Journal of Machine Learning Research 27 (2026) 1-39 Submitted 12/24; Revised 8/26; Published 9/26

Deep Neural Network for Estimating General Risk Scores with Right Censored Data

Dongxiao Han HANDONGXIAO@NANKAI.EDU.CN
School of Statistics and Data Science
LPMC, KLMDASR and LEBPS, Nankai University
Tianjin, China

Zan Li * LIZAN2022@NANKAI.EDU.CN
School of Statistics and Data Science
LPMC, KLMDASR and LEBPS, Nankai University
Tianjin, China

Xinyuan Song XYSONG@CUHK.EDU.HK
Department of Statistics
The Chinese University of Hong Kong
Hong Kong, China

Donglin Zeng DZENG@UMICH.EDU
Department of Biostatistics
University of Michigan
Ann Arbor, Michigan, USA

Liuquan Sun SLQ@AMT.AC.CN
SKLMS, Academy of Mathematics and Systems Science
Chinese Academy of Sciences
and School of Mathematical Sciences
University of Chinese Academy of Sciences
Beijing, China

Editor: Jie Peng

Abstract

In this article, we propose a methodology for learning a risk score based on a general loss function, in which the score function is parameterized by a rectified linear unit feedforward deep neural network. The non-asymptotic error bound of the loss function is established and three examples are provided to illustrate the proposed method. The finite sample behavior of the proposed method is evaluated through simulation studies and an application to a SEER colon cancer data is provided.

Keywords: deep neural network, non-asymptotic error bound, proportional hazards model, risk score, survival analysis

1. Introduction

Survival data are frequently encountered in biomedical applications. Examples include heart failure, node-negative breast cancer and primary biliary cirrhosis. The hallmark of such

*. Zan Li is the corresponding author.

©2026 Dongxiao Han, Zan Li, Xinyuan Song, Donglin Zeng and Liuquan Sun.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v27/24-2109.html.

data is that the observation duration for a subject might conclude prior to the occurrence of the event of interest. In this circumstance, the survival time of the subject is categorized as censored. Quantifying subjects' prospects (learning risk scores) based on their over risk profile plays a pivotal role in identifying individuals who may require heightened surveillance and more aggressive therapeutic interventions.

Numerous methodologies of learning risk scores have been established in the context of non-censored outcomes, including projection pursuit regression (Friedman and Stuetzle, 1981), kernel smoothing (Wand and Jones, 1994), support vector machines (Hearst et al., 1998), neural network (Anthony and Bartlett, 1999) and decision trees (Rokach and Maimon, 2005). Substantial advancements have been made in statistical methods tailored for learning risk scores in the case subject to censoring, with significant contributions from various researchers (Cox, 1972; Lin and Ying, 1994; Ripley and Ripley, 2001; Chen et al., 2002; Ripley et al., 2004; Zeng and Lin, 2006, 2007; Shivaswamy et al., 2007; Khan and Zubek, 2008; Chen et al., 2010; Van Belle et al., 2010, 2011; Wang et al., 2016). For instance, works by Cox (1972), Lin and Ying (1994), Chen et al. (2002), Zeng and Lin (2006), Zeng and Lin (2007) and Chen et al. (2010) centered around estimating survival function or association parameters via the hazard function. Specifically, Cox (1972) introduced the semi-parametric proportional hazards (PH) model and estimated regression parameters through partial likelihood maximization. Chen et al. (2010) considered the non-parametric proportional hazards (NPH) model and proposed a global partial likelihood method for regression function estimation. These studies presupposed a constant hazard ratio between subjects with differing covariate values over time. Meanwhile, several models, such as the additive hazards model (Lin and Ying, 1994), linear transformation models (Chen et al., 2002) and general transformation models (Zeng and Lin, 2006, 2007), have been explored under the contrary assumption. On another note, Shivaswamy et al. (2007), Khan and Zubek (2008), Van Belle et al. (2010, 2011) and Wang et al. (2016) developed support vector-based approaches. In particular, Shivaswamy et al. (2007) and Khan and Zubek (2008) adopted asymmetric modifications to the ϵ-insensitive loss function to accommodate censoring. Van Belle et al. (2010, 2011) introduced a least-squares support vector machine integrating ranking and regression. Wang et al. (2016) proposed a non-parametric decision function with an additive structure for the counting process, providing a desirable risk score and a time-varying offset to consider varying at-risk populations over time. Their estimating method for the risk score is based on predicting the counting process associated with the time-to-event outcomes in at-risk subjects via a series of support vector machines. In addition, Ripley and Ripley (2001) and Ripley et al. (2004) discussed estimating model parameters using shallow neural networks in the context of different survival models. Furthermore, Bou-Hamad et al. (2011) conducted a comprehensive review of survival tree methodologies.

Deep neural network methods have recently been developed for learning risk scores under the censoring framework (Katzman et al., 2018; Lee et al., 2018; Gensheimer and Narasimhan, 2019; Kvamme et al., 2019). For example, Katzman et al. (2018) estimated the regression functions of the NPH model by minimizing the average negative log partial likelihood. Lee et al. (2018) learned the distribution of survival time by minimizing a loss function. This function comprises the log-likelihood of the joint distribution of the first hitting time and event time and a combination of cause-specific ranking loss functions.

Gensheimer and Narasimhan (2019) proposed a discrete-time survival model and estimated it using the maximum likelihood method. Kvamme et al. (2019) introduced a loss function that can scale well to large data sets for the NPH model.

In this article, we propose a general loss function for learning a risk score and estimate the score function using a rectified linear unit (ReLU) feedforward deep neural network (DNN). The framework applies to margin-based, rank-based, and proportional-hazards losses, and could extend more broadly to any user-defined loss satisfying the stated regularity conditions. The non-asymptotic error bound of the loss function is developed. Beyond providing theoretical justification for applying neural networks to censored survival data, the proposed excess-risk bound offers interpretable insights into the sources of error that affect empirical risk minimization in this setting. Unlike standard empirical risk analysis with fully observed outcomes, where the loss is simply averaged over independent samples, the presence of right censoring introduces an additional layer of complexity: for each uncensored observation, the loss involves an empirical average over its risk set. This risk-set-based structure gives rise to a distinct source of uncertainty beyond the usual estimation error. Our bound explicitly decomposes the excess risk into three components: a risk-set estimation error specific to censored data, a classical estimation error, and a neural-network approximation error. The first two components increase with network complexity, as they are controlled by quantities that grow with network width and depth. The third component decreases as the network becomes more expressive, but worsens as the covariate dimension increases. This decomposition reveals a fundamental trade-off between the estimation errors and the approximation error, and clarifies the distinct roles played by censoring, network complexity, and covariate dimension in the overall excess risk. These findings offer practical guidance for controlling network complexity and highlight the potential benefits of dimension reduction or additional structural assumptions in high-dimensional settings.

Among the five most related works to the present study are Chen et al. (2010), Wang et al. (2016), Katzman et al. (2018), Kvamme et al. (2019) and Zhong et al. (2022). However, our proposed method diverges significantly in several key aspects:

(i) Our work considers a general loss function, while the five works mentioned above are restricted to specific model settings.

(ii) While our work, along with Katzman et al. (2018) and Kvamme et al. (2019), considers the NPH model and employs a neural network for regression function estimation, the latter two lack theoretical justification. In comparison, our method provides a non-asymptotic error bound for the loss function, thereby enhancing its theoretical rigor.

(iii) Chen et al. (2010) focused on the NPH model with a unidimensional covariate, estimating the regression function by maximizing a global partial likelihood derived through the local polynomial approach. Extending their method to accommodate multiple covariates presents substantial challenges. Our approach, conversely, allows the number of covariates to be large and employs a DNN for risk score estimation. This renders our theoretical proof fundamentally different from that of Chen et al. (2010).

(iv) Zhong et al. (2022) studied DNN estimation in a partially linear PH model using a sparse ReLU network with uniformly bounded weights and biases. In contrast, our

method considers a bounded-output ReLU network without explicit parameter-wise boundedness or sparsity constraints and develops a non-asymptotic excess-risk bound for the NPH model.

(v) Wang et al. (2016) is rooted in a search space framed within a reproducing kernel Hilbert space, a notable divergence from our method, which operates within a DNN. This distinction in search spaces leads to markedly different methodologies for estimation and proof verification.

The rest of the article is organized as follows. Section 2 provides the general loss function and the estimation method. Section 3 presents theoretical results and three examples to illustrate the proposed method. Section 4 reports results from simulation studies conducted to evaluate the proposed methods. An application to a SEER colon cancer data set is provided in Section 5. Some concluding remarks are made in Section 6. All proofs and additional simulation results are given in the Appendix.

2. Estimation Method

Let T, C and $\boldsymbol{X}$ be the failure time, censoring time and a p-dimensional covariate vector, respectively. Also let $Y = \min(T, C)$ be the event time and $\Delta = I(T \leq C)$ be the censoring indicator, where $I(\cdot)$ is the indicator function. We assume that the support of $\boldsymbol{X}$ is a bounded subset of $\mathbb{R}^p$. Without loss of generality, let the support of $\boldsymbol{X}$ be $[-a_1, a_1]^p$, where a_1 is a positive constant. The observations $(\boldsymbol{X}_i, \Delta_i, Y_i)$ $(i = 1, \ldots, n)$ are assumed to be independent replicates of $(\boldsymbol{X}, \Delta, Y)$. We first consider the following NPH model as an example before introducing our general loss function for learning a risk score. The NPH model assumes that the conditional hazard of T given $\boldsymbol{X} = \boldsymbol{x}$ takes the form of

$$\lambda(t \mid \boldsymbol{x}) = \lambda_0(t) \exp\{g^*(\boldsymbol{x})\},$$

where $\lambda_0(t)$ is an unspecified baseline hazard function and $g^*(\boldsymbol{x})$ is an unknown function. According to Cox (1972), we can estimate g^* by maximizing the following partial likelihood,

$$\prod_{i=1}^{n} \left[\frac{\exp\{g(\boldsymbol{X}_i)\}}{\sum_{k=1}^{n} I(Y_k \geq Y_i) \exp\{g(\boldsymbol{X}_k)\}} \right]^{\Delta_i},$$

which is equivalent to minimizing the following loss function,

$$\frac{1}{n} \sum_{i=1}^{n} \Delta_i \log \left[\frac{\sum_{k=1}^{n} I(Y_k \geq Y_i) \exp\{g(\boldsymbol{X}_k) - g(\boldsymbol{X}_i)\}}{\sum_{k=1}^{n} I(Y_k \geq Y_i)} \right]. \tag{1}$$

Let

$$\frac{1}{n} \sum_{i=1}^{n} \Delta_i f \left[\frac{\sum_{k=1}^{n} I(Y_k \geq Y_i) h\{g(\boldsymbol{X}_k) - g(\boldsymbol{X}_i)\}}{\sum_{k=1}^{n} I(Y_k \geq Y_i)} \right], \tag{2}$$

be the general loss function, where f and h are two specified functions. When $\{f(t), h(t)\} = \{\log(t), \exp(t)\}$, (2) becomes the loss function for the NPH model. In the case of $\{f(t), h(t)\}=$

$\{t, [a+t]_+\}$, (2) turns to the loss function for estimating g^* in Wang et al. (2016) as discussed in Section 3, where $[a+t]_+ = \max\{a+t, 0\}$ and a could be any positive number. Define

$$PR(g) \equiv E\left\{\Delta f\left(\frac{E[I(\tilde{Y} \geq Y)h\{g(\tilde{\boldsymbol{X}}) - g(\boldsymbol{X})\} \mid \boldsymbol{X}, Y]}{E\{I(\tilde{Y} \geq Y) \mid Y\}}\right)\right\}, \tag{3}$$

and

$$g^* \equiv \text{argmin}_g PR(g),$$

where g^* is an unknown risk score function assumed to be continuous and $(\tilde{\boldsymbol{X}}, \tilde{Y})$ is a random vector independent and identically distributed with $(\boldsymbol{X}, Y)$. We assume $g^*(\boldsymbol{0}) = 0$ for identifiability. In the case of $\{f(t), h(t)\} = \{\log(t), \exp(t)\}$, minimizing (3) yields g^* in the NPH model. Our objective is to estimate g^*. Let $\check{g}$ be the minimizer of the empirical version (2) in the search space $\mathcal{G}$, i.e.,

$$\check{g} \equiv \text{argmin}_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} \Delta_i f\left[\frac{\sum_{k=1}^{n} I(Y_k \geq Y_i)h\{g(\boldsymbol{X}_k) - g(\boldsymbol{X}_i)\}}{\sum_{k=1}^{n} I(Y_k \geq Y_i)}\right].$$

Because $g^*(\boldsymbol{0})$ is assumed to be 0, $\check{g}$ may be a biased estimator of g^*. We consider

$$\widehat{g}(\boldsymbol{x}) = \check{g}(\boldsymbol{x}) - \sum_{i=1}^{n} \check{g}(\boldsymbol{X}_i)/n,$$

as our final estimator of $g^*(\boldsymbol{x})$ to tackle this issue. Since DNN is a powerful method for solving large-scale real-world problems such as natural language processing (Collobert et al., 2011; Socher et al., 2013), human action recognition (Ji et al., 2012; Le et al., 2011) or physics (Montavon et al., 2013), we parameterize g in (2) by a ReLU feedforward DNN $\mathcal{H}_{\mathcal{D},\mathcal{W},\mathcal{B}}$ (i.e., $\mathcal{G} = \mathcal{H}_{\mathcal{D},\mathcal{W},\mathcal{B}}$). Here $\mathcal{D}$ is the depth of the network, $\mathcal{W}$ is the maximum width of all hidden layers, $\mathcal{B} = \sup\{||g||_\infty \mid g \in \mathcal{H}_{\mathcal{D},\mathcal{W},\mathcal{B}}\}$ and $||\cdot||_\infty$ is the maximal absolute value of a function. The parameterization for g can be briefly described as follows,

$$\boldsymbol{\phi}_0 \equiv \boldsymbol{X} \xrightarrow{\boldsymbol{W}_1,\, b_1} \boldsymbol{\psi}_1 \xrightarrow{\sigma} \boldsymbol{\phi}_1 \cdots \xrightarrow{\boldsymbol{W}_{\mathcal{D}},\, b_{\mathcal{D}}} \boldsymbol{\psi}_{\mathcal{D}} \xrightarrow{\sigma} \boldsymbol{\phi}_{\mathcal{D}} \xrightarrow{\boldsymbol{W}_{\mathcal{D}+1},\, b_{\mathcal{D}+1}} \boldsymbol{\psi}_{\mathcal{D}+1} \equiv g(\boldsymbol{X}),$$

where $\boldsymbol{W}_j \in \mathbb{R}^{d_j \times d_{j-1}}$ is the weight matrix, d_j is the width of the jth layer, $\sigma(t) \equiv \max\{t, 0\}$ is the ReLU activation function and $b_j \in \mathbb{R}^{d_j}$ is the bias vector in the jth linear transformation, i.e.,

$$\boldsymbol{\psi}_j \equiv \boldsymbol{W}_j\boldsymbol{\phi}_{j-1} + b_j, \quad \text{for } j = 1, \ldots, \mathcal{D}+1,$$

and

$$\boldsymbol{\phi}_j \equiv \sigma(\boldsymbol{\psi}_j), \quad \text{for } j = 1, \ldots, \mathcal{D}.$$

By replacing $\mathcal{G}$ with $\mathcal{H}_{\mathcal{D},\mathcal{W},\mathcal{B}}$, we get the estimator $\check{g}$ via minimizing (2).

3. Theoretical Result and Examples

We first establish the non-asymptotic error bound for $PR(\widehat{g})$. Some notations are needed. Let $||\cdot||_2$ be the Euclidean norm, $\boldsymbol{Z}_i = (\boldsymbol{X}_i^\top, Y_i)^\top$ for $i = 1, \ldots, n$, $\mathbb{Z}$ be the set of integers, $g_0 \equiv \text{argmin}_{g \in \mathcal{H}_{\mathcal{D},\mathcal{W},\mathcal{B}}} PR(g)$ and

$$\mathcal{G}_1 = \left\{ f\left(\frac{E[I(Y \geq y)h\{g(\boldsymbol{X}) - g(\boldsymbol{x})\}]}{E\{I(Y \geq y)\}} \right) \;\middle|\; g \in \mathcal{G} \right\}.$$

For any measurable function Φ on $[-a_1, a_1]^p$ and Φ' on $[-a_1, a_1]^p \times [0, \tau]$, define $P_n\Phi = \sum_{i=1}^n \Phi(\boldsymbol{X}_i)/n$, and $P_n\Phi' = \sum_{i=1}^n \Phi'(\boldsymbol{Z}_i)/n$. For any $\boldsymbol{x}' \in [-a_1, a_1]^p$, any $\epsilon > 0$ and any probability measures μ_1 on $[-a_1, a_1]^p$ and μ_2 on $[-a_1, a_1]^p \times [0, \tau]$, let $\mathcal{G}_{\boldsymbol{x}'} = \{h\{g(\boldsymbol{x}) - g(\boldsymbol{x}')\} \mid g \in \mathcal{H}_{\mathcal{D},\mathcal{W},\mathcal{B}}\}$, and $N(\epsilon, \mathcal{G}_{x'}, L_1(\mu_1))$ and $N(\epsilon, \mathcal{G}_1, L_1(\mu_2))$ be the covering numbers of $\mathcal{G}_{x'}$ relative to the distance $L_1(\mu_1)$ and that of $\mathcal{G}_1$ relative to the distance $L_1(\mu_2)$ (Van Der Vaart and Wellner, 1996, page 83), respectively. The following conditions are needed to prove Theorem 1.

(C1) There exist two positive numbers τ and $a_2 \geq 4(\log n/n)^{1/2}$ such that

$$P(Y \leq \tau) = 1, \quad \text{and} \quad P(Y = \tau) \geq a_2.$$

(C2) There exists a positive number a_3 such that $||g^*||_\infty \leq a_3$, and for any $\boldsymbol{x}_1, \boldsymbol{x}_2 \in [-a_1, a_1]^p$,

$$|g^*(\boldsymbol{x}_1) - g^*(\boldsymbol{x}_2)| \leq a_3||\boldsymbol{x}_1 - \boldsymbol{x}_2||_2.$$

The function f is differentiable, and $||f||_\infty$, $||\dot{f}||_\infty$ and $||h||_\infty$ are all finite, where $\dot{f}$ is the derivative of f.

