \mathbf{z}$ satisfies the Dobrushin's uniqueness condition with coupling matrix $\overline{\Phi}^{(y,y)}$. From Lemma 3.20, $\mathsf{x}_t|\mathbf{x}_{-t}, \mathbf{z}$ satisfies $\mathrm{LSI}_{\mathsf{x}_t|\mathbf{x}_{-t} = \boldsymbol{x}_{-t}, \mathbf{z} = \boldsymbol{z}}\Big(\frac{8x_{\max}^2 C_{3,\tau}^2}{\pi^2}\Big)$ for every $t \in [p_y]$. It remains to show that $f_{\min} > 0$. Consider any $t \in [p_y]$, any $\boldsymbol{x} \in \mathcal{X}^{\widetilde{p}}$, and any $\boldsymbol{z} \in \mathcal{Z}^{p_z}$. Let $\overline{x}_t = x_t^2 - x_{\max}^2/3$. We have

$$\begin{aligned} f_{\mathsf{x}_t|\mathbf{x}_{-t}, \mathbf{z}}(x_t|\boldsymbol{x}_{-t}, \boldsymbol{z}) &\overset{(a)}{=} \frac{\exp\Big([\theta_t(\boldsymbol{z}) + 2\Theta_{t,-t}^\top \boldsymbol{x}_{-t}]x_t + \Theta_{tt}\overline{x}_t\Big)}{\int_{\mathcal{X}} \exp\Big([\theta_t(\boldsymbol{z}) + 2\Theta_{t,-t}^\top \boldsymbol{x}_{-t}]x_t + \Theta_{tt}\overline{x}_t\Big) dx_t} \\ &\overset{(b)}{\geq} \frac{\exp\Big(-|\theta_t(\boldsymbol{z}) + 2\Theta_{t,-t}^\top \boldsymbol{x}_{-t}|x_{\max} - \Theta_{tt}x_{\max}^2\Big)}{\int_{\mathcal{X}} \exp\Big(|\theta_t(\boldsymbol{z}) + 2\Theta_{t,-t}^\top \boldsymbol{x}_{-t}|x_{\max} + \Theta_{tt}x_{\max}^2\Big) dx_t} \\ &\overset{(c)}{\geq} \frac{\exp\Big(-(|\theta(\boldsymbol{z})| + 2\|\Theta_{t,-t}\|_1\|\boldsymbol{x}\|_\infty)x_{\max} - \Theta_{tt}x_{\max}^2\Big)}{\int_{\mathcal{X}} \exp\Big((|\theta(\boldsymbol{z})| + 2\|\Theta_{t,-t}\|_1\|\boldsymbol{x}\|_\infty)x_{\max} + \Theta_{tt}x_{\max}^2\Big) dx_t} \\ &\overset{(d)}{\geq} \frac{\exp\Big(-(\alpha + 2(2\xi + \zeta)x_{\max})x_{\max}\Big)}{\int_{\mathcal{X}} \exp\Big((\alpha + 2(2\xi + \zeta)x_{\max})x_{\max}\Big) dx_t} \overset{(e)}{=} \frac{1}{2x_{\max}C_{3,\tau}^2},\end{aligned}$$

where (a) follows from Eq. (3.10), (b) and (d) follow from Definition 3.8, (c) follows by triangle inequality and Cauchy–Schwarz inequality, and (e) follows because $\int_{\mathcal{X}} dx_t = 2x_{\max}$. Therefore, $f_{\min} = \frac{1}{2x_{\max} C_{3,\tau}^2}$. Putting (i), (ii), and (iii) together, and using Proposition 3.5, we see that $\mathbf{y}|\mathbf{a}, \mathbf{v}, \mathbf{z}$ satisfies $\mathrm{LSI}_{\mathbf{x}}\Big(\frac{C_{5,\tau}}{(1-\|\overline{\Phi}^{(y,y)}\|_{\mathrm{op}})^2}\Big)$ where $C_{5,\tau}$ was defined in Eq. (3.222).

Now, we apply Proposition 3.6 to q_1 and q_2 one-by-one. The general strategy is to choose appropriate pseudo derivatives and pseudo Hessians for both q_1 and q_2, and evaluate the corresponding terms appearing in Proposition 3.6.

Concentration for q_1. Fix any $\boldsymbol{x} \in \mathcal{X}^{\widetilde{p}}$. We start by decomposing $q_1(\boldsymbol{x})$ as follows

$$q_1(\boldsymbol{x}) = \overline{\omega}^\top r(\boldsymbol{x}), \tag{3.224}$$

where $\overline{\omega} \triangleq (\omega_1^2, \cdots, \omega_{p_y}^2)$ and $r(\boldsymbol{x}) \triangleq (r_1(\boldsymbol{x}), \cdots, r_{p_y}(\boldsymbol{x}))$ with $r_t(\boldsymbol{x}) = x_t^2$ for every $t \in [p_y]$. Next, we define $H : \mathcal{X}^{\widetilde{p}} \to \mathbb{R}^{p_y \times p_y}$ such that

$$H_{tu}(\boldsymbol{x}) = \frac{dr_u(\boldsymbol{x})}{dx_t} \quad \text{for every } t, u \in [p_y]. \tag{3.225}$$

Pseudo derivative. We bound the ℓ_2 norm of the gradient of $q_1(\boldsymbol{x})$ as follows

$$\begin{aligned}\big\|\nabla q_1(\boldsymbol{x})\big\|_2^2 = \sum_{t\in[p_y]} \Big(\frac{dq_1(\boldsymbol{x})}{dx_t}\Big)^2 &\overset{Eq.\ (3.224)}{=} \sum_{t\in[p_y]} \Big(\frac{\overline{\omega}^\top dr(\boldsymbol{x})}{dx_t}\Big)^2 \\ &\overset{Eq.\ (3.225)}{=} \big\|H(\boldsymbol{x})\overline{\omega}\big\|_2^2 \\ &\overset{(a)}{\leq} \|H(\boldsymbol{x})\|_{\mathrm{op}}^2 \big\|\overline{\omega}\big\|_2^2 \overset{(b)}{\leq} \|H(\boldsymbol{x})\|_1 \|H(\boldsymbol{x})\|_\infty \big\|\overline{\omega}\big\|_2^2,\end{aligned} \tag{3.226}$$

where (a) follows because induced matrix norms are submultiplicative and (b) follows because the matrix operator norm is bounded by square root of the product of matrix one norm and matrix infinity norm. Now, we claim that the one norm and the infinity norm of $H(\boldsymbol{x})$ are bounded as follows

$$\max\Big\{\max_{\boldsymbol{x}\in\mathcal{X}^{\widetilde{p}}} \|H(\boldsymbol{x})\|_1, \max_{\boldsymbol{x}\in\mathcal{X}^{\widetilde{p}}} \|H(\boldsymbol{x})\|_\infty\Big\} \leq 2x_{\max}. \tag{3.227}$$

Taking this claim as given at the moment, we continue with our proof. Combining Eqs. (3.226) and (3.227), we have

$$\max_{\boldsymbol{x}\in\mathcal{X}^{\widetilde{p}}} \big\|\nabla q_1(\boldsymbol{x})\big\|_2^2 \leq 4x_{\max}^2 \big\|\overline{\omega}\big\|_2^2 = 4x_{\max}^2 \sum_{t\in[p_y]} \omega_t^4 \leq 4x_{\max}^2 \max_{u\in[p_y]} \omega_u^2 \sum_{t\in[p_y]} \omega_t^2 \overset{(a)}{\leq} 16x_{\max}^2 \alpha^2 \big\|\omega\big\|_2^2,$$

where (a) follows because $\omega \in 2\Lambda_\theta$. Therefore, we choose the pseudo derivative (see Definition 3.5) as follows

$$\widetilde{\nabla} q_1(\boldsymbol{x}) = 4x_{\max}\alpha\big\|\omega\big\|_2. \tag{3.228}$$

Pseudo Hessian. Fix any $\rho \in \mathbb{R}$. We bound $\|\nabla(\rho^\top \widetilde{\nabla} q_1(\boldsymbol{x}))\|_2^2$ (see Definition 3.5) as follows

$$\|\nabla(\rho^\top \widetilde{\nabla} q_1(\boldsymbol{x}))\|_2^2 = \sum_{u \in [p_y]} \Big(\frac{d\rho^\top \widetilde{\nabla} q_1(\boldsymbol{x})}{dx_u}\Big)^2 \overset{Eq.~(3.228)}{=} 0.$$

Therefore, we choose the pseudo Hessian (see Definition 3.5) as follows

$$\widetilde{\nabla}^2 q_1(\boldsymbol{x}) = 0. \tag{3.229}$$

The concentration result in Eq. (3.223) for q_1 follows by applying Proposition 3.6 with the pseudo discrete derivative defined in Eq. (3.228) and the pseudo discrete Hessian defined in Eq. (3.229).

It remains to show that the one-norm and the infinity-norm of $H(\boldsymbol{x})$ are bounded as in Eq. (3.227).

Bounds on the one-norm and the infinity-norm of $H(\boldsymbol{x})$. We have

$$H_{tu}(\boldsymbol{x}) = \begin{cases} 2x_t & \text{if} \quad t = u, \\ 0 & \text{otherwise.} \end{cases} \tag{3.230}$$

Therefore,

$$\|H(\boldsymbol{x})\|_1 = \max_{u \in [p_y]} \sum_{t \in [p_y]} |H_{tu}(\boldsymbol{x})| \overset{Eq.~(3.230)}{\leq} \max_{u \in [p_y]} 2|x_u| \overset{(a)}{\leq} 2x_{\max} \quad \text{and}$$
$$\|H(\boldsymbol{x})\|_\infty = \max_{t \in [p_y]} \sum_{u \in [p_y]} |H_{tu}(\boldsymbol{x})| \overset{Eq.~(3.230)}{\leq} \max_{t \in [p_y]} 2|x_t| \overset{(a)}{\leq} 2x_{\max},$$

where (a) follows from Definition 3.8.

Concentration for q_2. Fix any $\boldsymbol{x} \in \mathcal{X}^{\widetilde{p}}$. We start by decomposing $q_2(\boldsymbol{x})$ as follows

$$q_2(\boldsymbol{x}) = \omega^\top r(\boldsymbol{x}), \tag{3.231}$$

where $r(\boldsymbol{x}) \triangleq (r_1(\boldsymbol{x}), \cdots, r_{p_y}(\boldsymbol{x}))$ with $r_t(\boldsymbol{x}) = x_t \exp\big(-[\theta_t + 2\Theta_{t,-t}^\top \boldsymbol{x}_{-t}]x_t - \Theta_{tt}\overline{x}_t\big)$ for every $t \in [p_y]$. Next, we define $H : \mathcal{X}^{\widetilde{p}} \to \mathbb{R}^{p_y \times p_y}$ such that

$$H_{tu}(\boldsymbol{x}) = \frac{dr_u(\boldsymbol{x})}{dx_t} \quad \text{for every } t, u \in [p_y]. \tag{3.232}$$

Pseudo derivative. We bound the ℓ_2 norm of the gradient of $q_2(\boldsymbol{x})$ as follows

$$\big\|\nabla q_2(\boldsymbol{x})\big\|_2^2 = \sum_{t \in [p_y]} \Big(\frac{dq_2(\boldsymbol{x})}{dx_t}\Big)^2 \overset{Eq.~(3.231)}{=} \sum_{t \in [p_y]} \Big(\frac{\omega^\top dr(\boldsymbol{x})}{dx_t}\Big)^2$$
$$\overset{Eq.~(3.232)}{=} \big\|H(\boldsymbol{x})\omega\big\|_2^2$$

$$\overset{(a)}{\leq} \|H(\boldsymbol{x})\|_{\mathrm{op}}^2 \big\|\omega\big\|_2^2 \overset{(b)}{\leq} \|H(\boldsymbol{x})\|_1 \|H(\boldsymbol{x})\|_\infty \big\|\omega\big\|_2^2, \qquad (3.233)$$

where (a) follows because induced matrix norms are submultiplicative and (b) follows because the matrix operator norm is bounded by square root of the product of matrix one norm and matrix infinity norm. Now, we claim that the one norm and the infinity norm of $H(\boldsymbol{x})$ are bounded as follows

$$\max\left\{\max_{\boldsymbol{x}\in\mathcal{X}^{\widetilde{p}}} \|H(\boldsymbol{x})\|_1, \max_{\boldsymbol{x}\in\mathcal{X}^{\widetilde{p}}} \|H(\boldsymbol{x})\|_\infty\right\} \leq C_{3,\tau} C_{4,\tau}. \qquad (3.234)$$

where $C_{3,\tau}$ and $C_{4,\tau}$ were defined in Eq. (3.221) and Eq. (3.222) respectively. Taking this claim as given at the moment, we continue with our proof. Combining Eqs. (3.233) and (3.234), we have

$$\max_{\boldsymbol{x}\in\mathcal{X}^{\widetilde{p}}} \big\|\nabla q_2(\boldsymbol{x})\big\|_2^2 \leq C_{3,\tau}^2 C_{4,\tau}^2 \big\|\omega\big\|_2^2.$$

Therefore, we choose the pseudo derivative (see Definition 3.5) as follows

$$\widetilde{\nabla} q_2(\boldsymbol{x}) = C_{3,\tau} C_{4,\tau} \big\|\omega\big\|_2. \qquad (3.235)$$

Pseudo Hessian. Fix any $\rho \in \mathbb{R}$. We bound $\|\nabla(\rho^\top \widetilde{\nabla} q_2(\boldsymbol{x}))\|_2^2$ (see Definition 3.5) as follows

$$\|\nabla(\rho^\top \widetilde{\nabla} q_2(\boldsymbol{x}))\|_2^2 = \sum_{u\in[p_y]} \Big(\frac{d\rho^\top \widetilde{\nabla} q_2(\boldsymbol{x})}{dx_u}\Big)^2 \overset{Eq.~(3.235)}{=} 0.$$

Therefore, we choose the pseudo Hessian (see Definition 3.5) as follows

$$\widetilde{\nabla}^2 q_2(\boldsymbol{x}) = 0. \qquad (3.236)$$

The concentration result in Eq. (3.223) for q_1 follows by applying Proposition 3.6 with the pseudo discrete derivative defined in Eq. (3.235) and the pseudo discrete Hessian defined in Eq. (3.236).

It remains to show that the one-norm and the infinity-norm of $H(\boldsymbol{x})$ are bounded as in Eq. (3.234).

Bounds on the one-norm and the infinity-norm of H. We have

$$H_{tu}(\boldsymbol{x}) = \begin{cases} \big[1 - [\theta_u + 2\Theta_u^\top \boldsymbol{x}]x_u\big] \exp\big(-[\theta_u + 2\Theta_{u,-u}^\top \boldsymbol{x}_{-u}]x_u - \Theta_{uu}\overline{x}_u\big) & \text{if} \quad t = u, \\ -2\Theta_{tu} x_u^2 \exp\big(-[\theta_u + 2\Theta_{u,-u}^\top \boldsymbol{x}_{-u}]x_u - \Theta_{uu}\overline{x}_u\big) & \text{otherwise.} \end{cases} \qquad (3.237)$$

Therefore,

$$\|H(\boldsymbol{x})\|_1 = \max_{u\in[p_y]} \sum_{t\in[p_y]} |H_{tu}(\boldsymbol{x})|$$

$$\overset{Eq.\ (3.237)}{=} \max_{u\in[p_y]} \big|1-[\theta_u+2\Theta_u^\top \boldsymbol{x}]x_u\big| \exp\big(-[\theta_u+2\Theta_{u,-u}^\top \boldsymbol{x}_{-u}]x_u - \Theta_{uu}\overline{x}_u\big)$$
$$+ 2\max_{u\in[p_y]} x_u^2 \exp\big(-[\theta_u+2\Theta_{u,-u}^\top \boldsymbol{x}_{-u}]x_u - \Theta_{uu}\overline{x}_u\big) \sum_{t\neq u} |\Theta_{tu}|$$
$$\overset{(a)}{\leq} (1+\alpha x_{\max} + 4x_{\max}^2(\xi+\zeta)) \exp\left(x_{\max}(\alpha+2(2\xi+\zeta)x_{\max})\right) \overset{(b)}{=} C_{3,\tau}C_{4,\tau},$$

where (a) follows from Definition 3.8 along with triangle inequality and Cauchy–Schwarz inequality and (b) follows from Eqs. (3.221) and (3.222). Similarly, we have

$$\|H(\boldsymbol{x})\|_\infty = \max_{t\in[p_y]} \sum_{u\in[p]} |H_{tu}(\boldsymbol{x})|$$
$$\overset{Eq.\ (3.237)}{=} \max_{t\in[p_y]} \big|1-[\theta_t+2\Theta_t^\top \boldsymbol{x}]x_t\big| \exp\big(-[\theta_t+2\Theta_{t,-t}^\top \boldsymbol{x}_{-t}]x_t - \Theta_{tt}\overline{x}_t\big)$$
$$+ 2\max_{t\in[p_y]} \sum_{u\neq t} |\Theta_{tu}| x_u^2 \exp\big(-[\theta_u+2\Theta_{u,-u}^\top \boldsymbol{x}_{-u}]x_u - \Theta_{uu}\overline{x}_u\big)$$
$$\overset{(a)}{\leq} (1+\alpha x_{\max} + 4x_{\max}^2(\xi+\zeta)) \exp\left(x_{\max}(\alpha+2(2\xi+\zeta)x_{\max})\right) \overset{(b)}{=} C_{3,\tau}C_{4,\tau},$$

where (a) follows from Definition 3.8 along with triangle inequality and Cauchy–Schwarz inequality and (b) follows from Eqs. (3.221) and (3.222).

Chapter 4

Causal Inference via Latent Factor Modeling

4.1 Introduction

This chapter presents a novel framework using latent factor modeling to estimate treatment effects in modern data-rich environments in the presence of unobserved confounding. As a motivating example, consider an internet retailer. The platform collects not only information on purchases of many customers across many products or product categories, but also on glance views, impressions, conversions, engagement metrics, navigation paths, shipping choices, payment methods, returns, reviews, and more. While some variables, such as geo-location and type of device or browser, can be safely treated as pre-determined relative to the platform's treatments (advertisements, discounts, web-page design, etc.), most are outcomes affected by the treatments, latent customer preferences, and unobserved product features. We leverage the availability of many outcome measures in modern data-rich environments to estimate treatment effects in the presence of unobserved confounding. The core identification concept is that if each element of a high-dimensional outcome vector is influenced by a common low-dimensional vector of unobserved confounders, it becomes possible to remove the influence of the confounders and identify treatment effects.

Two primary approaches to the estimation of treatment effects are outcome-based and assignment-based methods. Consider again the example of an internet-retail platform where customers interact with various product categories. For each consumer-category pair, the platform makes decisions to either offer a discount or not, and records whether the consumer purchased a product in the category. Outcome-based methods operate by imputing the missing potential outcomes for each consumer-product category pair. This process involves predicting whether a consumer, who received a discount, would have made the purchase without the discount (i.e., the potential outcome without discount), and conversely, if a consumer who did not receive the discount would have purchased the product had they received the discount (i.e., the potential outcome with discount). In contrast, assignment-based methods estimate the probabilities of consumers receiving discounts in each product category and adjust for missing potential outcomes by weighting observed outcomes inversely to the probability of missingness.

A substantial body of literature has explored outcome-based methods, particularly in settings where all confounding factors are measured (see, e.g., Abadie and Imbens, 2006; Angrist, 1998; Cochran, 1968; Rosenbaum and Rubin, 1983b, among many others). Imputing potential outcomes in the presence of unobserved confounders poses a more complex challenge. In this context, a commonly adopted framework is the synthetic control method and its variants (see, e.g., Abadie et al., 2010a; Abadie and Gardeazabal, 2003a; Arkhangelsky et al., 2021; Cattaneo et al., 2021). An alternative but related approach to outcome imputation under unobserved confounding is the latent factor framework (Bai, 2009; Bai and Ng, 2002; Xiong and Pelger, 2023), wherein each element of the large-dimensional outcome vector is influenced by the same low-dimensional vector of unobserved confounders. Matrix completion methods (see, e.g., Agarwal et al., 2023; Athey et al., 2021; Bai and Ng, 2021; Chatterjee, 2015; Dwivedi et al., 2022a) which have found widespread applications in recommendation systems and panel data models, are closely related to latent factor models. Similarly, existing assignment-based procedures to estimate treatment effects rely on the assumption of no unmeasured confounding (see, e.g., Hirano et al., 2003; Robins et al., 2000; Wooldridge, 2007), common trends restrictions (Abadie, 2005), or the availability of an instrumental variable (Abadie, 2003; Sloczynski et al., 2024).

In this chapter, we propose a doubly-robust estimator (see Bang and Robins, 2005; Chernozhukov et al., 2018; Robins et al., 1994) of treatment effects in the presence of unobserved confounding. This estimator leverages information on both the outcome process and the treatment assignment mechanism under a latent factor framework. It combines outcome imputation and inverse probability weighting with a new cross-fitting approach for matrix completion. We show that the proposed doubly-robust estimator has better finite-sample guarantees than alternative outcome-based and assignment-based estimators. Furthermore, the doubly-robust estimator is approximately Gaussian, asymptotically unbiased, and converges at a parametric rate, under provably valid error rates for matrix completion, irrespective of other properties of the matrix completion algorithm used for estimation.

To our knowledge, this is the first work that leverages latent structures in both the assignment and the outcome processes to obtain a doubly-robust estimator of treatment effects in the presence of unobserved confounding. Arkhangelsky and Imbens (2022) study doubly-robust identification with longitudinal data under the assumption that conditioning of a function of the treatment assignments over time (e.g., the fraction of times an individual is exposed to treatment) is enough to remove confounding. Athey et al. (2021), Bai and Ng (2021), Dwivedi et al. (2022a), Agarwal et al. (2023), and Xiong and Pelger (2023) propose estimators that apply matrix completion techniques to impute potential outcomes. Although these studies utilize low-rank restrictions in the outcome process, they do not investigate the possibility of similar latent structures in the treatment assignment process. Our work addresses this question, and demonstrates substantial benefits from incorporating knowledge about the structure of the assignment mechanism.

4.1.1 Some terminology and notation

For any real number $b \in \mathbb{R}$, $\lfloor b \rfloor$ is the greatest integer less than or equal to b. For any positive integer b, $[b]$ denotes the set of integers from 1 to b, i.e., $[b] \triangleq \{1, \cdots, b\}$. We use c to denote any generic universal constant, whose value may change between instances. For any $c > 0$, $m(c) = \max\{c, \sqrt{c}\}$ and $\ell_c = \log(2/c)$. For any two deterministic sequences a_n and b_n where b_n is positive, $a_n = O(b_n)$ means that there exist a finite $c > 0$ and a finite $n_0 > 0$ such that $|a_n| \leq c\, b_n$ for all $n \geq n_0$. Similarly, $a_n = o(b_n)$ means that for every $c > 0$, there exists a finite $n_0 > 0$ such that $|a_n| < c\, b_n$ for all $n \geq n_0$. Further, $a_n = \Omega(b_n)$ means that there exist a finite $c > 0$ and a finite $n_0 > 0$ such that $|a_n| \geq c\, b_n$ for all $n \geq n_0$. For a sequence of random variables, $x_n = O_p(1)$ means that the sequence $|x_n|$ is stochastically bounded, i.e., for every $\varepsilon > 0$, there exists a finite $\delta > 0$ and a finite $n_0 > 0$ such that $\mathbb{P}\big(|x_n| > \delta\big) < \varepsilon$ for all $n \geq n_0$. Similarly, $x_n = o_p(1)$ means that the sequence $|x_n|$ converges to zero in probability, i.e., for every $\varepsilon > 0$ and $\delta > 0$, there exists a finite $n_0 > 0$ such that $\mathbb{P}\big(|x_n| > \delta\big) < \varepsilon$ for all $n \geq n_0$. For sequences of random variables x_n and b_n, $x_n = O_p(b_n)$ means $x_n = \overline{x}_n b_n$ where the sequence $\overline{x}_n = O_p(1)$. Likewise, $x_n = o_p(b_n)$ means $x_n = \overline{x}_n b_n$ where the sequence $\overline{x}_n = o_p(1)$.

A mean-zero random variable x is subGaussian if there exists some $b > 0$ such that $\mathbb{E}[\exp(sx)] \leq \exp(b^2 s^2/2)$ for all $s \in \mathbb{R}$. Then, the subGaussian norm of x is given by $\|x\|_{\psi_2} = \inf\{t > 0 : \mathbb{E}[\exp(x^2/t^2)] \leq 2\}$. A mean-zero random variable x is subExponential if there exist some $b_1, b_2 > 0$ such that $\mathbb{E}[\exp(sx)] \leq \exp(b_1^2 s^2/2)$ for all $-1/b_2 < s < 1/b_2$. Then, the subExponential norm of x is given by $\|x\|_{\psi_1} = \inf\{t > 0 : \mathbb{E}[\exp(|x|/t)] \leq 2\}$. $\texttt{Uniform}(a, b)$ denotes the uniform distribution over the interval $[a, b]$ for $a, b \in \mathbb{R}$ such that $a < b$. $\mathcal{N}(\mu, \sigma^2)$ denotes the Gaussian distribution with mean μ and variance σ^2.

For a vector $u \in \mathbb{R}^n$, we denote its t^{th} coordinate by u_t and its 2-norm $\|u\|_2$. For a matrix $U \in \mathbb{R}^{n_1 \times n_2}$, we denote the element in i^{th} row and j^{th} column by $u_{i,j}$, the i^{th} row by $U_{i,\cdot}$, the j^{th} column by $U_{\cdot,j}$, the largest eigenvalue by $\lambda_{\max}(U)$, and the smallest by $\lambda_{\min}(U)$. Given a set of indices $\mathcal{R} \subseteq [n_1]$ and $\mathcal{C} \subseteq [n_2]$, $U_{\mathcal{I}} \in \mathbb{R}^{|\mathcal{R}| \times |\mathcal{C}|}$ is a sub-matrix of U corresponding to the entries in $\mathcal{I} \triangleq \mathcal{R} \times \mathcal{C}$, and $U_{-\mathcal{I}} = \{u_{i,j} : (i,j) \in \{[n_1] \times [n_2]\} \setminus \mathcal{I}\}$. Further, we denote the Frobenius norm by $\|U\|_{\mathrm{F}} \triangleq \big(\sum_{i \in [n_1], j \in [n_2]} u_{i,j}^2\big)^{1/2}$, the $(1, 2)$ operator norm by $\|U\|_{1,2} \triangleq \max_{j \in [n_2]} \big(\sum_{i \in [n_1]} u_{i,j}^2\big)^{1/2}$, the $(2, \infty)$ operator norm by $\|U\|_{2,\infty} \triangleq \max_{i \in [n_1]} \big(\sum_{j \in [n_2]} u_{i,j}^2\big)^{1/2}$, and the maximum norm by $\|U\|_{\max} \triangleq \max_{i \in [n_1], j \in [n_2]} |u_{i,j}|$. Given two matrices $U, V \in \mathbb{R}^{n_1 \times n_2}$, the operators $\odot$ and $\oslash$ denote element-wise multiplication and division, respectively, i.e., $t_{i,j} = u_{i,j} \cdot v_{i,j}$ when $T = U \odot V$, and $t_{i,j} = u_{i,j}/v_{i,j}$ when $T = U \oslash V$. When V is a binary matrix, i.e., $V \in \{0, 1\}^{n_1 \times n_2}$, the operator $\otimes$ is defined such that $t_{i,j} = u_{i,j}$ if $v_{i,j} = 1$ and $t_{i,j} = ?$ if $v_{i,j} = 0$ for $T = U \otimes V$. Given two matrices $U \in \mathbb{R}^{n_1 \times n_2}$ and $V \in \mathbb{R}^{n_1 \times n_3}$, the operator $*$ denotes the (transposed column-wise) Khatri-Rao product of U and V, i.e., $T = U * V \in \mathbb{R}^{n_1 \times n_2 n_3}$ such that $t_{i,j} = u_{i,j-n_2\bar{j}} \cdot v_{i,1+\bar{j}}$ where $\bar{j} = \lfloor (j-1)/n_2 \rfloor$. For random objects U and V, $U \perp\!\!\!\perp V$ means that U is independent of V.

4.2 Problem Formulation

Consider a setting with N units and M measurements per unit. For each unit-measurement pair $i \in [N]$ and $j \in [M]$, we observe a treatment assignment $a_{i,j} \in \{0,1\}$ and the value of the outcome $y_{i,j} \in \mathbb{R}$. Although our results can be easily generalized to multi-ary treatments, for the ease of exposition, we focus on binary treatments.

We operate within the Neyman-Rubin potential outcomes framework and denote the potential outcome for unit $i \in [N]$ and measurement $j \in [M]$ under treatment $a \in \{0,1\}$ by $y_{i,j}^{(a)} \in \mathbb{R}$. A no-spillover assumption is implicit in the notation, i.e., the potential outcome $y_{i,j}^{(a)}$ does not depend on the treatment assignment for any other unit-measurement pair. In the context of online retail data, the assumption of no spillovers across measurements is justified if the cross-elasticity of demand across product categories, j, is low. Our framework allows for the possibility that the same treatment affects multiple outcomes (e.g., $a_{i,j} = a_{i,j'}$ with probability one, for some j and j' in $[M]$). Realized outcomes, $y_{i,j}$, depend on potential outcomes and treatment assignments,

$$y_{i,j} = y_{i,j}^{(0)}(1 - a_{i,j}) + y_{i,j}^{(1)} a_{i,j}, \tag{4.1}$$

for all $i \in [N]$ and $j \in [M]$. Section 4.4.4 and Appendix 4.I extend this framework to a panel data setting with lagged treatment effects.

4.2.1 Sources of stochastic variation

In the setup of this chapter, each unit $i \in [N]$ is characterized by a set of unknown parameters, $\{(\theta_{i,j}^{(0)}, \theta_{i,j}^{(1)}, p_{i,j}) \in \mathbb{R}^2 \times [0,1]\}_{j\in[M]}$, which we treat as fixed. Potential outcomes and treatment assignments are generated as follows: for all $i \in [N], j \in [M]$, and $a \in \{0,1\}$,

$$y_{i,j}^{(a)} = \theta_{i,j}^{(a)} + \varepsilon_{i,j}^{(a)} \tag{4.2}$$

and

$$a_{i,j} = p_{i,j} + \eta_{i,j}, \tag{4.3}$$

where $\varepsilon_{i,j}^{(a)}$ and $\eta_{i,j}$ are mean-zero random variables, and

$$\eta_{i,j} = \begin{cases} -p_{i,j} & \text{with probability} \quad 1 - p_{i,j} \\ 1 - p_{i,j} & \text{with probability} \quad p_{i,j}. \end{cases} \tag{4.4}$$

It follows that $\theta_{i,j}^{(a)}$ is the mean of the potential outcome $y_{i,j}^{(a)}$, and $p_{i,j}$ is the unknown assignment probability or latent propensity score. The matrices $\Theta^{(0)} \triangleq \{\theta_{i,j}^{(0)}\}_{i\in[N],j\in[M]}$, $\Theta^{(1)} \triangleq \{\theta_{i,j}^{(1)}\}_{i\in[N],j\in[M]}$, and $P \triangleq \{p_{i,j}\}_{i\in[N],j\in[M]}$ collect mean potential outcomes and assignment probabilities. Then, the matrices $E^{(0)} \triangleq \{\varepsilon_{i,j}^{(0)}\}_{i\in[N],j\in[M]}$, $E^{(1)} \triangleq \{\varepsilon_{i,j}^{(1)}\}_{i\in[N],j\in[M]}$, and $W \triangleq \{\eta_{i,j}\}_{i\in[N],j\in[M]}$ capture all sources of randomness in potential outcomes and treatment assignments.

Our setup allows $\Theta^{(0)}, \Theta^{(1)}$ to be arbitrarily associated with P, inducing unobserved confounding. The assumptions in Section 4.4 imply that $\Theta^{(0)}, \Theta^{(1)}$, and P include all confounding factors, and require $(\varepsilon_{i,j}^{(0)}, \varepsilon_{i,j}^{(1)}) \perp\!\!\!\perp \eta_{i,j}$ for every $i \in [N]$ and $j \in [M]$.

4.2.2 Target causal estimand

For any given unit $i \in [N]$, our goal is to estimate the effect of the treatment averaged over all measurements,

$$\mathrm{ATE}_{i,\cdot} \triangleq \mu_{i,\cdot}^{(1)} - \mu_{i,\cdot}^{(0)}$$

where

$$\mu_{i,\cdot}^{(a)} \triangleq \frac{1}{M} \sum_{j \in [M]} \theta_{i,j}^{(a)}.$$

Analogously, for any given measurement $j \in [M]$, we could aim to estimate the effect of the treatment averaged over all units,

$$\mathrm{ATE}_{\cdot,j} \triangleq \mu_{\cdot,j}^{(1)} - \mu_{\cdot,j}^{(0)} \tag{4.5}$$

where

$$\mu_{\cdot,j}^{(a)} \triangleq \frac{1}{N} \sum_{i \in [N]} \theta_{i,j}^{(a)}.$$

For consistency with the existing literature, we consider the latter estimand and note that it is straightforward to adapt the methods in this chapter to the former estimand as well as the estimation of alternative parameters, like the treatment effect over a subset of the measurements, $S \subset [M]$. We note that $\mathrm{ATE}_{\cdot,j}$ is akin to the conditional average treatment effect of Abadie and Imbens (2006), but based on the latent means, $\theta_{i,j}^{(a)}$, in Eq. (4.2) rather than on conditional means that depend on observed covariates only.

4.3 Learning Algorithm

In this section, we propose a procedure that uses the treatment assignment matrix A and the observed outcomes matrix Y to estimate $\mathrm{ATE}_{\cdot,j}$, where

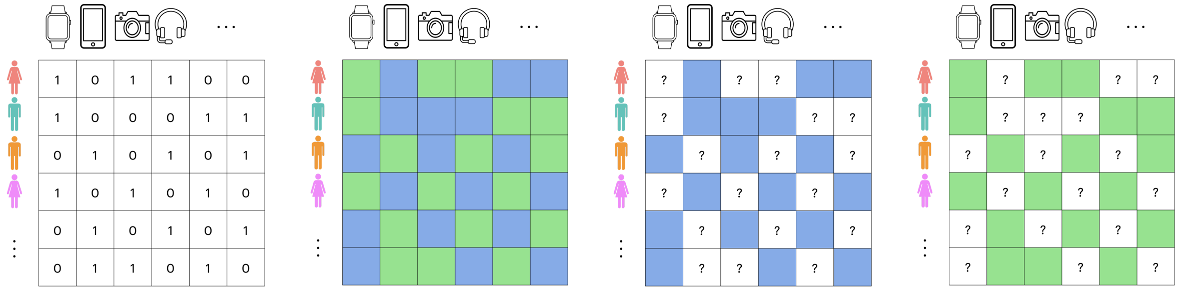
$$Y \triangleq \{y_{i,j}\}_{i \in [N], j \in [M]} \quad \text{and} \quad A \triangleq \{a_{i,j}\}_{i \in [N], j \in [M]}.$$

The estimator proposed in this section leverages matrix completion as a key subroutine. We start the section with a brief overview of matrix completion methods.

4.3.1 Matrix completion: A primer

Consider a matrix of parameters $T \in \mathbb{R}^{N \times M}$. While T is unobserved, we observe the matrix $S \in \{\mathbb{R} \cup \{\,?\}\}^{N \times M}$ where ? denotes a missing value. The relationship between S and T is given by

$$S = (T + H) \otimes F. \tag{4.6}$$



$$\begin{matrix} 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 & 0 \end{matrix}$$

(a) A (b) Y (c) $Y^{(0),\text{obs}}$ (d) $Y^{(1),\text{obs}}$

Figure 4.3.1: Schematic of the treatment assignment matrix A, the observed outcomes matrix Y (where green and blue fills indicate observations under $a = 1$ and $a = 0$, respectively), and the observed component of the potential outcomes matrices, i.e., $Y^{(0),\text{obs}}$ and $Y^{(1),\text{obs}}$ (where ? indicates a missing value). All matrices are $N \times M$ where N is the number of customers and M is the number of products.

Here, $H \in \mathbb{R}^{N\times M}$ is a noise matrix, and $F \in \{0,1\}^{N\times M}$ is a masking matrix with ones for the recorded entries of S and zeros for the missing entries.

A matrix completion algorithm, denoted by $\texttt{MC}$, takes the S as its input, and returns an estimate of T, which we denote by $\widehat{T}$ or $\texttt{MC}(S)$. In other words, $\texttt{MC}$ produces an estimate of a matrix from noisy observations of a subset of all the elements of the matrix.

The matrix completion literature is rich with algorithms $\texttt{MC}$ that provide error guarantees, namely bounds on $\|\texttt{MC}(S) - T\|$ for a suitably chosen norm/metric $\|\cdot\|$, under a variety of assumptions on the triplet (T, H, F). Typical assumptions are (i) T is low-rank, (ii) the entries of H are independent, mean-zero and sub-Gaussian random variables, and (iii) the entries of F are independent Bernoulli random variables. Though matrix completion is commonly associated with the imputation of missing values, a typically underappreciated aspect is that it also denoises the observed matrix. Even when each entry of S is observed, $\texttt{MC}(S)$ subtracts the effects of H from S, i.e., it performs matrix denoising. Nguyen et al. (2019) provide a survey of various matrix completion algorithms.

4.3.2 Key building blocks

We now define and express matrices that are related to the quantities of interest $\Theta^{(0)}, \Theta^{(1)}$, and P in a form similar to Eq. (4.6). See Figure 4.3.1 for a visual representation of these matrices.

- **Outcomes**: Let $Y^{(0),\text{obs}} = Y \otimes (\mathbf{1} - A) \in \{\mathbb{R} \cup \{\,?\}\}^{N\times M}$ be a matrix with (i,j)-th entry equal to $y_{i,j}$ if $a_{i,j} = 0$, and equal to ? otherwise. Here, $\mathbf{1}$ is the $N \times M$ matrix with all entries equal to one. Analogously, let $Y^{(1),\text{obs}} = Y \otimes A \in \{\mathbb{R} \cup \{\,?\}\}^{N\times M}$ be a matrix with (i,j)-th entry equal to $y_{i,j}$ if $a_{i,j} = 1$, and equal to ? otherwise. In other words, $Y^{(0),\text{obs}}$ and $Y^{(1),\text{obs}}$ capture the observed components of $\{y^{(0)}_{i,j}\}_{i\in[N],j\in[M]}$ and $\{y^{(1)}_{i,j}\}_{i\in[N],j\in[M]}$, respectively, with missing entries denoted by ?. Then, we can write

$$Y^{(0),\text{obs}} = (\Theta^{(0)} + E^{(0)}) \otimes (\mathbf{1} - A) \quad \text{and} \quad Y^{(1),\text{obs}} = (\Theta^{(1)} + E^{(1)}) \otimes A. \tag{4.7}$$

- **Treatments**: From Eq. (4.3), we can write

$$A = (P + W).$$

Building on the earlier discussion, the application of matrix completion yields the following estimates:

$$\widehat{\Theta}^{(0)} = \texttt{MC}(Y^{(0),\text{obs}}), \quad \widehat{\Theta}^{(1)} = \texttt{MC}(Y^{(1),\text{obs}}), \quad \text{and} \quad \widehat{P} = \texttt{MC}(A), \tag{4.8}$$

where the algorithm MC may vary for $\widehat{\Theta}^{(0)}$, $\widehat{\Theta}^{(1)}$, and $\widehat{P}$. Because all entries of A are observed, MC(A) denoises A but does not need to impute missing entries. From Eq. (4.7) and Eq. (4.8), it follows that $\widehat{\Theta}^{(0)}$ and $\widehat{\Theta}^{(1)}$ depend on A and Y, whereas $\widehat{P}$ depends only on A.

In this section, we deliberately leave the matrix completion algorithm MC as a "black-box". In Section 4.4, we establish finite-sample and asymptotic guarantees for our proposed estimator, contingent on specific properties for MC. In Section 4.5, we propose a novel end-to-end matrix completion algorithm that satifies these properties.

Given matrix completion estimates of $(\widehat{\Theta}^{(0)}, \widehat{\Theta}^{(1)}, \widehat{P})$, we formulate two preliminary estimators for $\text{ATE}_{\cdot,j}$: (i) an outcome imputation estimator, which uses $\widehat{\Theta}^{(0)}$ and $\widehat{\Theta}^{(1)}$ only, and (ii) an inverse probability weighting estimator, which uses $\widehat{P}$ only. Then, we combine these to obtain a doubly-robust estimator of $\text{ATE}_{\cdot,j}$.

Outcome imputation (OI) estimator. Let $\widehat{\theta}^{(a)}_{i,j}$ denote the (i,j)-th entry of $\widehat{\Theta}^{(a)}$ for $i \in [N], j \in [M]$, and $a \in \{0,1\}$. The OI estimator for $\text{ATE}_{\cdot,j}$ is defined as follows:

$$\widehat{\text{ATE}}^{\text{OI}}_{\cdot,j} \triangleq \widehat{\mu}^{(1,\text{OI})}_{\cdot,j} - \widehat{\mu}^{(0,\text{OI})}_{\cdot,j}, \tag{4.9}$$

where

$$\widehat{\mu}^{(a,\text{OI})}_{\cdot,j} \triangleq \frac{1}{N} \sum_{i \in [N]} \widehat{\theta}^{(a)}_{i,j} \quad \text{for} \quad a \in \{0,1\}.$$

That is, the OI estimator is obtained by taking the difference of the average value of the j-th column of the estimates $\widehat{\Theta}^{(0)}$ and $\widehat{\Theta}^{(1)}$. The quality of the OI estimator depends on how well $\widehat{\Theta}^{(0)}$ and $\widehat{\Theta}^{(1)}$ approximate the mean potential outcome matrices $\Theta^{(0)}$ and $\Theta^{(1)}$, respectively.

