Deep learning for individual heterogeneity with generated regressors by adversarial training

Yifei Ding and Ruoyao Shi

November 6, 2024

Abstract

We propose a semiparametric framework that combines machine learning with generated regressors via control function to capture individual heterogeneity while addressing endogeneity and sample selection bias in complex econometric models. This approach models individual heterogeneity through high-dimensional observable characteristics, with generated regressors supporting the control function to manage endogeneity or sample selection bias flexibly across various economic structures. Leveraging a tailored deep learning architecture, our framework integrates control functions and parameter functions seamlessly, enabling its adaptation to diverse econometric models. Using adversarial training, we achieve sup-norm convergence rates of parameter functions and control function at the optimal min-max rate, which enhances robustness and yields valid inferences for inferential structural parameters in high-dimensional settings. Extending the Double Machine Learning (DML) approach, we incorporate endogenous components and establish a new influence function that directly includes generated regressors, broadening the framework's applicability across econometric models. With automatic differentiation in PyTorch, the influence function applies directly to data, streamlining inference and supporting various structural parameters without additional calculations. This integration makes the framework particularly useful in applied settings where individual heterogeneity and endogeneity are critical, such as personalized policy-making, targeted economic interventions, and customized optimizations in technology. Our simulations demonstrate superior performance, validating this framework's practical use in econometric analysis where heterogeneity and endogeneity are key considerations.

1 Introduction

In contemporary econometrics, accounting for individual heterogeneity is critical for both predictive accuracy and causal inference. Economic agents, whether they be individuals, firms, or institutions, exhibit distinct behaviors driven by diverse underlying factors. Traditional econometric methods, which assume homogeneity in model parameters, can result in biased estimates and incorrect inferences, particularly when such heterogeneity is ignored. Understanding and modeling these differences is essential for making reliable economic predictions and drawing meaningful policy conclusions, particularly in complex, real-world environments.

As the availability of large, high-dimensional datasets has grown, the challenge of capturing individual heterogeneity has intensified. In many applied settings—ranging from personalized policy-making to firm-level decision-making—the variables that determine heterogeneity are often high-dimensional and interdependent. For instance, demographic factors, socio-economic variables, and behavioral indicators can all play crucial roles in driving heterogeneity, yet their interactions are difficult to model using traditional econometric tools. As a result, econometricians have increasingly turned to machine learning (ML) techniques, which are well-suited to dealing with high-dimensional data, for their flexibility and predictive power.

One of the most influential frameworks for handling high-dimensional data in econometrics is Double Machine Learning (DML), developed by Chernozhukov et al. (2018). DML leverages machine learning methods to model nuisance parameters, thereby enabling valid causal inference in settings where confounding factors are high-dimensional and complex. The framework's key strength lies in its ability to separate the estimation of treatment effects from the modeling of control variables, ensuring that high-dimensional confounders are appropriately accounted for without sacrificing interpretability. While DML has proven to be a powerful tool for causal inference, its traditional implementation assumes homogeneity in model parameters or relies on specific models with limited heterogeneity, which limits its ability to capture individual heterogeneity.

Deep learning architectures, known for their capacity to model non-linear relationships and capture complex patterns in high-dimensional setting, have emerged as promising tools for addressing these limitations. However, existing applications of deep learning to econometrics, such as the work by Farrell et al. (2021), typically assume that the model operates in an exogenous setting, where the variables of interest are not correlated with unobserved factors. This exogeneity assumption, while simplifying the modeling process, severely limits the framework's applicability in real-world settings where endogeneity—the correlation between

explanatory variables and error terms—is pervasive. Endogeneity is particularly common in policy-relevant contexts, where unobserved factors often influence both the treatment and outcome variables, leading to biased estimates if not properly addressed.

The primary contribution of my work is to extend the Double Machine Learning framework to account for both individual heterogeneity and endogeneity in high-dimensional settings. To achieve this, I introduce a novel deep learning architecture that incorporates generated regressors as part of a control function approach, enabling the model to mitigate endogeneity while maintaining flexibility in capturing heterogeneity. The control function method involves incorporating the residuals from the first stage regression as additional covariates in the second stage regression. This approach not only addresses endogeneity but also allows for more flexible modeling of the relationship between the endogenous variable and the outcome. By integrating this method within a deep learning framework, this approach marks a significant improvement over existing methods by allowing for accurate causal inference in settings where traditional econometric techniques often fail.