(C3) For any $0 < \epsilon < 1$, assume

$$\sup_{\boldsymbol{x}', \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in [-a_1, a_1]^p} N(\epsilon||h||_\infty, \mathcal{G}_{x'}, L_1(P_n)) \leq a_4 \Psi_1(\mathcal{W}, \mathcal{D})(16ea_5)^{\Psi_1(\mathcal{W},\mathcal{D})} \left(\frac{1}{\epsilon}\right)^{\Psi_1(\mathcal{W},\mathcal{D})-1},$$

$$\sup_{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in [-a_1, a_1]^p} N(\epsilon||f||_\infty, \mathcal{G}_1, L_1(P_n)) \leq a_4 \Psi_2(\mathcal{W}, \mathcal{D})(16ea_5)^{\Psi_2(\mathcal{W},\mathcal{D})} \left(\frac{1}{\epsilon}\right)^{\Psi_2(\mathcal{W},\mathcal{D})-1},$$

where e is Euler's number, a_4 and a_5 are two positive numbers, and Ψ_1 and Ψ_2 are increasing in $\mathcal{D}$ and $\mathcal{W}$ individually, with their forms depending on h and f.

(C4) $\mathcal{W} = 3^{p+3} \max\{p \lfloor N^{1/p} \rfloor, N + 2\}$, $\mathcal{D} = 11L + 18 + 2p$, and $\mathcal{B} \geq 3a_3$, where N and L could be any positive integers and $\lfloor t \rfloor = \max\{m : m \leq t, m \in \mathbb{Z}\}$.

(C5) There exists a positive constant a_6 such that

$$PR(g_0) - PR(g^*) \leq a_6||g_0 - g^*||_\infty.$$

The assumption $P(Y \leq \tau) = 1$, and $P(Y = \tau) \geq a_2$ in condition (C1) is used to bound $E\{I(Y \geq y)\}$. Moreover, both this assumption and the assumption $a_2 \geq 4(\log n/n)^{1/2}$ are used to bound the difference between $E\{I(Y \geq y)\}$ and its empirical version. Condition (C2) implies that g^* should be a Lipschitz continuous function, and that f and h are bounded

functions. It also implies that f is differentiable with a bounded derivative. We will show that condition (C3) is satisfied when f is Lipschitz continuous, and h is either Lipschitz continuous or an indicator function in the Appendix. This condition is for bounding the estimation error and capturing the uncertainty of estimating

$$f\bigg(\frac{E[I(Y \geq y)h\{g^*(\boldsymbol{X}) - g^*(\boldsymbol{x})\}]}{E\{I(Y \geq y)\}}\bigg),$$

by replacing the expectations within it with their empirical versions. Conditions (C2) and (C4) are for bounding $||g_0 - g^*||_\infty$ by utilizing Theorem 2.5 in Shen et al. (2022). Condition (C5) indicates that the approximation error can be controlled by $||g_0 - g^*||_\infty$ up to a constant. Conditions (C2), (C4) and (C5) ensure that the approximation error is controlled by a function of N and L. The following Theorem 1 provides the error bound for $PR(\widehat{g})$.

Theorem 1 *Assume that conditions (C1)-(C5) hold. Then, we have that with probability at least* $1 - 8/n$,

$$\begin{aligned} PR(\widehat{g}) - PR(g^*) \leq & 4||\dot{f}||_\infty\{2||h||_\infty(\log n/n)^{1/2} + l_1 f_1(\mathcal{D}, \mathcal{W}, n)\}/a_2^2 + 2l_2 f_2(\mathcal{D}, \mathcal{W}, n) \\ & + 262 a_1 a_3 a_6 \sqrt{p}\{N^2 L^2 \log_3(N+2)\}^{-1/p}, \end{aligned}$$

where

$$\begin{aligned} & l_1 = 2\sqrt{80||h||_\infty(1 + 4||h||_\infty)^2/3}\sqrt{(1 + ||h||_\infty)(9 + a_4 + a_5)}, \\ & l_2 = \sqrt{160||f||_\infty(1 + 8||f||_\infty)^2/3}\sqrt{(1 + ||f||_\infty)(8 + a_4 + a_5)}, \\ & f_1(\mathcal{D}, \mathcal{W}, n) = \sqrt{\psi_1(\mathcal{W}, \mathcal{D}) \log n/n}, \quad \textit{and} \quad f_2(\mathcal{D}, \mathcal{W}, n) = \sqrt{\psi_2(\mathcal{W}, \mathcal{D}) \log n/n}. \end{aligned}$$

The term $4||\dot{f}||_\infty\{2||h||_\infty(\log n/n)^{1/2} + l_1 f_1(\mathcal{D}, \mathcal{W}, n)\}/a_2^2$ captures the uncertainty of estimating

$$f_g^*(\boldsymbol{z}) = f\bigg(\frac{E[I(Y \geq y)h\{g^*(\boldsymbol{X}) - g^*(\boldsymbol{x})\}]}{E\{I(Y \geq y)\}}\bigg),$$

where $\boldsymbol{z} = (\boldsymbol{x}^\top, y)^\top$. This source of error is specific to the risk-set-based censored-data formulation because, for each uncensored observation i, the loss involves an empirical average over the corresponding risk set $\{k : Y_k \geq Y_i\}$. This differs from standard empirical-risk analysis with fully observed outcomes, where the empirical risk is obtained by directly averaging the per-observation loss terms. Because $f_1(\mathcal{D}, \mathcal{W}, n)$ depends on $\Psi_1(\mathcal{W}, \mathcal{D})$, which is increasing in the network width and depth, this error component generally increases as the network class becomes more complex. The second term, $2l_2 f_2(\mathcal{D}, \mathcal{W}, n)$, and $262 a_1 a_3 a_6 \sqrt{p}\{N^2 L^2 \log_3(N+2)\}^{-1/p}$ bounds the estimation error. Its dependence on $\Psi_2(\mathcal{W}, \mathcal{D})$ similarly shows that increasing network complexity may enlarge the estimation error. In contrast, the third term $262 a_1 a_3 a_6 \sqrt{p}\{N^2 L^2 \log_3(N+2)\}^{-1/p}$ bounds the neural-network approximation error and decreases as the network becomes more expressive. Theorem 1 therefore reveals a trade-off in which the risk-set estimation and statistical estimation errors increase with network complexity, whereas the approximation error decreases as the network becomes more expressive. It also explicitly shows that the approximation error deteriorates as the covariate dimension p increases. These observations provide practical

guidance for controlling network complexity and motivate dimension reduction or additional structural assumptions in high-dimensional applications.

We apply the proposed method to three examples to illustrate the proposed method.

Example **1.** We consider the risk score function proposed by Wang et al. (2016). For a subject with the covariates $\boldsymbol{X} = \boldsymbol{x}$, the prediction rule $U(t, \boldsymbol{x})$ at time t is given by

$$U(t, \boldsymbol{x}) = \alpha(t) + g^*(\boldsymbol{x}),$$

where $\alpha(t)$ and $g^*(\boldsymbol{x})$ are unknown functions, representing the time-varying offset and the subject-specific risk score, respectively. For a subject with the covariates $\boldsymbol{X} = \boldsymbol{x}$ who is still at risk at time t, the sign of $U(t, \boldsymbol{x})$ is used for immediate risk prediction: if $U(t, \boldsymbol{x}) > 0$, the subject is predicted to experience the event at the next time point; otherwise, it is predicted to remain event-free. According to Wang et al. (2016), we can obtain $\check{g}$ by minimizing the following $PR_n(g)$,

$$PR_n(g) \equiv \frac{1}{n} \sum_{i=1}^{n} \Delta_i \left\{ \frac{\sum_{k=1}^{n} I(Y_k \geq Y_i)[a + g(\boldsymbol{X}_k) - g(\boldsymbol{X}_i)]_+}{\sum_{k=1}^{n} I(Y_k \geq Y_i)} \right\}. \tag{4}$$

The expression (4) implies that $f(t)$ and $h(t)$ are specified as the functions t and $[a + t]_+$, respectively. Since $\sup_{g \in \mathcal{H}_{\mathcal{D},\mathcal{W},\mathcal{B}}} ||g||_\infty \leq \mathcal{B}$, we can regard t and $[a + t]_+$ as bounded functions. In Corollary 2 and simulation studies, we set $a = 2$ for simplicity. The results for other cases are similar. Wang et al. (2016) grounded their approach in a search space encapsulated within a reproducing kernel Hilbert space, marking a significant departure from our method that functions within a DNN framework. This essential difference in the search space employed by our respective method subsequently leads to distinctly divergent methodologies for both estimation and proof verification. Conditions (C3) and (C5) are satisfied for this pair of (f, h). The following Corollary 2 provides an upper bound for $PR(\widehat{g}) - PR(g^*)$.

Corollary 2 *Assume that conditions (C1), (C2) and (C4) are satisfied. Then, we have that with probability at least* $1 - (e + 10)/n$,

$$\begin{aligned} PR(\widehat{g}) - PR(g^*) \leq & 4\{(4 + 4\mathcal{B})(\log n/n)^{1/2} + l_1 f_1(\mathcal{D}, \mathcal{W}, n)\}/a_2^2 + 2l_2 f_2(\mathcal{D}, \mathcal{W}, n) \\ & + 524 a_1 a_3 \sqrt{p}\{N^2 L^2 \log_3(N + 2)\}^{-1/p}/a_2, \end{aligned}$$

where

$$\begin{aligned} l_1 &= 2\sqrt{160(1 + \mathcal{B})(9 + 8\mathcal{B})^2/3}\sqrt{(3 + 2\mathcal{B})(9 + c_1 + 1/a_2)}, \\ l_2 &= \sqrt{320(1 + \mathcal{B})(17 + 16\mathcal{B})^2/3}\sqrt{(3 + 2\mathcal{B})(8 + c_1 + 1/a_2)}, \\ f_1(\mathcal{D}, \mathcal{W}, n) &= \sqrt{c_2 \mathcal{D} \mathcal{S}_1 \log \mathcal{S}_1 \log n/n}, \\ f_2(\mathcal{D}, \mathcal{W}, n) &= \sqrt{c_2 \mathcal{D} \mathcal{S}_2 \log \mathcal{S}_2 \log n/n}, \\ \mathcal{S}_1 &= 2(p + 1)\mathcal{W} + (\mathcal{D} - 1)(4\mathcal{W}^2 + 2\mathcal{W}) + 2\mathcal{W} + 1, \\ \mathcal{S}_2 &= 2(2p + 1)\mathcal{W} + (\mathcal{D} - 1)(4\mathcal{W}^2 + 2\mathcal{W}) + 2\mathcal{W} + 1, \end{aligned}$$

and c_1 *and* c_2 *are two positive constants.*

In Corollary 2,

$$||\dot{f}||_\infty = 1, \quad ||f||_\infty = 2 + 2\mathcal{B}, \quad ||h||_\infty = 2 + 2\mathcal{B}, \quad a_4 = c_1, \quad a_5 = 1/a_2,$$
$$a_6 = 2/a_2, \quad \Psi_1(\mathcal{W}, \mathcal{D}) = c_2 \mathcal{D}\mathcal{S}_1 \log \mathcal{S}_1, \quad \text{and} \quad \Psi_2(\mathcal{W}, \mathcal{D}) = c_2 \mathcal{D}\mathcal{S}_2 \log \mathcal{S}_2.$$

Similar to the discussion following Theorem 1, $4\{(4 + 4\mathcal{B})(\log n/n)^{1/2} + l_1 f_1(\mathcal{D}, \mathcal{W}, n)\}/a_2^2$ captures the uncertainty of $f_g^*(\boldsymbol{z})$, and the other two terms are the upper bounds for the estimation and approximation errors, respectively. We can obtain the optimal convergence rate of the upper bound for $PR(\widehat{g}) - PR(g^*)$ by balancing these three terms. Let $N = O(n^{p/(2p+4)}/(\log n)^{(p+1)/(p+2)})$ and $L = O(1)$. Then the upper bound for $PR(\widehat{g}) - PR(g^*)$ achieves the optimal rate $(\log n/n)^{1/(p+2)}$.

Example **2.** If g^* is a good risk score, then $g^*(\boldsymbol{X}) \geq g^*(\tilde{\boldsymbol{X}})$ is prone to happen in the case of $\tilde{\boldsymbol{Y}} \geq \boldsymbol{Y}$. As a result, we consider minimizing the following probability to obtain the risk score g^*,

$$E[\Delta P\{g(\tilde{\boldsymbol{X}}) > g(\boldsymbol{X}) \mid \boldsymbol{Y}, \boldsymbol{X}, \tilde{\boldsymbol{Y}} \geq \boldsymbol{Y}\}].$$

Using the conditional probability formula, we have

$$P\{g(\tilde{\boldsymbol{X}}) > g(\boldsymbol{X}) \mid \boldsymbol{Y}, \boldsymbol{X}, \tilde{\boldsymbol{Y}} \geq \boldsymbol{Y}\} = \frac{P\{g(\tilde{\boldsymbol{X}}) > g(\boldsymbol{X}), \tilde{\boldsymbol{Y}} \geq \boldsymbol{Y} \mid \boldsymbol{Y}, \boldsymbol{X}\}}{P(\tilde{\boldsymbol{Y}} \geq \boldsymbol{Y} \mid \boldsymbol{Y}, \boldsymbol{X})}.$$

This implies

$$E[\Delta P\{g(\tilde{\boldsymbol{X}}) > g(\boldsymbol{X}) \mid \boldsymbol{Y}, \boldsymbol{X}, \tilde{\boldsymbol{Y}} \geq \boldsymbol{Y}\}] = E\left[\Delta \frac{P\{g(\tilde{\boldsymbol{X}}) > g(\boldsymbol{X}), \tilde{\boldsymbol{Y}} \geq \boldsymbol{Y} \mid \boldsymbol{Y}, \boldsymbol{X}\}}{P(\tilde{\boldsymbol{Y}} \geq \boldsymbol{Y} \mid \boldsymbol{Y}, \boldsymbol{X})}\right].$$

We obtain $\check{g}$ by minimizing the following empirical version $PR_n(g)$,

$$PR_n(g) \equiv \frac{1}{n} \sum_{i=1}^{n} \Delta_i \left[\frac{\sum_{k=1}^n I(Y_k \geq Y_i) I\{g(\boldsymbol{X}_k) - g(\boldsymbol{X}_i) > 0\}}{\sum_{k=1}^n I(Y_k \geq Y_i)}\right]. \tag{5}$$

The previous loss function is closely related to the partial-rank objective functions used in Khan and Tamer (2007) and Song et al. (2007). Khan and Tamer (2007) proposed a partial-rank criterion for censored transformation models, while Song et al. (2007) replaced the discontinuous indicator function with a smooth approximation to facilitate optimization. These methods and our approach all exploit pairwise orderings by comparing an individual with an observed event with individuals who remain at risk at its event time. The loss constructions, however, are different. The criteria of Khan and Tamer (2007) and Song et al. (2007) aggregate all comparable pairs in a single objective function. In contrast, our loss first normalizes, for each individual with an observed event, the number of incorrectly ordered pairwise comparisons by the size of the corresponding risk set and then averages the resulting individual-specific losses. Moreover, the previous methods use a linear index $\boldsymbol{X}^\top \boldsymbol{\beta}$ and focus on estimation and inference for the finite-dimensional coefficient $\boldsymbol{\beta}$, whereas we use a nonparametric DNN risk score $g(\boldsymbol{X})$ and derive the excess-risk bound.

The expression (5) indicates that $f(t)$ and $h(t)$ are specified as t and $I(t > 0)$, respectively. The upper bound for $PR(\widehat{g}) - PR(g^*)$ is derived in the next Corollary 3.

Corollary 3 *Suppose that conditions (C1), (C2) and (C4) are satisfied. Then, we have that with probability at least* $1 - 8/n$,

$$PR(\widehat{g}) - PR(g^*) \leq 4\{2(\log n/n)^{1/2} + l_1 f_1(\mathcal{D}, \mathcal{W}, n)\}/a_2^2 + 2l_2 f_2(\mathcal{D}, \mathcal{W}, n) + 524 a_1 a_3 c_4 \sqrt{p}\{N^2 L^2 \log_3(N+2)\}^{-1/p}/a_2,$$

where

$$l_1 = 40\sqrt{10(9 + c_1 + 1/a_2)/3}, \quad l_2 = 24\sqrt{15(8 + c_1 + 1/a_2)},$$
$$f_1(\mathcal{D}, \mathcal{W}, n) = \sqrt{c_3 \mathcal{S}_1 p(2\mathcal{W}\mathcal{D} + 1)\log n/n},$$
$$f_2(\mathcal{D}, \mathcal{W}, n) = \sqrt{2c_3 \mathcal{S}_2 p(2\mathcal{W}\mathcal{D} + 1)\log n/n},$$

c_3 *is a positive number and* c_4 *is the maximum value of the density function of* $g^*(\boldsymbol{X})$.

In Corollary 3,

$$||\dot{f}||_\infty = 1, \quad ||f||_\infty = 1, \quad ||h||_\infty = 1, \quad a_4 = c_1, \quad a_5 = 1/a_2, \quad a_6 = 2c_4/a_2,$$
$$\Psi_1(\mathcal{W}, \mathcal{D}) = c_3 \mathcal{S}_1 p(2\mathcal{W}\mathcal{D} + 1), \quad \text{and} \quad \Psi_2(\mathcal{W}, \mathcal{D}) = 2c_3 \mathcal{S}_2 p(2\mathcal{W}\mathcal{D} + 1).$$

The structure of the upper bound in Corollary 3 is the same as that in Corollary 2. We balance the three terms in Corollary 3 to obtain the optimal convergence rate of the upper bound for $PR(\widehat{g}) - PR(g^*)$. Let $N = O(1)$ and $L = O(\{n/(\log n)\}^{p/(2p+4)})$. Then the upper bound for $PR(\widehat{g}) - PR(g^*)$ achieves the optimal rate $(\log n/n)^{1/(p+2)}$.

Example **3.** We consider the NPH model. As discussed in Section 2, we can estimate $g^*(\boldsymbol{x})$ by minimizing the loss function (1), which is a special case of the expression (2) by replacing $\{f(t), h(t)\}$ with $\{\log(t), \exp(t)\}$. In the instance where the dimension of X is equal to 1, Chen et al. (2010) suggested estimating g^* through the maximization of a global partial likelihood, derived via the local polynomial approach. Extending their methodology to accommodate multiple covariates poses significant challenges. Conversely, our proposed method, along with Katzman et al. (2018) and Kvamme et al. (2019), which employs a ReLU DNN to parameterize g, readily accommodates a larger number of covariates. Apart from conditions (C1), (C2) and (C4), an additional condition (D1) is needed to prove Corollary 4.