Inverse probability weighting (IPW) estimator. Let $\widehat{p}_{i,j}$ denote the (i,j)-th entry of $\widehat{P}$ for $i \in [N]$ and $j \in [M]$. The IPW estimate for $\text{ATE}_{\cdot,j}$ is defined as follows:

$$\widehat{\text{ATE}}^{\text{IPW}}_{\cdot,j} \triangleq \widehat{\mu}^{(1,\text{IPW})}_{\cdot,j} - \widehat{\mu}^{(0,\text{IPW})}_{\cdot,j}, \tag{4.10}$$

where

$$\widehat{\mu}^{(0,\text{IPW})}_{\cdot,j} \triangleq \frac{1}{N} \sum_{i \in [N]} \frac{y_{i,j}\big(1 - a_{i,j}\big)}{1 - \widehat{p}_{i,j}} \quad \text{and} \quad \widehat{\mu}^{(1,\text{IPW})}_{\cdot,j} \triangleq \frac{1}{N} \sum_{i \in [N]} \frac{y_{i,j} a_{i,j}}{\widehat{p}_{i,j}}.$$

That is, the IPW estimator is obtained by taking the difference of the average value of the j-th column of the matrices $Y^{(0),\mathrm{obs}}$ and $Y^{(1),\mathrm{obs}}$, replacing unobserved entries with zeros, and weighting each outcome by the inverse of the estimated assignment probability to account for confounding. The quality of the IPW estimate depends on how well $\widehat{P}$ approximates the probability matrix P.

The matrix completion-based OI and IPW estimators in Eq. (4.9) and Eq. (4.10) have the same form as the classical OI and IPW estimators, which are derived for settings where all confounders are observed (e.g., Imbens and Rubin, 2015a). In contrast to the classical setting, our framework is one with unmeasured confounding.

4.3.3 Doubly-robust (DR) estimator

The DR estimator of $\mathrm{ATE}_{\cdot,j}$ combines the estimates $\widehat{\Theta}^{(0)}, \widehat{\Theta}^{(1)}$, and $\widehat{P}$ from Eq. (4.8). It is defined as follows:

$$\widehat{\mathrm{ATE}}_{\cdot,j}^{\mathrm{DR}} \triangleq \widehat{\mu}_{\cdot,j}^{(1,\mathrm{DR})} - \widehat{\mu}_{\cdot,j}^{(0,\mathrm{DR})}, \tag{4.11}$$

where

$$\widehat{\mu}_{\cdot,j}^{(0,\mathrm{DR})} \triangleq \frac{1}{N}\sum_{i\in[N]} \widehat{\theta}_{i,j}^{(0,\mathrm{DR})} \quad \text{with} \quad \widehat{\theta}_{i,j}^{(0,\mathrm{DR})} \triangleq \widehat{\theta}_{i,j}^{(0)} + \big(y_{i,j} - \widehat{\theta}_{i,j}^{(0)}\big)\frac{1-a_{i,j}}{1-\widehat{p}_{i,j}},$$

and

$$\widehat{\mu}_{\cdot,j}^{(1,\mathrm{DR})} \triangleq \frac{1}{N}\sum_{i\in[N]} \widehat{\theta}_{i,j}^{(1,\mathrm{DR})} \quad \text{with} \quad \widehat{\theta}_{i,j}^{(1,\mathrm{DR})} \triangleq \widehat{\theta}_{i,j}^{(1)} + \big(y_{i,j} - \widehat{\theta}_{i,j}^{(1)}\big)\frac{a_{i,j}}{\widehat{p}_{i,j}}. \tag{4.12}$$

In Section 4.4, we prove that $\widehat{\mathrm{ATE}}_{\cdot,j}^{\mathrm{DR}}$ consistently estimates $\mathrm{ATE}_{\cdot,j}$ as long as either $(\widehat{\Theta}^{(0)}, \widehat{\Theta}^{(1)})$ is consistent for $(\Theta^{(0)}, \Theta^{(1)})$ or $\widehat{P}$ is consistent for P, i.e., it is doubly-robust. Furthermore, we show that the DR estimator provides superior finite sample guarantees than the OI and IPW estimators, and that it satisfies a central limit theorem at a parametric rate under weak conditions on the convergence rate of the matrix completion routine. Using simulated data, Figure 4.3.2 demonstrates the improved performance of DR, relative to OI and IPW. Despite substantial biases observed in both OI and IPW estimates, the error of the DR estimate closely follows a mean-zero Gaussian distribution. We provide a detailed description of the simulation setup in Section 4.6.

4.4 Analysis and Main Results

This section presents the formal results of this chapter. Section 4.4.1 details assumptions, Section 4.4.2 discusses finite-sample guarantees, and Section 4.4.3 presents a central limit theorem for $\widehat{\mathrm{ATE}}_{\cdot,j}^{\mathrm{DR}}$.

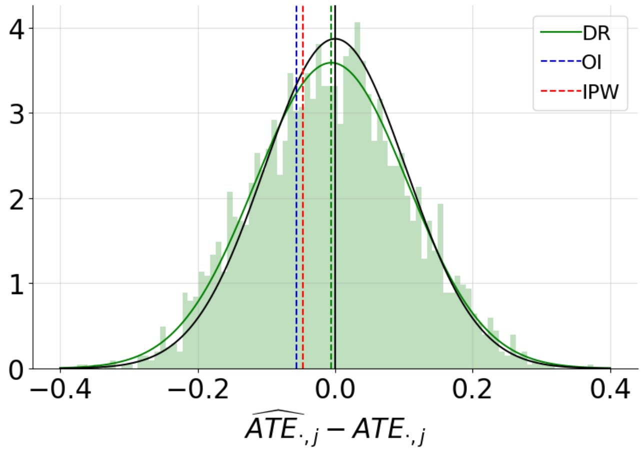


Figure 4.3.2: Simulation evidence of the convergence of the error of the doubly-robust (DR) estimator to a mean-zero Gaussian distribution. The histogram represents $\widehat{\mathrm{ATE}}^{\mathrm{DR}}_{\cdot,j} - \mathrm{ATE}_{\cdot,j}$, the green curve represents the (best) fitted Gaussian distribution, and the black curve represents the Gaussian approximation from Theorem 4.2 in Section 4.4. Histogram counts are normalized so that the area under the histogram integrates to one. Unlike DR, the outcome imputation (OI) and inverse probability weighting (IPW) estimators have non-trivial biases, as evidenced by the means of the distributions in dashed green, blue, and red, respectively. Section 4.6 reports complete simulation results.

4.4.1 Assumptions

Requirements on data generating process. We make two assumptions on how the data is generated. First, we impose a positivity condition on the assignment probabilities.

Assumption 4.1 (Positivity on true assignment probabilities)**.** *The unknown assignment probability matrix P is such that*

$$\lambda \leq p_{i,j} \leq 1 - \lambda, \tag{4.13}$$

for all $i \in [N]$ and $j \in [M]$, where $0 < \lambda \leq 1/2$.

Assumption 4.1 requires that the propensity score for each unit-outcome pair is bounded away from 0 and 1, implying that any unit-item pair can be assigned either of the two treatments. An analogous assumption is pervasive in causal inference models with no-unmeasured confounding. For simplicity of exposition and to avoid notational clutter, Assumption 4.1 requires Eq. (4.13) for all outcomes, $j \in [M]$. In practical applications, however, $\mathrm{ATE}_{\cdot,j}$ may be estimated for a select group of those outcomes. In that case, the positivity assumption applies only for the selected subset of outcomes for which $\mathrm{ATE}_{\cdot,j}$ is estimated.

Next, we formalize the requirements on the noise variables.

Assumption 4.2 (Zero-mean, independent, and subGaussian noise)**.** *Fix any $j \in [M]$. Then,* *(a)* $\{(\varepsilon_{i,j}^{(0)}, \varepsilon_{i,j}^{(1)}, \eta_{i,j}) : i \in [N]\}$ *are mean zero and independent (across i);*

(b) for every $i \in [N]$ and $j \in [M]$, $(\varepsilon_{i,j}^{(0)}, \varepsilon_{i,j}^{(1)}) \perp\!\!\!\perp \eta_{i,j}$; moreover, the distribution of $(\varepsilon_{i,j}^{(0)}, \varepsilon_{i,j}^{(1)})$ depends on $(\Theta^{(0)}, \Theta^{(1)}, P)$ only through $(\theta_{i,j}^{(0)}, \theta_{i,j}^{(1)})$, and the distribution of $\eta_{i,j}$ depends on $(\Theta^{(0)}, \Theta^{(1)}, P)$ only through $p_{i,j}$; and

(c) $\varepsilon_{i,j}^{(a)}$ has subGaussian norm bounded by a constant $\overline{\sigma}$ for every $i \in [N]$ and $a \in \{0, 1\}$.

Assumption 4.2(a) defines $(\Theta^{(0)}, \Theta^{(1)}, P)$ as matrices collecting the means of the potential outcomes and treatment assignments in Eqs. (4.2) and (4.3). Further, for every measurement, it imposes independence across units in the noise variables. Assumption 4.2(b) imposes independence between the noise in the potential outcomes and noise in treatment assignment, and implies that for each particular unit i and measurement j, confounding emerges only from the interplay between $(\theta_{i,j}^{(0)}, \theta_{i,j}^{(1)})$ and $p_{i,j}$. Finally, Assumption 4.2(c) is mild and useful to derive finite-sample guarantees. For the central limit theorem in Section 4.4.3, subGaussianity could be disposed of by restricting the moments of $\varepsilon_{i,j}^{(a)}$. Assumption 4.2 does not restrict the dependence between $\varepsilon_{i,j}^{(0)}$ and $\varepsilon_{i,j}^{(1)}$. Neither Assumption 4.2 restricts the dependence of $\eta_{i,j}$ across outcomes. In particular, Assumption 4.2 allows for the existence of pairs of outcomes (j, j') such that $\mathbb{E}[\eta_{i,j}^2] = \mathbb{E}[\eta_{i,j'}^2] = \mathbb{E}[\eta_{i,j}\eta_{i,j'}]$, in which case $a_{i,j} = a_{i,j'}$ with probability one.

Requirements on matrix completion estimators. First, we assume the estimate $\widehat{P}$ is consistent with Assumption 4.1.

Assumption 4.3 (Positivity on estimated assignment probabilities)**.** *The estimated probability matrix $\widehat{P}$ is such that*

$$\bar{\lambda} \leq \widehat{p}_{i,j} \leq 1 - \bar{\lambda},$$

for all $i \in [N]$ and $j \in [M]$, where $0 < \bar{\lambda} \leq \lambda$.

Assumption 4.3 holds when the entries of $\widehat{P}$ are truncated to the range $[\bar{\lambda}, 1 - \bar{\lambda}]$, provided $\bar{\lambda}$ is not greater than λ. Second, our theoretical analysis requires independence between certain elements of the estimates $(\widehat{P}, \widehat{\Theta}^{(0)}, \widehat{\Theta}^{(1)})$ from Eq. (4.8), and the noise matrices $(W, E^{(0)}, E^{(1)})$. We formally state this independence condition as an assumption below.

Assumption 4.4 (Independence between estimates and noise)**.** *Fix any $j \in [M]$. There exists a non-empty partition $(\mathcal{R}_0, \mathcal{R}_1)$ of the units $[N]$ such that*

$$\big\{\big(\widehat{p}_{i,j}, \widehat{\theta}_{i,j}^{(a)}\big)\big\}_{i \in \mathcal{R}_s} \perp\!\!\!\perp \big\{\eta_{i,j}\big\}_{i \in \mathcal{R}_s} \tag{4.14}$$

and

$$\big\{\widehat{p}_{i,j}\big\}_{i \in \mathcal{R}_s} \perp\!\!\!\perp \big\{\big(\eta_{i,j}, \varepsilon_{i,j}^{(a)}\big)\big\}_{i \in \mathcal{R}_s}, \tag{4.15}$$

for every $a \in \{0, 1\}$ and $s \in \{0, 1\}$.

Eq. (4.14) requires that within each of the two partitions of the units, estimated mean potential outcomes and estimated assignment probabilities are jointly independent of the error in assignment probabilities, for every measurement. Similarly, Eq. (4.15) requires that within each of the two partitions of the units, estimated assignment probabilities are independent jointly of the noise in assignment probabilities and potential outcomes, for every measurement. Conditions like Eq. (4.14) and Eq. (4.15) are familiar in the doubly-robust estimation literature. Chernozhukov et al. (2018) employ a cross-fitting device to enforce an assumption similar to Assumption 4.4 in a context with no unmeasured confounders. Section 4.5 provides a novel cross-fitting procedure for matrix estimation under which Assumption 4.4 holds for any `MC` algorithm (under additional assumptions on the noise variables).

Matrix completion error rates. The formal guarantees in this section depend on the normalized $(1,2)$-norms of the errors in estimating the unknown parameters $(\Theta^{(0)}, \Theta^{(1)}, P)$. We use the following notation for these errors:

$$\mathcal{E}\big(\widehat{P}\big) \triangleq \frac{\|\widehat{P}-P\|_{1,2}}{\sqrt{N}} \quad \text{and} \quad \mathcal{E}\big(\widehat{\Theta}\big) \triangleq \sum_{a\in\{0,1\}} \mathcal{E}\big(\widehat{\Theta}^{(a)}\big), \quad \text{with} \quad \mathcal{E}\big(\widehat{\Theta}^{(a)}\big) \triangleq \frac{\|\widehat{\Theta}^{(a)}-\Theta^{(a)}\|_{1,2}}{\sqrt{N}}. \tag{4.16}$$

A variety of matrix completion algorithms deliver $\mathcal{E}\big(\widehat{P}\big) = O_p(\min\{N,M\}^{-\alpha})$ and $\mathcal{E}\big(\widehat{\Theta}\big) = O_p(\min\{N,M\}^{-\beta})$, where $0 < \alpha, \beta \leq 1/2$. The conditions in this section track dependence on N only. We say that the normalized errors $\mathcal{E}\big(\widehat{P}\big)$ and $\mathcal{E}\big(\widehat{\Theta}\big)$ achieve the parametric rate when they have the same rate as $O_p(N^{-1/2})$. Section 4.5 explicitly characterizes how the rates of convergence $\mathcal{E}\big(\widehat{P}\big)$ and $\mathcal{E}\big(\widehat{\Theta}\big)$ depend on N and M for a particular matrix completion algorithm based on Bai and Ng (2021).

4.4.2 Non-asymptotic guarantees

The first main result of this section provides a non-asymptotic error bound for $\widehat{\mathrm{ATE}}^{\mathrm{DR}}_{\cdot,j} - \mathrm{ATE}_{\cdot,j}$ in terms of the errors $\mathcal{E}\big(\widehat{P}\big)$ and $\mathcal{E}\big(\widehat{\Theta}\big)$ defined in Eq. (4.16).

Theorem 4.1 (Finite Sample Guarantees for DR)**.** *Suppose Assumptions 4.1 to 4.4 hold. Fix $\delta \in (0,1)$ and $j \in [M]$. Then, with probability at least $1-\delta$, we have*

$$\big|\widehat{\mathrm{ATE}}^{\mathrm{DR}}_{\cdot,j} - \mathrm{ATE}_{\cdot,j}\big| \leq \mathrm{Err}^{\mathrm{DR}}_{N,\delta}, \tag{4.17}$$

where

$$\mathrm{Err}^{\mathrm{DR}}_{N,\delta} \triangleq \frac{2}{\bar{\lambda}}\bigg[\mathcal{E}\big(\widehat{\Theta}\big)\mathcal{E}\big(\widehat{P}\big) + \Big(\frac{\sqrt{c\ell_{\delta/12}}}{\sqrt{\ell_1}}\mathcal{E}\big(\widehat{\Theta}\big) + 2\bar{\sigma}\sqrt{c\ell_{\delta/12}} + \frac{2\bar{\sigma}m(c\ell_{\delta/12})}{\sqrt{\ell_1}}\Big)\frac{1}{\sqrt{N}}\bigg], \tag{4.18}$$

for $m(c)$ and ℓ_c as defined in Section 4.1.

The proof of Theorem 4.1 is given in Appendix 4.B. Eqs. (4.17) and (4.18) bound the absolute error of the DR estimator by the rate of $\mathcal{E}\big(\widehat{\Theta}\big)(\mathcal{E}\big(\widehat{P}\big) + N^{-0.5}) + N^{-0.5}$.

When $\mathcal{E}(\widehat{P})$ is lower bounded at the parametric rate of $N^{-0.5}$, $\mathrm{Err}^{\mathrm{DR}}_{N,\delta}$ has the same rate as $\mathcal{E}(\widehat{P})\mathcal{E}(\widehat{\Theta}) + N^{-0.5}$.

Doubly-robust behavior of $\widehat{\mathrm{ATE}}^{\mathrm{DR}}_{\cdot,j}$. The error rate of $\mathcal{E}(\widehat{P})\mathcal{E}(\widehat{\Theta}) + N^{-0.5}$ immediately reveals that the DR estimate is doubly-robust with respect to the error in estimating the mean potential outcomes $(\Theta^{(0)}, \Theta^{(1)})$ and the assignment probabilities P. First, the error $\mathrm{Err}^{\mathrm{DR}}_{N,\delta}$ decays at a parametric rate of $O_p(N^{-0.5})$ as long as the product of error rates, $\mathcal{E}(\widehat{P})\mathcal{E}(\widehat{\Theta})$, decays as $O_p(N^{-0.5})$. As a result, $\widehat{\mathrm{ATE}}^{\mathrm{DR}}_{\cdot,j}$ can exhibit a parametric error rate even when neither the mean potential outcomes nor the assignment probabilities are estimated at a parametric rate. Second, $\mathrm{Err}^{\mathrm{DR}}_{N,\delta}$ decays to zero as long as either of $\mathcal{E}(\widehat{P})$ or $\mathcal{E}(\widehat{\Theta})$ decays to zero, provided both errors are $O_p(1)$.

We next compare the performance of DR estimator with the OI and IPW estimators from Eqs. (4.9) and (4.10), respectively. Towards this goal, we characterize the $\mathrm{ATE}_{\cdot,j}$ estimation error of $\widehat{\mathrm{ATE}}^{\mathrm{OI}}_{\cdot,j}$ in terms of $\mathcal{E}(\widehat{\Theta})$ and of $\widehat{\mathrm{ATE}}^{\mathrm{IPW}}_{\cdot,j}$ in terms of $\mathcal{E}(\widehat{P})$.

Proposition 4.1 (Finite Sample Guarantees for OI and IPW)**.** *Fix any $j \in [M]$. For OI, we have*

$$\left|\widehat{\mathrm{ATE}}^{\mathrm{OI}}_{\cdot,j} - \mathrm{ATE}_{\cdot,j}\right| \le \mathrm{Err}^{\mathrm{OI}}_{N} \triangleq \mathcal{E}(\widehat{\Theta}). \tag{4.19}$$

For IPW, suppose Assumptions 4.1 to 4.4 hold. Define $\theta_{\max} \triangleq \sum_{a\in\{0,1\}} \|\Theta^{(a)}\|_{\max}$, and fix any $\delta \in (0,1)$. Then, with probability at least $1-\delta$, we have

$$\left|\widehat{\mathrm{ATE}}^{\mathrm{IPW}}_{\cdot,j} - \mathrm{ATE}_{\cdot,j}\right| \le \mathrm{Err}^{\mathrm{IPW}}_{N,\delta}, \tag{4.20}$$

where

$$\mathrm{Err}^{\mathrm{IPW}}_{N,\delta} \triangleq \frac{2}{\bar{\lambda}}\left[\theta_{\max}\,\mathcal{E}(\widehat{P}) + \Big(\frac{\sqrt{c\ell_{\delta/12}}}{\sqrt{\ell_1}}\theta_{\max} + 2\bar{\sigma}\sqrt{c\ell_{\delta/12}} + \frac{2\bar{\sigma}m(c\ell_{\delta/12})}{\sqrt{\ell_1}}\Big)\frac{1}{\sqrt{N}}\right],$$

for $m(c)$ and ℓ_c as defined in Section 4.1.

The proofs of Eq. (4.19) and Eq. (4.20) are given in Appendices 4.D and 4.E, respectively. Proposition 4.1 implies that in an asymptotic sequence with bounded $\theta_{\max}$, OI and IPW attain the parametric rate $O_p(N^{-0.5})$ provided $\mathcal{E}(\widehat{\Theta})$ and $\mathcal{E}(\widehat{P})$ are $O_p(N^{-0.5})$, respectively. The next corollary, proven in Appendix 4.C, compares these error rates with those obtained for the DR estimator in Theorem 4.1.

Corollary 4.1 (Gains of DR over OI and IPW)**.** *Suppose Assumptions 4.1 to 4.4 hold. Fix any $j \in [M]$. Consider an asymptotic sequence such that Suppose $\theta_{\max}$ is bounded. If $\mathcal{E}(\widehat{P}) = O_p(N^{-\alpha})$ and $\mathcal{E}(\widehat{\Theta}) = O_p(N^{-\beta})$ for $0 \le \alpha \le 0.5$ and $0 \le \beta \le 0.5$, then*

$$\left|\widehat{\mathrm{ATE}}^{\mathrm{OI}}_{\cdot,j} - \mathrm{ATE}_{\cdot,j}\right| = O_p(N^{-\beta}), \qquad \left|\widehat{\mathrm{ATE}}^{\mathrm{IPW}}_{\cdot,j} - \mathrm{ATE}_{\cdot,j}\right| = O_p(N^{-\alpha}),$$

and

$$\left|\widehat{\mathrm{ATE}}^{\mathrm{DR}}_{\cdot,j} - \mathrm{ATE}_{\cdot,j}\right| = O_p(N^{-\min\{\alpha+\beta,0.5\}}).$$

Corollary 4.1 shows that the DR estimate's error decay rate is consistently superior to that of the OI and IPW estimates across a variety of regimes for α, β. Specifically, the error $\mathrm{Err}_{N,\delta}^{\mathrm{DR}}$ scales strictly faster than both $\mathrm{Err}_{N}^{\mathrm{OI}}$ and $\mathrm{Err}_{N,\delta}^{\mathrm{IPW}}$ if the estimation errors of $\widehat{\Theta}^{(0)}$, $\widehat{\Theta}^{(1)}$, and $\widehat{P}$ converge slower than at the parametric rate $O_p(N^{-1/2})$. When the estimation errors of $\widehat{\Theta}^{(0)}$, $\widehat{\Theta}^{(1)}$, and $\widehat{P}$ all decay at a parametric rate, OI, IPW, and DR estimation errors decay also at a parametric rate.

4.4.3 Asymptotic guarantees

The next result, proven in Appendix 4.C as a corollary of Theorem 4.1, provides conditions on $\mathcal{E}\big(\widehat{P}\big)$ and $\mathcal{E}\big(\widehat{\Theta}\big)$ for consistency of $\widehat{\mathrm{ATE}}_{\cdot,j}^{\mathrm{DR}}$.

Corollary 4.2 (Consistency for DR)**.** *Suppose Assumptions 4.1 to 4.4 hold. As $N \to \infty$, if either (i) $\mathcal{E}\big(\widehat{P}\big) = o_p(1)$, $\mathcal{E}\big(\widehat{\Theta}\big) = O_p(1)$, or (ii) $\mathcal{E}\big(\widehat{\Theta}\big) = o_p(1)$, $\mathcal{E}\big(\widehat{P}\big) = O_p(1)$, it holds that*

$$\widehat{\mathrm{ATE}}_{\cdot,j}^{\mathrm{DR}} - \mathrm{ATE}_{\cdot,j} \xrightarrow{p} 0, \tag{4.21}$$

for all $j \in [M]$.

Corollary 4.2 states that $\widehat{\mathrm{ATE}}_{\cdot,j}^{\mathrm{DR}}$ is a consistent estimator for $\mathrm{ATE}_{\cdot,j}$ as long as either the mean potential outcomes or the assignment probabilities are estimated consistently.

The next theorem, proven in Appendix 4.F, establishes a Gaussian approximation for $\widehat{\mathrm{ATE}}_{\cdot,j}^{\mathrm{DR}}$ under mild conditions on error rates $\mathcal{E}\big(\widehat{P}\big)$ and $\mathcal{E}\big(\widehat{\Theta}\big)$.

Theorem 4.2 (Asymptotic Normality for DR)**.** *Suppose Assumptions 4.1 to 4.4 and the following conditions hold,*

(C1) $\mathcal{E}\big(\widehat{P}\big) = o_p(1)$ and $\mathcal{E}\big(\widehat{\Theta}\big) = o_p(1)$.

(C2) $\mathcal{E}\big(\widehat{P}\big)\mathcal{E}\big(\widehat{\Theta}\big) = o_p\big(N^{-1/2}\big)$.

(C3) For every $i \in [N]$ and $j \in [M]$, let $\sigma_{i,j}^{(0)}$ and $\sigma_{i,j}^{(1)}$ be the standard deviations of $\varepsilon_{i,j}^{(0)}$ and $\varepsilon_{i,j}^{(1)}$, respectively. The sequence

$$\overline{\sigma}_j^2 \triangleq \frac{1}{N}\sum_{i\in[N]} \frac{(\sigma_{i,j}^{(1)})^2}{p_{i,j}} + \frac{1}{N}\sum_{i\in[N]} \frac{(\sigma_{i,j}^{(0)})^2}{1-p_{i,j}}, \tag{4.22}$$

is bounded away from zero as N increases.

Then, for all $j \in [M]$,

$$\sqrt{N}\big(\widehat{\mathrm{ATE}}_{\cdot,j}^{\mathrm{DR}} - \mathrm{ATE}_{\cdot,j}\big)/\overline{\sigma}_j \xrightarrow{d} \mathcal{N}\big(0,1\big), \tag{4.23}$$

as $N \to \infty$.

Theorem 4.2 describes two simple requirements on the estimated matrices $\widehat{P}$ and $(\widehat{\Theta}^{(0)}, \widehat{\Theta}^{(1)})$, under which $\widehat{\mathrm{ATE}}^{\mathrm{DR}}_{\cdot,j}$ exhibits an asymptotic Gaussian distribution centered at $\mathrm{ATE}_{\cdot,j}$. Condition (C1) requires that the estimation errors of $\widehat{P}$ and $(\widehat{\Theta}^{(0)}, \widehat{\Theta}^{(1)})$ converge to zero in probability. Condition (C2) requires that the product of the errors decays sufficiently fast, at a rate $o_p\big(N^{-1/2}\big)$, ensuring that the bias of the normalized estimator in Eq. (4.23) converges to zero. Condition (C2) is similar to conditions in the literature on doubly-robust estimation of average treatment effects under observed confounding (e.g., Assumption 5.1 in Chernozhukov et al., 2018). Specifically, in that context, Chernozhukov et al. (2018) assume that the product of propensity estimation error and outcome regression error decays faster than $N^{-1/2}$.

Black-box asymptotic normality. We emphasize that Theorem 4.2 applies to any matrix completion algorithm `MC`, provided conditions (C1) and (C2) hold. This level of generality is useful because the product of $\mathcal{E}\big(\widehat{P}\big)$ and $\mathcal{E}\big(\widehat{\Theta}\big)$ is $o_p\big(N^{-1/2}\big)$ for a wide range of `MC` algorithms, under mild assumptions on $(\Theta^{(0)}, \Theta^{(1)}, P)$. In contrast, achieving such black-box asymptotic normality for OI or IPW estimates is challenging. Their biases are tied to the individual error rates, $\mathcal{E}\big(\widehat{\Theta}\big)$ and $\mathcal{E}\big(\widehat{P}\big)$, which are typically lower-bouded at the parametric rate of $N^{-0.5}$.

The next result, proven in Appendix 4.F.3, provides a consistent estimator for the asymptotic variance $\overline{\sigma}_j^2$ from Theorem 4.2.

Proposition 4.2 (Consistent variance estimation)**.** *Suppose Assumptions 4.1 to 4.3 and condition (C1) in Theorem 4.2 holds. Suppose the partition* $(\mathcal{R}_0, \mathcal{R}_1)$ *of the units* $[N]$ *from Assumption 4.4 is such that*

$$\{\big(\widehat{p}_{i,j}, \widehat{\theta}^{(a)}_{i,j}\big)\}_{i\in\mathcal{R}_s} \perp\!\!\!\perp \{(\eta_{i,j}, \varepsilon^{(a)}_{i,j})\}_{i\in\mathcal{R}_s}, \tag{4.24}$$

for every $j \in [M]$, $a \in \{0,1\}$ *and* $s \in \{0,1\}$. *Then, for all* $j \in [M]$, $\widehat{\sigma}_j^2 - \overline{\sigma}_j^2 \xrightarrow{p} 0$, *where*

$$\widehat{\sigma}_j^2 \triangleq \frac{1}{N}\sum_{i\in[N]} \frac{\big(y_{i,j} - \widehat{\theta}^{(1)}_{i,j}\big)^2 a_{i,j}}{\big(\widehat{p}_{i,j}\big)^2} + \frac{1}{N}\sum_{i\in[N]} \frac{\big(y_{i,j} - \widehat{\theta}^{(0)}_{i,j}\big)^2 (1-a_{i,j})}{\big(1-\widehat{p}_{i,j}\big)^2}. \tag{4.25}$$

4.4.4 Application to panel data with lagged treatment effects

Sections 4.4.2 and 4.4.3 considered a model where the outcome $y_{i,j}$ for unit $i \in [N]$ and measurement $j \in [M]$ depends on treatment assignment only for unit i and measurement j, i.e., $a_{i,j}$. The Appendix 4.I discusses how to extend the results of this section to a setting of panel data with lagged treatment effects. In a panel data setting, the M measurements correspond to T time periods, and t denotes the time index. Then, Appendix 4.I considers an auto-regressive setting, where the potential outcomes at time t depends on the treatment assignment at time t and the realized outcome at time $t-1$, i.e., for all $i \in [N], t \in [T]$, and $a \in \{0,1\}$,

$$y_{i,t}^{(a|y_{i,t-1})} = \alpha^{(a)} y_{i,t-1} + \theta^{(a)}_{i,t} + \varepsilon^{(a)}_{i,t},$$

and observed outcomes satisfy

$$y_{i,t} = y_{i,t}^{(0|y_{i,t-1})}(1 - a_{i,t}) + y_{i,t}^{(1|y_{i,t-1})} a_{i,t}.$$

The presence of lagged treatment effects in this model makes it crucial to define causal estimands for entire sequences of treatments. The supplementary appendix describes how the proposed doubly-robust estimation can be extended to treatment sequences and derives a generalization of Theorem 4.1.

4.5 Matrix Completion with Cross-Fitting

In this section, we introduce a novel algorithm designed to construct estimates $(\widehat{\Theta}^{(0)}, \widehat{\Theta}^{(1)}, \widehat{P})$ that adhere to Assumption 4.4 and satisfy conditions (C1) and (C2) in Theorem 4.2. We first explain why traditional matrix completion algorithms fail to deliver the properties required by Assumption 4.4. We then present `Cross-Fitted-MC`, a meta-algorithm that takes any matrix completion algorithm and uses it to construct $(\widehat{\Theta}^{(0)}, \widehat{\Theta}^{(1)}, \widehat{P})$ that satisfy Assumption 4.4, and the stronger independence condition in Proposition 4.2. Finally, we describe `Cross-Fitted-SVD`, an end-to-end algorithm obtained by combining `Cross-Fitted-MC` with the singular value decomposition (`SVD`)-based algorithm of Bai and Ng (2021), and establish that it also satisfies conditions (C1) and (C2) in Theorem 4.2.

Traditional matrix completion. Estimates $(\widehat{\Theta}^{(0)}, \widehat{\Theta}^{(1)}, \widehat{P})$ obtained from existing matrix completion algorithms need not satisfy Assumption 4.4. In particular, using the entire assignment matrix A to estimate each element of P typically results in a violation of $\{\widehat{p}_{i,j}\}_{i \in \mathcal{R}_s} \perp\!\!\!\perp \{\eta_{i,j}\}_{i \in \mathcal{R}_s}$ in Assumption 4.4, as each entry of $\widehat{P}$ is allowed to depend on the entire noise matrix W. For example, in spectral methods (e.g., Nguyen et al., 2019), $\widehat{P}$ is a function of the SVD of the entire matrix A, and

$$\widehat{p}_{i,j} \not\perp\!\!\!\perp a_{i',j'}, \tag{4.26}$$

for all $(i,j), (i',j') \in [N] \times [M]$ in general, which implies $\{\widehat{p}_{i,j}\}_{i \in \mathcal{R}_s} \not\perp\!\!\!\perp \{\eta_{i,j}\}_{i \in \mathcal{R}_s}$, for every $\mathcal{R}_s \subset [N]$. Similarly, in matching methods such as nearest neighbors (Li et al., 2019), $\widehat{P}$ is a function of the matches/neighbors estimated from the entire matrix A. Dependence structures such as $\widehat{p}_{i,j} \not\perp\!\!\!\perp a_{i,j}$ for any $i, j \in [N] \times [M]$—which is weaker than Eq. (4.26)—are enough to violate the $\{\widehat{p}_{i,j}\}_{i \in \mathcal{R}_s} \perp\!\!\!\perp \{\eta_{i,j}\}_{i \in \mathcal{R}_s}$ requirement in Assumption 4.4. Likewise, the requirement $\{\widehat{\theta}_{i,j}^{(a)}\}_{i \in \mathcal{R}_s} \perp\!\!\!\perp \{\eta_{i,j}\}_{i \in \mathcal{R}_s}$ in Assumption 4.4 can be violated, because $\widehat{\Theta}^{(0)}$ and $\widehat{\Theta}^{(1)}$ depend respectively on $Y^{(0),\text{obs}}$ and $Y^{(1),\text{obs}}$, which themselves depend on the entire matrix A.

4.5.1 `Cross-Fitted-MC`: A meta-cross-fitting algorithm for matrix completion

We now introduce `Cross-Fitted-MC`, a cross-fitting procedure that modifies any `MC` algorithm to produce $(\widehat{\Theta}^{(0)}, \widehat{\Theta}^{(1)}, \widehat{P})$ that satisfy Assumption 4.4. We employ the following assumption on the noise variables.

Assumption 4.5 (Block independence between noise)**.** *Let* $(\mathcal{R}_0, \mathcal{R}_1)$ *denote the partition of the units* $[N]$ *from Assumption 4.4. There exists partitions* $(\mathcal{C}_0, \mathcal{C}_1)$ *of the measurements* $[M]$*, such that for each block* $\mathcal{I} \in \mathcal{P} \triangleq \{\mathcal{R}_s \times \mathcal{C}_k : s, k \in \{0, 1\}\}$*,*

$$W_{\mathcal{I}} \perp\!\!\!\perp W_{-\mathcal{I}}, E_{-\mathcal{I}}^{(a)} \tag{4.27}$$

and

$$W_{-\mathcal{I}} \perp\!\!\!\perp W_{\mathcal{I}}, E_{\mathcal{I}}^{(a)}. \tag{4.28}$$

for every $a \in \{0, 1\}$*.*

For a given block $\mathcal{I}$, Eq. (4.27) requires the noise in the treatment assignments corresponding to $\mathcal{I}$ to be independent jointly of the noise in the treatment assignments and the potential outcomes corresponding to the remaining three blocks. Likewise, Eq. (4.28) requires the noise in the treatment assignments corresponding to the remaining three blocks to be independent jointly of the noise in the treatment assignments and the potential outcomes corresponding to $\mathcal{I}$. Assumption 4.5 leaves unrestricted the dependence of the noise variables across outcomes that belong to the same block.

For notational simplicity, Assumption 4.5 imposes independence conditions across blocks of outcomes in a partition of $[M]$ into two blocks only. It is important to note, however, that the results in this section hold under more general dependence patterns. In particular, at the cost of additional notational complexity, it is straightforward to extend the result in this section to partitions of outcomes $(\mathcal{C}_0, \mathcal{C}_1, \ldots, \mathcal{C}_m)$ such that for each $k \in \{0, 1, \ldots, m\}$, $s \in \{0, 1\}$ and $a \in \{0, 1\}$, there exists $k' \in \{0, 1, \ldots, m\} \setminus \{k\}$ with $\{\eta_{i,j}\}_{(i,j) \in \mathcal{R}_s \times \mathcal{C}_k} \perp\!\!\!\perp \{\eta_{i,j}, \varepsilon_{i,j}^{(a)}\}_{(i,j) \in \mathcal{R}_{1-s} \times \mathcal{C}_{k'}}$ and $\{\eta_{i,j}\}_{(i,j) \in \mathcal{R}_{1-s} \times \mathcal{C}_{k'}} \perp\!\!\!\perp \{\eta_{i,j}, \varepsilon_{i,j}^{(a)}\}_{(i,j) \in \mathcal{R}_s \times \mathcal{C}_k}$. This allows for rather general patterns of dependence across outcomes while preserving independence across specific sets of outcomes (e.g., certain product categories in the retail example of Section 4.1).

Recall the setup from Section 4.3.1: Given an observation matrix $S \in \{\mathbb{R} \cup \{\,?\}\}^{N \times M}$, a matrix completion algorithm $\texttt{MC}$ produces an estimate $\widehat{T} = \texttt{MC}(S) \in \mathbb{R}^{N \times M}$ of a matrix of interest T, where S and T are related via Eq. (4.6). With this background, we now describe the `Cross-Fitted-MC` meta-algorithm.

1. The inputs are (i) a matrix completion algorithm $\texttt{MC}$, (ii) an observation matrix $S \in \{\mathbb{R} \cup \{\,?\}\}^{N \times M}$, and (iii) a block partition $\mathcal{P}$ of the set $[N] \times [M]$ into four blocks as in Assumption 4.5.

2. For each block $\mathcal{I} \in \mathcal{P}$, construct $\widehat{T}_{\mathcal{I}}$ by applying $\texttt{MC}$ on $S \otimes \mathbf{1}^{-\mathcal{I}}$ where $\mathbf{1}^{-\mathcal{I}} \in \mathbb{R}^{N \times M}$ denotes a masking matrix with (i, j)-th entry equal to 0 if $(i, j) \in \mathcal{I}$ and 1 otherwise, and the operator $\otimes$ is as defined in Section 4.1. In other words,

$$\widehat{T}_{\mathcal{I}} = \overline{T}_{\mathcal{I}} \quad \text{where} \quad \overline{T} = \texttt{MC}(S \otimes \mathbf{1}^{-\mathcal{I}}). \tag{4.29}$$

3. Return $\widehat{T} \in \mathbb{R}^{N \times M}$ obtained by collecting together $\{\widehat{T}_{\mathcal{I}}\}_{\mathcal{I} \in \mathcal{P}}$, with each entry in its original position.

We represent this meta-algorithm succinctly as below:

$$\widehat{T} = \texttt{Cross-Fitted-MC}(\texttt{MC}, S, \mathcal{P}).$$

In summary, `Cross-Fitted-MC` produces an estimate $\widehat{T}$ such that for each block $\mathcal{I} \in \mathcal{P}$, the sub-matrix $\widehat{T}_{\mathcal{I}}$ is constructed only using the entries of S corresponding to the remaining three blocks of $\mathcal{P}$. Figure 4.5.1(a) provides a schematic of the block partition $\mathcal{P}$ for $\mathcal{R}_0 = [\lfloor N/2 \rfloor]$ and $\mathcal{C}_0 = [\lfloor M/2 \rfloor]$. See Figure 4.5.1$(b)$ for a visualization of $S \otimes \mathbf{1}^{-\mathcal{I}}$. The following result, proven in Appendix 4.G.1, establishes $(\widehat{\Theta}^{(0)}, \widehat{\Theta}^{(1)}, \widehat{P})$ generated by `Cross-Fitted-MC` satisfy Assumption 4.4.

Proposition 4.3 (Guarantees for `Cross-Fitted-MC`). *Suppose Assumptions 4.2 and 4.5 hold. Let* `MC` *be any matrix completion algorithm and* $\mathcal{P}$ *be the block partition of the set* $[N] \times [M]$ *into four blocks from Assumption 4.5. Let*

$$\widehat{\Theta}^{(0)} = \textit{Cross-Fitted-MC}(\textit{MC}, Y^{(0),\text{obs}}, \mathcal{P}), \tag{4.30}$$

$$\widehat{\Theta}^{(1)} = \textit{Cross-Fitted-MC}(\textit{MC}, Y^{(1),\text{obs}}, \mathcal{P}), \tag{4.31}$$

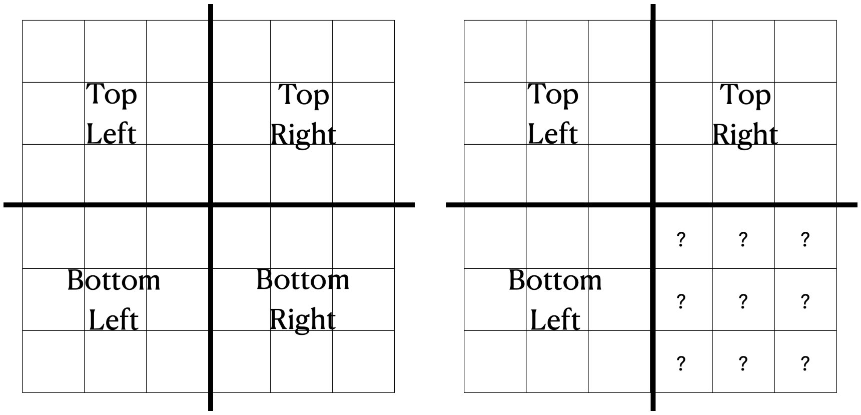
$$\widehat{P} = \textit{Cross-Fitted-MC}(\textit{MC}, A, \mathcal{P}), \tag{4.32}$$

where $Y^{(0),\text{obs}}$ *and* $Y^{(1),\text{obs}}$ *are defined in Eq. (4.7). Then, Assumption 4.4 holds for all* $j \in [M]$*. Further, suppose*

$$W_{\mathcal{I}}, E^{(a)}_{\mathcal{I}} \perp\!\!\!\perp W_{-\mathcal{I}}, E^{(a)}_{-\mathcal{I}}, \tag{4.33}$$

for every block $\mathcal{I} \in \mathcal{P}$ *and* $a \in \{0, 1\}$*. Then, Eq. (4.24) holds too.*

A host of `MC` algorithms are designed to de-noise and impute missing entries of matrices under random patterns of missingness; the most common missingness pattern studied is where each entry has the same probability of being missing, independent of everything else. In contrast, `Cross-Fitted-MC` generates patterns where all entries in one block are deterministically missing, as in Figure 4.5.1(b). A recent strand of research on the interplay between matrix completion methods and causal inference models—specifically, within the synthetic controls framework—has contributed matrix completion algorithms that allow for block missingness (see, e.g., Agarwal et al., 2023, 2020, 2021; Arkhangelsky et al., 2021; Athey et al., 2021; Bai and Ng, 2021; Dwivedi et al., 2022a,b). However, it is a challenge to apply known theoretical guarantees for these methods to the setting in this chapter because of: (i) the use of cross-fitting—which creates blocks where all observations are missing—and (ii) outside of the completely-missing blocks, there can still be missing observations with heterogeneous probabilities of missingness. In the next section, we show how to modify an `MC` algorithm designed for block missingness patterns so that it can be applied to our setting with cross-fitting and heterogeneous probabilities of missingness outside the folds. For concreteness, we work with the Tall-Wide matrix completion algorithm of Bai and Ng (2021).



