Statistical Learning for Large-Scale and Complex-Structured Data

by

Weijing Tang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in The University of Michigan
2022

Doctoral Committee:

Professor Ji Zhu, Chair
Professor Elizaveta Levina
Professor Qiaozhu Mei
Associate Professor Gongjun Xu

Weijing Tang

weijtang@umich.edu

ORCID iD: 0000-0002-3543-9924

To my grandparents and parents

ACKNOWLEDGEMENTS

I would like to first express my deepest gratitude to my advisor Ji Zhu for all his continuous and endless support during my graduate study. Taking me on as his student was one of the most fortunate things that could have ever happened to me. I am deeply indebted to Ji for his generous availability and patience to discuss research and advise me on all matters of life; for his encouragement and support to help me reach beyond what I thought are my limits and to build confidence and independence; for his inspiring comments and sharp questions in our meetings as well as connecting me with researchers in relevant disciplines, which greatly influenced the way I approach research questions; and for sharing valuable academic resources and information. At the same time, Ji always gave me space to make decisions on my own. Without Ji's support and guidance throughout my Ph.D. process, this dissertation would not have been conceivable.

I am also sincerely grateful to my other committee members Elizaveta Levina, Gongjun Xu, and Qiaozhu Mei for their constructive and insightful feedbacks on my work and for their generous encouragement and support to me during the past few years. Liza brought me to Michigan and, over our many interactions, I learned a lot from her deep insights and taste of different research problems during our joint group meeting, her support to women scientists in the broad academic community, and her experiences in both personal and academic perspectives. I got to know Gongjun in my first year when I was a student in STATS 601 and later started to collaborate with him on the projects for censored data; Gongjun was always flexible enough to

discuss and provide many helpful advice in our collaboration; his strong support and encouragement during my graduate study and my job search were greatly appreciated. I would also like to thank Qiaozhu for our discussion on deep learning early on in my graduate study, which inspired and led me to a fascinating area of research, and for his generous support and help for my job search.

I was extremely fortunate to be a member of such a collegial family in the Department of Statistics at Michigan and also to interact with many faculty across the campus. I would like to heartily thank Xuming He for his valuable time and efforts in helping me over my graduate study and for his constructive and helpful comments on my job application materials and job talk that greatly improved them. I sincerely thank Akbar Waljee and Brahmajee Nallamothu for insightful conversations that influenced my perception of what my research could be able to contribute in healthcare and for their strong support on my job search; I also benefited a lot from their helpful suggestions on writing and presentation. I also owe a debt of gratitude to Kevin He for offering domain knowledge, helpful advice, and valuable data in collaboration as well as his patience and willingness over many rounds of discussion. I would like to thank Kerby Shedden, Yuekai Sun, Kean Ming Tan, and Naisyin Wang for attending my practice job talk and for providing valuable feedbacks. I also want to show my great appreciation to Yinqiu He, Yuqi Gu, and Xu Shi who helped rehearsal my job talk, generously provided detailed comments page by page, shared their valuable experiences and encouraged me. I also thank Judy Dyer for her time and energy on improving my English communication skills and for her support during my job search.

Sincere and genuine gratitude is also due to all my wonderful friends I met here at Michigan. My friends have contributed far more to my academic achievement and my enjoyable and unforgettable Ph.D. journey than I could have imagined; especially for the past two years throughout this pandemic, your companion has been a huge source of delight and happiness.

Last but not least, I want to express my heartfelt gratitude and love to my grandparents and parents for their unconditional love and support, without which I would not and could not have gotten this far. Finally, I wish to extend my special thanks to Jiaqi Ma for always being on my side and sharing my tears and joys every moment of my way.

TABLE OF CONTENTS

CHAPTER

LIST OF TABLES

Table

LIST OF FIGURES

Figure

LIST OF APPENDICES

Appendix

ABSTRACT

Our modern era has seen an explosion in the amount of valuable information stored in large and complex datasets. The growing scale, diversity of data structures, and incomplete observations in these datasets pose new challenges for statistical learning. Motivated by these challenges, this dissertation addresses three important problems below.

(I) The first part of the dissertation presents how ordinary differential equations (ODE) can be novelly used to enhance modeling flexibility and computational efficiency in survival analysis for complex and incomplete *censored data.* Despite rich literature on survival analysis, most existing statistical models and estimation methods still suffer from practical limitations such as restricted model capacity and a lack of scalability for large-scale studies. We introduce a unified ODE framework for survival analysis that allows flexible modeling and enables a statistically efficient procedure for estimation and inference. In particular, the proposed estimation procedure is computationally efficient, easy-to-implement, and applicable to a wide range of survival models. Moreover, to accommodate data in diverse formats, we extend the ODE framework by leveraging deep neural networks for powerful prediction.

(II) The second part of the dissertation focuses on statistical models for *signed networks.* Statistical network models are useful for understanding the underlying formation mechanism and characteristics of complex networks. However, statistical models for *signed networks* have been largely unexplored. In signed networks, there exist both positive (e.g., like, trust) and negative (e.g., dislike, distrust) edges, which are commonly seen in real-world scenarios. The positive and negative edges in signed networks lead to unique structural patterns, which pose challenges for statistical modeling. In this part, we introduce a novel latent space approach for modeling signed networks and accommodating the well-known *balance theory* in social science, i.e., “the enemy of my enemy is my friend” and “the friend of my friend is my friend”. The proposed approach treats both edges and their signs as random variables, and characterizes the balance theory with a novel and natural notion of population-level balance. This approach guides us towards building a class of balanced inner-product models, and towards developing scalable algorithms via projected gradient descent to estimate the latent variables. We also establish non-asymptotic error rates for the estimates.

(III) The third part of the dissertation focuses on applications of statistical machine learning to healthcare. In particular, quick and accurate prediction of disease progression can provide valuable information for clinicians to provide appropriate care in a timely manner. The success of prediction models often relies on the availability of a large number of labeled training data. However, in many healthcare settings, only a small minority of available data is accurately labeled while unlabeled data is abundant. Further, input variables such as clinical events in the medical records are usually of a complex, longitudinal nature, which poses additional challenges. Motivated by the scarcity of annotated data, we propose a new semi-supervised joint learning method for classifying clinical events data, which requires fewer labeled training data while maintaining the same prediction performance when compared to the supervised method.

CHAPTER I

Introduction

The modern era has seen an explosion in the amount of valuable information stored in emerging datasets. For example, in healthcare, with the growing adoption of electronic health record (EHR) systems, clinical data, including laboratory measurements and results, diagnostic and procedure codes, and medication orders have become widely available (Goldstein et al., 2017). The wealth of EHR data is a clear benefit in that it has enormous potential for biomedical and healthcare research such as studying drug safety and effectiveness (Lin and Schneeweiss, 2016), clinical knowledge extraction (Hong et al., 2021), and risk predictive modeling (Bennett et al., 2012). More examples in other disciplines include the wide usage of social media, which provides rich relational information among users that can be used for recommender systems (Sun et al., 2015); the availability of protein-protein interactions could be analyzed to help assign protein function (Vazquez et al., 2003). Despite the valuable information, the complexity of modern data also poses many new challenges for statistical learning in scientific research and engineering applications, of which three important such challenges are

1. *Growing scale of datasets*: Recent advances in data collection and storage provide massive amounts of data for research. For instance, the organ procurement and transplantation network is a national transplant information database that

has collected and reported data on every organ donation, transplant event, and waiting list candidate. Analyzing such medical records of millions of patients with follow-up of over more than thirty years requires computationally efficient statistical methods (He et al., 2017).

2. *Diversity of data structures*: In addition to the large scale, present-day data can take diverse forms. As one example, EHR data in the medical information mart for intensive care (MIMIC) database contains not only patients' demographic data, but also radiology images, clinical notes in free texts, and irregularly time-stamped laboratory tests and physiological measurements (Johnson et al., 2016). Moreover, connectivity and interaction relationships among individuals of a complex system are often represented as *networks* such as social networks, biological networks, and traffic networks (Newman, 2010). Given such diversity of data structures, many traditional statistical methods originally developed for standard vector data have become less suitable for complex data analysis due to the lack of modeling capacity and flexibility.

3. *Incomplete observations*: In real-world data collections, the outcome of interest may not be observed for various reasons. For example, when analyzing the survival time until death, we frequently observe *censored data* since patients may opt out of the treatment during the study, in which case, instead of observing the survival time, we record the time to the last follow-up (Miller Jr., 2011). There are also cases in which the outcomes are completely unknown for a majority of samples due to technical reasons or high cost of annotation, such as accurate diagnostics of disease progression. Statistical methods that appropriately leverage these complex and incomplete data for research are desirable.

The above challenges arising from statistical learning for large-scale and complex-structured data lead to the following foremost research questions: methodologically, how to enhance the modeling flexibility given the diverse types of predictor variables within data points and the relationships among data points; how to make appropriate statistical inference and accurate predictions given incomplete observations that are often encountered in various applications; and computationally, how to improve the computational efficiency given the tremendous growth of available data points. In particular, this dissertation addresses these questions in the following contexts.

(I) Unified ODE Framework for Large-Scale Survival Analysis. Survival analysis is an important branch of statistical learning where the outcome of interest is the time until occurrence of an event. In practice, some events may not be observed due to a limited observation time window or missing follow-up, which is known as *censoring*. In this case, instead of observing an *event time*, we record a *censored time*, for example, the end of the observation window, to indicate that no event has occurred prior to it. Survival analysis methods take into account the partial information contained in the censored data and have crucial applications in various real-world problems, such as modeling solar activity in astronomy, disease occurrence in health, reliability of devices in engineering, and customer lifespan in marketing (Chen et al., 2009; Miller Jr., 2011; Modarres et al., 2016). Despite the rich literature on survival analysis, the growing scale and complexity of present-day data present new demands that existing statistical models and estimation methods cannot entirely satisfy. Most of these methods suffer from practical limitations including a lack of scalability for large-scale data, statistically inefficient estimators, and complicated numerical implementations.

To address these issues, in Chapter II, we introduce a novel ordinary differential equation (ODE) approach to survival analysis. In particular, we recognize that maximum likelihood estimation (MLE) with censored data can be viewed as an ODE constrained optimization problem. This key observation has inspired us to take a new perspective on modeling the time-to-event data – viewing the hazard as the dynamic change of the cumulative hazard, and quantifying them together using an ODE. Following this connection, we further propose a scalable procedure for estimation and inference founded upon well-established numerical solvers and sensitivity analysis tools for ODEs. Remarkably, this ODE approach has led to important gains in modeling, computation, and statistical efficiency. 1) The proposed ODE modeling framework is sufficiently general to unify many existing survival models, including the proportional hazards model, the linear transformation model, the accelerated failure time model, and the time-varying coefficient model. The generality of the proposed framework serves as the foundation of a widely applicable estimation procedure. 2) The proposed estimation procedure overcomes various practical limitations of existing estimation methods, such as scalability against maximum partial likelihood estimation for the time-varying Cox model, and statistical efficiency and numerical stability against rank-based methods for the linear transformation model, which have been demonstrated through extensive simulation studies. 3) We establish a new general sieve M-theorem to establish the asymptotic distributional theory for complicated bundled parameters induced by the ODE notion, and show the semi-parametric efficiency of the proposed regression estimate. The new M-theorem can also be useful for developing the asymptotic distribution of sieve estimators for other models.

To further accommodate data in diverse formats, such as longitudinal laboratory test and radiology images, a promising idea is to leverage deep neural networks into survival analysis due to their capability of automatic feature extraction from large-scale raw data. While scalable learning of neural networks in classification or regression tasks is straightforward, the existence of censoring in survival analysis leads to difficult-to-evaluate integrals in MLE, which imposes an intrinsic optimization challenge for training neural networks. To avoid computing this integral, most existing methods either made additional structural assumptions (Katzman et al., 2018) or discretized the continuous event time with the cost of potential information loss (Lee et al., 2018). However, the proposed ODE framework can be naturally extended with deep neural networks and address the optimization challenge by using an efficient numerical approach. Our approach shows strong prediction performance on large-scale real-world data examples over existing state-of-the-art methods.

The materials in this chapter are mainly adapted from our research papers (Tang et al., 2022a,b). Tang et al. (2022a) is a joint work with Kevin He, Gongjun Xu, and Ji Zhu; Tang et al. (2022a) is a joint work with Jiaqi Ma, Qiaozhu Mei, and Ji Zhu.

(II) Latent Space Approach for Signed Networks. Networks represent connectivity relationships between individuals in a complex system and are ubiquitous in diverse engineering and scientific fields, such as biology, computer science, economics and sociology. In a network, a node represents an individual and an edge between two nodes indicates the presence of a certain relation. Real-world networks are heterogeneous in their variety of edge and node attributes. For example, in *signed networks*, there exist both positive and negative edges, and such signed relationships are common. Examples include like and dislike relationships in social networks among users; and cooperation and competition in international relation networks among countries. While research on networks has steadily increased in recent years to understand the underlying formation mechanism and characteristics of complex networks, most statistical network models focus on binary networks and are thereby inadequate for modeling signed networks. Modeling signed networks is challenging not merely due to the additional sign for each edge, but because the presence of positive and negative edges depends on the other. In particular, the *balance theory* (Harary et al., 1953) in social science postulates that balanced triangles (three nodes connected to each other with a positive product of edge signs; see Figure III.1 in Chapter III for examples) should be more prevalent than unbalanced triangles in signed networks. This theory directly coincides with the proverb, “the enemy of my enemy is my friend” and “the friend of my friend is my friend”, for which recent studies have found empirical evidence (Leskovec et al., 2010; Feng et al., 2020).

In Chapter III, we introduce a novel statistical framework to provide generative models for signed networks while accommodating the balance theory. Specifically, the proposed framework statistically formalizes the balance theory by introducing the novel notion of *population-level balance*. This framework enables the investigation of the balance property for a general class of latent space models, where the presence of an edge and its sign depend on nodes’ positions in a low-dimensional latent space. Latent space models are particularly attractive due to their interpretable structure, their nature for network visualization, and their ability for downstream machine learning tasks, such as node clustering and classification, and network link prediction. For this general class of latent space models, we derive practical conditions for a model with inherent population-level balance, which has further guided us to propose a class of balanced inner-product models. We have also developed scalable estimation methods based on projected gradient descent algorithms and established their non-asymptotic error rates. In addition, we apply the proposed approach to an international relation network, which provides an informative and interpretable model-based visualization of countries during World War II. The materials in this chapter are adapted from joint work with Ji Zhu (Tang and Zhu, 2022+).

(III) Semi-Supervised Learning for Longitudinal Clinical Events Data. Classification is one of the most important statistical learning tasks, of which the goal is to predict the label, given input features using a previously trained model or rule. In general, the success of prediction models relies on the accessibility of a large number of labeled training data. However, in many healthcare settings, only a minority of available data is accurately labeled, while unlabeled data is abundant.

Thus, an attractive research direction is developing semi-supervised learning methods that can incorporate easy-to-collect unlabeled data to help improve prediction accuracy. While there have been many works following this direction, there are fewer works that can take longitudinal input such as laboratory tests and charted events that are commonly seen in EHR. Further, most existing approaches separate feature extraction using unlabeled data and building prediction models using labeled data into two steps.

In Chapter IV, we propose a semi-supervised learning method targeting for longitudinal clinical events data. Specifically, our model consists of a sequence generative model and a label prediction model, and the two parts are learned end to end using both labeled and unlabeled data in a joint manner to obtain better prediction performance. With this approach, related data can cluster well in the learned feature space under supervision from labeled data. We develop efficient algorithms based on variational inference techniques to estimate parameters. Using five mortality-related classification tasks on the MIMIC III database, we demonstrate that the proposed method outperforms the purely supervised method that uses labeled data only and existing two-step semi-supervised methods. Our approach can help reduce the cost of collecting clinical labels in building prediction models. The materials in this chapter are adapted from joint work with Jiaqi Ma, Akbar Waljee, and Ji Zhu (Tang et al., 2020).

CHAPTER II

Survival Analysis via Ordinary Differential Equations

2.1 Introduction

Survival analysis is an important branch of statistical modeling, where the primary outcome of interest is the time to a certain event. In practice, event times may not be observed due to a limited observation time window or missing follow-up during the study, which is referred to as censored data. Survival analysis methods handle these complex and incomplete censored data and have important applications in various real-world problems, such as rehospitalization, cancer survival in healthcare, reliability of devices, and customer lifetime (Chen et al., 2009; Miller Jr., 2011; Modarres et al., 2016).

Many statistical models have been developed to deal with censored data in the literature. For example, the Cox proportional hazard model is probably the most classical semi-parametric model for handling censored data (Cox, 1975), and it assumes that the covariates have a constant multiplicative effect on the hazard function. Although easy to interpret, the constant hazard ratio assumption is often considered as overly strong for real-world applications. As a result, many other semi-parametric models have been proposed as attractive alternatives, such as accelerated failure time (AFT) models, transformation models, and additive hazards models. See Aalen (1980), Buckley and James (1979), Gray (1994), Bennett (1983), Cheng et al. (1995), Fine et al. (1998), and Chen et al. (2002) for a sample of references. Given different assumptions made in these semi-parametric models, different estimation and inference procedures have also been developed accordingly, such as maximum partial likelihood based estimators (MPLE) (Zucker and Karr, 1990; Gray, 1994; Bagdonavicius and Nikulin, 2001; Chen et al., 2002), least square and rank-based methods (Buckley and James, 1979; Lai and Ying, 1991; Tsiatis, 1990; Jin et al., 2003, 2006), non-parametric maximum likelihood estimators (NPMLE) (Murphy et al., 1997; Zeng and Lin, 2007b), and sieve maximum likelihood estimators (MLE) (Huang, 1999; Shen and Wong, 1994; Ding and Nan, 2011; Zhao et al., 2017).

Despite successful applications of these existing models and their estimation methods, the growing scale and diversity of formates of modern data poses new challenges that are not entirely addressed. For example, electronic health records of millions of patients over several decades are readily available, and they include laboratory test results, radiology images, and doctors' clinical notes (Johnson et al., 2016; Goldstein et al., 2017). Research towards more flexible and scalable modeling of event times has attracted great attention in recent years.

In this chapter, we introduce a novel Ordinary Differential Equation (ODE) notion for survival analysis. We show that it provides a unified view of many existing survival models and opportunities to develop more flexible model structures such as neural networks and, more importantly, facilitates the development of a scalable and easy-to-implement estimation and inference procedure, which can be applied to a wide range of ODE survival models. We note that the proposed approach is founded upon well-established numerical solvers and sensitivity analysis tools for ODEs, and it overcomes various practical limitations of existing estimation methods when applied to different survival models for large-scale studies.

Specifically, the proposed framework models the dynamic change of the cumulative hazard function[1] through an ODE. Let T be the event time and X be covariates. Denote the conditional cumulative hazard function of T given $X = x$ as $\Lambda_x(t)$. Then $\Lambda_x(t)$ is characterized by the following ODE with a fixed initial value

$$\begin{cases} \Lambda_x'(t) = f(t, \Lambda_x(t), x) \\ \Lambda_x(t_0) = c(x) \end{cases}, \tag{2.1}$$

where the derivative is with respect to t, $f(\cdot)$ and $c(\cdot)$ are functions to be specified, and t_0 is a predefined initial time point. In particular, function $c(\cdot)$ determines the probability of an event occurring after t_0; for instance, $\Lambda_x(0) = 0$ corresponds to the case when no event occurs before time 0. Further, function $f(\cdot)$ determines how covariates x affect the hazard function at time t given an individual's own cumulative hazard. Thus, different specifications of the function $f(\cdot)$ lead to different ODE models.

Next, we comment on both benefits of the ODE approach in terms of modeling and computation and new theoretical challenges induced by the ODE notion, and summarize our contributions below.

- Firstly, the ODE modeling framework is general enough to unify many aforementioned existing survival models through different specifications of the function $f(\cdot)$, which serves as the foundation of a widely applicable estimation procedure that will be developed later. For example, the ODE (2.1) is equivalent to the Cox model when $f(\cdot)$ takes the form $\alpha(t)\exp\big(x^T\beta\big)$ for some function $\alpha(\cdot)$, and it is equivalent to the AFT model when $f(\cdot)$ takes the form $q(\Lambda_x(t))\exp\big(x^T\beta\big)$ for some function $q(\cdot)$. Similarly, we can obtain many more models such as the time-varying variants of the Cox model, the linear trans-

[1] The derivative of the cumulative hazard function describes the instantaneous rate at which the event occurs given survival, and is a popular modeling target in survival analysis.

formation model, and the additive hazards model to name a few (see Section 2.2 for details). We note that the ODE notion can provide new and sometimes more explicit interpretations in terms of the hazard by re-writing the existing models in the ODE form. In addition, the generality of the proposed framework offers an opportunity for designing more flexible model structures and model diagnostics.

- Secondly, and also more importantly, introducing the ODE notion facilitates the development of a general and easy-to-implement procedure for estimation and inference in large-scale survival analysis. In Section 2.3.1, we illustrate the proposed procedure by using a general class of ODE models as an example. In particular, this general class includes the most flexible linear transformation model, where both the transformation function and the error distribution are unspecified. Since the $f(\cdot)$ function for the general model contains both finite-dimensional and infinite-dimensional parameters, we propose a spline-based sieve MLE that directly maximizes the likelihood in a sieve space. We provide an easy-to-implement gradient-based optimization algorithm founded upon *local sensitivity analysis* tools for ODEs (Dickinson and Gelinas, 1976), where numerical ODE solvers are used to compute the log-likelihood function and its gradients. Since efficient implementations of both ODE solvers and splines are available in many software, the resultant algorithm is easy to carry out in practice. It is worth noting that, in comparison to existing estimation methods, the proposed procedure has advantages in various aspects, such as scalability against MPLE for the time-varying Cox model, optimization-parameter efficiency against NPMLE, statistical efficiency and numerical stability against rank-based methods for the linear transformation model. We demonstrate these advantages through extensive simulation studies. For example, when the sample size is $8,000$, it takes the proposed ODE approach about 6 seconds to estimate the semi-parametric ODE-AFT model while the rank-based method needs 350 seconds.

- Thirdly, we note that the ODE notion brings new challenges to asymptotic distributional theory. While many asymptotic distributional theories for M-estimation in semi-parametric models have been developed (see Huang (1999), Shen (1997), Ai and Chen (2003), Wellner and Zhang (2007), Zhang et al. (2010), He et al. (2010), Ding and Nan (2011) for a sample of references), they cannot be directly applied to our setting. Among them, the proposed theory in Ding and Nan (2011) considers bundled parameters where the infinite-dimensional parameter is an unknown function of the finite-dimensional Euclidean parameter and has been applied to the AFT model, and recently, to the accelerated hazards model in Zhao et al. (2017). However, for the general class of ODE models, the estimation criterion is parameterized with more general bundled parameters where the nuisance parameter is an unknown function of not only finite-dimensional regression parameters of interest but also other infinite-dimensional nuisance parameters. To accommodate this different and challenging scenario induced by the ODE notion, we develop a new sieve M-theorem for more general bundled parameters. By applying it to the general class of ODE models along with ODE related methodologies (Walter, 1998), we show consistency, asymptotic normality, and semi-parametric efficiency for the estimated regression parameters. The proposed theory can also be extended to develop the asymptotic normality of estimators for other ODE models.

- Finally, we recognize an intrinsic optimization challenge due to the existence of censoring when leveraging highly flexible model structures such as neural networks for survival analysis, and show how the proposed ODE framework can be naturally used to address this optimization challenge, so as to accommodate data in diverse formats and make powerful predictions. In particular, we model $f(\cdot)$ as a neural network taking the cumulative hazard $\Lambda_x(t)$, the time t, and the covariates x as inputs, and allow efficient estimation of the model in large-scale applications using stochastic gradient descent. Compared with existing methods in deep learning survival analysis (Faraggi and Simon, 1995; Ching et al., 2018; Katzman et al., 2018; Lee et al., 2018; Gensheimer and Narasimhan, 2019; Chapfuwa et al., 2018; Kvamme et al., 2019), our proposed method is able to provide a broad family of continuous-time survival distributions without strong structural assumptions and potential information loss from discretizing event times.

The rest of this chapter is organized as follows. We introduce the ODE modeling framework and present a general class of ODE models as special cases in Section 2.2. We provide the estimation and inference procedure and related theoretical results in Section 2.3. We present that the flexibility of this unified ODE framework allows us to design a neural-network-based survival model for powerful predictions on complex data in diverse formats in Section 2.4.

2.2 The Unified ODE Modeling Framework

To characterize the conditional distribution of T given X, the conditional hazard function, denoted as $\lambda_x(t) = \Lambda'_x(t)$, provides a popular modeling target as it describes the instantaneous rate at which the event occurs given survival. In this paper, we view the hazard function as the dynamic change of the cumulative hazard function and quantify them using an ODE.

In our ODE framework, the hazard function depends not only on the time and covariates but also on the cumulative hazard as shown in (2.1), where function $f(\cdot)$ specifies the dynamic change of $\Lambda_x(t)$ and covariates x serve as additional parameters in terms of the ODE. The initial value in (2.1) implies that, for an individual with covariates x, the probability for an event to occur after t_0 is controlled by $\exp(-c(x))$. For example, it is often the case that time 0 is defined prior to the occurrence of events, which implies that an event always occurs after time 0, i.e. the survival function $S_x(0) = 1$, and it follows that $\Lambda_x(0) = 0$. We use this initial value in the ODE framework hereafter for simplicity, while the estimation method and the theoretical properties established later can be extended to the general case where $c(x)$ can be a function of covariates. Under certain smoothness conditions (Walter, 1998, page 108), the initial value problem (2.1) has exactly one solution, which uniquely characterizes the conditional distribution of the event time.

Next, we present a general class of ODE models as an instantiation of the ODE framework. Suppose there are two groups of covariates denoted by $X \in \mathbf{R}^{d_1}$ and $Z \in \mathbf{R}^{d_2}$ respectively. We consider ODE models in the form of

$$\Lambda'_{x,z}(t) = \alpha(t) \exp\big(x^T \beta + z^T \boldsymbol{\eta}(t)\big) q(\Lambda_{x,z}(t)), \tag{2.2}$$

where $\alpha(\cdot)$ and $q(\cdot)$ are two unknown positive functions, and given an individual's own cumulative hazard, both covariates x and z have multiplicative effects on the hazard, one with time-independent coefficients $\beta \in \mathbf{R}^{d_1}$ and the other with time-varying coefficients $\boldsymbol{\eta}(t) \in \mathbf{R}^{d_2}$. Here $\boldsymbol{\eta}(\cdot) = (\eta_1(\cdot), \ldots, \eta_{d_2}(\cdot))^T$.[1] We note that this general class of ODE models is a specific example; other examples beyond this class are included in Remark II.2 to further illustrate the flexibility of the proposed ODE framework. In particular, this general class covers many existing models as special cases. As shown below, model (2.2) reduces to the time-varying Cox model when $q(\cdot) = 1$, to the linear transformation model when covariates z are not considered, and further reduces to the AFT model if $\alpha(\cdot) = 1$. In the following, we will also show that by rewriting many existing models under the format (2.1), the ODE framework

[1] Throughout this chapter, we bold vectors only when each element is a function.

brings them new interpretations in terms of the hazard function.

2.2.1 Cox Model and Time-Varying Cox Model

The Cox proportional hazard model assumes that the covariates have a multiplicative effect on the hazard function, i.e. $\lambda_x(t) = \alpha(t)\exp\bigl(x^T\beta\bigr)$, where $\alpha(t)$ is a baseline hazard function and $\exp\bigl(x^T\beta\bigr)$ is the relative risk, and extensions of the Cox model allow for time-varying coefficients (Zucker and Karr, 1990; Gray, 1994). Here we write the Cox model with both time-independent and time-varying effects as a simple ODE, whose right-hand side does not depend on the cumulative function, i.e.

$$\Lambda'_{x,z}(t) = \alpha(t)\exp\bigl(x^T\beta + z^T\boldsymbol{\eta}(t)\bigr), \tag{2.3}$$

which allows covariates x to have time-independent effects and covariates z to have time-varying effects on the hazard function. The baseline hazard function $\alpha(t)$ and time-varying effects $\boldsymbol{\eta}(t)$ can be specified in a parametric model or left unspecified in a semi-parametric model.

2.2.2 Accelerated Failure Time Model

The AFT model assumes that the log transformation of T is linearly correlated with covariates, i.e. $\log T = -X^T\beta + \epsilon$. In the proposed ODE framework, the AFT model can be written as

$$\Lambda'_x(t) = q(\Lambda_x(t))\exp\bigl(x^T\beta\bigr), \tag{2.4}$$

where the function $q(\cdot)$ uniquely determines the distribution of error ϵ in the following way. Let $H_q(u) = \int_0^{-\ln u} q^{-1}(v)dv$ and $G_q(u) = H_q^{-1}(u)$, then G_q is the survival function of $\delta = \exp(\epsilon)$ as shown in Bagdonavicius and Nikulin (2001). For example, if $q(t) = vk^{\frac{1}{v}}t^{1-\frac{1}{v}}$, then δ follows a Weibull distribution with $G_q(t) = \exp(-kt^v)$. When the error distribution is unknown (as in a semi-parametric AFT model), we can leave the function $q(\cdot)$ unspecified.

The ODE (2.4) provides a new and clear interpretation on how covariates affect the hazard for the AFT model. Specifically, it implies that given an individual's own cumulative hazard, covariates x have a multiplicative constant effect on the hazard function. Further, besides the direct effects of covariates, if $q(\cdot)$ is a monotonic increasing function, then an individual with a higher cumulative hazard at a particular time would have a higher "baseline" hazard. Note that although we can also present the hazard directly as a function of covariates and time, i.e. $\lambda_x(t) = \lambda_\delta(t \exp(x^T\beta)) \exp(x^T\beta)$, the covariate effects are entangled with the baseline hazard λ_δ in this representation, which is more difficult to interpret.

2.2.3 Linear Transformation Model

As an extension of the AFT model, the linear transformation model assumes that, after a monotonic increasing transformation $\varphi(\cdot)$, the event time T is linearly correlated with covariates, i.e. $\varphi(T) = -X^T\beta + \epsilon$. In the proposed ODE framework, it can be written as

$$\Lambda'_x(t) = q(\Lambda_x(t)) \exp(x^T\beta) \alpha(t), \tag{2.5}$$

where $q(\cdot)$ corresponds to the distribution of ϵ in the same way as in the AFT model, and $\alpha(\cdot)$ is uniquely determined by the equation $\varphi(t) = \log \int_0^t \alpha(s)ds$. In comparison to model (2.4), the hazard function at time t depends not only on the current cumulative hazard and covariates, but also on the current time t directly.

Different specifications of $\varphi(\cdot)$ and ϵ have been proposed in the literature for the linear transformation model. We consider the case where both the transformation and the error distribution are unknown. This specification is especially preferred when parametric assumptions on the transformation function or the error distribution cannot be properly justified. However, when both $q(\cdot)$ and $\alpha(\cdot)$ are unknown, they may not be identifiable. The equivalent linear regression representation, $\varphi(T) = -x^T\beta + \epsilon$, allows us to see the identifiability issue clearly. Note that, when no covariate is associated with survival, i.e., $\beta = 0$, non-identifiability issue arises because parameters (φ, ϵ) and $(f(\varphi), f(\epsilon))$ give the same event time distribution for any arbitrary function f. Therefore, we consider $\beta \neq 0$, in which case Horowitz (1996) showed that the model parameters are identifiable up to a scale and a location normalization under certain regularity conditions. Following that result, we have developed Proposition 2.2.1 that characterizes the identifiability of parameters in (2.5), while Proposition 2.2.2 provides necessary and sufficient degeneration conditions for AFT and Cox models. The proofs are given in the Supplemental Material.

Proposition 2.2.1. *Suppose at least one of the covariates in x is continuous and this covariate has a non-zero β coefficient, which without loss of generality is assumed to be positive. Let $(q(\cdot), \beta, \alpha(\cdot))$ specify the survival distribution through (2.5). Then for any other $(\tilde{q}(\cdot), \tilde{\beta}, \tilde{\alpha}(\cdot))$ that gives the same survival distribution, if and only if there exist positive constants c_1 and c_2 such that $\tilde{\beta} = c_1\beta$, $\int_0^t \tilde{\alpha}(s)ds = c_2(\int_0^t \alpha(s)ds)^{c_1}$, and $\int_0^t \tilde{q}^{-1}(s)ds = c_2(\int_0^t q^{-1}(s)ds)^{c_1}$ for any $t > 0$.*

Proposition 2.2.2. *Suppose the conditions in Proposition 2.2.1 hold, then the linear transformation model in (2.5) coincides with the Cox model if and only if there exist positive constants c_1 and c_2 such that $q(u) = c_2 u^{1-c_1}$, and it coincides with the AFT model if and only if there exist positive constants c_1 and c_2 such that $\alpha(t) = c_2 t^{c_1 - 1}$ for $t > 0$.*

Remark II.1. *Note that the original forms of the AFT model and the linear transformation model do not directly take time-varying coefficients. Existing works on the linear transformation model that consider varying coefficients choose to model them as a function of certain covariates rather than a function of time (Chen and Tong, 2010; Qiu and Zhou, 2015). In contrast, the equivalent ODE forms of the AFT model in*

(2.4) and the linear transformation model in (2.5) can naturally accommodate time-varying coefficients. For example, we can consider the generalization in (2.2), where given an individual's own cumulative hazard covariates z have time-varying multiplicative effects $\boldsymbol{\eta}(t)$ on the hazard. In particular, this generalization is equivalent to a covariate-dependent transformation model

$$\varphi_Z(T) = -X^T\beta + \epsilon,$$

where $\varphi_z(t) = \log \int_0^t \alpha(s) \exp\big(z^\top \boldsymbol{\eta}(s)\big)\, ds$, i.e., covariates z have multiplicative time-varying effect $\boldsymbol{\eta}(t)$ on the gradient of $\exp(\varphi_z(t))$.

Remark II.2. *The proposed ODE framework is general enough to cover other existing models as well. For example, both the additive hazard model (Aalen, 1980; Mckeague and Sasieni, 1994) and the additive-multiplicative hazard model (Lin and Ying, 1995) can be viewed as a specific ODE model, i.e. $\Lambda'_{x,z}(t) = r_1(x^T\beta) + \alpha(t)r_2(z^T\eta)$, where $r_1(\cdot)$ and $r_2(\cdot)$ are some known link functions. Subsequently, the generalized additive hazards model and the generalized additive-multiplicative hazards model (Bagdonavicius and Nikulin, 2001) can be written as $\Lambda'_x(t) = q(\Lambda_x(t))(r_1(x) + \alpha(t)r_2(x))$. The generalized Sedyakin's model (Bagdonavicius and Nikulin, 2001), which was proposed as an extension of the AFT model, can also be viewed as a special case of (2.1) with $\Lambda'_x(t) = f(\Lambda_x(t), x)$.*

Remark II.3. *Further, the proposed ODE framework and the estimation method in Section 2.3.1 can also be extended to deal with time-varying covariates. Suppose the covariate is a stochastic process $X(t), t \geq 0$ and $T_{X(\cdot)}$ is the failure time under $X(\cdot)$. Denote the conditional survival, the hazard function, and the cumulative function by $S_{x(\cdot)}(t) = P\{T_{X(\cdot)} \geq t | X(s) = x(s), 0 \leq s \leq t\}$, $\lambda_{x(\cdot)}(t) = -\frac{S'_{x(\cdot)}(t)}{S_{x(\cdot)}(t)}$, and $\Lambda_{x(\cdot)}(t) = -\log\big(S_{x(\cdot)}(t)\big)$, respectively. Then the ODE (2.1) can be extended to $\Lambda'_{x(\cdot)}(t) = f(t, \Lambda_{x(\cdot)}(t), x(t))$. This extension also covers many existing models as spe-*

cial cases. For example, the linear transformation model with time-varying covariates (Zeng and Lin, 2006) can be written as $\Lambda'_{x(\cdot)}(t) = q(\Lambda_{x(\cdot)}(t)) \exp\big(x(t)^T\beta\big)\alpha(t)$, and the Cox model with time-varying covariates can be viewed as a special case with $q(\cdot) \equiv 1$. For presentation simplicity, we focus on models in the form of (2.2) in this paper.

2.2.4 Related Estimation Methods and Their Limitations

The maximum partial likelihood estimator (MPLE) (Cox, 1975) was first proposed for the Cox model, and the asymptotic property of MPLE was established by Andersen and Gill (1982) via the counting process martingale theory. For time-varying Cox models, many different estimation methods have been developed while relying on maximizing the partial likelihood (Zucker and Karr, 1990; Gray, 1994). However, evaluating the partial likelihood for an uncensored individual requires access to all other observations who were in its risk set. This prevents parallel computing for partial likelihood-based methods, which is a drawback when analyzing large scale data.

For the linear transformation model, different specifications of the transformation and the error distribution along with different estimation methods have been proposed. For example, Cheng et al. (1995), Fine et al. (1998), Shen (1998), Chen et al. (2002), and Bagdonavicius and Nikulin (1999) have considered an unknown transformation with a known error distribution, which includes the Cox model and the proportional odds model (Bennett, 1983) as special cases. The corresponding modified MPLE (Chen et al., 2002; Bagdonavicius and Nikulin, 1999), sieve MLE (Shen, 1998), and NPMLE (Murphy et al., 1997; Zeng and Lin, 2007b) have also been developed. However, due to the large number of nuisance parameters, it is difficult to obtain NPMLE in practice, especially in large-scale applications. Alternatively, Cai et al. (2005) considered a parametric Box-Cox transformation with an unknown error distribution, which includes the semi-parametric AFT model as a special case, and least square and rank-based methods have been proposed to estimate the regression parameters (Buckley and James, 1979; Lai and Ying, 1991; Tsiatis, 1990; Jin et al., 2003, 2006). Nevertheless, they are not asymptotically efficient and may suffer additional numerical errors resulting from discrete objective functions. Subsequently, under the AFT model, Zeng and Lin (2007a) and Lin and Chen (2012) proposed efficient estimators based on a kernel-smoothed profile likelihood, and Ding and Nan (2011) developed an efficient sieve MLE. When both the transformation function and the error distribution are unknown, a partial rank-based method has been proposed (Khan and Tamer, 2007; Song et al., 2006), and its computation is analogous to that of the partial likelihood, where the rank of an uncensored individual is determined by all other individuals in its risk set, and thus the computational challenge for large-scale applications still remains.

As evident from the above discussion, many existing estimation methods suffer from important limitations in practice. In Section 2.3.1, we propose a scalable, easy-to-implement and efficient estimation method that can be applied to a wide range of models.

2.3 The Efficient Procedure for Estimation and Inference

In this section, we provide the estimation procedure in Section 2.3.1 and establish the consistency and asymptotic normality of the estimates for statistical inference in Section 2.3.2. Simulation studies and a real-world data example are presented in Sections 2.3.3 and 2.3.4 respectively.

2.3.1 Maximum Likelihood Estimation

We propose a general estimation procedure that can be applied to a wide range of ODE models. Here we use the ODE model in (2.2) as an illustrative example, and the proposed estimation method can also be applied to other models such as those mentioned in Remark II.2.

We denote the event time as T, the censoring time as C. Let $Y = \min\{T, C\}$ and $\Delta = \mathbb{1}(T \le C)$, where $\mathbb{1}(\cdot)$ denotes the indicator function. Our data consist of n independent and identically distributed observations $\{Y_i, \Delta_i, X_i, Z_i\},\ i = 1, \dots, n$. Since $\alpha(\cdot)$ and $q(\cdot)$ in (2.2) are positive, we set $\gamma(\cdot) = \log \alpha(\cdot)$ and $g(\cdot) = \log q(\cdot)$. Under the conditional independence between T and C given covariates (X, Z), the log-likelihood function of the parameters $(\beta, \gamma(\cdot), \boldsymbol{\eta}(\cdot), g(\cdot))$ is given by

$$l_n(\beta, \gamma(\cdot), g(\cdot), \boldsymbol{\eta}(\cdot)) = \frac{1}{n}\sum_{i=1}^{n}[\Delta_i\{\gamma(Y_i) + X_i^T\beta + Z_i^T\boldsymbol{\eta}(Y_i) + g(\Lambda_i(Y_i; \beta, \gamma, g, \boldsymbol{\eta}))\} \tag{2.6}$$

$$- \Lambda_i(Y_i; \beta, \gamma, \boldsymbol{\eta}, g)],$$

where $\Lambda_i(t; \beta, \gamma, \boldsymbol{\eta}, g)$ denotes the solution of ODE (2.2) parameterized by $(\beta, \gamma, \boldsymbol{\eta}, g)$ given covariates $X = X_i$ and $Z = Z_i$. The log-likelihood function (2.6) includes both finite-dimensional parameter β and infinite-dimensional parameters $\gamma, \boldsymbol{\eta}, g$.

We propose a sieve MLE that maximizes the log-likelihood over a sequence of finite-dimensional parameter spaces that are dense in the original parameter space as the sample size increases. The sieve space can be chosen as linear spans of many types of basis functions with desired properties (Chen, 2007). In particular, we construct the sieve space using polynomial splines due to their capacity in approximating complex functions and the simplicity of their construction. Under suitable smoothness conditions, $\gamma_0(\cdot)$, $\boldsymbol{\eta}_0(\cdot)$, and $g_0(\cdot)$, the true parameters associated with the data generating distribution, can be well approximated by some functions in the space of polynomial splines as defined in Schumaker (2007, page 108, Definition 4.1). Further, there exists a group of spline bases such that functions in the space of polynomial splines can be written as linear combinations of the spline bases (Schumaker, 2007, page 117, Corollary 4.10). Different groups of spline bases may be used for the estimation of different parameters $(\gamma, \boldsymbol{\eta})$ and g because of their different domains.

Specifically, we construct the proposed sieve estimator as follows. Let $\mathcal{B} \subset \mathbf{R}^{d_1}$ be the parameter space of β. Let $\{B_j^1, 1 \le j \le q_n^1\}$ and $\{B_j^2, 1 \le j \le q_n^2\}$ be two groups of spline bases that are used for the estimation of parameters $(\gamma, \boldsymbol{\eta})$ and g respectively. Here the number of spline bases, q_n^i, should grow sublinearly in rate $O(n^{v_i})$ for some $v_i \in (0, 0.5)$, $i = 1, 2$ for convergence guarantee (see Section 2.3.2 for rigorous definitions). Overall, we wish to find $d_2 + 1$ members $(\gamma, \eta_1, \cdots, \eta_{d_2})$ from the space of polynomial splines associated with $\{B_j^1\}$, one member g from that associated with $\{B_j^2\}$, along with $\beta \in \mathcal{B}$ to maximize the log-likelihood function (2.6). Let $Z_{i0} = 1$, $Z_i = (Z_{i1}, \cdots, Z_{id_2})^T$. Then the objective function can be written as

$$\begin{aligned} l_n(\beta, a, b) = \frac{1}{n} \sum_{i=1}^{n} \Big[\Delta_i \{ X_i^T \beta + \sum_{l=0}^{d_2} \sum_{j=1}^{q_n^1} a_j^l B_j^1(Y_i) Z_{il} + \sum_{j=1}^{q_n^2} b_j B_j^2(\Lambda_i(Y_i; \beta, a, b)) \} \\ - \Lambda_i(Y_i; \beta, a, b) \Big], \end{aligned} \tag{2.7}$$

where $a = \left(a_j^l\right)_{j=1,\cdots,q_n^1, l=0,\cdots,d_2}$ and $b = (b_j)_{j=1,\cdots,q_n^2}$ are the coefficients of the spline bases, and $\Lambda_i(t; \beta, a, b)$ is the solution of

$$\begin{cases} \Lambda_i'(t) = \exp\Big(X_i^T \beta + \sum_{l=0}^{d_2} \sum_{j=1}^{q_n^1} a_j^l B_j^1(t) Z_{il} + \sum_{j=1}^{q_n^2} b_j B_j^2(\Lambda_i(t)) \Big), \\ \Lambda_i(0) = 0. \end{cases} \tag{2.8}$$

The proposed sieve estimators are given by $\hat{\beta}_n = \hat{\beta}$, $\hat{\boldsymbol{\eta}}_n(\cdot) = \Big(\sum_{j=1}^{q_n^1} \hat{a}_j^1 B_j^1(\cdot), \ldots, \sum_{j=1}^{q_n^1} \hat{a}_j^{d_2} B_j^1(\cdot) \Big)$, $\hat{\gamma}_n(\cdot) = \sum_{j=1}^{q_n^1} \hat{a}_j^0 B_j^1(\cdot)$, and $\hat{g}_n(\cdot) = \sum_{j=1}^{q_n^2} \hat{b}_j B_j^2(\cdot)$, where $(\hat{\beta}, \hat{a}, \hat{b})$ maximizes the objective function (2.7).

Note that the objective function (2.7) contains the solution of a parameterized ODE (i.e. (2.8)), and this is different from most traditional optimization problems. In particular, it is nontrivial to evaluate the objective function and its gradient with respect to parameters when there is no closed-form solution for the ODE. To address this optimization challenge, we develop a gradient-based optimization algorithm by taking advantage of local sensitivity analysis (Dickinson and Gelinas, 1976; Petzold et al., 2006) and well-implemented ODE solvers. Specifically, we evaluate the objective function and its gradient as follows:

1. we numerically calculate $\Lambda_i(Y_i; \beta, a, b)$ by solving (2.8) given the current parameter estimates β, a, b and covariates X_i, Z_i, the initial value at $t_0 = 0$, and the evaluating time $t = Y_i$;

2. we evaluate the derivative of $\Lambda_i(Y_i; \beta, a, b)$ with respect to the parameters β, a, and b through solving another ODE which is derived by local sensitivity analysis, and calculate the gradient of the objective function by the chain rule.

We summarize the results of the local sensitivity analysis in the following, and provide detailed derivations in the Supplemental Material. The local sensitivity analysis is a technique that studies the rate of change in the solution of an ODE system with respect to the parameters. There are two ways to obtain the sensitivity: forward sensitivity analysis and adjoint sensitivity analysis. Both of them require solving another ODE with some fixed initial value. For example, we consider to compute the gradient of $\Lambda(y; \theta)$ with respect to its parameter θ, where $\Lambda(t; \theta)$ is the solution of (2.8) and θ consists of parameters β, a, and b in our case. For presentation simplicity, we denote the right-hand side of (2.8) by the function $f(t, \Lambda; \theta)$, i.e.

$$f(t, \Lambda; \theta) = \exp\left(X^T\beta + \sum_{l=0}^{d_2} \sum_{j=1}^{q_n^1} a_j^l B_j^1(t) Z_j + \sum_{j=1}^{q_n^2} b_j B_j^2(\Lambda) \right),$$

and its partial derivative with respect to θ and Λ by f'_θ and f'_Λ respectively. In forward sensitivity analysis, it can be shown that the partial derivative of $\Lambda(y; \theta)$ with respect to θ is given by the solution of (2.9) at $t = y$, i.e. $\Lambda'_\theta(y;\theta) = F_1(y)$ with F_1 satisfying

$$\begin{cases} F_1'(t) = f_\theta'(t,\Lambda;\theta) + f_\Lambda'(t,\Lambda;\theta)F_1, \\ F_1(0) = 0. \end{cases} \tag{2.9}$$

In the alternative adjoint sensitivity analysis, we can show that the partial derivative can also be obtained by evaluating the solution of (2.10) at $t = 0$, i.e. $\Lambda'_\theta(y;\theta) = F_2(0)$ with F_2 satisfying

$$\begin{cases} (\kappa(t); F_2'(t)) = (-\kappa \cdot f_\Lambda'(t,\Lambda;\theta); -\kappa \cdot f_\theta'(t,\Lambda;\theta)), \\ (\kappa(t); F_2(t))|_{t=y} = (1; \mathbf{0}). \end{cases} \tag{2.10}$$

Thus, after plugging the form of $f(t,\Lambda;\theta)$ into either (2.9) or (2.10), we can obtain the gradients through solving the corresponding ODE. In Remark II.4, we compare the computational complexity of forward and adjoint sensitivity analyses and provide a general guidance on which sensitivity analysis to use when computing gradients under survival ODE models.

It is worth noting that the proposed estimation method can be easily implemented using existing computing packages. For example, the “Optimization Toolbox” in MATLAB contains “fminunc” for unconstrained optimization and “fmincon” for constrained optimization; both require initialization and the objective function. In our implementation, we also provide evaluation of the gradient for faster and more reliable computations. In particular, we compute both the objective function and the gradient by well-implemented ODE solvers in MATLAB. In addition, we construct the sieve space using B-splines for its numerical simplicity, whose implementation is available in the “Curve Fitting Toolbox”.

Remark II.4. *In general, forward sensitivity analysis is computationally more efficient when the dimension of the ODE system is relatively large and the number of* *parameters is small, while adjoint sensitivity analysis is best suited in the complementary scenario. See Dickinson and Gelinas (1976) and Petzold et al. (2006) for more details. For a general ODE model such as (2.1) where the size of the ODE system is 1 and the number of parameters increases as the sample size n grows, we can use the adjoint sensitivity analysis along with parallel computing for n independent individuals. Alternatively, if the memory permits, we can combine ODEs for n individuals into a large ODE system with n dimensions, which is larger than the number of parameters, and then the forward sensitivity analysis is preferred.*

Remark II.5. *Moreover, we introduce a computational trick for the general class of ODE models in (2.2) that can significantly accelerate the evaluation of the objective and gradients, where we need to solve ODEs for n independent individuals. Specifically, the trick transforms the problem of solving n different ODEs at their respective observed times into a problem of solving a single ODE at n different time points. More generally, this trick can be applied to any ODE model where the right-hand side is separable in the way that $f(t, \Lambda_x; \theta, x) = f_1(t; \theta, x) f_2(\Lambda_x; \theta)$ with two functions f_1 and f_2. We refer to the Supplemental Material for more details about this computational trick.*

Remark II.6. *The proposed sieve MLE can also be applied to many existing models. For example, for the time-varying Cox model where $q(\cdot) = 1$, we can remove the function $g(\cdot)$ from the objective function (2.6). For the semi-parametric AFT model where Z is not considered and $\alpha(\cdot) = 1$, we can just keep parameters β and $g(\cdot)$ in (2.6). For the linear transformation model, if either $q(\cdot)$ or $\alpha(\cdot)$ is specified, we can replace the corresponding term in (2.6) with the specified finite-dimensional parametric form. Also note that in comparison to existing estimation methods in Section 2.2.4, the proposed estimation method allows parallel computing, which is especially important for large-scale applications. Specifically, since the log-likelihood of each individual only depends on its own observations, the evaluation for independent data points can be* *carried out simultaneously. Further, compared with the NPMLE where the number of optimization parameters is linear in n (Murphy et al., 1997; Zeng and Lin, 2007b), the number of optimization parameters used in sieve MLE increases more slowly with the sample size.*

Remark II.7. *The objective function (2.7) is convex with respect to β and a for the (time-varying) Cox model, where the parameter b is not included, and the global optimum can be achieved quickly. For the general case, the objective function is nonconvex and the optimization algorithm may converge to a local optimum. Nevertheless, based on our extensive simulation studies, the algorithm generally performs well with appropriately chosen initialization, such as initializing the algorithm with the estimates from the Cox model.*

Remark II.8. *Note that different identifiability conditions are required for different survival models. Thus, we need to add corresponding constraints in the optimization algorithm.*

- *For the general ODE model (2.2) where both covariates X (with time-independent effects) and Z (with at least one non-zero time-varying effect) are considered, two groups of parameters $(\beta, \gamma, g, \boldsymbol{\eta})$ and $(\tilde{\beta}, \tilde{\gamma}, \tilde{g}, \tilde{\boldsymbol{\eta}})$ give the same survival distribution if and only if $\beta = \tilde{\beta}$, $\gamma = \tilde{\gamma} + c$, $g = \tilde{g} - c$, and $\boldsymbol{\eta} = \tilde{\boldsymbol{\eta}}$ for some constant c. To guarantee the identifiability, we can constrain either the value of $\gamma(\cdot)$ at a fixed time point t^* or the norm of $\gamma(\cdot)$, in which the former leads to a linear constraint on the coefficients of spline bases.*

- *For the linear transformation model where the time-varying effects are not considered and at least one component of X has a non-zero coefficient, parameters (β, γ, g) are identifiable up to two scaling factors as shown in Proposition 2.2.1. To guarantee identifiability, we can put constraints on β and γ. For β, we can either constrain the first element of β to be 1 (Khan and Tamer, 2007; Song et al.,* *2006), which can be naturally achieved by arranging covariates if we know which covariate has a non-zero effect, or set $\|\beta\| = 1$. For γ, we can add a similar constraint as that for the general ODE model (2.2). Alternatively, we can put constraints on γ and g by setting $\int_0^{t^*} \exp(\gamma(s))ds = c_1$ and $\int_0^{t^*} \exp(-g(s))ds = c_2$, with some positive constants $c_1 \neq c_2 > 0$ and a fixed time point t^*. In our implementation, we choose to use two linear constraints, i.e. set the first element of β to 1 and $\gamma(t^*) = 0$ for simplicity in optimization.*

2.3.2 Theoretical Properties

In this section, we study the theoretical properties of the proposed sieve MLE. Although many works have investigated asymptotic distributional theories for M-estimation with bundled parameters (Ai and Chen, 2003; Chen et al., 2003; Ding and Nan, 2011), their results cannot be directly applied to our setting. In particular, the nuisance parameters in existing works often take the form of an unknown function of only some finite-dimensional Euclidean parameters of interest. However, our work focuses on a more general scenario, where the nuisance parameter is an unknown function of not only the Euclidean parameters but also some other infinite-dimensional nuisance parameters. To deal with theoretical challenges due to the additional functional nuisance parameters, we develop a new sieve M-theorem for the asymptotic theory of a general family of semi-parametric M-estimators. Moreover, we apply the proposed general theorem to establish the asymptotic normality and semi-parametric efficiency of the proposed sieve MLE $\hat{\beta}_n$ when the convergence rate of the sieve estimator of the nuisance parameter can be slower than $\sqrt{n}$. We present regularity conditions and main theorems in this section and give all the proofs in the Supplemental Material.

For the simplicity of notation, we focus on model (2.2) without covariates Z, i.e. the linear transformation model (2.5), and the results can be similarly extended to the general case with additional regularity conditions on Z (see Remark II.11). Recall that we have set $\gamma(\cdot) = \log \alpha(\cdot)$ and $g(\cdot) = \log q(\cdot)$ to ensure the positivity of $\alpha(\cdot)$ and $q(\cdot)$ in (2.5). Then we reformulate the ODE model as follows,

$$\begin{cases} \Lambda'(t) = \exp\big(x^T\beta + \gamma(t) + g(\Lambda(t))\big) \\ \Lambda(0) = 0 \end{cases} . \tag{2.11}$$

Note that the parameter β is identifiable when time-varying effects are considered, but in (2.11) it is identifiable only up to a scaling factor when both γ and g are unknown as shown in Proposition 2.2.1. To guarantee the identifiability, we constrain the first element of β to be 1 and $\gamma(t^*) = c$ with some constant c for simplicity in optimization. Specifically, denote $X = (X_{(1)}, X_{(-1)})$, $\beta = (1, \bar{\beta}^T)^T$, $\bar{\gamma}(\cdot) = \gamma(\cdot) - \gamma(t^*)$ with $\bar{\gamma}(t^*) \equiv 0$, and $\bar{X}_{(1)} = X_{(1)} + \gamma(t^*)$, then we have $X^T\beta + \gamma(t) = \bar{X}_{(1)} + X_{(-1)}^T\bar{\beta} + \bar{\gamma}(t)$. We substitute $\bar{\beta}$, $\bar{\gamma}$, and $\bar{X}_{(1)}$ by β, γ, and $X_{(1)}$ respectively for notational simplicity hereafter, and the ODE (2.11) is then equivalent to

$$\begin{cases} \Lambda'(t) = \exp\Big(x_{(1)} + x_{(-1)}^T\beta + \gamma(t) + g(\Lambda(t))\Big) \\ \Lambda(0) = 0 \end{cases} , \tag{2.12}$$

with $\gamma(t^*) \equiv 0$. Before stating the regularity conditions, we first introduce some notations. We denote the solution of (2.12) by $\Lambda(t, x, \beta, \gamma, g)$ to explicitly indicate that the solution of (2.12) depends on covariates x and parameters (β, γ, g). We denote the true parameters associated with the data generating distribution by (β_0, γ_0, g_0) and simplify $\Lambda(t, x, \beta_0, \gamma_0, g_0)$ as $\Lambda_0(t, x)$. In addition, some commonly used notations in the empirical process literature will be used in this section as well. Let $Pf = \int f(x)Pr(dx)$, where Pr is a probability measure, and denote the empirical probability measure as $\mathbb{P}_n$.

Then we assume the following regularity conditions.

(C1) The true parameter β_0 is an interior point of a compact set $\mathcal{B} \subset \mathbf{R}^d$.

(C2) The density of X is bounded below by a constant $c > 0$ over its domain $\mathcal{X}$, which is a compact subset of $\mathbf{R}^{d+1}$, and $P(X_{(-1)}X_{(-1)}^T)$ is nonsingular.

(C3) There exists a truncation time $\tau < \infty$ such that, for some positive constant δ_0, $Pr(Y > \tau|X) \geq \delta_0$ almost surely with respect to the probability measure of X. Then there is a constant $\mu = \sup_{x \in \mathcal{X}} \Lambda_0(\tau, x) \leq -\log \delta_0$ such that $\Lambda_0(\tau, X) = -\log Pr(T > \tau|X) \leq \mu$ almost surely with respect to the probability measure of X.

(C4) Let $S^p([a,b])$ be the collection of bounded functions f on $[a,b]$ with bounded derivatives $f^{(j)}$, $j = 1, \ldots, k$, where the kth derivative $f^{(k)}$ satisfies the m-Hölder continuity condition:

$$|f^{(k)}(s) - f^{(k)}(t)| \leq L|s-t|^m \qquad \text{for } s, t \in [a,b],$$

where k is a positive integer and $m \in (0,1]$ with $p = m + k$, and $L < \infty$ is a constant. The true function $\gamma_0(\cdot)$ belongs to $\Gamma^{p_1} = \{\gamma \in S^{p_1}([0,\tau]) : \gamma(t^*) = 0\}$ with $p_1 \geq 2$ and the true function $g_0(\cdot)$ belongs to $S^{p_2}([0, \mu + \delta_1]) = \mathcal{G}^{p_2}$ with some positive constant δ_1 and $p_2 \geq 3$.

(C5) Denote $R(t) = \int_0^t \exp(\gamma_0(s))ds$, $V = X_{(1)} + X_{(-1)}^T\beta_0$, and $U = e^V R(Y)$. There exists $\eta_1 \in (0,1)$ such that for all $u \in \mathbf{R}^d$ with $\|u\| = 1$,

$$u^T Var(X_{(-1)} \mid U, V)u \geq \eta_1 u^T P(X_{(-1)}X_{(-1)}^T \mid U, V)u \quad \text{almost surely.}$$

(C6) Let $\psi(t, x, \beta, \gamma, g) = x_{(1)} + x_{(-1)}^T\beta + \gamma(t) + g(\Lambda(t, x, \beta, \gamma, g))$ and denote its functional derivatives with respect to $\gamma(\cdot)$ and $g(\cdot)$ along the direction $v(\cdot)$ and $w(\cdot)$ at the true parameter by $\psi'_{0\gamma}(t,x)[v]$ and $\psi'_{0g}(t,x)[w]$ respectively, whose rigorous definitions are given by (A.19)-(A.20) in the Supplemental Material. For any $v(\cdot) \in \Gamma^{p_1}$ and $w(\cdot) \in \mathcal{G}^{p_2}$, there exists $\eta_2 \in (0,1)$ such that

$$\begin{aligned}&(P\{\psi'_{0\gamma}(Y,X)[v]\psi'_{0g}(Y,X)[w] \mid \Delta = 1\})^2\\ \leq&\eta_2 P\{(\psi'_{0\gamma}(Y,X)[v])^2 \mid \Delta = 1\} P\{(\psi'_{0g}(Y,X)[w])^2 \mid \Delta = 1\}\end{aligned}$$

almost surely.

Conditions (C1)-(C3) are common regularity assumptions in survival analysis. Condition (C4) requires $p_2 \geq 3$ to control the error rates of the spline approximation for the true function g_0 and its first and second derivatives. Moreover, together with $p_1 \geq 2$, (C4) will also be used to verify the assumptions (A4)-(A6) for the general M-theorem (Theorem 2.3.3) when we apply it to derive the asymptotic normality of the proposed sieve MLE (Theorem 2.3.2). A similar condition to (C5) was imposed by Wellner and Zhang (2007) for the panel count data, by Ding and Nan (2011) for the linear transformation model with a known transformation, and by Zhao et al. (2017) for the accelerated hazards model. When the transformation function is known, condition (C5) is equivalent to the assumption C7 in Ding and Nan (2011) and can be verified in many applications as shown in Wellner and Zhang (2007). For the general case where both the transformation function and the error distribution are unspecified, condition (C6) is assumed to avoid strong collinearity between $\psi'_{0\gamma}(Y,X)[v]$ and $\psi'_{0g}(Y,X)[w]$.

Note that the parameter $g(\cdot)$ takes $\Lambda(t,x,\beta,\gamma,g)$ as its argument in (2.12), which involves the other parameters β and $\gamma(\cdot)$. Thus, β, $\gamma(\cdot)$ and $g(\cdot)$ are bundled parameters. For any $g(\cdot) \in \mathcal{G}^{p_2}$, we directly consider the composite function $g(\Lambda(t,x,\beta,\gamma,g))$

as a function from $\mathcal{T} \times \mathcal{X} \times \mathcal{B} \times \Gamma^{p_1}$ to $\mathbf{R}$. And we define the collection of functions

$$\mathcal{H}^{p_2} = \{\zeta(\cdot, \beta, \gamma) : \zeta(t, x, \beta, \gamma) = g(\Lambda(t, x, \beta, \gamma, g)), t \in [0, \tau], x \in \mathcal{X}, \beta \in \mathcal{B}, \gamma \in \Gamma^{p_1}, \\ g \in \mathcal{G}^{p_2} \text{ such that } \sup_{t \in [0,\tau], x \in \mathcal{X}} |\Lambda(t, x, \beta, \gamma, g)| \leq \mu + \delta_1\},$$

with δ_1 given in condition (C4). For any $\zeta(\cdot, \beta, \gamma) \in \mathcal{H}^{p_2}$, we define its norm as

$$\|\zeta(\cdot, \beta, \gamma)\|_2 = \left[\int_{\mathcal{X}} \int_0^{\tau} [\zeta(t, x, \beta, \gamma)]^2 d\Lambda_0(t, x) dF_X(x) \right]^{1/2},$$

where $F_X(x)$ is the cumulative distribution function of X. Denote the parameter $\theta = (\beta, \gamma(\cdot), \zeta(\cdot, \beta, \gamma))$ and the true parameter $\theta_0 = (\beta_0, \gamma_0(\cdot), \zeta_0(\cdot, \beta_0, \gamma_0))$ with $\zeta_0(t, x, \beta_0, \gamma_0) = g_0(\Lambda(t, x, \beta_0, \gamma_0, g_0))$. Denote the parameter space by $\Theta = \mathcal{B} \times \Gamma^{p_1} \times \mathcal{H}^{p_2}$. For any θ_1 and θ_2 in Θ, we define the distance

$$d(\theta_1, \theta_2) = \left(\|\beta_1 - \beta_2\|^2 + \|\gamma_1 - \gamma_2\|_2^2 + \|\zeta_1(\cdot, \beta_1, \gamma_1) - \zeta_2(\cdot, \beta_2, \gamma_2)\|_2^2 \right)^{1/2},$$

where $\|\cdot\|$ is the Euclidean norm and $\|\gamma\|_2 = (\int_0^{\tau} (\gamma(t))^2 dt)^{1/2}$ is the L_2 norm.

Next, we construct the sieve space as follows. Let $0 = t_0 < t_1 < \cdots < t_{K_n^1} < t_{K_n^1+1} = \tau$ be a partition of $[0, \tau]$ with $K_n^1 = O(n^{\nu_1})$ and $\max_{1 \leq j \leq K_n^1+1} |t_j - t_{j-1}| = O(n^{-\nu_1})$ for some $\nu_1 \in (0, 0.5)$. Let $T_{K_n^1} = \{t_1, \cdots, t_{K_n^1}\}$ denote the set of partition points and $S_n(T_{K_n^1}, K_n^1, p_1)$ be the space of polynomial splines of order p_1 as defined in Schumaker (2007, page 108, Definition 4.1). Similarly, let $T_{K_n^2}$ be a set of partition points of $[0, \mu]$ with $K_n^2 = O(n^{\nu_2})$ and $\max_{1 \leq j \leq K_n^2+1} |t_j - t_{j-1}| = O(n^{-\nu_2})$ for some $\nu_2 \in (0, 0.5)$, and $S_n(T_{K_n^2}, K_n^2, p_2)$ be the space of polynomial splines of order p_2. According to Schumaker (2007, page 117, Corollary 4.10), there exist two sets of B-spline bases $\{B_j^1, 1 \leq j \leq q_n^1\}$ with $q_n^1 = K_n^1 + p_1$ and $\{B_j^2, 1 \leq j \leq q_n^2\}$ with $q_n^2 = K_n^2 + p_2$ such that for any $s_1 \in S_n(T_{K_n^1}, K_n^1, p_1)$ and $s_2 \in S_n(T_{K_n^2}, K_n^2, p_2)$, we can write $s_1(t) = \sum_{j=1}^{q_n^1} a_j B_j^1(t)$ and $s_2(t) = \sum_{j=1}^{q_n^2} b_j B_j^2(t)$. Let $\Gamma_n^{p_1} = \{\gamma \in S_n(T_{K_n^1}, K_n^1, p_1) : \gamma(0) = 0\}$, $\mathcal{G}_n^{p_2} = S_n(T_{K_n^2}, K_n^2, p_2)$, and

$$\begin{aligned}\mathcal{H}_n^{p_2} = \{&\zeta(\cdot, \beta, \gamma) : \\ &\zeta(t, x, \beta, \gamma) = g(\Lambda(t, x, \beta, \gamma, g)), g \in \mathcal{G}_n^{p_2}, t \in [0, \tau], x \in \mathcal{X}, \beta \in \mathcal{B}, \gamma \in \Gamma_n^{p_1}\}.\end{aligned}$$

Let $\Theta_n = \mathcal{B} \times \Gamma_n^{p_1} \times \mathcal{H}_n^{p_2}$ be the sieve space. It is not difficult to see that $\Theta_n \subset \Theta_{n+1} \subset \cdots \subset \Theta$. We consider the sieve estimator $\hat{\theta}_n = (\hat{\beta}_n, \hat{\gamma}_n(\cdot), \hat{\zeta}_n(\cdot, \hat{\beta}_n, \hat{\gamma}_n))$, where $\hat{\zeta}_n(t, x, \hat{\beta}_n, \hat{\gamma}_n) = \hat{g}_n(\Lambda(t, x, \hat{\beta}_n, \hat{\gamma}_n, \hat{g}_n))$, that maximizes the log-likelihood (2.6) (without covariates Z and parameter $\boldsymbol{\eta}$) over the sieve space Θ_n. The consistency and convergence rate of the sieve MLE $\hat{\theta}_n$ are then established in the following theorem.

Theorem 2.3.1. *(Convergence rate of $\hat{\theta}_n$.) Let ν_1 and ν_2 satisfy the restrictions* $\max\{\frac{1}{2(2+p_1)}, \frac{1}{2p_1} - \frac{\nu_2}{p_1}\} < \nu_1 < \frac{1}{2p_1}$, $\max\{\frac{1}{2(1+p_2)}, \frac{1}{2(p_2-1)} - \frac{2\nu_1}{p_2-1}\} < \nu_2 < \frac{1}{2p_2}$, *and* $2\min\{2\nu_1, \nu_2\} > \max\{\nu_1, \nu_2\}$. *Suppose conditions (C1)-(C6) hold, then we have*

$$d(\hat{\theta}_n, \theta_0) = O_p(n^{-\min\{p_1\nu_1, p_2\nu_2, \frac{1-\max\{\nu_1,\nu_2\}}{2}\}}).$$

Theorem 2.3.1 gives the convergence rate of the proposed estimator $\hat{\theta}_n$ to the true parameter θ_0, and its proof is provided in the Supplemental Material by verifying the conditions in Shen and Wong (1994, Theorem 1). Note the subscripts 1 and 2 correspond to the space of the spline approximation for two infinite-dimensional parameters γ and g, respectively. The restrictions on ν_1 and ν_2 are feasible for p_1 and p_2 not far away from each other. For example, if $p_1 = p_2 = p$ and $\nu_1 = \nu_2 = \nu$, the restriction on ν is equivalent to $\frac{1}{2(1+p)} < v < \frac{1}{2p}$, and the convergence rate becomes $d(\hat{\theta}_n, \theta_0) = O_p(n^{-\min\{p\nu, \frac{1-\nu}{2}\}})$, which is the same as the case when there is only one infinite-dimensional parameter (Ding and Nan, 2011; Zhao et al., 2017). Further, if $\nu = \frac{1}{1+2p}$, we have $d(\hat{\theta}_n, \theta_0) = O_p(n^{-\frac{p}{1+2p}})$, which achieves the optimal convergence rate in the nonparametric regression setting.

Although the convergence rate for the nuisance parameter is slower than the typical rate $n^{1/2}$, we will show that the sieve MLE of the regression parameter, i.e. $\hat{\beta}_n$, is still asymptotically normal and achieves the semi-parametric efficiency bound. First, we introduce two additional regularity conditions which are stated below.

(C7) There exist $\mathbf{v}^* = (v_1^*, \cdots, v_d^*)^T$ and $\mathbf{w}^* = (w_1^*, \cdots, w_d^*)^T$, where $v_j^* \in \Gamma^2$ and $w_j^* \in \mathcal{G}^2$ for $j = 1, \cdots, d$, such that $P\{\Delta \mathbf{A}^*(U, X)\psi'_{0\gamma}(Y, X)[v]\} = 0$ and $P\{\Delta \mathbf{A}^*(U, X)\psi'_{0g}(Y, X)[w]\} = 0$ hold for any $v \in \Gamma^{p_1}$ and $w \in \mathcal{G}^{p_2}$. Here U and V are defined the same as in condition (C5) and

$$\begin{aligned}\mathbf{A}^*(t, X) = &- \Big(g_0'(\tilde{\Lambda}_0(t)) \exp\Big(g_0(\tilde{\Lambda}_0(t))\Big)t + 1\Big) X_{(-1)} \\ &+ g_0'(\tilde{\Lambda}_0(t)) \exp\Big(g_0(\tilde{\Lambda}_0(t))\Big) \int_0^t \mathbf{v}^*(R^{-1}(se^{-V}))\,ds + \mathbf{v}^*(R^{-1}(te^{-V})) \\ &+ g_0'(\tilde{\Lambda}_0(t)) \exp\Big(g_0(\tilde{\Lambda}_0(t))\Big) \int_0^{\tilde{\Lambda}_0(t)} \exp(-g_0(s))\mathbf{w}^*(s)\,ds + \mathbf{w}^*(\tilde{\Lambda}_0(t)),\end{aligned}$$

where $\tilde{\Lambda}_0(t)$ is the solution of $\tilde{\Lambda}_0'(t) = \exp\Big(g_0(\tilde{\Lambda}_0)\Big)$ with $\tilde{\Lambda}_0(0) = 0$.

(C8) Let $\boldsymbol{l}^*(\beta_0, \gamma_0, \zeta_0; W) = \int \mathbf{A}^*(t, X)\,dM(t)$, where $M(t) = \Delta \mathbb{1}(U \leq t) - \int_0^t \mathbb{1}(U \geq s)\,d\tilde{\Lambda}_0(s)$ is the event counting process martingale. The information matrix $I(\beta_0) = P(\boldsymbol{l}^*(\beta_0, \gamma_0, \zeta_0; W)^{\otimes 2})$ is nonsingular. Here for a vector a, $a^{\otimes 2} = aa^T$.

The additional condition (C7) essentially requires the existence of the least favorable direction that is used to establish the semi-parametric efficiency bound. The directions $\mathbf{v}^*$ and $\mathbf{w}^*$ may be found through the equations in (C7). We illustrate how to construct $\mathbf{v}^*$ and $\mathbf{w}^*$ for the Cox model and the linear transformation model with a known transformation respectively in Remark II.10. Condition (C8) is a natural assumption that requires the information matrix to be invertible. The following theorem establishes the asymptotic normality and semi-parametric efficiency of the sieve MLE $\hat{\beta}_n$ of the regression parameter for the general linear transformation model.

Theorem 2.3.2. *(Asymptotic normality of $\hat{\beta}_n$) Suppose the conditions in Theorem 2.3.1 and (C7)-(C8) hold, then we have*

$$\sqrt{n}(\hat{\beta}_n - \beta_0) = \sqrt{n} I^{-1}(\beta_0) \mathbb{P}_n \boldsymbol{l}^*(\beta_0, \gamma_0, \zeta_0; W) + o_p(1) \to_d N(0, I^{-1}(\beta_0))$$

with $I(\beta_0)$ given in condition (C8) and $\to_d$ denoting convergence in distribution.

Theorem 2.3.2 states that $\hat{\beta}_n$ is asymptotically normal with variance as the inverse of the information matrix. In practice, the information matrix can be approximated by the estimated information matrix of all parameters including the coefficients of spline bases.

We note that the existing sieve M-theorem for bundled parameters (Ding and Nan, 2011; Zhao et al., 2017) cannot be directly applied to prove Theorem 2.3.2, because it does not allow the infinite-dimensional nuance parameter to be a function of other infinite-dimensional nuance parameters. Therefore, to study the asymptotic distribution of $\hat{\beta}_n$, we first establish a new general M-theorem for bundled parameters where the infinite-dimensional nuisance parameter is a function of not only the Euclidean parameter of interest but also other infinite-dimensional nuisance parameters. The established M-theorem under such a general scenario then enables us to prove Theorem 2.3.2 by verifying its assumptions for the linear transformation model. The details are provided in the Supplemental Material. Since the new M-theorem can be useful for developing the asymptotic normality of sieve estimators for other ODE models, we state it below for readers of interest.

We first introduce the general setting and notation for the proposed sieve M-theorem. Let $m(\theta; W)$ be an objective function of unknown parameters $\theta = (\beta, \boldsymbol{\gamma}(\cdot), \zeta(\cdot, \beta, \boldsymbol{\gamma}))$ given a single observation W, where β is the finite-dimensional parameter of interest, $\boldsymbol{\gamma}(\cdot) = (\gamma_1(\cdot), \ldots, \gamma_{d_2}(\cdot))$ denotes infinite-dimensional nuisance parameters, and $\zeta(\cdot, \beta, \boldsymbol{\gamma})$ is another infinite-dimensional nuisance parameter that can be a function of β and $\boldsymbol{\gamma}$. Here "$\cdot$" represents some components of W. Given i.i.d. observations $\{W_i\}_{i=1}^n$, the sieve estimator $\hat{\theta}_n = (\hat{\beta}_n, \hat{\boldsymbol{\gamma}}_n(\cdot), \hat{\zeta}_n(\cdot, \hat{\beta}_n, \hat{\boldsymbol{\gamma}}_n))$ maximizes the objective function, $\mathbb{P}_n m(\theta; W)$, over certain sieve space. For example, $\hat{\theta}_n$ becomes the sieve MLE if m is the log-likelihood function. We denote the derivative of m with respect to β as m'_β, the functional derivative of m with respect to γ_j along the direction $v(\cdot)$ as $m'_{\gamma_j}[v]$ for $1 \le j \le d_2$, and the functional derivative of m with respect to ζ along the direction $h(\cdot)$ as $m'_\zeta[h]$, whose rigorous definitions are given in the Supplemental Material. The following theorem then establishes the asymptotic normality of the sieve estimator, $\hat{\beta}_n$, under the above general setting.

Theorem 2.3.3. *(A general M-theorem for bundled parameters.) Under assumptions (A1)-(A6) in the Supplemental Material, we have*

$$\begin{aligned}\sqrt{n}(\hat{\beta}_n - \beta_0) &= A^{-1}\sqrt{n}\mathbb{P}_n \boldsymbol{m}^*(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0); W) + o_p(1)\\ &\to_d N(0, A^{-1}B(A^{-1})^T),\end{aligned}$$

where

$$\begin{aligned}\boldsymbol{m}^*(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0); W) = {}& m'_\beta(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0); W)\\ &- \sum_{j=1}^{d_2} m'_{\gamma_j}(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0); W)[\boldsymbol{v}_j^*]\\ &- m'_\zeta(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0); W)[\boldsymbol{h}^*(\cdot, \beta_0, \boldsymbol{\gamma}_0)],\end{aligned}$$

$$B = P\{\boldsymbol{m}^*(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0); W)\boldsymbol{m}^*(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0); W)^T\},$$

with $\boldsymbol{v}_j^* = (v_{j1}^*, \ldots, v_{jd_1}^*)^T$, $\boldsymbol{h}^* = (h_1^*, \ldots, h_d^*)^T$ and A given in the assumption (A3).

Remark II.9. *The assumptions needed in Theorem 2.3.3 are similar to those in Ding* *and Nan (2011) (see the Supplemental Material for details). However, our proposed theorem significantly differs from the main theorem in Ding and Nan (2011), because the latter considers $\zeta(\cdot, \beta)$ to be a function of only the finite-dimensional parameter β, while we consider a more general scenario of bundled parameters, where the nuisance parameter $\zeta(\cdot, \beta, \boldsymbol{\gamma})$ can be a function of both the finite-dimensional parameter β and other infinite-dimensional nuisance parameters $\boldsymbol{\gamma}$. The proposed theorem nontrivially extends the asymptotic distributional theories for M-estimation under this general scenario.*

Remark II.10. *We note that to find the least favorable directions $\boldsymbol{v}^*$ and $\boldsymbol{w}^*$ required in (C7), we may solve the equations in (C7), which can be simplified to equations* (A.*41) and* (A.*43) provided in the Supplemental Material. For illustration, we provide explicit constructions of the least favorable directions for the Cox model and for the linear transformation model with a known transformation respectively. Specifically, for the Cox model, we have $g_0 \equiv 0$ and $\boldsymbol{v}^*$ can be derived as*

$$\boldsymbol{v}^*(t) = \frac{P\{\mathbb{1}(Y \geq t) e^{X^T \beta_0} X\}}{P\{\mathbb{1}(Y \geq t) e^{X^T \beta_0}\}};$$

for the linear transformation model where γ_0 is known, $\boldsymbol{w}^*$ can be obtained as

$$\boldsymbol{w}^*(t) = \boldsymbol{\phi}(t) - g_0'(t) \int_0^t \boldsymbol{\phi}(s) ds,$$

where

$$\boldsymbol{\phi}(t) = \Big(g_0'(t) \exp(g_0(t)) \tilde{\Lambda}_0^{-1}(t) + 1 \Big) \frac{P\{\mathbb{1}(\Lambda_0(Y, X) \geq t) X\}}{P\{\mathbb{1}(\Lambda_0(Y, X) \geq t)\}}$$

with $\tilde{\Lambda}_0$ defined in (C7).

Given the above constructions of the least favorable directions, we can further simplify the non-singularity condition of the information matrix in (C8). For the

Cox model, the information matrix can be derived as

$$I(\beta_0) = \int\limits_0^\infty P\left([-X + \boldsymbol{\mu}(t)]^{\otimes 2}\, \mathbb{1}(U \geq t)\right) dt,$$

where $\boldsymbol{\mu}(t) = P\{\mathbb{1}(U \geq t)e^{X^T\beta_0}X\}/P\{\mathbb{1}(U \geq t)e^{X^T\beta_0}\}$ *with* U *defined in (C5). Respectively, for the linear transformation where* γ_0 *is known, the information matrix can be derived as*

$$I(\beta_0) = \int\limits_0^\infty m^2(t) \cdot Var(X|U \geq t) \cdot P(U \geq t) \cdot \exp\Big(g_0(\tilde{\Lambda}_0(t))\Big)\, dt,$$

where $m(t) = g_0'(\tilde{\Lambda}_0(t)) \exp\Big(g_0(\tilde{\Lambda}_0(t))\Big)t + 1$*. The non-singularity condition requires the integral of a covariance matrix to be positive definite.*

Remark II.11. *Moreover, for the general class of ODE models that include covariates* Z *with time-varying coefficients* $\boldsymbol{\eta}(\cdot)$ *in (2.2), we have further established the same convergence rate of the sieve estimator* $\hat{\theta}_n$ *in Theorem A.5.1 and the asymptotic normality of* $\hat{\beta}_n$ *in Theorem A.5.2 in the Supplemental Material. In particular, the conditions (C1)-(C8) have been revised to (C1′)-(C8′) with additional regularity conditions on covariates* Z*. We refer to the Supplemental Material for the full list of conditions, rigorous statements of theorems, and their proofs.*

2.3.3 Simulation Studies

In this section, we use simulation studies to show the finite sample performance of the sieve MLE under the time-varying Cox model and the general linear transformation model.

Time-varying Cox model We generate event times from the model

$$\Lambda_x'(t) = \alpha(t)\exp(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \eta(t)x_5),$$

where $(x_1, x_2, x_3, x_4, x_5)$ follows a multivariate normal distribution with mean 0 and autoregressive covariance truncated at ± 2, $\beta_1 = \beta_4 = 1$, and $\beta_2 = \beta_3 = -1$. Let $\eta(t) = \sin\left(\frac{3}{4}\pi t\right)$ be a time-varying coefficient for x_5 and the coefficients of all other covariates be time-independent. The baseline hazard $\alpha(t)$ is set to 0.5. The censoring times are generated from an independent uniform distribution $U(0, 3)$, which leads to a censoring rate around 50%. The sample size N varies from $1,000$ to $8,000$. We fit both the log-transformed baseline hazard function $\log\alpha(t)$ and time-varying coefficient $\eta(t)$ by cubic B-splines and set the number of knots $K_n = \lfloor N'^{\frac{1}{5}} \rfloor$, i.e., the largest integer smaller than $N'^{\frac{1}{5}}$, where N' is the number of distinct observation time points. The interior knots are located at the K_n quantiles of the N' distinct observation time points. We compare the estimation accuracy and the computing time of the proposed sieve MLE with those of the partial likelihood-based estimator implemented in the "coxph" function in R with the "tt" argument set as the same cubic B-spline transformation of time.

Table II.1 summarizes the estimates of regression coefficients β_1 and β_2 based on 1000 replications. The estimates of the other two regression coefficients β_3 and β_4 perform similarly, and the results are included in the Supplemental Material. For the time-varying coefficient $\eta(t)$, we report the integrated mean square error (IMSE), which is the weighted sum of mean square error (MSE) of pointwise estimates over simulated time points from 0 to 2. As one can see, the mean and standard deviation of IMSE of the proposed sieve estimator decrease as the sample size increases. Remarkably, they are consistently smaller than those of the partial likelihood-based estimator. For time-independent coefficients, the proposed sieve estimator performs

Table II.1: Simulation results under time-varying Cox model.

N	Method	$\beta_1 = 1$				$\beta_2 = -1$				IMSE($\eta(t)$)	
		Bias	SE	ESE	CP	Bias	SE	ESE	CP	Mean	SD
1000	ODE	.008	.070	.070	.958	-.012	.076	.078	.955	.053	.041
	Cox-MPLE	.006	.070	.068	.952	-.010	.075	.075	.950	.109	.094
2000	ODE	.004	.048	.048	.958	-.004	.053	.054	.957	.029	.021
	Cox-MPLE	.002	.048	.048	.956	-.003	.053	.053	.959	.053	.041
4000	ODE	.003	.033	.034	.952	-.003	.038	.038	.938	.016	.011
	Cox-MPLE	.003	.033	.034	.950	-.002	.038	.037	.936	.026	.020
8000	ODE	.000	.024	.024	.962	-.001	.026	.026	.938	.009	.006
	Cox-MPLE	.000	.023	.024	.959	-.001	.026	.026	.936	.013	.009

Bias is the difference between the mean of estimates and the true value, SE is the sample standard error of the estimates, Mean is the mean of IMSE, and SD is the standard deviation of IMSE. ESE is the mean of the standard error estimators by inverting the estimated information matrix of all parameters, including the coefficients of spline bases, and CP is the corresponding coverage proportion of 95% confidence intervals.

as well as the partial likelihood-based estimator. The mean of the standard error estimator, which is obtained by inverting the estimated information matrix of all parameters including the coefficients of spline bases, is approximate to the sample standard error, and the corresponding 95% confidence interval achieves a proper coverage proportion. From the left and middle panels of Figure II.1, we can see that the means of the estimated $\alpha(t)$ and $\eta(t)$ are close to the true functions, and the 95% pointwise confidence bands cover the true functions well.

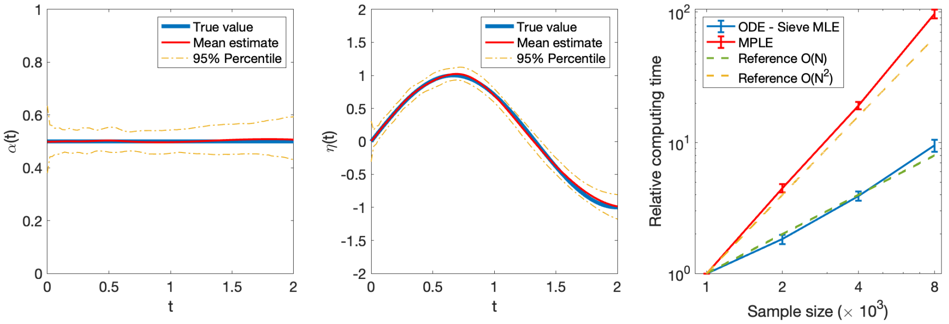

Figure II.1: True $\alpha_0(t)$ and mean of $\hat{\alpha}(t)$ (left); true $\eta(t)$ and mean of $\hat{\eta}(t)$ (middle) with the sample size $N = 8000$; log-log plot of mean relative computation time (right) with respect to the sample size under the time-varying Cox model.

It is also worth noting that, in comparison to the partial likelihood-based estimation method whose relative computing time with respect to that with the smallest sample size increases quickly as the sample size grows, the proposed estimation method is computationally more efficient, especially when the sample size is large (see the right panel of Figure II.1). When the number of knots increases with the sample size, the computation time of the proposed method grows at a rate slightly larger than the linear rate (but far below the quadratic rate).

Linear transformation model We generate event times from the model $\Lambda_x'(t) = q(\Lambda_x(t)) \exp(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)\alpha(t)$. The covariates are independent normal with mean 0 and standard deviation 0.5 truncated at ± 2. We consider four different settings for $q(\cdot)$ and $\alpha(\cdot)$: 1) a constant $q(t) = 1$ and a monotonic increasing $\alpha(t) = t^3$, in which case the Cox model is correctly specified; 2) a monotonic decreasing $q(t) = e^{-t}$ and a constant $\alpha(t) = 2$; 3) a monotonic decreasing $q(t) = 2/(1+t)$ and a constant $\alpha(t) = 1$; 4) an increasing $q(t) = \log(1+t) + 2$ and an increasing $\alpha(t) = log(1+t)$. In each setting, we generate the censoring time from an independent uniform distribution $U(0, c)$, where c is chosen to achieve approximately 25-30% censoring rates. The sample size N varies from $1,000$ to $8,000$.

In setting 1), we compare the proposed sieve MLE for the ODE-Cox model, where the function $q(\cdot)$ is set to 1, with the partial-likelihood based estimator implemented using the R package *survival*. We fit $\log \alpha(\cdot)$ by cubic B-splines with $\lfloor N'^{\frac{1}{5}} \rfloor$ interior knots that are located at the quantiles of the distinct observation time points. In setting 2), we compare the proposed sieve MLE for the ODE-LT model, where the function $q(\cdot)$ is set to e^{-t}, with the NPMLE for the equivalent logarithmic transformation model considered in Zeng and Lin (2007b). We fit $\log \alpha(\cdot)$ by cubic B-splines with the same placement of interior knots. In setting 3), we compare the proposed sieve MLE for the ODE-AFT model, where the function α is set to 1, with the rank-

Table II.2: Estimates of regression coefficients under correctly-specified ODE-Cox with $q(\cdot) \equiv 1$, ODE-LT with $q(t) = e^{-t}$, and ODE-AFT with $\alpha(\cdot) \equiv 1$. Bias, SE, ESE and CP contain the same meanings as those in Table II.1.

		$\beta_1 = 1$				$\beta_2 = 1$				$\beta_3 = 1$			
	Method	Bias	SE	ESE	CP	Bias	SE	ESE	CP	Bias	SE	ESE	CP
1)	MPLE	.002	.076	.075	.934	-.003	.075	.075	.941	-.001	.074	.075	.954
	ODE-Cox	.003	.076	.076	.936	-.002	.075	.076	.942	.000	.074	.076	.955
2)	NPMLE	.004	.117	.115	.949	-.001	.114	.115	.951	.003	.113	.115	.960
	ODE-LT	.005	.117	.115	.950	-.000	.114	.115	.951	.003	.113	.115	.961
3)	Rank-based	.004	.105	.102	.944	-.001	.102	.102	.950	.002	.100	.103	.954
	ODE-AFT	.000	.102	.097	.944	-.005	.100	.097	.944	-.002	.097	.097	.950

Setting 1): the Cox model is correctly specified. Setting 2): the logarithmic transformation model is correctly specified. Setting 3): the AFT model is correctly specified.

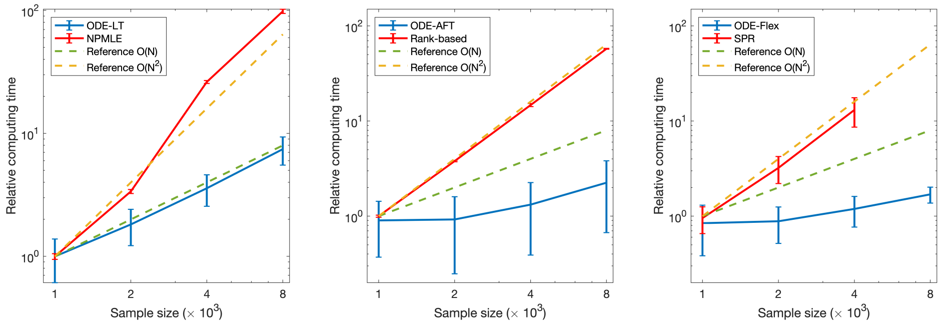

Figure II.2: The log-log plots of mean relative computing time with respect to the sample size under the ODE-LT, the ODE-AFT model, and the ODE-Flex model are provided from left to right respectively.

based estimation approach implemented using the R package *aftgee*. We fit $\log q(t)$ by cubic B-splines with $\lfloor N^{\frac{1}{7}} \rfloor$ interior knots that are located at the quantiles of the estimated cumulative hazards under the Cox model. In setting 4) (as well as settings 1)-3)), we fit the general linear transformation model (ODE-Flex) where both $q(\cdot)$ and $\alpha(\cdot)$ are unspecified, and compare the sieve MLE with the smoothed partial rank (SPR) method in Song et al. (2006). Both methods constrain $\beta_1 = 1$ for identifiability guarantee. For the sake of space, the results of the setting 4) are provided in the Supplemental Material.

Tables II.2 and II.3 summarize the estimates of regression coefficients with the

Table II.3: Estimates of regression coefficients under the general linear transformation model ODE-Flex with both $q(\cdot)$ and $\alpha(\cdot)$ unspecified. Bias, SE, ESE and CP contain the same meanings as those in Table II.1.

	$\beta_2 = 1$				$\beta_3 = 1$			
Setting	Bias	SE	ESE	CP	Bias	SE	ESE	CP
1)	.008	.106	.107	.947	.012	.104	.107	.959
2)	-.019	.161	.151	.927	-.016	.159	.151	.938
3)	-.014	.134	.131	.941	-.012	.131	.132	.945
4)	.001	.092	.090	.939	.005	.091	.090	.954

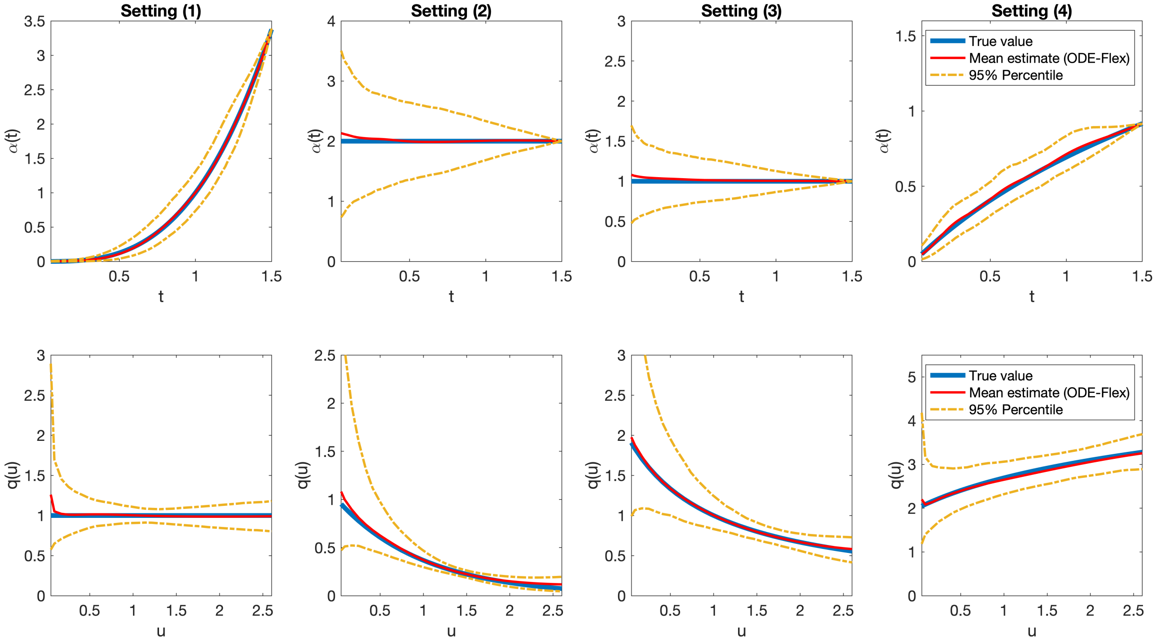

Figure II.3: The solid blue curves are the true $q(\cdot)$ (upper row) and $\alpha(\cdot)$ (lower row). The solid red curves are the means of corresponding estimated $\hat{q}(\cdot)$ and $\hat{\alpha}(\cdot)$ under the general linear transformation model. The dashed yellow curves represent 95% pointwise confidence bands over 1,000 replications. From left to right, the four columns correspond to settings (1)-(4) respectively.

sample size $N = 4,000$ based on 1000 replications. Full results for the other sample sizes are provided in the Supplemental Material. Table II.2 indicates that when any of the Cox model, the logarithmic transformation model, or the AFT model is correctly specified, the sieve estimator for the corresponding correctly specified ODE model achieves similar performance as the partial-likelihood based estimator for the Cox model, the NPMLE for the logarithmic transformation model, or the rank-based estimator for the AFT model. However, the relative computing time of the proposed ODE approach increases linearly as the sample size grows while that of the NPMLE for the logarithmic transformation model or the rank-based method for the AFT model increases in a quadratic rate as shown in Figure II.2.

For the general linear transformation model, we find that the proposed ODE-Flex method has advantages against the existing SPR method in terms of estimation accuracy, numerical stability, and computational efficiency. We refer to the Supplemental Material for detailed results and comparison with SPR. From Table II.3, we can see that the bias of the ODE-Flex estimator is nearly negligible in all settings. The standard error estimators are close to the sample standard errors, and the corresponding 95% confidence intervals achieve a reasonable coverage proportion. When the Cox model, the logarithmic transformation model, or the AFT model is correctly specified, their estimators (in Table II.2) achieve smaller standard errors than those for ODE-Flex (in Table II.3), which is expected because both $q(\cdot)$ and $\alpha(\cdot)$ are unspecified in ODE-Flex. Figure II.3 shows the mean of $\hat{\alpha}(\cdot)$ and $\hat{q}(\cdot)$ respectively. As one can see, the means of $\hat{\alpha}(\cdot)$ and $\hat{q}(\cdot)$ under the general linear transformation model are all close to the true functions. Moreover, the relative computing time of ODE-Flex increases in a much smaller rate than that of SPR as the sample size grows as shown in the right panel of Figure II.2.

Note we have also considered other alternative knots placements (see the Supplemental Material) and our numerical results suggest that knot selection does not appear critical for the proposed method.

2.3.4 Data Example

In this section, we apply the proposed method to a kidney post-transplantation mortality study. End-stage renal disease (ESRD) is one of the most deadly and costly diseases in the United States. From 2004 to 2016, ESRD incident cases increased from 345.6 to 373.4 per million people, with Medicare expenditures escalating from 18 to 35 billion dollars (Saran et al., 2017). Kidney transplantation is the renal replacement therapy for the majority of patients with ESRD. Successful kidney transplantation is associated with improved survival, improved quality of life, and health care cost savings when compared to dialysis. However, despite aggressive efforts to increase the number of donor kidneys, the demand far exceeds the supply of donor kidneys for transplantation and hence, the donor waiting list is very long. Currently about 130,000 patients are waiting for lifesaving organ transplants in the U.S., among whom 100,000 await kidney transplants and fewer than 15% of patients will receive transplants in their lifetime. To optimize the organ allocation, further research is essential to determine the risk factor associated with post-transplant mortality.

To better understand this problem, we considered the data obtained from the Organ Procurement and Transplantation Network (OPTN). There were 146,248 patients who received transplants between 1990 and 2008. Failure time (recorded in years) was defined as the time from transplantation to graft failure or death, whichever occurred first, where graft failure was considered to occur when the transplanted kidney ceased to function. Patient survival was censored 6 year post-transplant or at the end of study (2008). The median follow-up time was around 6 years and the censoring rate was 62%. Covariates included in this study were age at transplantation, race, gender, cold ischemic time, donation after cardiac death (DCD), BMI, expanded criteria donor (ECD), dialysis time, comorbidity conditions such as glomerulonephritis, polycystic kidney disease, diabetes, and hypertension. Detecting and accounting for time-varying effects are particularly important in the context of kidney transplantation, as non-proportional hazards have already been reported in the literature (Wolfe et al., 1999; He et al., 2017). Also, analyses with time-varying effects provide valuable clinical information that could be obscured otherwise.

However, existing statistical softwares become computationally infeasible when fitting a time-varying effects model on a data set as large as what we have here. Thus, to estimate the potential time-varying effects, we fit the time-varying Cox model using the proposed sieve MLE, which is computationally scalable. Specifically, based on previous studies, DCD, Polycystic, Diabetes and Hypertension are modeled with time-independent effects, and the remaining variables are estimated with time-varying effects. The time-varying effects are all implemented by cubic B-splines with 5 interior knots, which is chosen based on the Bayesian information criterion. Figure II.4 shows the estimated baseline hazard function. We can see that the post-transplant mortality is high in the short term after surgery, with a weakening association over time. Table II.4 summarizes the estimated time-independent effects, and Figure II.5 shows examples of fitted time-varying effects with 95% pointwise confidence intervals, where the standard error estimators were obtained by inverting the estimated information matrix of all parameters including time-independent coefficients and the coefficients of spline bases. As one can see, the effects of baseline age varied over time, resulting in an eventually strengthened association. Specifically, compared with the reference group (age at transplantation between 19-39), patients 40 to 49 years of age had a protective effect in the short term after transplantation. We can also see that the high cold ischemic time is a risk factor for mortality in the short run, with a weakening association over time. Thus, special care should be dedicated to improve the short-term outcome. As expected, longer waiting times on dialysis (greater than 5 years) negatively impact post-transplant survival, especially in the short run. Male gender was not significantly associated with mortality immediately after the renal transplantation but became a risk factor in the long run. As can be seen in Figure II.5, underweight shows a protective effect in the short run, and then a slightly weakening association over time, which confirms the previous finding of Lafranca et al. (2015). The results regarding high BMI should be interpreted with caution. Although higher levels of BMI in the general population are typically associated with high mortality, in chronic kidney diseases, such as patients with kidney dialysis and kidney transplantation, higher BMI has been associated with better survival, which has been labeled as reverse epidemiology (Dekker et al., 2008; Kovesdy et al., 2010). Our results show that both overweight and obesity improved survival in the short term after kidney transplantation, but obesity became a risk factor after long-term exposure. One possible explanation is that BMI is a complex marker of visceral and nonvisceral adiposity and also of nutritional status including muscle mass (Kovesdy et al., 2010), and the improved short-term outcome associated with higher BMI may be related to differential benefits by one or more of these components. Our findings indicate a need to critically reassess the role of BMI in the risk stratification of kidney transplantation. A further assessment (such as sub-group analysis) of high BMI that differentiates between visceral adiposity, nonvisceral adiposity and higher muscle mass may improve risk stratification in kidney transplant recipients. In addition, our results show that graft survival for patients with Glomerulonephritis is better than patients with other primary diseases. Regarding racial disparities, the long-term survival outcomes for African Americans continue to lag behind non-African Americans. Finally, as expected, the effect of expanded criteria donor (ECD) is not as good as optimal donor. When a sub-optimal organ becomes available, patients and physicians must decide whether to accept the offer and special care must be dedicated to improve the survival benefit.