(D1) $\inf_{0 \leq t \leq \tau} E\{I(Y \geq t \mid \boldsymbol{X})\} \geq a_2$.

Define $\boldsymbol{O}_i = (\boldsymbol{X}_i, \Delta_i, Y_i)$ $(i = 1, \ldots, n)$. Under condition (D1), $E[\{g^*(\boldsymbol{X}) - \widehat{g}(\boldsymbol{X})\}^2 \mid \boldsymbol{O}_1, \ldots, \boldsymbol{O}_n]$ can be bounded by $PR(\widehat{g}) - PR(g^*)$ up to a constant. The upper bounds for $E[\{g^*(\boldsymbol{X}) - \widehat{g}(\boldsymbol{X})\}^2 \mid \boldsymbol{O}_1, \ldots, \boldsymbol{O}_n]$, $E\{g^*(\boldsymbol{X}) - \widehat{g}(\boldsymbol{X})\}^2$, and $PR(\widehat{g}) - PR(g^*)$ are derived in the next Corollary 4, respectively.

Corollary 4 *Assume that conditions (C1), (C2), (C4) and (D1) hold. Then, we have that* (*i*) *with probability at least* $1 - 8/n$,

$$PR(\widehat{g}) - PR(g^*) \leq 4e^{2\mathcal{B}}\{2e^{2\mathcal{B}}(\log n/n)^{1/2} + l_1 f_1(\mathcal{D}, \mathcal{W}, n)\}/a_2^2 + 2l_2 f_2(\mathcal{D}, \mathcal{W}, n) + 524 a_1 a_3 \sqrt{p} e^{4\mathcal{B}}\{N^2 L^2 \log_3(N+2)\}^{-1/p}/a_2.$$

where

$$l_1 = 2\sqrt{80e^{2\mathcal{B}}(1+4e^{2\mathcal{B}})^2/3}\sqrt{(1+e^{2\mathcal{B}})(9+c_1+e^{4\mathcal{B}}/a_2)},$$
$$l_2 = \sqrt{320\mathcal{B}(1+16\mathcal{B})^2/3}\sqrt{(1+2\mathcal{B})(8+c_1+e^{4\mathcal{B}}/a_2)},$$
$$f_1(\mathcal{D},\mathcal{W},n) = \sqrt{c_2\mathcal{D}\mathcal{S}_1\log\mathcal{S}_1\log n/n},$$

and

$$f_2(\mathcal{D},\mathcal{W},n) = \sqrt{c_2\mathcal{D}\mathcal{S}_2\log\mathcal{S}_2\log n/n}.$$

(*ii*) *with probability at least* $1-10/n$,

$$\begin{aligned}&E[\{\widehat{g}(\boldsymbol{X})-g^*(\boldsymbol{X})\}^2 \mid \boldsymbol{O}_1,\ldots,\boldsymbol{O}_n]\\ \leq&\frac{4e^{4\mathcal{B}}}{a_2E(\Delta)}[4e^{2\mathcal{B}}\{2e^{2\mathcal{B}}(\log n/n)^{1/2}+l_1f_1(\mathcal{D},\mathcal{W},n)\}/a_2^2+2l_2f_2(\mathcal{D},\mathcal{W},n)\\ &+524a_1a_3\sqrt{p}e^{4\mathcal{B}}\{N^2L^2\log_3(N+2)\}^{-1/p}/a_2]+2l_3^2f_3^2(\mathcal{D},\mathcal{W},n),\end{aligned}$$

where

$$l_3 = \sqrt{80\mathcal{B}(1+4\mathcal{B})^2/3}\sqrt{(\mathcal{B}+1)(8+c_1)},$$
$$f_3(\mathcal{D},\mathcal{W},n) = \sqrt{c_2\mathcal{D}\mathcal{S}_3\log\mathcal{S}_3\log n/n},$$

and

$$\mathcal{S}_3 = (p+1)\mathcal{W}+(\mathcal{D}-1)(\mathcal{W}^2+\mathcal{W})+\mathcal{W}+1.$$

(*iii*)

$$\begin{aligned}&E\{\widehat{g}(\boldsymbol{X})-g^*(\boldsymbol{X})\}^2\\ \leq&\frac{4e^{4\mathcal{B}}}{a_2E(\Delta)}[4e^{2\mathcal{B}}\{2e^{2\mathcal{B}}(\log n/n)^{1/2}+l_1f_1(\mathcal{D},\mathcal{W},n)\}/a_2^2+2l_2f_2(\mathcal{D},\mathcal{W},n)\\ &+524a_1a_3\sqrt{p}e^{4\mathcal{B}}\{N^2L^2\log_3(N+2)\}^{-1/p}/a_2]+2l_3^2f_3^2(\mathcal{D},\mathcal{W},n)+\frac{10}{n}(2\mathcal{B}+a_5)^2.\end{aligned}$$

In Corollary 4,

$||\dot{f}||_\infty = e^{2\mathcal{B}}$, $||f||_\infty = 2\mathcal{B}$, $||h||_\infty = e^{2\mathcal{B}}$, $a_4 = c_1$, $a_5 = e^{4\mathcal{B}}/a_2$, $a_6 = 2e^{\mathcal{B}}/a_2$,
$\Psi_1(\mathcal{W},\mathcal{D}) = c_2\mathcal{D}\mathcal{S}_1\log\mathcal{S}_1$, and $\Psi_2(\mathcal{W},\mathcal{D}) = c_2\mathcal{D}\mathcal{S}_2\log\mathcal{S}_2$.

The structure of the upper bound in (i) is the same as that in Corollary 2. It can be shown that

$$E[\{\widehat{g}(\boldsymbol{X})-g^*(\boldsymbol{X})\}^2 \mid \boldsymbol{O}_1,\ldots,\boldsymbol{O}_n] \leq 2(P_n\widehat{g}-P\widehat{g})^2+\frac{4e^{4\mathcal{B}}}{a_2E(\Delta)}\{PR(\widehat{g})-PR(g^*)\}.$$

The term $2l_3^2 f_3^2(\mathcal{D}, \mathcal{W}, n)$ in (ii) represents an upper bound for $2(P_n\widehat{g} - P\widehat{g})^2$, while other terms in (ii) provide an upper bound of $4e^{4\mathcal{B}}\{PR(\widehat{g}) - PR(g^*)\}/\{a_2 E(\Delta)\}$. Moreover, since the event that $E[\{\widehat{g}(\boldsymbol{X}) - g^*(\boldsymbol{X})\}^2 \mid \boldsymbol{O}_1, \ldots, \boldsymbol{O}_n]$ is larger than the upper bound in (ii) may occur, an additional term $10(2\mathcal{B} + a_5)^2/n$ is required in (iii). Similar to the discussion following Corollary 2, we can obtain that these three upper bounds achieve the optimal rate $(\log n/n)^{1/(p+2)}$ in the case of $N = O(n^{p/(2p+4)}/(\log n)^{(p+1)/(p+2)})$ and $L = O(1)$.

Remark 5 *The neural-network class used in our analysis differs from that of Zhong et al. (2022): they used a sparse ReLU network with uniformly bounded weights and biases, whereas we consider a bounded-output ReLU network without explicit parameter-wise or sparsity constraints. Consequently, their network complexity is directly characterized by the sparsity level, while ours is controlled through $\Psi_2(\mathcal{W}, \mathcal{D})$, which jointly depends on the network width and depth.*

4. Simulation Studies

To assess the effectiveness of the proposed generalized loss function, simulations across examples 1, 2 and 3 were conducted, each representing a specific loss function to describe different situations. We conducted simulations using data generated from two types of models: the PH model and the accelerated failure time (AFT) model.

We set p at 5 and conducted a comparison between the three examples. The covariates $\boldsymbol{X}_0$ were generated from a multiple normal distribution that the marginal distribution followed $N(2, 1)$ and the covariance between X_{0i} and X_{0j} was 0.5, where X_{0i} and X_{0j} were the ith and jth elements of $\boldsymbol{X}_0$, respectively. We truncated each element of $\boldsymbol{X}_0$ to the interval $[-5, 5]$ to obtain $\boldsymbol{X}$. Two kinds of risk scores $q(\boldsymbol{x})$ were considered as follows, including linear and nonlinear functions:
$q_1(\boldsymbol{x}) = 0.05(x_1 + 2x_2 + 3x_3 + 4x_4 + 5x_5)$;
$q_2(\boldsymbol{x}) = 0.05\{x_1^2 x_2^2 + x_3 + (x_4 x_5)^{1/2} + \exp(x_5/2) - 1\}$.
We should emphasize that $q(\boldsymbol{x})$ is not $g^*(\boldsymbol{x})$. With the underlying risk score function $q(\boldsymbol{x})$, the failure time of the PH data was generated based on the model $\lambda(t \mid \boldsymbol{x}) = 0.1t \exp\{q(\boldsymbol{x})\}$. For the AFT data, we assumed that $\log T = -q(\boldsymbol{X}) + \epsilon$ to ensure that the failure time decreases as the risk score increases, where ϵ followed a standard logistic distribution. The censoring time C was uniformly distributed as $U(0, c)$, where c was chosen to achieve the specified censoring rate (CR). 2000 and 4000 samples were employed for the training of the network, while 20000 samples were applied as the validation data. The results below were based on 100 replications with a CR of 10% and 20%, respectively.

All neural networks were implemented in PyTorch. Each network consisted of 3 or 4 hidden layers, with a width of either 128 or 256 neurons. A dropout rate of 0.5 was applied to mitigate overfitting. Model parameters were optimized using the Adam optimizer with an initial learning rate selected from the set $\{0.1, 0.01, 0.005, 0.001, 0.0005, 0.0001\}$ based on the criterion of minimizing the training loss. In addition, a learning-rate scheduling strategy was employed as an implicit regularization mechanism: the learning rate was reduced by a factor of 0.5 if the training loss did not improve for 7 consecutive epochs. The default Adam momentum parameters $(\theta_1, \theta_2) = (0.9, 0.999)$ were adopted. Training was performed with a maximum of 2000 epochs and a mini-batch size of 200. Early stopping was applied when the training loss showed no improvement for 10 epochs.

Typically, a high score represents a high risk of failure. We denoted the estimated risk scores in the three examples as $\widehat{g}_1$, $\widehat{g}_2$ and $\widehat{g}_3$, respectively. For each $\widehat{g}_i, i = 1, 2, 3$, we separated the 20000 validation subjects into two groups (i.e., the high-risk and the low-risk groups) through their values of $\widehat{g}_i$. The cutting point was set as the 0.5 quantile of $\widehat{g}_i$ in this study. To evaluate the performance of $\widehat{g}_i$, the following three criteria were proposed: the logrank Chi-square statistic (Wang et al., 2016) (logrank) calculated based on the high-risk and the low-risk groups, the hazard ratio (ratio) between the high-risk and the low-risk groups, and the Spearman rank correlation (corre) between $\widehat{g}_i(\boldsymbol{x})$ and $q(\boldsymbol{x})$. A higher value of either logrank or ratio indicates a larger difference between the two groups, while a higher value of corre signifies a stronger positive correlation between $\widehat{g}_i(\boldsymbol{x})$ and $q(\boldsymbol{x})$. Moreover, we chose an "optimal risk score" $\widehat{g}_{\text{opt}}$ based on the training samples, the 2-fold cross-validation and the corre criterion. Specifically, we randomly divided the training samples into two parts in a $1:1$ ratio. For each loss function, the first part was used to train the network and the second part was adopted to calculate a corre value. Subsequently, the network was trained on the second part, and the corre value was calculated on the first part. The average of these two values was taken as the final corre value for the corresponding risk score of this loss function. Finally, the risk score with the highest corre value was chosen as the "optimal risk score" $\widehat{g}_{\text{opt}}$.

	$\widehat{g}_{\text{opt}}$			g_1			g_2			g_3		
						sample size 2000						
CR	corre	logrank	ratio	corre	logrank	ratio	corre	logrank	ratio	corre	logrank	ratio
$q_1, 10\%$	0.8588	149.1	2.185	0.8588	149.1	2.1849	0.2131	17.25	1.2015	0.8130	129.6	2.0720
$q_1, 20\%$	0.8537	145.4	2.165	0.8545	145.8	2.1672	0.1623	4.782	1.1467	0.7737	106.8	1.9349
$q_2, 10\%$	0.9053	623.1	5.237	0.9053	623.1	5.2371	0.1558	16.56	1.2927	0.8628	571.7	4.8534
$q_2, 20\%$	0.8984	597.0	5.033	0.9033	602.1	5.0715	0.1170	6.123	1.1678	0.8044	465.6	4.1031
						sample size 4000						
CR	corre	logrank	ratio	corre	logrank	ratio	corre	logrank	ratio	corre	logrank	ratio
$q_1, 10\%$	0.8753	151.6	2.200	0.9086	164.0	2.2751	0.2638	8.519	1.2999	0.8296	135.3	2.1060
$q_1, 20\%$	0.8906	156.4	2.230	0.8906	156.4	2.2304	0.2212	12.44	1.2493	0.8499	146.0	2.1694
$q_2, 10\%$	0.9187	673.9	5.651	0.9187	673.9	5.6513	0.1668	17.68	1.3040	0.8429	631.5	4.8344
$q_2, 20\%$	0.8997	652.8	5.478	0.8997	652.8	5.4781	0.1477	7.471	1.1874	0.8873	559.8	4.9068

Table 1: Simulation results for the PH data with bounded covariates.

The means of three criteria under each loss function and the outcomes selected based on the "optimal risk score" $\widehat{g}_{\text{opt}}$ were presented for comparison. As illustrated in Tables 1 and 2, across cross-validation, the results of $\widehat{g}_{\text{opt}}$ demonstrates satisfactory performance, with logrank exceeding or approaching 100 and corre surpassing 0.85 in most cases. Since the loss function $g_1(\boldsymbol{x})$ is most frequently selected during cross-validation, the results of $\widehat{g}_{\text{opt}}$ and the loss function $g_1(\boldsymbol{x})$ share the same outcomes in most cases. Apart from the influence of the loss function $g_1(\boldsymbol{x})$ on $\widehat{g}_{\text{opt}}$, $g_3(\boldsymbol{x})$ is also selected in some simulation iterations. In comparison, $g_2(x)$ is chosen rarely during the cross-validation step. The sub-optimal performance of $g_2(\boldsymbol{x})$ may be attributed to its discontinuous loss function. Fortunately, since our objective is to choose the optimal risk score, the poorly performed $g_2(x)$ hardly affects our method's performance. In addition, for each of the three loss functions, the proposed method performs more superiorly across all three criteria when applied to the PH data than

	$\widehat{g}_{\text{opt}}$			g_1			g_2			g_3		
					sample size 2000							
CR	corre	logrank	ratio	corre	logrank	ratio	corre	logrank	ratio	corre	logrank	ratio
$q_1, 10\%$	0.8207	84.43	1.7928	0.8207	84.43	1.7928	0.1561	14.95	1.1131	0.7175	55.91	1.6069
$q_1, 20\%$	0.8188	82.52	1.7801	0.8188	82.52	1.7801	0.0324	1.900	1.0345	0.6334	43.44	1.5182
$q_2, 10\%$	0.8890	429.3	3.8456	0.8890	429.3	3.8456	0.0936	16.11	1.1185	0.7325	271.0	2.8769
$q_2, 20\%$	0.8515	444.7	3.9314	0.8515	444.7	3.9314	−0.0163	2.726	0.9581	0.6394	144.7	2.1536
					sample size 4000							
CR	corre	logrank	ratio	corre	logrank	ratio	corre	logrank	ratio	corre	logrank	ratio
$q_1, 10\%$	0.8554	90.77	1.8330	0.8554	90.77	1.8330	0.1729	18.76	1.1287	0.7394	63.04	1.6556
$q_1, 20\%$	0.8521	90.77	1.8333	0.8521	90.77	1.8333	0.1910	25.13	1.1507	0.6630	50.65	1.5704
$q_2, 10\%$	0.8779	457.0	4.0201	0.8779	457.0	4.0201	0.1293	35.66	1.1826	0.7324	254.0	2.7803
$q_2, 20\%$	0.8660	453.7	4.1697	0.8760	460.2	4.0400	0.0114	0.592	0.9876	0.7027	239.5	2.6972

Table 2: Simulation results for the AFT data with bounded covariates.

to the AFT data. Moreover, as the sample size n increases, all three criteria increase under each loss function in most cases, as expected. We further investigated the precision of the proposed methods through variances. Since different criteria exhibit varying magnitudes, we considered the relative variances, i.e., the root of sample variances divided by the sample mean squares, resulting in most of them being less than 0.05. Furthermore, although results obtained across the three criteria demonstrate effective predictions, consistent outcomes are not always guaranteed. For instance, in the case of q_2 under 20% censoring in Table 2, logrank is superior substantially under sample size 2000, yet corre and ratio possess lower values, compared with the outcomes of sample size 4000. Despite this, generally speaking, the results in most cases show a consistent trend under all criteria.

We also compared the proposed method with that of Wang et al. (2016) in the Appendix. The results show that the proposed method has lower loss values, indicating its superior optimization performance.

5. Application

In this section, we analyzed a SEER colon cancer data set spanning from 1988 to 2015, which was publicly accessible at `https://seer.cancer.gov/data/`. After excluding all subjects with missing covariates and survival times, the data set comprised $213,900$ observations. The data set included 17 covariates: 2 binary variables, 9 categorical variables, and 6 continuous variables, which were detailed in Table 3. For the categorical variables, one-hot encoding was employed to map each category to a separate binary vector, ensuring these variables were appropriately pre-processed for the analysis. After encoding, the covariates were referred to as features to distinguish them from the original ones. The dimension of the features is 161. The original data provided the complete lifetime information. To align with our model assumption, the censoring time C was uniformly distributed as $U(0, c_A)$, with c_A selected to achieve the specified CR of 10% and 20%, which mirrors the settings used in the simulation study. 90% of the observations were used to train the network, while the remaining 10% served as the validation set. In addition, consistent with the simulation

studies, a 2-fold cross-validation was used to determine the optimal risk score. The training procedure was the same as in the simulation studies.

