

# Nonconvex Penalized LAD Estimation in Partial Linear Models with DNNs: Asymptotic Analysis and Proximal Algorithms

Lechen Feng<sup>\*</sup>      Haoran Li<sup>†</sup>      Lucky Li<sup>‡</sup>      Xingqiu Zhao<sup>§</sup>

## Abstract

This paper investigates the partial linear model by Least Absolute Deviation (LAD) regression. We parameterize the nonparametric term using Deep Neural Networks (DNNs) and formulate a penalized LAD problem for estimation. Specifically, our model exhibits the following challenges. First, the regularization term can be nonconvex and nonsmooth, necessitating the introduction of infinite dimensional variational analysis and nonsmooth analysis into the asymptotic normality discussion. Second, our network must expand (in width, sparsity level and depth) as more samples are observed, thereby introducing additional difficulties for theoretical analysis. Third, the oracle of the proposed estimator is itself defined through a ultra high-dimensional, nonconvex, and discontinuous optimization problem, which already entails substantial computational and theoretical challenges. Under such the challenges, we establish the consistency, convergence rate, and asymptotic normality of the estimator. Furthermore, we analyze the oracle problem itself and its continuous relaxation. We study the convergence of a proximal subgradient method for both formulations, highlighting their structural differences lead to distinct computational subproblems along the iterations. In particular, the relaxed formulation admits significantly cheaper proximal updates, reflecting an inherent trade-off between statistical accuracy and computational tractability.

**Keywords:** Partial Linear Model, Least Absolute Deviation, Deep Neural Network, Optimization, Stochastic Subgradient Descent

## 1 Introduction

Partial Linear Models (PLMs) have been extensively studied in classical multivariate regression; see *Hardle et al.* (2006) [18] for comprehensive survey of this framework. The main motivation is to allow different covariates to be modeled in different ways: through simple linear effects, or through more flexible nonparametric components. In general, PLMs achieve a balance between flexibility and robustness, retaining the adaptability of nonparametric methods while reducing the dimensionality burden of fully nonparametric models. Building on these advantages, PLMs have found important applications in biostatistics, computational public health, life sciences, environmental science, and economics; see *Engle et al.* (1986) [12], *Zeger and Diggle* (1994) [47] and *Peng et al.* (2006) [33].

---

<sup>\*</sup>Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong. Email: [fenglechen0326@163.com](mailto:fenglechen0326@163.com)

<sup>†</sup>Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong. Email: [hao316.li@connect.polyu.hk](mailto:hao316.li@connect.polyu.hk)

<sup>‡</sup>College of Computing, Data Science, and Society, University of California, Berkeley, CA 94720. Email: [luckyql@berkeley.edu](mailto:luckyql@berkeley.edu), [luckyql17@gmail.com](mailto:luckyql17@gmail.com)

<sup>§</sup>Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong. Email: [xingqiu.zhao@polyu.edu.hk](mailto:xingqiu.zhao@polyu.edu.hk)

In this paper, we consider the following PLM:

$$Y = \beta_0^\top X + g_0(Z) + \varepsilon \quad (1)$$

with covariates  $X \in \mathbb{R}^d$ ,  $Z \in \mathbb{R}^l$ , a vector of unknown parameters  $\beta_0$ , an unknown nonlinear function  $g_0$  and a random error  $\varepsilon$ . Consider  $N$  i.i.d. observations  $\{U_i \doteq (X_i, Y_i, Z_i)\}_{i=1}^N$ , while denote  $\mathbf{X} \doteq (X_1, \dots, X_N)$ ,  $\mathbf{Y} \doteq (Y_1, \dots, Y_N)$ ,  $\mathbf{Z} \doteq (Z_1, \dots, Z_N)$ , and  $\mathbf{U} \doteq (U_1, \dots, U_N)$ . We aim to solve the following penalized Least Absolute Deviation (LAD) regression problem for the estimation of unknown parameters  $\beta_0$  and  $g_0$ :

$$(\hat{\beta}_N, \hat{g}_N) \in \operatorname{argmin}_{\beta, g} \frac{1}{N} \sum_{i=1}^N |Y_i - \beta^\top X_i - g(Z_i)| + \lambda_N \mathcal{J}_{N,M}(\beta, g) \quad (2)$$

with given  $M > 0$ ,  $\lambda_N > 0$  and (possibly) nonconvex and nonsmooth regularization term  $\mathcal{J}_{N,M}(\beta, g)$  bounded by constant  $M$ , i.e.,  $\|\mathcal{J}_{N,M}\|_\infty < M$ . In this paper, we estimate the unknown function  $g_0$  through a sparse Deep Neural Network (DNN), and therefore focus on the following finite-dimensional optimization problem:

$$\min_{\beta, g(\mathbf{W})} \frac{1}{N} \sum_{i=1}^N |Y_i - \beta^\top X_i - g(\mathbf{W}; Z_i)| + \lambda_N \mathcal{J}_{N,M}(\beta, g(\mathbf{W})), \quad (3)$$

where  $g(\mathbf{W})$  belongs to a prescribed class of sparse DNNs; the readers may refer to Section 3 for the details.

Least Squares Estimation (LSE) has been the most widely studied and influential for PLMs (1), due to its simplicity and broad applicability. Specifically, LSE seeks to solve the following optimization problem:

$$\min_{\beta, g} \frac{1}{N} \sum_{i=1}^N (Y_i - \beta^\top X_i - g(Z_i))^2 + \mathcal{J}(\beta, g) \quad (4)$$

with various parametric policies of nonlinear function  $g$  and different choices of regularization term  $\mathcal{J}$ , as implemented in the respective literature. Broadly speaking, methods for handling the nonparametric function  $g_0$  fall into two main classes: estimating the linear and nonlinear components jointly, and disentangling the estimation of two components. The joint estimation approach primarily relies on smoothing techniques, e.g., cubic spline smoother (Engle *et al.* (1986) [12]), local polynomial smoother (Hamilton and Truong (1997) [17]) and B-splines smoother (McLean *et al.* (2014) [30]). Whereas the separate estimation approach mainly includes the profile likelihood method (Severini and Wong (1992) [40]), the partial residual approach (Cuzick (1992) [8], Ferraccioli *et al.* (2023) [15]), and the difference approach (Duran *et al.* (2012) [11]). Meanwhile, motivated by considerations such as sparsity, smoothness, robustness, and prevention of overfitting, a line of research has focused on the selection of regularization terms  $\mathcal{J}$  of problem (4). Henceforth, various  $\mathcal{J}$  has been proposed, including Lasso (Tibshirani (1996) [42]), SCAD (Smoothly Clipped Absolute Deviation, Fan and Li (2001) [13]), Elastic Net (Zou and Hastie (2005) [51]), MCP (Minimax Concave Penalty, Zhang (2010) [48]) and SACR (Smoothly Adaptively Centered Ridge, Belli (2022) [2]). In recent year, with the rapid growth of big data, data have become more diverse and voluminous, bringing the new challenges for LSE. Roozbeh and Arashi (2016) [36] introduce a biased estimator for shrinkage parameter which is of harmonic type mean of ridge estimators, aiming to tackle the problem of multicollinearity. In addition, Auerbach (2022) [1] introduces a matching pairs method to incorporate network data into econometric modeling.

Despite the aforementioned numerous efforts to improve LSE, it remains inherently sensitive to outliers, heavy-tailed errors, and high reliance on assumptions (e.g., linearity and homoscedasticity) cannot be fully overcome; see Cizek and Sadikoglu (2020) [7] for the detailed discussion of the limitations of LSE. To circumvent the aforementioned drawbacks, LAD estimation has been adopted as a robust alternative for analyzing PLMs, i.e., estimating the unknown parameters  $\beta_0$  and  $g_0$  by solving optimization problem (2). Since the LAD cost function is inherently non-differentiable, even ignoring the potential non-smoothness of

the regularization term  $\mathcal{J}_{N,M}$ , its theoretical analysis is challenging. In the early literature, to render the problem analytically tractable, unknown function  $g_0$  is often represented through a basis expansion

$$g_0(Z) = \sum_{k=1}^K \theta_k \phi_k(Z),$$

where the basis functions  $\{\phi_k\}_{k=1}^K$  are pre-specified, and identifying the coefficient vector  $\theta = (\theta_1, \dots, \theta_K)^\top$  is then equivalent to estimating  $g_0$  itself; see He and Shi (1994) [20] and Lee (2003) [28]. Remarkably, such early works primarily focus on establishing the consistency and asymptotic distribution of the parametric component  $\beta_0$ , rather than fully characterizing the nonparametric part. In more recent years, to ensure the consistency of the estimation of  $g_0$ , alternative structural conditions are additionally required. For instance, Lian (2012) [29] considers the following additive PLMs

$$Y = \beta_0^\top X + \sum_{k=1}^l g_{0,k}(Z_{(k)}) + \varepsilon$$

with  $Z = (Z_{(1)}, \dots, Z_{(l)})^\top$ , while Ben and Lan (2016) [41] further extend Lian's framework to the ultra high-dimensional setting. For a comprehensive introduction of LAD for PLMs, we refer to the monographs by Koenker *et al.* (2017) [25].

Over the past decade, deep learning has been widely applied in many domains and has achieved remarkable success, thereby being naturally incorporated into the traditional statistical field. Generally speaking, DNNs not only exhibit strong function approximation capabilities (see Hornik *et al.* (1989) [22] for the universal approximation theorem) but also help mitigate the curse of dimensionality, making them a valuable tool for estimating the nonlinear function  $g_0$  of PLMs (1). For instance, Farrell *et al.* (2019) [14] apply deep learning to semiparametric inference and establish nonasymptotic bounds of DNNs for nonparametric term, covering the standard LSE in particular. Additionally, Zhong and Wang (2024) [50] leverage deep learning for PLMs in quantile regression to achieve interpretable results and enable statistical inference, while they later extended these results to partially linear Cox models; see Zhong *et al.* (2022) [49]. Subsequently, deep learning for PLMs via quantile regression has been extended in multiple directions, including high-dimensional PLMs (Wang (2025) [45]) and dependent data PLMs (see Brown (2024) [5] for stationary  $\beta$ -mixing sequences). In a different direction, Wen *et al.* (2016) [46] introduce sparse DNNs with ReLU activation function to fit unknown nonlinear function  $g_0$ , yielding saving computational resources and mitigating overfitting. Further, Schmidt-Hieber (2020) [38] establishes several non-asymptotic properties of the DNNs with the aforementioned sparse structure, including upper bounds on covering numbers and approximation rates for Hölder smooth functions, which provides essential theoretical tools for this paper. In this paper, we adopt the DNN architecture of Wen *et al.* (2016) [46], while allowing it to expand in width, sparsity level, and depth as the sample size increases. For notational convenience in this section, we write  $\mathcal{M}_N$  for the DNN architecture associated with  $N$  samples.

In conclusion, the shortcomings of the aforementioned works are as follows:

- Methods based on LSE (e.g., [8, 11, 12, 14, 15, 17, 30, 40]) are so sensitive to outliers that a single outlier can lead to completely unreliable estimates (see Hubert and Ronchetti (2009) [23] for details).
- The existing methods of estimating unknown functions, e.g. [14, 45, 50], often let the regularization term exhibit very simple form (or even omit it). However, the estimator is usually apriori assumed to be sparse, flat, smooth and so on, leading to the nonconvex and nonsmooth regularization term, which is beyond the scope of the existing theoretical framework.
- Although the existing works such as [45, 49, 50] assume that the DNN architecture  $\mathcal{M}_N$  expands (in width, depth, and sparsity level) as the sample size increases, the proofs of consistency and asymp-

otic normality rely on the fixed network architecture. This creates an inconsistency of the existing theoretical framework.

- Methods assuming additivity of the nonlinear term  $g_0$  are not amenable to modeling the interaction among covariates (e.g., [20, 25, 28, 29, 41]).
- Nonparametric methods (e.g., [16, 19, 35]) do not leverage the known linear structure of  $\beta_0^\top X$ . As a result, such methods require many unnecessary parameters to approximate  $\beta_0^\top X$ , which can lead to the curse of dimensionality, especially when the dimension of  $X$  is high.

In this paper, we propose estimator (2) to address the above issues. Concretely, the contributions are as follows.

- We establish the consistency, convergence rate and asymptotic normality of estimator (2). Notably, the nonconvex and nonsmooth regularization term of (2) invalidates the use of classical differential calculus (e.g., the chain and sum rules) on the penalized LAD criterion. This is a critical issue because a key step in establishing asymptotic normality for M-estimators relies on analyzing the differential properties of the objective function; see [50]. Hence, we need to demonstrate that aforementioned regularization term exhibit the chain rule, additive properties and the projection theorem of partial limiting subgradient, necessitating the tools from infinite-dimensional variational and nonsmooth analysis (e.g. Mordukhovich subgradient, epi-convergence and generalized cone).
- The expansion (in width, depth and sparsity level) of  $\mathcal{M}_N$  architecture causes the covering number of candidate estimators to approach infinity, rendering the classic methodology (for proving consistency and asymptotic normality) inapplicable; see [43, 45, 50] for details. To address this issue, we characterize the growth rate of the covering number and the entropy of the candidate estimators, and demonstrate the universal convergence of the criterion function (2).
- The oracle of estimator (2) is equivalent to a nonconvex and discontinuous optimization problem. A significant computational challenge arises when using proximal gradient-type methods, as the computational cost mainly depends on the projection operator. To balance computational tractability and precision, we propose two approaches. First, for the primal formulation, we derive a closed-form solution for the projection onto the sparse constraint. This result allows us to directly analyze the computational complexity of the proximal algorithm. Second, we approximate the primal optimization problem with a sequence of coordinate convex relaxation problems, and prove that such approximated problem converges to the primal problem. The relaxed formulation admits significantly cheaper proximal updates, reflecting an inherent trade-off between statistical accuracy and computational tractability.
- To evaluate the optimization error, we establish the global convergence of the proximal stochastic subgradient method for both the primal and approximate formulations of the penalized LAD regression problem (3). Our proof leverages the Lyapunov framework developed by [4, 9, 10], and the core of our analysis is to show the Weak Sard Property. To prove this property, we employ tools from differential and algebraic geometry, including Whitney stratification, Sard's theorem, and the chain rule for locally Lipschitz functions. To establish the Weak Sard Property, we partition the feasible set into a collection of disjoint smooth manifolds, and the penalized LAD cost is smooth on each manifold. We then use the classical gradient of local mollifier to cover the Clarke subgradient of the penalized LAD cost. By applying the standard Sard's theorem to such the local mollifiers on each piece of the partition, the Weak Sard Property is deduced.

## 2 Preliminaries

**Notation:** For  $A \in \mathbb{R}^{m \times n}$ , we define  $\sigma_{\min}(A) := \min_{\|x\|_2=1} \|Ax\|_2$ . The graph of  $f$ , denoted by  $\mathbf{graph}(f)$ , is defined as  $\mathbf{graph}(f) := \{(x, f(x)) \in \mathbb{R}^n \times \mathbb{R}^m \mid x \in \text{dom}(f)\}$ . For  $n \geq 1$ , the set  $[n]$  denotes  $\{1, \dots, n\}$ .  $S$  be a subset of a topological space  $X$ . A point  $x$  is a cluster point of the set  $S$  if every neighbourhood of  $x$  contains infinitely many points of  $S$  different from  $x$  itself. Let  $\mathbf{Cluster}(S)$  denote the set of all the cluster points of  $S$ . For a set  $A$ , let  $\|A\| := \sup_{a \in A} \|a\|$ . Given a set of functions  $\mathcal{F}$ , we define  $\|G\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |G(f)|$ . For a sub-Gaussian random variable  $X$ , its  $\psi_2$ -norm is defined as  $\|X\|_{\psi_2} = \inf\{t > 0 : \mathbb{E}[\exp(X^2/t^2)] \leq 2\}$ . Let  $P$  denote the true distribution of the observations and  $\mathbb{P}_n$  the empirical measure based on a sample  $X_1, \dots, X_n$ , that is,  $\mathbb{P}_n f = \frac{1}{n} \sum_{i=1}^n f(X_i)$  and  $Pf = \mathbb{E}_P[f(X)]$ . The empirical process  $\mathbb{G}_n$  is then defined by  $\mathbb{G}_n f = \sqrt{n}(\mathbb{P}_n - P)f$ . For the empirical measure  $\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$ , the empirical  $L^p$  space is defined as

$$L^p(\mathbb{P}_n) = \left\{ f \text{ measurable} : \|f\|_{L^p(\mathbb{P}_n)} = \left( \int |f|^p d\mathbb{P}_n \right)^{1/p} = \left( \frac{1}{n} \sum_{i=1}^n |f(X_i)|^p \right)^{1/p} < \infty \right\}.$$

We use the soft O-notation  $\tilde{O}(\cdot)$  to suppress polylogarithmic factors in complexity bounds. Formally,  $f(n) = \tilde{O}(g(n))$  if there exists a constant  $k > 0$  such that  $f(n) = O(g(n) \log^k n)$ . The convex hull of the set  $C$ , defined as

$$\mathbf{conv}(C) := \left\{ \sum_{i=1}^k \lambda_i x_i \mid x_i \in C, \lambda_i \geq 0, \sum_{i=1}^k \lambda_i = 1, k \in \mathbb{N} \right\}.$$

We use the notation  $f(x) \lesssim g(x)$  to mean that there exists a constant  $C > 0$ , independent of the relevant variables, such that  $f(x) \leq Cg(x)$ . Given a probability space  $(\Omega, \mathcal{F}, P)$ , define

- $P^*(A)$ : Outer measure of a set  $A \subseteq \Omega$ , defined as  $P^*(A) := \inf\{P(B) \mid B \in \mathcal{A}, A \subseteq B\}$ .
- $\mathbb{E}^*[X]$ : Outer expectation of a function  $X : \Omega \rightarrow \mathbb{R}$ , defined as  $\mathbb{E}^*[X] := \inf\{\mathbb{E}[Y] \mid Y \text{ measurable}, Y \geq X\}$ .