A key theoretical innovation of my work is the generalization of Double Machine Learning to accommodate endogenous components through the control function approach. Unlike traditional DML methods, which either avoid directly handling endogeneity or treat it with simplified methods akin to 2SLS, my framework introduces generated regressors that effectively mitigate endogeneity while preserving the flexibility of the model to capture individual heterogeneity. This demonstrates that DML can be extended to remain useful even when endogenous components are present, making it applicable to a broader range of econometric models. This allows the framework to retain its flexibility in high-dimensional settings while ensuring valid causal inference.

Specifically, I introduce a novel influence function (Neyman orthogonal score) that explicitly accounts for the inclusion of generated regressors in the model. This goes beyond the standard orthogonal score developed by Chernozhukov et al. (2018); Farrell et al. (2021) by providing a more robust way to handle endogeneity, ensuring that causal estimates remain unbiased even in the presence of endogenous relationships. This extension is crucial for real-world applications, where ignoring endogeneity could lead to severely biased conclusions.

Another key innovation in my framework is the use of adversarial training within the deep learning architecture, which ensures sup-norm convergence for statistical inference. This provides stronger convergence guarantees compared to the traditional L2-norm convergence used in the original DML frameworks(Chernozhukov et al., 2018). By achieving sup-norm con-

vergence of nuisance functions, particularly with generated regressors, my framework enables more precise and reliable statistical inferences. Specifically, this methodological advancement provides sup-norm convergence rates and guarantee estimated parameter functions and control function with generated regressors convergent to its true values at a rate faster than $N^{-1/4}$, which actually achieves the optimal min-max rate. This significantly extends the theoretical guarantees of ML models in econometric settings, as the standard DML framework relies on weaker L2-norm conditions, which may not suffice in complex models with high-dimensional heterogeneity and endogenous components.

By leveraging the power of automatic differentiation engines like PyTorch, we can easily derive empirical influence functions (orthogonal scores) in existence of generated regressors. It conveniently facilitates valid inference for any second-stage smooth inferential parameters with first-stage parameter functions of different machine learning models under regularity conditions.

The innovations presented in this paper hold significant implications for a wide range of real-world applications, from policy analysis to business strategy. In the tech industry, for instance, my model can evaluate the causal impact of personalized recommendation systems on consumer behavior, accounting for both individual-level heterogeneity and the endogeneity that arises from user interactions with algorithms. This is highly relevant for designing targeted marketing strategies, allowing companies to optimize their algorithms and improve customer satisfaction and retention. Similarly, in labor economics, understanding how policy interventions affect different groups while controlling for endogenous factors can lead to more effective and targeted policies. By addressing these challenges, our framework offers a robust and flexible tool for econometric analysis in high-dimensional, heterogeneous, and endogenous settings.

The remainder of this paper is organized as follows: Section 2 provides an overview of the proposed semiparametric framework, summarizing the key methodological results. Section 3 presents the detailed model setting with examples that illustrate the framework. Section 4 explores the structural deep learning approach with generated regressors, emphasizing the role of adversarial training in achieving robust estimation. Section 5 elaborates on the semiparametric inference process, including the development of influence functions for models with generated regressors. Section 6 details the estimation procedures, including practical implementation and empirical results. Section 7 covers the simulation studies, demonstrating the performance of the proposed methods under various scenarios. Finally, Section 8 concludes the paper by summarizing the findings and suggesting directions for future research.

2 A Semiparametric Framework for Individual Characteristics with Generated Regressors

In this section, we describe our semiparametric framework that uses machine learning to capture individual heterogeneity and simultaneously estimate the control function using generated regressors. The semiparametric framework here follows with the model setting in Farrell et al. (2021) but generalize it with generated regressors through control function. Let's assume dependent variables, $\mathbf{Y} \in \mathbb{R}^{d_\mathbf{Y}}$, treatment variables of research interest ,$\mathbf{T} \in \mathbb{R}^{d_\mathbf{T}}$, and a vector of variables $\mathbf{Z} \in \mathbb{R}^{d_\mathbf{Z}}$ are observed data. In the contexts of endogeneity and sample selection bias, $\mathbf{Z}$ denotes instrumental variables and selection variables respectively. The data types of $\mathbf{T}$ and $\mathbf{Z}$ can include either continuous or discrete variables as long as generated regressors are continuous.