No.	Type	Category
1	continuous	Age recode with <1 year olds
2	categorical	Race recode
3	binary	Sex
4	continuous	Year of diagnosis
5	categorical	Site recode ICD-O-3/WHO 2008
6	categorical	Diagnostic confirmation
7	continuous	Regional nodes examined
8	continuous	Regional nodes positive
9	binary	First malignant primary indicator
10	continuous	Total number of in situ/malignant tumors for patient
11	continuous	Total number of benign/borderline tumors for patient
12	categorical	ICD-O-3 Hist/behav
13	categorical	RX Summ-Surg/Rad Seq
14	categorical	Reason no cancer(directed surgery)
15	categorical	COD to site recode
16	categorical	Marital status at diagnosis
17	categorical	SEER historic stage A

Table 3: Corvariates of the SEER data.

	g_1			g_2			g_3		
CR	corre	logrank	ratio	corre	logrank	ratio	corre	logrank	ratio
10%	0.7111	478.2	7.7895	0.2995	26.25	1.5617	0.7206	474.1	7.9446
20%	0.6625	462.0	6.7396	0.2056	27.93	1.5835	0.6819	425.5	6.8973

Table 4: Application results of our proposed method.

As determined through the cross-validation step, g_3 is chosen as the optimal function $\widehat{g}_{\text{opt}}$ for the two censoring rates. Table 4 presents the performance of three loss functions across three criteria, illustrating that $g_1(\boldsymbol{x})$ and $g_3(\boldsymbol{x})$ yield comparable results and outperform $g_2(\boldsymbol{x})$ across all criteria. Given the similarity in results between $g_1(\boldsymbol{x})$ and $g_3(\boldsymbol{x})$, $g_1(\boldsymbol{x})$ is recommended for this application to maintain consistency with the simulation finding that the loss function $g_1(\boldsymbol{x})$ is mainly chosen at the cross-validation stage. Besides, although neural networks are often regarded as black boxes in the prediction explanation, we used the integrated gradients (IG) approach to evaluate the importance of covariates in our study. The IG approach quantifies the influence of inputs on the output by calculating the average gradients of different inputs and is a widely used interpretive technique for DNN. By using $\widehat{g}_{\text{opt}}$ as the loss function and the trained neural network, Figure 1 displays importance scores of the features under the two CRs, respectively. By mapping features with an importance score greater than 0.1 to their corresponding covariates, we obtained

that the important covariates under both CRs are Nos. 1, 4, 6, 9, 10, 14, 15, and 17. The data set was also analyzed by using the semi-parametric PH model with 161 features, where the Lasso penalty was employed for feature selection. This model is appropriate because Grambsch and Therneau (1994) suggested that the data set satisfies the PH assumption. Similar to the proposed method, we mapped these features to the corresponding covariates, identifying covariates Nos. 1, 3, 4, 5, 6, 8, 9, 14, 15, 16, and 17 as significantly influencing the survival of colon cancer patients. A comparison of the results reveals that the covariates Nos. 1, 4, 6, 9, 14, 15, and 17 are selected by both methods, indicating that the proposed method demonstrates acceptable covariate selection capability. As mentioned in Section 4, we compared the proposed method with the support vector hazards machine method in Wang et al. (2016) in simulation studies. However, employing the method in Wang et al. (2016) for data analysis is impractical for the following reasons. During the quadratic programming stage, it is necessary to generate matrices of dimension $\mathcal{O}(n^2)$, which results in significant memory requirements. Standard desktop computers, and even high-performance servers, are generally inadequate to handle these demands. This limitation shows that the proposed method is more applicable in practical scenarios.

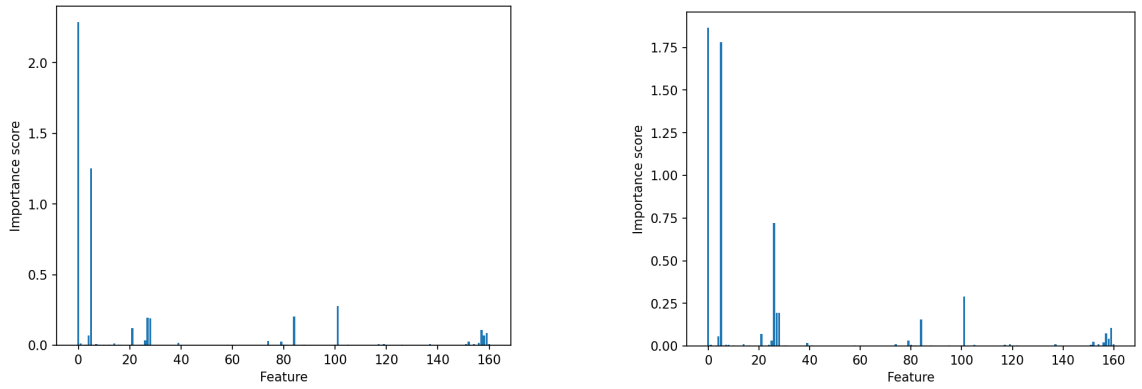


Figure 1: These two figures present the feature importance under our proposed method with the 10% CR (left) and 20% CR (right), respectively. The horizontal axis denotes the feature index, and the vertical axis represents the importance score.

We further compared the proposed method with the Cox-MLP approach in Kvamme et al. (2019). Their method performs parameter estimation based on a weighted partial log-likelihood and DNN. In Kvamme et al. (2019), three criteria were proposed to evaluate the performance of the Cox-MLP method: the concordance index (C-index), the integrated Brier score (IBS), and the negative binomial log-likelihood (nbll). The C-index is expected to be close to one, indicating better performance, whereas for IBS and nbll, smaller values are preferable. Table 5 compares the two methods by the three criteria mentioned above, and Table 6 by our three criteria. As shown in Table 5, our method performs comparably to Cox-MLP in terms of C-index and IBS, and outperforms it on nbll. Table 6 shows that our method surpasses Cox-MLP across all three criteria we proposed. We also examined the ability of Cox-MLP to identify significant covariates. Figure 2 displays the importance scores of the features obtained by Cox-MLP under the two CRs. Applying the aforementioned

thresholding rule, where features with importance scores exceeding 0.1 were mapped to the corresponding covariates, Cox-MLP identifies Nos. 1, 4, 7, 8, 13, and 14 as important covariates under both CRs. Comparing these with the covariates selected via the semi-parametric PH model with the Lasso penalty, the overlapping covariates are Nos. 1, 4, 8, and 14. These results suggest that Cox-MLP identifies fewer relevant covariates than the proposed method.

	Cox-MLP			our proposed		
CR	C-index	IBS	nbll	C-index	IBS	nbll
10%	0.5743	0.1639	0.4863	0.5769	0.1692	0.4211
20%	0.5490	0.1656	0.4885	0.5482	0.1616	0.4043

Table 5: Comparison of the Cox-MLP and our proposed methods using C-index, IBS, and nbll.

	Cox-MLP			our proposed		
CR	corre	logrank	ratio	corre	logrank	ratio
10%	-0.0784	153.9	0.7967	0.7206	474.1	7.9446
20%	-0.0994	134.8	0.7348	0.6819	425.5	6.8973

Table 6: Comparison of the Cox-MLP and our proposed methods using corre, logrank and ratio.

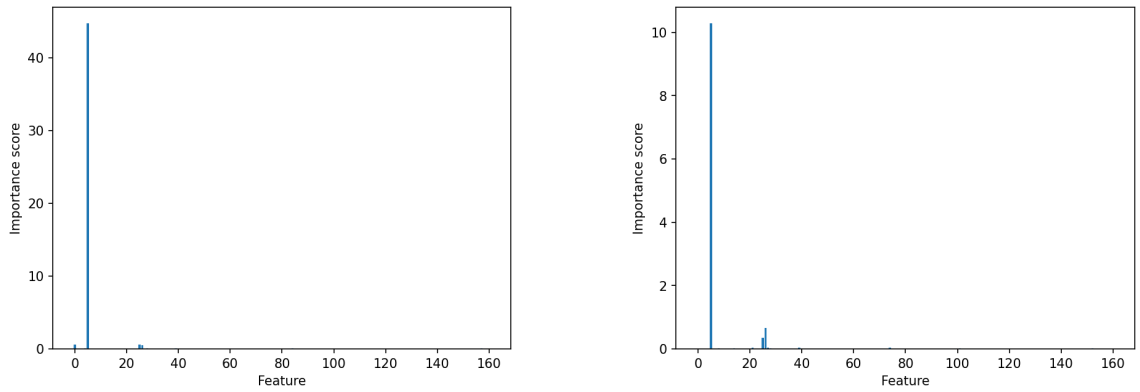


Figure 2: These two figures present the feature importance under the Cox-MLP method with the 10% CR (left) and 20% CR (right), respectively. The horizontal axis denotes the feature index, and the vertical axis represents the importance score.

6. Discussion

In this article, we proposed to learn the risk score based on a general loss function, in which the score function was parameterized by a ReLU DNN. The non-asymptotic error bound was established. Moreover, we provided three examples to illustrate the proposed method. The simulation results showed that our approach works well for the situations considered. An application to a SEER colon cancer data set was provided.

In this study, we confined our analysis to the scenario where the support of $\boldsymbol{X}$ was a bounded rectangle. In many applications, however, the dimension of $\boldsymbol{X}$ could be exceedingly large, potentially resulting in an extremely slow convergence rate, as indicated by Corollaries 2-4. We could impose some local manifold structure on the support of $\boldsymbol{X}$ to address this issue (Baraniuk and Wakin, 2009). In addition, we only assumed that g^* is a continuous function, without imposing any additional smoothness conditions. Deriving the upper bound for the loss difference and establishing the optimal convergence rate (Schmidt-Hieber, 2020; Lu et al., 2021; Zhong et al., 2022; Jiao et al., 2023) under the assumption that g^* is a smooth function merits future research. Moreover, our current theoretical framework does not yet provide insight into the precise conditions under which one loss function may outperform another, and it remains challenging to offer a similar interpretation for the risk score in the general framework as we did in examples $1- -3$. These represent interesting and important directions for future research. Nonetheless, from a practical standpoint, the proposed method is well grounded as the three examples already cover a reasonably wide spectrum of modeling needs. Example 1 constructs a prediction rule that links the risk score with a time-varying offset, example 2 is based on a rank-based loss, offering a robust alternative, and example 3 corresponds to the PH model, the most widely used model in survival analysis. Furthermore, we employed cross-validation to select the best-performing risk score among the three. Together, these features ensure that the proposed framework is practically useful, even in the absence of deeper theoretical guidance on when one loss may dominate another. Given the broad coverage of these examples, expanding the loss class by simply adding more candidate pairs for empirical selection is unnecessary. However, our method easily applies to any user-defined loss functions and selection using cross-validation.

The optimal convergence rate is determined by $\Psi_1(\mathcal{W}, \mathcal{D})$ and $\Psi_2(\mathcal{W}, \mathcal{D})$, which depend on the choices of f and h. When both f and h are Lipschitz continuous, or when f is Lipschitz continuous and $h(t) = I(t > 0)$, explicit forms of $\Psi_1(\mathcal{W}, \mathcal{D})$ and $\Psi_2(\mathcal{W}, \mathcal{D})$ can be obtained, yielding an optimal convergence rate of $(\log n/n)^{1/(p+2)}$ in both cases. For general choices of f and h, however, the explicit forms of $\Psi_1(\mathcal{W}, \mathcal{D})$ and $\Psi_2(\mathcal{W}, \mathcal{D})$ are unknown, and consequently the optimal convergence rate is also unknown. How general f and h influence the optimal rate therefore remains an open question worthy of further study. In addition, the theoretical guidance was not incorporated into our experimental setups, as the optimal configurations derived from theory require excessively large width or depth. In practice, adopting networks of such scale would lead to overfitting. Consequently, we selected the neural network structures based on empirical settings. Developing theories that can guide DNN configurations is a valuable direction for future research. Additionally, survival models with time-varying covariates have been widely studied in recent years (Huang et al., 1998; Ishwaran et al., 2008; Fu and Simonoff, 2017; Xie et al., 2024). However, the proposed method cannot be extended straightforwardly to address this case, making it a valuable

topic for future work. Besides, as dimension reduction has been a prominent focus in survival analysis (Chen et al., 1999; Xia et al., 2010; Sun et al., 2019; Molstad and Patra, 2023), integrating dimension reduction techniques with DNN warrants further investigation.

Acknowledgments

The authors thank the Editor, Professor Jie Peng and two reviewers for their insightful comments and suggestions that greatly improved the quality of the paper. Dongxiao Han's research is partly supported by the National Natural Science Foundation of China (12471259 and 12231011), Shenzhen Wukong Investment Management Co. Ltd, and Academy for Advanced Interdisciplinary Studies, Nankai University (NKQNJC202522). Zan Li's research is supported by the National Natural Science Foundation of China (12401354). Xinyuan Song's research is partly supported by the Research Grant Council of the Hong Kong Special Administration Region (14303622 and 14300425). Liuquan Sun's research is partly supported by the National Natural Science Foundation of China (12571299 and 12631011).

Appendix A. Technical proofs

This section contains the verification of condition (C3) and proofs for Theorem 1 and Corollaries 2-4.

A.1 Verification of Condition (C3)

In this subsection, we prove that condition (C3) is satisfied when f is Lipschitz continuous and h is either Lipschitz continuous or an indicator function. We first consider the case where both f and h are Lipschitz continuous functions. Assume that the following conditions are satisfied.

(E1) Both f and h are Lipschitz continuous functions with a common Lipschitz constant $b_1 \geq 1$.

(E2) $||f||_\infty \geq 2\mathcal{B}$ and $||h||_\infty \geq 2\mathcal{B}$.

For any $\boldsymbol{x}' \in [-a_1, a_1]^p$, let $\mathcal{G}^*_{\boldsymbol{x}'} = \{g(\boldsymbol{x}) - g(\boldsymbol{x}') \mid g \in \mathcal{H}_{\mathcal{D},\mathcal{W},\mathcal{B}}\}$. For any $\epsilon > 0$ and $\boldsymbol{x}' \in [-a_1, a_1]^p$, based on condition (E1), it can be shown that

$$N(\epsilon||h||_\infty, \mathcal{G}_{x'}, L_1(P_n)) \leq N(\epsilon||h||_\infty/b_1, \mathcal{G}^*_{x'}, L_1(P_n)). \tag{6}$$

For any $0 < \epsilon < 1$, by the facts that $b_1 \geq 1$ and $||h||_\infty \geq 2\mathcal{B}$, and Theorem 2.6.7 in Van Der Vaart and Wellner (1996), one can prove

$$\begin{aligned} N(\epsilon||h||_\infty/b_1, \mathcal{G}^*_{x'}, L_1(P_n)) \leq & N(2\epsilon\mathcal{B}/b_1, \mathcal{G}^*_{x'}, L_1(P_n)) \\ \leq & c_1 V(\mathcal{G}^*_{x'})(16eb_1)^{V(\mathcal{G}^*_{x'})} \left(\frac{1}{\epsilon}\right)^{V(\mathcal{G}^*_{x'})-1}, \end{aligned} \tag{7}$$

where c_1 is a positive constant and $V(\mathcal{G}^*_{x'})$ is the VC-index of $\mathcal{G}^*_{x'}$ (Van Der Vaart and Wellner, 1996, page 141). It is easy to show that any function in $\mathcal{G}^*_{x'}$ can be implemented

by a ReLU network with p input units, width $2\mathcal{W}$, and depth $\mathcal{D}$. For any $0 < \epsilon < 1$, in light of (6), (7) and Theorem 7 in Bartlett et al. (2019), we can show

$$N(\epsilon||h||_\infty, \mathcal{G}_{x'}, L_1(P_n)) \leq c_1 c_2 \mathcal{D}\mathcal{S}_1 \log \mathcal{S}_1 (16eb_1)^{c_2 \mathcal{D}\mathcal{S}_1 \log \mathcal{S}_1} \left(\frac{1}{\epsilon}\right)^{c_2 \mathcal{D}\mathcal{S}_1 \log \mathcal{S}_1 - 1}, \tag{8}$$

where c_2 is a positive constant and $\mathcal{S}_1 = 2(p+1)\mathcal{W} + (\mathcal{D}-1)(4\mathcal{W}^2 + 2\mathcal{W}) + 2\mathcal{W} + 1$. Define

$$\mathcal{G}^* = \{g'(\boldsymbol{x}_1, \boldsymbol{x}_2) = g(\boldsymbol{x}_1) - g(\boldsymbol{x}_2) \mid g \in \mathcal{H}_{\mathcal{D},\mathcal{W},\mathcal{B}}\}.$$

Based on conditions (C1) and (E1), and the fact that $||f||_\infty \geq 2\mathcal{B}$, it is easy to show

$$\begin{aligned} N(\epsilon||f||_\infty, \mathcal{G}_1, L_1(P_n)) \leq & N(a_2\epsilon||f||_\infty / b_1^2, \mathcal{G}^*, L_1(P^* \times P_n)) \\ \leq & N(2a_2\epsilon\mathcal{B}/b_1^2, \mathcal{G}^*, L_1(P^* \times P_n)), \end{aligned} \tag{9}$$

where P^* is the true probability measure generated by $\boldsymbol{X}$, and $P^* \times P_n$ is the product measurable of P^* and P_n on $[-a_1, a_1]^p \times [-a_1, a_1]^p$. It can be shown that any function in $\mathcal{G}^*$ can be implemented by a ReLU network with $2p$ input units, width $2\mathcal{W}$, and depth $\mathcal{D}$. For any $0 < \epsilon < 1$, according to (9), the fact that $b_1 \geq 1$, Theorem 2.6.7 in Van Der Vaart and Wellner (1996) and Theorem 7 in Bartlett et al. (2019), we can obtain

$$N(\epsilon||f||_\infty, \mathcal{G}_1, L_1(P_n)) \leq c_1 c_2 \mathcal{D}\mathcal{S}_2 \log \mathcal{S}_2 \left(\frac{16eb_1^2}{a_2}\right)^{c_2 \mathcal{D}\mathcal{S}_2 \log \mathcal{S}_2} \left(\frac{1}{\epsilon}\right)^{c_2 \mathcal{D}\mathcal{S}_2 \log \mathcal{S}_2 - 1}, \tag{10}$$

where $\mathcal{S}_2 = 2(2p+1)\mathcal{W} + (\mathcal{D}-1)(4\mathcal{W}^2 + 2\mathcal{W}) + 2\mathcal{W} + 1$.

Next we consider the case that f is Lipschitz continuous and h is an indicator function. Suppose that the following conditions hold.

(E3) f is a Lipschitz continuous function with the Lipschitz constant b_1 and $h(t) = I(t > 0)$.