Let  $L^2(\mathbf{m})$  be the space of square-integrable functions with respect to Lebesgue measure  $\mathbf{m}$ . The identity map on  $L^2(\mathbf{m})$  is a linear operator  $\mathbf{I} : L^2(\mathbf{m}) \rightarrow L^2(\mathbf{m})$  such that for every function  $f \in L^2(\mathbf{m})$ ,

$$\mathbf{I}(f) = f.$$

Furthermore, we define set-valued mapping

$$\mathbf{sign}^*(t) = \begin{cases} 1 & t > 0, \\ -1 & t < 0, \\ [-1, 1] & t = 0, \end{cases}$$

and signum function

$$\mathbf{sign}(t) = \begin{cases} 1 & t \geq 0, \\ -1 & t < 0. \end{cases}$$

**Definition 1** (Covering numbers, Definition 2.1.5 of [43]). *The covering number  $N(\varepsilon, \mathcal{F}, \|\cdot\|)$  is the minimal number of balls  $\{g : \|g - f\| < \varepsilon\}$  of radius  $\varepsilon$  needed to cover the set  $\mathcal{F}$ . The centers of the balls need not belong to  $\mathcal{F}$ , but they should have finite norms. The entropy (without bracketing) is the logarithm of the covering number.*

**Definition 2** (Bracketing numbers, Definition 2.1.6 of [43]). *Given two functions  $l$  and  $u$ , the bracket  $[l, u]$  is the set of all functions  $f$  with  $l \leq f \leq u$ . An  $\varepsilon$ -bracket is a bracket  $[l, u]$  with  $\|u - l\| < \varepsilon$ . The bracketing*

number  $N_{[\cdot]}(\varepsilon, \mathcal{F}, \|\cdot\|)$  is the minimum number of  $\varepsilon$ -brackets needed to cover  $\mathcal{F}$ . The entropy with bracketing is the logarithm of the bracketing number. In the definition of the bracketing number, the upper and lower bounds  $u$  and  $l$  of the brackets need not belong to  $\mathcal{F}$  themselves but are assumed to have finite norms.

**Definition 3** (Generalized normals, Definition 1.1 of [32]). Let  $\Omega$  be a nonempty subset of  $X$ . Given  $x \in \Omega$  and  $\varepsilon \geq 0$ , define the set of  $\varepsilon$ -normals to  $\Omega$  at  $x$  by

$$\widehat{N}_\varepsilon(x; \Omega) := \left\{ x^* \in X^* \left| \limsup_{\substack{u \xrightarrow{\Omega} x}} \frac{\langle x^*, u - x \rangle}{\|u - x\|} \leq \varepsilon \right. \right\}.$$

When  $\varepsilon = 0$ , elements of (1.2) are called Fréchet normals and their collection, denoted by  $\widehat{N}(x; \Omega)$ , is the prenormal cone to  $\Omega$  at  $x$ . If  $x \notin \Omega$ , we put  $\widehat{N}_\varepsilon(x; \Omega) := \emptyset$  for all  $\varepsilon \geq 0$ .

**Definition 4** (Sequential Normal Compactness, Definition 1.20 of [31]). A set  $\Omega \subset X$  is Sequentially Normally Compact (SNC) at  $\bar{x} \in \Omega$  if for any sequence  $(\varepsilon_k, x_k, x_k^*) \in [0, \infty) \times \Omega \times X^*$  satisfying

$$\varepsilon_k \downarrow 0, \quad x_k \rightarrow \bar{x}, \quad x_k^* \in \widehat{N}_{\varepsilon_k}(x_k; \Omega), \quad \text{and} \quad x_k^* \xrightarrow{w^*} 0$$

one has  $\|x_k^*\| \rightarrow 0$  as  $k \rightarrow \infty$ .

**Definition 5** (Sequential Normal Epi-Compactness of functions, Definition 1.116 of [31]). Let  $\varphi : X \rightarrow \overline{\mathbb{R}}$  be finite at  $\bar{x}$ . We say that  $\varphi$  is Sequentially Normally Epi-Compact (SNEC) at  $\bar{x}$  if its epigraph is sequentially normally compact at  $(\bar{x}, \varphi(\bar{x}))$ .

**Definition 6** (Subderivatives, Definition 8.1 of [34]). For a function  $f : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$  and a point  $\bar{x}$  with  $f(\bar{x})$  finite, the subderivative function  $\mathbf{d}f(\bar{x}) : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$  is defined by

$$\mathbf{d}f(\bar{x})(\bar{w}) := \liminf_{\substack{\tau \downarrow 0 \\ w \rightarrow \bar{w}}} \frac{f(\bar{x} + \tau w) - f(\bar{x})}{\tau}$$

**Definition 7** (Subdifferentials of extended-real-valued functions, Definition 1.32 of [32]). Let  $\varphi : X \rightarrow \overline{\mathbb{R}}$  be an extended-real-valued function on a Banach space  $X$ .

(i) Given  $\varepsilon \geq 0$  and  $x \in \text{dom } \varphi$ , the set

$$\widehat{\partial}_\varepsilon \varphi(x) := \left\{ x^* \in X^* \left| \liminf_{u \rightarrow x} \frac{\varphi(u) - \varphi(x) - \langle x^*, u - x \rangle}{\|u - x\|} \geq -\varepsilon \right. \right\}$$

is the  $\varepsilon$ -subdifferential of  $\varphi$  at  $x$ . The set  $\widehat{\partial}_0 \varphi(x)$  is denoted by

$$\widehat{\partial} \varphi(x) := \left\{ x^* \in X^* \left| \liminf_{u \rightarrow x} \frac{\varphi(u) - \varphi(x) - \langle x^*, u - x \rangle}{\|u - x\|} \geq 0 \right. \right\}$$

and is called the presubdifferential or the regular subdifferential of  $\varphi$  at this point. We put  $\widehat{\partial}_\varepsilon \varphi(x) := \emptyset$  for all  $\varepsilon \geq 0$  if  $x \notin \text{dom } \varphi$ .

(ii) Define the (basic, limiting) subdifferential of  $\varphi$  at  $\bar{x} \in \text{dom } \varphi$  by

$$\partial \varphi(\bar{x}) = \text{Lim sup}_{\substack{x \xrightarrow{\varphi} \bar{x} \\ \varepsilon \downarrow 0}} \widehat{\partial}_\varepsilon \varphi(x).$$

(iii) The singular subdifferential of  $\varphi$  at  $\bar{x} \in \text{dom } \varphi$  is defined by

$$\partial^\infty \varphi(\bar{x}) := \text{Lim sup}_{\substack{x \xrightarrow{\varphi} \bar{x} \\ \varepsilon, \lambda \downarrow 0}} \lambda \widehat{\partial}_\varepsilon \varphi(x).$$

We put  $\partial \varphi(\bar{x}) := \emptyset$  and  $\partial^\infty \varphi(\bar{x}) := \emptyset$  for  $\bar{x} \notin \text{dom } \varphi$ .

**Definition 8** (Clarke subdifferential, Definition 1 of [4]). *The Clarke subdifferential  $\partial^C f(x)$  of  $f$  at  $x$  is the set*

$$\partial^C f(x) = \begin{cases} \overline{\text{conv}} \{ \partial f(x) + \partial^\infty f(x) \} & \text{if } x \in \text{dom } f, \\ \emptyset & \text{if } x \notin \text{dom } f. \end{cases}$$

**Definition 9** (Constructions of second-order subdifferentials, Definition 1.46 of [32]). *Let  $\varphi : X \rightarrow \overline{\mathbb{R}}$  be an extended-real-valued function on a Banach space  $X$ , let  $\bar{x} \in \text{dom } \varphi$ , and let  $\bar{v} \in \partial\varphi(\bar{x})$  be a first-order subgradient from Definition 7. Define:*

(i) *The mapping  $\partial_N^2 \varphi(\bar{x}, \bar{v}) : X^{**} \rightrightarrows X^*$  with the values*

$$\partial_N^2 \varphi(\bar{x}, \bar{v})(u) := (D_N^* \partial\varphi)(\bar{x}, \bar{v})(u), \quad u \in X^{**},$$

*is the normal second-order subdifferential of  $\varphi$  at  $\bar{x}$  relative to  $\bar{v}$ .*

(ii) *The mapping  $\partial_M^2 \varphi(\bar{x}, \bar{v}) : X^{**} \rightrightarrows X^*$  with the values*

$$\partial_M^2 \varphi(\bar{x}, \bar{v})(u) := (D_M^* \partial\varphi)(\bar{x}, \bar{v})(u), \quad u \in X^{**},$$

*is the mixed second-order subdifferential of  $\varphi$  at  $\bar{x}$  relative to  $\bar{v}$ .*

(iii) *The mapping  $\check{\partial}^2 \varphi(\bar{x}, \bar{v}) : X^{**} \rightrightarrows X^*$  with the values*

$$\check{\partial}^2 \varphi(\bar{x}, \bar{v})(u) := (\hat{D}^* \partial\varphi)(\bar{x}, \bar{v})(u), \quad u \in X^{**},$$

*is the combined second-order subdifferential of  $\varphi$  at  $\bar{x}$  relative to  $\bar{v}$ .*

*For the definition of co-derivative  $D_N^*$ , see [32].*

**Definition 10** (Lower closure, Page 14 of [34]). *The function is lower semi-continuous and is the greatest of all the lower semi-continuous functions  $g$  such that  $g \leq f$ . It is called the lower closure of  $f$ , denoted by  $\text{clf}$ .*

**Definition 11** (Lower and upper epi-limits, Definition 7.1 of [34]). *For any sequence  $\{f^v\}_{v \in \mathbb{N}}$  of functions on  $\mathbb{R}^n$ , the lower epi-limit  $\text{e-liminf}_v f^v$  is the function having as its epigraph the outer limit of the sequence of sets  $\text{epi} f^v$ :*

$$\text{epi}(\text{e-liminf}_v f^v) \doteq \limsup_v (\text{epi} f^v).$$

*The upper epi-limit  $\text{e-limsup}_v f^v$  is the function having as its epigraph the inner limit of the sets  $\text{epi} f^v$ :*

$$\text{epi}(\text{e-limsup}_v f^v) \doteq \liminf_v (\text{epi} f^v).$$

*When these two functions coincide, the epi-limits function  $\text{e-lim} f^v$  is said to exist:*

$$\text{e-lim} f^v \doteq \text{e-liminf}_v f^v = \text{e-limsup}_v f^v.$$

**Definition 12** (O-minimal structure, Definition 6 of [4]). *An o-minimal structure is a sequence of Boolean algebras  $\mathcal{O}_d$  of subsets of  $\mathbb{R}^d$  such that for each  $d \in \mathbb{N}$ :*

(i) *if  $A$  belongs to  $\mathcal{O}_d$ , then  $A \times \mathbb{R}$  and  $\mathbb{R} \times A$  belong to  $\mathcal{O}_{d+1}$ ;*

(ii) *if  $\pi : \mathbb{R}^d \times \mathbb{R} \rightarrow \mathbb{R}^d$  denotes the coordinate projection onto  $\mathbb{R}^d$ , then for any  $A$  in  $\mathcal{O}_{d+1}$  the set  $\pi(A)$  belongs to  $\mathcal{O}_d$ ;*

(iii)  *$\mathcal{O}_d$  contains all sets of the form  $\{x \in \mathbb{R}^d : p(x) = 0\}$ , where  $p$  is a polynomial on  $\mathbb{R}^d$ ;*

(iv) the elements of  $\mathcal{O}_1$  are exactly the finite unions of intervals (possibly infinite) and points.

The sets  $A$  belonging to  $\mathcal{O}_d$ , for some  $d \in \mathbb{N}$ , are called *definable in the o-minimal structure*.

**Definition 13** (Lyapunov condition, Assumption B of [9]). *Let  $\mathcal{X}$  be a closed set and let  $G : \mathcal{X} \rightrightarrows \mathbb{R}^d$  be a set-valued map. Then an arc  $z : \mathbb{R}_+ \rightarrow \mathbb{R}^d$  is called a trajectory of  $G$  if it satisfies the differential inclusion  $\dot{z}(t) \in G(z(t))$  for a.e.  $t \geq 0$ . there exists a continuous function  $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}$ , which is bounded from below, and such that the following two properties hold.*

- (Weak Sard) *For a dense set of values  $r \in \mathbb{R}$ , the intersection  $\varphi^{-1}(r) \cap G^{-1}(0)$  is empty.*
- (Descent) *Whenever  $z : \mathbb{R}_+ \rightarrow \mathbb{R}^d$  is a trajectory of the differential inclusion and  $0 \notin G(z(0))$ , there exists a real  $T > 0$  satisfying*

$$\varphi(z(T)) < \sup_{t \in [0, T]} \varphi(z(t)) \leq \varphi(z(0)).$$

**Definition 14** (Chain rule, Definition 5.1 of [9]). *Consider a locally Lipschitz function  $f$  on  $\mathbb{R}^d$ . We will say that  $f$  admits a chain rule if for any absolutely continuous curves  $z : \mathbb{R}_+ \rightarrow \mathbb{R}^d$ , equality*

$$(f \circ z)'(t) = \langle \partial^C f(z(t)), z'(t) \rangle \text{ holds for a.e. } t \geq 0,$$

**Definition 15** (Smooth manifold, tangent space and normal space, Page 13, Page 51 and Page 138 of [27]). *A set  $M \subset \mathbb{R}^d$  is a  $C^p$  smooth manifold if there is an integer  $r \in \mathbb{N}$  such that around any point  $x \in M$ , there is a neighborhood  $U$  and a  $C^p$ -smooth map  $F : U \rightarrow \mathbb{R}^{d-r}$  with  $\nabla F(x)$  of full rank and satisfying  $M \cap U = \{y \in U : F(y) = 0\}$ . If this is the case, the tangent and normal spaces to  $M$  at  $x$  are defined to be  $T_M(x) := \text{Null}(\nabla F(x))$  and  $N_M(x) := (T_M(x))^\perp$ , respectively.*

**Definition 16** (Whitney stratification, Definition 5.6 of [9]). *A Whitney  $C^p$ -stratification  $\mathcal{A}$  of a set  $Q \subset \mathbb{R}^d$  is a partition of  $Q$  into finitely many nonempty  $C^p$  manifolds, called strata, satisfying the following compatibility conditions.*

- *Frontier condition: For any two strata  $L$  and  $M$ , the implication*

$$L \cap \text{cl } M \neq \emptyset \implies L \subset \text{cl } M \text{ holds.}$$

- *Whitney condition: For any sequence of points  $z_k$  in a stratum  $M$  converging to a point  $\bar{z}$  in a stratum  $L$ , if the corresponding normal vectors  $v_k \in N_M(z_k)$  converge to a vector  $v$ , then the inclusion  $v \in N_L(\bar{z})$  holds.*

A function  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  is *Whitney  $C^p$ -stratifiable* if its graph admits a Whitney  $C^p$ -stratification.

**Definition 17** (Hölder formulation class, Page 7 of [38]). *Let  $\gamma$  and  $B$  be two positive constants and  $\lfloor \gamma \rfloor$  denote the largest integer strictly less than  $\gamma$ . We call a function  $h : \mathbb{T} \subset \mathbb{R}^q \rightarrow \mathbb{R}$  a  $(\gamma, B)$ -Hölder smooth function if it satisfies*

$$\sup_{z \in \mathbb{T}} \left| \frac{\partial^{|\alpha|} h}{\partial^{\alpha_1} z_1 \dots \partial^{\alpha_q} z_q}(z) \right| \leq B, \text{ for all } \alpha = (\alpha_1, \dots, \alpha_q)^\top \in \mathbb{N}^q \text{ and } |\alpha| = \sum_{i=1}^q \alpha_i \leq \lfloor \gamma \rfloor,$$

and

$$\sup_{z, z^* \in \mathbb{T}} \left| \frac{\partial^{|\alpha|} h}{\partial^{\alpha_1} z_1 \dots \partial^{\alpha_q} z_q}(z) - \frac{\partial^{|\alpha|} h}{\partial^{\alpha_1} z_1 \dots \partial^{\alpha_q} z_q}(z^*) \right| \leq B \|z - z^*\|_2^{\gamma - \lfloor \gamma \rfloor}, \text{ for all } |\alpha| = \lfloor \gamma \rfloor.$$



Denote the class of all such  $(\gamma, B)$ -Hölder smooth functions as  $\mathcal{H}_q^\gamma(\mathbb{T}, B)$ . Let  $J \in \mathbb{N}$ ,  $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_J)^\top \in \mathbb{R}_+^J$ ,  $\mathbf{d} = (q, d_1, \dots, d_J)^\top \in \mathbb{N}^{J+1}$  and  $\bar{\mathbf{d}} = (\bar{d}_1, \dots, \bar{d}_J)^\top \in \mathbb{N}^J$  with  $\bar{d}_1 \leq q$  and  $\bar{d}_k \leq d_{k-1}, k = 2, \dots, J$ . We further define a composite function class:

$$\mathcal{H}(J, \boldsymbol{\gamma}, \mathbf{d}, \bar{\mathbf{d}}, B) = \left\{ h = h_J \circ \dots \circ h_1 : \mathbb{T} \rightarrow \mathbb{R} \mid h_k = (h_{k1}, \dots, h_{kd_k})^\top \text{ and } h_{kj} \in \mathcal{H}_{d_k}^{\gamma_k}([a_k, b_k]^{\bar{d}_k}, B) \text{ for some } |a_k|, |b_k| \leq B \right\}. \quad (5)$$

We call  $\bar{\mathbf{d}}$  the intrinsic dimension of the function  $h$  in  $\mathcal{H}(J, \boldsymbol{\gamma}, \mathbf{d}, \bar{\mathbf{d}}, B)$ .