Table II.4: Summary of estimates for time-independent effects in kidney post-transplantation mortality study

Variables	DCD	Polycystic	Diabetes	Hypertension
EST	-0.081	-0.511	0.333	-0.146
ESE	0.038	0.021	0.012	0.014
95% CI	$[-0.156, -0.007]$	$[-0.553, -0.469]$	$[\ 0.310,\ \ 0.357]$	$[-0.172, -0.119]$
p-value	0.033	< 0.001	< 0.001	< 0.001

[*] EST is the estimated time-independent effect, ESE is the estimated standard error by inverting the estimated information matrix of all parameters including the coefficients of spline basis, and CI is the confidence interval.

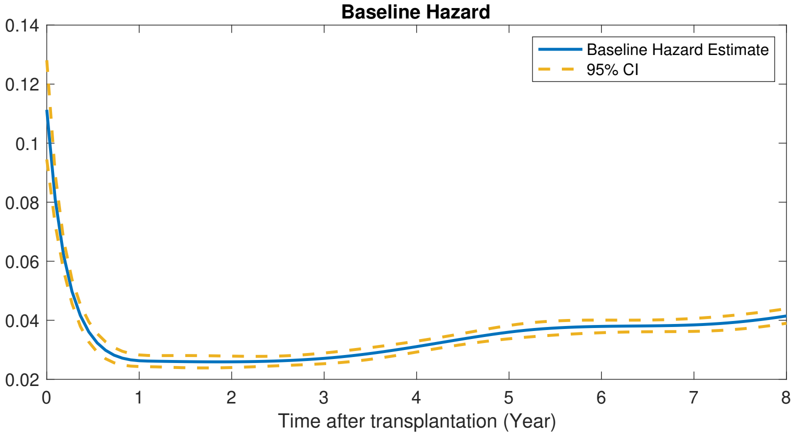

Figure II.4: Estimated baseline hazard $\hat{\alpha}(t)$ using the the proposed sieve MLE method for the kidney transplantation data.

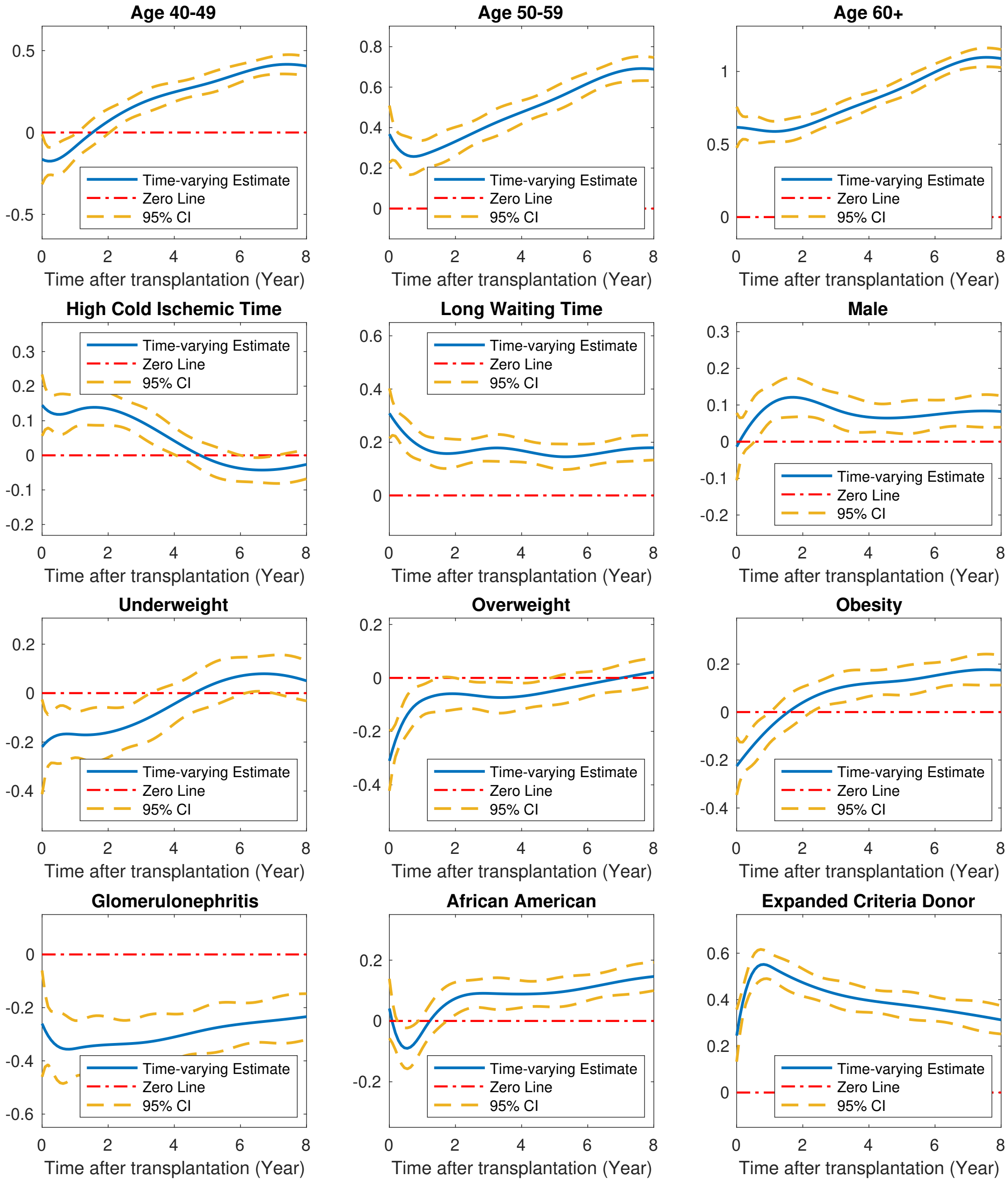

Figure II.5: Estimated time-varying effects using the proposed sieve MLE method for the kidney transplantation data.

2.4 Neural Network Extension for Powerful Prediction

Although the aforementioned models in Section 2.2 are useful, they often model the effect of features on the survival distribution in a simple, if not linear, way. These restrictions prevent the traditional models from being flexible enough to model complex data in diverse formats. To this end, various deep neural network models have been introduced into survival analysis due to their ability in automatically extracting useful features from large-scale raw data (Faraggi and Simon, 1995; Ching et al., 2018; Katzman et al., 2018; Lee et al., 2018; Gensheimer and Narasimhan, 2019; Chapfuwa et al., 2018; Kvamme et al., 2019; Steingrimsson and Morrison, 2020; Zhao, 2021).

As a natural choice for estimating a probabilistic model, likelihood-based methods have been widely used for both traditional and deep survival analysis. However, a major challenge for scalable maximum likelihood estimation of neural network models lies in difficult-to-evaluate integrals due to the existence of censoring. Specifically, for an uncensored observation i whose event time $T = t_i$ is recorded, the likelihood is the probability density function $p(t_i)$. But, for a censored observation j, only the censored time $C = t_j$ is recorded while the event time T is unknown. The likelihood of observation j is the survival function $S(t_j)$, which is the probability of no event occurring prior to t_j: $S(t_j) = \mathcal{P}\{T > t_j\} = 1 - \int_0^{t_j} p(s)ds$. This integral imposes an intrinsic difficulty for optimization: evaluating the likelihood and the gradient with respect to parameters requires the calculation of integrals, which usually has no closed forms for most flexible distribution families specified by neural networks.

To address this challenge, most existing works try to avoid the integrals in the following two ways: 1) making additional structural assumptions so that no integral is included in the objective function, such as partial-likelihood-based methods under the proportional hazard (PH) assumption (Cox, 1975), or making parametric assumption that leads to closed-form integration in the likelihood (Wei, 1992); 2) discretizing the continuous event time with pre-specified intervals so that the integral is simplified into a cumulative product. However, the structural and parametric assumptions are often restrictive and thus limit the flexibility of the model (Ng'andu, 1997; Zeng and Lin, 2007b); further, stochastic gradient descent algorithms cannot be directly applied to the partial-likelihood-based objective functions and thus limit the scalability of the model. As for discretization of the event time, it will likely cause information loss and introduce pre-specified time intervals as hyper-parameters.

In this section, we recognize that maximizing the likelihood function for censored data can be viewed as an optimization problem with differential equation (DE) constraints, and thereby tackle the aforementioned optimization challenges with an efficient numerical approach. On the basis of the unified ODE modeling framework in Section 2.2, we propose to specify $f(t, \Lambda_x(t), x) = h(\Lambda_x(t), t; x, \theta)$ in (2.1), where the function $h(\cdot, \cdot, \cdot, \theta)$ is modeled by a neural network taking the cumulative hazard $\Lambda_x(t)$, the time t, and the feature x as inputs and parameterized by θ. Since the likelihood given both uncensored and censored data can be re-written in a simple form of the hazard and the cumulative hazard, we can evaluate the likelihood function by solving the ODEs numerically. Moreover, the gradient of the likelihood with respect to θ can be efficiently calculated via adjoint sensitivity analysis, which is a general method for differentiating optimization objectives with DE constraints (Pontryagin et al., 1962; Plessix, 2006). We name the proposed method as SODEN, Survival model through Ordinary Differential Equation Networks.

In comparison to existing methods described above, the proposed SODEN is more flexible to handle event times allowing for a broad range of distributions without strong structural assumptions. Further, we directly learn a continuous-time survival model using an ODE network, which avoids potential information loss from discretizing event times. We empirically evaluate the effectiveness of SODEN through both simulation studies and experiments on real-world datasets, and demonstrate that SODEN outperforms state-of-the-art models in most scenarios.

The rest of this section is organized as follows. In Section 2.4.1, we summarize related work on deep learning survival analysis and the DE-constrained optimization. In Sections 2.4.2 and 2.4.3, we describe the proposed model and the corresponding learning approach respectively. We evaluate the proposed method using simulation studies in Section 2.4.4 and on real-world data examples in Section 2.4.5.

2.4.1 Related Work

Deep survival analysis There has been an increasing research interest on utilizing neural networks to improve feature representation in survival analysis. Earlier works (Faraggi and Simon, 1995; Ching et al., 2018; Katzman et al., 2018) are adapted from the Cox model. Recall that the Cox model (Cox, 1972) makes the proportional hazard (PH) assumption where the ratio of the hazard function is constant over time. Specifically, the hazard function consists of two terms: an unspecified baseline hazard function and a relative risk function, that is

$$\lambda_x(t) = \alpha(t)\exp(g(x;\theta)); \tag{2.13}$$

and the relative risk linearly depends on features, that is $g(x;\theta) = x^T\theta$. They adapted the Cox model to allow nonlinear dependence on features but still make the PH assumption. For example, Katzman et al. (2018) used neural networks to model the relative risk $g(x;\theta)$ in (2.13). Kvamme et al. (2019) further allowed the relative risk to vary with time, which resulted in a flexible model without the PH assumption. Specifically, they extended the relative risk as $g(t, x;\theta)$ to model interactions between features and time. These models are all trained by maximizing the partial likelihood (Cox, 1975) or its modified version, which does not need to compute the integrals included in the likelihood function. The partial likelihood function is given by

$$\mathcal{PL}(\theta; D) = \prod_{i:\Delta_i=1} \frac{\exp(g(y_i, x_i; \theta))}{\sum_{j \in R_i} \exp(g(y_i, x_j; \theta))}, \tag{2.14}$$

where $D = \{(y_i, \Delta_i, x_i) : i = 1, \cdots, N\}$, $R_i = \{j : y_j \geq y_i\}$ denotes the set of individuals who survived longer than the i^{th} individual, which is known as the *at-risk* set. Note that evaluation of the partial likelihood for an uncensored observation requires access to all other observations in the at-risk set. Hence, stochastic gradient descent (SGD) algorithms cannot be directly applied to partial likelihood-based objective functions, which is a serious limitation in training deep neural networks for large-scale applications. In the worst case, the risk set can be as large as the full data set. When the PH assumption holds, i.e., the numerators and denominators in (2.14) do not depend on y_i, evaluating the partial likelihood has a time complexity of $O(N)$ by computing $g(x_i; \theta)$ once and storing the cumulative sums. For flexible non-PH models, under which the likelihood has the form as (2.14), the time complexity further increases to $O(N^2)$. Although in practice one can naively restrict the at-risk set within each mini-batch, there is a lack of theoretical justification for this ad-hoc approach and the corresponding objective function is unclear.

On the other hand, SGD-based algorithms can be naturally applied to the original likelihood function. Following this direction, Lee et al. (2018) and Gensheimer and Narasimhan (2019) propose to discretize the continuous event time with pre-specified intervals, such that the integral in the likelihood is replaced by a cumulative product. This method scales well with large sample size and does not make strong structural assumptions. However, determining the break points for time intervals is non-trivial, since too many intervals may lead to unstable model estimation while too few intervals may cause information loss.

We note that there are works that also consider a continuous event time but they do not optimize the likelihood function. Instead, they target summary statistics of the event time distribution such as the restricted mean survival time or the survival

Model	Non-linear	No PH Assumption	Continuous-time	SGD
Cox	✗	✗	✓	?[1]
DeepSurv	✓	✗	✓	?
DeepHit	✓	✓	✗	✓
Nnet-survival	✓	✓	✗	✓
Cox-Time	✓	✓	✓	?
SODEN (proposed)	✓	✓	✓	✓

Table II.5: Comparison between the proposed method, SODEN, and related work, Cox (Cox, 1972), DeepSurv (Katzman et al., 2018), DeepHit (Lee et al., 2018), Nnet-survival (Gensheimer and Narasimhan, 2019), and Cox-Time (Kvamme et al., 2019).

probability at a fixed time point (Steingrimsson and Morrison, 2020; Zhao, 2021). We also note that (Groha et al., 2020) proposes a neural-network-based ODE approach to model the Kolmogorov forward equation that characterizes the transition probabilities for multi-state survival analysis.

The proposed SODEN is a flexible continuous-time model and is trained by maximizing the likelihood function, where SGD-based algorithms can be applied. Table II.5 summarizes the comparison between SODEN and several representative existing methods.

DE-constrained optimization DE-constrained optimization has wide and important applications in various areas, such as optimal control, inverse problems, and shape optimization (Antil and Leykekhman, 2018). One of the major contributions of this work is to recognize that the maximum likelihood estimation in survival analysis is essentially a DE-constrained optimization problem. Specifically, the maximum

[1] SGD algorithms for Cox, DeepSurv, and Cox-Time can be naively implemented in practice, but not theoretically justifiable due to the form of the objective functions.

likelihood estimation (MLE) for the proposed SODEN can be rewritten as

$$\max_{\theta} \sum_{i=1}^{N} \Delta_i \log h(\Lambda_{x_i}(y_i), y_i; x_i, \theta) - \Lambda_{x_i}(y_i) \tag{2.15}$$

$$\text{subject to } \Lambda'_{x_i}(t) = h(\Lambda_{x_i}(t), t; x_i, \theta)$$

$$\Lambda_{x_i}(0) = 0,\ i = 1, \ldots, N$$

where the constraint is a DE parameterized by θ and the objective contains the solution of the DE. Therefore, maximizing the likelihood function for censored data that contains the solution of the parameterized ODE can be viewed as an optimization problem with DE constraints as shown in (2.15)[2]. By bringing the strength of existing DE-constrained optimization techniques, we are able to develop novel numerical approaches for MLE in survival analysis without compromising the flexibility of models. There has been a rich literature on evaluating the gradient of the objective function in the DE-constrained optimization problem (Peto and Peto, 1972; Cao et al., 2003; Alexe and Sandu, 2009; Gerdts, 2011). Among them, the adjoint sensitivity analysis is computationally efficient when evaluating the gradient of a scalar function with respect to large number of model parameters (Cao et al., 2003). Therefore, we use the adjoint method to compute the gradient of (2.15), whose detailed derivation is provided in Section 2.4.3.

DE-constrained optimization has also found its applications in deep learning. Chen et al. (2018) and Dupont et al. (2019) recently used ODEs parameterized with neural networks to model continuous-depth neural networks, normalizing flows, and time series, which lead to DE-constrained optimization problems. Here, we share the merits of parameterizing the ODEs with neural networks but study a novel application of DE-constrained optimization in survival analysis.

[2] The optimization problem in (2.15) belongs to a subclass of DE-constrained optimization problems, with the generic form of $\min_\theta J(\Lambda, \theta)$, subject to $g_1(\Lambda(t), \Lambda'(t), t; \theta) = 0$ and $g_2(\Lambda(0); \theta) = 0$.

2.4.2 Survival ODE Model through Neural Networks

We consider the cumulative hazard function $\Lambda_x(\cdot)$ through an ODE (2.16) with a fixed initial value,

$$\begin{cases} \Lambda_x'(t) = h(\Lambda_x(t), t; x, \theta) \\ \Lambda_x(0) = 0 \end{cases}, \tag{2.16}$$

where the function h determines the dynamic change of $\Lambda_x(\cdot)$: the derivative of cumulative hazard at time t is determined by the current cumulative hazard $\Lambda_x(t)$, the current time t, and feature x through the function h parameterized by θ. The initial value implies that the event always occurs after time 0 since $S_x(0) = \exp(-\Lambda_x(0)) = 1$. Given an individual's feature vector x and the parameter vector θ, for any specific time point t^*, the cumulative hazard $\Lambda_x(t^*)$ can be obtained as the solution of the initial value problem (2.16) at the time t^*, and the hazard rate can be obtained as $\lambda_x(t^*) = h(\Lambda_x(t^*), t^*; x, \theta)$. Therefore, the function h fully determines the conditional distribution of the event time T. The existence and uniqueness of the solution can be guaranteed if h and its derivatives are Lipschitz continuous (Walter, 1998). In this chapter, we specify h as a neural network and the above guarantees hold as long as the neural network has finite weights and Lipschitz non-linearities. In practice, we do not require the initial value problem (2.16) to have a closed-form solution. We can obtain $\Lambda_x(t^*)$ numerically using any ODE solver given the derivative function h, initial value at $t_0 = 0$, evaluating time $t_1 = t^*$, parameters θ, and features x, that is

$$\Lambda_x(t^*) = \text{ODESolver}(h, \Lambda_x(0) = 0, t_1 = t^*, x, \theta). \tag{2.17}$$

We consider a general ODE form, where $h(\cdot, t; x, \theta)$ is a feed-forward neural network taking $\Lambda_x(t)$, t, and x as inputs, and θ represents all parameters in the neural network. Specifically, the Softplus activation function (Dugas et al., 2001) is used to constrain the output of the neural network, i.e. the hazard function, to be always positive. We refer this general form as SODEN; note that SODEN is a flexible survival model as it does not make strong assumptions on the family of the underlying distribution or how features x affect the event time.

Remark II.12. *Although there are other modeling alternatives that can uniquely characterize the event distribution such as the survival function, we choose to model the hazard in ODE (2.16) for three reasons. First, the hazard function has been widely used as the modeling target for summarizing survival data in the literature, due to its meaningful interpretation and informativeness about the underlying mechanism of events (Klein and Moeschberger, 2003, Chapter 2). Next, the hazard function is easier to model compared to the survival function, in the sense that it requires fewer constraints for the neural network structure under the ODE framework. For example, if we replace the cumulative hazard $\Lambda_x(t)$ with $S_x(t)$ in ODE (2.16), we need to make sure the solution not only being monotonically decreasing in t but also being within $[0,1]$ for any $t \geq 0$, which poses additional constraints on the structure of the neural network h. Last but not least, the hazard function itself is of direct interest in many applications. For example, recent works in operational planning requires knowledge of the hazard rate of the waiting time until the customer abandons the queue (Ibrahim and Whitt, 2009; Reed and Tezcan, 2012).*

2.4.3 Model Learning

We optimize SODEN by maximizing the likelihood function given i.i.d. observations. The negative log-likelihood function of the i^{th} observation can be written as

$$\mathcal{L}(\theta; D_i) \triangleq -\Delta_i \log h(\Lambda_{x_i}(y_i), y_i; x_i, \theta) + \Lambda_{x_i}(y_i), \tag{2.18}$$

where $D_i = (y_i, \Delta_i, x_i)$ for $i = 1, \cdots, N$, and $\Lambda_{x_i}(y_i)$, as given in (2.17), also depends on parameters θ. Our goal is to minimize $\sum_{i=1}^{N} \mathcal{L}(\theta; D_i)$ with respect to θ.

For large-scale applications, we propose to use mini-batch SGD to optimize the criterion, where the gradient of $\mathcal{L}$ with respect to θ is calculated through the adjoint method (Pontryagin et al., 1962). In comparison to naively applying the chain rule through all the operations used in computing the loss function, the adjoint method has the advantage of reducing memory usage and controlling numerical error explicitly in back-propagation.

Next, we demonstrate how the gradients can be obtained.

Back-propagation through adjoint sensitivity analysis In the forward pass, we need to evaluate $\mathcal{L}(\theta; D_i)$ for each i in a batch. While there might be no closed form for the solution of (2.16), $\Lambda_{x_i}(y_i)$ can be numerically calculated using a black-box ODESolver in (2.17) and all other calculations are straightforward. In the backward pass, the only non-trivial part in the calculation of the gradients of $\mathcal{L}$ with respect to θ is back-propagation through the black-box ODESolver in (2.17). We compute it by solving another augmented ODE introduced by adjoint sensitivity analysis. Specifically, let the adjoint $a(t)$ satisfy $a'(t) = -\frac{\partial h}{\partial \Lambda} a(t)$ with $a(y_i) = 1$, and then it follows that $\boldsymbol{\nabla}_\theta \Lambda_{x_i}(y_i) = \int_0^{y_i} a \frac{\partial h}{\partial \theta} dt$. Therefore, the gradient can be obtained by evaluating the following augmented ODE

$$\begin{cases} s'(t) = [h(\Lambda(t), t; x_i, \theta), -a(t)\frac{\partial h}{\partial \Lambda}, -a(t)\frac{\partial h}{\partial \theta}] \\ s(y_i) = [\Lambda_{x_i}(y_i), 1, \mathbf{0}_{|\theta|}] \end{cases}, \tag{2.19}$$

with $s(t) = [\Lambda(t), a(t), \bar{s}(t)]$ at $t = 0$, i.e., $\boldsymbol{\nabla}_\theta \Lambda_{x_i}(y_i) = \bar{s}(0)$. Note that this approach does not need to access internal operations of ODE solvers used in the forward pass. Moreover, modern ODE solvers allow one to control the trade-off between the computing time and accuracy. Also note that a GPU-based implementation of back-propagation following the above rule is available in the *torchdiffeq* library (Chen et al., 2018).

Mini-batching with time-rescaling trick We also provide a practical time-rescaling trick for mini-batching to better exploit the existing GPU-based implementation of ODE solvers. Concatenating ODEs of different observations in a mini-batch into a single combined ODE system is a useful trick for efficiently solving multiple ODEs on GPU. However, the existing GPU-based ODE solvers and the adjoint method in Chen et al. (2018) require that all the individual ODEs share the same initial point t_0 and the evaluating point t_1 in the ODESolver (2.17), which is unfortunately not the case in SODEN. For the i^{th} observation in a mini-batch, the ODE (2.16) in the forward pass needs to be evaluated at the corresponding observed time $t_1 = y_i$. To mitigate this discrepancy, we propose a time-rescaling trick that allows us to get the solution of individual ODEs at different time points by evaluating the combined ODE at only one time point. The key observation is that we can align the evaluating points of individual ODEs by variable transformation. Let $H_i(t) = \Lambda_{x_i}(t \cdot y_i)$, for which the dynamics is determined by

$$\begin{cases} H_i'(t) = h(H_i(t), ty_i; x_i, \theta)y_i \triangleq \tilde{h}(H_i(t), t; (x_i, y_i), \theta) \\ H_i(0) = \Lambda_{x_i}(0 \cdot y_i) = 0 \end{cases}.$$

Since $H_i(1) = \Lambda_{x_i}(y_i)$ for all i, evaluating the combined ODE of all $H_i(s)$ at $s = 1$ once will give us the values of $\Lambda_{x_i}(y_i)$ for all i. We therefore can take advantage of the existing GPU-based implementation for mini-batching by solving the combined ODE system of $H_i(s)$ with the time-rescaling trick[2].

[2] We note that some recently developed deep learning libraries (e.g., JAX (Bradbury et al., 2018)) could support mini-batching over complicated operations such as solving ODEs with different initial and evaluating time points without using the time-rescaling trick. However, the proposed rescaling trick provides an easy-to-implement extension for the *torchdiffeq* library and potentially other frameworks.

2.4.4 Simulation Study

In this section, we conduct a simulation study to illustrate that the proposed SODEN can fit well with data when the commonly used PH assumption does not hold. For ease of visualization, we consider events generated from two groups where their survival functions cross each other, thus the PH assumption is violated. Further, we also show the advantage of SODEN as a continuous-time model rather than a discrete-time model.

Set-up We generate event times from the conditional distribution defined by the survival function $S_x(t) = e^{-2t}I(x = 0) + e^{-2t^2}I(x = 1)$, where x follows a Bernoulli distribution with probability 0.5 and $I(\cdot)$ is the indicator function. The binary feature x can be viewed as an indicator for two groups of individuals. Note that the survival functions of the two groups, $S_0(t)$ and $S_1(t)$, cross at $t = 1$, hence the PH assumption does not hold. The censoring times were uniformly sampled between $(0, 2)$, which led to a censoring rate around 25%.

We apply the proposed SODEN and investigate the predicted survival functions and hazard functions under $x = 0$ and $x = 1$ respectively. We also provide the results of DeepHit (Lee et al., 2018), which is a discrete-time model without the PH assumption[3], to further illustrate the advantage of the continuous nature of SODEN. We train both models on the same simulated data with sample size 10,000. The reported results are based on 10 independent trials.

Results The results of SODEN are shown in the left column of Figure II.6. Note that the Kaplan-Meier (KM) estimate for each group can be considered a gold standard under our simulation setting, and we also plot them in Figure II.6 as the true survival functions corresponding to the data generating distribution. The predicted survival functions generally agree well with the true survival functions (the upper-left

[3] See Appendix A.8 for more details about this model.

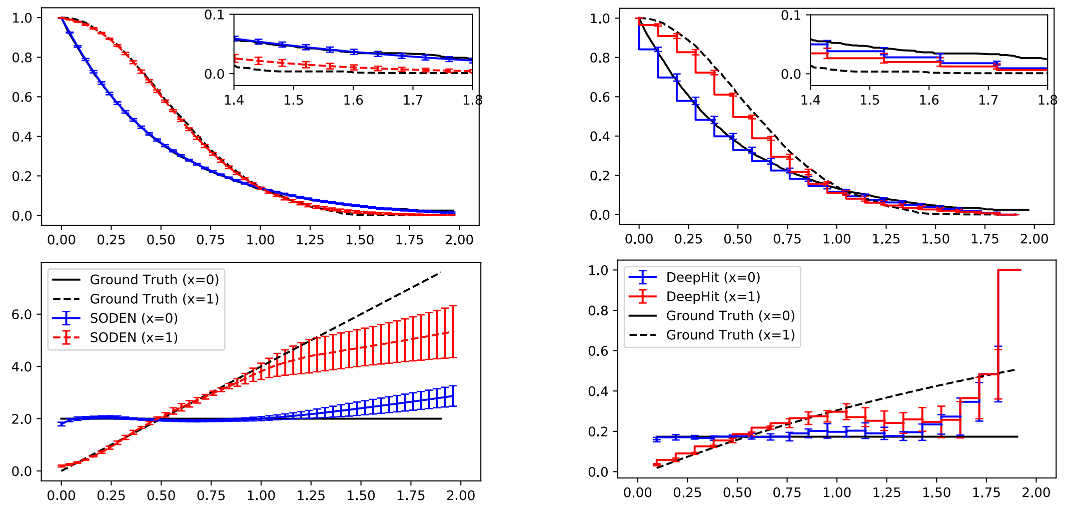

Figure II.6: The survival functions (top row) and hazard functions (bottom row) of two groups, $x = 0$ and $x = 1$. The left column shows the results of SODEN, and the right column shows the results of DeepHit. In all figures, the results are the average of 10 independent trials and error bars indicate the standard deviation. The red curve indicates the predicted function for group $x = 1$ and the blue curve indicates the predicted function for group $x = 0$. The survival (Kaplan-Meier curves) and hazard functions corresponding to the data generating distribution for the two groups are shown in black curves (solid curves for group $x = 0$ and dashed curves for group $x = 1$).

figure). The predicted survival functions of the two groups cross approximately at $t = 1$, indicating SODEN can fit well with data not under the PH assumption. The lower-left figure shows that the predicted hazard functions of SODEN agree well with the true hazard functions when time is relatively small, but deviate from the true hazard functions as time increases. This is anticipated as there are few data points when t is large and there are many more data points when t is small. As a side note, while the estimate of the survival function looks better than that of the hazard function when t is large, it is a visual artifact. As the survival function is monotonically decreasing and bounded between 0 and 1, the deviation (as indicated by the error bar) of the estimated survival function from the ground truth near the tail is visually diminished. Relatively, the estimate of the survival function actually becomes worse for larger time t.

The results of DeepHit are shown in the right column of Figure II.6. Due to the discrete nature of the model, both the survival functions and the hazard functions predicted by DeepHit are step functions. While the predicted survival functions (the upper-right figure) fit well with the true survival functions when t is small, the survival functions of the two groups are not well separated when t is large. As for the hazard function (the lower-right figure), similarly, the predicted hazard functions fit well when t is small but fluctuate wildly when t is large.

2.4.5 Real-world Examples

In this section, we demonstrate the effectiveness of SODEN by comparing it with five baseline models on three real-world datasets. We also conduct an ablation study to show the benefits of not making the PH assumption.

2.4.5.1 Datasets

We conduct experiments on the following three datasets: the Study to Understand Prognoses Preferences Outcomes and Risks of Treatment (**SUPPORT**), the Molecular Taxonomy of Breast Cancer International Consortium (**METABRIC**) (Katzman et al., 2018), and the Medical Information Mart for Intensive Care III (**MIMIC**) database (Johnson et al., 2016; Goldberger et al., 2000).

SUPPORT and METABRIC are two common survival analysis benchmark datasets, which have been used in many previous works (Katzman et al., 2018; Lee et al., 2018; Gensheimer and Narasimhan, 2019; Kvamme et al., 2019). We adopt the version preprocessed by Katzman et al. (2018) and refer readers there for more details. Despite their wide adoption in existing literature, we note that SUPPORT and METABRIC have relatively small sample sizes (8.8k for SUPPORT and 1.9k for METABRIC), which may not be ideal to evaluate deep survival analysis models.

In this paper, we further build a novel large-scale survival analysis benchmark

Dataset	N	p	Censoring rate	Censored time (Yrs)		Observed time (Yrs)	
				Mean	Median	Mean	Median
MIMIC (MIMIC-SEQ)	35,304	26 (5+15×24)	61%	0.21	0.02	1.50	0.42
SUPPORT	8,873	14	32%	2.90	2.51	0.56	0.16
METABRIC	1,904	9	42%	0.44	0.43	0.27	0.24

Table II.6: Summary statistics of three datasets. N is the sample size and p is the number of features. MIMIC-SEQ uses 5 time-static features and 15 time series features within the first 24 hours after admission.

dataset from the publicly available MIMIC database. The MIMIC database provides deidentified clinical data of patients admitted to an Intensive Care Unit (ICU) stay. We take adult patients who are alive 24 hours after the first admission to ICU. The event of interest is defined as the mortality after admission. The event time is observed if there is a record of death in the database; otherwise, the censored time is defined as the last time of being discharged from the hospital. In MIMIC dataset, we extract 26 features based on the first 24-hour clinical data following Purushotham et al. (2018). In addition, to further evaluate deep learning models on applications with more complex data structure, we consider another feature set involving time series for the same group of patients, which is named as **MIMIC-SEQ** for differentiation. MIMIC-SEQ contains 5 time-static features and 15 time series features within the first 24 hours after admission. Following the protocols described above, we are able to get a dataset with over 35k samples.

The detailed summary statistics of the three datasets are provided in Table II.6. In all datasets, the categorical features are encoded as dummy variables and all the features are standardized.

2.4.5.2 Models for Comparison

We compare the proposed method with the classical linear **Cox** model and four state-of-the-art neural-network-based models:

- **DeepSurv** is a PH model which replaces the linear feature combination in Cox with a neural network to improve feature extraction (Katzman et al., 2018).

- **Cox-Time** is a continuous-time model allowing non-PH, and is optimized by maximizing a modified partial-likelihood based loss function (Kvamme et al., 2019).

- **DeepHit** is a discrete-time survival model which estimates the probability mass at each pre-specified time interval, and is optimized by minimizing the linear combination of the negative log-likelihood and a differentiable surrogate ranking loss tailored for concordance index (Lee et al., 2018).

- **Nnet-Survival** also models discrete-time distribution via estimating the conditional hazard probability at each time interval (Gensheimer and Narasimhan, 2019).

Detailed model specifications and loss functions for the neural-network-based baselines can be found in Appendices A.7 and A.8. Note that on the MIMIC-SEQ dataset, we only compare neural-network-based models.

In Section 2.4.4, we have shown that the proposed model, because of its flexible parameterization, is able to fit well to the simulated data where the PH assumption does not hold. Here we further conduct an ablation study on real-world datasets to test the effect of the flexible parameterization. Specifically, we compare the general form of the proposed **SODEN**, with two of its degenerate variants, **SODEN-PH** and **SODEN-Cox**. SODEN-PH factorizes $h(\Lambda_x(t), t; x, \theta) = h_0(t; \theta)g(x; \theta)$ as a multiplication of two functions to satisfy the PH assumption, where both h_0 and g are specified as neural networks. SODEN-Cox is a linear version of SODEN-PH where $g(x) = e^{x\beta}$. Notably, SODEN-Cox and SODEN-PH are designed to have similar representation power as Cox and DeepSurv respectively.

2.4.5.3 Evaluation Metrics

Evaluating survival predictions needs to account for censoring. Here we describe several commonly used evaluation metrics (Kvamme et al., 2019; Wang et al., 2019a).

Time-dependent concordance index Concordance index (C-index) (Harrell Jr. et al., 1984) is a commonly used discriminative evaluation metric in survival analysis, and it measures the probability that, for a random pair of observations, the relative order of the two event times is consistent with that of the two predicted survival probabilities. The C-index was originally designed for models using the PH assumption, where the relative order of the predicted survival probabilities for two given individuals does not change with time. Antolini et al. (2005) further propose time-dependent C-index for models without PH assumption, where the relative order of the predicted survival probabilities may be different if evaluated at different time points. In addition, Uno et al. (2011) introduce inverse probability weights to the C-index such that it does not depend on the study-specific censoring distribution. Following Antolini et al. (2005) and Uno et al. (2011), we adopt the inverse probability weighted time dependent C-index in our evaluation, which is given by

$$C^{td} = \frac{\sum_{i:\Delta_i=1} \sum_{j:y_i<y_j} I(\hat{S}_{x_i}(y_i) < \hat{S}_{x_j}(y_i))/\hat{G}^2(y_i)}{\sum_{i:\Delta_i=1} \sum_{j:y_i<y_j} 1/\hat{G}^2(y_i)},$$

where x_i, y_i, and Δ_i are the features, observed time, and event indicator for individual i; $I(\cdot)$ is the indicator function; $\hat{S}_{x_i}(t)$ is the predicted survival function at time t given x_i; and $\hat{G}(t)$ is the Kaplan-Meier estimator for the survival function of the censoring time, i.e. $\mathcal{P}(C > t)$. Under the independence assumption between the censoring time and the event time, C^{td} converges to the discrimination measure $\mathcal{P}(S_{x_i}(T_i) < S_{x_j}(T_i) | T_i < T_j)$.

In practice, the estimation of $\hat{G}(t)$ as well as the model predictions are relatively unstable for large t due to limited number of observations, yet they lead to large inverse probability weights $1/\hat{G}(t)$. Following Uno et al. (2011), we implement a truncated version of time-dependent C-index within a pre-specified time interval $(0, \tau)$, i.e.,

$$C_{\tau}^{td} = \frac{\sum_{i:\Delta_i=1, y_i<\tau} \sum_{j:y_i<y_j} I(\hat{S}_{x_i}(y_i) < \hat{S}_{x_j}(y_i))/\hat{G}^2(y_i)}{\sum_{i:\Delta_i=1, y_i<\tau} \sum_{j:y_i<y_j} 1/\hat{G}^2(y_i)}.$$

We report results under various τ with $\hat{G}(\tau) = 10^{-8}, 0.2$, and 0.4. When $\hat{G}(\tau) = 10^{-8}$, it is almost identical to the non-truncated version. Note that C_{τ}^{td}=1 corresponds to a perfect ranking of predicted survival probabilities and C_{τ}^{td}=0.5 corresponds to a random ordering.

Integrated Brier score For a binary classifier, the Brier score (BS) is defined as the mean square difference between the predicted probability and the ground-truth binary label. The metric BS can be decomposed into two components measuring calibration and discriminative performance respectively. Given similar discriminative performance, a lower BS indicates the closer the predicted survival probability $\hat{S}_x(t)$ is to the true probability of experiencing the event after time t. We refer well calibrated models to those with good probability estimates. Graf et al. (1999) generalized BS to take account for censoring in survival analysis. Specifically, the BS for survival analysis at time t is defined as

$$\mathrm{BS}(t) = \frac{1}{N} \sum_{i=1}^{N} \left\{ \frac{(\hat{S}_{x_i}(t))^2 I(y_i \leq t, \Delta_i = 1)}{\hat{G}(y_i)} + \frac{(1 - \hat{S}_{x_i}(t))^2 I(y_i > t)}{\hat{G}(t)} \right\},$$

where the notations are the same as C^{td}. As the predicted survival probability depends on the time point of evaluation, we use integrated BS (IBS) to measure the overall BS on a time interval:

$$\text{IBS} = \frac{1}{t_{\max} - t_{\min}} \int_{t_{\min}}^{t_{\max}} \text{BS}(t)\, dt.$$

In practice, we choose the interval $[0, t_{\max}]$ with various $t_{\max}$ satisfying $\hat{G}(t_{\max}) = 10^{-8}, 0.2$, and 0.4, and compute this integral numerically by averaging over 100 grid points. The lower the IBS, the better the performance.

Integrated binomial log-likelihood Graf et al. (1999) also generalized the binomial log-likelihood (BLL), which is a binary classification evaluation metric measuring both discrimination and calibration, to survival analysis in a similar way as BS. The BLL for survival analysis at time t is defined as

$$\text{BLL}(t) = \frac{1}{N} \sum_{i=1}^{N} \left\{ \frac{\log\left(1 - \hat{S}_{x_i}(t)\right) I(y_i \le t, \Delta_i = 1)}{\hat{G}(y_i)} + \frac{\log\left(\hat{S}_{x_i}(t)\right) I(y_i > t)}{\hat{G}(t)} \right\},$$

where the notations are the same as BS. We can also define the integrated BLL (IBLL) to measure the overall performance from $t_{\min}$ to $t_{\max}$, where

$$\text{IBLL} = \frac{1}{t_{\max} - t_{\min}} \int_{t_{\min}}^{t_{\max}} \text{BLL}(t)\, dt.$$

The higher the IBLL, the better the performance. Note that the IBS takes the squared error in the loss, i.e., error^2, while the negative IBLL accounts for error with scale $-\log(1 - \text{error})$. Thus, in general, IBLL has larger magnitude than IBS and penalizes more for larger error.

Negative log-likelihood Negative log-likelihood (NLL) corresponds to $\mathcal{L}(\theta; D_i)$ in (2.18) and predictive NLL on held out data measures the goodness-of-fit of the model to the observed data. However, NLL is only applicable to models that provide likelihood, and it is not comparable between discrete-time and continuous-time models due to the difference in the likelihood definition. We use NLL to compare three variants of SODEN in the ablation study. The lower the NLL, the better the performance.

2.4.5.4 Experimental Setup

We randomly split each dataset into training, validation and testing sets with a ratio of 3:1:1. To make the evaluation more reliable, we take 5 independent random splits for MIMIC(-SEQ), 10 independent random splits for SUPPORT and METABRIC as their sizes are relatively small. For each split, we train the Cox model on the combination of training and validation sets. For neural-network-based models, we train each model on the training set, and apply early-stopping using the loss on the validation set with patience 10. The hyper-parameters of each model are tuned within each split through 100 independent trials using random search. We select the optimal hyper-parameter setting with the best score on the validation set. For continuous-time models, DeepSurv, Cox-Time, and SODEN, the validation score is set as the loss. For discrete-time models, DeepHit and Nnet-Survival, the loss functions (i.e., NLLs) across different pre-specified time intervals are not comparable so the validation score is set as C^{td} as was done in Kvamme et al. (2019).

For all neural networks, we use multilayer perceptrons (MLP) with ReLU activation in all layers except for the output layer. For SODEN, Softplus is used to constrain the output to be always positive; for DeepHit and Nnet-Survival, Softmax and Sigmoid are used respectively to return PMF and discrete hazard probability. For the MIMIC-SEQ dataset, we incorporate a one-layer Gated Recurrent Units (GRU) encoder into the model architecture of each deep survival model to learn feature representation from sequence data. We use the RMSProp (Tieleman and Hinton, 2012) optimizer and tune batch size, learning rate, weight decay, momentum, the number

$\mathcal{P}(C > \tau)$	Model	C^{td}_{τ} (↑)	IBLL (↑)	IBS (↓)
10^{-8}	DeepSurv	$0.685 \pm .002$	$-0.335 \pm .003$	$\mathbf{0.103} \pm .001$
	Cox-Time	$0.681 \pm .002$	$-0.332 \pm .003$	$\mathbf{0.103} \pm .001$
	Nnet-Survival	$0.679 \pm .003*$	$\underline{-0.331} \pm .003*$	$0.104 \pm .001$
	DeepHit	$\mathbf{0.688} \pm .002$	$-0.336 \pm .005*$	$0.106 \pm .001*$
	SODEN (ours)	$\underline{0.687} \pm .002$	**-0.328** $\pm .004$	$\mathbf{0.103} \pm .001$
0.2	DeepSurv	$0.685 \pm .002$	$-0.400 \pm .013$	$\mathbf{0.124} \pm .003$
	Cox-Time	$0.681 \pm .002$	$-0.397 \pm .011$	$\underline{0.125} \pm .003$
	Nnet-Survival	$0.679 \pm .003*$	$\underline{-0.396} \pm .011*$	$0.126 \pm .003$
	DeepHit	$\mathbf{0.688} \pm .002$	$-0.402 \pm .012*$	$0.128 \pm .004*$
	SODEN (ours)	$\underline{0.687} \pm .002$	**-0.391** $\pm .011$	$\underline{0.125} \pm .004$
0.4	DeepSurv	$0.740 \pm .002*$	$-0.386 \pm .013*$	$0.121 \pm .005*$
	Cox-Time	$0.744 \pm .003*$	$-0.382 \pm .014*$	$\underline{0.120} \pm .005*$
	Nnet-Survival	$0.737 \pm .004*$	$-0.391 \pm .015$	$0.123 \pm .006$
	DeepHit	$\mathbf{0.752} \pm .003$	$\underline{-0.381} \pm .014$	$\underline{0.120} \pm .005*$
	SODEN (ours)	$\mathbf{0.752} \pm .003$	**-0.374** $\pm .013$	$\mathbf{0.118} \pm .005$

Table II.7: Comparison of time dependent concordance index (C^{td}_{τ}), integrated binomial log-likelihood (IBLL), integrated brier score (IBS) on MIMIC-SEQ. The **bold** and underline markers denote the best and the second best performance respectively. The (±) error bar denotes the standard error of the mean. The asterisk (*) after a baseline model performance indicates a significant (either positive or negative) difference between that baseline model and the proposed SODEN, under pairwise t-test with p-value < 0.05.

of layers, and the number of neurons in each layer. The search ranges for the aforementioned hyper-parameters are shared across all neural-network-based models on each dataset. Additionally, we tune batch normalization and dropout for all neural-network-based baseline models. For DeepHit and Nnet-Survival, we tune the number of pre-specified time intervals. We also smooth the predicted survival function by interpolation, which is an important post-processing step to improve the performance of these discrete-time models. The tuning ranges of hyper-parameters are listed in Appendix A.9.

$\mathcal{P}(C > \tau)$	Model	C^{td}_{τ} (↑)	IBLL (↑)	IBS (↓)
10^{-8}	Cox	0.660 ± .001∗	−0.335 ± .003∗	0.105 ± .001∗
	DeepSurv	0.683 ± .001	−0.326 ± .005∗	$\underline{0.101}$ ± .001∗
	Cox-Time	0.680 ± .001	−0.326 ± .003∗	$\underline{0.101}$ ± .001∗
	Nnet-Survival	0.681 ± .001	$\underline{-0.321}$ ± .002	$\underline{0.101}$ ± .001
	DeepHit	**0.685** ± .002	−0.327 ± .003∗	0.102 ± .001∗
	SODEN (ours)	$\underline{0.684}$ ± .002	**-0.319** ± .003	**0.100** ± .001
0.2	Cox	0.660 ± .001∗	−0.413 ± .007∗	0.132 ± .003∗
	DeepSurv	0.683 ± .001	−0.402 ± .006∗	$\underline{0.127}$ ± .002∗
	Cox-Time	0.680 ± .001	−0.404 ± .007∗	0.128 ± .002∗
	Nnet-Survival	0.682 ± .001	$\underline{-0.398}$ ± .007	$\underline{0.127}$ ± .003
	DeepHit	**0.685** ± .002	−0.404 ± .008∗	0.128 ± .002∗
	SODEN (ours)	$\underline{0.684}$ ± .002	**-0.395** ± .006	**0.126** ± .002
0.4	Cox	0.706 ± .003∗	−0.399 ± .018∗	0.124 ± .007∗
	DeepSurv	0.739 ± .003∗	−0.387 ± .016	$\underline{0.120}$ ± .007∗
	Cox-Time	0.737 ± .003∗	−0.387 ± .020∗	$\underline{0.120}$ ± .007∗
	Nnet-Survival	0.741 ± .005	$\underline{-0.386}$ ± .019∗	$\underline{0.120}$ ± .007∗
	DeepHit	**0.747** ± .004	−0.404 ± .023	0.128 ± .009
	SODEN (ours)	$\underline{0.746}$ ± .003	**-0.379** ± .019	**0.118** ± .007

Table II.8: Comparison of performance on MIMIC. The notations share the same definitions as in Table II.7.

2.4.5.5 Results

Discriminative and calibration performance The comparison of model performances on MIMIC-SEQ, MIMIC, SUPPORT, and METABRIC are respectively reported in Tables II.7 to II.10.

We first consider the C-index metric, which measures the discriminative performance. We observe that the proposed SODEN outperforms other continuous-time models (Cox, DeepSurv, and Cox-Time). The differences in C-index are significant on all datasets, except for those with large τ on MIMIC-SEQ and MIMIC. The gain of SODEN against DeepSurv and Cox-Time demonstrates the benefits of not making the PH assumption and having a principled likelihood objective. We also observe that all neural network models significantly outperform the Cox model in almost all cases.

$\mathcal{P}(C > \tau)$	Model	C^{td}_{τ} (↑)	IBLL (↑)	IBS (↓)
10^{-8}	Cox	0.596 ± .002∗	−0.568 ± .001∗	0.194 ± .001∗
	DeepSurv	0.609 ± .003∗	**-0.559** ± .002∗	**0.190** ± .001∗
	Cox-Time	0.607 ± .004∗	−0.565 ± .002	$\underline{0.191}$ ± .001
	Nnet-Survival	0.624 ± .003	−0.570 ± .004	0.193 ± .001∗
	DeepHit	**0.631** ± .003	−0.583 ± .006∗	0.197 ± .001∗
	SODEN (ours)	$\underline{0.627}$ ± .003	$\underline{-0.563}$ ± .002	$\underline{0.191}$ ± .001
0.2	Cox	0.596 ± .002∗	−0.585 ± .001∗	0.201 ± .000∗
	DeepSurv	0.609 ± .003∗	**-0.577** ± .002	**0.197** ± .001
	Cox-Time	0.606 ± .004∗	−0.583 ± .002	0.199 ± .001
	Nnet-Survival	0.623 ± .003	−0.586 ± .003	0.201 ± .001∗
	DeepHit	**0.630** ± .003	−0.601 ± .006∗	0.205 ± .002∗
	SODEN (ours)	$\underline{0.627}$ ± .003	$\underline{-0.579}$ ± .002	$\underline{0.198}$ ± .001
0.4	Cox	0.595 ± .002∗	−0.602 ± .001∗	0.208 ± .001∗
	DeepSurv	0.608 ± .002∗	**-0.595** ± .002	**0.205** ± .001
	Cox-Time	0.605 ± .004∗	−0.601 ± .002	0.207 ± .001
	Nnet-Survival	0.623 ± .003	−0.602 ± .003	0.208 ± .001∗
	DeepHit	**0.630** ± .003	−0.619 ± .007∗	0.212 ± .002∗
	SODEN (ours)	$\underline{0.626}$ ± .003	$\underline{-0.597}$ ± .002	**0.205** ± .001

Table II.9: Comparison of performance on SUPPORT. The notations share the same definitions as in Table II.7.

For discrete-time models, Nnet-Survival and DeepHit show strong discriminative performance on the C-index metric compared to continuous-time models in general. This is not surprising due to the facts that 1) similar as SODEN, the discrete-time models do not make strong structural assumptions; 2) the discrete-time models are tuned with C-index as the validation metric, and DeepHit has an additional ranking loss tailored for C-index. However, we find their advantage diminishes on MIMIC-SEQ and MIMIC, where the data size is much larger. We suspect the information loss due to discretizing the event time becomes more severe as the data size grows, and will eventually turn to the discriminative performance bottleneck.

We then consider the IBLL and IBS metrics, which measure a combination of the discriminative performance and the calibration performance. Overall, most models are similarly well-calibrated. However, DeepHit is obviously less calibrated than most

$\mathcal{P}(C > \tau)$	Model	C^{td}_{τ} (↑)	IBLL$_{\tau}$ (↑)	IBS$_{\tau}$ (↓)
10^{-8}	Cox	$0.644 \pm .006*$	$\underline{-0.508} \pm .009*$	$\underline{0.169} \pm .002$
	DeepSurv	$0.635 \pm .007*$	$-0.517 \pm .011*$	$0.171 \pm .003*$
	Cox-Time	$0.648 \pm .007*$	$-0.511 \pm .009*$	$0.172 \pm .003*$
	Nnet-Survival	$\underline{0.666} \pm .005$	$-0.510 \pm .007$	$0.171 \pm .002*$
	DeepHit	$\mathbf{0.674} \pm .006*$	$-0.514 \pm .004*$	$0.174 \pm .002*$
	SODEN (ours)	$0.661 \pm .005$	$\mathbf{-0.498} \pm .008$	$\mathbf{0.167} \pm .003$
0.2	Cox	$0.639 \pm .006*$	$\underline{-0.521} \pm .006$	$\underline{0.176} \pm .002$
	DeepSurv	$0.635 \pm .006*$	$-0.530 \pm .005*$	$0.179 \pm .002*$
	Cox-Time	$0.647 \pm .005*$	$-0.531 \pm .007*$	$0.179 \pm .002*$
	Nnet-Survival	$\underline{0.662} \pm .004$	$-0.523 \pm .003$	$0.177 \pm .001$
	DeepHit	$\mathbf{0.671} \pm .004*$	$-0.533 \pm .003*$	$0.182 \pm .001*$
	SODEN (ours)	$0.659 \pm .003$	$\mathbf{-0.516} \pm .005$	$\mathbf{0.174} \pm .002$
0.4	Cox	$0.637 \pm .006*$	$-0.521 \pm .006$	$\underline{0.175} \pm .002$
	DeepSurv	$0.635 \pm .006*$	$-0.526 \pm .005*$	$0.178 \pm .002*$
	Cox-Time	$0.644 \pm .005*$	$-0.526 \pm .006*$	$0.178 \pm .002*$
	Nnet-Survival	$\underline{0.660} \pm .003$	$\underline{-0.519} \pm .003$	$0.176 \pm .001$
	DeepHit	$\mathbf{0.668} \pm .004*$	$-0.528 \pm .003*$	$0.180 \pm .001*$
	SODEN (ours)	$0.658 \pm .004$	$\mathbf{-0.513} \pm .005$	$\mathbf{0.173} \pm .002$

Table II.10: Comparison of performance on METABRIC. The notations share the same definitions as in Table II.7.

other models, given it has the worst IBLL and IBS and the best C-index metric in most settings. This may be due to the surrogate ranking loss used in DeepHit.

In summary, the proposed SODEN demonstrates significantly better discriminative performance than all continuous-time baseline methods on all datasets. On the larger datasets (MIMIC-SEQ and MIMIC), SODEN achieves better or similar C-index metric compared to the discrete models. The superior discriminative performance of DeepHit comes at the price of the inferior calibration performance.

Finally, we remark that the event time and censoring time in MIMIC both have heavily right-skewed distributions, as indicated by the large discrepancy between their mean and median in Table II.6. On MIMIC and MIMIC-SEQ, including more testing data near the tail in evaluation ($\hat{G}(\tau) = 10^{-8}$ or 0.2[4]) gives a worse C^{td}_{τ} compared

[4] On MIMIC and MIMIC-SEQ, both $\hat{G}(\tau) = 10^{-8}$ and $\hat{G}(\tau) = 0.2$ have a tiny number of samples being excluded due to the right-skewness of the censoring distribution, and thus are close to the

Dataset	Metric ($\mathcal{P}(C > \tau)$)	SODEN	SODEN-PH	SODEN-Cox
MIMIC-SEQ	NLL	$\mathbf{0.489} \pm .072$	$\underline{0.520} \pm .069*$	N/A
	C^{td} (10^{-8})	$\mathbf{0.687} \pm .002$	$\underline{0.682} \pm .001*$	N/A
	C^{td} (0.2)	$\mathbf{0.687} \pm .002$	$\underline{0.683} \pm .001*$	N/A
	C^{td} (0.4)	$\mathbf{0.752} \pm .003$	$\underline{0.739} \pm .005*$	N/A
MIMIC	NLL	$\mathbf{0.411} \pm .007$	$\underline{0.436} \pm .007*$	$0.450 \pm .006*$
	C^{td} (10^{-8})	$\mathbf{0.684} \pm .002$	$\underline{0.679} \pm .002*$	$0.659 \pm .001*$
	C^{td} (0.2)	$\mathbf{0.684} \pm .002$	$\underline{0.679} \pm .002*$	$0.659 \pm .001*$
	C^{td} (0.4)	$\mathbf{0.746} \pm .003$	$\underline{0.734} \pm .003*$	$0.706 \pm .003*$
SUPPORT	NLL	$\mathbf{0.676} \pm .008$	$\underline{0.702} \pm .008*$	$0.761 \pm .022*$
	C^{td} (10^{-8})	$\mathbf{0.627} \pm .003$	$\underline{0.608} \pm .003*$	$0.591 \pm .003*$
	C^{td} (0.2)	$\mathbf{0.627} \pm .003$	$\underline{0.608} \pm .002*$	$0.590 \pm .004*$
	C^{td} (0.4)	$\mathbf{0.626} \pm .003$	$\underline{0.607} \pm .002*$	$0.589 \pm .004*$
METABRIC	NLL	$\mathbf{0.149} \pm .015$	$0.176 \pm .013*$	$\underline{0.167} \pm .010*$
	C^{td} (10^{-8})	$\mathbf{0.661} \pm .005$	$0.640 \pm .005*$	$\underline{0.642} \pm .006*$
	C^{td} (0.2)	$\mathbf{0.659} \pm .003$	$\underline{0.639} \pm .004*$	$0.638 \pm .005*$
	C^{td} (0.4)	$\mathbf{0.658} \pm .004$	$\underline{0.639} \pm .005*$	$0.636 \pm .006*$

Table II.11: Comparison of negative log-likelihood (NLL) and time dependent concordance index (C^{td}_{τ}) between SODEN and its degenerate variants, SODEN-Cox and SODEN-PH, for ablation study. The **bold**, underline, and (±) error bar share the same definitions as in Table II.7. The asterisk (*) indicates a significant difference between the proposed SODEN and its degenerate variants, under pairwise t-test with p-value < 0.05.

to including less tail samples ($\hat{G}(\tau) = 0.4$). This is because models tend to have poor prediction performance near the tail due to limited number of observations, yet these tail samples get large inverse probability weights. This also explains why the differences in C^{td}_{τ} among different models are less significant when including more tail samples.

Ablation study While the trend over Cox, DeepSurv, and SODEN has supported our conjecture that flexible parameterization by introducing non-linearity and not making the PH assumption is important for practical survival analysis on modern datasets, we further verify this conjecture by the ablation study with SODEN-PH

non-truncated version C^{td}.

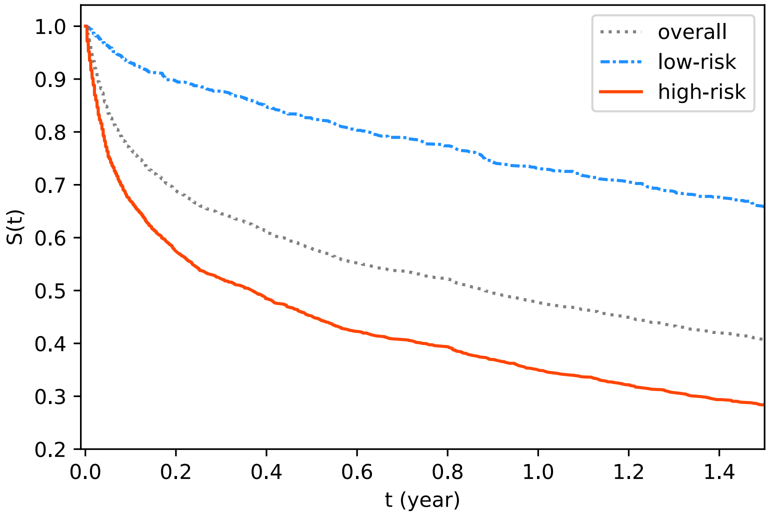

Figure II.7: Kaplan-Meier curves of high/low-risk groups for SODEN on MIMIC.

and SODEN-Cox (see Table II.11).

First, we observe that the relative differences in the C-index metric among SODEN-Cox, SODEN-PH, and SODEN are similar as those among Cox, DeepSurv, and SODEN. In fact, we can see that the C_τ^{td}'s of SODEN-Cox and SODEN-PH in Table II.11 are respectively similar with those of their partial-likelihood counterparts Cox and DeepSurv in Tables II.7 to II.10. This observation implies that 1) neural networks can approximate the baseline hazard function as well as the non-parametric Breslow's estimator (Lin, 2007); 2) maximizing the likelihood function with numerical approximation approaches, where SGD based algorithms can be naturally applied, can perform as well as maximizing the partial likelihood for PH models.

Second, SODEN outperforms SODEN-PH and SODEN-Cox in terms of NLL by a large margin. The major difference between SODEN-PH and SODEN is that the former is restricted by the PH assumption while the latter is not. The comparison of NLL between SODEN-PH and SODEN provides a strong evidence that the PH assumption may not hold on these datasets. Further, SODEN-Cox often being the worst verifies again that both non-linearity and the flexibility of non-PH models matter.

Risk discriminating visualization We further provide visualization of risk discrimination. We show the Kaplan-Meier curves (Kaplan and Meier, 1958) of high-risk and low-risk groups identified by SODEN on the MIMIC dataset. We first obtain the predicted survival probability for each individual at the median of all observed survival times in the test set. We then split the test set into high-risk and low-risk groups evenly based on their predicted survival probabilities. The Kaplan-Meier curves for the high-risk group, the low-risk group, and the entire test set are shown in Figure II.7. The difference between high-risk and low-risk groups is statistically significant where the p-value of the log rank test (Peto and Peto, 1972) is smaller than 0.001.

2.5 Discussion

In this chapter, we have proposed a novel ODE framework for survival analysis. We revisit the rich literature of survival analysis and provide a unified view of many existing survival models in Section 2.2. This unification merit serves as the foundation of the proposed widely applicable estimation procedure. In particular, the proposed estimation procedure is scalable and easy to implement based on well-developed numerical solvers and local sensitivity analysis tools for ODEs. We have demonstrated the effectiveness of the proposed method on both simulation studies and real-world data examples.

In Section 2.3, we focus on estimation and inference for a general class of semi-parametric ODE models, in which case the effects of certain covariates are often of interest. We have established the consistency and semi-parametric efficiency of the proposed sieve estimator, with a new general sieve M-theorem. The proposed general theory derives the asymptotic distribution of bundled parameters, where the nuisance parameter is a function of not only the regression parameters of interest but also other infinite-dimensional nuisance parameters. Though we have only illustrated the efficient estimation in the linear transformation model as an example to motivate such a theoretical development, the proposed general theory can be extended to other models.

Further, the proposed ODE framework and the estimation method offer new opportunities for investigating more flexible model structures as well. In Section 2.4, we have developed survival models with powerful representation learning via neural networks to improve prediction performance. The proposed SODEN can model a broad range family of continuous event time distributions without strong structural assumptions and the algorithm scales well by allowing direct use of mini-batch SGD.

In addition, an interesting application of the unified ODE framework is to check the model specification. In particular, the estimation and inference for a general ODE model can help test whether a nested model is appropriate for a dataset. For example, Proposition 2.2.2 implies that the function $q(\cdot)$ or $\alpha(\cdot)$ in the linear transformation model (2.5) should be a power function when it coincides with the Cox or the AFT model. Though we have established the consistency of the functional parameters $q(\cdot)$ and $\alpha(\cdot)$ in the nonparametric linear transformation model, it is worthwhile to further investigate their asymptotic distributional theory for model diagnostics as future work. As a preliminary study, we have explored a heuristic parametric approach for model diagnostics and provided its finite sample performance in the Supplemental Material.

Finally, we note that a few recent works have tried to address the computation burden of certain estimation methods for specific models on massive time-to-event data. In particular, Wang et al. (2019b) proposed an efficient divide-and-conquer (DAC) algorithm for the sparse Cox model. Kawaguchi et al. (2020) developed an algorithm for reducing the computation cost of fitting the Fine-Gray (Fine and Gray, 1999) proportional subdistributional hazards model by exploiting its special structure. Zuo et al. (2021) proposed a subsampling procedure to approximate the full-data estimator for the additive hazard model. Note that most of these methods are tailored for a specific model while our method can be applied more broadly. Further, our estimation procedure and these methods are not competitors. In contrast, some of the techniques used in these methods, such as DAC, can be naturally integrated into the proposed estimation procedure, which is an interesting future direction to be explored.

CHAPTER III

Latent Space Approach for Signed Networks

3.1 Introduction

Networks characterize connectivity relationships between individuals of a complex system and are ubiquitous in various fields, such as social science, biology, transportation, and information technology (Newman, 2010). In a network, a node represents an individual and an edge between two nodes indicates the presence of certain interaction or relation. Given the unique relational information represented by networks, many statistical models have been developed to understand the underlying mechanism of the system and help explain the observed phenomenon on networks; see for example Goldenberg et al. (2010) for a comprehensive overview. One important class of statistical models is the latent variable model, where the presence/absence of an edge depends on the node latent variables. For example, stochastic block models use latent categorical variables to describe the block structure among nodes (Abbe, 2018); latent space models map nodes into a low-dimensional metric space while accounting for transitivity, homophily for node attributes, node heterogeneity and clustering (Hoff et al., 2002; Krivitsky et al., 2009). Such latent variable models are attractive due to their interpretable structure, their nature for network visualization, and their ability for downstream network-assisted learning such as node clustering, node classification, and network link prediction.

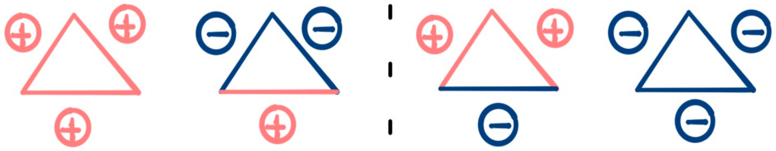

Figure III.1: Four types of triangles in signed networks, where the left two are balanced and the right two are unbalanced.

Nonetheless, most statistical network models only focus on the presence/absence of edges while ignoring different types of edges, which makes them inadequate for modeling *signed networks*. A signed network consists of two types of edges, positive edges and negative edges, and such polarized relationships are common in real-world networks. For example, positive and negative edges may respectively correspond to relationships of like and dislike in social networks, collaboration and competition in trading networks, or alliance and militarized dispute in international relation networks. Modeling signed networks has its own unique challenges not merely due to the additional sign for each edge, but more importantly, because the presence of positive and negative edges affect each other in certain ways. There have been various social theories that describe the structural pattern of signed networks (Guha et al., 2004; Leskovec et al., 2010; Knoke, 2013), an important one being the structural balance theory (Heider, 1946; Harary et al., 1953). Specifically, the balance theory describes the distribution of different types of triangles (i.e. three nodes that are connected with each other). A triangle in a signed network is called *balanced* if the product of its three edge signs is positive; and it is called *unbalanced* otherwise (see Figure III.1 for examples). The balance theory postulates that balanced triangles should be more prevalent than unbalanced triangles in signed networks, which directly coincides with the proverb, “the enemy of my enemy is my friend” and “the friend of my friend is my friend”. Moreover, recent studies have found empirical evidence of the balance property in many real-world signed networks (Leskovec et al., 2010; Kirkley et al., 2019; Feng et al., 2020).

On the other hand, there have been very few statistical models for signed networks that incorporate the balance theory into modeling. To the best of our knowledge, Derr et al. (2018) is the only recent work; specifically, it extends the configuration model (Chung and Lu, 2006) to signed networks with a focus on matching not only the node degree distribution but also the sign distribution and proportion of balanced triangles. Besides statistical models, there is a collection of work using low-rank matrix completion approaches induced by the balance theory for learning tasks such as sign prediction and clustering (Hsieh et al., 2012; Chiang et al., 2014). These works assume that there are underlying signed edges (not allowing for zeros) between *all* possible pairs of n nodes, and view the network as a fixed $n \times n$ adjacency matrix with entries of $\{+1, -1\}$. In comparison, statistical network models can provide statistically interpretable structures and account for noise in both signs and edges by modeling network distributions that precisely quantify the randomness in the observed data.

In this chapter, we propose a latent space approach to accommodate the balance theory for signed networks in a statistically principled way. Specifically, we introduce a novel definition of balance at the population level, which matches the balance theory in nature while viewing an observed network as the realization of a random quantity. For concreteness, we consider an undirected signed network with n nodes denoted by a symmetric signed adjacency matrix A, with $A_{ij} = A_{ji} = 1$ if node i and node j are linked by a positive edge, $A_{ij} = A_{ji} = -1$ if node i and node j are linked by a negative edge, and $A_{ij} = A_{ji} = 0$ if there is no edge between i and j. We assume there is no self-loop and thereby the diagonal elements of A are zeros. We assume A_{ij} to be random variables taking values in $\{-1, 0, 1\}$ and define the notion of balance at the population level as follows.

Definition 3.1.1 (Population-level balanced network)**.** A network is population-level balanced if

$$E(A_{ij}A_{j\ell}A_{\ell i}\big||A_{ij}A_{j\ell}A_{\ell i}| = 1) > 0, \text{ for any three different nodes } (i, j, \ell).$$

This definition suggests the expected product of signs on any triangle to be positive but does not require all triangles to be balanced in an observed signed network. Furthermore, the stochastic notion in Definition 3.1.1 allows us to investigate what generating mechanisms of signed networks are inherently of population-level balance. Specifically, we will focus on a general class of latent space models, due to aforementioned merits of latent space models. Rigorous descriptions are provided in Section 3.2. The key finding is that, if there exists a partition of the latent space such that edges tend to be positive within the same subset and negative between different subsets, then the network generated from such a latent space model is population-level balanced.

Based on this finding, we further propose a class of balanced inner-product models that directly capture the population-level balance. A unique difference from latent space models for unsigned networks is that we introduce an additional *latent polar variable* for each node. In particular, when the product of latent polar variables of two nodes has a large positive value, the sign of an edge between them is more likely to be positive; for a node with the latent polar variable being zero, it has no preference on the signs when forming edges with other nodes. We note that it is this novel introduction of latent polar variables that enables modeling signed networks with the population-level balance.

The rest of this chapter is organized as follows. We introduce the latent space approach for signed networks in Section 3.2, where we also provide a sufficient condition for the population-level balance. We present the proposed balanced inner product models in Section 3.3. We develop two scalable estimation methods in Section 3.4 and establish their non-asymptotic error rates in Section 3.5, which are further validated by simulation studies in Section 3.6. We demonstrate the effectiveness of the proposed approach in modeling a real-world signed network for international relations in Section 3.7. All proofs are given in the Appendices.

3.2 A Latent Space Approach for Signed Networks

In this section, we propose a probabilistic generative process for undirected signed networks with n nodes. Recall that $A \in \{1, 0, -1\}^{n\times n}$ is the signed adjacency matrix. Suppose the latent space $\mathcal{U}_0$ is endowed with the probability measure P_u; $B(\cdot,\cdot) : \mathcal{U}_0 \times \mathcal{U}_0 \to (0,1)$ is a function symmetric in its two arguments; $f(\cdot,\cdot) : \mathcal{U}_0 \times \mathcal{U}_0 \to (-\infty,\infty)$ is also a function symmetric in its two arguments.

Definition 3.2.1 (A general latent space model for signed network $G(n, \mathcal{U}_0, P_u, B, f)$)**.** For $1 \le i \le n$, let $u_i \in \mathcal{U}_0$ be the latent vector independently sampled from the distribution P_u. Given the latent vectors of a pair of nodes i and j, independently of other pairs, an edge between node i and node j is drawn with probability $B(u_i, u_j)$, i.e.,

$$|A_{ij}| \overset{\text{ind.}}{\sim} \text{Bernoulli}(P_{ij}) \;\; \text{with} \;\; P_{ij} = B(u_i, u_j);$$

then for each edge (i.e. $|A_{ij}| = 1$), independently of all others, it takes the positive sign with logit $f(u_i, u_j)$ and the negative sign otherwise, i.e.,

$$\text{logit}(A_{ij} = 1 \Big| |A_{ij}| = 1) = f(u_i, u_j).$$

We write $A \sim G(n, \mathcal{U}_0, P_u, B, f)$ to denote a signed network with n nodes generated from the above procedure.

Note that in the network generative process in Definition 3.2.1, the first part for generating edges covers many existing latent variable models for unsigned networks as special cases by specifying different functions $B(\cdot,\cdot)$; the second part for generating signs further models the sign distribution through specifying the function $f(\cdot,\cdot)$. Given this general class of latent space models for signed networks, next we identify the connection between the population-level balance and the key components of the model (P_u, B, f). As we will see, this connection serves as the foremost step for incorporating the balance theory into modeling signed networks. The following proposition provides a sufficient condition for the symmetric function $f(\cdot,\cdot)$ such that the generated network is population-level balanced.

Proposition 3.2.1. *Suppose a symmetric function $f(\cdot,\cdot)$ satisfies that*

$$f(a,b)\cdot f(b,c)\cdot f(c,a) > 0, \quad \textit{for any } a,b,c \in \mathcal{U}, \tag{3.1}$$

where $\mathcal{U}$ is a subset of $\mathcal{U}_0$ with probability 1, i.e., $P_u(\mathcal{U}) = 1$. Then, a network $A \sim G(n, \mathcal{U}_0, P_u, B, f)$ is population-level balanced.

The proof of Proposition 3.2.1 is based on the fact that $E(A_{ij}|u_i, u_j) > 0$ if and only if the logit $f(u_i, u_j) > 0$ when the probability that an edge appears between node i and node j is nonzero. The details are provided in the Supplemental Material. We note that though there is room for relaxation of the requirement (3.1), its simplicity provides a feasible direction for further analyses on the form of f.

Since not any arbitrary symmetric function f would satisfy (3.1), it is desirable to study what characteristics the function f should have. To this end, we have established the necessary and sufficient conditions in Theorem 3.2.2 for the function f to satisfy (3.1).

Theorem 3.2.2. *For a symmetric function $f(\cdot,\cdot) : \mathcal{U}_0 \times \mathcal{U}_0 \to (-\infty, \infty)$, $f(a,b)\cdot f(b,c)\cdot f(c,a) > 0$ holds for any $a,b,c \in \mathcal{U}$, where $\mathcal{U} \subset \mathcal{U}_0$ with $P_u(\mathcal{U}) = 1$, if and only if*

(i) the function f is positive on $\mathcal{U}\times\mathcal{U}$, i.e., $f(a,b) > 0$ for any $a,b \in \mathcal{U}$; or

(ii) there exists two nonempty subsets S and T, with $S \cup T = \mathcal{U}$ and $S \cap T = \varnothing$, such that $sign(f(a,b)) = \mathbb{1}(a \in S)\cdot\mathbb{1}(b \in S)$ for any $a,b \in \mathcal{U}$, where $\mathbb{1}(event)$ is not the usual indicator function, but rather equals 1 if the event holds and -1 otherwise.

Theorem 3.2.2 implies that, for a function f to satisfy (3.1), if it is not always positive, then $\mathcal{U}$ can be divided into two nonempty disjoint subsets such that the function f is positive when the two arguments belong to the same subset and negative otherwise. On the other hand, if the function f is always positive, it corresponds to a trivial case in which the expected signs between all pairs of nodes are positive. Note that when $\mathcal{U}$ is discrete and finite, Theorem 3.2.2 is a direct result of Harary et al. (1953), while our theorem can be applied to more general latent spaces.

Next, we further illustrate the implication of Theorem 3.2.2 in choosing the function f to describe the population-level balance by taking commonly used latent spaces as examples.

Example 1 The latent space $\mathcal{U}_0$ can be a finite set as in stochastic block models (Abbe, 2018). Let $\mathcal{U}_0 = \{1, \cdots, K\}$ and u_i denotes the community that node i belongs to. Theorem 3.2.2 implies that the K communities can be further combined into two groups and edges tend to be positive within the same group and negative between different groups, as shown in the left side of Figure III.2. We provide a rigorous description for the above result in the following corollary.

Corollary 3.2.1. *For a finite set $\mathcal{U}_0 = \{1, \ldots, K\}$, the symmetric function $f(\cdot, \cdot) : \mathcal{U}_0 \times \mathcal{U}_0 \to (-\infty, \infty)$ satisfies that $f(a, b) \cdot f(b, c) \cdot f(c, a) > 0$ for any $a, b, c \in \mathcal{U}_0$, if and only if there exists a grouping function $g : \{1, \ldots, K\} \to \{-1, 1\}$ and some constants $q_{ab} = q_{ba} > 0$ such that $f(a, b) = q_{ab} \cdot g(a) \cdot g(b)$ holds for $1 \leq a, b \leq K$.*

Note that here the grouping function g identifies two antagonistic groups in the signed network, where nodes from different groups tend to "dislike" each other.

Example 2 The latent space can also be a Euclidean space as in the latent distance model and the latent projection model (Hoff et al., 2002). The following proposition provides an important class of continuous symmetric functions for which the require-

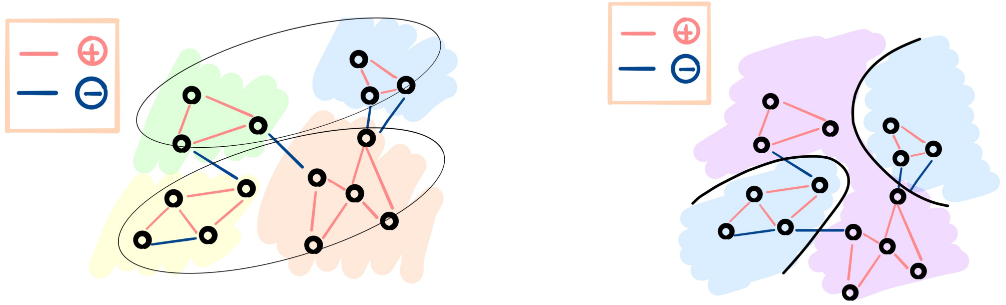

Figure III.2: Illustration of the latent space partition in Theorem 3.2.2. Left: $\mathcal{U}_0$ is a finite set, where each color corresponds to a possible state in $\mathcal{U}_0$ and the two ellipses correspond to the partition. Right: $\mathcal{U}_0$ is a Euclidean space, where the two colors correspond to the partition.

ment (3.1) is satisfied.

Proposition 3.2.2. *For the Euclidean space $\mathcal{U}_0 = \boldsymbol{R}^k$, the requirement (3.1) holds if $f(a,b) = \phi(a)\phi(b)$, where $\phi(\cdot)$ is a real-valued continuous function and $P_u(u : \phi(u) \neq 0) = 1$.*

From the mathematical perspective, Proposition 3.2.2 is an obvious result based on the inequality in (3.1). However, in combination with Theorem 3.2.2, it leads us to a useful and interesting interpretation for the function ϕ. Specifically, $\phi(\cdot)$ can be viewed as the logit (or score) of any binary "classifier" that separates $\mathcal{U}_0$ into two disjoint regions; the function f is positive if the two arguments are classified into the same region and negative otherwise. As shown in the right panel of Figure III.2, the boundary of the classifier tries to cut as many negative edges as possible while retaining most positive edges within the same region.

Remark III.1. *Moreover, our findings in this section can be generalized beyond triangles.*

First, though the balance theory describes patterns for triangles, the notion of population-level balance in Definition 3.1.1 can be generalized to any ℓ-loops. Specifically, an ℓ-loop is defined as a path from a node to itself with length ℓ and is balanced if the product of signs on the loop is positive. We say a network is population-level *loop-balanced if for any $\ell \geq 3$ different nodes $(i_1, \ldots, i_\ell)$,*

$$E(A_{i_1 i_2} \cdots A_{i_{\ell-1} i_\ell} A_{i_\ell i_1} \big| |A_{i_1 i_2} \cdots A_{i_{\ell-1} i_\ell} A_{i_\ell i_1}| = 1) > 0.$$

Second, it is not difficult to extend Proposition 3.2.1 to loop-balance. We can show that a network $A \sim G(n, \mathcal{U}_0, P_u, B, f)$ is population-level loop-balanced if the function f satisfies that for any $\ell \geq 3$,

$$f(a_1, a_2) \cdots f(a_{\ell-1}, a_\ell) f(a_\ell, a_1) > 0, \quad \textit{for any } a_i \in \mathcal{U}, 1 \leq i \leq \ell, \tag{3.2}$$

where $\mathcal{U}$ is a subset of $\mathcal{U}_0$ with probability 1, i.e., $P_u(\mathcal{U}) = 1$. Moreover, as a direct result of Theorem 3.2.2, the necessary and sufficient conditions for a symmetric function f to satisfy (3.2) are the same as those for satisfying (3.1). This implies that, for a function f satisfying (3.1), the network $A \sim G(n, \mathcal{U}, P_u, B, f)$ is not only triangle-balanced but also loop-balanced at the population level.

Third, the definition of population-level (loop-)balance can be further generalized to a general weighted network as the balance theory only focuses on the sign of the product. Correspondingly, we require that $E(sign(A_{ij} A_{j\ell} A_{\ell i}) \big| A_{ij} A_{j\ell} A_{\ell i} \neq 0) > 0$ for any three different nodes (i, j, ℓ). These generalizations provide flexibility for modeling real-world networks.

3.3 Balanced Inner-Product Models

Motivated by the key finding in Theorem 3.2.2, we propose two inner-product models for signed networks that fall within the general class of latent space models in Definition 3.2.1. Both models are inherently of population-level balance. We first present the separate inner-product model and then introduce the joint inner-product model by adding an additional structural assumption. We will demonstrate the usefulness of this structural assumption in estimating the latent variables both theoretically (if it is correctly specified) and empirically in subsequent sections.

3.3.1 Separate Inner-Product Model

We assume that for any $1 \leq i < j \leq n$, we have

$$|A_{ij}| = |A_{ji}| \overset{\text{ind.}}{\sim} \text{Bernoulli}(P_{ij}), \quad \text{with} \;\; \text{logit}(P_{ij}) = \Theta_{ij} = \alpha_i + \alpha_j + z_i^\top z_j, \qquad (3.3)$$

where $\alpha_i \in \mathbf{R}$ and $z_i \in \mathbf{R}^k$ (for $i = 1, \ldots, n$) are latent variables. Further, independently of all others, we also assume

$$\text{logit}(A_{ij} = 1 \Big| |A_{ij}| = 1) = \eta_{ij} = v_i v_j, \qquad (3.4)$$

where $v_i \in \mathbf{R}$ for $i = 1, \ldots, n$ are also latent variables.

The proposed separate inner-product model has the capacity to capture various commonly observed characteristics of signed networks. Specifically, the parameter α_i enables modeling node degree heterogeneity, of which a larger value leads to higher probability of connecting with other nodes given other parameters fixed. Thus, we call $\{\alpha_i\}_{i=1}^n$ *degree heterogeneity parameters*. Next, the inner-product formation between the *latent position vectors* z_i and z_j inherently models transitivity, i.e., nodes with common neighbors (regardless of friend or enemy) are more likely to be linked. Because the closer the latent position vectors of two nodes are in the latent space, the higher inner product it is and more likely to connect with each other. Finally, the parameters $\{v_i\}_{i=1}^n$ model the distribution of signs through their product, which satisfies the sufficient condition (3.1) for the population-level balance. In particular, an edge between two nodes tends to have a positive sign when their latent variables v_i and v_j have the same sign and a negative sign otherwise. Moreover, the magnitude of v_i controls the discrepancy level between positive and negative signs. Therefore, we name them as *latent polar variables*. When all latent polar variables are zeros, negative and positive signs are exchangeable.

Following Ma et al. (2020), which is for unsigned networks, we impose no distributional assumptions (prior) on the latent variables (α_i, z_i, v_i) for the sake of modeling flexibility and estimation scalability, in comparison to treating them as random and using Bayesian estimation approaches as in existing works (Krivitsky et al., 2009).

For presentation simplicity, we rewrite the model in matrix form. Specifically, we have

$$\Theta = \alpha 1_n^\top + 1_n \alpha^\top + ZZ^\top, \quad \eta = vv^\top,$$

where $\alpha = (\alpha_1, \cdots, \alpha_n)$, 1_n is the all one vector in $\mathbf{R}^n$, $Z = (z_1, \cdots, z_n)^\top \in \mathbf{R}^{n\times k}$, and $v = (v_1, \cdots, v_n)^\top \in \mathbf{R}^n$.

Identifiability To ensure identifiability of parameters (α, Z, v), we provide additional constraints in Proposition 3.3.1. Given centered latent position variables, that is $J_n Z = Z$, where $J_n = I_n - \frac{1}{n} 1_n 1_n^\top$, the parameters are identifiable up to an orthogonal transformation and a sign flipping.

Proposition 3.3.1. *Suppose two sets of parameters (α, Z, v) and $(\bar{\alpha}, \bar{Z}, \bar{v})$ satisfy that A1) $J_n Z = Z$ and $J_n \bar{Z} = \bar{Z}$; A2) $Z \in \boldsymbol{R}^{n\times k}$ is of full rank. Then, they specify the same network distribution through (3.3) and (3.4) if and only if there exist an orthogonal matrix $O \in \boldsymbol{R}^{k\times k}$ with $O^\top O = OO^\top = I_k$ and $\kappa \in \{-1, 1\}$ such that $\alpha = \bar{\alpha}, Z = \bar{Z}O, v = \kappa \bar{v}$.*

3.3.2 Joint Inner-Product Model

Based on the above separate inner-product model, we further consider the dependency of the latent polar variable v_i on the latent position variable z_i. The idea of introducing their relationship originates naturally from Proposition 3.2.2, where we view the latent polar variable v_i as a function of the latent position variable z_i, i.e., $v_i = \phi(z_i)$ with some link function ϕ. Modeling such a link function ϕ can provide more structural information of the network. On the other hand, there are flexibilities in choosing the family of link function ϕ, which would lead to different shapes of the latent space partition derived by ϕ. For the scope of this chapter, we assume ϕ is a linear function in z_i in the joint inner-product model, i.e., $v_i = w^\top z_i + \gamma$ with $w \in \mathbf{R}^k$

and $\gamma \in \mathbf{R}$, and discuss other nonlinear alternatives in Remark III.2. Specifically, the joint inner-product model is given by (3.3) and replaces (3.4) with

$$\operatorname{logit}(A_{ij} = 1 \Big| |A_{ij}| = 1) = \eta_{ij} = (w^\top z_i + \gamma)(w^\top z_j + \gamma). \tag{3.5}$$

In particular, the hyperplane $\{z \in \mathbf{R}^k : w^\top z + \gamma = 0\}$ separates the latent space into two regions. A pair of nodes tend to have a positive edge when their latent positions are located on the same side of the hyperplane and have a negative edge when their latent positions are located on different sides of the hyperplane. If $w = 0$ and $\gamma \neq 0$, the sign of each edge has a homogeneous logit γ^2 to be positive.

Identifiability For the joint inner-product model, the identifiability condition for parameters (α, Z, w, γ) is established correspondingly in Proposition 3.3.2.

Proposition 3.3.2. *Suppose two sets of parameters (α, Z, w, γ) and $(\bar{\alpha}, \bar{Z}, \bar{w}, \bar{\gamma})$ satisfy that A1) $J_n Z = Z$ and $J_n \bar{Z} = \bar{Z}$; A2) $Z \in \boldsymbol{R}^{n\times k}$ is of full rank. Then, they specify the same network distribution through (3.3) and (3.5) if and only if there exist an orthogonal matrix $O \in \boldsymbol{R}^{k\times k}$ with $O^\top O = OO^\top = I_k$ and $\kappa \in \{-1, 1\}$ such that $\alpha = \bar{\alpha}, Z = \bar{Z}O, w = \kappa O^\top \bar{w}, \gamma = \kappa\bar{\gamma}$.*

Remark III.2. *Though we use a linear link function in the joint inner-product model (3.5), more flexible nonlinear functions can be considered. For example, we may assume ϕ belongs to a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ associated with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ under which $\mathcal{H}$ is complete. There is a positive semidefinite kernel function $\mathbb{K}(\cdot, \cdot) : \boldsymbol{R}^k \times \boldsymbol{R}^k \to \boldsymbol{R}_+$ such that $\phi(z_i) = \langle \phi, \mathbb{K}(\cdot, z_i) \rangle_{\mathcal{H}}$. Multiple choices of RKHS are available for practical use, including those with polynomial kernel, Gaussian kernel, and Laplacian kernel (Scholkopf and Smola, 2018).*

3.4 Model Estimation

In this section, we develop two methods for fitting the proposed models (3.3)-(3.5). Both methods minimize the negative log-likelihood function of the balanced inner-product models through projected gradient descent.

Under balanced inner-product models, the negative log-likelihood function consists of two parts. The first part is derived from the probability of forming edges:

$$\mathcal{L}_e(\alpha, Z) = \sum_{i<j} \Big\{ |A_{ij}|\Theta_{ij} + \log(1 - \sigma(\Theta_{ij})) \Big\},$$

where $\Theta = \alpha 1_n^\top + 1_n \alpha^\top + ZZ^\top$, and $\sigma(x) = 1/(1 + \exp(-x))$ is the sigmoid function, which is the inverse of the logit function. The second part is derived from the probability of assigning signs:

$$\mathcal{L}_s(v) = \sum_{i<j} \Big\{ |A_{ij}| \frac{1 + A_{ij}}{2} \eta_{ij} + |A_{ij}| \log(1 - \sigma(\eta_{ij})) \Big\},$$

where $\eta = vv^\top$, and when under the joint inner-product model, we further have $v = Zw + \gamma 1_n$, or equivalently, v belongs to the column space of $(1_n, Z)$.

The first method estimates parameters (α, Z) and v separately by minimizing $\mathcal{L}_e(\alpha, Z)$ and $\mathcal{L}_s(v)$ respectively. Hence we name it the separate estimation method. Note that the separate estimation method does not depend on a specific relationship between the latent polar variables v and the latent position vectors Z. Therefore, the separate estimation method can always be applied regardless of the underlying link function ϕ. Alternatively, we also propose a joint estimation method tailored for the joint inner-product model, which exploits the structural assumption for more accurate estimation. Specifically, we jointly estimate parameters (α, Z, v) by minimizing a weighted sum of $\mathcal{L}_e(\alpha, Z)$ and $\mathcal{L}_s(v)$, while constraining v to be in the column space of $(1_n, Z)$.

Notation Before presenting the algorithm details, we first introduce the following general notations to be used hereafter. For any $X \in \mathbf{R}^{d_1 \times d_2}$, X_{i*} and X_{*j} denote the

i-th row and j-th column of matrix X respectively, and for any function $\omega(\cdot)$, $\omega(X)$ represents applying the function $\omega(\cdot)$ element-wisely to X, that is $\omega(X) \in \mathbf{R}^{d_1 \times d_2}$ and $[\omega(X)]_{ij} = \omega(X_{ij})$. We use $\circ$ to denote the Hadamard product, that is, for any two matrices $X, Y \in \mathbf{R}^{d_1 \times d_2}$, $X \circ Y \in \mathbf{R}^{d_1 \times d_2}$ and $[X \circ Y]_{ij} = X_{ij} Y_{ij}$. Moreover, we use $\|X\|_F$, $\|X\|_{op}$, $\|X\|_*$, and $\|X\|_{\max}$ to denote the Frobenius norm, the operator norm, the nuclear norm, and the max norm of a matrix respectively. We use $col(X)$ to denote the column space of X. For a vector $x \in \mathbf{R}^d$, we use $\|x\|$ to denote the Euclidean norm.

3.4.1 Separate Estimation Method

First, to estimate parameters (α, Z), we solve the non-convex optimization problem below:

$$\min_{\alpha \in \mathbf{R}, Z \in \mathbf{R}^{n \times k}} -\sum_{i,j} \Big\{ |A_{ij}| \Theta_{ij} + \log(1 - \sigma(\Theta_{ij})) \Big\}, \tag{3.6}$$

subject to $\Theta = \alpha 1_n^\top + 1_n \alpha^\top + ZZ^\top$ and $Z = J_n Z$. In particular, the signed adjacency matrix enters the objective function through its absolute value, which leads to the same optimization problem studied in Ma et al. (2020) when there is no edge covariate. Here we adopt the projected gradient descent algorithm along with the initialization method proposed in Ma et al. (2020) because of their theoretical guarantee and scalability to large networks. We provide the detailed description of the method in Algorithm B.1 and the initialization algorithm in the Supplemental Material.

Next, to estimate the latent polar variables v, we solve another non-convex optimization problem, i.e.,

$$\min_{v \in \mathbf{R}^n} -\sum_{i,j} |A_{ij}| \Big\{ \frac{1 + A_{ij}}{2} \eta_{ij} + \log(1 - \sigma(\eta_{ij})) \Big\} \text{ subject to } \eta = vv^\top. \tag{3.7}$$

Similarly, we develop a fast gradient descent algorithm, which is summarized in Algorithm B.2. We also use an initialization algorithm based on the universal singular value thresholding proposed by Chatterjee (2015) (see the Supplemental Material).

Remark III.3. *We note that, although we use gradient descent algorithms for estimating both (α, Z) and v, the subtle difference in their objectives makes the theory in Ma et al. (2020) not directly applicable to (3.7). Specifically, unlike the objective in (3.6), not all elements of the signed adjacency matrix contribute to the objective in (3.7). Instead, only nonzero entries, i.e., $\{(i,j) : |A_{ij}| = 1\}$, are used for inferring the latent polar variables through (3.7). In this case, one key step in building the improvement in errors of iterates in Ma et al. (2020) no longer holds. Therefore, we establish a new error bound for Algorithm B.2 (see Section 3.5).*

Remark III.4. *We also note that our optimization problem in (3.7) is closely related to the line of research on low-rank matrix estimation. See Koltchinskii et al. (2011); Candes et al. (2013); Davenport et al. (2014); Chen and Wainwright (2015); Zheng and Lafferty (2016); Wang et al. (2017) for a sample of references. In particular, (3.7) can be viewed as a one-bit matrix completion problem, where we observe a random subset of binary entries generated from a distribution determined by a low-rank matrix. To solve this problem, Davenport et al. (2014) considered a convex relaxation that replaces the low-rank constraint by the nuclear norm penalization. Though it becomes a convex optimization problem, in general, solving such a nuclear-norm penalized optimization problem requires singular value decomposition at each iteration, which is computationally expensive for large matrices. Alternatively, gradient descent algorithms have been used for improving the computational efficiency. Chen and Wainwright (2015) and Wang et al. (2017) have established convergence guarantees and statistical errors for the gradient descent algorithms in application to low-rank matrix estimation problems, which particularly cover the one-bit matrix completion problem. However, theories in aforementioned works are based on the uniform random sampling assumption, i.e., each entry of the matrix is observed independently with a uniform probability p, while in our case, entries are observed with different probabilities P_{ij}. Thus, our theoretical analysis of the proposed gradient descent algorithm* *in Section 3.5 provides new results for one-bit matrix completion under non-uniform random sampling.*

3.4.2 Joint Estimation Method

Under the joint inner-product model (3.4)-(3.5), we propose to jointly estimate parameters (α, Z, v) by re-parameterizing $v = Zw + \gamma 1_n$ with $w \in \mathbf{R}^k$ and $\gamma \in \mathbf{R}$. By introducing a hyperparameter λ, we minimize the following weighted negative log-likelihood,

$$\mathcal{L}_\lambda(\alpha, Z, w, \gamma) = -\sum_{i,j} \Big\{ (1-\lambda)\big[|A_{ij}|\Theta_{ij} + \log(1-\sigma(\Theta_{ij}))\big] \\ + \lambda |A_{ij}|\big[\frac{1+A_{ij}}{2}\eta_{ij} + \log(1-\sigma(\eta_{ij}))\big]\Big\},$$

subject to $\Theta = \alpha 1_n^\top + 1_n \alpha^\top + ZZ^\top$, $Z = JZ$, and $\eta = (Zw + \gamma 1_n)(Zw + \gamma 1_n)^\top$.

Here λ controls the weight of relative information from the edge formation and the sign assignment respectively. In particular, when $\lambda = 0$, no information from the edge signs is used and the joint estimation reduces to the separate estimation for (α, Z) in (3.6). Later in Section 3.5.2, we will theoretically show that, under certain conditions, any positive λ below some threshold yields more accurate estimation of latent position variables Z than the separate estimation (i.e., $\lambda = 0$), but the magnitude of the improvement depends on the choice of λ. In principle, we can select λ in a data-driven manner by performing cross-validation on the observed signed adjacency matrix, where we randomly mask a subset of entries, fit the joint inner-product model by using the remaining entries, repeat the process multiple times, and then select λ from a candidate set with the best average prediction performance on the holdout entries. In practice, we find simply setting $\lambda = 1/2$ also works generally well, in which case the solution becomes the usual maximum likelihood estimator.

To solve the above constrained minimization problem, we develop a projected gradient descent algorithm, whose details are given in Algorithm III.1. Similar to

Algorithm III.1: The projected gradient descent algorithm for joint estimation

```
  Input: signed adjacency network $A \in R^{n\times n}$, latent space dimension $K \geq 1$,
         number of iterations $T$, initial values $(\alpha_0, Z_0, w_0, \gamma_0)$ with
         $(w_0, \gamma_0) = \arg\min_{w\in\mathbf{R}^k, \gamma\in\mathbf{R}} \mathcal{L}_\lambda(\alpha_0, Z_0, w, \gamma)$, step sizes $(\tau_\alpha, \tau_z)$
1 for $t = 0, \ldots, T-1$ do
2     $\tilde{Z}_{t+1} = Z_t - \tau_z \boldsymbol{\nabla}_Z \mathcal{L}_\lambda =$
        $Z_t + 2\tau_z\big[(1-\lambda)(|A| - \sigma(\Theta_t))Z_t + \lambda(|A| \circ (A+1)/2 - |A| \circ \sigma(\eta_t))v_t w_t^\top\big]$;
3     $\alpha_{t+1} = \alpha_t - \tau_\alpha \boldsymbol{\nabla}_\alpha \mathcal{L}_\lambda = \alpha_t + 2\tau_\alpha(1-\lambda)(|A| - \sigma(\Theta_t))1_n$;
4     $Z_{t+1} = J\tilde{Z}_{t+1}$;
5     $(w_{t+1}, \gamma_{t+1}) = \arg\min_{w\in\mathbf{R}^k, \gamma\in\mathbf{R}} \mathcal{L}_\lambda(\alpha_{t+1}, Z_{t+1}, w, \gamma)$;
6     $v_{t+1} = Z_{t+1}w_{t+1} + \gamma_{t+1}1_n$;
7 end
  Output: $(\hat{\alpha}, \hat{Z}, \hat{v}) = (\alpha_T, Z_T, v_T)$
```

Algorithm B.1, we first update α and Z by moving against their gradients. In particular, when updating Z in line 2, the gradient comes from not only the edge formation likelihood but also the sign assignment likelihood, whose weights are adjusted by λ. Then, after centering Z's columns, we update (w, γ) by minimizing the objective while fixing the current estimate of Z. Though the algorithm involves an inner non-convex optimization in line 5, we find that, in practice, a simple one-step gradient descent provides an effective approximation, i.e., replacing line 5 by $w_{t+1} = w_t + 2\tau_w\lambda Z_t^\top(|A| \circ (A+1)/2 - |A| \circ \sigma(\eta_t))v_t$ and $\gamma_{t+1} = \gamma_t + 2\tau_\gamma\lambda 1_n^\top(|A| \circ (A+1)/2 - |A| \circ \sigma(\eta_t))v_t$. We initialize the algorithm by $(\alpha_0, Z_0) = (\hat{\alpha}, \hat{Z})$ obtained from Algorithms B.1 and $(w_0, \gamma_0) = \arg\min_{w\in\mathbf{R}^k, \gamma\in\mathbf{R}} \mathcal{L}_\lambda(\alpha_0, Z_0, w, \gamma)$, and set $\tau_Z = \tau / \max\{\|Z_0\|_{op}^2, \|v_0\|^2\}$, $\tau_\alpha = \tau/(2n)$, and set $\tau_w = \tau/(\|Z_0\|_{op}^2\|v_0\|^2)$, $\tau_\gamma = \tau/(n\|v_0\|^2)$ when using one-step gradient descent approximation for line 5.

3.5 Theoretical Results

In this section, we establish high probability error bounds for the proposed two estimation methods. Note for the separate estimation method, the error bound for estimating latent position vectors Z under model (3.3) and that for estimating latent polar variables v under model (3.4) are derived separately. Thus, the separate estimation method is robust in the sense that, when one of models (3.3) and (3.4) is mis-specified, our theoretical results still hold for the other. On the other hand, for the joint estimation method that utilizes the relationship between v and Z, we further discuss how incorporating their dependency can help reduce the estimation error of latent variables under the joint inner-product model (3.5).

3.5.1 Results for the Separate Estimation Method

We present theoretical guarantees of Algorithms B.1 and B.2 under the separate inner-product model (3.3) and model (3.4) respectively. Note that the error bound for the outputs of Algorithm B.1 is a straightforward result of Ma et al. (2020, Theorem 9) when there is no edge covariate; we adjust it in Proposition 3.5.1 for presentation coherence. Nonetheless, their theory cannot be directly applied to the setting of Algorithm B.2, because only nonzero entries of the signed adjacency matrix are included in the objective (3.7), which breaks an important step towards establishing the estimation improvements for successive iterations in their proof. Therefore, our established error bound for the outputs of Algorithm B.2 is a new result for a more general setting, where entries are observed with non-uniform probabilities.

We describe error bounds for the outputs of Algorithms B.1 and B.2 with details below. We firstly define the parameter spaces as

$$\begin{aligned}\mathcal{F}_\theta(n,k,M_1,M_2) =& \Big\{\alpha \in \mathbf{R}^n, Z \in \mathbf{R}^{n\times k}, \Theta \in \mathbf{R}^{n\times n} \mid \Theta = \alpha 1_n^\top + 1_n\alpha^\top + ZZ^\top, J_n Z = Z, \\ & \max_{1\le i\le n} \|Z_{i*}\|^2, 2\|\alpha\|_{\max} \le \frac{M_1}{2}, \max_{1\le i\neq j\le n} \Theta_{ij} \le -M_2\Big\} \qquad (3.8)\end{aligned}$$

and

$$\mathcal{F}_\eta(n,M_3) = \Big\{v \in \mathbf{R}^n, \eta \in \mathbf{R}^{n\times n} \mid \eta = vv^\top, \|v\|_{\max}^2 \le M_3\Big\}. \qquad (3.9)$$

We allow k, M_1, M_2, and M_3 in (3.8)-(3.9) to change with the network size n similarly as in Ma et al. (2020). Note that, given the inequalities in (3.8), it is straightforward to see that, for any $\Theta \in \mathcal{F}_\theta(n, k, M_1, M_2)$, we have $-M_1 \leq \Theta_{ij} \leq -M_2$ for $1 \leq i \neq j \leq n$. Therefore, M_2, as the upper bound of logit-transformed probabilities of observing edges, controls the network sparsity, of which a larger value leads to a sparser network. The true parameters are denoted by (α^*, Z^*, v^*), $\Theta^* = \alpha^* 1_n^\top + 1_n {\alpha^*}^\top + Z^* {Z^*}^\top$, and $\eta^* = v^* {v^*}^\top$.