(E4) $||f||_\infty \geq 1$.

By the previous proof, any function in $\mathcal{G}^*_{x'}$ can be implemented by a ReLU network with p input units, width $2\mathcal{W}$, and depth $\mathcal{D}$. Then according to Theorem 10 in Bartlett et al. (2019), one can show

$$V(\mathcal{G}_{x'}) \leq c_3 \mathcal{S}_1 p(2\mathcal{W}\mathcal{D} + 1),$$

where c_3 is a positive constant. For any $0 < \epsilon < 1$, it follows from Theorem 2.6.7 in Van Der Vaart and Wellner (1996) that

$$\begin{aligned} & N(\epsilon||h||_\infty, \mathcal{G}_{x'}, L_1(P_n)) \\ \leq & c_1 c_3 \mathcal{S}_1 p(2\mathcal{W}\mathcal{D} + 1)(16e)^{c_3 \mathcal{S}_1 p(2\mathcal{W}\mathcal{D}+1)} \left(\frac{1}{\epsilon}\right)^{c_3 \mathcal{S}_1 p(2\mathcal{W}\mathcal{D}+1) - 1}. \end{aligned} \tag{11}$$

Define

$$\mathcal{G}^{**} = \{g''(\boldsymbol{x}_1, \boldsymbol{x}_2) = I\{g(\boldsymbol{x}_1) - g(\boldsymbol{x}_2) > 0\} \mid g \in \mathcal{H}_{\mathcal{D},\mathcal{W},\mathcal{B}}\}.$$

Based on conditions (C1), (E3) and (E4), it is easy to show

$$\begin{aligned} N(\epsilon||f||_\infty, \mathcal{G}_1, L_1(P_n)) \leq & N(a_2\epsilon||f||_\infty/b_1, \mathcal{G}^{**}, L_1(P^* \times P_n)) \\ \leq & N(a_2\epsilon/b_1, \mathcal{G}^{**}, L_1(P^* \times P_n)). \end{aligned} \tag{12}$$

According to the previous proof, we have that any function in $\mathcal{G}^*$ can be implemented by a ReLU network with $2p$ input units, width $2\mathcal{W}$, and depth $\mathcal{D}$. Then by Theorem 10 in Bartlett et al. (2019), one can prove

$$V(\mathcal{G}^{**}) \leq 2c_3\mathcal{S}_2p(2\mathcal{W}\mathcal{D}+1).$$

For any $0 < \epsilon < 1$, it follows from (12), the fact that $b_1 \geq 1$, and Theorem 2.6.7 in Van Der Vaart and Wellner (1996) that

$$\begin{aligned} & N(\epsilon||f||_\infty, \mathcal{G}_1, L_1(P_n)) \\ \leq & 2c_1c_3\mathcal{S}_2p(2\mathcal{W}\mathcal{D}+1)\left(\frac{16eb_1}{a_2}\right)^{2c_3\mathcal{S}_2p(2\mathcal{W}\mathcal{D}+1)}\left(\frac{1}{\epsilon}\right)^{2c_3\mathcal{S}_2p(2\mathcal{W}\mathcal{D}+1)-1}. \end{aligned} \tag{13}$$

A.2 Proof of Theorem 1

Let $\boldsymbol{O}_{n+1}$ be a random vector distributed as $\boldsymbol{O}$ and independent of $\{\boldsymbol{O}_1, \ldots, \boldsymbol{O}_n\}$. With some slight abuse of notations, for any measurable function Φ on $[-a_1, a_1]^p \times [0, \tau] \times \{0, 1\}$, define $P_n\Phi = \sum_{i=1}^n \Phi(\boldsymbol{O}_i)/n$, $P_n'\Phi = \sum_{i=2}^{n+1} \Phi(\boldsymbol{O}_i)/n$ and $P\Phi = E\Phi(\boldsymbol{O})$. For any $\boldsymbol{x} \in [-a_1, a_1]^p$, and any $y \in [0, \tau]$, let

$$\boldsymbol{z} = (\boldsymbol{x}^\top, y)^\top, \text{ and } f_g(\boldsymbol{z}) = f\left(\frac{E[I(Y \geq y)h\{g(\boldsymbol{X}) - g(\boldsymbol{x})\}]}{E\{I(Y \geq y)\}}\right).$$

Define

$$PR_n(g) = \frac{1}{n}\sum_{i=1}^n \Delta_i f\left[\frac{\sum_{k=1}^n I(Y_k \geq Y_i)h\{g(\boldsymbol{X}_k) - g(\boldsymbol{X}_i)\}}{\sum_{k=1}^n I(Y_k \geq Y_i)}\right], \text{ and } PR_n'(g) = P_n\Delta f_g.$$

According to the fact that $\check{g}$ is a minimizer of $PR_n(g)$, we have

$$\begin{aligned} & PR(\check{g}) - PR(g^*) \\ = & PR(\check{g}) - PR(g^*) - \{PR_n'(\check{g}) - PR_n'(g^*)\} + \{PR_n'(\check{g}) - PR_n'(g_0) + PR_n'(g_0) - PR_n'(g^*)\} \\ = & [PR(\check{g}) - PR(g^*) - \{PR_n'(\check{g}) - PR_n'(g^*)\}] + \{PR_n'(\check{g}) - PR_n'(g_0)\} \\ & + [\{PR_n'(g_0) - PR_n'(g^*)\} - \{PR(g_0) - PR(g^*)\}] + \{PR(g_0) - PR(g^*)\} \\ = & [PR(\check{g}) - PR(g^*) - \{PR_n'(\check{g}) - PR_n'(g^*)\}] \\ & + \{PR_n'(\check{g}) - PR_n(\check{g}) + PR_n(\check{g}) - PR_n(g_0) + PR_n(g_0) - PR_n'(g_0)\} \\ & + [\{PR_n'(g_0) - PR_n'(g^*)\} - \{PR(g_0) - PR(g^*)\}] + \{PR(g_0) - PR(g^*)\} \\ \leq & [PR(\check{g}) - PR(g^*) - \{PR_n'(\check{g}) - PR_n'(g^*)\}] + \{PR_n'(\check{g}) - PR_n(\check{g}) + PR_n(g_0) - PR_n'(g_0)\} \\ & + [\{PR_n'(g_0) - PR_n'(g^*)\} - \{PR(g_0) - PR(g^*)\}] + \{PR(g_0) - PR(g^*)\}. \end{aligned} \tag{14}$$

We first provide an upper bound for

$$PR_n'(\check{g}) - PR_n(\check{g}) + PR_n(g_0) - PR_n'(g_0).$$

By the triangle inequality and the mean-value theorem, we can obtain

$$\begin{aligned}
&PR_n'(\check{g}) - PR_n(\check{g}) + PR_n(g_0) - PR_n'(g_0) \\
\leq & 2 \sup_{g \in \mathcal{H}_{\mathcal{D},\mathcal{W},\mathcal{B}}} |PR_n'(g) - PR_n(g)| \\
\leq & 2||\dot{f}||_\infty \sup_{g \in \mathcal{H}_{\mathcal{D},\mathcal{W},\mathcal{B}}} \max_{1 \leq i \leq n} \left| \frac{\frac{1}{n}\sum_{k=1}^n I(Y_k \geq Y_i) h\{g(\boldsymbol{X}_k) - g(\boldsymbol{X}_i)\}}{\frac{1}{n}\sum_{k=1}^n I(Y_k \geq Y_i)} \right. \\
& \qquad \left. - \frac{E[I(Y \geq Y_i) h\{g(\boldsymbol{X}) - g(\boldsymbol{X}_i)\} \mid \boldsymbol{X}_i, Y_i]}{E\{I(Y \geq Y_i) \mid Y_i\}} \right|. \qquad (15)
\end{aligned}$$

Let $\boldsymbol{Z} = (\boldsymbol{X}^\top, Y)^\top$, and $\boldsymbol{Z}_{n+1}$ be a copy of $\boldsymbol{Z}$ and independent of $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$. For any $t > 0$, based on the Law of total expectation, the triangle inequality and the Hoeffding inequality (Hoeffding, 1963), we can show

$$\begin{aligned}
& P\bigg(\max_{1 \leq i \leq n} \left| \frac{1}{n}\sum_{k=1}^n I(Y_k \geq Y_i) - E\{I(Y \geq Y_i) \mid Y_i\} \right| > t \bigg) \\
\leq & nP\bigg(\left| \frac{1}{n}\sum_{k=1}^n I(Y_k \geq Y_1) - E\{I(Y \geq Y_1) \mid Y_1\} \right| > t \bigg) \\
= & nE\bigg\{ P\bigg(\left| \frac{1}{n}\sum_{k=1}^n I(Y_k \geq Y_1) - E\{I(Y \geq Y_1) \mid Y_1\} \right| > t \mid Y_1 \bigg) \bigg\} \\
\leq & nE\bigg\{ P\bigg(\left| \frac{1}{n}\sum_{k=2}^{n+1} I(Y_k \geq Y_1) - E\{I(Y \geq Y_1) \mid Y_1\} \right| > t - \frac{1}{n} \mid Y_1 \bigg) \bigg\} \\
\leq & 2n \exp\bigg\{ -2(t - \frac{1}{n})^2 n \bigg\}.
\end{aligned}$$

Let $t = 2(\log n/n)^{1/2}$. Then we have that with probability at least $1 - 2/n$,

$$\max_{1 \leq i \leq n} \left| \frac{1}{n}\sum_{k=1}^n I(Y_k \geq Y_i) - E\{I(Y \geq Y_i) \mid Y_i\} \right| \leq 2(\log n/n)^{1/2}. \qquad (16)$$

Using (16), we have that with probability at least $1 - 2/n$,

$$\begin{aligned}
& \sup_{g \in \mathcal{H}_{\mathcal{D},\mathcal{W},\mathcal{B}}} \max_{1 \leq i \leq n} \left| \frac{1}{n}\sum_{j=1}^n I(Y_j \geq Y_i) h\{g(\boldsymbol{X}_j) - g(\boldsymbol{X}_i)\} \bigg[E\{I(Y \geq Y_i) \mid Y_i\} - \frac{1}{n}\sum_{k=1}^n I(Y_k \geq Y_i) \bigg] \right| \\
\leq & ||h||_\infty \max_{1 \leq i \leq n} \left| \frac{1}{n}\sum_{k=1}^n I(Y_k \geq Y_i) - E\{I(Y \geq Y_i) \mid Y_i\} \right| \\
\leq & 2||h||_\infty (\log n/n)^{1/2}. \qquad (17)
\end{aligned}$$

In light of the triangle inequality and the Law of total expectation, one can show

$$
\begin{aligned}
&P\Bigg(\sup_{g\in\mathcal{H}_{\mathcal{D},\mathcal{W},\mathcal{B}}} \max_{1\le i\le n} \bigg| \frac{1}{n}\sum_{j=1}^{n} I(Y_j \ge Y_i) \Bigg(\frac{1}{n}\sum_{k=1}^{n} I(Y_k \ge Y_i) h\{g(\boldsymbol{X}_k) - g(\boldsymbol{X}_i)\} \\
&\qquad\qquad - E[I(Y \ge Y_i) h\{g(\boldsymbol{X}) - g(\boldsymbol{X}_i)\} \mid \boldsymbol{X}_i, Y_i] \Bigg) \bigg| > t \Bigg) \\
\le & nP\Bigg(\sup_{g\in\mathcal{H}_{\mathcal{D},\mathcal{W},\mathcal{B}}} \bigg| \frac{1}{n}\sum_{k=1}^{n} I(Y_k \ge Y_1) h\{g(\boldsymbol{X}_k) - g(\boldsymbol{X}_1)\} \\
&\qquad\qquad - E[I(Y \ge Y_1) h\{g(\boldsymbol{X}) - g(\boldsymbol{X}_1)\} \mid \boldsymbol{X}_1, Y_1] \bigg| > t \Bigg) \\
\le & nP\Bigg(\sup_{g\in\mathcal{H}_{\mathcal{D},\mathcal{W},\mathcal{B}}} \bigg| \frac{1}{n}\sum_{k=2}^{n+1} I(Y_k \ge Y_1) h\{g(\boldsymbol{X}_k) - g(\boldsymbol{X}_1)\} \\
&\qquad\qquad - E[I(Y \ge Y_1) h\{g(\boldsymbol{X}) - g(\boldsymbol{X}_1)\} \mid \boldsymbol{X}_1, Y_1] \bigg| > t - \frac{2||h||_\infty}{n} \Bigg) \\
\le & nP\Bigg\{ \sup_{g\in\mathcal{H}_{\mathcal{D},\mathcal{W},\mathcal{B}}} \Bigg(\frac{1}{n}\sum_{k=2}^{n+1} I(Y_k \ge Y_1) h\{g(\boldsymbol{X}_k) - g(\boldsymbol{X}_1)\} \\
&\qquad\qquad - E[I(Y \ge Y_1) h\{g(\boldsymbol{X}) - g(\boldsymbol{X}_1)\} \mid \boldsymbol{X}_1, Y_1] \Bigg) > t - \frac{2||h||_\infty}{n} \Bigg\} \\
&+ nP\Bigg\{ \sup_{g\in\mathcal{H}_{\mathcal{D},\mathcal{W},\mathcal{B}}} \Bigg(-\frac{1}{n}\sum_{k=2}^{n+1} I(Y_k \ge Y_1) h\{g(\boldsymbol{X}_k) - g(\boldsymbol{X}_1)\} \\
&\qquad\qquad + E[I(Y \ge Y_1) h\{g(\boldsymbol{X}) - g(\boldsymbol{X}_1)\} \mid \boldsymbol{X}_1, Y_1] \Bigg) > t - \frac{2||h||_\infty}{n} \Bigg\} \\
\le & nE\Bigg[P\Bigg\{ \sup_{g\in\mathcal{H}_{\mathcal{D},\mathcal{W},\mathcal{B}}} \Bigg(\frac{1}{n}\sum_{k=2}^{n+1} I(Y_k \ge Y_1) h\{g(\boldsymbol{X}_k) - g(\boldsymbol{X}_1)\} \\
&\qquad\qquad - E[I(Y \ge Y_1) h\{g(\boldsymbol{X}) - g(\boldsymbol{X}_1)\} \mid \boldsymbol{X}_1, Y_1] \Bigg) > t - \frac{2||h||_\infty}{n} \;\Bigg|\; \boldsymbol{X}_1, Y_1 \Bigg\} \Bigg] \\
&+ nE\Bigg[P\Bigg\{ \sup_{g\in\mathcal{H}_{\mathcal{D},\mathcal{W},\mathcal{B}}} \Bigg(-\frac{1}{n}\sum_{k=2}^{n+1} I(Y_k \ge Y_1) h\{g(\boldsymbol{X}_k) - g(\boldsymbol{X}_1)\} \\
&\qquad\qquad + E[I(Y \ge Y_1) h\{g(\boldsymbol{X}) - g(\boldsymbol{X}_1)\} \mid \boldsymbol{X}_1, Y_1] \Bigg) > t - \frac{2||h||_\infty}{n} \;\Bigg|\; \boldsymbol{X}_1, Y_1 \Bigg\} \Bigg].
\end{aligned} \tag{18}
$$

For any $\boldsymbol{x} \in [-a_1, a_1]^p$, $y \in [0, \tau]$ and any $g \in \mathcal{H}_{\mathcal{D},\mathcal{W},\mathcal{B}}$, define

$$
r_{1,g}(\boldsymbol{x}, y) = I(y \ge Y_1) h\{g(\boldsymbol{x}) - g(\boldsymbol{X}_1)\}, \quad \text{and} \quad \mathcal{R}_1 = \{r_{1,g} \mid g \in \mathcal{H}_{\mathcal{D},\mathcal{W},\mathcal{B}}\}.
$$

Let

$$
\begin{aligned}
l_1 &= 2\sqrt{80||h||_\infty(1+4||h||_\infty)^2/3}\sqrt{(1+||h||_\infty)(9+a_4+a_5)}, \\
f_1(\mathcal{D}, \mathcal{W}, n) &= \sqrt{\psi_1(\mathcal{W}, \mathcal{D}) \log n / n},
\end{aligned}
$$

and

$$t = l_1 f_1(\mathcal{D}, \mathcal{W}, n).$$

It is easy to show $t > 2||h||_\infty/n$. It can be checked that

$$\begin{aligned}
&P\Bigg\{ \sup_{g\in\mathcal{H}_{\mathcal{D},\mathcal{W},\mathcal{B}}} \Bigg(\frac{1}{n}\sum_{k=2}^{n+1} I(Y_k \geq Y_1)h\{g(\boldsymbol{X}_k) - g(\boldsymbol{X}_1)\} \\
&\qquad\qquad - E[I(Y \geq Y_1)h\{g(\boldsymbol{X}) - g(\boldsymbol{X}_1)\} \mid \boldsymbol{X}_1, Y_1]\Bigg) > t - \frac{2||h||_\infty}{n} \;\Bigg|\; \boldsymbol{X}_1, Y_1\Bigg\} \\
=&P\Bigg(\sup_{g\in\mathcal{H}_{\mathcal{D},\mathcal{W},\mathcal{B}}} P_n' r_{1,g} - P r_{1,g} > t - \frac{2||h||_\infty}{n} \;\Bigg|\; \boldsymbol{X}_1, Y_1\Bigg) \\
=&P\Bigg\{ \sup_{g\in\mathcal{H}_{\mathcal{D},\mathcal{W},\mathcal{B}}} P_n'(r_{1,g} + ||h||_\infty) - (P r_{1,g} + ||h||_\infty) > t - \frac{2||h||_\infty}{n} \;\Bigg|\; \boldsymbol{X}_1, Y_1\Bigg\} \\
\leq&P\Bigg\{ \sup_{g\in\mathcal{H}_{\mathcal{D},\mathcal{W},\mathcal{B}}} \frac{P_n'(r_{1,g} + ||h||_\infty) - (P r_{1,g} + ||h||_\infty)}{1 + P_n'(r_{1,g} + ||h||_\infty) + (P r_{1,g} + ||h||_\infty)} > \frac{t - 2||h||_\infty/n}{1 + 4||h||_\infty} \;\Bigg|\; \boldsymbol{X}_1, Y_1\Bigg\}. \qquad (19)
\end{aligned}$$