**Definition 18** (Dual operator, Theorem 5.11-1 of [6]). Let  $X$  and  $Y$  be two normed vector spaces over the same field  $\mathbb{K}$ . Given any operator  $A \in \mathcal{L}(X; Y)$ , there exists one and only one operator  $A^* \in \mathcal{L}(Y^*; X^*)$ , called the dual operator of  $A$ , or simply the dual of  $A$ , such that

$$A^* y^*(x) = y^*(Ax) \quad \text{for all } x \in X \text{ and all } y^* \in Y^*.$$

Besides,  $\|A^*\|_{\mathcal{L}(Y^*; X^*)} = \|A\|_{\mathcal{L}(X; Y)}$ .

**Notations of Deep Neural Networks (DNN).** Let  $L \geq 2$  be an integer representing the number of layers, and let  $\mathbf{q} = (q_0, q_1, \dots, q_L)^\top \in \mathbb{N}^{L+1}$  define the number of neurons in each layer. An  $L$ -layer neural network is a function  $g : \mathbb{R}^{q_0} \rightarrow \mathbb{R}^{q_L}$  that maps a  $q_0$ -dimensional input to a  $q_L$ -dimensional output. It is defined by the following composition of functions:

$$\begin{aligned} m_0(z) &= z, \\ m_1(z) &= \sigma_1(W_1 m_0(z) + b_1), \\ &\dots, \\ m_{L-1}(z) &= \sigma_{L-1}(W_{L-1} m_{L-2}(z) + b_{L-1}), \\ g(z) &= W_L m_{L-1}(z) + b_L, \end{aligned} \quad (6)$$

where for each layer  $k = 1, \dots, L$ ,  $W_k$  is a  $q_k \times q_{k-1}$  weight matrix and  $b_k$  is a  $q_k$ -dimensional bias vector. The term  $m_k$  for  $1 \leq k \leq L-1$  represents the output of the  $k$ -th *hidden layer*, and  $L$  is the *depth* of the network. The functions  $\sigma_k : \mathbb{R}^{q_k} \rightarrow \mathbb{R}^{q_k}$  for  $k = 1, \dots, L-1$  are activation functions that operate element-wise on their input vectors. That is, for a vector  $v = (v_1, \dots, v_{q_k})^\top$ ,  $\sigma_k(v) = (\sigma(v_1), \dots, \sigma(v_{q_k}))^\top$ , where  $\sigma : \mathbb{R} \rightarrow \mathbb{R}$  is a scalar activation function (e.g., ReLU, Sigmoid). Note that the final layer (layer  $L$ ) has no activation function, a common configuration for regression tasks. To simplify the notation, we can absorb the bias vector  $b_k$  into the weight matrix  $W_k$ . This is achieved by defining an augmented weight matrix  $\tilde{W}_k = (W_k, b_k) \in \mathbb{R}^{q_k \times (q_{k-1} + 1)}$  and appending a 1 to the input of each layer. For instance, the network's input  $z$  is augmented to  $\tilde{z} = (z^\top, 1)^\top$ . This requires a corresponding modification of the activation functions. For each hidden layer  $k = 1, \dots, L-1$ , we define an operator  $\phi_k$  that first applies the activation  $\sigma_k$  and then appends a 1 to the resulting vector:

$$\phi_k(v) = (\sigma_k(v)^\top, 1)^\top.$$

With these definitions, the neural network in (6) can be expressed more compactly as a composition of matrix-vector products and activation operators:

$$g(z) = \tilde{W}_L \circ \phi_{L-1} \circ \tilde{W}_{L-1} \circ \dots \circ \phi_1 \circ \tilde{W}_1(\tilde{z}). \quad (7)$$

Note that the total number of parameters in (7) is  $\sum_{k=1}^L q_k(q_{k-1} + 1)$ , which can be very large and may lead to overfitting. For  $s \in \mathbb{N}$ ,  $L \geq 2$ ,  $A > 0$  and  $\mathbf{q} = (q_0, q_1, \dots, q_L)^\top$ , we consider a sparsely connected neural network class

$$\begin{aligned} \mathcal{M}(s, L, \mathbf{q}, A) = \left\{ g(z) = W_L \phi_{L-1} \circ \dots \circ W_2 \phi_1(W_1 \tilde{z}) \mid W_k \in \mathbb{R}^{q_k \times (q_{k-1} + 1)}, \|W_k\|_\infty \leq 1 \text{ for } \right. \\ \left. k = 1, \dots, L, \sum_{k=1}^L \|W_k\|_0 \leq s \text{ and } \|g\|_\infty \leq A \right\}, \end{aligned} \quad (8)$$

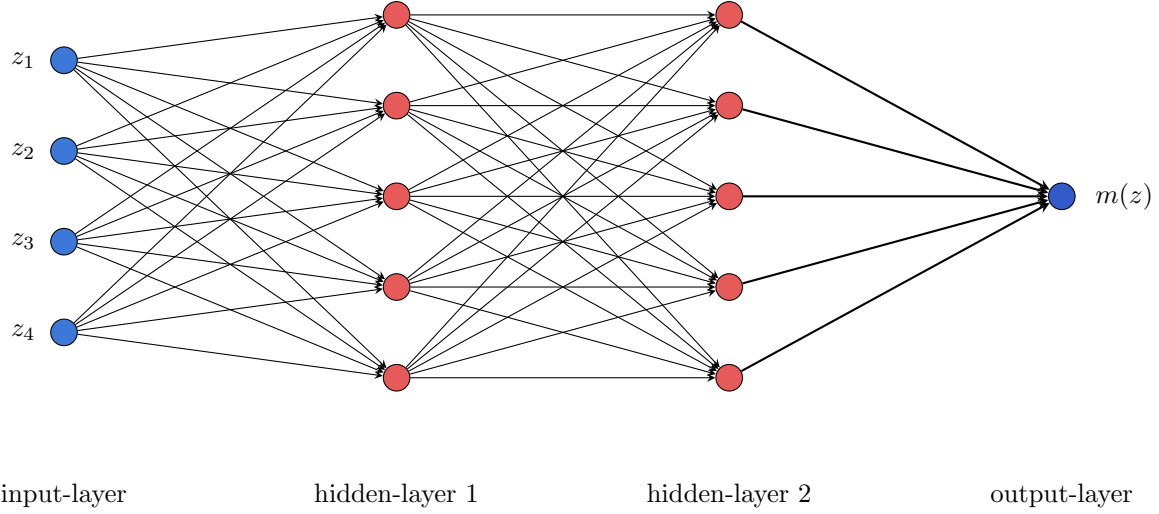


Figure 1: A 3-layer neural network with four input variables and one output.

where  $\|\cdot\|_\infty$  is the sup-norm of a matrix or function and  $\|\cdot\|_0$  is the number of non-zero elements of a matrix.

### 3 Statistical Perspective

To solve (2), we parametrize the nonlinear term  $g$  by DNN and rewrite the problem (2) as

$$\min_{\beta \in \mathbb{R}_C^d, g \in \mathcal{M}(s, L, q, C)} \frac{1}{N} \sum_{i=1}^N |Y_i - \beta^\top X_i - g(Z_i)| + \lambda_N \mathcal{J}_{N, M}(\beta, g) \quad (9)$$

with parameters of DNN  $\mathbf{W} := (W_1, \dots, W_L)$ . Due to the DNN structure of  $\mathcal{M}$ , (9) can be rewritten into the following finite-dimension form:

$$\begin{aligned} \min_{\beta, \mathbf{W}} \quad & \frac{1}{N} \sum_{i=1}^N |Y_i - \beta^\top X_i - g(\mathbf{W}; Z_i)| + \lambda_N \mathcal{J}_{N, M}(\beta, g(\mathbf{W})), \\ \text{s.t.} \quad & \|W_k\|_\infty \leq 1, \text{ for } k = 1, \dots, L, \\ & \sum_{k=1}^L \|W_k\|_0 \leq s \end{aligned} \quad (10)$$

for given  $L, s$ .  $\mathcal{J}_{N, M}(\beta, g(\mathbf{W}))$  (or equivalently denoted as  $\mathcal{J}_{N, M}(\beta, \mathbf{W})$ ) can be any bounded regulation term. For example  $\mathcal{J}_{N, M}(\beta, g(\mathbf{W})) = |\partial_Z g(\mathbf{W}; Z)| \wedge M$ .

**Remark 1.** *The norm of the Jacobian of  $g$  with respect to the input  $Z$  is used as a penalty term in DNN training to prevent overfitting and improve robustness against input data corruption [21]. Furthermore, this Jacobian term can be efficiently computed using the standard backpropagation algorithm [24].*

For some  $J \in \mathbb{N}$ ,  $\gamma = (\gamma_1, \dots, \gamma_J) \in \mathbb{R}_+^J$ ,  $\mathbf{d} = (q, d_1, \dots, d_J)^\top \in \mathbb{N}^{J+1}$  and  $\bar{\mathbf{d}} = (\bar{d}_1, \dots, \bar{d}_J)^\top \in \mathbb{N}^J$  with  $\bar{d}_1 \leq q$  and  $\bar{d}_k \leq d_{k-1}, k = 2, \dots, J$ , we define the *effective smoothness*  $\bar{\gamma}_k = \gamma_k \prod_{i=k+1}^J (\gamma_i \wedge 1)$  of a function  $h$  in  $\mathcal{H}(J, \gamma, \mathbf{d}, \bar{\mathbf{d}}, B)$ , and write

$$\zeta = \min_{k \in \{1, \dots, J\}} \frac{\bar{\gamma}_k}{2\bar{\gamma}_k + \bar{d}_k} \text{ and } r_N = N^{-\zeta}.$$

For the covariate  $X = (X_{(1)}, \dots, X_{(d)})^\top$ , we define

$$\varphi_k^* = \arg \min_{\varphi \in L^2(P_Z)} \mathbb{E}[f_\varepsilon(0|V)\{X_{(k)} - \varphi(Z)\}^2], k = 1, \dots, d, \quad (11)$$

where  $L^2(P_Z) = \{\varphi \mid \mathbb{E}\varphi^2(Z) < \infty\}$ . And denote  $\varphi^*(Z) = (\varphi_1^*(Z), \dots, \varphi_d^*(Z))^\top$ ,  $\Sigma_1 = \mathbb{E}[\{X - \varphi^*(Z)\}\{X - \varphi^*(Z)\}^\top]$  and  $\Sigma_2 = \mathbb{E}[f_\varepsilon(0|X, Z)\{X - \varphi^*(Z)\}\{X - \varphi^*(Z)\}^\top]$ .

We consider the joint probability space  $(\Omega, \mathcal{F}, \hat{P})$  for the random variables  $(X, Z, \varepsilon)$ . Here, the sample space  $\Omega = \Omega_X \times \Omega_Z \times \Omega_\varepsilon$  is the product of their sample spaces,  $\mathcal{F}$  is the corresponding product  $\sigma$ -algebra, and  $\hat{P}$  is their joint probability measure. Furthermore, we define the probability space  $(\Omega^\infty := \prod_{i=1}^\infty \Omega^i, \mathcal{F}^\infty := \sigma(\prod_{i=1}^\infty \mathcal{F}^i), P)$  for the sequence of samples  $((X_i, Z_i, \varepsilon_i))_{i=1}^\infty$ , where  $\Omega^\infty$  denotes the product space and  $\mathcal{F}^\infty$  is the corresponding product  $\sigma$ -algebra. Furthermore, we assume  $P(\varepsilon \leq 0) = \frac{1}{2}$ . For simplicity in the subsequent analysis, we let  $\mathbb{E}[\cdot] = \mathbb{E}_U[\cdot]$ .

Let  $\mathbb{R}_C^d = \{\beta \in \mathbb{R}^d \mid \|\beta\|_\infty < C\}$ , we define

$$(\hat{\beta}_N, \hat{g}_N) \in \arg \min_{\beta \in \mathbb{R}_C^d, g \in \mathcal{M}_C^N} L_N(\theta) + \lambda_N \mathcal{J}_{N,M}(\beta, g) \quad (12)$$

with  $L_N(\theta) := \frac{1}{N} \sum_{i=1}^N |Y_i - \beta^\top X_i - g(Z_i)|$ .

**Assumption 1.** *We introduce the following assumptions.*

(A1) *The true vector parameter  $\beta_0$  belongs to a compact subset  $\mathbb{R}_C^d := \{\beta \in \mathbb{R}^d \mid \|\beta\|_\infty < C\}$  and the true nonparametric function  $g_0$  satisfies  $\|g_0\|_\infty < C$  and belongs to  $\mathcal{H} = \mathcal{H}(J, \gamma, \mathbf{d}, \bar{\mathbf{d}}, B)$ .*

(A2) *There exists a constant  $A_0 > 0$  s.t.  $\sigma_{\min}(\Sigma_2) > A_0$  and  $\sigma_{\min}(\mathbb{E}[(X - \mathbb{E}[X|Z])(X - \mathbb{E}[X|Z])^\top]) > A_0$ .*

(A3) *The covariates  $V = (X, Z)$  take values in a compact subset of  $\mathbb{R}^{d+l}$  that, without loss of generality, will be assumed to be  $[0, 1]^{d+l}$ . In addition, the probability density function (PDF) of  $Z$  is bounded away from zero and from infinity.*

(A4)  *$L = O(\log N)$ ,  $s = O(Nr_N^2 \log N)$ ,  $\lambda_N = o(1)$  and*

$$Nr_N^2 \lesssim \min_{k=1, \dots, L} \{q_k\} \leq \max_{k=1, \dots, L} \{q_k\} \lesssim N.$$

(A5) *The conditional PDF  $f_\varepsilon(\cdot|v)$  of the random error  $\varepsilon$  given the covariate  $V = v$ , has continuous derivative  $f'_\varepsilon(\cdot|v)$ , and there exist positive constants  $b_0$  and  $c_0$  such that  $1/c_0 < f_\varepsilon(t|v) < c_0$  and  $|f'_\varepsilon(t|v)| < d_0$  for all  $|t| \leq b_0, v \in [0, 1]^{d+l}$ . Furthermore, we assume  $\mathbb{E}[|\varepsilon||V = v] < \infty$  for any  $v \in [0, 1]^{d+l}$ .*

(A6) *For any  $k \in \{1, \dots, J\}$ ,  $\bar{\gamma}_k > \bar{d}_k/2$ , and  $\mathbb{E}[\|X\|^2] < \infty$ .*

(A7) *In addition, there exists  $B_f > 0$  such that  $f(t|v) \leq B_f$  for all  $t \in \mathbb{R}$  and  $v \in [0, 1]^{d+l}$ .*

(A8)  *$\mathcal{J}_{N,M}$  is separable i.e.  $\mathcal{J}_{N,M}(\beta, g) = \mathcal{J}_{N,1}(\beta) + \mathcal{J}_{N,2}(g)$ ;  $\mathcal{J}_{N,M}$  is lower semi continuous (l.s.c.) and SNEC on  $\mathbb{R}_C^d \times \mathcal{M}_C^N$ , and let the qualification condition*

$$\left[ (0, \mathbf{v}^*) \in \partial_{(\xi, h)}^\infty \mathcal{J}_{N,M}(\beta, g) \right] \implies \mathbf{v}^* = 0$$

*with  $\xi = \beta - \beta_0$ ,  $h(Z) = g(Z) - g_0(Z) + (\beta - \beta_0)^\top \varphi^*(Z)$  for any  $(\xi, h) \in \mathbb{R}_C^d \times \mathcal{M}_C^N$ ;  $\lambda_N (\|\partial_g \mathcal{J}_{N,2}\|_* + \|\partial_\beta \mathcal{J}_{N,1}\|) = o_p\left(\frac{1}{\sqrt{N}}\right)$ . Here,  $\|\cdot\|_*$  denotes the operator norm in  $L_2(\mathbf{m})$  space with Lebesgue measure  $\mathbf{m}$ .*

**Theorem 1.** Suppose Assumptions (A1)-(A5) hold. Then the estimators  $\hat{\beta}_N$  and  $\hat{g}_N$  from optimization problem (12) exhibit the following rates of convergence:

$$\begin{aligned}\|\hat{\beta}_N - \beta_0\|_\infty &= O_p(r_N \log^2 N + \lambda_N), \\ \|\hat{g}_N - g_0\|_{L^2(P)} &= O_p(r_N \log^2 N + \lambda_N).\end{aligned}$$

*Proof.* Let  $\hat{\theta}_N = (\hat{\beta}_N, \hat{g}_N)$ ,  $\theta_0 = (\beta_0, g_0)$  and  $d(\theta_1, \theta_2) = [\mathbb{E}\{X^\top \beta_1 + g_1(Z) - X^\top \beta_2 - g_2(Z)\}^2]^{1/2}$ , for any  $\theta_1 = (\beta_1, g_1)$  and  $\theta_2 = (\beta_2, g_2)$ . We first show that

$$d(\hat{\theta}_N, \theta_0) \xrightarrow{P} 0, \text{ as } N \rightarrow \infty. \quad (13)$$

We first show that

$$\sup_{\theta \in \mathbb{R}_C^d \times \mathcal{M}_C^N} |L_N(\theta) + \lambda_N \mathcal{J}_{N,M}(\theta) - L_0(\theta)| \xrightarrow{P} 0, \text{ as } N \rightarrow \infty. \quad (14)$$

with  $L_0(\theta) := \mathbb{E}[Y - \beta^\top X - g(Z)]$ . Based on the assumption of  $\lambda_N$  and the definition of  $\mathcal{J}_{N,M}(\theta)$ , it suffices to show

$$\sup_{\theta \in \mathbb{R}_C^d \times \mathcal{M}_C^N} |L_N(\theta) - L_0(\theta)| \xrightarrow{P} 0, \text{ as } N \rightarrow \infty. \quad (15)$$