(a) S (b) $S \otimes \mathbf{1}^{-\text{Bottom Right}}$

Figure 4.5.1: Panel (a): A matrix S partitioned into four blocks when $\mathcal{R}_0 = [N/2]$ and $\mathcal{C}_0 = [M/2]$ in Assumption 4.5, i.e., $\mathcal{P} = \{$Top Left, Top Right, Bottom Left, Bottom Right$\}$. Panel (b): The matrix $S \otimes \mathbf{1}^{-\text{Bottom Right}}$ obtained from the matrix S by masking the entries corresponding to the Bottom Right block with $?$.

4.5.2 The `Cross-Fitted-SVD` algorithm

`Cross-Fitted-SVD` is an end-to-end `MC` algorithm obtained by instantiating the `Cross-Fitted-MC` meta-algorithm with the Tall-Wide algorithm of Bai and Ng (2021), which we denote as `TW`. For completeness, we detail the `TW` algorithm in Section 4.5.2.1, and then use it to describe `Cross-Fitted-SVD` in Section 4.5.2.2.

4.5.2.1 The `TW` algorithm of Bai and Ng (2021).

Bai and Ng (2021) propose `TW` to impute missing values in matrices with a set of rows and a set of columns without missing entries. More concretely, for any matrix $S \in \{\mathbb{R} \cup \{?\}\}^{N \times M}$, let $\mathcal{R}_{\text{obs}} \subseteq [N]$ and $\mathcal{C}_{\text{obs}} \subseteq [M]$ denote the set of all rows and all columns, respectively, with all entries observed. Then, all missing entries of S belong to the block $\mathcal{I} = \mathcal{R}_{\text{miss}} \times \mathcal{C}_{\text{miss}}$, where $\mathcal{R}_{\text{miss}} \triangleq [N] \setminus \mathcal{R}_{\text{obs}}$ and $\mathcal{C}_{\text{miss}} \triangleq [M] \setminus \mathcal{C}_{\text{obs}}$.

Given a rank hyper-parameter $r \in [\min\{|\mathcal{R}_{\text{obs}}|, |\mathcal{C}_{\text{obs}}|\}]$, $\texttt{TW}_r$ produces an estimate of T as follows:

1. Run `SVD` separately on $S^{(\text{tall})} \triangleq S_{[N] \times \mathcal{C}_{\text{obs}}}$ and $S^{(\text{wide})} \triangleq S_{\mathcal{R}_{\text{obs}} \times [M]}$, i.e.,

$$\texttt{SVD}(S^{(\text{tall})}) = (U^{(\text{tall})} \in \mathbb{R}^{N \times \overline{r}_N}, \Sigma^{(\text{tall})} \in \mathbb{R}^{\overline{r}_N \times \overline{r}_N}, V^{(\text{tall})} \in \mathbb{R}^{|\mathcal{C}_{\text{obs}}| \times \overline{r}_N})$$

and

$$\texttt{SVD}(S^{(\text{wide})}) = (U^{(\text{wide})} \in \mathbb{R}^{|\mathcal{R}_{\text{obs}}| \times \overline{r}_M}, \Sigma^{(\text{wide})} \in \mathbb{R}^{\overline{r}_M \times \overline{r}_M}, V^{(\text{wide})} \in \mathbb{R}^{M \times \overline{r}_M})$$

where $\overline{r}_N \triangleq \min\{N, |\mathcal{C}_{\text{obs}}|\}$ and $\overline{r}_M \triangleq \min\{|\mathcal{R}_{\text{obs}}|, M\}$. The columns of $U^{(\text{tall})}$ and $U^{(\text{wide})}$ are the left singular vectors of $S^{(\text{tall})}$ and $S^{(\text{wide})}$, respectively, and the columns of $V^{(\text{tall})}$ and $V^{(\text{wide})}$ are the right singular vectors of $S^{(\text{tall})}$ and $S^{(\text{wide})}$, respectively. The diagonal entries of $\Sigma^{(\text{tall})}$ and $\Sigma^{(\text{wide})}$ are the singular values of $S^{(\text{tall})}$ and $S^{(\text{wide})}$, respectively, and the off-diagonal entries are zeros. This step of

TW requires the existence of the fully observed blocks $S^{(\text{tall})}$ and $S^{(\text{wide})}$, i.e., $\mathcal{R}_{\text{obs}}$ and $\mathcal{C}_{\text{obs}}$ cannot be empty.

2. Let $\widetilde{V}^{(\text{tall})} \in \mathbb{R}^{|\mathcal{C}_{\text{obs}}| \times r}$ be the sub-matrix of $V^{(\text{tall})}$ that keeps the columns corresponding to the r largest singular values only. Let $\widetilde{V}^{(\text{wide})} \in \mathbb{R}^{|\mathcal{C}_{\text{obs}}| \times r}$ be the sub-matrix of $V^{(\text{wide})}$ that keeps the columns corresponding to the r largest singular values only and the rows corresponding to the indices in $\mathcal{C}_{\text{obs}}$ only. Obtain a rotation matrix $R \in \mathbb{R}^{r \times r}$ as follows:

$$R \triangleq \widetilde{V}^{(\text{tall})\top} \widetilde{V}^{(\text{wide})} \big(\widetilde{V}^{(\text{wide})\top} \widetilde{V}^{(\text{wide})}\big)^{-1}.$$

   That is, R is obtained by regressing $\widetilde{V}^{(\text{tall})}$ on $\widetilde{V}^{(\text{wide})}$. In essence, R aligns the right singular vectors of $S^{(\text{tall})}$ and $S^{(\text{wide})}$ using the entries that are common between these two matrices, i.e., the entries corresponding to indices $\mathcal{R}_{\text{obs}} \times \mathcal{C}_{\text{obs}}$. The formal guarantees of the TW algorithm remains unchanged if one alternatively regresses $\widetilde{V}^{(\text{wide})}$ on $\widetilde{V}^{(\text{tall})}$, or uses the left singular vectors of $S^{(\text{tall})}$ and $S^{(\text{wide})}$ for alignment.

3. Let $\overline{\Sigma}^{(\text{tall})} \in \mathbb{R}^{\overline{r}_N \times r}$ be the sub-matrix of $\Sigma^{(\text{tall})}$ that keeps the columns corresponding to the r largest singular values only. Let $\overline{V}^{(\text{wide})} \in \mathbb{R}^{M \times r}$ be the sub-matrix of $V^{(\text{wide})}$ that keeps the columns corresponding to the r largest singular values only. Return $\widehat{T} \triangleq U^{(\text{tall})} \overline{\Sigma}^{(\text{tall})} R \overline{V}^{(\text{wide})\top}$ as an estimate for T.

4.5.2.2 `Cross-Fitted-SVD` algorithm.

1. The inputs are (i) $A \in \mathbb{R}^{N \times M}$, (ii) $Y^{(a),\text{obs}} \in \{\mathbb{R} \cup \{\,?\}\}^{N \times M}$ for $a \in \{0, 1\}$, (iii) a block partition $\mathcal{P}$ of the set $[N] \times [M]$ into four blocks as in Assumption 4.5, and (iv) hyper-parameters r_1, r_2, r_3, and $\bar{\lambda}$ such that $r_1, r_2, r_3 \in [\min\{N, M\}]$ and $0 < \bar{\lambda} \leq 1/2$.

2. Return $\widehat{P} = \texttt{Proj}_{\bar{\lambda}}\big(\texttt{Cross-Fitted-MC}(\texttt{TW}_{r_1}, A, \mathcal{P})\big)$ where $\texttt{Proj}_{\bar{\lambda}}(\cdot)$ projects each entry of its input to the interval $[\bar{\lambda}, 1 - \bar{\lambda}]$.

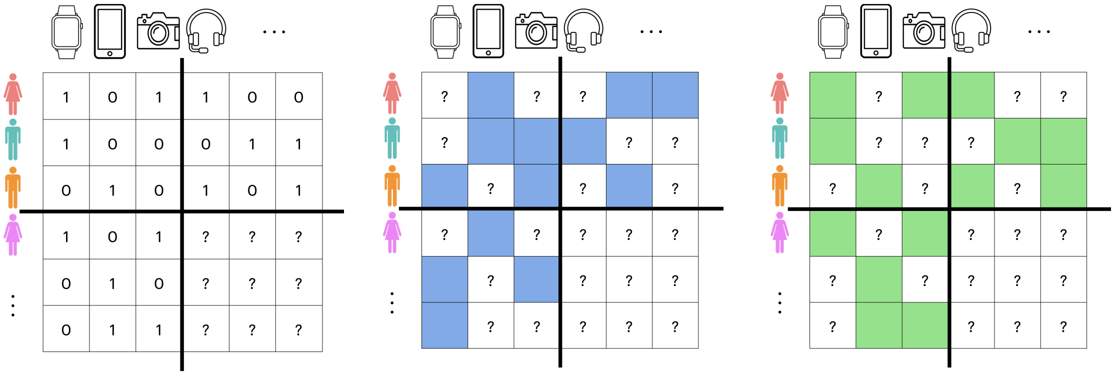
3. Define $Y^{(0),\text{full}}$ as equal to $Y^{(0),\text{obs}}$, but with all missing entries in $Y^{(0),\text{obs}}$ set to zero. Define $Y^{(1),\text{full}}$ analogously with respect to $Y^{(1),\text{obs}}$.

4. Return $\widehat{\Theta}^{(0)} = \texttt{Cross-Fitted-MC}(\texttt{TW}_{r_2}, Y^{(0),\text{full}}, \mathcal{P}) \oslash (\mathbf{1} - \widehat{P})$.

5. Return $\widehat{\Theta}^{(1)} = \texttt{Cross-Fitted-MC}(\texttt{TW}_{r_3}, Y^{(1),\text{full}}, \mathcal{P}) \oslash \widehat{P}$.

We provide intuition on the key steps of the `Cross-Fitted-SVD` algorithm next.

Computing $\widehat{P}$. The estimate $\widehat{P}$ comes from applying `Cross-Fitted-MC` with TW on A and truncating the entries of the resulting matrix to the range $[\bar{\lambda}, 1 - \bar{\lambda}]$, in accordance with Assumption 4.3. The TW sub-routine is directly applicable to A, because for any block $\mathcal{I} = \mathcal{R}_s \times \mathcal{C}_k \in \mathcal{P}$ the masked matrix $A \otimes \mathbf{1}^{-\mathcal{I}}$ has $[N] \setminus \mathcal{R}_s$ fully observed rows and $[M] \setminus \mathcal{C}_k$ fully observed columns. See Figure 4.5.2(a) for a visualization of $A \otimes \mathbf{1}^{-\mathcal{I}}$.



1	0	1	1	0	0
1	0	0	0	1	1
0	1	0	1	0	1
1	0	1	?	?	?
0	1	0	?	?	?
0	1	1	?	?	?

(*a*) $A \otimes \mathbf{1}^{-\text{Bottom Right}}$

?	🟦	?	?	🟦	🟦
?	🟦	🟦	🟦	?	?
🟦	?	🟦	?	🟦	?
?	🟦	?	?	?	?
🟦	?	🟦	?	?	?
🟦	?	?	?	?	?

(*b*) $Y^{(0),\text{obs}} \otimes \mathbf{1}^{-\text{Bottom Right}}$

🟩	?	🟩	🟩	?	?
🟩	?	?	?	🟩	🟩
?	🟩	?	🟩	?	🟩
🟩	?	🟩	?	?	?
?	🟩	?	?	?	?
?	🟩	🟩	?	?	?

(*c*) $Y^{(1),\text{obs}} \otimes \mathbf{1}^{-\text{Bottom Right}}$

Figure 4.5.2: Panels (*a*), (*b*), and (*c*) illustrate the matrices $A \otimes \mathbf{1}^{-\mathcal{I}}$, $Y^{(0),\text{obs}} \otimes \mathbf{1}^{-\mathcal{I}}$, and $Y^{(1),\text{obs}} \otimes \mathbf{1}^{-\mathcal{I}}$ obtained from A, $Y^{(0),\text{obs}}$ and $Y^{(1),\text{obs}}$, respectively, for the block partition $\mathcal{P}$ in Figure 4.5.1(*a*) and the block $\mathcal{I} =$ Bottom Right. Unlike Panels (*b*) and (*c*), there exists rows and columns with all entries observed in Panel (*a*). To enable the application of `TW` for Panels (*b*) and (*c*), we replace missing entries in blocks Top Left, Top Right, and Bottom Left with zeros.

Computing $\widehat{\Theta}^{(0)}$ and $\widehat{\Theta}^{(1)}$. The estimates $\widehat{\Theta}^{(0)}$ and $\widehat{\Theta}^{(1)}$ are constructed by applying `Cross-Fitted-MC` with `TW` on $Y^{(0),\text{full}}$ and $Y^{(1),\text{full}}$, which do not have missing entries. `TW` is not directly applicable on $Y^{(0),\text{obs}}$ and $Y^{(1),\text{obs}}$, as both matrices may not have any rows and columns that are fully observed. See Figure 4.5.2(*b*) and Figure 4.5.2(*c*) for visualizations of $Y^{(0),\text{obs}} \otimes \mathbf{1}^{-\mathcal{I}}$ and $Y^{(1),\text{obs}} \otimes \mathbf{1}^{-\mathcal{I}}$, respectively. However, notice that, due to Assumption 4.2(a) and Assumption 4.2(b),

$$\mathbb{E}[Y^{(0),\text{full}}] = \mathbb{E}[Y \odot (\mathbf{1} - A)] = \Theta^{(0)} \odot (\mathbf{1} - P),$$

and

$$\mathbb{E}[Y^{(1),\text{full}}] = \mathbb{E}[Y \odot A] = \Theta^{(1)} \odot P.$$

As a result, $\texttt{MC}(Y^{(0),\text{full}})$ and $\texttt{MC}(Y^{(1),\text{full}})$ provide estimates of $\Theta^{(0)} \odot (\mathbf{1} - P)$ and $\Theta^{(1)} \odot P$, respectively—recall the discussion in Section 4.3.1. To construct $\widehat{\Theta}^{(0)}$ and $\widehat{\Theta}^{(1)}$, we divide the entries of $\texttt{MC}(Y^{(0),\text{full}})$ and $\texttt{MC}(Y^{(1),\text{full}})$ by the entries of $(\mathbf{1} - \widehat{P})$ and $\widehat{P}$, respectively, to adjust for heterogeneous probabilities of missingness (see, e.g., Bhattacharya and Chatterjee, 2022; Jin et al., 2021; Xiong and Pelger, 2023, for related procedures). This inverse probability of treatment weighting adjustment to estimate $\widehat{\Theta}^{(0)}$ and $\widehat{\Theta}^{(1)}$ is distinct and in addition to the augmented IPW procedure that generates $\widehat{\text{ATE}}^{\text{DR}}_{\cdot,j}$ from estimates $\widehat{\Theta}^{(0)}$, $\widehat{\Theta}^{(1)}$ and $\widehat{P}$.

4.5.3 Theoretical guarantees for `Cross-Fitted-SVD`

To establish theoretical guarantees for `Cross-Fitted-SVD`, we adopt three assumptions from Bai and Ng (2021). The first assumption imposes a low-rank structure on the matrices P, $\Theta^{(0)}$, and $\Theta^{(1)}$, namely that their entries are given by an inner product of latent factors.

Assumption 4.6 (Linear latent factor model on the confounders)**.** *There exist constants* $r_p, r_{\theta_0}, r_{\theta_1} \in [\min\{N, M\}]$ *and a collection of latent factors*

$$U \in \mathbb{R}^{N\times r_p}, \quad V \in \mathbb{R}^{M\times r_p}, \quad U^{(a)} \in \mathbb{R}^{N\times r_{\theta_a}}, \quad \textit{and} \quad V^{(a)} \in \mathbb{R}^{M\times r_{\theta_a}} \quad \textit{for} \quad a \in \{0,1\},$$

such that the unobserved confounders $(\Theta^{(0)}, \Theta^{(1)}, P)$ *satisfy the following factorization:*

$$P = UV^\top \quad \textit{and} \quad \Theta^{(a)} = U^{(a)}V^{(a)\top} \quad \textit{for} \quad a \in \{0,1\}. \tag{4.34}$$

Assumption 4.6 decomposes each of the unobserved confounders (P, $\Theta^{(0)}$, and $\Theta^{(0)}$) into low-dimensional unit-dependent latent factors (U, $U^{(0)}$, and $U^{(1)}$) and measurement-dependent latent factors (V, $V^{(0)}$, and $V^{(1)}$). In particular, every unit $i \in [N]$ is associated with three low-dimensional factors: (i) $U_i \in \mathbb{R}^{r_p}$, (ii) $U_i^{(0)} \in \mathbb{R}^{r_{\theta_0}}$, and (iii) $U_i^{(1)} \in \mathbb{R}^{r_{\theta_1}}$. Similarly, every measurement $j \in [M]$ is associated with three factors: (i) $V_j \in \mathbb{R}^{r_p}$, (ii) $V_j^{(0)} \in \mathbb{R}^{r_{\theta_0}}$, and (iii) $V_j^{(1)} \in \mathbb{R}^{r_{\theta_1}}$. Low-rank assumptions are standard in the matrix completion literature.

The second assumption requires that the factors that determine P, $\Theta^{(0)} \odot (\mathbf{1} - P)$, and $\Theta^{(1)} \odot P$ explain a sufficiently large amount of the variation in the data. This assumption is made on the factors of $\Theta^{(0)} \odot (\mathbf{1} - P)$ and $\Theta^{(1)} \odot P$ instead of $\Theta^{(0)}$ and $\Theta^{(1)}$ as the `TW` algorithm is applied on $Y^{(0),\text{full}} = Y \odot (\mathbf{1} - A)$ and $Y^{(1),\text{full}} = Y \odot A$, instead of $Y^{(0),\text{obs}}$ and $Y^{(1),\text{obs}}$ (see steps 4 and 5 of `Cross-Fitted-SVD`). To determine the factors of $\Theta^{(0)} \odot (\mathbf{1} - P)$ and $\Theta^{(1)} \odot P$, let

$$\overline{U} \triangleq [\mathbf{1}_N, -U] \in \mathbb{R}^{N\times(r_p+1)} \quad \text{and} \quad \overline{V} \triangleq [\mathbf{1}_M, V] \in \mathbb{R}^{M\times(r_p+1)},$$

where $\mathbf{1}_N \in \mathbb{R}^N$ and $\mathbf{1}_M \in \mathbb{R}^M$ are vectors of all 1's. Then,

$$\Theta^{(0)} \odot (\mathbf{1} - P) = \overline{U}^{(0)}\overline{V}^{(0)\top} \quad \text{and} \quad \Theta^{(1)} \odot P = \overline{U}^{(1)}\overline{V}^{(1)\top}, \tag{4.35}$$

where $\overline{U}^{(0)} \triangleq \overline{U} * U^{(0)} \in \mathbb{R}^{N\times r_{\theta_0}(r_p+1)}$, $\overline{V}^{(0)} \triangleq \overline{V} * V^{(0)} \in \mathbb{R}^{M\times r_{\theta_0}(r_p+1)}$, $\overline{U}^{(1)} \triangleq U * U^{(1)} \in \mathbb{R}^{N\times r_{\theta_1} r_p}$, and $\overline{V}^{(1)} \triangleq V * V^{(1)} \in \mathbb{R}^{M\times r_{\theta_1} r_p}$, with the operator $*$ denoting the Khatri-Rao product (see Section 4.1). We provide details of the derivation of these factors in Appendix 4.G.2.3.

Assumption 4.7 (Strong factors)**.** *There exists a positive constant* c *such that*

$$\|U\|_{2,\infty} \leq c, \quad \|V\|_{2,\infty} \leq c, \quad \|U^{(a)}\|_{2,\infty} \leq c, \quad \textit{and} \quad \|V^{(a)}\|_{2,\infty} \leq c \quad \textit{for} \quad a \in \{0,1\}.$$

Further, the matrices defined below exist and are positive definite:

$$\lim_{N\to\infty} \frac{U^\top U}{N}, \quad \lim_{M\to\infty} \frac{V^\top V}{M}, \quad \lim_{N\to\infty} \frac{\overline{U}^{(a)\top}\overline{U}^{(a)}}{N}, \quad \textit{and} \quad \lim_{M\to\infty} \frac{\overline{V}^{(a)\top}\overline{V}^{(a)}}{M} \quad \textit{for} \quad a \in \{0,1\}.$$

Assumption 4.7, a classic assumption in the literature on latent factor models, ensures that the factor structure is strong. Specifically, it ensures that each eigenvector of P, $\Theta^{(0)} \odot (\mathbf{1} - P)$, and $\Theta^{(1)} \odot P$ carries sufficiently large signal.

The third assumption requires a strong factor structure on the sub-matrices of P, $\Theta^{(0)} \odot (\mathbf{1} - P)$, and $\Theta^{(1)} \odot P$ corresponding to every block $\mathcal{I}$ in the block partition $\mathcal{P}$ from Assumption 4.5. Further, it also requires that the size $\mathcal{I}$ grows linearly in N and M.

Assumption 4.8 (Strong block factors)**.** *Consider the block partition $\mathcal{P} \triangleq \{\mathcal{R}_s \times \mathcal{C}_k : s, k \in \{0,1\}\}$ from Assumption 4.5. For every $s \in \{0,1\}$, let $U_{(s)} \in \mathbb{R}^{|\mathcal{R}_s| \times r_p}$, $\overline{U}^{(0)}_{(s)} \in \mathbb{R}^{|\mathcal{R}_s| \times r_{\theta_0}(r_p+1)}$, and $\overline{U}^{(1)}_{(s)} \in \mathbb{R}^{|\mathcal{R}_s| \times r_{\theta_1} r_p}$ be the sub-matrices of U, $\overline{U}^{(0)}$, and $\overline{U}^{(1)}$, respectively, that keeps the rows corresponding to the indices in $\mathcal{R}_s$. For every $k \in \{0,1\}$, let $V_{(k)} \in \mathbb{R}^{|\mathcal{C}_k| \times r_p}$, $\overline{V}^{(0)}_{(k)} \in \mathbb{R}^{|\mathcal{C}_k| \times r_{\theta_0}(r_p+1)}$, and $\overline{V}^{(1)}_{(k)} \in \mathbb{R}^{|\mathcal{C}_k| \times r_{\theta_1} r_p}$ be the sub-matrices of V, $\overline{V}^{(0)}$, and $\overline{V}^{(1)}$, respectively, that keeps the rows corresponding to the indices in $\mathcal{C}_k$. Then, for every $s, k \in \{0,1\}$, the matrices defined below exist and are positive definite:*

$$\lim_{N\to\infty} \frac{U_{(s)}^\top U_{(s)}}{|\mathcal{R}_s|}, \quad \lim_{M\to\infty} \frac{V_{(k)}^\top V_{(k)}}{|\mathcal{C}_k|}, \quad \lim_{N\to\infty} \frac{\overline{U}^{(a)\top}_{(s)} \overline{U}^{(a)}_{(s)}}{|\mathcal{R}_s|}, \quad \textit{and} \quad \lim_{M\to\infty} \frac{\overline{V}^{(a)\top}_{(k)} \overline{V}^{(a)}_{(k)}}{|\mathcal{C}_k|} \quad \textit{for } a \in \{0,1\}.$$

Further, for every $s, k \in \{0,1\}$, $|\mathcal{R}_s| = \Omega(N)$ and $|\mathcal{C}_k| = \Omega(M)$.

The subsequent assumption introduces additional conditions on the noise variables in Bai and Ng (2021) than those specified in Assumptions 4.2 and 4.5.

Assumption 4.9 (Weak dependence in noise across measurements and independence in noise across units)**.**

(a) $\sum_{j' \in [M]} \left|\mathbb{E}[\eta_{i,j}\eta_{i,j'}]\right| \leq c$ *for every $i \in [N]$ and $j \in [M]$,*

(b) $\sum_{j' \in [M]} \left|\mathbb{E}[\overline{\varepsilon}^{(a)}_{i,j}\overline{\varepsilon}^{(a)}_{i,j'}]\right| \leq c$ *for every $i \in [N]$, $j \in [M]$, and $a \in \{0,1\}$, where $\overline{\varepsilon}^{(a)}_{i,j} \triangleq \theta_{i,j}\eta_{i,j} + \varepsilon^{(a)}_{i,j} p_{i,j} + \varepsilon^{(a)}_{i,j}\eta_{i,j}$, and*

(c) The elements of $\{(E^{(a)}_{i,\cdot}, W_{i,\cdot}) : i \in [N]\}$ are mutually independent (across i) for $a \in \{0,1\}$.

Assumption 4.9(a) and Assumption 4.9(b) requires the noise variables to exhibit only weak dependency across measurements. Still, these assumptions allow the existence of pairs of perfectly correlated outcomes (e.g., $j, j' \in [M]$ such that $a_{i,j} = a_{i,j'}$). Assumption 4.9(c) requires the noise $(E^{(a)}, W)$ to be jointly independent across units, for every $a \in \{0,1\}$. We are now ready to provide guarantees on the estimates produced by `Cross-Fitted-SVD`. The proof can be found in Appendix 4.G.2.

Proposition 4.4 (Guarantees for `Cross-Fitted-SVD`)**.** *Suppose Assumptions 4.1, 4.2, and 4.6 to 4.9 hold. Consider an asymptotic sequence such that $\theta_{\max}$ is bounded as both N and M increase. Suppose $\theta_{\max}$ is bounded. Let $\widehat{P}$, $\widehat{\Theta}^{(0)}$, and $\widehat{\Theta}^{(1)}$ be the estimates returned by `Cross-Fitted-SVD` with the block partition $\mathcal{P}$ from Assumption 4.5, $r_1 = r_p$, $r_2 = r_{\theta_0}(r_p+1)$, $r_3 = r_{\theta_1} r_p$, and any $\bar{\lambda}$ such that $0 < \bar{\lambda} \leq \lambda$ with λ denoting the constant from Assumption 4.1. Then, as $N, M \to \infty$,*

$$\mathcal{E}\big(\widehat{P}\big) = O_p\bigg(\frac{1}{\sqrt{N}} + \frac{1}{\sqrt{M}}\bigg) \quad \textit{and} \quad \mathcal{E}\big(\widehat{\Theta}\big) = O_p\bigg(\frac{1}{\sqrt{N}} + \frac{1}{\sqrt{M}}\bigg).$$

Proposition 4.4 implies that the conditions (C1) and (C2) in Theorem 4.2 hold whenever $N^{1/2}/M = o(1)$. Then, the DR estimator from Eq. (4.11) constructed using `Cross-Fitted-SVD` estimates $\widehat{\Theta}^{(0)}$, $\widehat{\Theta}^{(1)}$, and $\widehat{P}$ exhibits an asymptotic Gaussian distribution centered at the target causal estimand. Further, Proposition 4.4 implies that the estimation errors $\mathcal{E}\big(\widehat{P}\big)$ and $\mathcal{E}\big(\widehat{\Theta}\big)$ achieve the parametric rate whenever $N/M = O(1)$.

4.5.4 Application to panel data with staggered adoption

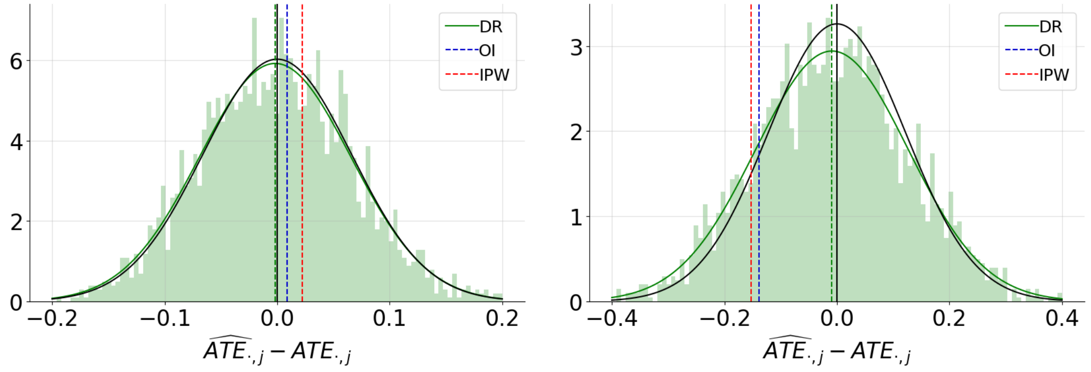
Section 4.5.1 considered a setting with block independence between noise (formalized in Assumption 4.5). Appendix 4.J discusses how to extend the proposed doubly-robust framework to a setting of panel data with staggered adoption, where this assumption may not hold. Recall (from Section 4.4.4) that in the panel data setting M measurements correspond to T time periods, and t denotes the time index. Then, Appendix 4.J considers a setting where a unit remains under control for some period of time, after which it deterministically remains under treatment. In other words, for every unit $i \in [N]$, there exists a time point $t_i \in [T]$ such that $a_{i,t} = 0$ for $t \leq t_i$, and $a_{i,t} = 1$ for $t > t_i$. Such a treatment assignment pattern leads to a heavy dependence in the noise $\{\eta_{i,t}\}_{t\in[T]}$ for every unit $i \in [N]$. Appendix 4.J describes an alternative approach to the `Cross-Fitted-SVD` algorithm and shows that Assumption 4.4 still holds for a suitable staggered adoption model.

4.6 Simulations

This section reports simulation results on the performance of the DR estimator of Eq. (4.11) and the OI and IPW estimators of Eqs. (4.9) and (4.10), respectively.

Data Generating Process (DGP). We now briefly describe the DGP for our simulations; Appendix 4.H provides details. All simulations set $N = M$. To generate, P, $\Theta^{(0)}$, and $\Theta^{(1)}$, we use the latent factor model given in Eq. (4.34). To introduce unobserved confounding, we set the unit-specific latent factors to be the same across P, $\Theta^{(0)}$, and $\Theta^{(1)}$, i.e., $U = U^{(0)} = U^{(1)}$. The entries of U and the measurement-specific latent factors, $V, V^{(0)}, V^{(1)}$ are each sampled independently from a uniform distribution, with hyperparameter r_p equal to the dimension of U and V, and hyperparameter r_p equal to the dimension of $U^{(a)}$ and $V^{(a)}$ for $a = 0, 1$. Further, the entries of the noise matrices $E^{(0)}$ and $E^{(1)}$ are sampled independently from a normal distribution, and the entries of W are sampled independently as in Eq. (4.4). Then, $y_{i,j}^{(a)}$, $a_{i,j}$, and $y_{i,j}$ are determined from Eqs. (4.1) to (4.3), respectively. The simulation generates P, $\Theta^{(0)}$, and $\Theta^{(1)}$ once. Given the fixed values of P, $\Theta^{(0)}$, and $\Theta^{(1)}$, the simulation generates 2500 realizations of (Y, A)—that is, only the noise matrices $E^{(0)}, E^{(1)}, W$ are resampled for each of the 2500 realizations. For each simulation realization, we apply the `Cross-Fitted-SVD` algorithm with hyper-parameters as in Proposition 4.4 and $\bar{\lambda} = \lambda = 0.05$ to obtain $\widehat{P}$, $\widehat{\Theta}^{(0)}$, and $\widehat{\Theta}^{(1)}$, and compute $\text{ATE}_{\cdot,j}$ from Eq. (4.5), and $\widehat{\text{ATE}}_{\cdot,j}^{\text{OI}}$, $\widehat{\text{ATE}}_{\cdot,j}^{\text{IPW}}$ and $\widehat{\text{ATE}}_{\cdot,j}^{\text{DR}}$ from Eqs. (4.9) to (4.11).

Results. Figure 4.6.1 reports simulation results for $N = 1000$, with $r_p = 3$, $r_\theta = 3$ in Panel (a), and $r_p = 5$, $r_\theta = 3$ in Panel (b). Figure 4.3.2 in Section 4.3 reports simulation results for $r_p = 3$, $r_\theta = 5$. In each case, the figure shows a histogram of the distribution of $\widehat{\text{ATE}}_{\cdot,j}^{\text{DR}} - \text{ATE}_{\cdot,j}$ across 2500 simulation instances for a fixed j, along with the best fitting Gaussian distribution (green curve). The histogram counts are normalized so that the area under the histogram integrates to one. Figure 4.6.1 plots the Gaussian distribution in the result of Theorem 4.2 (black curve). The dashed blue, red and green



(a) $r_p = 3$, $r_\theta = 3$ (b) $r_p = 5$, $r_\theta = 3$

Figure 4.6.1: Empirical illustration of the asymptotic performance of DR as in Theorem 4.2. The histogram corresponds to the errors of 2500 independent instances of DR estimates, the green curve represents the (best) fitted Gaussian distribution, and the black curve represents the Gaussian approximation from Theorem 4.2. The dashed green, blue, and red lines represent the biases of DR, OI, and IPW estimators.

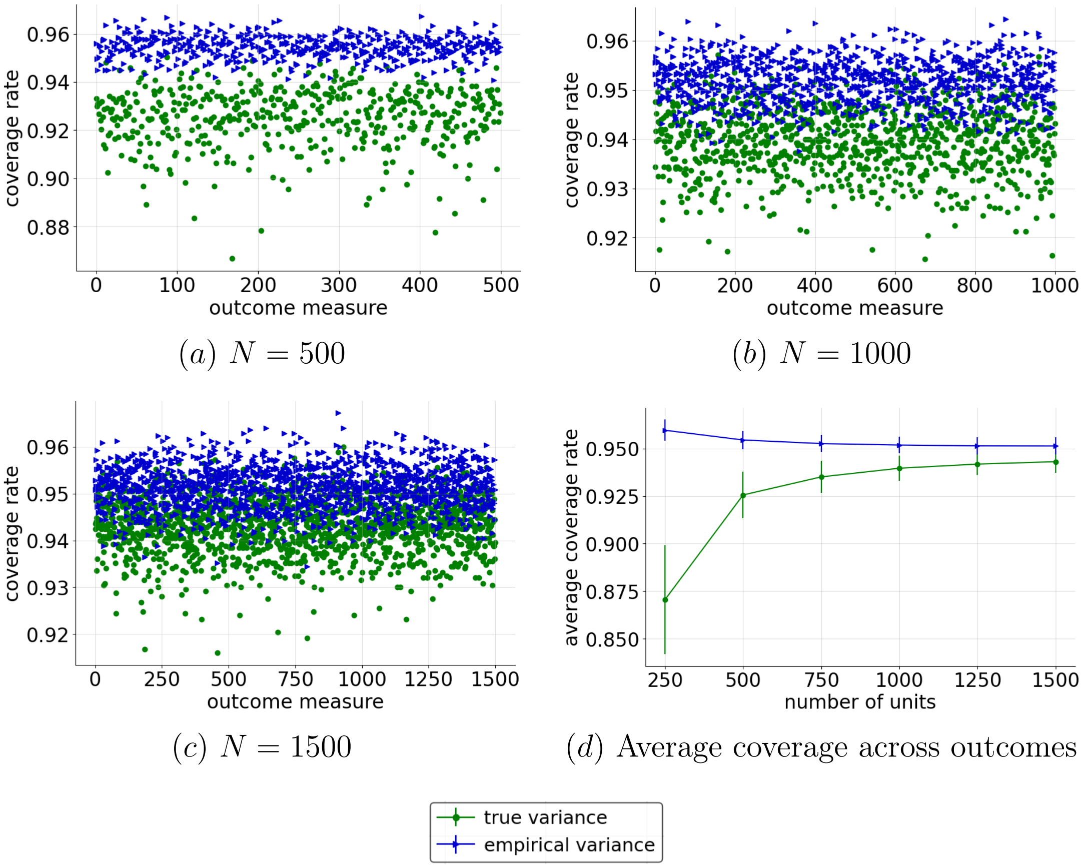
lines in Figures 4.3.2 and 4.6.1 indicate the values of the means of the OI, IPW, and DR error, respectively, across simulation instances. For reference, we place a black solid line at zero. The DR estimator has minimal bias and a close-to-Gaussian distribution. The biases of OI and IPW are non-negligible. In Appendix 4.H, we compare the biases and the standard deviations of OI, IPW, and DR across many j.

Panels (a), (b), and (c) of Figure 4.6.2 report coverage rates over the 2500 simulations for $\widehat{\mathrm{ATE}}^{\mathrm{DR}}_{\cdot,j}$-centered nominal 95% confidence intervals with $N = 500$, $N = 1000$, and $N = 1500$, respectively, all with $M = N$ and $r_p = r_\theta = 3$. For every $j \in [M]$, panels (a), (b) and (c) show $\widehat{\mathrm{c}}_j$, the percentage of times $[\widehat{\mathrm{ATE}}^{\mathrm{DR}}_{\cdot,j} \pm 1.96\widehat{\sigma}_j/\sqrt{N}]$ covers $\mathrm{ATE}_{\cdot,j}$ (in blue), and c_j, the percentage of times $[\widehat{\mathrm{ATE}}^{\mathrm{DR}}_{\cdot,j} \pm 1.96\sigma_j/\sqrt{N}]$ covers $\mathrm{ATE}_{\cdot,j}$ (in green). Panel (d) shows the means and standard deviations of $\{\widehat{\mathrm{c}}_j\}_{j\in[M]}$ and $\{\mathrm{c}_j\}_{j\in[M]}$ for different values of N. Confidence intervals based on the large-sample approximation results of Section 4.4 exhibit small size distortion even for fairly small values of N.

4.7 Concluding Remarks

This chapter introduces a new framework to estimate treatment effects in the presence unobserved confounding. We consider modern data-rich environments, where there are many units, and outcomes of interest per unit. We show it is possible to control for the confounding effects of a set of latent variables when this set is low-dimensional relative to the number of observed treatments and outcomes.

Our proposed estimator is doubly-robust, combining outcome imputation and inverse probability weighting with matrix completion. Analytical tractability of its distribution is gained through a novel cross-fitting procedure for causal matrix completion. We study the properties of the doubly-robust estimator, along with the outcome imputation



(a) $N = 500$

(b) $N = 1000$

(c) $N = 1500$

(d) Average coverage across outcomes

Figure 4.6.2: Panels (a), (b), and (c) report coverage rates for nominal 95% confidence intervals constructed using the estimated variance from Eq. (4.25) (in blue) and the true variance from Eq. (4.22) (in green) for $N \in \{500, 1000, 1500\}$ and $M = N$. Panel (d) shows the means and standard deviations of coverage rates across outcomes for different values of N.

and inverse probability weighting-based estimators under black-box matrix completion error rates. We show that the decay rate of the error of the doubly-robust estimator dominates those of the outcome imputation and the inverse probability weighting estimators. Moreover, we establish a Gaussian approximation to the distribution of the doubly-robust estimator. Simulation results demonstrate the practical relevance of the formal properties of the doubly-robust estimator.

Appendix

4.A Supporting concentration and convergence results

This section presents known results on subGaussian, subExponential, and subWeibull random variables (defined below), along with few basic results on convergence of random variables.

We use subGaussian(σ) to represent a subGaussian random variable, where σ is a bound on the subGaussian norm; and subExponential(σ) to represent a subExponential random variable, where σ is a bound on the subExponential norm. Recall the definitions of the norms from Section 4.1.

Lemma 4.1 (subGaussian concentration: Theorem 2.6.3 of Vershynin (2018))**.** *Let $x \in \mathbb{R}^n$ be a random vector whose entries are independent, zero-mean, subGaussian(σ) random variables. Then, for any $b \in \mathbb{R}^n$ and $t \geq 0$,*

$$\mathbb{P}\Big\{\big|b^\top x\big| \geq t\Big\} \leq 2\exp\Big(\frac{-ct^2}{\sigma^2\|b\|_2^2}\Big).$$

The following corollary expresses the bound in Lemma 4.1 in a convenient form.

Corollary 4.3 (subGaussian concentration)**.** *Let $x \in \mathbb{R}^n$ be a random vector whose entries are independent, zero-mean, subGaussian(σ) random variables. Then, for any $b \in \mathbb{R}^n$ and any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\big|b^\top x\big| \leq \sigma\sqrt{c\ell_\delta} \cdot \|b\|_2.$$

Proof. The proof follows from Lemma 4.1 by choosing $\delta \triangleq 2\exp(-ct^2/\sigma^2\|b\|_2^2)$. □

Lemma 4.2 (subExponential concentration: Theorem 2.8.2 of Vershynin (2018))**.** *Let $x \in \mathbb{R}^n$ be a random vector whose entries are independent, zero-mean, subExponential(σ) random variables. Then, for any $b \in \mathbb{R}^n$ and $t \geq 0$,*

$$\mathbb{P}\Big\{\big|b^\top x\big| \geq t\Big\} \leq 2\exp\Bigg(-c\min\Big(\frac{t^2}{\sigma^2\|b\|_2^2}, \frac{t}{\sigma\|b\|_\infty}\Big)\Bigg).$$

The following corollary expresses the bound in Lemma 4.2 in a convenient form.

Corollary 4.4 (subExponential concentration). *Let $x \in \mathbb{R}^n$ be a random vector whose entries are independent, zero-mean, subExponential(σ) random variables. Then, for any $b \in \mathbb{R}^n$ and any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\left|b^\top x\right| \le \sigma m(c\ell_\delta) \cdot \|b\|_2,$$

where recall that $m(c\ell_\delta) = \max\left(c\ell_\delta, \sqrt{c\ell_\delta}\right)$.

Proof. Choosing $t = t_0\sigma\|b\|_2$ in Lemma 4.2, we have

$$\mathbb{P}\Big\{\left|b^\top x\right| \ge t_0\sigma\|b\|_2\Big\} \le 2\exp\Big(-ct_0\min\Big(t_0, \frac{\|b\|_2}{\|b\|_\infty}\Big)\Big) \le 2\exp\Big(-ct_0\min\big(t_0,1\big)\Big),$$

where the second inequality follows from $\min\{t_0, c\} \ge \min\{t_0, 1\}$ for any $c \ge 1$ and $\|b\|_2 \ge \|b\|_\infty$. Then, the proof follows by choosing $\delta \triangleq 2\exp\big(-ct_0\min\big(t_0,1\big)\big)$ which fixes $t_0 = \max\{\sqrt{c\ell_\delta}, c\ell_\delta\} = m(c\ell_\delta)$.

□

Lemma 4.3 (Product of subGaussians is subExponential: Lemma. 2.7.7 of Vershynin (2018)). *Let x_1 and x_2 be subGaussian(σ_1) and subGaussian(σ_2) random variables, respectively. Then, x_1x_2 is subExponential($\sigma_1\sigma_2$) random variable.*

Next, we provide the definition of a subWeibull random variable.