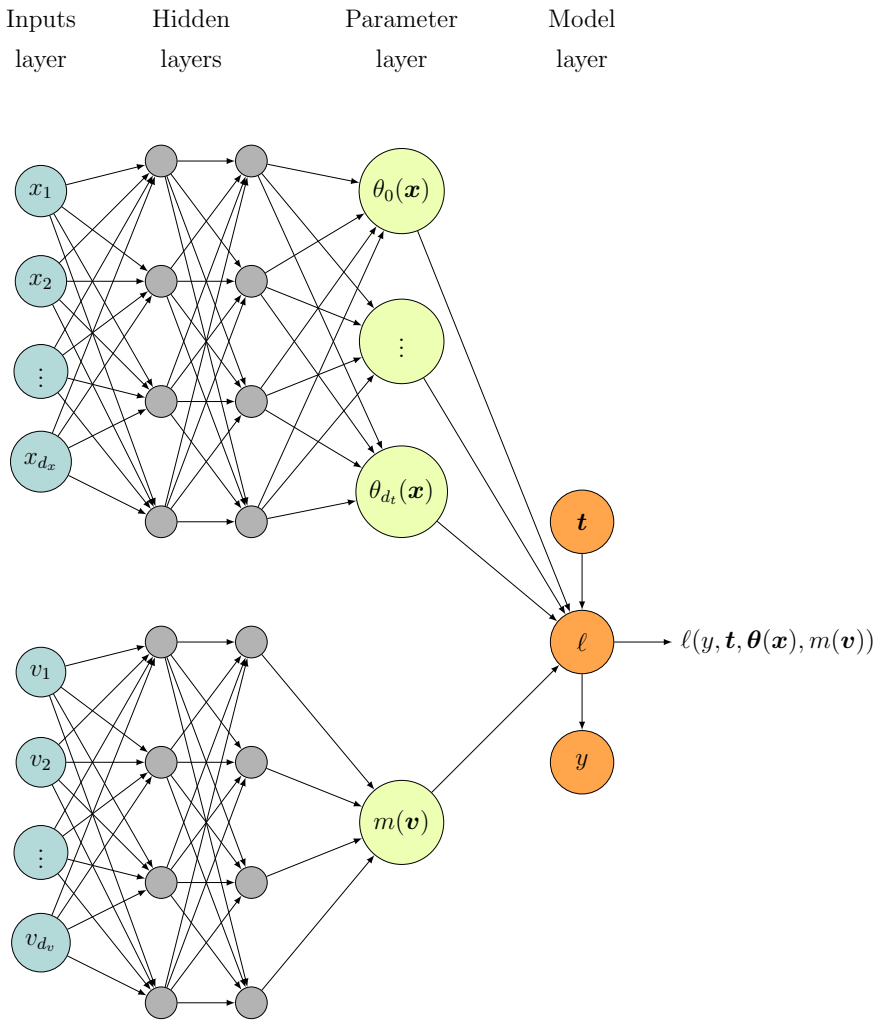


Figure 1: Illustration of the structured deep neural network estimation of the parameter functions $\boldsymbol{\theta}(\boldsymbol{x})$ and control function m($\boldsymbol{v}$) for generic structured model (1)

The structural relationship between $\mathbf{Y}$ and $\mathbf{T}$ is following with an economic model and statistically connected by a parametric loss function $\ell(\mathbf{Y}, \mathbf{T}, \boldsymbol{\theta})$ where parameter $\boldsymbol{\theta} \in \mathbb{R}^{d_{\boldsymbol{\theta}}}$ solved by $\min_{\theta \in \Theta} \mathbb{E}[\ell(\boldsymbol{Y}, \boldsymbol{T}, \boldsymbol{\theta})]$ of a standard M-estimation. In this framework, the parameters $\boldsymbol{\theta}$ are set as heterogeneity by recasting $\boldsymbol{\theta}$ into individual parameter functions $\boldsymbol{\theta}(\boldsymbol{X})$: $\mathbb{R}^{d_{\boldsymbol{X}}} \to \mathbb{R}^{d_{\boldsymbol{\theta}}}$, as in Farrell et al. (2021), where $\boldsymbol{X}$ describes individual characteristics. Intuitively, $\boldsymbol{\theta}$ is the typical causal parameters in parametric models as constant slopes in linear regression model. $\boldsymbol{\theta}$ represents homogeneous causal effects for any individual while $\boldsymbol{\theta}(\boldsymbol{x})$ represents for individual causal effect based on individual characteristics. $\boldsymbol{\theta}(\boldsymbol{x})$ can be regarded as functions of causal effects based on individual characteristics. It can allow us to provide individual policy targeting

based on different type of persons.