Denote  $\mathcal{F}_N := \{f(x, y, z) := |y - \beta^\top x - g(z)|, \forall (x, y, z) \in \mathbb{R} \times [0, 1]^{d+l} \mid (\beta, g) \in \mathbb{R}_C^d \times \mathcal{M}_C^N\}$ . Notably,  $F(x, y, z) := |y| + 2C$  is an envelope function of  $\mathcal{F}_N$  with  $\mathbb{E}F < \infty$ . Based on Theorem 2.4.6 of [43], it holds that

$$\begin{aligned}& \mathbb{E}^* \|\mathbb{P}_N - P\|_{\mathcal{F}_N} \\& \stackrel{(I)}{\leq} \mathbb{E}_{U,e}^* \left\| \frac{1}{N} \sum_{i=1}^N e_i f(U_i) \right\|_{\mathcal{F}_N} \\& \stackrel{(II)}{=} 2\mathbb{E}\mathbb{E}_e \left\| \frac{1}{N} \sum_{i=1}^N e_i f(U_i) \right\|_{\mathcal{F}_N} \\& \stackrel{(III)}{\leq} 2\mathbb{E}\mathbb{E}_e \left\| \frac{1}{N} \sum_{i=1}^N e_i f(U_i) \right\|_{\mathcal{F}_N \wedge q} + \mathbb{E}F\mathbf{1}\{F > q\} \\& \stackrel{(IV)}{\leq} 2\mathbb{E} \left\{ \sqrt{1 + \log N(\epsilon, \mathcal{F}_N \wedge q, L_1(\mathbb{P}_N))} \sup_{f \in \mathcal{G}_N} \left\| \frac{1}{N} \sum_{i=1}^N e_i f(U_i) \right\|_{\psi_2|U} \right\} + \underbrace{2\epsilon + \epsilon_F}_{:=\hat{\epsilon}} \\& \leq 2\mathbb{E} \left\{ \sqrt{1 + \log N(\epsilon, \mathcal{F}_N \wedge q, L_1(\mathbb{P}_N))} \sqrt{\frac{6}{N}q} \right\} + \hat{\epsilon}\end{aligned} \quad (16)$$

with  $\mathcal{F}_N \wedge q = \{f \wedge q \mid f \in \mathcal{F}_N\}$ . Here, inequality (I) holds due to Lemma 2.3.1 of [43], where  $e_1, \dots, e_N$  are i.i.d. Rademacher random variables. The inequality (II) holds due to the measurability of  $\left\| \frac{1}{N} \sum_{i=1}^N e_i f(U_i) \right\|_{\mathcal{F}_N}$  and Fubini Theorem. Indeed, Schmidt-Hieber (2020) proves

$$\log N(\epsilon, \mathcal{M}_C^N, \|\cdot\|_\infty) \leq (s+1) \log \left( \frac{2H^2(L+1)}{\epsilon} \right) \quad (17)$$

with  $H := \prod_{k=1}^L (q_k + 1)$ ; see Lemma 5 of [39] for details. Combining with the fact that  $\beta \in \mathbb{R}_C^d$ , we may conclude that there exists a countable dense subset  $\{\tilde{f}_i\}_{i \geq 0}$  of  $\mathcal{F}_N$ . Thus,  $\left\| \frac{1}{N} \sum_{i=1}^N e_i f(U_i) \right\|_{\mathcal{F}_N}$  equals  $\left\| \frac{1}{N} \sum_{i=1}^N e_i f(U_i) \right\|_{\{\tilde{f}_i\}_{i \geq 0}}$ , and is of course  $P$ -measurable. Therefore, in this paper, we no need to distinguish the outer measure (expectation) and classic measure (expectation). Furthermore, we denote  $\mathcal{G}_N$  is an  $\epsilon$ -net in  $L_1(\mathbb{P}_N)$  over  $\mathcal{F}_N \wedge q$ , while inequality (IV) holds for any  $\epsilon_F > 0$  by selecting sufficiently large  $q$ .

Due to the triangle inequality,  $\forall f_1, f_2 \in \mathcal{F}_N \wedge q$ , we have

$$\|f_1 - f_2\|_\infty \leq \sup_{\forall x, z \in [0,1]^{d+l}} |g_1(z) - g_2(z) + (\beta_1 - \beta_2)^\top x| \leq \|g_1 - g_2\|_\infty + \|\beta_1 - \beta_2\|_\infty.$$

Hence, by (17), it is easy to show that

$$\begin{aligned} N(\epsilon, \mathcal{F}_N \wedge q, L_1(\mathbb{P}_N)) &\leq N(\epsilon, \mathcal{F}_N \wedge q, \|\cdot\|_\infty) \\ &\leq N\left(\frac{\epsilon}{2}, \mathcal{M}_C^N, \|\cdot\|_\infty\right) N\left(\frac{\epsilon}{2}, \mathbb{R}_C^d, \|\cdot\|_\infty\right) \\ &\leq K_\epsilon N\left(\frac{\epsilon}{2}, \mathcal{M}_C^N, \|\cdot\|_\infty\right), \end{aligned} \tag{18}$$

where the first inequality holds due to  $\|f\|_{L^1(\mathbb{P}_N)} \leq \|f\|_\infty$ , for any  $f \in \mathcal{F}_N \wedge q$  and discrete measure  $\mathbb{P}_N$  combining with the nature of  $\epsilon$ -net. Moreover, the second inequality holds due to Heine–Borel Theorem in which  $K_\epsilon = N\left(\frac{\epsilon}{2}, \mathbb{R}_C^d, \|\cdot\|_\infty\right)$ . Substituting (18) into (16), it holds that

$$\begin{aligned} \mathbb{E}\|\mathbb{P}_N - P\|_{\mathcal{F}_N} &\leq 2\mathbb{E}\left\{\sqrt{1 + \log\left(K_\epsilon N\left(\frac{\epsilon}{2}, \mathcal{M}_C^N, \|\cdot\|_\infty\right)\right)} \sqrt{\frac{6}{N}q}\right\} + \hat{\epsilon} \\ &\leq 2\mathbb{E}\left\{\sqrt{1 + \log(K_\epsilon) + (s+1)\log\left(\frac{4H^2(L+1)}{\epsilon}\right)} \sqrt{\frac{6}{N}q}\right\} + \hat{\epsilon} \quad (\text{By (17)}). \end{aligned}$$

Due to Assumption (A4), the integrand is

$$\sqrt{1 + \log(K_\epsilon) + (s+1)\log\left(\frac{4H^2(L+1)}{\epsilon}\right)} \sqrt{\frac{6}{N}q} = O(r_N \log^{\frac{3}{2}} N) = o(1).$$

This completes the proof of (15).

We now prove

$$\inf_{d(\theta, \theta_0) > \epsilon, \theta \in \mathbb{R}_C^d \times \mathcal{M}_C^\infty} L_0(\theta) > L_0(\theta_0) \tag{19}$$

with  $\mathcal{M}_C^\infty = \cup_{i=1}^\infty \mathcal{M}_C^i$ . According to the equation (C.46) of [3], for any two scalars  $a, b$ , it holds that

$$|a - b| - |a| = -b \left( \frac{1}{2} - \mathbf{1}\{a \leq 0\} \right) + \int_0^b (\mathbf{1}\{a \leq t\} - \mathbf{1}\{a \leq 0\}) dt. \tag{20}$$

For any  $\theta \in \mathbb{R}_C^d \times \mathbf{conv}(\mathcal{M}_C^\infty)$ , we denote  $\Lambda(\theta; V) := X^\top \beta + g(Z)$  and  $\Lambda(\theta_0; V) := X^\top \beta_0 + g_0(Z)$ . Taking  $a = Y - \Lambda(\theta_0; V)$  and  $b = \Lambda(\theta; V) - \Lambda(\theta_0; V)$  into (20), we have

$$\begin{aligned} &L_0(\theta) - L_0(\theta_0) \\ &= \mathbb{E} \left[ -b \left( \frac{1}{2} - \mathbf{1}\{a \leq 0\} \right) + \int_0^b (\mathbf{1}\{a \leq t\} - \mathbf{1}\{a \leq 0\}) dt \right] \\ &\stackrel{(V)}{=} \mathbb{E} \left[ \int_0^b (\mathbf{1}\{a \leq t\} - \mathbf{1}\{a \leq 0\}) dt \right] \\ &= \mathbb{E}_V \left[ \mathbb{E}_Y \left[ \int_0^{\Lambda(\theta; V) - \Lambda(\theta_0; V)} \mathbf{1}\{Y - \Lambda(\theta_0; V) \leq t\} - \mathbf{1}\{Y - \Lambda(\theta_0; V) \leq 0\} dt \middle| V \right] \right] \\ &= \mathbb{E}_V \left[ \int_0^{\Lambda(\theta; V) - \Lambda(\theta_0; V)} F_{Y|V}(\Lambda(\theta_0; V) + t) - F_{Y|V}(\Lambda(\theta_0; V)) dt \right] \\ &= \mathbb{E}_V \left[ \int_0^{\Lambda(\theta; V) - \Lambda(\theta_0; V)} t f_{Y|V}(\Lambda(\theta_0; V)) + \frac{t^2}{2} f'_{Y|V}(\Lambda(\theta_0; V) + \bar{t}_{V,t}) dt \right] \\ &\stackrel{(VI)}{\geq} \frac{1}{2c_0} \mathbb{E}_V \left[ |\Lambda(\theta; V) - \Lambda(\theta_0; V)|^2 \right] - \frac{d_0}{6} \mathbb{E}_V \left[ |\Lambda(\theta; V) - \Lambda(\theta_0; V)|^3 \right] \end{aligned} \tag{21}$$

with  $\bar{t}_{V,t}$  between 0 and  $t$ . The equality (V) holds due to

$$\begin{aligned}
& \mathbb{E} \left[ -b \left( \frac{1}{2} - \mathbf{1}\{a \leq 0\} \right) \right] \\
&= \mathbb{E}_V \left[ \mathbb{E}_Y \left[ (\Lambda(\theta; V) - \Lambda(\theta_0; V)) \left( \frac{1}{2} - \mathbf{1}\{Y - \Lambda(\theta_0; V) \leq 0\} \right) \middle| V \right] \right] \\
&= \mathbb{E}_V \left[ \mathbb{E}_Y \left[ (\Lambda(\theta; V) - \Lambda(\theta_0; V)) \left( \frac{1}{2} - \mathbf{1}\{Y - \Lambda(\theta_0; V) \leq 0\} \right) \middle| V \right] \right] \\
&= \mathbb{E}_V \left[ (\Lambda(\theta; V) - \Lambda(\theta_0; V)) \mathbb{E}_Y \left[ \left( \frac{1}{2} - \mathbf{1}\{\varepsilon \leq 0\} \right) \middle| V \right] \right] \\
&= 0,
\end{aligned}$$

while the inequality (VI) holds due to Assumption (A1) and (A5).

Let

$$\bar{q}(\theta) := \frac{\left( \frac{1}{c_0} \right)^{\frac{3}{2}} \mathbb{E}_V \left[ |\Lambda(\theta; V) - \Lambda(\theta_0; V)|^2 \right]^{\frac{3}{2}}}{d_0 \mathbb{E}_V \left[ |\Lambda(\theta; V) - \Lambda(\theta_0; V)|^3 \right]}$$

and consider the case  $\left( \frac{1}{c_0} \mathbb{E}_V [|\Lambda(\theta; V) - \Lambda(\theta_0; V)|^2] \right)^{\frac{1}{2}} \leq \bar{q}(\theta)$ . It holds that

$$d_0 \mathbb{E}_V \left[ |\Lambda(\theta; V) - \Lambda(\theta_0; V)|^3 \right] \leq \frac{1}{c_0} \mathbb{E}_V \left[ |\Lambda(\theta; V) - \Lambda(\theta_0; V)|^2 \right].$$

Then we have

$$L_0(\theta) - L_0(\theta_0) \geq \frac{1}{3c_0} \mathbb{E}_V \left[ |\Lambda(\theta; V) - \Lambda(\theta_0; V)|^2 \right] = \frac{1}{3c_0} d(\theta, \theta_0)^2. \quad (22)$$

Next, we consider the case  $\left( \frac{1}{c_0} \mathbb{E}_V \left[ |\Lambda(\theta; V) - \Lambda(\theta_0; V)|^2 \right] \right)^{\frac{1}{2}} > \bar{q}(\theta)$ . Let  $\tilde{\theta} = ((1 - \alpha)\beta + \alpha\beta_0, (1 - \alpha)g + \alpha g_0)$  such that  $\left( \frac{1}{c_0} \mathbb{E}_V [|\Lambda(\tilde{\theta}; V) - \Lambda(\theta_0; V)|^2] \right)^{\frac{1}{2}} = \bar{q}(\theta)$ . Then it holds that  $1 - \alpha = \frac{\bar{q}(\theta)}{\sqrt{\frac{1}{c_0} \mathbb{E}_V [|\Lambda(\tilde{\theta}; V) - \Lambda(\theta_0; V)|^2]}}$ . On the other hand, we have

$$L_0(\theta) - L_0(\theta_0) \geq \frac{L_0(\tilde{\theta}) - L_0(\theta_0)}{1 - \alpha} = \frac{\sqrt{\frac{1}{c_0} \mathbb{E}_V \left[ |\Lambda(\tilde{\theta}; V) - \Lambda(\theta_0; V)|^2 \right]}}{\bar{q}(\theta)} (L_0(\tilde{\theta}) - L_0(\theta_0)). \quad (23)$$

Note that

$$\begin{aligned}
\bar{q}(\theta) &= \frac{\left( \frac{1}{c_0} \right)^{\frac{3}{2}} \mathbb{E}_V \left[ |\Lambda(\theta; V) - \Lambda(\theta_0; V)|^2 \right]^{\frac{3}{2}}}{d_0 \mathbb{E}_V \left[ |\Lambda(\theta; V) - \Lambda(\theta_0; V)|^3 \right]} \\
&= \frac{\left( \frac{1}{c_0} \right)^{\frac{3}{2}} \mathbb{E}_V \left[ |\Lambda(\tilde{\theta}; V) - \Lambda(\theta_0; V)|^2 \right]^{\frac{3}{2}}}{d_0 \mathbb{E}_V \left[ |\Lambda(\tilde{\theta}; V) - \Lambda(\theta_0; V)|^3 \right]} \\
&= \frac{\bar{q}^3(\theta)}{d_0 \mathbb{E}_V \left[ |\Lambda(\tilde{\theta}; V) - \Lambda(\theta_0; V)|^3 \right]}.
\end{aligned}$$

Then we have  $d_0 \mathbb{E}_V \left[ \left| \Lambda(\tilde{\theta}; V) - \Lambda(\theta_0; V) \right|^3 \right] = \bar{q}^2(\theta)$ . Furthermore, by (21), it holds that

$$L_0(\tilde{\theta}) - L_0(\theta_0) \geq \frac{1}{2c_0} \mathbb{E}_V \left[ |\Lambda(\tilde{\theta}; V) - \Lambda(\theta_0; V)|^2 \right] - \frac{d_0}{6} \mathbb{E}_V \left[ |\Lambda(\tilde{\theta}; V) - \Lambda(\theta_0; V)|^3 \right] = \frac{1}{3} \bar{q}^2(\theta). \quad (24)$$

Substituting (24) into (23), we have

$$L_0(\theta) - L_0(\theta_0) \geq \frac{\bar{q}(\theta)}{3\sqrt{c_0}} d(\theta, \theta_0). \quad (25)$$

Combining (22) and (25), we complete the proof of (19).

We now finish the proof of the consistency. For the function  $g_0$ , let

$$g_N^* := \operatorname{argmin}_{g \in \mathcal{M}_C^N} \|g - g_0\|_{L_2} \text{ and } \theta_N^* := (\beta_0, g_N^*),$$

while Schmidt-Hieber (2020) proves

$$d(\theta_N^*, \theta_0) = O(r_N) \rightarrow 0, \text{ as } N \rightarrow \infty; \quad (26)$$

see Equation (26) of [39] for details. Due to the definition of  $d(\cdot, \cdot)$ , it holds that

$$\begin{aligned} |L_0(\theta_N^*) - L_0(\theta_0)| &= |\mathbb{E} [Y - \beta_0^\top X - g_N^*(Z)] - \mathbb{E} [Y - \beta_0^\top X - g_0(Z)]| \\ &\leq \mathbb{E} [|g_N^*(Z) - g_0(Z)|] \\ &\leq d(\theta_N^*, \theta_0). \end{aligned}$$

Combining with (26), we have

$$L_0(\theta_N^*) \leq L_0(\theta_0) + o(1). \quad (27)$$

On the other hand, it is easy to show that

$$\begin{aligned} L_N(\hat{\theta}_N) + \lambda_N \mathcal{J}_{N,M}(\hat{\theta}_N) &\leq L_N(\theta_N^*) + \lambda_N \mathcal{J}_{N,M}(\theta_N^*) \\ L_0(\hat{\theta}_N) + L_N(\hat{\theta}_N) + \lambda_N \mathcal{J}_{N,M}(\hat{\theta}_N) - L_0(\hat{\theta}_N) &\leq L_0(\theta_N^*) + L_N(\theta_N^*) + \lambda_N \mathcal{J}_{N,M}(\theta_N^*) - L_0(\theta_N^*). \end{aligned} \quad (28)$$

Applying (15) to (28), we have

$$\begin{aligned} L_0(\hat{\theta}_N) &\leq L_0(\theta_N^*) + |L_N(\hat{\theta}_N) + \lambda_N \mathcal{J}_{N,M}(\hat{\theta}_N) - L_0(\hat{\theta}_N)| + |L_N(\theta_N^*) + \lambda_N \mathcal{J}_{N,M}(\theta_N^*) - L_0(\theta_N^*)| \\ &\leq L_0(\theta_N^*) + 2 \sup_{\theta \in \mathbb{R}_C^d \times \mathcal{M}_C^N} |L_N(\theta) + \lambda_N \mathcal{J}_{N,M}(\theta) - L_0(\theta)| \\ &\stackrel{(15)}{\leq} L_0(\theta_N^*) + o_p(1). \end{aligned}$$

Combining with (27), it holds that

$$L_0(\hat{\theta}_N) \leq L_0(\theta_0) + o(1) + o_p(1) \leq L_0(\theta_0) + o_p(1). \quad (29)$$

If there exists  $\epsilon_1 > 0$  such that  $P(d(\hat{\theta}_N, \theta_0) > \epsilon_1) > 0$ , then, based on (19),

$$P(L_0(\hat{\theta}_N) > L_0(\theta_0) + \epsilon_2) > P(d(\hat{\theta}_N, \theta_0) > \epsilon_1) > 0$$

holds for some  $\epsilon_2 > 0$ , which contradicts with (29). This completes the proof of (13).