Error bound for Algorithm B.1 Let (α_t, Z_t) be the updated parameters at the t-th iteration in Algorithm B.1 and $\Theta_t = \alpha_t 1_n^\top + 1_n {\alpha_t}^\top + Z_t {Z_t}^\top$. Since the latent position vectors $Z \in \mathbf{R}^{n \times k}$ are identifiable up to an orthogonal transformation, we define the distance between two latent matrices Z_1 and Z_2 as $dist(Z_1, Z_2) = \min_{O \in O(k)} \|Z_1 - Z_2 O\|_F$, where $O(k)$ is the collection of all orthogonal matrices in $\mathbf{R}^k$. Let $O_t = \arg\min_{O \in O(k)} \|Z_t - Z^* O\|_F$, $\Delta_{Z_t} = Z_t - Z^* O_t$, and $\Delta_{\Theta_t} = \Theta_t - \Theta^*$.

For theoretical justification, in Algorithm B.1, we further assume projection onto the constraint sets $\mathcal{C}_Z = \{Z \in \mathbf{R}^{n \times k}, J_n Z = Z, \max_{1 \leq i \leq n} \|Z_{i*}\|^2 \leq M_1/2\}$ and $\mathcal{C}_\alpha = \{\alpha \in \mathbf{R}^n, 2\|\alpha\|_{\max} \leq M_1/2\}$ at each iteration. The following proposition establishes the high probability error bounds for estimating both the latent position matrix Z and the logit-transformed probability matrix Θ.

Proposition 3.5.1. *Set the step sizes as $\tau_z = \tau/\|Z_0\|_{op}^2$, $\tau_\alpha = \tau/(2n)$ for any $\tau \leq c$ where $c > 0$ is a universal constant. Suppose 1) the initializers α_0, Z_0 in Algorithm B.1 satisfy $\|Z^*\|_{op}^2 \|\Delta_{Z_0}\|_F^2 + \|\Delta_{\alpha_0} 1_n^\top\|_F^2 \leq c_0 e^{-2M_1} \|Z^*\|_{op}^4 / \kappa_{Z^*}^4$ for a sufficiently small positive constant c_0, where κ_{Z^*} is the conditional number of Z^*; and 2) $\|Z^*\|_{op}^2 \geq C_1 \kappa_{Z^*}^2 \sqrt{n} e^{M_1 - M_2/2} \max\{\sqrt{\tau k} e^{M_1}, 1\}$ for a sufficiently large constant C_1. Then there exist positive constants ρ, c_1, and C uniformly over $\mathcal{F}_\theta(n, k, M_1, M_2)$ such that, with probability at least $1 - n^{-c_1}$, we have*

$$\|Z^*\|_{op}^2 \|\Delta_{Z_T}\|_F^2,\ \|\Delta_{\Theta_T}\|_F^2 \leq C \kappa_{Z^*}^2 e^{2M_1} nk \cdot \max\{e^{-M_2}, \frac{\log n}{n}\},$$

for some $T \leq \log\Big(\frac{M_1^2}{\kappa_{Z^*}^2 e^{4M_1 - M_2}} \frac{n}{k^3}\Big) / \log\Big(1 - \frac{\tau}{e^{M_1} \kappa_{Z^*}^2} \rho\Big)^{-1}$.

Error bound for Algorithm B.2 Let v_t be the updated parameters at the t-th iteration in Algorithm B.2 and $\eta_t = v_t {v_t}^\top$. Similarly, as the latent polar variables $v \in \mathbf{R}^n$ are identifiable up to a sign, we define the distance between two latent vectors v_1 and v_2 as $dist(v_1, v_2) = \min_{\kappa \in \{-1,1\}} \|v_1 - \kappa v_2\|$. Let $\kappa_t = \arg\min_{\kappa \in \{-1,1\}} \|v_t - \kappa v^*\|$ and $\Delta_{v_t} = v_t - \kappa_t v^*$, and further let $\Delta_{\eta_t} = \eta_t - \eta^*$.

Although the error bound presented below does not rely on a specific generating process of edges such as in model (3.3) and the parameter space $\mathcal{F}_\theta(n, k, M_1, M_2)$ in (3.8), it depends on the lower bound of the probability of observing an edge. For notation consistency, we use M_1 to denote the lower bound of the logit-transformed probability matrix, i.e., $\Theta_{ij} \geq -M_1$ for $1 \leq i, j \leq n$. Similarly, for theoretical justification, we constrain v to be in the set $\mathcal{C}_v = \{v \in \mathbf{R}^n, \|v\|_{\max}^2 \leq M_3\}$ at each iteration in Algorithm B.2. The following theorem establishes the high probability error bounds for estimating the latent polar variables v and the logit-transformed probability matrix η.

Theorem 3.5.1. *Set the step size as $\tau_v = \tau/\|v_0\|^2$ for any $\tau \leq c$, where $c > 0$ is a universal constant. Suppose 1) the initializer v_0 in Algorithm B.2 satisfy $\|\Delta_{v_0}\| \leq c_0 e^{-(M_1+M_3)/2}\|v^*\|$ for a sufficiently small positive constant c_0; and 2) $\|v^*\|^2 \geq C_1\sqrt{n}$ $e^{M_1+M_3}\max\{\sqrt{\tau e^{M_1+M_3}}, 1\}$ for a sufficiently large constant C_1. Then there exist positive constants ρ, c_1, and C uniformly over $\mathcal{F}_\eta(n, M_3)$ and M_1 such that, with probability at least $1 - n^{-c_1}$, we have*

$$\|v^*\|^2\|\Delta_{v_T}\|^2,\ \|\Delta_{\eta_T}\|_F^2 \leq Ce^{2(M_1+M_3)}n,$$

for some $T \leq \log\left(\frac{M_3^2}{e^{3(M_1+M_3)}}n\right) / \log\left(1 - \frac{\tau}{e^{M_1+M_3}}\rho\right)^{-1}$.

Theorem 3.5.1 implies that the mean square error $\|\Delta_{\eta_T}\|_F^2/n^2$ is of order $\mathcal{O}(1/n)$, which coincides with the existing error rate for one-bit rank-1 matrix completion problems (Davenport et al., 2014; Chen and Wainwright, 2015), while our result can be viewed as their extension to the case where entries of the one-bit matrix are randomly observed with non-uniform probabilities. In particular, for the more general non-uniform case, the key ingredient in our proof is to derive a lower bound of the sampling operator $|A| \in \{0,1\}^{n\times n}$. We prove that the sampling operator $|A|$ has a positive lower bound, i.e., $\||A| \circ \eta\|_F \geq c\|\eta\|_F$ with some $c > 0$, as long as η belongs to a specific data-dependent set. This positive lower bound enables us to extend the proof in Ma et al. (2020) when establishing iterative improvements. The proof of Theorem 3.5.1 is given in the Supplemental Material.

Remark III.5. *Note that the error bounds for v and η in Theorem 3.5.1 hold regardless of the concrete form of model (3.3). Therefore, the above results still hold even if model (3.3) is mis-specified. But, it still depends on the lower bound of the probability matrix of observing edges. The error bound implies that as M_1 gets larger, the error bound also becomes larger. Intuitively, when the lower bound of Θ decreases, there might be fewer observed edges in the network and thereby the estimation errors for v and η would be larger due to the lack of observations.*

Remark III.6. *The assumptions in both Proposition 3.5.1 and Theorem 3.5.1 require relatively good initializations of (α, Z, v). We note that the conditions for α_0 and Z_0 can be achieved with theoretical justification by the universal-singular-value-thresholding (Chatterjee, 2015) based initialization algorithm proposed in Ma et al. (2020). We further extend this algorithm to initialize v_0. Based on our simulation studies (see the Supplemental Material), we find that simple random initialization also achieves similar estimation errors after the algorithm converges while requiring more iterations for algorithm convergence.*

3.5.2 Results for the Joint Estimation Method

We first present the convergence guarantee and the error bound for the estimators obtained by Algorithm III.1. Then we further investigate how the joint estimation method could improve the estimation of Z on top of the separate estimation.

Under the joint inner-product model, we redefine the parameter space as

$$\mathcal{F}(n,k,M_1,M_2,M_3) = \Big\{\alpha, v \in \mathbf{R}^n, Z \in \mathbf{R}^{n\times k}, \Theta, \eta \in \mathbf{R}^{n\times n} \;\Big|$$
$$\Theta = \alpha 1_n^\top + 1_n\alpha^\top + ZZ^\top, J_n Z = Z, \eta = vv^\top, v = Zw + \gamma 1_n,$$
$$\max_{1\le i\le n} \|Z_{i*}\|^2, \|\alpha\|_{\max} \le \frac{M_1}{2}, \max_{1\le i\ne j\le n} \Theta_{ij} \le -M_2, \|v\|_{\max}^2 \le M_3, \|w\| \le M, |\gamma| \le M'\Big\},$$

where k, M_1, M_2, and M_3 are allowed to change with the network size n. Let (α^*, Z^*, v^*) be the true parameters, where $v^* = Z^*w^* + \gamma^* 1_n$ with some $w^* \in \mathbf{R}^k$ and $\gamma^* \in \mathbf{R}$.

Error bound for Algorithm III.1 Let (α_t, Z_t, v_t) be the updated parameters at the t-th iteration in Algorithm III.1. We assume the projection onto the same constraint sets $\mathcal{C}_\alpha$, $\mathcal{C}_Z$, and $\mathcal{C}_v$ at the end of each iteration as those for Algorithms B.1 and B.2. The following theorem first guarantees that the error of iterates $\{(\alpha_t, Z_t)\}_{t\ge 0}$ converges up to a statistical error and then gives the high probability error bounds for the estimators of Z and Θ.

Theorem 3.5.2. *Set the step sizes as $\tau_Z = r_0\tau/\|Z_0\|_{op}^2$, $\tau_\alpha = \tau/(2n)$, and the weight $\lambda = \tilde{\lambda} r_0 / e^{M_1}\kappa_{Z^*}^2$ with $r_0 = \min\{1, \|Z_0\|_{op}^2/\|v_0\|^2\}$ for any $\tau \le c_\tau$, $\tilde{\lambda} \le c_\lambda$, where c_τ and c_λ are universal constants. Let $\zeta_n = \max\{\||A| - P\|_{op}, 1\}$ and $\varphi_n = \max\{\||A| \circ ((1+A)/2 - Q)\|_{op}, 1\}$. Denote the error metric for iterates as $\tilde{e}_t^Z = \|\Delta_{Z_t}\|_F^2\|Z_0\|_{op}^2 + \|\Delta_{\alpha_t} 1_n^\top\|_F^2$. Suppose the initializers α_0, Z_0 in Algorithm III.1 satisfy $\tilde{e}_0^Z \le c_0 e^{-2M_1 - 3M_3}\|Z^*\|_{op}^4/\kappa_{Z^*}^4$ for a sufficiently small positive constant c_0, where κ_{Z^*} is the conditional number of Z^*. Then, we have*

1. *(Deterministic bounds for iterative errors) If $\|Z^*\|_{op}^2 \ge C_0 e^{M_1}\kappa_{Z^*}^2\zeta_n \max\{1, \sqrt{\tau k} e^{M_1 + 3M_3/2}\kappa_{Z^*}\}$ and $\|v^*\|^2 \ge C_0 e^{M_1+M_3}\varphi_n \max\{\sqrt{\tau} e^{M_1/2 + M_3}, 1\}$ for a sufficiently large constant C_0, then there exist universal positive constants ρ_1, ρ_2,*

C', and C'' such that for all $t \geq 0$

$$\tilde{e}^Z_{t+1} \leq \left(1 - \frac{r_0\tau\rho_1}{e^{M_1}\kappa^2_{Z^*}}\right)\tilde{e}^Z_t - \lambda\frac{r_0\tau\rho_2}{e^{M_3}}\min\{\||A| \circ \Delta_{\eta_t}\|_F^2, e^{-M_1}\|\Delta_{\eta_t}\|_F^2\}$$
$$+ r_0\tau C' e^{M_1}\zeta_n^2 k + \lambda r_0 \tau C'' e^{M_1+M_3}\varphi_n^2.$$

2. *(High-probability bounds) Suppose $\|Z^*\|_{op}^2 \geq C_0 e^{M_1 - M_2/2}\kappa^2_{Z^*}\sqrt{n}\max\{1, \sqrt{\tau k}e^{M_1+3M_3/2}\kappa_{Z^*}\}$ and $\|v^*\|^2 \geq C_0 e^{M_1+M_3}\sqrt{n}\max\{\sqrt{\tau}e^{M_1/2+M_3}, 1\}$ for a sufficiently large constant C_0. Then there exist positive constants ρ_1, c, and C uniformly over $\mathcal{F}(n, k, M_1, M_2, M_3)$ such that, with probability at least $1 - n^{-c}$, we have*

$$\|Z^*\|_{op}^2\|\Delta_{Z_T}\|_F^2,\ \|\Delta_{\Theta_T}\|_F^2 \leq C\kappa^2_{Z^*}e^{2M_1}nk \cdot \max\{e^{-M_2}, \frac{\log n}{n}, e^{M_3-M_1}\frac{1}{\kappa^2_{Z^*}k}\},$$

for some $T \leq \log\left(\frac{M_1^2}{\kappa^2_{Z^*}e^{4M_1+3M_3-M_2}}\frac{n}{k^3}\right) / \log\left(1 - \frac{r_0\tau\rho_1}{e^{M_1}\kappa^2_{Z^*}}\right)^{-1}$.

The first part of Theorem 3.5.2 indicates that, compared to the separate estimation method, the joint method involving the gradient of the sign likelihood leads to an extra improvement on the error bound of iterates, which depends on Δ_{η_t}, while introducing another statistical error term φ_n. As a result, in the second part, the high probability error bounds depend on the maximum of three terms, among which the first two are the same as in Proposition 3.5.1 and the third is resulted from φ_n. When $M_3 \leq M_1 - M_2 + \log(\kappa^2_{Z^*}k)$, the maximum multiplier reduces to the one in Proposition 3.5.1. Overall, the error bounds of Z and Θ for the joint estimation method are still in the order $\mathcal{O}(nk)$, which is the same as that for the separate estimation method. In addition, the following corollary gives the error bounds of v_T and η_T obtained from the line 5 in Algorithm III.1.

Corollary 3.5.1. *For $v_t = Z_t w_t + \gamma_t 1_n$ with $(w_t, \gamma_t) = \arg\min_{w \in \boldsymbol{R}^k, \gamma \in \boldsymbol{R}} \mathcal{L}_\lambda(\alpha_t, Z_t, w, \gamma)$, we have $\|\Delta_{\eta_t}\|_F \leq 16 e^{M_1+M_3}\max\{\zeta_n, \varphi_n\} + e^{M_1/2+M_3}(2 + \|\Delta_{Z_t}\|_F\|w^*\|)\|v^*\|\|\Delta_{Z_t}\|_F\|w^*\|$ for $t \geq 0$. Suppose the conditions for high probability bounds in Theorem 3.5.2 hold, then there exist positive constants ρ_1, c, and C uniformly over $\mathcal{F}(n, k, M_1, M_2, M_3)$*

such that, with probability at least $1 - n^{-c}$*, we have*

$$\|v^*\|^2\|\Delta_{v_T}\|^2, \|\Delta_{\eta_T}\|^2 \le Ce^{3M_1+2M_3}nk \cdot \max\{\frac{e^{M_3-M_1}}{k}, \kappa_{Z^*}^2 \max\{e^{-M_2}, \frac{\log n}{n}\}\},$$

for some $T \le \log\Big(\frac{M_1^2}{\kappa_{Z^*}^2 e^{4M_1+3M_3-M_2}}\frac{n}{k^3}\Big)/\log\Big(1 - \frac{r_0\tau\rho_1}{e^{M_1}\kappa_{Z^*}^2}\Big)^{-1}$.

In particular, the deterministic error bound for Δ_{η_t} consists of the statistical error term $\max\{\zeta_n, \varphi_n\}$ and the estimation error of Z_t, and with high probability $\|\Delta_{\eta_T}\|^2$ is dominated by the estimation error of Z_t and thus is also in the order $\mathcal{O}(nk)$.

One-step improvement Although Theorem 3.5.2 guarantees the convergence of Algorithm III.1 up to certain statistical errors, the achieved error rate is in the same order as that for the separate estimation method. To further investigate how exploiting extra structural information in the joint inner-product model would help estimate the latent variables, we consider the estimation error moving against the gradient one step from the estimators obtained by the separate method below.

Suppose we are given estimators $(\bar{\alpha}, \bar{Z})$ of latent variables obtained from the separate estimation Algorithm B.1 and an estimator $\bar{v} = \bar{Z}\bar{w} + \bar{\gamma}1_n$. Then we update the estimator of Z by one step through Algorithm III.1 as below:

$$\hat{Z} = \bar{Z} - 2\tau_z(1-\lambda)(\sigma(\bar{\Theta}) - |A|)\bar{Z} - 2\tau_z\lambda\big(|A| \circ \sigma(\bar{\eta}) - B\big)\bar{v}\bar{w}^\top, \tag{3.10}$$

where $\bar{\Theta} = \bar{\alpha}1_n^\top + 1_n\bar{\alpha}^\top + \bar{Z}\bar{Z}^\top$, $\bar{\eta} = \bar{v}\bar{v}^\top$, and $B = |A| \circ (A+1)/2$. Note that $[B]_{ij} = |A_{ij}|b_{ij}$ with b_{ij} independently following Bernoulli$(\sigma(\eta_{ij}^*))$ conditional on $|A|$. The following proposition provides insights on under what scenarios the one-step update in the joint estimation method could lead to better estimates of the latent position vectors Z. In below, for ease of derivation, we consider the parameter space $\mathcal{F}(n, k, M_1, M_2, M_3)$ with fixed M_i $(i = 1, 2, 3)$ and k.

Proposition 3.5.2. *Given the estimators* $(\bar{\alpha}, \bar{Z})$ *obtained from Algorithm B.1 and the estimators* $(\bar{w}, \bar{\gamma})$ *that are independent of* B *conditional on* $|A|$ *and satisfy* $\|\bar{w} - w^*\|^2 + \|\bar{\gamma} - \gamma^*\|^2 = \mathcal{O}(1/n)$*. We update* $\bar{Z}$ *for one step by (3.10) and obtain* $\hat{Z}$*.*

Suppose the conditions in Proposition 3.5.1 hold, and the singular values of the sample covariance $Z^{*\top}Z^*/n$ *are of constant order. Then there exists an optimal* λ_{opt} *that minimizes* $\mathbb{E}\|\hat{Z} - Z^*\|_F^2$*. Furthermore, if* $\lambda_{opt} > 0$*, we have* $\mathbb{E}\|\Delta_{\hat{Z}}\|_F^2 < \mathbb{E}\|\Delta_{\bar{Z}}\|_F^2$ *for any* $\lambda \in (0, 2\lambda_{opt})$*, and the improvement* $\mathbb{E}\|\Delta_{\bar{Z}}\|_F^2 - \mathbb{E}\|\Delta_{\hat{Z}}\|_F^2$ *with* λ_{opt} *is at least*

$$\frac{\big\||A| \circ \xi \circ T_1\big\|_F^2 \left(\big\||A| \circ \xi \circ T_1\big\|_F - \big\||A| \circ \xi \circ T_2\big\|_F - \big\||A| \circ \xi \circ T_3\big\|_F\right)^2}{16\left(\big\||A| \circ \xi \circ \xi \circ (T_1 + T_2 - T_3)\big\|_{op}^2 + \mathbb{E}\big\|B - |A| \circ \sigma(\eta^*)\big\|_{op}^2/(\|\bar{Z}\|_{op}^2\|\bar{w}\|^4)\right)},$$

where T_i*'s are given in (*B*.41)-(*B*.43) respectively for* $i = 1, 2, 3$ *in the Supplemental Material with* $\|T_1\|_F = \mathcal{O}(1)$*,* $\|T_2\|_F = \mathcal{O}(1)/\|w^*\|$*, and* $\|T_3\|_F = \mathcal{O}(1/\sqrt{n})$*, and* ξ *is an element-wise positive constant matrix. Here* $\mathbb{E}$ *represents the conditional expectation of* B *given* $|A|$*.*

We provide the expression of the optimal λ_{opt} that minimizes $\mathbb{E}\|\hat{Z} - Z^*\|_F^2$, the proof of Proposition 3.5.2, and discuss when the conditional independence assumption and the prerequisite error rate of $(\bar{w}, \bar{\gamma})$ in Proposition 3.5.2 hold in the Supplemental Material. Since a positive λ_{opt} implies a strict decrease in the mean square error of Z after one-step update, we further investigate in which case λ_{opt} tends to be positive. In particular, $\lambda_{opt} > 0$ if and only if $\big\||A| \circ \xi \circ T_1\big\|_F - \big\||A| \circ \xi \circ T_2\big\|_F - \big\||A| \circ \xi \circ T_3\big\|_F$ is strictly positive. Our analysis on the upper bounds of the three terms suggests that the first two terms are the dominating terms and a larger $\|w^*\|$ more likely results in a positive λ_{opt}. Therefore, when the signal from the edge sign distribution is strong, incorporating information from observed signs in the joint estimation method is useful for improving the estimation of Z. Moreover, the magnitude of improvement also depends on the levels of the signal and the noise in the sign distribution. Specifically, as $\|w^*\| \asymp \|\bar{w}\|$ increases, the difference between the upper bounds of the two dominating terms in the numerator increases while the upper bound of the denominator decreases, therefore overall the improvement is likely to increase. This implies that larger signals in the edge signs would lead to greater improvement in estimating Z. On the other hand, we find that the improvement decreases when the noise $\mathbb{E}\big\|B - |A| \circ \sigma(\eta^*)\big\|_{op}^2/\|\bar{Z}\|_{op}^2$ in the edge sign distribution increases in the denominator.

3.6 Simulation Studies

In this section, we conduct simulation studies to investigate how estimation errors of the proposed methods depend on: 1) the network size and the dimension of latent position vectors; 2) the network density; and 3) the proportion of positive edges.

Estimation methods We compare three estimation methods. In addition to the separate estimation method and the joint estimation method introduced in Section 3.4, we further add an intermediate method, *one-step-joint* estimation, to illustrate the one-step improvement discussed in Proposition 3.5.2. Specifically, given $\bar{Z}$ and $\tilde{v}$ obtained from Algorithms B.1 and B.2 respectively, we compute the one-step-joint estimators $(J_n\hat{Z}, \bar{v})$ by first updating $\bar{v} = \bar{Z}\bar{w}+\bar{\gamma}1_n$ with $\bar{w}, \bar{\gamma} = \arg\min_{w\in\mathbf{R}^k,\gamma\in\mathbf{R}} \|\tilde{v} - \bar{Z}w - \gamma 1_n\|$ and then obtaining $\hat{Z}$ by plugging $(\bar{Z}, \bar{v})$ into (3.10). We set $\lambda = 1/2$, so that the joint estimation is the same as the maximum likelihood estimation.

Simulation settings For a given network size n and a latent position vector dimension k, we set the model parameters as follows. We first generate the latent positions $Z_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$ from the standard normal distribution, for $1 \leq i \leq n, 1 \leq j \leq k$. By centering columns of Z, we get $Z^* = J_n Z$, where $J_n = I_n - \frac{1}{n}1_n 1_n^\top$. We further normalize Z^* element-wise such that $\|Z^* Z^{*\top}\|_F = n$. Next, we generate the node degree heterogeneity parameters $\alpha_i^* = -\alpha_i/\sum_{i=1}^n \alpha_i$, where $\alpha_i \overset{\text{iid}}{\sim} U(1, 3)$ is uniformly distributed for $1 \leq i \leq n$. Finally, we set $w^* = 1/\sqrt{k}\cdot 1_k$, $\gamma^* = 0$, and $v_i^* = w^{*\top} Z_i^* + \gamma^*$.

Given the true latent variables Z^*, α^*, and v^*, we randomly generate 20 replications of the signed adjacency matrix following (3.3) and (3.4), and fit models by three estimation methods. For each method, we measure the relative errors for Z, v, Θ, and η. Due to the identifiability conditions in Proposition 3.3.2, we define the relative error for Z as $\|\hat{Z} - Z^*\bar{Q}\|_F / \|Z^*\|_F$, where $\bar{Q} = \arg\min_{Q \in O(k)} \|\hat{Z} - Z^*Q\|_F$ and $O(k)$ is the collection of all orthogonal matrices in $\mathbf{R}^k$. We define the error for v as $\|\hat{v} - \bar{\kappa}v^*\| / \|v^*\|$, where $\bar{\kappa} = \arg\min_{\kappa \in \{1,-1\}} \|\hat{v} - \kappa v^*\|$. The relative errors for Θ and η are defined as $\|\hat{\Theta} - \Theta^*\|_F / \|\Theta^*\|_F$ and $\|\hat{\eta} - \eta^*\|_F / \|\eta^*\|_F$ respectively, where $\hat{\Theta} = \hat{\alpha}1_n^\top + 1_n\hat{\alpha}^\top + \hat{Z}\hat{Z}^\top$ and $\hat{\eta} = \hat{v}\hat{v}^\top$.

3.6.1 Varying the Network Size and the Dimension of the Latent Space

In Figure III.3, we summarize how estimation errors vary with different network sizes. We fix $k = 2$ and vary $n \in \{500, 1000, 2000, 4000\}$. We can see that, for a fixed dimension of the latent space, the relative errors of all three estimation methods decrease in the rate of $1/\sqrt{n}$ as the network size n grows, which align well with the theoretical error rates given in Section 3.5. Next, compared to the separate estimation method, the joint estimation method consistently achieves smaller estimation errors on all four quantities of interest across different network sizes. In addition, the one-step-joint estimation that simply updates estimates by one-step gradient descent is able to reduce the estimation errors compared to the separate estimation method.

In Figure III.4, we further summarize how estimation errors of Z and Θ vary with different dimensions of the latent position vector. We fix $n = 2000$ and vary $k \in \{2, 4, 8\}$. We find that, for a fixed network size, the relative errors increase in the rate of $\sqrt{k}$ as the dimension of latent position vector k grows. This also agrees well with our theoretical results. The relative trend among the three estimation methods for different k is similar as that in Figure III.3, where the joint estimation method is consistently the best.

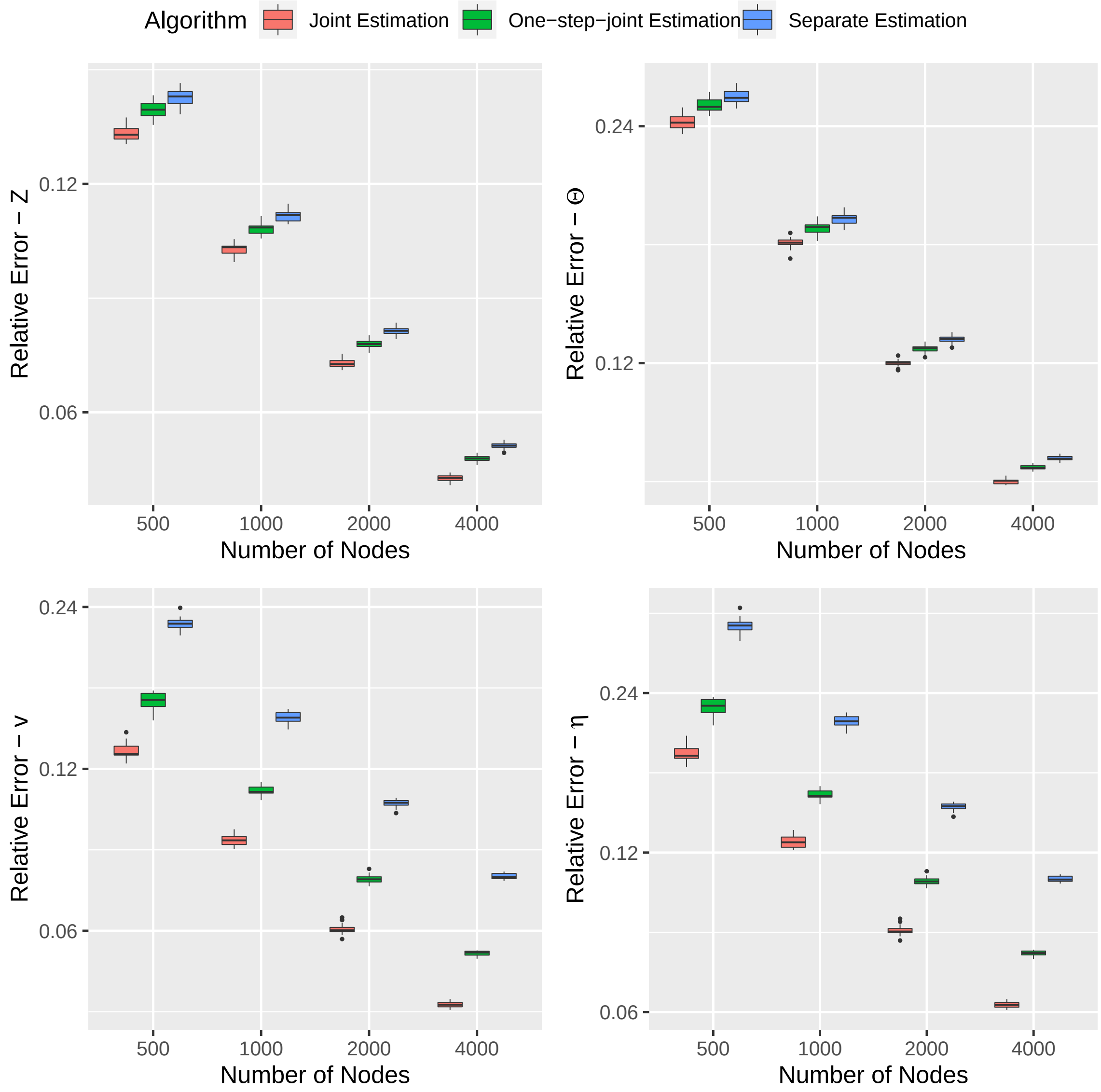

Figure III.3: Log-log plots of relative errors with respect to the network size n. The dimension of the latent position vector is fixed as $k = 2$.

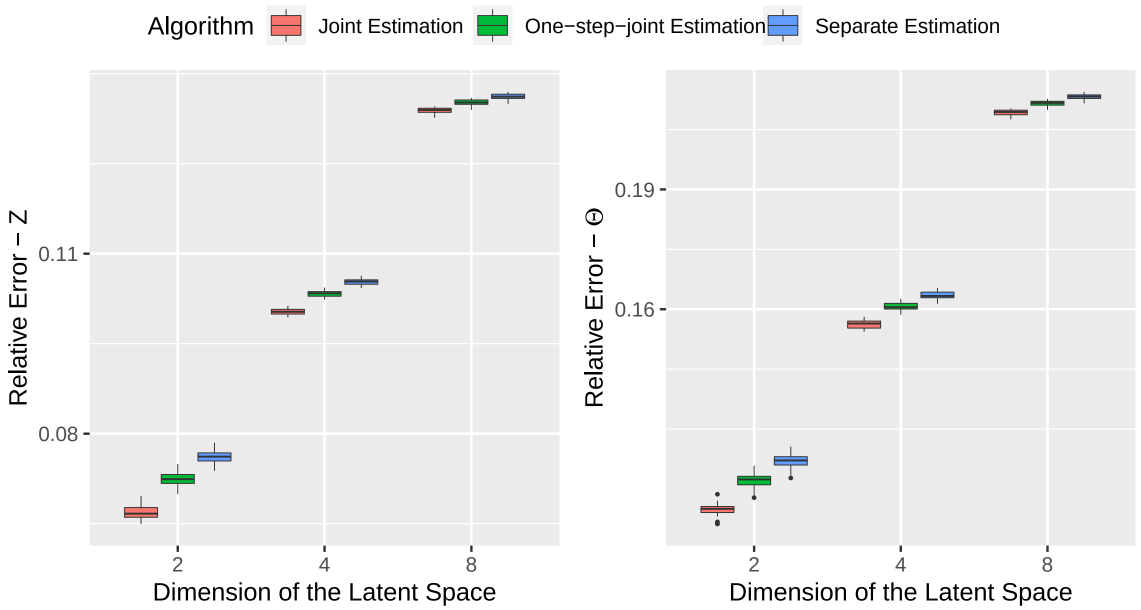

Figure III.4: Log-log plots of relative errors with respect to the dimension of the latent position vector k. The network size is fixed as $n = 2000$.

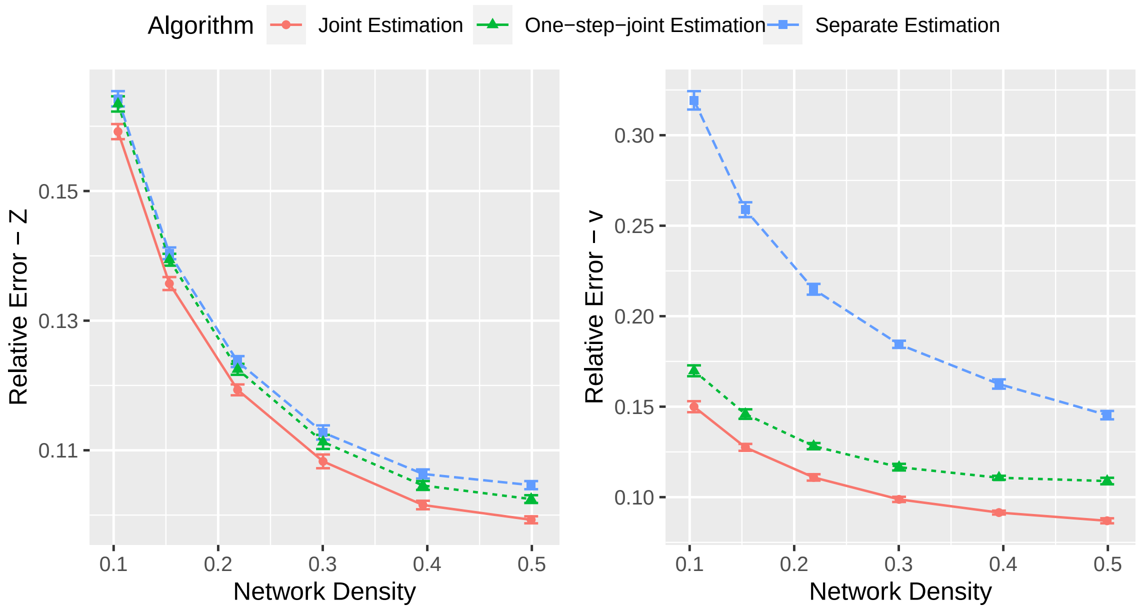

Figure III.5: Relative errors with respect to the network density. The network size is fixed as $n = 2000$ and the dimension of the latent position vector is fixed as $k = 4$.

3.6.2 Varying the Network Density

We investigate how estimation errors for three estimation methods vary with the network density. To this end, we generate the node degree heterogeneity parameters $\alpha_i^* = -\bar{\alpha} - \alpha_i / \sum_{i=1}^n \alpha_i$ where $\alpha_i \overset{\text{iid}}{\sim} U(1,3)$ is uniformly distributed for $1 \leq i \leq n$. We fix $n = 2000$ and $k = 4$, and vary $\bar{\alpha} \in \{0, 0.25, 0.5, 0.75, 1, 1.25\}$, which leads to the network density ranging from 0.1 to 0.5.

Figure III.5 summarizes the relative estimation errors of Z and v over 20 replications under different network densities. We can see that both estimation errors of Z and v for all three estimation methods decrease as the network gets denser, and the joint estimation method achieves lower estimation errors than the other two methods consistently across various network densities. In particular, when the network is dense, the improvement in estimating Z from the joint estimation against the separate estimation increases, which is expected because in joint estimation, the observed edges' signs are also useful for inferring Z and denser networks provide more information. In addition, regarding the estimation error of v, the joint and one-step-joint estimation methods that use the additional structural information between Z and v perform more stably than the separate estimation method as the network gets sparser.

3.6.3 Varying the Proportion of Positive Edges

We also investigate the effect of the proportion of positive edges on three estimation methods. For this purpose, we change the simulation setting. Specifically, we fix $n = 2000$ and $k = 4$, and vary $\gamma^* \in \{0, 0.3, 0.6, 0.9, 1.2, 1.5, 1.8\}$, which results in the proportion of positive edges ranging from 0.52 to 0.91. To eliminate the artificial effect resulting from varying γ^* when evaluating the estimation error of v, we focus on the estimation error of the centered v, i.e., $v_{cen} = J_n v$. We define the relative error for v_{cen} as $\|J_n \hat{v} - \bar{\kappa} J_n v^*\| / \|J_n v^*\| = \|J_n \hat{v} - \bar{\kappa} Z^* w^*\| / \|Z^* w^*\|$, where

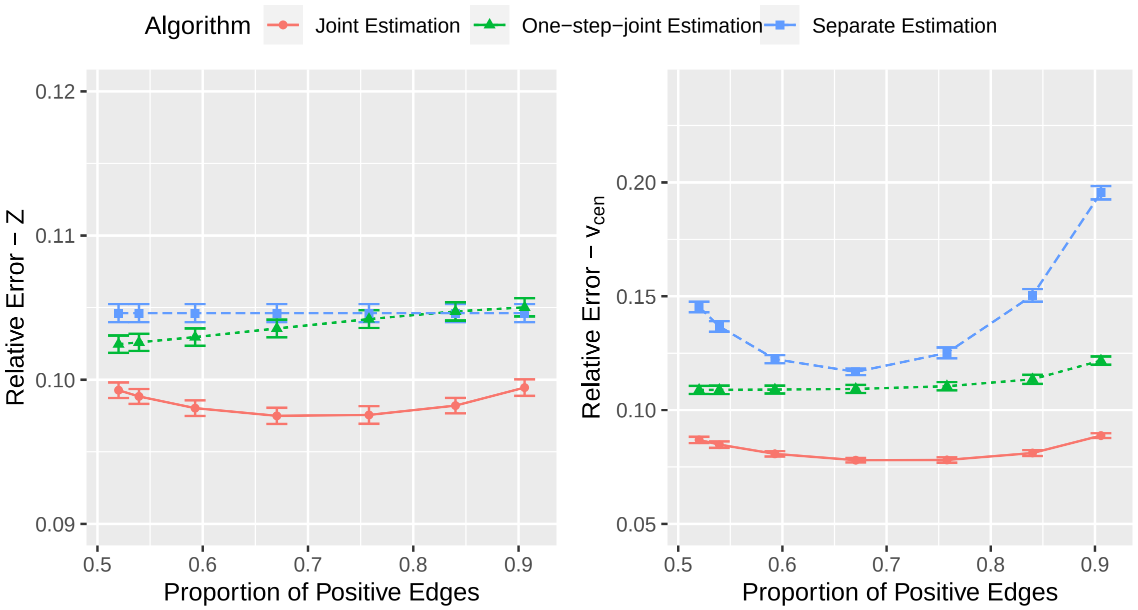

Figure III.6: Relative errors with respect to the proportion of positive edges. The network size is fixed as $n = 2000$ and the dimension of the latent position vector is fixed as $k = 4$.

$\bar{\kappa} = \arg\min_{\kappa \in \{1,-1\}} \|J_n \hat{v} - J_n \kappa v^*\|$.

Figure III.6 summarizes the relative estimation errors of Z and v_{cen} over 20 replications under different proportions of positive edges. Overall, the joint estimation method performs consistently the best among three methods and is robust across different sign distributions. We note that the estimation error for Z for the separate method does not change, because the generated absolute adjacency matrix $|A|$ does not change when varying γ^* and thereby the separate estimates $\hat{Z}$ stay the same. We also observe that the optimal performance of the separate method for estimating v is not achieved around 50% positive signs. This is because since positive and negative signs are not exchangeable under the balance structural theory, the optimum should not be expected to be 50% as in the standard binary classification setting where class labels are exchangeable.

3.7 International Relation Data

In this section, we apply the proposed method to an international relation data, i.e. the Correlates of War (COW) (Izmirlioglu, 2017), to demonstrate how the proposed method can be used to make informative interpretation and visualization of signed networks. The COW dataset records various types of international relations among countries, such as wars, alliances, and militarized interstate disputes. Similarly as Kirkley et al. (2019), we construct a signed network of countries, where the positive edges represent alliance relationships, and the negative edges represent the existence of militarized disputes between countries. We take the snapshot of the records during World War II (WWII), i.e., from 1939 to 1945. In particular, if two countries were involved in both alliances and militarized disputes, we set the sign of their edge to positive if the number of years of alliances is larger than that of the militarized disputes, and we set the sign to negative otherwise. According to the COW records, there are 68 countries that were involved in alliances or militarized disputes during WWII, and the resulted signed network contains 566 positive edges and 519 negative edges.

We fit two models to the COW dataset, one corresponding to the joint estimation method and the other corresponding to the separate estimation method. For both models, we set the dimension of the latent position vectors as $k = 2$, such that the estimated $\hat{Z}$ can be directly visualized on a 2-dimensional plane. The models fitted by the joint estimation method and the separate estimation method are visualized in Figure III.7 and Figure III.8 respectively. In both figures, each node represents a country and their coordinates are given by $\hat{Z}$. The size of each node is determined by the estimated degree heterogeneity parameter $\hat{\alpha}$, with larger nodes corresponding to larger $\hat{\alpha}_i$ values. The color and the shape of each node i distinguish the estimated latent polar variable $\hat{v}_i$. Specifically, if $\hat{v}_i > 0$, the node is visualized as a red circular point; and if $\hat{v}_i < 0$, the node is visualized as a blue square point. For both the red and the blue points, the larger the absolute magnitude $|\hat{v}_i|$, the darker the color. The sign of each edge is also indicated in the figure, with dashed green being positive and solid purple being negative.

Figure III.7 shows that the proposed model fitted by the joint estimation method is able to capture important information in the signed network of countries during WWII. First, in terms of the estimated node degree heterogeneity $\hat{\alpha}$, the top 11 countries are Germany, Italy, Japan, the United Kingdom (UK), Romania, the United States (USA), Brazil, Bulgaria, Hungary, France, the Soviet Union (USSR). In particular, UK, USA, USSR were the 3 leading countries of the Allies of WWII. France played important and complicated roles in both the Allies and the Axis. Brazil was the only South American country that actively participated in WWII. All other countries were members of the Axis. Since members of the Axis were more active than those of the Allies on average, it is reasonable that even small members have high values of $\hat{\alpha}$. Second, regarding the estimated latent polar variable $\hat{v}$, the model is able to divide the countries into two groups (blue and red) that mostly align with the division between the Axis and the Allies. As we assume a linear transformation from Z to v, the plane (the space of Z) can be linearly separated into two areas with the boundary illustrated by the grey dash-dotted line. We can see that edges crossing the boundary are mainly negative (purple), while edges within the same side are mainly positive (green). Finally, the estimated latent position vectors $\hat{Z}$ also capture various other interesting aspects. Within the Allies, $\hat{Z}$ further clusters countries in America together (see the right part of the figure), among which most edges are positive (green). The Middle East countries form a cluster as well (at the top of the figure), though not as tight as countries in America. It is also interesting to see that France, one of the major Allied powers in WWII, is positioned on the same side of Germany, Italy, and Japan. This is because, after occupied by Germany, France was divided into two political powers, Free France and Vichy France, with the latter collaborating

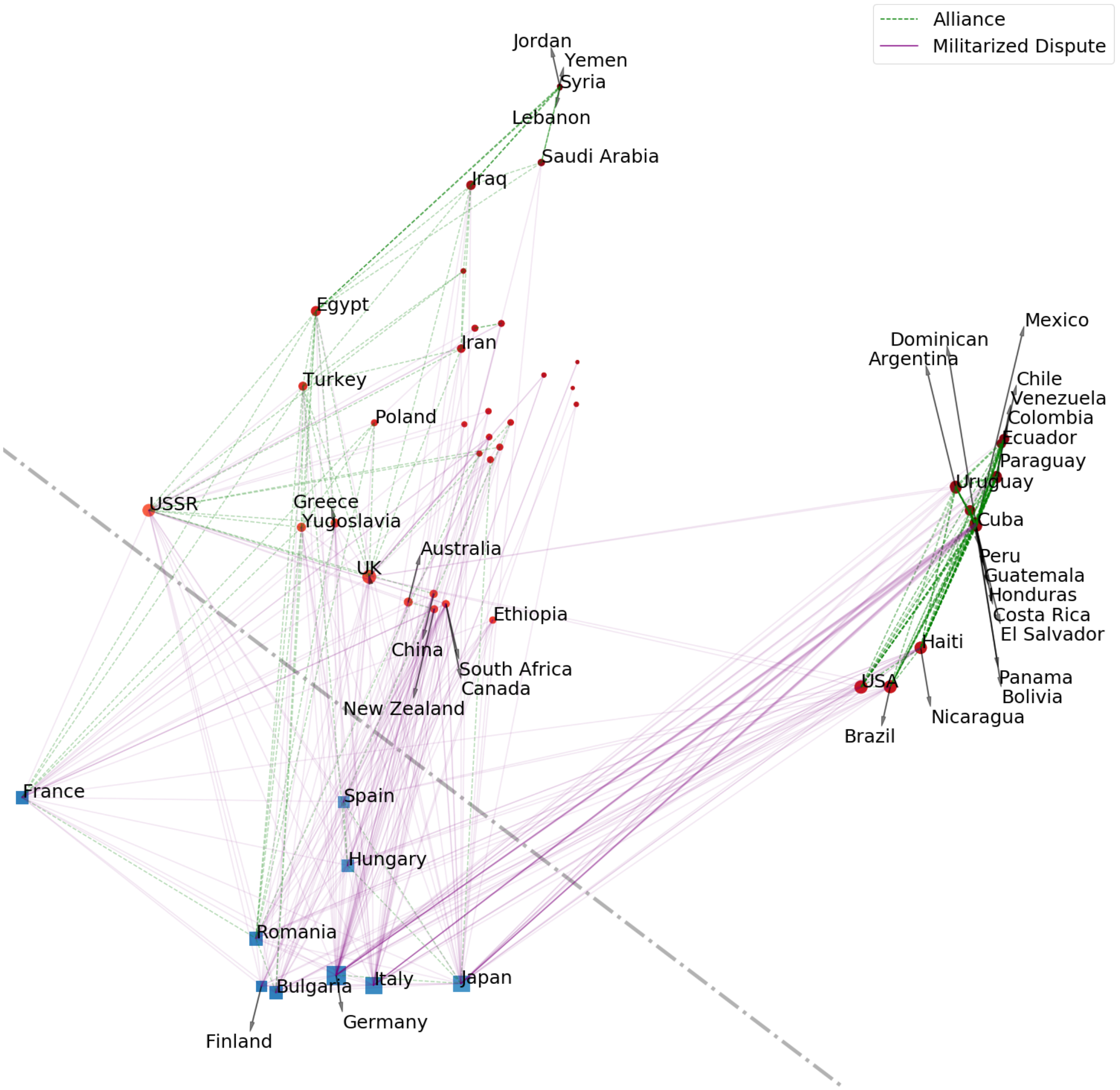

Figure III.7: Visualization of the model fitted by the joint estimation method on the COW dataset. The nodes are countries involved in WWII. The green dashed lines represent positive edges (alliance) and the purple solid lines represent negative edges. The node sizes are determined by the estimated node degree heterogeneity parameters $\hat{\alpha}$. The node colors and shapes are determined by the estimated latent polar variables $\hat{v}$: for each node i, the node is a red circle if $\hat{v}_i > 0$ and a blue square otherwise; and the larger absolute magnitude of $\hat{v}_i$, the darker the color. The grey dash-dotted line represents the linear boundary between the red and blue nodes. To make the visualization easier, we only labeled countries with degree greater than 5.

with Germany and fighting against the Allies in several campaigns from 1940 to 1944.

On the other hand, the model fitted by the separate estimation method captures useful information from the network (see Figure III.8) but not as interpretable as that by the joint estimation method.

Specifically, the model fitted by the separate estimation method roughly captures the division between the two alliances, but it mistakenly colors UK in blue together with the Axis powers. Compared to Figure III.7, another country that is flipped in terms of $\hat{v}$ is Romania, which was considered as a member of the Axis during most of the time of WWII. In addition, as the latent position vectors $\hat{Z}$ and the latent polar variables $\hat{v}$ are estimated separately, the estimation of $\hat{Z}$ is not aware of the signs of the edges. As a result, the node positions in Figure III.8 do not reflect the two alliances. In particular, Germany and Japan are placed in the center of the figure as they have the most number of edges (counting both positive and negative edges), and they separate countries in America (at the right of the figure) from other members of the Allies (at the top-left corner).

3.8 Discussion

In this chapter, we propose a latent space approach that accommodates the structural balance theory for modeling signed networks. In particular, we introduce a novel notion of population-level balance, which is a natural choice to characterize the structural balance theory when we treat both the edges and their signs as random variables. We develop sufficient conditions for a latent space model to be balanced at the population level, and propose two balanced inner-product models following the conditions. We also provide scalable estimation algorithms with theoretical guarantees.

There are a few directions we may continue to explore in the future. First, the joint inner-product model could be extended to have a nonlinear link function ϕ,

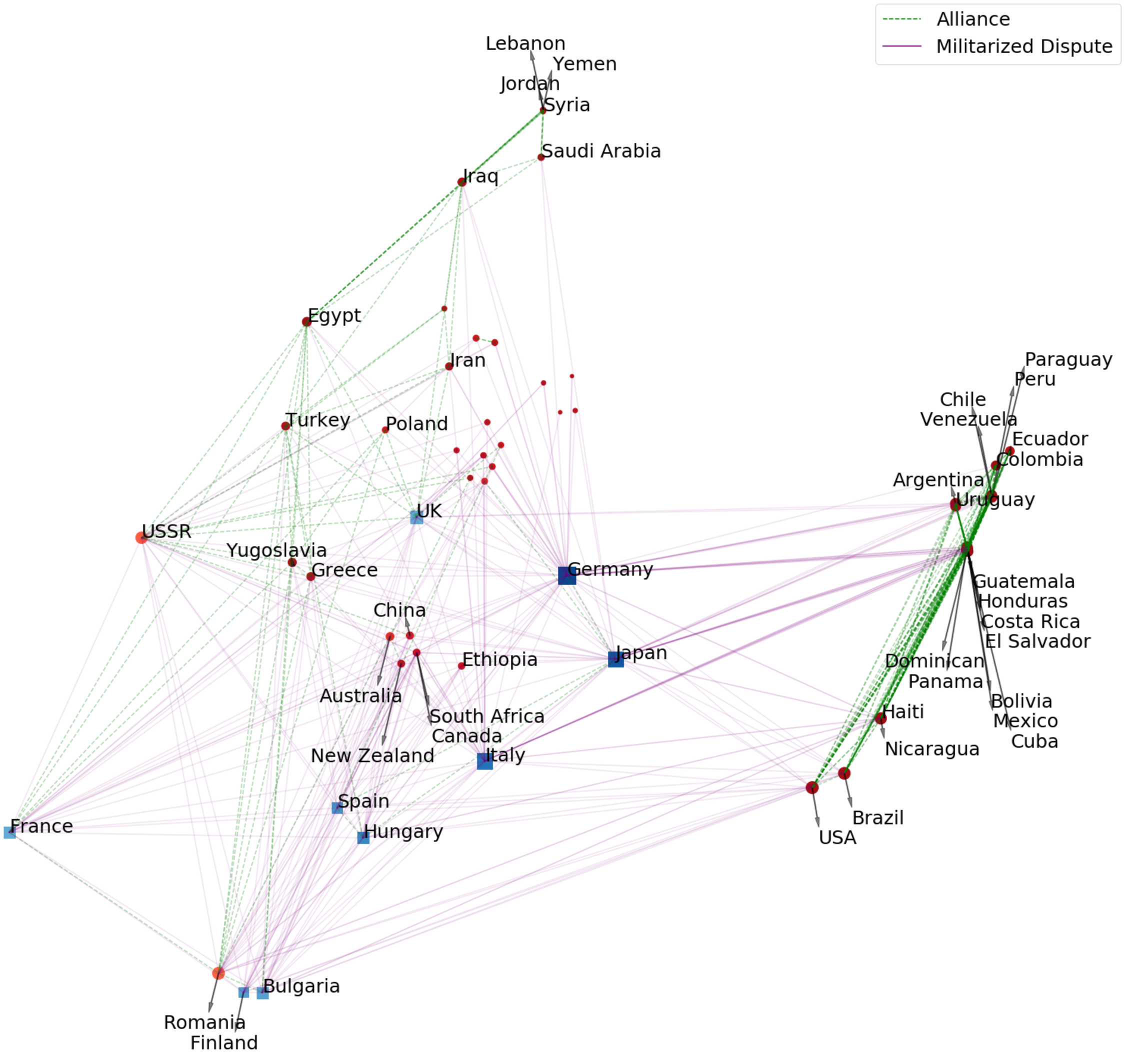

Figure III.8: Visualization of the fitted model by separate estimation. The visual semantics are the same as Figure III.7, except for there is no linear boundary between the red and blue nodes.

which would increase the flexibility of the model. Second, we may generalize the proposed approach to weighted signed networks to better leverage richer edge information available in real-world networks. Finally, it is desirable to extend the latent space approach for undirected networks to directed networks, which can be potentially used to model other interesting social theories, such as the social status theory, for signed networks.

CHAPTER IV

Semi-Supervised Learning for Longitudinal Clinical Events

4.1 Introduction

Deep neural network models have been increasingly used to analyze large-scale electronic health records (EHR) and have shown superior prediction performances in several medical tasks including automatic detection of diabetic retinopathy using medical images (Gulshan et al., 2016) and clinical text classification (Yao et al., 2019). As opposed to medical images and clinicians' text notes, input features such as clinical events are usually of longitudinal nature. Specifically, sensor recordings, laboratory test results, medications, and new diagnosis codes are recorded on each clinical visit and may change over time. Such longitudinal nature is often accompanied by additional modeling challenges such as irregular time gaps between visits, varying lengths of follow-ups, and complex missing patterns. Recurrent neural networks (RNNs), given their clear advantages in taking sequential inputs and successes in natural language processing (NLP) (Wu et al., 2016), are a natural choice for handling longitudinal inputs, and in recent years, they have been successfully used to analyze clinical events data in different applications such as early detection of heart failure (Choi et al., 2016), kidney failure after transplantation (Esteban et al., 2016), and daily sepsis and myocardial infarction (Kaji et al., 2019).

Despite many existing successful applications of RNNs on the classification of clinical events data, most of them rely on the accessibility of a large number of accurately labeled training data. However, in many healthcare settings, qualified graders and disease/domain experts are required to make an accurate diagnosis. Moreover, invasive measurements may result in additional risk to patients and non-invasive measurement may not be ubiquitous and may result in substantial cost. Therefore, it is often difficult to collect a large number of accurate labels, which limits further applications of deep learning models on clinical events data when labels are scarce. On the other hand, with the availability of routinely collected EHR, there usually exists abundant and easy-to-collect unlabeled data. Therefore, our goal is to develop semi-supervised learning methods for longitudinal clinical events which can incorporate unlabeled data to help improve classification performance. Successful implementation of such methodology will help reduce costs of collecting clinical labels when building prediction models.

Although there have been many works on semi-supervised learning in the field of deep learning (Kingma et al., 2014; Odena, 2016; Narayanaswamy et al., 2017; Socher et al., 2013), there are few works that take longitudinal input such as laboratory tests and charted events that are commonly seen in EHR. Further, most existing approaches treat feature extraction using unlabeled data and building prediction models using labeled data as two separate steps (Dai and Le, 2015; Che et al., 2017; Ballinger et al., 2018). The potential drawback of such a two-step approach is that the learned feature representation in the first step receives no supervised guidance from labeled data and, therefore, may not be specific to the desired task.

To overcome the lack of supervision in the first step, we propose to jointly learn feature representation from both labeled and unlabeled data. Our model consists of two parts: a sequence generative network for modeling longitudinal clinical events and a label prediction network which takes the hidden feature representation of the sequence generative network as inputs. The two parts are learned end to end using both labeled and unlabeled training data in a joint manner, such that the data could be well separated in the shared feature space. We empirically show that the proposed joint learning method significantly outperforms the two-step method when labels are scarce. Furthermore, we consider two different generative models for modeling longitudinal clinical events. In addition to the RNNs that have been used in the aforementioned works, where all recurrent layers are deterministic, we also adopt stochastic RNNs which contain an additional stochastic latent recurrent layer. Based on our numerical experiments, taking stochastic RNNs as the generative model could further improve the prediction performance in most cases.

The rest of this chapter is organized as follows. We introduce related work in Section 4.2 and present the proposed semi-supervised joint learning approach with technical details in Section 4.3. We demonstrate the effectiveness of the proposed method in Section 4.4 and conclude this chapter with discussions in Section 4.5.

4.2 Related Work

Many semi-supervised learning methods have been proposed for deep learning models (Kingma et al., 2014; Odena, 2016; Narayanaswamy et al., 2017; Socher et al., 2013; Dai and Le, 2015). In particular, deep generative models have made great progress on learning feature representations with little or no supervised information in recent years (Cho et al., 2014; Chung et al., 2015; Kingma and Welling, 2014; Goodfellow et al., 2014), and have shown their advantages on unsupervised and semi-supervised tasks. For instance, Kingma et al. (2014) proposed a two-step semi-supervised learning method by first learning a low-dimensional feature representation from unlabeled images via variational autoencoder (VAE) (Kingma and Welling, 2014) and then learning an image classifier from labeled data. However, of the semi-supervised learning methods, only a few can be applied to accommodate longitudinal clinical events (Dai and Le, 2015; Che et al., 2017; Ballinger et al., 2018). Among them, Dai and Le (2015) proposed to pre-train parameters in a RNN encoder with large amounts of unlabeled data and then learn specific text classification tasks starting with pre-trained initialization. Other unsupervised representation learning algorithms such as word2vec (Mikolov et al., 2013) can also be used in the pre-training step. They showed that such pre-training procedure using unlabeled data could provide stable initialization and could be generalized well in different text classification tasks. Following this approach, DeepHeart (Ballinger et al., 2018) also pre-trained parameters in RNNs on unsupervised and weakly supervised tasks and then built a prediction model for four conditions associated with cardiovascular risks using labeled data. More recently, the ehrGAN (Che et al., 2017) was developed to generate realistic patients' clinical events via unsupervised learning. Based on the implicit belief that the generated samples from ehrGAN with input x are likely to have the same label as x, they were further used to produce pseudo labeled data for supervised learning. However, this assumption may not hold in general since the learning procedure of ehrGAN in the first step does not use any label information.

All of the aforementioned semi-supervised learning methods for classifying longitudinal clinical events separate the learning process into two steps: (1) learn a deep generative model using unlabeled data to either pre-train the parameters or augment data; (2) learn a classifier for a specific classification task using labeled data based on the pre-trained initialization or augmented labeled data obtained in the first step. The key potential limitation of such two-step methods is that there is no or weak supervision from labels in the first step. Although data points may cluster well in the feature space by learning the intrinsic structure from unlabeled data, the clusters do not necessarily correspond to the labels of interest. The joint learning approach in our proposed method would make use of label information to help learn feature representation that can better separate data corresponding to the labels, and therefore, obtain better prediction performance.

4.3 Semi-Supervised Joint Learning with Longitudinal Features via Neural Networks

In this section, we first describe the problem setup of semi-supervised classification for longitudinal features, using clinical events as a specific example. Then we introduce the neural network models, and present the joint learning approach.

4.3.1 Problem Setup

We focus on two different types of features that are commonly seen in EHR: longitudinal features and time-static features. Longitudinal features may include multiple laboratory measurements, charted observations, and active treatments. These features are recorded every time a patient comes for a clinical visit or new laboratory tests or medications are ordered. We denote longitudinal features by $x = (x_1, \cdots, x_\ell)$, where $x_j \in R^{d_1}$ for $j = 1, \cdots, \ell$, and ℓ is the length of the sequence and can be different for different individuals. Time-static features may include gender, race, admission type, and age at enrollment, which are constant throughout the entire study. We denote time-static features by $w \in R^{d_2}$. The label $y \in \{1, \cdots, K\}$ could be the corresponding class associated with mortality or progression of diseases. In the semi-supervised learning setting, we observe only a small number of labeled data (x^i, w^i, y^i) for $i = 1, \cdots, n$ and a large number of unlabeled data (x^i, w^i) for $i = n+1, \cdots, n+m$, where m is usually much greater than n. We aim to learn a classifier that maps (x, w) to a class label y and incorporate both unlabeled and labeled data to improve prediction performance.

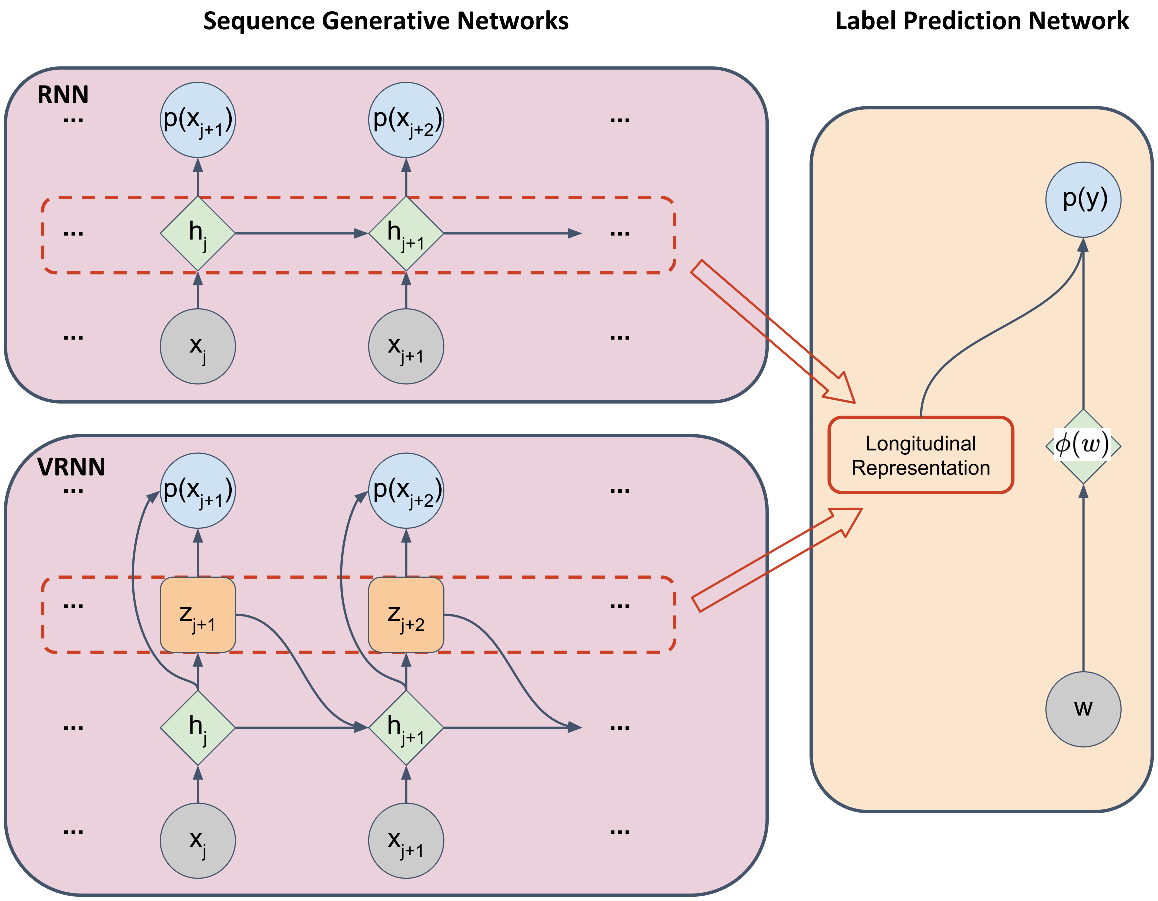

Figure IV.1: Model structure of two sequence generative networks and the label prediction network, where circles represent the inputs and outputs, diamonds represent deterministic hidden layers, and squares represent the stochastic latent recurrent layer in VRNN.

4.3.2 Model Structure

We propose two neural network models whose architectures are given in Figure IV.1. Each model consists of two parts: (1) a sequence probabilistic generative network for longitudinal features, which takes any sequence of longitudinal features $(x_1, x_2, \cdots, x_{j-1})$ as inputs and models the distribution of features at the next time step, i.e., $p(x_j|x_1, \cdots, x_{j-1})$; (2) a label prediction network which takes the hidden recurrent layer of the sequence generative model and the time-static features as inputs and outputs the probability for each class.

Sequence generative network We consider two different generative models to model what comes next in a sequence. The first one is a Gated Recurrent Units (GRU) neural network. We choose GRU because it can better capture long-term dependency due to the additional gate mechanisms compared with vanilla RNN (Cho et al., 2014). The second one is a variational recurrent neural network (VRNN) which contains an additional stochastic recurrent layer. It has been shown that introducing a latent stochastic recurrent layer can provide significant improvements in natural speech processing (Chung et al., 2015), and we adopt it here to examine its potential advantages in modeling clinical events. We describe the two probabilistic models in detail below.

RNN: As shown in the upper left corner of Figure IV.1, the hidden units in the recurrent layer $h = (h_1, \cdots, h_\ell)$ leverage historical information through the recurrent connection $h_j = f(x_j, h_{j-1})$, where f is a nonlinear transformation introduced in GRU. The historical information stored in h_{j-1} determines the distribution of longitudinal features at next time step. Specifically, the conditional density of x_j is given by $p(x_j; h_{j-1}) = \psi(x_j, h_{j-1})$, where ψ is an appropriate density function. For example, if x_j is continuous, we can use a multivariate Gaussian distribution $x_j \sim \mathcal{N}(\mu_{x,j}, \mathrm{diag}(\sigma^2_{x,j}))$, where $[\mu_{x,j}, \sigma_{x,j}] = \xi(h_{j-1})$ and ξ is modeled by a fully connected neural network. Here we assume different components of x_j are uncorrelated conditional on h_{j-1} in $p(x_j; h_{j-1})$, but they can be correlated in the marginal distribution $p(x_j)$. Since all transformations are deterministic, the joint probability density of longitudinal features is given by

$$p(x) = \prod_{j=1}^{\ell} p(x_j | x_1, \cdots, x_{j-1}) = \prod_{j=1}^{\ell} p(x_j; h_{j-1}),$$

where h_0 is usually set as a zero vector in practice.

VRNN: As shown in the bottom left corner of Figure IV.1, there is an additional stochastic recurrent layer $z = (z_1, \cdots, z_\ell)$ compared to the RNN. In particular, layer z is different from the standard hidden layer h since z_j's are random variables while h_j's take deterministic values. The conditional distribution of z_j accesses the historical information through the hidden state h_{j-1}. Specifically, the variable z_j is assumed to follow a multivariate Gaussian distribution with mean $\mu_{z,j}$ and variance diag$(\sigma^2_{z,j})$, which are determined through a fully connected neural network taking h_{j-1} as inputs. Moreover, the distribution of x_j will not only be conditioned on h_{j-1} but also on the latent z_j, i.e. $p(x_j|z_j; h_{j-1}) = \psi(x_j, \rho(z_j), h_{j-1})$, where ψ is an appropriate density function and ρ is a feature extractor with a two-layer fully connected neural network. Note that, in contrast to the assumption in RNN, now different components of x_j can be correlated conditional on h_{j-1} in $p(x_j; h_{j-1})$, after marginalizing z_j in $p(x_j|z_j; h_{j-1})$. Overall, the joint distribution of longitudinal features x and latent recurrent features z is given by

$$p(x, z) = \prod_{j=1}^{\ell} p(x_j, z_j | x_1, \cdots, x_{j-1}, z_1, \cdots, z_{j-1}) = \prod_{j=1}^{\ell} p(x_j|z_j; h_{j-1}) p(z_j; h_{j-1}),$$

where, similarly as RNN, h_0 can be set as a zero vector. The hidden units are updated through the recurrence equation $h_j = f(x_j, [z_j, h_{j-1}])$, where f is a GRU module treating the concatenation of z_j and h_j as the hidden state.

Label prediction network The label prediction network takes the recurrent layer of the sequence generative network and the time-static features as inputs and returns the probability of belonging to each class. As shown on the right of Figure IV.1, we first use a feature extractor ϕ for time-static features, where ϕ is a fully connected neural network taking w as inputs. Then we merge the information from both longitudinal features and time-static features by concatenating the hidden feature representation of the sequence generative network and the extracted features $\phi(w)$. Specifically, we utilize the recurrent hidden layer h for RNN and $\tilde{\mu}_z(x) = E_{z \sim q(z|x)} z$, the expectation of the approximate posterior $q(z|x)$ (to be specified later in Section 4.3.3) of the stochastic recurrent layer z for VRNN. When different individuals have varying lengths of longitudinal features, we can apply a max pooling layer on $h(x)$ or $\tilde{\mu}_z(x)$ over the time steps before we concatenate them with $\phi(w)$. After merging the feature representations, another fully connected neural network along with a Softmax output layer φ is used to output the probability scores, i.e. $p(y|x,w) = \varphi(y; h(x), \phi(w))$ for RNN and $\varphi(y; \tilde{\mu}_z(x), \phi(w))$ for VRNN.

4.3.3 Joint Learning

The sequence generative network and the label prediction network are learned jointly end to end through shared parameters in the representation of longitudinal features. Specifically, we minimize an objective function that consists of an unsupervised loss and a supervised loss.

The unsupervised loss is constructed by using the negative log-likelihood for longitudinal features x. For RNN, the unsupervised loss is given by

$$\mathcal{L}_g(\theta_g; x) \triangleq -\log p(x) = -\sum_{j=1}^{\ell} \log p(x_j; h_{j-1}) \tag{4.1}$$

where θ_g represents all parameters of RNN. For VRNN, however, the marginal density function $p(x)$ is intractable due to the highly non-linear dependency between x and z. Thus, following Kingma and Welling (2014) and Chung et al. (2015), we consider a variational lower bound of the marginal likelihood function by introducing an approximate posterior model $q(z|x)$, i.e.

$$\log p(x) \geq E_{z\sim q(z|x)} \log p(x|z) - KL(q(z|x)\|p(z)),$$

where KL is the Kullback–Leibler divergence. When the approximate posterior $q(z|x)$ equals the true posterior $p(z|x)$, the gap between the log-likelihood and the lower bound becomes zero. In practice, the approximate posterior model $q(z|x)$ is chosen to be

$$q(z|x) = \prod_{j=1}^{\ell} q(z_j|x_1, \cdots, x_j, z_1, \cdots, z_{j-1}) = \prod_{j=1}^{\ell} q(z_j|x_j; h_{j-1}),$$

where $q(z_j|x_j; h_{j-1})$ follows a multivariate Gaussian distribution whose mean and covariance are parameterized by a neural network taking x_j and h_{j-1} as inputs. Overall, the generative model $p(x, z)$ and the approximate posterior model $q(z|x)$ are learned simultaneously by minimizing the negative lower bound

$$\mathcal{L}_g(\theta_g; x) = -E_{z\sim q(z|x)} \sum_{j=1}^{\ell} \left(\log p(x_j|z_j; h_{j-1}) - KL(q(z_j|x_j; h_{j-1}) \| p(z_j; h_{j-1}))\right), \tag{4.2}$$

where θ_g represents all parameters of VRNN.

The supervised loss is given by the cross entropy between the true class label y and the class probabilities returned by the label prediction network

$$\mathcal{L}_d(\theta_d, \tilde{\theta}_g; y, x, w) = -\log p(y|x, w), \tag{4.3}$$

where θ_d represents all parameters used in the time-static feature extractor ϕ and classifier φ, and $\tilde{\theta}_g$ are the parameters used in h or the approximate posterior model $q(z|x)$ that is a subset of θ_g.

The overall objective function is a weighted sum of the unsupervised loss and the supervised loss

$$\mathcal{L}(\theta_g, \theta_d) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_d(\theta_d, \tilde{\theta}_g; y^i, x^i, w^i) + \eta \cdot \frac{1}{n+m} \sum_{i=1}^{n+m} \mathcal{L}_g(\theta_g; x^i), \tag{4.4}$$

where the first term is an average over the labeled data, the second term is an average over both labeled and unlabeled data, and η is a weight hyperparameter. The parameters $\tilde{\theta}_g$ are included in both $\mathcal{L}_d$ and $\mathcal{L}_g$ and are iteratively updated using both unlabeled and labeled data during training. Therefore, the representation of longitudinal features is learned not only by the unsupervised generative task but also under the supervision from the labeled data. Further, the hyperparameter η controls the trade-off between the unsupervised learning and the supervised learning. A lower value of η leads to a stronger supervision from labeled data but weaker unsupervised learning from unlabeled data. For example, when η equals to zero, it is equivalent to supervised learning using labeled data alone. Based on our numerical experiments, η

is an important hyperparameter that needs to be tuned carefully.

4.4 Numerical Experiments

To demonstrate the effectiveness of joint learning, we evaluate the proposed method by using five mortality-related classification tasks on the Medical Information Mart for Intensive Care III (MIMIC) database (Johnson et al., 2016; Goldberger et al., 2000). We aim to examine: 1) whether additional unlabeled longitudinal features can help improve the prediction performance through semi-supervised learning approaches; 2) how the proposed joint learning method performs in comparison with existing two-step semi-supervised learning methods; 3) how using the stochastic RNN as the sequence generative model differs from using the deterministic RNN.

4.4.1 Datasets

The MIMIC database provides deidentified clinical data of patients admitted to an Intensive Care Unit (ICU) stay. It has been used to benchmark the performance of deep learning models for predicting the length of stay, phenotyping, ICD-9 code group, in-hospital mortality (Harutyunyan et al., 2019), short-term mortality, and long-term mortality (Purushotham et al., 2018). Nonetheless, the evaluation of semi-supervised learning methods on MIMIC is still lacking. In this chapter, we predict five mortality related tasks: in-hospital mortality, 2-day and 3-day mortality (short-term mortality), and 30-day and 1-year mortality (long-term mortality). We focus on adult patients who were alive the first 24 hours after the first admission to ICU, which results in an analytic sample of 35,643 patients. Table IV.1 summarizes the proportion of mortality for each task.

Following Purushotham et al. (2018), we take 15 longitudinal features from the first 24 hours after admission to ICU. Specifically, they are the 3 types of Glasgow Coma Scale scores, systolic blood pressure, heart rate, body temperature, PaO2,

Table IV.1: The proportion of in-hospital mortality, 2-day mortality, 3-day mortality, 30-day mortality, and 1-year mortality in the admissions where adult patients were alive the first 24 hours.

Total number of admissions	**In-hospital**	**2-day**	**3-day**	**30-day**	**1-year**
35643	0.105	0.018	0.029	0.147	0.250

FiO2, urine output, white blood cells count, serum urea nitrogen level, serum bicarbonate level, sodium level, potassium level, and bilirubin level. Each longitudinal feature is sampled hourly. We also use 5 time-static features: age, admission type, and three chronic diseases diagnosis including metastatic cancer, hematologic malignancy, and acquired immunodeficiency syndrome.

4.4.2 Methods for Comparison

Overall, we consider a supervised learning method that uses only labeled data and three semi-supervised learning methods that can use both labeled and unlabeled data. The supervised learning method MMDL combines an RNN for the longitudinal features and a fully connected neural network for the time-static features, and is trained by minimizing $\mathcal{L}_d$ in (4.3) with labeled data only (Purushotham et al., 2018). We use Two-Step to refer to the two-step semi-supervised sequence learning method used in Dai and Le (2015) and Ballinger et al. (2018). Specifically, it shares the same architecture as MMDL, where the RNN is first trained by minimizing $\mathcal{L}_g$ in (4.1) and then the label prediction network is learned by minimizing $\mathcal{L}_d$ in (4.3) starting with the pre-trained initialization of $\tilde{\theta}_g$. The proposed methods are referred to as Joint-RNN and Joint-VRNN respectively, and they are trained by minimizing the overall loss $\mathcal{L}$ in (4.4) jointly. Joint-RNN shares the same architecture with MMDL and Two-Step, while Joint-VRNN substitutes the RNN with an VRNN.

For the comparison to be fair, in MMDL, Two-Step, and Joint-RNN, we adopt exactly the same neural network architecture as used by Purushotham et al. (2018). In Joint-VRNN, we also make the architecture choices as close to the former three models as possible. Specifically, we use GRU for all recurrent units and the sigmoid activation for non-linear transformations except for using a Softmax output layer to return probability scores. Dropout is applied with rate 0.1 after each sigmoid activation in the fully connected neural network. The numbers of layers and hidden units in the recurrent layer h and the fully connected neural networks are the same as those used by Purushotham et al. (2018). For VRNN, we fix the dimension of z_j as 8 and the number of hidden units in the feature extractor $\rho(z_j)$ as 32. Finally, as all patients in this dataset have the same length of longitudinal features, we simply concatenate $h(x)$ or $\tilde{\mu}_z(x)$ over the time steps when sending them as the inputs to the label prediction network, following the implementation of MMDL.

4.4.3 Experiment Setting

We split the dataset into five folds for stratified cross-validation, among which we use three folds for training, one fold for validation, and the remaining fold for testing. To examine semi-supervised learning methods with various proportions of labeled data, we randomly select a subset of the training folds as labeled data and mask labels of the remaining training folds as unlabeled data. The proportion of labeled training data varies from 1% to 100%. For each classification task, MMDL is learned using only labeled training data and the other three semi-supervised learning methods are learned using both the labeled and unlabeled training data. We report the mean and standard error of the Area under the Receiver Operating Characteristic curve (AuROC) across five testing folds to evaluate the prediction performance.

For the two joint learning methods (Joint-RNN and Joint-VRNN), we grid search the weight hyperparameter η from $\{0.001, 0.01, 0.1, 1, 10\}$ and choose the one with the highest AuROC on the validation fold separately during each round of cross-validation to avoid information leakage. For better pre-training in the Two-Step method, we further tune the non-architecture-specific hyperparameters, including the learning rate and the dropout rate, in the first step. We grid search the optimal learning rate from $\{0.001, 0.005, 0.01\}$ and the optimal dropout rate from $\{0.1, 0.2, 0.5\}$ with the lowest $\mathcal{L}_g$ on a validation set. In the second step, we initialize $\tilde{\theta}_g$ in the label prediction network with the pre-trained values and train it using labeled data.

All models are implemented in PyTorch and trained with the RMSProp optimizer. We fix the learning rate as 0.001 (except for the pre-training step of Two-Step) and the batch size as 100, following the implementation of MMDL. We use early stopping for all models when reaching the highest AuROC on the validation fold to prevent overfitting.

4.4.4 Results

Figure IV.2 shows the AuROC of the four methods under various proportions of labeled training data on five mortality-related classification tasks. First, we observe that when labels are scarce, semi-supervised learning methods significantly outperform the supervised method (MMDL) on the five tasks in most cases. This implies that semi-supervised learning methods which incorporate unlabeled data can help improve prediction performance compared to the supervised method which uses labeled data only. Further, we notice that, even in the fully labeled case, i.e. when the label percentage is 100%, modeling what comes next in a sequence as an auxiliary task (as joint learning methods do) could further improve the performance on classification tasks.

Second, we observe that the joint learning methods obtain a higher AuROC by a large margin compared to the existing two-step method, especially when predicting short-term mortality. Moreover, the gain of the joint learning methods increases

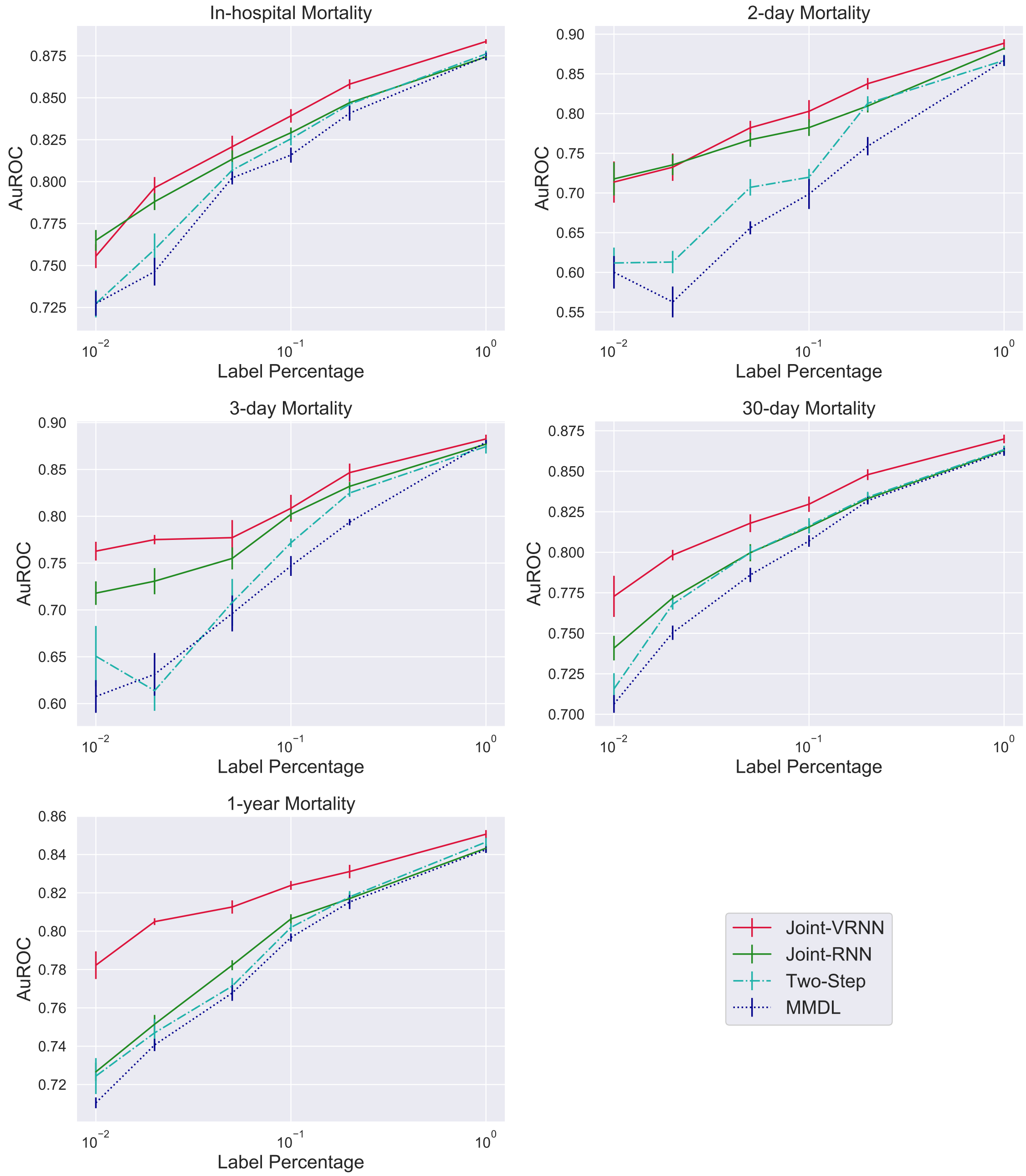

Figure IV.2: AuROC of the proposed joint learning methods (Joint-RNN and Joint-VRNN), the two-step method (Two-Step), and the supervised method (MMDL) vs the proportion of labeled training data on five tasks. The horizontal axis is in the logarithmic scale with base 10. The results are averaged over five testing folds and the error bars indicate the standard error of the mean.

as the label percentage decreases. This implies that, although the pre-training step of the two-step method might provide a potentially good initialization, the lack of supervision from labels in the pre-training step would lead to limited improvement on prediction performance in the second step. Instead, the proposed joint learning methods can take advantage of available labels and learn representations of the longitudinal features under supervision from both labeled and unlabeled data.

Third, as shown in Figure IV.2, the Joint-VRNN that contains the stochastic recurrent layer further improves the prediction performance in comparison with the Joint-RNN. The gain is especially obvious for the long-term mortality prediction. This extends the observation of the benefit of using latent random recurrent layers in previous literature to modeling longitudinal features.

4.5 Discussion

In this chapter, we propose a semi-supervised joint learning method for classifying longitudinal features, 2 with an application to clinical events. With joint learning, the feature representation of the longitudinal information is learned under supervision from both unlabeled and labeled data so that related data can be separated well corresponding to the labels. We compare the proposed methods with the existing supervised learning method and two-step semi-supervised learning method. Our experimental results verify that, by incorporating unlabeled data, semi-supervised learning methods outperform the supervised method when labels are scarce, and among the semi-supervised learning methods, the proposed joint learning methods can further improve the prediction accuracy compared to the two-step method in most cases.

Notably, the horizontal difference between the curves of semi-supervised and supervised methods indicates the difference on the usage of labeled training data to maintain the same prediction performance. For example, as shown in Figure IV.2, the Joint-VRNN method uses 2% labels to obtain 80% AuROC for 1-year mortality prediction while the supervised method needs 10% labels to achieve the same performance. Therefore, the usage of semi-supervised learning met 2hods could help reduce the cost of collecting clinical labels when building prediction models for applications in healthcare.

We should note a few remarks on the proposed joint learning methods. First, when there are multiple prediction tasks, the two-step method has the advantage that the pre-training step only needs to be done once while the joint learning methods require the training of the sequence generative network for each task. Second, compared to the two-step method, the joint learning methods have one additional hyperparameter η to be tuned. In practice, though, a simple grid search of η is enough to obtain good performance as shown in our experiments. Third, Joint-VRNN has a higher computational cost than Joint-RNN due to the sequential sampling. However, Joint-VRNN demonstrates promising improvements in prediction accuracy compared to Joint-RNN, which is especially important in healthcare applications.

CHAPTER V

Conclusion and Future Directions

Motivated by the growing scale of datasets, the diversity of data structures, and incomplete observations that are often encountered in various real-world applications, this dissertation has focused on three directions. First, in Chapter II, we proposed an ODE approach for censored data in survival analysis, which provides a unified modeling framework and a scalable procedure for estimation and inference based on well-established numerical solvers and tools for ODEs. Remarkably, the proposed ODE approach improves computational efficiency and model expression power while maintaining statistical efficiency. Next, in Chapter III, we developed a latent space method for modeling heterogeneous network data. In particular, we focus on signed networks with edge heterogeneity and the proposed latent space method accommodates the important balance theory in social science and provides interpretable and informative embeddings for network data. Finally, in Chapter IV, we proposed a semi-supervised learning method for risk prediction on longitudinal clinical events. Motivated by the longitudinal nature of EHR and the scarcity of annotated data, the proposed method requires fewer labeled training data to obtain the same prediction performance than using labeled data alone.

Moreover, there are some intriguing directions that are worth of future exploration.

Bridging Differential Equations and Statistical Learning The differential equation (DE) is a fundamental tool for describing dynamic systems in many disciplines, such as explaining the laws of physics, understanding the growth of diseases, and demonstrating the motion of economic systems. In particular, recent work in deep learning has forged a connection with DE. Such work is built in two ways: using deep neural networks for solving high-dimensional partial differential equations, where finite difference methods become infeasible (Sirignano and Spiliopoulos, 2018); and using ODEs for continuous modeling, such as modeling continuous-depth neural networks to improve memory and parameter efficiency (Chen et al., 2018), and modeling continuous normalizing flows to reduce the computational burden while increasing expressiveness (Grathwohl et al., 2019). Chapter II further identifies a new direction to leverage DE to statistical learning for enhancing modeling flexibility and computational efficiency while maintaining statistical efficiency, which makes DE a promising and versatile tool for statistical learning under various modern data settings.

One future direction is exploiting the ODE notion for general counting processes. For example, we can analyze the recurrent event data in longitudinal studies and the repeated directed interactions in dynamic networks through multivariate counting processes, of which the intensity function can be modeled by DE. This approach may inherit the merits of Chapter II in terms of modeling and computation and can be potentially generalized to handle *competing risks*, interval-censored data, and dynamic risk predictions using time-dependent covariates.

Embedding Learning and Statistical Inference for Network Data To analyze heterogeneous network data, the network embedding method provides a low-dimensional vector representation for each node that preserves the connectivity patterns of networks. Learned embedding is useful for identifying underlying patterns, visualizing networks, and downstream learning tasks (e.g., node clustering and classification, and link prediction). Chapter III shows that the latent space approach is flexible to develop interpretable and statistically principled network embedding via model-based learning and inference. It would be worthwhile to further investigate how the latent space approach can be used to model heterogeneous networks with node attributes and network data beyond pairwise interactions.

In addition, while network embedding approaches have been successfully applied to many downstream tasks, little is done to theoretically investigate how learned embeddings affect the downstream task performance. The unsupervised embedding such as Chapter III is entirely task-independent and may be too generic due to the lack of supervision. One promising direction is to develop supervised or semi-supervised network embedding learning methods through joint modeling when labeled training data for the learning task is available.

APPENDICES

APPENDIX A

Appendix of Chapter II

This Appendix is structured as follows. We provide the detailed derivation of the local sensitivity analysis and optimization algorithm in Section A.1. We present the proposed general M-theorem for bundled parameters (Theorem 2.3.3) and its proof in Section A.2. The proofs of Theorems 2.3.1 and 2.3.2 are given in Section A.3, those of Propositions 2.2.1 and 2.2.2 are given in Section A.4. We further establish the convergence rate and the asymptotic normality of the proposed sieve estimator for the general class of ODE models in the presence of covariates Z with time-varying coefficients in Section A.5. Additional simulation studies are provided in Section A.6. A brief introduction to the partial likelihood-based method and discrete-time survival models are given in Sections A.7 and A.8 respectively. The tuning ranges of hyperparameters for the neural-network-based models are listed in Section A.9.

A.1 Optimization Algorithm With Local Sensitivity Analysis

In this section, we first derive two types of local sensitivity analysis that can be used to compute the gradient of the log-likelihood function when it contains the solution of a general ODE. When the ODE is separable in the model formulation, we introduce a trick to further accelerate the evaluation of the objective for n independent observations in subsection A.1.1.

We consider any parameterized survival model in the form of

$$\begin{cases} d\Lambda_x(t)/dt = f(t, \Lambda_x(t); x, \theta) \\ \Lambda_x(t_0) = c(x, \theta) \end{cases}, \tag{A.1}$$

where θ denotes all the parameters. For example, for the general class of ODE models in (2.8), the function f is given by the right hand side of (2.8), the parameter θ consists of β, a, and b, the initial time point $t_0 = 0$, and the initial value $c(\cdot)$ equals to zero. Denote the solution of (A.1) by $\Lambda_x(t; \theta)$. Then under the non-informative censoring, the log-likelihood function is given by

$$l_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left[\Delta_i \log f\left(Y_i, \Lambda_{X_i}(Y_i; \theta); X_i, \theta\right) - \Lambda_{X_i}(Y_i; \theta) \right].$$

To obtain the maximum likelihood estimator, we propose a gradient-based optimization algorithm which utilizes the local sensitivity analysis to compute the gradient. By applying the chain rule, the gradient is given by

$$\frac{d\, l_n(\theta)}{d\theta} = \frac{1}{n} \sum_{i=1}^{n} \Bigg\{ \left[\Delta_i \frac{f_2'\left(Y_i, \Lambda_{X_i}(Y_i; \theta); X_i, \theta\right)}{f\left(Y_i, \Lambda_{X_i}(Y_i; \theta); X_i, \theta\right)} - 1 \right] \frac{\partial \Lambda_{x_i}(Y_i; \theta)}{\partial \theta} + \Delta_i \frac{f_4'\left(Y_i, \Lambda_{X_i}(Y_i; \theta); X_i, \theta\right)}{f\left(Y_i, \Lambda_{X_i}(Y_i; \theta); X_i, \theta\right)} \Bigg\},$$

where we use the subscript 2 and 4 in the derivatives to indicate that the derivatives are taken with respect to the first and the fourth argument of the function f respectively. Then as long as we can derive the gradient of $\Lambda_x(y; \theta)$ with respect to θ for a given y, we can obtain the gradient of the likelihood function for faster gradient-based computations.

There are two commonly used types of local sensitivity analyses: forward sensitivity analysis and adjoint sensitivity analysis (Dickinson and Gelinas, 1976; Petzold et al., 2006). We first derive the corresponding ODE for the forward sensitivity analysis. Denote the partial derivatives of $f(t, \Lambda; x, \theta)$ with respect to θ and Λ by f_θ' and f_Λ', respectively. Under certain smoothness condition of f, there is one unique solution $\Lambda_x(t;\theta)$ of (A.1) and it satisfies

$$\Lambda_x(t;\theta) = \int_{t_0}^{t} f(s,\Lambda_x(s;\theta);x,\theta)ds + c(x,\theta).$$

By interchanging the integral and partial differential operators, it follows that

$$\begin{aligned}\frac{\partial\Lambda_x(t;\theta)}{\partial\theta} &= \frac{\partial}{\partial\theta}\int_{t_0}^{t} f(s,\Lambda_x(s;\theta);x,\theta)ds + c'_\theta(x,\theta) \qquad \text{(A.2)}\\ &= \int_{t_0}^{t}\left(f'_\theta(s,\Lambda_x(s;\theta);x,\theta) + f'_\Lambda(s,\Lambda_x(s;\theta);x,\theta)\frac{\partial\Lambda_x(s;\theta)}{\partial\theta}\right)ds + c'_\theta(x,\theta),\end{aligned}$$

where $c'_\theta(x,\theta)$ is the derivative of $c(x,\theta)$ with respect to θ. Therefore, $\partial\Lambda_x(y;\theta)/\partial\theta = F_1(y)$ with F_1 satisfying

$$\begin{cases} dF_1(t)/dt = f'_\theta(t,\Lambda_x(t;\theta);x,\theta) + f'_\Lambda(t,\Lambda_x(t;\theta);x,\theta)\cdot F_1 \\ F_1(t_0) = c'_\theta(x,\theta) \end{cases}. \qquad \text{(A.3)}$$

After plugging $t_0 = 0$ and $c(\cdot) = 0$, (A.3) becomes the initial value problem (2.9) in Section 2.3.1.

Next, we derive the corresponding ODE for the adjoint sensitivity analysis. Since $\Lambda_x(t;\theta)$ is solution of (A.1), for some appropriately chosen differentiable function $\kappa(t,\theta)$ to be specified later, we have

$$\Lambda_x(t;\theta) = \Lambda_x(t;\theta) - \int_{t_0}^{t}\kappa(s,\theta)\Big[\frac{\partial\Lambda_x(s;\theta)}{\partial s} - f(s,\Lambda_x(s,\theta);x,\theta)\Big]ds.$$

By taking derivatives with respect to θ on both sides, it follows that

$$\begin{aligned}\frac{\partial\Lambda_x(t;\theta)}{\partial\theta} &= \frac{\partial\Lambda_x(t;\theta)}{\partial\theta} - \frac{\partial}{\partial\theta}\int_{t_0}^{t}\kappa(s,\theta)\Big[\frac{\partial\Lambda_x(s;\theta)}{\partial s} - f(s,\Lambda_x(s,\theta);x,\theta)\Big]ds\\ &= \frac{\partial\Lambda_x(t;\theta)}{\partial\theta} - \int_{t_0}^{t}\kappa(s,\theta)\frac{\partial}{\partial\theta}\Big[\frac{\partial\Lambda_x(s;\theta)}{\partial s} - f(s,\Lambda_x(s,\theta);x,\theta)\Big]ds\\ &= \int_{t_0}^{t}(1+\kappa(s,\theta))\frac{\partial}{\partial\theta}f(s,\Lambda_x(s;\theta);x,\theta)ds + c'_\theta(x,\theta)\\ &\quad - \int_{t_0}^{t}\kappa(s,\theta)\frac{\partial}{\partial s}\Big[\frac{\partial\Lambda_x(s;\theta)}{\partial\theta}\Big]ds,\end{aligned}$$

where the second equality holds because

$$\int_{t_0}^{t} \frac{\partial \kappa(s,\theta)}{\partial \theta}\Big[\frac{\partial \Lambda_x(s;\theta)}{\partial s} - f(s,\Lambda_x(s,\theta);x,\theta)\Big]\,ds = 0,$$

and the last equality holds by plugging (A.2) and exchanging the order of derivatives.

Using integral by parts, we have

$$\int_{t_0}^{t} \kappa(s,\theta)\frac{d}{ds}\Big[\frac{d\Lambda_x(s;\theta)}{d\theta}\Big]\,ds + \int_{t_0}^{t} \frac{d\kappa(s,\theta)}{ds}\frac{d\Lambda_x(s;\theta)}{d\theta}\,ds = \Big(\kappa(s,\theta)\frac{d\Lambda_x(s;\theta)}{d\theta}\Big)\Big|_{t_0}^{t}.$$

Then it follows that

$$\begin{aligned}\frac{d\Lambda_x(t;\theta)}{d\theta} &= \int_{t_0}^{t} (1+\kappa(s,\theta))\left(f'_\theta(s,\Lambda_x(s;\theta);x,\theta) + f'_\Lambda(s,\Lambda_x(s;\theta);x,\theta)\frac{d\Lambda_x(s;\theta)}{d\theta}\right)ds \\ &\quad + c'_\theta(x,\theta) + \int_{t_0}^{t} \frac{d\kappa(s,\theta)}{ds}\frac{d\Lambda_x(s;\theta)}{d\theta}\,ds - \Big(\kappa(s,\theta)\frac{d\Lambda_x(s;\theta)}{d\theta}\Big)\Big|_{t_0}^{t} \\ &= \int_{t_0}^{t} (1+\kappa(s,\theta))f'_\theta(s,\Lambda_x(s;\theta);x,\theta)\,ds \\ &\quad + \int_{t_0}^{t} \frac{d\Lambda_x(s;\theta)}{d\theta}\left(\frac{d\kappa(s,\theta)}{ds} + (1+\kappa(s,\theta))f'_\Lambda(s,\Lambda_x(s;\theta);x,\theta)\right)ds \\ &\quad + (1+\kappa(t_0,\theta))c'_\theta(x,\theta) - \kappa(t,\theta)\frac{d\Lambda_x(t;\theta)}{d\theta}.\end{aligned}$$

Denote $\tilde{\kappa}(t,\theta) \triangleq \kappa(t,\theta)+1$ and choose proper $\tilde{\kappa}(t,\theta)$ that satisfies

$$\begin{cases} d\tilde{\kappa}(t,\theta)/dt = -\tilde{\kappa}(t;\theta)f'_\Lambda(t,\Lambda_x(t,\theta);x,\theta) \\ \tilde{\kappa}(y;\theta) = 1 \end{cases}, \tag{A.4}$$

then the gradient of $\Lambda_x(t,\theta)$ with respect to θ is given by

$$\frac{d\Lambda_x(y;\theta)}{d\theta} = \int_{t_0}^{y} \tilde{\kappa}(s,\theta)f'_\theta(s,\Lambda_x(s;\theta);x,\theta)\,ds + \tilde{\kappa}(t_0,\theta)c'_\theta(x,\theta).$$

After plugging $t_0 = 0$ and $c(\cdot) = 0$, the above equation becomes

$$\frac{d\Lambda_x(y;\theta)}{d\theta} = \int_{0}^{y} \tilde{\kappa}(s,\theta)f'_\theta(s,\Lambda_x(s;\theta);x,\theta)\,ds.$$

Together with (A.4), it shows that the solution of (2.10) at $t=0$ gives the gradient of $\Lambda_x(y;\theta)$ with respect to θ. Note that to solve (2.10) at $t=0$, it requires evaluating the entire trajectory of $\Lambda_x(t, \theta)$ from y to 0. In our implementation, we combine ODEs (A.1) and (2.10) into a larger ODE system, i.e.,

$$\begin{cases} (\Lambda'(t); \kappa'(t); F_2'(t)) = (f(t, \Lambda; \theta); -\kappa \cdot f_\Lambda'(t, \Lambda; \theta); -\kappa \cdot f_\theta'(t, \Lambda; \theta)) \\ (\Lambda(t); \kappa(t); F_2(t))|_{t=y} = (\Lambda_x(y; \theta); 1; \mathbf{0}) \end{cases},$$

and evaluate it at $t = 0$, where $\Lambda_x(y; \theta)$ is available when computing the likelihood function. As discussed in Section 3 of the main text, the proposed estimation methods can be easily implemented using existing computing packages.

A.1.1 Acceleration trick for simultaneously solving separable ODEs for n independent observations

Recall that evaluating the log-likelihood function requires solving ODEs for n independent observations. For a general ODE model, as suggested in Remark II.4, we can use either the adjoint method along with parallel computing or the forward method by combining n ODEs into a large ODE system with n dimensions. The complexity of both methods scales linearly with the sample size. We further introduce a trick to reduce the absolute magnitude of computing time for separable ODEs, which cover the general class of ODE models in (2.2) as a special case.