If $(t - 2||h||_\infty/n)/(1 + 4||h||_\infty) < 1$ and $(t - 2||h||_\infty/n)/(5||h||_\infty + 20||h||_\infty^2) < 1$, by using (19), Theorem 11.6 in Györfi et al. (2002) and condition (C3), we have

$$\begin{aligned}
&nE\Bigg[P\Bigg\{ \sup_{g\in\mathcal{H}_{\mathcal{D},\mathcal{W},\mathcal{B}}} \Bigg(\frac{1}{n}\sum_{k=2}^{n+1} I(Y_k \geq Y_1)h\{g(\boldsymbol{X}_k) - g(\boldsymbol{X}_1)\} \\
&\qquad\qquad - E[I(Y \geq Y_1)h\{g(\boldsymbol{X}) - g(\boldsymbol{X}_1)\} \mid \boldsymbol{X}_1, Y_1]\Bigg) > t - \frac{2||h||_\infty}{n} \;\Bigg|\; \boldsymbol{X}_1, Y_1\Bigg\}\Bigg] \\
\leq&nE\Bigg[P\Bigg\{ \sup_{g\in\mathcal{H}_{\mathcal{D},\mathcal{W},\mathcal{B}}} \frac{P_n'(r_{1,g} + ||h||_\infty) - (P r_{1,g} + ||h||_\infty)}{1 + P_n'(r_{1,g} + ||h||_\infty) + (P r_{1,g} + ||h||_\infty)} > \frac{t - 2||h||_\infty/n}{1 + 4||h||_\infty} \;\Bigg|\; \boldsymbol{X}_1, Y_1\Bigg\}\Bigg] \\
\leq&4nE\{PN((t - 2||h||_\infty/n)/(5 + 20||h||_\infty), \mathcal{R}_1, L_1(P_n'))\} \exp\Bigg\{ - \frac{3(t - 2||h||_\infty/n)^2 n}{80||h||_\infty(1 + 4||h||_\infty)^2}\Bigg\} \\
\leq&4na_4\Psi_1(\mathcal{W},\mathcal{D})(16ea_5)^{\Psi_1(\mathcal{W},\mathcal{D})}\left(\frac{5||h||_\infty + 20||h||_\infty^2}{t - 2||h||_\infty/n}\right)^{\Psi_1(\mathcal{W},\mathcal{D})-1} \\
&\times \exp\Bigg\{ - \frac{3(t - 2||h||_\infty/n)^2 n}{80||h||_\infty(1 + 4||h||_\infty)^2}\Bigg\} \\
\leq&\frac{1}{n}. \qquad (20)
\end{aligned}$$

When $(t-2||h||_\infty/n)/(1+4||h||_\infty) < 1$ and $(t-2||h||_\infty/n)/(5||h||_\infty+20||h||_\infty^2) \geq 1$, according to (19) and Theorem 11.6 in Györfi et al. (2002), one can show

$$\begin{aligned}
&nE\bigg[P\bigg\{\sup_{g\in\mathcal{H}_{\mathcal{D},\mathcal{W},\mathcal{B}}}\bigg(\frac{1}{n}\sum_{k=2}^{n+1}I(Y_k\geq Y_1)h\{g(\boldsymbol{X}_k)-g(\boldsymbol{X}_1)\}\\
&\qquad\qquad -E[I(Y\geq Y_1)h\{g(\boldsymbol{X})-g(\boldsymbol{X}_1)\}\mid \boldsymbol{X}_1,Y_1]\bigg)>t-\frac{2||h||_\infty}{n}\;\bigg|\;\boldsymbol{X}_1,Y_1\bigg\}\bigg]\\
\leq& nE\bigg[P\bigg\{\sup_{g\in\mathcal{H}_{\mathcal{D},\mathcal{W},\mathcal{B}}}\frac{P_n'(r_{1,g}+||h||_\infty)-(Pr_{1,g}+||h||_\infty)}{1+P_n'(r_{1,g}+||h||_\infty)+(Pr_{1,g}+||h||_\infty)}>\frac{t-2||h||_\infty/n}{1+4||h||_\infty}\;\bigg|\;\boldsymbol{X}_1,Y_1\bigg\}\bigg]\\
\leq& 4nE\{PN((t-2||h||_\infty/n)/(5+20||h||_\infty),\mathcal{R}_1,L_1(P_n'))\}\exp\bigg\{-\frac{3(t-2||h||_\infty/n)^2n}{80||h||_\infty(1+4||h||_\infty)^2}\bigg\}\\
=&4n\exp\bigg\{-\frac{3(t-2||h||_\infty/n)^2n}{80||h||_\infty(1+4||h||_\infty)^2}\bigg\}\\
\leq&\frac{1}{n}.
\end{aligned}\tag{21}$$

In the case that $(t-2||h||_\infty/n)/(1+4||h||_\infty)\geq 1$, based on (19), we can obtain

$$\begin{aligned}
&nE\bigg[P\bigg\{\sup_{g\in\mathcal{H}_{\mathcal{D},\mathcal{W},\mathcal{B}}}\bigg(\frac{1}{n}\sum_{k=2}^{n+1}I(Y_k\geq Y_1)h\{g(\boldsymbol{X}_k)-g(\boldsymbol{X}_1)\}\\
&\qquad\qquad -E[I(Y\geq Y_1)h\{g(\boldsymbol{X})-g(\boldsymbol{X}_1)\}\mid \boldsymbol{X}_1,Y_1]\bigg)>t-\frac{2||h||_\infty}{n}\;\bigg|\;\boldsymbol{X}_1,Y_1\bigg\}\bigg]\\
\leq& nE\bigg[P\bigg\{\sup_{g\in\mathcal{H}_{\mathcal{D},\mathcal{W},\mathcal{B}}}\frac{P_n'(r_{1,g}+||h||_\infty)-(Pr_{1,g}+||h||_\infty)}{1+P_n'(r_{1,g}+||h||_\infty)+(Pr_{1,g}+||h||_\infty)}>\frac{t-2||h||_\infty/n}{1+4||h||_\infty}\;\bigg|\;\boldsymbol{X}_1,Y_1\bigg\}\bigg]\\
\leq& nE\bigg[P\bigg\{\sup_{g\in\mathcal{H}_{\mathcal{D},\mathcal{W},\mathcal{B}}}\frac{P_n'(r_{1,g}+||h||_\infty)-(Pr_{1,g}+||h||_\infty)}{1+P_n'(r_{1,g}+||h||_\infty)+(Pr_{1,g}+||h||_\infty)}>1\;\bigg|\;\boldsymbol{X}_1,Y_1\bigg\}\bigg]\\
=&0.
\end{aligned}\tag{22}$$

(20)-(22) indicate

$$\begin{aligned}
&nE\bigg[P\bigg\{\sup_{g\in\mathcal{H}_{\mathcal{D},\mathcal{W},\mathcal{B}}}\bigg(\frac{1}{n}\sum_{k=2}^{n+1}I(Y_k\geq Y_1)h\{g(\boldsymbol{X}_k)-g(\boldsymbol{X}_1)\}\\
&\qquad -E[I(Y\geq Y_1)h\{g(\boldsymbol{X})-g(\boldsymbol{X}_1)\}\mid \boldsymbol{X}_1,Y_1]\bigg)>l_1f_1(\mathcal{D},\mathcal{W},n)-\frac{2||h||_\infty}{n}\;\bigg|\;\boldsymbol{X}_1,Y_1\bigg\}\bigg]\\
\leq&\frac{1}{n}.
\end{aligned}\tag{23}$$

Similarly, one can prove

$$\begin{aligned}
&nE\bigg[P\bigg\{\sup_{g\in\mathcal{H}_{\mathcal{D},\mathcal{W},\mathcal{B}}}\bigg(-\frac{1}{n}\sum_{k=2}^{n+1}I(Y_k\geq Y_1)h\{g(\boldsymbol{X}_k)-g(\boldsymbol{X}_1)\}\\
&\qquad +E[I(Y\geq Y_1)h\{g(\boldsymbol{X})-g(\boldsymbol{X}_1)\}\mid \boldsymbol{X}_1,Y_1]\bigg)>l_1f_1(\mathcal{D},\mathcal{W},n)-\frac{2||h||_\infty}{n}\;\bigg|\;\boldsymbol{X}_1,Y_1\bigg\}\bigg]\\
\leq&\frac{1}{n}.
\end{aligned}\tag{24}$$

Based on (18), (23) and (24), we have that with probability at least $1-2/n$,

$$\sup_{g\in\mathcal{H}_{\mathcal{D},\mathcal{W},\mathcal{B}}}\max_{1\le i\le n}\left|\frac{1}{n}\sum_{j=1}^{n}I(Y_j\ge Y_i)\bigg(\frac{1}{n}\sum_{k=1}^{n}I(Y_k\ge Y_i)h\{g(\boldsymbol{X}_k)-g(\boldsymbol{X}_i)\} - E[I(Y\ge Y_i)h\{g(\boldsymbol{X})-g(\boldsymbol{X}_i)\}\mid \boldsymbol{X}_i,Y_i]\bigg)\right|\le l_1 f_1(\mathcal{D},\mathcal{W},n). \tag{25}$$

Using (15), (16), (17) and (25), we can show that with probability at least $1-6/n$,

$$PR'_n(\check{g})-PR_n(\check{g})+PR_n(g_0)-PR'_n(g_0)\le 4||\dot{f}||_\infty\{2||h||_\infty(\log n/n)^{1/2}+l_1f_1(\mathcal{D},\mathcal{W},n)\}/a_2^2. \tag{26}$$

Now we provide an upper bound for

$$PR(\check{g})-PR(g^*)-\{PR'_n(\check{g})-PR'_n(g^*)\}.$$

For any $\boldsymbol{x}\in[-a_1,a_1]^p$, any $y\in[0,\tau]$, and any $\delta\in\{0,1\}$, define $r_{2,g}(\boldsymbol{x}^\top,y,\delta)=\delta\{f_{g^*}(\boldsymbol{z})-f_g(\boldsymbol{z})\}$, and $\mathcal{R}_2=\{r_{2,g}\mid g\in\mathcal{H}_{\mathcal{D},\mathcal{W},\mathcal{B}}\}$. It can be shown that

$$\begin{aligned}
&P(PR(\check{g})-PR(g^*)-\{PR'_n(\check{g})-PR'_n(g^*)\}>t)\\
\le&P\bigg(\sup_{g\in\mathcal{H}_{\mathcal{D},\mathcal{W},\mathcal{B}}}[PR(g)-PR(g^*)-\{PR'_n(g)-PR'_n(g^*)\}]>t\bigg)\\
=&P\bigg\{\sup_{g\in\mathcal{H}_{\mathcal{D},\mathcal{W},\mathcal{B}}}(P_nr_g-Pr_g)>t\bigg\}\\
=&P\bigg[\sup_{g\in\mathcal{H}_{\mathcal{D},\mathcal{W},\mathcal{B}}}\{P_n(r_g+2||f||_\infty)-P(r_g+2||f||_\infty)\}>t\bigg]\\
\le&P\bigg\{\sup_{g\in\mathcal{H}_{\mathcal{D},\mathcal{W},\mathcal{B}}}\frac{P_n(r_g+2||f||_\infty)-P(r_g+2||f||_\infty)}{1+P_n(r_g+2||f||_\infty)+P(r_g+2||f||_\infty)}>\frac{t}{1+8||f||_\infty}\bigg\}.
\end{aligned} \tag{27}$$

Define

$$\begin{aligned}
&l_2=\sqrt{160||f||_\infty(1+8||f||_\infty)^2/3}\sqrt{(1+||f||_\infty)(8+a_4+a_5)},\\
&f_2(\mathcal{D},\mathcal{W},n)=\sqrt{\psi_2(\mathcal{W},\mathcal{D})\log n/n},
\end{aligned}$$

and

$$t=l_2f_2(\mathcal{D},\mathcal{W},n).$$

If $t/(1+8||f||_\infty) < 1$ and $t/(5||f||_\infty + 40||f||_\infty^2) < 1$, by using (27), Theorem 11.6 in Györfi et al. (2002) and condition (C3), we have

$$\begin{aligned}
&P(PR(\check{g}) - PR(g^*) - \{PR'_n(\check{g}) - PR'_n(g^*)\} > t) \\
\leq& P\left\{ \sup_{g\in\mathcal{H}_{\mathcal{D},\mathcal{W},\mathcal{B}}} \frac{P_n(r_g + 2||f||_\infty) - P(r_g + 2||f||_\infty)}{1 + P_n(r_g + 2||f||_\infty) + P(r_g + 2||f||_\infty)} > \frac{t}{1+8||f||_\infty} \right\} \\
\leq& 4EN(t/(5+40||f||_\infty), \mathcal{R}_2, L_1(P_n)) \exp\left\{ -\frac{3t^2 n}{160||f||_\infty(1+8||f||_\infty)^2} \right\} \\
\leq& 4EN(t/(5+40||f||_\infty), \mathcal{G}_1, L_1(P_n)) \exp\left\{ -\frac{3t^2 n}{160||f||_\infty(1+8||f||_\infty)^2} \right\} \\
\leq& 4a_4\Psi_2(\mathcal{W},\mathcal{D})(16ea_5)^{\Psi_2(\mathcal{W},\mathcal{D})} \left(\frac{5||f||_\infty + 40||f||_\infty^2}{t}\right)^{\Psi_2(\mathcal{W},\mathcal{D})-1} \exp\left\{ -\frac{3t^2 n}{160||f||_\infty(1+8||f||_\infty)^2} \right\} \\
\leq& \frac{1}{n}.
\end{aligned} \tag{28}$$

When $t/(1+8||f||_\infty) < 1$ and $t/(5||f||_\infty + 40||f||_\infty^2) \geq 1$, according to (27) and Theorem 11.6 in Györfi et al. (2002), one can show

$$\begin{aligned}
&P(PR(\check{g}) - PR(g^*) - \{PR'_n(\check{g}) - PR'_n(g^*)\} > t) \\
\leq& P\left\{ \sup_{g\in\mathcal{H}_{\mathcal{D},\mathcal{W},\mathcal{B}}} \frac{P_n(r_g + 2||f||_\infty) - P(r_g + 2||f||_\infty)}{1 + P_n(r_g + 2||f||_\infty) + P(r_g + 2||f||_\infty)} > \frac{t}{1+8||f||_\infty} \right\} \\
\leq& 4EN(t/(5+40||f||_\infty), \mathcal{R}_2, L_1(P_n)) \exp\left\{ -\frac{3t^2 n}{160||f||_\infty(1+8||f||_\infty)^2} \right\} \\
\leq& 4EN(t/(5+40||f||_\infty), \mathcal{G}_1, L_1(P_n)) \exp\left\{ -\frac{3t^2 n}{160||f||_\infty(1+8||f||_\infty)^2} \right\} \\
=& 4\exp\left\{ -\frac{3t^2 n}{160||f||_\infty(1+8||f||_\infty)^2} \right\} \\
\leq& \frac{1}{n}.
\end{aligned} \tag{29}$$

In the case that $t/(1+8||f||_\infty) \geq 1$, based on (27), we can obtain

$$\begin{aligned}
&P(PR(\check{g}) - PR(g^*) - \{PR'_n(\check{g}) - PR'_n(g^*)\} > t) \\
\leq& P\left\{ \sup_{g\in\mathcal{H}_{\mathcal{D},\mathcal{W},\mathcal{B}}} \frac{P_n(r_g + 2||f||_\infty) - P(r_g + 2||f||_\infty)}{1 + P_n(r_g + 2||f||_\infty) + P(r_g + 2||f||_\infty)} > \frac{t}{1+8||f||_\infty} \right\} \\
\leq& P\left\{ \sup_{g\in\mathcal{H}_{\mathcal{D},\mathcal{W},\mathcal{B}}} \frac{P_n(r_g + 2||f||_\infty) - P(r_g + 2||f||_\infty)}{1 + P_n(r_g + 2||f||_\infty) + P(r_g + 2||f||_\infty)} > 1 \right\} \\
=& 0.
\end{aligned} \tag{30}$$

According to (28)-(30), one can prove that with probability at least $1 - 1/n$,

$$PR(\check{g}) - PR(g^*) - \{PR'_n(\check{g}) - PR'_n(g^*)\} \leq l_2 f_2(\mathcal{D}, \mathcal{W}, n). \tag{31}$$

Similar to (31), we have that with probability at least $1-1/n$,

$$\{PR'_n(g_0) - PR'_n(g^*)\} - \{PR(g_0) - PR(g^*)\} \leq l_2 f_2(\mathcal{D}, \mathcal{W}, n). \tag{32}$$

Using conditions (C2), (C4) and (C5), and Theorem 2.5 in Shen et al. (2022), one can show

$$PR(g_0) - PR(g^*) \leq a_6 ||g_0 - g^*||_\infty \leq 262 a_1 a_3 a_6 \sqrt{p} \{N^2 L^2 \log_3(N+2)\}^{-1/p}. \tag{33}$$

According to (14), (26), (31)-(33) and condition (C1), we can obtain that with probability at least $1-8/n$,

$$\begin{aligned} PR(\widehat{g}) - PR(g^*) =& PR(\check{g}) - PR(g^*) \\ \leq& 4||\dot{f}||_\infty \{2||h||_\infty (\log n/n)^{1/2} + l_1 f_1(\mathcal{D}, \mathcal{W}, n)\}/a_2^2 + 2 l_2 f_2(\mathcal{D}, \mathcal{W}, n) \\ &+ 262 a_1 a_3 a_6 \sqrt{p} \{N^2 L^2 \log_3(N+2)\}^{-1/p}. \end{aligned}$$

The proof of Theorem 1 is completed. ■

A.3 Proof of Corollary 2

Using the fact that $f(t) = t$, $h(t) = [2+t]_+$ and $\sup_{g \in \mathcal{H}_{\mathcal{D},\mathcal{W},\mathcal{B}}} ||g||_\infty \leq \mathcal{B}$, we have

$$||\dot{f}||_\infty = 1, \quad ||f||_\infty = 2 + 2\mathcal{B}, \quad \text{and} \quad ||h||_\infty = 2 + 2\mathcal{B}. \tag{34}$$