We now prove that

$$d(\hat{\theta}_N, \theta_0) = O_p(r_N \log^2 N + \lambda_N), \quad (30)$$

following the line of Theorem 3.4.6 of [43]. We set the parameters  $(\{\theta_n\}_{n=1}^\infty, \{\theta_{n,0}\}_{n=1}^\infty, c, \{\underline{\delta}_n\}_{n=1}^\infty, \{\lambda_n\}_{n=1}^\infty)$  of Theorem 3.4.6 of [43] to our counterparts  $(\{\theta_N^*\}_{N=1}^\infty, \{\theta_0\}_{N=1}^\infty, 0, \{0\}_{N=1}^\infty, \{\lambda_N\}_{N=1}^\infty)$ , respectively. Furthermore, we write  $R := 2H^2(L+1)$  and

$$\mathcal{A}_\delta^N = \{\theta \in \mathbb{R}_C^d \times \mathcal{M}_C^N \mid d(\theta, \theta_0) \leq \delta\}. \quad (31)$$

By Theorem 3.4.6 of [43], it suffices to verify that, for any  $\delta > 0$ ,

$$\delta^2 \lesssim \inf_{\theta \in \mathbb{R}_C^d \times \mathcal{M}_C^N : \frac{\delta}{2} < d(\theta, \theta_0) \leq \delta} L_0(\theta) - L_0(\theta_0), \quad (32)$$

$$\mathbb{E} \left[ \sup_{\substack{\theta \in \mathcal{A}_\delta^N, \\ \mathcal{J}_{N,M}(\theta) < \delta/\lambda_N}} \sqrt{N} |(L_N - L_0)(\theta_N^*) - (L_N - L_0)(\theta)| \right] \lesssim \phi_N(\delta) \quad (33)$$

with  $\phi_N(\delta) = \delta \sqrt{s \log \frac{R}{\delta}} + \frac{s}{\sqrt{N}} \log \frac{R}{\delta}$ .

Indeed, for every  $\tilde{M} > 0$  there exists a constant  $\gamma_{\tilde{M}} > 0$  such that  $\mathbb{E}_\varepsilon[|\varepsilon|] - \mathbb{E}_\varepsilon[|\varepsilon + \mu|] \leq -\gamma_{\tilde{M}} |\mu|^2$  for  $|\mu| \leq \tilde{M}$ . Then for any  $(\beta, g) \in \mathbb{R}_C^d \times \mathcal{M}_C^\infty$ , we have

$$L_0(\theta_0) - L_0(\theta) \lesssim -\gamma_{\tilde{M}} d^2(\theta, \theta_0) \quad (34)$$

with  $\tilde{M} := \sup_{(\beta, g) \in \mathbb{R}_C^d \times \mathcal{M}_C^\infty} 2\|\beta\|_\infty + 2\|g\|_\infty$ . Then, (32) holds by taking  $\frac{\delta}{2} < d(\theta, \theta_0)$  into (34).

We now verify (33). Denote  $\rho(\theta; U) := |Y - \beta^\top X - g(Z)|, \forall U = (X, Y, Z) \in \mathbb{R} \times [0, 1]^{d+l}$  and  $\mathcal{B}_\delta^N = \{\rho(\theta_N^*; U) - \rho(\theta; U) \mid \theta \in \mathcal{A}_\delta^N\}$ . For any  $\theta, \theta_1 \in \mathcal{A}_\delta^N$ , we have  $\mathbb{E}|\rho(\theta; U) - \rho(\theta_1; U)|^2 \leq 4d^2(\theta, \theta_1)$ . Lemma 5 of [39] then implies that

$$\begin{aligned} \log(N_{[\cdot]}(\epsilon, \mathcal{B}_\delta^N, L^2(P))) &\leq \log(N_{[\cdot]}(\epsilon, \mathcal{B}_\delta^N, \|\cdot\|_\infty)) \\ &\stackrel{\text{(VII)}}{\leq} \log(N(\epsilon, \mathcal{B}_\delta^N, \|\cdot\|_\infty)) \\ &\leq \log(N(\epsilon, \mathcal{F}_N, \|\cdot\|_\infty)) \\ &\leq \log\left(K_\epsilon N\left(\frac{\epsilon}{2}, \mathcal{M}_C^N, \|\cdot\|_\infty\right)\right) \quad (\text{by (18)}) \\ &\lesssim \log(K_\epsilon) + s \log \frac{R}{\epsilon} \quad (\text{By (17)}) \end{aligned} \quad (35)$$

where  $N_{[\cdot]}(\epsilon, \mathcal{B}_\delta^N, L^2(P))$  ( $N_{[\cdot]}(\epsilon, \mathcal{B}_\delta^N, \|\cdot\|_\infty)$ ) is the bracket number of  $\mathcal{B}_\delta^N$  with  $L^2(P)$  norm ( $L^\infty$  norm). The inequality (VII) holds due to Page 132 of [43]. Henceforth, it follows that

$$J_{[\cdot]}(\delta, \mathcal{B}_\delta^N) = \int_0^\delta \sqrt{1 + \log(N_{[\cdot]}(\epsilon, \mathcal{B}_\delta^N, L^2(P)))} d\epsilon \lesssim \delta \sqrt{s \log \frac{R}{\delta}},$$

where the last inequality holds by noticing

$$\int_0^\delta \sqrt{\log \frac{R}{\epsilon}} d\epsilon = \delta \sqrt{\log \frac{R}{\delta}} + \frac{R\sqrt{\pi}}{2} \operatorname{erfc}\left(\sqrt{\log \frac{R}{\delta}}\right) \lesssim \delta \sqrt{\log \frac{R}{\delta}} \quad (36)$$

with  $\operatorname{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-t^2} dt$ . By Lemma 3.4.2 of [43], we conclude that

$$\begin{aligned} &\mathbb{E} \left[ \sup_{\substack{\theta \in \mathcal{A}_\delta^N, \\ \mathcal{J}_{N,M}(\theta) < \delta/\lambda_N}} \sqrt{N} |(L_N - L_0)(\theta_N^*) - (L_N - L_0)(\theta)| \right] \\ &= \mathbb{E} \left[ \sup_{\substack{\theta \in \mathcal{A}_\delta^N, \\ \mathcal{J}_{N,M}(\theta) < \delta/\lambda_N}} |\mathbb{G}_N(\rho(\theta_N^*; U) - \rho(\theta; U))| \right] \\ &\lesssim J_{[\cdot]}(\delta, \mathcal{B}_\delta^N) \left\{ \frac{J_{[\cdot]}(\delta, \mathcal{B}_\delta^N)}{\delta^2 \sqrt{N}} + 1 \right\} \\ &= \phi_N(\delta). \end{aligned}$$



Setting  $\delta_N = \eta_N = r_N \log^2 N$  in Theorem 3.4.6 of [43], it can be verified that

$$\frac{1}{\eta_N^2} \phi_N(\eta_N) \lesssim \sqrt{N} \text{ and } L_N(\hat{\theta}_N) + \lambda_N \mathcal{J}_{N,M}(\hat{\theta}_N) \leq L_N(\theta_N^*) + \lambda_N \mathcal{J}_{N,M}(\theta_N^*). \quad (37)$$

Then, by Theorem 3.4.6 of [43], we obtain  $d(\hat{\theta}_N, \theta_0) = O_p(r_N \log^2 N + \lambda_N)$  and  $\mathcal{J}_{N,M}(\hat{\theta}_N) = O_p(\eta_N / \lambda_N + 1)$ .

Furthermore, by Assumption (A5),

$$\begin{aligned} d^2(\hat{\theta}_N, \theta_0) &= \mathbb{E}\{X^\top (\hat{\beta}_N - \beta_0) + \hat{g}_N(Z) - g_0(Z)\}^2 \\ &= \mathbb{E}[\{(X - \mathbb{E}[X|Z])^\top (\hat{\beta}_N - \beta_0) + (\hat{\beta}_N - \beta_0)^\top \mathbb{E}[X|Z] + \hat{g}_N(Z) - g_0(Z)\}^2] \\ &= \mathbb{E}[\{(X - \mathbb{E}[X|Z])^\top (\hat{\beta}_N - \beta_0)\}^2] \\ &\quad + \mathbb{E}[(\hat{\beta}_N - \beta_0)^\top \mathbb{E}[X|Z] + \hat{g}_N(Z) - g_0(Z)]^2. \end{aligned} \quad (38)$$

Since the matrix  $\mathbb{E}[\{X - \mathbb{E}[X|Z]\}\{X - \mathbb{E}[X|Z]\}^\top]$  is positive definite, it follows that  $\|\hat{\beta}_N - \beta_0\|_\infty = O_p(r_N \log^2 N + \lambda_N)$  and thus  $\|\hat{g}_N - g_0\|_{L^2(P)} = O_p(r_N \log^2 N + \lambda_N)$ . This completes the proof.  $\square$

**Theorem 2.** Let  $\mathcal{F}_{\delta_N} := \{f - g : f, g \in \mathcal{F}_N, \|f - g\|_{L^2(P)} \leq \delta_N\}$  with  $\delta_N = O(r_N \log^2 N)$ . Under Assumptions (A1)-(A7), it holds that

$$\|\mathbb{G}_N\|_{\mathcal{F}_{\delta_N}} \xrightarrow{P} 0, \quad \mathbb{G}_N(\hat{f}_N) \rightsquigarrow \mathcal{N}(0, \Sigma)$$

with  $\hat{f}_N = |Y - \hat{\beta}_N^\top X - \hat{g}_N(Z)|$  and  $\Sigma = \text{Var}_U(|Y - \beta_0^\top X - g_0(Z)|)$ .

*Proof.* By the Markov inequality and Lemma 2.3.1 of [43], it holds that

$$\begin{aligned} P(\|\mathbb{G}_N\|_{\mathcal{F}_{\delta_N}} > x) &\leq \frac{2}{x} \mathbb{E}_{U,e} \left\| \frac{1}{\sqrt{N}} \sum_{i=1}^N e_i f(U_i) \right\|_{\mathcal{F}_{\delta_N}} \\ &= \frac{2}{x} \mathbb{E} \mathbb{E}_e \left\| \frac{1}{\sqrt{N}} \sum_{i=1}^N e_i f(U_i) \right\|_{\mathcal{F}_{\delta_N}} \\ &\lesssim \frac{2}{x} \mathbb{E} \left[ \int_0^\infty \sqrt{\log D(\epsilon, \mathcal{F}_{\delta_N}, L^2(\mathbb{P}_N))} d\epsilon \right] \quad (\text{by Corollary 2.2.9 of [43]}) \\ &\lesssim \frac{2}{x} \mathbb{E} \left[ \int_0^\infty \sqrt{\log N(\epsilon, \mathcal{F}_{\delta_N}, L^2(\mathbb{P}_N))} d\epsilon \right] \\ &= \frac{2}{x} \mathbb{E} \left[ \int_0^{\Xi_N^2} \sqrt{\log N(\epsilon, \mathcal{F}_{\delta_N}, L^2(\mathbb{P}_N))} d\epsilon \right] \quad (\text{with } \Xi_N^2 := \left\| \frac{1}{N} \sum_{i=1}^N f^2(U_i) \right\|_{\mathcal{F}_{\delta_N}}) \\ &\stackrel{(I)}{\lesssim} \frac{2}{x} \mathbb{E} \left[ \int_0^{\Xi_N^2} \sqrt{\log N(\epsilon, \mathcal{F}_N, \|\cdot\|_\infty)} d\epsilon \right] \\ &\lesssim \frac{2}{x} \mathbb{E} \left[ \int_0^{\Xi_N^2} \sqrt{s \log \left( \frac{R}{\epsilon} \right)} d\epsilon \right], \end{aligned} \quad (39)$$

where according to  $N(\epsilon, \mathcal{F}_{\delta_N}, L^2(\mathbb{P}_N)) \lesssim N^2(\epsilon, \mathcal{F}_N, L^2(\mathbb{P}_N)) \leq N^2(\epsilon, \mathcal{F}_N, \|\cdot\|_\infty)$ , the inequality (I) holds.

Note that

$$\Xi_N^2 = \|\mathbb{P}_N f^2\|_{\mathcal{F}_{\delta_N}} = \|P f^2\|_{\mathcal{F}_{\delta_N}} + \|\mathbb{P}_N f^2 - P f^2\|_{\mathcal{F}_{\delta_N}} \leq \delta_N^2 + \|\mathbb{P}_N f^2 - P f^2\|_{\mathcal{F}_{\delta_N}}. \quad (40)$$

Taking (40) into (39), we obtain

$$\begin{aligned}
P(\|\mathbb{G}_N\|_{\mathcal{F}_{\delta_N}} > x) &= \frac{2}{x} \mathbb{E} \left[ \int_0^{\delta_N^2 + \|\mathbb{P}_N f^2 - P f^2\|_{\mathcal{F}_{\delta_N}}} \sqrt{s \log \left( \frac{R}{\epsilon} \right)} d\epsilon \right] \\
&\lesssim \frac{2}{x} \mathbb{E} \left[ (\delta_N^2 + \|\mathbb{P}_N f^2 - P f^2\|_{\mathcal{F}_{\delta_N}}) \sqrt{s \log \left( \frac{R}{\delta_N^2 + \|\mathbb{P}_N f^2 - P f^2\|_{\mathcal{F}_{\delta_N}}} \right)} \right] \\
&\leq \frac{2}{x} \mathbb{E} \left[ \delta_N^2 \sqrt{s \log \left( \frac{R}{\delta_N^2} \right)} \right] + \frac{2}{x} \mathbb{E} \left[ \|\mathbb{P}_N f^2 - P f^2\|_{\mathcal{F}_{\delta_N}} \sqrt{s \log \left( \frac{R}{\delta_N^2} \right)} \right] \\
&\leq \frac{2}{x} \tilde{O}(r_N^3 \sqrt{N}) + \frac{2}{x} \tilde{O}(r_N^2 \sqrt{N}) \quad (\text{by Assumption (A7)}) \\
&= o(1) \quad (\text{by Assumption (A3)})
\end{aligned} \tag{41}$$

Based on Slutsky's theorem, we have  $\mathbb{G}_N(\hat{f}_N) \rightsquigarrow \mathcal{N}(0, \Sigma)$  with  $\hat{f}_N = |Y - \hat{\beta}_N^\top X - \hat{g}_N(Z)|$  and  $\Sigma = \text{Var}_U(|Y - \beta_0^\top X - g_0(Z)|)$ .  $\square$

**Theorem 3.** *Under the assumptions (A1)-(A8), it holds that*

$$\sqrt{N}(\hat{\beta}_N - \beta_0) \rightsquigarrow \mathcal{N}(0, \Sigma_2^{-1} \Sigma_1 \Sigma_2^{-1})$$

*Proof.* For  $\hat{\theta}_N = (\hat{\beta}_N, \hat{g}_N)$ , we introduce the following notations:  $\xi = \beta - \beta_0$ ,  $\hat{\xi}_N = \hat{\beta}_N - \beta_0$ ,  $h(Z) = g(Z) - g_0(Z) + (\beta - \beta_0)^\top \varphi^*(Z)$ ,  $\hat{h}_N(Z) = \hat{g}_N(Z) - g_0(Z) + (\hat{\beta}_N - \beta_0)^\top \varphi^*(Z)$  and  $\tilde{X} = X - \varphi^*(Z)$ . These imply that

$$\frac{1}{N} \sum_{i=1}^N |Y_i - \beta^\top X_i - g(Z_i)| = \frac{1}{N} \sum_{i=1}^N |\varepsilon_i - \xi^\top \tilde{X}_i - h(Z_i)|.$$

Denote  $M_N(\xi, h) = \frac{1}{N} \sum_{i=1}^N |\varepsilon_i - \xi^\top \tilde{X}_i - h(Z_i)|$ , and we may calculate the subgradient of the loss function  $M_N$  at  $\xi$  as

$$\partial_\xi M_N(\xi, h) = \frac{1}{N} \sum_{i=1}^N \left( -\text{sign}^*(\varepsilon_i - \xi^\top \tilde{X}_i - h(Z_i)) \tilde{X}_i \right)$$

with Clearly, letting  $\Psi_N(\xi, h) := \mathbb{P}_N \psi(\xi, h)$  with  $\psi(\xi, h) = -\text{sign}(\varepsilon - \xi^\top \tilde{X} - h(Z)) \tilde{X}$ , we have  $\Psi_N(\xi, h) \in \partial_\xi M_N(\xi, h)$ . We further denote

$$\begin{aligned}
(\xi_0, h_0(Z)) &= (0, 0) \in \mathbb{R}^d \times L^2(P), \\
\Psi_0(\xi, h) &= \mathbb{E} \psi(\xi, h), \\
\tilde{\mathcal{A}}_\delta^N &= \{(\xi, h) \mid \xi = \beta - \beta_0, h(Z) = g(Z) - g_0(Z) + (\beta - \beta_0)^\top \varphi^*(Z), (\beta, g) \in \mathcal{A}_\delta^N\}, \\
\mathcal{C}_\delta^N &= \{\psi(\xi, h) - \psi(\xi_0, h_0) \mid (\xi, h) \in \tilde{\mathcal{A}}_\delta^N\},
\end{aligned}$$

for the convenience of the following discussions. Although  $(\xi_0, h_0) = (0, 0)$  is constant, we still introduce such notation to articulate the fields  $((\beta, g)$  or  $(\xi, h))$  to analysis  $\Psi_0$  and  $\Psi_N$ .