Definition 4.1 (subWeibull random variable: Definition 1 of Zhang and Wei (2022)). *For $\rho > 0$, a random variable x is subWeibull with index ρ if it has a bounded subWeibull norm defined as follows:*

$$\|x\|_{\psi_\rho} \triangleq \inf\{t > 0 : \mathbb{E}[\exp(|x|^\rho/t^\rho)] \le 2\}.$$

We use subWeibull$_\rho(\sigma)$ to represent a subWeibull random variable with index ρ, where σ is a bound on the subWeibull norm. Note that subGaussian and subExponential random variables are subWeibull random variable with indices 2 and 1, respectively.

Lemma 4.4 (Product of subWeibulls is subWeibull: Proposition 2 of Zhang and Wei (2022)). *For $i \in [d]$, let x_i be a subWeibull$_{\rho_i}(\sigma_i)$ random variable. Then, $\Pi_{i\in[d]}x_i$ is subWeibull$_\rho(\sigma)$ random variable where*

$$\sigma = \Pi_{i\in[d]}\sigma_i \quad \textit{and} \quad \rho = \Bigg(\sum_{i\in[d]} 1/\rho_i\Bigg)^{-1}.$$

Next set of lemmas provide useful intermediate results on stochastic convergence.

Lemma 4.5. *Let X_n and $\overline{X}_n$ be sequences of random variables. Let δ_n be a deterministic sequence such that $0 \le \delta_n \le 1$ and $\delta_n \to 0$. Suppose $X_n = o_p(1)$ and $\mathbb{P}(|\overline{X}_n| \le |X_n|) \ge 1-\delta_n$. Then, $\overline{X}_n = o_p(1)$.*

Proof. We need to show that for any $\epsilon > 0$ and $\delta > 0$, there exist finite $\overline{n}$, such that

$$\mathbb{P}(|\overline{X}_n| > \delta) < \epsilon$$

for all $n \geq \overline{n}$. Fix any $\epsilon > 0$. As δ_n converges to zero, there exists a finite n_0 such that $\delta_n < \epsilon/2$, for all $n \geq n_0$. As X_n is converges to zero in probability, there exists finite n_1, such that $\mathbb{P}(|X_n| > \delta) < \epsilon/2$ for all $n \geq n_1$. Now, the event $\{|\overline{X}_n| > \delta\}$ belongs to the union of $\{|\overline{X}_n| > |X_n|\}$ and $\{|X_n| > \delta\}$. As a result, we obtain

$$\mathbb{P}(|\overline{X}_n| > \delta) \leq \mathbb{P}(|\overline{X}_n| > |X_n|) + \mathbb{P}(|X_n| > \delta) \leq \delta_n + \mathbb{P}(|X_n| > \delta) < \epsilon,$$

for $n \geq \overline{n} = \max\{n_0, n_1\}$. Therefore, $\overline{X}_n = o_p(1)$. □

Lemma 4.6. *Let X_n and $\overline{X}_n$ be sequences of random variables. Suppose $\mathbb{E}\big[|X_n|\big|\overline{X}_n\big] = o_p(1)$. Then, $X_n = o_p(1)$.*

Proof. Fix any $\delta > 0$. Markov's inequality implies

$$\mathbb{P}\Big(|X_n| \geq \delta\Big|\overline{X}_n\Big) \leq \frac{1}{\delta}\mathbb{E}\Big[|X_n|\Big|\overline{X}_n\Big] = o_p(1).$$

The law of total probability and the boundedness of conditional probabilities yield

$$\mathbb{P}\Big(|X_n| \geq \delta\Big) = \mathbb{E}\Big[\mathbb{P}\Big(|X_n| \geq \delta\Big|\overline{X}_n\Big)\Big] \longrightarrow 0.$$

□

Lemma 4.7. *Let X_n and $\overline{X}_n$ be sequences of random variables. Suppose $X_n = O_p(1)$ and $\mathbb{P}\big(|\overline{X}_n| \geq |X_n| + f(\epsilon)\big) < \epsilon$ for some positive function f and every $\epsilon \in (0,1)$. Then, $\overline{X}_n = O_p(1)$.*

Proof. We need to show that for any $\epsilon > 0$, there exist finite $\overline{\delta} > 0$ and $\overline{n} > 0$, such that

$$\mathbb{P}(|\overline{X}_n| > \overline{\delta}) < \epsilon$$

for all $n \geq \overline{n}$. Fix any $\epsilon > 0$. Because X_n is bounded in probability, there exist finite δ and n_0, such that $\mathbb{P}(|X_n| > \delta) < \epsilon/2$ for all $n \geq n_0$. Further, we have $\mathbb{P}\big(|\overline{X}_n| \geq |X_n| + f(\epsilon/2)\big) < \epsilon/2$. Now, the event $\{|\overline{X}_n| > \delta + f(\epsilon/2)\}$ belongs to the union of $\{|\overline{X}_n| > |X_n| + f(\epsilon/2)\}$ and $\{|X_n| > \delta\}$. As a result, we obtain

$$\mathbb{P}\big(|\overline{X}_n| > \delta + f(\epsilon/2)\big) \leq \mathbb{P}\big(|\overline{X}_n| > |X_n| + f(\epsilon/2)\big) + \mathbb{P}\big(|X_n| > \delta\big) < \epsilon.$$

for all $n \geq n_0$. In other words, $\mathbb{P}(|\overline{X}_n| > \overline{\delta}) < \epsilon$ for all $n \geq \overline{n}$, where $\overline{\delta} = \delta + f(\epsilon/2) > 0$ and $\overline{n} = n_0$. Therefore, $\overline{X}_n = O_p(1)$. □

4.B Proof of Theorem 4.1: Finite Sample Guarantees for DR

Fix any $j \in [M]$. Recall the definitions of the parameter $\mathrm{ATE}_{\cdot,j}$ and corresponding doubly-robust estimate $\widehat{\mathrm{ATE}}_{\cdot,j}^{\mathrm{DR}}$ from Eqs. (4.5) and (4.11), respectively. The error $\Delta\mathrm{ATE}_{\cdot,j}^{\mathrm{DR}} = \widehat{\mathrm{ATE}}_{\cdot,j}^{\mathrm{DR}} - \mathrm{ATE}_{\cdot,j}$ can be re-expressed as

$$\begin{aligned}\Delta\mathrm{ATE}_{\cdot,j}^{\mathrm{DR}} &= \frac{1}{N}\sum_{i\in[N]}\Big(\widehat{\theta}_{i,j}^{(1,\mathrm{DR})} - \widehat{\theta}_{i,j}^{(0,\mathrm{DR})}\Big) - \frac{1}{N}\sum_{i\in[N]}\Big(\theta_{i,j}^{(1)} - \theta_{i,j}^{(0)}\Big)\\ &= \frac{1}{N}\sum_{i\in[N]}\Big(\big(\widehat{\theta}_{i,j}^{(1,\mathrm{DR})} - \theta_{i,j}^{(1)}\big) - \big(\widehat{\theta}_{i,j}^{(0,\mathrm{DR})} - \theta_{i,j}^{(0)}\big)\Big)\\ &\overset{(a)}{=} \frac{1}{N}\sum_{i\in[N]}\big(\mathbb{T}_{i,j}^{(1,\mathrm{DR})} + \mathbb{T}_{i,j}^{(0,\mathrm{DR})}\big),\end{aligned} \tag{4.36}$$

where (a) follows after defining $\mathbb{T}_{i,j}^{(1,\mathrm{DR})} \triangleq \big(\widehat{\theta}_{i,j}^{(1,\mathrm{DR})} - \theta_{i,j}^{(1)}\big)$ and $\mathbb{T}_{i,j}^{(0,\mathrm{DR})} \triangleq -\big(\widehat{\theta}_{i,j}^{(0,\mathrm{DR})} - \theta_{i,j}^{(0)}\big)$ for every $(i,j) \in [N]\times[M]$. Then, we have

$$\begin{aligned}\mathbb{T}_{i,j}^{(1,\mathrm{DR})} &= \widehat{\theta}_{i,j}^{(1,\mathrm{DR})} - \theta_{i,j}^{(1)}\\ &\overset{(a)}{=} \widehat{\theta}_{i,j}^{(1)} + \big(y_{i,j} - \widehat{\theta}_{i,j}^{(1)}\big)\frac{a_{i,j}}{\widehat{p}_{i,j}} - \theta_{i,j}^{(1)}\\ &\overset{(b)}{=} \widehat{\theta}_{i,j}^{(1)} + \big(\theta_{i,j}^{(1)} + \varepsilon_{i,j}^{(1)} - \widehat{\theta}_{i,j}^{(1)}\big)\frac{p_{i,j}+\eta_{i,j}}{\widehat{p}_{i,j}} - \theta_{i,j}^{(1)} \qquad (4.37)\\ &= (\widehat{\theta}_{i,j}^{(1)} - \theta_{i,j}^{(1)})\Big(1 - \frac{p_{i,j}+\eta_{i,j}}{\widehat{p}_{i,j}}\Big) + \varepsilon_{i,j}^{(1)}\Big(\frac{p_{i,j}+\eta_{i,j}}{\widehat{p}_{i,j}}\Big)\\ &= \frac{(\widehat{\theta}_{i,j}^{(1)} - \theta_{i,j}^{(1)})(\widehat{p}_{i,j} - p_{i,j})}{\widehat{p}_{i,j}} - \frac{(\widehat{\theta}_{i,j}^{(1)} - \theta_{i,j}^{(1)})\eta_{i,j}}{\widehat{p}_{i,j}} + \frac{\varepsilon_{i,j}^{(1)}p_{i,j}}{\widehat{p}_{i,j}} + \frac{\varepsilon_{i,j}^{(1)}\eta_{i,j}}{\widehat{p}_{i,j}}, \qquad (4.38)\end{aligned}$$

where (a) follows from Eq. (4.12), and (b) follows from Eqs. (4.1) to (4.3). A similar derivation for $a = 0$ implies that

$$\begin{aligned}\mathbb{T}_{i,j}^{(0,\mathrm{DR})} &= -\frac{(\widehat{\theta}_{i,j}^{(0)} - \theta_{i,j}^{(0)})(1-\widehat{p}_{i,j} - (1-p_{i,j}))}{1-\widehat{p}_{i,j}} + \frac{(\widehat{\theta}_{i,j}^{(0)} - \theta_{i,j}^{(0)})(-\eta_{i,j})}{1-\widehat{p}_{i,j}} - \frac{\varepsilon_{i,j}^{(0)}(1-p_{i,j})}{1-\widehat{p}_{i,j}}\\ &\quad - \frac{\varepsilon_{i,j}^{(0)}(-\eta_{i,j})}{1-\widehat{p}_{i,j}}\\ &= \frac{(\widehat{\theta}_{i,j}^{(0)} - \theta_{i,j}^{(0)})(\widehat{p}_{i,j} - p_{i,j})}{1-\widehat{p}_{i,j}} - \frac{(\widehat{\theta}_{i,j}^{(0)} - \theta_{i,j}^{(0)})\eta_{i,j}}{1-\widehat{p}_{i,j}} - \frac{\varepsilon_{i,j}^{(0)}(1-p_{i,j})}{1-\widehat{p}_{i,j}} + \frac{\varepsilon_{i,j}^{(0)}\eta_{i,j}}{1-\widehat{p}_{i,j}}. \qquad (4.39)\end{aligned}$$

Consider any $a \in \{0,1\}$ and any $\delta \in (0,1)$. We claim that, with probability at least $1-6\delta$,

$$\frac{1}{N}\Big|\sum_{i\in[N]}\mathbb{T}_{i,j}^{(\mathrm{a},\mathrm{DR})}\Big| \le \frac{2}{\bar{\lambda}}\mathcal{E}\big(\widehat{\Theta}^{(a)}\big)\mathcal{E}\big(\widehat{P}\big) + \frac{2\sqrt{c\ell_\delta}}{\bar{\lambda}\sqrt{\ell_1 N}}\mathcal{E}\big(\widehat{\Theta}^{(a)}\big) + \frac{2\overline{\sigma}\sqrt{c\ell_\delta}}{\bar{\lambda}\sqrt{N}} + \frac{2\overline{\sigma}m(c\ell_\delta)}{\bar{\lambda}\sqrt{\ell_1 N}}, \tag{4.40}$$

where recall that $m(c\ell_\delta) = \max\left(c\ell_\delta, \sqrt{c\ell_\delta}\right)$. We provide a proof of this claim at the end of this section. Applying triangle inequality in Eq. (4.36) and using Eq. (4.40) with a union bound, we obtain that

$$\left|\Delta\mathrm{ATE}^{\mathrm{DR}}_{\cdot,j}\right| \leq \frac{2}{\bar{\lambda}}\mathcal{E}\big(\widehat{\Theta}\big)\mathcal{E}\big(\widehat{P}\big) + \frac{2\sqrt{c\ell_\delta}}{\bar{\lambda}\sqrt{\ell_1 N}}\mathcal{E}\big(\widehat{\Theta}\big) + \frac{4\overline{\sigma}\sqrt{c\ell_\delta}}{\bar{\lambda}\sqrt{N}} + \frac{4\overline{\sigma}m(c\ell_\delta)}{\bar{\lambda}\sqrt{\ell_1 N}},$$

with probability at least $1 - 12\delta$. The claim in Eq. (4.18) follows by re-parameterizing δ.

Proof of bound Eq. (4.40). Recall the partitioning of the units $[N]$ into $\mathcal{R}_0$ and $\mathcal{R}_1$ from Assumption 4.4. Now, to enable the application of concentration bounds, we split the summation over $i \in [N]$ in the left hand side of Eq. (4.40) into two parts—one over $i \in \mathcal{R}_0$ and the other over $i \in \mathcal{R}_1$—such that the noise terms are independent of the estimates of $\Theta^{(0)}, \Theta^{(1)}, P$ in each of these parts as in Eqs. (4.14) and (4.15).

Fix $a = 1$ and note that $|\sum_{i\in[N]} \mathbb{T}^{(1,\mathrm{DR})}_{i,j}| \leq |\sum_{i\in\mathcal{R}_0} \mathbb{T}^{(1,\mathrm{DR})}_{i,j}| + |\sum_{i\in\mathcal{R}_1} \mathbb{T}^{(1,\mathrm{DR})}_{i,j}|$. Fix any $s \in \{0, 1\}$. Then, Eq. (4.38) and triangle inequality imply

$$\begin{aligned}\Big|\sum_{i\in\mathcal{R}_s} \mathbb{T}^{(1,\mathrm{DR})}_{i,j}\Big| \leq \Big|\sum_{i\in\mathcal{R}_s} \frac{\big(\widehat{\theta}^{(1)}_{i,j} - \theta^{(1)}_{i,j}\big)\big(\widehat{p}_{i,j} - p_{i,j}\big)}{\widehat{p}_{i,j}}\Big| + \Big|\sum_{i\in\mathcal{R}_s} \frac{\big(\widehat{\theta}^{(1)}_{i,j} - \theta^{(1)}_{i,j}\big)\eta_{i,j}}{\widehat{p}_{i,j}}\Big| \\ + \Big|\sum_{i\in\mathcal{R}_s} \frac{\varepsilon^{(1)}_{i,j} p_{i,j}}{\widehat{p}_{i,j}}\Big| + \Big|\sum_{i\in\mathcal{R}_s} \frac{\varepsilon^{(1)}_{i,j} \eta_{i,j}}{\widehat{p}_{i,j}}\Big|. \qquad (4.41)\end{aligned}$$

Applying the Cauchy-Schwarz inequality to bound the first term yields that

$$\begin{aligned}\Big|\sum_{i\in\mathcal{R}_s} \frac{\big(\widehat{\theta}^{(1)}_{i,j} - \theta^{(1)}_{i,j}\big)\big(\widehat{p}_{i,j} - p_{i,j}\big)}{\widehat{p}_{i,j}}\Big| &\leq \sqrt{\sum_{i\in\mathcal{R}_s}\left(\frac{\widehat{\theta}^{(1)}_{i,j} - \theta^{(1)}_{i,j}}{\widehat{p}_{i,j}}\right)^2 \sum_{i\in\mathcal{R}_s}\big(\widehat{p}_{i,j} - p_{i,j}\big)^2} \\ &\leq \big\|\big(\widehat{\Theta}^{(1)}_{\cdot,j} - \Theta^{(1)}_{\cdot,j}\big) \oslash \widehat{P}_{\cdot,j}\big\|_2 \big\|\widehat{P}_{\cdot,j} - P_{\cdot,j}\big\|_2. \qquad (4.42)\end{aligned}$$

To bound the second term in Eq. (4.41), note that $\eta_{i,j}$ is subGaussian$(1/\sqrt{\ell_1})$ (see Example 2.5.8 in Vershynin (2018)) as well as zero-mean and independent across all $i \in [N]$ due to Assumption 4.2(a). By Assumption 4.4, $\{(\widehat{p}_{i,j}, \widehat{\theta}^{(1)}_{i,j})\}_{i\in\mathcal{R}_s} \perp\!\!\!\perp \{\eta_{i,j}\}_{i\in\mathcal{R}_s}$. The subGaussian concentration result in Corollary 4.3 yields

$$\Big|\sum_{i\in\mathcal{R}_s} \frac{\big(\widehat{\theta}^{(1)}_{i,j} - \theta^{(1)}_{i,j}\big)\eta_{i,j}}{\widehat{p}_{i,j}}\Big| \leq \frac{\sqrt{c\ell_\delta}}{\sqrt{\ell_1}}\sqrt{\sum_{i\in\mathcal{R}_s}\left(\frac{\widehat{\theta}^{(1)}_{i,j} - \theta^{(1)}_{i,j}}{\widehat{p}_{i,j}}\right)^2} \leq \frac{\sqrt{c\ell_\delta}}{\sqrt{\ell_1}}\big\|\big(\widehat{\Theta}^{(1)}_{\cdot,j} - \Theta^{(1)}_{\cdot,j}\big) \oslash \widehat{P}_{\cdot,j}\big\|_2, \quad (4.43)$$

with probability at least $1 - \delta$.

To bound the third term in Eq. (4.41), note that $\varepsilon^{(1)}_{i,j}$ is subGaussian$(\overline{\sigma})$, zero-mean, and independent across all $i \in [N]$ due to Assumption 4.2. By Assumption Assumption 4.4, $\{\widehat{p}_{i,j}\}_{i\in\mathcal{R}_s} \perp\!\!\!\perp \{\varepsilon^{(1)}_{i,j}\}_{i\in\mathcal{R}_s}$. The subGaussian concentration result in Corollary 4.3 yields

$$\Big|\sum_{i\in\mathcal{R}_s} \frac{\varepsilon^{(1)}_{i,j} p_{i,j}}{\widehat{p}_{i,j}}\Big| \leq \overline{\sigma}\sqrt{c\ell_\delta}\sqrt{\sum_{i\in\mathcal{R}_s}\left(\frac{p_{i,j}}{\widehat{p}_{i,j}}\right)^2} \leq \overline{\sigma}\sqrt{c\ell_\delta}\big\|P_{\cdot,j} \oslash \widehat{P}_{\cdot,j}\big\|_2, \qquad (4.44)$$

with probability at least $1-\delta$.

To bound the fourth term in Eq. (4.41), note that $\varepsilon^{(1)}_{i,j}\eta_{i,j}$ is subExponential$(\overline{\sigma}/\sqrt{\ell_1})$ because of Lemma 4.3 as well as zero-mean and independent across all $i \in [N]$ due to Assumption 4.2. By Assumption 4.4, $\{\widehat{p}_{i,j}\}_{i\in\mathcal{R}_s} \perp\!\!\!\perp \{(\eta_{i,j}, \varepsilon^{(1)}_{i,j})\}_{i\in\mathcal{R}_s}$. The subExponential concentration result in Corollary 4.4 yields that

$$\Big|\sum_{i\in\mathcal{R}_s}\frac{\varepsilon^{(1)}_{i,j}\eta_{i,j}}{\widehat{p}_{i,j}}\Big| \le \frac{\overline{\sigma}m(c\ell_\delta)}{\sqrt{\ell_1}}\|\mathbf{1}_N \oslash \widehat{P}_{\cdot,j}\|_2, \tag{4.45}$$

with probability at least $1-\delta$. Putting together Eqs. (4.41) to (4.45), we conclude that, with probability at least $1-3\delta$,

$$\begin{aligned}\frac{1}{N}\Big|\sum_{i\in\mathcal{R}_s}\mathbb{T}^{(1,\mathrm{DR})}_{i,j}\Big| \le \frac{1}{N}\big\|\big(\widehat{\Theta}^{(1)}_{\cdot,j}-\Theta^{(1)}_{\cdot,j}\big)\oslash\widehat{P}_{\cdot,j}\big\|_2\big\|\widehat{P}_{\cdot,j}-P_{\cdot,j}\big\|_2 + \frac{\sqrt{c\ell_\delta}}{\sqrt{\ell_1}N}\big\|\big(\widehat{\Theta}^{(1)}_{\cdot,j}-\Theta^{(1)}_{\cdot,j}\big)\oslash\widehat{P}_{\cdot,j}\big\|_2 \\ + \frac{\overline{\sigma}\sqrt{c\ell_\delta}}{N}\big\|P_{\cdot,j}\oslash\widehat{P}_{\cdot,j}\big\|_2 + \frac{\overline{\sigma}m(c\ell_\delta)}{\sqrt{\ell_1}N}\big\|\mathbf{1}_N\oslash\widehat{P}_{\cdot,j}\big\|_2.\end{aligned}$$

Then, noting that $1/\widehat{p}_{i,j} \le 1/\bar{\lambda}$ for every $i \in [N]$ and $j \in [M]$ from Assumption 4.3, and consequently that $\|B_{\cdot,j}\oslash\widehat{P}_{\cdot,j}\|_2 \le \|B\|_{1,2}/\bar{\lambda}$ for any matrix B and every $j \in [M]$, we obtain the following bound, with probability at least $1-3\delta$,

$$\begin{aligned}\frac{1}{N}\Big|\sum_{i\in\mathcal{R}_s}\mathbb{T}^{(1,\mathrm{DR})}_{i,j}\Big| &\le \frac{1}{\bar{\lambda}N}\|\widehat{\Theta}^{(1)}-\Theta^{(1)}\|_{1,2}\|\widehat{P}-P\|_{1,2} + \frac{\sqrt{c\ell_\delta}}{\bar{\lambda}\sqrt{\ell_1}N}\|\widehat{\Theta}^{(1)}-\Theta^{(1)}\|_{1,2} \\ &\qquad + \frac{\overline{\sigma}\sqrt{c\ell_\delta}}{\bar{\lambda}N}\|P\|_{1,2} + \frac{\overline{\sigma}m(c\ell_\delta)}{\bar{\lambda}\sqrt{\ell_1}N}\|\mathbf{1}\|_{1,2} &(4.46)\\ &\overset{(a)}{\le} \frac{1}{\bar{\lambda}}\mathcal{E}\big(\widehat{\Theta}^{(1)}\big)\mathcal{E}\big(\widehat{P}\big) + \frac{\sqrt{c\ell_\delta}}{\bar{\lambda}\sqrt{\ell_1 N}}\mathcal{E}\big(\widehat{\Theta}^{(1)}\big) + \frac{\overline{\sigma}\sqrt{c\ell_\delta}}{\bar{\lambda}\sqrt{N}} + \frac{\overline{\sigma}m(c\ell_\delta)}{\bar{\lambda}\sqrt{\ell_1 N}}, &(4.47)\end{aligned}$$

where (a) follows from Eq. (4.16) and because $\|P\|_{1,2} \le \sqrt{N}$ and $\|\mathbf{1}\|_{1,2} = \sqrt{N}$. Then, the claim in Eq. (4.40) follows for $a=1$ by using Eq. (4.47) and applying a union bound over $s \in \{0,1\}$. The proof of Eq. (4.40) for $a=0$ follows similarly.

4.C Proofs of Corollaries 4.1 and 4.2

4.C.1 Proof of Corollary 4.1: Gains of DR over OI and IPW

Fix any $j \in [M]$ and any $\delta \in (0,1)$. First, consider IPW. Take any $\alpha \in [0, 1/2]$. From Eq. (4.20), with probability at least $1-\delta$,

$$N^{\alpha}\big|\widehat{\mathrm{ATE}}^{\mathrm{IPW}}_{\cdot,j} - \mathrm{ATE}_{\cdot,j}\big| \le \frac{2\theta_{\max}}{\bar{\lambda}}N^{\alpha}\mathcal{E}\big(\widehat{P}\big) + f_1(\delta)N^{\alpha-1/2} \le \frac{2\theta_{\max}}{\bar{\lambda}}N^{\alpha}\mathcal{E}\big(\widehat{P}\big) + f_1(\delta),$$

where

$$f_1(\delta) \triangleq \frac{2}{\bar{\lambda}}\bigg(\frac{\sqrt{c\ell_{\delta/12}}}{\sqrt{\ell_1}}\theta_{\max} + 2\overline{\sigma}\sqrt{c\ell_{\delta/12}} + \frac{2\overline{\sigma}m(c\ell_{\delta/12})}{\sqrt{\ell_1}}\bigg),$$

for $m(c)$ and ℓ_c as defined in Section 4.1. Then, if $\mathcal{E}\big(\widehat{P}\big) = O_p\big(N^{-\alpha}\big)$, Lemma 4.7 implies

$$\big|\widehat{\mathrm{ATE}}^{\mathrm{IPW}}_{\cdot,j} - \mathrm{ATE}_{\cdot,j}\big| = O_p\big(N^{-\alpha}\big).$$

Next, consider DR. From Eq. (4.17), with probability at least $1-\delta$,

$$\big|\widehat{\mathrm{ATE}}^{\mathrm{DR}}_{\cdot,j} - \mathrm{ATE}_{\cdot,j}\big| \leq \frac{2}{\bar{\lambda}}\mathcal{E}\big(\widehat{\Theta}\big)\mathcal{E}\big(\widehat{P}\big) + f_2(\delta)N^{-1/2},$$

where

$$f_2(\delta) \triangleq \frac{2}{\bar{\lambda}}\bigg(\frac{\sqrt{c\ell_{\delta/12}}}{\sqrt{\ell_1}}\mathcal{E}\big(\widehat{\Theta}\big) + 2\bar{\sigma}\sqrt{c\ell_{\delta/12}} + \frac{2\bar{\sigma}m(c\ell_{\delta/12})}{\sqrt{\ell_1}}\bigg).$$

Suppose $\mathcal{E}\big(\widehat{P}\big) = O_p\big(N^{-\alpha}\big)$ and $\mathcal{E}\big(\widehat{\Theta}\big) = O_p\big(N^{-\beta}\big)$. Consider two cases. First, suppose $\alpha+\beta \leq 0.5$. Then, with probability at least $1-\delta$,

$$\begin{aligned} N^{\alpha+\beta}\big|\widehat{\mathrm{ATE}}^{\mathrm{DR}}_{\cdot,j} - \mathrm{ATE}_{\cdot,j}\big| &\leq \frac{2}{\bar{\lambda}}N^{\alpha+\beta}\mathcal{E}\big(\widehat{\Theta}\big)\mathcal{E}\big(\widehat{P}\big) + f_2(\delta)N^{\alpha+\beta-1/2} \\ &\leq \frac{2}{\bar{\lambda}}N^{\alpha+\beta}\mathcal{E}\big(\widehat{\Theta}\big)\mathcal{E}\big(\widehat{P}\big) + f_2(\delta). \end{aligned}$$

Lemma 4.7 implies $\big|\widehat{\mathrm{ATE}}^{\mathrm{DR}}_{\cdot,j} - \mathrm{ATE}_{\cdot,j}\big| = O_p(N^{-(\alpha+\beta)})$. Next, suppose $\alpha+\beta > 0.5$. With probability at least $1-\delta$,

$$N^{1/2}\big|\widehat{\mathrm{ATE}}^{\mathrm{DR}}_{\cdot,j} - \mathrm{ATE}_{\cdot,j}\big| \leq \frac{2}{\bar{\lambda}}N^{1/2}\mathcal{E}\big(\widehat{\Theta}\big)\mathcal{E}\big(\widehat{P}\big) + f_2(\delta) \leq \frac{2}{\bar{\lambda}}N^{\alpha+\beta}\mathcal{E}\big(\widehat{\Theta}\big)\mathcal{E}\big(\widehat{P}\big) + f_2(\delta).$$

Lemma 4.7 implies $\big|\widehat{\mathrm{ATE}}^{\mathrm{DR}}_{\cdot,j} - \mathrm{ATE}_{\cdot,j}\big| = O_p(N^{-1/2})$.

4.C.2 Proof of Corollary 4.2: Consistency for DR

Fix any $j \in [M]$. Then, choose $\delta = 1/N$ in Eq. (4.18) and note that every term in the right hand side of Eq. (4.18) is $o_p(1)$ under the conditions on $\mathcal{E}\big(\widehat{\Theta}\big)$ and $\mathcal{E}\big(\widehat{P}\big)$. Then, Eq. (4.21) follows from Lemma 4.5.

4.D Proof of Proposition 4.1 (4.19): Finite Sample Guarantees for OI

Fix any $j \in [M]$. Recall the definitions of the parameter $\mathrm{ATE}_{\cdot,j}$ and corresponding outcome imputation estimate $\widehat{\mathrm{ATE}}^{\mathrm{OI}}_{\cdot,j}$ from Eqs. (4.5) and (4.9), respectively. The error $\Delta\mathrm{ATE}^{\mathrm{OI}}_{\cdot,j} = \widehat{\mathrm{ATE}}^{\mathrm{OI}}_{\cdot,j} - \mathrm{ATE}_{\cdot,j}$ can be re-expressed as

$$\Delta\mathrm{ATE}^{\mathrm{OI}}_{\cdot,j} = \frac{1}{N}\sum_{i\in[N]}\Big(\widehat{\theta}^{(1)}_{i,j} - \widehat{\theta}^{(0)}_{i,j}\Big) - \frac{1}{N}\sum_{i\in[N]}\Big(\theta^{(1)}_{i,j} - \theta^{(0)}_{i,j}\Big) = \frac{1}{N}\sum_{i\in[N]}\Big(\big(\widehat{\theta}^{(1)}_{i,j} - \theta^{(1)}_{i,j}\big) - \big(\widehat{\theta}^{(0)}_{i,j} - \theta^{(0)}_{i,j}\big)\Big).$$

Using the triangle inequality, we have

$$\big|\Delta \mathrm{ATE}^{\mathrm{OI}}_{\cdot,j}\big| \leq \frac{1}{N}\Big|\sum_{i\in[N]}\big(\widehat{\theta}^{(1)}_{i,j}-\theta^{(1)}_{i,j}\big)\Big| + \frac{1}{N}\Big|\sum_{i\in[N]}\big(\widehat{\theta}^{(0)}_{i,j}-\theta^{(0)}_{i,j}\big)\Big|. \tag{4.48}$$

Consider any $a \in \{0,1\}$. We claim that

$$\frac{1}{N}\Big|\sum_{i\in[N]}\big(\widehat{\theta}^{(a)}_{i,j}-\theta^{(a)}_{i,j}\big)\Big| \leq \mathcal{E}\big(\widehat{\Theta}^{(a)}\big). \tag{4.49}$$

The proof is complete by putting together Eqs. (4.48) and (4.49).

Proof of Eq. (4.49) Fix any $a \in \{0,1\}$. Using the Cauchy-Schwarz inequality, we have

$$\frac{1}{N}\Big|\sum_{i\in[N]}\big(\widehat{\theta}^{(1)}_{i,j}-\theta^{(1)}_{i,j}\big)\Big| \leq \frac{1}{N}\|\mathbf{1}_N\|_2\|\widehat{\Theta}^{(1)}_{\cdot,j}-\Theta^{(1)}_{\cdot,j}\|_2 \leq \frac{1}{\sqrt{N}}\|\widehat{\Theta}^{(1)}-\Theta^{(1)}\|_{1,2}.$$

4.E Proof of Proposition 4.1 (4.20): Finite Sample Guarantees for IPW

Fix any $j \in [M]$. Recall the definitions of the parameter $\mathrm{ATE}_{\cdot,j}$ and corresponding inverse probability weighting estimate $\widehat{\mathrm{ATE}}^{\mathrm{IPW}}_{\cdot,j}$ from Eqs. (4.5) and (4.10), respectively. The error $\Delta\mathrm{ATE}^{\mathrm{IPW}}_{\cdot,j} = \widehat{\mathrm{ATE}}^{\mathrm{IPW}}_{\cdot,j} - \mathrm{ATE}_{\cdot,j}$ can be re-expressed as

$$\begin{aligned}\Delta\mathrm{ATE}^{\mathrm{IPW}}_{\cdot,j} &= \frac{1}{N}\sum_{i\in[N]}\Big(\frac{y_{i,j}a_{i,j}}{\widehat{p}_{i,j}} - \frac{y_{i,j}(1-a_{i,j})}{1-\widehat{p}_{i,j}}\Big) - \frac{1}{N}\sum_{i\in[N]}\Big(\theta^{(1)}_{i,j}-\theta^{(0)}_{i,j}\Big)\\ &= \frac{1}{N}\sum_{i\in[N]}\Big(\Big(\frac{y_{i,j}a_{i,j}}{\widehat{p}_{i,j}} - \theta^{(1)}_{i,j}\Big) - \Big(\frac{y_{i,j}(1-a_{i,j})}{1-\widehat{p}_{i,j}} - \theta^{(0)}_{i,j}\Big)\Big)\\ &\overset{(a)}{=} \frac{1}{N}\sum_{i\in[N]}\Big(\mathbb{T}^{(1,\mathrm{IPW})}_{i,j} + \mathbb{T}^{(0,\mathrm{IPW})}_{i,j}\Big),\end{aligned} \tag{4.50}$$

where (a) follows after defining $\mathbb{T}^{(1,\mathrm{IPW})}_{i,j} \triangleq y_{i,j}a_{i,j}/\widehat{p}_{i,j} - \theta^{(1)}_{i,j}$ and $\mathbb{T}^{(0,\mathrm{IPW})}_{i,j} \triangleq \theta^{(0)}_{i,j} - y_{i,j}(1-a_{i,j})/(1-\widehat{p}_{i,j})$. Then, we have

$$\begin{aligned}\mathbb{T}^{(1,\mathrm{IPW})}_{i,j} &= \frac{y_{i,j}a_{i,j}}{\widehat{p}_{i,j}} - \theta^{(1)}_{i,j}\\ &\overset{(a)}{=} \frac{\big(\theta^{(1)}_{i,j}+\varepsilon^{(1)}_{i,j}\big)\big(p_{i,j}+\eta_{i,j}\big)}{\widehat{p}_{i,j}} - \theta^{(1)}_{i,j}\\ &= \theta^{(1)}_{i,j}\Big(\frac{p_{i,j}+\eta_{i,j}}{\widehat{p}_{i,j}} - 1\Big) + \varepsilon^{(1)}_{i,j}\Big(\frac{p_{i,j}+\eta_{i,j}}{\widehat{p}_{i,j}}\Big)\\ &= \frac{\theta^{(1)}_{i,j}\big(p_{i,j}-\widehat{p}_{i,j}\big)}{\widehat{p}_{i,j}} + \frac{\theta^{(1)}_{i,j}\eta_{i,j}}{\widehat{p}_{i,j}} + \frac{\varepsilon^{(1)}_{i,j}p_{i,j}}{\widehat{p}_{i,j}} + \frac{\varepsilon^{(1)}_{i,j}\eta_{i,j}}{\widehat{p}_{i,j}},\end{aligned} \tag{4.51}$$

where (a) follows from Eqs. (4.1) to (4.3). A similar derivation for $a=0$ implies that

$$\begin{aligned}
\mathbb{T}^{(0,\mathrm{IPW})}_{i,j} &= \theta^{(0)}_{i,j} - \frac{y_{i,j}(1-a_{i,j})}{1-\widehat{p}_{i,j}}\\
&= -\frac{\theta^{(0)}_{i,j}\big(1-p_{i,j}-\big(1-\widehat{p}_{i,j}\big)\big)}{1-\widehat{p}_{i,j}} - \frac{\theta^{(0)}_{i,j}(-\eta_{i,j})}{1-\widehat{p}_{i,j}} - \frac{\varepsilon^{(0)}_{i,j}\big(1-p_{i,j}\big)}{1-\widehat{p}_{i,j}} - \frac{\varepsilon^{(0)}_{i,j}(-\eta_{i,j})}{1-\widehat{p}_{i,j}}\\
&= \frac{\theta^{(0)}_{i,j}\big(p_{i,j}-\widehat{p}_{i,j}\big)}{1-\widehat{p}_{i,j}} + \frac{\theta^{(0)}_{i,j}\eta_{i,j}}{1-\widehat{p}_{i,j}} - \frac{\varepsilon^{(0)}_{i,j}\big(1-p_{i,j}\big)}{1-\widehat{p}_{i,j}} + \frac{\varepsilon^{(0)}_{i,j}\eta_{i,j}}{1-\widehat{p}_{i,j}}.
\end{aligned}$$

Consider any $a \in \{0,1\}$ and $\delta \in (0,1)$. We claim that, with probability at least $1-6\delta$,

$$\frac{1}{N}\Big|\sum_{i\in[N]}\mathbb{T}^{(\mathrm{a,IPW})}_{i,j}\Big| \le \frac{2}{\bar{\lambda}}\|\Theta^{(a)}\|_{\max}\,\mathcal{E}\big(\widehat{P}\big) + \frac{2\sqrt{c\ell_\delta}}{\bar{\lambda}\sqrt{\ell_1 N}}\|\Theta^{(a)}\|_{\max} + \frac{2\bar{\sigma}\sqrt{c\ell_\delta}}{\bar{\lambda}\sqrt{N}} + \frac{2\bar{\sigma}m(c\ell_\delta)}{\bar{\lambda}\sqrt{\ell_1 N}}. \quad (4.52)$$

where recall that $m(c\ell_\delta) = \max\big(c\ell_\delta, \sqrt{c\ell_\delta}\big)$. We provide a proof of this claim at the end of this section. Applying triangle inequality in Eq. (4.50) and using Eq. (4.52) with a union bound, we obtain that

$$\big|\Delta\mathrm{ATE}^{\mathrm{IPW}}_{\cdot,j}\big| \le \frac{2}{\bar{\lambda}}\theta_{\max}\,\mathcal{E}\big(\widehat{P}\big) + \frac{2\sqrt{c\ell_\delta}}{\bar{\lambda}\sqrt{\ell_1 N}}\theta_{\max} + \frac{4\bar{\sigma}\sqrt{c\ell_\delta}}{\bar{\lambda}\sqrt{N}} + \frac{4\bar{\sigma}m(c\ell_\delta)}{\bar{\lambda}\sqrt{\ell_1 N}},$$

with probability at least $1-12\delta$. The claim in Eq. (4.20) follows by re-parameterizing δ.

Proof of Eq. (4.52). This proof follows a very similar road map to that used for establishing the inequality in Eq. (4.40). Recall the partitioning of the units $[N]$ into $\mathcal{R}_0$ and $\mathcal{R}_1$ from Assumption 4.4. Now, to enable the application of concentration bounds, we split the summation over $i \in [N]$ in the left hand side of Eq. (4.52) into two parts—one over $i \in \mathcal{R}_0$ and the other over $i \in \mathcal{R}_1$—such that the noise terms are independent of the estimates of $\Theta^{(0)}, \Theta^{(1)}, P$ in each of these parts as in Eqs. (4.14) and (4.15).

Fix $a=1$ and note that $|\sum_{i\in[N]}\mathbb{T}^{(1,\mathrm{IPW})}_{i,j}| \le |\sum_{i\in\mathcal{R}_0}\mathbb{T}^{(1,\mathrm{IPW})}_{i,j}| + |\sum_{i\in\mathcal{R}_1}\mathbb{T}^{(1,\mathrm{IPW})}_{i,j}|$. Fix any $s \in \{0,1\}$. Then, Eq. (4.51) and triangle inequality imply that

$$\Big|\sum_{i\in\mathcal{R}_s}\mathbb{T}^{(1,\mathrm{IPW})}_{i,j}\Big| \le \Big|\sum_{i\in\mathcal{R}_s}\frac{\theta^{(1)}_{i,j}\big(p_{i,j}-\widehat{p}_{i,j}\big)}{\widehat{p}_{i,j}}\Big| + \Big|\sum_{i\in\mathcal{R}_s}\frac{\theta^{(1)}_{i,j}\eta_{i,j}}{\widehat{p}_{i,j}}\Big| + \Big|\sum_{i\in\mathcal{R}_s}\frac{\varepsilon^{(1)}_{i,j}p_{i,j}}{\widehat{p}_{i,j}}\Big| + \Big|\sum_{i\in\mathcal{R}_s}\frac{\varepsilon^{(1)}_{i,j}\eta_{i,j}}{\widehat{p}_{i,j}}\Big|. \quad (4.53)$$

Next, note that the decomposition in Eq. (4.53) is identical to the one in Eq. (4.41), except for the fact when compared to Eq. (4.41), the first two terms in Eq. (4.53) have a factor of $\theta^{(1)}_{i,j}$ instead of $\big(\widehat{\theta}^{(1)}_{i,j} - \theta^{(1)}_{i,j}\big)$. As a result, mimicking steps used to derive Eq. (4.46), we obtain the following bound, with probability at least $1-3\delta$,

$$\frac{1}{N}\Big|\sum_{i\in\mathcal{R}_s}\mathbb{T}^{(1,\mathrm{IPW})}_{i,j}\Big| \le \frac{1}{\bar{\lambda}N}\|\Theta^{(1)}\|_{1,2}\|\widehat{P}-P\|_{1,2} + \frac{\sqrt{c\ell_\delta}}{\bar{\lambda}\sqrt{\ell_1}N}\|\Theta^{(1)}\|_{1,2} + \frac{\bar{\sigma}\sqrt{c\ell_\delta}}{\bar{\lambda}N}\|P\|_{1,2} + \frac{\bar{\sigma}m(c\ell_\delta)}{\bar{\lambda}\sqrt{\ell_1}N}\|\mathbf{1}\|_{1,2}$$

$$\overset{(a)}{\leq} \frac{1}{\bar{\lambda}\sqrt{N}}\|\Theta^{(1)}\|_{\max}\|\widehat{P}-P\|_{1,2}+\frac{\sqrt{c\ell_\delta}}{\bar{\lambda}\sqrt{\ell_1 N}}\|\Theta^{(1)}\|_{\max}+\frac{\overline{\sigma}\sqrt{c\ell_\delta}}{\bar{\lambda}\sqrt{N}}+\frac{\overline{\sigma}m(c\ell_\delta)}{\bar{\lambda}\sqrt{\ell_1 N}}$$

$$\overset{(b)}{\leq} \frac{1}{\bar{\lambda}}\|\Theta^{(1)}\|_{\max}\mathcal{E}\big(\widehat{P}\big)+\frac{\sqrt{c\ell_\delta}}{\bar{\lambda}\sqrt{\ell_1 N}}\|\Theta^{(1)}\|_{\max}+\frac{\overline{\sigma}\sqrt{c\ell_\delta}}{\bar{\lambda}\sqrt{N}}+\frac{\overline{\sigma}m(c\ell_\delta)}{\bar{\lambda}\sqrt{\ell_1 N}}, \tag{4.54}$$

where (a) follows because $\|\Theta^{(1)}\|_{1,2} \leq \sqrt{N}\|\Theta^{(1)}\|_{\max}$, $\|P\|_{1,2} \leq \sqrt{N}$ and $\|\mathbf{1}\|_{1,2} = \sqrt{N}$, and (b) follows from Eq. (4.16). Then, the claim in Eq. (4.52) follows for $a = 1$ by using Eq. (4.54) and applying a union bound over $s \in \{0, 1\}$. The proof of Eq. (4.52) for $a = 0$ follows similarly.