We first verify that condition (C3) is satisfied. For any $t_1, t_2 \in \mathbb{R}$, we have

$$|[t_1]_+ - [t_2]_+| \leq |t_1 - t_2|. \tag{35}$$

For any $\epsilon > 0$ and $\boldsymbol{x}' \in [-a_1, a_1]^p$, based on (8), (10) and (35), it can be shown that

$$N(\epsilon||h||_\infty, \mathcal{G}_{x'}, L_1(P_n)) \leq c_1 c_2 \mathcal{D}\mathcal{S}_1 \log \mathcal{S}_1 (16e)^{c_2 \mathcal{D}\mathcal{S}_1 \log \mathcal{S}_1} \left(\frac{1}{\epsilon}\right)^{c_2 \mathcal{D}\mathcal{S}_1 \log \mathcal{S}_1 - 1},$$

and

$$N(\epsilon||f||_\infty, \mathcal{G}_1, L_1(P_n)) \leq c_1 c_2 \mathcal{D}\mathcal{S}_2 \log \mathcal{S}_2 \left(\frac{16e}{a_2}\right)^{c_2 \mathcal{D}\mathcal{S}_2 \log \mathcal{S}_2} \left(\frac{1}{\epsilon}\right)^{c_2 \mathcal{D}\mathcal{S}_2 \log \mathcal{S}_2 - 1}. \tag{36}$$

(36) implies that condition (C3) is satisfied with $a_4 = c_1$ and $a_5 = 1/a_2$. It follows from (34) and the expressions of l_1 and l_2, we can obtain

$$l_1 = 2\sqrt{160(1+\mathcal{B})(9+8\mathcal{B})^2/3}\sqrt{(3+2\mathcal{B})(9+c_1+1/a_2)},$$

and

$$l_2 = \sqrt{320(1+\mathcal{B})(17+16\mathcal{B})^2/3}\sqrt{(3+2\mathcal{B})(8+c_1+1/a_2)}. \tag{37}$$

Now we show that condition (C5) is satisfied. By condition (C1), (35) and the triangle inequality, we have

$$\begin{aligned}
&PR(g_0) - PR(g^*)\\
=&E\left[\Delta\frac{E\{I(\tilde{Y}\geq Y)[g_0(\tilde{\boldsymbol{X}})-g_0(\boldsymbol{X})+2]_+ \mid \boldsymbol{X},Y\}}{E\{I(\tilde{Y}\geq Y)\mid Y\}}\right]\\
&-E\left[\Delta\frac{E\{I(\tilde{Y}\geq Y)[g^*(\tilde{\boldsymbol{X}})-g^*(\boldsymbol{X})+2]_+ \mid \boldsymbol{X},Y\}}{E\{I(\tilde{Y}\geq Y)\mid Y\}}\right]\\
\leq&E\left[\Delta\frac{E\{I(\tilde{Y}\geq Y)\mid g_0(\tilde{\boldsymbol{X}})-g_0(\boldsymbol{X})-g^*(\tilde{\boldsymbol{X}})+g^*(\boldsymbol{X})|\mid \boldsymbol{X},Y\}}{E\{I(\tilde{Y}\geq Y)\mid Y\}}\right]\\
\leq&\frac{2}{a_2}||g_0-g^*||_\infty. \qquad (38)
\end{aligned}$$

(38) indicates that condition (C5) is satisfied. Using (34), (36)-(38) and Theorem 1, we can prove that with probability at least $1-8/n$,

$$\begin{aligned}
PR(\widehat{g})-PR(g^*)\leq&4\{(4+4\mathcal{B})(\log n/n)^{1/2}+l_1f_1(\mathcal{D},\mathcal{W},n)\}/a_2^2+2l_2f_2(\mathcal{D},\mathcal{W},n)\\
&+524a_1a_3\sqrt{p}\{N^2L^2\log_3(N+2)\}^{-1/p}/a_2,
\end{aligned}$$

where

$$f_1(\mathcal{D},\mathcal{W},n)=\sqrt{c_2\mathcal{D}\mathcal{S}_1\log\mathcal{S}_1\log n/n},\quad\text{and}\quad f_2(\mathcal{D},\mathcal{W},n)=\sqrt{c_2\mathcal{D}\mathcal{S}_2\log\mathcal{S}_2\log n/n}.$$

We finish the proof of Corollary 2. ■

A.4 Proof of Corollary 3

Based on the fact that $f(t)=t$ and $h(t)=I(t>0)$, we have

$$||\dot{f}||_\infty=1,\quad ||f||_\infty=1,\quad\text{and}\quad ||h||_\infty=1. \qquad (39)$$

Now we verify that condition (C3) is satisfied. According to (11), (13) and (39), one can prove

$$\begin{aligned}
&N(\epsilon||h||_\infty,\mathcal{G}_{x'},L_1(P_n))\\
\leq&c_1c_3\mathcal{S}_1p(2\mathcal{W}\mathcal{D}+1)(16e)^{c_3\mathcal{S}_1p(2\mathcal{W}\mathcal{D}+1)}\left(\frac{1}{\epsilon}\right)^{c_3\mathcal{S}_1p(2\mathcal{W}\mathcal{D}+1)-1},
\end{aligned}$$

and

$$\begin{aligned}
&N(\epsilon||f||_\infty,\mathcal{G}_1,L_1(P_n))\\
\leq&2c_1c_3\mathcal{S}_2p(2\mathcal{W}\mathcal{D}+1)\left(\frac{16e}{a_2}\right)^{2c_3\mathcal{S}_2p(2\mathcal{W}\mathcal{D}+1)}\left(\frac{1}{\epsilon}\right)^{2c_3\mathcal{S}_2p(2\mathcal{W}\mathcal{D}+1)-1}. \qquad (40)
\end{aligned}$$

(40) implies that condition (C2) is satisfied. By (39) and the expressions of l_1 and l_2, we can obtain

$$l_1=40\sqrt{10(9+c_1+1/a_2)/3},\quad\text{and}\quad l_2=24\sqrt{15(8+c_1+1/a_2)}. \qquad (41)$$

Using condition (C1) and the triangle inequality, we have

$$\begin{aligned}
&PR(g_0) - PR(g^*) \\
=&E\left[\Delta \frac{E\{I(\tilde{Y} \geq Y) I\{g_0(\tilde{\boldsymbol{X}}) > g_0(\boldsymbol{X})\} \mid \boldsymbol{X}, Y\}}{E\{I(\tilde{Y} \geq Y) \mid Y\}}\right] \\
&- E\left[\Delta \frac{E\{I(\tilde{Y} \geq Y) I\{g^*(\tilde{\boldsymbol{X}}) > g^*(\boldsymbol{X})\} \mid \boldsymbol{X}, Y\}}{E\{I(\tilde{Y} \geq Y) \mid Y\}}\right] \\
\leq&\frac{1}{a_2} P\{g_0(\tilde{\boldsymbol{X}}) > g_0(\boldsymbol{X}), g^*(\tilde{\boldsymbol{X}}) \leq g^*(\boldsymbol{X})\} \\
\leq&\frac{1}{a_2} P\{g_0(\boldsymbol{X}) - ||g_0 - g^*||_\infty < g^*(\tilde{\boldsymbol{X}}) \leq g_0(\boldsymbol{X}) + ||g_0 - g^*||_\infty\} \\
\leq&\frac{2c_4}{a_2} ||g_0 - g^*||_\infty, \qquad (42)
\end{aligned}$$

where c_4 is the maximum value of the density function of $g^*(\boldsymbol{X})$. Using (39)-(42) and Theorem 1, we can prove that with probability at least $1 - 8/n$,

$$\begin{aligned}
PR(\widehat{g}) - PR(g^*) \leq&4\{2(\log n/n)^{1/2} + l_1 f_1(\mathcal{D}, \mathcal{W}, n)\}/a_2^2 + 2l_2 f_2(\mathcal{D}, \mathcal{W}, n) \\
&+ 524 a_1 a_3 c_4 \sqrt{p}\{N^2 L^2 \log_3(N+2)\}^{-1/p}/a_2,
\end{aligned}$$

where

$$f_1(\mathcal{D}, \mathcal{W}, n) = \sqrt{c_3 \mathcal{S}_1 p(2\mathcal{W}\mathcal{D} + 1) \log n/n},$$

and

$$f_2(\mathcal{D}, \mathcal{W}, n) = \sqrt{2c_3 \mathcal{S}_2 p(2\mathcal{W}\mathcal{D} + 1) \log n/n}.$$

The proof of Corollary 3 is completed. ■

A.5 Proof of Corollary 4

By the fact that $f(t) = \log(t)$, $h(t) = \exp(t)$ and $\sup_{g \in \mathcal{H}_{\mathcal{D},\mathcal{W},\mathcal{B}}} ||g||_\infty \leq \mathcal{B}$, we have

$$||\dot{f}||_\infty = e^{2\mathcal{B}}, \quad ||f||_\infty = 2\mathcal{B}, \quad \text{and} \quad ||h||_\infty = e^{2\mathcal{B}}. \qquad (43)$$

Mimicking the arguments of (36) and (38), we can obtain that for any $0 < \epsilon < 1$,

$$\begin{aligned}
N(\epsilon||h||_\infty, \mathcal{G}_{x'}, L_1(P_n)) \leq& c_1 c_2 \mathcal{D}\mathcal{S}_1 \log \mathcal{S}_1 (16 e^{2\mathcal{B}+1})^{c_2 \mathcal{D}\mathcal{S}_1 \log \mathcal{S}_1} \left(\frac{1}{\epsilon}\right)^{c_2 \mathcal{D}\mathcal{S}_1 \log \mathcal{S}_1 - 1}, \\
N(\epsilon||f||_\infty, \mathcal{G}_1, L_1(P_n)) \leq& c_1 c_2 \mathcal{D}\mathcal{S}_2 \log \mathcal{S}_2 \left(\frac{16 e^{4\mathcal{B}+1}}{a_2}\right)^{c_2 \mathcal{D}\mathcal{S}_2 \log \mathcal{S}_2} \left(\frac{1}{\epsilon}\right)^{c_2 \mathcal{D}\mathcal{S}_2 \log \mathcal{S}_2 - 1},
\end{aligned}$$

and

$$\begin{aligned}
&PR(g_0) - PR(g^*)\\
=&E\bigg(\Delta \log\bigg[\frac{E\{I(\tilde{Y}\geq Y)\exp\{g_0(\tilde{\boldsymbol{X}})-g_0(\boldsymbol{X})\}\mid \boldsymbol{X},Y\}}{E\{I(\tilde{Y}\geq Y)\mid Y\}}\bigg]\bigg)\\
&-E\bigg(\Delta \log\bigg[\frac{E\{I(\tilde{Y}\geq Y)\exp\{g^*(\tilde{\boldsymbol{X}})-g^*(\boldsymbol{X})\}\mid \boldsymbol{X},Y\}}{E\{I(\tilde{Y}\geq Y)\mid Y\}}\bigg]\bigg)\\
\leq&\frac{e^{4\mathcal{B}}}{a_2}E\bigg[\Delta E\{I(\tilde{Y}\geq Y)\mid g_0(\tilde{\boldsymbol{X}})-g_0(\boldsymbol{X})-g^*(\tilde{\boldsymbol{X}})+g^*(\boldsymbol{X})|\mid \boldsymbol{X},Y\}\bigg]\\
\leq&\frac{2e^{4\mathcal{B}}}{a_2}||g_0-g^*||_\infty. \qquad (44)
\end{aligned}$$

By (43), (44) and the expressions of l_1 and l_2, we can obtain

$$l_1 = 2\sqrt{80e^{2\mathcal{B}}(1+4e^{2\mathcal{B}})^2/3}\sqrt{(1+e^{2\mathcal{B}})(9+c_1+e^{4\mathcal{B}}/a_2)}, \qquad (45)$$

and

$$l_2 = \sqrt{320\mathcal{B}(1+16\mathcal{B})^2/3}\sqrt{(1+2\mathcal{B})(8+c_1+e^{4\mathcal{B}}/a_2)}. \qquad (46)$$

According to (43), (44) and Theorem 1, one can prove that with probability at least $1-8/n$,

$$\begin{aligned}
PR(\widehat{g})-PR(g^*)\leq&4e^{2\mathcal{B}}\{2e^{2\mathcal{B}}(\log n/n)^{1/2}+l_1f_1(\mathcal{D},\mathcal{W},n)\}/a_2^2+2l_2f_2(\mathcal{D},\mathcal{W},n)\\
&+524a_1a_3\sqrt{p}e^{4\mathcal{B}}\{N^2L^2\log_3(N+2)\}^{-1/p}/a_2, \qquad (47)
\end{aligned}$$

where

$$f_1(\mathcal{D},\mathcal{W},n)=\sqrt{c_2\mathcal{D}\mathcal{S}_1\log\mathcal{S}_1\log n/n},\quad \text{and}\quad f_2(\mathcal{D},\mathcal{W},n)=\sqrt{c_2\mathcal{D}\mathcal{S}_2\log\mathcal{S}_2\log n/n}.$$

Based on condition (D1) and the second-order Taylor expansion of $PR(\widehat{g})$ at g^*, one can show

$$\begin{aligned}
PR(\widehat{g})-PR(g^*)\geq&\frac{a_2E(\Delta)}{2e^{4\mathcal{B}}}E[\{\widehat{g}(\boldsymbol{X})-g^*(\boldsymbol{X})-P\widehat{g}\}^2\mid \boldsymbol{O}_1,\ldots,\boldsymbol{O}_n]\\
=&\frac{a_2E(\Delta)}{2e^{4\mathcal{B}}}E[\{\widehat{g}(\boldsymbol{X})-g^*(\boldsymbol{X})+P_n\widehat{g}-P\widehat{g}\}^2\mid \boldsymbol{O}_1,\ldots,\boldsymbol{O}_n]\\
\geq&\frac{a_2E(\Delta)}{2e^{4\mathcal{B}}}\big\{E[\{\widehat{g}(\boldsymbol{X})-g^*(\boldsymbol{X})\}^2|\boldsymbol{O}_1,\ldots,\boldsymbol{O}_n]+(P_n\widehat{g}-P\widehat{g})^2\\
&-2|P_n\widehat{g}-P\widehat{g}|\big(E[\{\widehat{g}(\boldsymbol{X})-g^*(\boldsymbol{X})\}^2\mid \boldsymbol{O}_1,\ldots,\boldsymbol{O}_n]\big)^{1/2}\big\}\\
=&\frac{a_2E(\Delta)}{2e^{4\mathcal{B}}}\big\{\big(E[\{\widehat{g}(\boldsymbol{X})-g^*(\boldsymbol{X})\}^2\mid \boldsymbol{O}_1,\ldots,\boldsymbol{O}_n]\big)^{1/2}-|P_n\widehat{g}-P\widehat{g}|\big\}^2.
\end{aligned}$$

This implies

$$E[\{\widehat{g}(\boldsymbol{X})-g^*(\boldsymbol{X})\}^2\mid \boldsymbol{O}_1,\ldots,\boldsymbol{O}_n]\leq 2(P_n\widehat{g}-P\widehat{g})^2+\{PR(\widehat{g})-PR(g^*)\}\frac{4e^{4\mathcal{B}}}{a_2E(\Delta)}. \qquad (48)$$

It can be shown that

$$\begin{aligned} P(P_n\widehat{g} - P\widehat{g} > t) \leq & P\bigg(\sup_{g\in\mathcal{H}_{\mathcal{D},\mathcal{W},\mathcal{B}}} P_n g - Pg > t \bigg) \\ = & P\bigg(\sup_{g\in\mathcal{H}_{\mathcal{D},\mathcal{W},\mathcal{B}}} P_n(g+\mathcal{B}) - P(g+\mathcal{B}) > t \bigg) \\ \leq & P\bigg(\sup_{g\in\mathcal{H}_{\mathcal{D},\mathcal{W},\mathcal{B}}} \frac{P_n(g+\mathcal{B}) - P(g+\mathcal{B})}{1+P_n(g+\mathcal{B}) + P(g+\mathcal{B})} > \frac{t}{1+4\mathcal{B}} \bigg). \end{aligned} \tag{49}$$

Define

$$\begin{aligned} & l_3 = \sqrt{80\mathcal{B}(1+4\mathcal{B})^2/3}\sqrt{(\mathcal{B}+1)(8+c_1)}, \\ & \mathcal{S}_3 = (p+1)\mathcal{W} + (\mathcal{D}-1)(\mathcal{W}^2+\mathcal{W}) + \mathcal{W} + 1, \\ & f_3(\mathcal{D},\mathcal{W},n) = \sqrt{c_2\mathcal{D}\mathcal{S}_3 \log \mathcal{S}_3 \log n/n}, \quad \text{and} \quad t = l_3 f_3(\mathcal{D},\mathcal{W},n). \end{aligned}$$

If $t/(1+4\mathcal{B}) < 1$ and $t/(5\mathcal{B}+20\mathcal{B}^2) < 1$, applying (49), Theorem 11.6 in Györfi et al. (2002), Theorem 7 in Bartlett et al. (2019), and Theorem 2.6.7 in Van Der Vaart and Wellner (1996), we have

$$\begin{aligned} P(P_n\widehat{g} - P\widehat{g} > t) \leq & P\bigg(\sup_{g\in\mathcal{H}_{\mathcal{D},\mathcal{W},\mathcal{B}}} \frac{P_n(g+\mathcal{B}) - P(g+\mathcal{B})}{1+P_n(g+\mathcal{B}) + P(g+\mathcal{B})} > \frac{t}{1+4\mathcal{B}} \bigg) \\ \leq & 4EN(t/(5+20\mathcal{B}), \mathcal{H}_{\mathcal{D},\mathcal{W},\mathcal{B}}, L_1(P_n)) \exp\bigg\{ -\frac{3t^2 n}{80\mathcal{B}(1+4\mathcal{B})^2} \bigg\} \\ \leq & 4c_1c_2\mathcal{D}\mathcal{S}_3 \log \mathcal{S}_3 \{16e\}^{c_2\mathcal{D}\mathcal{S}_3\log\mathcal{S}_3} \bigg(\frac{5\mathcal{B}+20\mathcal{B}^2}{t} \bigg)^{c_2\mathcal{D}\mathcal{S}_3\log\mathcal{S}_3 - 1} \\ & \times \exp\bigg\{ -\frac{3t^2 n}{80\mathcal{B}(1+4\mathcal{B})^2} \bigg\} \\ \leq & \frac{1}{n}. \end{aligned} \tag{50}$$