Specifically, we consider the separable ODE model in the form of

$$\begin{cases} d\Lambda_x(t)/dt = f_1(t; x, \theta_1) \cdot f_2(\Lambda_x; \theta_2) \\ \Lambda_x(t_0) = c \end{cases}, \tag{A.5}$$

with two functions f_1 and f_2. In particular, for the general class of ODE models in (2.8),

$$f_1(t; x, z, \theta_1) = \exp\left(x^T\beta + \sum_{l=0}^{d_2} \sum_{j=1}^{q_n^1} a_j^l B_j^1(t) z_l \right)$$

and

$$f_2(\Lambda_{x,z}; \theta_2) = \exp\left(\sum_{j=1}^{q_n^2} b_j B_j^2(\Lambda_{x,z}(t)) \right).$$

For n independent observations $\{\Delta_i, X_i, Y_i\}_{i=1}^n$, we need to evaluate the solution of

n different ODEs in (A.5), each of which is associated with X_i, at their respective observed times Y_i. The acceleration trick is based on the key observation that solving (A.5) at y is equivalent to solving the problem

$$\begin{cases} dG(t)/dt = f_2(G;\theta_2) \\ G(t_0) = c \end{cases} \tag{A.6}$$

at $\int_{t_0}^{y} f_1(t;x,\theta_1)\,dt + t_0$, i.e.,

$$\Lambda_x(y;\theta_1,\theta_2) = G(\int\limits_{t_0}^{y} f_1(t;x,\theta_1)\,dt + t_0;\theta_2).$$

Therefore, we can instead solve a single ODE (A.6) at n different points $\{t_0 + \int_{t_0}^{Y_i} f_1(t;X_i,\theta_1)\,dt\}_{i=1}^{n}$ to compute $\Lambda_{X_i}(Y_i;\theta_1,\theta_2)$ for $1 \leq i \leq n$. Moreover, given $\Lambda_{X_i}(Y_i;\theta_1,\theta_2)$, the gradient of $\Lambda_{X_i}(Y_i;\theta_1,\theta_2)$ with respect to θ_1 can be computed by

$$\frac{\partial \Lambda_{X_i}(Y_i;\theta_1,\theta_2)}{\partial \theta_1} = f_2(\Lambda_{X_i}(Y_i;\theta_1,\theta_2);\theta_2) \int\limits_{t_0}^{Y_i} \frac{\partial f_1(t;X_i,\theta_1)}{\partial \theta_1} dt.$$

And we can obtain the gradient of $\Lambda_{X_i}(Y_i;\theta_1,\theta_2)$ with respect to θ_2 by solving another single ODE at n different points:

$$\frac{\partial \Lambda_{X_i}(Y_i;\theta_1,\theta_2)}{\partial \theta_2} = G_2(\int\limits_{t_0}^{y} f_1(t;x,\theta_1)\,dt + t_0;\theta_2),$$

where $\tilde{G}(\cdot;\theta_2)$ is the solution of

$$\begin{cases} d\tilde{G}(t)/dt = f_2{}'_{\theta_2}(G;\theta_2) + f_2{}'_{G}(G;\theta_2) \cdot \tilde{G} \\ \tilde{G}(t_0) = 0 \end{cases}.$$

Based on our experiments, the proposed acceleration trick can significantly reduce the absolute computing time of simultaneously solving separable ODEs for n independent observations.

A.2 The General Sieve M-theorem for Bundled Parameters (Theorem 2.3.3) and Its Proof

In this section, we establish a new general sieve M-theorem for studying the asymptotic normality of M-estimators when the estimation criterion is parameterized with more general bundled parameters. Note that the proposed M-theorem significantly differs from Theorem 2.1 in Ding and Nan (2011) and Theorem 6.1 in Wellner and Zhang (2007). They consider either well-separated parameters (Wellner and Zhang, 2007) or bundled parameters where the nuisance parameter can be a function of only the finite-dimensional parameters (Ding and Nan, 2011); while we consider a more general scenario of bundled parameters where the nuisance parameter can be a function of both the finite-dimensional parameter β and other infinite-dimensional parameters. Therefore, the proposed theorem nontrivially extends the asymptotic distributional theories for M-estimation under this general scenario and is crucial for studying the asymptotic normality of the sieve MLE for the general ODE model in (2).

Specifically, given i.i.d. observations $W_1, \cdots, W_n \in \mathcal{W}$, we maximize an objective function

$$\frac{1}{n}\sum_{1}^{n} m(\beta, \boldsymbol{\gamma}(\cdot), \zeta(\cdot, \beta, \boldsymbol{\gamma}); W_i)$$

to estimate the unknown parameters $(\beta, \boldsymbol{\gamma}(\cdot), \zeta(\cdot, \beta, \boldsymbol{\gamma}))$. Here $\beta \in \mathbf{R}^{d_1}$ denotes the finite-dimensional parameter of interest, $\boldsymbol{\gamma}(\cdot) = (\gamma_1(\cdot), \ldots, \gamma_{d_2}(\cdot))$ denotes nuisance infinite-dimensional parameters and $\zeta(\cdot, \beta, \boldsymbol{\gamma})$ denotes another nuisance infinite-dimensional parameter that is a function of β and $\boldsymbol{\gamma}(\cdot)$. To accommodate this different and challenging scenario bundled parameters, we develop a new general sieve M-theorem. We firstly introduce notation in Section A.2.1, and establish the asymptotic normality of the sieve estimator that maximizes the objective function over some sieve parameter space in Section A.2.2.

A.2.1 Notation

Here we follow notation used in Ding and Nan (2011) and Wellner and Zhang (2007). Let $\theta = (\beta, \boldsymbol{\gamma}(\cdot), \zeta(\cdot, \beta, \boldsymbol{\gamma}))$, $\beta \in \mathcal{B} \subset \mathbf{R}^{d_1}$, $\boldsymbol{\gamma} \in \Gamma^{d_2}$, and $\zeta \in \mathcal{H}$, where $\mathcal{B}$ is the parameter space of β, Γ is a class of functions mapping from $\mathcal{W}$ to $\mathbf{R}$ and $\mathcal{H}$ is a class of functions mapping from $\mathcal{W} \times \mathcal{B} \times \Gamma^{d_2}$ to $\mathbf{R}$. Let $\Theta = \mathcal{B} \times \Gamma^{d_2} \times \mathcal{H}$ be the parameter space of θ. The distance between θ_1 and $\theta_2 \in \Theta$ is defined as

$$d(\theta_1, \theta_2) = \{\|\beta_1 - \beta_2\|^2 + \sum_{j=1}^{d_2} \|\gamma_{1j} - \gamma_{2j}\|_\Gamma^2 + \|\zeta_1(\cdot, \beta_1, \boldsymbol{\gamma}_1) - \zeta_2(\cdot, \beta_2, \boldsymbol{\gamma}_2)\|_{\mathcal{H}}^2\}^{1/2},$$

where $\|\cdot\|$ is the Euclidean norm, $\|\cdot\|_\Gamma$ is some norm of Γ, and $\|\cdot\|_{\mathcal{H}}$ is some norm of $\mathcal{H}$. Let Θ_n be the sieve parameter space, where $\Theta_n \subset \Theta_{n+1} \subset \cdots \subset \Theta$ and the sequence becomes dense as $n \to \infty$. We obtain the sieve M-estimator $\hat{\theta}_n = (\hat{\beta}_n, \hat{\boldsymbol{\gamma}}_n, \hat{\zeta}_n(\cdot, \hat{\beta}_n, \hat{\boldsymbol{\gamma}}_n)) \in \Theta_n$ by maximizing the objective function over the sieve parameter space. We study the asymptotic normality of the sieve M-estimator of the Euclidean parameter of interest, $\hat{\beta}_n$, as follows.

For any fixed $\gamma(\cdot) \in \Gamma$, let $\{\gamma_\eta(\cdot) : \eta$ in a neighborhood of $0 \in \mathbf{R}\}$ be a smooth curve in Γ running through $\gamma(\cdot)$ at $\eta = 0$, that is $\gamma_\eta(\cdot)|_{\eta=0} = \gamma(\cdot)$. Similarly, for any fixed $\zeta(\cdot, \beta, \boldsymbol{\gamma}) \in \mathcal{H}$, let $\{\zeta_\eta(\cdot, \beta, \boldsymbol{\gamma}) : \eta$ in a neighborhood of $0 \in \mathbf{R}\}$ be a smooth curve in $\mathcal{H}$ running through $\zeta(\cdot, \beta, \boldsymbol{\gamma})$ at $\eta = 0$, that is $\zeta_\eta(\cdot, \beta, \boldsymbol{\gamma})|_{\eta=0} = \zeta(\cdot, \beta, \boldsymbol{\gamma})$. Assume all $\zeta(\cdot, \beta, \boldsymbol{\gamma}) \in \mathcal{H}$ are twice Frechet differentiable with respect to β and $\boldsymbol{\gamma}$, and denote

$$\mathbb{V} = \{v : v(\cdot) = \frac{\partial \gamma_\eta(\cdot)}{\partial \eta}|_{\eta=0}, \gamma_\eta \in \Gamma\},$$

$$\mathbb{H} = \{h : h(\cdot, \beta, \boldsymbol{\gamma}) = \frac{\partial \zeta_\eta(\cdot, \beta, \boldsymbol{\gamma})}{\partial \eta}|_{\eta=0}, \zeta_\eta \in \mathcal{H}, \beta \in \mathcal{B}, \boldsymbol{\gamma} \in \Gamma^{d_2}\}.$$

Assume the objective function m is twice Frechet differentiable. For $1 \le j \le d_2$, we use the subscript 1, $2^{(j)}$ or 3 in the derivatives to indicate that the derivatives are taken with respect to the first, the j-th component of the second or the third argument of the function, respectively. We use function v or h inside the square brackets to denote the direction of the functional derivative with respect to γ_j or ζ. Since for a small δ, we have $\zeta(\cdot, \beta + \delta, \boldsymbol{\gamma}) - \zeta(\cdot, \beta, \boldsymbol{\gamma}) = \zeta'_\beta(\cdot, \beta, \boldsymbol{\gamma})\delta + o(\delta)$, where

$\zeta'_\beta(\cdot,\beta,\boldsymbol{\gamma}) = \partial\zeta(\cdot,\beta,\boldsymbol{\gamma})/\partial\beta$; then as shown in Ding and Nan (2011) on page 3036, it follows that

$$\lim_{\delta\to 0}\frac{1}{\delta}\{m(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta+\delta,\boldsymbol{\gamma});W) - m(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)\}$$
$$= m'_3(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[\zeta'_\beta(\cdot,\beta,\boldsymbol{\gamma})],$$

$$\lim_{\delta\to 0}\frac{1}{\delta}\{m'_3(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta+\delta,\boldsymbol{\gamma});W)[h(\cdot,\beta,\boldsymbol{\gamma})] - m'_3(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[h(\cdot,\beta,\boldsymbol{\gamma})]\}$$
$$= m''_{33}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[h(\cdot,\beta,\boldsymbol{\gamma}),\zeta'_\beta(\cdot,\beta,\boldsymbol{\gamma})],$$

$$\lim_{\delta\to 0}\frac{1}{\delta}\{m'_{2(j)}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta+\delta,\boldsymbol{\gamma});W)[v] - m'_2(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[v]\}$$
$$= m''_{2(j)3}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[v,\zeta'_\beta(\cdot,\beta,\boldsymbol{\gamma})], \quad \text{for } 1\le j\le d_2,$$

and

$$\lim_{\delta\to 0}\frac{1}{\delta}\{m'_3(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[h(\cdot,\beta+\delta,\boldsymbol{\gamma})] - m'_3(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[h(\cdot,\beta,\boldsymbol{\gamma})]\}$$
$$= m'_3(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[h'_\beta(\cdot,\beta,\boldsymbol{\gamma})].$$

Let $e_j = (0,\ldots,1,\ldots,0)\in\mathbf{R}^{d_2}$ with the j-th element being 1. For $1\le j\le d_2$, we have $\zeta(\cdot,\beta,\boldsymbol{\gamma}+v\cdot e_j) - \zeta(\cdot,\beta,\boldsymbol{\gamma}) = \zeta'_{\gamma_j}(\cdot,\beta,\boldsymbol{\gamma})[v] + o(\|v\|_\Gamma)$ for a small v; then by the definition of functional derivatives, it follows that, for $1\le j\le d_2$,

$$\begin{aligned}
&m(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma}+v\cdot e_j);W) - m(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)\\
&= m(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma}) + \zeta'_{\gamma_j}(\cdot,\beta,\boldsymbol{\gamma})[v] + o(\|v\|_\Gamma);W) - m(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)\\
&= \{m(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma}) + \zeta'_{\gamma_j}(\cdot,\beta,\boldsymbol{\gamma})[v] + o(\|v\|_\Gamma);W)\\
&\quad - m(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma}) + \zeta'_{\gamma_j}(\cdot,\beta,\boldsymbol{\gamma})[v];W)\}\\
&\quad + \{m(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma}) + \zeta'_{\gamma_j}(\cdot,\beta,\boldsymbol{\gamma})[v];W) - m(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)\}\\
&= m'_3(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma}) + \zeta'_{\gamma_j}(\cdot,\beta,\boldsymbol{\gamma})[v];W)[o(\|v\|_\Gamma)] +\\
&\quad m'_3(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[\zeta'_{\gamma_j}(\cdot,\beta,\boldsymbol{\gamma})[v]] + o(\|\zeta'_{\gamma_j}(\cdot,\beta,\boldsymbol{\gamma})[v]\|_\Gamma)\\
&= m'_3(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[\zeta'_{\gamma_j}(\cdot,\beta,\boldsymbol{\gamma})[v]] + o(\|v\|_\Gamma),
\end{aligned}$$

where the last equality holds because

$$\lim_{v\to 0} m(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma}) + \zeta'_{\gamma_j}(\cdot,\beta,\boldsymbol{\gamma})[v];W)\Big[\frac{o(\|v\|_\Gamma)}{\|v\|_\Gamma}\Big] = 0,$$

and $o(\|\zeta'_{\gamma_j}(\cdot,\beta,\boldsymbol{\gamma})[v]\|_\Gamma) = o(\|v\|_\Gamma)$ for bounded functional derivatives. Similarly we have for $1 \leq j, \ell \leq d_2$,

$$
\begin{aligned}
&m'_{2(j)}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma}+v\cdot e_\ell);W)[v_1] - m'_{2(j)}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[v_1]\\
&\quad = m''_{2(j)3}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[v_1,\zeta'_{\gamma_\ell}(\cdot,\beta,\boldsymbol{\gamma})[v]] + o(\|v\|_\Gamma),
\end{aligned}
$$

$$
\begin{aligned}
&m'_3(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma}+v\cdot e_j);W)[h(\cdot,\beta,\boldsymbol{\gamma})] - m'_3(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[h(\cdot,\beta,\boldsymbol{\gamma})]\\
&= m''_{33}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[h(\cdot,\beta,\boldsymbol{\gamma}),\zeta'_{\gamma_j}(\cdot,\beta,\boldsymbol{\gamma})[v]] + o(\|v\|_\Gamma),
\end{aligned}
$$

$$
\begin{aligned}
&m'_3(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[h(\cdot,\beta,\boldsymbol{\gamma}+v\cdot e_j)] - m'_3(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[h(\cdot,\beta,\boldsymbol{\gamma})]\\
&= m'_3(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[h'_{\gamma_j}(\cdot,\beta,\boldsymbol{\gamma})[v]] + o(\|v\|_\Gamma).
\end{aligned}
$$

Based on the chain rule of the functional derivative, we have for $1 \leq j, \ell \leq d_2$,

$$
\begin{aligned}
m'_\beta(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W) &= \frac{\partial m(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)}{\partial\beta}\\
&= m'_1(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)\\
&\quad + m'_3(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[\zeta'_\beta(\cdot,\beta,\boldsymbol{\gamma})],\\
m'_{\gamma_j}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[v] &= m'_{2(j)}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[v]\\
&\quad + m'_3(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[\zeta'_{\gamma_j}(\cdot,\beta,\boldsymbol{\gamma})[v]],\\
m'_\zeta(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[h] &= m'_3(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[h(\cdot,\beta,\boldsymbol{\gamma})],\\
m''_{\beta\beta}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W) &= \frac{\partial m'_\beta(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)}{\partial\beta}\\
&= m''_{11}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)\\
&\quad + m''_{13}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[\zeta'_\beta(\cdot,\beta,\boldsymbol{\gamma})]\\
&\quad + m''_{31}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[\zeta'_\beta(\cdot,\beta,\boldsymbol{\gamma})]\\
&\quad + m''_{33}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[\zeta'_\beta(\cdot,\beta,\boldsymbol{\gamma}),\zeta'_\beta(\cdot,\beta,\boldsymbol{\gamma})]\\
&\quad + m'_3(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[\zeta''_{\beta\beta}(\cdot,\beta,\boldsymbol{\gamma})],\\
m''_{\gamma_j\gamma_\ell}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[v_1,v_2] &= m''_{2(j)2(\ell)}\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[v_1,v_2]\\
&\quad +m''_{2(j)3}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[v_1,\zeta'_{\gamma_\ell}(\cdot,\beta,\boldsymbol{\gamma})[v_2]]
\end{aligned}
$$

$$\begin{aligned}
&+m''_{32^{(\ell)}}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[\zeta'_{\gamma_j}(\cdot,\beta,\boldsymbol{\gamma})[v_1],v_2]\\
&+m''_{33}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[\zeta'_{\gamma_j}(\cdot,\beta,\boldsymbol{\gamma})[v_1],\zeta'_{\gamma_\ell}(\cdot,\beta,\boldsymbol{\gamma})[v_2]]\\
&+m'_3(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[\zeta''_{\gamma_j\gamma_\ell}(\cdot,\beta,\boldsymbol{\gamma})[v_1,v_2]],
\end{aligned}$$

$$\begin{aligned}
m''_{\zeta\zeta}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[h_1,h_2] &= m''_{33}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[h_1(\cdot,\beta,\boldsymbol{\gamma}),h_2(\cdot,\beta,\boldsymbol{\gamma})],\\
m''_{\gamma_j\beta}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[v] &= \frac{\partial m'_{\gamma_j}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[v]}{\partial\beta}\\
&= m''_{2^{(j)}1}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[v]\\
&\quad + m''_{2^{(j)}3}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[v,\zeta'_\beta(\cdot,\beta,\boldsymbol{\gamma})]\\
&\quad + m''_{31}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[\zeta'_{\gamma_j}(\cdot,\beta,\boldsymbol{\gamma})[v]]\\
&\quad + m''_{33}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[\zeta'_{\gamma_j}(\cdot,\beta,\boldsymbol{\gamma})[v],\zeta'_\beta(\cdot,\beta,\boldsymbol{\gamma})]\\
&\quad + m'_3(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[\zeta''_{\gamma_j\beta}(\cdot,\beta,\boldsymbol{\gamma})[v]]\\
m''_{\zeta\beta}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[h] &= \frac{\partial m'_\zeta(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[h]}{\partial\beta}\\
&= m''_{31}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[h(\cdot,\beta,\boldsymbol{\gamma})]\\
&\quad + m''_{33}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[h(\cdot,\beta,\boldsymbol{\gamma}),\zeta'_\beta(\cdot,\beta,\boldsymbol{\gamma})]\\
&\quad + m'_3(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[h'_\beta(\cdot,\beta,\boldsymbol{\gamma})],\\
m''_{\zeta\gamma_j}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[h,v] &= m''_{32^{(j)}}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[h(\cdot,\beta,\boldsymbol{\gamma}),v]\\
&\quad + m''_{33}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[h(\cdot,\beta,\boldsymbol{\gamma}),\zeta'_{\gamma_j}(\cdot,\beta,\boldsymbol{\gamma})[v]]\\
&\quad + m'_3(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[h'_{\gamma_j}(\cdot,\beta,\boldsymbol{\gamma})[v]],\\
m''_{\gamma_j\zeta}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[v,h] &= m''_{2^{(j)}3}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[v,h(\cdot,\beta,\boldsymbol{\gamma})]\\
&\quad + m''_{33}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[\zeta'_{\gamma_j}(\cdot,\beta,\boldsymbol{\gamma})[v],h(\cdot,\beta,\boldsymbol{\gamma})]\\
&\quad + m'_3(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[h'_{\gamma_j}(\cdot,\beta,\boldsymbol{\gamma})[v]].
\end{aligned}$$

Following Wellner and Zhang (2007), we further define

$$\begin{aligned}
S'_\beta(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma})) &= Pm'_\beta(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W),\\
S'_{\gamma_j}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma}))[v] &= Pm'_{\gamma_j}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[v],
\end{aligned}$$

$$S'_{\zeta}(\beta, \boldsymbol{\gamma}(\cdot), \zeta(\cdot, \beta, \boldsymbol{\gamma}))[h] = Pm'_{\zeta}(\beta, \boldsymbol{\gamma}(\cdot), \zeta(\cdot, \beta, \boldsymbol{\gamma}); W)[h],$$

$$S'_{\beta,n}(\beta, \boldsymbol{\gamma}(\cdot), \zeta(\cdot, \beta, \boldsymbol{\gamma})) = \mathbb{P}_n m'_{\beta}(\beta, \boldsymbol{\gamma}(\cdot), \zeta(\cdot, \beta, \boldsymbol{\gamma}); W),$$

$$S'_{\gamma_j,n}(\beta, \boldsymbol{\gamma}(\cdot), \zeta(\cdot, \beta, \boldsymbol{\gamma}))[v] = \mathbb{P}_n m'_{\gamma_j}(\beta, \boldsymbol{\gamma}(\cdot), \zeta(\cdot, \beta, \boldsymbol{\gamma}); W)[v],$$

$$S'_{\zeta,n}(\beta, \boldsymbol{\gamma}(\cdot), \zeta(\cdot, \beta, \boldsymbol{\gamma}))[h] = \mathbb{P}_n m'_{\zeta}(\beta, \boldsymbol{\gamma}(\cdot), \zeta(\cdot, \beta, \boldsymbol{\gamma}); W)[h],$$

$$S''_{\beta\beta}(\beta, \boldsymbol{\gamma}(\cdot), \zeta(\cdot, \beta, \boldsymbol{\gamma})) = Pm''_{\beta\beta}(\beta, \boldsymbol{\gamma}(\cdot), \zeta(\cdot, \beta, \boldsymbol{\gamma}); W),$$

$$S''_{\gamma_j\gamma_\ell}(\beta, \boldsymbol{\gamma}(\cdot), \zeta(\cdot, \beta, \boldsymbol{\gamma}))[v_1, v_2] = Pm''_{\gamma_j\gamma_\ell}(\beta, \boldsymbol{\gamma}(\cdot), \zeta(\cdot, \beta, \boldsymbol{\gamma}); W)[v_1, v_2],$$

$$S''_{\zeta\zeta}(\beta, \boldsymbol{\gamma}(\cdot), \zeta(\cdot, \beta, \boldsymbol{\gamma}))[h_1, h_2] = Pm''_{\zeta\zeta}(\beta, \boldsymbol{\gamma}(\cdot), \zeta(\cdot, \beta, \boldsymbol{\gamma}); W)[h_1, h_2],$$

$$S''_{\gamma_j\beta}(\beta, \boldsymbol{\gamma}(\cdot), \zeta(\cdot, \beta, \boldsymbol{\gamma}))[v] = S''_{\beta\gamma_j}(\beta, \boldsymbol{\gamma}(\cdot), \zeta(\cdot, \beta, \boldsymbol{\gamma}))[v] = Pm''_{\gamma_j\beta}(\beta, \boldsymbol{\gamma}(\cdot), \zeta(\cdot, \beta, \boldsymbol{\gamma}); W)[v],$$

$$S''_{\zeta\beta}(\beta, \boldsymbol{\gamma}(\cdot), \zeta(\cdot, \beta, \boldsymbol{\gamma}))[h] = S''_{\beta\zeta}(\beta, \boldsymbol{\gamma}(\cdot), \zeta(\cdot, \beta, \boldsymbol{\gamma}))[h] = Pm''_{\zeta\beta}(\beta, \boldsymbol{\gamma}(\cdot), \zeta(\cdot, \beta, \boldsymbol{\gamma}); W)[h],$$

$$S''_{\zeta\gamma_j}(\beta, \boldsymbol{\gamma}(\cdot), \zeta(\cdot, \beta, \boldsymbol{\gamma}))[h, v] = Pm''_{\zeta\gamma_j}(\beta, \boldsymbol{\gamma}(\cdot), \zeta(\cdot, \beta, \boldsymbol{\gamma}); W)[h, v],$$

$$S''_{\gamma_j\zeta}(\beta, \boldsymbol{\gamma}(\cdot), \zeta(\cdot, \beta, \boldsymbol{\gamma}))[v, h] = Pm''_{\gamma_j\zeta}(\beta, \boldsymbol{\gamma}(\cdot), \zeta(\cdot, \beta, \boldsymbol{\gamma}); W)[v, h].$$

Furthermore, for $\mathbf{h} = (h_1, \cdots, h_{d_1})^T \in \mathbb{H}^{d_1}$ and $\mathbf{v} = (v_1, \cdots, v_{d_1})^T \in \mathbb{V}^{d_1}$, denote that

$$\begin{aligned}
m'_{\gamma_j}(\beta, \boldsymbol{\gamma}(\cdot),&\zeta(\cdot, \beta, \boldsymbol{\gamma}); W)[\mathbf{v}] = \\
&(m'_{\gamma_j}(\beta, \boldsymbol{\gamma}(\cdot), \zeta(\cdot, \beta, \boldsymbol{\gamma}); W)[v_1], \cdots, m'_{\gamma_j}(\beta, \boldsymbol{\gamma}(\cdot), \zeta(\cdot, \beta, \boldsymbol{\gamma}); W)[v_{d_1}])^T, \\
m'_{\zeta}(\beta, \boldsymbol{\gamma}(\cdot),&\zeta(\cdot, \beta, \boldsymbol{\gamma}); W)[\mathbf{h}] = \\
&(m'_{\zeta}(\beta, \boldsymbol{\gamma}(\cdot), \zeta(\cdot, \beta, \boldsymbol{\gamma}); W)[h_1], \cdots, m'_{\zeta}(\beta, \boldsymbol{\gamma}(\cdot), \zeta(\cdot, \beta, \boldsymbol{\gamma}); W)[h_{d_1}])^T, \\
m''_{\gamma_j\gamma_\ell}(\beta, \boldsymbol{\gamma}(\cdot),&\zeta(\cdot, \beta, \boldsymbol{\gamma}); W)[\mathbf{v}, v] = \\
&(m''_{\gamma_j\gamma_\ell}(\beta, \boldsymbol{\gamma}(\cdot), \zeta(\cdot, \beta, \boldsymbol{\gamma}); W)[v_1, v], \cdots, m''_{\gamma_j\gamma_\ell}(\beta, \boldsymbol{\gamma}(\cdot), \zeta(\cdot, \beta, \boldsymbol{\gamma}); W)[v_{d_1}, v])^T, \\
m''_{\zeta\zeta}(\beta, \boldsymbol{\gamma}(\cdot),&\zeta(\cdot, \beta, \boldsymbol{\gamma}); W)[\mathbf{h}, h] = \\
&(m''_{\zeta\zeta}(\beta, \boldsymbol{\gamma}(\cdot), \zeta(\cdot, \beta, \boldsymbol{\gamma}); W)[h_1, h], \cdots, m''_{\zeta\zeta}(\beta, \boldsymbol{\gamma}(\cdot), \zeta(\cdot, \beta, \boldsymbol{\gamma}); W)[h_{d_1}, h])^T, \\
m''_{\gamma_j\beta}(\beta, \boldsymbol{\gamma}(\cdot),&\zeta(\cdot, \beta, \boldsymbol{\gamma}); W)[\mathbf{v}] = \\
&(m''_{\gamma_j\beta}(\beta, \boldsymbol{\gamma}(\cdot), \zeta(\cdot, \beta, \boldsymbol{\gamma}); W)[v_1], \cdots, m''_{\gamma_j\beta}(\beta, \boldsymbol{\gamma}(\cdot), \zeta(\cdot, \beta, \boldsymbol{\gamma}); W)[v_{d_1}])^T, \\
m''_{\zeta\beta}(\beta, \boldsymbol{\gamma}(\cdot),&\zeta(\cdot, \beta, \boldsymbol{\gamma}); W)[\mathbf{h}] = \\
&(m''_{\zeta\beta}(\beta, \boldsymbol{\gamma}(\cdot), \zeta(\cdot, \beta, \boldsymbol{\gamma}); W)[h_1], \cdots, m''_{\zeta\beta}(\beta, \boldsymbol{\gamma}(\cdot), \zeta(\cdot, \beta, \boldsymbol{\gamma}); W)[h_{d_1}])^T,
\end{aligned}$$

$$m''_{\zeta\gamma_j}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[\mathbf{h},v] = \\ (m''_{\zeta\gamma_j}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[h_1,v],\cdots,m''_{\zeta\gamma_j}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[h_{d_1},v])^T,$$

$$m''_{\gamma_j\zeta}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[\mathbf{v},h] = \\ (m''_{\gamma_j\zeta}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[v_1,h],\cdots,m''_{\gamma_j\zeta}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[v_{d_1},h])^T.$$

We define correspondingly

$$S'_{\gamma_j}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma}))[\mathbf{v}] = Pm'_{\gamma_j}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[\mathbf{v}],$$

$$S'_{\zeta}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma}))[\mathbf{h}] = Pm'_{\zeta}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[\mathbf{h}],$$

$$S'_{\gamma_j,n}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma}))[\mathbf{v}] = \mathbb{P}_n m'_{\gamma_j}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[\mathbf{v}],$$

$$S'_{\zeta,n}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma}))[\mathbf{h}] = \mathbb{P}_n m'_{\zeta}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[\mathbf{h}],$$

$$S''_{\gamma_j\gamma_\ell}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma}))[\mathbf{v},v] = Pm''_{\gamma_j\gamma_\ell}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[\mathbf{v},v],$$

$$S''_{\zeta\zeta}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma}))[\mathbf{h},h] = Pm''_{\zeta\zeta}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[\mathbf{h},h],$$

$$S''_{\gamma_j\beta}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma}))[\mathbf{v}] = Pm''_{\gamma_j\beta}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[\mathbf{v}],$$

$$S''_{\zeta\beta}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma}))[\mathbf{h}] = Pm''_{\zeta\beta}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[\mathbf{h}],$$

$$S''_{\zeta\gamma_j}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma}))[\mathbf{h},v] = Pm''_{\zeta\gamma_j}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[\mathbf{h},v],$$

$$S''_{\gamma_j\zeta}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma}))[\mathbf{v},h] = Pm''_{\gamma_j\zeta}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma});W)[\mathbf{v},h].$$

A.2.2 The general sieve M-theorem

Recall that the sieve M-estimator $\hat{\theta}_n = (\hat{\beta}_n, \hat{\boldsymbol{\gamma}}_n, \hat{\zeta}_n(\cdot,\hat{\beta}_n,\hat{\boldsymbol{\gamma}}_n)) \in \Theta_n$ maximizes the objective function over the sieve parameter space Θ_n. Next, we establish the asymptotic normality of the sieve estimator $\hat{\beta}_n$. The key difference between the proposed new sieve M-theorem in this paper and Theorem 2.1 in Ding and Nan (2011) is that the nuisance parameter $\zeta(\cdot,\beta,\boldsymbol{\gamma})$ can be a function of not only Euclidean parameter β but also other nuisance parameters $\boldsymbol{\gamma}(\cdot)$.

To establish the asymptotic normality, we assume the following assumptions.

(A1) (Rate of convergence) For an estimator $\hat{\theta}_n = (\hat{\beta}_n, \hat{\boldsymbol{\gamma}}_n(\cdot), \hat{\zeta}_n(\cdot,\hat{\beta}_n,\hat{\boldsymbol{\gamma}}_n)) \in \Theta_n$ and the true parameter $\theta_0 = (\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0)) \in \Theta$, $d(\hat{\theta}_n, \theta_0) = O_p(n^{-\xi})$ for some positive ξ.

(A2) $S'_\beta(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0)) = 0$, $S'_{\gamma_j}(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0))[v] = 0$ for all $v \in \Gamma^{p_1}$ and $1 \leq j \leq d_2$, and $S'_\zeta(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0))[h] = 0$ for all $h \in \mathbb{H}$.

(A3) (Positive information) There exists $\mathbf{v}_j^* = (v_{j1}^*, \cdots, v_{jd_1}^*)^T \in \mathbb{V}^{d_1}$, $1 \leq j \leq d_2$, and $\mathbf{h}^* = (h_1^*, \cdots, h_{d_1}^*)^T \in \mathbb{H}^{d_1}$ such that for any $v \in \mathbb{V}$ and $h \in \mathbb{H}$, $1 \leq \ell \leq d_2$

$$\begin{aligned} S''_{\beta\gamma_\ell}(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0))[v] &= \sum_{j=1}^{d_2} S''_{\gamma_j\gamma_\ell}(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0))[\mathbf{v}_j^*, v] \\ &\quad + S''_{\zeta\gamma_\ell}(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0))[\mathbf{h}^*, v], \\ S''_{\beta\zeta}(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0))[h] &= \sum_{j=1}^{d_2} S''_{\gamma_j\zeta}(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0))[\mathbf{v}_j^*, h] \\ &\quad + S''_{\zeta\zeta}(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0))[\mathbf{h}^*, h]. \end{aligned}$$

Furthermore, the matrix

$$\begin{aligned} A &= -S''_{\beta\beta}(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0)) + \sum_{j=1}^{d_2} S''_{\gamma_j\beta}(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0))[\mathbf{v}_j^*] \\ &\quad + S''_{\zeta\beta}(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0))[\mathbf{h}^*] \\ &= -P\{m''_{\beta\beta}(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0); W) + \sum_{j=1}^{d_2} m''_{\gamma_j\beta}(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0); W)[\mathbf{v}_j^*] \\ &\qquad + m''_{\zeta\beta}(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0); W)[\mathbf{h}^*]\} \end{aligned}$$

is nonsingular.

(A4) The estimator $\hat{\theta}_n = (\hat{\beta}_n, \hat{\boldsymbol{\gamma}}_n(\cdot), \hat{\zeta}_n(\cdot, \hat{\beta}_n, \hat{\boldsymbol{\gamma}}_n))$ satisfies

$$S'_{\beta,n}(\hat{\beta}_n, \hat{\boldsymbol{\gamma}}_n(\cdot), \hat{\zeta}_n(\cdot, \hat{\beta}_n, \hat{\boldsymbol{\gamma}}_n)) = o_p(n^{-1/2}),$$

$$S'_{\gamma_j,n}(\hat{\beta}_n, \hat{\boldsymbol{\gamma}}_n(\cdot), \hat{\zeta}_n(\cdot, \hat{\beta}_n, \hat{\boldsymbol{\gamma}}_n))[\mathbf{v}_j^*] = o_p(n^{-1/2}),$$

for $1 \leq j \leq d_2$, and

$$S'_{\zeta,n}(\hat{\beta}_n, \hat{\boldsymbol{\gamma}}_n(\cdot), \hat{\zeta}_n(\cdot, \hat{\beta}_n, \hat{\boldsymbol{\gamma}}_n))[\mathbf{h}^*] = o_p(n^{-1/2}).$$

(A5) (Stochastic equicontinuity) For some positive C,

$$\sup_{d(\theta,\theta_0)\le Cn^{-\xi},\theta\in\Theta_n} \|\sqrt{n}(S'_{\beta,n}-S'_{\beta})(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma})) - \sqrt{n}(S'_{\beta,n}-S'_{\beta})(\beta_0,\boldsymbol{\gamma}_0(\cdot),\zeta_0(\cdot,\beta_0,\boldsymbol{\gamma}_0))\| = o_p(1),$$

$$\sup_{d(\theta,\theta_0)\le Cn^{-\xi},\theta\in\Theta_n} |\sqrt{n}(S'_{\gamma_j,n}-S'_{\gamma_j})(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma}))[\mathbf{v}_j^*] - \sqrt{n}(S'_{\gamma_j,n}-S'_{\gamma_j})(\beta_0,\boldsymbol{\gamma}_0(\cdot),\zeta_0(\cdot,\beta_0,\boldsymbol{\gamma}_0))[\mathbf{v}_j^*]| = o_p(1),$$

for $1\le j\le d_2$, and

$$\sup_{d(\theta,\theta_0)\le Cn^{-\xi},\theta\in\Theta_n} |\sqrt{n}(S'_{\zeta,n}-S'_{\zeta})(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma}))[\mathbf{h}^*(\cdot,\beta,\boldsymbol{\gamma})] - \sqrt{n}(S'_{\zeta,n}-S'_{\zeta})(\beta_0,\boldsymbol{\gamma}_0(\cdot),\zeta_0(\cdot,\beta_0,\boldsymbol{\gamma}_0))[\mathbf{h}^*(\cdot,\beta_0,\boldsymbol{\gamma}_0)]| = o_p(1).$$

(A6) (Smoothness of the model) For some $\alpha > 1$ with $\alpha\xi > \frac{1}{2}$, and for $\theta\in\Theta_n$ satisfying $d(\theta,\theta_0)\le Cn^{-\xi}$,

$$\begin{aligned}
&\|S'_{\beta}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma})) - S'_{\beta}(\beta_0,\boldsymbol{\gamma}_0(\cdot),\zeta_0(\cdot,\beta_0,\boldsymbol{\gamma}_0))\\
&\quad - S''_{\beta\beta}(\beta_0,\boldsymbol{\gamma}_0(\cdot),\zeta_0(\cdot,\beta_0,\boldsymbol{\gamma}_0))(\beta-\beta_0)\\
&\quad - \sum_{j=1}^{d_2} S''_{\beta\gamma_j}(\beta_0,\boldsymbol{\gamma}_0(\cdot),\zeta_0(\cdot,\beta_0,\boldsymbol{\gamma}_0))[e_j(\boldsymbol{\gamma}-\boldsymbol{\gamma}_0)^T]\\
&\quad - S''_{\beta\zeta}(\beta_0,\boldsymbol{\gamma}_0(\cdot),\zeta_0(\cdot,\beta_0,\boldsymbol{\gamma}_0))[\zeta(\cdot,\beta,\boldsymbol{\gamma})-\zeta_0(\cdot,\beta_0,\boldsymbol{\gamma}_0)]\|\\
&\qquad = O(d^{\alpha}(\theta,\theta_0)),
\end{aligned}$$

$$\begin{aligned}
&|S'_{\gamma_j}(\beta,\boldsymbol{\gamma}(\cdot),\zeta(\cdot,\beta,\boldsymbol{\gamma}))[\mathbf{v}_j^*] - S'_{\gamma_j}(\beta_0,\boldsymbol{\gamma}_0(\cdot),\zeta_0(\cdot,\beta_0,\boldsymbol{\gamma}_0))[\mathbf{v}_j^*]\\
&\quad - S''_{\gamma_j\beta}(\beta_0,\boldsymbol{\gamma}_0(\cdot),\zeta_0(\cdot,\beta_0,\boldsymbol{\gamma}_0))[\mathbf{v}_j^*](\beta-\beta_0)\\
&\quad - \sum_{\ell=1}^{d_2} S''_{\gamma_j\gamma_\ell}(\beta_0,\boldsymbol{\gamma}_0(\cdot),\zeta_0(\cdot,\beta_0,\boldsymbol{\gamma}_0))[\mathbf{v}_j^*, e_\ell(\boldsymbol{\gamma}-\boldsymbol{\gamma}_0)^T]\\
&\quad - S''_{\gamma_j\zeta}(\beta_0,\boldsymbol{\gamma}_0(\cdot),\zeta_0(\cdot,\beta_0,\boldsymbol{\gamma}_0))[\mathbf{v}_j^*, \zeta(\cdot,\beta,\boldsymbol{\gamma})-\zeta_0(\cdot,\beta_0,\boldsymbol{\gamma}_0)]|\\
&\qquad = O(d^{\alpha}(\theta,\theta_0)),\quad \text{for } 1\le j\le d_2,
\end{aligned}$$

and

$$
\begin{aligned}
|S'_{\zeta}(\beta, \boldsymbol{\gamma}(\cdot), \zeta(\cdot, \beta, \boldsymbol{\gamma}))[\mathbf{h}^*(\cdot, \beta, \boldsymbol{\gamma})] &- S'_{\zeta}(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0))[\mathbf{h}^*(\cdot, \beta_0, \boldsymbol{\gamma}_0)] \\
&- S''_{\zeta\beta}(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0))[\mathbf{h}^*(\cdot, \beta_0, \boldsymbol{\gamma}_0)](\beta - \beta_0) \\
&- \sum_{j=1}^{d_2} S''_{\zeta\gamma_j}(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0))[\mathbf{h}^*(\cdot, \beta_0, \boldsymbol{\gamma}_0), e_j(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)^T] \\
&- S''_{\zeta\zeta}(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0))[\mathbf{h}^*(\cdot, \beta_0, \boldsymbol{\gamma}_0), \zeta(\cdot, \beta, \boldsymbol{\gamma}) - \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0)]| \\
&\quad = O(d^{\alpha}(\theta, \theta_0)).
\end{aligned}
$$

The convergence rate in (A1) is a prerequisite for the asymptotic normality. Assumption (A2) is a common regularity assumption when m is the likelihood function, and it usually holds for the score functions. The direction $\mathbf{v}_j^*$ and $\mathbf{h}^*$ in (A3) are the least favorable directions for maximum likelihood estimation, which may be found through solving the equations in (A3). Assumptions (A4) and (A5) can be obtained by the maximal inequality in Lemma 3.4.2 of (Billingsley, 2008, page 324) and the Markov's inequality. Assumption (A6) can be usually verified by the Taylor expansion. We repeat Theorem 2.3.3 below for readers' convenience, which is a general sieve M-theorem for bundled parameters where the nuisance parameter $\zeta(\cdot, \beta, \boldsymbol{\gamma})$ is a function of the Euclidean parameter β and other nuisance parameters $\boldsymbol{\gamma}(\cdot)$.

Theorem. *Suppose that assumptions (A1)-(A6) hold, then*

$$
\begin{aligned}
\sqrt{n}(\hat{\beta}_n - \beta_0) &= A^{-1}\sqrt{n}\mathbb{P}_n \boldsymbol{m}^*(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0); W) + o_p(1) \\
&\to_d N(0, A^{-1}B(A^{-1})^T),
\end{aligned}
$$

where

$$
\begin{aligned}
\boldsymbol{m}^*(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0); W) = m'_{\beta}&(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0); W) \\
&- \sum_{j=1}^{d_2} m'_{\gamma_j}(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0); W)[\boldsymbol{v}_j^*] \\
&- m'_{\zeta}(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0); W)[\boldsymbol{h}^*(\cdot, \beta_0, \boldsymbol{\gamma}_0)],
\end{aligned}
$$

$$
B = P\{\boldsymbol{m}^*(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0); W)\boldsymbol{m}^*(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0); W)^T\},
$$

and A is given in the assumption (A3).

Proof of Theorem 2.3.3. We prove the theorem by following the proof of Theorem 6.1 in Wellner and Zhang (2007) and Theorem 2.1 in Ding and Nan (2011). Assumptions (A1) and (A5) lead to

$$\sqrt{n}(S'_{\beta,n} - S'_{\beta})(\hat{\beta}_n, \hat{\boldsymbol{\gamma}}_n(\cdot), \hat{\zeta}_n(\cdot, \hat{\beta}_n, \hat{\boldsymbol{\gamma}}_n)) - \sqrt{n}(S'_{\beta,n} - S'_{\beta})(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0)) = o_p(1).$$

Note that $S'_{\beta}(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0)) = 0$ by (A2), $S'_{\beta,n}(\hat{\beta}_n, \hat{\boldsymbol{\gamma}}_n(\cdot), \hat{\zeta}_n(\cdot, \hat{\beta}_n, \hat{\boldsymbol{\gamma}}_n)) = o_p(n^{-1/2})$ by (A4), we have

$$\sqrt{n}S'_{\beta}(\hat{\beta}_n, \hat{\boldsymbol{\gamma}}_n(\cdot), \hat{\zeta}_n(\cdot, \hat{\beta}_n, \hat{\boldsymbol{\gamma}}_n)) + \sqrt{n}S'_{\beta,n}(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0)) = o_p(1). \tag{A.7}$$

After combining the equation (A.7) and the equations in assumptions (A2) and (A6), we have

$$\begin{aligned}
&S'_{\beta,n}(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0)) + S''_{\beta\beta}(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0))(\hat{\beta}_n - \beta_0)\\
&\quad + \sum_{j=1}^{d_2} S''_{\beta\gamma_j}(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0))[e_j(\hat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_0)^T]\\
&\quad + S''_{\beta\zeta}(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0))[\hat{\zeta}_n(\cdot, \hat{\beta}_n, \hat{\boldsymbol{\gamma}}_n) - \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0)]\\
&= O(d^{\alpha}(\hat{\theta}_n, \theta_0)) + o_p(n^{-1/2}) = o_p(n^{-1/2}).
\end{aligned}$$

The last equation holds because for $\alpha > 1$ with $\alpha\xi > \frac{1}{2}$, assumption (A1) implies that

$$O(d^{\alpha}(\hat{\theta}_n, \theta_0)) = O_p(n^{-\alpha\xi}) = o_p(n^{-1/2}).$$

Similarly, we have for $1 \leq j \leq d_2$

$$\begin{aligned}
&S'_{\gamma_j,n}(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0))[\mathbf{v}_j^*] + S''_{\gamma_j\beta}(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0))[\mathbf{v}_j^*](\hat{\beta}_n - \beta_0)\\
&\quad + \sum_{\ell=1}^{d_2} S''_{\gamma_j\gamma_\ell}(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0))[\mathbf{v}_j^*, e_\ell(\hat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_0)^T]\\
&\quad + S''_{\gamma_j\zeta}(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0))[\mathbf{v}_j^*, \hat{\zeta}_n(\cdot, \hat{\beta}_n, \hat{\boldsymbol{\gamma}}_n) - \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0)]\\
&= o_p(n^{-1/2})
\end{aligned}$$

and

$$
\begin{aligned}
&S'_{\zeta,n}(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0))[\mathbf{h}^*(\cdot, \beta_0, \boldsymbol{\gamma}_0)] + S''_{\zeta\beta}(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0))[\mathbf{h}^*(\cdot, \beta_0, \boldsymbol{\gamma}_0)](\hat{\beta}_n - \beta_0) \\
&\quad + \sum_{j=1}^{d_2} S''_{\zeta\gamma_j}(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0))[\mathbf{h}^*(\cdot, \beta_0, \boldsymbol{\gamma}_0), e_j(\hat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_0)^T] \\
&\quad + S''_{\zeta\zeta}(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0))[\mathbf{h}^*(\cdot, \beta_0, \boldsymbol{\gamma}_0), \hat{\zeta}_n(\cdot, \hat{\beta}_n, \hat{\gamma}_n) - \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0)] \\
&= o_p(n^{-1/2}).
\end{aligned}
$$

Combining these equations with assumption (A3) leads to

$$
\begin{aligned}
&\{S''_{\beta\beta}(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0)) - \sum_{j=1}^{d_2} S''_{\gamma_j\beta}(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0))[\mathbf{v}^*_j] \\
&\qquad - S''_{\zeta\beta}(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0))[\mathbf{h}^*(\cdot, \beta_0, \boldsymbol{\gamma}_0)]\}(\hat{\beta}_n - \beta_0) \\
&= -\{S'_{\beta,n}(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0)) - \sum_{j=1}^{d_2} S'_{\gamma_j,n}(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0))[\mathbf{v}^*_j] \\
&\qquad - S'_{\zeta,n}(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0))[\mathbf{h}^*(\cdot, \beta_0, \boldsymbol{\gamma}_0)]\} \\
&\quad + o_p(n^{-1/2}),
\end{aligned}
$$

and equivalently,

$$
-A(\hat{\beta}_n - \beta_0) = -\mathbb{P}_n\mathbf{m}^*(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0); W) + o_p(n^{-1/2}).
$$

Then under assumptions (A4) and (A5),

$$
\begin{aligned}
\sqrt{n}(\hat{\beta}_n - \beta_0) &= A^{-1}\sqrt{n}\mathbb{P}_n\mathbf{m}^*(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0); W) + o_p(1) \\
&\to_d N(0, A^{-1}B(A^{-1})^T).
\end{aligned}
$$

□

A.3 Proof of Theorems 2.3.1 and 2.3.2

Without loss of generality, we prove Theorems 2.3.1 and 2.3.2 in the case that $X_{(1)}$ is not included in (2.12). The results in this section still hold if $X_{(1)}$ is included due to the boundedness of $X_{(1)}$. For notational simplicity, we further replace $X_{(-1)}$ by X in (2.12), which then becomes equivalent to the ODE in (2.11).

We first introduce some common notations that will be used in the proof hereafter. For any fixed $\gamma(\cdot) \in \Gamma^{p_1}$, let $\{\gamma_\eta(\cdot) : \eta$ in a neighborhood of $0 \in \mathbf{R}\}$ be a smooth curve in Γ^{p_1} running through $\gamma(\cdot)$ at $\eta = 0$, that is $\gamma_\eta(\cdot)|_{\eta=0} = \gamma(\cdot)$. Similarly, for any fixed $g(\cdot) \in \mathcal{G}^{p_2}$, let $\{g_\eta(\cdot) : \eta$ in a neighborhood of $0 \in \mathbf{R}\}$ be a smooth curve in $\mathcal{G}^{p_2}$ running through $g(\cdot)$ at $\eta = 0$, that is $g_\eta(\cdot)|_{\eta=0} = g(\cdot)$. Denote

$$\mathbb{V} = \{v : v(\cdot) = \frac{\partial \gamma_\eta(\cdot)}{\partial \eta}|_{\eta=0}, \gamma_\eta \in \Gamma^{p_1}\}$$

and

$$\mathbb{W} = \{w : w(\cdot) = \frac{\partial g_\eta(\cdot)}{\partial \eta}|_{\eta=0}, g_\eta \in \mathcal{G}^{p_2}\}.$$

Recall that $\Lambda_0(t,x) = \Lambda(t,x,\beta_0,\gamma_0,g_0)$ and $R(t) = \int_0^t \exp(\gamma_0(s))ds$. Let $\tilde{\Lambda}_0(t)$ denote the solution of $\tilde{\Lambda}_0'(t) = \exp\Big(g_0(\tilde{\Lambda}_0)\Big)$ with $\tilde{\Lambda}_0(0) = 0$. It is straightforward to show that $\tilde{\Lambda}_0(\cdot)$ is the cumulative hazard function of $R(T)e^{X^T\beta_0}$ and $\Lambda_0(t,X) = \tilde{\Lambda}_0(R(t)e^{X^T\beta_0})$. We use symbol $\gtrsim$ to denote that the left side is bounded below by a constant times the right side. We also use symbol $\lesssim$ to denote that the left side is bounded above by a constant times the right side. If without further explanation, by default, the L_2 norm of a function $f(\cdot)$ of t and x is given by

$$\|f(\cdot)\|_2 = \left[\int\limits_{\mathcal{X}} \int\limits_0^\tau (f(t,x))^2 d\Lambda_0(t,x) dF_X(x)\right]^{1/2},$$

and the supreme norm is given by $\|f(\cdot)\|_\infty = \sup_{t\in[0,\tau],x\in\mathcal{X}} |f(t,x)|$. For any $g \in \mathcal{G}^{p_2}$, the L_2 norm is given by $\|g\|_2 = (\int_0^\mu (g(t))^2 dt)^{1/2}$ and the supreme norm is given by $\|g\|_\infty = \sup_{t\in[0,\mu]} |g(t)|$.

The rest of this section is structured as follows. Subsection A.3.1 introduces several lemmas which will be used to prove Theorem 2.3.1 and 2.3.2. Subsections A.3.2 and A.3.3 provide the proof of Theorem 2.3.1 by checking the conditions C1-3 in Shen and Wong (1994, Theorem 1) and the proof of Theorem 2.3.2 by verifying assumptions (A1)-(A6) of the proposed general M-theorem, respectively. Furthermore, we derive in subsection A.3.4 the equivalent but more feasible equations for finding the least favorable directions required in condition (C7) and provide explicit constructions for the Cox model and the linear transformation model with a known transformation as illustration. Subsequently, we simplify the non-regularity assumption in Condition (C8) in subsection A.3.5.

A.3.1 Lemmas

Lemma A.3.1. *(Existence and uniqueness theorem.) For any $x \in \mathcal{X}, \beta \in \mathcal{B}, \gamma \in \Gamma^{p_1}, g \in \mathcal{G}^{p_2}$ under conditions (C1)-(C4), the initial value problem (2.11) has exactly one bounded and continuous solution $\Lambda(t, x, \beta, \gamma, g)$ on $[0, \tau]$. And its first and second derivatives with respect to $\beta \in \mathcal{B}, \gamma \in \Gamma^{p_1}$ and the first derivative with respect to $g \in \mathcal{G}^{p_2}$ are also bounded and continuous on $[0, \tau]$.*

Proof of Lemma A.3.1. Let $f(t, \Lambda) = \exp\big(x^T\beta + \gamma(t) + g(\Lambda)\big)$, then by the mean value theorem

$$|f(t,\Lambda) - f(t,\tilde{\Lambda})| \leq \exp\big(x^T\beta + \gamma(t) + g(c)\big)|g'(c)| \cdot |\Lambda - \tilde{\Lambda}| \leq L|\Lambda - \tilde{\Lambda}|$$

holds for any (t, Λ) and $(t, \tilde{\Lambda})$ in $[0, \tau] \times [0, \mu]$, where $c \in [\Lambda, \tilde{\Lambda}]$ and $L < \infty$ under conditions (C1)-(C4). This implies that $f(t, \Lambda)$ satisfies the Lipschitz condition with respect to Λ in $[0, \tau] \times [0, \mu]$. By Theorem 10.VI in Walter (1998, page 108), there is exactly one solution to the initial value problem (2.11). The solution $\Lambda(t, x, \beta, \gamma, g)$ is bounded, continuous, and satisfies

$$\Lambda(t, x, \beta, \gamma, g) = \int_0^t \exp\big(x^T\beta + \gamma(s) + g(\Lambda(s, x, \beta, \gamma, g))\big)\, ds. \tag{A.8}$$

In the following, we write $\Lambda(t) = \Lambda(t, x, \beta, \gamma, g)$ for simplicity. Similarly to the above derivation, for any $\beta \in \mathcal{B}, v \in \mathbb{V}, w \in \mathbb{W}$, we have unique, bounded, and continuous solutions of the following initial value problems:

$$\frac{d\Lambda'_\beta(t)}{dt} = \exp\big(x^T\beta + \gamma(t) + g(\Lambda(t))\big)\{x + g'(\Lambda(t))\Lambda'_\beta(t)\}, \quad \Lambda'_\beta(0) = 0, \tag{A.9}$$

$$\frac{d\Lambda'_\gamma(t)[v]}{dt} = \exp\big(x^T\beta + \gamma(t) + g(\Lambda(t))\big)\{v(t) + g'(\Lambda(t))\Lambda'_\gamma(t)[v]\}, \Lambda'_\gamma(0)[v] = 0, \tag{A.10}$$

$$\begin{aligned}\frac{d\Lambda_g'(t)[w]}{dt} &= \exp\bigl(x^T\beta+\gamma(t)+g(\Lambda(t))\bigr)\{w(\Lambda(t))\\ &\qquad + g'(\Lambda(t))\Lambda_g'(t)[w]\},\quad \Lambda_g'(0)[w]=0, \quad (A.11)\\ \frac{d\Lambda_{\beta\beta}''(t)}{dt} &= \exp\bigl(x^T\beta+\gamma(t)+g(\Lambda(t))\bigr)\{[x+g'(\Lambda(t))\Lambda_\beta'(t)][x+g'(\Lambda(t))\Lambda_\beta'(t)]^T\\ &\quad + g''(\Lambda(t))\Lambda_\beta'(t)\Lambda_\beta'(t)^T + g'(\Lambda(t))\Lambda_{\beta\beta}''(t)\},\quad \Lambda_{\beta\beta}''(0)=0, \qquad (A.12)\\ \frac{d\Lambda_{\gamma\gamma}''(t)[v_1,v_2]}{dt} &= \exp\bigl(x^T\beta+\gamma(t)+g(\Lambda(t))\bigr)\cdot\\ &\quad \{(v_1(t)+g'(\Lambda(t))\Lambda_\gamma'(t)[v_1])(v_2(t)+g'(\Lambda(t))\Lambda_\gamma'(t)[v_2])\\ &\quad + g''(\Lambda(t))\Lambda_\gamma'(t)[v_1]\Lambda_\gamma'(t)[v_2]\\ &\quad + g'(\Lambda(t))\Lambda_{\gamma\gamma}''(t)[v_1,v_2]\},\quad \Lambda_{\gamma\gamma}''(0)[v_1,v_2]=0, \qquad (A.13)\\ \frac{d\Lambda_{\beta\gamma}''(t)[v]}{dt} &= \exp\bigl(x^T\beta+\gamma(t)+g(\Lambda(t))\bigr)\cdot\\ &\quad \{(v(t)+g'(\Lambda(t))\Lambda_\gamma'(t)[v])(x+g'(\Lambda(t))\Lambda_\beta'(t))\\ &\quad + g''(\Lambda(t))\Lambda_\beta'(t)\Lambda_\gamma'(t)[v]\\ &\quad + g'(\Lambda(t))\Lambda_{\beta\gamma}''(t)[v]\},\quad \Lambda_{\beta\gamma}''(0)[v]=0, \qquad (A.14)\\ \frac{d\Lambda_{g\beta}''(t)[w]}{dt} &= \exp\bigl(x^T\beta+\gamma(t)+g(\Lambda(t))\bigr)\cdot\\ &\quad \{(w(\Lambda(t))+g'(\Lambda(t))\Lambda_g'(t)[w])(x+g'(\Lambda(t))\Lambda_\beta'(t))\\ &\quad + w'(\Lambda(t))\Lambda_\beta'(t) + g''(\Lambda(t))\Lambda_\beta'(t)\Lambda_g'(t)[w]\\ &\quad + g'(\Lambda(t))\Lambda_{g\beta}''(t)[w]\},\quad \Lambda_{g\beta}''(0)[w]=0, \qquad (A.15)\\ \frac{d\Lambda_{g\gamma}''(t)[w,v]}{dt} &= \exp\bigl(x^T\beta+\gamma(t)+g(\Lambda(t))\bigr)\cdot\\ &\quad \{(w(\Lambda(t))+g'(\Lambda(t))\Lambda_g'(t)[w])(v(t)+g'(\Lambda(t))\Lambda_\gamma'(t)[v])\\ &\quad + w'(\Lambda(t))\Lambda_\gamma'(t)[v] + g''(\Lambda(t))\Lambda_\gamma'(t)[v]\Lambda_g'(t)[w]\\ &\quad + g'(\Lambda(t))\Lambda_{g\gamma}''(t)[w,v]\},\quad \Lambda_{g\gamma}''(0)[w,v]=0. \qquad (A.16)\end{aligned}$$

Next we verify that the derivative of $\Lambda(t,x,\beta,\gamma,g)$ with respect to β follows the ODE (A.9). By plugging in Equation (A.8) and (A.9), it follows that

$$\limsup_{\delta\to 0}\frac{1}{|\delta|}|\Lambda(t,x,\beta+\delta,\gamma,g)-\Lambda(t)-\Lambda_\beta'(t)^T\delta|$$

$$
\begin{aligned}
&= \limsup_{\delta\to 0} \frac{1}{|\delta|} \Big| \int_0^t \exp\big(x^T(\beta+\delta) + \gamma(s) + g(\Lambda(s,x,\beta+\delta,\gamma,g))\big) \\
&\qquad - \exp\big(x^T\beta + \gamma(s) + g(\Lambda(s))\big) \\
&\qquad - \exp\big(x^T\beta + \gamma(s) + g(\Lambda(s))\big)\big(x^T\delta + g'(\Lambda(s))\Lambda'_\beta(s)^T\delta\big)\,ds\Big| \\
&\le \limsup_{\delta\to 0} \frac{1}{|\delta|} \int_0^t \Big| \exp\big(x^T(\beta+\delta) + \gamma(s) + g(\Lambda(s,x,\beta+\delta,\gamma,g))\big) \\
&\qquad - \exp\big(x^T\beta + \gamma(s) + g(\Lambda(s))\big) \\
&\qquad - \exp\big(x^T\beta + \gamma(s) + g(\Lambda(s))\big)\big(x^T\delta + g'(\Lambda(s))\Lambda'_\beta(s)^T\delta\big)\Big|\,ds \\
&\le \int_0^t \limsup_{\delta\to 0} \frac{1}{|\delta|} \Big| \exp\big(x^T(\beta+\delta) + \gamma(s) + g(\Lambda(s,x,\beta+\delta,\gamma,g))\big) \\
&\qquad - \exp\big(x^T\beta + \gamma(s) + g(\Lambda(s))\big) \\
&\qquad - \exp\big(x^T\beta + \gamma(s) + g(\Lambda(s))\big)\big(x^T\delta + g'(\Lambda(s))\Lambda'_\beta(s)^T\delta\big)\Big|\,ds \\
&= \int_0^t \exp\big(x^T\beta + \gamma(s) + g(\Lambda(s))\big) \\
&\qquad \cdot g'(\Lambda(s))\{\limsup_{\delta\to 0} \frac{1}{|\delta|}|\Lambda(s,x,\beta+\delta,\gamma,g) - \Lambda(s) - \Lambda'_\beta(s)^T\delta|\}\,ds,
\end{aligned}
$$

where the second inequality holds due to the reverse Fatou's lemma. Using the Gronwall's inequality, we have that

$$
\limsup_{\delta\to 0} \frac{1}{|\delta|}|\Lambda(t,x,\beta+\delta,\gamma,g) - \Lambda(t) - \Lambda'_\beta(t)^T\delta| \le 0,
$$

which implies that the solution $\Lambda'_\beta(t)$ of (A.9) is the derivative of $\Lambda(t,x,\beta,\gamma,g)$ with respect to β. The other first and second derivatives of of $\Lambda(t,x,\beta,\gamma,g)$ with respect to β,γ,g can be similarly derived, and we omit the details here. □

Lemma A.3.2. *Let $\psi(t,x,\beta,\gamma,g) = \log\lambda(t,x,\beta,\gamma,g) = x^T\beta+\gamma(t)+g(\Lambda(t,x,\beta,\gamma,g))$, and denote the first derivatives of $\psi(t,x,\beta,\gamma,g)$ with respect to γ and g at the true parameter (β_0,γ_0,g_0) by $\psi'_{0\gamma}(t,x)[v]$ and $\psi'_{0g}(t,x)[w]$, respectively. For any $\psi'_{0\gamma}(\cdot)[v] \in \mathcal{E}_\gamma = \{\psi'_{0\gamma}(\cdot)[v] : \psi'_{0\gamma}(t,x)[v], t \in [0,\tau], x \in \mathcal{X}, v \in \Gamma^{p_1}\}$, the L_2 norm of $\psi'_{0\gamma}(\cdot)[v]$ is* *defined as*

$$\|\psi'_{0\gamma}(\cdot)[v]\|_2 = \left[\int_{\mathcal{X}} \int_0^\tau (\psi'_{0\gamma}(t,x)[v])^2 d\Lambda_0(t,x) dF_X(x) \right]^{1/2}.$$

The L_2 norm of $\psi'_{0g}(\cdot)[w] \in \mathcal{E}_g = \{\psi'_{0g}(\cdot)[w] : \psi'_{0g}(t,x)[w], t \in [0,\tau], x \in \mathcal{X}, w \in \mathcal{G}^{p_2}\}$ is similarly defined. Under conditions (C2)-(C4), $\psi'_{0\gamma}[\cdot] : v \to \psi'_{0\gamma}(\cdot)[v]$ and $\psi'_{0g}[\cdot] : w \to \psi'_{0g}(\cdot)[w]$ are bounded linear operators (from Γ^{p_1} to $\mathcal{E}_\gamma$ and from $\mathcal{G}^{p_2}$ to $\mathcal{E}_g$). In particular, the operators $\psi'_{0\gamma}[\cdot]$ and $\psi'_{0g}[\cdot]$ are bounded from below, i.e.,

$$\|\psi'_{0\gamma}(\cdot)[v]\|_2 \gtrsim \|v\|_2, \text{ for any } v \in \Gamma^{p_1}, \tag{A.17}$$

and

$$\|\psi'_{0g}(\cdot)[w]\|_2 \gtrsim \|w\|_2, \text{ for any } w \in \mathcal{G}^{p_2}. \tag{A.18}$$

Proof of Lemma A.3.2. By solving initial value problems in (A.10)-(A.11), the first derivatives of $\psi(t,x,\beta,\gamma,g)$ with respect to γ and g at the true parameter (β_0,γ_0,g_0) are given by

$$\begin{aligned}\psi'_{0\gamma}(t,x)[v] &= g'_0(\Lambda_0(t,x))\Lambda'_{0\gamma}(t,x)[v] + v(t)\\ &= g'_0(\Lambda_0(t,x)) \exp(g_0(\Lambda_0(t,x)))e^{x^T\beta_0} \int_0^t \exp(\gamma_0(s))v(s)ds + v(t), \end{aligned}\tag{A.19}$$

$$\begin{aligned}\psi'_{0g}(t,x)[w] &= g'_0(\Lambda_0(t,x))\Lambda'_{0g}(t,x)[w] + w(\Lambda_0(t,x))\\ &= g'_0(\Lambda_0(t,x)) \exp(g_0(\Lambda_0(t,x))) \int_0^{\Lambda_0(t,x)} \exp(-g_0(s))w(s)ds + w(\Lambda_0(t,x)),\end{aligned}\tag{A.20}$$

We first verify that $\psi'_{0\gamma}[\cdot]$ is a bounded linear operator. Using $(a+b)^2 \le 2(a^2+b^2)$, the L_2 norm of $\psi'_{0\gamma}(\cdot)[v]$ is bounded by

$$\|\psi'_{0\gamma}(\cdot)[v]\|_2^2$$

$$\le 2 \int_{\mathcal{X}} \int_0^\tau \left(g'_0(\Lambda_0(t,x)) \exp(g_0(\Lambda_0(t,x)))e^{x^T\beta_0} \int_0^t \exp(\gamma_0(s))v(s)ds \right)^2 d\Lambda_0(t,x)dF_X(x)$$

$$+ 2\int_{\mathcal{X}}\int_0^\tau v(t)^2\,d\Lambda_0(t,x)\,dF_X(x)$$

$$= 2\int_{\mathcal{X}}\int_0^\tau (g_0'(\Lambda_0(t,x)))^2\exp(2g_0(\Lambda_0(t,x)))e^{2x^T\beta_0}\left(\int_0^t \exp(\gamma_0(s))v(s)\,ds\right)^2 d\Lambda_0(t,x)\,dF_X(x)$$

$$+ 2\int_{\mathcal{X}}\int_0^\tau v(t)^2\,d\Lambda_0(t,x)\,dF_X(x). \tag{A.21}$$

By the Cauchy-Schwarz inequality, we have for $t \in [0,\tau]$

$$\begin{aligned}\left(\int_0^t \exp(\gamma_0(s))v(s)\,ds\right)^2 &\le \int_0^t (v(s))^2\,ds\int_0^t \exp(2\gamma_0(s))\,ds\\ &\le \int_0^\tau (v(s))^2\,ds\int_0^\tau \exp(2\gamma_0(s))\,ds\\ &\le \|v\|_2^2\tau e^{2c_1},\end{aligned}$$

where $c_1 = \max_{s\in[0,\tau]}\gamma_0(s) < \infty$ under (C4). It follows that the first term in (A.21) is bounded above by

$$2\|v\|_2^2\tau e^{2c_1}\cdot\int_{\mathcal{X}}\int_0^\tau (g_0'(\Lambda_0(t,x)))^2\exp(2g_0(\Lambda_0(t,x)))e^{2x^T\beta_0}\,d\Lambda_0(t,x)\,dF_X(x) \lesssim \|v\|_2^2,$$

because the integral is finite under (C2)-(C4). The second term in (A.21) is also bounded by

$$\begin{aligned}&2\int_{\mathcal{X}}\int_0^\tau v(t)^2\,d\Lambda_0(t,x)\,dF_X(x)\\ =&2\int_{\mathcal{X}}\int_0^\tau \exp\big(x^T\beta_0+\gamma_0(t)+g_0(\Lambda_0(t,x))\big)v(t)^2\,dt\,dF_X(x)\\ \le&2\int_{\mathcal{X}}\int_0^\tau c_2 v(t)^2\,dt\,dF_X(x) = 2c_2\|v\|_2^2,\end{aligned}$$

where $c_2 = \max_{t\in[0,\tau],x\in\mathcal{X}}\exp\big(x^T\beta_0+\gamma_0(t)+g_0(\Lambda_0(t,x))\big) < \infty$ under (C2)-(C4). Therefore, $\|\psi_{0\gamma}'(\cdot)[v]\|_2 \lesssim \|v\|_2$ for any $v \in \Gamma^{p_1}$.

Similarly, we can show that $\psi'_{0g}[\cdot]$ is a bounded linear operator by

$$\|\psi'_{0g}(\cdot)[w]\|_2^2 = \int_{\mathcal{X}}\int_0^{\tau}(\psi'_{0g}(t,x)[w])^2 d\Lambda_0(t,x)\, dF_X(x)$$

$$= \int_{\mathcal{X}}\int_0^{\Lambda_0(\tau,x)}\left(g_0'(t)\exp(g_0(t))\int_0^t \exp(-g_0(s))w(s)\, ds + w(t)\right)^2 dt\, dF_X(x)$$

$$\lesssim \int_{\mathcal{X}}\int_0^{\Lambda_0(\tau,x)}\left(g_0'(t)\exp(g_0(t))\int_0^t \exp(-g_0(s))w(s)\, ds\right)^2 dt\, dF_X(x)$$

$$+ \int_{\mathcal{X}}\int_0^{\Lambda_0(\tau,x)}(w(t))^2\, dt\, dF_X(x)$$

$$\lesssim \int_{\mathcal{X}}\int_0^{\Lambda_0(\tau,x)}(g_0'(t))^2\exp(2g_0(t))\left(\int_0^t \exp(-2g_0(s))\, ds\right)\left(\int_0^t (w(s))^2 ds\right) dt\, dF_X(x)$$

$$+ \int_{\mathcal{X}}\int_0^{\Lambda_0(\tau,x)}(w(t))^2\, dt\, dF_X(x)$$

$$\lesssim \int_0^{\mu}(g_0'(t))^2\exp(2g_0(t))\left(\int_0^t \exp(-2g_0(s))\, ds\right)\left(\int_0^t (w(s))^2 ds\right) dt$$

$$+ \int_0^{\mu}(w(t))^2\, dt,$$

where the second last inequality holds by the Cauchy-Schwarz inequality and $\mu = \max_{x\in\mathcal{X}}\Lambda_0(\tau,x)$ given in condition (C3). The first term is further bounded by

$$\left(\int_0^{\mu}(w(t))^2 dt\right)\int_0^{\mu}(g_0'(t))^2\exp(2g_0(t))\left(\int_0^t \exp(-2g_0(s))\, ds\right) dt \lesssim \int_0^{\mu}(w(t))^2 dt = \|w\|_2^2,$$

since the second integral is finite under conditions (C2)-(C4). Thus, $\|\psi'_{0g}(\cdot)[w]\|_2 \lesssim \|w\|_2$ for any $w \in \mathcal{G}^{p_2}$.

Next, we show that linear operators $\psi'_{0\gamma}[\cdot]$ and $\psi'_{0g}[\cdot]$ are bijective functions. Suppose that $\psi'_{0\gamma}(\cdot)[v_1] = \psi'_{0\gamma}(\cdot)[v_2] \in \mathcal{E}_\gamma$ holds almost surely with respect to the measure $\rho(t,x) = \Lambda_0(t,x) \times F_X(x)$. Using the ODE in (A.10), we have

$$v_i(t) = \psi'_{0\gamma}(t,x)[v_i] - g_0'(\Lambda_0(t,x))\int_0^t \psi'_{0\gamma}(s,x)[v_i]\, d\Lambda_0(s,x), \text{ for } i = 1, 2,$$

and then $v_1 = v_2$ almost surely with respect to ρ, i.e., $\int_{\mathcal{X}} \int_0^\tau (v_1(t) - v_2(t))^2 d\rho(t,x) = 0$. It follows that

$$\begin{aligned}
&\int_{\mathcal{X}} \int_0^\tau (v_1(t) - v_2(t))^2 d\rho(t,x) \\
&= \int_{\mathcal{X}} \int_0^\tau \exp\big(x^T\beta_0 + \gamma_0(t) + g_0(\Lambda_0(t,x))\big)(v_1(t) - v_2(t))^2 dt dF_X(x) \\
&\geq \int_{\mathcal{X}} \int_0^\tau c_3(v_1(t) - v_2(t))^2 dt dF_X(x) = c_3\|v_1 - v_2\|_2^2,
\end{aligned}$$

where $c_3 = \min_{t\in[0,\tau],x\in\mathcal{X}} \exp\big(x^T\beta_0 + \gamma_0(t) + g_0(\Lambda_0(t,x))\big) < \infty$ under (C2)-(C4), which implies that $\psi'_{0\gamma}[\cdot]$ is a bijective function from Γ^{p_1} to $\mathcal{E}_\gamma$.

Similarly, suppose that $\psi'_{0g}(\cdot)[w_1] = \psi'_{0g}(\cdot)[w_2] \in \mathcal{E}_g$ holds almost surely with respect to the measure $\rho(t,x)$. Using the ODE in (A.11), we have

$$w_i(\Lambda_0(t,x)) = \psi'_{0g}(t,x)[w_i] - g'_0(\Lambda_0(t,x)) \int_0^t \psi'_{0g}(s,x)[w_i] d\Lambda_0(s,x), \text{ for } i = 1,2,$$

and then $w_1(\Lambda_0(t,x)) = w_2(\Lambda_0(t,x))$ almost surely with respect to ρ. It follows that

$$\begin{aligned}
0 &= \int_{\mathcal{X}} \int_0^\tau (w_1(\Lambda_0(t,x)) - w_2(\Lambda_0(t,x)))^2 d\rho(t,x) \\
&= \int_{\mathcal{X}} \int_0^{\Lambda_0(\tau,x)} (w_1(t) - w_2(t))^2 dt dF_X(x) \\
&\gtrsim \int_0^{\sup_{x\in\mathcal{X}} \Lambda_0(\tau,x)} (w_1(t) - w_2(t))^2 dt = \|w_1 - w_2\|_2^2,
\end{aligned}$$

where the last inequality holds under condition (C2). So $w_1 = w_2 \in \mathcal{G}^{p_2}$ and $\psi'_{0g}[\cdot]$ is a bijective function from $\mathcal{G}^{p_1}$ to $\mathcal{E}_g$.

By bounded inverse theorem, it follows that the bijective bounded linear operators $\psi'_{0\gamma}[\cdot]$ and $\psi'_{0g}[\cdot]$ have bounded inverse operator $(\psi'_{0\gamma})^{-1}[\cdot]$ and $(\psi'_{0g})^{-1}[\cdot]$. Then, there is a constant $0 < L < \infty$ such that

$$\|v\|_2 = \|(\psi'_{0\gamma})^{-1}\left[\psi'_{0\gamma}(\cdot)[v]\right]\|_2 \leq L\|\psi'_{0\gamma}(\cdot)[v]\|_2,$$

which implies that $\psi'_{0\gamma}[\cdot]$ is bounded from below since $\|\psi'_{0\gamma}(\cdot)[v]\|_2 \geq 1/L\|v\|_2$. Analogously, $\psi'_{0g}[\cdot]$ is also bounded from below, which can be obtained using the same argument as above. □

Lemma A.3.3. *Let $\zeta_\eta(\cdot,\beta,\gamma)$ be a smooth curve in $\mathcal{H}^{p_2}$ running through $\zeta(\cdot,\beta,\gamma)$ at $\eta = 0$, that is $\zeta_\eta(\cdot,\beta,\gamma)|_{\eta=0} = \zeta(\cdot,\beta,\gamma)$. For any score function $h(\cdot,\beta,\gamma)$ with $\zeta(\cdot,\beta,\gamma) = g(\Lambda(\cdot,\beta,\gamma,g))$ in*

$$\mathbb{H} = \left\{ h : h(\cdot,\beta,\gamma) = \frac{\partial \zeta_\eta(\cdot,\beta,\gamma)}{\partial \eta}|_{\eta=0}, \zeta_\eta \in \mathcal{H}^{p_2} \right\},$$

under conditions (C1)-(C4), there exists $w \in \mathbb{W}$ such that

$$h(\cdot,\beta,\gamma) = w(\Lambda(\cdot,\beta,\gamma,g)) + g'(\Lambda(\cdot,\beta,\gamma,g))\Lambda'_g(\cdot,\beta,\gamma,g)[w].$$

Proof of Lemma A.3.3. Since $\zeta_\eta(\cdot,\beta,\gamma)$ is a smooth curve in $\mathcal{H}^{p_2}$ running through $\zeta(\cdot,\beta,\gamma)$ at $\eta = 0$, we can rewrite it in the form of $\zeta_\eta(\cdot,\beta,\gamma) = g_\eta(\Lambda(\cdot,\beta,\gamma,g_\eta))$ where g_η is a smooth curve in $\mathcal{G}^{p_2}$ running through g at $\eta = 0$. For a small η, we have $g_\eta = g + \eta w + o(\eta)$ with $w = \frac{\partial g_\eta}{\partial \eta}|_{\eta=0} \in \mathbb{W}$. It follows that

$$\lim_{\eta \to 0} \frac{g_\eta(\Lambda(\cdot,\beta,\gamma,g_\eta)) - g(\Lambda(\cdot,\beta,\gamma,g_\eta))}{\eta} = w(\Lambda(\cdot,\beta,\gamma,g)).$$

Also, by the definition of functional derivatives, we have

$$\begin{aligned} g(\Lambda(\cdot,\beta,\gamma,g_\eta)) - g(\Lambda(\cdot,\beta,\gamma,g)) &= g(\Lambda(\cdot,\beta,\gamma,g+\eta w + o(\eta))) - g(\Lambda(\cdot,\beta,\gamma,g_\eta)) \\ &= g'(\Lambda(\cdot,\beta,\gamma,g))\Lambda'_g(\cdot,\beta,\gamma,g)[\eta w + o(\eta)] \\ &\quad + o(\|\eta w + o(\eta)\|) \\ &= \eta g'(\Lambda(\cdot,\beta,\gamma,g))\Lambda'_g(\cdot,\beta,\gamma,g)[w] + o(\eta), \end{aligned}$$

where the last equality holds because

$$\lim_{\eta \to 0} \frac{g'(\Lambda(\cdot,\beta,\gamma,g))\Lambda'_g(\cdot,\beta,\gamma,g)[o(\eta)]}{\eta} = g'(\Lambda(\cdot,\beta,\gamma,g))\Lambda'_g(\cdot,\beta,\gamma,g)\Big[\lim_{\eta\to 0}\frac{o(\eta)}{\eta}\Big] = 0.$$

Combining these two equations together, we have,

$$\begin{aligned} h(\cdot,\beta,\gamma) &= \lim_{\eta \to 0} \frac{g_\eta(\Lambda(\cdot,\beta,\gamma,g_\eta)) - g(\Lambda(\cdot,\beta,\gamma,g))}{\eta} \\ &= \lim_{\eta \to 0} \frac{g_\eta(\Lambda(\cdot,\beta,\gamma,g_\eta)) - g(\Lambda(\cdot,\beta,\gamma,g_\eta)) + g(\Lambda(\cdot,\beta,\gamma,g_\eta)) - g(\Lambda(\cdot,\beta,\gamma,g))}{\eta} \\ &= w(\Lambda(\cdot,\beta,\gamma,g)) + g'(\Lambda(\cdot,\beta,\gamma,g))\Lambda'_g(\cdot,\beta,\gamma,g)[w]. \end{aligned}$$

□

Lemma A.3.4. *Denote*

$$\begin{aligned} l(\beta,\gamma,\zeta(\cdot,\beta,\gamma);W) &= \Delta\{X^T\beta+\gamma(Y)+g(\Lambda(Y,X,\beta,\gamma,g))\} - \Lambda(Y,X,\beta,\gamma,g) \\ &= \Delta\{X^T\beta+\gamma(Y)+\zeta(Y,X,\beta,\gamma)\} \\ &\quad - \int_0^Y \exp\bigl(X^T\beta+\gamma(t)+\zeta(t,X,\beta,\gamma)\bigr)\,dt. \end{aligned}$$

Under conditions (C1)-(C4) , $l(\beta,\gamma,\zeta(\cdot,\beta,\gamma);W)$ has bounded and continuous first and second derivatives with respect to $\beta \in \mathcal{B}$, $\gamma \in \Gamma^{p_2}$, and $\zeta(\cdot,\beta,\gamma) \in \mathcal{H}^{p_1}$.

Proof of Lemma A.3.4. The derivatives with respect to the first, the second, and the third argument of the objective function are

$$l_1'(\beta,\gamma,\zeta;W) = \Delta X - X\int_0^Y \exp\bigl(X^T\beta+\gamma(t)+\zeta(t,X,\beta,\gamma)\bigr)\,dt,$$

$$l_2'(\beta,\gamma,\zeta;W)[v] = \Delta v(Y) - \int_0^Y \exp\bigl(X^T\beta+\gamma(t)+\zeta(t,X,\beta,\gamma)\bigr)v(t)\,dt,$$

$$l_3'(\beta,\gamma,\zeta;W)[h] = \Delta h(Y,X,\beta,\gamma) - \int_0^Y \exp\bigl(X^T\beta+\gamma(t)+\zeta(t,X,\beta,\gamma)\bigr)h(t,X,\beta,\gamma)\,dt,$$

$$l_{11}''(\beta,\gamma,\zeta;W) = -XX^T\int_0^Y \exp\bigl(X^T\beta+\gamma(t)+\zeta(t,X,\beta,\gamma)\bigr)\,dt,$$

$$l_{12}''(\beta,\gamma,\zeta;W)[v] = -X\int_0^Y \exp\bigl(X^T\beta+\gamma(t)+\zeta(t,X,\beta,\gamma)\bigr)v(t)\,dt,$$

$$l_{13}''(\beta,\gamma,\zeta;W)[h] = -X\int_0^Y \exp\bigl(X^T\beta+\gamma(t)+\zeta(t,X,\beta,\gamma)\bigr)h(t,X,\beta,\gamma)\,dt,$$

$$l_{23}''(\beta,\gamma,\zeta;W)[v,h] = -\int_0^Y \exp\bigl(X^T\beta+\gamma(t)+\zeta(t,X,\beta,\gamma)\bigr)v(t)h(t,X,\beta,\gamma)\,dt,$$

$$l_{22}''(\beta,\gamma,\zeta;W)[v_1,v_2] = -\int_0^Y \exp\bigl(X^T\beta+\gamma(t)+\zeta(t,X,\beta,\gamma)\bigr)v_1(t)v_2(t)\,dt,$$

$$l''_{33}(\beta,\gamma,\zeta;W)[h_1,h_2] = -\int_0^Y \exp\big(X^T\beta+\gamma(t)+\zeta(t,X,\beta,\gamma)\big)h_1(t,X,\beta,\gamma)h_2(t,X,\beta,\gamma)\,dt.$$

The derivatives with respect to β and γ of $\zeta(\cdot,\beta,\gamma)$ are

$$\begin{aligned}
\zeta'_\beta(\cdot,\beta,\gamma) &= g'(\Lambda(\cdot,\beta,\gamma,g))\Lambda'_\beta(\cdot,\beta,\gamma,g),\\
\zeta'_\gamma(\cdot,\beta,\gamma)[v] &= g'(\Lambda(\cdot,\beta,\gamma,g))\Lambda'_\gamma(\cdot,\beta,\gamma,g)[v],\\
\zeta''_{\beta\beta}(\cdot,\beta,\gamma) &= g''(\Lambda(\cdot,\beta,\gamma,g))\Lambda'_\beta(\cdot,\beta,\gamma,g)\Lambda'_\beta(\cdot,\beta,\gamma,g)^T\\
&\quad + g'(\Lambda(\cdot,\beta,\gamma,g))\Lambda''_{\beta\beta}(\cdot,\beta,\gamma,g),\\
\zeta''_{\gamma\gamma}(\cdot,\beta,\gamma)[v_1,v_2] &= g''(\Lambda(\cdot,\beta,\gamma,g))\Lambda'_\gamma(\cdot,\beta,\gamma,g)[v_1]\Lambda'_\gamma(\cdot,\beta,\gamma,g)[v_2]\\
&\quad + g'(\Lambda(\cdot,\beta,\gamma,g))\Lambda''_{\gamma\gamma}(\cdot,\beta,\gamma,g)[v_1,v_2],\\
\zeta''_{\beta\gamma}(\cdot,\beta,\gamma)[v] &= g''(\Lambda(\cdot,\beta,\gamma,g))\Lambda'_\beta(\cdot,\beta,\gamma,g)\Lambda'_\gamma(\cdot,\beta,\gamma,g)[v]\\
&\quad + g'(\Lambda(\cdot,\beta,\gamma,g))\Lambda''_{\beta\gamma}(\cdot,\beta,\gamma,g)[v].
\end{aligned}$$

After some calculations using the chain rule, we have

$$\begin{aligned}
&l'_\beta(\beta,\gamma,\zeta(\cdot,\beta,\gamma);W)\\
&\qquad = \Delta\{X+\zeta'_\beta(Y,X,\beta,\gamma)\}\\
&\qquad\quad - \int_0^Y \exp\big(X^T\beta+\gamma(t)+\zeta(t,X,\beta,\gamma)\big)[\zeta'_\beta(t,X,\beta,\gamma)+X]\,dt\\
&\qquad = \Delta\{X+g'(\Lambda(Y,X,\beta,\gamma,g))\Lambda'_\beta(Y,X,\beta,\gamma,g)\} - \Lambda'_\beta(Y,X,\beta,\gamma,g),\\
&l'_\gamma(\beta,\gamma,\zeta(\cdot,\beta,\gamma);W)[v]\\
&\qquad = \Delta\{v(Y)+\zeta'_\gamma(Y,X,\beta,\gamma)[v]\}\\
&\qquad\quad - \int_0^Y \exp\big(X^T\beta+\gamma(t)+\zeta(t,X,\beta,\gamma)\big)\{v(t)+\zeta'_\gamma(t,X,\beta,\gamma)[v]\}\,dt\\
&\qquad = \Delta\{v(Y)+g'(\Lambda(Y,X,\beta,\gamma,g))\Lambda'_\gamma(Y,X,\beta,\gamma,g)[v]\} - \Lambda'_\gamma(Y,X,\beta,\gamma,g)[v],\\
&l'_\zeta(\beta,\gamma,\zeta(\cdot,\beta,\gamma);W)[h]\\
&\qquad = \Delta h(Y,X,\beta,\gamma) - \int_0^Y \exp\big(X^T\beta+\gamma(t)+\zeta(t,X,\beta,\gamma)\big)h(t,X,\beta,\gamma)\,dt,
\end{aligned}$$

$$
\begin{aligned}
&l''_{\beta\beta}(\beta, \gamma, \zeta(\cdot, \beta, \gamma); W) \\
&\qquad = \Delta \zeta''_{\beta\beta}(Y, X, \beta, \gamma) - \int_0^Y \exp\bigl(X^T\beta + \gamma(t) + \zeta(t, X, \beta, \gamma)\bigr) \\
&\qquad\quad \cdot [(X + \zeta'_\beta(t, X, \beta, \gamma))(X + \zeta'_\beta(t, X, \beta, \gamma))^T + \zeta''_{\beta\beta}(t, X, \beta, \gamma)]\, dt \\
&\qquad = \Delta\{g''(\Lambda(Y, X, \beta, \gamma, g))\Lambda'_\beta(Y, X, \beta, \gamma, g)\Lambda'_\beta(Y, X, \beta, \gamma, g)^T \\
&\qquad\quad + g'(\Lambda(Y, X, \beta, \gamma, g))\Lambda''_{\beta\beta}(Y, X, \beta, \gamma, g)\} - \Lambda''_{\beta\beta}(Y, X, \beta, \gamma, g), \\
&l''_{\gamma\gamma}(\beta, \gamma, \zeta(\cdot, \beta, \gamma); W)[v_1, v_2] \\
&\qquad = \Delta \zeta''_{\gamma\gamma}(Y, X, \beta, \gamma)[v_1, v_2] - \int_0^Y \exp\bigl(X^T\beta + \gamma(t) + \zeta(t, X, \beta, \gamma)\bigr) \\
&\qquad\quad \cdot \{(v_1(t) + \zeta'_\gamma(t, X, \beta, \gamma)[v_1])(v_2(t) + \zeta'_\gamma(t, X, \beta, \gamma)[v_2]) \\
&\qquad\quad + \zeta''_{\gamma\gamma}(t, X, \beta, \gamma)[v_1, v_2]\}\, dt \\
&\qquad = \Delta\{g''(\Lambda(Y, X, \beta, \gamma, g))\Lambda'_\gamma(Y, X, \beta, \gamma, g)[v_1]\Lambda'_\gamma(Y, X, \beta, \gamma, g)[v_2] \\
&\qquad\quad + g'(\Lambda(Y, X, \beta, \gamma, g))\Lambda''_{\gamma\gamma}(Y, X, \beta, \gamma, g)[v_1, v_2]\} - \Lambda''_{\gamma\gamma}(Y, X, \beta, \gamma, g)[v_1, v_2], \\
&l''_{\gamma\beta}(\beta, \gamma, \zeta(\cdot, \beta, \gamma); W)[v] \\
&\qquad = \Delta \zeta''_{\gamma\beta}(Y, X, \beta, \gamma)[v] - \int_0^Y \exp\bigl(X^T\beta + \gamma(t) + \zeta(t, X, \beta, \gamma)\bigr) \\
&\qquad\quad \cdot \{(v(t) + \zeta'_\gamma(t, X, \beta, \gamma)[v])(X + \zeta'_\beta(t, X, \beta, \gamma)) \\
&\qquad\quad + \zeta''_{\gamma\beta}(t, X, \beta, \gamma)[v]\}\, dt \\
&\qquad = \Delta\{g''(\Lambda(Y, X, \beta, \gamma, g))\Lambda'_\gamma(Y, X, \beta, \gamma, g)[v]\Lambda'_\beta(Y, X, \beta, \gamma, g) \\
&\qquad\quad + g'(\Lambda(Y, X, \beta, \gamma, g))\Lambda''_{\gamma\beta}(Y, X, \beta, \gamma, g)[v]\} - \Lambda''_{\gamma\beta}(Y, X, \beta, \gamma, g)[v], \\
&l''_{\zeta\beta}(\beta, \gamma, \zeta(\cdot, \beta, \gamma); W)[h] \\
&\qquad = \Delta h'_\beta(Y, X, \beta, \gamma) - \int_0^Y \exp\bigl(X^T\beta + \gamma(t) + \zeta(t, X, \beta, \gamma)\bigr) \\
&\qquad\quad \cdot \{(h(t, X, \beta, \gamma))(X + \zeta'_\beta(t, X, \beta, \gamma)) + h'_\beta(t, X, \beta, \gamma)\}\, dt \\
&\qquad = \Delta\{w'(\Lambda(Y, X, \beta, \gamma, g))\Lambda'_\beta(Y, X, \beta, \gamma, g)
\end{aligned}
$$

$$
\begin{aligned}
&\qquad + g''(\Lambda(Y,X,\beta,\gamma,g))\Lambda'_g(Y,X,\beta,\gamma,g)[w]\Lambda'_\beta(Y,X,\beta,\gamma,g)\\
&\qquad + g'(\Lambda(Y,X,\beta,\gamma,g))\Lambda''_{g\beta}(Y,X,\beta,\gamma,g)[w]\}\\
&\quad - \Lambda''_{g\beta}(Y,X,\beta,\gamma,g)[w],
\end{aligned}
$$

$$
\begin{aligned}
&l''_{\zeta\gamma}(\beta,\gamma,\zeta(\cdot,\beta,\gamma);W)[h,v]\\
&\quad = \Delta h'_\gamma(Y,X,\beta,\gamma)[v] - \int_0^Y \exp\bigl(X^T\beta+\gamma(t)+\zeta(t,X,\beta,\gamma)\bigr)\\
&\qquad \cdot \{(h(t,X,\beta,\gamma))(v(t)+\zeta'_\gamma(t,X,\beta,\gamma)[v]) + h'_\gamma(t,X,\beta,\gamma)[v]\}\,dt\\
&\quad = \Delta\{w'(\Lambda(Y,X,\beta,\gamma,g))\Lambda'_\gamma(Y,X,\beta,\gamma,g)[v]\\
&\qquad + g''(\Lambda(Y,X,\beta,\gamma,g))\Lambda'_g(Y,X,\beta,\gamma,g)[w]\Lambda'_\gamma(Y,X,\beta,\gamma,g)[v]\\
&\qquad + g'(\Lambda(Y,X,\beta,\gamma,g))\Lambda''_{g\gamma}(Y,X,\beta,\gamma,g)[w,v]\}\\
&\quad - \Lambda''_{g\gamma}(Y,X,\beta,\gamma,g)[w,v],
\end{aligned}
$$

$$
\begin{aligned}
&l''_{\zeta\zeta}(\beta,\gamma,\zeta(\cdot,\beta,\gamma);W)[h_1,h_2]\\
&\quad = -\int_0^Y \exp\bigl(X^T\beta+\gamma(t)+\zeta(t,X,\beta,\gamma)\bigr)h_1(t,X,\beta,\gamma)h_2(t,X,\beta,\gamma)\,dt,\\
&\quad = -\int_0^Y \exp\bigl(X^T\beta+\gamma(t)+\zeta(t,X,\beta,\gamma)\bigr)\\
&\qquad \cdot \{w_1(\Lambda(t,X,\beta,\gamma,g)) + g'(\Lambda(t,X,\beta,\gamma,g))\Lambda'_g(t,X,\beta,\gamma,g)[w_1]\}\\
&\qquad \cdot \{w_2(\Lambda(t,X,\beta,\gamma,g)) + g'(\Lambda(t,X,\beta,\gamma,g))\Lambda'_g(t,X,\beta,\gamma,g)[w_2]\}\,dt,
\end{aligned}
$$

All the above derivatives are bounded and continuous under conditions (C1)-(C4) by Lemma A.3.1. □

Lemma A.3.5. *(Spline approximation) For $\gamma_0 \in \Gamma^{p_1}$, there exists a function $\gamma_{0n} \in \Gamma_n^{p_1}$ such that*

$$\|\gamma_{0n} - \gamma_0\|_\infty = O(n^{-p_1\nu_1}).$$

For $g_0 \in \mathcal{G}^{p_2}$, there exists a function $g_{0n} \in \mathcal{G}_n^{p_2}$ such that

$$\|g_{0n} - g_0\|_\infty = O(n^{-p_2\nu_2}).$$

Proof of Lemma A.3.5. Since $\gamma_0 \in \Gamma^{p_1} \subset S^{p_1}([0,\tau])$, by Corollary 6.21 in Schumaker (2007), there exists a function in the polynomial space with order p_1, i.e., $\tilde{\gamma}_{0n} \in S_n(T_{K_n^1}, K_n^1, p_1)$, such that $\|\tilde{\gamma}_{0n} - \gamma_0\|_\infty = O(n^{-p_1\nu_1})$. It follows that

$$\begin{aligned}\|(\tilde{\gamma}_{0n}(\cdot) - \tilde{\gamma}_{0n}(t^*)) - \gamma_0\|_\infty &\le \|\tilde{\gamma}_{0n} - \gamma_0\|_\infty + |\tilde{\gamma}_{0n}(t^*)| \\ &= \|\tilde{\gamma}_{0n} - \gamma_0\|_\infty + |\tilde{\gamma}_{0n}(t^*) - \gamma_0(t^*)| \\ &\le 2\|\tilde{\gamma}_{0n} - \gamma_0\|_\infty = O(n^{-p_1\nu_1}),\end{aligned}$$

where the second equality holds because $\gamma_0(t^*) = 0$ for $\gamma_0 \in \Gamma^{p_1}$. Let $\gamma_{0n}(\cdot) = \tilde{\gamma}_{0n}(\cdot) - \tilde{\gamma}_{0n}(t^*)$, then $\gamma_{0n}(t^*) = 0$ and thereby we find $\gamma_{0n} \in \Gamma_n^{p_1}$ such that $\|\gamma_{0n} - \gamma_0\|_\infty = O(n^{-p_1\nu_1})$. The second part is a direct result of Corollary 6.21 in Schumaker (2007). □

Lemma A.3.6. *(Bracket number of $l(\theta; W)$) Let $\theta_{0n} = (\beta_0, \gamma_{0n}(\cdot), \zeta_{0n}(\cdot, \beta_0, \gamma_{0n}))$ with*

$$\zeta_{0n}(t, x, \beta_0, \gamma_{0n}) = g_{0n}(\Lambda(t, x, \beta_0, \gamma_{0n}, g_{0n})),$$

where γ_{0n} and g_{0n} are defined in Lemma A.3.5. Denote $\mathcal{F}_n = \{l(\theta; W) - l(\theta_{0n}; W) : \theta \in \Theta_n\}$. Under conditions (C1)-(C4), the ϵ-bracketing number associated with $\|\cdot\|_\infty$ for $\mathcal{F}_n$, denoted by $N_{[\,]}(\epsilon, \mathcal{F}_n, \|\cdot\|_\infty)$, has the following upper bound for some constants c_1 and c_2,

$$N_{[\,]}(\epsilon, \mathcal{F}_n, \|\cdot\|_\infty) \lesssim \left(\frac{1}{\epsilon}\right)^{c_1 q_{n_1} + c_2 q_{n_2} + d}.$$

Proof of Lemma A.3.6. Denote the ceiling of x by $\lceil x \rceil$. Following the calculation in Shen and Wong (1994, Page 597), we have that, for any $\epsilon > 0$, there exists a set of ϵ-brackets

$$\left\{[\gamma_l^L, \gamma_l^U] : \|\gamma_l^U - \gamma_l^L\|_\infty \le \epsilon, l = 1, \cdots, \left\lceil (\frac{1}{\epsilon})^{c_1 q_{n_1}} \right\rceil \right\}$$

such that for any $\gamma \in \Gamma_n^{p_1}$, $\gamma_l^L(t) \le \gamma(t) \le \gamma_l^U(t)$ holds on $[0,\tau]$ for some $1 \le l \le \lceil (\frac{1}{\epsilon})^{c_1 q_{n_1}} \rceil$. Similarly, there exists another set of ϵ-brackets

$$\left\{[g_i^L, g_i^U] : \|g_i^U - g_i^L\|_\infty \le \epsilon, i = 1, \cdots, \left\lceil (\frac{1}{\epsilon})^{c_2 q_{n_2}} \right\rceil \right\}$$

such that for any $g \in \mathcal{G}_n^{p_1}$, $g_i^L(t) \le g(t) \le g_i^U(t)$ holds on $[0,\mu]$ for some $1 \le i \le$

$\lceil (\frac{1}{\epsilon})^{c_2 q_{n_2}} \rceil$. Since $\mathcal{B} \subset \mathbf{R}^d$ is compact, it can be covered by $\lceil c_3(\frac{1}{\epsilon})^d \rceil$ balls with radius ϵ, i.e. for any $\beta \in \mathcal{B}$, there exists $1 \leq k \leq \lceil c_3(\frac{1}{\epsilon})^d \rceil$ such that $\|\beta_k - \beta\| \leq \epsilon$. Hence, under condition (C2), $|X^T\beta - X^T\beta_k| \leq c_4\epsilon$ for some constant $c_4 > 0$ and any $X \in \mathcal{X}$. By the mean value theorem, we have that

$$\begin{aligned}
&|\exp\bigl(g(\Lambda) + X^T\beta + \gamma(t)\bigr) - \exp\bigl(g_i^L(\Lambda) + X^T\beta_k + \gamma_l^L(t)\bigr)| \\
&= \exp\Bigl(\tilde{\psi}(t,\Lambda)\Bigr)|g(\Lambda) + X^T\beta + \gamma(t) - g_i^L(\Lambda) + X^T\beta_k + \gamma_l^L(t)| \\
&\leq \exp\Bigl(\tilde{\psi}(t,\Lambda)\Bigr)(|g(\Lambda) - g_i^L(\Lambda)| + |X^T\beta - X^T\beta_k| + |\gamma(t) - \gamma_l^L(t)|) \\
&\leq \exp\Bigl(\tilde{\psi}(t,\Lambda)\Bigr)(\|g - g_i^L\|_\infty + |X^T\beta - X^T\beta_k| + \|\gamma - \gamma_l^L\|_\infty),
\end{aligned}$$

where $\tilde{\psi}(t,\Lambda) = g_i^L(\Lambda) + X^T\beta_k + \gamma_l^L(t) + \xi(g(\Lambda) - g_i^L(\Lambda) + X^T(\beta - \beta_k) + \gamma(t) - \gamma_l^L(t))$ for some $\xi \in (0,1)$ and is bounded under conditions (C1)-(C4). Hence,

$$|\exp\bigl(g(\Lambda) + X^T\beta + \gamma(t)\bigr) - \exp\bigl(g_i^L(\Lambda) + X^T\beta_k + \gamma_l^L(t)\bigr)| \lesssim \epsilon$$

over $(t,\Lambda) \in [0,\tau] \times [0,b]$. Employing Theorem 12.V of continuous dependence in Walter (1998, page 145), we have $|\Lambda(t,X,\beta,\gamma,g) - \Lambda(t,X,\beta_k,\gamma_l^L,g_i^L)| \leq c_5\epsilon$ for some constant $c_5 > 0$ and any $t \in [0,\tau]$. Denote $\Lambda_{ilk}(t,x) = \Lambda(t,x,\beta_k,\gamma_l^L,g_i^L)$. Define

$$\begin{aligned}
m(\theta;W) &= l(\theta;W) - l(\theta_{0n};W) \\
&= \Delta\{X^T\beta + \gamma(Y) + g(\Lambda(Y,X,\beta,\gamma,g))\} - \Lambda(Y,X,\beta,\gamma,g) - l(\theta_{0n};W),
\end{aligned}$$

$$m_{ilk}^L(W) = \Delta\{X^T\beta_k - c_4\epsilon + \gamma_l^L(Y) + g_i^L(\xi_{ilk}^L)\} - \Lambda_{ilk}(Y,X) - c_5\epsilon - l(\theta_{0n};W),$$

and

$$m_{ilk}^U(W) = \Delta\{X^T\beta_k + c_4\epsilon + \gamma_l^U(Y) + g_i^U(\xi_{ilk}^U)\} - \Lambda_{ilk}(Y,X) + c_5\epsilon - l(\theta_{0n};W),$$

where $\xi_{ilk}^L = \arg\min_{|s|\leq c_5\epsilon} g_i^L(\Lambda_{ilk}(Y,X) + s)$ and $\xi_{ilk}^U = \arg\max_{|s|\leq c_5\epsilon} g_i^U(\Lambda_{ilk}(Y,X) + s)$.