By analogy to the proof of Theorem 1, we have

$$\log N_{[\cdot]}(\epsilon, \mathcal{A}_\delta^N, \|\cdot\|_\infty) \lesssim s \log \frac{R}{\epsilon}.$$

Let  $\{[l_i, u_i]: l_i = (\beta_{i;l}, g_{i;l}), u_i = (\beta_{i;u}, g_{i;u}), i = 1, \dots, K\}$  be the  $\epsilon$ -brackets of  $\mathcal{A}_\delta^N$  with  $K = N(\epsilon, \mathcal{A}_\delta^N, L^\infty(P))$ , and for any  $(\beta, g) \in \mathcal{A}_\delta^N$ , without loss of generality, we assume that  $[l_1, u_1]$  is the  $\epsilon$ -

bracket of  $(\beta, g)$ . We may notice that

$$\begin{aligned}
& \int \left| \mathbf{sign} \left( \varepsilon - \xi_{1;l}^\top \tilde{X} - h_{1;l}(Z) \right) - \mathbf{sign} \left( \varepsilon - \xi_{1;u}^\top \tilde{X} - h_{1;u}(Z) \right) \right|^2 dP_{\varepsilon, \tilde{X}, Z} \\
& \stackrel{(\Delta_1)}{=} \int \left| \mathbf{sign} \left( Y - \beta_{1;l}^\top X - g_{1;l}(Z) \right) - \mathbf{sign} \left( Y - \beta_{1;u}^\top X - g_{1;u}(Z) \right) \right|^2 dP_{Y, X, Z} \\
& \stackrel{(\Delta_2)}{\leq} 4P \left( Y - \beta_{1;l}^\top X - g_{1;l}(Z) \geq 0, Y - \beta_{1;u}^\top X - g_{1;u}(Z) < 0 \right) \\
& \stackrel{(\Delta_3)}{\lesssim} \sup_{\tau \in \mathbb{R}} P(Y \in [\tau, \tau + (C+1)\epsilon]) \stackrel{(\Delta_4)}{\lesssim} \epsilon,
\end{aligned}$$

where we denote  $\xi_{1;\nu} = \beta_{1;\nu} - \beta_0$  and  $h_{1;\nu}(Z) = g_{1;\nu}(Z) - g_0(Z) + (\beta_{1;\nu} - \beta_0)^\top \varphi^*(Z)$  in equality  $(\Delta_1)$  with  $\nu \in \{l, u\}$ . Moreover, inequality  $(\Delta_2)$  holds due to  $\mathbf{sign}(Y - \beta_{1;l}^\top X - g_{1;l}(Z)) \geq \mathbf{sign}(Y - \beta_{1;u}^\top X - g_{1;u}(Z))$ , and inequalities  $(\Delta_3)$  and  $(\Delta_4)$  hold by Assumption (A3) and (A7). Hence, we can deduce that

$$\log N_{[\cdot]}(\epsilon, \mathcal{C}_\delta^N, L^2(P)) \lesssim s \log \frac{R}{\epsilon};$$

thus for any  $\delta > 0$ ,

$$J_{[\cdot]}(\delta, \mathcal{C}_\delta^N) = \int_0^\delta \sqrt{1 + N_{[\cdot]}(\epsilon, \mathcal{C}_\delta^N, L^2(P))} d\epsilon \lesssim \delta \sqrt{s \log \frac{R}{\delta}}.$$

Let  $\delta_N = O(r_N \log^2 N + \lambda_N)$ , it follows

$$\begin{aligned}
& \mathbb{E} \left\{ \sup_{(\xi, h) \in \mathcal{C}_{\delta_N}^N} \left\| \sqrt{N} [(\Psi_N - \Psi_0)(\xi, h) - (\Psi_N - \Psi_0)(\xi_0, h_0)] \right\| \right\} \\
& = \mathbb{E} \left\{ \sup_{(\xi, h) \in \mathcal{C}_{\delta_N}^N} \left\| \sqrt{N} (\mathbb{P}_N - P) [\psi_\tau(\xi, h) - \psi_\tau(\xi_0, h_0)] \right\| \right\} \\
& \stackrel{(\Delta_5)}{\lesssim} J_{[\cdot]}(\delta_N, \mathcal{C}_{\delta_N}^N) \left\{ \frac{J_{[\cdot]}(\delta_N, \mathcal{C}_{\delta_N}^N)}{\delta_N^2 \sqrt{N}} + 1 \right\} \\
& = o(1),
\end{aligned}$$

where the inequality  $(\Delta_5)$  holds by Theorem 2.14.18' of [43]. Since  $\|\hat{\xi}_N\| \vee \|\hat{h}_N\|_{L^2([0,1]^d)} = O_P(r_N \log^2 N + \lambda_N)$ , we have

$$\mathbb{E} \left\| \sqrt{N} [(\Psi_N - \Psi_0)(\hat{\xi}_N, \hat{h}_N) - (\Psi_N - \Psi_0)(\xi_0, h_0)] \right\| = o(1),$$

or, written alternatively,

$$\sqrt{N} \{ \Psi_0(\hat{\xi}_N, \hat{h}_N) + \Psi_N(\xi_0, h_0) \} = \sqrt{N} \{ \Psi_N(\hat{\xi}_N, \hat{h}_N) + \Psi_0(\xi_0, h_0) \} + o_p(1). \quad (42)$$

Let  $\tilde{Y}_{i,N} = \varepsilon_i - \hat{h}_N(Z_i)$ ,  $i = 1, \dots, N$ . Then  $\hat{\xi}_N$  is the minimizer of  $M_N^*(\xi) = \frac{1}{N} \sum_{i=1}^N |\tilde{Y}_{i,N} - \xi^\top \tilde{X}_i|$  with respect to  $\xi$  and

$$\Psi_N(\hat{\xi}_N, \hat{h}_N) = -\frac{1}{N} \sum_{i=1}^N \mathbf{sign}(\tilde{Y}_{i,N} - \hat{\xi}_N^\top \tilde{X}_i) \tilde{X}_i. \quad (43)$$

Since  $M_N^*$  is a continuous piecewise function of  $\xi$ , it follows that the limiting subgradient is bounded by the difference between the right and left derivatives. Thus, we have

$$\begin{aligned}
0 & \stackrel{(\Delta_6)}{\in} \partial_\xi (M_N^* + \lambda_N \mathcal{J}_{N,M})|_{(\xi, h) = (\hat{\xi}_N, \hat{h}_N)} \\
& \stackrel{(\Delta_7)}{\subseteq} \partial_\xi M_N^*|_{(\xi, h) = (\hat{\xi}_N, \hat{h}_N)} + \lambda_N \partial_\xi \mathcal{J}_{N,M}|_{(\xi, h) = (\hat{\xi}_N, \hat{h}_N)}.
\end{aligned}$$

where  $(\Delta_6)$  holds by Theorem 10.1 of [34], and  $(\Delta_7)$  holds by Exercise 10.10 and Equation 10(6) of [34]. Thus, we have

$$\begin{aligned}
& \partial_{(\xi, h)} \mathcal{J}_{N, M} \\
&= \left. \frac{\partial \theta}{\partial(\xi, h)} \right|_{(\xi, h)=(\hat{\xi}_N, \hat{h}_N)}^* \partial_{\theta} \mathcal{J}_{N, M} \Big|_{(\beta, g)=(\hat{\beta}_N, \hat{g}_N)} \quad (\text{by Proposition 1.37 of [32]}) \\
&\stackrel{(\Delta_8)}{=} \left. \frac{\partial \theta}{\partial(\xi, h)} \right|_{(\xi, h)=(\hat{\xi}_N, \hat{h}_N)}^* \begin{bmatrix} \left. \partial_{\beta} \mathcal{J}_{N, 1} \right|_{(\beta, g)=(\hat{\beta}_N, \hat{g}_N)} \\ \left. \partial_g \mathcal{J}_{N, 2} \right|_{(\beta, g)=(\hat{\beta}_N, \hat{g}_N)} \end{bmatrix} \\
&= \begin{bmatrix} I & 0 \\ \boldsymbol{\varphi}^*(Z) & \mathbf{I} \end{bmatrix}^* \begin{bmatrix} \left. \partial_{\beta} \mathcal{J}_{N, 1} \right|_{(\beta, g)=(\hat{\beta}_N, \hat{g}_N)} \\ \left. \partial_g \mathcal{J}_{N, 2} \right|_{(\beta, g)=(\hat{\beta}_N, \hat{g}_N)} \end{bmatrix} \quad (\text{by the definition of } \xi, h) \\
&= \begin{bmatrix} \left. \partial_{\beta} \mathcal{J}_{N, 1} \right|_{(\beta, g)=(\hat{\beta}_N, \hat{g}_N)} + \underbrace{\left( \left. \partial_g \mathcal{J}_{N, 2} \right|_{(\beta, g)=(\hat{\beta}_N, \hat{g}_N)} (\boldsymbol{\varphi}_1^*(Z)), \dots, \left. \partial_g \mathcal{J}_{N, 2} \right|_{(\beta, g)=(\hat{\beta}_N, \hat{g}_N)} (\boldsymbol{\varphi}_d^*(Z)) \right)}_{\mathbf{G}_N} \\ \left. \partial_g \mathcal{J}_{N, 2} \right|_{(\beta, g)=(\hat{\beta}_N, \hat{g}_N)} \end{bmatrix}^{\top}
\end{aligned}$$

where  $(\Delta_8)$  holds by Proposition 10.5 of [34] since  $\mathcal{J}_{N, M}(\theta) = \mathcal{J}_{N, 1}(\beta) + \mathcal{J}_{N, 2}(g)$  and  $(\mathbf{d}\mathcal{J}_{N, 1}(\hat{\beta}_N)(0), \mathbf{d}\mathcal{J}_{N, 2}(\hat{g}_N)(0)) = 0$ . Here,  $\mathcal{M}_C^N$  can be embedded in  $L^2(\mathbf{m})$  space with Lebesgue measure  $\mathbf{m}$ , and  $\partial_g \mathcal{J}_{N, 2} \Big|_{(\beta, g)=(\hat{\beta}_N, \hat{g}_N)} \in (L^2(\mathbf{m}))^*$  is the differential operator with norm  $\|\cdot\|_{L^2(\mathbf{m})}$ . By Riesz representation theorem, we can represent  $\partial_g \mathcal{J}_{N, 2} \Big|_{(\beta, g)=(\hat{\beta}_N, \hat{g}_N)} (\boldsymbol{\varphi}_i^*(Z)) = \int f_{\mathcal{J}}(Z) \boldsymbol{\varphi}_i^*(Z) \mathbf{m}(dZ), i = 1, \dots, d$  with a potential element  $f_{\mathcal{J}}(Z) \in L^2(\mathbf{m})$ . Finally, by Assumption (A8) and Corollary 3.44 of [31], we have

$$\partial_{\xi} \mathcal{J}_{N, M} \Big|_{(\xi, h)=(\hat{\xi}_N, \hat{h}_N)} \subseteq \partial_{\beta} \mathcal{J}_{N, 1} \Big|_{(\beta, g)=(\hat{\beta}_N, \hat{g}_N)} + \mathbf{G}_N.$$

Let  $\mathcal{I}_0 := \{i | \tilde{Y}_{i, N} = \hat{\xi}_N^{\top} \tilde{X}_i\}$ , it holds that

$$0 = \frac{1}{N} \left[ \sum_{i=1, i \in \mathcal{I}_0}^N g_i \tilde{X}_i - \sum_{i=1, i \notin \mathcal{I}_0}^N \mathbf{sign}(\tilde{Y}_{i, N} - \hat{\xi}_N^{\top} \tilde{X}_i) \tilde{X}_i \right] + \lambda_N (\mathcal{Q}_1 + \mathcal{Q}_2), \quad (44)$$

where  $g_i \in [-1, 1], i \in \mathcal{I}_0$  are some subgradients of the absolute value function and  $\mathcal{Q}_1 \in \partial_{\beta} \mathcal{J}_{N, 1} \Big|_{(\beta, g)=(\hat{\beta}_N, \hat{g}_N)}$  and  $\mathcal{Q}_2 \in \mathbf{G}_N$ . According to Assumption (A8), we have  $\lambda_N \|\mathcal{Q}_1 + \mathcal{Q}_2\| = o_p(\frac{1}{\sqrt{N}})$ . Taking (44) into (43), it holds that

$$\begin{aligned}
\left\| \frac{\partial M_N^*(\xi)}{\partial \xi} \Big|_{\xi=\hat{\xi}_N} \right\|_1 &\leq \frac{1}{N} \left[ \sum_{i=1, i \in \mathcal{I}_0}^N \left\| \mathbf{sign}(\tilde{Y}_{i, N} - \hat{\xi}_N^{\top} \tilde{X}_i) \tilde{X}_i \right\| + \sum_{i=1, i \in \mathcal{I}_0}^N \left\| g_i \tilde{X}_i \right\| \right] + \lambda_N \|\mathcal{Q}_1 + \mathcal{Q}_2\| \\
&\leq \frac{1}{N} \left[ \sum_{i=1, i \in \mathcal{I}_0}^N \|\tilde{X}_i\| + \|g_i\| \|\tilde{X}_i\| \right] + \lambda_N \|\mathcal{Q}_1 + \mathcal{Q}_2\| \\
&\leq \frac{2}{N} \sum_{i=1, i \in \mathcal{I}_0}^N \|\tilde{X}_i\| + \lambda_N \|\mathcal{Q}_1 + \mathcal{Q}_2\| \\
&\leq \frac{2}{N} \sum_{i=1}^N \mathbf{1}\{\tilde{Y}_{i, N} = \hat{\xi}_N^{\top} \tilde{X}_i\} \|\tilde{X}_i\| + \lambda_N \|\mathcal{Q}_1 + \mathcal{Q}_2\| \\
&\leq \left\{ 2 \sum_{i=1}^N \mathbf{1}\{\tilde{Y}_{i, N} = \hat{\xi}_N^{\top} \tilde{X}_i\} \right\} \max_{i=1, \dots, N} \left( \frac{\|\tilde{X}_i\|}{N} \right) + \lambda_N \|\mathcal{Q}_1 + \mathcal{Q}_2\|
\end{aligned}$$

$$\begin{aligned}
&\stackrel{(\Delta_9)}{\leq} d \max_{i=1,\dots,N} \left( \frac{\|\tilde{X}_i\|}{N} \right) + \lambda_N \|\mathcal{Q}_1 + \mathcal{Q}_2\| \\
&\leq d \max_{i=1,\dots,N} \left( \frac{\|\tilde{X}_i\|}{N} \right) + o_p \left( \frac{1}{\sqrt{N}} \right) \\
&= o_p \left( \frac{1}{\sqrt{N}} \right),
\end{aligned}$$

where the inequality  $(\Delta_9)$  holds due to Theorem 3.3 of [50]. Furthermore, the last equality holds by Assumptions (A3), (A6) and (A7). Moreover, a calculation yields  $\Psi_0(\xi_0, h_0) = 0$ , so the left hand side of (42) satisfies

$$\sqrt{N} \{\Psi_0(\hat{\xi}_N, \hat{h}_N) + \Psi_N(\xi_0, h_0)\} = o_p(1),$$

or equivalently,

$$\sqrt{N} \Psi_0(\hat{\xi}_N, \hat{h}_N) = -\sqrt{N} \Psi_N(\xi_0, h_0) + o_p(1).$$

Applying the Taylor's expansion for  $\Psi_0(\xi, h)$  at  $(\xi_0, h_0)$ , we obtain

$$\begin{aligned}
&\Psi_0(\hat{\xi}_N, \hat{h}_N) \\
&= \mathbb{E}[-\mathbf{sign}(\varepsilon - \xi_{(t)}^\top \tilde{X} - h_{(t)}(Z)) \tilde{X}] \Big|_{t=1} \\
&= \mathbb{E}_V[(2F_\varepsilon(\xi_{(t)}^\top \tilde{X} + h_{(t)}(Z)|V) - 1) \tilde{X}] \Big|_{t=1} \\
&= \mathbb{E}[-\mathbf{sign}(\varepsilon) \tilde{X}] + \frac{\partial \mathbb{E}_V[(2F_\varepsilon(\xi_{(t)}^\top \tilde{X} + h_{(t)}(Z)|V) - 1) \tilde{X}]}{\partial t} \Big|_{t=0} + O(d^2(\hat{\theta}_N, \theta_0)) \\
&= 2\mathbb{E}_V\{f_\varepsilon(0|V) \tilde{X} \tilde{X}^\top\}(\hat{\xi}_N - \xi_0) + 2\mathbb{E}_V\{f_\varepsilon(0|V)(\hat{h}_N - h_0) \tilde{X}^\top\} + O(d^2(\hat{\theta}_N, \theta_0)) \\
&= 2\mathbb{E}_V\{f_\varepsilon(0|V) \tilde{X} \tilde{X}^\top\}(\hat{\xi}_N - \xi_0) + 2\mathbb{E}_V\{f_\varepsilon(0|V)(\hat{h}_N - h_0)(X - \varphi^*(Z))^\top\} + O(d^2(\hat{\theta}_N, \theta_0)) \\
&= 2\mathbb{E}_V\{f_\varepsilon(0|V) \tilde{X} \tilde{X}^\top\}(\hat{\xi}_N - \xi_0) + O(d^2(\hat{\theta}_N, \theta_0)).
\end{aligned}$$