4.F Proof of Theorem 4.2: Asymptotic Normality for DR

For every $(i, j) \in [N] \times [M]$, recall the definitions of $\mathbb{T}^{(1,\mathrm{DR})}_{i,j}$ and $\mathbb{T}^{(0,\mathrm{DR})}_{i,j}$ from Eq. (4.38) and Eq. (4.39), respectively. Then, define

$$\mathbb{X}^{(1,\mathrm{DR})}_{i,j} \triangleq \mathbb{T}^{(1,\mathrm{DR})}_{i,j} - \varepsilon^{(1)}_{i,j} - \frac{\varepsilon^{(1)}_{i,j}\eta_{i,j}}{p_{i,j}} \tag{4.55}$$

$$\mathbb{X}^{(0,\mathrm{DR})}_{i,j} \triangleq \mathbb{T}^{(0,\mathrm{DR})}_{i,j} + \varepsilon^{(0)}_{i,j} - \frac{\varepsilon^{(0)}_{i,j}\eta_{i,j}}{1-p_{i,j}},$$

and

$$\mathbb{Z}^{\mathrm{DR}}_{i,j} \triangleq \varepsilon^{(1)}_{i,j} + \frac{\varepsilon^{(1)}_{i,j}\eta_{i,j}}{p_{i,j}} - \varepsilon^{(0)}_{i,j} + \frac{\varepsilon^{(0)}_{i,j}\eta_{i,j}}{1-p_{i,j}}. \tag{4.56}$$

Then, $\Delta\mathrm{ATE}^{\mathrm{DR}}_{\cdot,j}$ in Eq. (4.36) can be expressed as

$$\Delta\mathrm{ATE}^{\mathrm{DR}}_{\cdot,j} = \frac{1}{N}\sum_{i\in[N]}\Big(\mathbb{X}^{(1,\mathrm{DR})}_{i,j} + \mathbb{X}^{(0,\mathrm{DR})}_{i,j} + \mathbb{Z}^{\mathrm{DR}}_{i,j}\Big).$$

We obtain the following convergence results.

Lemma 4.8 (Convergence of $\mathbb{X}^{\mathrm{DR}}_{j}$)**.** *Fix any $j \in [M]$. Suppose Assumptions 4.1 to 4.4 and conditions (C1) to (C3) in Theorem 4.2 hold. Then,*

$$\frac{1}{\overline{\sigma}_j\sqrt{N}}\sum_{i\in[N]}\Big(\mathbb{X}^{(1,\mathrm{DR})}_{i,j} + \mathbb{X}^{(0,\mathrm{DR})}_{i,j}\Big) = o_p(1).$$

Lemma 4.9 (Convergence of $\mathbb{Z}^{\mathrm{DR}}_{j}$)**.** *Fix any $j \in [M]$. Suppose Assumptions 4.1 and 4.2 hold and condition (C3) in Theorem 4.2 hold. Then,*

$$\frac{1}{\overline{\sigma}_j\sqrt{N}}\sum_{i\in[N]}\mathbb{Z}^{\mathrm{DR}}_{i,j} \xrightarrow{d} \mathcal{N}(0,1).$$

Now, the result in Theorem 4.2 follows from Slutsky's theorem.

4.F.1 Proof of Lemma 4.8

Fix any $j \in [M]$. Consider any $a \in \{0,1\}$. We claim that

$$\frac{1}{\sqrt{N}} \sum_{i\in[N]} \mathbb{X}_{i,j}^{(a,\mathrm{DR})} \leq O\Big(\sqrt{N}\mathcal{E}\big(\widehat{\Theta}^{(a)}\big)\mathcal{E}\big(\widehat{P}\big)\Big) + o_p(1). \tag{4.57}$$

We provide a proof of this claim at the end of this section. Then, using Eq. (4.57) and the fact that $\overline{\sigma}_j \geq c > 0$ as per condition (C3), we obtain the following,

$$\begin{aligned}\frac{1}{\overline{\sigma}_j\sqrt{N}} \sum_{i\in[N]}\Big(\mathbb{X}_{i,j}^{(1,\mathrm{DR})} + \mathbb{X}_{i,j}^{(0,\mathrm{DR})}\Big) &\leq \frac{1}{c}\Big(O\Big(\sqrt{N}\mathcal{E}\big(\widehat{\Theta}\big)\mathcal{E}\big(\widehat{P}\big)\Big) + o_p(1)\Big) \\ &\overset{(a)}{=} \frac{1}{c}\Big(\sqrt{N}o_p(N^{-1/2}) + o_p(1)\Big) \overset{(b)}{=} o_p(1),\end{aligned}$$

where (a) follows from (C2), and (b) follows because $o_p(1) + o_p(1) = o_p(1)$.

Proof of Eq. (4.57) Recall the partitioning of the units $[N]$ into $\mathcal{R}_0$ and $\mathcal{R}_1$ from Assumption 4.4. Now, to enable the application of concentration bounds, we split the summation over $i \in [N]$ in the left hand side of Eq. (4.57) into two parts—one over $i \in \mathcal{R}_0$ and the other over $i \in \mathcal{R}_1$—such that the noise terms are independent of the estimates of $\Theta^{(0)}, \Theta^{(1)}, P$ in each of these parts as in Eqs. (4.14) and (4.15).

Fix $a = 1$. Then, Eqs. (4.38) and (4.55) imply that

$$\begin{aligned}\mathbb{X}_{i,j}^{(1,\mathrm{DR})} &= \frac{\big(\widehat{\theta}_{i,j}^{(1)} - \theta_{i,j}^{(1)}\big)\big(\widehat{p}_{i,j} - p_{i,j}\big)}{\widehat{p}_{i,j}} - \frac{\big(\widehat{\theta}_{i,j}^{(1)} - \theta_{i,j}^{(1)}\big)\eta_{i,j}}{\widehat{p}_{i,j}} + \frac{\varepsilon_{i,j}^{(1)}p_{i,j}}{\widehat{p}_{i,j}} + \frac{\varepsilon_{i,j}^{(1)}\eta_{i,j}}{\widehat{p}_{i,j}} - \varepsilon_{i,j}^{(1)} - \frac{\varepsilon_{i,j}^{(1)}\eta_{i,j}}{p_{i,j}} \\ &= \frac{\big(\widehat{\theta}_{i,j}^{(1)} - \theta_{i,j}^{(1)}\big)\big(\widehat{p}_{i,j} - p_{i,j}\big)}{\widehat{p}_{i,j}} - \frac{\big(\widehat{\theta}_{i,j}^{(1)} - \theta_{i,j}^{(1)}\big)\eta_{i,j}}{\widehat{p}_{i,j}} - \frac{\varepsilon_{i,j}^{(1)}\big(\widehat{p}_{i,j} - p_{i,j}\big)}{\widehat{p}_{i,j}} - \frac{\varepsilon_{i,j}^{(1)}\eta_{i,j}\big(\widehat{p}_{i,j} - p_{i,j}\big)}{\widehat{p}_{i,j}p_{i,j}}.\end{aligned}$$

Now, note that $|\sum_{i\in[N]}\mathbb{X}_{i,j}^{(1,\mathrm{DR})}| \leq |\sum_{i\in\mathcal{R}_0}\mathbb{X}_{i,j}^{(1,\mathrm{DR})}| + |\sum_{i\in\mathcal{R}_1}\mathbb{X}_{i,j}^{(1,\mathrm{DR})}|$. Fix any $s \in \{0,1\}$. Then, triangle inequality implies that

$$\begin{aligned}\frac{1}{\sqrt{N}}\Big|\sum_{i\in\mathcal{R}_s}\mathbb{X}_{i,j}^{(1,\mathrm{DR})}\Big| &\leq \frac{1}{\sqrt{N}}\Big|\sum_{i\in\mathcal{R}_s}\frac{\big(\widehat{\theta}_{i,j}^{(1)} - \theta_{i,j}^{(1)}\big)\big(\widehat{p}_{i,j} - p_{i,j}\big)}{\widehat{p}_{i,j}}\Big| + \frac{1}{\sqrt{N}}\Big|\sum_{i\in\mathcal{R}_s}\frac{\big(\widehat{\theta}_{i,j}^{(1)} - \theta_{i,j}^{(1)}\big)\eta_{i,j}}{\widehat{p}_{i,j}}\Big| \\ &+ \frac{1}{\sqrt{N}}\Big|\sum_{i\in\mathcal{R}_s}\frac{\varepsilon_{i,j}^{(1)}\big(\widehat{p}_{i,j} - p_{i,j}\big)}{\widehat{p}_{i,j}}\Big| + \frac{1}{\sqrt{N}}\Big|\sum_{i\in\mathcal{R}_s}\frac{\varepsilon_{i,j}^{(1)}\eta_{i,j}\big(\widehat{p}_{i,j} - p_{i,j}\big)}{\widehat{p}_{i,j}p_{i,j}}\Big|. \end{aligned}\tag{4.58}$$

To control the first term in Eq. (4.58), we use the Cauchy-Schwarz inequality and Assumption 4.3 as in Appendix 4.B (see Eqs. (4.42), (4.46), and (4.47)).

To control the second term in Eq. (4.58), we condition on $\{(\widehat{p}_{i,j}, \widehat{\theta}_{i,j}^{(1)})\}_{i\in\mathcal{R}_s}$. Then, Assumption 4.4 (i.e., Eq. (4.14)) provides that $\{(\widehat{p}_{i,j}, \widehat{\theta}_{i,j}^{(1)})\}_{i\in\mathcal{R}_s} \perp\!\!\!\perp \{\eta_{i,j}\}_{i\in\mathcal{R}_s}$. As a result, $\sum_{i\in\mathcal{R}_s}\big(\widehat{\theta}_{i,j}^{(1)} - \theta_{i,j}^{(1)}\big)\eta_{i,j}/\widehat{p}_{i,j}$ is subGaussian$\big(\big[\sum_{i\in\mathcal{R}_s}\big(\widehat{\theta}_{i,j}^{(1)} - \theta_{i,j}^{(1)}\big)^2/\big(\widehat{p}_{i,j}\big)^2\big]^{1/2}/\sqrt{\ell_1}\big)$ because

$\eta_{i,j}$ is subGaussian$(1/\sqrt{\ell_1})$ (see Example 2.5.8 in Vershynin (2018)) as well as zero-mean and independent across all $i \in [N]$ due to Assumption 4.2(a). Then, we have

$$\begin{aligned}\frac{1}{\sqrt{N}}\mathbb{E}\Big[\Big|\sum_{i\in\mathcal{R}_s}\frac{\big(\widehat{\theta}^{(1)}_{i,j}-\theta^{(1)}_{i,j}\big)\eta_{i,j}}{\widehat{p}_{i,j}}\Big|\Big|\{(\widehat{p}_{i,j},\widehat{\theta}^{(1)}_{i,j})\}_{i\in\mathcal{R}_s}\Big] &\overset{(a)}{\le} \frac{c}{\sqrt{N}}\sqrt{\sum_{i\in\mathcal{R}_s}\Big(\frac{\widehat{\theta}^{(1)}_{i,j}-\theta^{(1)}_{i,j}}{\widehat{p}_{i,j}}\Big)^2}\\ &\le \frac{c}{\sqrt{N}}\big\|\big(\widehat{\Theta}^{(1)}_{\cdot,j}-\Theta^{(1)}_{\cdot,j}\big)\oslash\widehat{P}_{\cdot,j}\big\|_2\\ &\overset{(b)}{\le} \frac{c}{\bar{\lambda}}\mathcal{E}\big(\widehat{\Theta}^{(1)}\big)\le\frac{c}{\bar{\lambda}}\mathcal{E}\big(\widehat{\Theta}\big)\overset{(c)}{=}o_p(1), \end{aligned} \tag{4.59}$$

where (a) follows as the first moment of subGaussian(σ) is $O(\sigma)$, (b) follows from Assumption 4.3 and Eq. (4.16), and (c) follows from (C1).

To control the third term in Eq. (4.58), we condition on $\{\widehat{p}_{i,j}\}_{i\in\mathcal{R}_s}$. Then, Assumption 4.4 (i.e., Eq. (4.15)) provides that $\{\widehat{p}_{i,j}\}_{i\in\mathcal{R}_s} \perp\!\!\!\perp \{\varepsilon^{(1)}_{i,j}\}_{i\in\mathcal{R}_s}$. As a result, $\sum_{i\in\mathcal{R}_s}\varepsilon^{(1)}_{i,j}\big(\widehat{p}_{i,j}-p_{i,j}\big)/\widehat{p}_{i,j}$ is subGaussian$\big(\bar{\sigma}\big[\sum_{i\in\mathcal{R}_s}\big(\widehat{p}_{i,j}-p_{i,j}\big)^2/\big(\widehat{p}_{i,j}\big)^2\big]^{1/2}\big)$ because $\varepsilon^{(1)}_{i,j}$ is subGaussian$(\bar{\sigma})$, zero-mean, and independent across all $i \in [N]$ due to Assumption 4.2. Then, we have

$$\begin{aligned}\frac{1}{\sqrt{N}}\mathbb{E}\Big[\Big|\sum_{i\in\mathcal{R}_s}\frac{\varepsilon^{(1)}_{i,j}\big(\widehat{p}_{i,j}-p_{i,j}\big)}{\widehat{p}_{i,j}}\Big|\Big|\{\widehat{p}_{i,j}\}_{i\in\mathcal{R}_s}\Big] &\overset{(a)}{\le} \frac{c\bar{\sigma}}{\sqrt{N}}\sqrt{\sum_{i\in\mathcal{R}_s}\Big(\frac{\widehat{p}_{i,j}-p_{i,j}}{\widehat{p}_{i,j}}\Big)^2}\\ &\le \frac{c\bar{\sigma}}{\sqrt{N}}\big\|\big(\widehat{P}_{\cdot,j}-P_{\cdot,j}\big)\oslash\widehat{P}_{\cdot,j}\big\|_2\\ &\overset{(b)}{\le} \frac{c\bar{\sigma}}{\bar{\lambda}}\mathcal{E}\big(\widehat{P}\big)\overset{(c)}{=}o_p(1), \end{aligned} \tag{4.60}$$

where (a) follows as the first moment of subGaussian(σ) is $O(\sigma)$, (b) follows from Assumption 4.3 and Eq. (4.16), and (c) follows from (C1).

To control the fourth term in Eq. (4.58), we condition on $\{\widehat{p}_{i,j}\}_{i\in\mathcal{R}_s}$. Then, Assumption 4.4 (i.e., Eq. (4.15)) provides that $\{\widehat{p}_{i,j}\}_{i\in\mathcal{R}_s} \perp\!\!\!\perp \{(\eta_{i,j},\varepsilon^{(1)}_{i,j})\}_{i\in\mathcal{R}_s}$. As a result, $\sum_{i\in\mathcal{R}_s}\varepsilon^{(1)}_{i,j}\eta_{i,j}\big(\widehat{p}_{i,j}-p_{i,j}\big)/\widehat{p}_{i,j}p_{i,j}$ is subExponential$\big(\bar{\sigma}\big[\sum_{i\in\mathcal{R}_s}\big(\widehat{p}_{i,j}-p_{i,j}\big)^2/\big(\widehat{p}_{i,j}p_{i,j}\big)^2\big]^{1/2}/\sqrt{\ell_1}\big)$ because $\varepsilon^{(1)}_{i,j}\eta_{i,j}$ is subExponential$(\bar{\sigma}/\sqrt{\ell_1})$ due to Lemma 4.3 as well as zero-mean and independent across all $i \in [N]$ due to Assumption 4.2. Then, we have

$$\begin{aligned}\frac{1}{\sqrt{N}}\mathbb{E}\Big[\Big|\sum_{i\in\mathcal{R}_s}\frac{\varepsilon^{(1)}_{i,j}\eta_{i,j}\big(\widehat{p}_{i,j}-p_{i,j}\big)}{\widehat{p}_{i,j}p_{i,j}}\Big|\Big|\{\widehat{p}_{i,j}\}_{i\in\mathcal{R}_s}\Big] &\overset{(a)}{\le} \frac{c\bar{\sigma}}{\sqrt{N}}\sqrt{\sum_{i\in\mathcal{R}_s}\Big(\frac{\widehat{p}_{i,j}-p_{i,j}}{\widehat{p}_{i,j}p_{i,j}}\Big)^2}\\ &\le \frac{c\bar{\sigma}}{\sqrt{N}}\big\|\big(\widehat{P}_{\cdot,j}-P_{\cdot,j}\big)\oslash\big(\widehat{P}_{\cdot,j}\odot P_{\cdot,j}\big)\big\|_2\\ &\overset{(b)}{\le} \frac{c\bar{\sigma}}{\bar{\lambda}\lambda}\mathcal{E}\big(\widehat{P}\big)\overset{(c)}{=}o_p(1), \end{aligned} \tag{4.61}$$

where (a) follows as the first moment of subExponential(σ) is $O(\sigma)$, (b) follows from Assumption 4.3 and Eq. (4.16), and (c) follows from (C1).

Putting together Eqs. (4.58) to (4.61) using Lemma 4.6, we have

$$\frac{1}{\sqrt{N}}\Big|\sum_{i\in\mathcal{R}_s}\mathbb{X}_{i,j}^{(1,\mathrm{DR})}\Big| \leq O\Big(\sqrt{N}\mathcal{E}\big(\widehat{\Theta}^{(1)}\big)\mathcal{E}\big(\widehat{P}\big)\Big)+o_p(1).$$

Then, the claim in Eq. (4.57) follows for $a=1$ by using $|\sum_{i\in[N]}\mathbb{X}_{i,j}^{(1,\mathrm{DR})}| \leq |\sum_{i\in\mathcal{R}_0}\mathbb{X}_{i,j}^{(1,\mathrm{DR})}| + |\sum_{i\in\mathcal{R}_1}\mathbb{X}_{i,j}^{(1,\mathrm{DR})}|$. The proof of Eq. (4.57) for $a=0$ follows similarly.

4.F.2 Proof of Lemma 4.9

To prove this result, we invoke Lyapunov central limit theorem (CLT).

Lemma 4.10 (Lyapunov CLT, see Theorem 27.3 of Billingsley (2017))**.** *Consider a sequence $x_1, x_2, \cdots$ of mean-zero independent random variables such that the moments $\mathbb{E}[|x_i|^{2+\omega}]$ are finite for some $\omega > 0$. Moreover, assume that the Lyapunov's condition is satisfied, i.e.,*

$$\sum_{i=1}^{N}\mathbb{E}[|x_i|^{2+\omega}]\Big/\Big(\sum_{i=1}^{N}\mathbb{E}[x_i^2]\Big)^{\frac{2+\omega}{2}} \longrightarrow 0, \tag{4.62}$$

as $N\to\infty$. Then,

$$\sum_{i=1}^{N}x_i\Big/\Big(\sum_{i=1}^{N}\mathbb{E}[x_i^2]\Big)^{\frac{1}{2}} \xrightarrow{d} \mathcal{N}(0,1),$$

as $N\to\infty$.

Fix any $j\in[M]$. We apply Lyapunov CLT in Lemma 4.10 on the sequence $\mathbb{Z}_{1,j}^{\mathrm{DR}}, \mathbb{Z}_{2,j}^{\mathrm{DR}}, \cdots$ where $\mathbb{Z}_{i,j}^{\mathrm{DR}}$ is as defined in Eq. (4.56). Note that this sequence is zero-mean from Assumption 4.2(a) and Assumption 4.2(b), and independent from Assumption 4.2(b). First, we show in Appendix 4.F.2.1 that

$$\mathbb{V}\mathrm{ar}(\mathbb{Z}_{i,j}^{\mathrm{DR}}) = \frac{(\sigma_{i,j}^{(1)})^2}{p_{i,j}} + \frac{(\sigma_{i,j}^{(0)})^2}{1-p_{i,j}}, \tag{4.63}$$

for each $i\in[N]$. Next, we show in Appendix 4.F.2.2 that Lyapunov's condition Eq. (4.62) holds for the sequence $\mathbb{Z}_{1,j}^{\mathrm{DR}}, \mathbb{Z}_{2,j}^{\mathrm{DR}}, \cdots$ with $\omega = 1$. Finally, applying Lemma 4.10 and using the definition of $\overline{\sigma}_j$ from Eq. (4.22) yields Lemma 4.9.

4.F.2.1 Proof of Eq. (4.63)

Fix any $i\in[N]$ and consider $\mathbb{V}\mathrm{ar}(\mathbb{Z}_{i,j}^{\mathrm{DR}})$. We have

$$\mathbb{V}\mathrm{ar}\Big(\mathbb{Z}_{i,j}^{\mathrm{DR}}\Big) = \mathbb{V}\mathrm{ar}\bigg(\varepsilon_{i,j}^{(1)}\Big(1+\frac{\eta_{i,j}}{p_{i,j}}\Big) - \varepsilon_{i,j}^{(0)}\Big(1-\frac{\eta_{i,j}}{1-p_{i,j}}\Big)\bigg). \tag{4.64}$$

We claim the following:

$$\mathbb{V}\mathrm{ar}\Big(\varepsilon_{i,j}^{(1)}\Big(1+\frac{\eta_{i,j}}{p_{i,j}}\Big)\Big) = \frac{(\sigma_{i,j}^{(1)})^2}{p_{i,j}}, \tag{4.65}$$

$$\mathbb{V}\mathrm{ar}\Big(\varepsilon_{i,j}^{(0)}\Big(1-\frac{\eta_{i,j}}{1-p_{i,j}}\Big)\Big) = \frac{(\sigma_{i,j}^{(0)})^2}{1-p_{i,j}}, \tag{4.66}$$

and

$$\mathbb{C}\mathrm{ov}\Big(\varepsilon_{i,j}^{(1)}\Big(1+\frac{\eta_{i,j}}{p_{i,j}}\Big), \varepsilon_{i,j}^{(0)}\Big(1-\frac{\eta_{i,j}}{1-p_{i,j}}\Big)\Big) = 0, \tag{4.67}$$

with Eq. (4.63) following from Eqs. (4.64) to (4.67).

To establish Eq. (4.65), notice that Assumption 4.2(a) and (b) imply $\varepsilon_{i,j}^{(1)} \perp\!\!\!\perp \eta_{i,j}$ and $\mathbb{E}[\varepsilon_{i,j}^{(1)}] = \mathbb{E}[\eta_{i,j}] = 0$ so that $\mathbb{E}[\varepsilon_{i,j}^{(1)}(1+\eta_{i,j}/p_{i,j})] = 0$. Then,

$$\begin{aligned}\mathbb{V}\mathrm{ar}\Big(\varepsilon_{i,j}^{(1)}\Big(1+\frac{\eta_{i,j}}{p_{i,j}}\Big)\Big) &= \mathbb{E}\Big[\Big(\varepsilon_{i,j}^{(1)}\Big(1+\frac{\eta_{i,j}}{p_{i,j}}\Big)\Big)^2\Big] = \mathbb{E}\Big[\Big(\varepsilon_{i,j}^{(1)}\Big)^2\Big]\mathbb{E}\Big[\Big(1+\frac{\eta_{i,j}}{p_{i,j}}\Big)^2\Big] \\ &= \mathbb{E}\Big[\Big(\varepsilon_{i,j}^{(1)}\Big)^2\Big]\Big[1+\mathbb{E}\Big[\frac{\eta_{i,j}^2}{p_{i,j}^2}\Big]\Big] \overset{(a)}{=} (\sigma_{i,j}^{(1)})^2\Big[1+\frac{p_{i,j}(1-p_{i,j})}{p_{i,j}^2}\Big] \\ &= \frac{(\sigma_{i,j}^{(1)})^2}{p_{i,j}},\end{aligned}$$

where (a) follows because $\mathbb{E}[\eta_{i,j}^2] = \mathbb{V}\mathrm{ar}(\eta_{i,j}) = p_{i,j}(1-p_{i,j})$ from Eq. (4.3), and $\mathbb{E}\big[(\varepsilon_{i,j}^{(1)})^2\big] = \mathbb{V}\mathrm{ar}(\varepsilon_{i,j}^{(1)}) = (\sigma_{i,j}^{(1)})^2$ from condition (C3). A similar argument establishes Eq. (4.66). Eq. (4.67) follows from,

$$\begin{aligned}\mathbb{C}\mathrm{ov}\Big(\varepsilon_{i,j}^{(1)}\Big(1+\frac{\eta_{i,j}}{p_{i,j}}\Big), \varepsilon_{i,j}^{(0)}\Big(1-\frac{\eta_{i,j}}{1-p_{i,j}}\Big)\Big) &= \mathbb{E}\Big[\varepsilon_{i,j}^{(1)}\Big(1+\frac{\eta_{i,j}}{p_{i,j}}\Big)\varepsilon_{i,j}^{(0)}\Big(1-\frac{\eta_{i,j}}{1-p_{i,j}}\Big)\Big] \\ &\overset{(a)}{=} \mathbb{E}\Big[\Big(1+\frac{\eta_{i,j}}{p_{i,j}}\Big)\Big(1-\frac{\eta_{i,j}}{1-p_{i,j}}\Big)\Big]\mathbb{E}[\varepsilon_{i,j}^{(1)}\varepsilon_{i,j}^{(0)}] \\ &= \Big(1-\mathbb{E}\Big[\frac{\eta_{i,j}^2}{p_{i,j}(1-p_{i,j})}\Big]\Big)\mathbb{E}[\varepsilon_{i,j}^{(1)}\varepsilon_{i,j}^{(0)}] \\ &\overset{(b)}{=} 0\cdot\mathbb{E}[\varepsilon_{i,j}^{(1)}\varepsilon_{i,j}^{(0)}] = 0,\end{aligned}$$

where (a) follows because $(\varepsilon_{i,j}^{(0)}, \varepsilon_{i,j}^{(1)}) \perp\!\!\!\perp \eta_{i,j}$ from Assumption 4.2(b) and (b) follows because $\mathbb{E}[\eta_{i,j}^2] = \mathbb{V}\mathrm{ar}(\eta_{i,j}) = p_{i,j}(1-p_{i,j})$.

4.F.2.2 Proof of Lyapunov's condition with $\omega = 1$

We have

$$\frac{\sum_{i\in[N]}\mathbb{E}\big[|\mathbb{Z}_{i,j}^{\mathrm{DR}}|^3\big]}{\big(\sum_{i\in[N]}\mathbb{V}\mathrm{ar}(\mathbb{Z}_{i,j}^{\mathrm{DR}})\big)^{3/2}} = \frac{1}{N^{3/2}}\frac{\sum_{i\in[N]}\mathbb{E}\big[|\mathbb{Z}_{i,j}^{\mathrm{DR}}|^3\big]}{\big(\frac{1}{N}\sum_{i\in[N]}\mathbb{V}\mathrm{ar}(\mathbb{Z}_{i,j}^{\mathrm{DR}})\big)^{3/2}}$$

$$
\begin{aligned}
&\overset{(a)}{=} \frac{1}{N^{3/2}} \frac{\sum_{i\in[N]} \mathbb{E}\big[|\mathbb{Z}_{i,j}^{\mathrm{DR}}|^3\big]}{\big(\overline{\sigma}_j\big)^{3/2}} \\
&\overset{(b)}{\leq} \frac{1}{N^{3/2}} \frac{\sum_{i\in[N]} \mathbb{E}\big[|\mathbb{Z}_{i,j}^{\mathrm{DR}}|^3\big]}{c_1^{3/2}} \overset{(c)}{\leq} \frac{1}{N^{1/2}} \frac{c_2}{c_1^{3/2}},
\end{aligned} \tag{4.68}
$$

where (a) follows by putting together Eqs. (4.22) and (4.63), (b) follows because $\overline{\sigma}_j \geq c_1 > 0$ as per condition (C3), (c) follows because the absolute third moments of subExponential random variables are bounded, after noting that $\mathbb{Z}_{i,j}^{\mathrm{DR}}$ is a subExponential random variable. Then, condition Eq. (4.62) holds for $\omega = 1$ as the right hand side of Eq. (4.68) goes to zero as $N \to \infty$.

4.F.3 Proof of Proposition 4.2: Consistent variance estimation

Fix any $j \in [M]$ and recall the definitions of $\overline{\sigma}_j^2$ and $\widehat{\sigma}_j^2$ from Eqs. (4.22) and (4.25), respectively. The error $\Delta_j = \widehat{\sigma}_j^2 - \overline{\sigma}_j^2$ can be expressed as

$$
\begin{aligned}
\Delta_j &= \frac{1}{N} \sum_{i\in[N]} \bigg(\frac{\big(\widehat{\theta}_{i,j}^{(1)} - y_{i,j}\big)^2 a_{i,j}}{\big(\widehat{p}_{i,j}\big)^2} + \frac{\big(\widehat{\theta}_{i,j}^{(0)} - y_{i,j}\big)^2 (1-a_{i,j})}{\big(1-\widehat{p}_{i,j}\big)^2} \bigg) - \bigg(\frac{(\sigma_{i,j}^{(1)})^2}{p_{i,j}} + \frac{(\sigma_{i,j}^{(0)})^2}{1-p_{i,j}} \bigg) \\
&= \frac{1}{N} \sum_{i\in[N]} \bigg(\frac{\big(\widehat{\theta}_{i,j}^{(1)} - y_{i,j}\big)^2 a_{i,j}}{\big(\widehat{p}_{i,j}\big)^2} - \frac{(\sigma_{i,j}^{(1)})^2}{p_{i,j}} \bigg) + \bigg(\frac{\big(\widehat{\theta}_{i,j}^{(0)} - y_{i,j}\big)^2 (1-a_{i,j})}{\big(1-\widehat{p}_{i,j}\big)^2} - \frac{(\sigma_{i,j}^{(0)})^2}{1-p_{i,j}} \bigg) \\
&\overset{(a)}{=} \frac{1}{N} \sum_{i\in[N]} \Big(\mathbb{T}_{i,j}^{(1)} + \mathbb{T}_{i,j}^{(0)} \Big),
\end{aligned} \tag{4.69}
$$

where (a) follows after defining

$$
\mathbb{T}_{i,j}^{(1)} \triangleq \frac{\big(\widehat{\theta}_{i,j}^{(1)} - y_{i,j}\big)^2 a_{i,j}}{\big(\widehat{p}_{i,j}\big)^2} - \frac{(\sigma_{i,j}^{(1)})^2}{p_{i,j}} \quad \text{and} \quad \mathbb{T}_{i,j}^{(0)} \triangleq \frac{\big(\widehat{\theta}_{i,j}^{(0)} - y_{i,j}\big)^2 (1-a_{i,j})}{\big(1-\widehat{p}_{i,j}\big)^2} - \frac{(\sigma_{i,j}^{(0)})^2}{1-p_{i,j}}.
$$

for every $(i,j) \in [N] \times [M]$. Then, we have

$$
\begin{aligned}
\mathbb{T}_{i,j}^{(1)} &\overset{(a)}{=} \frac{\big(\widehat{\theta}_{i,j}^{(1)} - \theta_{i,j}^{(1)} - \varepsilon_{i,j}^{(1)}\big)^2 \big(p_{i,j} + \eta_{i,j}\big)}{\big(\widehat{p}_{i,j}\big)^2} - \frac{(\sigma_{i,j}^{(1)})^2}{p_{i,j}} \\
&= \frac{\big(\widehat{\theta}_{i,j}^{(1)} - \theta_{i,j}^{(1)}\big)^2 a_{i,j}}{\big(\widehat{p}_{i,j}\big)^2} - - \frac{2\varepsilon_{i,j}^{(1)} p_{i,j} \big(\widehat{\theta}_{i,j}^{(1)} - \theta_{i,j}^{(1)}\big)}{\big(\widehat{p}_{i,j}\big)^2} - \frac{2\varepsilon_{i,j}^{(1)} \eta_{i,j} \big(\widehat{\theta}_{i,j}^{(1)} - \theta_{i,j}^{(1)}\big)}{\big(\widehat{p}_{i,j}\big)^2} \\
&\qquad + \frac{\big(\varepsilon_{i,j}^{(1)}\big)^2 p_{i,j}}{\big(\widehat{p}_{i,j}\big)^2} + \frac{\big(\varepsilon_{i,j}^{(1)}\big)^2 \eta_{i,j}}{\big(\widehat{p}_{i,j}\big)^2} - \frac{(\sigma_{i,j}^{(1)})^2}{p_{i,j}},
\end{aligned}
$$

where (a) follows from Eqs. (4.1) to (4.3). A similar derivation for $a = 0$ implies that

$$
\mathbb{T}_{i,j}^{(0)} = \frac{\big(\widehat{\theta}_{i,j}^{(0)} - \theta_{i,j}^{(0)} - \varepsilon_{i,j}^{(0)}\big)^2 (1 - p_{i,j} - \eta_{i,j})}{\big(1-\widehat{p}_{i,j}\big)^2} - \frac{(\sigma_{i,j}^{(0)})^2}{1-p_{i,j}}
$$

$$= \frac{\big(\widehat{\theta}^{(0)}_{i,j} - \theta^{(0)}_{i,j}\big)^2 \big(1 - a_{i,j}\big)}{\big(1 - \widehat{p}_{i,j}\big)^2} - \frac{2\varepsilon^{(0)}_{i,j} \big(1 - p_{i,j}\big) \big(\widehat{\theta}^{(0)}_{i,j} - \theta^{(0)}_{i,j}\big)}{\big(1 - \widehat{p}_{i,j}\big)^2} + \frac{2\varepsilon^{(0)}_{i,j} \eta_{i,j} \big(\widehat{\theta}^{(0)}_{i,j} - \theta^{(0)}_{i,j}\big)}{\big(1 - \widehat{p}_{i,j}\big)^2}$$
$$+ \frac{\big(\varepsilon^{(0)}_{i,j}\big)^2 \big(1 - p_{i,j}\big)}{\big(1 - \widehat{p}_{i,j}\big)^2} - \frac{\big(\varepsilon^{(0)}_{i,j}\big)^2 \eta_{i,j}}{\big(1 - \widehat{p}_{i,j}\big)^2} - \frac{(\sigma^{(0)}_{i,j})^2}{1 - p_{i,j}}.$$

Consider any $a \in \{0, 1\}$. We claim that

$$\frac{1}{N} \Big| \sum_{i \in [N]} \mathbb{T}^{(\mathrm{a})}_{i,j} \Big| = o_p(1). \tag{4.70}$$

We provide a proof of this claim at the end of this section. Then, applying triangle inequality in Eq. (4.69), we obtain the following

$$\Delta_j \leq o_p(1) + o_p(1) \overset{(a)}{=} o_p(1),$$

where (a) follows because $o_p(1) + o_p(1) = o_p(1)$.

Proof of bound Eq. (4.70). This proof follows a very similar road map to that used for establishing the inequality in Eq. (4.57). Recall the partitioning of the units $[N]$ into $\mathcal{R}_0$ and $\mathcal{R}_1$ from Assumption 4.4. Now, to enable the application of concentration bounds, we split the summation over $i \in [N]$ in the left hand side of Eq. (4.70) into two parts—one over $i \in \mathcal{R}_0$ and the other over $i \in \mathcal{R}_1$—such that the noise terms are independent of the estimates of $\Theta^{(0)}, \Theta^{(1)}, P$ in each of these parts as in Eqs. (4.14) and (4.15).

Fix $a = 1$. Now, note that $|\sum_{i \in [N]} \mathbb{T}^{(1)}_{i,j}| \leq |\sum_{i \in \mathcal{R}_0} \mathbb{T}^{(1)}_{i,j}| + |\sum_{i \in \mathcal{R}_1} \mathbb{T}^{(1)}_{i,j}|$. Fix any $s \in \{0, 1\}$. Then, triangle inequality implies that

$$\frac{1}{N} \Big| \sum_{i \in \mathcal{R}_s} \mathbb{T}^{(1)}_{i,j} \Big| \leq \frac{1}{N} \Big| \sum_{i \in \mathcal{R}_s} \frac{\big(\widehat{\theta}^{(1)}_{i,j} - \theta^{(1)}_{i,j}\big)^2 a_{i,j}}{\big(\widehat{p}_{i,j}\big)^2} \Big| + \frac{1}{N} \Big| \sum_{i \in \mathcal{R}_s} \frac{2\varepsilon^{(1)}_{i,j} p_{i,j} \big(\widehat{\theta}^{(1)}_{i,j} - \theta^{(1)}_{i,j}\big)}{\big(\widehat{p}_{i,j}\big)^2} \Big|$$
$$+ \frac{1}{N} \Big| \sum_{i \in \mathcal{R}_s} \frac{2\varepsilon^{(1)}_{i,j} \eta_{i,j} \big(\widehat{\theta}^{(1)}_{i,j} - \theta^{(1)}_{i,j}\big)}{\big(\widehat{p}_{i,j}\big)^2} \Big| + \frac{1}{N} \Big| \sum_{i \in \mathcal{R}_s} \frac{\big(\varepsilon^{(1)}_{i,j}\big)^2 \eta_{i,j}}{\big(\widehat{p}_{i,j}\big)^2} \Big| + \frac{1}{N} \Big| \sum_{i \in \mathcal{R}_s} \frac{\big(\varepsilon^{(1)}_{i,j}\big)^2 p_{i,j}}{\big(\widehat{p}_{i,j}\big)^2} - \frac{(\sigma^{(1)}_{i,j})^2}{p_{i,j}} \Big|. \tag{4.71}$$

To bound the first term in Eq. (4.71), we have

$$\begin{aligned} \frac{1}{N} \Big| \sum_{i \in \mathcal{R}_s} \frac{\big(\widehat{\theta}^{(1)}_{i,j} - \theta^{(1)}_{i,j}\big)^2 a_{i,j}}{\big(\widehat{p}_{i,j}\big)^2} \Big| &\overset{(a)}{\leq} \frac{1}{N} \Big| \sum_{i \in \mathcal{R}_s} \frac{\big(\widehat{\theta}^{(1)}_{i,j} - \theta^{(1)}_{i,j}\big)^2}{\big(\widehat{p}_{i,j}\big)^2} \Big| \\ &\overset{(b)}{\leq} \frac{1}{\bar{\lambda}^2 N} \big\| \widehat{\Theta}^{(1)}_{\cdot,j} - \Theta^{(1)}_{\cdot,j} \big\|_2^2 \\ &\overset{(c)}{=} \frac{1}{\bar{\lambda}^2} \Big[\mathcal{E}\big(\widehat{\Theta}^{(1)}\big) \Big]^2 \leq \frac{1}{\bar{\lambda}^2} \Big[\mathcal{E}\big(\widehat{\Theta}\big) \Big]^2 \overset{(d)}{=} o_p(1) o_p(1) \overset{(e)}{=} o_p(1), \end{aligned} \tag{4.72}$$

where (a) follows as $a_{i,j} \in \{0, 1\}$, (b) follows from Assumption 4.3, (c) follows from Eq. (4.16), (d) follows from (C1), and (e) follows because $o_p(1) o_p(1) = o_p(1)$.

To control second term in Eq. (4.71), we condition on $\{(\widehat{p}_{i,j}, \widehat{\theta}^{(1)}_{i,j})\}_{i\in\mathcal{R}_s}$. Then, Eq. (4.24) provides that $\{(\widehat{p}_{i,j}, \widehat{\theta}^{(1)}_{i,j})\}_{i\in\mathcal{R}_s} \perp\!\!\!\perp \{\varepsilon^{(1)}_{i,j}\}_{i\in\mathcal{R}_s}$. As a result, $\sum_{i\in\mathcal{R}_s}\varepsilon^{(1)}_{i,j} p_{i,j}\big(\widehat{\theta}^{(1)}_{i,j} - \theta^{(1)}_{i,j}\big)/\big(\widehat{p}_{i,j}\big)^2$ is subGaussian$\big(\overline{\sigma}\big[\sum_{i\in\mathcal{R}_s}\big(p_{i,j}\big)^2\big(\widehat{\theta}^{(1)}_{i,j} - \theta^{(1)}_{i,j}\big)^2/\big(\widehat{p}_{i,j}\big)^4\big]^{1/2}\big)$ because $\varepsilon^{(1)}_{i,j}$ is subGaussian$(\overline{\sigma})$, zero-mean and independent across all $i \in [N]$ due to Assumption 4.2. Then, we have

$$
\begin{aligned}
&\frac{1}{N}\mathbb{E}\Big[\Big|\sum_{i\in\mathcal{R}_s}\frac{2\varepsilon^{(1)}_{i,j}p_{i,j}\big(\widehat{\theta}^{(1)}_{i,j}-\theta^{(1)}_{i,j}\big)}{\big(\widehat{p}_{i,j}\big)^2}\Big|\Big|\{(\widehat{p}_{i,j},\widehat{\theta}^{(1)}_{i,j})\}_{i\in\mathcal{R}_s}\Big]\\
&\overset{(a)}{\leq}\frac{c\overline{\sigma}}{N}\sqrt{\sum_{i\in\mathcal{R}_s}\bigg(\frac{p_{i,j}\big(\widehat{\theta}^{(1)}_{i,j}-\theta^{(1)}_{i,j}\big)}{\big(\widehat{p}_{i,j}\big)^2}\bigg)^2}\\
&\overset{(b)}{\leq}\frac{c\overline{\sigma}}{\overline{\lambda}^2N}\big\|\widehat{\Theta}^{(1)}_{\cdot,j}-\Theta^{(1)}_{\cdot,j}\big\|_2\overset{(c)}{=}\frac{c\overline{\sigma}}{\overline{\lambda}^2}\frac{\mathcal{E}\big(\widehat{\Theta}^{(1)}\big)}{\sqrt{N}}\leq\frac{c\overline{\sigma}}{\overline{\lambda}^2}\frac{\mathcal{E}\big(\widehat{\Theta}\big)}{\sqrt{N}}\overset{(d)}{=}o_p(1),
\end{aligned}
\tag{4.73}
$$

where (a) follows as the first moment of subGaussian(σ) is $O(\sigma)$, (b) follows from Assumptions 4.1 and 4.3, (c) follows from Eq. (4.16), and (d) follows from (C1).