Note that $[m_{ilk}^L(W), m_{ilk}^U(W)]$ is a ϵ-bracket because

$$\begin{aligned}
&|m_{ilk}^U(W) - m_{ilk}^L(W)| \\
&\qquad = |\Delta\{2c_4\epsilon + \gamma_l^U(Y) - \gamma_l^L(Y) + g_i^U(\xi_{ilk}^U) - g_i^L(\xi_{ilk}^L)\} + 2c_5\epsilon| \\
&\qquad \leq 2c_4\epsilon + |\gamma_l^U(Y) - \gamma_l^L(Y)| + |g_i^U(\xi_{ilk}^U) - g_i^L(\xi_{ilk}^U)| + |g_i^L(\xi_{ilk}^U) - g_i^L(\xi_{ilk}^L)| + 2c_5\epsilon
\end{aligned}$$

$$\leq 2c_4\epsilon + \|\gamma_l^U - \gamma_l^L\|_\infty + \|g_i^U - g_i^L\|_\infty + c_7|\xi_{ilk}^U - \xi_{ilk}^L| + 2c_5\epsilon$$

$$\leq 2c_4\epsilon + \epsilon + \epsilon + 2c_7c_5\epsilon + 2c_5\epsilon \lesssim \epsilon,$$

where $c_7 = \max_{t\in[0,b]} |(g_i^L)'(t)|$ in the second inequality. Hence $\|m_{ilk}^U - m_{ilk}^L\|_\infty \lesssim \epsilon$.

For any $\theta = (\beta, \gamma(\cdot), \zeta(\cdot, \beta, \gamma))$ with $\zeta(t, x, \beta, \gamma) = g(\Lambda(t, x, \beta, \gamma, g))$, there exits $1 \leq i \leq \lceil (\frac{1}{\epsilon})^{c_2 q_{n_2}} \rceil, 1 \leq l \leq \lceil (\frac{1}{\epsilon})^{c_1 q_{n_1}} \rceil, 1 \leq k \leq \lceil c_3 (\frac{1}{\epsilon})^d \rceil$ such that $g_i^L(t) \leq g(t) \leq g_i^U(t)$ on $t \in [0, \mu]$, $\gamma_l^L(t) \leq \gamma(t) \leq \gamma_l^U(t)$ on $t \in [0, \tau]$, and $|X^T\beta_k - X^T\beta| \leq c_4\epsilon$. It follows that

$$\begin{aligned} m_{ilk}^U(W) &= \Delta\{(X^T\beta_k + c_4\epsilon) + \gamma_l^U(Y) + g_i^U(\xi_{ilk}^U)\} + (c_5\epsilon - \Lambda_{ilk}(Y, X)) - l(\theta_{0n}; W) \\ &\geq \Delta\{X^T\beta + \gamma(Y) + g_i^U(\xi_{ilk}^U)\} + (c_5\epsilon - \Lambda_{ilk}(Y, X)) - l(\theta_{0n}; W) \\ &\geq \Delta\{X^T\beta + \gamma(Y) + g_i^U(\Lambda(t, X, \beta, \gamma, g))\} - \Lambda(t, X, \beta, \gamma, g) - l(\theta_{0n}; W) \\ &\geq \Delta\{X^T\beta + \gamma(Y) + g(\Lambda(t, X, \beta, \gamma, g))\} - \Lambda(t, X, \beta, \gamma, g) - l(\theta_{0n}; W) \\ &= m(\theta; W), \end{aligned}$$

where the second inequality holds because $|\Lambda(Y, X, \beta, \gamma, g) - \Lambda_{ilk}(Y, X)| \leq c_5\epsilon$. The other side can be verified similarly. Therefore, we have

$$N_{[\,]}(\epsilon, \mathcal{F}_n, \|\cdot\|_\infty) \lesssim \left(\frac{1}{\epsilon}\right)^{c_1 q_{n_1}} \left(\frac{1}{\epsilon}\right)^{c_2 q_{n_2}} \left(\frac{1}{\epsilon}\right)^{d} = \left(\frac{1}{\epsilon}\right)^{c_1 q_{n_1} + c_2 q_{n_2} + d},$$

which completes the proof. □

Lemma A.3.7. *For $1 \leq j \leq d$, denote $\mathcal{F}_{n,j}^\gamma(\eta) = \{l_\gamma'(\theta; W)[v_j^* - v_j] : \theta \in \Theta_n, v_j \in \Gamma_n^1, d(\theta, \theta_0) \leq \eta, \|v_j^* - v_j\|_\infty \leq \eta\}$ and $\mathcal{F}_{n,j}^\zeta(\eta) = \{l_\zeta'(\theta; W)[h_j^* - h_j] : \theta \in \Theta_n, h_j \in \mathcal{H}_n^2, d(\theta, \theta_0) \leq \eta, \|w_j^* - w_j\|_\infty \leq \eta\}$, where v_j^* is defined in condition (C7) and $h_j^*(\cdot, \beta, \gamma) = w_j^*(\Lambda(\cdot, \beta, \gamma, g)) + g'(\Lambda(\cdot, \beta, \gamma, g))\Lambda_g'(\cdot, \beta, \gamma, g)[w_j^*]$ with w_j^* given in condition (C7). Then under conditions (C1)-(C4) and (C7), we have*

$$N_{[\,]}(\epsilon, \mathcal{F}_{n,j}^\gamma(\eta), \|\cdot\|_\infty) \lesssim \left(\frac{\eta}{\epsilon}\right)^{c_1 q_{n_1} + c_2 q_{n_2} + d}$$

and

$$N_{[\,]}(\epsilon, \mathcal{F}_{n,j}^\zeta(\eta), \|\cdot\|_\infty) \lesssim \left(\frac{\eta}{\epsilon}\right)^{c_3 q_{n_1} + c_4 q_{n_2} + d}$$

for some constants $c_1, c_2, c_3,$ and c_4.

Lemma A.3.8. *For $1 \leq j \leq d$, denote*

$$\begin{aligned}\mathcal{F}^{*\beta}_{n,j}(\eta) = \{l'_{\beta_j}(\theta;W) - &l'_{\beta_j}(\theta_0;W) : \\ &\theta \in \Theta_n, d(\theta,\theta_0) \leq \eta, \|g'(\Lambda(\cdot,\beta,\gamma,g)) - g'_0(\Lambda_0(\cdot))\|_2 \leq \eta\},\end{aligned}$$

$$\begin{aligned}\mathcal{F}^{*\gamma}_{n,j}(\eta) = \{l'_{\gamma}(\theta;W)[v^*_j] - &l'_{\gamma}(\theta_0;W)[v^*_j] : \\ &\theta \in \Theta_n, d(\theta,\theta_0) \leq \eta, \|g'(\Lambda(\cdot,\beta,\gamma,g)) - g'_0(\Lambda_0(\cdot))\|_2 \leq \eta\},\end{aligned}$$

and

$$\mathcal{F}^{*\zeta}_{n,j}(\eta) = \{l'_{\zeta}(\theta;W)[h^*_j] - l'_{\zeta}(\theta_0;W)[h^*_j] : \theta \in \Theta_n, d(\theta,\theta_0) \leq \eta\},$$

where v^*_j is defined in condition (C7) and

$$h^*_j(\cdot,\beta,\gamma) = w^*_j(\Lambda(\cdot,\beta,\gamma,g)) + g'(\Lambda(\cdot,\beta,\gamma,g))\Lambda'_g(\cdot,\beta,\gamma,g)[w^*_j]$$

with w^*_j given in condition (C7). Then under conditions (C1)-(C4) and (C7), we have

$$\begin{aligned}N_{[\,]}(\epsilon, \mathcal{F}^{*\beta}_{n,j}(\eta), \|\cdot\|_\infty) &\lesssim \left(\frac{\eta}{\epsilon}\right)^{c_1 q_{n_1} + c_2 q_{n_2} + d}, \\ N_{[\,]}(\epsilon, \mathcal{F}^{*\gamma}_{n,j}(\eta), \|\cdot\|_\infty) &\lesssim \left(\frac{\eta}{\epsilon}\right)^{c_3 q_{n_1} + c_4 q_{n_2} + d},\end{aligned}$$

and

$$N_{[\,]}(\epsilon, \mathcal{F}^{*\zeta}_{n,j}(\eta), \|\cdot\|_\infty) \lesssim \left(\frac{\eta}{\epsilon}\right)^{c_5 q_{n_1} + c_6 q_{n_2} + d}$$

for some constants c_i, $i = 1, \ldots, 6$.

The proofs of Lemma A.3.7 and A.3.8 follow a similar calculation as in Lemma A.3.6 and therefore are omitted here.

A.3.2 Proof of Theorem 2.3.1

Proof of Theorem 2.3.1. We prove the theorem by checking the conditions C1-3 in Shen and Wong (1994, Theorem 1). Using the fact $P\{\int_0^Y f(t,X)d\Lambda_0(t,X)\} = P\{\Delta f(Y,X)\}$, we have

$$\begin{aligned}Pl(\beta,\gamma,&\zeta(\cdot,\beta,\gamma);W) = P\{\Delta[X^T\beta + \gamma(Y) + g(\Lambda(Y,X,\beta,\gamma,g)) \\ &- \exp\bigl(X^T\beta + \gamma(Y) + g(\Lambda(Y,X,\beta,\gamma,g)) - X^T\beta_0 - \gamma_0(Y) - g_0(\Lambda_0(Y,X))\bigr)]\}.\end{aligned}$$

It follows that, by the Taylor expansion,

$$\begin{aligned}
&Pl(\beta_0,\gamma_0,\zeta_0(\cdot,\beta_0,\gamma_0);W)-Pl(\beta,\gamma,\zeta(\cdot,\beta,\gamma);W)\\
&=P\{\Delta[\exp\big(X^T\beta+\gamma(Y)+g(\Lambda(Y,X,\beta,\gamma,g))-X^T\beta_0-\gamma_0(Y)-g_0(\Lambda_0(Y,X))\big)\\
&\qquad -1-(X^T\beta+\gamma(Y)+g(\Lambda(Y,X,\beta,\gamma,g))-X^T\beta_0-\gamma_0(Y)-g_0(\Lambda_0(Y,X)))]\}\\
&=\frac{1}{2}A+o(A), \qquad\qquad (A.22)
\end{aligned}$$

where $A=P\{\Delta[X^T\beta+\gamma(Y)+g(\Lambda(Y,X,\beta,\gamma,g))-X^T\beta_0-\gamma_0(Y)-g_0(\Lambda_0(Y,X))]^2\}$. After subtracting and adding the term $g(\Lambda_0(Y,X))$, we have

$$\begin{aligned}
A&=P\{\Delta[X^T(\beta-\beta_0)+\gamma(Y)-\gamma_0(Y)+g(\Lambda(Y,X,\beta,\gamma,g))-g(\Lambda_0(Y,X))\\
&\qquad +g(\Lambda_0(Y,X))-g_0(\Lambda_0(Y,X))]^2\}\\
&=P\{\Delta[(g'(\Lambda_0(Y,X))\Lambda'_{0\beta}(Y,X)+X)^T(\beta-\beta_0)\\
&\qquad +g'(\Lambda_0(Y,X))\Lambda'_{0\gamma}(Y,X)[\gamma-\gamma_0]+\gamma(Y)-\gamma_0(Y)\\
&\qquad +g'(\Lambda_0(Y,X))\Lambda'_{0g}(Y,X)[g-g_0]+g(\Lambda_0(Y,X))-g_0(\Lambda_0(Y,X))\\
&\qquad +o(\|\beta-\beta_0\|)+o(\|\gamma-\gamma_0\|_2)+o(\|g-g_0\|_2)]^2\},
\end{aligned}$$

where the second equality is obtained by using the Taylor expansion. Since $\Lambda'_{0\beta}(t,x)$ is bounded by Lemma A.3.1 and $\Lambda'_{0\gamma}(\cdot)[v]$ and $\Lambda'_{0g}(\cdot)[w]$ are bounded linear operators, which can be verified using the same arguments as in Lemma A.3.2, we have

$$\begin{aligned}
g'(\Lambda_0(Y,X))\Lambda'_{0\beta}(Y,X)^T(\beta-\beta_0)&=g'_0(\Lambda_0(Y,X))\Lambda'_{0\beta}(Y,X)^T(\beta-\beta_0)\\
&\qquad +o(\|\beta-\beta_0\|)+o(\|g-g_0\|_2),\\
g'(\Lambda_0(Y,X))\Lambda'_{0\gamma}(Y,X)[\gamma-\gamma_0]&=g'_0(\Lambda_0(Y,X))\Lambda'_{0\gamma}(Y,X)[\gamma-\gamma_0]\\
&\qquad +o(\|\gamma-\gamma_0\|_2)+o(\|g-g_0\|_2),\\
g'(\Lambda_0(Y,X))\Lambda'_{0g}(Y,X)[g-g_0]&=g'_0(\Lambda_0(Y,X))\Lambda'_{0g}(Y,X)[g-g_0]+o(\|g-g_0\|_2).
\end{aligned}$$

Note that under conditions (C1)-(C4), we have

$$d^2(\theta,\theta_0)\lesssim\|\beta-\beta_0\|^2+\|\gamma-\gamma_0\|_2^2+\|g-g_0\|_2^2\lesssim d^2(\theta,\theta_0). \qquad (A.23)$$

Plugging these equations above into A, it follows that

$$\begin{aligned} A \gtrsim P\{\Delta[(& g_0'(\Lambda_0(Y,X))\Lambda_{0\beta}'(Y,X) + X)^T(\beta - \beta_0) \\ &+ g_0'(\Lambda_0(Y,X))\Lambda_{0\gamma}'(Y,X)[\gamma - \gamma_0] + \gamma(Y) - \gamma_0(Y) \\ &+ g_0'(\Lambda_0(Y,X))\Lambda_{0g}'(Y,X)[g - g_0] + g(\Lambda_0(Y,X)) - g_0(\Lambda_0(Y,X))]^2\} \\ + o(d^2&(\theta,\theta_0)). \end{aligned} \tag{A.24}$$

Then, by solving the initial value problem in (A.9), we have

$$\begin{aligned} g_0'(\Lambda_0(Y,X))\Lambda_{0\beta}'(Y,X) + X &= (g_0'(\Lambda_0(Y,X))\exp(g_0(\Lambda_0(Y,X)))R(Y)e^{X^T\beta_0} + 1)X \\ &= (g_0'(\tilde{\Lambda}_0(U))\exp\Big(g_0(\tilde{\Lambda}_0(U))\Big)U + 1)X \\ &\triangleq \epsilon_1(U)X, \end{aligned} \tag{A.25}$$

with U given in condition (C5) and ϵ_1 is a deterministic function.

Note that using equations (A.19) and (A.20) in Lemma A.3.2, we also have

$$\begin{aligned} \psi_{0\gamma}'(Y,X)[\gamma - \gamma_0] &= g_0'(\Lambda_0(Y,X))\Lambda_{0\gamma}'(Y,X)[\gamma - \gamma_0] + \gamma(Y) - \gamma_0(Y) \\ &= g_0'(\tilde{\Lambda}_0(U)\exp\Big(g_0(\tilde{\Lambda}_0(U))\Big)\int_0^U (\gamma - \gamma_0)(R^{-1}(se^{-V}))\,ds \\ &\quad + (\gamma - \gamma_0)(R^{-1}(Ue^{-V})) \qquad \text{(A.26)} \\ &\triangleq \epsilon_2(U,V)[\gamma - \gamma_0], \qquad \text{(A.27)} \end{aligned}$$

which is a deterministic function of U and V given in condition (C5), and

$$\begin{aligned} \psi_{0g}'(Y,X)[g - g_0] &= g_0'(\Lambda_0(Y,X))\Lambda_{0g}'(Y,X)[g - g_0] + g(\Lambda_0(Y,X)) - g_0(\Lambda_0(Y,X)) \\ &= g_0'(\tilde{\Lambda}_0(U)\exp\Big(g_0(\tilde{\Lambda}_0(U))\Big)\int_0^{\tilde{\Lambda}_0(U)} \exp(-g_0(s))(g - g_0)(s)\,ds \\ &\quad + (g - g_0)(\tilde{\Lambda}_0(U)) \qquad \text{(A.28)} \\ &\triangleq \epsilon_3(U)[g - g_0], \qquad \text{(A.29)} \end{aligned}$$

which is a deterministic function of U.

Then, it follows from (A.24)

$$\begin{aligned}
A &\gtrsim P\{\Delta[\epsilon_1(U)X^T(\beta-\beta_0)+\epsilon_2(U,V)[\gamma-\gamma_0]+\epsilon_3(U)[g-g_0]]^2\}+o(d^2(\theta,\theta_0))\\
&= P\{\Delta(\epsilon_1(U)X^T(\beta-\beta_0))^2\}+P\{\Delta(\epsilon_2(U,V)[\gamma-\gamma_0]+\epsilon_3(U)[g-g_0])^2\}\\
&\quad +2P\{\Delta(\epsilon_1(U)X^T(\beta-\beta_0))(\epsilon_2(U,V)[\gamma-\gamma_0]+\epsilon_3(U)[g-g_0])\}+o(d^2(\theta,\theta_0))\\
&\geq P\{\Delta(\epsilon_1(U)X^T(\beta-\beta_0))^2\}+P\{\Delta(\epsilon_2(U,V)[\gamma-\gamma_0]+\epsilon_3(U)[g-g_0])^2\}\\
&\quad -2|P\{\Delta(\epsilon_1(U)X^T(\beta-\beta_0))(\epsilon_2(U,V)[\gamma-\gamma_0]+\epsilon_3(U)[g-g_0])\}|+o(d^2(\theta,\theta_0)).
\end{aligned} \tag{A.30}$$

By using the fact that

$$P\{\Delta f(U,X)\}=P\{\int_0^Y f(R(t)e^{X^T\beta_0},X)\,d\Lambda_0(t,X)\}=P\{\int_0^U f(t,X)\,d\tilde{\Lambda}_0(t)\},$$

we have

$$\begin{aligned}
&|P\{\Delta(\epsilon_1(U)X^T(\beta-\beta_0))(\epsilon_2(U,V)[\gamma-\gamma_0]+\epsilon_3(U)[g-g_0])\}|^2\\
&=\left(P\left\{\int_0^U \epsilon_1(t)X^T(\beta-\beta_0)(\epsilon_2(t,V)[\gamma-\gamma_0]+\epsilon_3(t)[g-g_0])\,d\tilde{\Lambda}_0(t)\right\}\right)^2\\
&=\left(P\left\{\int_0^U \epsilon_1(t)P\{X^T(\beta-\beta_0)|U,V\}(\epsilon_2(t,V)[\gamma-\gamma_0]+\epsilon_3(t)[g-g_0])\,d\tilde{\Lambda}_0(t)\right\}\right)^2\\
&\leq P\left\{\int_0^U (\epsilon_1(t))^2\left(P\{X^T(\beta-\beta_0)|U,V\}\right)^2 d\tilde{\Lambda}_0(t)\right\}\\
&\qquad\cdot P\left\{\int_0^U (\epsilon_2(t,V)[\gamma-\gamma_0]+\epsilon_3(t)[g-g_0])^2 d\tilde{\Lambda}_0(t)\right\},
\end{aligned}$$

where the last step is obtained using the Cauchy-Schwartz inequality. Under Condition (C5), there exists $\eta_1\in(0,1)$ such that

$$(1-\eta_1)(\beta-\beta_0)^TP\{XX^T|U,V\}(\beta-\beta_0)\geq(P\{X^T(\beta-\beta_0)|U,V\})^2,$$

since the first element of $\beta-\beta_0$ is zero with the identifiability constraint. Thus, we have

$$|P\{\Delta(\epsilon_1(U)X^T(\beta-\beta_0))(\epsilon_2(U,V)[\gamma-\gamma_0]+\epsilon_3(U)[g-g_0])\}|^2$$

$$
\begin{aligned}
\leq &(1-\eta_1) P\left\{\int_0^U (\epsilon_1(t))^2(\beta-\beta_0)^T P\{XX^T|U,V\}(\beta-\beta_0)\, d\tilde{\Lambda}_0(t)\right\}\\
&\qquad\cdot P\left\{\int_0^U (\epsilon_2(t,V)[\gamma-\gamma_0]+\epsilon_3(t)[g-g_0])^2\, d\tilde{\Lambda}_0(t)\right\}\\
= &(1-\eta_1) P\left\{\int_0^U (\epsilon_1(t)X^T(\beta-\beta_0))^2\, d\tilde{\Lambda}_0(t)\right\}\\
&\qquad\cdot P\left\{\int_0^U (\epsilon_2(t,V)[\gamma-\gamma_0]+\epsilon_3(t)[g-g_0])^2\, d\tilde{\Lambda}_0(t)\right\}\\
= &(1-\eta_1) P\{\Delta(\epsilon_1(U)X^T(\beta-\beta_0))^2\} P\{\Delta(\epsilon_2(U,V)[\gamma-\gamma_0]+\epsilon_3(U)[g-g_0])^2\},
\end{aligned}
$$

and it yields from (A.30) that

$$
\begin{aligned}
A \geq\, & P\{\Delta(\epsilon_1(U)X^T(\beta-\beta_0))^2\} + P\{\Delta(\epsilon_2(U,V)[\gamma-\gamma_0]+\epsilon_3(U)[g-g_0])^2\}\\
& - 2(1-\eta_1)^{1/2}(P\{\Delta(\epsilon_2(U,V)[\gamma-\gamma_0]+\epsilon_3(U)[g-g_0])^2\})^{1/2}\cdot\\
& (P\{\Delta(\epsilon_1(U)X^T(\beta-\beta_0))^2\})^{1/2}\\
\geq\, & (1-(1-\eta_1)^{1/2})\{P\{\Delta(\epsilon_1(U)X^T(\beta-\beta_0))^2\}\\
& \qquad\qquad + P\{\Delta(\epsilon_2(U,V)[\gamma-\gamma_0]+\epsilon_3(U)[g-g_0])^2\}\}\\
\gtrsim\, & P\{\Delta(\epsilon_1(U)X^T(\beta-\beta_0))^2\} + P\{\Delta(\epsilon_2(U,V)[\gamma-\gamma_0]+\epsilon_3(U)[g-g_0])^2\}\\
=\, & A_1 + A_2,
\end{aligned}
$$

where the second inequality is obtained by $2ab \leq a^2+b^2$.

For A_1, under condition (C3), we have for $t \in [0,\tau]$,

$$P\{\mathbb{1}(Y>t)|X\} \geq P\{\mathbb{1}((Y>\tau)|X\} \geq \delta_0.$$

Then it follows that,

$$
\begin{aligned}
A_1 &= P\{\int_0^Y \exp\big(X^T\beta_0+\gamma_0(t)+g(\Lambda_0(t,X))\big)\left(\epsilon_1(R(t)e^{X^T\beta_0})X^T(\beta-\beta_0)\right)^2 dt\}\\
&= P\{\int_0^\tau P\{\mathbb{1}(Y>t)|X\}
\end{aligned}
$$

$$
\begin{aligned}
&\cdot \exp\big(X^T\beta_0 + \gamma_0(t) + g(\Lambda_0(t,X))\big) \Big(\epsilon_1(R(t)e^{X^T\beta_0})X^T(\beta-\beta_0)\Big)^2 dt\} \\
\geq{}& \delta_0 P\{\int_0^{\tau} \exp\big(X^T\beta_0 + \gamma_0(t) + g(\Lambda_0(t,X))\big) \Big(\epsilon_1(R(t)e^{X^T\beta_0})X^T(\beta-\beta_0)\Big)^2 dt\} \\
={}& \delta_0 P\{\int_0^{R(\tau)e^{X^T\beta_0}} \exp\Big(g(\tilde{\Lambda}_0(s))\Big) \big(\epsilon_1(s)X^T(\beta-\beta_0)\big)^2 ds\} \\
\geq{}& \delta_0 P\{\int_0^{cR(\tau)} \exp\Big(g(\tilde{\Lambda}_0(s))\Big) \big(\epsilon_1(s)X^T(\beta-\beta_0)\big)^2 ds\} \\
={}& \delta_0(\beta-\beta_0)^T P\{XX^T\}(\beta-\beta_0) \int_0^{cR(\tau)} \exp\Big(g(\tilde{\Lambda}_0(s))\Big)(\epsilon_1(s))^2 ds,
\end{aligned}
$$

where the fourth equality is derived by variable transformation $s = R(t)e^{X^T\beta_0}$ and $c = \min_{x\in\mathcal{X}} e^{x^T\beta_0}$, which is positive since $\mathcal{X}$ is bounded under condition (C2). As condition (C2) implies that the smallest eigenvalue of $P\{XX^T\}$, denoted by λ_1, is positive as well, we have $(\beta-\beta_0)^T P\{XX^T\}(\beta-\beta_0) \geq \lambda_1\|\beta-\beta_0\|^2$. Also, by definition $\epsilon_1(s)$ satisfies the equation $g_0'(\tilde{\Lambda}_0(t)) \int_0^t \exp\Big(g_0(\tilde{\Lambda}_0(s))\Big)\epsilon_1(s)\,ds + 1 = \epsilon_1(t)$, thus it can not be a constant zero and $\int_0^{cR(\tau)} \exp\Big(g(\tilde{\Lambda}_0(s))\Big)(\epsilon_1(s))^2 ds$ is bounded away from 0 below. Hence, $A_1 \gtrsim \|\beta-\beta_0\|^2$.

For A_2, it is bounded below by

$$
\begin{aligned}
& P\{\Delta(\epsilon_2(U,V)[\gamma-\gamma_0] + \epsilon_3(U)[g-g_0])^2\} \\
\geq{}& P\{\Delta(\epsilon_2(U,V)[\gamma-\gamma_0])^2\} + P\{\Delta(\epsilon_3(U)[g-g_0])^2\} \\
& - 2|P\{\Delta(\epsilon_2(U,V)[\gamma-\gamma_0])(\epsilon_3(U)[g-g_0])\}| \\
\geq{}& P\{\Delta(\epsilon_2(U,V)[\gamma-\gamma_0])^2\} + P\{\Delta(\epsilon_3(U)[g-g_0])^2\} \\
& - 2\eta_2^{1/2} P\{\Delta\}(P\{\Delta(\epsilon_2(U,V)[\gamma-\gamma_0])^2\})^{1/2}(P\{\Delta(\epsilon_3(U)[g-g_0])^2\})^{1/2} \\
\geq{}& (1-\eta_2^{1/2}P\{\Delta\})\{P\{\Delta(\epsilon_2(U,V)[\gamma-\gamma_0])^2\} + P\{\Delta(\epsilon_3(U)[g-g_0])^2\}\} \\
\gtrsim{}& P\{\Delta(\epsilon_2(U,V)[\gamma-\gamma_0])^2\} + P\{\Delta(\epsilon_3(U)[g-g_0])^2\},
\end{aligned}
$$

where the second inequality holds under condition (C6) because there exists some

$\eta_2 \in (0,1)$ such that

$$(P\{\epsilon_2(U,Y)[\gamma-\gamma_0]\epsilon_3(U)[g-g_0]|\Delta=1\})^2 \leq \\ \eta_2 P\{(\epsilon_2(U,Y)[\gamma-\gamma_0])^2|\Delta=1\}P\{(\epsilon_3(U)[g-g_0])^2|\Delta=1\}.$$

Furthermore, the first term is bounded under condition (C3)

$$\begin{aligned} P\{\Delta(\epsilon_2(U,V)[\gamma-\gamma_0])^2\} &= P\{\Delta(\psi'_{0\gamma}(Y,X)[\gamma-\gamma_0])^2\} \\ &= P\{\int_0^\tau P\{\mathbb{1}(Y>t)|X\}(\psi'_{0\gamma}(t,X)[\gamma-\gamma_0])^2 d\Lambda_0(t,X)\} \\ &\geq \delta_0 P\{\int_0^\tau (\psi'_{0\gamma}(t,X)[\gamma-\gamma_0])^2 d\Lambda_0(t,X)\} \\ &\gtrsim \|\gamma-\gamma_0\|_2^2, \end{aligned}$$

where the second inequality is obtained by Lemma A.3.2 because $\gamma-\gamma_0 \in \Gamma^{p_1}$. Using the same argument, we have $P\{\Delta(\epsilon_3(U)[g-g_0])^2\} \gtrsim \|g-g_0\|_2^2$. Therefore,

$$\begin{aligned} Pl(\beta_0,\gamma_0,\zeta_0(\cdot,\beta_0,\gamma_0);W) - Pl(\beta,\gamma,\zeta(\cdot,\beta,\gamma);W) &= \frac{1}{2}A + o(A) \\ &\gtrsim \|\beta-\beta_0\|^2 + \|\gamma-\gamma_0\|_2^2 + \|g-g_0\|_2^2 \\ &\gtrsim d^2(\theta,\theta_0), \end{aligned}$$

which implies that

$$\inf_{d(\theta,\theta_0)\geq\epsilon, \theta\in\Theta_n} Pl(\beta_0,\gamma_0,\zeta_0(\cdot,\beta_0,\gamma_0);W) - Pl(\beta,\gamma,\zeta(\cdot,\beta,\gamma);W) \gtrsim \epsilon^2.$$

Hence the condition C1 in Shen and Wong (1994, Theorem 1) holds with $\alpha = 1$ in their notation.

Next, we verify the condition C2 in Shen and Wong (1994, Theorem 1). It follows that

$$\begin{aligned} &(l(\beta,\gamma,\zeta(\cdot,\beta,\gamma);W) - l(\beta_0,\gamma_0,\zeta_0(\cdot,\beta_0,\gamma_0);W))^2 \\ &= \{\Delta X^T(\beta-\beta_0) + \Delta[\gamma(Y)-\gamma_0(Y)] + \Delta[g(\Lambda(Y,X,\beta,\gamma,g)) - g_0(\Lambda_0(Y,X))] \\ &\quad - \int_0^Y [\exp\big(X^T\beta+\gamma(t)+g(\Lambda(t,X,\beta,\gamma,g))\big) - \exp\big(X^T\beta_0+\gamma_0(t)+g_0(\Lambda_0(t,X))\big)]dt\}^2 \end{aligned}$$

$$
\begin{aligned}
&\lesssim (X^T(\beta-\beta_0))^2 + \Delta(\gamma(Y)-\gamma_0(Y))^2 + \Delta[g(\Lambda(Y,X,\beta,\gamma,g)) - g_0(\Lambda_0(Y,X))]^2 \\
&\quad + \{\int_0^Y [\exp\bigl(X^T\beta+\gamma(t)+g(\Lambda(t,X,\beta,\gamma,g))\bigr) - \exp\bigl(X^T\beta_0+\gamma_0(t)+g_0(\Lambda_0(t,X))\bigr)]\,dt\}^2 \\
&\lesssim \|\beta-\beta_0\|^2 + \Delta(\gamma(Y)-\gamma_0(Y))^2 + \Delta[g(\Lambda(Y,X,\beta,\gamma,g)) - g_0(\Lambda_0(Y,X))]^2 \\
&\quad + \int_0^\tau [\exp\bigl(X^T\beta+\gamma(t)+g(\Lambda(t,X,\beta,\gamma,g))\bigr) - \exp\bigl(X^T\beta_0+\gamma_0(t)+g_0(\Lambda_0(t,X))\bigr)]^2\,dt,
\end{aligned}
\tag{A.31}
$$

where the second inequality is obtained by the condition (C2) and the Cauchy-Schwartz inequality

$$
\begin{aligned}
&\{\int_0^Y [\exp\bigl(X^T\beta+\gamma(t)+g(\Lambda(t,X,\beta,\gamma,g))\bigr) - \exp\bigl(X^T\beta_0+\gamma_0(t)+g_0(\Lambda_0(t,X))\bigr)]\,dt\}^2 \\
&= \{\int_0^\tau [1(Y\ge t)\exp\bigl(X^T\beta+\gamma(t)+g(\Lambda(t,X,\beta,\gamma,g))\bigr) \\
&\qquad - \exp\bigl(X^T\beta_0+\gamma_0(t)+g_0(\Lambda_0(t,X))\bigr)]\,dt\}^2 \\
&\le \int_0^\tau 1(Y\ge t)\,dt \int_0^\tau [\exp\bigl(X^T\beta+\gamma(t)+g(\Lambda(t,X,\beta,\gamma,g))\bigr) \\
&\qquad - \exp\bigl(X^T\beta_0+\gamma_0(t)+g_0(\Lambda_0(t,X))\bigr)]^2\,dt \\
&\le \tau \int_0^\tau [\exp\bigl(X^T\beta+\gamma(t)+g(\Lambda(t,X,\beta,\gamma,g))\bigr) - \exp\bigl(X^T\beta_0+\gamma_0(t)+g_0(\Lambda_0(t,X))\bigr)]^2\,dt.
\end{aligned}
$$

For the second term in (A.31), we have

$$
\begin{aligned}
&P\{\Delta(\gamma(Y)-\gamma_0(Y))^2\} \\
&= P\int_0^\tau 1(Y\ge t)\exp\bigl(X^T\beta_0+\gamma_0(t)+g_0(\Lambda_0(t,X))\bigr)(\gamma(t)-\gamma_0(t))^2\,dt \\
&\le \int_0^\tau P\{\exp\bigl(X^T\beta_0+\gamma_0(t)+g_0(\Lambda_0(t,X))\bigr)\}(\gamma(t)-\gamma_0(t))^2\,dt \\
&\lesssim \|\gamma-\gamma_0\|_2^2,
\end{aligned}
\tag{A.32}
$$

where the last inequality holds because $\exp\bigl(X^T\beta_0+\gamma_0(t)+g_0(\Lambda_0(t,X))\bigr)$ is bounded under conditions (C1)-(C4). For the third term in (A.31), we have

$$
\begin{aligned}
&P\{\Delta[g(\Lambda(Y,X,\beta,\gamma,g)) - g_0(\Lambda_0(Y,X))]^2\} \\
&= P\int_0^Y [g(\Lambda(t,X,\beta,\gamma,g)) - g_0(\Lambda_0(t,X))]^2 d\Lambda_0(t,X) \\
&= P\int_0^\tau 1(Y \geq t)[g(\Lambda(t,X,\beta,\gamma,g)) - g_0(\Lambda_0(t,X))]^2 d\Lambda_0(t,X) \\
&\leq P\int_0^\tau [g(\Lambda(t,X,\beta,\gamma,g)) - g_0(\Lambda_0(t,X))]^2 d\Lambda_0(t,X) \\
&= \|\zeta(\cdot,\beta,\gamma) - \zeta_0(\cdot,\beta_0,\gamma_0)\|_2^2, \qquad \text{(A.33)}
\end{aligned}
$$

For the fourth term in (A.31), using the mean value theorem, it follows that

$$
\begin{aligned}
&P\{\int_0^\tau [\exp\big(X^T\beta + \gamma(t) + g(\Lambda(t,X,\beta,\gamma,g))\big) - \exp\big(X^T\beta_0 + \gamma_0(t) + g_0(\Lambda_0(t,X))\big)]^2 dt\} \\
&= P\{\int_0^\tau \exp\Big(2\tilde{\psi}(t,X)\Big)[X^T(\beta-\beta_0) + (\gamma(t) - \gamma_0(t)) \\
&\qquad\qquad + g(\Lambda(t,X,\beta,\gamma,g)) - g_0(\Lambda_0(t,X))]^2 dt\} \\
&\lesssim P\{\int_0^\tau \exp\Big(2\tilde{\psi}(t,X)\Big)\{[X^T(\beta-\beta_0)]^2 + [\gamma(t) - \gamma_0(t)]^2 \\
&\qquad\qquad + [g(\Lambda(t,X,\beta,\gamma,g)) - g_0(\Lambda_0(t,X))]^2\} dt\} \\
&= I_1 + I_2 + I_3,
\end{aligned}
$$

where $\tilde{\psi}(t,X) = X^T\beta_0 + \gamma_0(t) + g_0(\Lambda_0(t,X)) + \xi(X^T(\beta-\beta_0) + \gamma(t) - \gamma_0(t) + g(\Lambda(t,X,\beta,\gamma,g))$ $-g_0(\Lambda_0(t,X)))$ for some $\xi \in (0,1)$ and is bounded under conditions (C1)-(C4). Hence,

$$I_1 \lesssim (\beta-\beta_0)^T P(XX^T)(\beta-\beta_0) \leq \lambda_d\|\beta-\beta_0\|^2,$$

where λ_d is the largest eigenvalue of $P(XX^T)$,

$$I_2 \lesssim \|\gamma - \gamma_0\|_2^2,$$

and

$$\begin{aligned}I_3 &= P\{\int_0^\tau \exp\Big(2\tilde{\psi}(t,X) - X^T\beta_0 - \gamma_0(t) - g_0(\Lambda_0(t,X))\Big) \\ &\qquad\qquad \cdot [g(\Lambda(t,X,\beta,\gamma,g)) - g_0(\Lambda_0(t,X))]^2 d\Lambda_0(t,X)\} \\ &\lesssim P\{\int_0^\tau [g(\Lambda(t,X,\beta,\gamma,g)) - g_0(\Lambda_0(t,X))]^2 d\Lambda_0(t,X)\} \\ &= \|\zeta(\cdot,\beta,\gamma) - \zeta_0(\cdot,\beta_0,\gamma_0)\|_2^2.\end{aligned}$$

Therefore, we have

$$\begin{aligned}P(l(\beta,\gamma,&\zeta(\cdot,\beta,\gamma);W) - l(\beta_0,\gamma_0,\zeta_0(\cdot,\beta_0,\gamma_0);W))^2 \\ &\lesssim \|\beta-\beta_0\|^2 + \|\gamma-\gamma_0\|_2^2 + \|\zeta(\cdot,\beta,\gamma) - \zeta_0(\cdot,\beta_0,\gamma_0)\|_2^2 \\ &\lesssim d^2(\theta,\theta_0),\end{aligned}$$

which implies that

$$\begin{aligned}&\sup_{d(\theta,\theta_0)\le\epsilon,\theta\in\Theta_n} Var\{l(\beta,\gamma,\zeta(\cdot,\beta,\gamma);W) - l(\beta_0,\gamma_0,\zeta_0(\cdot,\beta_0,\gamma_0);W)\} \\ \le &\sup_{d(\theta,\theta_0)\le\epsilon,\theta\in\Theta_n} P\{l(\beta,\gamma,\zeta(\cdot,\beta,\gamma);W) - l(\beta_0,\gamma_0,\zeta_0(\cdot,\beta_0,\gamma_0);W)\}^2 \lesssim \epsilon^2.\end{aligned}$$

Thus the condition C2 in Shen and Wong (1994, Theorem 1) holds with $\beta = 1$ in their notation.

Next we verify the condition C3 in Shen and Wong (1994, Theorem 1). By Lemma A.3.6, we have

$$\begin{aligned}H(\epsilon,\mathcal{F}_n,\|\cdot\|_\infty) = \log(N(\epsilon,\mathcal{F}_n,\|\cdot\|_\infty)) &\lesssim (c_1 q_{n_1} + c_2 q_{n_2} + d)\log(1/\epsilon) \\ &\lesssim n^{\max\{\nu_1,\nu_2\}}\log(1/\epsilon).\end{aligned}$$

So the condition C3 holds with constants $2r_0 = \max\{\nu_1,\nu_2\}$ and $r = 0^+$ in their notations, which leads to $\tau = \frac{1-\max\{\nu_1,\nu_2\}}{2} - \frac{\log\log n}{2\log n}$ in their main result. We can select slightly large $\tilde{\nu}_1$ and $\tilde{\nu}_2$ such that $\frac{1-\max\{\tilde{\nu}_1,\tilde{\nu}_2\}}{2} \le \frac{1-\max\{\nu_1,\nu_2\}}{2} - \frac{\log\log n}{2\log n}$ for sufficiently large n and still denote $\tilde{\nu}_i$ by ν_i for $i = 1, 2$. Then, $\tau = \frac{1-\max\{\nu_1,\nu_2\}}{2}$. Also, since the sieve estimator $\hat{\theta}_n$ maximizes the empirical log-likelihood over the sieve space Θ_n, the inequality (1.1) in Shen and Wong (1994) holds with $\eta_n = 0$. Therefore, by Theorem 1 in Shen and Wong (1994), we have

$$d(\hat{\theta}_n, \theta_0) = O_p(\max\{n^{-\frac{1-\max\{\nu_1,\nu_2\}}{2}}, d(\theta_{0n}, \theta_0), K^{1/2}(\theta_{0n}, \theta_0)\}),$$

where $K(\theta_{0n}, \theta_0) = P\{l(\theta_0; W) - l(\theta_{0n}; W)\}$. Further, using the Taylor expansion for $P\{l(\theta_0; W) - l(\theta_{0n}; W)\}$ in (A.22), we have

$$\begin{aligned}
K(\theta_{0n}, \theta_0) &= \frac{1}{2} P\{\Delta[\gamma_0(Y) + g_0(\Lambda(Y, X, \beta_0, \gamma_0, g_0) \\
&\qquad - \gamma_{0n}(Y) - g_{0n}(\Lambda(Y, X, \beta_0, \gamma_{0n}, g_{0n}))]^2\} + o(d^2(\theta_{0n}, \theta_0)) \\
&\leq P\{\Delta[g_0(\Lambda(Y, X, \beta_0, \gamma_0, g_0)) - g_{0n}(\Lambda(Y, X, \beta_0, \gamma_{0n}, g_{0n}))]^2\} \\
&\quad + P\{\Delta(\gamma_0(Y) - \gamma_{0n}(Y))^2\} + o(d^2(\theta_{0n}, \theta_0)) \\
&\lesssim \|\zeta_0(\cdot, \beta_0, \gamma_0) - \zeta_{0n}(\cdot, \beta_{0n}, \gamma_{0n})\|_2^2 + \|\gamma_0 - \gamma_{0n}\|_2^2 + o(d^2(\theta_{0n}, \theta_0)) \\
&= O(d^2(\theta_{0n}, \theta_0)),
\end{aligned}$$

where the first inequality is obtained by the fact $(a+b)^2 \leq 2(a^2+b^2)$ and the second inequality holds by using the same argument as in (A.32) and (A.33). Moreover, $d^2(\theta_{0n}, \theta_0) \lesssim \|\gamma_0 - \gamma_{0n}\|_2^2 + \|g_0 - g_{0n}\|_2^2 \lesssim \|\gamma_0 - \gamma_{0n}\|_\infty^2 + \|g_0 - g_{0n}\|_\infty^2 = O(n^{-2\min\{p_1\nu_1, p_2\nu_2\}})$ due to inequality (A.23) and Lemma A.3.5. Thus, we have

$$d(\hat{\theta}_n, \theta_0) = O_p(\max\{n^{-\frac{1-\max\{\nu_1,\nu_2\}}{2}}, n^{-\min\{p_1\nu_1, p_2\nu_2\}}\}) = O_p(n^{-\min\{p_1\nu_1, p_2\nu_2, \frac{1-\max\{\nu_1,\nu_2\}}{2}\}}),$$

which completes the proof. □

A.3.3 Proof of Theorem 2.3.2

Proof of Theorem 2.3.2. We prove the theorem by verifying assumptions (A1)-(A6) in Appendix A.2. By Theorem 2.3.1 we know that assumption (A1) holds with $\xi = \min\{p_1\nu_1, p_2\nu_2, \frac{1-\max\{\nu_1,\nu_2\}}{2}\}$. It is straightforward to show that assumption (A2) holds based on the fact that score functions have zero mean. To verify assumption (A3), first, we will find $\mathbf{v}^* = (v_1^*, \cdots, v_d^*)'$ and $\mathbf{h}^* = (h_1^*, \cdots, h_d^*)'$ with $\mathbf{h}^*(\cdot) = \mathbf{w}^*(\Lambda_0(\cdot)) + g_0'(\Lambda_0(\cdot))\Lambda_{0g}'(\cdot)[\mathbf{w}^*]$ such that for any $v \in \mathbb{V}$ and $h \in \mathbb{H}$ with

$h(\cdot) = w(\Lambda_0(\cdot)) + g_0'(\Lambda_0(\cdot))\Lambda_{0g}'(\cdot)[w]$,

$$S_{\beta\gamma}''(\beta_0, \gamma_0(\cdot), \zeta_0(\cdot, \beta_0, \gamma_0))[v] = S_{\gamma\gamma}''(\beta_0, \gamma_0(\cdot), \zeta_0(\cdot, \beta_0, \gamma_0))[\mathbf{v}^*, v] + S_{\zeta\gamma}''(\beta_0, \gamma_0(\cdot), \zeta_0(\cdot, \beta_0, \gamma_0))[\mathbf{h}^*, v], \qquad \text{(A.34)}$$

$$S_{\beta\zeta}''(\beta_0, \gamma_0(\cdot), \zeta_0(\cdot, \beta_0, \gamma_0))[h] = S_{\gamma\zeta}''(\beta_0, \gamma_0(\cdot), \zeta_0(\cdot, \beta_0, \gamma_0))[\mathbf{v}^*, h] + S_{\zeta\zeta}''(\beta_0, \gamma_0(\cdot), \zeta_0(\cdot, \beta_0, \gamma_0))[\mathbf{h}^*, h]. \qquad \text{(A.35)}$$

By Lemma A.3.4 and the property $P\{\int_0^Y f(t,X)\,d\Lambda_0(t,X)\} = P\{\Delta f(Y,X)\}$, for any $\mathbf{v} \in \mathbb{V}^d, v \in \mathbb{V}$ and $\mathbf{h} \in \mathbb{H}^d$ with $\mathbf{h}(\cdot) = \mathbf{w}(\Lambda_0(\cdot)) + g_0'(\Lambda_0(\cdot))\Lambda_{0g}'(\cdot)[\mathbf{w}]$, we have

$$\begin{aligned}
&S_{\beta\gamma}''(\beta_0, \gamma_0, \zeta_0)[v] - S_{\gamma\gamma}''(\beta_0, \gamma_0, \zeta_0)[\mathbf{v}, v] - S_{\zeta\gamma}''(\beta_0, \gamma_0, \zeta_0)[\mathbf{h}, v] \\
=&P\{l_{\beta\gamma}''(\beta_0, \gamma_0, \zeta_0; W)[v] - l_{\gamma\gamma}''(\beta_0, \gamma_0, \zeta_0; W)[\mathbf{v}, v] - l_{\zeta\gamma}''(\beta_0, \gamma_0, \zeta_0; W)[\mathbf{h}, v]\} \\
=&P\{\Delta \left[g_0'(\Lambda_0(Y,X))\Lambda_{0\beta}'(Y,X) + X - g_0'(\Lambda_0(Y,X))\Lambda_{0\gamma}'(Y,X)[\mathbf{v}] - \mathbf{v}(Y)\right. \\
&\quad \left. - g_0'(\Lambda_0(Y,X))\Lambda_{0g}'(Y,X)^T[\mathbf{w}] - \mathbf{w}(\Lambda_0(Y,X))\right] \left(g_0'(\Lambda_0(Y,X))\Lambda_{0\gamma}'(Y,X)[v] + v(Y)\right)\} \\
=&P\{\Delta \left(\epsilon_1(U)X - \epsilon_2(U,V)[\mathbf{v}] - \epsilon_3(U)[\mathbf{w}]\right) \psi_{0\gamma}'(Y,X)[v]\}, \qquad \text{(A.36)}
\end{aligned}$$

where the last equality holds with $\epsilon_1, \epsilon_2, \epsilon_3, \psi_{0\gamma}'$ given in (A.25)-(A.29) and U given in the condition (C5). Similarly, for any $\mathbf{v} \in \mathbb{V}^d$, $\mathbf{h} \in \mathbb{H}^d$ and $h \in \mathbb{H}$ with $h(\cdot) = w(\Lambda_0(\cdot)) + g_0'(\Lambda_0(\cdot))\Lambda_{0g}'(\cdot)[w]$, we have

$$\begin{aligned}
&S_{\beta\zeta}''(\beta_0, \gamma_0, \zeta_0)[h] - S_{\gamma\zeta}''(\beta_0, \gamma_0, \zeta_0)[\mathbf{v}, h] - S_{\zeta\zeta}''(\beta_0, \gamma_0, \zeta_0)[\mathbf{h}, h] \\
=&P\{\Delta \left(\epsilon_1(U)X - \epsilon_2(U,V)[\mathbf{v}] - \epsilon_3(U)[\mathbf{w}]\right) \psi_{0g}'(Y,X)[w]\}. \qquad \text{(A.37)}
\end{aligned}$$

Note that under condition (C7), there exists $\mathbf{v}^* = (v_1^*, \cdots, v_d^*)^T$ and $\mathbf{w}^* = (w_1^*, \cdots, w_d^*)^T$, where $v_j^* \in \Gamma^2$ and $w_j^* \in \mathcal{G}^2$ for $j = 1, \cdots, d$, such that $P\{\Delta\mathbf{A}^*(U,X)\psi_{0\gamma}'(Y,X)[v]\} = 0$ and $P\{\Delta\mathbf{A}^*(U,X)\psi_{0g}'(Y,X)[w]\} = 0$ hold for any $v \in \Gamma^{p_1}$ and $w \in \mathcal{G}^{p_2}$. Since $\mathbf{A}^*(U,X) = \epsilon_1(U)X - \epsilon_2(U,V)[\mathbf{v}^*] - \epsilon_3(U)[\mathbf{w}^*]$, plugging $\mathbf{v} = \mathbf{v}^*$ in (A.36) and $\mathbf{w} = \mathbf{w}^*$ in (A.37) we have equations (A.34) and (A.35) hold with $\mathbf{v}^*$ and $\mathbf{w}^*$ given in condition (C7). Then it follows that

$$l_\beta'(\beta_0, \gamma_0, \zeta_0; W) - l_\gamma'(\beta_0, \gamma_0, \zeta_0; W)[\mathbf{v}^*] - l_\zeta'(\beta_0, \gamma_0, \zeta_0; W)[\mathbf{h}^*(\cdot, \beta_0, \gamma_0)]$$

$$
\begin{aligned}
&= \Delta \mathbf{A}^*(U, X) - \int_0^Y \mathbf{A}^*(R(t)e^{X^T\beta_0}, X)\, d\Lambda_0(t, X) \\
&= \Delta \mathbf{A}^*(U, X) - \int_0^{R(Y)e^{X^T\beta_0}} \mathbf{A}^*(t, X)\, d\tilde{\Lambda}_0(t) \\
&= \int \mathbf{A}^*(t, X)\, dM(t) = \boldsymbol{l}^*(\beta_0, \gamma_0, \zeta_0; W),
\end{aligned}
$$

with $M(t)$ and $\boldsymbol{l}^*$ given in condition (C8). Based on the zero-mean property of score function together with the facts in (A.34) and (A.35), the matrix A in assumption (A3) is given by

$$
\begin{aligned}
A = &-S''_{\beta\beta}(\beta_0, \gamma_0, \zeta_0) + S''_{\gamma\beta}(\beta_0, \gamma_0, \zeta_0)[\mathbf{v}^*] + S''_{\zeta\beta}(\beta_0, \gamma_0, \zeta_0)[\mathbf{h}^*] \\
&- S''_{\gamma\gamma}(\beta_0, \gamma_0, \zeta_0)[\mathbf{v}^*, \mathbf{v}^*] + S''_{\beta\gamma}(\beta_0, \gamma_0, \zeta_0)[\mathbf{v}^*] - S''_{\zeta\gamma}(\beta_0, \gamma_0, \zeta_0)[\mathbf{h}^*, \mathbf{v}^*] \\
&- S''_{\zeta\zeta}(\beta_0, \gamma_0, \zeta_0)[\mathbf{h}^*, \mathbf{h}^*] + S''_{\beta\zeta}(\beta_0, \gamma_0, \zeta_0)[\mathbf{h}^*] - S''_{\gamma\zeta}(\beta_0, \gamma_0, \zeta_0)[\mathbf{v}^*, \mathbf{h}^*] \\
= &\, P\{(l'_\beta(\beta_0, \gamma_0, \zeta_0; W) - l'_\gamma(\beta_0, \gamma_0, \zeta_0; W)[\mathbf{v}^*] - l'_\zeta(\beta_0, \gamma_0, \zeta_0; W)[\mathbf{h}^*])^{\otimes 2}\} \\
= &\, P\{\boldsymbol{l}^*(\beta_0, \gamma_0, \zeta_0; W)^{\otimes 2}\},
\end{aligned}
$$

which is the information matrix for β_0 and is nonsingular under condition (C8). Thus, assumption (A3) holds.

To verify assumption (A4), we first note that the first part holds because $\hat{\beta}_n$ satisfies $S'_{\beta,n}(\hat{\theta}_n) = 0$ where $\hat{\theta}_n = (\hat{\beta}_n, \hat{\gamma}_n(\cdot), \hat{\zeta}_n(\cdot, \hat{\beta}_n, \hat{\gamma}_n))$. Next we need to show that $S'_{\gamma,n}(\hat{\theta}_n)[v_j^*] = o_p(n^{-1/2})$. Since $v_j^* \in \Gamma^2$, by Lemma A.3.5 there exists $v_{j,n}^* \in \Gamma_n^2$ such that $\|v_{j,n}^* - v_j^*\|_\infty = O(n^{-2\nu_1})$. Based on the fact that $v_{j,n}^*$ can be written as the linear combination of basis functions B_k^1 for $k = 1, \ldots, q_n^1$, we have $S'_{\gamma,n}(\hat{\theta}_n)[v_{j,n}^*] = 0$.[1] Since

[1] Note that we constrain the parameter $\gamma(t^*) = 0$ for identifiability guarantee. For any $\gamma \in \Gamma_n^{p_1}$ in the sieve space, the constraint can be achieved by fixing the coefficient of one specific B-spline basis (suppose it is indexed as the first basis and let $a_1 \equiv 0$) and leaving coefficients of other bases as free optimization parameters. Since $\hat{\theta}_n$ maximizes $l_n(\theta)$ in the sieve space and $v_{j,n}^* \in \Gamma_n^2$ can be written as the linear combination of bases with the first coefficient a_1 fixed as 0, we have the gradient of $l_n(\theta)$ with respect to γ along the direction $v_{j,n}^*$ at $\hat{\theta}_n$ equal to zero, i.e., $S'_{\gamma,n}(\hat{\theta}_n)[v_{j,n}^*] = 0$.

$S'_\gamma(\beta_0, \gamma_0(\cdot), \zeta_0(\cdot, \beta_0, \gamma_0))[v_j^* - v_{j,n}^*] = 0$, it suffices to show that for each $1 \le j \le d$,

$$P\{l'_\gamma(\hat{\theta}_n; W)[v_j^* - v_{j,n}^*] - l'_\gamma(\theta_0; W)[v_j^* - v_{j,n}^*]\} + (\mathbb{P}_n - P)\{l'_\gamma(\hat{\theta}_n; W)[v_j^* - v_{j,n}^*]\}$$
$$= I_{1n} + I_{2n} = o_p(n^{-1/2}).$$

We will first show that I_{1n} is $o_p(n^{-1/2})$. Using the Taylor expansion for $l'_\gamma(\hat{\theta}_n)[v_j^* - v_{j,n}^*]$ at θ_0, we have

$$\begin{aligned} I_{1n} =& P\{(\hat{\beta}_n - \beta_0)^T l''_{\beta\gamma}(\tilde{\beta}_n, \tilde{\gamma}_n(\cdot), \tilde{\zeta}_n(\cdot, \tilde{\beta}_n, \tilde{\gamma}_n); W)[v_j^* - v_{j,n}^*] \\ &+ l''_{\gamma\gamma}(\tilde{\beta}_n, \tilde{\gamma}_n(\cdot), \tilde{\zeta}_n(\cdot, \tilde{\beta}_n, \tilde{\gamma}_n); W)[v_j^* - v_{j,n}^*, \hat{\gamma}_n - \gamma_0] \\ &+ l''_{\gamma\zeta}(\tilde{\beta}_n, \tilde{\gamma}_n(\cdot), \tilde{\zeta}_n(\cdot, \tilde{\beta}_n, \tilde{\gamma}_n); W)[v_j^* - v_{j,n}^*, \hat{\zeta}_n - \zeta_0]\}, \end{aligned}$$

where $(\tilde{\beta}_n, \tilde{\gamma}_n(\cdot), \tilde{\zeta}_n(\cdot, \tilde{\beta}_n, \tilde{\gamma}_n))$ is some point between θ_0 and $\hat{\theta}_n$. Let $\tilde{\Lambda}(t,x) = \Lambda(t, x, \tilde{\beta}_n, \tilde{\gamma}_n, \tilde{g}_n)$. Note that by solving initial value problems in Lemma A.3.1, we have $\tilde{\Lambda}'_\beta(t,x)$ and $\tilde{\Lambda}''_{\beta\beta}(t,x)$ are bounded on $t \in [0,\tau]$ and $x \in \mathcal{X}$ based on the boundedness of $\tilde{\gamma}_n$, $\tilde{g}_n$, $\tilde{g}'_n$ and $\tilde{g}''_n$. Also, we have $\|\tilde{\Lambda}'_\gamma(\cdot)[v]\|_\infty \lesssim \|v\|_\infty$ and $\sup_{t\in[0,\tau], x\in\mathcal{X}} \|\tilde{\Lambda}''_{\beta\gamma}(t,x)[v]\| \lesssim \|v\|_\infty$. It follows that

$$\begin{aligned} &\sup_{t\in[0,\tau], x\in\mathcal{X}} \|\tilde{\zeta}''_{\beta\gamma}(t, x, \tilde{\beta}_n, \tilde{\gamma}_n)[v_j^* - v_{j,n}^*]\| \\ =& \sup_{t\in[0,\tau], x\in\mathcal{X}} \|\tilde{g}''_n(\tilde{\Lambda}(t,x))\tilde{\Lambda}'_\beta(t,x)\tilde{\Lambda}'_\gamma(t,x)[v_j^* - v_{j,n}^*] + \tilde{g}'_n(\tilde{\Lambda}(t,X))\tilde{\Lambda}''_{\beta\gamma}(t,x)[v_j^* - v_{j,n}^*]\| \\ \lesssim& \|v_j^* - v_{j,n}^*\|_\infty, \end{aligned}$$

and

$$\begin{aligned} &P\{\|l''_{\beta\gamma}(\tilde{\beta}_n, \tilde{\gamma}_n(\cdot), \tilde{\zeta}_n(\cdot, \tilde{\beta}_n, \tilde{\gamma}_n); W)[v_j^* - v_{j,n}^*]\|\} \\ =& P\Big\{\Big\| \int_0^\tau \tilde{\zeta}''_{\beta\gamma}(t, X, \tilde{\beta}_n, \tilde{\gamma}_n)[v_j^* - v_{j,n}^*] \mathbb{1}(Y \ge t)\, d\Lambda_0(t, X) \\ &\quad - \int_0^\tau \exp\Big(X^T \tilde{\beta}_n + \tilde{\gamma}_n(t) + \tilde{g}_n(\tilde{\Lambda}(t,X))\Big) \\ &\qquad \cdot \{(v_j^*(t) - v_{j,n}^*(t) + \tilde{g}'_n(\tilde{\Lambda}(t,X))\tilde{\Lambda}'_\gamma(t,X)[v_j^* - v_{j,n}^*])(X + \tilde{g}'_n(\tilde{\Lambda}(t,X))\tilde{\Lambda}'_\beta(t,X)) \\ &\qquad\quad + \tilde{\zeta}''_{\beta\gamma}(t, X, \tilde{\beta}_n, \tilde{\gamma}_n)[v_j^* - v_{j,n}^*]\} \mathbb{1}(Y \ge t)\, dt \Big\|\Big\} \end{aligned}$$

$$\lesssim \sup_{t\in[0,\tau],x\in\mathcal{X}} \|\tilde{\zeta}''_{\beta\gamma}(t,x,\tilde{\beta}_n,\tilde{\gamma}_n)[v_j^*-v_{j,n}^*]\| + \|v_j^*-v_{j,n}^*\|_\infty \lesssim \|v_j^*-v_{j,n}^*\|_\infty.$$

Therefore, the first term in I_{1n} is dominated by

$$\begin{aligned}
&P\{|(\hat{\beta}_n-\beta_0)^T l''_{\beta\gamma}(\tilde{\beta}_n,\tilde{\gamma}_n(\cdot),\tilde{\zeta}_n(\cdot,\tilde{\beta}_n,\tilde{\gamma}_n);W)[v_j^*-v_{j,n}^*]|\}\\
\le&\|\hat{\beta}_n-\beta_0\|P\{\|l''_{\beta\gamma}(\tilde{\beta}_n,\tilde{\gamma}_n(\cdot),\tilde{\zeta}_n(\cdot,\tilde{\beta}_n,\tilde{\gamma}_n);W)[v_j^*-v_{j,n}^*]\|\}\\
\lesssim&\|\hat{\beta}_n-\beta_0\|\|v_j^*-v_{j,n}^*\|_\infty \le d(\hat{\theta}_n,\theta_0)\|v_j^*-v_{j,n}^*\|_\infty\\
=&O_p(n^{-\min\{p_1\nu_1,p_2\nu_2,\frac{1-\max\{\nu_1,\nu_2\}}{2}\}})\cdot O(n^{-2\nu_1})\\
=&O_p(n^{-\min\{(p_1+2)\nu_1,p_2\nu_2+2\nu_1,\frac{1-\max\{\nu_1,\nu_2\}}{2}+2\nu_1\}}).
\end{aligned}$$

By solving initial value problems in (A.10) and (A.13) and the Cauchy-Schwarz inequality (similar arguments are used in Lemma A.3.2 to prove that linear operators are bounded above), we have $\|\tilde{\Lambda}'_\gamma(\cdot)[v]\|_2 \lesssim \|v\|_2$ and $\|\tilde{\Lambda}''_{\gamma\gamma}(\cdot)[\nu_1,\nu_2]\|_2 \lesssim \|\nu_1\|_\infty\|\nu_2\|_2$. It follows that

$$\begin{aligned}
&\|\tilde{\zeta}''_{\gamma\gamma}(\cdot,\tilde{\beta}_n,\tilde{\gamma}_n)[v_j^*-v_{j,n}^*,\hat{\gamma}_n-\gamma_0]\|_2\\
=&\|\tilde{g}''_n(\tilde{\Lambda}(\cdot))\tilde{\Lambda}'_\gamma(\cdot)[\hat{\gamma}_n-\gamma_0]\tilde{\Lambda}'_\gamma(\cdot)[v_j^*-v_{j,n}^*]+\tilde{g}'_n(\tilde{\Lambda}(\cdot))\tilde{\Lambda}''_{\gamma\gamma}(\cdot)[v_j^*-v_{j,n}^*,\hat{\gamma}_n-\gamma_0]\|_2\\
\lesssim&\|\tilde{\Lambda}'_\gamma(\cdot)[v_j^*-v_{j,n}^*]\|_\infty\|\tilde{\Lambda}'_\gamma(\cdot)[\hat{\gamma}_n-\gamma_0]\|_2+\|\tilde{\Lambda}''_{\gamma\gamma}(\cdot)[v_j^*-v_{j,n}^*,\hat{\gamma}_n-\gamma_0]\|_2\\
\lesssim&\|v_j^*-v_{j,n}^*\|_\infty\cdot\|\hat{\gamma}_n-\gamma_0\|_2,
\end{aligned}$$

and by the Cauchy-Schwarz inequality the second term in I_{1n} is bounded by

$$\begin{aligned}
&(P\{\big|l''_{\gamma\gamma}(\tilde{\beta}_n,\tilde{\gamma}_n(\cdot),\tilde{\zeta}_n(\cdot,\tilde{\beta}_n,\tilde{\gamma}_n);W)[v_j^*-v_{j,n}^*,\hat{\gamma}_n-\gamma_0]\big|\})^2\\
\le&P\{\big|l''_{\gamma\gamma}(\tilde{\beta}_n,\tilde{\gamma}_n(\cdot),\tilde{\zeta}_n(\cdot,\tilde{\beta}_n,\tilde{\gamma}_n);W)[v_j^*-v_{j,n}^*,\hat{\gamma}_n-\gamma_0]\big|^2\}\\
=&P\Big\{\Big|\Delta\tilde{\zeta}''_{\gamma\gamma}(Y,X,\tilde{\beta}_n,\tilde{\gamma}_n)[v_j^*-v_{j,n}^*,\hat{\gamma}_n-\gamma_0]\\
&-\int_0^\tau \mathbb{1}(Y\ge t)\exp\Big(X^T\tilde{\beta}_n+\tilde{\gamma}_n(t)+\tilde{g}_n(\tilde{\Lambda}(t,X))\Big)\cdot\{\tilde{\zeta}''_{\gamma\gamma}(t,X,\tilde{\beta}_n,\tilde{\gamma}_n)[v_j^*-v_{j,n}^*,\hat{\gamma}_n-\gamma_0]\\
&\quad+((v_j^*-v_{j,n}^*)(t)+\tilde{g}'_n(\tilde{\Lambda}(t,X))\tilde{\Lambda}'_\gamma(t,X)[v_j^*-v_{j,n}^*])\\
&\quad\cdot((\hat{\gamma}_n-\gamma_0)(t)+\tilde{g}'_n(\tilde{\Lambda}(t,X))\tilde{\Lambda}'_\gamma(t,X)[\hat{\gamma}_n-\gamma_0])\}dt\Big|^2\Big\}\\
\lesssim&\|\tilde{\zeta}''_{\gamma\gamma}(\cdot,\tilde{\beta}_n,\tilde{\gamma}_n)[v_j^*-v_{j,n}^*,\hat{\gamma}_n-\gamma_0]\|_2^2+\|v_j^*-v_{j,n}^*\|_\infty^2\cdot(\|\hat{\gamma}_n-\gamma_0\|_2^2+\|\tilde{\Lambda}'_\gamma(\cdot)[\hat{\gamma}_n-\gamma_0]\|_2^2)
\end{aligned}$$

$$\lesssim \|v_j^* - v_{j,n}^*\|_\infty^2 \cdot \|\hat{\gamma}_n - \gamma_0\|_2^2 \leq \|v_j^* - v_{j,n}^*\|_\infty^2 \cdot d^2(\hat{\theta}_n, \theta_0).$$

So we have

$$\begin{aligned} &P\{\big|l''_{\gamma\gamma}(\tilde{\beta}_n, \tilde{\gamma}_n(\cdot), \tilde{\zeta}_n(\cdot, \tilde{\beta}_n, \tilde{\gamma}_n); W)[v_j^* - v_{j,n}^*, \hat{\gamma}_n - \gamma_0]\big|\} \\ &\qquad = O_p(n^{-\min\{(p_1+2)\nu_1, p_2\nu_2+2\nu_1, \frac{1-\max\{\nu_1,\nu_2\}}{2}+2\nu_1\}}). \end{aligned}$$

Also, by subtracting and adding some terms and using $\|a+b\|_2 \leq \|a\|_2 + \|b\|_2$, we have

$$\begin{aligned} &\|\hat{\zeta}'_{n,\gamma}(\cdot, \hat{\beta}_n, \hat{\gamma}_n)[v_j^* - v_{j,n}^*] - \zeta'_{0\gamma}(\cdot, \beta_0, \gamma_0)[v_j^* - v_{j,n}^*]\|_2 \\ =&\|\hat{g}'_n(\Lambda(\cdot, \hat{\beta}_n, \hat{\gamma}_n, \hat{g}_n))\Lambda'_\gamma(\cdot, \hat{\beta}_n, \hat{\gamma}_n, \hat{g}_n)[v_j^* - v_{j,n}^*] - g'_0(\Lambda_0(\cdot))\Lambda'_{0\gamma}(\cdot)[v_j^* - v_{j,n}^*]\|_2 \\ \leq&\|\hat{g}'_n(\Lambda(\cdot, \hat{\beta}_n, \hat{\gamma}_n, \hat{g}_n))\Lambda'_\gamma(\cdot, \hat{\beta}_n, \hat{\gamma}_n, \hat{g}_n)[v_j^* - v_{j,n}^*] - g'_0(\Lambda_0(\cdot))\Lambda'_\gamma(\cdot, \hat{\beta}_n, \hat{\gamma}_n, \hat{g}_n)[v_j^* - v_{j,n}^*]\|_2 \\ &+ \|g'_0(\Lambda_0(\cdot))\Lambda'_\gamma(\cdot, \hat{\beta}_n, \hat{\gamma}_n, \hat{g}_n)[v_j^* - v_{j,n}^*] - g'_0(\Lambda_0(\cdot))\Lambda'_\gamma(\cdot, \beta_0, \hat{\gamma}_n, \hat{g}_n)[v_j^* - v_{j,n}^*]\|_2 \\ &+ \|g'_0(\Lambda_0(\cdot))\Lambda'_\gamma(\cdot, \beta_0, \hat{\gamma}_n, \hat{g}_n)[v_j^* - v_{j,n}^*] - g'_0(\Lambda_0(\cdot))\Lambda'_\gamma(\cdot, \beta_0, \gamma_0, \hat{g}_n)[v_j^* - v_{j,n}^*]\|_2 \\ &+ \|g'_0(\Lambda_0(\cdot))\Lambda'_\gamma(\cdot, \beta_0, \gamma_0, \hat{g}_n)[v_j^* - v_{j,n}^*] - g'_0(\Lambda_0(\cdot))\Lambda'_{0\gamma}(\cdot)[v_j^* - v_{j,n}^*]\|_2 \\ =&J_1 + J_2 + J_3 + J_4. \end{aligned}$$

For J_1, since $\hat{\gamma}_n$, $\hat{g}_n$ and $\hat{g}'_n$ are bounded, we have $\|\Lambda'_\gamma(\cdot, \hat{\beta}_n, \hat{\gamma}_n, \hat{g}_n)[v_j^* - v_{j,n}^*]\|_\infty \lesssim \|v_j^* - v_{j,n}^*\|_\infty$ and it follows that

$$\begin{aligned} J_1 &\leq \|\hat{g}'_n(\Lambda(\cdot, \hat{\beta}_n, \hat{\gamma}_n, \hat{g}_n)) - g'_0(\Lambda_0(\cdot))\|_2 \cdot \|\Lambda'_\gamma(\cdot, \hat{\beta}_n, \hat{\gamma}_n, \hat{g}_n)[v_j^* - v_{j,n}^*]\|_\infty \\ &\lesssim \|\hat{g}'_n(\Lambda(\cdot, \hat{\beta}_n, \hat{\gamma}_n, \hat{g}_n)) - g'_0(\Lambda_0(\cdot))\|_2 \cdot \|v_j^* - v_{j,n}^*\|_\infty \\ &= O_p(n^{-\min\{p_1\nu_1, (p_2-1)\nu_2, \frac{1-\max\{\nu_1,\nu_2\}}{2}\}}) \cdot O(n^{-2\nu_1}) \\ &= O_p(n^{-\min\{(p_1+2)\nu_1, (p_2-1)\nu_2+2\nu_1, \frac{1-\max\{\nu_1,\nu_2\}}{2}+2\nu_1\}}), \end{aligned}$$

where the third equality holds based on the same argument of Ding and Nan (2011) on their page 3058. For J_2, by using the mean value theorem, it follows that

$$\begin{aligned} J_2 &= \|g'_0(\Lambda_0(\cdot))(\Lambda''_{\gamma\beta}(\cdot, \tilde{\beta}_n, \hat{\gamma}_n, \hat{g}_n)[v_j^* - v_{j,n}^*])^T(\hat{\beta}_n - \beta_0)\|_2 \\ &\lesssim \|\Lambda''_{\gamma\beta}(\cdot, \tilde{\beta}_n, \hat{\gamma}_n, \hat{g}_n)[v_j^* - v_{j,n}^*]\|_2\|\hat{\beta}_n - \beta_0\| \\ &\lesssim \|v_j^* - v_{j,n}^*\|_\infty \cdot \|\hat{\beta}_n - \beta_0\| \end{aligned}$$

$$= O_p(n^{-\min\{(p_1+2)\nu_1, p_2\nu_2+2\nu_1, \frac{1-\max\{\nu_1,\nu_2\}}{2}+2\nu_1\}}),$$

where $\tilde{\beta}_n$ is a point between $\hat{\beta}_n$ and β_0, the second inequality is based on the boundedness of g_0', and the third inequality is obtained by solving the initial value problem in (A.14) along with the boundedness of $\hat{\gamma}_n$, $\hat{g}_n$, $\hat{g}_n'$, $\hat{g}_n''$ and $\Lambda_\beta'(\cdot, \tilde{\beta}_n, \hat{\gamma}_n, \hat{g}_n)$. By a similar argument that we used for the second term in I_{1n}, we have for J_3,

$$\begin{aligned}J_3 &= \|g_0'(\Lambda_0(\cdot))\Lambda_{\gamma\gamma}''(\cdot, \beta_0, \tilde{\gamma}_n, \hat{g}_n)[v_j^* - v_{j,n}^*, \hat{\gamma}_n - \gamma_0]\|_2 \\ &\lesssim \|v_j^* - v_{j,n}^*\|_\infty \cdot \|\hat{\gamma}_n - \gamma_0\|_2 = O_p(n^{-\min\{(p_1+2)\nu_1, p_2\nu_2+2\nu_1, \frac{1-\max\{\nu_1,\nu_2\}}{2}+2\nu_1\}}),\end{aligned}$$

and for J_4

$$\begin{aligned}J_4 &= \|g_0'(\Lambda_0(\cdot))\Lambda_{\gamma g}''(\cdot, \beta_0, \gamma_0, \tilde{g}_n)[v_j^* - v_{j,n}^*, \hat{g}_n - g_0]\|_2 \\ &\lesssim \|\Lambda_{\gamma g}''(\cdot, \beta_0, \gamma_0, \tilde{g}_n)[v_j^* - v_{j,n}^*, \hat{g}_n - g_0]\|_2 \\ &\lesssim (\|\hat{g}_n(\Lambda(\cdot, \hat{\beta}_n, \hat{\gamma}_n, \hat{g}_n)) - g_0(\Lambda_0(\cdot))\|_2 + \|\hat{g}_n'(\Lambda(\cdot, \hat{\beta}_n, \hat{\gamma}_n, \hat{g}_n)) - g_0'(\Lambda_0(\cdot))\|_2) \cdot \|v_j^* - v_{j,n}^*\|_\infty \\ &= O_p(n^{-\min\{p_1\nu_1, (p_2-1)\nu_2, \frac{1-\max\{\nu_1,\nu_2\}}{2}\}}) \cdot O(n^{-2\nu_1}),\end{aligned}$$

where $\tilde{\gamma}_n$ is a point between $\hat{\gamma}_n$ and γ_0 and $\tilde{g}_n$ is a point between $\hat{g}_n$ and g_0. Thus, we have

$$\begin{aligned}&\|\hat{\zeta}_{n,\gamma}'(\cdot, \hat{\beta}_n, \hat{\gamma}_n)[v_j^* - v_{j,n}^*] - \zeta_{0\gamma}'(\cdot, \beta_0, \gamma_0)[v_j^* - v_{j,n}^*]\|_2 \\ \lesssim &O_p(n^{-\min\{(p_1+2)\nu_1, (p_2-1)\nu_2+2\nu_1, \frac{1-\max\{\nu_1,\nu_2\}}{2}+2\nu_1\}}),\end{aligned}$$

and it follows that for the third term in I_{1n} is bounded by

$$\begin{aligned}&(P\{\left|l_{\gamma\zeta}''(\tilde{\beta}_n, \tilde{\gamma}_n(\cdot), \tilde{\zeta}_n(\cdot, \tilde{\beta}_n, \tilde{\gamma}_n); W)[v_j^* - v_{j,n}^*, \hat{\zeta}_n - \zeta_0]\right|\})^2 \\ \leq &P\{\left|l_{\gamma\zeta}''(\tilde{\beta}_n, \tilde{\gamma}_n(\cdot), \tilde{\zeta}_n(\cdot, \tilde{\beta}_n, \tilde{\gamma}_n); W)[v_j^* - v_{j,n}^*, \hat{\zeta}_n - \zeta_0]\right|^2\} \\ =&P\Big\{\Big|\Delta(\hat{\zeta}_{n,\gamma}'(Y, X, \hat{\beta}_n, \hat{\gamma}_n)[v_j^* - v_{j,n}^*] - \zeta_{0\gamma}'(Y, X, \beta_0, \gamma_0)[v_j^* - v_{j,n}^*]) \\ &- \int_0^\tau \mathbb{1}(Y \geq t) \exp\Big(X^T\tilde{\beta}_n + \tilde{\gamma}_n(t) + \tilde{g}_n(\tilde{\Lambda}(t, X))\Big) \cdot \{(\hat{\zeta}_n(t, X, \hat{\beta}_n, \hat{\gamma}_n) - \zeta_0(t, X, \beta_0, \gamma_0)) \\ &\quad \cdot ((v_j^* - v_{j,n}^*)(t) + \tilde{g}_n'(\tilde{\Lambda}(t, X))\tilde{\Lambda}_\gamma'(t, X)[v_j^* - v_{j,n}^*]) \\ &\quad + \hat{\zeta}_{n,\gamma}'(t, X, \hat{\beta}_n, \hat{\gamma}_n)[v_j^* - v_{j,n}^*] - \zeta_{0\gamma}'(t, X, \beta_0, \gamma_0)[v_j^* - v_{j,n}^*]\} dt\Big|^2\Big\}\end{aligned}$$

$$
\begin{aligned}
&\lesssim \|\hat{\zeta}'_{n,\gamma}(\cdot,\hat{\beta}_n,\hat{\gamma}_n)[v_j^* - v_{j,n}^*] - \zeta'_{0\gamma}(\cdot,\beta_0,\gamma_0)[v_j^* - v_{j,n}^*]\|_2^2 \\
&\qquad + \|\hat{\zeta}_n(\cdot,\hat{\beta}_n,\hat{\gamma}_n) - \zeta_0(\cdot,\beta_0,\gamma_0)\|_2^2 \cdot \|v_j^* - v_{j,n}^*\|_\infty^2 \\
&= O_p(n^{-2\min\{(p_1+2)\nu_1,(p_2-1)\nu_2+2\nu_1,\frac{1-\max\{\nu_1,\nu_2\}}{2}+2\nu_1\}}).
\end{aligned}
$$

Thus, we have $I_{1n} = O_p(n^{-\min\{(p_1+2)\nu_1,(p_2-1)\nu_2+2\nu_1,\frac{1-\max\{\nu_1,\nu_2\}}{2}+2\nu_1\}}) = o_p(n^{-1/2})$, because $(p_1+2)\nu_1 > 1/2$, $(p_2-1)\nu_2 + 2\nu_1 > 1/2$, and $4\nu_1 > \max\{\nu_1,\nu_2\}$ under the restrictions listed in Theorem 2.3.1.

Next we will use the maximal inequality in Lemma 3.4.2 of Van Der Vaart and Wellner (1996) (on page 324) and the Markov's inequality to show that $I_{2n} = o_p(n^{-1/2})$. By Lemma A.3.7, the ϵ-bracketing number associated with $\|\cdot\|_\infty$ norm for the class $\mathcal{F}^\gamma_{n,j}(\eta)$ is bounded by $(\eta/\epsilon)^{c_1 q_{n_1} + c_2 q_{n_2} + d}$, which implies that

$$
\log N_{[\,]}(\epsilon,\mathcal{F}^\gamma_{n,j}(\eta),L_2(P)) \le \log N_{[\,]}(\epsilon,\mathcal{F}^\gamma_{n,j}(\eta),\|\cdot\|_\infty) \lesssim (c_1 q_{n_1} + c_2 q_{n_2})\log(\eta/\epsilon).
$$

It follows that the bracketing integral satisfies

$$
J_{[\,]}(\epsilon,\mathcal{F}^\gamma_{n,j}(\eta),L_2(P)) = \int_0^\eta \sqrt{1 + \log N_{[\,]}(\epsilon,\mathcal{F}^\gamma_{n,j}(\eta),L_2(P))}\,d\epsilon \lesssim (c_1 q_{n_1} + c_2 q_{n_2})^{1/2}\eta.
$$

Here we choose $\eta_n = O(n^{-\min\{2\nu_1,p_2\nu_2,\frac{1-\max\{\nu_1,\nu_2\}}{2}\}})$ such that $\|v_j^* - v_{j,n}^*\|_\infty = O(n^{-2\nu_1}) \le \eta_n$ and $d(\hat{\theta}_n,\theta_0) = O_p(n^{-\min\{p_1\nu_1,p_2\nu_2,\frac{1-\max\{\nu_1,\nu_2\}}{2}\}}) \le \eta_n$ for $p_1 \ge 2$, then $l'_\gamma(\hat{\theta}_n); W)[v_j^* - v_{j,n}^*] \in \mathcal{F}^\gamma_{n,j}(\eta_n)$. For any $l'_\gamma(\theta;W)[v_j^* - v_j] \in \mathcal{F}^\gamma_{n,j}(\eta_n)$, we have

$$
\begin{aligned}
&P\{l'_\gamma(\theta;W)[v_j^* - v_j]\}^2 \\
=&P\{\Delta((v_j^* - v_j)(Y) + g'(\Lambda(Y,X,\beta,\gamma,g))\Lambda'_\gamma(Y,X,\beta,\gamma,g)[v_j^* - v_j]) \\
&\quad - \int_0^Y \exp\big(X^T\beta + \gamma(t) + g(\Lambda(t,X,\beta,\gamma,g))\big)\{(v_j^* - v_j)(t) + \zeta'_\gamma(t,X,\beta,\gamma)[v_j^* - v_j]\}\,dt\}^2 \\
\lesssim&\|v_j^* - v_j\|_\infty^2 + \|\Lambda'_\gamma(\cdot,\beta,\gamma,g)[v_j^* - v_j]\|_\infty^2 \\
\lesssim&\|v_j^* - v_j\|_\infty^2.
\end{aligned}
$$

Also, $\sup_{\theta:d(\theta,\theta_0)\le\eta_n;v_j:\|v_j^* - v_j\|_\infty\le\eta_n} |l'_\gamma(\theta;W)[v_j^* - v_j]|$ is bounded by some constant $0 < M < \infty$ (or slowly growing with n and it can be treated as bounded by the same argument used in Shen and Wong (1994, page 591)). By the maximal inequality, it follows that

$$
\begin{aligned}
E_P\|\mathbb{G}_n\|_{\mathcal{F}^{\gamma}_{n,j}(\eta_n)} &\lesssim J_{[\,]}(\epsilon, \mathcal{F}^{\gamma}_{n,j}(\eta_n), L_2(P))\left(1 + \frac{J_{[\,]}(\epsilon, \mathcal{F}^{\gamma}_{n,j}(\eta_n), L_2(P))}{\eta_n^2\sqrt{n}}M\right) \\
&\lesssim (c_1 q_{n_1} + c_2 q_{n_2})^{1/2}\eta_n + (c_1 q_{n_1} + c_2 q_{n_2})n^{-1/2} \\
&= O(n^{\frac{\max\{\nu_1,\nu_2\}}{2}}) \cdot O(n^{-\min\{2\nu_1, p_2\nu_2, \frac{1-\max\{\nu_1,\nu_2\}}{2}\}}) + O(n^{\max\{\nu_1,\nu_2\}-1/2}) \\
&= O(n^{-\min\{2\nu_1 - \frac{\max\{\nu_1,\nu_2\}}{2}, p_2\nu_2 - \frac{\max\{\nu_1,\nu_2\}}{2}, 1/2-\max\{\nu_1,\nu_2\}\}}) + O(n^{\max\{\nu_1,\nu_2\}-1/2}) \\
&= o(1),
\end{aligned}
$$

where $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P)$ and the last equality holds because $0 < \nu_1, \nu_2 < 1/2$, $4\nu_1 > \max\{\nu_1,\nu_2\}$, and $p_2\nu_2 > 2\nu_2 > \max\{\nu_1,\nu_2\}$. Then by the Markov's inequality, we have

$$I_{2n} = n^{-1/2}\mathbb{G}_n l'_{\gamma}(\hat{\theta}_n; W)[v_j^* - v_{j,n}^*] = o_p(n^{-1/2}).$$

By combining $I_{1n} = o_p(n^{-1/2})$ and $I_{2n} = o_p(n^{-1/2})$, we have $S'_{\gamma,n}(\hat{\theta}_n)[v_j^*] = o_p(n^{-1/2})$.

Next, to verify the last part of (A4), we need to show that $S'_{\zeta,n}(\hat{\theta}_n)[h_j^*] = o_p(n^{-1/2})$ with $h_j^*(\cdot, \hat{\beta}_n, \hat{\gamma}_n) = w_j^*(\hat{\Lambda}(\cdot)) + \hat{g}'_n(\hat{\Lambda}(\cdot))\hat{\Lambda}'_g(\cdot)[w_j^*]$, where we write $\hat{\Lambda}(\cdot) = \Lambda(\cdot, \hat{\beta}_n, \hat{\gamma}_n, \hat{g}_n)$ for notational simplicity. Since $w_j^* \in \mathcal{G}^2$, by Lemma A.3.5 there exists $w_{j,n}^* \in \mathcal{G}_n^2$ such that $\|w_{j,n}^* - w_j^*\|_\infty = O(n^{-2\nu_2})$. It follows that $S'_{\zeta,n}(\hat{\beta}_n, \hat{\gamma}_n(\cdot), \hat{\zeta}_n(\cdot, \hat{\beta}_n, \hat{\gamma}_n))[h_{j,n}^*] = 0$ with $h_{j,n}^*(\cdot, \hat{\beta}_n, \hat{\gamma}_n) = w_{j,n}^*(\hat{\Lambda}(\cdot)) + \hat{g}'_n(\hat{\Lambda}(\cdot))\hat{\Lambda}'_g(\cdot)[w_{j,n}^*]$. Then it suffices to show that for each $1 \le j \le d$,

$$
\begin{aligned}
S'_{\zeta,n}(\hat{\theta}_n)[h_j^*] &= S'_{\zeta,n}(\hat{\theta}_n)[h_j^* - h_{j,n}^*] \\
&= P\{l'_{\zeta}(\hat{\theta}_n; W)[h_j^* - h_{j,n}^*] - l'_{\zeta}(\theta_0; W)[h_j^* - h_{j,n}^*]\} \\
&\quad + (\mathbb{P}_n - P)\{l'_{\zeta}(\hat{\zeta}_n; W)[h_j^* - h_{j,n}^*]\} \\
&= I_{3n} + I_{4n} = o_p(n^{-1/2}),
\end{aligned}
$$

since $S'_{\zeta}(\theta_0)[h_j^* - h_{j,n}^*] = 0$. We will take the similar arguments used in the proof of $S'_{\gamma,n}(\hat{\theta}_n)[v_j^*] = o_p(n^{-1/2})$ to show that both I_{3n} and I_{4n} equal to $o_p(n^{-1/2})$.

For I_{3n}, using the Taylor expansion for $l'_{\zeta}(\hat{\theta}_n)[h_j^* - h_{j,n}^*]$ at θ_0, we have

$$I_{3n} = P\{(\hat{\beta}_n - \beta_0)^T l''_{\beta\zeta}(\tilde{\beta}_n, \tilde{\gamma}_n(\cdot), \tilde{\zeta}_n(\cdot, \tilde{\beta}_n, \tilde{\gamma}_n); W)[h_j^* - h_{j,n}^*]$$

$$+ l''_{\zeta\gamma}(\tilde{\beta}_n, \tilde{\gamma}_n(\cdot), \tilde{\zeta}_n(\cdot, \tilde{\beta}_n, \tilde{\gamma}_n); W)[h_j^* - h_{j,n}^*, \hat{\gamma}_n - \gamma_0]$$

$$+ l''_{\zeta\zeta}(\tilde{\beta}_n, \tilde{\gamma}_n(\cdot), \tilde{\zeta}_n(\cdot, \tilde{\beta}_n, \tilde{\gamma}_n); W)[h_j^* - h_{j,n}^*, \hat{\zeta}_n - \zeta_0]\},$$

where $(\tilde{\beta}_n, \tilde{\gamma}_n(\cdot), \tilde{\zeta}_n(\cdot, \tilde{\beta}_n, \tilde{\gamma}_n))$ is some point between θ_0 and $\hat{\theta}_n$. Let $\tilde{\Lambda}(t, x) = \Lambda(t, x, \tilde{\beta}_n, \tilde{\gamma}_n, \tilde{g}_n)$. Note that by solving initial value problems in Lemma A.3.1, we have $\tilde{\Lambda}'_\beta(t, x)$ is bounded on $t \in [0, \tau]$ and $x \in \mathcal{X}$ based on the boundedness of $\tilde{\gamma}_n$, $\tilde{g}_n$, and $\tilde{g}'_n$. Also, we have $\|\tilde{\Lambda}'_g(\cdot)[w]\|_\infty \lesssim \|w\|_\infty$, $\|\tilde{\Lambda}'_\gamma(\cdot)[v]\|_2 \lesssim \|v\|_2$, and furthermore, $\sup_{t\in[0,\tau], x\in\mathcal{X}} \|\tilde{\Lambda}''_{g\beta}(t, x)[w]\| \lesssim \|w\|_\infty + \|w'\|_\infty$ and $\|\tilde{\Lambda}''_{g\gamma}(\cdot)[w, v]\|_2 \lesssim (\|w\|_\infty + \|w'\|_\infty)\|v\|_2$. Using the triangle inequality, it follows that

$$\begin{aligned}
\|(h_j^* - h_{j,n}^*)(\cdot, \tilde{\beta}_n, \tilde{\gamma}_n)\|_\infty &= \|(w_j^* - w_{j,n}^*)(\tilde{\Lambda}(\cdot)) + \tilde{g}'_n(\tilde{\Lambda}(\cdot))\tilde{\Lambda}'_g(\cdot)[w_j^* - w_{j,n}^*]\|_\infty \\
&\lesssim \|w_j^* - w_{j,n}^*\|_\infty, \\
\sup_{t\in[0,\tau], x\in\mathcal{X}} \|(h_j^* - h_{j,n}^*)'_\beta(t, x, \tilde{\beta}_n, \tilde{\gamma}_n)\| &= \sup_{t\in[0,\tau], x\in\mathcal{X}} \|(w_j^* - w_{j,n}^*)'(\tilde{\Lambda}(t, x))\tilde{\Lambda}'_\beta(t, x) \\
&\qquad + \tilde{g}'_n(\tilde{\Lambda}(t, x))\tilde{\Lambda}''_{g\beta}(t, x)[w_j^* - w_{j,n}^*] \\
&\qquad + \tilde{g}''_n(\tilde{\Lambda}(t, x))\tilde{\Lambda}'_g(t, x)[w_j^* - w_{j,n}^*]\tilde{\Lambda}'_\beta(t, x)\| \\
&\lesssim \|w_j^* - w_{j,n}^*\|_\infty + \|(w_j^* - w_{j,n}^*)'\|_\infty,
\end{aligned}$$

and

$$\begin{aligned}
\|(h_j^* - h_{j,n}^*)'_\gamma(\cdot, \tilde{\beta}_n, \tilde{\gamma}_n)[\hat{\gamma}_n - \gamma_0]\|_2 &= \|(w_j^* - w_{j,n}^*)'(\tilde{\Lambda}(\cdot))\tilde{\Lambda}'_\gamma(\cdot)[\hat{\gamma}_n - \gamma_0] \\
&\qquad + \tilde{g}'_n(\tilde{\Lambda}(\cdot))\tilde{\Lambda}''_{g\gamma}(\cdot)[w_j^* - w_{j,n}^*, \hat{\gamma}_n - \gamma_0] \\
&\qquad + \tilde{g}''_n(\tilde{\Lambda}(\cdot))\tilde{\Lambda}'_g(\cdot)[w_j^* - w_{j,n}^*]\tilde{\Lambda}'_\gamma(\cdot)[\hat{\gamma}_n - \gamma_0]\|_2 \\
&\lesssim (\|w_j^* - w_{j,n}^*\|_\infty + \|(w_j^* - w_{j,n}^*)'\|_\infty)\|\hat{\gamma}_n - \gamma_0\|_2.
\end{aligned}$$

Therefore, by plugging the derivatives in Lemma A.3.4 and using the triangle inequality and the Cauchy-Schwarz inequality, I_{3n} is dominated by

$$I_{3n} \lesssim \sup_{t\in[0,\tau], x\in\mathcal{X}} |(\hat{\beta}_n - \beta_0)^T (h_j^* - h_{j,n}^*)'_\beta(t, x, \tilde{\beta}_n, \tilde{\gamma}_n)| + \|(h_j^* - h_{j,n}^*)'_\gamma(\cdot, \tilde{\beta}_n, \tilde{\gamma}_n)[\hat{\gamma}_n - \gamma_0]\|_2$$

$$+ \|(h_j^* - h_{j,n}^*)(\cdot, \tilde{\beta}_n, \tilde{\gamma}_n)\|_\infty \cdot P\Big\{ \int_0^\tau \exp\Big(X^T\tilde{\beta}_n + \tilde{\gamma}_n(t) + \tilde{g}_n(\tilde{\Lambda}(t, X))\Big)$$

$$
\begin{aligned}
&\quad \cdot \big((X + \tilde{g}_n'(\tilde{\Lambda}(t,X))^T(\hat{\beta}_n - \beta_0) + \hat{\gamma}_n(t) - \gamma_0(t) + \tilde{g}_n'(\tilde{\Lambda}(t,X))\tilde{\Lambda}_\gamma'(t,X)[\hat{\gamma}_n - \gamma_0] \\
&\qquad + \hat{\zeta}_n(t,X) - \zeta_0(t,X)\big)^2 dt\Big\}^{1/2} \\
&\lesssim \|\hat{\beta}_n - \beta_0\| \cdot \sup_{t\in[0,\tau],x\in\mathcal{X}} \|(h_j^* - h_{j,n}^*)_\beta'(t,x,\tilde{\beta}_n,\tilde{\gamma}_n)\| + \|(h_j^* - h_{j,n}^*)_\gamma'(\cdot,\tilde{\beta}_n,\tilde{\gamma}_n)[\hat{\gamma}_n - \gamma_0]\|_2 \\
&\quad + \|(h_j^* - h_{j,n}^*)(\cdot,\tilde{\beta}_n,\tilde{\gamma}_n)\|_\infty \cdot \Big(\|\hat{\beta}_n - \beta_0\|^2 + \|\hat{\gamma}_n - \gamma_0\|_2^2 + \|\hat{\zeta}_n - \zeta_0\|_2^2\Big)^{1/2} \\
&\lesssim (\|w_j^* - w_{j,n}^*\|_\infty + \|(w_j^* - w_{j,n}^*)'\|_\infty) d(\hat{\theta}_n, \theta_0).
\end{aligned}
$$

Based on the Corollary 6.21 in Schumaker (2007), we have $\|(w_j^* - w_{j,n}^*)'\|_\infty = O(n^{-\nu_2})$ and

$$
\begin{aligned}
I_{3n} &= O(n^{-\nu_2}) \cdot O_p(n^{-\min\{p_1\nu_1, p_2\nu_2, \frac{1-\max\{\nu_1,\nu_2\}}{2}\}}) \\
&= O_p(n^{-\min\{p_1\nu_1+\nu_2, (p_2+1)\nu_2, \frac{1-\max\{\nu_1,\nu_2\}}{2}+\nu_2\}}) \\
&= o_p(n^{-1/2}),
\end{aligned}
$$

where the last equality holds because $p_1\nu_1 + \nu_2 > 1/2$, $(p_2+1)\nu_2 > 1/2$, and $2\nu_2 > \max\{\nu_1, \nu_2\}$.

Next, we use the maximal inequality and the Markov's inequality to show that $I_{4n} = o_p(n^{-1/2})$. By Lemma A.3.7, the ϵ-bracketing number associated with $\|\cdot\|_\infty$ norm for the class $\mathcal{F}_{n,j}^\zeta(\eta)$ is bounded by $(\eta/\epsilon)^{c_3 q_{n_1} + c_4 q_{n_2} + d}$, which implies that

$$
\log N_{[\,]}(\epsilon, \mathcal{F}_{n,j}^\zeta(\eta), L_2(P)) \le \log N_{[\,]}(\epsilon, \mathcal{F}_{n,j}^\zeta(\eta), \|\cdot\|_\infty) \lesssim (c_3 q_{n_1} + c_4 q_{n_2}) \log(\eta/\epsilon).
$$

It follows that the bracketing integral satisfies

$$
J_{[\,]}(\epsilon, \mathcal{F}_{n,j}^\zeta(\eta), L_2(P)) = \int_0^\eta \sqrt{1 + \log N_{[\,]}(\epsilon, \mathcal{F}_{n,j}^\zeta(\eta), L_2(P))}\, d\epsilon \lesssim (c_3 q_{n_1} + c_4 q_{n_2})^{1/2}\eta.
$$

Here we choose $\eta_n = O(n^{-\min\{p_1\nu_1, 2\nu_2, \frac{1-\max\{\nu_1,\nu_2\}}{2}\}})$ such that $\|w_j^* - w_{j,n}^*\|_\infty = O(n^{-2\nu_2}) \le \eta_n$ and $d(\hat{\theta}_n, \theta_0) = O_p(n^{-\min\{p_1\nu_1, p_2\nu_2, \frac{1-\max\{\nu_1,\nu_2\}}{2}\}}) \le \eta_n$ for $p_2 \ge 3$, then $l_\zeta'(\hat{\theta}_n; W)[h_j^* - h_{j,n}^*] \in \mathcal{F}_{n,j}^\zeta(\eta_n)$. For any $l_\zeta'(\theta; W)[h_j^* - h_j] \in \mathcal{F}_{n,j}^\zeta(\eta_n)$, we have

$$
\begin{aligned}
&P\{l_\zeta'(\theta; W)[h_j^* - h_j]\}^2 \\
=&P\{\Delta((w_j^* - w_j)(Y) + g'(\Lambda(Y,X,\beta,\gamma,g))\Lambda_g'(Y,X,\beta,\gamma,g)[w_j^* - w_j])
\end{aligned}
$$

$$
\begin{aligned}
&- \int_0^Y \exp\bigl(X^T\beta + \gamma(t) + g(\Lambda(t, X, \beta, \gamma, g))\bigr)\{(w_j^* - w_j)(t) + \zeta_g'(t, X, \beta, \gamma)[w_j^* - w_j]\}dt\}^2 \\
&\lesssim \|w_j^* - w_j\|_\infty^2 + \|\Lambda_g'(\cdot, \beta, \gamma, g)[w_j^* - w_j]\|_\infty^2 \\
&\lesssim \|w_j^* - w_j\|_\infty^2 \le \eta_n.
\end{aligned}
$$

Also, $\sup_{\theta:d(\theta,\theta_0)\le\eta_n; w_j:\|w_j^*-w_j\|_\infty\le\eta_n} |l_\zeta'(\theta; W)[h_j^* - h_j]|$ is bounded by some constant $0 < M < \infty$. By the maximal inequality, it follows that

$$
\begin{aligned}
E_P\|\mathbb{G}_n\|_{\mathcal{F}_{n,j}^\zeta(\eta_n)} &\lesssim J_{[\,]}(\epsilon, \mathcal{F}_{n,j}^\zeta(\eta_n), L_2(P)) \left(1 + \frac{J_{[\,]}(\epsilon, \mathcal{F}_{n,j}^\zeta(\eta_n), L_2(P))}{\eta_n^2\sqrt{n}} M\right) \\
&\lesssim (c_3 q_{n_1} + c_4 q_{n_2})^{1/2}\eta_n + (c_3 q_{n_1} + c_4 q_{n_2}) n^{-1/2} \\
&= O(n^{\frac{\max\{\nu_1,\nu_2\}}{2}}) \cdot O(n^{-\min\{p_1\nu_1, 2\nu_2, \frac{1-\max\{\nu_1,\nu_2\}}{2}\}}) + O(n^{\max\{\nu_1,\nu_2\}-1/2}) \\
&= O(n^{-\min\{p_1\nu_1 - \frac{\max\{\nu_1,\nu_2\}}{2}, 2\nu_2 - \frac{\max\{\nu_1,\nu_2\}}{2}, 1/2-\max\{\nu_1,\nu_2\}\}}) + O(n^{\max\{\nu_1,\nu_2\}-1/2}) \\
&= o(1),
\end{aligned}
$$

where the last equality holds because $0 < \nu_1, \nu_2 < 1/2$, $2\nu_2 > \max\{\nu_1, \nu_2\} > \max\{\nu_1, \nu_2\}/2$, and $p_1\nu_1 \ge 2\nu_1 > \max\{\nu_1, \nu_2\}/2$ for $p_1 \ge 2$. Then by the Markov's inequality, we have

$$
I_{4n} = n^{-1/2}\mathbb{G}_n l_\zeta'(\hat\theta_n; W)[h_j^* - h_{j,n}^*] = o_p(n^{-1/2}).
$$

By combining $I_{3n} = o_p(n^{-1/2})$ and $I_{4n} = o_p(n^{-1/2})$, we verify that $S_{\zeta,n}'(\hat\theta_n)[h_j^*] = o_p(n^{-1/2})$. This completes the verification of the assumption (A4).