Here the derivative w.r.t.  $h$  and  $\xi$  are based on the derivatives of the line  $h_{(t)} = (1-t)h_0 + t\hat{h}_N, t \in [0, 1]$  and  $\xi_{(t)} = (1-t)\xi_0 + t\hat{\xi}_N, t \in [0, 1]$  w.r.t.  $t$ . Furthermore, the last equality holds by the orthogonality of  $X - \varphi^*(Z)$  w.r.t.  $f(Z) \in L^2(P_Z)$  from the definition of  $\varphi^*(Z)$ . Since  $\hat{\xi}_N - \xi_0 = \hat{\beta}_N - \beta_0$  and Assumption (A6), it follows that

$$\sqrt{N}(\hat{\beta}_N - \beta_0) = \frac{1}{2} [\mathbb{E}_V\{f_\varepsilon(0|V) \tilde{X} \tilde{X}^\top\}]^{-1} \sqrt{N} \Psi_N(\xi_0, h_0) + o_p(1) \rightsquigarrow \mathcal{N}(0, \Sigma_2^{-1} \Sigma_1 \Sigma_2^{-1}).$$

Therefore, the result follows.  $\square$

## 4 Optimization Perspective

### 4.1 Continuous Approximation Approach

In this subsection, we design an efficient algorithm for solving optimization problem (10), which can be rewritten into the following form:

$$\begin{aligned}
&\min_{\beta, \mathbf{W}} \quad \frac{1}{N} \sum_{i=1}^N |Y_i - \beta^\top X_i - g(\mathbf{W}; Z_i)| + \sum_{k=1}^L \delta_{\|W_k\|_\infty \leq 1}(W_k) + \lambda_N \mathcal{J}_{N,M}(\beta, \mathbf{W}), \\
&\text{s.t.} \quad \sum_{k=1}^L \|W_k\|_0 \leq s,
\end{aligned} \tag{45}$$

where for any subset  $C \subseteq \mathbb{R}^n$ ,

$$\delta_C(x) = \begin{cases} 0, & \text{if } x \in C, \\ +\infty, & \text{if } x \notin C. \end{cases}$$

Due to the introduction of coupled nonconvex constraint  $\sum_{k=1}^L \|W_k\|_0 \leq s$ , problem (45) is very challenging; hence we relax the abovementioned constraint into the following separable form:

$$\begin{aligned} \min_{\beta, \mathbf{W}} \mathcal{L}_N(\beta, \mathbf{W}; \mathbf{U}) &\doteq \underbrace{\frac{1}{N} \sum_{i=1}^N |Y_i - \beta^\top X_i - g(\mathbf{W}; Z_i)|}_{\mathcal{R}_N(\beta, \mathbf{W}; \mathbf{U})} + \sum_{k=1}^L (\delta_{\|W_k\|_\infty \leq 1}(W_k) + \gamma_k \|W_k\|_0) \\ &\quad + \lambda_N \mathcal{J}_{N,M}(\beta, \mathbf{W}) \end{aligned} \tag{46}$$

weighting parameter  $\{\gamma_k\}_{k=1}^L$ .

**Definition 19.** Let  $f: \mathbb{R} \rightarrow \mathbb{R}$  and define  $f_\sigma(y) \doteq f(y/\sigma)$  for any  $\sigma > 0$ . The function  $f$  is said to possess Property  $\mathcal{D}$ , if

1.  $f$  is real analytic on  $(y_0, \infty)$  for some  $y_0 < 0$ ,
2.  $\forall y \geq 0$ ,  $f''(y) \geq -\mu_0$ , where  $\mu_0 > 0$  is some constant,
3.  $f$  is concave on  $\mathbb{R}$ ,
4.  $f(y) = 0 \Leftrightarrow y = 0$ ,
5.  $\lim_{y \rightarrow \infty} f(y) = 1$ .

It is obvious that if  $f$  possesses Property  $\mathcal{D}$ , then

$$\lim_{\sigma \downarrow 0^+} f_\sigma(|y|) = I(y) = \begin{cases} 0, & y = 0, \\ 1, & \text{otherwise.} \end{cases}$$

In fact, there are a plenty of functions that satisfy Property  $\mathcal{D}$ , for instance,  $f(y) = 1 - e^{-y}$ . For  $x = (x_1, \dots, x_n)^\top \in \mathbb{R}^n$ , denote  $f_\sigma(x) = \sum_{i=1}^n f_\sigma(x_i)$ . Hence, problem (46) may be approximated by the following continuous optimization problem

$$\min_{\beta, \mathbf{W}} \mathcal{L}_{N,\sigma}(\beta, \mathbf{W}; \mathbf{U}) \doteq \mathcal{R}_N(\beta, \mathbf{W}; \mathbf{U}) + \sum_{k=1}^L (\delta_{\|W_k\|_\infty \leq 1}(W_k) + \gamma_k f_\sigma(|W_k|)) + \lambda_N \mathcal{J}_{N,M}(\beta, \mathbf{W}). \tag{47}$$

**Theorem 4.** Let  $\sigma_k \downarrow 0$ , then the following statements hold.

1.  $\inf_{\beta, \mathbf{W}} \mathcal{L}_{N,\sigma_k}(\beta, \mathbf{W}; \mathbf{U}) \rightarrow \inf_{\beta, \mathbf{W}} \mathcal{L}_N(\beta, \mathbf{W}; \mathbf{U})$ .
2. For  $v$  in some index set  $N \in \mathcal{N}_\infty$ , the sets  $\text{argmin } \mathcal{L}_{N,\sigma_k}$  are nonempty and form a bounded sequence with

$$\limsup_k (\text{argmin } \mathcal{L}_{N,\sigma_k}) \subseteq \text{argmin } \mathcal{L}_N.$$

3. For any choice of  $\epsilon_k \downarrow 0$  and  $(\beta_k, \mathbf{W}_k) \in \epsilon_k\text{-argmin } \mathcal{L}_{N,\sigma_k}$ , the sequence  $\{(\beta_k, \mathbf{W}_k)\}_{k \in \mathbb{N}}$  is bounded and such that all its cluster points belong to  $\text{argmin } \mathcal{L}_N$ .

*Proof.* Since for  $x = (x_1, \dots, x_n)^\top \in \mathbb{R}^n$

$$f_\sigma(x) = \sum_{i=1}^n f_\sigma(x_i),$$

we have  $f_{\sigma_{k+1}}(x) \geq f_{\sigma_k}(x)$  for every  $x \geq 0$ . Hence, for every  $(\beta, \mathbf{W})$ , it follows that  $\mathcal{L}_{N, \sigma_{k+1}}(\beta, \mathbf{W}; \mathbf{U}) \geq \mathcal{L}_{N, \sigma_k}(\beta, \mathbf{W}; \mathbf{U})$ , and  $\{\mathcal{L}_{N, \sigma_k}(\beta, \mathbf{W}; \mathbf{U})\}_{k \in \mathbb{N}}$  is nondecreasing. By Proposition 7.4 of [34],  $\text{e-lim}_k \mathcal{L}_{N, \sigma_k}$  exists and equals  $\sup_k [\text{cl} \mathcal{L}_{N, \sigma_k}]$ . Based on the fact  $\lim_{\sigma \downarrow 0^+} f_\sigma(|x|) = \|x\|_0$ , it follows that  $\sup_k [\text{cl} \mathcal{L}_{N, \sigma_k}](\beta, \mathbf{W}; \mathbf{U}) = \mathcal{L}_N(\beta, \mathbf{W}; \mathbf{U})$ . Obviously, for every  $\sigma_k$ ,  $\mathcal{L}_{N, \sigma_k}(\beta, \mathbf{W}; \mathbf{U})$  is a coercive function. According to Exercise 7.32 of [34], the sequence  $\{\mathcal{L}_{N, \sigma_k}\}_{k \in \mathbb{N}}$  is eventually level-bounded. By noticing that  $\mathcal{L}_{N, \sigma_k}$  and  $\mathcal{L}_N$  are l.s.c. and proper, we finish the proof by Theorem 7.33 of [34].  $\square$

Problem (47) can be rewritten into the following form with bounded feasible set:

$$\begin{aligned} \min_{\beta, \mathbf{W}} \quad & G(\beta, \mathbf{W}; \mathbf{U}) \doteq \mathcal{R}_N(\beta, \mathbf{W}; \mathbf{U}) + \sum_{k=1}^L \gamma_k f_\sigma(|W_k|) + \lambda_N \mathcal{J}_{N, M}(\beta, \mathbf{W}) \\ \text{s.t.} \quad & W_k \in [-1, 1]^{q_k \times (q_{k-1} + 1)}, \quad k = 1, \dots, L, \\ & \beta \in [-C, C]^d. \end{aligned} \tag{48}$$

We fix a probability space  $(\Omega', \mathcal{F}', \mathbb{P}')$  and equip  $\mathcal{X}$  with the Borel  $\sigma$ -algebra with

$$\mathcal{X} \doteq [-C, C]^d \times \prod_{k=1}^L [-1, 1]^{q_k \times (q_{k-1} + 1)}.$$

We suppose that there exists a measurable mapping  $\zeta: \mathcal{X} \times \Omega' \rightarrow \mathbb{R} \times \prod_{k=1}^L \mathbb{R}^{q_k \times (q_{k-1} + 1)}$  satisfying:

$$\mathbb{E}_{\omega'}[\zeta(\beta, \mathbf{W}, \omega')] \in \partial_{(\beta, \mathbf{W})}^C G(\beta, \mathbf{W}; \mathbf{U}) \quad \text{for all } (\beta, \mathbf{W}) \in \mathcal{X}.$$

In this section, we aim to analysis the proximal stochastic subgradient method that performs the following update rule

$$\begin{cases} \text{Sample } \omega'_k \sim \mathbb{P}', \\ (\beta_{k+1}, \mathbf{W}_{k+1}^\top)^\top \in \mathbf{Proj}_{\mathcal{X}}((\beta_k, \mathbf{W}_k^\top)^\top - \alpha_k \zeta(\beta_k, \mathbf{W}_k, \omega'_k)) \end{cases} \tag{49}$$

with given an iterate  $(\beta_k, \mathbf{W}_k) \in \mathcal{X}$ .

**Assumption 2.** *We assume the following assumptions hold.*

- The sequence  $\{\alpha_k\}_{k \geq 1}$  is nonnegative, square summable, but not summable:

$$\alpha_k \geq 0, \quad \sum_{k=1}^{\infty} \alpha_k = \infty, \quad \text{and} \quad \sum_{k=1}^{\infty} \alpha_k^2 < \infty.$$

- There exists a function  $p: \mathcal{X} \rightarrow \mathbb{R}_+$ , that is bounded on bounded sets, such that

$$\mathbb{E}_{\omega'}[\|\zeta(\beta, \mathbf{W}, \omega')\|^2] \leq p(\beta, \mathbf{W}) \quad \text{for all } (\beta, \mathbf{W}) \in \mathcal{X}.$$

- For every convergent sequence  $\{z_k\}_{k \geq 1}$ , we have

$$\mathbb{E}_{\omega'} \left[ \sup_{k \geq 1} \|\zeta(\beta_k, \mathbf{W}_k, \omega')\| \right] < \infty.$$

**Theorem 5.** Let  $\{(\beta_k, \mathbf{W}_k)\}_{k \geq 1}$  be the iterates produced by the proximal stochastic subgradient method (49). The almost surely, for all  $(\beta^*, \mathbf{W}^*) \in \mathbf{Cluster}(\{(\beta_k, \mathbf{W}_k)\}_{k \geq 1})$ , it holds that

$$0 \in \partial_{(\beta, \mathbf{W})}^C G(\beta^*, \mathbf{W}^*; \mathbf{U}) + N_{\mathcal{X}}(\beta^*, \mathbf{W}^*),$$

and the function values  $\{G(\beta_k, \mathbf{W}_k; \mathbf{U})\}_{k \geq 1}$  converge.

*Proof.* To prove this theorem, by Theorem 6.2 of [9] and Assumption 2, it suffices to show the descent property and weak Sard property. By Example 2.4 of [26],  $G$  is definable in an o-minimal structure. Then, by Theorem 5.8 of [9],  $G$  and  $1_{\mathcal{X}}$  admit the chain rule. Therefore, the descent property holds by Lemma 6.3 of [9]. Thus we only argue the weak Sard property. Since  $G$ , and  $1_{\mathcal{X}}$  are definable in an o-minimal structure, there exist Whitney  $C^{d+H}$ -stratifications  $\mathcal{A}_G$ , and  $\mathcal{A}_{\mathcal{X}}$  of  $\mathbf{graph}(G)$ , and  $\mathcal{X}$ , respectively with  $H = \prod_{k=1}^L (q_k + 1)$ . Let  $\mathbf{Proj}(\mathcal{A}_G)$  be the Whitney stratifications of  $\mathbb{R}^{d+H}$  obtained by applying the coordinate projection  $(\beta, \mathbf{W}, r) \mapsto (\beta, \mathbf{W})$  to each stratum in  $\mathcal{A}_G$ . Appealing to Theorem 4.8 of [44], we obtain a Whitney  $C^{d+H}$ -stratification  $\mathcal{A}$  of  $\mathbb{R}^{d+H}$  such that for every strata  $M \in \mathcal{A}$  and  $L \in \mathbf{Proj}(\mathcal{A}_G) \cup \mathcal{A}_{\mathcal{X}}$ , either  $M \cap L = \emptyset$  or  $M \subseteq L$ .

Consider an arbitrary stratum  $M \in \mathcal{A}$  with  $M \cap \mathcal{X} \neq \emptyset$  and a point  $x \in M$ . Obviously, we have  $M \subseteq \mathcal{X}$ . Select the unique strata  $M_G \in \mathbf{Proj}(\mathcal{A}_G)$ , and  $M_{\mathcal{X}} \in \mathcal{A}_{\mathcal{X}}$  containing  $x$ . Let  $\hat{G}$  be  $C^{d+H}$ -smooth functions agreeing with  $G$  on a neighborhood of  $x$  in  $M_G$ . By Proposition 4 of [4], we conclude

$$\partial^C G(\beta, \mathbf{W}) \subseteq \nabla \hat{G}(\beta, \mathbf{W}) + N_{M_G}(\beta, \mathbf{W}), \quad N_{\mathcal{X}}(\beta, \mathbf{W}) \subseteq N_{M_{\mathcal{X}}}(\beta, \mathbf{W}).$$

Hence summing yields

$$\begin{aligned} \partial^C G(\beta, \mathbf{W}) + N_{\mathcal{X}}(\beta, \mathbf{W}) &\subseteq \nabla(\hat{G})(\beta, \mathbf{W}) + N_{M_G}(\beta, \mathbf{W}) + N_{M_{\mathcal{X}}}(\beta, \mathbf{W}) \\ &\subseteq \nabla(\hat{G})(\beta, \mathbf{W}) + N_M(\beta, \mathbf{W}), \end{aligned}$$

where the last inclusion follows from  $M \subseteq M_G$  and  $M \subseteq M_{\mathcal{X}}$ . Notice that  $\hat{G}$  agrees with  $G$  on a neighborhood of  $(\beta, \mathbf{W})$  in  $M$ . Hence if the inclusion,  $0 \in \partial^C G(\beta, \mathbf{W}) + N_{\mathcal{X}}(\beta, \mathbf{W})$ , holds it must be that  $(\beta, \mathbf{W})$  is a critical point of the  $C^{d+H}$ -smooth function  $G$  restricted to  $M$ , in the classical sense. Applying the Theorem 6.10 (classical Sard's theorem) of [27] to each manifold  $M$ , weak Sard's property holds. Hence, we finish the proof by Theorem 6.2 of [9].  $\square$

**Remark 2.** The core idea of the proof above is to establish the Weak Sard property, which allows us to follow the line of [9] to complete the proof. First, we partition the feasible set  $\mathcal{X}$  into a collection of disjoint smooth manifolds  $\mathcal{A}$ , such that objective function (48) is smooth on each manifold  $M \in \mathcal{A}$ . We then apply the projection theorem (Proposition 4 of [4]) to show  $\partial^C G(\beta, \mathbf{W}) \subseteq \nabla \hat{G}(\beta, \mathbf{W}) + N_{M_G}(\beta, \mathbf{W})$  and  $N_{\mathcal{X}}(\beta, \mathbf{W}) \subseteq N_{M_{\mathcal{X}}}(\beta, \mathbf{W})$ . Subsequently, we demonstrate that the summation of the classical gradient of the local mollifier  $\nabla \hat{G}(\beta, \mathbf{W})$  and the normal cone  $N_M(\beta, \mathbf{W})$  covers  $\partial^C G(\beta, \mathbf{W}) + N_{\mathcal{X}}(\beta, \mathbf{W})$ . Finally, the Weak Sard Property is deduced by applying the standard Sard's theorem to the local mollifier  $\hat{G}(\beta, \mathbf{W})$  on each manifold in the partition.