To control third term in Eq. (4.71), we condition on $\{(\widehat{p}_{i,j}, \widehat{\theta}^{(1)}_{i,j})\}_{i\in\mathcal{R}_s}$. Then, Eq. (4.24) provides that $\{(\widehat{p}_{i,j}, \widehat{\theta}^{(1)}_{i,j})\}_{i\in\mathcal{R}_s} \perp\!\!\!\perp \{(\eta_{i,j}, \varepsilon^{(1)}_{i,j})\}_{i\in\mathcal{R}_s}$. As a result, $\sum_{i\in\mathcal{R}_s}\varepsilon^{(1)}_{i,j}\eta_{i,j}\big(\widehat{\theta}^{(1)}_{i,j} - \theta^{(1)}_{i,j}\big)/\big(\widehat{p}_{i,j}\big)^2$ is subExponential$\big(\overline{\sigma}\big[\sum_{i\in\mathcal{R}_s}\big(\widehat{\theta}^{(1)}_{i,j} - \theta^{(1)}_{i,j}\big)^2/\big(\widehat{p}_{i,j}\big)^4\big]^{1/2}/\sqrt{\ell_1}\big)$ because $\varepsilon^{(1)}_{i,j}\eta_{i,j}$ is subExponential$(\overline{\sigma}/\sqrt{\ell_1})$ due to Lemma 4.3 as well as zero-mean and independent across all $i \in [N]$ due to Assumption 4.2. Then, we have

$$
\begin{aligned}
&\frac{1}{N}\mathbb{E}\Big[\Big|\sum_{i\in\mathcal{R}_s}\frac{2\varepsilon^{(1)}_{i,j}\eta_{i,j}\big(\widehat{\theta}^{(1)}_{i,j}-\theta^{(1)}_{i,j}\big)}{\big(\widehat{p}_{i,j}\big)^2}\Big|\Big|\{(\widehat{p}_{i,j},\widehat{\theta}^{(1)}_{i,j})\}_{i\in\mathcal{R}_s}\Big]\\
&\overset{(a)}{\leq}\frac{c\overline{\sigma}}{N}\sqrt{\sum_{i\in\mathcal{R}_s}\bigg(\frac{\widehat{\theta}^{(1)}_{i,j}-\theta^{(1)}_{i,j}}{\big(\widehat{p}_{i,j}\big)^2}\bigg)^2}\\
&\overset{(b)}{\leq}\frac{c\overline{\sigma}}{\overline{\lambda}^2N}\big\|\widehat{\Theta}^{(1)}_{\cdot,j}-\Theta^{(1)}_{\cdot,j}\big\|_2\overset{(c)}{=}\frac{c\overline{\sigma}}{\overline{\lambda}^2}\frac{\mathcal{E}\big(\widehat{\Theta}^{(1)}\big)}{\sqrt{N}}\leq\frac{c\overline{\sigma}}{\overline{\lambda}^2}\frac{\mathcal{E}\big(\widehat{\Theta}\big)}{\sqrt{N}}\overset{(d)}{=}o_p(1),
\end{aligned}
\tag{4.74}
$$

where (a) follows as the first moment of subExponential(σ) is $O(\sigma)$ (Zhang and Wei, 2022, Corollary 3), (b) follows from Assumption 4.3, (c) follows from Eq. (4.16), and (d) follows from (C1).

To control fourth term in Eq. (4.71), we condition on $\{\widehat{p}_{i,j}\}_{i\in\mathcal{R}_s}$. Then, Eq. (4.24) provides that $\{\widehat{p}_{i,j}\}_{i\in\mathcal{R}_s} \perp\!\!\!\perp \{(\eta_{i,j}, \varepsilon^{(1)}_{i,j})\}_{i\in\mathcal{R}_s}$. As a result, $\sum_{i\in\mathcal{R}_s}\big(\varepsilon^{(1)}_{i,j}\big)^2\eta_{i,j}/\big(\widehat{p}_{i,j}\big)^2$ is subWeibull$_{2/3}$ $\big(\overline{\sigma}^2\big[\sum_{i\in\mathcal{R}_s}1/\big(\widehat{p}_{i,j}\big)^4\big]^{1/2}/\sqrt{\ell_1}\big)$ because $(\varepsilon^{(1)}_{i,j})^2\eta_{i,j}$ is subWeibull$_{2/3}(\overline{\sigma}^2/\sqrt{\ell_1})$ due to Lemma 4.4 as well as zero-mean and independent across all $i \in [N]$ due to Assumption 4.2. Then, we have

$$
\frac{1}{N}\mathbb{E}\Big[\Big|\sum_{i\in\mathcal{R}_s}\frac{\big(\varepsilon^{(1)}_{i,j}\big)^2\eta_{i,j}}{\big(\widehat{p}_{i,j}\big)^2}\Big|\Big|\{\widehat{p}_{i,j}\}_{i\in\mathcal{R}_s}\Big]\overset{(a)}{\leq}\frac{c\overline{\sigma}^2}{N}\sqrt{\sum_{i\in\mathcal{R}_s}\frac{1}{\big(\widehat{p}_{i,j}\big)^4}}\overset{(b)}{\leq}\frac{c\overline{\sigma}^2}{\overline{\lambda}^2\sqrt{N}}=o_p(1),
\tag{4.75}
$$

where (a) follows as the first moment of subWeibull$_{2/3}(\sigma)$ is $O(\sigma)$ (Zhang and Wei, 2022, Corollary 3) and (b) follows from Assumption 4.3.

To control fifth term in Eq. (4.71), we have

$$\begin{aligned}\Big|\sum_{i\in\mathcal{R}_s}\Big(\frac{\big(\varepsilon^{(1)}_{i,j}\big)^2 p_{i,j}}{\big(\widehat{p}_{i,j}\big)^2}-\frac{(\sigma^{(1)}_{i,j})^2}{p_{i,j}}\Big)\Big| &= \Big|\sum_{i\in\mathcal{R}_s}\Big(\frac{\big(\varepsilon^{(1)}_{i,j}\big)^2 p_{i,j}}{\big(\widehat{p}_{i,j}\big)^2}-\frac{\big(\sigma^{(1)}_{i,j}\big)^2 p_{i,j}}{\big(\widehat{p}_{i,j}\big)^2}+\frac{\big(\sigma^{(1)}_{i,j}\big)^2 p_{i,j}}{\big(\widehat{p}_{i,j}\big)^2}-\frac{(\sigma^{(1)}_{i,j})^2}{p_{i,j}}\Big)\Big| \\ &\overset{(a)}{\le} \Big|\sum_{i\in\mathcal{R}_s}\Big(\frac{\big[\big(\varepsilon^{(1)}_{i,j}\big)^2-\big(\sigma^{(1)}_{i,j}\big)^2\big]p_{i,j}}{\big(\widehat{p}_{i,j}\big)^2}\Big)\Big|+\Big|\sum_{i\in\mathcal{R}_s}\Big(\frac{\big(\sigma^{(1)}_{i,j}\big)^2 p_{i,j}}{\big(\widehat{p}_{i,j}\big)^2}-\frac{(\sigma^{(1)}_{i,j})^2}{p_{i,j}}\Big)\Big|,\end{aligned} \tag{4.76}$$

where (a) follows from the triangle inequality. To control the first term in Eq. (4.76), we condition on $\{\widehat{p}_{i,j}\}_{i\in\mathcal{R}_s}$. Then, Eq. (4.24) provides that $\{\widehat{p}_{i,j}\}_{i\in\mathcal{R}_s} \perp\!\!\!\perp \{\varepsilon^{(1)}_{i,j}\}_{i\in\mathcal{R}_s}$. Further, $\mathbb{E}[(\varepsilon^{(1)}_{i,j})^2-(\sigma^{(1)}_{i,j})^2]=0$ due to (C3) and Assumption 4.2. As a result, $\sum_{i\in\mathcal{R}_s}\big[\big(\varepsilon^{(1)}_{i,j}\big)^2-\big(\sigma^{(1)}_{i,j}\big)^2\big]p_{i,j}/\big(\widehat{p}_{i,j}\big)^2$ is subExponential$\big(\bar\sigma^2\big[\sum_{i\in\mathcal{R}_s}\big(p_{i,j}\big)^2/\big(\widehat{p}_{i,j}\big)^4\big]^{1/2}\big)$ because $(\varepsilon^{(1)}_{i,j})^2-(\sigma^{(1)}_{i,j})^2$ is subExponential$(\bar\sigma^2)$ and independent across all $i\in[N]$ due to Lemma 4.3. Then, we have

$$\frac{1}{N}\mathbb{E}\Big[\Big|\sum_{i\in\mathcal{R}_s}\frac{\big[\big(\varepsilon^{(1)}_{i,j}\big)^2-\big(\sigma^{(1)}_{i,j}\big)^2\big]p_{i,j}}{\big(\widehat{p}_{i,j}\big)^2}\Big|\Big|\{\widehat{p}_{i,j}\}_{i\in\mathcal{R}_s}\Big] \overset{(a)}{\le} \frac{c\bar\sigma^2}{N}\sqrt{\sum_{i\in\mathcal{R}_s}\Big(\frac{p_{i,j}}{\big(\widehat{p}_{i,j}\big)^2}\Big)^2} \overset{(b)}{\le} \frac{c\bar\sigma^2}{\bar\lambda^2\sqrt{N}} = o_p(1), \tag{4.77}$$

where (a) follows as the first moment of subExponential(σ) is $O(\sigma)$ and (b) follows from Assumption 4.3. To bound the second term in Eq. (4.76), applying the Cauchy-Schwarz inequality yields that

$$\begin{aligned}\frac{1}{N}\Big|\sum_{i\in\mathcal{R}_s}\Big(\frac{\big(\sigma^{(1)}_{i,j}\big)^2 p_{i,j}}{\big(\widehat{p}_{i,j}\big)^2}-\frac{(\sigma^{(1)}_{i,j})^2}{p_{i,j}}\Big)\Big| &= \frac{1}{N}\Big|\sum_{i\in\mathcal{R}_s}\frac{\big(\sigma^{(1)}_{i,j}\big)^2\big(\big(p_{i,j}\big)^2-\big(\widehat{p}_{i,j}\big)^2\big)}{\big(\widehat{p}_{i,j}\big)^2 p_{i,j}}\Big| \\ &\overset{(a)}{\le} \frac{2}{N}\sum_{i\in\mathcal{R}_s}\frac{\big(\sigma^{(1)}_{i,j}\big)^2|p_{i,j}-\widehat{p}_{i,j}|}{\big(\widehat{p}_{i,j}\big)^2 p_{i,j}} \\ &\overset{(b)}{\le} \frac{2\bar\sigma^2}{\lambda\bar\lambda^2 N}\sum_{i\in\mathcal{R}_s}\big|p_{i,j}-\widehat{p}_{i,j}\big| \\ &\overset{(c)}{\le} \frac{2\bar\sigma^2}{\lambda\bar\lambda^2\sqrt{N}}\big\|P_{\cdot,j}-\widehat{P}_{\cdot,j}\big\|_2 \overset{(d)}{=} \frac{2\bar\sigma^2}{\lambda\bar\lambda^2}\mathcal{E}\big(\widehat{P}\big) \overset{(e)}{=} o_p(1),\end{aligned} \tag{4.78}$$

where (a) follows by using $\big(p_{i,j}\big)^2-\big(\widehat{p}_{i,j}\big)^2=(p_{i,j}+\widehat{p}_{i,j})(p_{i,j}-\widehat{p}_{i,j})\le 2|p_{i,j}-\widehat{p}_{i,j}|$, (b) follows from Assumptions 4.1 and 4.3, and because the variance of a subGaussian random variable is upper bounded by the square of its subGaussian norm, (c) follows by the relationship between ℓ_1 and ℓ_2 norms of a vector, (d) follows from Eq. (4.16), and (e) follows from (C1).

Putting together Eqs. (4.71) to (4.78) using Lemma 4.6,

$$\frac{1}{N}\Big|\sum_{i\in\mathcal{R}_s}\mathbb{T}^{(1)}_{i,j}\Big| = o_p(1).$$

Then, the claim in Eq. (4.70) follows for $a = 1$ by using $|\sum_{i \in [N]} \mathbb{T}_{i,j}^{(1)}| \leq |\sum_{i \in \mathcal{R}_0} \mathbb{T}_{i,j}^{(1)}| + |\sum_{i \in \mathcal{R}_1} \mathbb{T}_{i,j}^{(1)}|$. The proof of Eq. (4.70) for $a = 0$ follows similarly.

4.G Proofs of Propositions 4.3 and 4.4

In Section 4.G.1, we prove Proposition 4.3, i.e., we show that the estimates of P, $\Theta^{(0)}$, and $\Theta^{(1)}$ generated by `Cross-Fitted-MC` satisfy Assumption 4.4. Next, we prove Proposition 4.4 implying that the estimates of P, $\Theta^{(0)}$, and $\Theta^{(1)}$ generated by `Cross-Fitted-SVD` satisfy the condition (C2) in Theorem 4.2 as long as $\sqrt{N}/M = o(1)$.

4.G.1 Proof of Proposition 4.3: Guarantees for `Cross-Fitted-MC`

Consider any matrix completion algorithm `MC`. We show that

$$\widehat{P}_{\mathcal{I}}, \widehat{\Theta}_{\mathcal{I}}^{(a)} \perp\!\!\!\perp W_{\mathcal{I}} \tag{4.79}$$

and

$$\widehat{P}_{\mathcal{I}} \perp\!\!\!\perp W_{\mathcal{I}}, E_{\mathcal{I}}^{(a)}, \tag{4.80}$$

for every $\mathcal{I} \in \mathcal{P}$ and $a \in \{0, 1\}$, where $\mathcal{P}$ is the block partition of $[N] \times [M]$ into four blocks from Assumption 4.5. Then, Eqs. (4.14) and (4.15) in Assumption 4.4 follow from Eqs. (4.79) and (4.80), respectively.

Consider $\widehat{\Theta}^{(0)}, \widehat{\Theta}^{(1)}$, and $\widehat{P}$ as in Eqs. (4.30) to (4.32). Fix any $a \in \{0, 1\}$. From Eq. (4.29), note that $\widehat{P}_{\mathcal{I}}$ depends only on $A \otimes \mathbf{1}^{-\mathcal{I}}$ and $\widehat{\Theta}_{\mathcal{I}}^{(a)}$ depends on $Y^{(a),\text{obs}} \otimes \mathbf{1}^{-\mathcal{I}}$. In other words, the randomness in $\big(\widehat{P}_{\mathcal{I}}, \widehat{\Theta}_{\mathcal{I}}^{(a)}\big)$ stems from the randomness in $\big(A_{-\mathcal{I}}, Y_{-\mathcal{I}}^{(a),\text{obs}}\big)$ which in turn stems from the randomness in $\big(W_{-\mathcal{I}}, E_{-\mathcal{I}}^{(a)}\big)$. Then, Eq. (4.79) follows from Eq. (4.27). Likewise, the randomness in $\widehat{P}_{\mathcal{I}}$ stems from the randomness in $A_{-\mathcal{I}}$ which in turn stems from the randomness in $W_{-\mathcal{I}}$. Then, Eq. (4.80) follows from Eq. (4.28).

To prove Eq. (4.24), we show that

$$\widehat{P}_{\mathcal{I}}, \widehat{\Theta}_{\mathcal{I}}^{(a)} \perp\!\!\!\perp W_{\mathcal{I}}, E_{\mathcal{I}}^{(a)}, \tag{4.81}$$

for every $\mathcal{I} \in \mathcal{P}$ and $a \in \{0, 1\}$. As mentioned above, the randomness in $\big(\widehat{P}_{\mathcal{I}}, \widehat{\Theta}_{\mathcal{I}}^{(a)}\big)$ stems from the randomness in $\big(A_{-\mathcal{I}}, Y_{-\mathcal{I}}^{(a),\text{obs}}\big)$ which in turn stems from the randomness in $\big(W_{-\mathcal{I}}, E_{-\mathcal{I}}^{(a)}\big)$. Then, Eq. (4.81) follows from Eq. (4.33).

4.G.2 Proof of Proposition 4.4: Guarantees for `Cross-Fitted-SVD`

To prove this result, we first derive a corollary of Lemma A.1 in Bai and Ng (2021) for a generic matrix of interest T, such that $S = (T + H) \otimes F$, and apply it to P, $\Theta^{(0)} \odot (\mathbf{1} - P)$, and $\Theta^{(1)} \odot P$. We impose the following restrictions on T, H, and F.

Assumption 4.10 (Strong linear latent factors)**.** *There exist a constant $r_T \in [\min\{N, M\}]$ and a collection of latent factors*

$$\widetilde{U} \in \mathbb{R}^{N \times r_T} \quad \textit{and} \quad \widetilde{V} \in \mathbb{R}^{M \times r_T},$$

such that,

(a) T satisfies the factorization: $T = \widetilde{U}\widetilde{V}^\top$,

(b) $\|\widetilde{U}\|_{2,\infty} \leq c$ and $\|\widetilde{V}\|_{2,\infty} \leq c$ for some positive constant c, and

(c) The matrices defined below exist and are positive definite:

$$\lim_{N\to\infty} \frac{\widetilde{U}^\top \widetilde{U}}{N} \quad \textit{and} \quad \lim_{M\to\infty} \frac{\widetilde{V}^\top \widetilde{V}}{M}.$$

Assumption 4.11 (Zero-mean, weakly dependent, and subExponential noise)**.** *The noise matrix H is such that,*

(a) $\{h_{i,j} : i \in [N], j \in [M]\}$ are zero-mean subExponential with the subExponential norm bounded by a constant $\overline{\sigma}$,

(b) $\sum_{j' \in [M]} \left|\mathbb{E}[h_{i,j} h_{i,j'}]\right| \leq c$ for every $i \in [N]$ and $j \in [M]$, and

(c) The elements of $\{H_{i,\cdot} : i \in [N]\}$ are mutually independent (across i).

Assumption 4.12 (Strong block factors)**.** *Consider the latent factors $\widetilde{U} \in \mathbb{R}^{N \times r_T}$ and $\widetilde{V} \in \mathbb{R}^{M \times r_T}$ from Assumption 4.10. Let $\mathcal{R}_{\text{obs}} \subseteq [N]$ and $\mathcal{C}_{\text{obs}} \subseteq [M]$ denote the set of rows and columns of S, respectively, with all entries observed, and $\mathcal{R}_{\text{miss}} \triangleq [N] \setminus \mathcal{R}_{\text{obs}}$ and $\mathcal{C}_{\text{miss}} \triangleq [M] \setminus \mathcal{C}_{\text{obs}}$. Let $\widetilde{U}^{\text{obs}} \in \mathbb{R}^{|\mathcal{R}_{\text{obs}}| \times r_T}$ and $\widetilde{U}^{\text{miss}} \in \mathbb{R}^{|\mathcal{R}_{\text{miss}}| \times r_T}$ be the sub-matrices of $\widetilde{U}$ that keeps the rows corresponding to the indices in $\mathcal{R}_{\text{obs}}$ and $\mathcal{R}_{\text{miss}}$, respectively. Let $\widetilde{V}^{\text{obs}} \in \mathbb{R}^{|\mathcal{C}_{\text{obs}}| \times r_T}$ and $\widetilde{V}^{\text{miss}} \in \mathbb{R}^{|\mathcal{C}_{\text{miss}}| \times r_T}$ be the sub-matrices of $\widetilde{V}$ that keeps the rows corresponding to the indices in $\mathcal{C}_{\text{obs}}$ and $\mathcal{C}_{\text{miss}}$, respectively. Then, the matrices defined below exist and are positive definite:*

$$\lim_{N\to\infty} \frac{\widetilde{U}^{\text{obs}\top} \widetilde{U}^{\text{obs}}}{|\mathcal{R}_{\text{obs}}|}, \ \lim_{M\to\infty} \frac{\widetilde{U}^{\text{miss}\top} \widetilde{U}^{\text{miss}}}{|\mathcal{R}_{\text{miss}}|}, \ \lim_{N\to\infty} \frac{\widetilde{V}^{\text{obs}\top} \widetilde{V}^{\text{obs}}}{|\mathcal{C}_{\text{obs}}|}, \ \textit{and} \ \lim_{M\to\infty} \frac{\widetilde{V}^{\text{miss}\top} \widetilde{V}^{\text{miss}}}{|\mathcal{C}_{\text{miss}}|}. \tag{4.82}$$

Further, the mask matrix F is such that

$$|\mathcal{R}_{\text{obs}}| = \Omega(N), \quad |\mathcal{R}_{\text{miss}}| = \Omega(N), \quad |\mathcal{C}_{\text{obs}}| = \Omega(M), \quad \textit{and} \quad |\mathcal{C}_{\text{miss}}| = \Omega(M). \tag{4.83}$$

The next result characterizes the entry-wise error in recovering the missing entries of a matrix where all entries in one block are deterministically missing (see the discussion in Section 4.5.1) using the `TW` algorithm (summarized in Section 4.5.2.1). Its proof, essentially established as a corollary of Bai and Ng (2021, Lemma A.1), is provided in Section 4.G.3.

Corollary 4.5. *Consider a matrix of interest T, a noise matrix H, and a mask matrix F such that that Assumptions 4.10 to 4.12 hold. Let $S \in \{\mathbb{R} \cup \{?\}\}^{N \times M}$ be the observed* *matrix as in Eq. (4.6). Let $\mathcal{R}_{\text{obs}} \subseteq [N]$ and $\mathcal{C}_{\text{obs}} \subseteq [M]$ denote the set of rows and columns of S, respectively, with all entries observed. Let $\mathcal{I} = \mathcal{R}_{\text{miss}} \times \mathcal{C}_{\text{miss}}$ where $\mathcal{R}_{\text{miss}} \triangleq [N] \setminus \mathcal{R}_{\text{obs}}$ and $\mathcal{C}_{\text{miss}} \triangleq [M] \setminus \mathcal{C}_{\text{obs}}$. Then, $\texttt{TW}_{r_T}$ produces an estimate $\widehat{T}_{\mathcal{I}}$ of $T_{\mathcal{I}}$ such that*

$$\|\widehat{T}_{\mathcal{I}} - T_{\mathcal{I}}\|_{\max} = O_p\Big(\frac{1}{\sqrt{N}} + \frac{1}{\sqrt{M}}\Big),$$

as $N, M \to \infty$.

Given this corollary, we now complete the proof of Proposition 4.4. Consider the partition $\mathcal{P}$ from Assumption 4.5 and fix any $\mathcal{I} \in \mathcal{P}$. Recall that `Cross-Fitted-SVD` applies `TW` on $P \otimes \mathbf{1}^{-\mathcal{I}}$, $Y^{(0),\text{full}} \otimes \mathbf{1}^{-\mathcal{I}}$, and $Y^{(1),\text{full}} \otimes \mathbf{1}^{-\mathcal{I}}$, and note that the mask matrix $\mathbf{1}^{-\mathcal{I}}$ satisfies the requirement in Assumption 4.12, i.e., Eq. (4.83) under Assumption 4.8.

4.G.2.1 Estimating P.

Consider estimating P using `Cross-Fitted-SVD`. To apply Corollary 4.5, we use Assumptions 4.6 and 4.7 to note that P satisfies Assumption 4.10 with rank parameter r_p. Then, we use Eq. (4.4), Assumption 4.2, and Assumption 4.9 to note that W satisfies Assumption 4.11. Finally, we use Assumption 4.8 to note that Assumption 4.12 holds. Step 2 of `Cross-Fitted-SVD` can be rewritten as $\widehat{P} = \texttt{Proj}_{\bar{\lambda}}(\overline{P})$ and $\overline{P} = \texttt{Cross-Fitted-MC}(\texttt{TW}_{r_1}, A, \mathcal{P})$ where $r_1 = r_p$. Then,

$$\|\widehat{P}_{\mathcal{I}} - P_{\mathcal{I}}\|_{\max} \overset{(a)}{\leq} \|\overline{P}_{\mathcal{I}} - P_{\mathcal{I}}\|_{\max} \overset{(b)}{=} O_p\Big(\frac{1}{\sqrt{N}} + \frac{1}{\sqrt{M}}\Big),$$

where (a) follows from Assumption 4.1, the choice of $\bar{\lambda}$, and the definition of $\texttt{Proj}_{\bar{\lambda}}(\cdot)$, and (b) follows from Corollary 4.5. Applying a union bound over all $\mathcal{I} \in \mathcal{P}$, we have

$$\mathcal{E}\big(\widehat{P}\big) \overset{(a)}{\leq} \|\widehat{P} - P\|_{\max} = O_p\Big(\frac{1}{\sqrt{N}} + \frac{1}{\sqrt{M}}\Big), \tag{4.84}$$

where (a) follows from the definition of $(1, 2)$ operator norm.

4.G.2.2 Estimating $\Theta^{(0)}$ and $\Theta^{(1)}$.

For every $a \in \{0, 1\}$, we show that

$$\mathcal{E}\big(\widehat{\Theta}^{(a)}\big) = O_p\Big(\frac{1}{\sqrt{N}} + \frac{1}{\sqrt{M}}\Big). \tag{4.85}$$

We focus on $a = 1$ noting that the proof for $a = 0$ is analogous. We split the proof in two cases: (i) $\|\big(\widehat{\Theta}^{(1)} - \Theta^{(1)}\big) \odot \widehat{P}\|_{\max} \leq \|\Theta^{(1)} \odot \big(\widehat{P} - P\big)\|_{\max}$ and (ii) $\|\big(\widehat{\Theta}^{(1)} - \Theta^{(1)}\big) \odot \widehat{P}\|_{\max} \geq \|\Theta^{(1)} \odot \big(\widehat{P} - P\big)\|_{\max}$.

In the first case, we have

$$\bar{\lambda}\|\widehat{\Theta}^{(1)} - \Theta^{(1)}\|_{\max} \overset{(a)}{\leq} \|\big(\widehat{\Theta}^{(1)} - \Theta^{(1)}\big) \odot \widehat{P}\|_{\max} \leq \|\Theta^{(1)} \odot \big(\widehat{P} - P\big)\|_{\max} \overset{(b)}{\leq} \|\Theta^{(1)}\|_{\max}\|\widehat{P} - P\|_{\max}, \tag{4.86}$$

where (a) follows from Assumption 4.3 and (b) follows from the definition of $\|\Theta^{(1)}\|_{\max}$. Then,

$$\mathcal{E}\big(\widehat{\Theta}^{(1)}\big) \overset{(a)}{\leq} \|\widehat{\Theta}^{(1)} - \Theta^{(1)}\|_{\max} \overset{(b)}{\leq} \frac{\|\Theta^{(1)}\|_{\max}}{\bar{\lambda}}\|\widehat{P} - P\|_{\max} \overset{(c)}{=} \frac{\|\Theta^{(1)}\|_{\max}}{\bar{\lambda}} O_p\bigg(\frac{1}{\sqrt{N}} + \frac{1}{\sqrt{M}}\bigg),$$

where (a) follows from the definition of $(1,2)$ operator norm, (b) follows from Eq. (4.86), and (c) follows from Eq. (4.84). Then, Eq. (4.85) follows as $1/\bar{\lambda}$ and $\|\Theta^{(1)}\|_{\max}$ are assumed to be bounded.

In the second case, using Eqs. (4.2) and (4.3) to expand $Y^{(1),\text{full}}$, we have

$$Y^{(1),\text{full}} = \Theta^{(1)} \odot P + \Theta^{(1)} \odot W + E^{(1)} \odot P + E^{(1)} \odot W.$$

Next, we utilize two claims proven in Sections 4.G.2.3 and 4.G.2.4 respectively: $\Theta^{(1)} \odot P$ satisfies Assumption 4.10 with rank parameter $r_{\theta_1} r_p$ and

$$\bar{\varepsilon}^{(1)} \triangleq \Theta^{(1)} \odot W + E^{(1)} \odot P + E^{(1)} \odot W,$$

satisfies Assumption 4.11. Finally, Assumption 4.8 in Section 4.5 implies that Assumption 4.12 holds.

Now, note that step 5 of `Cross-Fitted-SVD` can be rewritten as $\widehat{\Theta}^{(1)} = \overline{\Theta}^{(1)} \oslash \widehat{P}$ and $\overline{\Theta}^{(1)} = \texttt{Cross-Fitted-MC}(\texttt{TW}_{r_3}, Y^{(1),\text{full}}, \mathcal{P})$ where $r_3 = r_{\theta_1} r_p$. Then, from Corollary 4.5,

$$\|\overline{\Theta}^{(1)}_{\mathcal{I}} - \Theta^{(1)}_{\mathcal{I}} \odot P_{\mathcal{I}}\|_{\max} = O_p\bigg(\frac{1}{\sqrt{N}} + \frac{1}{\sqrt{M}}\bigg).$$

Applying a union bound over all $\mathcal{I} \in \mathcal{P}$ and noting that $\overline{\Theta}^{(1)} = \widehat{\Theta}^{(1)} \odot \widehat{P}$, we have

$$\|\widehat{\Theta}^{(1)} \odot \widehat{P} - \Theta^{(1)} \odot P\|_{\max} = O_p\bigg(\frac{1}{\sqrt{N}} + \frac{1}{\sqrt{M}}\bigg). \tag{4.87}$$

The left hand side of Eq. (4.87) can be written as,

$$\begin{aligned}\|\widehat{\Theta}^{(1)} \odot \widehat{P} - \Theta^{(1)} \odot P\|_{\max} &= \|\widehat{\Theta}^{(1)} \odot \widehat{P} - \Theta^{(1)} \odot \widehat{P} + \Theta^{(1)} \odot \widehat{P} - \Theta^{(1)} \odot P\|_{\max} \\ &\overset{(a)}{\geq} \|\big(\widehat{\Theta}^{(1)} - \Theta^{(1)}\big) \odot \widehat{P}\|_{\max} - \|\Theta^{(1)} \odot \big(\widehat{P} - P\big)\|_{\max} \\ &\overset{(b)}{\geq} \bar{\lambda}\|\widehat{\Theta}^{(1)} - \Theta^{(1)}\|_{\max} - \|\Theta^{(1)}\|_{\max}\|\widehat{P} - P\|_{\max}, \end{aligned} \tag{4.88}$$

where (a) follows from triangle inequality as $\|\big(\widehat{\Theta}^{(1)} - \Theta^{(1)}\big) \odot \widehat{P}\|_{\max} \geq \|\Theta^{(1)} \odot \big(\widehat{P} - P\big)\|_{\max}$ and (b) follows from the choice of $\bar{\lambda}$ and the definition of $\|\Theta^{(1)}\|_{\max}$. Then,

$$\begin{aligned}\mathcal{E}\big(\widehat{\Theta}^{(1)}\big) \overset{(a)}{\leq} \|\widehat{\Theta}^{(1)} - \Theta^{(1)}\|_{\max} &\overset{(b)}{\leq} \frac{1}{\bar{\lambda}}\|\widehat{\Theta}^{(1)} \odot \widehat{P} - \Theta^{(1)} \odot P\|_{\max} + \frac{\|\Theta^{(1)}\|_{\max}}{\bar{\lambda}}\|\widehat{P} - P\|_{\max} \\ &\overset{(c)}{=} \frac{1}{\bar{\lambda}} O_p\bigg(\frac{1}{\sqrt{N}} + \frac{1}{\sqrt{M}}\bigg) + \frac{\|\Theta^{(1)}\|_{\max}}{\bar{\lambda}} O_p\bigg(\frac{1}{\sqrt{N}} + \frac{1}{\sqrt{M}}\bigg),\end{aligned}$$

where (a) follows from the definition of $L_{1,2}$ norm, (b) follows from Eq. (4.88), and (c) follows from Eqs. (4.84) and (4.87). Then, Eq. (4.85) follows as $1/\bar{\lambda}$ and $\|\Theta^{(1)}\|_{\max}$ are assumed to be bounded.

4.G.2.3 Proof that $\boldsymbol{\Theta}^{(0)} \odot (\mathbf{1} - \boldsymbol{P})$ and $\boldsymbol{\Theta}^{(1)} \odot \boldsymbol{P}$ satisfy Assumption 4.10.

First, we show that $\overline{U}^{(0)} \in \mathbb{R}^{N \times r_{\theta_0}(r_p+1)}$ and $\overline{V}^{(0)} \in \mathbb{R}^{N \times r_{\theta_0}(r_p+1)}$ are factors of $\Theta^{(0)} \odot (\mathbf{1} - P)$, and $\overline{U}^{(1)} \in \mathbb{R}^{N \times r_{\theta_1} r_p}$ and $\overline{V}^{(1)} \in \mathbb{R}^{N \times r_{\theta_1}}$ are factors of $\Theta^{(1)} \odot P$ as claimed in Eq. (4.35). We have

$$\Theta^{(1)} \odot P = \Big(\sum_{i \in [r_{\theta_1}]} U^{(1)}_{i,\cdot} V^{(1)^\top}_{i,\cdot} \Big) \odot \Big(\sum_{j \in [r_p]} U_{j,\cdot} V_{j,\cdot}^\top \Big) = \sum_{i \in [r_{\theta_1}]} \sum_{j \in [r_p]} \Big(U^{(1)}_{i,\cdot} \odot U_{j,\cdot} \Big) \Big(V^{(1)}_{i,\cdot} \odot V_{j,\cdot} \Big)^\top$$
$$\overset{(a)}{=} \big(U * U^{(1)} \big) \big(V * V^{(1)} \big)^\top \overset{(b)}{=} \overline{U}^{(1)} \overline{V}^{(1)^\top},$$

where (a) follows from the definition of Khatri-Rao product (see Section 4.1) and (b) follows from the definitions of $\overline{U}^{(1)}$ and $\overline{V}^{(1)}$. The proof for $\Theta^{(0)} \odot (\mathbf{1} - P)$ follows similarly. Then, Assumption 4.10(a) holds from Eq. (4.35). Next, we note that

$$\|\overline{U}^{(1)}\|_{2,\infty} = \|U * U^{(1)}\|_{2,\infty} \overset{(a)}{=} \max_{i \in [N]} \sqrt{\sum_{j \in [r_p]} u_{i,j}^2 \sum_{j' \in [r_{\theta_1}]} (u^{(1)}_{i,j'})^2} \leq \|U\|_{2,\infty} \|U^{(1)}\|_{2,\infty} \overset{(b)}{\leq} c,$$

where (a) follows from the definition of Khatri-Rao product (see Section 4.1), and (b) follows from Assumption 4.7. Then, $\Theta^{(1)} \odot P$ satisfies Assumption 4.10(b) by using similar arguments on $\overline{V}^{(1)}$. Further, $\Theta^{(0)} \odot (\mathbf{1} - P)$ satisfies Assumption 4.10(b) by noting that $\|\overline{U}\|_{2,\infty}$ and $\|\overline{V}\|_{2,\infty}$ are bounded whenever $\|U\|_{2,\infty}$ and $\|V\|_{2,\infty}$ are bounded, respectively. Finally, Assumption 4.10(c) holds from Assumption 4.7.

4.G.2.4 Proof that $\overline{\varepsilon}^{(1)}$ satisfies Assumption 4.11

Recall that $\overline{\varepsilon}^{(1)} \triangleq \Theta^{(1)} \odot W + E^{(1)} \odot P + E^{(1)} \odot W$. Then, Assumption 4.11(a) holds as $\overline{\varepsilon}^{(1)}_{i,j}$ is zero-mean from Assumption 4.2 and Eq. (4.3), and $\overline{\varepsilon}^{(1)}_{i,j}$ is subExponential because $\varepsilon^{(1)}_{i,j} \eta_{i,j}$ is a subExponential random variable Lemma 4.3, every subGaussian random variable is subExponential random variable, and sum of subExponential random variables is a subExponential random variable. Finally, Assumption 4.11(b) and Assumption 4.11(b) hold from Assumption 4.9(b) and Assumption 4.9(c), respectively.

4.G.3 Proof of Corollary 4.5

Corollary 4.5 is a direct application of Bai and Ng (2021, Lemma A.1), specialized to our setting. Notably, Bai and Ng (2021) make three assumptions numbered A, B, and C in their paper to establish the corresponding result. It remains to establish that the conditions assumed in Corollary 4.5 imply the necessary conditions used in the proof of Bai and Ng (2021, Lemma A.1). First, note that certain assumptions in Bai and Ng (2021) are not actually used in their proof of Lemma A.1 (or in the proof of other results used in that proof), namely, the distinct eigenvalue condition in Assumption A(a)(iii), the asymptotic normality conditions in Assumption A(c) and the asymptotic normality conditions in Assumption C. Next, Eq. (4.83) in Assumption 4.12

implies Assumption B and Eq. (4.82) in Assumption 4.12 is equivalent to the remaining conditions in Assumption C.

It remains to show how Assumptions 4.10 and 4.11 imply the remainder of conditions in Bai and Ng (2021, Assumptions A). For completeness, these conditions are collected in the following assumption.

Assumption 4.13. *The noise matrix H is such that,*

(a) $\max_{j\in[M]} \frac{1}{N} \sum_{j'\in[M]} \big| \sum_{i\in[N]} \mathbb{E}[h_{i,j}h_{i,j'}]\big| \leq c,$

(b) $\max_{j\in[M]} \big|\mathbb{E}[h_{i,j}h_{i',j}]\big| \leq c_{i,i'}$ *and* $\max_{i\in[N]} \sum_{i'\in[N]} c_{i,i'} \leq c,$

(c) $\frac{1}{NM} \sum_{i,i'\in[N]} \sum_{j,j'\in[M]} \big|\mathbb{E}[h_{i,j}h_{i',j'}]\big| \leq c,$ *and*

(d) $\max_{j,j'\in[M]} \frac{1}{N^2}\mathbb{E}\big[\big| \sum_{i\in[N]} \big(h_{i,j}h_{i,j'} - \mathbb{E}[h_{i,j}h_{i,j'}]\big)\big|^4\big].$

Assumption 4.13 is a restatement of the subset of conditions from Bai and Ng (2021, Assumption A) necessary in Bai and Ng (2021, proof of Lemma A.1) and it essentially requires weak dependence in the noise across measurements and across units. In particular, Assumption 4.13(a), (b), (c), and (d) correspond to Assumption A(b)(ii), (iii), (iv), (v), respectively, of Bai and Ng (2021). For the other conditions in Bai and Ng (2021, Assumption A), note that Assumption 4.10 above is equivalent to their Assumption A(a)(i) and (ii) of Bai and Ng (2021) when the factors are non-random as in this work. Similarly, Assumption 4.11(a) above is analogous to Assumption A(b)(i) of Bai and Ng (2021). Assumption A(b)(vi) of Bai and Ng (2021) is implied by their other Assumptions for non-random factors as stated in Bai (2003).

To establish Corollary 4.5, it remains to establish that Assumption 4.13 holds, which is done in Section 4.G.3.1 below.

4.G.3.1 Assumption 4.13 holds

First, Assumption 4.13(a) holds as follows,

$$\max_{j\in[M]} \frac{1}{N} \sum_{j'\in[M]} \Big| \sum_{i\in[N]} \mathbb{E}\big[h_{i,j}h_{i,j'}\big]\Big| \overset{(a)}{\leq} \max_{j\in[M]} \frac{1}{N} \sum_{i\in[N]} \sum_{j'\in[M]} \Big|\mathbb{E}\big[h_{i,j}h_{i,j'}\big]\Big| \overset{(b)}{\leq} \max_{j\in[M]} \frac{1}{N} \sum_{i\in[N]} c = c,$$

where (a) follows from triangle inequality and (b) follows from Assumption 4.11(b). Next, from Assumption 4.11(a) and Assumption 4.11(c), we have

$$\max_{j\in[M]} \big|\mathbb{E}[h_{i,j}h_{i',j}]\big| = \begin{cases} 0 & \text{if } i \neq i' \\ \max_{j\in[M]} \big|\mathbb{E}[h_{i,j}^2]\big| \leq c & \text{if } i = i' \end{cases}$$

Then, Assumption 4.13(b) holds as $\max_{i\in[N]} \max_{j\in[M]} \sum_{i'\in[N]} \big|\mathbb{E}[h_{i,j}h_{i',j}]\big| \leq c$. Next, Assumption 4.13(c) holds as follows,

$$\frac{1}{NM} \sum_{i,i'\in[N]} \sum_{j,j'\in[M]} \big|\mathbb{E}[h_{i,j}h_{i',j'}]\big| \overset{(a)}{=} \frac{1}{NM} \sum_{i\in[N]} \sum_{j,j'\in[M]} \big|\mathbb{E}[h_{i,j}h_{i,j'}]\big| \overset{(b)}{\leq} \frac{1}{NM} \sum_{i\in[N]} \sum_{j\in[M]} c = c,$$

where (a) follows from Assumption 4.11(c) and (b) follows from Assumption 4.11(b). Next, let $\gamma_{i,j,j'} \triangleq h_{i,j}h_{i,j'} - \mathbb{E}[h_{i,j}h_{i,j'}]$ and fix any $j, j' \in [M]$. Then, Assumption 4.13(d) holds as follows,

$$\frac{1}{N^2}\mathbb{E}\Big[\Big(\sum_{i\in[N]}\gamma_{i,j,j'}\Big)^4\Big] = \frac{1}{N^2}\mathbb{E}\Big[\Big(\sum_{i_1\in[N]}\gamma_{i_1,j,j'}\Big)\Big(\sum_{i_2\in[N]}\gamma_{i_2,j,j'}\Big)\Big(\sum_{i_3\in[N]}\gamma_{i_3,j,j'}\Big)\Big(\sum_{i_4\in[N]}\gamma_{i_4,j,j'}\Big)\Big]$$
$$\overset{(a)}{=} \frac{1}{N^2}\sum_{i\in[N]}\mathbb{E}\Big[\gamma_{i,j,j'}^4\Big] + \frac{3}{N^2}\sum_{i\neq i'\in[N]}\mathbb{E}\Big[\gamma_{i,j,j'}^2\gamma_{i',j,j'}^2\Big] \leq c,$$

where (a) follows from linearity of expectation and Assumption 4.11(c) after by noting that $\mathbb{E}[\gamma_{i,j,j'}] = 0$ for all $i, j, j' \in [N] \times [M] \times [M]$ and (b) follows because $\gamma_{i,j,j'}$ has bounded moments due to Assumption 4.11(a).