Now we verify assumption (A5). Since the proofs of three stochastic equicontinuity equations are essentially based on the identical arguments, we only present the proof of the first equation as follows. First, by Lemma A.3.8, the ϵ-bracketing number associated with $\|\cdot\|_\infty$ norm for the class $\mathcal{F}_{n,j}^{*\beta}(\eta)$ is bounded by $(\eta/\epsilon)^{c_1 q_{n_1} + c_2 q_{n_2} + d}$, which implies that the bracketing integral is bounded by $(c_1 q_{n_1} + c_2 q_{n_2})^{1/2}\eta$, i.e.

$$
J_{[\,]}(\epsilon, \mathcal{F}_{n,j}^{*\beta}(\eta), L_2(P)) \lesssim (c_1 q_{n_1} + c_2 q_{n_2})^{1/2}\eta.
$$

For any $l_{\beta_j}'(\theta; W) - l_{\beta_j}'(\theta_0; W) \in \mathcal{F}_{n,j}^{*\beta}(\eta_n)$, by taking the Taylor expansion at θ_0, it follows that

$$\begin{aligned} l'_{\beta_j}(\theta; W) - l'_{\beta_j}(\theta_0; W) =& (\beta - \beta_0)^T l''_{\beta_j \beta}(\tilde{\beta}, \tilde{\gamma}(\cdot), \tilde{\zeta}(\cdot, \tilde{\beta}, \tilde{\gamma}); W) \\ & + l''_{\beta_j \gamma}(\tilde{\beta}, \tilde{\gamma}(\cdot), \tilde{\zeta}(\cdot, \tilde{\beta}, \tilde{\gamma}); W)[\gamma - \gamma_0] \\ & + l''_{\beta_j \zeta}(\tilde{\beta}, \tilde{\gamma}(\cdot), \tilde{\zeta}(\cdot, \tilde{\beta}, \tilde{\gamma}); W)[\zeta - \zeta_0] \end{aligned}$$

where $(\tilde{\beta}, \tilde{\gamma}(\cdot), \tilde{\zeta}(\cdot, \tilde{\beta}, \tilde{\gamma}))$ is some point between θ_0 and θ. By applying the triangle inequality and the Cauchy-Schwarz inequality, we have

$$\begin{aligned} P\{l'_{\beta_j}(\theta; W) - l'_{\beta_j}(\theta_0; W)\}^2 \leq & \|\beta - \beta_0\|^2 P\{\|l''_{\beta_j \beta}(\tilde{\beta}, \tilde{\gamma}(\cdot), \tilde{\zeta}(\cdot, \tilde{\beta}, \tilde{\gamma}); W)\|^2\} \\ & + P\{l''_{\beta_j \gamma}(\tilde{\beta}, \tilde{\gamma}(\cdot), \tilde{\zeta}(\cdot, \tilde{\beta}, \tilde{\gamma}); W)[\gamma - \gamma_0]\}^2 \\ & + P\{l''_{\beta_j \zeta}(\tilde{\beta}, \tilde{\gamma}(\cdot), \tilde{\zeta}(\cdot, \tilde{\beta}, \tilde{\gamma}); W)[\zeta - \zeta_0]\}^2 \\ = & B_1 + B_2 + B_3. \end{aligned}$$

For B_1, by Lemma A.3.4, $l''_{\beta_j \beta}(\tilde{\theta}; W)$ is bounded and it follows that $B_1 \lesssim \|\beta - \beta_0\|^2$. For B_2, since $\tilde{g}$, $\tilde{g}'$, $\tilde{g}''$, $\tilde{\Lambda}'_{\beta_j}(t, x)$ are bounded and $\|\tilde{\Lambda}''_{\beta_j \gamma}(\cdot)[v]\|_2 \lesssim \|v\|_2$, by applying the Cauchy-Schwarz inequality and the same arguments that are used in Lemma A.3.2 to prove that linear operators are bounded above, it follows that

$$\begin{aligned} B_2 =& P\Big\{\Delta \tilde{\zeta}''_{\beta_j \gamma}(Y, X)[\gamma - \gamma_0] - \int_0^Y \Big(\tilde{\zeta}''_{\beta_j \gamma}(t, X)[\gamma - \gamma_0] + (X_j + \tilde{g}'(\tilde{\Lambda}(t, X))\tilde{\Lambda}'_{\beta_j}(t, X)) \\ & \qquad \cdot (\gamma(t) - \gamma_0(t) + \tilde{g}'(\tilde{\Lambda}(t, X))\tilde{\Lambda}'_{\gamma}(t, X)[\gamma - \gamma_0])\Big) d\tilde{\Lambda}(t, X)\Big\}^2 \\ \lesssim & P\{\int_0^Y (\tilde{\zeta}''_{\beta_j \gamma}(t, X)[\gamma - \gamma_0])^2 d\Lambda_0(t, X)\} \\ & + P\{\int_0^Y (\gamma(t) - \gamma_0(t) + \tilde{g}'(\tilde{\Lambda}(t, X))\tilde{\Lambda}'_{\gamma}(t, X)[\gamma - \gamma_0])^2 d\tilde{\Lambda}(t, X)\} \\ \lesssim & P\{\int_0^Y (\tilde{g}'(\tilde{\Lambda}(t, X))\tilde{\Lambda}''_{\beta_j \gamma}(t, X)[\gamma - \gamma_0])^2 d\Lambda_0(t, X)\} \\ & + P\{\int_0^Y (\tilde{g}''(\tilde{\Lambda}(t, X))\tilde{\Lambda}'_{\gamma}(t, X)[\gamma - \gamma_0]\tilde{\Lambda}'_{\beta_j}(t, X))^2 d\Lambda_0(t, X)\} \end{aligned}$$

$$
\begin{aligned}
&+ P\{\int_0^Y (\gamma(t) - \gamma_0(t) + \tilde{g}'(\tilde{\Lambda}(t,X))\tilde{\Lambda}'_\gamma(t,X)[\gamma - \gamma_0])^2 d\tilde{\Lambda}(t,X)\} \\
&\lesssim \|\gamma - \gamma_0\|_2^2 \leq \eta^2.
\end{aligned}
$$

For B_2, similarly, we can show that

$$
\begin{aligned}
B_3 =& P\Big\{ - \int_0^Y \Big((\zeta'_{\beta_j} - \zeta'_{0\beta_j})(t,X) + (X_j + \tilde{g}'(\tilde{\Lambda}(t,X))\tilde{\Lambda}'_{\beta_j}(t,X))(\zeta - \zeta_0)(t,X) d\tilde{\Lambda}(t,X) \\
&+ \Delta(\zeta'_{\beta_j} - \zeta'_{0\beta_j})(Y,X)\Big\}^2 \\
\lesssim& \|\zeta - \zeta_0\|_2^2 + \|\zeta'_{\beta_j} - \zeta'_{0\beta_j}\|_2^2 \leq \eta^2 + \|\zeta'_{\beta_j} - \zeta'_{0\beta_j}\|_2^2.
\end{aligned}
$$

Furthermore, by using the triangle inequality together with the boundedness of Λ'_{β_j} and g'_0, it follows that

$$
\begin{aligned}
\|\zeta'_{\beta_j} - \zeta'_{0\beta_j}\|_2^2 &= \|g'(\Lambda(\cdot,\beta,\gamma,g))\Lambda'_{\beta_j}(\cdot,\beta,\gamma,g) - g'_0(\Lambda_0(\cdot))\Lambda'_{0\beta_j}(\cdot)\|_2^2 \\
&\leq \|g'(\Lambda(\cdot,\beta,\gamma,g))\Lambda'_{\beta_j}(\cdot,\beta,\gamma,g) - g'_0(\Lambda_0(\cdot))\Lambda'_{\beta_j}(\cdot,\beta,\gamma,g)\|_2^2 \\
&\quad + \|g'_0(\Lambda_0(\cdot))\Lambda'_{\beta_j}(\cdot,\beta,\gamma,g) - g'_0(\Lambda_0(\cdot))\Lambda'_{0\beta_j}(\cdot)\|_2^2 \\
&\lesssim \|g'(\Lambda(\cdot,\beta,\gamma,g)) - g'_0(\Lambda_0(\cdot))\|_2^2 + \|\Lambda'_{\beta_j}(\cdot,\beta,\gamma,g) - \Lambda'_{0\beta_j}(\cdot)\|_2^2 \\
&\lesssim \|g'(\Lambda(\cdot,\beta,\gamma,g)) - g'_0(\Lambda_0(\cdot))\|_2^2 + d^2(\theta,\theta_0) \leq \eta^2.
\end{aligned}
$$

Therefore, we have $P\{l'_{\beta_j}(\theta;W) - l'_{\beta_j}(\theta_0;W)\}^2 \lesssim \eta^2$. By Lemma A.3.4, we also have $\|l'_{\beta_j}(\theta;W) - l'_{\beta_j}(\theta_0;W)\|_\infty$ is bounded. We choose $\eta_n = O(n^{-\min\{p_1\nu_1,(p_2-1)\nu_2,\frac{1-\max\{\nu_1,\nu_2\}}{2}\}})$. Then by the maximal inequality, it follows that

$$
\begin{aligned}
E_P\|\mathbb{G}_n\|_{\mathcal{F}^{*\beta}_{n,j}(\eta_n)} &\lesssim (c_1q_{n_1} + c_2q_{n_2})^{1/2}\eta_n + (c_1q_{n_1} + c_2q_{n_2})n^{-1/2} \\
&= O(n^{\frac{\max\{\nu_1,\nu_2\}}{2}}) \cdot O(n^{-\min\{p_1\nu_1,(p_2-1)\nu_2,\frac{1-\max\{\nu_1,\nu_2\}}{2}\}}) + O(n^{\max\{\nu_1,\nu_2\}-1/2}) \\
&= O(n^{-\min\{p_1\nu_1-\frac{\max\{\nu_1,\nu_2\}}{2},(p_2-1)\nu_2-\frac{\max\{\nu_1,\nu_2\}}{2},1/2-\max\{\nu_1,\nu_2\}\}}) + O(n^{\max\{\nu_1,\nu_2\}-1/2}) \\
&= o(1),
\end{aligned}
$$

where the last equality holds because $p_1\nu_1 \geq \nu_1 > \max\{\nu_1,\nu_2\}/2$, $(p_2-1)\nu_2 \geq 2\nu_2 > \max\{\nu_1,\nu_2\}/2$ for $p_2 \geq 3$, and $0 < \nu_1, \nu_2 < 1/2$. Thus, for $\xi = \min\{p_1\nu_1, p_2\nu_2, \frac{1-\max\{\nu_1,\nu_2\}}{2}\}$

and $Cn^{-\xi} = O(n^{-\min\{p_1\nu_1, p_2\nu_2, \frac{1-\max\{\nu_1,\nu_2\}}{2}})$, by Markov's inequality, we have

$$\sup_{d(\theta,\theta_0)\le Cn^{-\xi}, \theta\in\Theta_n} |\mathbb{G}_n\{l'_{\beta_j}(\theta; W) - l'_{\beta_j}(\theta_0; W)\}| = o_p(1),$$

which completes the verification of the first equation in the assumption (A5). The other two stochastic equicontinuity equations in (A5) can be verified using the same arguments.

Finally, we verify assumption (A6) using the Taylor expansion. Similarly, we just prove the first equation, since the proofs of the other two equations are based on the same arguments. By taking the Taylor expansion of $l'_\beta(\theta; W)$ at θ_0, it follows that

$$l'_\beta(\theta; W) - l'_\beta(\theta_0; W) = l''_{\beta\beta}(\tilde\theta; W)(\beta - \beta_0) + l''_{\beta\gamma}(\tilde\theta; W)[\gamma - \gamma_0] + l''_{\beta\zeta}(\tilde\theta; W)[\zeta - \zeta_0]$$

where $\tilde\theta = (\tilde\beta, \tilde\gamma(\cdot), \tilde\zeta(\cdot, \tilde\beta, \tilde\gamma))$ is a point between θ and θ_0. Thus,

$$\begin{aligned}
&P\{l'_\beta(\theta; W) - l'_\beta(\theta_0; W) - l''_{\beta\beta}(\theta_0; W)(\beta - \beta_0) - l''_{\beta\gamma}(\theta_0; W)[\gamma - \gamma_0] - l''_{\beta\zeta}(\theta_0; W)[\zeta - \zeta_0]\} \\
&\quad = P\Big\{(l''_{\beta\beta}(\tilde\theta; W) - l''_{\beta\beta}(\theta_0; W))(\beta - \beta_0)\Big\} + P\Big\{l''_{\beta\gamma}(\tilde\theta; W)[\gamma - \gamma_0] - l''_{\beta\gamma}(\theta_0; W)[\gamma - \gamma_0]\Big\} \\
&\qquad + P\Big\{l''_{\beta\zeta}(\tilde\theta; W)[\zeta - \zeta_0] - l''_{\beta\zeta}(\theta_0; W)[\zeta - \zeta_0]\Big\}
\end{aligned}$$

After some direct calculation, we have

$$\begin{aligned}
&\Big|P\Big\{l''_{\beta\beta}(\tilde\theta; W) - l''_{\beta\beta}(\theta_0; W)\Big\}\Big| \\
\le& P\Big\{\int_0^Y \Big|\Big(\exp\big(X^T\beta_0 + \gamma_0(t) + \zeta_0(t, X)\big) - \exp\Big(X^T\tilde\beta + \tilde\gamma(t) + \tilde\zeta(t, X)\Big)\Big)\, \tilde\zeta''_{\beta\beta}(t, X)\Big| dt\Big\} \\
&+ P\Big\{\Big|\int_0^Y (X + \zeta'_{0\beta}(t, X))(X + \zeta'_{0\beta}(t, X))^T - (X + \tilde\zeta'_\beta(t, X))(X + \tilde\zeta'_\beta(t, X))^T d\Lambda_0(t, X)\Big|\Big\} \\
&+ P\Big\{\int_0^Y \Big|\Big(\exp\big(X^T\beta_0 + \gamma_0(t) + \zeta_0(t, X)\big) - \exp\Big(X^T\tilde\beta + \tilde\gamma(t) + \tilde\zeta(t, X)\Big)\Big) \\
&\qquad\qquad \cdot (X + \tilde\zeta'_\beta(t, X))(X + \tilde\zeta'_\beta(t, X))^T\Big| dt\Big\} \\
=& K_1 + K_2 + K_3.
\end{aligned}$$

For K_1, by the mean value theorem and the Cauchy-Schwarz inequality, it follows that

$$
\begin{aligned}
K_1 &= P\Big\{\int_0^Y \Big|\exp\Big(\tilde{\psi}(t,X)\Big)\Big(X^T(\beta_0-\tilde{\beta})+(\gamma_0-\tilde{\gamma})(t)+\zeta_0(t,X)-\tilde{\zeta}(t,X)\Big)\,\tilde{\zeta}''_{\beta\beta}(t,X)\Big|\,dt\Big\}\\
&\lesssim \|\beta_0-\tilde{\beta}\|+\|\gamma_0-\tilde{\gamma}\|_2+\|\zeta_0-\tilde{\zeta}\|_2 \le d(\theta_0,\theta)\\
&= O(n^{-\min\{p_1\nu_1,p_2\nu_2,\frac{1-\max\{\nu_1,\nu_2\}}{2}\}}),
\end{aligned}
$$

where $\tilde{\psi}(t,X)=X^T\beta_0+\gamma_0(t)+\zeta_0(t,X)+\xi(X^T(\tilde{\beta}-\beta_0)+\tilde{\gamma}(t)-\gamma_0(t)+\tilde{\zeta}(t,X)-\zeta_0(t,X))$ for some $\xi\in(0,1)$ and is bounded. For K_2, by the Cauchy-Schwarz inequality and the same arguments that are used to verify assumption (A4), we have

$$
\begin{aligned}
K_2 &\lesssim P\Big\{\int_0^\tau \Big|(\zeta'_{0\beta}(t,X)-\tilde{\zeta}'_\beta(t,X))(X+\zeta'_{0\beta}(t,X)+\tilde{\zeta}'_\beta(t,X))^T\Big|^2\,d\Lambda_0(t,X)\Big\}^{1/2}\\
&\lesssim \|\zeta'_{0\beta}(\cdot)-\tilde{\zeta}'_\beta(\cdot)\|_2\\
&\lesssim d(\theta_0,\theta)+\|g'_0(\Lambda_0(\cdot))-g(\Lambda(\cdot,\beta,\gamma,g))\|_2\\
&= O(n^{-\min\{p_1\nu_1,(p_2-1)\nu_2,\frac{1-\max\{\nu_1,\nu_2\}}{2}\}}).
\end{aligned}
$$

For K_3, by applying the same arguments for K_1, we can show that

$$
K_3 \lesssim \|\beta_0-\tilde{\beta}\|+\|\gamma_0-\tilde{\gamma}\|_2+\|\zeta_0-\tilde{\zeta}\|_2 = O(n^{-\min\{p_1\nu_1,p_2\nu_2,\frac{1-\max\{\nu_1,\nu_2\}}{2}\}}).
$$

Therefore,

$$
\begin{aligned}
&P\Big\{\Big|(l''_{\beta\beta}(\tilde{\theta};W)-l''_{\beta\beta}(\theta_0;W))(\beta-\beta_0)\Big|\Big\}\\
&\quad= O(n^{-\min\{p_1\nu_1,(p_2-1)\nu_2,\frac{1-\max\{\nu_1,\nu_2\}}{2}\}})\cdot O(n^{-\min\{p_1\nu_1,p_2\nu_2,\frac{1-\max\{\nu_1,\nu_2\}}{2}\}})\\
&\quad= O(n^{-\min\{2p_1\nu_1,p_1\nu_1+(p_2-1)\nu_2,(2p_2-1)\nu_2,\frac{1}{2}+p_1\nu_1-\frac{\max\{\nu_1,\nu_2\}}{2},\frac{1}{2}+(p_2-1)\nu_2-\frac{\max\{\nu_1,\nu_2\}}{2},1-\max\{\nu_1,\nu_2\}\}})\\
&\quad= o(n^{-1/2}),
\end{aligned}
$$

where the last equality holds because $p_1\ge 2$ and $p_2\ge 3$, thus $2p_1\nu_1 > p_1/(p_1+2)\ge 1/2$, $p_1\nu_1+(p_2-1)\nu_2 > \frac{p_1}{2(p_1+2)}+\frac{p_2-1}{2(p_2+1)} \ge \frac{1}{2\cdot 2}+\frac{1}{2\cdot 2}=\frac{1}{2}$, $(2p_2-1)\nu_2 > \frac{2p_2-1}{2(p_2+1)} > \frac{1}{2}$, $p_1\nu_1 \ge 2\nu_1 > \frac{\max\{\nu_1,\nu_2\}}{2}$, $(p_2-1)\nu_2 > \nu_2 > \frac{\max\{\nu_1,\nu_2\}}{2}$, and $\max\{\nu_1,\nu_2\}<1/2$. Similarly, we can show that

$$
P\Big\{\Big|l''_{\beta\gamma}(\tilde{\theta};W)[\gamma-\gamma_0]-l''_{\beta\gamma}(\theta_0;W)[\gamma-\gamma_0]\Big|\Big\}
$$

$$= O(n^{-\min\{2p_1\nu_1, p_1\nu_1+(p_2-1)\nu_2, (2p_2-1)\nu_2, \frac{1}{2}+p_1\nu_1-\frac{\max\{\nu_1,\nu_2\}}{2}, \frac{1}{2}+(p_2-1)\nu_2-\frac{\max\{\nu_1,\nu_2\}}{2}, 1-\max\{\nu_1,\nu_2\}\}})$$

$$= o(n^{-1/2})$$

and

$$P\Big\{\Big|l''_{\beta\zeta}(\tilde{\theta}; W)[\zeta-\zeta_0] - l''_{\beta\zeta}(\theta_0; W)[\zeta-\zeta_0]\Big|\Big\}$$

$$= O(n^{-\min\{2p_1\nu_1, p_1\nu_1+(p_2-1)\nu_2, (2p_2-1)\nu_2, \frac{1}{2}+p_1\nu_1-\frac{\max\{\nu_1,\nu_2\}}{2}, \frac{1}{2}+(p_2-1)\nu_2-\frac{\max\{\nu_1,\nu_2\}}{2}, 1-\max\{\nu_1,\nu_2\}\}})$$

$$= o(n^{-1/2}).$$

Thus, it follows that

$$P\{l'_\beta(\theta; W) - l'_\beta(\theta_0; W) - l''_{\beta\beta}(\theta_0; W)(\beta-\beta_0) - l''_{\beta\gamma}(\theta_0; W)[\gamma-\gamma_0] - l''_{\beta\zeta}(\theta_0; W)[\zeta-\zeta_0]\}$$

$$= O(n^{-\min\{2p_1\nu_1, p_1\nu_1+(p_2-1)\nu_2, (2p_2-1)\nu_2, \frac{1}{2}+p_1\nu_1-\frac{\max\{\nu_1,\nu_2\}}{2}, \frac{1}{2}+(p_2-1)\nu_2-\frac{\max\{\nu_1,\nu_2\}}{2}, 1-\max\{\nu_1,\nu_2\}\}})$$

$$= O(n^{-\alpha\xi})$$

where $\alpha = \min\{2p_1\nu_1, p_1\nu_1 + (p_2-1)\nu_2, (2p_2-1)\nu_2, \frac{1}{2} + p_1\nu_1 - \frac{\max\{\nu_1,\nu_2\}}{2}, \frac{1}{2} + (p_2 - 1)\nu_2 - \frac{\max\{\nu_1,\nu_2\}}{2}, 1 - \max\{\nu_1,\nu_2\}\}/\min\{p_1\nu_1, p_2\nu_2, \frac{1-\max\{\nu_1,\nu_2\}}{2}\} > 1$ and $\alpha\xi > 1/2$. This completes the verification of (A6).

Therefore, we have verified (A1)-(A6) and by Theorem 2.3.3, we have

$$\sqrt{n}(\hat{\beta}_n - \beta_0) = A^{-1}\sqrt{n}\mathbb{P}_n\boldsymbol{l}^*(\beta_0, \gamma_0, \zeta_0; W) + o_p(1) \to_d N(0, A^{-1}B(A^{-1})^T),$$

where

$$\boldsymbol{l}^*(\beta_0, \gamma_0, \zeta_0; W) = l'_\beta(\beta_0, \gamma_0, \zeta_0; W) - l'_\gamma(\beta_0, \gamma_0, \zeta_0; W)[\mathbf{v}^*] - l'_\zeta(\beta_0, \gamma_0, \zeta_0; W)[\mathbf{h}^*(\cdot, \beta_0, \gamma_0)]$$

and A is given by $P\{\boldsymbol{l}^*(\beta_0, \gamma_0, \zeta_0; W)^{\otimes 2}\} = I(\beta_0)$, as shown in the above verification of (A3). Thus, $A = B = I(\beta_0)$ and $A^{-1}B(A^{-1})^T = I^{-1}(\beta_0)$. Therefore, we have

$$\sqrt{n}(\hat{\beta}_n - \beta_0) = \sqrt{n}I^{-1}(\beta_0)\mathbb{P}_n\boldsymbol{l}^*(\beta_0, \gamma_0, \zeta_0; W) + o_p(1) \to_d N(0, I^{-1}(\beta_0)),$$

which completes the proof. □

A.3.4 Explanation of Condition (C7)

Condition (C7) assumes the existence of the least favorable directions which are essential for semi-parametric efficiency. We may find $\mathbf{v}^*$ and $\mathbf{w}^*$ through equations in (C7). Specifically, $\mathbf{v}^*$ and $\mathbf{w}^*$ need to satisfy $P\{\Delta\mathbf{A}^*(U,X)\psi'_{0\gamma}(Y,X)[v]\} = 0$ and $P\{\Delta\mathbf{A}^*(U,X)\psi'_{0g}(Y,X)[w]\} = 0$ for any $v \in \Gamma^{p_1}$ and $w \in \mathcal{G}^{p_2}$.

For the first equation, using the fact of $P\{\int_0^Y f(t,X)\,d\Lambda_0(t,X)\} = P\{\Delta f(Y,X)\}$ and the equations in (A.19), we have for any $v \in \Gamma^{p_1}$

$$
\begin{align}
&P\{\Delta\mathbf{A}^*(U,X)\psi'_{0\gamma}(Y,X)[v]\} \nonumber\\
=&P\{\int_0^Y \mathbf{A}^*(R(t)e^{X^T\beta_0},X) \nonumber\\
&\cdot\left(g_0'(\Lambda_0(t,X))\exp(g_0(\Lambda_0(t,X)))e^{X^T\beta_0}\int_0^t \exp(\gamma_0(s))v(s)\,ds + v(t)\right)d\Lambda_0(t,X)\} \nonumber\\
=&P\{\int_0^{R(Y)e^{X^T\beta_0}} \exp\Big(g_0(\tilde\Lambda_0(t))\Big)\mathbf{A}^*(t,X) \nonumber\\
&\cdot\left(g_0'(\tilde\Lambda_0(t))\exp\Big(g_0(\tilde\Lambda_0(t))\Big)\int_0^t v(R^{-1}(e^{-X^T\beta_0}s))\,ds + v(R^{-1}(e^{-X^T\beta_0}t)\right)dt\} \nonumber\\
=&P\{\int_0^{R(Y)e^{X^T\beta_0}} v(R^{-1}(e^{-X^T\beta_0}s))\int_s^{R(Y)e^{X^T\beta_0}} g_0'(\tilde\Lambda_0(t))\exp\Big(2g_0(\tilde\Lambda_0(t))\Big)\mathbf{A}^*(t,X)\,dt\,ds \nonumber\\
&+\int_0^{R(Y)e^{X^T\beta_0}} v(R^{-1}(e^{-X^T\beta}s))\exp\Big(g_0(\tilde\Lambda_0(s))\Big)\mathbf{A}^*(s,X)\,ds\} \nonumber\\
=&P\{\int_0^\infty \mathbb{1}(R(Y)\geq s)\cdot v(R^{-1}(s))\cdot e^{X^T\beta_0} \nonumber\\
&\cdot\Big(\int_{se^{X^T\beta_0}}^{R(Y)e^{X^T\beta_0}} g_0'(\tilde\Lambda_0(t))\exp\Big(2g_0(\tilde\Lambda_0(t))\Big)\mathbf{A}^*(t,X)\,dt \tag{A.38}\\
&\qquad\qquad+\exp\Big(g_0(\tilde\Lambda_0(se^{X^T\beta_0}))\Big)\mathbf{A}^*(se^{X^T\beta_0},X)\Big)ds\} \nonumber\\
=&\int_0^\infty v(R^{-1}(s))\cdot P\Big\{\mathbb{1}(R(Y)\geq s)\cdot e^{X^T\beta_0} \nonumber\\
&\cdot\Big(\int_{se^{X^T\beta_0}}^{R(Y)e^{X^T\beta_0}} g_0'(\tilde\Lambda_0(t))\exp\Big(2g_0(\tilde\Lambda_0(t))\Big)\mathbf{A}^*(t,X)\,dt \tag{A.39}
\end{align}
$$

$$+\exp\Big(g_0(\tilde{\Lambda}_0(se^{X^T\beta_0}))\Big)\mathbf{A}^*(se^{X^T\beta_0}, X)\Big)\Big\}ds, \tag{A.40}$$

where the second equality is obtained by the variable transformation $\tilde{t} = R(t)x^{X^T\beta_0}$ and further replacing the notation $\tilde{t}$ with t in the integral, and the third equality holds by switching the order of integration. To make the equation (A.40) equal to zero for any $v \in \Gamma^{p_1}$, we can take $\mathbf{v}^*$ and $\mathbf{w}^*$ satisfying

$$\begin{aligned} &P\{\int_{se^{X^T\beta_0}}^{R(Y)e^{X^T\beta_0}} g_0'(\tilde{\Lambda}_0(t))\exp\Big(2g_0(\tilde{\Lambda}_0(t))\Big)\mathbf{A}^*(t,X)e^{X^T\beta_0}\,dt\} \\ &\qquad = -P\{\mathbb{1}(R(Y) \geq s)\exp\Big(g_0(\tilde{\Lambda}_0(se^{X^T\beta_0}))\Big)\mathbf{A}^*(se^{X^T\beta_0}, X)e^{X^T\beta_0}\}. \end{aligned} \tag{A.41}$$

For the second equation in (C7), similarly, we have

$$\begin{aligned} &P\{\Delta\mathbf{A}^*(U,X)\psi_{0g}'(Y,X)[w]\} \\ =&P\{\int_0^Y \mathbf{A}^*(R(t)e^{X^T\beta_0}, X) \\ &\cdot\left(g_0'(\Lambda_0(t,X))\exp(g_0(\Lambda_0(t,X)))\int_0^{\Lambda_0(t,X)}\exp(-g_0(s))w(s)\,ds + w(\Lambda_0(t,X))\right)d\Lambda_0(t,X)\} \\ =&P\{\int_0^{R(Y)e^{X^T\beta_0}} \exp\Big(g_0(\tilde{\Lambda}_0(t))\Big)\mathbf{A}^*(t,X) \\ &\qquad\cdot\left(g_0'(\tilde{\Lambda}_0(t))\exp\Big(g_0(\tilde{\Lambda}_0(t))\Big)\int_0^t w(\tilde{\Lambda}_0(s))\,ds + w(\tilde{\Lambda}_0(t))\right)dt\} \\ =&P\{\int_0^U\left(\int_s^U g_0'(\tilde{\Lambda}_0(t))\exp\Big(2g_0(\tilde{\Lambda}_0(t))\Big)\mathbf{A}^*(t,X)\,dt + \exp\Big(g_0(\tilde{\Lambda}_0(\eta))\Big)\mathbf{A}^*(t,X)\right) \\ &\qquad\cdot w(\tilde{\Lambda}_0(s))\,ds\} \\ =&\int_0^\infty w(\tilde{\Lambda}_0(s)) \\ &\cdot P\{\int_s^U g_0'(\tilde{\Lambda}_0(t))\exp\Big(2g_0(\tilde{\Lambda}_0(t))\Big)\mathbf{A}^*(t,X)\,dt + \mathbb{1}(U \geq s)\exp\Big(g_0(\tilde{\Lambda}_0(s))\Big)\mathbf{A}^*(s,X)\}ds. \end{aligned}$$

To make it equal to zero for any $w \in \mathcal{G}^{p_2}$, we can take $\mathbf{v}^*$ and $\mathbf{w}^*$ such that, for any

η, $\mathbf{A}^*(t,x)$ satisfies

$$\begin{aligned}\int\limits_s^\infty & P\{\mathbb{1}(U \geq t)\mathbf{A}^*(t,X)\}g_0'(\tilde{\Lambda}_0(t))\exp\Big(2g_0(\tilde{\Lambda}_0(t))\Big)\,dt \\ &= -\exp\Big(g_0(\tilde{\Lambda}_0(s))\Big)P\{\mathbb{1}(U \geq s)\mathbf{A}^*(s,X)\}. \qquad \text{(A.42)}\end{aligned}$$

By taking derivatives with respect to s on both sides, we have

$$\exp\Big(g_0(\tilde{\Lambda}_0(s))\Big)\frac{dP\{\mathbb{1}(U \geq s)\mathbf{A}^*(s,X)\}}{ds} = 0,$$

which implies that $P\{\mathbb{1}(U \geq s)\mathbf{A}^*(s,X)\}$ is a constant. Then equation (A.42) holds only if

$$P\{\mathbb{1}(U \geq s)\mathbf{A}^*(s,X)\} = 0. \qquad \text{(A.43)}$$

Therefore, we can take $\mathbf{v}^*$ and $\mathbf{w}^*$ such that $\mathbf{A}^*(t,x)$ satisfies equations (A.41) and (A.43).

Next, we provide solutions for the Cox model and the linear transformation model with a known transformation function as illustration.

For the Cox model where $g_0 \equiv 0$, it suffices to find $\mathbf{v}^*$ such that the equation in (A.41) holds with $\mathbf{A}^*(t,x) = -x + \mathbf{v}^*(R^{-1}(te^{-x^T\beta_0}))$, which implies that $P\{\mathbb{1}(R(Y) \geq t)e^{X^T\beta_0}(\mathbf{v}^*(R^{-1}(t)) - X)\} = 0$. We can take

$$\mathbf{v}^*(t) = \frac{P\{\mathbb{1}(Y \geq t)e^{X^T\beta_0}X\}}{P\{\mathbb{1}(Y \geq t)e^{X^T\beta_0}\}}.$$

For the linear transformation model where γ_0 is known, it suffices to find $\mathbf{w}^*$ such that the equation in (A.43) holds with

$$\begin{aligned}\mathbf{A}^*(t,x) = &- \big(g_0'(\tilde{\Lambda}_0(t))\exp\Big(g_0(\tilde{\Lambda}_0(t))\Big)t + 1\big)x \\ &+ g_0'(\tilde{\Lambda}_0(t)\exp\Big(g_0(\tilde{\Lambda}_0(t))\Big)\int\limits_0^{\tilde{\Lambda}_0(t)}\exp(-g_0(s))\mathbf{w}^*(s)\,ds + \mathbf{w}^*(\tilde{\Lambda}_0(t)).\end{aligned}$$

It follows that $\mathbf{w}^*$ satisfies

$$\begin{aligned}g_0'(\tilde{\Lambda}_0(t)\exp\Big(g_0(\tilde{\Lambda}_0(t))\Big)&\int\limits_0^{\tilde{\Lambda}_0(t)}\exp(-g_0(s))\mathbf{w}^*(s)\,ds + \mathbf{w}^*(\tilde{\Lambda}_0(t)) \\ &= \big(g_0'(\tilde{\Lambda}_0(t))\exp\Big(g_0(\tilde{\Lambda}_0(t))\Big)t + 1\big)\frac{P\{\mathbb{1}(U \geq t)X\}}{P\{\mathbb{1}(U \geq t)\}}.\end{aligned}$$

By taking the variable transformation $\tilde{t} = \tilde{\Lambda}_0(t)$ and further replacing $\tilde{t}$ with t, it is sufficient to take $\mathbf{w}^*$ such that $g_0'(t)\exp(g_0(t))\int_0^t \exp(-g_0(s))\mathbf{w}^*(s)\,ds + \mathbf{w}^*(t) = \boldsymbol{\phi}(t)$ where $\boldsymbol{\phi}(t)$ is given by

$$\boldsymbol{\phi}(t) = \left(g_0'(t)\exp(g_0(t))\tilde{\Lambda}_0^{-1}(t) + 1\right)\frac{P\{\mathbb{1}(\Lambda_0(Y,X) \geq t)X\}}{P\{\mathbb{1}(\Lambda_0(Y,X) \geq t)\}}.$$

It is straightforward to verify that $\mathbf{w}^*$ can be taken as $\mathbf{w}^*(t) = \boldsymbol{\phi}(t) - g_0'(t)\int_0^t \boldsymbol{\phi}(s)\,ds$.

A.3.5 Simplification of Condition (C8)

Condition (C8) assumes non-singularity assumption of the information matrix. We may simplify it to some sufficient conditions if we can find the least favorable directions required in the condition (C7). Recall that we have provided explicit constructions of the least favorable directions for the Cox model and for the linear transformation model with a known transformation respectively in Section A.3.4. We further reduce the non-singularity assumption for the above two cases as follows.

For the Cox model, we have $g_0 \equiv 0$, $\tilde{\Lambda}_0(t) \equiv t$, and the least favorable function $\mathbf{v}^*$ can be derived as

$$\mathbf{v}^*(t) = \frac{P\{\mathbb{1}(Y \geq t)e^{X^T\beta_0}X\}}{P\{\mathbb{1}(Y \geq t)e^{X^T\beta_0}\}}.$$

It follows that the efficient score for β is

$$\boldsymbol{l}^*(\beta_0,\gamma_0;W) = \int_0^\infty \mathbf{A}^*(t,x)\,dM(t) = \int_0^\infty [-X + \frac{P\{\mathbb{1}(U \geq t)e^{X^T\beta_0}X\}}{P\{\mathbb{1}(U \geq t)e^{X^T\beta_0}\}}]\,dM(t),$$

where $U = e^{X^T\beta_0}\int_0^Y \exp(\gamma_0(s))\,ds$ as defined in (C5) and $M(t) = \Delta\mathbb{1}(U \leq t) - \int_0^t \mathbb{1}(U \geq s)\,ds$ is the event counting process martingale. Let $\boldsymbol{\mu}(t) = \frac{P\{\mathbb{1}(U\geq t)e^{X^T\beta_0}X\}}{P\{\mathbb{1}(U\geq t)e^{X^T\beta_0}\}}$. Then by the property of martingale, the information matrix is given by

$$\begin{aligned} I(\beta_0) = P(\boldsymbol{l}^*(\beta_0,\gamma_0;W)^{\otimes 2}) &= P\left(\int_0^\infty [-X + \boldsymbol{\mu}(t)]^{\otimes 2}\,\mathbb{1}(U \geq t)\,dt\right) \\ &= \int_0^\infty P\left([-X + \boldsymbol{\mu}(t)]^{\otimes 2}\,\mathbb{1}(U \geq t)\right)dt, \end{aligned}$$

which reduces to the same information matrix of the MPLE for the Cox model. The above information matrix is similarly assumed to be positive definite in Kalbfleisch and Prentice (2011, page 175). The non-singularity condition in (C8) can be satisfied if $P\left([-X+\boldsymbol{\mu}(t)]^{\otimes 2}\,\mathbb{1}(U \geq t)\right)$ is positive definite over a set of t with non-zero measure.

For the linear transformation model with a known transformation, i.e. γ_0 is known, given the least favorable direction $\mathbf{w}^*$ in Remark 8, the efficient score for β is

$$\boldsymbol{l}^*(\beta_0, \zeta_0(\cdot, \beta_0); W) = \int_0^\infty m(t)\,[P(X|U \geq t) - X]\, dM(t),$$

with $m(t) = g_0'(\tilde{\Lambda}_0(t)) \exp\Big(g_0(\tilde{\Lambda}_0(t))\Big) t + 1$, and the information matrix is

$$\begin{aligned} I(\beta_0) &= P\left(\int_0^\infty m(t)^2\,[P(X|U \geq t) - X]^{\otimes 2}\,\mathbb{1}(U \geq t)\ d\tilde{\Lambda}_0(t)\right) \\ &= \int_0^\infty m^2(t)\cdot P\left([P(X|U \geq t) - X]^{\otimes 2}\,\mathbb{1}(U \geq t)\right)\cdot \exp\Big(g_0(\tilde{\Lambda}_0(t))\Big)\,dt \\ &= \int_0^\infty m^2(t)\cdot Var(X|U \geq t)\cdot P(U \geq t)\cdot \exp\Big(g_0(\tilde{\Lambda}_0(t))\Big)\,dt. \end{aligned}$$

The information matrix takes a similar form as that in Ding and Nan (2011), where it is assumed to be positive definite. Here we further investigate some sufficient conditions for its non-singularity. The condition (C8) can be satisfied if $m^2(t)\cdot Var(X|U \geq t)\cdot P(U \geq t)$ is positive definite over a set of t with non-zero measure. In particular, when the event time follows the AFT model with a Weibull error, i.e., $\gamma_0 \equiv 0$ and $\tilde{\Lambda}_0(t) = kt^v$, the information matrix becomes

$$I(\beta_0) = \int_0^\infty v^2\cdot Var(X|Ce^{X^T\beta_0} \geq t)\cdot P(Ce^{X^T\beta_0} \geq t)\ dF_0(t),$$

where $F_0(t) = 1 - \exp(-kt^v)$ and C is the censoring time. This information matrix is nonsingular if the conditional variance $Var(X|Ce^{X^T\beta_0} \geq t)$ is positive definite for t over certain interval.

A.4 Proof of Propositions 2.2.1 and 2.2.2

The proof of Proposition 2.2.1 is based on the existing identifiability conditions for the linear transformation model (Horowitz, 1996) when both the transformation function and the error distribution are unknown.

Proof of Proposition 2.2.1. Suppose two groups of parameters $(q_i(\cdot), \beta_i, \alpha_i(\cdot))$ for $i = 1, 2$ give the same survival distribution. Let $H_i(u) = \int_0^{-\ln u} q_i^{-1}(v)\,dv$, $G_i(u) = H_i^{-1}(u)$, and $\varphi_i(t) = \log \int_0^t \alpha_i(s)\,ds$ for $i = 1, 2$. In the equivalent linear regression representation, we have that $\varphi_i(T) = -x^T\beta_i + \epsilon_i$ specifies the same distribution of event time T for $i = 1, 2$, where the survival function of $\exp(\epsilon_i)$ is given by G_i. Note that, for the linear transformation model $\varphi(T) = -x^\top\beta + \epsilon$ with both φ and the distribution of ϵ unspecified, Horowitz (1996, page 105) stated that the model parameters are identifiable up to a scale and a location normalization when at least one of the covariates x has a non-zero β coefficient and the conditional probability distribution of this covariate given the remaining covariates is absolutely continuous with respect to Lebesgue measure. Since we assume that there is at least one of the covariates in x is continuous and this covariate has a non-zero coefficient, following the identifiability conditions stated in Horowitz (1996), there exist constants $c_1 > 0$ and c_2 such that $\beta_1 = c_1\beta_2$, $\varphi_1(t) = c_1\varphi_2(t) + c_2$ for any $t > 0$, and ϵ_1 has the same distribution as $c_1\epsilon_2 + c_2$, i.e.,

$$\begin{aligned} G_1(t) &= Pr(\exp(\epsilon_1) > t) = Pr(\exp(c_1\epsilon_2 + c_2) > t) \\ &= Pr(\exp(\epsilon_2) > (te^{-c_2})^{1/c_1})) = G_2((te^{-c_2})^{1/c_1}). \end{aligned}$$

After plugging the definitions of φ_i along with some calculations, we have for any $t > 0$

$$\int_0^t \alpha_1(s)\,ds = e^{c_2}\left(\int_0^t \alpha_2(s)\,ds\right)^{c_1}.$$

Let $\exp(-s) = G_1(t) = G_2((te^{-c_2})^{1/c_1})$. Then by the definitions of G_i we have

$$t = H_1(\exp(-s)) = \int_0^s q_1^{-1}(v)\,dv \quad \text{and} \quad (te^{-c_2})^{1/c_1} = H_2(\exp(-s)) = \int_0^s q_2^{-1}(v)\,dv.$$

It follows that $\int_0^s q_1^{-1}(v)\,dv = e^{c_2}\left(\int_0^s q_2^{-1}(v)\,dv\right)^{c_1}$ for any $s > 0$, which completes the proof. □

As a direct result of Proposition 2.2.1, Proposition 2.2.2 provides the necessary and sufficient degeneration condition for AFT and Cox models.

Proof of Proposition 2.2.2. The linear transformation model in (2.5) coincides with the Cox model if and only if there exists some positive function $\tilde{\alpha}$ such that parameters $(1, \tilde{\beta}, \tilde{\alpha}(\cdot))$ and $(q(\cdot), \beta, \alpha(\cdot))$ give the same survival distribution. By Proposition 2.2.1, there exists positive constants c_1 and c_2 such that

$$\int_0^t q^{-1}(s)\,ds = c_2 t^{c_1}, \beta = c_1\tilde{\beta}, \text{ and } \int_0^t \alpha(s)\,ds = c_2\left(\int_0^t \tilde{\alpha}(s)\,ds\right)^{c_1}.$$

It implies that the function q satisfies $q(t) = \frac{1}{c_1 c_2} t^{1-c_1}$. Similarly, when the linear transformation model coincides with the AFT model, there exists some positive function $\tilde{q}$ such that parameters $(\tilde{q}(\cdot), \tilde{\beta}, 1)$ and $(q(\cdot), \beta, \alpha(\cdot))$ give the same survival distribution. By Proposition 2.2.1, there exists positive constants c_1 and c_2 such that

$$\int_0^t q^{-1}(s)\,ds = c_2\left(\int_0^t \tilde{q}(s)\,ds\right)^{c_1}, \beta = c_1\tilde{\beta}, \text{ and } \int_0^t \alpha(s)\,ds = c_2 t^{c_1}.$$

It follows that the function α takes the form $\alpha(t) = c_1 c_2 t^{c_1 - 1}$, which completes the proof. □

A.5 Theoretical Properties for the General Class of ODE Models and Their Proofs

In this section, we further establish the convergence rate and the asymptotic normality of the proposed sieve estimator for the general class of ODE models in the presence of covariates Z with time-varying coefficients. We reformulate the model to ensure the positivity of $\alpha(\cdot)$ and $q(\cdot)$ in (2.2) below,

$$\begin{cases} \Lambda'(t) = \exp\big(x^T\beta + \gamma(t) + z^T\boldsymbol{\eta}(t) + g(\Lambda(t))\big) \\ \Lambda(0) = 0 \end{cases}, \tag{A.44}$$

where $\gamma(\cdot) = \log\alpha(\cdot)$ and $g(\cdot) = \log q(\cdot)$. Recall that, when there is at least one non-zero time-varying effect, i.e., $\boldsymbol{\eta}(t) \neq 0$, two groups of parameters $(\beta, \gamma, g, \boldsymbol{\eta})$ and $(\tilde{\beta}, \tilde{\gamma}, \tilde{g}, \tilde{\boldsymbol{\eta}})$ give the same survival distribution if only if $\beta = \tilde{\beta}$, $\gamma = \tilde{\gamma}+c$, $g = \tilde{g}-c$, and $\boldsymbol{\eta} = \tilde{\boldsymbol{\eta}}$ for some constant c. To guarantee the identifiability, we constrain $\gamma(t^*) = 0$ with some fixed time point t^*.

Before stating the regularity conditions and main theorems, we firstly update the notation to make them consistent with the model in (A.44). Let $Z \in \mathbf{R}^{d_2+1}$ substitute $(1, Z^T)^T$ and $\boldsymbol{\gamma}(\cdot)$ substitute $(\gamma(\cdot), \eta_1(\cdot), \ldots, \eta_{d_2}(\cdot))^T$ for notational simplicity, then the general class of ODE models is equivalent to

$$\begin{cases} \Lambda'(t) = \exp\big(x^T\beta + z^T\boldsymbol{\gamma}(t) + g(\Lambda(t))\big) \\ \Lambda(0) = 0 \end{cases}, \tag{A.45}$$

with the first component of $\boldsymbol{\gamma}$ fixed at the time point t^*, i.e., $\gamma_1(t^*) = c$. We denote the solution of (A.45) by $\Lambda(t, x, z, \beta, \boldsymbol{\gamma}, g)$ and the true parameters associated with the data generating distribution by (β_0, γ_0, g_0) and simplify $\Lambda(t, x, z, \beta_0, \boldsymbol{\gamma}_0, g_0)$ as $\Lambda_0(t, x, z)$.

To accommodate covariates Z with time-varying coefficients, we update the conditions (C1)-(C8) to (C1$'$)-(C8$'$) with additional regularity conditions on covariates Z and provide the theorem statements and the sketch of proof in the following subsections.

A.5.1 Regularity conditions and main theorems

We assume additional regularity conditions on Z and list the updated conditions below.

(C1$'$) The true parameter β_0 is an interior point of a compact set $\mathcal{B} \subset \mathbf{R}^{d_1}$.

(C2$'$) The joint density of X and Z is bounded below by a constant $c > 0$ over the compact domain $\mathcal{X} \times \mathcal{Z} \subset \mathbf{R}^{d_1+d_2+1}$. $P(XX^T)$ and $P(ZZ^T)$ are nonsingular.

(C3$'$) There exists a truncation time $\tau < \infty$ such that, for some positive constant δ_0, $Pr(Y > \tau | X, Z) \geq \delta_0$ almost surely with respect to the joint probability measure of X and Z. Then there is a constant $\mu = \sup_{x\in\mathcal{X}, z\in\mathcal{Z}} \Lambda_0(\tau, x, z) \leq -\log \delta_0$ such that $\Lambda_0(\tau, X, Z) = -\log Pr(T > \tau | X, Z) \leq \mu$ almost surely with respect to the joint probability measure of X and Z.

(C4$'$) Let $S^p([a, b])$ denote the collection of bounded functions f on $[a, b]$ defined in (C4). The true function $\boldsymbol{\gamma}_0(\cdot)$ belongs to $\Gamma^{p_1}_{t^*} \times \underbrace{\Gamma^{p_1} \times \cdots \times \Gamma^{p_1}}_{d_2}$, where $\Gamma^{p_1} := S^{p_1}([0, \tau])$ and $\Gamma^{p_1}_{t^*} := \{\gamma \in S^{p_1}([0, \tau]) : \gamma(t^*) = 0\}$ with $p_1 \geq 2$, and the true function $g_0(\cdot)$ belongs to $\mathcal{G}^{p_2} := S^{p_2}([0, \mu + \delta_1])$ with some positive constant δ_1 and $p_2 \geq 3$.

(C5$'$) Denote $R_z(t) = \int_0^t \exp\big(z^T \boldsymbol{\gamma}_0(s)\big)\, ds$, $V = X^T \beta_0$, and $U = e^V R_Z(Y)$. There exists $\eta_1 \in (0, 1)$ such that for all $u \in \mathbf{R}^{d_1}$ with $\|u\| = 1$,

$$u^T Var(X \mid U, V, Z)u \geq \eta_1 u^T P(XX^T \mid U, V, Z)u \quad \text{almost surely.}$$

(C6$'$) Let $\psi(t, x, z, \beta, \boldsymbol{\gamma}, g) = x^T\beta + z^T\boldsymbol{\gamma}(t) + g(\Lambda(t, x, z, \beta, \boldsymbol{\gamma}, g))$ and denote its functional derivatives with respect to the entirety $\bar{\gamma}(\cdot) = z^T\boldsymbol{\gamma}(\cdot)$ and $g(\cdot)$ along the direction $v(\cdot)$ and $w(\cdot)$ at the true parameter by $\psi'_{0\bar{\gamma}}(t, x, z)[v]$ and $\psi'_{0g}(t, x, z)[w]$ respectively, whose rigorous definitions are given by (A.49)-(A.50). For any $\boldsymbol{v}(\cdot) = (v_1, \ldots, v_{d_2+1})^T$ with $v_j \in \Gamma^{p_1}$, $1 \leq j \leq d_2 + 1$, and $w(\cdot) \in \mathcal{G}^{p_2}$, there exists $\eta_2 \in (0, 1)$ such that

$$\begin{aligned}&(P\{\psi'_{0\bar{\gamma}}(Y, X, Z)[Z^T\boldsymbol{v}]\psi'_{0g}(Y, X, Z)[w] \,|\, \Delta = 1\})^2\\ &\qquad \leq \eta_2 P\{(\psi'_{0\bar{\gamma}}(Y, X, Z)[Z^T\boldsymbol{v}])^2 \,|\, \Delta = 1\} P\{(\psi'_{0g}(Y, X, Z)[w])^2 \,|\, \Delta = 1\}\end{aligned}$$

almost surely.

(C7$'$) There exist $\mathbf{v}_j^* = (v_{j1}^*, \cdots, v_{jd_1}^*)^T$ and $\mathbf{w}^* = (w_1^*, \cdots, w_{d_1}^*)^T$, where $v_{jk}^* \in \Gamma^2$ and $w_k^* \in \mathcal{G}^2$ for $1 \leq j \leq d_2 + 1, 1 \leq k \leq d_1$, such that

$P\{\Delta \mathbf{A}^*(U, X, Z)\psi'_{0\gamma_\ell}(Y, X, Z)[v]\} = 0$ and $P\{\Delta \mathbf{A}^*(U, X, Z)\psi'_{0g}(Y, X, Z)[w]\} = 0$

hold for any $v \in \Gamma^{p_1}$, $1 \leq \ell \leq d_2 + 1$, and $w \in \mathcal{G}^{p_2}$. Here $\psi'_{0\gamma_\ell}(t, x, z)[v]$ denotes the functional derivative with respect to the ℓ-th component of $\boldsymbol{\gamma}$ along the direction $v(\cdot)$ at the true parameter, U and V are defined in condition (C5$'$), and

$$\begin{aligned}\mathbf{A}^*(t, X, Z) = &- \Big(g_0'(\tilde{\Lambda}_0(t)) \exp\Big(g_0(\tilde{\Lambda}_0(t))\Big)t + 1\Big)\, X \\ &+ \sum_{j=1}^{d_2+1} \Bigg[g_0'(\tilde{\Lambda}_0(t)) \exp\Big(g_0(\tilde{\Lambda}_0(t))\Big) \int_0^t Z_j \mathbf{v}_j^*(R_Z^{-1}(se^{-V}))\,ds + Z_j \mathbf{v}_j^*(R_Z^{-1}(te^{-V}))\Bigg] \\ &+ g_0'(\tilde{\Lambda}_0(t)) \exp\Big(g_0(\tilde{\Lambda}_0(t))\Big) \int_0^{\tilde{\Lambda}_0(t)} \exp(-g_0(s))\mathbf{w}^*(s)\,ds + \mathbf{w}^*(\tilde{\Lambda}_0(t)),\end{aligned}$$

where $\tilde{\Lambda}_0(t)$ is the solution of $\tilde{\Lambda}_0'(t) = \exp\Big(g_0(\tilde{\Lambda}_0)\Big)$ with $\tilde{\Lambda}_0(0) = 0$.

(C8$'$) Let $\boldsymbol{l}^*(\beta_0, \boldsymbol{\gamma}_0, \zeta_0; W) = \int \mathbf{A}^*(t, X, Z)\, dM(t)$, where $M(t) = \Delta \mathbb{1}(U \leq t) - \int_0^t \mathbb{1}(U \geq s)\, d\tilde{\Lambda}_0(s)$ is the event counting process martingale. The information matrix $I(\beta_0) = P(\boldsymbol{l}^*(\beta_0, \boldsymbol{\gamma}_0, \zeta_0; W)^{\otimes 2})$ is nonsingular. Here for a vector a, $a^{\otimes 2} = aa^T$.

In the presence of covariates Z with time-varying coefficients, conditions (C2$'$)-(C3$'$) contain additional common regularity assumptions for Z in survival analysis. Condition (C4$'$) controls the error rates of the spline approximation for the true time-varying coefficients. The expectation in condition (C5$'$) is further conditioned on covariates Z. Condition (C6$'$) is similarly assumed to avoid strong collinearity between $\psi'_{0\bar{\gamma}}(Y, X, Z)[v]$ and $\psi'_{0g}(Y, X, Z)[w]$ while $\bar{\gamma}$ denotes the linear combination $z^T\boldsymbol{\gamma}$. Condition (C7$'$) additionally requires the existence of the least favorable directions for time-varying coefficients and the information matrix in (C8$'$) also depends on the additional least favorable directions. In particular, conditions (C1$'$)-(C8$'$) are equivalent to conditions (C1)-(C8) respectively when Z only contains the intercept.

Given the above regularity conditions, for the general class of ODE models in (2.2), we can establish the same convergence rate of the sieve estimator as that in Theorem 2.3.1 and the asymptotic normality as in Theorem 2.3.2. Since the theory is investigated with the fixed number of covariates d_1 and d_2 as the sample size n grows, including additional covariates Z with time-varying coefficients does not change the nature of the proof. For presentation integrity, we provide rigorous definitions of the corresponding parameter space, the sieve space, theorem statements, and a sketch of proof that summarizes the main steps in the following subsection.

First, we define the parameter space and the associated distance when including covariates Z with time-varying coefficients. We similarly define the collection of functions

$\mathcal{H}^{p_2} = \{\zeta(\cdot, \beta, \boldsymbol{\gamma}) : \zeta(t, x, z, \beta, \boldsymbol{\gamma}) = g(\Lambda(t, x, z, \beta, \boldsymbol{\gamma}, g)), t \in [0, \tau], x \in \mathcal{X}, z \in \mathcal{Z}, \beta \in \mathcal{B},$

$\boldsymbol{\gamma} \in \Gamma^{p_1}_{t^*} \times \underbrace{\Gamma^{p_1} \times \cdots \times \Gamma^{p_1}}_{d_2}, g \in \mathcal{G}^{p_2}$ such that $\sup\limits_{t\in[0,\tau], x\in\mathcal{X}, z\in\mathcal{Z}} |\Lambda(t, x, z, \beta, \boldsymbol{\gamma}, g)| \leq \mu + \delta_1\},$

with δ_1 given in condition (C4′). For any $\zeta(\cdot, \beta, \boldsymbol{\gamma}) \in \mathcal{H}^{p_2}$, we define its norm as

$$\|\zeta(\cdot, \beta, \boldsymbol{\gamma})\|_2 = \left[\int\limits_{\mathcal{X}\times\mathcal{Z}} \int\limits_0^\tau [\zeta(t, x, z, \beta, \boldsymbol{\gamma})]^2 d\Lambda_0(t, x, z) dF_{X,Z}(x, z) \right]^{1/2},$$

where $F_{X,Z}(x, z)$ is the cumulative distribution function of (X, Z). Denote the parameter $\theta = (\beta, \boldsymbol{\gamma}(\cdot), \zeta(\cdot, \beta, \boldsymbol{\gamma}))$ and the true parameter $\theta_0 = (\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0))$ with

$$\zeta_0(t, x, z, \beta_0, \boldsymbol{\gamma}_0) = g_0(\Lambda(t, x, z, \beta_0, \boldsymbol{\gamma}_0, g_0)).$$

Denote the parameter space by $\Theta = \mathcal{B} \times \Gamma^{p_1}_{t^*} \times \underbrace{\Gamma^{p_1} \times \cdots \times \Gamma^{p_1}}_{d_2} \times \mathcal{H}^{p_2}$. For any θ_1 and θ_2 in Θ, we define the distance

$$d(\theta_1, \theta_2) = \left(\|\beta_1 - \beta_2\|^2 + \|\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_2\|_2^2 + \|\zeta_1(\cdot, \beta_1, \boldsymbol{\gamma}_1) - \zeta_2(\cdot, \beta_2, \boldsymbol{\gamma}_2)\|_2^2\right)^{1/2},$$

where $\|\cdot\|$ is the Euclidean norm and $\|\boldsymbol{\gamma}\|_2 = (\sum_{j=1}^{d_2+1} \int_0^\tau (\gamma_j(t))^2 dt)^{1/2}$.

Next, we construct the sieve space by using the space of polynomial splines in a similar way. Let $\Gamma^{p_1}_n = S_n(T_{K^1_n}, K^1_n, p_1)$, $\Gamma^{p_1}_{t^*,n} = \{\gamma \in S_n(T_{K^1_n}, K^1_n, p_1) : \gamma(t^*) = 0\}$,

$\mathcal{G}_n^{p_2} = S_n(T_{K_n^2}, K_n^2, p_2)$, and

$$\mathcal{H}_n^{p_2} = \{\zeta(\cdot,\beta,\boldsymbol{\gamma}) : \zeta(t,x,z,\beta,\boldsymbol{\gamma}) = g(\Lambda(t,x,z,\beta,\boldsymbol{\gamma},g)),\ t \in [0,\tau], x \in \mathcal{X}, z \in \mathcal{Z}, \beta \in \mathcal{B},$$
$$\boldsymbol{\gamma} \in \Gamma^{p_1}_{t^*,n} \times \underbrace{\Gamma^{p_1}_n \times \cdots \times \Gamma^{p_1}_n}_{d_2}, g \in \mathcal{G}_n^{p_2}\}.$$

Let $\Theta_n = \mathcal{B} \times \Gamma^{p_1}_{t^*,n} \times \underbrace{\Gamma^{p_1}_n \times \cdots \times \Gamma^{p_1}_n}_{d_2} \times \mathcal{H}_n^{p_2}$ be the sieve space. The sieve estimator $\hat{\theta}_n = (\hat{\beta}_n, \hat{\boldsymbol{\gamma}}_n(\cdot), \hat{\zeta}_n(\cdot, \hat{\beta}_n, \hat{\boldsymbol{\gamma}}_n))$ maximizes the log-likelihood (2.6) over the sieve space Θ_n. The convergence rate of the sieve MLE $\hat{\theta}_n$ and the asymptotic normality of the sieve MLE $\hat{\beta}_n$ of the regression parameter are then established in Theorem A.5.1 and Theorem A.5.2 respectively.

Theorem A.5.1. *(Convergence rate of $\hat{\theta}_n$.) Let ν_1 and ν_2 satisfy the restrictions* $\max\{\frac{1}{2(2+p_1)}, \frac{1}{2p_1} - \frac{\nu_2}{p_1}\} < \nu_1 < \frac{1}{2p_1}$, $\max\{\frac{1}{2(1+p_2)}, \frac{1}{2(p_2-1)} - \frac{2\nu_1}{p_2-1}\} < \nu_2 < \frac{1}{2p_2}$, *and* $2\min\{2\nu_1, \nu_2\} > \max\{\nu_1, \nu_2\}$. *Suppose conditions (C1′)-(C6′) hold, then we have*

$$d(\hat{\theta}_n, \theta_0) = O_p(n^{-\min\{p_1\nu_1, p_2\nu_2, \frac{1-\max\{\nu_1,\nu_2\}}{2}\}}).$$

Theorem A.5.2. *(Asymptotic normality of $\hat{\beta}_n$) Suppose the conditions in Theorem A.5.1 and (C7′)-(C8′) hold, then we have*

$$\sqrt{n}(\hat{\beta}_n - \beta_0) = \sqrt{n} I^{-1}(\beta_0) \mathbb{P}_n \boldsymbol{l}^*(\beta_0, \gamma_0, \zeta_0; W) + o_p(1) \to_d N(0, I^{-1}(\beta_0))$$

with $I(\beta_0)$ given in condition (C8′) and $\to_d$ denoting convergence in distribution.

A.5.2 Sketch of proof

Given the updated conditions (C1′)-(C8′), the proof of Theorems A.5.1 and A.5.2 is based on the similar techniques and arguments as that of Theorems 2.3.1 and 2.3.2. We provide the sketch of proof and highlight their main differences below.

Lemmas. The corresponding Lemmas A.3.1-A.3.8 in the presence of covariates Z still hold under new conditions (C1′)-(C7′), which are used to prove Theorems A.5.1 and A.5.2. Specifically,

- The existence and uniqueness of the solution $\Lambda(t, x, z, \beta, \boldsymbol{\gamma}, g)$ of the initial value problem in (A.44) along with its derivatives in Lemma A.3.1, and the boundedness and continuity of derivatives of $l(\beta, \boldsymbol{\gamma}, \zeta; W)$ in Lemma A.3.4 both hold due to the boundedness of Z and the smoothness of $\boldsymbol{\eta}$ under conditions (C1′)-(C4′). In particular, the derivatives are characterized by the corresponding updated initial value problems with covariates Z. For example, initial value problems (A.9)-(A.11) become (A.46)-(A.48) respectively as follows

$$\frac{d\Lambda'_{\beta}(t)}{dt} = \exp\bigl(x^T\beta + z^T\boldsymbol{\gamma}(t) + g(\Lambda(t))\bigr)\{x + g'(\Lambda(t))\Lambda'_{\beta}(t)\}, \tag{A.46}$$

$$\frac{d\Lambda'_{\gamma_j}(t)[v]}{dt} = \exp\bigl(x^T\beta + z^T\boldsymbol{\gamma}(t) + g(\Lambda(t))\bigr)\{z_j v(t) + g'(\Lambda(t))\Lambda'_{\gamma_j}(t)[v]\}, \tag{A.47}$$

$$\frac{d\Lambda'_{g}(t)[w]}{dt} = \exp\bigl(x^T\beta + z^T\boldsymbol{\gamma}(t) + g(\Lambda(t))\bigr)\{w(\Lambda(t)) + g'(\Lambda(t))\Lambda'_{g}(t)[w]\}. \tag{A.48}$$

- In Lemma A.3.2, we show that the operators $\psi'_{0\bar{\gamma}}[\cdot]$ and $\psi'_{0g}[\cdot]$ are bounded from below by the continuous dependence of the IVP solution on parameters in Walter (1998, page 145), where $\psi'_{0\bar{\gamma}}[\cdot]$ denotes the functional derivatives with respect to the entirety $\bar{\gamma}(\cdot) = z^T\boldsymbol{\gamma}(\cdot)$. By solving initial value problem in (A.47), the first derivatives of $\psi(t, x, \beta, \gamma, g)$ with respect to $\bar{\gamma}$ and g at the true parameter (β_0, γ_0, g_0) are updated as

$$\begin{aligned}&\psi'_{0\bar{\gamma}}(t, x, z)[v] = g'_0(\Lambda_0(t, x, z))\Lambda'_{0\bar{\gamma}}(t, x)[v] + v(t)\\ &= g'_0(\Lambda_0(t, x, z))\exp(g_0(\Lambda_0(t, x, z)))e^{x^T\beta_0}\int\limits_0^t \exp\bigl(z^T\boldsymbol{\gamma}_0(s)\bigr)v(s)\,ds + v(t),\end{aligned} \tag{A.49}$$

$$\begin{aligned}&\psi'_{0g}(t, x, z)[w] = g'_0(\Lambda_0(t, x, z))\Lambda'_{0g}(t, x, z)[w] + w(\Lambda_0(t, x, z))\\ &= g'_0(\Lambda_0(t, x, z))\exp(g_0(\Lambda_0(t, x, z)))\int\limits_0^{\Lambda_0(t,x,z)} \exp(-g_0(s))w(s)\,ds + w(\Lambda_0(t, x, z)).\end{aligned} \tag{A.50}$$

- The upper bounds of the ϵ-bracketing numbers associated with $\mathcal{F}_n$, $\mathcal{F}^{\gamma_\ell}_{n,j}(\eta)$,

$\mathcal{F}^{\zeta}_{n,j}(\eta)$, $\mathcal{F}^{*\beta}_{n,j}(\eta)$, $\mathcal{F}^{*\gamma_\ell}_{n,j}(\eta)$, $\mathcal{F}^{*\zeta}_{n,j}(\eta)$ for $1 \le \ell \le d_2{+}1, 1 \le j \le d_1$ in Lemmas A.3.6-A.3.8 are updated as $(\frac{1}{\epsilon})^{c_1 q_{n_1}(d_2+1)+c_2 q_{n_2}+d_1}$ and $(\frac{\eta}{\epsilon})^{c_1 q_{n_1}(d_2+1)+c_2 q_{n_2}+d_1}$, where d_1 and d_2 are dimensions of covariates X and Z respectively. Since we consider the number of covariates d_i fixed as the sample size increases, the updated upper bounds in the presence of Z would not change the convergence rate of the sieve estimator and the nature of the proof.

Proof of Theorem A.5.1. To establish the overall convergence rate of the sieve MLE $\hat{\theta}_n$ in Theorem A.5.1, we verify three conditions C1-C3 required in the main theorem in Shen and Wong (1994). Specifically,

- The condition C1 in Shen and Wong (1994) specifies the increasing rate of the expected log-likelihood ratio as the parameter θ moves away from the true value θ_0. We will prove that

$$\inf_{d(\theta,\theta_0)\ge\epsilon,\theta\in\Theta_n} Pl(\beta_0,\gamma_0,\zeta_0(\cdot,\beta_0,\gamma_0);W) - Pl(\beta,\gamma,\zeta(\cdot,\beta,\gamma);W) \gtrsim \epsilon^2.$$

In the presence of covariate Z, we update

$$\begin{aligned} Pl(\beta,\boldsymbol{\gamma},\zeta(\cdot,\beta,\boldsymbol{\gamma});W) = P\{&\Delta[X^T\beta + Z^T\boldsymbol{\gamma}(Y) + g(\Lambda(Y,X,Z,\beta,\boldsymbol{\gamma},g)) \\ &- \exp\big(X^T\beta + Z^T\boldsymbol{\gamma}(Y) - X^T\beta_0 - Z^T\boldsymbol{\gamma}_0(Y)\big) \\ &\cdot \exp(g(\Lambda(Y,X,Z,\beta,\boldsymbol{\gamma},g)) - g_0(\Lambda_0(Y,X,Z)))]\}. \end{aligned}$$

Using the Taylor expansion along with the same arguments, we have

$$\begin{aligned} &Pl(\beta_0,\boldsymbol{\gamma}_0,\zeta_0(\cdot,\beta_0,\boldsymbol{\gamma}_0);W) - Pl(\beta,\boldsymbol{\gamma},\zeta(\cdot,\beta,\boldsymbol{\gamma});W) \\ \gtrsim\ & P\{\Delta[(g_0'(\Lambda_0(Y,X,Z))\Lambda_{0\beta}'(Y,X,Z) + X)^T(\beta-\beta_0) \\ &+ \sum_{j=1}^{d_2+1} g_0'(\Lambda_0(Y,X,Z))\Lambda_{0\gamma_j}'(Y,X,Z)[(\boldsymbol{\gamma}-\boldsymbol{\gamma}_0)^T e_j] + Z^T(\boldsymbol{\gamma}(Y)-\boldsymbol{\gamma}_0(Y)) \\ &+ g_0'(\Lambda_0(Y,X,Z))\Lambda_{0g}'(Y,X,Z)[g-g_0] + g(\Lambda_0(Y,X,Z)) - g_0(\Lambda_0(Y,X,Z))]^2\} \\ &+ o(d^2(\theta,\theta_0)) \\ =\ & P\{\Delta[\epsilon_1(U)X^T(\beta-\beta_0) + \epsilon_2(U,V,Z)[(\boldsymbol{\gamma}(Y)-\boldsymbol{\gamma}_0(Y))^T Z] + \epsilon_3(U)[g-g_0]]^2\} \end{aligned}$$

$$+ o(d^2(\theta, \theta_0)),$$

where ϵ_1, ϵ_2, and ϵ_3 are deterministic functions of Z, U, V given in condition (C5$'$). Under the updated conditions (C5$'$)-(C6$'$), we can similarly derive that

$$\begin{aligned}
&Pl(\beta_0, \boldsymbol{\gamma}_0, \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0); W) - Pl(\beta, \boldsymbol{\gamma}, \zeta(\cdot, \beta, \boldsymbol{\gamma}); W) \\
&\gtrsim P\{\Delta(\epsilon_1(U) X^T(\beta - \beta_0))^2\} + P\{\Delta(\epsilon_2(U, V, Z)[(\boldsymbol{\gamma}(Y) - \boldsymbol{\gamma}_0(Y))^T Z])^2\} \\
&\quad + P\{\Delta(\epsilon_3(U)[g - g_0])^2\} + o(d^2(\theta, \theta_0)).
\end{aligned}$$

Given the boundedness of Z, the first and third terms are similarly bounded below by $\|\beta - \beta_0\|^2$ and $\|g - g_0\|_2^2$ respectively. The second term is bounded below by

$$\begin{aligned}
P\{\Delta(\epsilon_2(U, V, Z)[(\boldsymbol{\gamma}(Y) - \boldsymbol{\gamma}_0(Y))^T Z])^2\} &\gtrsim \|(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)^T Z\|_2^2 \\
&= \int_0^\tau (\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)(t)^T P\{ZZ^T\}(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)(t)\, dt \\
&\geq \int_0^\tau \lambda_1^{(Z)} (\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)(t)^T (\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)(t)\, dt \\
&= \lambda_1^{(Z)} \|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\|_2^2,
\end{aligned}$$

where $\lambda_1^{(Z)}$ is the smallest eigenvalue of $P\{ZZ^T\}$, which is positive due to the nonsingularity in the updated condition (C2$'$). Therefore, we have

$$\begin{aligned}
Pl(\beta_0, \boldsymbol{\gamma}_0,& \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0); W) - Pl(\beta, \boldsymbol{\gamma}, \zeta(\cdot, \beta, \boldsymbol{\gamma}); W) \\
&\gtrsim \|\beta - \beta_0\|^2 + \|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\|_2^2 + \|g - g_0\|_2^2 \\
&\gtrsim d^2(\theta, \theta_0).
\end{aligned}$$

- The condition C2 in Shen and Wong (1994) controls the decreasing rate of the variance of the log-likelihood ratio as the parameter θ approaches the true value θ_0. We use the same arguments to show that

$$\sup_{d(\theta,\theta_0) \leq \epsilon, \theta \in \Theta_n} Var\{l(\beta, \boldsymbol{\gamma}, \zeta(\cdot, \beta, \boldsymbol{\gamma}); W) - l(\beta_0, \boldsymbol{\gamma}_0, \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0); W)\} \lesssim \epsilon^2.$$

Note that the second term in (A.31) is replaced and upper bounded by

$$
\begin{aligned}
&P\{\Delta\left(Z^T(\boldsymbol{\gamma}(Y)-\boldsymbol{\gamma}_0(Y))\right)^2\} \\
&= P\int_0^\tau 1(Y\geq t)\exp\big(X^T\beta_0+Z^T\boldsymbol{\gamma}_0(t)+g_0(\Lambda_0(t,X,Z))\big)\left(Z^T(\boldsymbol{\gamma}(t)-\boldsymbol{\gamma}_0(t))\right)^2 dt \\
&\leq \int_0^\tau \sup_{x\in\mathcal{X},z\in\mathcal{Z},t\in[0,\tau]}\{\exp\big(x^T\beta_0+Z^T\boldsymbol{\gamma}_0(t)+g_0(\Lambda_0(t,x,z))\big)\}P\left(Z^T(\boldsymbol{\gamma}(t)-\boldsymbol{\gamma}_0(t))\right)^2 dt \\
&\leq \int_0^\tau \sup_{x\in\mathcal{X},z\in\mathcal{Z},t\in[0,\tau]}\{\exp\big(x^T\beta_0+Z^T\boldsymbol{\gamma}_0(t)+g_0(\Lambda_0(t,x,z))\big)\}\lambda_{d_2+1}^{(Z)}\|\boldsymbol{\gamma}(t)-\boldsymbol{\gamma}_0(t)\|^2 dt \\
&\lesssim \|\boldsymbol{\gamma}-\boldsymbol{\gamma}_0\|_2^2,
\end{aligned}
$$

where $\lambda_{d_2+1}^{(Z)}$ is the largest eigenvalue of $P(ZZ^T)$.

- The condition C3 in Shen and Wong (1994) bounds the size of the space of log-likelihood ratio induced by θ, i.e., $\mathcal{F}_n=\{l(\theta;W)-l(\theta_{0n};W):\theta\in\Theta_n\}$. By Lemma 6, we have the L_∞-metric entropy of the space $\mathcal{F}_n$ bounded by

$$
\begin{aligned}
H(\epsilon,\mathcal{F}_n,\|\cdot\|_\infty)=\log(N(\epsilon,\mathcal{F}_n,\|\cdot\|_\infty)) &\lesssim c_1q_{n_1}(d_2+1)+c_2q_{n_2}+d_1 \\
&\lesssim n^{\max\{\nu_1,\nu_2\}}\log(1/\epsilon),
\end{aligned}
$$

  as the number of covariates d_i is considered as fixed.

After verifying the conditions C1-C3, by Theorem 1 in Shen and Wong (1994), we have for the sieve MLE $\hat{\theta}_n$

$$
d(\hat{\theta}_n,\theta_0)=O_p(\max\{n^{-\frac{1-\max\{\nu_1,\nu_2\}}{2}},d(\theta_{0n},\theta_0),K^{1/2}(\theta_{0n},\theta_0)\}),
$$

where $K(\theta_{0n},\theta_0)=P\{l(\theta_0;W)-l(\theta_{0n};W)\}$. We can similarly show that $K(\theta_{0n},\theta_0)\lesssim O(d^2(\theta_{0n},\theta_0))$ by the Taylor expansion, so the convergence rate of $\hat{\theta}_n$ depends on the sieve approximation error $d(\theta_{0n},\theta_0)$. Here $\theta_{0n}=(\beta_0,\boldsymbol{\gamma}_{0n}(\cdot),\zeta_{0n}(\cdot,\beta_0,\boldsymbol{\gamma}_{0n}))\in\Theta_n$ with $\zeta_{0n}(t,x,z,\beta_0,\boldsymbol{\gamma}_{0n})=g_{0n}(\Lambda(t,x,z,\beta_0,\boldsymbol{\gamma}_{0n},g_{0n}))$. Note that $\gamma_{0n,j}\in\Gamma_n^{p_1}$ and $g_{0n}\in\mathcal{G}^{p_2}$ are defined in Lemma 5 such that $\|\gamma_{0n,j}-\gamma_{0,j}\|_\infty=O(n^{-p_1\nu_1})$ and $\|g_{0n}-g_0\|_\infty=O(n^{-p_2\nu_2})$, which is based on the existing spline approximation error in Corollary 6.21 in Schumaker (2007). Since $d^2(\theta_{0n},\theta_0)\lesssim\|\beta_0-\beta_0\|^2+\|\boldsymbol{\gamma}_{0n}-\boldsymbol{\gamma}_0\|_2^2+\|g_{0n}-g_0\|_2^2\lesssim$

$\|\boldsymbol{\gamma}_0 - \boldsymbol{\gamma}_{0n}\|_\infty^2 + \|g_0 - g_{0n}\|_\infty^2 = O(n^{-2\min\{p_1\nu_1, p_2\nu_2\}})$, it follows that

$$d(\hat{\theta}_n, \theta_0) = O_p(n^{-\min\{p_1\nu_1, p_2\nu_2, \frac{1-\max\{\nu_1,\nu_2\}}{2}\}}).$$

Proof of Theorem A.5.2. To establish the asymptotic normality in Theorem A.5.2, we similarly verify the assumptions (A1)-(A6) for the proposed general M-theorem in Theorem 2.3.3 under the updated conditions (C1′)-(C8′). For example, to verify assumption (A3), first, we need to find $\mathbf{v}_j^* = (v_{j1}^*, \cdots, v_{jd_1}^*)'$, $1 \le j \le d_2 + 1$, and $\mathbf{h}^* = (h_1^*, \cdots, h_{d_1}^*)'$ with $\mathbf{h}^*(\cdot) = \mathbf{w}^*(\Lambda_0(\cdot)) + g_0'(\Lambda_0(\cdot))\Lambda_{0g}'(\cdot)[\mathbf{w}^*]$ such that for any $v \in \mathbb{V}$ and $h \in \mathbb{H}$ with $h(\cdot) = w(\Lambda_0(\cdot)) + g_0'(\Lambda_0(\cdot))\Lambda_{0g}'(\cdot)[w]$,

$$\begin{aligned} S''_{\beta\gamma_\ell}(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0))[v] = \sum_{j=1}^{d_2+1} & S''_{\gamma_j\gamma_\ell}(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0))[\mathbf{v}_j^*, v] \\ & + S''_{\zeta\gamma_l}(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0))[\mathbf{h}^*, v], \end{aligned} \tag{A.51}$$

$$\begin{aligned} S''_{\beta\zeta}(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0))[h] = \sum_{j=1}^{d_2+1} & S''_{\gamma_j\zeta}(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0))[\mathbf{v}_j^*, h] \\ & + S''_{\zeta\zeta}(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0))[\mathbf{h}^*, h]. \end{aligned} \tag{A.52}$$

By Lemma A.3.4 and the property $P\{\int_0^Y f(t, X, Z)\,d\Lambda_0(t, X, Z)\} = P\{\Delta f(Y, X, Z)\}$, for any $\mathbf{v}_j \in \mathbb{V}^{d_1}, v \in \mathbb{V}$ and $\mathbf{h} \in \mathbb{H}^{d_1}$ with $\mathbf{h}(\cdot) = \mathbf{w}(\Lambda_0(\cdot)) + g_0'(\Lambda_0(\cdot))\Lambda_{0g}'(\cdot)[\mathbf{w}]$, we have for $1 \le \ell \le d_2 + 1$

$$\begin{aligned} & S''_{\beta\gamma_\ell}(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0)[v] - \sum_{j=1}^{d_2+1} S''_{\gamma_j\gamma_\ell}(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0)[\mathbf{v}_j, v] - S''_{\zeta\gamma_l}(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0)[\mathbf{h}, v] \\ =& P\{l''_{\beta\gamma_\ell}(\beta_0, \boldsymbol{\gamma}_0, \zeta_0; W)[v] - \sum_{j=1}^{d_2+1} l''_{\gamma_j\gamma_\ell}(\beta_0, \boldsymbol{\gamma}_0, \zeta_0; W)[\mathbf{v}_j, v] - l''_{\zeta\gamma_\ell}(\beta_0, \boldsymbol{\gamma}_0, \zeta_0; W)[\mathbf{h}, v]\} \\ =& P\Bigg\{\Delta\Bigg[g_0'(\Lambda_0(Y, X, Z))\Lambda_{0\beta}'(Y, X, Z) + X \\ & \quad - \sum_{j=1}^{d_2+1}\Big(g_0'(\Lambda_0(Y, X, Z))\Lambda_{0\gamma_j}'(Y, X, Z)[\mathbf{v}_j] + \mathbf{v}_j(Y)Z_j\Big) \\ & \quad - g_0'(\Lambda_0(Y, X, Z))\Lambda_{0g}'(Y, X, Z)^T[\mathbf{w}] - \mathbf{w}(\Lambda_0(Y, X, Z))\Bigg] \end{aligned}$$

$$\cdot \left(g_0'(\Lambda_0(Y,X,Z))\Lambda_{0\gamma_\ell}'(Y,X,Z)[v] + v(Y)Z_\ell\right)\Bigg\}.$$

Under the updated condition (C7′), there exist $\mathbf{v}_j^* = (v_{j1}^*, \cdots, v_{jd_1}^*)^T$ and $\mathbf{w}^* = (w_1^*, \cdots, w_{d_1}^*)^T$, where $v_{jk}^* \in \Gamma^2$ and $w_k^* \in \mathcal{G}^2$ for $1 \leq j \leq d_2+1, 1 \leq k \leq d_1$, such that $P\{\Delta \mathbf{A}^*(U,X,Z)\psi_{0\gamma_\ell}'(Y,X,Z)[v]\} = 0$ hold for any $v \in \Gamma^{p_1}$, $1 \leq \ell \leq d_2+1$. Therefore, we have that the equation (A.51) holds with $\mathbf{v}_j^*$ and $\mathbf{w}^*$ given in condition (C7′). Similarly, we can show that the equation (A.52) holds as well.

A.6 Additional Simulation Studies

In this section, we provide full results of simulation studies with various sample sizes and investigate 1) how the numerical performance of the proposed method depends on the knot selection by comparing multiple natural knot selections; 2) a heuristic parametric approach that applies the unified ODE framework along with the proposed estimation and inference procedure for model diagnostics.

A.6.1 Time-varying Cox model

Table A.1 summarizes the estimates of regression coefficients β_3 and β_4 in the time-varying Cox model that is considered in subsection 2.3.3. The proposed sieve estimators for β_3 and β_4 perform similarly to those for β_1 and β_2 as shown in Table II.1. The bias of the estimators for β_3 and β_4 decreases and becomes negligible as the sample size increases. The estimated standard error by inverting the estimated information matrix for all parameters including the coefficients of spline basis are close to the sample standard error and the corresponding 95% confidence intervals obtain reasonable coverage proportion.

Table A.1: Simulation results under time-varying Cox model.

N	Method	$\beta_3 = -1$				$\beta_4 = 1$			
		Bias	SE	ESE	CP	Bias	SE	ESE	CP
1000	ODE	-.009	.076	.078	.942	.009	.068	.070	.948
	Cox-MPLE	-.007	.076	.075	.938	.007	.068	.068	.943
2000	ODE	-.004	.052	.054	.965	.005	.047	.048	.955
	Cox-MPLE	-.003	.052	.053	.966	.004	.047	.048	.952
4000	ODE	-.003	.037	.038	.951	.004	.034	.034	.951
	Cox-MPLE	-.003	.037	.037	.950	.003	.034	.034	.950
8000	ODE	.000	.026	.026	.959	-.001	.024	.024	.947
	Cox-MPLE	.000	.026	.026	.952	-.001	.024	.024	.949

Bias is the difference between mean of estimates and the true value; SE is the sample standard error of the estimates; ESE is the mean of the standard error estimators by inverting the estimated information matrix of all parameters including the coefficients of spline basis, and CP is the corresponding coverage proportion of 95% confidence intervals.

A.6.2 Comparison with the method in Royston and Parmar (2002) under the Cox model

In setting 1), we compare the proposed sieve MLE under the Cox model with the parametric method in Royston and Parmar (2002), where the log-transformed baseline cumulative hazard is modeled as a natural cubic spline function of the log-transformed time. We implement it using the "flexsurvspline" function in the R package *flexsurv* with the same number of interior knots, i.e., $\lfloor N'^{\frac{1}{5}} \rfloor$. The sample size N varies from 1000 to 8000.

Table A.2 summarizes the estimates of regression coefficients based on 1000 replicates. We can see that both the proposed estimation method (ODE-Cox) and the method in Royston and Parmar (2002) (flexsurv) perform similarly to maximum partial likelihood estimation (MPLE) in terms of estimation accuracy. As shown in Figure A.1, the proposed method ODE-Cox achieves comparable integrated mean square errors (IMSE) of the estimated cumulative hazard function to those of "flexsurv". In addition, the relative computing time (the computing time with respect to that with the smallest sample size 1000) of proposed method ODE-Cox increases slowly than that of “flexsurv” as the sample size grows. We note that the increasing rate of the relative computing time of the ODE-Cox is even slower than the linear rate, which may be benefited from efficient implementation of existing numerical ODE solvers.

Table A.2: Simulation results under the Cox model.

N	Method	$\beta_1 = 1$				$\beta_2 = 1$				$\beta_3 = 1$			
		Bias	SE	ESE	CP	Bias	SE	ESE	CP	Bias	SE	ESE	CP
1000	MPLE	.006	.153	.152	.948	.010	.157	.152	.944	.004	.152	.152	.950
	ODE-Cox	.009	.153	.157	.952	.013	.157	.157	.952	.007	.152	.158	.961
	Flexsurv	.007	.153	.152	.948	.011	.156	.151	.943	.005	.151	.152	.952
2000	MPLE	.005	.106	.107	.954	-.002	.107	.107	.949	.006	.105	.107	.958
	ODE-Cox	.007	.106	.109	.956	-.001	.107	.109	.955	.007	.105	.109	.961
	Flexsurv	.006	.105	.107	.956	-.001	.107	.107	.950	.007	.105	.107	.955
4000	MPLE	.002	.076	.075	.934	-.003	.075	.075	.941	-.001	.074	.075	.954
	ODE-Cox	.003	.076	.076	.936	-.002	.075	.076	.942	.000	.074	.076	.955
	Flexsurv	.002	.076	.075	.934	-.002	.075	.075	.942	-.001	.074	.075	.953
8000	MPLE	-.002	.053	.053	.953	.000	.052	.053	.954	-.001	.053	.053	.944
	ODE-Cox	-.002	.053	.054	.953	-.000	.052	.054	.957	-.002	.054	.054	.947
	Flexsurv	-.001	.053	.053	.954	.000	.052	.053	.952	-.001	.053	.053	.944

Bias is the difference between the mean of estimates and the true value, and SE is the sample standard error of the estimates. ESE is the mean of the standard error estimators, and CP is the corresponding coverage proportion of 95% confidence intervals.

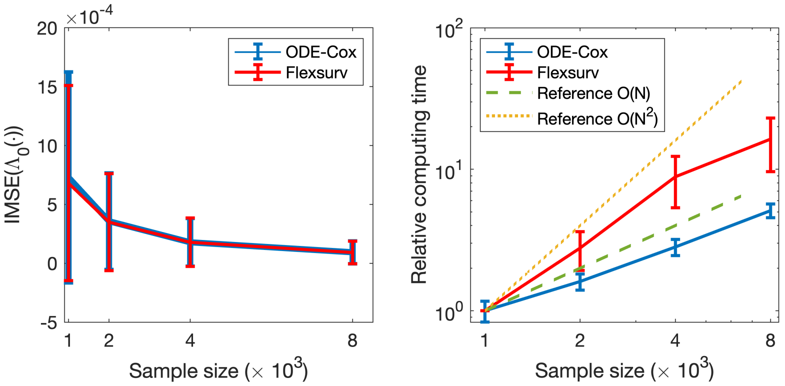

Figure A.1: Integrated mean square error (IMSE) of estimated baseline cumulative hazard functions and the log-log plot of mean relative computing time with respect to the sample size under the Cox model are provided from left to right.

A.6.3 Comparison with the NPMLE (Zeng and Lin, 2007b) under the linear transformation model

We have compared the proposed ODE approach and the NPMLE for the logarithmic transformation model in (Zeng and Lin, 2007b). Specifically, in the simulation setting (2), we generate event times from the ODE

$$\Lambda_x'(t) = q(\Lambda_x(t)) \exp(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)\alpha(t),$$

where functions $q(t) = \exp(-t)$ and $\alpha(t) = 2$. It is equivalent to generate event times with the cumulative hazard function

$$\Lambda_x(t) = G\{\exp(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)\Lambda_0(t)\},$$

where $G(u) = \log(1+u)$ and $\Lambda_0(t) = \int_0^t \alpha(s)ds = 2t$. For the NPMLE in Zeng and Lin (2007b), note that the function $G(\cdot)$ is known and the baseline cumulative hazard $\Lambda_0(\cdot)$ is unknown. An EM algorithm was implemented in Matlab to compute the NPMLE. To make fair comparison, we set the function $q(\cdot)$ known, i.e., $q(t) = \exp(-t)$, and the function $\alpha(\cdot)$ unknown for the ODE-LT. We fit $\log\alpha(\cdot)$ by cubic B-splines and set the number of knots K_n as the largest integer below $N'^{\frac{1}{5}}$, where N' is the number of distinct observation time points. The sample size N varies from $1,000$ to $8,000$.

Table A.3 summarizes the estimates of regression coefficients β based on 1000 replicates. The proposed estimation method (ODE-LT) achieves similar estimation accuracy of both β and the cumulative hazard (shown in the left panel of Figure A.2) as the NPMLE. However, the relative computing time of the proposed method ODE-LT increase linearly as the sample size grows while that of the NPMLE increases in a quadratic rate as shown in the right panel of Figure A.2.

Table A.3: Simulation results under the linear transformation model.

N	Method	$\beta_1 = 1$				$\beta_2 = 1$				$\beta_3 = 1$			
		Bias	SE	ESE	CP	Bias	SE	ESE	CP	Bias	SE	ESE	CP
1000	NPMLE	.003	.227	.230	.954	.003	.236	.230	.949	.003	.229	.230	.954
	ODE-LT	.005	.227	.231	.956	.005	.237	.231	.949	.004	.229	.231	.955
2000	NPMLE	-.002	.159	.162	.946	.003	.169	.162	.933	.006	.157	.162	.963
	ODE-LT	-.001	.159	.163	.947	.003	.169	.163	.933	.007	.157	.163	.961
4000	NPMLE	.004	.117	.115	.949	-.001	.114	.115	.951	.003	.113	.115	.960
	ODE-LT	.005	.117	.115	.950	-.000	.114	.115	.951	.003	.113	.115	.961
8000	NPMLE	-.005	.079	.081	.956	.000	.078	.081	.963	-.001	.079	.081	.950
	ODE-LT	-.004	.079	.081	.957	.001	.078	.081	.963	-.000	.079	.081	.951

Bias is the difference between the mean of estimates and the true value, SE is the sample standard error of the estimates, and Mean is the mean of IMSE. ESE is the mean of the standard error estimators, and CP is the corresponding coverage proportion of 95% confidence intervals.

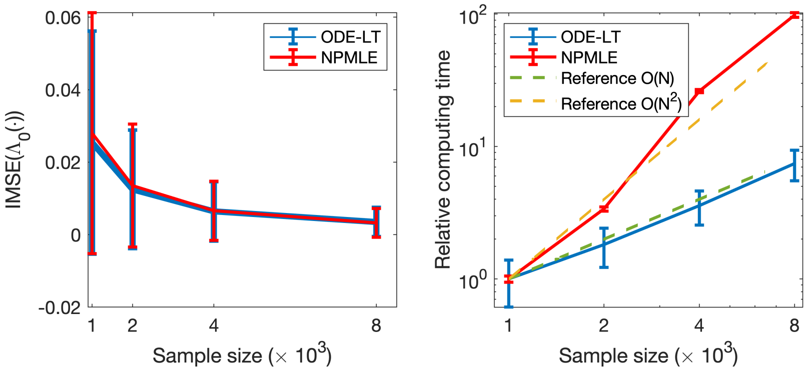

Figure A.2: Integrated mean square error (IMSE) of estimated baseline cumulative hazard functions and the log-log plot of mean relative computing time with respect to the sample size under the linear transformation model are provided from left to right.

A.6.4 Comparison with the rank-based method under the AFT model

In setting 3), we compare the proposed sieve MLE for the ODE-AFT model, where the function α is set to 1, with the rank-based estimation approach implemented using the R package *aftgee*. For the ODE-AFT model, we fit $\log q(t)$ by cubic B-splines with $\lfloor N^{\frac{1}{7}} \rfloor$ interior knots. Note that the argument of the function $q(\cdot)$ is the cumulative hazard. Unlike fitting the function $\alpha(\cdot)$ whose argument is the event time in the ODE-Cox model, we do not observe the corresponding cumulative hazard directly. Therefore, we use the estimated cumulative hazards under the Cox model as a remedy. Let $\hat{\Lambda}_i^{Cox}$ denote the estimated cumulative hazard for individual i under the Cox model. The interior knots are located at the quantiles of $\{\hat{\Lambda}_i^{Cox}\}_{i=1}^n$.

Table A.4 summarizes the estimates of regression coefficients β with varying sample sizes. Although the bias of the ODE approach is relatively greater than that of the rank-based method when the sample size is small, the bias of the estimates becomes negligible as the sample size increases. As shown in Figure A.3, the relative computing time of the proposed ODE approach increases in a slower rate than that of the rank-based method for the semi-parametric ODE-AFT model. Remarkably, the proposed ODE approach takes just 6 seconds for estimating the ODE-AFT model but the rank-based method takes 349 seconds when the sample size is $8,000$.

Table A.4: Simulation results under the AFT model.

N	Method	$\beta_1 = 1$				$\beta_2 = 1$				$\beta_3 = 1$			
		Bias	SE	ESE	CP	Bias	SE	ESE	CP	Bias	SE	ESE	CP
1000	Rank-based	-.000	.204	.206	.952	-.009	.213	.205	.925	-.013	.200	.206	.942
	ODE-AFT	-.014	.197	.191	.944	-.024	.209	.192	.931	-.032	.199	.192	.932
2000	Rank-based	-.002	.147	.145	.938	.005	.147	.145	.951	.004	.146	.146	.945
	ODE-AFT	-.010	.144	.137	.932	-.006	.144	.137	.937	-.005	.142	.137	.943
4000	Rank-based	.004	.105	.102	.944	-.001	.102	.102	.950	.002	.100	.103	.954
	ODE-AFT	.000	.102	.097	.944	-.005	.100	.097	.944	-.002	.097	.097	.950
8000	Rank-based	-.003	.071	.073	.956	.001	.071	.073	.962	.000	.072	.073	.949
	ODE-AFT	-.006	.070	.069	.950	-.003	.068	.069	.967	-.004	.071	.069	.945

Bias is the difference between the mean of estimates and the true value, and SE is the sample standard error of the estimates. ESE is the mean of the standard error estimators, and CP is the corresponding coverage proportion of 95% confidence intervals.

A.6.5 Comparison with the smoothed partial rank method under the general linear transformation model

In settings 1)-4), we compare the sieve MLE for the general linear transformation model (ODE-Flex), where both $q(\cdot)$ and $\alpha(\cdot)$ are unspecified, with the smoothed partial rank (SPR) estimation method in Song et al. (2006), which is a rank-based

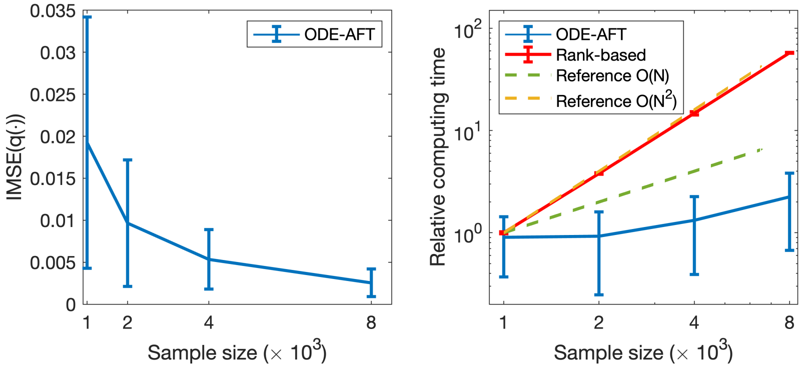

Figure A.3: Integrated mean square error (IMSE) of estimated baseline cumulative hazard functions and the log-log plot of mean relative computing time with respect to the sample size under the AFT model are provided from left to right.

estimation method for censored data. As the original code of SPR is not available, we implement the SPR estimation and inference methods by our own, and we verify that our implementations are able to reproduce the simulation results in Song et al. (2006). Note that SPR introduces an additional parameter c in the objective function to improve the estimation accuracy. We evaluate SPR with various values of the parameter c and the sample size N under our data settings 1)-4). We observe that SPR may return extreme estimates, so we count estimates with more than 5 deviation from the truth as failed replications.

Tables A.5-A.6 summarize the estimates of β_2 under settings 1)-4) over $1,000$ replications. (We observe similar performance for β_3 and so we omit its results here.) In terms of estimation accuracy, both the SPR estimator and ODE-Flex estimator show negligible biases when the sample size is large. However, two inference methods in Song et al. (2006) are sensitive to the choices of the parameter c : the sandwich estimator seriously underestimates the standard deviation for various values of the parameter c and the corresponding coverage proportion is far below the nominal level; the weighted bootstrap estimator overestimates the standard deviation for small values of c and underestimates it for relatively large values of c. In contrast, the proposed ODE-Flex method performs well across various sample sizes: the standard

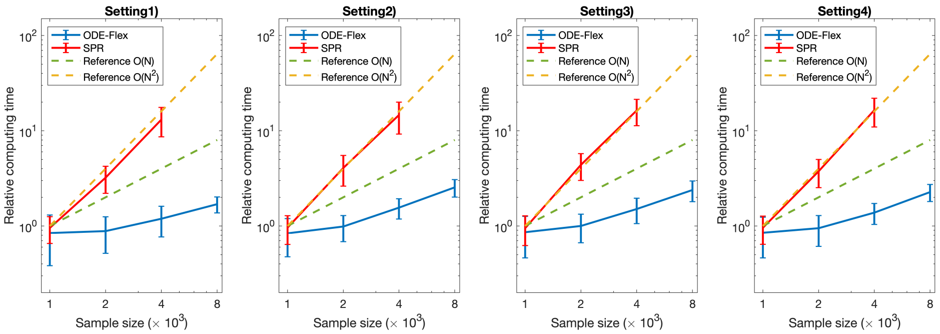

Figure A.4: The log-log plot of mean relative computing time with respect to the sample size under the nonparametric linear transformation model.

error estimators approximate the empirical standard deviations well and the coverage proportions are close to the nominal level. In terms of numerical stability, the proposed ODE-Flex method can stably return good estimates over $1,000$ replications, especially for large sample sizes: only less than 1% replications meet with numerical errors when $N = 4,000$ and 100% replications successfully return accurate estimates when $N = 8,000$. We note that this result is reported under a universal precision for ODE solvers and we find that these failed replications can be easily fixed by adjusting the precision of the ODE solver. However, the SPR method fails to return a reasonable point estimator for more than 12% realized resampling on average when computing the standard error estimator by the weighted bootstrap. We also observe that it is difficult to obtain the SPR point estimator for larger sample size such as $N = 8,000$ or larger parameter c such as 10^{-1} and 1 (success rate less than 10%). In terms of computation efficiency, as shown in Figure A.4, the computing time of ODE-Flex increases in a much smaller rate than that of SPR as the sample size grows, which implies that the proposed estimation method is computationally more efficient for large sample size.