However, the individual parameter functions $\theta(\boldsymbol{X})$ may not be identified as endogeneity or sample bias issues. The key aspect of methodology is we ensure accurate identifications of structural parameter functions $\theta(\boldsymbol{X})$ with generated regressors, $\mathbf{V} \in \mathbb{R}^{d_{\mathbf{v}}}$, through control function. Hence, we identify true structural parameter functions and control function, $\boldsymbol{b}^*(\boldsymbol{X}, \boldsymbol{V}) = (m^*(\boldsymbol{V}), \theta^*(\boldsymbol{X})')'$, by solving we assume that the true parameter functions solve

$$\boldsymbol{b}^*(\cdot) = \arg\min_{\boldsymbol{b}\in\mathcal{F}} \mathbb{E}[\ell(Y, T, m(V), \boldsymbol{\theta}(\boldsymbol{X})] = \mathbb{E}[\ell(Y, T, m(V), \boldsymbol{b}(\boldsymbol{X}, \boldsymbol{V})]. \tag{1}$$

where $\mathcal{F}$ denotes an general functional class. With this framework, we not only retain the interpretable economic meaning of functions $\boldsymbol{\theta}(\boldsymbol{X})$ but also causally identify parameter functions in various context of endogeneity or sample bias. We guarantee revaling of the structural parameter functions $\boldsymbol{\theta}(\boldsymbol{X})$ to capture heterogeneity based on individual characteristics and allowing for individual policy targeting. Section (4) demonstrate a generalized deep learning framework with generated regressors is well-suited for revealing the structural parameter functions and incorporating control function with generated regressors providing the functionality of correction endogeneity or sample bias. The implementation of this new deep learning archictture inherit the advantages of recovering structurally meaningful functions going beyond the traditional roles of machine learning on prediction task and truthfully move toward estimation of well-identified parameter functions.

To ensure the implementation of structural deep learning architecture to allow for inference, we adopt a prepocessed adversarial training scheme to estimate parameter functions and control function with generated regressors. Importantly, it allows us to achieve sup-norm convergence at a fast enough rate of $\mathcal{O}(N^{-1/4})$. With sup-norm convergent property, the convergence of parameter functions and control function with inputs of generated regressors can be easily taken cared. For estimators $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}^*$ and $\hat{m}$ of m^*, our theorem proves

$$\mathbb{E}\left[\|m - m^*\|_{L^\infty}^2 + \sum_{k=1}^{d_\theta} C_k \|\theta_k - \theta_k^*\|_{L^\infty}^2\right] = O\left(n^{-2\beta/(2\beta+d_x+d_v)} \log^{2\vee\beta^*} n\right)$$

or for $\boldsymbol{b}^*(\boldsymbol{X}, V) = (m^*(V), \theta^*(\boldsymbol{X})')'$, we have

$$\mathbb{E}\left[\|\boldsymbol{b} - \boldsymbol{b}^*\|_{L^\infty}^2\right] = O\left(n^{-2\beta/(2\beta+d_x+d_v)} \log^{2\vee\beta^*} n\right),$$

when loss function is sufficiently smooth and curved near the truth. This result devried from our innovative structural deep learning architecture with generated regressors and together the adversarial training scheme with prepocessed technique.

With the ingredients of heterogeneous structural parameter functions $\theta^*(\boldsymbol{X})$ and control function $m(\boldsymbol{V})$, we construct smooth inferential function $\boldsymbol{H} : \{\mathbb{R}^{d_{\boldsymbol{X}}} \times \mathbb{R}^{d_{\boldsymbol{\theta}}} \times \mathbb{R}^{d_m}\} \to \mathbb{R}^{d_{\boldsymbol{\mu}}}$, and perform inference on

$$\boldsymbol{\mu}^* = \mathbb{E}\big[\boldsymbol{H}(\boldsymbol{X}, m^*(V), \theta^*(\boldsymbol{X}); \boldsymbol{t}^*)\big] = \mathbb{E}\big[\boldsymbol{H}(\boldsymbol{X}, \boldsymbol{b}^*(\boldsymbol{V}, \boldsymbol{X}); \boldsymbol{t}^*)\big], \tag{2}$$

where $\boldsymbol{t}^*$ is certain fixed value chosen from researcher. The inferential parameters of interest are constructed according to interesting statistics from economics models or straightforward reduced form. The key of this framework is to ensure the parameter functions are correctly identified and still remain their economics meaningfulness to construct inferential parameters according to economics models. The function form of $\boldsymbol{\mu}^*$ is chosen by researchers and deliver the flexible various inferential parameters as regularity conditions are satisfied.