4.H Data generating process for the simulations

The inputs of the data generating process (DGP) are: the probability bound λ; two positive constants $c^{(0)}$ and $c^{(1)}$; and the standard deviations $\sigma_{i,j}^{(a)}$ for every $i \in [N], j \in [M], a \in \{0,1\}$. The DGP is:

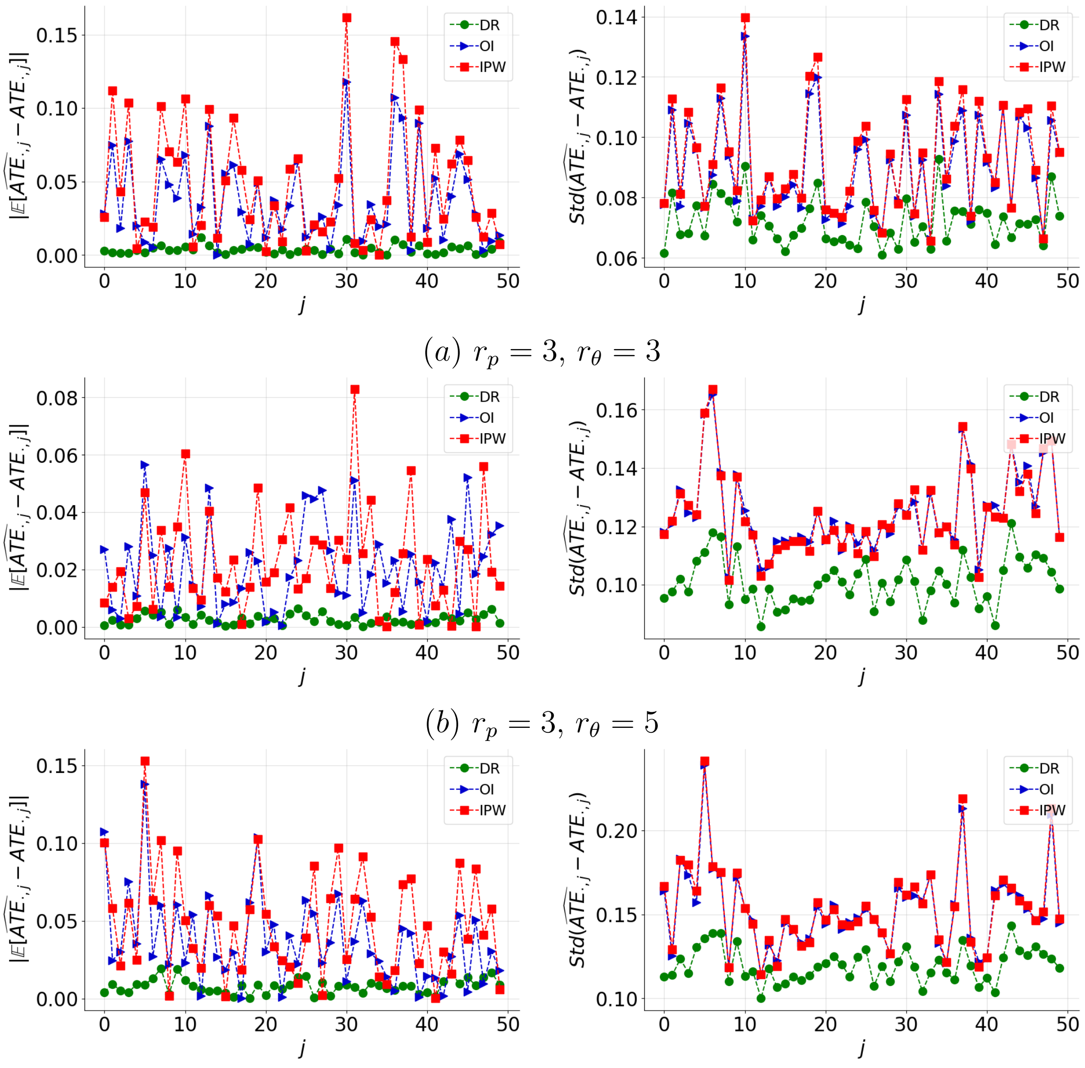
1. For positive integers r_p, r_θ and $r = \max\{r_p, r_\theta\}$, generate a proxy for the common unit-level latent factors $U^{\text{shared}} \in \mathbb{R}^{N\times r}$, such that, for all $i \in [N]$ and $j \in [r]$, $u_{i,j}^{\text{shared}}$ is independently sampled from a $\texttt{Uniform}(\sqrt{\lambda}, \sqrt{1-\lambda})$ distribution, with $\lambda \in (0,1)$.

2. Generate proxies for the measurement-level latent factors $V, V^{(0)}, V^{(1)} \in \mathbb{R}^{M\times r}$, such that, for all $i \in [M]$ and $j \in [r]$, $v_{i,j}, v_{i,j}^{(0)}, v_{i,j}^{(1)}$ are independently sampled from a $\texttt{Uniform}(\sqrt{\lambda}, \sqrt{1-\lambda})$ distribution.

3. Generate the treatment assignment probability matrix P
$$P = \frac{1}{r_p}U^{\text{shared}}_{[N]\times[r_p]}V^{\top}_{[M]\times[r_p]}.$$

4. For $a \in \{0,1\}$, run $\texttt{SVD}$ on $U^{\text{shared}}V^{(a)\top}$, i.e.,
$$\texttt{SVD}(U^{\text{shared}}V^{(a)\top}) = (U^{(a)}, \Sigma^{(a)}, W^{(a)}).$$
Then, generate the mean potential outcome matrices $\Theta^{(0)}$ and $\Theta^{(1)}$:
$$\Theta^{(a)} = \frac{c^{(a)}\texttt{Sum}(\Sigma^{(a)})}{r_\theta}U^{(a)}_{[N]\times[r_\theta]}W^{(a)\top}_{[M]\times[r_\theta]},$$
where $\texttt{Sum}(\Sigma^{(a)})$ denotes the sum of all entries of $\Sigma^{(a)}$.

5. Generate the noise matrices $E^{(0)}$ and $E^{(1)}$, such that, for all $i \in [N], j \in [M], a \in \{0,1\}$, $\varepsilon_{i,j}^{(a)}$ is independently sampled from a $\mathcal{N}(0, (\sigma_{i,j}^{(a)})^2)$ distribution. Then, determine $y_{i,j}^{(a)}$ from Eq. (4.2).



(a) $r_p = 3$, $r_\theta = 3$

(b) $r_p = 3$, $r_\theta = 5$

(c) $r_p = 5$, $r_\theta = 3$

Figure 4.H.1: Empirical illustration of the biases and the standard deviations of DR, OI, and IPW estimators for different j, and for different r_p and r_θ.

6. Generate the noise matrix W, such that, for all $i \in [N], j \in [M]$, $\eta_{i,j}$ is independently sampled as per Eq. (4.4). Then, determine $a_{i,j}$ and $y_{i,j}$ from Eq. (4.3) and Eq. (4.1), respectively.

In our simulations, we set $\lambda = 0.05$, $c^{(0)} = 1$ and $c^{(1)} = 2$. In practice, instead of choosing the values of $\sigma_{i,j}^{(a)}$ as ex-ante inputs, we make them equal to the standard deviation of all the entries in $\Theta^{(a)}$ for every i and j, separately for $a \in \{0, 1\}$.

In Figure 4.H.1, we compare the absolute biases and the standard deviations of OI, IPW, and DR across the first 50 values of j for $N = 1000$, with $r_p = 3$, $r_\theta = 3$ in Panel (a), $r_p = 3$, $r_\theta = 5$ in Panel (b), and $r_p = 5$, $r_\theta = 3$ in Panel (c). For each j, the estimate of the biases of OI, IPW, and DR is the average of $\widehat{\mathrm{ATE}}_{\cdot,j}^{\mathrm{OI}} - \mathrm{ATE}_{\cdot,j}$, $\widehat{\mathrm{ATE}}_{\cdot,j}^{\mathrm{IPW}} - \mathrm{ATE}_{\cdot,j}$ and $\widehat{\mathrm{ATE}}_{\cdot,j}^{\mathrm{DR}} - \mathrm{ATE}_{\cdot,j}$ across the Q simulation instances. Likewise, the estimate of the standard deviation of OI, IPW, and DR is the standard deviation of $\widehat{\mathrm{ATE}}^{\mathrm{OI}}_{\cdot,j} - \mathrm{ATE}_{\cdot,j}$, $\widehat{\mathrm{ATE}}^{\mathrm{IPW}}_{\cdot,j} - \mathrm{ATE}_{\cdot,j}$ and $\widehat{\mathrm{ATE}}^{\mathrm{DR}}_{\cdot,j} - \mathrm{ATE}_{\cdot,j}$ across the Q simulation instances. The DR estimator consistently outperforms the OI and IPW estimators in reducing both absolute biases and standard deviations.

4.I Doubly-robust estimation in panel data with lagged effects

This section describes how the doubly-robust framework of this article can be generalized to a panel data setting with lagged treatment effects. We highlight that, as is the convention in a panel data setting, t denotes the column (time) index and T denotes the total number of columns (time periods).

4.I.1 Formulation

As described in Section 4.4.4, potential outcomes are generated as follows: for all $i \in [N], t \in [T]$, and $a \in \{0, 1\}$,

$$y_{i,t}^{(a|y_{i,t-1})} = \alpha^{(a)} y_{i,t-1} + \theta_{i,t}^{(a)} + \varepsilon_{i,t}^{(a)}, \tag{4.89}$$

where $y_{i,t}^{(a|y_{i,t-1})}$ is the potential outcome for unit i at time t given treatment $a \in \{0, 1\}$ and lagged outcome $y_{i,t-1}$. This model combines unobserved confounding and lagged treatment effects, where the lagged effect is carried over via the auto-regressive term, $\alpha^{(a)} y_{i,t-1}$, with $\alpha^{(a)}$ being the auto-regressive parameter for treatment $a \in \{0, 1\}$. The treatment possibly starts at $t = 1$, and $y_{i,0}$ is assumed to not be affected by any future exposure to the treatment. Treatment assignments are continually assumed to be generated via Eq. (4.3). As in Eq. (4.1), realized outcomes, $y_{i,t}$, depend on potential outcomes and treatment assignments,

$$y_{i,t} = y_{i,t}^{(0|y_{i,t-1})}(1 - a_{i,t}) + y_{i,t}^{(1|y_{i,t-1})} a_{i,t}, \tag{4.90}$$

for all $i \in [N]$ and $t \in [T]$.

4.I.2 Target causal estimand

The lagged effects in Eq. (4.89) imply that the treatment effects need to be defined for sequences of treatments. For concreteness, consider the effect at time T for an always-treat policy, i.e., $a_{i,t} = 1$, versus never-treat, i.e., $a_{i,t} = 0$, for $i \in [N]$ and $j \in [T]$. Let $y_{i,T}^{[1]}$ be the potential outcome for unit i at time T under always-treat and $y_{i,T}^{[0]}$ be the potential outcome for unit i at time T under never-treat. We aim to estimate the difference in the expected potential outcomes under these two treatment policies averaged over all units,

$$\mathrm{ATE}_{\cdot,T} \triangleq \mu_{\cdot,T}^{[1]} - \mu_{\cdot,T}^{[0]}, \quad \text{where} \quad \mu_{\cdot,T}^{[a]} \triangleq \frac{1}{N} \sum_{i \in [N]} \mathbb{E}[y_{i,T}^{[a]}],$$

with the expectation taken over the distribution of $\{\varepsilon_{i,t}^{(a)}\}_{i\in[N],t\in[T]}$, conditioned on the initial outcomes $\{y_{i,0}\}_{i\in[N]}$. We make the following assumption about the noise in potential outcomes.

Assumption 4.14 (Zero-mean noise conditioned on the initial outcomes)**.** $\{\varepsilon_{i,t}^{(a)} : i \in [N], t \in [T], a \in \{0,1\}\}$ *are mean zero conditioned on* $\{y_{i,0}\}_{i\in[N]}$.

Assumption 4.14 holds whenever Assumption 4.2(a) holds conditioned on the initial outcomes $\{y_{i,0}\}_{i\in[N]}$. Another sufficient condition for Assumption 4.14 is that $(\varepsilon_{i,t}^{(0)}, \varepsilon_{i,t}^{(1)})$ are independent in time. Given this, the time dependence in the expected potential outcome $\mathbb{E}[y_{i,T}^{[a]}]$ is captured as follows: for $a \in \{0,1\}$

$$\mathbb{E}[y_{i,T}^{[a]}] = (\alpha^{(a)})^T y_{i,0} + \sum_{s=0}^{T-1} (\alpha^{(a)})^s \theta_{i,T-s}^{(a)}. \tag{4.91}$$

Eq. (4.91) forms the basis of our doubly-robust estimator of $\mathrm{ATE}_{\cdot,T}$.

We chose the contrast between always-treat and never-treat for concreteness. However, the framework and the results in this section can be generalized in a straightforward manner to contrast any two pre-specified sequences of treatments, where the treatment can also be chosen stochastically with pre-specified probabilities. For the remainder of this section, we condition on the initial outcomes $\{y_{i,0}\}_{i\in[N]}$ but omit it from our notation for brevity.

4.I.3 Doubly-robust estimator

The DR estimator of $\mathrm{ATE}_{\cdot,T}$ combines the estimates of $(\alpha^{(0)}, \alpha^{(1)})$, $(\Theta^{(0)}, \Theta^{(1)})$, and P. First, we obtain the estimates $(\widehat{\alpha}^{(0)}, \widehat{\alpha}^{(1)})$. These estimates can be computed using the likelihood approach of Bai (2024) whenever there exists some units such that they all have treatment a for some consecutive time points, for $a \in \{0,1\}$.

Next, we define the residual matrices $\widetilde{Y}^{(0),\mathrm{obs}}$ and $\widetilde{Y}^{(1),\mathrm{obs}}$. Let $\widetilde{Y}^{(0),\mathrm{obs}} \in \{\mathbb{R} \cup \{\,?\}\}^{N\times T}$ be a matrix with (i,t)-th entry equal to $y_{i,t} - \widehat{\alpha}^{(0)} y_{i,t-1}$ if $a_{i,t} = 0$, and equal to ? otherwise. Analogously, let $\widetilde{Y}^{(1),\mathrm{obs}} \in \{\mathbb{R} \cup \{\,?\}\}^{N\times T}$ be a matrix with (i,t)-th entry equal to $y_{i,t} - \widehat{\alpha}^{(1)} y_{i,t-1}$ if $a_{i,t} = 1$, and equal to ? otherwise. Then, similar to Eq. (4.8), the application of matrix completion yields the following estimates:

$$\widehat{\Theta}^{(0)} = \mathtt{MC}(\widetilde{Y}^{(0),\mathrm{obs}}), \quad \widehat{\Theta}^{(1)} = \mathtt{MC}(\widetilde{Y}^{(1),\mathrm{obs}}), \quad \text{and} \quad \widehat{P} = \mathtt{MC}(A). \tag{4.92}$$

Then, the DR estimate is defined as follows:

$$\widehat{\mathrm{ATE}}_{\cdot,T,J}^{\mathrm{DR}} \triangleq \widehat{\mu}_{\cdot,T,J}^{[1,\mathrm{DR}]} - \widehat{\mu}_{\cdot,T,J}^{[0,\mathrm{DR}]} \quad \text{where} \quad \widehat{\mu}_{\cdot,T,J}^{[a,\mathrm{DR}]} = \frac{1}{N} \sum_{i\in[N]} \left[(\widehat{\alpha}^{(a)})^T y_{i,0} + \sum_{s=0}^{J-1} (\widehat{\alpha}^{(a)})^s \widehat{\theta}_{i,T-s}^{[a,\mathrm{DR}]} \right], \tag{4.93}$$

where

$$\widehat{\theta}_{i,T-s}^{[0,\mathrm{DR}]} \triangleq \widehat{\theta}_{i,T-s}^{(0)} + \big(y_{i,T-s} - \widehat{\alpha}^{(0)} y_{i,T-s-1} - \widehat{\theta}_{i,T-s}^{(0)}\big) \frac{1 - a_{i,T-s}}{1 - \widehat{p}_{i,T-s}},$$

and

$$\widehat{\theta}^{[1,\mathrm{DR}]}_{i,T-s} \triangleq \widehat{\theta}^{(1)}_{i,T-s} + \big(y_{i,T-s} - \widehat{\alpha}^{(1)} y_{i,T-s-1} - \widehat{\theta}^{(1)}_{i,T-s}\big)\frac{a_{i,T-s}}{\widehat{p}_{i,T-s}}$$

The estimator is parameterized by an integer J, which denotes the contiguous number of time periods preceding time T that are used to estimate the expectations at time T (see the summation in Eq. (4.91)). Notably, using preceding J terms instead of $T-1$ terms allows us to adapt cross-fitting for the setting with lagged treatment effects. Let us briefly elaborate: suppose $(\widehat{\alpha}^{(0)}, \widehat{\alpha}^{(1)})$ are estimated from entries of Y in $[N] \times [L]$ for some $L < T-J$. Consider the column partitions $\mathcal{C}_0 = \{L+1, \ldots, T-J\}$ and $\mathcal{C}_1 = \{T-J+1, \ldots, T\}$ of times $[T] \setminus [L]$. Suppose Eqs. (4.27) and (4.28) in Assumption 4.5 hold for $\mathcal{I} = \mathcal{R}_0 \times \mathcal{C}_1$ and $\mathcal{I} = \mathcal{R}_1 \times \mathcal{C}_1$ for some row partitions $\mathcal{R}_0$ and $\mathcal{R}_1$ of units $[N]$. Then, applying `Cross-Fitted-MC` on the residual matrices $\widetilde{Y}^{(0),\mathrm{obs}}$ and $\widetilde{Y}^{(1),\mathrm{obs}}$ with row partitions $(\mathcal{R}_0, \mathcal{R}_1)$ and column partitions $(\mathcal{C}_0, \mathcal{C}_1)$ ensures that Assumption 4.4 holds for every column in $\mathcal{C}_1$ with row partitions $(\mathcal{R}_0, \mathcal{R}_1)$.

4.I.4 Non-asymptotic guarantees

Recall the notation for $\mathcal{E}\big(\widehat{\Theta}\big)$ and $\mathcal{E}\big(\widehat{P}\big)$ from Eq. (4.16) and define

$$\mathcal{E}(\widehat{\alpha}) \triangleq \sum_{a\in\{0,1\}} \mathcal{E}(\widehat{\alpha}^{(a)}) \quad \text{where} \quad \mathcal{E}(\widehat{\alpha}^{(a)}) \triangleq |\widehat{\alpha}^{(a)} - \alpha^{(a)}|. \tag{4.94}$$

Our analysis makes two additional assumptions to state a non-asymptotic error bound for $\widehat{\mathrm{ATE}}^{\mathrm{DR}}_{\cdot,T,J} - \mathrm{ATE}_{\cdot,T}$.

Assumption 4.15 (Bounded auto-regressive parameters and estimates)**.** *The auto-regressive parameters and their estimates are such that* $|\alpha^{(a)}| \leq \overline{\alpha}$ *and* $|\widehat{\alpha}^{(a)}| \leq \overline{\alpha}$*, for all* $a \in \{0,1\}$*, where* $\overline{\alpha} \in [0,1)$*.*

Assumption 4.15 requires the regression parameters to be bounded by a fixed constant less than 1. This condition is standard for auto-regressive models, as it implies stability of the outcome process in Eq. (4.89). The analogous condition on the estimated parameters can be ensured by truncating the estimates to $[0, \overline{\alpha}]$.

Assumption 4.16 (Bounded observed outcomes, mean potential outcomes, and estimated mean potential outcomes)**.** *The observed outcomes, the mean potential outcomes, and the estimates of the mean potential outcomes are such that* $|y_{i,t}| \leq C_1$*,* $|\theta^{(a)}_{i,t}| \leq C_2$*, and* $|\widehat{\theta}^{(a)}_{i,t}| \leq C_3$*, for all* $i \in [N]$*,* $j \in [M]$*, and* $a \in \{0,1\}$*, where* C_1*,* C_2*, and* C_3 *are universal constants.*

Assumption 4.16 requires the observed outcomes, the mean potential outcomes, and the estimates of the mean potential outcomes to be bounded to simplify our proof. With a more delicate analysis, Assumption 4.16 can be relaxed to require the average observed outcomes over $i \in [N]$, the average mean potential outcomes over $i \in [N]$, and the average estimated mean potential outcomes over $i \in [N]$ to be bounded.

Theorem 4.3 (Finite Sample Guarantees for DR with lagged effects)**.** *Consider the panel data model with lagged effects defined via Eqs.* (4.89) *and* (4.90)*. Suppose Assumptions 4.1 to 4.3, 4.15, and 4.16 hold and Assumption 4.4 holds for $t \in \{T-J+1, \ldots, T\}$ for some integer $J \in [T]$. Fix $\delta \in (0,1)$. Then, with probability at least $1-\delta$, we have*

$$\left|\widehat{\mathrm{ATE}}^{\mathrm{DR}}_{\cdot,T,J} - \mathrm{ATE}_{\cdot,T}\right| \leq \frac{\mathrm{Err}^{\mathrm{DR}}_{N,\delta/J}}{1-\overline{\alpha}} + C\left[\frac{\overline{\alpha}^{J}}{1-\overline{\alpha}} + \mathcal{E}(\widehat{\alpha})\Big(T\overline{\alpha}^{T-1} + \frac{1}{1-\overline{\alpha}}\Big)\right], \tag{4.95}$$

for $\mathrm{Err}^{\mathrm{DR}}_{N,\delta}$ as defined in Eq. (4.18) *in Theorem 4.1 and a universal constant C.*

The proof of Theorem 4.3 is given in Section 4.I.5. For brevity, the finite sample guarantees above uses $\mathcal{E}(\widehat{\Theta})$ and $\mathcal{E}(\widehat{P})$ as defined in Eq. (4.16), but the proof can be easily modified to replace the $\max_{j\in[T]}$ appearing in the definition of $\|\cdot\|_{1,2}$ in Eq. (4.16) with $\max_{j\in\{T-J+1,\cdots,T\}}$.

Next, we remark that Theorem 4.3 is a strict generalization of Theorem 4.1. To this end, note that when $\alpha^{(a)} = 0$ for all $a \in \{0,1\}$, the model considered in Theorem 4.3 simplifies to the model considered in Theorem 4.1. For this setting, the assumptions in Theorem 4.1 imply that the assumptions in Theorem 4.3 hold with $J = 1$. First, Assumption 4.15 holds with $\overline{\alpha} = 0$ when $\alpha^{(a)} = 0$ for all $a \in \{0,1\}$. Second, the proof of Theorem 4.3 can be easily modified to drop the requirement of Assumption 4.16 when $J = 1$ and $\overline{\alpha} = 0$. Substituting $\overline{\alpha} = 0$, $\mathcal{E}(\widehat{\alpha}) = 0$ (i.e., the auto-regressive parameters are known to be 0), and $J = 1$ in Eq. (4.95) recovers the guarantee stated in Theorem 4.1.

Doubly-robust behavior of $\widehat{\mathrm{ATE}}^{\mathrm{DR}}_{\cdot,T,J}$. When $\overline{\alpha} \neq 0$ and bounded away from one, Eq. (4.95) bounds the absolute error of the DR estimator by the rate of

$$\mathcal{E}(\widehat{\Theta})\Big(\mathcal{E}(\widehat{P}) + \sqrt{\frac{\log J}{N}}\Big) + \frac{1}{\sqrt{N}} + \overline{\alpha}^{J} + \mathcal{E}(\widehat{\alpha}).$$

Then, if the conditions of Theorem 4.3 are satisfied for some J such that $C\log N \geq J \geq \log N/(2\log(1/\overline{\alpha}))$, the error rate of the DR estimator is bounded by

$$\mathcal{E}(\widehat{\Theta})\Big(\mathcal{E}(\widehat{P}) + \sqrt{\frac{\log\log N}{N}}\Big) + \frac{1}{\sqrt{N}} + \mathcal{E}(\widehat{\alpha}),$$

which decays a parametric rate of $O_p(N^{-0.5})$ as long as

$$\mathcal{E}(\widehat{\Theta})\mathcal{E}(\widehat{P}) = O_p\Big(\frac{1}{\sqrt{N}}\Big), \quad \mathcal{E}(\widehat{\Theta}) = O_p\Big(\frac{1}{\sqrt{\log\log N}}\Big), \quad \text{and} \quad \mathcal{E}(\widehat{\alpha}) = O_p\Big(\frac{1}{\sqrt{N}}\Big).$$

Note that Proposition 4.4 still implies that `Cross-Fitted-SVD` achieves $\mathcal{E}(\widehat{P}) = O_p(N^{-0.5} + T^{-0.5})$ under suitable conditions. To estimate the auto-regressive parameter $\alpha^{(a)}$ for $a \in \{0,1\}$, Bai (2024, Section 5) shows that whenever there exist K units such that they all have treatment a for L consecutive time points, a full information maximum likelihood estimator provides $|\alpha^{(a)} - \widehat{\alpha}^{(a)}| = O_p((KL)^{-0.5})$. Next, establishing a matrix completion guarantee for the mean potential outcomes by residualizing as in Eq. (4.92) can be reduced to deriving a matrix completion guarantee for an approximately low-rank matrix. To this end, Agarwal and Singh (2024, Theorem 5) suggests that, up to logarithmic factors, an error rate of $N^{-0.5} + T^{-0.5} + \mathcal{E}(\widehat{\alpha})$ is plausible for $\mathcal{E}\big(\widehat{\Theta}\big)$ for our setting. A complete derivation of error guarantees for $\mathcal{E}(\widehat{\alpha})$ and $\mathcal{E}\big(\widehat{\Theta}\big)$ in the dynamic model is an interesting venue for future work.

4.I.5 Proof of Theorem 4.3: Finite Sample Guarantees for DR with lagged effects

The error $\Delta\mathrm{ATE}^{\mathrm{DR}}_{\cdot,T} = \widehat{\mathrm{ATE}}^{\mathrm{DR}}_{\cdot,T,J} - \mathrm{ATE}_{\cdot,T}$ can be re-expressed as

$$\Delta\mathrm{ATE}^{\mathrm{DR}}_{\cdot,T} = \big(\widehat{\mu}^{[1,\mathrm{DR}]}_{\cdot,T,J} - \widehat{\mu}^{[0,\mathrm{DR}]}_{\cdot,T,J}\big) - \big(\mu^{[1]}_{\cdot,T} - \mu^{[0]}_{\cdot,T}\big) = \big(\widehat{\mu}^{[1,\mathrm{DR}]}_{\cdot,T,J} - \mu^{[1]}_{\cdot,T}\big) - \big(\widehat{\mu}^{[0,\mathrm{DR}]}_{\cdot,T,J} - \mu^{[0]}_{\cdot,T}\big). \tag{4.96}$$

We claim that, with probability at least $1-\delta$,

$$\begin{aligned}\Big|\widehat{\mu}^{[1,\mathrm{DR}]}_{\cdot,T,J} - \mu^{[1]}_{\cdot,T}\Big| &\leq C\left[\frac{|\alpha^{(1)}|^J - |\alpha^{(1)}|^T}{1-|\alpha^{(1)}|} + \mathcal{E}(\widehat{\alpha}^{(1)})\Big(T\overline{\alpha}^{T-1} + \frac{1-|\alpha^{(1)}|^J}{1-|\alpha^{(1)}|} + \frac{1}{(1-|\alpha^{(1)}|)^2}\Big)\right] \\ &+ \frac{2}{(1-|\alpha^{(1)}|)\bar{\lambda}}\left[\mathcal{E}(\widehat{\Theta}^{(1)})\mathcal{E}\big(\widehat{P}\big) + \frac{1}{\sqrt{N}}\Big(\frac{\sqrt{c\ell_{\delta/(12J)}}}{\sqrt{\ell_1}}\mathcal{E}(\widehat{\Theta}^{(1)}) + 2\overline{\sigma}\sqrt{c\ell_{\delta/(12J)}} + \frac{2\overline{\sigma}m(c\ell_{\delta/(12J)})}{\sqrt{\ell_1}}\Big)\right],\end{aligned} \tag{4.97}$$

and

$$\begin{aligned}\Big|\widehat{\mu}^{[0,\mathrm{DR}]}_{\cdot,T,J} - \mu^{[0]}_{\cdot,T}\Big| &\leq C\left[\frac{|\alpha^{(0)}|^J - |\alpha^{(0)}|^T}{1-|\alpha^{(0)}|} + \mathcal{E}(\widehat{\alpha}^{(0)})\Big(T\overline{\alpha}^{T-1} + \frac{1-|\alpha^{(0)}|^J}{1-|\alpha^{(0)}|} + \frac{1}{(1-|\alpha^{(0)}|)^2}\Big)\right] \\ &+ \frac{2}{(1-|\alpha^{(0)}|)\bar{\lambda}}\left[\mathcal{E}(\widehat{\Theta}^{(0)})\mathcal{E}\big(\widehat{P}\big) + \frac{1}{\sqrt{N}}\Big(\frac{\sqrt{c\ell_{\delta/(12J)}}}{\sqrt{\ell_1}}\mathcal{E}(\widehat{\Theta}^{(0)}) + 2\overline{\sigma}\sqrt{c\ell_{\delta/(12J)}} + \frac{2\overline{\sigma}m(c\ell_{\delta/(12J)})}{\sqrt{\ell_1}}\Big)\right].\end{aligned} \tag{4.98}$$

Then, the claim in Eq. (4.95) follows by applying triangle inequality in Eq. (4.96) and using Assumption 4.15. We prove the bound (4.97) in Section 4.I.5.1, and also provide an expression for C. The proof of Eq. (4.98) follows similarly.

4.I.5.1 Proof of Eq. (4.97)

We start by decomposing $\mu^{[1]}_{\cdot,T}$ as follows:

$$\mu^{[1]}_{\cdot,T} = \frac{1}{N}\left[\sum_{i\in[N]}(\alpha^{(1)})^T y_{i,0} + \sum_{s=0}^{T-1}(\alpha^{(1)})^s\sum_{i\in[N]}\theta^{(1)}_{i,T-s}\right] = \mathbb{T}^{(1)}_J + \mathbb{U}^{(1)}_J + \mathbb{V}^{(1)},$$

where

$$\mathbb{T}^{(1)}_J \triangleq \frac{1}{N}\sum_{s=0}^{J-1}(\alpha^{(1)})^s\sum_{i\in[N]}\theta^{(1)}_{i,T-s}, \quad \mathbb{U}^{(1)}_J \triangleq \frac{1}{N}\sum_{s=J}^{T-1}(\alpha^{(1)})^s\sum_{i\in[N]}\theta^{(1)}_{i,T-s}, \tag{4.99}$$

and

$$\mathbb{V}^{(1)} \triangleq (\alpha^{(1)})^T \frac{1}{N} \sum_{i\in[N]} y_{i,0}. \tag{4.100}$$

Next, we decompose $\widehat{\mu}^{[1,\mathrm{DR}]}_{\cdot,T,J}$ in Eq. (4.93) as $\widehat{\mu}^{[1,\mathrm{DR}]}_{\cdot,T,J} = \widehat{\mathbb{T}}^{(1)}_J + \widehat{\mathbb{V}}^{(1)}$, where

$$\widehat{\mathbb{T}}^{(1)}_J \triangleq \frac{1}{N}\sum_{s=0}^{J-1}(\widehat{\alpha}^{(1)})^s \sum_{i\in[N]} \widehat{\theta}^{[1,\mathrm{DR}]}_{i,T-s}, \quad \text{and} \quad \widehat{\mathbb{V}}^{(1)} \triangleq (\widehat{\alpha}^{(1)})^T \frac{1}{N}\sum_{i\in[N]} y_{i,0}. \tag{4.101}$$

Finally, we define

$$\widetilde{\mathbb{T}}^{(1)}_J \triangleq \frac{1}{N}\sum_{s=0}^{J-1}(\alpha^{(1)})^s \sum_{i\in[N]} \left[\widehat{\theta}^{(1)}_{i,T-s} + \big(y_{i,T-s} - \alpha^{(1)} y_{i,T-s-1} - \widehat{\theta}^{(1)}_{i,T-s}\big)\frac{a_{i,T-s}}{\widehat{p}_{i,T-s}}\right], \tag{4.102}$$

which is similar to $\widehat{\mathbb{T}}^{(1)}_J$ except that $\widehat{\alpha}^{(1)}$ is replaced by $\alpha^{(1)}$. The proof proceeds by bounding each term in the following fundamental decomposition:

$$\widehat{\mu}^{[1,\mathrm{DR}]}_{\cdot,T,J} - \mu^{[1]}_{\cdot,T} = (\widehat{\mathbb{V}}^{(1)} - \mathbb{V}^{(1)}) + (\widetilde{\mathbb{T}}^{(1)}_J - \mathbb{T}^{(1)}_J) + (\widehat{\mathbb{T}}^{(1)}_J - \widetilde{\mathbb{T}}^{(1)}_J) - \mathbb{U}^{(1)}_J. \tag{4.103}$$

With $C_0 \triangleq \max_{i\in[N]} |y_{i,0}|$ and $C_{\mathrm{DR}} \triangleq C_3 + (2C_1 + C_3)/\bar{\lambda}$, we claim that the bounds

$$\big|\widehat{\mathbb{V}}^{(1)} - \mathbb{V}^{(1)}\big| \le C_0 T \mathcal{E}(\widehat{\alpha}^{(1)})\overline{\alpha}^{T-1}, \qquad |\mathbb{U}^{(1)}_J| \le C_2 \frac{|\alpha^{(1)}|^J - |\alpha^{(1)}|^T}{1 - |\alpha^{(1)}|}, \tag{4.104}$$

and

$$|\widehat{\mathbb{T}}^{(1)}_J - \widetilde{\mathbb{T}}^{(1)}_J| \le \mathcal{E}(\widehat{\alpha}^{(1)})\bigg(\frac{C_1}{\lambda}\frac{(1 - |\alpha^{(1)}|^J)}{1 - |\alpha^{(1)}|} + C_{\mathrm{DR}}\frac{1}{(1 - |\alpha^{(1)}|)^2}\bigg), \tag{4.105}$$

hold deterministically (conditioned on $\widehat{\alpha}^{(1)}$), and that the bound

$$\begin{aligned}|\widetilde{\mathbb{T}}^{(1)}_J - \mathbb{T}^{(1)}_J| \le \frac{2}{(1 - |\alpha^{(1)}|)\bar{\lambda}}\bigg[&\mathcal{E}(\widehat{\Theta}^{(1)})\mathcal{E}(\widehat{P}) \\ &+ \Big(\frac{\sqrt{c\ell_{\delta/(12J)}}}{\sqrt{\ell_1}}\mathcal{E}(\widehat{\Theta}^{(1)}) + 2\overline{\sigma}\sqrt{c\ell_{\delta/(12J)}} + \frac{2\overline{\sigma}m(c\ell_{\delta/(12J)})}{\sqrt{\ell_1}}\Big)\frac{1}{\sqrt{N}}\bigg],\end{aligned} \tag{4.106}$$

holds with probability at least $1 - \delta/2$. The claim in Eq. (4.97) follows by applying triangle inequality in Eq. (4.103) and using the above bounds.

It remains to establish the intermediate claims Eqs. (4.104) to (4.106). Throughout the rest of the proof, we repeatedly use the inequality below that holds for all $s \in [T]$:

$$\begin{aligned}\big|(\widehat{\alpha}^{(1)})^s - (\alpha^{(1)})^s\big| = \Big|(\widehat{\alpha}^{(1)} - \alpha^{(1)})\big(\sum_{l\in[s]}(\widehat{\alpha}^{(1)})^{s-l}(\alpha^{(1)})^{l-1}\big)\Big| &\overset{(a)}{\le} s\big|(\widehat{\alpha}^{(1)} - \alpha^{(1)})\big|\overline{\alpha}^{s-1} \\ &\overset{(b)}{=} s\mathcal{E}(\widehat{\alpha}^{(1)})\overline{\alpha}^{s-1},\end{aligned} \tag{4.107}$$

where (a) follows from Assumption 4.15 and (b) follows from Eq. (4.94).

Proof of Eq. (4.104) First, from Eq. (4.99), we have

$$|\mathbb{U}_J^{(1)}| = \Big|\frac{1}{N}\sum_{s=J}^{T-1}(\alpha^{(1)})^s \sum_{i\in[N]} \theta^{(1)}_{i,T-s}\Big| \overset{(a)}{\leq} C_2 \sum_{s=J}^{T-1}\big|\alpha^{(1)}\big|^s \overset{(b)}{=} C_2\frac{|\alpha^{(1)}|^J - |\alpha^{(1)}|^T}{1-|\alpha^{(1)}|},$$

where (a) follows from Assumption 4.16 and (b) follows from the sum of geometric series. Next, from Eqs. (4.100) and (4.101), we have

$$\big|\widehat{\mathbb{V}}^{(1)} - \mathbb{V}^{(1)}\big| = \Big|\Big((\widehat{\alpha}^{(1)})^T - (\alpha^{(1)})^T\Big)\frac{1}{N}\sum_{i\in[N]} y_{i,0}\Big| \overset{(a)}{\leq} C_0 T\mathcal{E}(\widehat{\alpha}^{(1)})\overline{\alpha}^{T-1},$$

where (a) follows from the definition of C_0 and Eq. (4.107).

Proof of Eq. (4.105) From Eqs. (4.101) and (4.102), and triangle inequality, we have

$$\begin{aligned}
\big|\widehat{\mathbb{T}}_J^{(1)} - \widetilde{\mathbb{T}}_J^{(1)}\big| \leq \frac{1}{N}\sum_{i\in[N]}\sum_{s=0}^{J-1}\Bigg|&(\widehat{\alpha}^{(1)})^s\Big(\widehat{\theta}^{(1)}_{i,T-s} + \big(y_{i,T-s} - \widehat{\alpha}^{(1)}y_{i,T-s-1} - \widehat{\theta}^{(1)}_{i,T-s}\big)\frac{a_{i,T-s}}{\widehat{p}_{i,T-s}}\Big) \\
&- (\alpha^{(1)})^s\left(\widehat{\theta}^{(1)}_{i,T-s} + \big(y_{i,T-s} - \alpha^{(1)}y_{i,T-s-1} - \widehat{\theta}^{(1)}_{i,T-s}\big)\frac{a_{i,T-s}}{\widehat{p}_{i,T-s}}\right)\Bigg| \\
= \frac{1}{N}\sum_{i\in[N]}\sum_{s=0}^{J-1}\Bigg|&(\alpha^{(1)})^s(\alpha^{(1)} - \widehat{\alpha}^{(1)})y_{i,T-s-1}\frac{a_{i,T-s}}{\widehat{p}_{i,T-s}} + \Big((\widehat{\alpha}^{(1)})^s - (\alpha^{(1)})^s\Big) \\
&\cdot\Big(\widehat{\theta}^{(1)}_{i,T-s} + \big(y_{i,T-s} - \widehat{\alpha}^{(1)}y_{i,T-s-1} - \widehat{\theta}^{(1)}_{i,T-s}\big)\frac{a_{i,T-s}}{\widehat{p}_{i,T-s}}\Big)\Bigg| \\
\overset{(a)}{\leq} \frac{1}{N}\sum_{i\in[N]}\sum_{s=0}^{J-1}\Big|&\frac{C_1}{\lambda}|\alpha^{(1)}|^s\mathcal{E}(\widehat{\alpha}^{(1)}) + C_{\mathrm{DR}}s\mathcal{E}(\widehat{\alpha}^{(1)})\overline{\alpha}^{s-1}\Big| \\
= \mathcal{E}(\widehat{\alpha}^{(1)})\bigg(&\frac{C_1}{\lambda}\frac{(1-|\alpha^{(1)}|^J)}{1-|\alpha^{(1)}|} + C_{\mathrm{DR}}\frac{1}{(1-|\alpha^{(1)}|)^2}\bigg),
\end{aligned}$$

where (a) follows from Eq. (4.94), Assumptions 4.3 and 4.16, and because $\max_{i\in[N],t\in[T]}$ $\big|\widehat{\theta}^{[1,\mathrm{DR}]}_{i,t}\big| \leq C_{\mathrm{DR}}$ from Assumptions 4.3, 4.15, and 4.16, and (b) follows from the sum of geometric and arithmetico-geometric sequences.

Proof of Eq. (4.106) We start by defining

$$\widetilde{\theta}^{[1,\mathrm{DR}]}_{i,T-s} \triangleq \widehat{\theta}^{(1)}_{i,T-s} + \big(y_{i,T-s} - \alpha^{(1)}y_{i,T-s-1} - \widehat{\theta}^{(1)}_{i,T-s}\big)\frac{a_{i,T-s}}{\widehat{p}_{i,T-s}}.$$

Then, from Eqs. (4.99) and (4.102), we have

$$|\widetilde{\mathbb{T}}_J^{(1)} - \mathbb{T}_J^{(1)}| = \Big|\sum_{s=0}^{J-1}(\alpha^{(1)})^s\frac{1}{N}\sum_{i\in[N]}(\widetilde{\theta}^{[1,\mathrm{DR}]}_{i,T-s} - \theta^{(1)}_{i,T-s})\Big| \overset{(a)}{\leq} \sum_{s=0}^{J-1}|\alpha^{(1)}|^s\frac{1}{N}\Big|\sum_{i\in[N]}(\widetilde{\theta}^{[1,\mathrm{DR}]}_{i,T-s} - \theta^{(1)}_{i,T-s})\Big|,$$

where (a) follows from triangle inequality. From Eqs. (4.3) and (4.89), we have

$$\widetilde{\theta}^{[1,\mathrm{DR}]}_{i,T-s} - \theta^{(1)}_{i,T-s} = \widehat{\theta}^{(1)}_{i,T-s} + (\theta^{(1)}_{i,T-s} + \varepsilon^{(1)}_{i,T-s} - \widehat{\theta}^{(1)}_{i,T-s})\frac{p_{i,T-s} + \eta_{i,T-s}}{\widehat{p}_{i,T-s}} - \theta^{(1)}_{i,T-s}.$$

Then, the term $\widetilde{\theta}^{[1,\mathrm{DR}]}_{i,T-s} - \theta^{(1)}_{i,T-s}$ is analogous to the display Eq. (4.37) in the proof of Theorem 4.1. Following similar algebra as in Appendix 4.B, we first obtain

$$\widetilde{\theta}^{[1,\mathrm{DR}]}_{i,T-s} - \theta^{(1)}_{i,T-s} = \frac{(\widehat{\theta}^{(1)}_{i,T-s} - \theta^{(1)}_{i,T-s})(\widehat{p}_{i,T-s} - p_{i,T-s})}{\widehat{p}_{i,T-s}} - \frac{(\widehat{\theta}^{(1)}_{i,T-s} - \theta^{(1)}_{i,T-s})\eta_{i,T-s}}{\widehat{p}_{i,T-s}} \\ + \frac{\varepsilon^{(1)}_{i,T-s} p_{i,T-s}}{\widehat{p}_{i,T-s}} + \frac{\varepsilon^{(1)}_{i,T-s}\eta_{i,T-s}}{\widehat{p}_{i,T-s}}.$$

Now, note that Assumption 4.4 holds for $j = T - s$ for all $s \in \{0, \ldots, J-1\}$. Hence, for any such s and for any $\delta \in (0,1)$, mimicking the derivation of Eq. (4.40) from Appendix 4.B, we obtain, with probability at least $1 - \delta/(2J)$,

$$\frac{1}{N}\Big|\sum_{i\in[N]}(\widetilde{\theta}^{[1,\mathrm{DR}]}_{i,T-s} - \theta^{(1)}_{i,T-s})\Big| \leq \frac{2}{\bar{\lambda}}\mathcal{E}(\widehat{\Theta}^{(1)})\mathcal{E}(\widehat{P}) + \frac{2\sqrt{c\ell_{\delta/(12J)}}}{\bar{\lambda}\sqrt{\ell_1 N}}\mathcal{E}(\widehat{\Theta}^{(1)}) + \frac{2\overline{\sigma}\sqrt{c\ell_{\delta/(12J)}}}{\bar{\lambda}\sqrt{N}} + \\ \frac{2\overline{\sigma}m(c\ell_{\delta/(12J)})}{\bar{\lambda}\sqrt{\ell_1 N}}. \tag{4.108}$$

Finally, multiplying both sides of Eq. (4.108) by $(\alpha^{(1)})^s$, summing it over $s \in \{0, \ldots, J-1\}$, and using a union bound argument yields that the bound in Eq. (4.106) holds with probability at least $1 - \delta/2$.