When $t/(1+4\mathcal{B}) < 1$ and $t/(5\mathcal{B}+20\mathcal{B}^2) \geq 1$, based on (49) and Theorem 11.6 in Györfi et al. (2002), one can prove

$$\begin{aligned} P(P_n\widehat{g} - P\widehat{g} > t) \leq & P\bigg(\sup_{g\in\mathcal{H}_{\mathcal{D},\mathcal{W},\mathcal{B}}} \frac{P_n(g+\mathcal{B}) - P(g+\mathcal{B})}{1+P_n(g+\mathcal{B}) + P(g+\mathcal{B})} > \frac{t}{1+4\mathcal{B}} \bigg) \\ \leq & 4EN(t/(5+20\mathcal{B}), \mathcal{H}_{\mathcal{D},\mathcal{W},\mathcal{B}}, L_1(P_n)) \exp\bigg\{ -\frac{3t^2 n}{80\mathcal{B}(1+4\mathcal{B})^2} \bigg\} \\ = & 4\exp\bigg\{ -\frac{3t^2 n}{80\mathcal{B}(1+4\mathcal{B})^2} \bigg\} \\ \leq & \frac{1}{n}. \end{aligned} \tag{51}$$

In the case that $t/(1+4\mathcal{B}) \geq 1$, according to (49), we can show

$$P(P_n\widehat{g} - P\widehat{g} > t) \leq P\bigg(\sup_{g\in\mathcal{H}_{\mathcal{D},\mathcal{W},\mathcal{B}}} \frac{P_n(g+\mathcal{B}) - P(g+\mathcal{B})}{1+P_n(g+\mathcal{B}) + P(g+\mathcal{B})} > \frac{t}{1+4\mathcal{B}} \bigg) = 0. \tag{52}$$

It follows from (50)-(52) that with probability at least $1 - 1/n$ that

$$\sup_{g \in \mathcal{H}_{\mathcal{D},\mathcal{W},\mathcal{B}}} P_n g - Pg \leq l_3 f_3(\mathcal{D}, \mathcal{W}, n). \tag{53}$$

Similar to (53), we have that with probability at least $1 - 1/n$ that

$$\sup_{g \in \mathcal{H}_{\mathcal{D},\mathcal{W},\mathcal{B}}} Pg - P_n g \leq l_3 f_3(\mathcal{D}, \mathcal{W}, n). \tag{54}$$

Using (47), (48), (53) and (54), we can obtain that with probability at least $1 - 10/n$,

$$\begin{aligned}
&E[\{\widehat{g}(\boldsymbol{X}) - g^*(\boldsymbol{X})\}^2 \mid \boldsymbol{O}_1, \ldots, \boldsymbol{O}_n] \\
\leq &\frac{4e^{4\mathcal{B}}}{a_2 E(\Delta)}[4e^{2\mathcal{B}}\{2e^{2\mathcal{B}}(\log n/n)^{1/2} + l_1 f_1(\mathcal{D}, \mathcal{W}, n)\}/a_2^2 + 2l_2 f_2(\mathcal{D}, \mathcal{W}, n) \\
&+ 524 a_1 a_3 \sqrt{p} e^{4\mathcal{B}}\{N^2 L^2 \log_3(N+2)\}^{-1/p}/a_2] + 2l_3^2 f_3^2(\mathcal{D}, \mathcal{W}, n).
\end{aligned} \tag{55}$$

This implies

$$\begin{aligned}
&E\{\widehat{g}(\boldsymbol{X}) - g^*(\boldsymbol{X})\}^2 \\
\leq &\frac{4e^{4\mathcal{B}}}{a_2 E(\Delta)}[4e^{2\mathcal{B}}\{2e^{2\mathcal{B}}(\log n/n)^{1/2} + l_1 f_1(\mathcal{D}, \mathcal{W}, n)\}/a_2^2 + 2l_2 f_2(\mathcal{D}, \mathcal{W}, n) \\
&+ 524 a_1 a_3 \sqrt{p} e^{4\mathcal{B}}\{N^2 L^2 \log_3(N+2)\}^{-1/p}/a_2] + 2l_3^2 f_3^2(\mathcal{D}, \mathcal{W}, n) + \frac{10}{n}(2\mathcal{B} + a_5)^2.
\end{aligned}$$

We accomplish the proof of Corollary 4. ■

Appendix B. Simulation results

In this section, we compared the proposed method with that in Wang et al. (2016). We set p at 5 and 10. In example 1, we first considered the PH data. The covariate $\boldsymbol{X}$ was drawn from a p-dimensional truncated normal distribution. For the case of $p = 5$, we generated $\boldsymbol{X}_0$ from a multivariate normal distribution that the marginal distribution followed $N(3, 0.25)$ and the correlation matrix was $\boldsymbol{H}$ with $(\boldsymbol{H})_{ij} = 0.75^{|i-j|}$. Then we obtained $\boldsymbol{X}$ by truncating each element of $\boldsymbol{X}_0$ on $[-5, 5]$. For $p = 10$, the generation procedure for $\boldsymbol{X}$ was largely similar to that for the $p = 5$ case. The only difference was that the marginal distribution for each element of $\boldsymbol{X}$ was $N(3, 1)$. We set $\{\lambda_0(t), \boldsymbol{\beta}\} = \{0.1, (0.2, 0.4, 0.1, 0.3, 0.5)^\top\}$ and $\{\lambda_0(t), \boldsymbol{\beta}\} = \{0.01t, (0.002, 0.001, 0.0025, 0.0005, 0.0015, 0.0045, 0.0015, 0.001, 0.002, 0.0035)^\top\}$ for the $p = 5$ and $p = 10$ cases, respectively. C was generated from the uniform distribution $U(0, c)$, where c was chosen to control a specified CR. The failure time of the AFT data with 5-dimensional covariates was simulated to follow

$$\log T = (0.3, 0.5, 0.6, 0.3, 0.4)^\top \boldsymbol{X} - 0.1X_1X_4 - 0.2X_2X_4 + 0.1X_3X_4 - 0.1X_5 + N(0, 1).$$

$\boldsymbol{X}$ followed the same distribution as $\boldsymbol{X}$ in the PH data with 5-dimension. C was also generated from a uniform distribution $U(0, c_T)$ with a positive number c_T to control a specified censoring ratio. In the case of $p = 10$, we first generated $\boldsymbol{X}_0$ from a multivariate normal distribution, where the marginal distribution was $N(3, 1)$ and the correlation matrix

$\boldsymbol{H}$ had elements $(\boldsymbol{H})ij = 0.75^{|i-j|}$. Then we obtained $\boldsymbol{X}$ by truncating each element of $\boldsymbol{X}_0$ to the interval $[-5, 5]$. The failure time T and censoring time C were generated from the following two models, respectively:

$$\log T = \boldsymbol{X}^\top \boldsymbol{\beta}_T + 0.1X_1X_5 + 0.2X_2X_6 - 0.1X_3X_7 + N(0,1),$$

and

$$\log C = \boldsymbol{X}^\top \boldsymbol{\beta}_C + 0.9X_1X_5 + 0.9X_2X_6 + 0.9X_3X_7 + N(0,1),$$

where $\boldsymbol{\beta}_T = (0.3, 0.5, 0.6, 0.3, 0.4, 0.5, 0.3, 0.6, -0.1, -0.4)^\top$, and $\boldsymbol{\beta}_C$ was chosen to control the CR. 100 and 200 samples were employed for training the network, while 20000 samples were applied as the validation data. The results presented below were based on 100 replications with a CR of 10% and 20%, resulting in the mean of loss function value $PR_n(g)$.

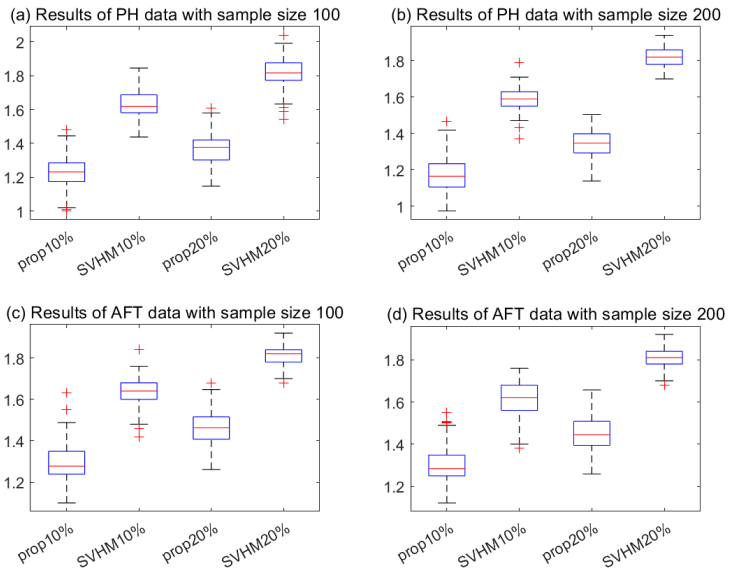


Figure 3: Comparison bewteen the proposed method (denoted as prop) and the support vector hazards machine (denoted as SVHM) approach in the case of 5-dimensional covariates. The red line in the middle of the blue box represents the median quantile of the results. The red cross represents the outlier of the results.

The proposed method was compared with the support vector hazards machine approach (Wang et al., 2016). Such method shares the same loss function with that in example 1 but incorporates the support vector machine framework for optimization. The comparison between these two methods mainly focused on the performance of optimization. Box plots are presented in Figures 3 and 4 for comparison. The findings demonstrate that, across most scenarios, the proposed method yields lower loss values than the support vector hazards machine approach in different dimensions, CRs and sample sizes. The differences in the median quantile of loss between the proposed method and the support vector hazards

machine approach become more pronounced in the 5-dimensional case. The bottom and top lines of the blue box represent the 25% and 75% quantiles of the simulation results, respectively. Given that the distances between the 25% and 75% quantiles are less than or not significantly exceeding 0.1, the precision of the proposed method is acceptable. Furthermore, the proposed method yields a lower total number of outliers compared to the support vector hazards machine approach. In addition, the results of the proposed and support vector hazards machine methods align with expectations, showing that the median quantile of loss increases with a rise in CR and decreases with a reduction in sample size. The influence is more pronounced for CR than sample size.

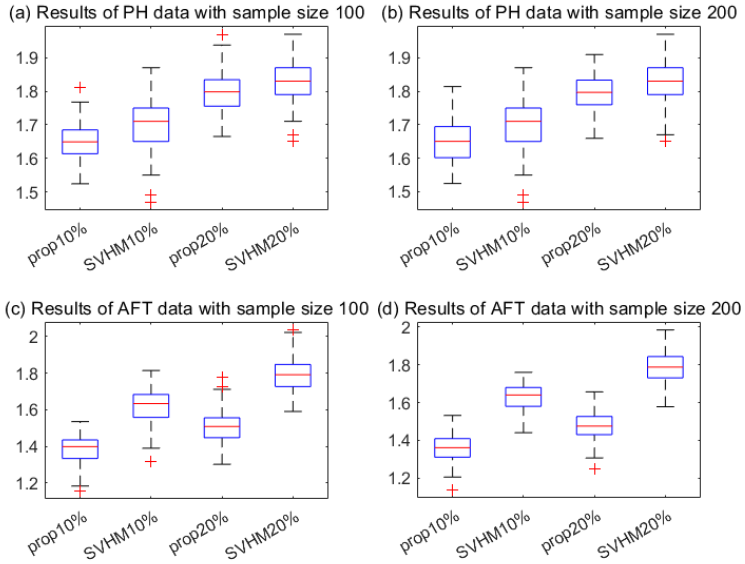


Figure 4: Comparison bewteen the proposed method and the support vector hazards machine approach in the case of 10-dimensional covariates.

References

Martin Anthony and Peter L Bartlett. *Neural network learning: Theoretical foundations*. Cambridge University Press, New York, 1999.

Richard G Baraniuk and Michael B Wakin. Random projections of smooth manifolds. *Foundations of Computational Mathematics*, 9(1):51–77, 2009.

Peter L Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(1):1–17, 2019.

Imad Bou-Hamad, Denis Larocque, and Hatem Ben-Ameur. A review of survival trees. *Statistics Surveys*, 5:44–71, 2011.

Chun-Houh Chen, Ker-Chau Li, and Jane-Ling Wang. Dimension reduction for censored regression data. *The Annals of Statistics*, 27(1):1–23, 1999.

Kani Chen, Zhezhen Jin, and Zhiliang Ying. Semiparametric analysis of transformation models with censored data. *Biometrika*, 89(3):659–668, 2002.

Kani Chen, Shaojun Guo, Liuquan Sun, and Jane-Ling Wang. Global partial likelihood for nonparametric proportional hazards models. *Journal of the American Statistical Association*, 105(490):750–760, 2010.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537, 2011.

David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

Jerome H Friedman and Werner Stuetzle. Projection pursuit regression. *Journal of the American Statistical Association*, 76(376):817–823, 1981.

Wei Fu and Jeffrey S Simonoff. Survival trees for left-truncated and right-censored data, with application to time-varying covariate data. *Biostatistics*, 18(2):352–369, 2017.

Michael F Gensheimer and Balasubramanian Narasimhan. A scalable discrete-time survival model for neural networks. *PeerJ*, 7:1–19, 2019.

Patricia M Grambsch and Terry M Therneau. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81(3):515–526, 1994.

László Györfi, Michael Köhler, Adam Krzyżak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer, New York, 2002.

Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. Support vector machines. *IEEE Intelligent Systems and Their Applications*, 13(4):18–28, 1998.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Xin Huang, Shande Chen, and Seng-jaw Soong. Piecewise exponential survival trees with time-dependent covariates. *Biometrics*, 54(4):1420–1433, 1998.

Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. Random survival forests. *The Annals of Applied Statistics*, 2(3):841–860, 2008.

Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3D convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35 (1):221–231, 2012.

Yuling Jiao, Guohao Shen, Yuanyuan Lin, and Jian Huang. Deep nonparametric regression on approximate manifolds: Nonasymptotic error bounds with polynomial prefactors. *The Annals of Statistics*, 51(2):691–716, 2023.

Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(24): 1–12, 2018.

Faisal M Khan and Valentina Bayer Zubek. Support vector regression for censored data (svrc): a novel tool for survival analysis. In *2008 Eighth IEEE International Conference on Data Mining*, pages 863–868. IEEE, 2008.

Shakeeb Khan and Elie Tamer. Partial rank estimation of duration models with general forms of censoring. *Journal of Econometrics*, 136(1):251–280, 2007.

Håvard Kvamme, Ørnulf Borgan, and Ida Scheel. Time-to-event prediction with neural networks and Cox regression. *Journal of Machine Learning Research*, 20(129):1–30, 2019.

Quoc V Le, Will Y Zou, Serena Y Yeung, and Andrew Y Ng. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3361–3368, 2011.

Changhee Lee, William Zame, Jinsung Yoon, and Mihaela Van Der Schaar. Deephit: a deep learning approach to survival analysis with competing risks. *Proceedings of the 2018 Conference on Artificial Intelligence*, pages 2314–2321, 2018.

Dan Yu Lin and Zhiliang Ying. Semiparametric analysis of the additive risk model. *Biometrika*, 81(1):61–71, 1994.

Jianfeng Lu, Zuowei Shen, Haizhao Yang, and Shijun Zhang. Deep network approximation for smooth functions. *SIAM Journal on Mathematical Analysis*, 53(5):5465–5506, 2021.

Aaron J Molstad and Rohit K Patra. Dimension reduction for integrative survival analysis. *Biometrics*, 79(3):1610–1623, 2023.

Grégoire Montavon, Matthias Rupp, Vivekanand Gobre, Alvaro Vazquez-Mayagoitia, Katja Hansen, Alexandre Tkatchenko, Klaus-Robert Müller, and O Anatole Von Lilienfeld. Machine learning of molecular electronic properties in chemical compound space. *New Journal of Physics*, 15(9):1–16, 2013.

Brian D Ripley and Ruth M Ripley. Neural networks as statistical methods in survival analysis. In *Clinical Application of Artificial Neural Network*, pages 237–255. Cambridge University Press, New York, 2001.

Ruth M Ripley, Adrian L Harris, and Lionel Tarassenko. Non-linear survival analysis using neural networks. *Statistics in Medicine*, 23(5):825–842, 2004.

Lior Rokach and Oded Maimon. *Decision trees*. Springer, Boston, 2005.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 48(4):1875–1897, 2020.

Zuowei Shen, Haizhao Yang, and Shijun Zhang. Optimal approximation rate of ReLU networks in terms of width and depth. *Journal de Mathématiques Pures et Appliquées*, 157:101–135, 2022.

Pannagadatta K Shivaswamy, Wei Chu, and Martin Jansche. A support vector approach to censored targets. *Proceedings of the 2007 Seventh IEEE nternational Conference on Data Mining*, pages 655–660, 2007.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, 2013.

Xiao Song, Shuangge Ma, Jian Huang, and Xiao-Hua Zhou. A semiparametric approach for the nonparametric transformation survival model with multiple covariates. *Biostatistics*, 8(2):197–211, 2007.

Qiang Sun, Ruoqing Zhu, Tao Wang, and Donglin Zeng. Counting process-based dimension reduction methods for censored outcomes. *Biometrika*, 106(1):181–196, 2019.

Vanya Van Belle, Kristiaan Pelckmans, Johan AK Suykens, and Sabine Van Huffel. Additive survival least-squares support vector machines. *Statistics in Medicine*, 29(2):296–308, 2010.

Vanya Van Belle, Kristiaan Pelckmans, Sabine Van Huffel, and Johan AK Suykens. Support vector methods for survival analysis: a comparison between ranking and regression approaches. *Artificial Intelligence in Medicine*, 53(2):107–118, 2011.

Aad W. Van Der Vaart and Jon A. Wellner. *Weak convergence and empirical processes: with applications to statistics*. Springer, New York, 1996.

Matt P Wand and M Chris Jones. *Kernel smoothing*. CRC press, 1994.

Yuanjia Wang, Tianle Chen, and Donglin Zeng. Support vector hazards machine: A counting process framework for learning risk scores for censored outcomes. *Journal of Machine Learning Research*, 17(1):1–37, 2016.

Yingcun Xia, Dixin Zhang, and Jinfeng Xu. Dimension reduction and semiparametric estimation of survival models. *Journal of the American Statistical Association*, 105(489): 278–290, 2010.

Wenyi Xie, Donglin Zeng, and Yuanjia Wang. Support vector machine for dynamic survival prediction with time-dependent covariates. *The Annals of Applied Statistics*, 18(3):2166–2186, 2024.

Donglin Zeng and Danyu Lin. Efficient estimation of semiparametric transformation models for counting processes. *Biometrika*, 93(3):627–640, 2006.

Donglin Zeng and Danyu Lin. Maximum likelihood estimation in semiparametric regression models with censored data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69(4):507–564, 2007.

Qixian Zhong, Jonas Mueller, and Jane-Ling Wang. Deep learning for the partially linear Cox model. *The Annals of Statistics*, 50(3):1348–1375, 2022.