Table A.5: Simulation results of β_2 under the general linear transformation model with both $q(\cdot)$ and $\alpha(\cdot)$ unspecified in settings 1) and 2).

	Method	N	c			Sandwich		Bootstrap			Bootstrap
				Bias	SE	ESE	CP	ESE	CP	Succ. %	Succ. %
1)	SPR	1000	10^{-4}	.030	.331	.000	.000	.697	.974	98.3	87.3
			10^{-3}	.034	.250	.000	.003	.478	.960	97.4	84.0
			10^{-2}	.048	.295	.002	.020	.103	.432	80.1	75.5
		2000	10^{-4}	-.003	.313	.000	.000	.668	.989	98.4	85.0
			10^{-3}	.013	.210	.000	.003	.314	.906	94.5	80.4
			10^{-2}	.007	.159	.002	.022	.033	.279	71.8	70.9
		4000	10^{-4}	.007	.153	.000	.001	.552	.994	97.9	83.1
			10^{-3}	.008	.120	.000	.000	.136	.762	95.2	77.7
			10^{-2}	.005	.105	.002	.022	.016	.222	67.7	67.8
		N		Bias	SE	ESE	CP			Succ. %	
	ODE-Flex	1000		.067	.248	.243	.958			93.6	
		2000		.024	.162	.158	.950			98.4	
		4000		.008	.106	.107	.947			99.5	
		8000		.012	.076	.075	.946			100.0	
	Method	N	c			Sandwich		Bootstrap			Bootstrap
				Bias	SE	ESE	CP	ESE	CP	Succ. %	Succ. %
2)	SPR	1000	10^{-4}	.082	.522	.000	.000	.739	.949	97.8	87.2
			10^{-3}	.091	.449	.000	.002	.538	.910	96.5	84.2
			10^{-2}	.104	.464	.003	.015	.166	.457	81.8	74.1
		2000	10^{-4}	.020	.347	.000	.000	.702	.988	98.3	85.5
			10^{-3}	.015	.320	.000	.000	.393	.895	95.5	80.3
			10^{-2}	.044	.337	.002	.005	.052	.262	75.7	69.3
		4000	10^{-4}	.014	.244	.000	.000	.585	.995	98.5	83.9
			10^{-3}	.019	.191	.000	.001	.183	.709	93.6	77.4
			10^{-2}	.022	.171	.002	.010	.019	.158	67.1	65.2
		N		Bias	SE	ESE	CP			Succ. %	
	ODE-Flex	1000		.024	.357	.312	.918			98.5	
		2000		.009	.246	.218	.931			99.5	
		4000		-.019	.161	.151	.927			100.0	
		8000		-.020	.113	.107	.939			100.0	

Bias is the difference between the mean of estimates and the true value, and SE is the sample standard error of the estimates. ESE is the mean of the standard error estimators, and CP is the corresponding coverage proportion of 95% confidence intervals.

Table A.6: Simulation results of β_2 under the general linear transformation model with both $q(\cdot)$ and $\alpha(\cdot)$ unspecified in settings 3) and 4).

	Method	N	c			Sandwich		Bootstrap			Bootstrap
				Bias	SE	ESE	CP	ESE	CP	Succ. %	Succ. %
3)	SPR	1000	10^{-4}	.053	.369	.000	.000	.779	.980	97.3	86.1
			10^{-3}	.056	.386	.000	.000	.529	.945	95.7	82.9
			10^{-2}	.079	.372	.004	.010	.128	.454	79.9	73.4
		2000	10^{-4}	.004	.304	.000	.000	.721	.992	97.8	84.4
			10^{-3}	.010	.308	.000	.000	.357	.888	96.0	79.2
			10^{-2}	.010	.222	.002	.016	.040	.251	74.8	68.3
		4000	10^{-4}	.005	.194	.000	.000	.602	.996	97.5	82.4
			10^{-3}	.007	.146	.000	.001	.154	.732	92.4	76.0
			10^{-2}	.011	.141	.002	.025	.020	.194	68.1	63.4
		N		Bias	SE	ESE	CP			Succ. %	
	ODE-Flex	1000		.016	.293	.270	.940			95.9	
		2000		.014	.197	.191	.948			99.0	
		4000		-.014	.134	.131	.941			99.7	
		8000		-.019	.088	.092	.957			100.0	
	Method	N	c			Sandwich		Bootstrap			Bootstrap
				Bias	SE	ESE	CP	ESE	CP	Succ. %	Succ. %
4)	SPR	1000	10^{-4}	.023	.349	.000	.000	.756	.987	97.1	84.3
			10^{-3}	.030	.226	.000	.003	.473	.963	95.4	80.5
			10^{-2}	.032	.227	.003	.022	.083	.417	77.9	71.8
		2000	10^{-4}	-.006	.253	.000	.000	.719	.993	97.5	82.0
			10^{-3}	.006	.147	.000	.002	.274	.902	95.2	76.8
			10^{-2}	.007	.136	.004	.034	.027	.275	73.8	66.9
		4000	10^{-4}	.001	.146	.000	.000	.574	.995	96.9	79.5
			10^{-3}	.004	.089	.000	.004	.108	.781	94.2	73.9
			10^{-2}	.000	.086	.002	.029	.019	.240	66.3	64.1
		N		Bias	SE	ESE	CP			Succ. %	
	ODE-Flex	1000		.020	.182	.191	.954			96.7	
		2000		.016	.132	.131	.958			98.8	
		4000		.001	.092	.090	.938			99.9	
		8000		.008	.062	.064	.960			100.0	

Bias is the difference between the mean of estimates and the true value, and SE is the sample standard error of the estimates. ESE is the mean of the standard error estimators, and CP is the corresponding coverage proportion of 95% confidence intervals.

A.6.6 Dependence on knots selection

To investigate how the numerical performance of the proposed method depends on the knot selection, we have done several simulation studies to compare two natural placements of knots for the ODE-Cox model, the ODE-AFT model, and the general linear transformation model. Specifically,

- For the ODE-Cox model, we compare the following two placements of knots when using the B-spline to fit the function $\log \alpha(\cdot)$: (K1) the interior knots are located at the $K_n = \lfloor N^{\frac{1}{5}} \rfloor$ quantiles of the distinct observation time points; (K2) the interior knots equally separate the time interval from 0 to the maximum of observed times.

- For the ODE-AFT model, we compare the following two placements of knots when using the B-spline to fit the function $\log q(\cdot)$: (K1) the interior knots are located at the $K_n = \lfloor N^{\frac{1}{7}} \rfloor$ quantiles of the estimated cumulative hazards $\{\hat{\Lambda}_i^{Cox}\}_{i=1}^n$ under the Cox model; (K2) the interior knots equally separate the interval from 0 to $2 \max_{1 \le i \le n} \{\hat{\Lambda}_i^{Cox}\}$.

- For the general linear transformation model, we compare combinations of the above knots placements when using the B-spline to fit both functions $\log \alpha(\cdot)$ and $\log q(\cdot)$: (K1) the interior knots for both functions are located at the corresponding quantiles; (K2) the interior knots for both functions equally separate the corresponding intervals.

Tables A.7-A.9 compare the estimates of regression coefficients β with two natural placements of knots for the ODE-Cox model, the ODE-AFT model, and the general linear transformation model respectively. Figures A.5-A.6 compare the integrated mean square errors (IMSE) of estimated functions, and the computing time associated with $K1$ and $K2$ from left to right for the ODE-Cox model and the ODE-AFT

Table A.7: Simulation results for two placements of knots under the Cox model.

N	Knots	$\beta_1 = 1$				$\beta_2 = 1$				$\beta_3 = 1$			
		Bias	SE	ESE	CP	Bias	SE	ESE	CP	Bias	SE	ESE	CP
1000	K1	.009	.153	.157	.952	.013	.157	.157	.952	.007	.152	.158	.961
	K2	.009	.153	.157	.953	.013	.157	.157	.951	.007	.152	.158	.960
2000	K1	.007	.106	.109	.956	-.001	.107	.109	.955	.007	.105	.109	.961
	K2	.006	.106	.110	.958	-.000	.107	.109	.956	.007	.105	.109	.960
4000	K1	.003	.076	.076	.936	-.002	.075	.076	.942	.000	.074	.076	.955
	K2	.002	.076	.076	.937	-.002	.075	.077	.944	-.000	.074	.077	.955
8000	K1	-.002	.053	.054	.953	-.000	.052	.054	.957	-.002	.054	.054	.947
	K2	-.001	.053	.054	.957	.001	.053	.054	.955	-.001	.053	.054	.949

Bias is the difference between the mean of estimates and the true value, and SE is the sample standard error of the estimates. ESE is the mean of the standard error estimators, and CP is the corresponding coverage proportion of 95% confidence intervals. In (K1), the interior knots are located at the $K_n = \lfloor N^{\frac{1}{5}} \rfloor$ quantiles of the distinct observation time points. In (K2), the interior knots equally separate the time interval from 0 to the maximum of observed times.

model. We can see that both two types of knot locations $K1$ and $K2$ return good estimates of parameters and standard errors. Overall, our numerical results suggest that knot selection does not appear critical for the proposed method in various simulation settings.

A.6.7 Model diagnostics

In this section, we use the linear transformation model as an example to illustrate how the unification of the proposed ODE framework along with the proposed estimation and inference procedure can be applied to model diagnostics and provide preliminary numerical results.

Recall that, under certain regularity conditions in Proposition 2.2.2, the linear transformation model, i.e.,

$$\begin{cases} \Lambda'(t) = \exp\bigl(x^T\beta + \gamma(t) + g(\Lambda(t))\bigr) \\ \Lambda(0) = 0 \end{cases},$$

reduces to the Cox model if and only if there exist positive constants c_1 and c_2 such that $g(t) = \log c_2 + (1 - c_1)\log t$, and it reduces to the AFT model if and only if there

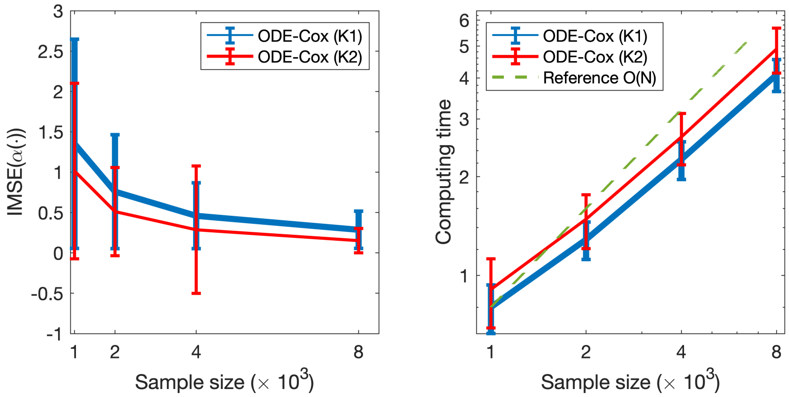

Figure A.5: Integrated mean square error (IMSE) of estimated $\alpha(\cdot)$ and the log-log plot of the computing time with respect to the sample size under the Cox model are provided from left to right.

Table A.8: Simulation results for two placements of knots under the AFT model.

N	Knots	$\beta_1 = 1$				$\beta_2 = 1$				$\beta_3 = 1$			
		Bias	SE	ESE	CP	Bias	SE	ESE	CP	Bias	SE	ESE	CP
1000	K1	-.014	.197	.191	.944	-.024	.209	.192	.931	-.032	.199	.192	.932
	K2	-.001	.194	.197	.954	-.010	.203	.197	.943	-.017	.195	.197	.945
2000	K1	-.010	.144	.137	.932	-.006	.144	.137	.937	-.005	.142	.137	.943
	K2	-.005	.143	.139	.941	.000	.143	.139	.942	-.001	.141	.139	.953
4000	K1	.000	.102	.097	.944	-.005	.100	.097	.944	-.002	.097	.097	.950
	K2	.002	.102	.098	.936	-.002	.100	.098	.938	.001	.097	.098	.950
8000	K1	-.006	.070	.069	.950	-.003	.068	.069	.967	-.004	.071	.069	.945
	K2	-.005	.070	.069	.951	-.001	.068	.069	.958	-.004	.071	.069	.942

Bias is the difference between the mean of estimates and the true value, and SE is the sample standard error of the estimates. ESE is the mean of the standard error estimators, and CP is the corresponding coverage proportion of 95% confidence intervals. In (K1), the interior knots are located at the $K_n = \lfloor N^{\frac{1}{7}} \rfloor$ quantiles of the estimated cumulative hazards under the Cox model. In (K2), the interior knots equally separate the interval from 0 to two times the maximum of the estimated cumulative hazards.

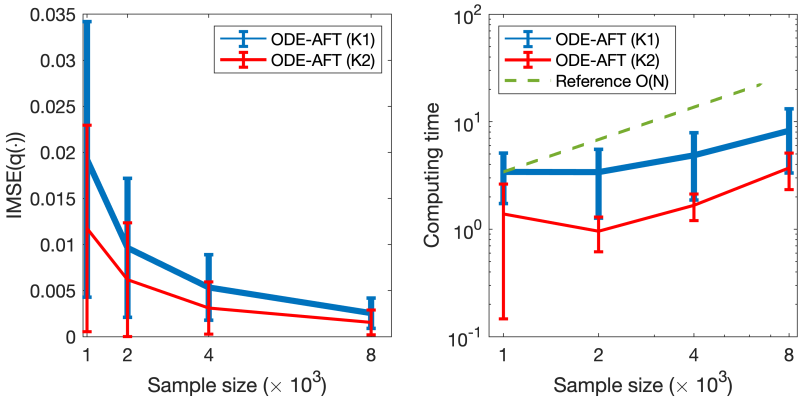

Figure A.6: Integrated mean square error (IMSE) of estimated $\alpha(\cdot)$ and the log-log plot of the computing time with respect to the sample size under the AFT model are provided from left to right.

Table A.9: Simulation results for two placements of knots under the general linear transformation model where both functions $\alpha(\cdot)$ and $q(\cdot)$ are unknown.

Setting	Knots	$\beta_2 = 1$				$\beta_3 = 1$			
		Bias	SE	ESE	CP	Bias	SE	ESE	CP
1)	K1	.008	.106	.107	.947	.012	.104	.107	.959
	K2	-.002	.098	.097	.946	.000	.095	.097	.955
2)	K1	-.019	.161	.151	.927	-.016	.159	.151	.938
	K2	.005	.152	.142	.936	.009	.155	.142	.931
3)	K1	-.014	.134	.131	.941	-.012	.131	.132	.945
	K2	.002	.131	.124	.936	.004	.131	.128	.939
4)	K1	.001	.092	.090	.939	.005	.091	.090	.954
	K2	-.002	.087	.084	.940	.002	.085	.084	.957

Bias is the difference between the mean of estimates and the true value, and SE is the sample standard error of the estimates. ESE is the mean of the standard error estimators, and CP is the corresponding coverage proportion of 95% confidence intervals.

exist positive constants c_1 and c_2 such that $\alpha(t) = \log c_2 + (c_1 - 1)\log t$ for $t > 0$. Therefore, to check whether the Cox or the AFT model is correctly specified, we can artificially create an additional basis function, $B(t)$, that does not belong to the linear span of $\{1, \log t\}$ and make inference about its coefficient.

Specifically, for checking the Cox model, we consider the following linear transformation model

$$\Lambda_x'(t) = \exp\bigl(a_1 \log(\Lambda_x) + a_2 B(\Lambda_x) + x^T\beta + \gamma(t)\bigr), \tag{A.53}$$

with unspecified $\gamma(\cdot)$. Then a local test of the null hypothesis $H_0 : a_2 = 0$ is a test for checking the Cox model specification. Correspondingly, a local test of the null hypothesis $H_0 : b_2 = 0$ under the model with unspecified $g(\cdot)$:

$$\Lambda_x'(t) = \exp\bigl(g(\Lambda_x) + x^T\beta + b_1 \log(t) + b_2 B(t)\bigr) \tag{A.54}$$

is a test for checking the AFT model specification. We note that, under H_0, the models (A.53) and (A.54) are identifiable up to a constant respectively, which is a direct result of Proposition 2.2.2. Thus, to guarantee the identifiability, we constrain $a_1 = 0$ and $b_1 = 0$ in the models (A.53) and (A.54) respectively. The proposed estimation and inference procedure can be applied to obtain the estimates of $(a_2, \beta, \gamma(\cdot))$ or $(b_2, \beta, g(\cdot))$ along with the local test of the corresponding H_0.

Next, we examine the above method under the simulation settings (1) and (3) in the main text, where the Cox and the AFT model are correctly specified respectively. We consider two choices of the known basis function: $B(t) = t$ and $B(t) = \log(1 + t)$. And we fit the unknown functions $\gamma(\cdot)$ and $g(\cdot)$ by cubic B-splines with the same placements of knots as described in the main text. The sample size varies from 1000 to 8000.

Table A.10 summarizes the estimates of the coefficients of interests based on 1000 replications. We can see that the bias of the estimator is nearly negligible in all settings. When the sample size is large, the coverage proportion of 95% confidence intervals, where the standard error estimator is obtained by inverting the estimated

Table A.10: Simulation results for checking the Cox and the AFT model specification.

Setting		$B(t) = t$				$B(t) = \log(1 + t)$			
	N	1000	2000	4000	8000	1000	2000	4000	8000
Cox is correctly specified: $a_2 = 0$	Bias	.023	.007	.006	.003	.041	.000	.009	.007
	SE	.133	.087	.056	.033	.459	.314	.206	.112
	ESE	.136	.092	.062	.043	.467	.325	.226	.159
	CP	.949	.954	.956	.968	.943	.945	.956	.977
AFT is correctly specified: $b_2 = 0$	Bias	-.003	.000	.003	.000	-.011	.003	.002	-.002
	SE	.182	.130	.083	.067	.423	.295	.198	.156
	ESE	.214	.155	.111	.079	.515	.376	.271	.195
	CP	.968	.960	.978	.961	.975	.963	.979	.964

Bias is the difference between the mean of estimates and the true value, and SE is the sample standard error of the estimates. ESE is the mean of the standard error estimators, and CP is the corresponding coverage proportion of 95% confidence intervals.

information matrix of all parameters including the coefficients of spline bases, is slightly greater than the nominal level. The corresponding t-statistics would lead to a conservative local test for H_0. We also find that the sample standard errors of the estimates vary with the choice of the basis $B(\cdot)$, and the ability to detect the model specification depends on $B(\cdot)$ as well. It may be preferable to make both functions $\gamma(\cdot)$ and $g(\cdot)$ unknown in the nonparametric linear transformation model for model diagnostics, which requires the asymptotic distributional theory for the functional parameters. We leave this interesting direction for future work.

A.7 Partial Likelihood Based Methods

The Cox model (Cox, 1972) and its extensions such as DeepSurv (Katzman et al., 2018) and Cox-Time (Kvamme et al., 2019) consider the hazard function in a semi-parametric way. Specifically, the conditional hazard function is factorized into two terms: a non-parametric baseline hazard function and a parametric relative risk function, that is

$$\lambda_x(t) = h_0(t) \exp(g(t, x; \theta)).$$

The **Cox** model assumes a time-invariant linear relative risk where $g(t, x; \theta) = x^T\theta$. Subsequently, **DeepSurv** allows the relative risk to be a nonlinear function of feature x, i.e. $g(t, x; \theta) = g(x; \theta)$, but the proportional hazard assumption still holds; **Cox-Time** further allows the relative risk function to depend on time, which can handle the non-proportional hazard. In particular, DeepSurv and Cox-Time use neural networks to model $g(x; \theta)$ and $g(t, x; \theta)$.

All the above models are fitted in two steps: the parameters in the relative risk function are learned through maximizing the partial likelihood function (Cox, 1975); the non-parametric cumulative baseline hazard function is obtained through the Breslow's estimator (Lin, 2007) given the fitted relative risk in the first step.

Partial likelihood. The partial likelihood function is defined as

$$\mathcal{PL}(\theta; D) = \prod_{i:\Delta_i=1} \frac{\exp(g(y_i, x_i; \theta))}{\sum_{j\in R_i} \exp(g(y_i, x_j; \theta))},$$

where $R_i = \{j : y_j \geq y_i\}$ denotes the set of individuals who survived longer than the i-th individual, which is the so called *at-risk* set. The estimator of θ is obtained by minimizing the negative log-partial likelihood function, that is

$$\min_\theta \sum_{i:\Delta_i=1} [-g(y_i, x_i; \theta) + \log \sum_{j\in R_i} \exp(g(y_i, x_j; \theta))].$$

The partial likelihood function of each individual requires the access to the data of all individuals in the at-risk set. Hence, stochastic gradient decent (SGD) based algorithms cannot be directly applied. Although we can naively sample a mini-batch and restrict the at-risk set to individuals who are included in the current mini-batch in practice, there is a lack of theoretical justification.

Breslow's estimator. In order to obtain the predicted survival function, we need to estimate the cumulative hazard function. For models with the proportional hazard assumption, the estimated cumulative hazard function can be written as

$$\hat{\Lambda}_x(t) = \int\limits_0^t \hat{h}_0(s)\,ds \cdot \exp\Big(g(x;\hat{\theta})\Big) = \hat{H}_0(t)\exp\Big(g(x;\hat{\theta})\Big),$$

where $\hat{H}_0$ is the estimated cumulative baseline hazard function. The Breslow's estimator for H_0 is given by

$$\hat{H}_0(t) = \sum_{i:y_i \le t} \frac{\Delta_i}{\sum_{j\in R_i}\exp\Big(g(x_j;\hat{\theta})\Big)}.$$

For Cox-Time with non-proportional hazard, the estimated cumulative hazard function is given by

$$\hat{\Lambda}_x(t) = \sum_{i:y_i \le t} \frac{\Delta_i}{\sum_{j\in R_i}\exp\Big(g(y_i, x_j;\hat{\theta})\Big)}\exp\Big(g(y_i, x;\hat{\theta})\Big).$$

The survival function can then be estimated by $\hat{S}_x(t) = \exp\Big(-\hat{\Lambda}_x(t)\Big)$.

A.8 Discrete-Time Methods

In the discrete-time setting, the range of possible values of the event time T is divided into a set of disjoint intervals through pre-specified break points $\{t_0 = 0, t_1, \cdots, t_L\}$. Denote the intervals by $I_l = (t_{l-1}, t_l], l = 1, \cdots, L$. Suppose the probability of occurrence of the event in time interval I_l is $p_l(x) \ge 0$ with $\sum_{l=1}^{L} p_l(x) = 1$. The cumulative distribution F_l and survival functions S_l are, respectively

$$F_l(x) = \mathcal{P}\{T \le t_l | X = x\} = \sum_{l=1}^{l} p_j(x),$$

$$S_l(x) = \mathcal{P}\{T > t_l | X = x\} = 1 - F_l = 1 - \sum_{j=1}^{l} p_j(x).$$

The conditional hazard probability $\lambda_l(x)$ is the probability that the event occurs in interval I_l conditional on the survival up to the beginning of I_l, which could also determine the survival function through

$$\lambda_l(x) = \mathcal{P}\{T \in I_l | T \ge t_{l-1}, X = x\} = \frac{p_l(x)}{S_{l-1}(x)}, \quad S_l(x) = \prod_{j=1}^{l}(1 - \lambda_j(x)).$$

Under the conditional independence assumption of the event time and the censoring time given features, the likelihood function is proportional to

$$\prod_i p_{l_i}(x_i)^{\Delta_i}(1-\sum_{j=1}^{l_i-1}p_j(x_i))^{1-\Delta_i} = \prod_i[\lambda_{l_i}(x_i)\prod_{j=1}^{l_i-1}(1-\lambda_j(x_i))],$$

where l_i is the index of time interval satisfying $t_{l_i-1} < y_i \leq t_{l_i}$.

DeepHit (Lee et al., 2018) models the probability mass function where the output of the neural network is a vector $[p_1(x), \cdots, p_L(x)]$. In addition to the negative log-likelihood (NLL) loss function, DeepHit considers another differentiable surrogate ranking loss tailored for time dependent concordance index, that is

$$\mathcal{L}_2 = \sum_{i:\Delta_i=1}\sum_{j:l_i<l_j}\eta(F_{l_i}(x_i), F_{l_i}(x_j)),$$

where $\eta(x,y) = \exp\left(\frac{-(x-y)}{\sigma}\right)$ and σ is a hyperparameter. They introduce another hyperparameter α to control the trade-off between the ranking loss and the NLL loss.

Nnet-Survival (Gensheimer and Narasimhan, 2019) models the conditional hazard probability where the output of the neural network is a vector $[\lambda_1(x), \cdots, \lambda_L(x)]$, and it is learned by maximizing the likelihood function.

A.9 Hyperparameter Tuning

We list the tuning ranges of hyperparameters for all neural network based models on three datasets in Table A.11, where $\{\cdot\}$ represents the discrete search space and $[\cdot]$ represents the continuous search space[1]. Specifically, we tune the rate of dropout and batch normalization for DeepSurv, DeepHit, Nnet-Survival, and Cox-Time. We treat the number of time intervals as a hyperparameter for DeepHit and Nnet-Survival. We also tune two hyperparameters, α and σ, associated with the surrogate ranking loss in DeepHit. Since the three datasets are of different sizes, we use different

[1] For the number of neurons, a real number is first sampled from the continuous space and then rounded to the closest integer.

search ranges for the batch size: $\{32, 64, 128, 256\}$ for METABRIC, $\{128, 256, 512\}$ for SUPPORT, and $\{512, 1024\}$ for MIMIC and MIMIC-SEQ. The discrete models (DeepHit and Nnet-Survival) appear to be sensitive to the number of time intervals on different datasets. Therefore we search the number of time intervals for these two discrete models from $\{10, 50, 100, 200, 400\}$ for the smaller datasets, METABRIC and SUPPORT, and from $\{50, 100, 200, 400, 800\}$ for the larger datasets, MIMIC and MIMIC-SEQ.

Number of dense hidden layers	$\{1, 2, 4\}$
Number of neurons in each dense hidden layer	$[2^2, 2^7]$
Number of neurons in each GRU hidden layer	$[2^3, 2^8]$
Learning rate	$[10^{-4.5}, 10^{-1.5}]$
Weight decay	$[10^{-9}, 10^{-4}]$
Momentum	$[0.85, 0.99]$
Dropout (DeepHit, DeepSurv, Nnet-Survival, Cox-Time)	$\{0, 0.1, 0.5\}$
Batch normalization (DeepHit, DeepSurv, Nnet-Survival, Cox-Time)	$\{\text{True}, \text{False}\}$
α (Surrogate ranking loss in DeepHit)	$[0, 1]$
σ (Surrogate ranking loss in DeepHit)	$\{0.25, 1, 5\}$

Table A.11: Tuning ranges of hyperparameters for experiments on the real-world datasets.

APPENDIX B

Appendix of Chapter III

This Appendix is structured as follows. We provide the detailed description of the separate estimation method in Algorithms B.1 and B.2 and their initialization algorithms in Section B.1. In addition, we present the proof of Proposition 3.2.1 in Section B.2; the proof of Theorem 3.2.2 in Section B.3; the proof of the identifiability conditions in Proposition 3.3.1 and 3.3.2 in Section B.4; the proof of the error bounds for estimating the latent polar variables v in Theorem 3.5.1 in Section B.5; the proof of the error bounds for the joint estimation method in Theorem 3.5.2 and Corollary 3.5.1 in Section B.6; and the proof of the one-step improvement of the joint estimation method against the separate estimation method in Proposition 3.5.2 and the discussion of when the conditional independence assumption and the prerequisite error rate of $(\bar{w}, \bar{\gamma})$ in Proposition 3.5.2 hold in Section B.7.

B.1 Algorithms for the Separate Estimation Method

The separate estimation method estimates parameters (α, Z) and v separately by minimizing $\mathcal{L}_e(\alpha, Z)$ and $\mathcal{L}_s(v)$ respectively. Below, we present two gradient descent algorithms to solve the non-convex optimization problems in (3.6) and (3.7) respectively.

First, to estimate parameters (α, Z), we adopt the projected gradient descent algorithm proposed in Ma et al. (2020) because of the theoretical guarantee and scalability to large networks. Specifically, we first update parameters α and Z towards the opposite direction of the gradient, and then project the updated Z by centering its columns at each iteration. Although another projection constraint is needed for retaining the range of parameters for theoretical justification in Section 3.5, our simulation studies suggest that Algorithm B.1 itself works well in practice. We provide the detailed description of the method in Algorithm B.1. We set the step size $\tau_z = \tau/\|Z_0\|_{op}^2$ and $\tau_\alpha = \tau/(2n)$ with some small constant τ.

Algorithm B.1: The projected gradient descent algorithm for estimating (α, Z)

```
Input: absolute adjacency network $|A| \in \{0,1\}^{n\times n}$, latent space dimension
       $k \geq 1$, number of iterations $T$, initial values $(\alpha_0, Z_0)$, step sizes
       $(\tau_\alpha, \tau_z)$
for $t = 0, \ldots, T-1$ do
    $\tilde{Z}_{t+1} = Z_t - \tau_z \boldsymbol{\nabla}_Z \mathcal{L}_e = Z_t + 2\tau_z(|A| - \sigma(\Theta_t))Z_t$;
    $\alpha_{t+1} = \alpha_t - \tau_\alpha \boldsymbol{\nabla}_\alpha \mathcal{L}_e = \alpha_t + 2\tau_\alpha(|A| - \sigma(\Theta_t))1_n$;
    $Z_{t+1} = J_n \tilde{Z}_{t+1}$;
end
Output: $(\hat{\alpha}, \hat{Z}) = (\alpha_T, Z_T)$
```

Next, to estimate the latent polar variables v, we develop a fast gradient descent algorithm, which is summarized in Algorithm B.2. We set the step size as $\tau_v = \tau/\|v_0\|^2$ with some small constant τ.

Algorithm B.2: The gradient descent algorithm for estimating v

```
Input: signed adjacency network $A \in R^{n\times n}$, number of iterations $T$, initial
       values $v_0$, step sizes $\tau_v$
for $t = 0, \ldots, T-1$ do
    $v_{t+1} = v_t - \tau_v \boldsymbol{\nabla}_v \mathcal{L}_s = v_t + 2\tau_v(|A| \circ (A+1)/2 - |A| \circ \sigma(\eta_t))v_t$;
end
Output: $\hat{v} = v_T$
```

Proposition 3.5.1 and Theorem 3.5.1 indicate that Algorithms B.1 and B.2 require relatively good initializations of (α, Z, v) for the guarantee of convergence and reaching statistical precision. In the rest of this section, we first introduce two theoretically justified initialization algorithms and then present simulation studies to examine the impact of initialization on the proposed non-convex optimization algorithms.

B.1.1 Initialization algorithms

For the initialization of Algorithm B.1, we adopt the theoretically justified initialization algorithm in Ma et al. (2020), which first estimates the edge probability matrix P by universal singular value thresholding (USVT) (Chatterjee, 2015) and then inverts the logit-transformed probability matrix to obtain the initial estimates (α_0, Z_0). We adapt the algorithm so that no edge covariates are involved and summarize it in Algorithm B.3. The following proposition indicates that Algorithm B.3 obtains good enough initial estimates with proper choice of the threshold δ.

Proposition B.1.1 (Ma et al. (2020))**.** *Suppose that $\|Z^*Z^{*\top}\|_F \geq c_0 n$ for some constant $c_0 > 0$, then there exists constant c_1 such that with probability at least $1-n^{-c_1}$, for any $n \geq C(k, M_1, \kappa_{Z^*})$, the output of Algorithm B.3 with $\delta \geq 1.1\sqrt{n}$ satisfies the initialization condition in Proposition 3.5.1.*

Algorithm B.3: Initialization of Algorithm B.1 by USVT

```
Input: absolute adjacency network $|A| \in \{0,1\}^{n\times n}$, latent space dimension
       $k \geq 1$, threshold $\delta$
Let $\tilde{P} = \sum_{s_i \geq \delta} s_i u_i v_i^\top$ where $\sum_{i=1}^n s_i u_i v_i^\top$ is the SVD of $|A|$; obtain $\hat{P}$ by
  elementwisely projecting $\tilde{P}$ to the interval $[e^{-M_1}/2, 1/2]$; let
  $\hat{\Theta} = \mathrm{logit}((\hat{P} + \hat{P}^\top)/2)$;
Let $\alpha_0 = \hat{\Theta}1_n/n - 1_n 1_n^\top \hat{\Theta} 1_n/(2n^2)$ that minimizes $\|\hat{\Theta} - \alpha 1_n^\top - 1_n \alpha^\top\|_F$ w.r.t.
  $\alpha$;
Let $\hat{G} = P_{\mathbb{S}_+^n}(R)$ where $R = J_n(\hat{\Theta} - \alpha_0 1_n^\top - 1_n \alpha_0^\top)J_n = J_n \hat{\Theta} J_n$;
Let $Z_0 = U_k D_k^{1/2}$ where $U_k D_k U_k^\top$ is the top-k singular value components of $\hat{G}$;
Output: $\alpha_0, Z_0$
```

For the initialization of Algorithm B.2, we propose an initialization algorithm based on USVT with missing entries. In particular, we apply USVT to the matrix B, where $B_{ij} = (1 + A_{ij})/2$ if there is an edge between node i and node j and $B_{ij} = 0$ otherwise. Then we similarly invert the logit-transformed probability matrix. We summarize it in Algorithm B.4 below. The following proposition justifies the proposed initialization algorithm when each edge is observed independently with uniform probability p, the proof of which is given in Subsection B.1.3.

Proposition B.1.2. *Suppose that each edge is observed independently with uniform probability* $p \asymp e^{-M_1}$ *and* $\|v^*\|^2 \geq c_0 n$ *for some constant* $c_0 > 0$*, then there exists constant* c_1 *such that with probability at least* $1 - n^{-c_1}$*, for any* $n \geq C(M_3)$*, the output of Algorithm B.4 with* $\delta \geq 2.01\sqrt{n\hat{p}}$ *satisfies the initialization condition in Theorem 3.5.1. Here* $\hat{p}$ *is the proportion of observed edges.*

```
Algorithm B.4: Initialization of Algorithm B.2 by USVT
  Input: signed adjacency network $A \in R^{n \times n}$, threshold $\delta$
  Let $\tilde{Q} = \sum_{s_i \geq \delta} s_i u_i v_i^\top$ where $\sum_{i=1}^n s_i u_i v_i^\top$ is the SVD of $B = |A| \circ (1 + A)/2$;
    obtain $\hat{Q}$ by elementwisely projecting $\tilde{Q}$ to the interval $[e^{-M_3}/2, 1 - e^{M_3}/2]$;
  Let $\tilde{\eta} = \mathrm{logit}((\hat{Q} + \hat{Q}^\top)/2)$ and $\hat{\eta} = P_{\mathbb{S}_+^n}(\tilde{\eta})$;
  Let $v_0 = U_1 D_1^{1/2}$ where $U_1 D_1 U_1^\top$ is the top-1 singular value components of $\hat{\eta}$;
  Output: $v_0$
```

Although, in our case, an edge between nodes i and j is observed with non-uniform probability P_{ij}, the proposed algorithm works well in all the simulation studies.

B.1.2 Analyzing impact of initialization through simulation studies

In this section, we analyze the impact of initialization on the estimation errors and the number of iterations till convergence. Here we compare the proposed initialization algorithm and random initialization for the separate estimation Algorithms B.1 and B.2 . To this end, we fix $k = 2$ and vary $n \in \{500, 1000, 2000, 4000\}$. For Algorithm B.3, we choose $\delta = 2.01\sqrt{n}$ and set $M_1 = \log(10)$. For Algorithm B.4, we choose $\delta = 2.01\sqrt{n\hat{p}}$ with $\hat{p} = \sum_{ij} A_{ij}^2/n^2$ and set $M_3 = \log(5)$. We summarize the relative estimation errors of Z and v after convergence of the algorithm, and the number of iterations required for convergence in Figure B.1.

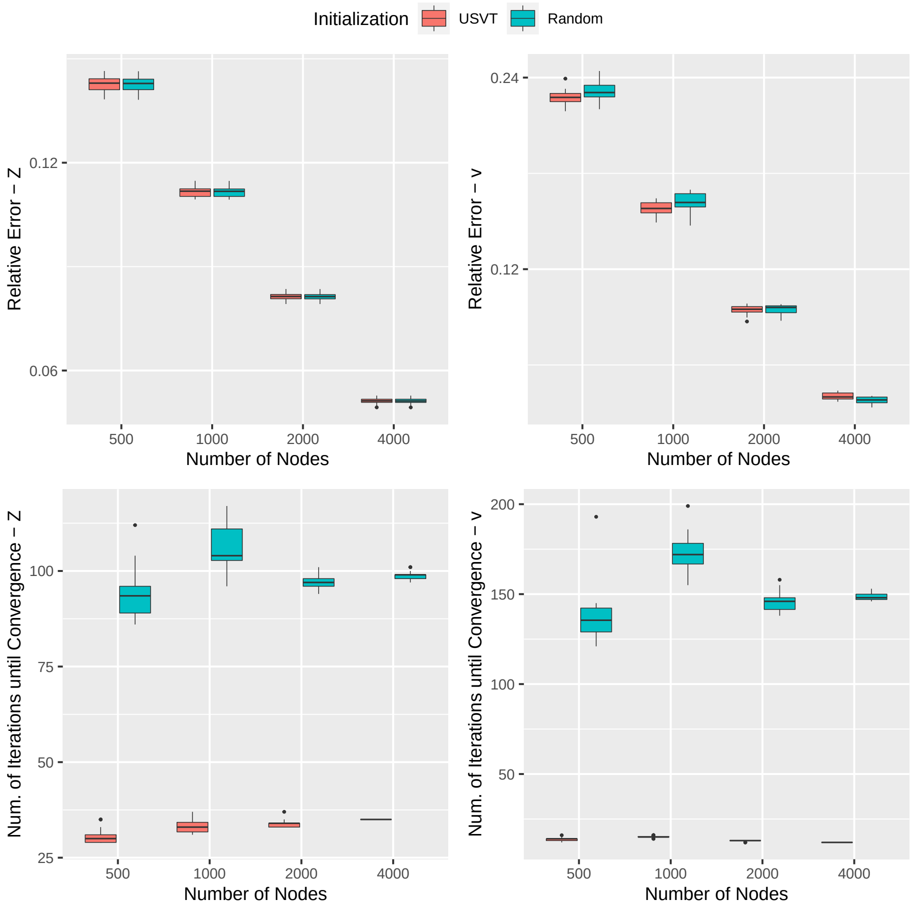

Figure B.1: Upper row: log-log plots of relative errors after convergence with respect to the network size n. Lower row: the number of iterations until convergence with respect to the network size n. The latent position vector dimension is fixed as $k = 2$.

We can see from Figure B.1 that the proposed initialization algorithms and random initialization both achieve similar estimation errors for Z and v when the algorithms converge, while the proposed algorithms requires fewer iterations to converge.

B.1.3 Proof of Proposition B.1.2

Proof of Proposition B.1.2. Based on the established error rate of USVT in Chatterjee (2015, Theorem 2.7), we have

$$\frac{1}{n^2}\|\hat{Q} - Q^*\|_F^2 \le C(M_1, M_3)n^{-1/3},$$

where $C(M_1, M_3)$ is a constant that depends on M_1 and M_3. Since the function logit$(\cdot)$ is $\frac{4e^{M_3}}{2-e^{-M_3}}$-Lipchitz continuous on the interval $[e^{-M_3}/2, 1-e^{M_3}/2]$ and Q^* is symmetric, we obtain that

$$\frac{1}{n^2}\|\tilde{\eta} - \eta^*\|_F^2 \le \frac{C'(M_3)}{n^2}\|(\hat{Q} + \hat{Q}^\top)/2 - Q^*\|_F^2 \le \tilde{C}(M_1, M_3)n^{-1/3}.$$

Further, by definition of $\eta_0 = v_0 v_0^\top$ and $\hat{\eta}$, and $\eta^* \in \mathbb{S}_+^n$ and $\operatorname{rank}(\eta^*) = 1$, we have $\|\hat{\eta} - \tilde{\eta}\|_F \le \|\eta^* - \tilde{\eta}\|_F$ and $\|\eta_0 - \hat{\eta}\|_F \le \|\eta^* - \hat{\eta}\|_F$. Therefore, it follows that

$$\|\eta_0 - \eta^*\|_F \le \|\eta_0 - \hat{\eta}\|_F + \|\hat{\eta} - \eta^*\|_F \le 2\|\hat{\eta} - \eta^*\|_F \le 2\|\hat{\eta} - \tilde{\eta}\|_F + 2\|\tilde{\eta} - \eta^*\|_F \le 4\|\tilde{\eta} - \eta^*\|_F.$$

Then, by Lemma B.5.6, we have

$$\begin{aligned}\|\Delta_{v_0}\|^2 &\le \frac{\|\eta_0 - \eta^*\|_F^2}{2(\sqrt{2}-1)\|v^*\|^2} \le \frac{8\|\tilde{\eta} - \eta^*\|_F^2}{(\sqrt{2}-1)\|v^*\|^2} \le \frac{8\tilde{C}(M_1, M_3)n^{-1/3}}{(\sqrt{2}-1)\|v^*\|^2}n^2 \\ &\le \frac{8\tilde{C}(M_1, M_3)n^{-1/3}}{(\sqrt{2}-1)\|v^*\|^2}\frac{\|v^*\|^4}{c_0^2} = C_1(M_1, M_3)n^{-1/3}\|v^*\|^2,\end{aligned}$$

where $C_1(M_1, M_3)$ is a constant that depends on M_1 and M_3. We can choose large enough n such that the initialization condition in Theorem 3.5.1 holds. □

B.2 Proof of Proposition 3.2.1

To prove Proposition 3.2.1, we need the following lemma.

Lemma B.2.1. *For any $\ell \ge 1$ random variables A_i, $1 \le i \le \ell$, taking values in $\{1, 0, -1\}$, $E(A_1 \cdots A_\ell \big| |A_1 \cdots A_\ell| = 1) > 0$ if and only if $E(A_1 \cdots A_\ell) > 0$.*

Proof. Note that $A_1 \cdots A_\ell$ can only take values in $\{1, 0, -1\}$, it follows that

$$\begin{aligned}&E(A_1 \cdots A_\ell) = P(A_1 \cdots A_\ell = 1) - P(A_1 \cdots A_\ell = -1) \\ =&[P(A_1 \cdots A_\ell = 1\big||A_1 \cdots A_\ell| = 1) - P(A_1 \cdots A_\ell = -1\big||A_1 \cdots A_\ell| = 1)]P(|A_1 \cdots A_\ell| = 1)\end{aligned}$$

$=E(A_1 \cdots A_\ell \big| |A_1 \cdots A_\ell| = 1) P(|A_1 \cdots A_\ell| = 1).$

Therefore, $E(A_1 \cdots A_\ell \big| |A_1 \cdots A_\ell| = 1) > 0$ if $E(A_1 \cdots A_\ell) > 0$. The other direction holds automatically because $E(A_1 \cdots A_\ell \big| |A_1 \cdots A_\ell| = 1) > 0$ implies $P(|A_1 \cdots A_\ell| = 1) > 0$. □

Proof of Proposition 3.2.1. For a network $A \sim G(n, \mathcal{U}, P_z, B, f)$, we have

$$\begin{aligned} E(A_{ij}|u_i, u_j) &= P(A_{ij} = 1|u_i, u_j) - P(A_{ij} = -1|u_i, u_j) \\ &= \left[P(A_{ij} = 1|u_i, u_j, |A_{ij}| = 1) - P(A_{ij} = -1|u_i, u_j, |A_{ij}| = 1)\right] P(|A_{ij}| = 1|u_i, u_j) \\ &= \left[\sigma(f(u_i, u_j)) - (1 - \sigma(f(u_i, u_j)))\right] B(u_i, u_j) \\ &= [2\sigma(f(u_i, u_j)) - 1] B(u_i, u_j). \end{aligned}$$

Given that $B(u_i, u_j) > 0$, we have

$$sign(E(A_{ij}|u_i, u_j)) = sign(2\sigma(f(u_i, u_j)) - 1) = sign(f(u_i, u_j)).$$

Therefore, for a symmetric function f satisfying (3.1), it follows that

$$E(A_{ij}|u_i, u_j) \cdot E(A_{j\ell}|u_j, u_\ell) \cdot E(A_{\ell i}|u_\ell, u_i) > 0, \text{ with probability 1.}$$

By the conditional independence among edges given their latent vectors, we have

$$\begin{aligned} E(A_{ij} A_{j\ell} A_{\ell i}) &= E\big(E(A_{ij} A_{j\ell} A_{\ell i}|u_i, u_j, u_\ell)\big) \\ &= E\big(E(A_{ij}|u_i, u_j) E(A_{j\ell}|u_j, u_\ell) E(A_{\ell i}|u_\ell, u_i)\big) > 0. \end{aligned}$$

By Lemma B.2.1, the requirement in Definition 3.1.1 holds, which completes the proof of Proposition 3.2.1. □

B.3 Proof of Theorem 3.2.2

Proof of Theorem 3.2.2. We first prove the direct part. Suppose the symmetric function f satisfies that $f(a, b) \cdot f(b, c) \cdot f(c, a) > 0$ holds for any $a, b, c \in \mathcal{U}$, where $\mathcal{U} \subset \mathcal{U}_0$ with $P_u(\mathcal{U}) = 1$. Then, it follows directly that, for any $a, b \in \mathcal{U}$, $f(a, b) \neq 0$ and $f(a, a) > 0$ since $[f(a, a)]^3 > 0$. If the function f is not always positive on $\mathcal{U} \times \mathcal{U}$, then there exists $a_1 \neq a_2 \in \mathcal{U}$ such that $f(a_1, a_2) < 0$. We construct two subsets of $\mathcal{U}$ by

$$S = \{s \in \mathcal{U} : f(a_1, s) > 0\}, \quad T = \{t \in \mathcal{U} : f(a_1, t) < 0\}.$$

Since $f(a_1, a_1) > 0$ and $f(a_1, a_2) < 0$, we have $a_1 \in S$ and $a_2 \in T$, and thus they are nonempty. It is obvious that S and T are disjoint. For any $u \in \mathcal{U}$, since $f(a_1, u) \neq 0$, u belongs to either S or T. Therefore, $S \cup T = \mathcal{U}$ and $S \cap T = \varnothing$.

Then, we show that $sign(f(a, b)) = \mathbb{1}(a \in S) \cdot \mathbb{1}(b \in S)$ for any $a, b \in \mathcal{U}$. For any $s_1, s_2 \in S$, we have $f(s_1, s_2) \cdot f(s_2, a_1) \cdot f(a_1, s_1) > 0$. By the construction of S, we have $f(s_2, a_1) > 0$ and $f(a_1, s_1) > 0$, and it follows that $f(s_1, s_2) > 0$. Similarly, we can show that $f(t_1, t_2) > 0$ holds for any $t_1, t_2 \in T$. For any $s \in S$ and $t \in T$, we have $f(s, t) \cdot f(t, a_1) \cdot f(a_1, s) > 0$. By the construction of S and T, we have $f(a_1, s) > 0$ and $f(t, a_1) < 0$, and thereby $f(s, t) < 0$. This has completed the proof of the direct part.

Next, we prove the converse part. In case (i), a positive function f on $\mathcal{U} \times \mathcal{U}$ automatically satisfies that $f(a, b) \cdot f(b, c) \cdot f(c, a) > 0$ for any $a, b, c \in \mathcal{U}$. In case (ii), suppose there exists two nonempty subsets S and T, with $S \cup T = \mathcal{U}$ and $S \cap T = \varnothing$, such that $sign(f(a, b)) = \mathbb{1}(a \in S) \cdot \mathbb{1}(b \in S)$ for any $a, b \in \mathcal{U}$. Then for any $a, b, c \in \mathcal{U}$,

$$sign(f(a, b)) \cdot sign(f(b, c)) \cdot sign(f(c, a)) = [\mathbb{1}(a \in S)]^2 [\mathbb{1}(b \in S)]^2 [\mathbb{1}(c \in S)]^2 > 0,$$

or saying that $f(a, b) \cdot f(b, c) \cdot f(c, a) > 0$, which completes the proof. □

B.4 Proof of Proposition 3.3.1 and 3.3.2 (Identifiability Condition)

The proof of Proposition 3.3.1 is almost identical to that of Proposition 3.3.2 and so we omit it here. Before presenting the proof of Proposition 3.3.2, we need the following lemma.

Lemma B.4.1. *For any two vectors $v_1, v_2 \in \boldsymbol{R}^n$ with $v_1 v_1^\top = v_2 v_2^\top$, we have $v_1 = v_2$*

or $v_1 = -v_2$.

Proof. If $\|v_1\| = 0$, after right multiplying v_2 on both sides of $v_1 v_1^\top = v_2 v_2^\top$, we have $\mathbf{0}_n = v_2\|v_2\|^2$. So $v_2 = v_1 = \mathbf{0}_n$. If $\|v_1\| \neq 0$, after right multiplying v_1 on both sides of $v_1 v_1^\top = v_2 v_2^\top$, we have $v_1\|v_1\|^2 = v_2(v_2^\top v_1)$. This implies that $v_1 = \kappa v_2$ with some constant κ. By plugging $v_1 = \kappa v_2$ into $v_1 v_1^\top = v_2 v_2^\top$, we obtain $\kappa^2 v_2 v_2^\top = v_2 v_2^\top$. So $\kappa^2 = 1$ and then $v_1 = v_2$ or $v_1 = -v_2$. □

Proof of Proposition 3.3.2. Suppose two sets of parameters (α, Z, w, u) and $(\bar{\alpha}, \bar{Z}, \bar{w}, \bar{u})$ yield the same edge connection probability and the same edge sign probability, i.e.,

$$\alpha 1_n^\top + 1_n \alpha^\top + ZZ^\top = \bar{\alpha} 1_n^\top + 1_n \bar{\alpha}^\top + \bar{Z}\bar{Z}^\top, \tag{B.1}$$

$$(Zw + u1_n)(Zw + u1_n)^\top = (\bar{Z}\bar{w} + \bar{u}1_n)(\bar{Z}\bar{w} + \bar{u}1_n)^\top. \tag{B.2}$$

First, we show that $\alpha = \bar{\alpha}$. By Assumption A1, we have $J_n Z = Z$ with $J_n = I_n - 1_n 1_n^\top/n$. Thus, $ZZ^\top 1_n = ZZ^\top J_n 1_n = ZZ^\top(1_n - 1_n) = 0_n$. Similarly, we have $\bar{Z}\bar{Z}^\top 1_n = 0_n$. Then, left multiplying 1_n to both sides in (B.1) yields

$$\alpha 1_n^\top 1_n + 1_n \alpha^\top 1_n = \bar{\alpha} 1_n^\top 1_n + 1_n \bar{\alpha}^\top 1_n,$$

which is equivalent to $\alpha - \bar{\alpha} = -\sum_{i=1}^n (\alpha_i - \bar{\alpha}_i)/n \cdot 1_n$, i.e.,

$$\alpha_1 - \bar{\alpha}_1 = \cdots = \alpha_n - \bar{\alpha}_n = -\frac{1}{n}\sum_{i=1}^n (\alpha_i - \bar{\alpha}_i).$$

Therefore, $\alpha_i - \bar{\alpha}_i = 0$, for $1 \leq i \leq n$.

Now, we have $ZZ^\top = \bar{Z}\bar{Z}^\top$ from (B.1). By the fact that, for any real matrix M, $rank(M) = rank(MM^\top)$, we obtain $rank(\bar{Z}) = rank(\bar{Z}\bar{Z}^\top) = rank(ZZ^\top) = rank(Z)$. By Assumption A2, it follow that $\bar{Z}$ is also full rank, i.e., $rank(\bar{Z}) = rank(Z) = k$. Denote the compact singular value decomposition of Z and $\bar{Z}$ by $Z = U_1\Sigma_1 V_1^\top$ and $\bar{Z} = U_2 \Sigma_2 V_2^\top$ respectively, where $U_i \in \mathbf{R}^{n\times k}$ with $U_i^\top U_i = I_k$, $\Sigma_i \in \mathbf{R}^{k\times k}$ is an invertible diagonal matrix, and $V_i \in O(k)$, for $i = 1, 2$. Since $ZZ^\top = \bar{Z}\bar{Z}^\top$, it follows that $U_1\Sigma_1^2 U_1^\top = U_2 \Sigma_2^2 U_2^\top$. By right multiplying $U_1\Sigma_1^{-1}$ to both sides, we have

$$U_1\Sigma_1 = U_2\Sigma_2^2 U_2^\top U_1 \Sigma_1^{-1}. \tag{B.3}$$

Together with $V_2^\top V_2 = I_k$, we obtain that

$$Z = (U_1\Sigma_1)V_1^\top = U_2\Sigma_2(V_2^\top V_2)\Sigma_2 U_2^\top U_1\Sigma_1^{-1}V_1^\top = \bar{Z}V_2QV_1^\top,$$

where $Q = \Sigma_2 U_2^\top U_1 \Sigma_1^{-1}$. By further left multiplying $\Sigma_1^{-1}U_1^\top$ to both sides in (B.3), we have

$$I_k = \Sigma_1^{-1}U_1^\top \cdot U_2\Sigma_2^2U_2^\top U_1\Sigma_1^{-1} = (\Sigma_1^{-1}U_1^\top U_2\Sigma_2)(\Sigma_2U_2^\top U_1\Sigma_1^{-1}) = Q^\top Q.$$

So $Q \in O(k)$. It follows that $Z = \bar{Z}O$, where $O = V_2QV_1^\top \in O(k)$.

Next, by Lemma B.4.1 and Equation (B.2), there exists $\kappa \in \{-1, 1\}$ such that $Zw + \gamma 1_n = \kappa(\bar{Z}\bar{w} + \bar{\gamma}1_n)$. Since $Z = \bar{Z}O$, we have $(\kappa\bar{\gamma} - \gamma)1_n = \bar{Z}(Ow - \kappa\bar{w})$. We left multiply $1_n^\top$ to both sides and obtain that

$$n(\kappa\bar{\gamma} - \gamma) = (\kappa\bar{\gamma} - \gamma)1_n^\top 1_n = 1_n^\top \bar{Z}(Ow - \kappa\bar{w}) = 1_n^\top J_n\bar{Z}(Ow - \kappa\bar{w}) = 0,$$

where the last two equalities hold because $J_n\bar{Z} = \bar{Z}$ and $1_n^\top J_n = 0_n^\top$. Thus, $\gamma = \kappa\bar{\gamma}$ and $\bar{Z}(Ow - \kappa\bar{w}) = 0_n$. Since $\bar{Z}$ is full rank, we have $Ow - \kappa\bar{w} = 0_k$, i.e., $w = \kappa O^\top\bar{w}$. This has completed the proof of the direct part.

Finally, for the converse part, the proof is straightforward by verifying Equations (B.1) and (B.2) given that $\alpha = \bar{\alpha}, Z = \bar{Z}O, w = \kappa O^\top\bar{w}$, and $\gamma = \kappa\bar{\gamma}$. □

B.5 Proof of Theorem 3.5.1

In this section, we establish the high probability error rate for estimating the latent vectors v in Algorithm B.2. Throughout the proof, for any matrix $X \in \mathbf{R}^{n\times n}$, let $X^0 \in \mathbf{R}^{n\times n}$ and $[X^0]_{ij} = X_{ij}\mathbb{1}(i \neq j)$. Let $P = \sigma(\Theta^*)$ and $Q = \sigma(\eta^*)$, then $E(|A|) = P^0$ and $E(\frac{1+A}{2} \mid |A|) = |A| \circ Q$. Denote the error metric $\|v^*\|^2\|\Delta_{v_t}\|^2$ at the t-th iteration by e_t. For the convenience of analysis, we further let $\tilde{e}_t = \|v_0\|^2\|\Delta_{v_t}\|^2$. Under the assumption on the initialization, we have

$$\|\Delta_{v_0}\| \leq \delta\|v^*\|, \quad (1-\delta)\tilde{e}_t \leq e_t \leq (1+\delta)\tilde{e}_t,$$

for some sufficiently small $\delta \in (0,1)$.

The proof of Theorem 3.5.1 relies on Lemmas B.5.1-B.5.4, whose proofs are subsequently given in the subsection B.5.1. Other technical lemmas are provided in the subsection B.5.2.

Lemma B.5.1. *Let $A \in \boldsymbol{R}^{n\times n}$ be the symmetric binary matrix. For any matrices $P, X \in \boldsymbol{R}^{n\times n}$ satisfying that $\|X\|_F \geq \frac{2r}{P_{min}}\|A-P\|_{op}\|X\|_{\max}$,*

$$\|A \circ X\|_F^2 \geq \frac{1}{2}P_{min}\|X\|_F^2,$$

where r is the rank of X and $P_{min} = \min_{i,j} P_{ij}$.

Lemma B.5.2. *Let $\varphi_n = \max\{\||A|\circ((1+A)/2-Q)\|_{op}, 1\}$. If $\|\Delta_{v_t}\| \leq c_0 e^{-(M_1+M_3)/2}\|v^*\|$ and $\|v^*\|^2 \geq Ce^{M_1+M_3}\varphi_n$ for sufficiently small constant $c_0 > 0$ and sufficiently large constant $C > 0$, and $\||A|\circ\Delta_{\eta_t}\|_F^2 \geq \frac{1}{4}e^{-M_1}\|\Delta_{\eta_t}\|_F^2$, then there exists universal positive constants ρ and C' such that*

$$\tilde{e}_{t+1} \leq (1 - \frac{\tau}{e^{M_1+M_3}}\rho)\tilde{e}_t + \tau C' e^{M_1+M_3}\varphi_n^2.$$

Lemma B.5.3. *Let $\varphi_n = \max\{\||A|\circ((1+A)/2-Q)\|_{op}, 1\}$. If $\|\Delta_{v_0}\| \leq c_0 e^{-(M_1+M_3)/2}\|v^*\|$ and $\|v^*\|^2 \geq Ce^{M_1+M_3}\varphi_n \cdot \max\{\sqrt{\tau e^{M_1+M_3}}, 1\}$ for sufficiently small constant $c_0 > 0$ and sufficiently large constant $C > 0$, and $\||A|\circ\Delta_{\eta_t}\|_F^2 \geq \frac{1}{4}e^{-M_1}\|\Delta_{\eta_t}\|_F^2$ for $t \leq t_0$, then for any $0 \leq t \leq t_0$,*

$$\|\Delta_{v_t}\| \leq c_0 e^{-(M_1+M_3)/2}\|v^*\|.$$

Lemma B.5.4. *Let $\varphi_n = \max\{\||A|\circ((1+A)/2-Q)\|_{op}, 1\}$ and $\zeta_n = \max\{\|A - P\|_{op}, 1\}$. There exists constants c and $C = C(c)$ such that, uniformly over the parameter space with probability at least $1 - n^{-c}$, for any $\Theta \in \mathcal{F}_\theta(n,k,M_1,M_2)$, we have*

$$\varphi_n \leq C\sqrt{n} \quad \textit{and} \quad \zeta_n \leq C\sqrt{n}\sqrt{\max\{e^{-M_2}, \log n/n\}}.$$

Proof of Theorem 3.5.1. We first derive the deterministic error bound under the assumption that $\|v^*\|^2 \geq C_1 e^{M_1+M_3}\varphi_n \cdot \max\{\sqrt{\tau e^{M_1+M_3}}, 1\}$ for sufficiently large constant C_1. We consider two cases, depending on whether Δ_{η_t} belongs to the set $\{\Delta : \frac{\|\Delta\|_F}{\|\Delta\|_{\max}} \leq 8e^{M_1}\||A| - P\|_{op}\}$ or not.

Case 1. Suppose that there exists some $T \leq T_n \triangleq \log\left(\frac{M_3^2}{e^{3(M_1+M_3)}}n\right) / \log\left(1 - \frac{\tau}{e^{M_1+M_3}}\rho\right)^{-1}$ such that $\|\Delta_{\eta_T}\|_F \leq 8e^{M_1}\||A| - P\|_{op}\|\Delta_{\eta_T}\|_{\max}$. Since $\|\Delta_{\eta_T}\|_{\max} \leq \|v_T v_T^\top\|_{\max} + \|v^* v^{*\top}\|_{\max} \leq 2M_3$, and by Lemma B.5.6, it follows that

$$e_T \leq \frac{1}{2(\sqrt{2}-1)}\|\Delta_{\eta_T}\|_F^2 \leq \frac{128}{\sqrt{2}-1}M_3^2 e^{2M_1}\||A| - P\|_{op}^2. \tag{B.4}$$

Case 2. Suppose that $\|\Delta_{\eta_t}\|_F \geq 8e^{M_1}\||A| - P\|_{op}\|\Delta_{\eta_t}\|_{\max}$ holds for any $t \leq T_n$. Since the rank of $\Delta_{\eta_t} = v_t v_t^\top - v^* v^{*\top}$ is at most 2 and $P_{min} = \min_{i,j} P_{ij} \geq \frac{1}{2}e^{-M_1}$, it follows that

$$\frac{2\,\mathrm{rank}(\Delta_{\eta_t})}{P_{min}}\||A| - P\|_{op}\|\Delta_{\eta_t}\|_{\max} \leq 8e^{M_1}\||A| - P\|_{op}\|\Delta_{\eta_t}\|_{\max} \leq \|\Delta_{\eta_t}\|_F.$$

By Lemma B.5.1, we have for any $t \leq T_n$,

$$\||A| \circ \Delta_{\eta_t}\|_F^2 \geq \frac{1}{2}P_{min}\|\Delta_{\eta_t}\|_F^2 \geq \frac{1}{4}e^{-M_1}\|\Delta_{\eta_t}\|_F^2.$$

Then by Lemma B.5.3, $\|\Delta_{v_t}\| \leq c_0 e^{-(M_1+M_3)/2}\|v^*\|$ holds for any $t \leq T_n$, and we further apply Lemma B.5.2, there exists universal positive constants ρ and C' such that for any $t \leq T_n$

$$\tilde{e}_{t+1} \leq (1 - \frac{\tau}{e^{M_1+M_3}}\rho)\tilde{e}_t + \tau C' e^{M_1+M_3}\varphi_n^2.$$

This implies that

$$\begin{aligned}
\tilde{e}_T &\leq \left(1 - \frac{\tau}{e^{M_1+M_3}}\rho\right)^T \tilde{e}_0 + \sum_{t=0}^{T-1} \tau C' e^{M_1+M_3}\varphi_n^2 \left(1 - \frac{\tau}{e^{M_1+M_3}}\rho\right)^t \\
&\leq \left(1 - \frac{\tau}{e^{M_1+M_3}}\rho\right)^T \tilde{e}_0 + \tau C' e^{M_1+M_3}\varphi_n^2 \cdot \frac{e^{M_1+M_3}}{\tau\rho} \\
&= \left(1 - \frac{\tau}{e^{M_1+M_3}}\rho\right)^T \tilde{e}_0 + \frac{1}{\rho}C' e^{2(M_1+M_3)}\varphi_n^2.
\end{aligned}$$

Since $(1-\delta)\tilde{e}_t \leq e_t \leq (1+\delta)\tilde{e}_t$ for sufficiently small δ, we have

$$\begin{aligned}
e_T &\leq 2\left(1 - \frac{\tau}{e^{M_1+M_3}}\rho\right)^T e_0 + \frac{1}{\rho}C' e^{2(M_1+M_3)}\varphi_n^2 \\
&= 2e^{3(M_1+M_3)}\frac{1}{M_3^2 n}e_0 + \frac{1}{\rho}C' e^{2(M_1+M_3)}\varphi_n^2,
\end{aligned}$$

where the second equality is obtained by plugging the definition of T. Note that

$$e_0 \leq c_0^2 e^{-(M_1+M_3)}\|v^*\|^4 \leq c_0^2 e^{-(M_1+M_3)} n^2 \|v^*\|_{\max}^4 \leq c_0^2 e^{-(M_1+M_3)} n^2 M_3^2,$$

it follows that

$$e_{T_n} \leq 2c_0^2 e^{2(M_1+M_3)} n + \frac{1}{\rho} C' e^{2(M_1+M_3)} \varphi_n^2. \tag{B.5}$$

Thus, by combining the deterministic error bounds (B.4) and (B.5) in two cases respectively, there always exists some $T \leq T_n$ with

$$e_T \leq C'' e^{2(M_1+M_3)} \cdot \max\{\frac{M_3^2}{e^{2M_3}} \||A| - P\|_{op}^2, \varphi_n^2, n\}, \tag{B.6}$$

where C'' is some universal constant and $\varphi_n = \max\{\||A| \circ ((1+A)/2 - Q)\|_{op}, 1\}$.

Next, we establish the high probabilistic error bound as follows. By Lemma B.5.4, there exists c_1 and $\tilde{C} = C(c_1)$ such that uniformly over the parameter space

$$P\left(\varphi_n \leq \tilde{C}\sqrt{n},\ \zeta_n \leq \tilde{C}\sqrt{n}\sqrt{\max\{e^{-M_2}, \log n/n\}}\right) \geq 1 - n^{-c_1}.$$

Denote the above event as $\mathcal{E}$. Then on $\mathcal{E}$, the assumption $\|v^*\|^2 \geq C_1 e^{M_1+M_3}\varphi_n \cdot \max\{\sqrt{\tau e^{M_1+M_3}}, 1\}$ holds for sufficiently large constant C_1, and thereby from (B.6) we obtain that

$$e_T \leq C'' e^{2(M_1+M_3)} \cdot \max\{\frac{M_3^2}{e^{2M_3}} \tilde{C}^2 n \max\{e^{-M_2}, \log n/n\}, \tilde{C}^2 n, n\} \leq C e^{2(M_1+M_3)} n,$$

where $e^{M_3} \geq M_3$ for $M_3 > 0$ and $C = C''\tilde{C}^2$.

Finally, by Lemma B.5.3 and Lemmas B.5.6-B.5.7, the error metric e_T is equivalent to $\|\Delta_{\eta_T}\|_F^2$, i.e., there exists universal constants a_1 and a_2 such that $a_1 e_T \leq \|\Delta_{\eta_T}\|_F^2 \leq a_2 e_T$. So the above high probability bound also holds for the logit-transformed probability matrix of signs $\|\Delta_{\eta_T}\|_F^2$, which completes the proof. □

B.5.1 Proof of Lemmas B.5.1-B.5.4

Proof of Lemma B.5.1. Since $\|A \circ X\|_F^2 = \sum_{i,j} A_{ij}^2 X_{ij}^2 = \sum_{i,j} A_{ij} X_{ij}^2$ and $\sum_{i,j}(P_{ij} - \frac{1}{2}P_{min})X_{ij}^2 \geq \frac{1}{2}P_{min}\|X\|_F^2$, it suffices to prove that

$$\sum_{i,j}(P_{ij} - A_{ij})X_{ij}^2 \leq \frac{1}{2}P_{min}\|X\|_F^2,$$

that is to say, $\langle P - A, X \circ X \rangle \leq \frac{1}{2} P_{min} \|X\|_F^2$. To bound the left-hand side, note that $\text{rank}(X \circ X) \leq \text{rank}(X)^2$ and $\|X \circ X\|_F \leq \|X\|_F \|X\|_{\max}$, we have

$$\begin{aligned}\langle P - A, X \circ X \rangle &\leq \|A - P\|_{op} \|X \circ X\|_* \\ &\leq \|A - P\|_{op} \cdot \sqrt{\text{rank}(X \circ X)} \|X \circ X\|_F \\ &\leq \|A - P\|_{op} \cdot r \|X\|_F \|X\|_{\max} \\ &\leq \frac{1}{2} P_{min} \|X\|_F^2.\end{aligned}$$

The last inequality holds because $\|X\|_F \geq \frac{2r}{P_{min}} \|A - P\|_{op} \|X\|_{\max}$, which completes the proof. □

In the following proof, for presentation simplicity, we let

$$h(\eta) = -\sum_{i,j} |A_{ij}| \frac{1 + A_{ij}}{2} \eta_{ij} + |A_{ij}| \log(1 - \sigma(\eta_{ij})), \tag{B.7}$$

and then $\boldsymbol{\nabla} h(\eta) = |A| \circ (\sigma(\eta) - \frac{1+A}{2})$.

Proof of Lemma B.5.2. By the definition of $k_{t+1} = \arg\min_{k \in \{-1,1\}} \|v_{t+1} - kv^*\|$, we have

$$\|\Delta_{v_{t+1}}\|^2 = \|v_{t+1} - k_{t+1} v^*\|^2 \leq \|v_{t+1} - k_t v^*\|^2.$$

After plugging the definition of v_{t+1}, it follows that

$$\begin{aligned}\|\Delta_{v_{t+1}}\|^2 \leq &\|v_t - 2\tau_v \boldsymbol{\nabla} h(\eta_t) v_t - k_t v^*\|^2 \\ = &\|v_t - k_t v^*\|^2 - 4\tau_v \langle v_t - k_t v^*, \boldsymbol{\nabla} h(\eta_t) v_t \rangle + (2\tau_v)^2 \|\boldsymbol{\nabla} h(\eta_t) v_t\|^2 \\ = &\|v_t - k_t v^*\|^2 - 4\tau_v \langle (v_t - k_t v^*) v_t^\top, \boldsymbol{\nabla} h(\eta_t) \rangle + (2\tau_v)^2 \|\boldsymbol{\nabla} h(\eta_t) v_t\|^2.\end{aligned}$$

Note that $(v_t - k_t v^*) v_t^\top = \frac{1}{2}(v_t v_t^\top - v^* v^{*\top}) + \frac{1}{2}(v_t v_t^\top + v^* v^{*\top}) - k_t v^* v_t^\top$, and by the symmetry of $\boldsymbol{\nabla} h(\eta_t)$, we have

$$\langle (\frac{1}{2}(v_t v_t^\top + v^* v^{*\top}) - k_t v^* v_t^\top, \boldsymbol{\nabla} h(\eta_t) \rangle = \frac{1}{2} \langle \Delta_{v_t} \Delta_{v_t}^\top, \boldsymbol{\nabla} h(\eta_t) \rangle.$$

After combining the above three equations, we obtain that

$$\|\Delta_{v_{t+1}}\|^2 \leq \|\Delta_{v_t}\|^2 - 2\tau_v \langle \boldsymbol{\nabla} h(\eta_t), \Delta_{v_t} \Delta_{v_t}^\top \rangle - 2\tau_v \langle \boldsymbol{\nabla} h(\eta_t), \Delta_{\eta_t} \rangle + (2\tau_v)^2 \|\boldsymbol{\nabla} h(\eta_t) v_t\|^2.$$

Recall that $\tau_v = \tau/2\|v_0\|^2$. Multiplying both sides with $\|v_0\|^2$, we have

$$\tilde{e}_{t+1} \leq \tilde{e}_t - \tau\langle \boldsymbol{\nabla} h(\eta_t), \Delta_{v_t}\Delta_{v_t}^\top\rangle - \tau\langle \boldsymbol{\nabla} h(\eta_t), \Delta_{\eta_t}\rangle + \tau^2 \frac{1}{\|v_0\|^2}\|\boldsymbol{\nabla} h(\eta_t) v_t\|^2.$$

Let $H(\eta) = -\sum_{i,j} |A_{ij}|Q_{ij}\eta_{ij} + |A_{ij}|\log(1-\sigma(\eta_{ij}))$ and then $\boldsymbol{\nabla} H(\eta) = |A| \circ (\sigma(\eta) - Q)$. We further bound

$$\begin{aligned}
\tilde{e}_{t+1} \leq& \tilde{e}_t - \tau\langle \boldsymbol{\nabla} H(\eta_t), \Delta_{\eta_t}\rangle - \tau\langle \boldsymbol{\nabla} h(\eta_t) - \boldsymbol{\nabla} H(\eta_t), \Delta_{\eta_t}\rangle - \tau\langle \boldsymbol{\nabla} h(\eta_t), \Delta_{v_t}\Delta_{v_t}^\top\rangle \\
&+ \frac{\tau^2}{\|v_0\|^2}\|\boldsymbol{\nabla} h(\eta_t) v_t\|^2 \\
\leq& \tilde{e}_t - \tau\langle \boldsymbol{\nabla} H(\eta_t), \Delta_{\eta_t}\rangle + \tau|\langle \boldsymbol{\nabla} h(\eta_t) - \boldsymbol{\nabla} H(\eta_t), \Delta_{\eta_t}\rangle| + \tau|\langle \boldsymbol{\nabla} h(\eta_t), \Delta_{v_t}\Delta_{v_t}^\top\rangle| \\
&+ \frac{\tau^2}{\|v_0\|^2}\|\boldsymbol{\nabla} h(\eta_t) v_t\|^2 \\
:=& \tilde{e}_t - \tau D_1 + \tau D_2 + \tau D_3 + \tau^2 D_4. \qquad \text{(B.8)}
\end{aligned}$$

We first bound D_1 from below. Note that

$$D_1 = \langle \boldsymbol{\nabla} H(\eta_t), \Delta_{\eta_t}\rangle = \langle |A| \circ (\sigma(\eta_t) - Q), \Delta_{\eta_t}\rangle = \langle \sigma(\eta_t) - Q, |A| \circ \Delta_{\eta_t}\rangle. \qquad \text{(B.9)}$$

Let $\tilde{H}(\eta) = -\sum_{i,j} Q_{ij}\eta_{ij} + \log(1-\sigma(\eta_{ij}))$. It is straightforward to verify that $\boldsymbol{\nabla}\tilde{H}(\eta) = \sigma(\eta) - Q$ and for any $\eta \in \mathcal{F}_\eta(n, M_3)$

$$\frac{1}{4} I_{n^2\times n^2} \succeq \boldsymbol{\nabla}^2 \tilde{H}(\eta) = diag\Big(vec\big(\sigma(\eta) \circ (1-\sigma(\eta))\big)\Big) \succeq \mu I_{n^2\times n^2},$$

where $\mu = \frac{e^{M_3}}{(1+e^{M_3})^2} \asymp e^{-M_3}$. Thus, $\tilde{H}(\cdot)$ is μ-strongly convex and $\frac{1}{4}$-smooth. By applying Lemma B.5.5 along with $\boldsymbol{\nabla}\tilde{H}(\eta^*) = 0$, we obtain

$$\begin{aligned}
\langle \boldsymbol{\nabla}\tilde{H}(\eta_t), |A| \circ \Delta_{\eta_t}\rangle &\geq \frac{\mu/4}{\mu + 1/4}\||A| \circ \Delta_{\eta_t}\|_F^2 + \frac{1}{\mu + 1/4}\|\sigma(|A| \circ \eta_t) - \sigma(|A| \circ \eta^*)\|_F^2 \\
&= \frac{\mu/4}{\mu + 1/4}\||A| \circ \Delta_{\eta_t}\|_F^2 + \frac{1}{\mu + 1/4}\||A| \circ (\sigma(\eta_t) - Q)\|_F^2 \qquad \text{(B.10)}
\end{aligned}$$

Combining (B.9), (B.10), and $\||A| \circ \Delta_{\eta_t}\|_F^2 \geq \frac{1}{4}e^{-M_1}\|\Delta_{\eta_t}\|_F^2$, we bound D_1 from below,

$$D_1 \geq \frac{\mu}{\mu + 1/4}\frac{e^{-M_1}}{16}\|\Delta_{\eta_t}\|_F^2 + \frac{1}{\mu + 1/4}\||A| \circ (\sigma(\eta_t) - Q)\|_F^2.$$

To bound D_2, recall that $\varphi_n = \max\{\||A| \circ ((1+A)/2 - Q)\|_{op}, 1\}$, we have

$$D_2 = |\langle |A| \circ ((1+A)/2 - Q), \Delta_{\eta_t}\rangle| \leq \varphi_n\|\Delta_{\eta_t}\|_* \leq \varphi_n \cdot \sqrt{2}\|\Delta_{\eta_t}\|_F.$$

By Cauchy-Schwarz inequality, we have for any positive constant c_1 to be specified later,

$$D_2 \leq c_1 \|\Delta_{\eta_t}\|_F^2 + \frac{1}{2c_1}\varphi_n^2. \tag{B.11}$$

Therefore, by Lemma B.5.6,

$$\begin{aligned} D_1 - D_2 &\geq (\frac{\mu}{\mu + 1/4}\frac{e^{-M_1}}{16} - c_1)\|\Delta_{\eta_t}\|_F^2 + \frac{1}{\mu + 1/4}\||A| \circ (\sigma(\eta_t) - Q)\|_F^2 - \frac{1}{2c_1}\varphi_n^2 \\ &\geq 2(\sqrt{2} - 1)(\frac{\mu}{\mu + 1/4}\frac{e^{-M_1}}{16} - c_1)\|v^*\|^2\|\Delta_{v_t}\|^2 + \frac{1}{\mu + 1/4}\||A| \circ (\sigma(\eta_t) - Q)\|_F^2 \\ &\quad - \frac{1}{2c_1}\varphi_n^2 \\ &= 2(\sqrt{2} - 1)(\frac{\mu}{\mu + 1/4}\frac{e^{-M_1}}{16} - c_1)e_t + \frac{1}{\mu + 1/4}\||A| \circ (\sigma(\eta_t) - Q)\|_F^2 - \frac{1}{2c_1}\varphi_n^2 \end{aligned} \tag{B.12}$$

Next, we bound D_3. Note that $\Delta_{v_t}\Delta_{v_t}^\top$ is a positive semi-definite matrix, we have

$$\begin{aligned} D_3 &= |\langle \nabla h(\eta_t), \Delta_{v_t}\Delta_{v_t}^\top \rangle| \leq \|\nabla h(\eta_t)\|_{op}\|\Delta_{v_t}\Delta_{v_t}^\top\|_* \\ &= \|\nabla h(\eta_t)\|_{op} \operatorname{Tr}(\Delta_{v_t}\Delta_{v_t}^\top) = \|\nabla h(\eta_t)\|_{op}\|\Delta_{v_t}\|^2 \\ &\leq \|\nabla H(\eta_t)\|_{op}\|\Delta_{v_t}\|^2 + \|\nabla h(\eta_t) - \nabla H(\eta_t)\|_{op}\|\Delta_{v_t}\|^2 \\ &\leq \||A| \circ (\sigma(\eta_t) - Q)\|_{op}\|\Delta_{v_t}\|^2 + \varphi_n\|\Delta_{v_t}\|^2. \end{aligned}$$

Since $\|\Delta_{v_t}\| \leq c_0 e^{-(M_1+M_3)/2}\|v^*\|$, by Cauchy-Schwarz inequality,

$$\begin{aligned} \||A| \circ (\sigma(\eta_t) - Q)\|_{op}\|\Delta_{v_t}\|_F^2 &\leq c_0 e^{-(M_1+M_3)/2}\||A| \circ (\sigma(\eta_t) - Q)\|_{op}\|v^*\|\|\Delta_{v_t}\|_F \\ &\leq c_2\||A| \circ (\sigma(\eta_t) - Q)\|_{op}^2 + \frac{c_0^2}{4c_2 e^{M_1+M_3}}\|v^*\|^2\|\Delta_{v_t}\|_F^2, \end{aligned}$$

where c_2 is a positive constant to be specified later. And given that $\|v^*\|^2 \geq Ce^{M_1+M_3}\varphi_n$ for sufficiently large constant C, we have

$$\varphi_n\|\Delta_{v_t}\|^2 \leq \frac{\|v^*\|^2\|\Delta_{v_t}\|^2}{Ce^{M_1+M_3}} = \frac{e_t}{Ce^{M_1+M_3}}.$$

Thus, it follows that

$$D_3 \leq (\frac{1}{Ce^{M_1+M_3}} + \frac{c_0^2}{4c_2 e^{M_1+M_3}})e_t + c_2\||A| \circ (\sigma(\eta_t) - Q)\|_{op}^2. \tag{B.13}$$

Finally, to bound D_4, by the assumption that $\|\Delta_{v_t}\| \leq c_0 e^{-(M_1+M_3)/2}\|v^*\|$ for sufficiently small constant c_0, there exists constant C_3 such that

$$D_4 = \frac{1}{\|v_0\|^2}\|\nabla h(\eta_t)v_t\|^2 \leq \frac{1}{\|v_0\|^2}\|\nabla h(\eta_t)\|_{op}^2\|v_t\|^2$$

$$
\begin{aligned}
&\leq \frac{1}{\|v_0\|^2}\left(\|\boldsymbol{\nabla} h(\eta_t)-\boldsymbol{\nabla} H(\eta_t)\|_{op}^2\|v_t\|^2+\|\boldsymbol{\nabla} H(\eta_t)\|_{op}^2\|v_t\|^2\right)\\
&\leq \frac{1}{\|v_0\|^2}\left(\varphi_n^2\|v_t\|^2+\||A|\circ(\sigma(\eta_t)-Q)\|_{op}^2\|v_t\|^2\right)\\
&\leq C_3(\varphi_n^2+\||A|\circ(\sigma(\eta_t)-Q)\|_{op}^2). \qquad \text{(B.14)}
\end{aligned}
$$

By plugging (B.12)-(B.14) into (B.8), it follows that

$$
\begin{aligned}
\tilde{e}_{t+1} \leq \tilde{e}_t &- \tau\left(2(\sqrt{2}-1)(\frac{\mu}{\mu+1/4}\frac{e^{-M_1}}{16}-c_1)-\frac{1}{Ce^{M_1+M_3}}-\frac{c_0^2}{4c_2e^{M_1+M_3}}\right)e_t\\
&-\tau\left(\frac{1}{\mu+1/4}-c_2-\tau C_3\right)\||A|\circ(\sigma(\eta_t)-Q)\|_F^2+\tau\frac{1}{2c_1}\varphi_n^2+\tau^2C_3\varphi_n^2,
\end{aligned}
$$

where c_1, c_2 are arbitrary positive constants, c_0 is a sufficiently small constant, and C is a sufficiently large constant. Given that $\mu \asymp e^{-M_3}$, we choose $c_1 = e^{-(M_1+M_3)}c$, $c_2 = c_0$, and c, τ small enough such that there exists some universal positive constants $\tilde{\rho}$ and C',

$$
2(\sqrt{2}-1)(\frac{\mu}{\mu+1/4}\frac{e^{-M_1}}{16}-c_1)-\frac{1}{Ce^{M_1+M_3}}-\frac{c_0^2}{4c_2e^{M_1+M_3}} > \tilde{\rho}e^{-(M_1+M_3)},
$$

$$
\frac{1}{\mu+1/4}-c_2-\tau C_3>0, \text{ and } \frac{1}{2c_1}+\tau C_3 < C'e^{M_1+M_3}.
$$

Then, we obtain

$$
\tilde{e}_{t+1} \leq \tilde{e}_t-\tau\tilde{\rho}e^{-(M_1+M_3)}e_t+\tau C'e^{M_1+M_3}\varphi_n^2.
$$

Recall that $e_t \geq (1-\delta)\tilde{e}_t$. Let $\rho=(1-\delta)\tilde{\rho}$, then we have

$$
\tilde{e}_{t+1} \leq \tilde{e}_t-\tau\tilde{\rho}(1-\delta)e^{-(M_1+M_3)}\tilde{e}_t+\tau C'e^{M_1+M_3}\varphi_n^2 = (1-\frac{\tau\rho}{e^{M_1+M_3}})\tilde{e}_t+\tau C'e^{M_1+M_3}\varphi_n^2,
$$

which completes the proof. □

Proof of Lemma B.5.3. By the definition of $\tilde{e}_t$, it is equivalent to show that

$$
\tilde{e}_t \leq c_0^2e^{-(M_1+M_3)}\|v^*\|^2\|v_0\|^2
$$

for any $0 \leq t \leq t_0$. We prove it by induction as below. For $t=0$, it follows by the initialization condition that

$$
\tilde{e}_0 = \|\Delta_{v_0}\|^2\|v_0\|^2 \leq c_0^2e^{-(M_1+M_3)}\|v^*\|^2\|v_0\|^2.
$$

Suppose that we have $\|\Delta_{v_t}\| \leq c_0e^{-(M_1+M_3)/2}\|v^*\|$, then by Lemma B.5.2 we have for

$(t+1)$-th iteration,

$$\begin{aligned}\tilde{e}_{t+1} &\le (1 - \frac{\tau\rho}{e^{M_1+M_3}})\tilde{e}_t + \tau C' e^{M_1+M_3}\varphi_n^2 \\ &\le (1 - \frac{\tau\rho}{e^{M_1+M_3}}) c_0^2 e^{-(M_1+M_3)}\|v^*\|^2\|v_0\|^2 + \tau C' e^{M_1+M_3}\varphi_n^2 \\ &= c_0^2 e^{-(M_1+M_3)}\|v^*\|^2\|v_0\|^2 \left(1 - \frac{\tau\rho}{e^{M_1+M_3}} + \frac{\tau C' e^{2(M_1+M_3)}\varphi_n^2}{c_0^2\|v^*\|^2\|v_0\|^2}\right),\end{aligned}$$

where C' and ρ are some positive constants. Since $\|v^*\|^2 \ge C e^{M_1+M_3}\varphi_n \cdot \sqrt{\tau e^{M_1+M_3}}$, we have

$$e^{2(M_1+M_3)}\varphi_n^2 \le \frac{\|v^*\|^4}{\tau C^2 e^{M_1+M_3}} \le \frac{4\|v^*\|^2\|v_0\|^2}{\tau C^2 e^{M_1+M_3}},$$

where the last inequality holds because $\|v_0\| \ge \|v^*\| - \|\Delta_{v_0}\| \ge (1 - c_0 e^{-(M_1+M_3)/2})\|v^*\| \ge 1/2\|v^*\|$ for sufficiently small c_0. Then it follows that

$$\tilde{e}_{t+1} \le c_0^2 e^{-(M_1+M_3)}\|v^*\|^2\|v_0\|^2 \left(1 - \frac{\tau\rho}{e^{M_1+M_3}} + \frac{4C'}{c_0^2 C^2 e^{M_1+M_3}}\right) \le c_0^2 e^{-(M_1+M_3)}\|v^*\|^2\|v_0\|^2,$$

where we choose C large enough such that $C^2 \ge \frac{4C'}{c_0^2\tau\rho}$. This completes the proof. □

Proof of Lemma B.5.4. Let $B \in \{0,1\}^{n\times n}$ denote the symmetric matrix $|A| \circ (1 + A)/2$. Note that, conditional on the absolute adjacency matrix $|A|$, the elements of the matrix B independently follow Bernoulli distribution with $E(B_{ij}\big||A_{ij}|) = |A_{ij}|Q_{ij}$ for all $i > j$. Since $|A_{ij}|Q_{ij} \le 1$ always holds, by applying Lemma B.5.8, there exists c and $C = C(c)$ such that uniformly over $\mathcal{F}_\eta(n, M_3)$ and the value of $|A|$

$$P\left(\|B - |A|\circ Q\|_{op} \le C\sqrt{n}\Big||A|\right) \ge 1 - \frac{n^{-c}}{2}.$$

It follows that uniformly over $\mathcal{F}_\eta(n, M_3)$ and M_1

$$\begin{aligned}P\left(\left\||A| \circ \left(\frac{1+A}{2} - Q\right)\right\|_{op} \le C\sqrt{n}\right) &= E_{|A|}\left[P\Big(\|B - |A|\circ Q\|_{op} \le C\sqrt{n}\Big||A|\Big)\right] \\ &\ge 1 - \frac{n^{-c}}{2},\end{aligned}$$

where $|A_{ij}|$ independently follows Bernoulli distribution with $E(|A_{ij}|) = P_{ij}$. This implies that $P(\varphi_n \le C\sqrt{n}) \ge 1 - \frac{1}{2}n^{-c}$. In addition, by Lemma 22 in Ma et al. (2020), we have $P(\||A| - P\|_{op} \le C\sqrt{n}\sqrt{\max\{e^{-M_2}, \log n/n\}}) \ge 1 - \frac{1}{2}n^{-c}$. Consider the intersection of the above two events $\mathcal{E}$, then $P(\mathcal{E}) \ge 1 - n^{c_1}$. □

B.5.2 Additional technical lemmas

Lemma B.5.5 (Nesterov (2013)). *For a continuously differentiable function f, if it is μ-strongly convex and L-smooth on a convex domain $\mathcal{D}$, i.e., for any $x, y \in \mathcal{D}$,*

$$\frac{\mu}{2}\|x-y\|^2 \leq f(y) - f(x) - \langle f'(x), y - x\rangle \leq \frac{L}{2}\|x-y\|^2,$$

then we have

$$\langle f'(x) - f'(y), x - y\rangle \geq \frac{\mu L}{\mu + L}\|x-y\|^2 + \frac{1}{\mu+L}\|f'(x) - f'(y)\|^2,$$

and

$$\langle f'(x) - f'(y), x - y\rangle \geq \mu\|x-y\|^2.$$

Lemma B.5.6 (Tu et al. (2016)). *For any $X_1, X_2 \in \boldsymbol{R}^{n\times k}$, then we have*

$$dist(X_1, X_2)^2 \leq \frac{1}{2(\sqrt{2}-1)\sigma_k^2(X_1)}\|X_1X_1^\top - X_2X_2^\top\|_F^2,$$

where $\sigma_k(X)$ is the smallest singular value of X.

Lemma B.5.7 (Tu et al. (2016)). *For any $X_1, X_2 \in \boldsymbol{R}^{n\times k}$ such that $dist(X_1, X_2) \leq c\|X_1\|_{op}$ with some constant c, then we have*

$$\|X_1X_1^\top - X_2X_2^\top\|_F \leq (2+c)\|X_1\|_{op}\, dist(X_1, X_2).$$

Lemma B.5.8 (Lei and Rinaldo (2015); Gao et al. (2017)). *Let $A \in \{0,1\}^{n\times n}$ be a symmetric adjacency matrix with $A_{ii} = 0$ for all i and $P \in [0,1]^{n\times n}$ be a symmetric matrix, where A_{ij} independently follows Bernoulli distribution with $E(A_{ij}) = P_{ij}$ for all $i > j$. Then, for any C_0, there is a constant $C = C(C_0)$ such that*

$$\|A - P\|_{op} \leq C\sqrt{nP_{max} + \log n}$$

with probability at least $1 - n^{-C_0}$, where $P_{max} = \max_{i,j} P_{ij}$.

B.6 Proof of Theorem 3.5.2 and Corollary 3.5.1

In this section, we establish the high probability error bounds for estimating the latent vectors Z and v in Algorithm III.1. We first introduce some notations. Let

$O_t = \arg\min_{O \in O(k)} \|Z_t - Z^* O\|_F$, $\Delta_{Z_t} = Z_t - Z^* O_t$, and $\Delta_{\alpha_t} = \alpha_t - \alpha^*$. To differentiate the error metrics for v and Z, we rewrite the error metric $e_t = e_t^v$ and further define the error metrics for Z as $e_t^Z = \|\Delta_{Z_t}\|_F^2 \|Z^*\|_{op}^2 + 2n\|\Delta_{\alpha_t}\|^2$ and $\tilde{e}_t^Z = \|\Delta_{Z_t}\|_F^2 \|Z_0\|_{op}^2 + 2n\|\Delta_{\alpha_t}\|^2$. Given (w_t, γ_t) obtained from the line 5 in Algorithm III.1, we denote $v_t = Z_t w_t + \gamma_t 1_n$, $\eta_t = v_t v_t^\top$, and $\Delta_{\eta_t} = \eta_t - \eta^*$. The proof of Theorem 3.5.2 and Corollary 3.5.1 relies on Lemmas B.6.1-B.6.4, whose proofs are subsequently given in the subsection B.6.1.