4.J Doubly-robust estimation in panel data with staggered adoption

This section shows how to extend the doubly-robust framework of this article to a setting with panel data and staggered adoption. Recall (from Section 4.I) that for panel data, t denotes the column (time) index and T denotes the total number of columns (time periods). In a staggered adoption setting, for every unit $i \in [N]$, there exists a time point $t_i \in [T]$ such that $a_{i,t} = 0$ for $t \leq t_i$, and $a_{i,t} = 1$ for $t > t_i$. This defines the observed treatment assignment matrix A. As mentioned in Section 4.5.4 and illustrated in the example below, a staggered treatment assignment leads to a heavy time-series dependence in $\{\eta_{i,t}\}_{t\in[T]}$.

Example 4.1 (Single adoption time)**.** *Consider a panel data setting where all units remain in the control group until time T_0. At time $T_0 + 1$, each unit $i \in [N]$ receives treatment with probability p_i, and remains in treatment until time T. With probability $1 - p_i$, each unit $i \in [N]$ stays in the control group until time T. In other words, for each unit $i \in [N]$*

$$p_{i,t} = 0 \quad \textit{for all} \quad t \leq T_0 \quad \textit{and} \quad p_{i,t} = p_i \quad \textit{for all} \quad T_0 < t \leq T.$$

Further, for units remaining in control,

$$\eta_{i,t} = 0 \quad \textit{for all} \quad t \le T_0 \quad \textit{and} \quad \eta_{i,t} = -p_i \quad \textit{for all} \quad T_0 < t \le T,$$

and for units receiving treatment,

$$\eta_{i,t} = 0 \quad \textit{for all} \quad t \le T_0 \quad \textit{and} \quad \eta_{i,t} = 1 - p_i \quad \textit{for all} \quad T_0 < t \le T.$$

The strong time-series dependence in $\eta_{i,t}$ above implies that Assumption 4.8 or Assumption 4.9(a) do not hold, which in turn implies that the guarantees for `Cross-Fitted-SVD`, as in Proposition 4.4, may not hold. To see this, first note that to ensure Assumption 4.5, the set of column partitions $\{\mathcal{C}_0, \mathcal{C}_1\}$ must be equal to $\{[T_0], [T] \setminus [T_0]\}$ due to the dependence in the noise W. Now, for Assumption 4.8 to hold, we need $|\mathcal{C}_k| = \Omega(T)$ for every $k \in \{0, 1\}$. However, for Assumption 4.9(a) to hold, we need $T - T_0$ to be a constant with respect to T as, for any $t \in [T] \setminus [T_0]$ and $i \in [N]$, $\sum_{t' \in [T]} \left|\mathbb{E}[\eta_{i,t}\eta_{i,t'}]\right| = (T - T_0)c_i$ where $c_i \in \{p_i^2, (1 - p_i)^2\}$.

Moreover, in Example 4.1, $t_i = T_0$ for all treated units. This allows the choice of $\{[T_0], [T] \setminus [T_0]\}$ as the set of column partitions $\{\mathcal{C}_0, \mathcal{C}_1\}$ in Assumption 4.5. More generally, if treatment adoption times $\{t_i\}_{i \in [N]}$ differ across units, then it may not be feasible to obtain a partition of $[T]$ into $\{\mathcal{C}_0, \mathcal{C}_1\}$ such that Assumption 4.5 holds.

In this section, we propose an alternative approach to the `Cross-Fitted-SVD` algorithm such that Assumption 4.4 still holds for a suitable staggered adoption model.

Assumption 4.17 (Staggered adoption and common unit factors)**.** *We consider a panel data setting with staggered adoption where*

1. *all units remain under control till time T_0, i.e., for every unit $i \in [N]$, there exists a time point $t_i \ge T_0$ such that $a_{i,t} = 0$ for $t \le t_i$, and $a_{i,t} = 1$ for $t > t_i$, and*
2. *the unit-dependent latent factors corresponding to P, $\Theta^{(0)}$, and $\Theta^{(1)}$ are the same, i.e., $U = U^{(0)} = U^{(1)} \in \mathbb{R}^{N \times r}$. In other words, for every $i \in [N]$ and $t \in [T]$, $p_{i,t} = g(U_i, V_t)$, $\theta_{i,t}^{(0)} = \langle U_i, V_t^{(0)} \rangle$, and $\theta_{i,t}^{(1)} = \langle U_i, V_t^{(1)} \rangle$ for some known function $g : \mathbb{R}^r \times \mathbb{R}^r \to \mathbb{R}$, with $\langle \cdot, \cdot \rangle$ denoting the inner product.*

For Example 4.1, the function g corresponds to the inner product, the unit-dependent latent factors are 1-dimensional (i.e., $r = 1$) with $U_i = p_i$ for every $i \in [N]$, and the time-dependent latent factors for the assignment probability are such that $V_t = 0$ for every $t \in [T_0]$ and $V_t = 1$ for every $t \in [T] \setminus [T_0]$. Consequently, Example 4.1 is consistent with Assumption 4.17 if $U_i^{(a)} = p_i$ for every $a \in \{0, 1\}$ and $i \in [N]$. Next, we provide a more flexible version of Example 4.1 that allows different adoption times for different units.

Example 4.2 (Different adoption times)**.** *Consider a panel data setting where all units remain in the control group until time T_0. At every time $t \in [T] \setminus [T_0]$, each unit $i \in [N]$ receives treatment with probability p_i, and remains in treatment until time T. Therefore, for $t \in [T] \setminus [T_0]$ and $i \in [N]$, $a_{i,t} = 1$ if the adoption time point $t_i \in \{T_0 + 1, \cdots, t\}$, which occurs with probability $\sum_{t' \in [t - T_0 - 1]} (1 - p_i)^{t' - 1} p_i$. In other words, for each unit $i \in [N]$,*

$$p_{i,t} = 0 \quad \textit{for all} \quad t \le T_0 \quad \textit{and} \quad p_{i,t} = 1 - (1 - p_i)^{t - T_0} \quad \textit{for all} \quad T_0 < t \le T.$$

For Example 4.2, the unit-dependent latent factors are 1-dimensional (i.e., $r = 1$) with $U_i = p_i$ for every $i \in [N]$, and the time-dependent latent factors for the assignment probability are such that $V_t = 0$ for every $t \in [T_0]$ and $V_t = t - T_0$ for every $t \in [T] \setminus [T_0]$. Further the function g is such that $g(U_i, V_t) = 1 - (1 - U_i)^{V_t}$. Consequently, Example 4.2 is consistent with Assumption 4.17 if $U_i^{(a)} = p_i$ for every $a \in \{0, 1\}$ and $i \in [N]$.

We now describe `Cross-Fitted-Regression`, an algorithm that generates estimates of $(\Theta^{(0)}, \Theta^{(1)}, P)$ for the staggered adoption model in Assumption 4.17 such that Assumption 4.4 holds.

1. The inputs are (i) $A \in \mathbb{R}^{N \times T}$, (ii) $Y^{(a),\text{obs}} \in \{\mathbb{R} \cup \{\,?\}\}^{N \times T}$ for $a \in \{0, 1\}$, (iii) the rank r of the unit-dependent latent factors, (iv) the time period T_0 until which all units remain under control, (v) the time period $t \in [T] \setminus [T_0]$ for which we want to estimate the average treatment effect, and (vi) the function g.
2. Let $Y^{(0),\text{pre}} \in \mathbb{R}^{N \times T_0}$ be the sub-matrix of $Y^{(0),\text{obs}}$ that keeps the first T_0 columns only. Run `SVD` on $Y^{(0),\text{pre}}$, i.e.,
$$\texttt{SVD}(Y^{(0),\text{pre}}) = (\widehat{U} \in \mathbb{R}^{N \times r}, \widehat{\Sigma} \in \mathbb{R}^{r \times r}, \widehat{V} \in \mathbb{R}^{|T_0| \times r}).$$
3. Let $\mathcal{R}^{(0)}$ and $\mathcal{R}^{(1)}$ be the set of units receiving control and treatment at time t, respectively. In other words, for every $a \in \{0, 1\}$, $\mathcal{R}^{(a)} \triangleq \{i \in [N] : a_{i,t} = a\}$. Next, randomly partition $\mathcal{R}^{(a)}$ into two nearly equal parts $\mathcal{R}_0^{(a)}$ and $\mathcal{R}_1^{(a)}$. For every $s \in \{0, 1\}$, define $\mathcal{R}_s = \mathcal{R}_s^{(0)} \cup \mathcal{R}_s^{(1)}$.
4. For every $s \in \{0, 1\}$, regress $\{a_{i,t}\}_{i \in \mathcal{R}_s}$ on $\{\widehat{U}_i\}_{i \in \mathcal{R}_s}$ using g to obtain $\widehat{V}_{1-s}$. For every $s \in \{0, 1\}$ and $i \in \mathcal{R}_s$, return $\widehat{p}_{i,t} = g(\widehat{U}_i, \widehat{V}_s)$.
5. For every $a \in \{0, 1\}$ and $s \in \{0, 1\}$, regress $\{y_{i,t}\}_{i \in \mathcal{R}_s^{(a)}}$ on $\{\widehat{U}_i\}_{i \in \mathcal{R}_s^{(a)}}$ to obtain $\widehat{V}_{1-s}^{(a)}$. For every $a \in \{0, 1\}$, $s \in \{0, 1\}$, and $i \in \mathcal{R}_s$, return $\widehat{\theta}_{i,t}^{(a)} = \widehat{U}_i \widehat{V}_s^{(a)^\top}$.

In summary, `Cross-Fitted-Regression` estimates the shared unit-dependent latent factors using the observed outcomes for all units until time period T_0. Then, for every $s \in \{0, 1\}$, the time-dependent latent factors $\widehat{V}_s$, $\widehat{V}_s^{(0)}$, and $\widehat{V}_s^{(1)}$ are estimated using the treatment assignments and the observed outcomes for units in $\mathcal{R}_{1-s}$.

To establish guarantees for `Cross-Fitted-Regression`, we adopt the subsequent assumption on the noise variables.

Assumption 4.18 (Independence across units and with respect to pre-adoption noise)**.**

(a) $\{(\eta_{i,t}, \varepsilon_{i,t}^{(a)}) : i \in [N]\}$ *are mutually independent (across* i*) given* $\{\varepsilon_{i,t}^{(0)}\}_{i \in [N], t \in [T_0]}$ *for every* $t \in [T] \setminus [T_0]$ *and* $a \in \{0, 1\}$*.*

(b) $\{\varepsilon_{i,t}^{(0)}\}_{i \in [N], t \in [T_0]} \perp\!\!\!\perp \{\eta_{i,t}, \varepsilon_{i,t}^{(a)}\}_{i \in [N]}$ *for every* $t \in [T] \setminus [T_0]$ *and* $a \in \{0, 1\}$*.*

Assumption 4.18(a) requires the noise $(E^{(a)}, W)$ corresponding to a time period $t \in T \setminus [T_0]$ to be jointly independent across units given the noise $E^{(0)}$ corresponding to time periods $[T_0]$, for every $a \in \{0, 1\}$. Assumption 4.18(b) is satisfied if, for instance, the noise variables follow a moving average model of order $t - T_0 - 1$. The following result, proven in Section 4.J.1, establishes that the estimates generated by `Cross-Fitted-Regression` satisfy Assumption 4.4. Deriving error bounds, i.e., $\mathcal{E}\big(\widehat{P}\big)$ and $\mathcal{E}\big(\widehat{\Theta}\big)$, for the estimates generated by `Cross-Fitted-Regression` for the staggered adoption model is an interesting direction for future research.

Proposition 4.5 (Guarantees for `Cross-Fitted-Regression`)**.** *Consider the staggered adoption model in Assumption 4.17 and suppose Assumption 4.18 holds. Fix any $t \in [T] \setminus [T_0]$, and $\{\widehat{\theta}^{(0)}_{i,t}, \widehat{\theta}^{(1)}_{i,t}, \widehat{p}_{i,t}\}_{i\in[N]}$ be the estimates returned by `Cross-Fitted-Regression`. Then, Assumption 4.4 holds.*

4.J.1 Proof of Proposition 4.5: Guarantees for `Cross-Fitted-Regression`

Fix any $s \in \{0,1\}$. Then, Assumption 4.18(a) and Assumption 4.18(b) imply that

$$\{\varepsilon^{(0)}_{i,t}\}_{i\in[N],t\in[T_0]} \cup \{\eta_{i,t}, \varepsilon^{(a)}_{i,t}\}_{i\in\overline{\mathcal{R}}_{1-s}} \perp\!\!\!\perp \{\eta_{i,t}, \varepsilon^{(a)}_{i,t}\}_{i\in\overline{\mathcal{R}}_{s}}, \tag{4.109}$$

for every partition $(\overline{\mathcal{R}}_0, \overline{\mathcal{R}}_1)$ of the units $[N]$.

`Cross-Fitted-Regression` estimates $\{\widehat{p}_{i,t}\}_{i\in\mathcal{R}_s}$ using $\{\widehat{U}_i\}_{i\in\mathcal{R}_s}$ and $\widehat{V}_s$, where $\widehat{V}_s$ is estimated using $\{\widehat{U}_i\}_{i\in\mathcal{R}_{1-s}}$ and $\{a_{i,t}\}_{i\in\mathcal{R}_{1-s}}$. Therefore, the randomness in $\{\widehat{p}_{i,t}\}_{i\in\mathcal{R}_s}$ stems from the randomness in $Y^{(0),\mathrm{pre}}$ and $\{a_{i,t}\}_{i\in\mathcal{R}_{1-s}}$ which in turn stems from the randomness in $\{\varepsilon^{(0)}_{i,t}\}_{i\in[N],t\in[T_0]}$ and $\{\eta_{i,t}\}_{i\in\mathcal{R}_{1-s}}$. Then, Eq. (4.15) follows from Eq. (4.109).

Next, fix any $a \in \{0,1\}$. Then, `Cross-Fitted-Regression` estimates $\{\widehat{\theta}^{(a)}\}_{i\in\mathcal{R}_s}$ using $\{\widehat{U}_i\}_{i\in\mathcal{R}_s}$ and $\widehat{V}^{(a)}_s$, where $\widehat{V}^{(a)}_s$ is estimated using $\{\widehat{U}_i\}_{i\in\mathcal{R}^{(a)}_{1-s}}$ and $\{y_{i,t}\}_{i\in\mathcal{R}^{(a)}_{1-s}}$. Therefore, the randomness in $\{\widehat{\theta}^{(a)}\}_{i\in\mathcal{R}_s}$ stems from the randomness in $Y^{(0),\mathrm{pre}}$ and $\{y_{i,t}\}_{i\in\mathcal{R}^{(a)}_{1-s}}$ which in turn stems from the randomness in $\{\varepsilon^{(0)}_{i,t}\}_{i\in[N],t\in[T_0]}$ and $\{\varepsilon^{(a)}_{i,t}\}_{i\in\mathcal{R}^{(a)}_{1-s}}$. Then, Eq. (4.14) follows from Eq. (4.109).

Bibliography

Abadie, A. (2003). Semiparametric instrumental variable estimation of treatment response models. *Journal of econometrics*, 113(2):231–263.

Abadie, A. (2005). Semiparametric difference-in-differences estimators. *The Review of Economic Studies*, 72(1):1–19.

Abadie, A., Agarwal, A., Dwivedi, R., and Shah, A. (2023). Doubly robust inference for causal latent factor models. *arXiv preprint arXiv:2402.11652*.

Abadie, A., Diamond, A., and Hainmueller, J. (2010a). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*, 105(490):493–505.

Abadie, A., Diamond, A., and Hainmueller, J. (2010b). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*, 105(490):493–505.

Abadie, A. and Gardeazabal, J. (2003a). The economic costs of conflict: A case study of the Basque country. *American economic review*, 93(1):113–132.

Abadie, A. and Gardeazabal, J. (2003b). The economic costs of conflict: A case study of the Basque country. *American economic review*, 93(1):113–132.

Abadie, A. and Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267.

Agarwal, A., Dahleh, M., Shah, D., and Shen, D. (2023). Causal matrix completion. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 3821–3826. PMLR.

Agarwal, A., Negahban, S., and Wainwright, M. J. (2010). Fast global convergence rates of gradient methods for high-dimensional statistical recovery. *Advances in Neural Information Processing Systems*, 23.

Agarwal, A., Shah, D., and Shen, D. (2020). Synthetic interventions. *arXiv preprint arXiv:2006.07691*.

Agarwal, A., Shah, D., Shen, D., and Song, D. (2021). On robustness of principal component regression. *Journal of the American Statistical Association*, pages 1–34.

Agarwal, A. and Singh, R. (2024). Causal inference with corrupted data: Measurement error, missing values, discretization, and differential privacy. *arXiv preprint arXiv:2107.02780*.

Aida, S. and Stroock, D. (1994). Moment estimates derived from Poincaré and logarithmic Sobolev inequalities. *Mathematical Research Letters*, 1(1):75–86.

Amemiya, T. (1985). *Advanced econometrics*. Harvard university press.

Angrist, J. D. (1998). Estimating the labor market impact of voluntary military service using social security data on military applicants. *Econometrica*, 66(2):249–288.

Angrist, J. D. and Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Arkhangelsky, D., Athey, S., Hirshberg, D. A., Imbens, G. W., and Wager, S. (2021). Synthetic difference-in-differences. *American Economic Review*, 111(12):4088–4118.

Arkhangelsky, D. and Imbens, G. (2018). The role of the propensity score in fixed effect models. Technical report, National Bureau of Economic Research.

Arkhangelsky, D. and Imbens, G. W. (2022). Doubly robust identification for causal panel data models. *The Econometrics Journal*, 25(3):649–674.

Athey, S., Bayati, M., Doudchenko, N., Imbens, G., and Khosravi, K. (2021). Matrix completion methods for causal panel data models. *Journal of the American Statistical Association*, 116(536):1716–1730.

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

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

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

Bai, J. (2024). Likelihood approach to dynamic panel models with interactive effects. *Journal of Econometrics*, 240(1):105636.

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

Bai, J. and Ng, S. (2021). Matrix completion, counterfactuals, and factor analysis of missing data. *Journal of the American Statistical Association*, 116(536):1746–1763.

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–972.

Barndorff-Nielsen, O. (2014). *Information and exponential families: in statistical theory*. John Wiley & Sons.

Barp, A., Briol, F.-X., Duncan, A., Girolami, M., and Mackey, L. (2019). Minimum Stein discrepancy estimators. *Advances in Neural Information Processing Systems*, 32.

Bertrand, M., Duflo, E., and Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *The Quarterly journal of economics*, 119(1):249–275.

Besag, J. (1975a). Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 24(3):179–195.

Besag, J. (1975b). Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 24(3):179–195.

Bhatia, R. (2007). *Perturbation bounds for matrix eigenvalues*. SIAM.

Bhattacharya, B. B. and Mukherjee, S. (2018). Inference in Ising models. *Bernoulli*, 24(1):493–525.

Bhattacharya, S. and Chatterjee, S. (2022). Matrix completion with data-dependent missingness probabilities. *IEEE Transactions on Information Theory*, 68(10):6762–6773.

Billingsley, P. (2017). *Probability and measure*. John Wiley & Sons.

Bresler, G. (2015). Efficiently learning Ising models on arbitrary graphs. In *Proceedings of the 47th Annual ACM Symposium on Theory of Computing*, pages 771–782.

Bresler, G. and Buhai, R.-D. (2020). Learning restricted Boltzmann machines with sparse latent variables. *Advances in Neural Information Processing Systems*, 33:7020–7030.

Bresler, G., Koehler, F., and Moitra, A. (2019). Learning restricted Boltzmann machines via influence maximization. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 828–839.

Brown, L. D. (1986). Fundamentals of statistical exponential families: with applications in statistical decision theory. Ims.

Busa-Fekete, R., Fotakis, D., Szörényi, B., and Zampetakis, M. (2019). Optimal learning of Mallows block model. In *Conference on Learning Theory*, pages 529–532. PMLR.

Cai, T., Liu, W., and Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607.

Caruana, R. (1997). Multitask learning. *Machine learning*, 28(1):41–75.

Cattaneo, M. D., Feng, Y., and Titiunik, R. (2021). Prediction intervals for synthetic control methods. *Journal of the American Statistical Association*, 116(536):1865–1880.

Chandrasekaran, V., Parrilo, P. A., and Willsky, A. S. (2012). Latent variable graphical model selection via convex optimization. *The Annals of Statistics*, 40(4):1935–1967.

Chatterjee, S. (2007). Estimation in spin glasses: A first step. *The Annals of Statistics*, 35(5):1931–1946.

Chatterjee, S. (2015). Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 43(1):177 – 214.

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

Chwialkowski, K., Strathmann, H., and Gretton, A. (2016). A kernel test of goodness of fit. In *International conference on machine learning*, pages 2606–2615. PMLR.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, pages 295–313.

Dagan, Y., Daskalakis, C., Dikkala, N., and Kandiros, A. V. (2021). Learning Ising models from one or multiple samples. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 161–168.

Dai, B., Liu, Z., Dai, H., He, N., Gretton, A., Song, L., and Schuurmans, D. (2019). Exponential family estimation via adversarial dynamics embedding. *Advances in Neural Information Processing Systems*, 32.

Darmois, G. (1935). Sur les lois de probabilitéa estimation exhaustive. *CR Acad. Sci. Paris*, 260(1265):85.

Daskalakis, C., Dikkala, N., and Panageas, I. (2019). Regression from dependent observations. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 881–889.

Daskalakis, C., Gouleakis, T., Tzamos, C., and Zampetakis, M. (2018). Efficient statistics, in high dimensions, from truncated samples. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 639–649. IEEE.

Diakonikolas, I., Kane, D. M., Stewart, A., and Sun, Y. (2021). Outlier-robust learning of Ising models under Dobrushin's condition. In *Conference on Learning Theory*, pages 1645–1682. PMLR.

Dwivedi, R., Tian, K., Tomkins, S., Klasnja, P., Murphy, S., and Shah, D. (2022a). Counterfactual inference for sequential experiments. *arXiv preprint arXiv:2202.06891*.

Dwivedi, R., Tian, K., Tomkins, S., Klasnja, P., Murphy, S., and Shah, D. (2022b). Doubly robust nearest neighbors in factor models. *arXiv preprint arXiv:2211.14297*.

Ferguson, T. S. (2017). *A course in large sample theory*. Routledge.

Fernández-Val, I. and Weidner, M. (2018). Fixed effects estimation of large-T panel data models. *Annual Review of Economics*, 10(1):109–138.

Fisher, R. (1931). Properties and applications of Hh functions. *Mathematical tables*, 1:815–852.

Fisher, R. A. (1934). Two new properties of mathematical likelihood. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 144(852):285–307.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Galton, F. (1898). An examination into the registered speeds of American trotting horses, with remarks on their value as hereditary data. *Proceedings of the Royal Society of London*, 62(379-387):310–315.

Ghang, W., Martin, Z., and Waruhiu, S. (2014). The sharp log-Sobolev inequality on a compact interval. *Involve*, 7:181–186.

Ghosal, P. and Mukherjee, S. (2020). Joint estimation of parameters in Ising model. *The Annals of Statistics*, 48(2):785–810.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378.

Goel, S. (2020). Learning Ising and Potts models with latent variables. In *International Conference on Artificial Intelligence and Statistics*, pages 3557–3566. PMLR.

Goodd, I. and Gaskins, R. A. (1971). Nonparametric roughness penalties for probability densities. *Biometrika*, 58(2):255–277.

Gorham, J. and Mackey, L. (2015). Measuring sample quality with Stein's method. *Advances in neural information processing systems*, 28.

Gorham, J. and Mackey, L. (2017). Measuring sample quality with kernels. In *International Conference on Machine Learning*, pages 1292–1301. PMLR.

Grünwald, P. D. and Dawid, A. P. (2004). Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. *the Annals of Statistics*, 32(4):1367–1433.

Gu, C. (1993). Smoothing spline density estimation: A dimensionless automatic algorithm. *Journal of the American Statistical Association*, 88(422):495–504.

Gu, C., Jeon, Y., and Lin, Y. (2013). Nonparametric density estimation in high-dimensions. *Statistica Sinica*, pages 1131–1153.

Gu, C. and Qiu, C. (1993). Smoothing spline density estimation: Theory. *The Annals of Statistics*, pages 217–234.

Gu, C. and Wang, J. (2003). Penalized likelihood density estimation: Direct cross-validation and scalable approximation. *Statistica Sinica*, pages 811–826.

Gutmann, M. and Hyvärinen, A. (2010). Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings.

Gutmann, M. U. and Hyvärinen, A. (2012). Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of machine learning research*, 13(2).

Hartford, J., Lewis, G., Leyton-Brown, K., and Taddy, M. (2017). Deep IV: A flexible approach for counterfactual prediction. In *International Conference on Machine Learning*, pages 1414–1423. PMLR.

Hernán, M. and Robins, J. (2020). *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC.

Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800.

Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189.

Hogg, R. V. and Craig, A. T. (1956). Sufficient statistics in elementary distribution theory. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, 17(3):209–216.

Holley, R. and Stroock, D. (1987). Logarithmic Sobolev inequalities and stochastic Ising models. *Journal of Statistical Physics*, 46:1159–1194.

Holmquist, B. (1988). Moments and cumulants of the multivariate normal distribution. *Stochastic Analysis and Applications*, 6(3):273–278.

Hyvärinen, A. (2007). Some extensions of score matching. *Computational statistics & data analysis*, 51(5):2499–2512.

Hyvärinen, A. and Dayan, P. (2005). Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4).

Imbens, G. W. and Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica: journal of the Econometric Society*, pages 467–475.

Imbens, G. W. and Rubin, D. B. (2015a). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge University Press.

Imbens, G. W. and Rubin, D. B. (2015b). *Causal inference in statistics, social, and biomedical sciences.* Cambridge University Press.

Jalali, A., Ravikumar, P., Vasuki, V., and Sanghavi, S. (2011). On learning discrete graphical models using group-sparse regularization. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2011, Fort Lauderdale, USA, April 11-13, 2011*, pages 378–387.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical review*, 106(4):620.

Jennrich, R. I. (1969). Asymptotic properties of non-linear least squares estimators. *Ann. Math. Statist.*, 40(2):633–643.

Jeon, Y. and Lin, Y. (2006). An effective method for high-dimensional log-density ANOVA estimation, with application to nonparametric graphical model building. *Statistica Sinica*, pages 353–374.

Jerrum, M. and Sinclair, A. (1989). Approximating the permanent. *SIAM journal on computing*, 18(6):1149–1178.

Jesson, A., Mindermann, S., Gal, Y., and Shalit, U. (2021). Quantifying ignorance in individual-level causal-effect estimates under hidden confounding. In *International Conference on Machine Learning*, pages 4829–4838. PMLR.

Jin, S., Miao, K., and Su, L. (2021). On factor models with random missing: EM estimation, inference, and cross validation. *Journal of Econometrics*, 222(1):745–777.

Jin, Y., Ren, Z., and Candès, E. J. (2023). Sensitivity analysis of individual treatment effects: A robust conformal inference approach. *Proceedings of the National Academy of Sciences*, 120(6):e2214889120.

Kallus, N., Mao, X., and Zhou, A. (2019). Interval estimation of individual-level causal effects under unobserved confounding. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2281–2290. PMLR.

Kandiros, V., Dagan, Y., Dikkala, N., Goel, S., and Daskalakis, C. (2021). Statistical estimation from dependent data. In *International Conference on Machine Learning*, pages 5269–5278. PMLR.

Kelner, J., Koehler, F., Meka, R., and Moitra, A. (2020). Learning some popular Gaussian graphical models without condition number bounds. *Advances in Neural Information Processing Systems*, 33:10986–10998.

Kennedy, E. H. (2016). Semiparametric theory and empirical processes in causal inference. *Statistical causal inferences and their applications in public health research*, pages 141–167.

Klivans, A. and Meka, R. (2017). Learning graphical models using multiplicative weights. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 343–354. IEEE.

Koopman, B. O. (1936). On distributions admitting a sufficient statistic. *Transactions of the American Mathematical society*, 39(3):399–409.

Ledoux, M. (2001). Logarithmic Sobolev inequalities for unbounded spin systems revisited. In *Séminaire de Probabilités XXXV*, pages 167–194. Springer.

Lee, A. (1914). Table of the Gaussian 'tail' functions; when the 'tail' is larger than the body. *Biometrika*, 10(2/3):208–214.

Leonard, T. (1978). Density estimation, stochastic processes and prior information. *Journal of the Royal Statistical Society: Series B (Methodological)*, 40(2):113–132.

Li, Y., Shah, D., Song, D., and Yu, C. L. (2019). Nearest neighbors for matrix estimation interpreted as blind regression for latent variable model. *IEEE Transactions on Information Theory*, 66(3):1760–1784.

Lin, L., Drton, M., and Shojaie, A. (2016). Estimation of high-dimensional graphical models using regularized score matching. *Electronic journal of statistics*, 10(1):806.

Liu, Q., Lee, J., and Jordan, M. (2016). A kernelized Stein discrepancy for goodness-of-fit tests. In *International conference on machine learning*, pages 276–284. PMLR.

Liu, S., Kanamori, T., Jitkrittum, W., and Chen, Y. (2019). Fisher efficient inference of intractable models. *Advances in Neural Information Processing Systems*, 32:8793–8803.

Liu, S., Kanamori, T., and Williams, D. J. (2022). Estimating density models with truncation boundaries using score matching. *The Journal of Machine Learning Research*, 23(1):8448–8485.

Ma, J. and Michailidis, G. (2016). Joint structural estimation of multiple graphical models. *The Journal of Machine Learning Research*, 17(1):5777–5824.

Ma, S., Xue, L., and Zou, H. (2013). Alternating direction methods for latent variable Gaussian graphical model selection. *Neural computation*, 25(8):2172–2198.

Marton, K. (2015). Logarithmic Sobolev inequalities in discrete product spaces: a proof by a transportation cost distance. *arXiv preprint arXiv:1507.02803*.

Meinshausen, N., Bühlmann, P., et al. (2006). High-dimensional graphs and variable selection with the lasso. *Annals of statistics*, 34(3):1436–1462.

Misra, S., Vuffray, M., and Lokhov, A. Y. (2020). Information theoretic optimal learning of Gaussian graphical models. In *Conference on Learning Theory*, pages 2888–2909. PMLR.

Mukherjee, S., Halder, S., Bhattacharya, B. B., and Michailidis, G. (2021). High dimensional logistic regression under network dependence. *arXiv preprint arXiv:2110.03200*.

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

Neyman, J. (1923). Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10(1):1–51.

Nguyen, L. T., Kim, J., and Shim, B. (2019). Low-rank matrix completion: A contemporary survey. *IEEE Access*, 7:94215–94237.

Ning, Y., Zhao, T., Liu, H., et al. (2017). A likelihood ratio framework for high-dimensional semiparametric regression. *Annals of Statistics*, 45(6):2299–2327.

O'Sullivan, F. (1988). Fast computation of fully automated log-density and log-hazard estimators. *SIAM Journal on scientific and statistical computing*, 9(2):363–379.

Painsky, A. and Wornell, G. W. (2019). Bregman divergence bounds and universality properties of the logarithmic loss. *IEEE Transactions on Information Theory*, 66(3):1658–1673.

Pearl, J. (2009). *Causality*. Cambridge university press.

Pearl, J., Glymour, M., and Jewell, N. P. (2016). *Causal inference in statistics: A primer*. John Wiley & Sons.

Pearson, K. (1902). On the systematic fitting of frequency curves. *Biometrika*, 2:2–7.

Pearson, K. and Lee, A. (1908). On the generalised probable error in multiple normal correlation. *Biometrika*, 6(1):59–68.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12:2825–2830.

Pitman, E. J. G. (1936). Sufficient statistics and intrinsic accuracy. In *Mathematical Proceedings of the cambridge Philosophical society*, volume 32, pages 567–579. Cambridge University Press.

Ren, C. X., Misra, S., Vuffray, M., and Lokhov, A. Y. (2021). Learning continuous exponential families beyond Gaussian. *arXiv preprint arXiv:2102.09198*.

Rhodes, B., Xu, K., and Gutmann, M. U. (2020). Telescoping density-ratio estimation. *Advances in neural information processing systems*, 33:4905–4916.

Robert, C. and Casella, G. (2013). *Monte Carlo statistical methods*. Springer Science & Business Media.

Robins, J. M., Hernan, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866.

Rosenbaum, P. R. and Rubin, D. B. (1983a). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society: Series B (Methodological)*, 45(2):212–218.

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

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

Rubin, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593.

Ryu, J. J., Shah, A., and Wornell, G. W. (2024). A unified view on learning unnormalized distributions via noise-contrastive estimation. *arXiv preprint arXiv:2409.18209*.

Santhanam, N. P. and Wainwright, M. J. (2012). Information-theoretic limits of selecting binary graphical models in high dimensions. *IEEE Trans. Information Theory*, 58(7):4117–4134.

Semenova, V. and Chernozhukov, V. (2021). Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal*, 24(2):264–289.

Shah, A., Dwivedi, R., Shah, D., and Wornell, G. W. (2022). On counterfactual inference with unobserved confounding. *arXiv preprint arXiv:2211.08209*.

Shah, A., Shah, D., and Wornell, G. (2021a). A computationally efficient method for learning exponential family distributions. *Advances in Neural Information Processing Systems*.

Shah, A., Shah, D., and Wornell, G. (2021b). A computationally efficient method for learning exponential family distributions. *Advances in Neural Information Processing Systems*, 34:15841–15854.

Shah, A., Shah, D., and Wornell, G. (2021c). On learning continuous pairwise markov random fields. In *International Conference on Artificial Intelligence and Statistics*, pages 1153–1161. PMLR.

Shah, A., Shah, D., and Wornell, G. (2021d). On learning continuous pairwise Markov random fields. In *International Conference on Artificial Intelligence and Statistics*, pages 1153–1161. PMLR.

Shah, A., Shah, D., and Wornell, G. (2023). On computationally efficient learning of exponential family distributions. *arXiv preprint arXiv:2309.06413*.

Shah, A., Shah, D., and Wornell, G. W. (2024). On computationally efficient learning of exponential family distributions. *IEEE Transactions on Information Theory*.

Sherrington, D. and Kirkpatrick, S. (1975). Solvable model of a spin-glass. *Physical review letters*, 35(26):1792.

Silverman, B. W. (1982). On the estimation of a probability density function by the maximum penalized likelihood method. *The Annals of Statistics*, pages 795–810.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A sparse-group lasso. *Journal of computational and graphical statistics*, 22(2):231–245.

Singh, R., Sahani, M., and Gretton, A. (2019). Kernel instrumental variable regression. *Advances in Neural Information Processing Systems*, 32:4593—-4605.

Sloczynski, T., Uysal, S. D., and Wooldridge, J. M. (2024). Abadie's kappa and weighting estimators of the local average treatment effect. *Journal of Business & Economic Statistics*. Forthcoming.

Sly, A. and Sun, N. (2012). The computational hardness of counting in two-spin models on d-regular graphs. In *2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*, pages 361–369. IEEE.

Suggala, A. S., Kolar, M., and Ravikumar, P. (2017). The expxorcist: Nonparametric graphical models via conditional exponential densities. In *Advances in Neural Information Processing Systems*, pages 4446–4456.

Sun, S., Kolar, M., and Xu, J. (2015). Learning structured densities via infinite dimensional exponential families. In *Advances in Neural Information Processing Systems*, pages 2287–2295.

Syrgkanis, V., Lei, V., Oprescu, M., Hei, M., Battocchi, K., and Lewis, G. (2019). Machine learning estimation of heterogeneous treatment effects with instruments. *Advances in Neural Information Processing Systems*, 32.

Taeb, A., Shah, P., and Chandrasekaran, V. (2020). Learning exponential family graphical models with latent variables using regularized conditional likelihood. *arXiv preprint arXiv:2010.09386*.

Tan, V. Y., Anandkumar, A., and Willsky, A. S. (2010). Learning Gaussian tree models: Analysis of error exponents and extremal structures. *IEEE Transactions on Signal Processing*, 58(5):2701–2714.

Tansey, W., Padilla, O. H. M., Suggala, A. S., and Ravikumar, P. (2015). Vector-space Markov random fields via exponential families. In *International Conference on Machine Learning*, pages 684–692.

Trench, W. F. (1999). Asymptotic distribution of the spectra of a class of generalized Kac–Murdock–Szegö matrices. *Linear algebra and its applications*, 294(1-3):181–192.

Valiant, L. G. (1979). The complexity of enumeration and reliability problems. *SIAM Journal on Computing*, 8(3):410–421.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.

Villani, C. (2009). *Optimal transport: old and new*, volume 338. Springer.

Vinyes, M. and Obozinski, G. (2018). Learning the effect of latent variables in Gaussian graphical models with unobserved variables. *arXiv preprint arXiv:1807.07754*.

Vuffray, M., Misra, S., Lokhov, A., and Chertkov, M. (2016a). Interaction screening: Efficient and sample-optimal learning of Ising models. *Advances in Neural Information Processing Systems*, 29.

Vuffray, M., Misra, S., and Lokhov, A. Y. (2022a). Efficient learning of discrete graphical models. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124017.

Vuffray, M., Misra, S., and Lokhov, A. Y. (2022b). Efficient learning of discrete graphical models. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124017.

Vuffray, M., Misra, S., Lokhov, A. Y., and Chertkov, M. (2016b). Interaction screening: Efficient and sample-optimal learning of Ising models. In *Advances in Neural Information Processing Systems*, pages 2595–2603.

Wainwright, M. J., Jaakkola, T. S., and Willsky, A. S. (2003). Tree-reweighted belief propagation algorithms and approximate ML estimation by pseudo-moment matching. In *AISTATS*, volume 3, page 3.

Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305.

Wainwright, M. J., Jordan, M. I., et al. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305.

Wainwright, M. J., Ravikumar, P., and Lafferty, J. D. (2006). High-dimensional graphical model selection using ℓ_1-regularized logistic regression. In *Advances in Neural Information Processing Systems*, pages 1465–1472.

Wang, G., Li, J., and Hopp, W. J. (2022). An instrumental variable forest approach for detecting heterogeneous treatment effects in observational studies. *Management Science*, 68(5):3399–3418.

Wang, K., Franks, A., and Oh, S.-Y. (2023). Learning Gaussian graphical models with latent confounders. *Journal of Multivariate Analysis*, 198:105213.

Wang, W., Wainwright, M. J., and Ramchandran, K. (2010). Information-theoretic bounds on model selection for Gaussian Markov random fields. In *2010 IEEE International Symposium on Information Theory*, pages 1373–1377. IEEE.

Wilhelm, S. and Manjunath, B. (2010a). tmvtnorm: A package for the truncated multivariate normal distribution. *sigma*, 2(2):1–25.

Wilhelm, S. and Manjunath, B. (2010b). tmvtnorm: A package for the truncated multivariate normal distribution. *SIGMA*, 2(2):1–25.

Won, J. H. and Kim, S.-J. (2006). Maximum likelihood covariance estimation with a condition number constraint. In *2006 Fortieth Asilomar Conference on Signals, Systems and Computers*, pages 1445–1449. IEEE.

Wooldridge, J. M. (2007). Inverse probability weighted estimation for general missing data problems. *Journal of econometrics*, 141(2):1281–1301.

Xiong, R. and Pelger, M. (2023). Large dimensional latent factor modeling with missing observations and applications to causal inference. *Journal of Econometrics*, 233(1):271–301.

Xu, L., Chen, Y., Srinivasan, S., de Freitas, N., Doucet, A., and Gretton, A. (2020). Learning deep features in instrumental variable regression. *arXiv preprint arXiv:2010.07154*.

Yadlowsky, S., Namkoong, H., Basu, S., Duchi, J., and Tian, L. (2022). Bounds on the conditional and average treatment effect with unobserved confounding factors. *The Annals of Statistics*, 50(5):2587–2615.

Yang, E., Ravikumar, P., Allen, G. I., and Liu, Z. (2015). Graphical models via univariate exponential family distributions. *J. Mach. Learn. Res.*, 16:3813–3847.

Yang, Z., Ning, Y., and Liu, H. (2018). On semiparametric exponential family graphical models. *The Journal of Machine Learning Research*, 19(1):2314–2372.

Yin, M., Shi, C., Wang, Y., and Blei, D. M. (2022). Conformal sensitivity analysis for individual treatment effects. *Journal of the American Statistical Association*, pages 1–14.

Yuan, X., Li, P., Zhang, T., Liu, Q., and Liu, G. (2016). Learning additive exponential family graphical models via $\ell_{2,1}$-norm regularized M-estimation. In *Advances in Neural Information Processing Systems*, pages 4367–4375.

Zhang, H. and Wei, H. (2022). Sharper sub-Weibull concentrations. *Mathematics*, 10(13):2252.

Zhou, S., Rütimann, P., Xu, M., and Bühlmann, P. (2011). High-dimensional covariance estimation based on Gaussian graphical models. *The Journal of Machine Learning Research*, 12:2975–3026.