## 4.2 Non-Approximation Approach

In this subsection, we directly solve the following optimization problem induced by  $\ell_0$ -norm:

$$\begin{aligned} \min_{\beta, \mathbf{W}} \quad & \mathcal{H}(\beta, \mathbf{W}; \mathbf{U}) \doteq \frac{1}{N} \sum_{i=1}^N |Y_i - \beta^\top X_i - g(\mathbf{W}; Z_i)| + \lambda_N \mathcal{J}_{N,M}(\beta, \mathbf{W}), \\ \text{s.t.} \quad & \beta \in [-C, C]^d, \\ & \mathbf{W} \in \mathcal{W} \doteq \left\{ \mathbf{W} : \sum_{k=1}^L \|W_k\|_0 \leq s \right\} \cap \prod_{k=1}^L [-1, 1]^{q_k \times (q_{k-1} + 1)}. \end{aligned} \tag{50}$$



Again, we fix a probability space  $(\Omega'', \mathcal{F}'', \mathbb{P}'')$  and equip  $[-C, C]^d \times \mathcal{W}$  with the Borel  $\sigma$ -algebra. We suppose that there exists a measurable mapping  $\tilde{\zeta}: [-C, C]^d \times \mathcal{W} \times \Omega'' \rightarrow \mathbb{R} \times \prod_{k=1}^L \mathbb{R}^{q_k \times (q_{k-1}+1)}$  satisfying:

$$\mathbb{E}_{\omega''}[\tilde{\zeta}(\beta, \mathbf{W}, \omega'')] \in \partial_{(\beta, \mathbf{W})}^C G(\beta, \mathbf{W}; \mathbf{U}) \quad \text{for all } (\beta, \mathbf{W}) \in [-C, C]^d \times \mathcal{W}.$$

We still consider the proximal stochastic subgradient method that performs the following update rule

$$\begin{cases} \text{Sample } \omega_k'' \sim \mathbb{P}'', \\ \beta_{k+1} \in \mathbf{Proj}_{[-C, C]^d}(\beta_k - \alpha_k \mathbf{Proj}_1(\tilde{\zeta}(\beta_k, \mathbf{W}_k, \omega_k''))), \\ \mathbf{W}_{k+1} \in \mathbf{Proj}_{\mathcal{W}}(\mathbf{W}_k - \alpha_k \mathbf{Proj}_2(\tilde{\zeta}(\beta_k, \mathbf{W}_k, \omega_k''))), \end{cases} \quad (51)$$

with given an iterate  $(\beta_k, \mathbf{W}_k) \in \mathcal{X}$ .

**Lemma 1.** *The sub-routine*

$$\mathbf{W}^* \in \mathbf{Proj}_{\mathcal{W}}(\mathbf{W} - \alpha(\tilde{\zeta}(\beta, \mathbf{W}, \omega_k''))) \quad (52)$$

*admits the following closed-form solution. Denoting  $G_i \doteq W_i - \alpha_k \mathbf{Proj}_2[\tilde{\zeta}(\beta, \mathbf{W}, \omega'')]_i$ , we compute local benefit values*

$$\Delta_{i;j,k} = \begin{cases} [G_i]_{jk}^2, & |[G_i]_{jk}| \leq 1, \\ 2|[G_i]_{jk}| - 1, & |[G_i]_{jk}| > 1, \end{cases}$$

*for each entry  $(i, j, k) \in [L] \times [q_i] \times [q_{i-1} + 1]$ . Let  $\mathbb{T} \subseteq [L] \times [q_i] \times [q_{i-1} + 1]$  contains the indices of the  $s$  largest values of  $\Delta_{i;j,k}$  (ties may be broken arbitrarily). Then the projection  $\mathbf{W}^*$  can be selected as*

$$[\mathbf{W}^*]_{jk} = \begin{cases} \mathbf{clip}([G_i]_{jk}, -1, 1), & (i, j, k) \in \mathbb{T}, \\ 0, & (i, j, k) \notin \mathbb{T}, \end{cases}$$

where  $\mathbf{clip}(y, -1, 1) = \min\{1, \max\{-1, y\}\}$ .

*Proof.* The sub-routine (52) is equivalent to the following optimization problem

$$\begin{aligned} \min_{\mathbf{W}} \quad & \sum_{i=1}^L \|W_i - G_i\|^2 = \sum_{i=1}^L \sum_{j=1}^{q_i} \sum_{k=1}^{q_{i-1}+1} |[W_i]_{jk} - [G_i]_{jk}|^2 \\ \text{s.t.} \quad & \mathbf{W} \in \mathcal{W} \end{aligned}$$

with given  $\mathbf{G} = (G_1, \dots, G_L)$ . For the sake of simplicity, we denote objective function as

$$\sum_{i=1}^L \sum_{j=1}^{q_i} \sum_{k=1}^{q_{i-1}+1} |[W_i]_{jk} - [G_i]_{jk}|^2 \doteq \sum_{(i,j,k)} |[W_i]_{jk} - [G_i]_{jk}|^2.$$

For any fixed set  $\mathbb{T}' \subseteq [L] \times [q_i] \times [q_{i-1} + 1]$  with  $|\mathbb{T}'| \leq s$ , we consider the restricted feasible set

$$\mathcal{W}_{\mathbb{T}'} = \{\mathbf{W}: [W_i]_{ij} = 0 \text{ for } (i, j, k) \notin \mathbb{T}', |[W_i]_{ij}| \leq 1 \ \forall (i, j, k) \in \mathbb{T}'\}.$$

Then

$$\begin{aligned} \min_{\mathbf{W} \in \mathcal{W}_{\mathbb{T}'}} \sum_{(i,j,k)} |[W_i]_{jk} - [G_i]_{jk}|^2 &= \sum_{(i,j,k) \in \mathbb{T}'} \min_{|[W_i]_{jk}| \leq 1} ([W_i]_{jk} - [G_i]_{jk})^2 + \sum_{(i,j,k) \notin \mathbb{T}'} [G_i]_{jk}^2 \\ &= \sum_{(i,j,k) \in \mathbb{T}'} ([W_i]_{jk}^* - [G_i]_{jk})^2 + \sum_{(i,j,k) \notin \mathbb{T}'} [G_i]_{jk}^2 \\ &= \sum_{(i,j,k)} [G_i]_{jk}^2 - \sum_{(i,j,k) \in \mathbb{T}'} \Delta_{i;j,k} \end{aligned}$$

with  $[W_i]_{jk}^* \doteq \text{clip}([G_i]_{jk}, -1, 1)$ . Consequently, it holds that

$$\min_{\mathbf{W} \in \mathcal{W}} \sum_{(i,j,k)} |[W_i]_{jk} - [G_i]_{jk}|^2 = \sum_{(i,j,k)} [G_i]_{jk}^2 - \max_{\mathbb{T}' : |\mathbb{T}'| \leq s} \sum_{(i,j,k) \in \mathbb{T}'} \Delta_{i,j,k}$$

To maximize  $\sum_{(i,j,k) \in \mathbb{T}'} \Delta_{i,j,k}$  subject to  $|\mathbb{T}'| \leq s$ , the optimal  $\mathbb{T}$  consists of the indices corresponding to the  $s$  largest  $\Delta_{i,j,k}$  (ties broken arbitrarily). Here, we complete the proof.  $\square$

**Theorem 6.** *Let  $\{(\beta_k, \mathbf{W}_k)\}_{k \geq 1}$  be the iterates produced by the proximal stochastic subgradient method (51). The almost surely, for all  $(\beta^*, \mathbf{W}^*) \in \text{Cluster}(\{(\beta_k, \mathbf{W}_k)\}_{k \geq 1})$ , it holds that*

$$0 \in \partial_{(\beta, \mathbf{W})}^C \mathcal{H}(\beta^*, \mathbf{W}^*; \mathbf{U}) + N_{[-C, C]^d \times \mathcal{W}}(\beta^*, \mathbf{W}^*),$$

and the function values  $\{\mathcal{H}(\beta_k, \mathbf{W}_k; \mathbf{U})\}_{k \geq 1}$  converge.

*Proof.* We only need to prove  $\mathcal{W}$  is semi-algebraic; the remaining argument is identical to the proof of Theorem 5. Write all entries of  $\mathbf{W}$  as a single vector  $x = (x_1, \dots, x_N) \in \mathbb{R}^H$ , where  $N$  is the total number of scalar elements in  $(W_1, \dots, W_L)$ . For any index set  $T \subset \{1, \dots, N\}$ , define

$$A_T = \{x \in \mathbb{R}^H : x_i = 0 \text{ for all } i \notin T\}.$$

Each  $A_T$  is the zero set of finitely many polynomial equations  $\{x_i = 0 : i \notin T\}$ , so  $A_T$  is an algebraic (hence semi-algebraic) subset of  $\mathbb{R}^N$ . The condition  $\sum_{k=1}^L \|W_k\|_0 \leq s$  is equivalent to saying that the total number of nonzero coordinates of  $x$  is at most  $s$ . Hence,

$$\mathcal{W} = \bigcup_{\substack{T \subset \{1, \dots, N\} \\ |T| \leq s}} A_T.$$

This is a finite union since there are only finitely many subsets  $T$  with  $|T| \leq s$ . A finite union of semi-algebraic sets is again semi-algebraic, and each  $A_T$  is semi-algebraic. Therefore,  $\mathcal{W}$  is a semi-algebraic subset of  $\mathbb{R}^H$ .  $\square$

**Remark 3.** *Under some additional mild assumptions, almost surely, the sequence  $\{\beta_k, \mathbf{W}_k\}_{k \geq 1}$  converges to a local minimum of  $\mathcal{H}$ , i.e., the proximal stochastic subgradient method (51) can escape active strict saddles and sharply repulsive critical points of  $\mathcal{H}$ ; the readers may refer to [37] for the details.*

**Remark 4.** *In fact, both the continuous approximation approach (proposed in Section 4.1) and the non-approximation approach (proposed in Section 4.2) admit independent research interests. On the one hand, although there exists a gap between continuous relaxation problem (46) and primal penalized LAD problem (10), proximal stochastic subgradient update (49) for problem (46) is very cheap, as it only requires projections onto a boxed set. Additionally, by Theorem 4, relaxation problem (46) can approximate penalized LAD problem (10) to arbitrary accuracy, thereby exhibiting independent interest beyond serving as a computational surrogate. On the other hand, the non-approximation approach aims to solve problem (50) which is completely equivalent to penalized LAD problem (10), henceforth enjoying high statistical accuracy. Nevertheless, the proximal stochastic subgradient update (51) involves a sorting operation with complexity  $O(H \log H)$ , and is therefore relatively more computationally demanding and unsuitable for ultra large-scale networks. Overall, these two approaches illustrate a fundamental trade-off between computational efficiency and statistical fidelity.*

# References

- [1] Eric Auerbach. Identification and estimation of a partially linear regression model using network data. *Econometrica*, 90(1):347–365, 2022.
- [2] Edoardo Belli. Smoothly adaptively centered ridge estimator. *Journal of Multivariate Analysis*, 189:104882, 2022.
- [3] Alexandre Belloni, Victor Chernozhukov, Denis Chetverikov, and Iván Fernández-Val. Conditional quantile processes based on series or many regressors. *Journal of Econometrics*, 213(1):4–29, 2019. Annals: In Honor of Roger Koenker.
- [4] Jérôme Bolte, Aris Daniilidis, Adrian Lewis, and Masahiro Shiota. Clarke subgradients of stratifiable functions. *SIAM Journal on Optimization*, 18(2):556–572, 2007.
- [5] Chad Brown. Inference in partially linear models under dependent data with deep neural networks. *arXiv preprint arXiv:2410.22574*, 2024.
- [6] Philippe G Ciarlet. *Linear and nonlinear functional analysis with applications*. SIAM, 2025.
- [7] Pavel Čížek and Serhan Sadıkoğlu. Robust nonparametric regression: A review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 12(3):e1492, 2020.
- [8] Jack Cuzick. Efficient estimates in semiparametric additive regression models with unknown error distribution. *The Annals of Statistics*, 20(2):1129–1136, 1992.
- [9] Damek Davis, Dmitriy Drusvyatskiy, Sham Kakade, and Jason D Lee. Stochastic subgradient method converges on tame functions. *Foundations of computational mathematics*, 20(1):119–154, 2020.
- [10] John C Duchi and Feng Ruan. Stochastic methods for composite and weakly convex optimization problems. *SIAM Journal on Optimization*, 28(4):3229–3259, 2018.
- [11] Esra Akdeniz Duran, Wolfgang Karl Härdle, and Maria Osipenko. Difference based ridge and liu type estimators in semiparametric regression models. *Journal of Multivariate Analysis*, 105(1):164–175, 2012.
- [12] Robert F. Engle, C. W. J. Granger, John Rice, and Andrew Weiss. Semiparametric estimates of the relation between weather and electricity sales. *Journal of the American Statistical Association*, 81(394):310–320, 1986.
- [13] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.
- [14] Max H Farrell, Tengyuan Liang, and Sanjog Misra. Deep neural networks for estimation and inference: application to causal effects and other semiparametric estimands. *arXiv preprint arXiv:1809.09953*, 20:1, 2018.
- [15] Federico Ferraccioli, Laura M Sangalli, and Livio Finos. Nonparametric tests for semiparametric regression models. *Test*, 32(3):1106–1130, 2023.
- [16] Dahua Gan, Yi Wang, Shuo Yang, and Chongqing Kang. Embedding based quantile regression neural network for probabilistic load forecasting. *Journal of Modern Power Systems and Clean Energy*, 6(2):244–254, 2018.
- [17] Scott A Hamilton and Young K Truong. Local linear estimation in partly linear models. *Journal of Multivariate Analysis*, 60(1):1–19, 1997.

- [18] Wolfgang Härdle, Yuichi Mori, and Philippe Vieu. *Statistical methods for biostatistics and related fields*. Springer Science & Business Media, 2006.
- [19] Kostas Hatalis, Alberto J Lamadrid, Katya Scheinberg, and Shalinee Kishore. Smooth pinball neural network for probabilistic forecasting of wind power. *arXiv preprint arXiv:1710.01720*, 2017.
- [20] Xuming He and Peide Shi. Convergence rate of b-spline estimators of nonparametric conditional quantile functions. *Journal of Nonparametric Statistics*, 3(3-4):299–308, 1994.
- [21] Judy Hoffman, Daniel A Roberts, and Sho Yaida. Robust learning with jacobian regularization. *arXiv preprint arXiv:1908.02729*, 2019.
- [22] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.
- [23] Peter J Huber and Elvezio M Ronchetti. Robust statistics, ser. *Wiley Ser Probab Math Stat New York, NY, USA Wiley-IEEE*, 52:54, 1981.
- [24] Shun ichi Amari. Backpropagation and stochastic gradient descent method. *Neurocomputing*, 5(4):185–196, 1993.
- [25] Roger Koenker, Victor Chernozhukov, Xuming He, and Limin Peng. *Handbook of quantile regression*. CRC Press: Boca Raton, FL, USA, 2017.
- [26] Julian Kranz, Davide Gallon, Steffen Dereich, and Arnulf Jentzen. Sad neural networks: Divergent gradient flows and asymptotic optimality via o-minimal structures. *arXiv preprint arXiv:2505.09572*, 2025.
- [27] John M Lee. Smooth manifolds. In *Introduction to smooth manifolds*, pages 1–29. Springer, 2003.
- [28] Sokbae Lee. Efficient semiparametric estimation of a partially linear quantile regression model. *Econometric theory*, 19(1):1–31, 2003.
- [29] Heng Lian. Semiparametric estimation of additive quantile regression models by two-fold penalty. *Journal of Business & Economic Statistics*, 30(3):337–350, 2012.
- [30] Mathew W McLean, Giles Hooker, Ana-Maria Staicu, Fabian Scheipl, and David Ruppert. Functional generalized additive models. *Journal of Computational and Graphical Statistics*, 23(1):249–269, 2014.
- [31] B.S. Mordukhovich. *Variational Analysis and Generalized Differentiation II: Applications*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2006.
- [32] B.S. Mordukhovich. *Second-Order Variational Analysis in Optimization, Variational Stability, and Control: Theory, Algorithms, Applications*. Springer Series in Operations Research and Financial Engineering. Springer International Publishing, 2024.
- [33] Roger D. Peng, Francesca Dominici, and Thomas A. Louis. Model choice in time series studies of air pollution and mortality. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 169(2):179–203, 02 2006.
- [34] R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.
- [35] Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.

- [36] Mahdi Roozbeh and Mohammad Arashi. New ridge regression estimator in semiparametric regression models. *Communications in Statistics-Simulation and Computation*, 45(10):3683–3715, 2016.
- [37] Sholom Schechtman. Stochastic subgradient descent on a generic definable function converges to a minimizer. *arXiv preprint arXiv:2109.02455*, 2021.
- [38] Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with relu activation function. 2020.
- [39] Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 48(4):1875 – 1897, 2020.
- [40] Thomas A Severini and Wing Hung Wong. Profile likelihood and conditionally parametric models. *The Annals of statistics*, pages 1768–1802, 1992.
- [41] Ben Sherwood and Lan Wang. Partially linear additive quantile regression in ultra-high dimension. *The Annals of Statistics*, 44(1):288 – 317, 2016.
- [42] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.
- [43] A. W. van der Vaart and Jon A. Wellner. *Empirical Processes*, pages 127–384. Springer International Publishing, Cham, 2023.
- [44] Lou Van den Dries and Chris Miller. Geometric categories and  $o$ -minimal structures. 1996.
- [45] Shuoyang Wang. Partially linear quantile regression for complex nonlinear component in ultra-high dimension. *Electronic Journal of Statistics*, 19(2):4054–4082, 2025.
- [46] Wei Wen, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. Learning structured sparsity in deep neural networks. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.
- [47] Scott L. Zeger and Peter J. Diggle. Semiparametric models for longitudinal data with application to cd4 cell numbers in hiv seroconverters. *Biometrics*, 50(3):689–699, 1994.
- [48] Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. 2010.
- [49] 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.
- [50] Qixian Zhong and Jane-Ling Wang. Neural networks for partially linear quantile regression. *Journal of Business & Economic Statistics*, 42(2):603–614, 2024.
- [51] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, 2005.