Lemma B.6.1 (Iterative errors for Z_{t+1})**.** *Set the step sizes as* $\tau_Z = r_0\tau / \|Z_0\|_{op}^2$, $\tau_\alpha = \tau/(2n)$, *and the weight* $\lambda = \tilde{\lambda} r_0 / e^{M_1}\kappa_{Z^*}^2$ *with* $r_0 = \min\{1, \|Z_0\|_{op}^2 / \|v_0\|^2\}$ *for any* $\tau \le c_\tau$, $\tilde{\lambda} \le c_\lambda$, *where* c_τ *and* c_λ *are universal constants. Let* $\zeta_n = \max\{\||A| - P\|_{op}, 1\}$ *and* $\varphi_n = \max\{\||A| \circ ((1+A)/2 - Q)\|_{op}, 1\}$. *If 1)* $\|\Delta_{Z_t}\|_F \le c_0 e^{-M_1}\|Z^*\|_{op}/\kappa_{Z^*}^2$ *and* $\|\Delta_{v_t}\| \le c_0 e^{-(M_1+M_3)/2}\|v^*\|$ *for a sufficiently small constant* c_0*; and 2)* $\|Z^*\|_{op}^2 \ge C_0 e^{M_1}\kappa_{Z^*}^2 \zeta_n$ *and* $\|v^*\|^2 \ge C_0 e^{M_1+M_3}\varphi_n$ *for a sufficiently large constant* C_0*, then there exist universal positive constants* ρ_1, ρ_2, C', *and* C'' *such that*

$$\tilde{e}_{t+1}^Z \le \left(1 - \frac{r_0\tau\rho_1}{e^{M_1}\kappa_{Z^*}^2}\right)\tilde{e}_t^Z - \lambda\frac{r_0\tau\rho_2}{e^{M_3}}\min\{\||A| \circ \Delta_{\eta_t}\|_F^2, e^{-M_1}\|\Delta_{\eta_t}\|_F^2\}$$
$$+ r_0\tau C' e^{M_1}\zeta_n^2 k + \lambda r_0 \tau C'' e^{M_1+M_3}\varphi_n^2,$$

Lemma B.6.2 (Bound for η_t given Z_t)**.** *Let* $\zeta_n = \max\{\||A| - P\|_{op}, 1\}$ *and* $\varphi_n = \max\{\||A| \circ ((1+A)/2 - Q)\|_{op}, 1\}$. *We have for* $t \ge 0$

$$\|\Delta_{\eta_t}\|_F \le 16 e^{M_1+M_3}\max\{\zeta_n, \varphi_n\} + e^{M_1/2+M_3}(2 + \|\Delta_{Z_t}\|_F\|w^*\|)\|v^*\|\|\Delta_{Z_t}\|_F\|w^*\|.$$

Lemma B.6.3. *Let* $\zeta_n = \max\{\||A| - P\|_{op}, 1\}$ *and* $\varphi_n = \max\{\||A| \circ ((1+A)/2 - Q)\|_{op}, 1\}$. *If 1)* $\|\Delta_{Z_t}\|_F \le c_0 e^{-M_1 - 3M_3/2}\|Z^*\|_{op}/\kappa_{Z^*}^2$ *for a sufficiently small constant* c_0*; and 2)* $\|Z^*\|_{op}^2 \ge C e^{M_1}\kappa_{Z^*}^2\zeta_n \max\{\sqrt{\tau k}e^{M_1+3M_3/2}\kappa_{Z^*}, 1\}$ *and* $\|v^*\|^2 \ge C e^{M_1+M_3}\varphi_n$ $\max\{\sqrt{\tau}e^{M_1/2+M_3}, 1\}$ *for a sufficiently large constant* C. *Then there exists a sufficiently small constant* c *such that*

$$\|\Delta_{v_t}\| \le c e^{-(M_1+M_3)/2}\|v^*\|.$$

Lemma B.6.4. *Set the step sizes as $\tau_Z = r_0\tau/\|Z_0\|_{op}^2$, $\tau_\alpha = \tau/(2n)$, and the weight $\lambda = \tilde\lambda r_0/e^{M_1}\kappa_{Z^*}^2$ with $r_0 = \min\{1, \|Z_0\|_{op}^2/\|v_0\|^2\}$ for any $\tau \le c_\tau$, $\tilde\lambda \le c_\lambda$, where c_τ and c_λ are universal constants. Suppose 1) $\tilde e_0^Z \le c_0^2 e^{-2M_1-3M_3}\|Z^*\|_{op}^4/4\kappa_{Z^*}^4$ for a sufficiently small constant c_0; and 2) $\|Z^*\|_{op}^2 \ge Ce^{M_1}\kappa_{Z^*}^2\zeta_n \max\{\sqrt{\tau k}e^{M_1+3M_3/2}\kappa_{Z^*}, 1\}$ and $\|v^*\|^2 \ge Ce^{M_1+M_3}\varphi_n \max\{\sqrt{\tau}e^{M_1/2+M_3}, 1\}$ for a sufficiently large constant C. Then for all $t \ge 0$,*

$$\|\Delta_{Z_t}\|_F \le c_0 e^{-M_1-3M_3/2}\|Z^*\|_{op}/\kappa_{Z^*}^2.$$

Proof of Theorem 3.5.2. We first prove the deterministic bounds. Given the initialization assumption, by Lemma B.6.4 and Lemma B.6.3, we have $\|\Delta_{Z_t}\|_F \le c_0 e^{-M_1-3M_3/2}\|Z^*\|_{op}/\kappa_{Z^*}^2$ and $\|\Delta_{v_t}\| \le c_0 e^{-(M_1+M_3)/2}\|v^*\|$ hold with a sufficiently small constant c_0 for all $t \ge 0$. Then, by Lemma B.6.1, it follows that for any $t \ge 0$

$$\begin{aligned}\tilde e_{t+1}^Z \le{}& \left(1 - \frac{r_0\tau\rho_1}{e^{M_1}\kappa_{Z^*}^2}\right)\tilde e_t^Z - \lambda\frac{r_0\tau\rho_2}{e^{M_3}}\min\{\||A|\circ\Delta_{\eta_t}\|_F^2, e^{-M_1}\|\Delta_{\eta_t}\|_F^2\}\\ &+ r_0\tau C' e^{M_1}\zeta_n^2 k + \lambda r_0\tau C'' e^{M_1+M_3}\varphi_n^2,\end{aligned}$$

which proves the deterministic error bounds for iterates. This further implies that

$$\begin{aligned}\tilde e_T^Z &\le \left(1 - \frac{r_0\tau\rho_1}{e^{M_1}\kappa_{Z^*}^2}\right)^T \tilde e_0^Z + \left(r_0\tau C' e^{M_1}\zeta_n^2 k + \lambda r_0\tau C'' e^{M_1+M_3}\varphi_n^2\right)\sum_{t=0}^{T-1}\left(1 - \frac{r_0\tau\rho_1}{e^{M_1}\kappa_{Z^*}^2}\right)^t\\ &\le \left(1 - \frac{r_0\tau\rho_1}{e^{M_1}\kappa_{Z^*}^2}\right)^T \tilde e_0^Z + \left(r_0\tau C' e^{M_1}\zeta_n^2 k + \lambda r_0\tau C'' e^{M_1+M_3}\varphi_n^2\right)\cdot\frac{e^{M_1}\kappa_{Z^*}^2}{r_0\tau\rho_1}\\ &= \left(1 - \frac{r_0\tau\rho_1}{e^{M_1}\kappa_{Z^*}^2}\right)^T \tilde e_0^Z + C' e^{2M_1}\zeta_n^2 k\frac{\kappa_{Z^*}^2}{\rho_1} + \tilde\lambda r_0 C'' e^{M_1+M_3}\varphi_n^2\frac{1}{\rho_1}\\ &= \frac{\kappa_{Z^*}^2 e^{4M_1+3M_3-M_2}}{M_1^2}\frac{k^3}{n}\tilde e_0^Z + C' e^{2M_1}\zeta_n^2 k\frac{\kappa_{Z^*}^2}{\rho_1} + \tilde\lambda r_0 C'' e^{M_1+M_3}\varphi_n^2\frac{1}{\rho_1},\end{aligned}$$

where the last equality is obtained by plugging the definition of T. Note that

$$\begin{aligned}\frac{\kappa_{Z^*}^2 e^{4M_1+3M_3-M_2}}{M_1^2}\frac{k^3}{n}\tilde e_0^Z &\le c_0\frac{\kappa_{Z^*}^2 e^{2M_1-M_2}}{M_1^2}\frac{k^3}{n}\frac{\|Z^*\|_{op}^4}{\kappa_{Z^*}^4} \le c_0\frac{\kappa_{Z^*}^2 e^{2M_1-M_2}}{M_1^2}\frac{k^3}{n}\frac{n^2M_1^2}{k^2}\\ &= c_0\kappa_{Z^*}^2 e^{2M_1-M_2}nk,\end{aligned}$$

where the first inequality is given by the initialization assumption and the second inequality is based on the fact that $k\|Z^*\|_{op}^2/\kappa_{Z^*}^2 = k\sigma_1(Z^*)^2 \le \|Z^*\|_F^2 \le nM_1$. In addition, since $e_T^Z \le (1+\delta)\tilde{e}_T^Z$ for sufficiently small δ, we obtain that

$$e_T^Z \le 2\tilde{e}_T^Z \le c_0\kappa_{Z^*}^2 e^{2M_1-M_2}nk + C'e^{2M_1}\zeta_n^2 k\frac{\kappa_{Z^*}^2}{\rho_1} + \tilde{\lambda} r_0 C'' e^{M_1+M_3}\varphi_n^2\frac{1}{\rho_1}. \tag{B.15}$$

Next, we prove the high-probability bounds. By Lemma B.5.4, there exists c_1 and $\tilde{C} = C(c_1)$ such that uniformly over the parameter space

$$P\left(\varphi_n \le \tilde{C}\sqrt{n},\ \zeta_n \le \tilde{C}\sqrt{n}\sqrt{\max\{e^{-M_2}, \log n/n\}}\right) \ge 1 - n^{-c_1}.$$

Denote the above event as $\mathcal{E}$. Then on $\mathcal{E}$, the assumptions in the deterministic bounds, i.e., $\|Z^*\|_{op}^2 \ge C_0 e^{M_1}\kappa_{Z^*}^2\zeta_n \max\{\sqrt{\tau k}e^{M_1+3M_3/2}\kappa_{Z^*}, 1\}$ and $\|v^*\|^2 \ge C_0 e^{M_1+M_3}\varphi_n \cdot \max\{\sqrt{\tau}e^{M_1/2+M_3}, 1\}$ hold for a sufficiently large constant C_0, and thereby from (B.15) we obtain that

$$\begin{aligned} e_T^Z &\le \frac{\tilde{C}^2}{\rho_1}\kappa_{Z^*}^2 e^{2M_1}nk \cdot \max\{C'e^{-M_2}, C'\frac{\log n}{n}, \tilde{\lambda} r_0 C'' e^{M_3-M_1}\frac{1}{\kappa_{Z^*}^2 k}\} \\ &\le C_1\kappa_{Z^*}^2 e^{2M_1}nk \cdot \max\{e^{-M_2}, \frac{\log n}{n}, e^{M_3-M_1}\frac{r_0}{\kappa_{Z^*}^2 k}\} \end{aligned} \tag{B.16}$$

where $C_1 = \max\{C', \tilde{\lambda}C''\}\frac{\tilde{C}^2}{\rho_1}$. □

Proof of Corollary 3.5.1. By Lemma B.6.2, we have

$$\|\Delta_{\eta_t}\|_F \le 16e^{M_1+M_3}\max\{\zeta_n, \varphi_n\} + e^{M_1/2+M_3}(2 + \|\Delta_{Z_t}\|_F\|w^*\|)\|v^*\|\|\Delta_{Z_t}\|_F\|w^*\|.$$

Furthermore, with a sufficiently small c_0, we have

$$2 + \|\Delta_{Z_t}\|_F\|w^*\| \le 2 + c_0 e^{-M_1-3M_3/2}\|Z^*\|_{op}\|w^*\|/\kappa_{Z^*} \le 2 + c_0 e^{-M_1-3M_3/2}\|v^*\| \le 3,$$

where the last inequality is based on $\|v^*\| \ge \|Z^*\|_{op}\|w^*\|/\kappa_{Z^*}$ and $\kappa_{Z^*} \ge 1$. It follows that

$$\|\Delta_{\eta_T}\|_F \le 16e^{M_1+M_3}\max\{\zeta_n, \varphi_n\} + 3e^{M_1/2+M_3}\|v^*\|\|\Delta_{Z_T}\|_F\|w^*\|.$$

Then, on the event $\mathcal{E}$,

$$\begin{aligned} \|\Delta_{\eta_T}\|_F^2 &\le 512e^{2(M_1+M_3)}\tilde{C}^2 n + 18e^{M_1+2M_3}\frac{\|v^*\|^2\|w^*\|^2}{\|Z^*\|_{op}^2}\|Z^*\|_{op}^2\|\Delta_{Z_T}\|_F^2 \\ &\le 512e^{2(M_1+M_3)}\tilde{C}^2 n + 18e^{M_1+2M_3}\frac{M^2}{r_0}e_T^Z, \end{aligned}$$

since $\|w^*\| \le M$, $r_0 \le \|Z_0\|_{op}^2/\|v_0\|^2 \asymp \|Z^*\|_{op}^2/\|v^*\|^2$, and $\|Z^*\|_{op}^2\|\Delta_{Z_T}\|_F^2 \le e_T^Z$. By plugging (B.16), we have

$$\|\Delta_{\eta_T}\|_F^2 \leq C_2 e^{3M_1+2M_3} nk \cdot \max\{\frac{e^{M_3-M_1}}{k}, \kappa_{Z^*}^2 \max\{e^{-M_2}, \frac{\log n}{n}\}\},$$

where $C_2 = \max\{512\tilde{C}^2, 18C_1M^2/r_0\}$. This completes the proof. □

B.6.1 Proof of Lemmas B.6.1-B.6.4

Proof of Lemma B.6.1. Let $O_{t+1} = \arg\min_{O\in O(k)} \|Z_{t+1} - Z^*O\|_F$ and $\tilde{O}_{t+1} = \arg\min_{O\in O(k)} \|\tilde{Z}_{t+1} - Z^*O\|_F$. For presentation simplicity, we define

$$g(\Theta) = -\sum_{i,j} |A_{ij}|\theta_{ij} + \log(1 - \sigma(\theta_{ij})).$$

By definition, we have

$$\begin{aligned}\|\Delta_{Z_{t+1}}\|_F^2 &= \|Z_{t+1} - Z^*O_{t+1}\|_F^2 \leq \|Z_{t+1} - Z^*\tilde{O}_{t+1}\|_F^2 = \|J\tilde{Z}_{t+1} - JZ^*\tilde{O}_{t+1}\|_F^2 \\ &\leq \|\tilde{Z}_{t+1} - Z^*\tilde{O}_{t+1}\|_F^2 \leq \|\tilde{Z}_{t+1} - Z^*O_t\|_F^2, \qquad \text{(B.17)}\end{aligned}$$

where the second inequality holds due to the column-wise centralization in the projection step. For $\tilde{Z}_{t+1}$ defined in line 2 in Algorithm III.1, we have

$$\begin{aligned}\|\Delta_{Z_{t+1}}\|_F^2 &\leq \|(1-\lambda)(Z_t - 2\tau_Z \nabla g(\Theta_t)Z_t - Z^*O_t) + \lambda(Z_t - 2\tau_Z\nabla h(\eta_t)v_tw_t^\top - Z^*O_t)\|_F^2 \\ &\leq (1-\lambda)\|Z_t - 2\tau_Z\nabla g(\Theta_t)Z_t - Z^*O_t\|_F^2 + \lambda\|Z_t - 2\tau_Z\nabla h(\eta_t)v_tw_t^\top - Z^*O_t\|_F^2,\end{aligned}$$

where the second inequality is due to the Jensen inequality. Similarly, we have

$$\begin{aligned}\|\Delta_{\alpha_{t+1}}\|^2 &= \|(1-\lambda)(\alpha_t - 2\tau_\alpha\nabla g(\Theta_t)1_n - \alpha^*) + \lambda(\alpha_t - \alpha^*)\|^2 \\ &\leq (1-\lambda)\|\alpha_t - 2\tau_\alpha\nabla g(\Theta_t)1_n - \alpha^*\|^2 + \lambda\|\alpha_t - \alpha^*\|^2.\end{aligned}$$

It follows that

$$\begin{aligned}\tilde{e}_{t+1}^Z &= \|\Delta_{Z_{t+1}}\|_F^2\|Z_0\|_{op}^2 + 2n\|\Delta_{\alpha_{t+1}}\|^2 \\ &\leq (1-\lambda)\left(\|Z_t - 2\tau_Z\nabla g(\Theta_t)Z_t - Z^*O_t\|_F^2\|Z_0\|_{op}^2 + 2n\|\alpha_t - 2\tau_\alpha\nabla g(\Theta_t)1_n - \alpha^*\|^2\right) \\ &\quad + \lambda\|Z_t - 2\tau_Z\nabla h(\eta_t)v_tw_t^\top - Z^*O_t\|_F^2\|Z_0\|_{op}^2 \\ &\quad + \lambda\cdot 2n\|\alpha_t - \alpha^*\|^2 \\ &:= (1-\lambda)J_1 + \lambda\cdot J_2\cdot\|Z_0\|_{op}^2 + \lambda\cdot 2n\|\alpha_t - \alpha^*\|^2. \qquad \text{(B.18)}\end{aligned}$$

Note that the term J_1 is equivalent to the error at $(t+1)$-th iteration of Algorithm B.1, where only the likelihood of observing edges is involved. Recall that we choose $\tau_Z = r_0\tau/\|Z_0\|_{op}^2$ with $r_0 = \min\{1, \|Z_0\|_{op}^2/\|v_0\|^2\} \leq 1$. Therefore, by Lemma 25 in Ma et al. (2020), there exist a constant c such that, for any $r_0\tau \leq \tau \leq c$, there exist universal positive constants ρ and C' such that

$$J_1 \leq \left(1 - \frac{r_0\tau}{e^{M_1}\kappa_{Z^*}^2}\rho\right)\tilde{e}_t^Z + r_0\tau C' e^{M_1}\zeta_n^2 k. \tag{B.19}$$

For the term J_2, we have

$$\begin{aligned}
J_2 &= \|\Delta_{Z_t}\|_F^2 - 4\tau_Z\langle Z_t - Z^*O_t, \nabla h(\eta_t)v_t w_t^\top\rangle + 4\tau_Z^2\|\nabla h(\eta_t)v_t w_t^\top\|_F^2 \\
&= \|\Delta_{Z_t}\|_F^2 - 4\tau_Z\langle Z_t w_t - Z^*O_t w_t, \nabla h(\eta_t)v_t\rangle + 4\tau_Z^2\|\nabla h(\eta_t)v_t\|^2\|w_t\|^2 \\
&= \|\Delta_{Z_t}\|_F^2 - 4\tau_Z\left\langle v_t - k_t v^* + Z^*O_t(k_t O_t^\top w^* - w_t) + (k_t\gamma^* - \gamma_t)1_n, \nabla h(\eta_t)v_t\right\rangle \\
&\quad + 4\tau_Z^2\|\nabla h(\eta_t)v_t\|^2\|w_t\|^2,
\end{aligned}$$

where $k_t = \arg\min_{k\in\{-1,1\}}\|v_t - kv^*\|$. The second equality is due to $\|v_1 v_2{}^\top\|_F = \|v_1\|\|v_2\|$ for any two vectors v_1, v_2; the third equality is obtained by adding and subtracting the term $k_t v^* = k_t Z^* w^* - k_t\gamma^* 1_n$. Note that, by the definition of (w_t, γ_t) in line 5 in Algorithm III.1, the gradients of $\mathcal{L}_\lambda(\alpha_{t+1}, Z_{t+1}, w, \gamma)$ with respect to w and γ equal to zero, i.e., $Z_t^\top\nabla h(\eta_t)v_t = 0$ and $1_n^\top\nabla h(\eta_t)v_t = 0$. Therefore, we have that

$$\begin{aligned}
\langle Z^*O_t(k_t O_t^\top w^* - w_t), \nabla h(\eta_t)v_t\rangle &= \langle (Z^*O_t - Z_t)(k_t O_t^\top w^* - w_t), \nabla h(\eta_t)v_t\rangle, \\
\langle (k_t\gamma^* - \gamma_t)1_n, \nabla h(\eta_t)v_t\rangle &= 0.
\end{aligned}$$

It follows that

$$J_2 = \|\Delta_{Z_t}\|_F^2 - 4\tau_Z\left\langle \Delta_{v_t} - \Delta_{Z_t}(k_t O_t^\top w^* - w_t), \nabla h(\eta_t)v_t\right\rangle + 4\tau_Z^2\|\nabla h(\eta_t)v_t\|^2\|w_t\|^2 \tag{B.20}$$

After plugging (B.19)-(B.20) and $\tau_Z = r_0\tau/\|Z_0\|_{op}^2$ into (B.18), we obtain that

$$\begin{aligned}
\tilde{e}_{t+1}^Z \leq &(1-\lambda)\left(\left(1 - \frac{r_0\tau}{e^{M_1}\kappa_{Z^*}^2}\rho\right)\tilde{e}_t^Z + r_0\tau C' e^{M_1}\zeta_n^2 k\right) + \lambda\left(\|\Delta_{Z_t}\|_F^2\|Z_0\|_{op}^2 + 2n\|\alpha_t - \alpha^*\|^2\right) \\
&- 4\lambda r_0\tau\left\langle \Delta_{v_t}, \nabla h(\eta_t)v_t\right\rangle
\end{aligned}$$

$$
\begin{aligned}
&+ 4\lambda r_0 \tau \left\langle \Delta_{Z_t}(k_t O_t^\top w^* - w_t), \boldsymbol{\nabla} h(\eta_t) v_t \right\rangle \\
&+ 4\lambda r_0^2 \tau^2 \frac{1}{\|Z_0\|_{op}^2} \|\boldsymbol{\nabla} h(\eta_t) v_t\|^2 \|w_t\|^2 \\
:= &\left(1 - \frac{(1-\lambda) r_0 \tau}{e^{M_1} \kappa_{Z^*}^2} \rho\right) \tilde{e}_t^Z + (1-\lambda) r_0 \tau C' e^{M_1} \zeta_n^2 k - 4\lambda r_0 \tau \cdot J_3 \\
&+ 4\lambda r_0 \tau \cdot J_4 + 4\lambda r_0 \tau^2 \cdot J_5.
\end{aligned} \tag{B.21}
$$

We first bound J_3 and J_5. Note that, in the proof of Lemma B.5.2, the terms have been shown to be bounded by

$$
-2J_3 = -2 \left\langle \Delta_{v_t}, \boldsymbol{\nabla} h(\eta_t) v_t \right\rangle \leq -D_1 + D_2 + D_3,
$$

and

$$
J_5 = r_0 \frac{\|v_0\|^2}{\|Z_0\|_{op}^2} \|w_t\|^2 D_4 = \min\{\frac{\|v_0\|^2}{\|Z_0\|_{op}^2}, 1\} \cdot \|w_t\|^2 D_4 \leq M^2 D_4,
$$

with $D_i, 1 \leq i \leq 4$, given in (B.8). By plugging the bounds established in (B.10), (B.11), (B.13), (B.14), and by Lemma B.5.6 $e_t^v \leq \|\Delta_{\eta_t}\|_F^2 / 2(\sqrt{2} - 1)$, it follows that

$$
\begin{aligned}
-4\lambda r_0 \tau J_3 + 4\lambda r_0 \tau^2 J_5 \leq &- 2\lambda r_0 \tau \frac{\mu/4}{\mu + 1/4} \||A| \circ \Delta_{\eta_t}\|_F^2 \\
&+ 2\lambda r_0 \tau \left(c_1 + \frac{1}{C e^{M_1 + M_3}} + \frac{c_0^2}{4 c_2 e^{M_1 + M_3}}\right) \|\Delta_{\eta_t}\|_F^2 \\
&- 2\lambda r_0 \tau \left(\frac{1}{\mu + 1/4} - c_2 - 2\tau M^2 C_3\right) \||A| \circ (\sigma(\eta_t) - Q)\|_F^2 \\
&+ 2\lambda r_0 \tau \left(\frac{1}{2c_1} + 2\tau M^2 C_3\right) \varphi_n^2.
\end{aligned} \tag{B.22}
$$

where c_1, c_2 are arbitrary positive constants, C_3 is a universal constant, c_0 is a sufficiently small constant, and C is a sufficiently large constant.

Next, to bound J_4, we have

$$
\begin{aligned}
J_4 &= \left\langle \Delta_{Z_t}(k_t O_t^\top w^* - w_t) v_t^\top, \boldsymbol{\nabla} h(\eta_t) \right\rangle \\
&\leq \|\boldsymbol{\nabla} h(\eta_t)\|_{op} \|\Delta_{Z_t}(k_t O_t^\top w^* - w_t) v_t^\top\|_* \\
&\leq \|\boldsymbol{\nabla} h(\eta_t)\|_{op} \|\Delta_{Z_t}(k_t O_t^\top w^* - w_t) v_t^\top\|_F \\
&= \|\boldsymbol{\nabla} h(\eta_t)\|_{op} \|\Delta_{Z_t}(k_t O_t^\top w^* - w_t)\| \|v_t\| \\
&\leq \left(\||A| \circ (\sigma(\eta_t) - Q)\|_{op} + \varphi_n\right) \|\Delta_{Z_t}(k_t O_t^\top w^* - w_t)\| \cdot 2\|v_0\|,
\end{aligned}
$$

where the second inequality is due to $\mathrm{rank}(\Delta_{Z_t}(k_tO_t^\top w^* - w_t)v_t^\top) = 1$. The last inequality is based on the triangle inequality and the assumptions $\|\Delta_{v_t}\| \le \delta\|v^*\|$ and $\|\Delta_{v_0}\| \le \delta\|v^*\|$. Furthermore, since

$\|\Delta_{Z_t}(k_tO_t^\top w^* - w_t)\| \le \|\Delta_{Z_t}\|_{op}\|k_tO_t^\top w^* - w_t\| \le \|\Delta_{Z_t}\|_F(\|w^*\| + \|w_t\|) \le 2M\|\Delta_{Z_t}\|_F$,

we obtain that

$$\begin{aligned}
4\lambda r_0\tau J_4 &\le 16\lambda r_0\tau M\left(\||A|\circ(\sigma(\eta_t) - Q)\|_{op} + \varphi_n\right)\|\Delta_{Z_t}\|_F\|Z_0\|_{op}\cdot\frac{\|v_0\|}{\|Z_0\|_{op}}\\
&\le 16\lambda r_0\tau M\left(\||A|\circ(\sigma(\eta_t) - Q)\|_{op} + \varphi_n\right)\|\Delta_{Z_t}\|_F\|Z_0\|_{op}\cdot\frac{1}{\sqrt{r_0}}\\
&\le 16\lambda\sqrt{r_0}\tau M\left(\||A|\circ(\sigma(\eta_t) - Q)\|_F + \varphi_n\right)(\tilde{e}_t^Z)^{\frac{1}{2}}\\
&\le 8\lambda r_0\tau M^2c_4\||A|\circ(\sigma(\eta_t) - Q)\|_F^2 + 8\lambda r_0\tau M^2c_4\varphi_n^2 + 8\lambda\tau(\frac{1}{c_4} + \frac{1}{c_4})\tilde{e}_t^Z,
\end{aligned} \tag{B.23}$$

where the last inequality is due to the triangle inequality with some constant c_4 to be specified later. By plugging (B.22) and (B.23) into (B.21), we obtain that

$$\begin{aligned}
\tilde{e}_{t+1}^Z \le{}& \left(1 - \frac{(1-\lambda)r_0\tau}{e^{M_1}\kappa_{Z^*}^2}\rho + \frac{16\lambda\tau}{c_4}\right)\tilde{e}_t^Z + (1-\lambda)r_0\tau C'e^{M_1}\zeta_n^2k\\
&- 2\lambda r_0\tau\frac{\mu/4}{\mu + 1/4}\||A|\circ\Delta_{\eta_t}\|_F^2 + 2\lambda r_0\tau\left(c_1 + \frac{1}{Ce^{M_1+M_3}} + \frac{c_0^2}{4c_2e^{M_1+M_3}}\right)\|\Delta_{\eta_t}\|_F^2\\
&- 2\lambda r_0\tau\left(\frac{1}{\mu + 1/4} - c_2 - 2\tau M^2C_3 - 4M^2c_4\right)\||A|\circ(\sigma(\eta_t) - Q)\|_F^2\\
&+ 2\lambda r_0\tau\left(\frac{1}{2c_1} + 2\tau M^2C_3 + 4M^2c_4\right)\varphi_n^2,
\end{aligned}$$

where $r_0 = \min\{1, \|Z_0\|_{op}^2/\|v_0\|^2\}$, c_1, c_2, and c_4 are arbitrary positive constants; ρ, C_3, and C' are universal constant; c_0 is a sufficiently small constant; C is a sufficiently large constant; and $\mu \asymp e^{-M_3}$. Given that $\lambda = \tilde{\lambda}r_0/e^{M_1}\kappa_{Z^*}^2$, we choose $c_1 = e^{-(M_1+M_3)}c_0/4$, $c_2 = c_0$, $c_4 = \sqrt{c}$, and $\tilde{\lambda}, \tau$ small enough such that there exists some universal positive constants ρ_1, ρ_2, and C'',

$$\frac{(1-\lambda)r_0\tau}{e^{M_1}\kappa_{Z^*}^2}\rho - \frac{16\lambda\tau}{c_4} > \frac{r_0\tau}{e^{M_1}\kappa_{Z^*}^2}\rho_1,$$

$$c_1 + \frac{1}{Ce^{M_1+M_3}} + \frac{c_0^2}{4c_2e^{M_1+M_3}} < \frac{c_0}{e^{M_1+M_3}},$$

$$
\begin{aligned}
\frac{\mu/4}{4(\mu+1/4)}e^{-M_1} - c_1 - \frac{1}{Ce^{M_1+M_3}} - \frac{c_0^2}{4c_2e^{M_1+M_3}} &> \rho_2 e^{-(M_1+M_3)}/8,\\
\frac{1}{\mu+1/4} - c_2 - 2\tau M^2 C_3 - 4M^2 c_4 &> 0,\\
\frac{1}{2c_1} + 2\tau M^2 C_3 + 4M^2 c_4 &< C'' e^{M_1+M_3}/2.
\end{aligned}
$$

Then we obtain that

$$
\begin{aligned}
\tilde{e}_{t+1}^Z \leq& \left(1 - \frac{r_0\tau}{e^{M_1}\kappa_{Z^*}^2}\rho_1\right)\tilde{e}_t^Z + (1-\lambda)r_0\tau C' e^{M_1}\zeta_n^2 k\\
&- 2\lambda r_0\tau\frac{\mu/4}{\mu+1/4}\||A|\circ\Delta_{\eta_t}\|_F^2 + 2\lambda r_0\tau\left(c_1 + \frac{1}{Ce^{M_1+M_3}} + \frac{c_0^2}{4c_2e^{M_1+M_3}}\right)\|\Delta_{\eta_t}\|_F^2\\
&+ \lambda r_0\tau C'' e^{M_1+M_3}\varphi_n^2.
\end{aligned}
$$

To bound the second line, we similarly discuss two cases

1. Suppose that $\|\Delta_{\eta_t}\|_F \leq 8e^{M_1}\||A|-P\|_{op}\|\Delta_{\eta_t}\|_{\max}$, then $\|\Delta_{\eta_t}\|_F \leq 16M_3e^{M_1}\zeta_n$ as $\|\Delta_{\eta_t}\|_{\max} \leq \|v_t v_t^\top\|_{\max} + \|v^* v^{*\top}\|_{\max} \leq 2M_3$. It follows that
$$
\begin{aligned}
&2\lambda r_0\tau\left(c_1 + \frac{1}{Ce^{M_1+M_3}} + \frac{c_0^2}{4c_2e^{M_1+M_3}}\right)\|\Delta_{\eta_t}\|_F^2\\
&\leq 2\lambda r_0\tau\frac{c_0}{e^{M_1+M_3}}\cdot 16^2 M_3^2 e^{2M_1}\zeta_n^2 \leq \lambda r_0\tau e^{M_1}\zeta_n^2,
\end{aligned}
$$
because $M_3^2 \leq e^{M_3}$ and c_0 is a sufficiently small constant. Then, we have
$$
\begin{aligned}
\tilde{e}_{t+1}^Z \leq& \left(1 - \frac{r_0\tau}{e^{M_1}\kappa_{Z^*}^2}\rho_1\right)\tilde{e}_t^Z + r_0\tau C' e^{M_1}\zeta_n^2 k - \lambda r_0\tau\rho_2 e^{-M_3}\||A|\circ\Delta_{\eta_t}\|_F^2\\
&+ \lambda r_0\tau C'' e^{M_1+M_3}\varphi_n^2 \qquad \text{(B.24)}
\end{aligned}
$$

2. Suppose that $\|\Delta_{\eta_t}\|_F \geq 8e^{M_1}\||A|-P\|_{op}\|\Delta_{\eta_t}\|_{\max}$. Since $\operatorname{rank}(\Delta_{\eta_t})$ is at most 2 and $P_{min} = \min_{i,j} P_{ij} \geq \frac{1}{2}e^{-M_1}$, it follows that $\frac{2\operatorname{rank}(\Delta_{\eta_t})}{P_{min}}\||A|-P\|_{op}\|\Delta_{\eta_t}\|_{\max} \leq 8e^{M_1}\||A|-P\|_{op}\|\Delta_{\eta_t}\|_{\max} \leq \|\Delta_{\eta_t}\|_F$. By Lemma B.5.1, we have
$$
\||A|\circ\Delta_{\eta_t}\|_F^2 \geq \frac{1}{2}P_{min}\|\Delta_{\eta_t}\|_F^2 \geq \frac{1}{4}e^{-M_1}\|\Delta_{\eta_t}\|_F^2.
$$
Then it follows that
$$
\begin{aligned}
&-2\lambda r_0\tau\frac{\mu/4}{\mu+1/4}\||A|\circ\Delta_{\eta_t}\|_F^2 + 2\lambda r_0\tau\left(c_1 + \frac{1}{Ce^{M_1+M_3}} + \frac{c_0^2}{4c_2e^{M_1+M_3}}\right)\|\Delta_{\eta_t}\|_F^2\\
\leq& -2\lambda r_0\tau\left(\frac{\mu/4}{4(\mu+1/4)}e^{-M_1} - c_1 - \frac{1}{Ce^{M_1+M_3}} - \frac{c_0^2}{4c_2e^{M_1+M_3}}\right)\|\Delta_{\eta_t}\|_F^2
\end{aligned}
$$

$$< -\lambda r_0 \tau \rho_2 e^{-(M_1+M_3)} \|\Delta_{\eta_t}\|_F^2,$$

and

$$\tilde{e}_{t+1}^Z \leq \left(1 - \frac{r_0\tau}{e^{M_1}\kappa_{Z^*}^2}\rho_1\right)\tilde{e}_t^Z + (1-\lambda)r_0\tau C' e^{M_1}\zeta_n^2 k - \lambda r_0\tau\rho_2 e^{-(M_1+M_3)}\|\Delta_{\eta_t}\|_F^2$$
$$+ \lambda r_0 \tau C'' e^{M_1+M_3}\varphi_n^2 \tag{B.25}$$

Combining (B.24) and (B.25) in two cases, we have

$$\tilde{e}_{t+1}^Z \leq \left(1 - \frac{r_0\tau}{e^{M_1}\kappa_{Z^*}^2}\rho_1\right)\tilde{e}_t^Z - \lambda r_0\tau\rho_2 e^{-M_3}\min\{\||A|\circ\Delta_{\eta_t}\|_F^2, e^{-M_1}\|\Delta_{\eta_t}\|_F^2\}$$
$$+ r_0\tau C' e^{M_1}\zeta_n^2 k + \lambda r_0\tau C'' e^{M_1+M_3}\varphi_n^2,$$

which completes the proof. □

Proof of Lemma B.6.2. Given (w_t, γ_t) obtained from the line 5 in Algorithm III.1, we denote $v_t = Z_t w_t + \gamma_t 1_n$ and $v_t^* = Z_t O_t w^* + \gamma^* 1_n$. Let $\eta_t = v_t v_t^\top$ and $\eta_t^* = v_t^* v_t^{*\top}$. Since (w_t, γ_t) minimizes the loss function $\mathcal{L}_\lambda(\alpha, Z, w, \gamma)$ with (α, Z) fixed as (α_t, Z_t), we have $\mathcal{L}_\lambda(\alpha_t, Z_t, O_t w^*, \gamma^*) \geq \mathcal{L}_\lambda(\alpha_t, Z_t, w_t, \gamma_t)$. As the first part in the weighted loss is constant given (α, Z) fixed, it follows that

$$0 \geq \mathcal{L}_\lambda(\alpha_t, Z_t, w_t, \gamma_t) - \mathcal{L}_\lambda(\alpha_t, Z_t, O_t w^*, \gamma^*)$$
$$= \lambda(h(\eta_t) - h(\eta_t^*)), \tag{B.26}$$

with h given in (B.7). It is straightforward to verify that for any $\eta \in \mathcal{F}_\eta(n, M_3)$

$$\boldsymbol{\nabla}^2 h(\eta) = \operatorname{diag}\Big(vec\big(|A|\circ\sigma(\eta)\circ(1-\sigma(\eta))\big)\Big) \succeq \mu\cdot\operatorname{diag}(vec(|A|))$$

with $\mu = \frac{e^{M_3}}{(1+e^{M_3})^2} \asymp e^{-M_3}$. Therefore, we have

$$h(\eta_t) - h(\eta_t^*) \geq \langle \boldsymbol{\nabla} h(\eta_t^*), \eta_t - \eta_t^*\rangle + \frac{\mu}{2}\||A|\circ(\eta_t - \eta_t^*)\|_F^2. \tag{B.27}$$

Combing (B.26) and (B.27), it follows that

$$\frac{\mu}{2}\||A|\circ(\eta_t - \eta_t^*)\|_F^2 \leq -\langle \boldsymbol{\nabla} h(\eta_t^*), \eta_t - \eta_t^*\rangle$$
$$\leq |\langle |A|\circ((1+A)/2 - Q), \eta_t - \eta_t^*\rangle| + |\langle |A|\circ(\sigma(\eta_t^*) - \sigma(\eta^*)), \eta_t - \eta_t^*\rangle|$$
$$= |\langle |A|\circ((1+A)/2 - Q), \eta_t - \eta_t^*\rangle| + |\langle \sigma(\eta_t^*) - \sigma(\eta^*), |A|\circ(\eta_t - \eta_t^*)\rangle|$$

$$\leq \varphi_n \left\|\eta_t - \eta_t^*\right\|_* + \|\sigma(\eta_t^*) - \sigma(\eta^*)\|_F \cdot \||A| \circ (\eta_t - \eta_t^*)\|_F$$
$$\leq \sqrt{2}\varphi_n \left\|\eta_t - \eta_t^*\right\|_F + \|\sigma(\eta_t^*) - \sigma(\eta^*)\|_F \cdot \||A| \circ (\eta_t - \eta_t^*)\|_F,$$

where the second last inequality is due to the Holder's inequality; the last inequality holds because $\mathrm{rank}(\eta_t - \eta_t^*)$ is at most 2. By solving the above quadratic inequality in terms of $\||A| \circ (\eta_t - \eta_t^*)\|_F$ and plugging $\mu \asymp e^{-M_3}$, we have

$$\||A| \circ (\eta_t - \eta_t^*)\|_F \leq e^{M_3}\|\sigma(\eta_t^*) - \sigma(\eta^*)\|_F + e^{M_3/2}\sqrt{\varphi_n \left\|\eta_t - \eta_t^*\right\|_F}. \tag{B.28}$$

Following the proof of Theorem 3.5.1, we consider two cases:

1. Suppose that $\|\eta_t - \eta_t^*\|_F \leq 8e^{M_1}\||A| - P\|_{op}\|\eta_t - \eta_t^*\|_{\max}$. Since $\|\eta_t - \eta_t^*\|_{\max} \leq \|v_t {v_t}^\top\|_{\max} + \|v_t^* {v_t^*}^\top\|_{\max} \leq 2M_3$, it follows that
$$\|\eta_t - \eta_t^*\|_F \leq 16M_3 e^{M_1}\zeta_n. \tag{B.29}$$

2. Suppose that $\|\eta_t - \eta_t^*\|_F \geq 8e^{M_1}\||A| - P\|_{op}\|\eta_t - \eta_t^*\|_{\max}$. Since $\mathrm{rank}(\eta_t - \eta_t^*)$ is at most 2 and $P_{min} = \min_{i,j} P_{ij} \geq \frac{1}{2}e^{-M_1}$, it follows that
$$\frac{2\,\mathrm{rank}(\eta_t - \eta_t^*)}{P_{min}}\||A| - P\|_{op}\|\eta_t - \eta_t^*\|_{\max} \leq 8e^{M_1}\||A| - P\|_{op}\|\eta_t - \eta_t^*\|_{\max} \leq \|\eta_t - \eta_t^*\|_F.$$
By Lemma B.5.1, we have
$$\||A| \circ (\eta_t - \eta_t^*)\|_F^2 \geq \frac{1}{2}P_{min}\|\eta_t - \eta_t^*\|_F^2 \geq \frac{1}{4}e^{-M_1}\|\eta_t - \eta_t^*\|_F^2.$$
Plugging the above inequality into (B.28), we have
$$\|\eta_t - \eta_t^*\|_F \leq 2e^{M_1/2 + M_3}\|\sigma(\eta_t^*) - \sigma(\eta^*)\|_F + 2e^{(M_1+M_3)/2}\sqrt{\varphi_n} \cdot \sqrt{\|\eta_t - \eta_t^*\|_F}.$$
By solving the above quadratic inequality in terms of $\sqrt{\|\eta_t - \eta_t^*\|_F}$, we have
$$\sqrt{\|\eta_t - \eta_t^*\|_F} \leq 2e^{(M_1+M_3)/2}\sqrt{\varphi_n} + \sqrt{2}e^{M_1/4 + M_3/2}\sqrt{\|\sigma(\eta_t^*) - \sigma(\eta^*)\|_F},$$
and it follows that
$$\begin{aligned}\|\eta_t - \eta_t^*\|_F &\leq 8e^{M_1+M_3}\varphi_n + 4e^{M_1/2+M_3}\|\sigma(\eta_t^*) - \sigma(\eta^*)\|_F \\ &\leq 8e^{M_1+M_3}\varphi_n + e^{M_1/2+M_3}\|\eta_t^* - \eta^*\|_F,\end{aligned} \tag{B.30}$$
where the last inequality holds because $\sigma'(\cdot) \leq 1/4$.

Combing the bounds (B.29) and (B.30) in two cases together, it follows that

$$\begin{aligned}\|\Delta_{\eta_t}\|_F &\leq \|\eta_t - \eta_t^*\|_F + \|\eta_t^* - \eta^*\|_F \\ &\leq 16e^{M_1} \max\{M_3\zeta_n, e^{M_3}\varphi_n\} + e^{M_1/2+M_3}\|\eta_t^* - \eta^*\|_F \\ &\leq 16e^{M_1+M_3} \max\{\zeta_n, \varphi_n\} + e^{M_1/2+M_3}\|\eta_t^* - \eta^*\|_F.\end{aligned}$$

Since we have $dist(v_t^*, v^*) \leq \|v_t^* - v^*\| = \|(Z_tO_t - Z^*)w^*\| \leq \|\Delta_{Z_t}\|_{op}\|w^*\| \leq \|\Delta_{Z_t}\|_F\|w^*\|$, then, by Lemma B.5.7, it follows that

$$e^{M_1/2+M_3}\|\eta_t^* - \eta^*\|_F \leq e^{M_1/2+M_3}(2 + \|\Delta_{Z_t}\|_F\|w^*\|)\|v^*\| \cdot dist(v_t^*, v^*),$$

and $\|\Delta_{\eta_t}\|_F \leq 16e^{M_1+M_3} \max\{\zeta_n, \varphi_n\} + e^{M_1/2+M_3}(2 + \|\Delta_{Z_t}\|_F\|w^*\|)\|v^*\|\|\Delta_{Z_t}\|_F\|w^*\|$.

□

Proof of Lemma B.6.3. By Lemma B.6.2, we have

$$\|\Delta_{\eta_t}\|_F \leq 16e^{M_1+M_3} \max\{\zeta_n, \varphi_n\} + e^{M_1/2+M_3}(2 + \|\Delta_{Z_t}\|_F\|w^*\|)\|v^*\|\|\Delta_{Z_t}\|_F\|w^*\|. \tag{B.31}$$

Note that, since $\|Z^*\|_{op}^2 \geq C\kappa_{Z^*}^3\zeta_n\sqrt{\tau k}e^{2M_1+3M_3/2}$ and $\|v^*\|^2 \geq C\varphi_n\sqrt{\tau}e^{3M_1/2+2M_3}$ for a sufficiently large constant C, then we choose C large enough such that

$$\frac{16}{C\sqrt{\tau}\min\{1, \kappa_{Z^*}\sqrt{k}\|w^*\|^2\}} \leq c_0.$$

By combining $\|v^*\|^2 = \|Z^*w^*\|^2 + \gamma^{*2}n \geq \|Z^*\|_{op}^2\|w^*\|^2/\kappa_{Z^*}^2$, it follows that

$$16e^{M_1+M_3}\zeta_n \leq \frac{16e^{-(M_1+M_3/2)}}{C\kappa_{Z^*}\sqrt{\tau k}}\frac{\|Z^*\|_{op}^2}{\kappa_{Z^*}^2} \leq \frac{16e^{-(M_1+M_3)/2}}{C\kappa_{Z^*}\sqrt{\tau k}}\frac{\|v^*\|^2}{\|w^*\|^2} \leq c_0e^{-(M_1+M_3)/2}\|v^*\|^2, \tag{B.32}$$

$$16e^{M_1+M_3}\varphi_n \leq \frac{16\|v^*\|^2}{C\sqrt{\tau}e^{M_1/2+M_3}} \leq c_0e^{-(M_1+M_3)/2}\|v^*\|^2. \tag{B.33}$$

Furthermore, since with a sufficiently small c_0

$$\|\Delta_{Z_t}\|_F\|w^*\| \leq c_0e^{-M_1-3M_3/2}\|Z^*\|_{op}\|w^*\|/\kappa_{Z^*}^2 \leq c_0e^{-M_1-3M_3/2}\|v^*\|,$$

where the last inequality is based on $\|v^*\| \geq \|Z^*\|_{op}\|w^*\|/\kappa_{Z^*}$ and $\kappa_{Z^*} \geq 1$, then we have

$$e^{M_1/2+M_3}(2 + \|\Delta_{Z_t}\|_F\|w^*\|)\|v^*\|\|\Delta_{Z_t}\|_F\|w^*\| \leq 3c_0e^{-(M_1+M_3)/2}\|v^*\|^2. \tag{B.34}$$

By plugging the bounds (B.32)-(B.34) into (B.31) and Lemma B.5.6, we have

$$\|\Delta_{v_t}\| \le \frac{1}{2(\sqrt{2}-1)\|v^*\|}\|\Delta_{\eta_t}\|_F \le \frac{2}{(\sqrt{2}-1)} c_0 e^{-(M_1+M_3)/2}\|v^*\|,$$

which completes the proof. □

Proof of Lemma B.6.4. We first prove that it suffices to show $\tilde{e}_t^Z \le c_0^2 e^{-2M_1-3M_3}\|Z^*\|_{op}^4/4\kappa_{Z^*}^4$ for all $t \ge 0$. Because suppose the above bound for the error metric holds, then we have that

$$\begin{aligned}\|\Delta_{Z_t}\|_F(\|Z^*\|_{op} - \|\Delta_{Z_t}\|_F) &\le \|\Delta_{Z_t}\|_F(\|Z^*\|_{op} - \|\Delta_{Z_t}\|_{op}) \le \|\Delta_{Z_t}\|_F\|Z_t\|_{op} \\ &\le (\tilde{e}_t^Z)^{\frac{1}{2}} \le c_0 e^{-M_1-3M_3/2}\|Z^*\|_{op}^2/2\kappa_{Z^*}^2.\end{aligned}$$

By solving the above quadratic inequality in terms of $\|\Delta_{Z_t}\|_F$, we obtain that

$$\begin{aligned}\|\Delta_{Z_t}\|_F &\le \frac{\|Z^*\|_{op}}{2} - \sqrt{\frac{\|Z^*\|_{op}^2}{4} - \frac{c_0\|Z^*\|_{op}^2}{2e^{M_1+3M_3/2}\kappa_{Z^*}^2}} \\ &= \frac{\frac{c_0\|Z^*\|_{op}^2}{2e^{M_1+3M_3/2}\kappa_{Z^*}^2}}{\frac{\|Z^*\|_{op}}{2} + \sqrt{\frac{\|Z^*\|_{op}^2}{4} - \frac{c_0\|Z^*\|_{op}^2}{2e^{M_1+3M_3/2}\kappa_{Z^*}^2}}} \\ &\le \frac{\frac{c_0\|Z^*\|_{op}^2}{2e^{M_1+3M_3/2}\kappa_{Z^*}^2}}{\frac{\|Z^*\|_{op}}{2}} = c_0 e^{-M_1-3M_3/2}\|Z^*\|_{op}/\kappa_{Z^*}^2.\end{aligned}$$

Therefore, next, we prove $\tilde{e}_t^Z \le c_0^2 e^{-2M_1-3M_3}\|Z^*\|_{op}^4/4\kappa_{Z^*}^4$ for all $t \ge 0$ by induction as below. The initialization assumption makes it hold for $t = 0$. Suppose $\tilde{e}_t^Z \le c_0^2 e^{-2M_1-3M_3}\|Z^*\|_{op}^4/4\kappa_{Z^*}^4$ hold, then we have $\|\Delta_{Z_t}\|_F \le c_0 e^{-M_1-3M_3/2}\|Z^*\|_{op}/\kappa_{Z^*}^2$, and further by Lemma B.6.3 we have $\|\Delta_{v_t}\| \le c e^{-(M_1+M_3)/2}\|v^*\|$ with a sufficiently small constant c. Then by Lemma B.6.1, it follows that

$$\begin{aligned}\tilde{e}_{t+1}^Z &\le \left(1 - \frac{r_0\tau\rho_1}{e^{M_1}\kappa_{Z^*}^2}\right)\tilde{e}_t^Z + r_0\tau C' e^{M_1}\zeta_n^2 k + \lambda r_0\tau C'' e^{M_1+M_3}\varphi_n^2 \\ &\le \left(1 - \frac{r_0\tau\rho_1}{e^{M_1}\kappa_{Z^*}^2}\right) c_0^2 e^{-2M_1-3M_3}\frac{\|Z^*\|_{op}^4}{4\kappa_{Z^*}^4} + r_0\tau C' e^{M_1}\zeta_n^2 k + \tilde{\lambda} r_0^2\tau C'' e^{M_3}\varphi_n^2\frac{1}{\kappa_{Z^*}^2} \\ &= c_0^2 e^{-2M_1-3M_3}\frac{\|Z^*\|_{op}^4}{4\kappa_{Z^*}^4} \\ &\quad\cdot\left(1 - \frac{r_0\tau\rho_1}{e^{M_1}\kappa_{Z^*}^2} + \frac{4r_0\tau C' e^{3(M_1+M_3)}\zeta_n^2\kappa_{Z^*}^4 k}{c_0^2\|Z^*\|_{op}^4} + \frac{4\tilde{\lambda} r_0^2\tau C'' e^{2M_1+4M_3}\varphi_n^2\kappa_{Z^*}^2}{c_0^2\|Z^*\|_{op}^4}\right)\end{aligned}$$

$$\leq c_0^2 e^{-2M_1-3M_3}\frac{\|Z^*\|_{op}^4}{4\kappa_{Z^*}^4}\left(1-\frac{r_0\tau\rho_1}{e^{M_1}\kappa_{Z^*}^2}+\frac{4r_0C'}{C^2c_0^2e^{M_1}\kappa_{Z^*}^2}+\frac{4\tilde{\lambda}r_0^2C''\kappa_{Z^*}^2\|v^*\|^4}{C^2c_0^2e^{M_1}\|Z^*\|_{op}^4}\right),$$

where the last inequality is obtained by plugging $\|Z^*\|_{op}^2 \geq Ce^{M_1}\kappa_{Z^*}^2\zeta_n\sqrt{\tau k}e^{M_1+3M_3/2}\kappa_{Z^*}$ and $\|v^*\|^2 \geq Ce^{M_1+M_3}\varphi_n\sqrt{\tau}e^{M_1/2+M_3}$. Given that $r_0 \leq \|Z_0\|_{op}^2/\|v_0\|^2 \asymp \|Z^*\|_{op}^2/\|v^*\|^2$, we choose C large enough such that

$$\frac{4C'}{C^2c_0^2} < \frac{\tau\rho_1}{2} \quad\text{and}\quad \frac{4\tilde{\lambda}r_0^2C''\kappa_{Z^*}^2\|v^*\|^4}{C^2c_0^2\|Z^*\|_{op}^4} < \frac{4\tilde{\lambda}C''\kappa_{Z^*}^2}{C^2c_0^2} < \frac{r_0\tau\rho_1}{2\kappa_{Z^*}^2},$$

then it follows that

$$\tilde{e}_{t+1}^Z \leq c_0^2 e^{-2M_1-3M_3}\frac{\|Z^*\|_{op}^4}{4\kappa_{Z^*}^4},$$

which completes the proof. □

B.7 Proof of Proposition 3.5.2 and Discussion on the Assumptions

From below, we consider the parameter space $\mathcal{F}(n, k, M_1, M_2, M_3)$ with fixed M_i and k.

B.7.1 Proof of Proposition 3.5.2

The following two lemmas are direct results of Proposition 3.5.1 and Theorem 3.5.1 respectively, which will be used to prove Proposition 3.5.2.

Lemma B.7.1. *Given the estimators $(\bar{\alpha}, \bar{Z})$ obtained from Algorithm B.1. Suppose the conditions in Proposition 3.5.1 hold, and the singular values of the sample covariance $Z^{*\top}Z^*/n$ are of constant order, then $\|\Delta_{\bar{Z}}\|_F = \mathcal{O}(1)$.*

Lemma B.7.2. *Given the estimators $\tilde{v}$ obtained from Algorithm B.2. Suppose the conditions in Theorem 3.5.1 hold, and $\|v^*\|^2/n$ is of constant order, then $\|\Delta_{\tilde{v}}\|_F = \mathcal{O}(1)$.*

Proof of Proposition 3.5.2. Recall that $\Delta_{\bar{Z}} = \bar{Z} - Z^*\bar{O}$ with $\bar{O} = \arg\min_{O\in O(k)} \|\bar{Z} - Z^*O\|_F$. Without loss of generality, we assume $\bar{O} = I_k$ in the proof, otherwise we replace the output of Algorithm B.1 by $\bar{Z}\bar{O}^\top$. Therefore, $\Delta_{\bar{Z}} = \bar{Z} - Z^*$. Based on the definition of $\hat{Z}$, we have

$$\begin{aligned}\|\hat{Z} - Z^*\|_F^2 &= \|\bar{Z} - 2\tau_z(1-\lambda)\nabla g(\bar{\Theta})\bar{Z} - 2\tau_z\lambda\nabla h(\bar{\eta})\bar{v}\bar{w}^\top - Z^*\|_F^2 \\ &= \|\Delta_{\bar{Z}} - 2\tau_z\lambda\nabla h(\bar{\eta})\bar{v}\bar{w}^\top\|_F^2 \\ &= \|\Delta_{\bar{Z}}\|_F^2 - 2\tau_z\lambda\langle\Delta_{\bar{Z}}\bar{w}, \nabla h(\bar{\eta})\bar{v}\rangle + 4\tau_z^2\lambda^2\|\nabla h(\bar{\eta})\bar{v}\|^2\|\bar{w}\|^2,\end{aligned}$$

where the second equality is due to the definition of $\bar{Z}$. Recall that $(\bar{\alpha}, \bar{Z})$ minimizes $g(\Theta)$ subject to $\Theta = \alpha 1_n^\top + 1_n\alpha^\top + ZZ^\top$, which implies that $\nabla g(\bar{\Theta})\bar{Z} = 0$. Note that $\bar{v}$ is independent of $B = |A| \circ (A+1)/2$ conditional on $|A|$, and $\bar{Z}$ only depends on $|A|$. Therefore, we have

$$\begin{aligned}\mathbb{E}\langle\Delta_{\bar{Z}}\bar{w}, \nabla h(\bar{\eta})\bar{v}\rangle &= \Big\langle\Delta_{\bar{Z}}\bar{w}, \nabla H(\bar{\eta})\bar{v}\Big\rangle + \mathbb{E}\Big\langle\Delta_{\bar{Z}}\bar{w}, \big(\nabla h(\bar{\eta}) - \nabla H(\bar{\eta})\big)\bar{v}\Big\rangle \\ &= \Big\langle\Delta_{\bar{Z}}\bar{w}, \nabla H(\bar{\eta})\bar{v}\Big\rangle + \mathbb{E}\Big\langle\Delta_{\bar{Z}}\bar{w}, \big(|A| \circ \sigma(\eta^*) - B\big)\bar{v}\Big\rangle \\ &= \Big\langle\Delta_{\bar{Z}}\bar{w}, \nabla H(\bar{\eta})\bar{v}\Big\rangle + \Big\langle\Delta_{\bar{Z}}\bar{w}, \mathbb{E}\big(|A| \circ \sigma(\eta^*) - B\big)\bar{v}\Big\rangle \\ &= \Big\langle\Delta_{\bar{Z}}\bar{w}, \nabla H(\bar{\eta})\bar{v}\Big\rangle,\end{aligned}$$

where $H(\eta) = -\sum_{i,j}|A_{ij}|Q_{ij}\eta_{ij} + |A_{ij}|\log(1-\sigma(\eta_{ij}))$ and $\nabla H(\eta) = |A| \circ (\sigma(\eta) - \sigma(\eta^*))$. And it follows that

$$\mathbb{E}\|\hat{Z} - Z^*\|_F^2 = \|\Delta_{\bar{Z}}\|_F^2 - 2\tau_z\lambda\,\Big\langle\Delta_{\bar{Z}}\bar{w}, \nabla H(\bar{\eta})\bar{v}\Big\rangle + 4\tau_z^2\lambda^2\,\mathbb{E}\|\nabla h(\bar{\eta})\bar{v}\|^2\|\bar{w}^\top\|^2. \tag{B.35}$$

The above term is quadratic in terms of λ, therefore $\mathbb{E}\|\hat{Z} - Z^*\|_F^2$ is minimized at

$$\lambda_{opt} := \frac{\Big\langle\Delta_{\bar{Z}}\bar{w}, \nabla H(\bar{\eta})\bar{v}\Big\rangle}{4\tau_z\mathbb{E}\|\nabla h(\bar{\eta})\bar{v}\|^2\|\bar{w}^\top\|^2}. \tag{B.36}$$

Note that if λ_{opt} is strictly positive, then for any $\lambda \in (0, 2\lambda_{opt})$, we have

$$-2\tau_z\lambda\,\Big\langle\Delta_{\bar{Z}}\bar{w}, \nabla H(\bar{\eta})\bar{v}\Big\rangle + 4\tau_z^2\lambda^2\,\mathbb{E}\|\nabla h(\bar{\eta})\bar{v}\|^2\|\bar{w}^\top\|^2 < 0,$$

and it follows that

$$\mathbb{E}\|\Delta_{\hat{Z}}\|_F^2 \le \mathbb{E}\|\hat{Z} - Z^*\|_F^2 < \mathbb{E}\|\Delta_{\bar{Z}}\|_F^2,$$

where the first inequality is based on the definition $\|\Delta_{\hat{Z}}\|_F^2 = \min_{O\in O(k)}\|\hat{Z} - Z^*O\|_F^2$.

Next, we analyze under which scenario λ_{opt} is more likely to be positive. For notational simplicity, we substitute $[1_n, \bar{Z}]$, $[1_n, \hat{Z}]$, $[\bar{\gamma}, \bar{w}^\top]^\top$, $[\hat{\gamma}, \hat{w}^\top]^\top$, and $\bar{w}$ by $\bar{Z}$, $\hat{Z}$, $\bar{w}$, $\hat{w}$,and $\bar{w}_{(-1)}$ respectively hereafter in the proof. Note that, after this substitution, the values of $\|\Delta_{\bar{Z}}\|_F$, $\|\Delta_{\hat{Z}}\|_F$, and $\Delta_{\bar{Z}}\bar{w}$ keep the same and we have $\bar{v} = \bar{Z}\bar{w}$. We decompose $\bar{\eta} - \eta^*$ into terms

$$\begin{aligned}
\bar{\eta} - \eta^* &= (\Delta_{\bar{Z}}\bar{w}\bar{w}^\top \bar{Z}^\top + \bar{Z}\bar{w}\bar{w}^\top \Delta_{\bar{Z}}^\top) + Z^*(\bar{w}\bar{w}^\top - w^* w^{*\top})Z^{*\top} - \Delta_{\bar{Z}}\bar{w}\bar{w}^\top \Delta_{\bar{Z}}^\top \\
&:= \tilde{T}_1 + \tilde{T}_2 - \tilde{T}_3.
\end{aligned}$$

Then we have

$$\begin{aligned}
2\Big\langle \Delta_{\bar{Z}}\bar{w}, \boldsymbol{\nabla} H(\bar{\eta})\bar{v} \Big\rangle &= 2\Big\langle \Delta_{\bar{Z}}\bar{w}\bar{w}^\top \bar{Z}^\top, \boldsymbol{\nabla} H(\bar{\eta}) \Big\rangle \\
&= \Big\langle \Delta_{\bar{Z}}\bar{w}\bar{w}^\top \bar{Z}^\top + \bar{Z}\bar{w}\bar{w}^\top \Delta_{\bar{Z}}^\top, \boldsymbol{\nabla} H(\bar{\eta}) \Big\rangle = \Big\langle \tilde{T}_1, \boldsymbol{\nabla} H(\bar{\eta}) \Big\rangle,
\end{aligned}$$

where the second equality holds due to the symmetry of $\boldsymbol{\nabla} H(\bar{\eta})$. Moreover, $\boldsymbol{\nabla} H(\bar{\eta}) = |A| \circ (\sigma(\bar{\eta}) - \sigma(\eta^*)) = |A| \circ \sigma'(\tilde{\eta}) \circ (\bar{\eta} - \eta^*)$ with some $[\tilde{\eta}]_{ij} = \tilde{\eta}_{ij}$ located between $\bar{\eta}_{ij}$ and η^*_{ij}. Let $\xi_{ij}^2 = \sigma'(\tilde{\eta}_{ij}) = \sigma(\tilde{\eta}_{ij})(1 - \sigma(\tilde{\eta}_{ij})) \geq \frac{e^{M_3}}{(1+e^{M_3})^2} > 0$ and $\xi = [\xi_{ij}]$, then $\boldsymbol{\nabla} H(\bar{\eta}) = |A| \circ \xi \circ \xi \circ (\tilde{T}_1 + \tilde{T}_2 - \tilde{T}_3)$. It follows that

$$\begin{aligned}
2\Big\langle \Delta_{\bar{Z}}\bar{w}, \boldsymbol{\nabla} H(\bar{\eta})\bar{v} \Big\rangle &= \Big\langle \tilde{T}_1, |A| \circ \xi \circ \xi \circ (\tilde{T}_1 + \tilde{T}_2 - \tilde{T}_3) \Big\rangle \\
&= \big\| |A| \circ \xi \circ \tilde{T}_1 \big\|_F^2 + \Big\langle |A| \circ \xi \circ \tilde{T}_1, |A| \circ \xi \circ (\tilde{T}_2 - \tilde{T}_3) \Big\rangle \\
&\geq \big\| |A| \circ \xi \circ \tilde{T}_1 \big\|_F \Big(\big\| |A| \circ \xi \circ \tilde{T}_1 \big\|_F - \big\| |A| \circ \xi \circ (\tilde{T}_2 - \tilde{T}_3) \big\|_F \Big) \\
&\geq \big\| |A| \circ \xi \circ \tilde{T}_1 \big\|_F \Big(\big\| |A| \circ \xi \circ \tilde{T}_1 \big\|_F - \big\| |A| \circ \xi \circ \tilde{T}_2 \big\|_F - \big\| |A| \circ \xi \circ \tilde{T}_3 \big\|_F \Big),
\end{aligned} \tag{B.37}$$

where the second equality holds because the elements of $|A|$ are binary and thereby $|A| \circ |A| = |A|$. By Lemma B.7.1, we have $\big\|\Delta_{\bar{Z}}\big\|_F = \mathcal{O}(1)$ and further $\big\|\bar{Z}\big\|_{op} \leq \big\|Z^*\big\|_{op} + \big\|\Delta_{\bar{Z}}\big\|_{op} \leq \boldsymbol{\sigma}_1(Z^*) + \big\|\Delta_{\bar{Z}}\big\|_F = \mathcal{O}(\sqrt{n})$, where $\boldsymbol{\sigma}_1(Z^*)$ is the largest singular value of Z^*. Together with $\|\Delta_{\bar{w}}\| = \mathcal{O}(1/\sqrt{n})$ and $\|w^*\| \leq \|\bar{w}\| + \mathcal{O}(1/\sqrt{n})$, we can bound each term below

$$\big\| |A| \circ \xi \circ \tilde{T}_1 \big\|_F \leq \frac{1}{2} \big\|\tilde{T}_1\big\|_F \leq \big\|\bar{Z}\bar{w}\bar{w}^\top \Delta_{\bar{Z}}^\top\big\|_F \leq \big\|\bar{Z}\big\|_{op} \big\|\Delta_{\bar{Z}}\big\|_F \big\|\bar{w}\big\|^2 = \mathcal{O}(\sqrt{n}) \big\|\bar{w}\big\|^2, \tag{B.38}$$

$$\begin{aligned}\big\||A|\circ\xi\circ\tilde{T}_2\big\|_F &\le \frac{1}{2}\big\|\tilde{T}_2\big\|_F \le \frac{1}{2}\big\|Z^*\big\|_{op}^2\big\|\bar{w}\bar{w}^\top - w^*w^{*\top}\big\|_F \\ &\le \frac{1}{2}\big\|Z^*\big\|_{op}^2\big\|\Delta_{\bar{w}}\big\|(\|\bar{w}\| + \|w^*\|) \le \mathcal{O}(\sqrt{n})\|\bar{w}\|, \qquad \text{(B.39)} \\ \big\||A|\circ\xi\circ\tilde{T}_3\big\|_F &\le \frac{1}{2}\big\|\tilde{T}_3\big\|_F \le \frac{1}{2}\big\|\Delta_{\bar{Z}}\big\|_F^2\big\|\bar{w}\big\|^2 = \mathcal{O}(1)\big\|\bar{w}\big\|^2. \qquad \text{(B.40)}\end{aligned}$$

The order in (B.40) suggest that the first two terms in (B.38) and (B.39) are the dominating terms. Since the denominator in (B.36) is always positive, λ_{opt} is positive if and only if the numerator $\mathbb{E}\Big\langle \Delta_{\bar{Z}}\bar{w}, \boldsymbol{\nabla} H(\bar{\eta})\bar{v}\Big\rangle$ is positive. The upper bounds in (B.38) and (B.39) suggest that, with large enough $\|\bar{w}\|$, the upper bound of $\big\||A|\circ\xi\circ\tilde{T}_1\big\|_F$ is greater than that of $\big\||A|\circ\xi\circ\tilde{T}_2\big\|_F$, which more likely leads to a positive $\big\||A|\circ\xi\circ\tilde{T}_1\big\|_F - \big\||A|\circ\xi\circ\tilde{T}_2\big\|_F - \big\||A|\circ\xi\circ\tilde{T}_3\big\|_F$ and thereby a positive λ_{opt} based on (B.37).

Finally, if $\lambda_{opt} > 0$, we choose $\lambda = \lambda_{opt}$. By plugging λ_{opt} into (B.35) and the bound in (B.37), the improvement is at least

$$\begin{aligned}\mathbb{E}\|\Delta_{\bar{Z}}\|_F^2 - \mathbb{E}\|\Delta_{\hat{Z}}\|_F^2 \ge \mathbb{E}\|\Delta_{\bar{Z}}\|_F^2 - \mathbb{E}\|\hat{Z} - Z^*\|_F^2 &= \frac{(2\,\Big\langle \Delta_{\bar{Z}}\bar{w}, \boldsymbol{\nabla} H(\bar{\eta})\bar{v}\Big\rangle)^2}{16\ \mathbb{E}\|\boldsymbol{\nabla} h(\bar{\eta})\bar{v}\|^2\|\bar{w}_{(-1)}\|^2} \\ &\ge \frac{\big\||A|\circ\xi\circ\tilde{T}_1\big\|_F^2\Big(\big\||A|\circ\xi\circ\tilde{T}_1\big\|_F - \big\||A|\circ\xi\circ\tilde{T}_2\big\|_F - \big\||A|\circ\xi\circ\tilde{T}_3)\big\|_F\Big)^2}{16\ \mathbb{E}\|\boldsymbol{\nabla} h(\bar{\eta})\bar{v}\|^2\|\bar{w}_{(-1)}\|^2}.\end{aligned}$$

Further, we have

$$\begin{aligned}\mathbb{E}\|\boldsymbol{\nabla} h(\bar{\eta})\bar{v}\|^2 &\le \mathbb{E}\|\boldsymbol{\nabla} h(\bar{\eta})\|_{op}^2\|\bar{Z}\|_{op}^2\|\bar{w}\|^2 \\ &\le \|\boldsymbol{\nabla} H(\bar{\eta})\|_{op}^2\|\bar{Z}\|_{op}^2\|\bar{w}\|^2 + \mathbb{E}\big\|B - |A|\circ\sigma(\eta^*)\big\|_{op}^2\|\bar{Z}\|_{op}^2\|\bar{w}\|^2 \\ &= \big\||A|\circ\xi\circ\xi\circ(\tilde{T}_1 + \tilde{T}_2 - \tilde{T}_3)\big\|_{op}^2\|\bar{Z}\|_{op}^2\|\bar{w}\|^2 \\ &\quad + \mathbb{E}\big\|B - |A|\circ\sigma(\eta^*)\big\|_{op}^2\|\bar{Z}\|_{op}^2\|\bar{w}\|^2,\end{aligned}$$

and $\|\bar{w}_{(-1)}\| \le \|\bar{w}\|$. Then it follows that

$$\begin{aligned}&\mathbb{E}\|\Delta_{\bar{Z}}\|_F^2 - \mathbb{E}\|\Delta_{\hat{Z}}\|_F^2 \\ &\quad\ge \frac{\big\||A|\circ\xi\circ\tilde{T}_1\big\|_F^2\Big(\big\||A|\circ\xi\circ\tilde{T}_1\big\|_F - \big\||A|\circ\xi\circ\tilde{T}_2\big\|_F - \big\||A|\circ\xi\circ\tilde{T}_3\big\|_F\Big)^2}{16\|\bar{Z}\|_{op}^2\|\bar{w}\|^4\Big(\big\||A|\circ\xi\circ\xi\circ(\tilde{T}_1 + \tilde{T}_2 - \tilde{T}_3)\big\|_{op}^2 + \mathbb{E}\big\|B - |A|\circ\sigma(\eta^*)\big\|_{op}^2\Big)}.\end{aligned}$$

For better implication of the above improvement, we define $T_i = \tilde{T}_i/\|\bar{Z}\|_{op}\|\bar{w}\|^2$ for

$i = 1, 2, 3$, and based on the arguments in (B.38)-(B.40), they are bounded by

$$\big\|T_1\big\|_F = \frac{\big\|\tilde{T}_1\big\|_F}{\|\bar{Z}\|_{op}\|\bar{w}\|^2} \le 2\frac{\big\|\bar{Z}\bar{w}\bar{w}^\top\Delta_{\bar{Z}}^\top\big\|_F}{\|\bar{Z}\|_{op}\|\bar{w}\|^2} \le 2\big\|\Delta_{\bar{Z}}\big\|_F = \mathcal{O}(1), \tag{B.41}$$

$$\big\|T_2\big\|_F = \frac{\big\|\tilde{T}_2\big\|_F}{\|\bar{Z}\|_{op}\|\bar{w}\|^2} \le \frac{\big\|Z^*\big\|_{op}^2\|\Delta_{\bar{w}}\|(\|\bar{w}\| + \|w^*\|)}{\|\bar{Z}\|_{op}\|\bar{w}\|^2} = \mathcal{O}(1)\frac{1}{\big\|\bar{w}\big\|}, \tag{B.42}$$

$$\big\|T_3\big\|_F = \frac{\big\|\tilde{T}_3\big\|_F}{\|\bar{Z}\|_{op}\|\bar{w}\|^2} \le \frac{\big\|\Delta_{\bar{Z}}\big\|_F^2\big\|\bar{w}\big\|^2}{\|\bar{Z}\|_{op}\|\bar{w}\|^2} = \mathcal{O}(1/\sqrt{n}), \tag{B.43}$$

respectively. Then the improvement is at least

$$\begin{aligned}&\mathbb{E}\|\Delta_{\bar{Z}}\|_F^2 - \mathbb{E}\|\Delta_{\hat{Z}}\|_F^2\\ &\quad\ge \frac{\big\||A| \circ \xi \circ T_1\big\|_F^2\left(\big\||A| \circ \xi \circ T_1\big\|_F - \big\||A| \circ \xi \circ T_2\big\|_F - \big\||A| \circ \xi \circ T_3\big\|_F\right)^2}{16\left(\big\||A| \circ \xi \circ \xi \circ (T_1 + T_2 - T_3)\big\|_{op}^2 + \mathbb{E}\big\|B - |A| \circ \sigma(\eta^*)\big\|_{op}^2/\|\bar{Z}\|_{op}^2\|\bar{w}\|^4\right)},\end{aligned}$$

which completes the proof. In particular, when $\|w^*\| \asymp \|\bar{w}\|$ increases, the upper bound of numerator in the improvement increases and that of the denominator decreases, therefore the improvement more likely increases This implies that larger signal in the edge signs would lead to greater improvement in estimating Z. □

B.7.2 Discussion on the Assumptions in Proposition 3.5.2

We note that the prerequisite error rate of $(\bar{w}, \bar{\gamma})$ in Proposition 3.5.2 can be achieved through first randomly sampling a subset of observed edges, then running Algorithm B.2 to obtain the separate estimate $\tilde{v}$, and finally regressing $\tilde{v}$ on $\bar{Z}$ to obtain $(\bar{w}, \bar{\gamma})$. The conditional independence assumption also holds in sequence if we use the remaining observed edges for the one-step update. The following proposition theoretically justifies the above procedure.

Proposition B.7.1. *Suppose the conditions in Proposition 3.5.1 and Theorem 3.5.1 hold, and the singular values of the sample covariance $Z^{*\top}Z^*/n$ and $\|v^*\|^2/n$ are of constant order. Given the estimators $\bar{Z}$ and $\tilde{v}$ obtained from Algorithms B.1 and B.2 respectively, if we regress $\tilde{v}$ on $\bar{Z}$ to obtain $(\bar{w}, \bar{\gamma})$, then we have $\|\bar{w} - w^*\|^2 + \|\bar{\gamma} - \gamma^*\|^2 = \mathcal{O}(1/n)$.*

Proof of Proposition B.7.1. Without loss of generality, we assume $\bar{O} = \arg\min_{O \in O(k)} \|\bar{Z} - Z^* O\|_F = I_k$ and $\tilde{k} = \arg\min_{k \in \{1,-1\}} \|\tilde{v} - kv^*\| = 1$ in the proof, otherwise we replace the outputs of Algorithms B.1 and B.2 by $\bar{Z}\bar{O}^\top$ and $\tilde{k}\tilde{v}$ respectively. Therefore, $\Delta_{\bar{Z}} = \bar{Z} - Z^*$ and $\Delta_{\tilde{v}} = \tilde{v} - v^*$. By Lemmas B.7.1 and B.7.2, we have $\|\bar{Z} - Z^*\| = \mathcal{O}(1)$ and $\|\tilde{v} - v^*\| = \mathcal{O}(1)$. We further substitute $[1_n, \bar{Z}]$, $[1_n, Z^*]$, $[\bar{\gamma}, \bar{w}^\top]^\top$, and $[\gamma^*, {w^*}^\top]^\top$ by $\bar{Z}$, Z^*, $\bar{w}$, and w^* respectively, then the value of $\|\Delta_{\bar{Z}}\|$ does not change and the singular values of ${Z^*}^\top Z^*/n$ are still of constant order. By the definition of $\bar{w}$, we have

$$0 = \bar{Z}^\top(\tilde{v} - \bar{Z}\bar{w}) = \bar{Z}^\top(\tilde{v} - Z^* w^* + Z^* w^* - \bar{Z} w^* + \bar{Z} w^* - \bar{Z}\bar{w}). \tag{B.44}$$

Since

$$\begin{aligned}\|\bar{Z}^\top(\tilde{v} - Z^* w^*)\| &= \|\bar{Z}^\top(\tilde{v} - v^*)\| \le \|\bar{Z}\|_{op}\|\tilde{v} - v^*\| \le (\|Z^*\|_{op} + \|\Delta_{\bar{Z}}\|_{op})\|\tilde{v} - v^*\| \\ &\le (\|Z^*\|_{op} + \|\Delta_{\bar{Z}}\|_F)\|\tilde{v} - v^*\| \le (\mathcal{O}(\sqrt{n}) + \mathcal{O}(1))\mathcal{O}(1) = \mathcal{O}(\sqrt{n}),\end{aligned}$$

$$\|\bar{Z}^\top(Z^* - \bar{Z})w^*\| \le \|\bar{Z}\|_{op}\|\Delta_{\bar{Z}}\|_F\|w^*\| = \mathcal{O}(\sqrt{n}) \cdot \mathcal{O}(1) \cdot \mathcal{O}(1) = \mathcal{O}(\sqrt{n}),$$

and, by Lemma B.5.7,

$$\begin{aligned}\|(\bar{Z}^\top\bar{Z} - {Z^*}^\top Z^*)(w^* - \bar{w})\| &\le \|\bar{Z}^\top\bar{Z} - {Z^*}^\top Z^*\|_F\|w^* - \bar{w}\| \\ &\le 3\|Z^*\|_{op}\|\Delta_{\bar{Z}}\|_F\|w^* - \bar{w}\| = \mathcal{O}(\sqrt{n}),\end{aligned}$$

we obtain from (B.44) that $\|{Z^*}^\top Z^*(w^* - \bar{w})\| = \mathcal{O}(\sqrt{n})$. Note that $\|{Z^*}^\top Z^*(w^* - \bar{w})\| \ge \boldsymbol{\sigma}_k^2(Z^*)\|w^* - \bar{w}\|$, where $\boldsymbol{\sigma}_k(Z^*)$ is the smallest singular value of Z^*, and by assumption there exist positive constants $C_1 < C_2$ such that $C_1 n \le \boldsymbol{\sigma}_k^2(Z^*) \le \boldsymbol{\sigma}_1^2(Z^*) \le C_2 n$. Therefore, we have

$$\|w^* - \bar{w}\| \le \mathcal{O}(\sqrt{n})/\boldsymbol{\sigma}_k^2(Z^*) = \mathcal{O}(1/\sqrt{n}),$$

which completes the proof. □

BIBLIOGRAPHY

BIBLIOGRAPHY

Aalen, O. (1980), A model for nonparametric regression analysis of counting processes, in *Mathematical Statistics and Probability Theory*, pp. 1–25, Springer, New York, NY, doi:10.1007/978-1-4615-7397-5_1.

Abbe, E. (2018), Community detection and stochastic block models: Recent developments, *Journal of Machine Learning Research*, *18*(177), 1–86.

Ai, C., and X. Chen (2003), Efficient estimation of models with conditional moment restrictions containing unknown functions, *Econometrica*, *71*(6), 1795–1843, doi:10.1111/1468-0262.00470.

Alexe, M., and A. Sandu (2009), Forward and adjoint sensitivity analysis with continuous explicit Runge–Kutta schemes, *Applied Mathematics and Computation*, *208*(2), 328–346, doi:https://doi.org/10.1016/j.amc.2008.11.035.

Andersen, P. K., and R. D. Gill (1982), Cox's regression model for counting processes: A large sample study, *The Annals of Statistics*, *10*(4), 1100–1120, doi:10.1214/aos/1176345976.

Antil, H., and D. Leykekhman (2018), A brief introduction to PDE-constrained optimization, in *Frontiers in PDE-Constrained Optimization*, pp. 3–40, Springer New York, New York, NY, doi:10.1007/978-1-4939-8636-1_1.

Antolini, L., P. Boracchi, and E. Biganzoli (2005), A time-dependent discrimination index for survival data, *Statistics in Medicine*, *24*(24), 3927–3944, doi:10.1002/sim.2427.

Bagdonavicius, V., and M. Nikulin (2001), *Accelerated Life Models: Modeling and Statistical Analysis*, Chapman and Hall/CRC, New York, NY, doi:10.1201/9781420035872.

Bagdonavicius, V. B., and M. S. Nikulin (1999), Generalized proportional hazards model based on modified partial likelihood, *Lifetime Data Analysis*, *5*(4), 329–350.

Ballinger, B., et al. (2018), DeepHeart: semi-supervised sequence learning for cardiovascular risk prediction, in *Thirty-Second AAAI Conference on Artificial Intelligence.*

Bennett, C. C., T. W. Doub, and R. Selove (2012), Ehrs connect research and practice: Where predictive modeling, artificial intelligence, and clinical decision support intersect, *Health Policy and Technology*, *1*(2), 105–114.

Bennett, S. (1983), Analysis of survival data by the proportional odds model, *Statistics in Medicine*, *2*(2), 273–277, doi:10.1002/sim.4780020223.

Billingsley, P. (2008), *Convergence of Probability Measures*, John Wiley & Sons, Ltd., doi:10.1002/9780470316962.

Bradbury, J., et al. (2018), JAX: composable transformations of Python+NumPy programs.

Buckley, J., and I. James (1979), Linear regression with censored data, *Biometrika*, *66*(3), 429–436.

Cai, T., L. Tian, and L. J. Wei (2005), Semiparametric Box-Cox power transformation models for censored survival observations, *Biometrika*, *92*(3), 619–632, doi:10.1093/biomet/92.3.619.

Candes, E. J., Y. C. Eldar, T. Strohmer, and V. Voroninski (2013), Phase retrieval via matrix completion, *SIAM Journal on Imaging Sciences*, *6*(1), 199–225, doi:10.1137/110848074.

Cao, Y., S. Li, L. Petzold, and R. Serban (2003), Adjoint sensitivity analysis for differential-algebraic equations: The adjoint DAE system and its numerical solution, *SIAM Journal on Scientific Computing*, *24*(3), 1076–1089, doi:10.1137/S1064827501380630.

Chapfuwa, P., C. Tao, C. Li, C. Page, B. Goldstein, L. Carin, and R. Henao (2018), Adversarial time-to-event modeling, in *Proceedings of the 35th International Conference on Machine Learning*, vol. 80, pp. 735–744.

Chatterjee, S. (2015), Matrix estimation by universal singular value thresholding, *Annals of Statistics*, *43*(1), 177–214, doi:10.1214/14-AOS1272.

Che, Z., Y. Cheng, S. Zhai, Z. Sun, and Y. Liu (2017), Boosting deep learning risk prediction with generative adversarial networks for electronic health records, in *2017 IEEE International Conference on Data Mining (ICDM)*, pp. 787–792, IEEE, doi:10.1109/ICDM.2017.93.

Chen, K., and X. Tong (2010), Varying coefficient transformation models with censored data, *Biometrika*, *97*(4), 969–976.

Chen, K., Z. Jin, and Z. Ying (2002), Semiparametric analysis of transformation models with censored data, *Biometrika*, *89*(3), 659–668, doi:10.1093/biomet/89.3.659.

Chen, R. T. Q., Y. Rubanova, J. Bettencourt, and D. K. Duvenaud (2018), Neural ordinary differential equations, in *Advances in Neural Information Processing Systems 31*, pp. 6571–6583.

Chen, X. (2007), Large sample sieve estimation of semi-nonparametric models, in *Handbook of Econometrics*, vol. 6B, 1 ed., chap. 76, Elsevier.

Chen, X., O. Linton, and I. Van Keilegom (2003), Estimation of semiparametric models when the criterion function is not smooth, *Econometrica*, *71*(5), 1591–1608, doi:10.1111/1468-0262.00461.

Chen, Y., and M. J. Wainwright (2015), Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees, *arXiv preprint arXiv:1509.03025*.

Chen, Y., H. Zhang, and P. Zhu (2009), Study of customer lifetime value model based on survival-analysis methods, in *2009 WRI World Congress on Computer Science and Information Engineering*, pp. 266–270, doi:10.1109/CSIE.2009.313.

Cheng, S. C., L. J. Wei, and Z. Ying (1995), Analysis of transformation models with censored data, *Biometrika*, *82*(4), 835–845, doi:10.1093/biomet/82.4.835.

Chiang, K.-Y., C.-J. Hsieh, N. Natarajan, I. S. Dhillon, and A. Tewari (2014), Prediction and clustering in signed networks: A local to global perspective, *Journal of Machine Learning Research*, *15*(34), 1177–1213.

Ching, T., X. Zhu, and L. X. Garmire (2018), Cox-nnet: An artificial neural network method for prognosis prediction of high-throughput omics data, *PLOS Computational Biology*, *14*(4), e1006,076, doi:10.1371/journal.pcbi.1006076.

Cho, K., B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio (2014), Learning phrase representations using RNN encoder–decoder for statistical machine translation, in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, doi:10.3115/v1/D14-1179.

Choi, E., A. Schuetz, W. F. Stewart, and J. Sun (2016), Using recurrent neural network models for early detection of heart failure onset, *Journal of the American Medical Informatics Association*, *24*(2), 361–370, doi:10.1093/jamia/ocw112.

Chung, F., and L. Lu (2006), *Complex Graphs and Networks (CBMS Regional Conference Series in Mathematics)*, American Mathematical Society, USA.

Chung, J., K. Kastner, L. Dinh, K. Goel, A. C. Courville, and Y. Bengio (2015), A recurrent latent variable model for sequential data, in *Advances in Neural Information Processing Systems*, pp. 2980–2988.

Cox, D. R. (1972), Regression models and life-tables, *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *34*(2), 187–220.

Cox, D. R. (1975), Partial likelihood, *Biometrika*, *62*(2), 269–276.

Dai, A. M., and Q. V. Le (2015), Semi-supervised sequence learning, in *Advances in Neural Information Processing Systems*, pp. 3079–3087.

Davenport, M. A., Y. Plan, E. Van Den Berg, and M. Wootters (2014), 1-bit matrix completion, *Information and Inference: A Journal of the IMA*, *3*(3), 189–223.

Dekker, F., R. Mutsert, P. Dijk, C. Zoccali, and K. Jager (2008), Survival analysis: time-dependent effects and time-varying risk factors, *Kidney International*, *74*, 994–997, doi:10.1038/ki.2008.328.

Derr, T., C. Aggarwal, and J. Tang (2018), Signed network modeling based on structural balance theory, in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pp. 557–566.

Dickinson, R. P., and R. J. Gelinas (1976), Sensitivity analysis of ordinary differential equation systems—A direct method, *Journal of Computational Physics*, *21*(2), 123–143, doi:10.1016/0021-9991(76)90007-3.

Ding, Y., and B. Nan (2011), A sieve M-theorem for bundled parameters in semiparametric models, with application to the efficient estimation in a linear model for censored data, *The Annals of Statistics*, *39*(6), 3032–3061, doi:10.1214/11-AOS934.

Dugas, C., Y. Bengio, F. Bélisle, C. Nadeau, and R. Garcia (2001), Incorporating second-order functional knowledge for better option pricing, in *Advances in Neural Information Processing Systems 13*, pp. 472–478.

Dupont, E., A. Doucet, and Y. W. Teh (2019), Augmented neural ODEs, in *Advances in Neural Information Processing Systems 32*, pp. 3140–3150.

Esteban, C., O. Staeck, S. Baier, Y. Yang, and V. Tresp (2016), Predicting clinical events by combining static and dynamic information using recurrent neural networks, in *2016 IEEE International Conference on Healthcare Informatics (ICHI)*, pp. 93–101, IEEE, doi:10.1109/ICHI.2016.16.

Faraggi, D., and R. Simon (1995), A neural network model for survival data, *Statistics in Medicine*, *14*(1), 73–82, doi:10.1002/sim.4780140108.

Feng, D., R. Altmeyer, D. Stafford, N. A. Christakis, and H. H. Zhou (2020), Testing for balance in social networks, *Journal of the American Statistical Association*, *0*(0), 1–19, doi:10.1080/01621459.2020.1764850.

Fine, J. P., and R. J. Gray (1999), A proportional hazards model for the subdistribution of a competing risk, *Journal of the American Statistical Association*, *94*(446), 496–509.

Fine, J. P., Z. Ying, and L. J. Wei (1998), On the linear transformation model for censored data, *Biometrika*, *85*(4), 980–986, doi:10.1093/biomet/85.4.980.

Gao, C., Z. Ma, A. Y. Zhang, and H. H. Zhou (2017), Achieving optimal misclassification proportion in stochastic block models, *Journal of Machine Learning Research*, *18*(1), 1980–2024.

Gensheimer, M. F., and B. Narasimhan (2019), A simple discrete-time survival model for neural networks, *PeerJ*, *7*, e6257, doi:10.7717/peerj.6257.

Gerdts, M. (2011), *Optimal Control of ODEs and DAEs*, De Gruyter, Berlin, Boston, doi:https://doi.org/10.1515/9783110249996.

Goldberger, A. L., et al. (2000), Physiobank, Physiotoolkit, and Physionet: components of a new research resource for complex physiologic signals, *Circulation*, *101*(23), e215–e220, doi:10.1161/01.cir.101.23.e215.

Goldenberg, A., A. X. Zheng, S. E. Fienberg, and E. M. Airoldi (2010), A survey of statistical network models, *Foundations and TrendsÂ® in Machine Learning*, *2*(2), 129–233, doi:10.1561/2200000005.

Goldstein, B. A., A. M. Navar, M. J. Pencina, and J. Ioannidis (2017), Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review, *Journal of the American Medical Informatics Association*, *24*(1), 198–208.

Goodfellow, I., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio (2014), Generative adversarial nets, in *Advances in Neural Information Processing Systems*, pp. 2672–2680.

Graf, E., C. Schmoor, W. Sauerbrei, and M. Schumacher (1999), Assessment and comparison of prognostic classification schemes for survival data, *Statistics in Medicine*, *18*(17-18), 2529–2545, doi:10.1002/(SICI)1097-0258(19990915/30)18:17/18⟨2529::AID-SIM274⟩3.0.CO;2-5.

Grathwohl, W., R. T. Q. Chen, J. Bettencourt, and D. Duvenaud (2019), Scalable reversible generative models with free-form continuous dynamics, in *International Conference on Learning Representations*.

Gray, R. J. (1994), Spline-based tests in survival analysis, *Biometrics*, *50*(3), 640, doi:10.2307/2532779.

Groha, S., S. M. Schmon, and A. Gusev (2020), Neural ODEs for multi-state survival analysis, *arXiv preprint arXiv:2006.04893*.

Guha, R., R. Kumar, P. Raghavan, and A. Tomkins (2004), Propagation of trust and distrust, in *Proceedings of the 13th International Conference on World Wide Web*, pp. 403–412.

Gulshan, V., et al. (2016), Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs, *Journal of the American Medical Association (JAMA)*, *316*(22), 2402–2410, doi:10.1001/jama.2016.17216.

Harary, F., et al. (1953), On the notion of balance of a signed graph., *The Michigan Mathematical Journal*, *2*(2), 143–146.

Harrell Jr., F. E., K. L. Lee, R. M. Califf, D. B. Pryor, and R. A. Rosati (1984), Regression modeling strategies for improved prognostic prediction, *Statistics in Medicine*, *3*(2), 143–152, doi:10.1002/sim.4780030207.

Harutyunyan, H., H. Khachatrian, D. C. Kale, G. Ver Steeg, and A. Galstyan (2019), Multitask learning and benchmarking with clinical time series data, *Scientific Data*, *6*(1), 1–18, doi:10.1038/s41597-019-0103-9.

He, K., Y. Yang, Y. Li, J. Zhu, and Y. Li (2017), Modeling time-varying effects with large-scale survival data: an efficient quasi-newton approach, *Journal of Computational and Graphical Statistics*, *26*(3), 635–645.

He, X., H. Xue, and N. Shi (2010), Sieve maximum likelihood estimation for doubly semiparametric zero-inflated poisson models, *Journal of Multivariate Analysis*, *101*(9), 2026–2038, doi:10.1016/j.jmva.2010.05.003.

Heider, F. (1946), Attitudes and cognitive organization, *The Journal of Psychology*, *21*(1), 107–112.

Hoff, P. D., A. E. Raftery, and M. S. Handcock (2002), Latent space approaches to social network analysis, *Journal of the American Statistical Association*, *97*(460), 1090–1098, doi:10.1198/016214502388618906.

Hong, C., et al. (2021), Clinical knowledge extraction via sparse embedding regression (keser) with multi-center large scale electronic health record data, *NPJ Digital Medicine*, *4*(1), 1–11.

Horowitz, J. L. (1996), Semiparametric estimation of a regression model with an unknown transformation of the dependent variable, *Econometrica*, *64*(1), 103–137, doi:10.2307/2171926.

Hsieh, C.-J., K.-Y. Chiang, and I. S. Dhillon (2012), Low rank modeling of signed networks, in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 507–515.

Huang, J. (1999), Efficient estimation of the partly linear additive Cox model, *The Annals of Statistics*, *27*(5), 1536–1563.

Ibrahim, R., and W. Whitt (2009), Real-time delay estimation in overloaded multiserver queues with abandonments, *Management Science*, *55*(10), 1729–1742, doi:10.1287/mnsc.1090.1041.

Izmirlioglu, A. (2017), The correlates of war dataset, *Journal of World-Historical Information*, *3*(1).

Jin, Z., D. Y. Lin, L. J. Wei, and Z. Ying (2003), Rank-based inference for the accelerated failure time model, *Biometrika*, *90*(2), 341–353, doi:10.1093/biomet/90.2.341.

Jin, Z., D. Y. Lin, and Z. Ying (2006), On least-squares regression with censored data, *Biometrika*, *93*(1), 147–161.

Johnson, A. E., et al. (2016), MIMIC-III, a freely accessible critical care database, *Scientific Data*, *3*(1), 1–9, doi:10.1038/sdata.2016.35.

Kaji, D. A., J. R. Zech, J. S. Kim, S. K. Cho, N. S. Dangayach, A. B. Costa, and E. K. Oermann (2019), An attention based deep learning model of clinical events in the intensive care unit, *PLOS One*, *14*(2), e0211,057, doi:10.1371/journal.pone.0211057.

Kalbfleisch, J. D., and R. L. Prentice (2011), *The statistical analysis of failure time data*, vol. 360, John Wiley & Sons.

Kaplan, E. L., and P. Meier (1958), Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association*, *53*(282), 457–481, doi:10.2307/2281868.

Katzman, J. L., U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger (2018), DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network, *BMC Medical Research Methodology*, *18*(1), 24, doi:10.1186/s12874-018-0482-1.

Kawaguchi, E. S., J. I. Shen, M. A. Suchard, and G. Li (2020), Scalable algorithms for large competing risks data, *Journal of Computational and Graphical Statistics*, *0*(0), 1–9, doi:10.1080/10618600.2020.1841650.

Khan, S., and E. Tamer (2007), Partial rank estimation of duration models with general forms of censoring, *Journal of Econometrics*, *136*(1), 251–280, doi:https://doi.org/10.1016/j.jeconom.2006.03.003.

Kingma, D. P., and M. Welling (2014), Auto-encoding variational bayes, in *2nd International Conference on Learning Representations (ICLR)*.

Kingma, D. P., S. Mohamed, D. J. Rezende, and M. Welling (2014), Semi-supervised learning with deep generative models, in *Advances in Neural Information Processing Systems*, pp. 3581–3589.

Kirkley, A., G. T. Cantwell, and M. E. J. Newman (2019), Balance in signed networks, *Physical Review E*, *99*, 012,320, doi:10.1103/PhysRevE.99.012320.

Klein, J. P., and M. L. Moeschberger (2003), *Basic Quantities and Models*, pp. 21–61, Springer New York, New York, NY, doi:10.1007/0-387-21645-6_2.

Knoke, D. (2013), Understanding social networks: Theories, concepts, and findings, *Contemporary Sociology*, *42*(2), 249–251, doi:10.1177/0094306113477381y.

Koltchinskii, V., K. Lounici, and A. B. Tsybakov (2011), Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion, *Annals of Statistics*, *39*(5), 2302–2329, doi:10.1214/11-AOS894.

Kovesdy, C., M. Czira, A. Rudas, A. Ujszaszi, L. Rosivall, M. Novak, K. Kalantar-Zadeh, M. Molnar, and I. Mucsi (2010), Survival analysis: time-dependent effects and time-varying risk factors, *American Journal of Transplantation*, *10*(12), 2644–2651.

Krivitsky, P. N., M. S. Handcock, A. E. Raftery, and P. D. Hoff (2009), Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models, *Social Networks*, *31*(3), 204 – 213, doi: https://doi.org/10.1016/j.socnet.2009.04.001.

Kvamme, H., Ørnulf Borgan, and I. Scheel (2019), Time-to-event prediction with neural networks and Cox regression, *Journal of Machine Learning Research*, *20*(129), 1–30.

Lafranca, J., J. IJermans, M. Betjes, and J. Frank (2015), Body mass index and outcome in renal transplant recipients: a systematic review and meta-analysis, *BMC Medicine*, *13*(111).

Lai, T. L., and Z. Ying (1991), Large sample theory of a modified Buckley-James estimator for regression analysis with censored data, *The Annals of Statistics*, *19*(3), 1370–1402.

Lee, C., W. R. Zame, J. Yoon, and M. van der Schaar (2018), DeepHit: A deep learning approach to survival analysis with competing risks, in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18)*, pp. 2314–2321.

Lei, J., and A. Rinaldo (2015), Consistency of spectral clustering in stochastic block models, *Annals of Statistics*, *43*(1), 215–237, doi:10.1214/14-AOS1274.

Leskovec, J., D. Huttenlocher, and J. Kleinberg (2010), Signed networks in social media, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1361–1370.

Lin, D. Y. (2007), On the Breslow estimator, *Lifetime Data Analysis*, *13*(4), 471–480, doi:10.1007/s10985-007-9048-y.

Lin, D. Y., and Z. Ying (1995), Semiparametric analysis of general additive-multiplicative hazard models for counting processes, *The Annals of Statistics*, *23*(5), 1712–1734.

Lin, K. J., and S. Schneeweiss (2016), Considerations for the analysis of longitudinal electronic health records linked to claims data to study the effectiveness and safety of drugs, *Clinical Pharmacology & Therapeutics*, *100*(2), 147–159.

Lin, Y., and K. Chen (2012), Efficient estimation of the censored linear regression model, *Biometrika*, *100*(2), 525–530, doi:10.1093/biomet/ass073.

Ma, Z., Z. Ma, and H. Yuan (2020), Universal latent space model fitting for large networks with edge covariates., *Journal of Machine Learning Research*, *21*(4), 1–67.

Mckeague, I. W., and P. D. Sasieni (1994), A partly parametric additive risk model, *Biometrika*, *81*(3), 501–514, doi:10.1093/biomet/81.3.501.

Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013), Efficient estimation of word representations in vector space, *arXiv preprint arXiv:1301.3781*.

Miller Jr., R. G. (2011), *Survival Analysis*, John Wiley & Sons.

Modarres, M., M. P. Kaminskiy, and V. Krivtsov (2016), *Reliability Engineering and Risk Analysis: A Practical Guide*, 3rd ed., CRC press, Boca Raton, doi:10.1201/9781315382425.

Murphy, S. A., A. J. Rossini, and A. W. van der Vaart (1997), Maximum likelihood estimation in the proportional odds model, *Journal of the American Statistical Association*, *92*(439), 968–976, doi:10.1080/01621459.1997.10474051.

Narayanaswamy, S., T. B. Paige, J.-W. Van de Meent, A. Desmaison, N. Goodman, P. Kohli, F. Wood, and P. Torr (2017), Learning disentangled representations with semi-supervised deep generative models, in *Advances in Neural Information Processing Systems*, pp. 5925–5935.

Nesterov, Y. (2013), *Introductory lectures on convex optimization: A basic course*, vol. 87, Springer Science & Business Media.

Newman, M. (2010), *Networks: An Introduction*, Oxford University Press.

Ng'andu, N. H. (1997), An empirical comparison of statistical tests for assessing the proportional hazards assumption of Cox's model, *Statistics in Medicine*, *16*(6), 611–626, doi:10.1002/(SICI)1097-0258(19970330)16:6⟨611::AID-SIM437⟩3.0.CO;2-T.

Odena, A. (2016), Semi-supervised learning with generative adversarial networks, *arXiv preprint arXiv:1606.01583*.

Peto, R., and J. Peto (1972), Asymptotically efficient rank invariant test procedures, *Journal of the Royal Statistical Society. Series A (General)*, *135*(2), 185–207, doi:10.2307/2344317.

Petzold, L., S. Li, Y. Cao, and R. Serban (2006), Sensitivity analysis of differential-algebraic equations and partial differential equations, *Computers and Chemical Engineering*, *30*(10), 1553–1559, doi:10.1016/j.compchemeng.2006.05.015.

Plessix, R.-E. (2006), A review of the adjoint-state method for computing the gradient of a functional with geophysical applications, *Geophysical Journal International*, *167*(2), 495–503, doi:10.1111/j.1365-246X.2006.02978.x.

Pontryagin, L. S., E. Mishchenko, V. Boltyanskii, and R. Gamkrelidze (1962), *Mathematical Theory of Optimal Processes*, Routledge, London, doi:10.1201/9780203749319.

Purushotham, S., C. Meng, Z. Che, and Y. Liu (2018), Benchmarking deep learning models on large healthcare datasets, *Journal of Biomedical Informatics*, *83*, 112 – 134, doi:https://doi.org/10.1016/j.jbi.2018.04.007.

Qiu, Z., and Y. Zhou (2015), Partially linear transformation models with varying coefficients for multivariate failure time data, *Journal of Multivariate Analysis*, *142*, 144–166, doi:https://doi.org/10.1016/j.jmva.2015.08.008.

Reed, J., and T. Tezcan (2012), Hazard rate scaling of the abandonment distribution for the gi/m/n + gi queue in heavy traffic, *Operations Research*, *60*(4), 981–995, doi:10.1287/opre.1120.1069.

Royston, P., and M. K. B. Parmar (2002), Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects, *Statistics in Medicine*, *21*(15), 2175–2197, doi:https://doi.org/10.1002/sim.1203.

Saran, R., B. Robinson, K. C. Abbott, L. Y. Agodoa, J. Bragg-Gresham, R. Balkrishnan, N. Bhave, et al. (2017), US renal data system 2016 annual data report: Epidemiology of kidney disease in the United States, *American Journal of Kidney Diseases*, *65*(5), A7–A8.

Scholkopf, B., and A. J. Smola (2018), *Learning with kernels: support vector machines, regularization, optimization, and beyond*, Adaptive Computation and Machine Learning series.

Schumaker, L. (2007), *Spline Functions: Basic Theory*, Cambridge Mathematical Library, 3rd ed., Cambridge University Press, Cambridge, doi:10.1017/CBO9780511618994.

Shen, X. (1997), On methods of sieves and penalization, *The Annals of Statistics*, *25*(6), 2555–2591, doi:10.1214/aos/1030741085.

Shen, X. (1998), Propotional odds regression and sieve maximum likelihood estimation, *Biometrika*, *85*(1), 165–177, doi:10.1093/biomet/85.1.165.

Shen, X., and W. H. Wong (1994), Convergence rate of sieve estimates, *The Annals of Statistics*, *22*(2), 580–615.

Sirignano, J., and K. Spiliopoulos (2018), DGM: A deep learning algorithm for solving partial differential equations, *Journal of Computational Physics*, *375*, 1339–1364.

Socher, R., A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts (2013), Recursive deep models for semantic compositionality over a sentiment treebank, in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1631–1642.

Song, X., S. Ma, J. Huang, and X. Zhou (2006), A semiparametric approach for the nonparametric transformation survival model with multiple covariates, *Biostatistics*, *8*(2), 197–211, doi:10.1093/biostatistics/kxl001.

Steingrimsson, J. A., and S. Morrison (2020), Deep learning for survival outcomes, *Statistics in Medicine*, *39*(17), 2339–2349, doi:https://doi.org/10.1002/sim.8542.

Sun, Z., L. Han, W. Huang, X. Wang, X. Zeng, M. Wang, and H. Yan (2015), Recommender systems based on social networks, *Journal of Systems and Software*, *99*, 109–119.

Tang, W., and J. Zhu (2022+), Population-level balance in signed networks, *(Under review at the Journal of the American Statistical Association)*.

Tang, W., J. Ma, A. K. Waljee, and J. Zhu (2020), Semi-supervised joint learning for longitudinal clinical events classification using neural network models, *Stat*, *9*(1), e305, doi:10.1002/sta4.305, (Special issue on deep learning).

Tang, W., K. He, G. Xu, and J. Zhu (2022a), Survival analysis via ordinary differential equations, *Journal of the American Statistical Association*, just accepted.

Tang, W., J. Ma, Q. Mei, and J. Zhu (2022b), SODEN: A scalable continuous-time survival model through ordinary differential equation networks, *Journal of Machine Learning Research*, *23*(34), 1–29.

Tieleman, T., and G. Hinton (2012), Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude, *Coursera: Neural Networks for Machine Learning*, *4*(2), 26–31.

Tsiatis, A. A. (1990), Estimating regression parameters using linear rank tests for censored data, *The Annals of Statistics*, *18*(1), 354–372.

Tu, S., R. Boczar, M. Simchowitz, M. Soltanolkotabi, and B. Recht (2016), Low-rank solutions of linear matrix equations via procrustes flow, in *Proceedings of the 33rd International Conference on Machine Learning*, pp. 964–973.

Uno, H., T. Cai, M. J. Pencina, R. B. D'Agostino, and L.-J. Wei (2011), On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data, *Statistics in Medicine*, *30*(10), 1105–1117.

Van Der Vaart, A. W., and J. A. Wellner (1996), Weak convergence, in *Weak convergence and empirical processes*, pp. 16–28, Springer.

Vazquez, A., A. Flammini, A. Maritan, and A. Vespignani (2003), Global protein function prediction from protein-protein interaction networks, *Nature Biotechnology*, *21*(6), 697–700.

Walter, W. (1998), First order systems. Equations of higher order, in *Ordinary Differential Equations*, pp. 105–157, Springer New York, New York, NY, doi: 10.1007/978-1-4612-0601-9_4.

Wang, L., X. Zhang, and Q. Gu (2017), A unified computational and statistical framework for nonconvex low-rank matrix estimation, in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pp. 981–990, PMLR.

Wang, P., Y. Li, and C. K. Reddy (2019a), Machine learning for survival analysis: A survey, *Association for Computing Machinery (ACM) Computing Surveys*, *51*(6), doi:10.1145/3214306.

Wang, Y., C. Hong, N. Palmer, Q. Di, J. Schwartz, I. Kohane, and T. Cai (2019b), A fast divide-and-conquer sparse Cox regression, *Biostatistics*, doi: 10.1093/biostatistics/kxz036.

Wei, L. J. (1992), The accelerated failure time model: A useful alternative to the Cox regression model in survival analysis, *Statistics in Medicine*, *11*(14-15), 1871–1879, doi:10.1002/sim.4780111409.

Wellner, J. A., and Y. Zhang (2007), Two likelihood-based semiparametric estimation methods for panel count data with covariates, *The Annals of Statistics*, *35*(5), 2106–2142, doi:10.1214/009053607000000181.

Wolfe, R., V. Ashbyv, E. Milfordv, A. Ojov, R. Ettengerv, L. Agodoav, P. Heldv, and F. Portv (1999), Comparison of mortality in all patients on dialysis, patients on dialysis awaiting transplantation, and recipients of a first cadaveric transplant, *The New England Journal of Medicine*, *341*(23), 1725–1730.

Wu, Y., et al. (2016), Google's neural machine translation system: Bridging the gap between human and machine translation, *arXiv preprint arXiv:1609.08144*.

Yao, L., C. Mao, and Y. Luo (2019), Clinical text classification with rule-based features and knowledge-guided convolutional neural networks, *BMC Medical Informatics and Decision Making*, *19*(Suppl 3), 71, doi:10.1186/s12911-019-0781-4.

Zeng, D., and D. Y. Lin (2006), Efficient estimation of semiparametric transformation models for counting processes, *Biometrika*, *93*(3), 627–640, doi:10.1093/biomet/93.3.627.

Zeng, D., and D. Y. Lin (2007a), Efficient estimation for the accelerated failure time model, *Journal of the American Statistical Association*, *102*(480), 1387–1396.

Zeng, D., and D. Y. Lin (2007b), Maximum likelihood estimation in semiparametric regression models with censored data, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *69*(4), 507–564, doi:10.1111/j.1369-7412.2007.00606.x.

Zhang, Y., L. Hua, and J. Huang (2010), A spline-based semiparametric maximum likelihood estimation method for the Cox model with interval-censored data, *Scandinavian Journal of Statistics*, *37*(2), 338–354, doi:10.1111/j.1467-9469.2009.00680.x.

Zhao, L. (2021), Deep neural networks for predicting restricted mean survival times, *Bioinformatics*, *36*(24), 5672–5677, doi:10.1093/bioinformatics/btaa1082.

Zhao, X., Y. Wu, and G. Yin (2017), Sieve maximum likelihood estimation for a general class of accelerated hazards models with bundled parameters, *Bernoulli*, *23*(4B), 3385–3411, doi:10.3150/16-bej850.

Zheng, Q., and J. Lafferty (2016), Convergence analysis for rectangular matrix completion using burer-monteiro factorization and gradient descent, *arXiv preprint arXiv:1605.07051.*

Zucker, D. M., and A. F. Karr (1990), Nonparametric survival analysis with time-dependent covariate effects: A penalized partial likelihood approach, *The Annals of Statistics*, *18*(1), 329–353, doi:10.1214/aos/1176347503.

Zuo, L., H. Zhang, H. Wang, and L. Liu (2021), Sampling-based estimation for massive survival data with additive hazards model, *Statistics in Medicine*, *40*(2), 441–450, doi:https://doi.org/10.1002/sim.8783.