To perform the third stage of inference, we derive the influence function that accounts for the presence of generated regressors, which is critical when estimating parameters in models that exhibit endogeneity or sample selection bias. Specifically, we derive a general Neyman orthogonal score that serves as the foundation for conducting valid inference in this complex setting. The Neyman orthogonal score allows us to implement inference procedures directly, accommodating various general structural models that are characterized by smooth inferential functions and smooth loss functions.

Theorem (4) in subsection of influence function demonstrates the Neyman orthogonal score on inferential parameters $\boldsymbol{\mu}^*$ with sufficient regularities. Let $H(\boldsymbol{x}, \boldsymbol{b}^*(\boldsymbol{x}, \boldsymbol{v}^*); \boldsymbol{t}^*)$ and $\boldsymbol{\ell_b}(\boldsymbol{w}, \boldsymbol{b}^*(\boldsymbol{x}, \boldsymbol{v}^*)))$ denotes the gradiants of inferential functions $\boldsymbol{H}$ and loss function ℓ with respect to $\boldsymbol{b}$ and $\boldsymbol{\Lambda}(\boldsymbol{x}, \boldsymbol{v}) := \mathbb{E}\left[\ell_{\boldsymbol{bb}}(\boldsymbol{Y}, \boldsymbol{T}, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v})) \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{V} = \boldsymbol{v}\right]$ denotes the conditional expectation of Hessian matrix function of loss function ℓ, conditioning on $\boldsymbol{b} = \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v})$. The Neyman orthogonal score is $\boldsymbol{\psi}(\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{x}, \boldsymbol{b}^*, \boldsymbol{\Lambda}) - \boldsymbol{\mu}^*$ where

$$\boldsymbol{\psi}(\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{x}, \boldsymbol{b}^*, \boldsymbol{\Lambda}) = \boldsymbol{H}(\boldsymbol{x}, \boldsymbol{b}^*(\boldsymbol{x}, \boldsymbol{v}^*); \boldsymbol{t}^*) - \boldsymbol{H_b}(\boldsymbol{x}, \boldsymbol{b}^*(\boldsymbol{x}, \boldsymbol{v}^*); \boldsymbol{t}^*)\boldsymbol{\Lambda}(\boldsymbol{x}, \boldsymbol{v}^*)^{-1}\boldsymbol{\ell_b}(\boldsymbol{w}, \boldsymbol{b}^*(\boldsymbol{x}, \boldsymbol{v}^*))).$$

The influence function we develop is particularly novel because it explicitly incorporates the impact of the estimation process for the generated regressors compared to Chernozhukov et al. (2018); Farrell et al. (2021). The usual influence functions assume that all regressors are observed and fixed in the context of DML framework, but in our framework, some regressors are generated through an initial estimation procedure. This additional layer of complexity requires a careful adjustment of the influence function to ensure that it remains valid and robust. One of the key advantages of our influence function is its convenience in implementation across a wide

range of econometric models. By integrating the effect of generated regressors into the influence function, we provide a unified framework that can be applied to various models without the need for substantial modifications. With knowing smooth function $\boldsymbol{H}(\boldsymbol{x}, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}^*); t^*)$ and three pieces of bias correction term including $\boldsymbol{H_b}(\boldsymbol{x}, \boldsymbol{b}^*(\boldsymbol{x}, \boldsymbol{v}^*); \boldsymbol{t}^*)$, $\boldsymbol{\Lambda}(\boldsymbol{x}, \boldsymbol{v}^*)$, $\boldsymbol{\ell_b}(\boldsymbol{w}, \boldsymbol{b}^*(\boldsymbol{x}, \boldsymbol{v}^*)))$, we can simply calculate each piece value through automatic differentiation without full derivation of analytical forms as in Farrell et al. (2021) and obtain inference. This versatility is crucial for practitioners who need to apply sophisticated econometric methods to real-world data, where endogeneity and high-dimensional covariates are common challenges.

With Neyman orthogonal score and estimates from the first and second stages, we can obtain the estimator of the inferential parameters $\widehat{\boldsymbol{\mu}}$ and estimator of the asymptotic variance $\widehat{\boldsymbol{\Psi}}$. Finally, the standard error $\widehat{\mathrm{SE}}(\hat{\boldsymbol{\mu}})$ is derived from $\hat{\boldsymbol{\Psi}}$, and the confidence intervals for $\boldsymbol{\mu}^*$ are constructed as:

$$\hat{\boldsymbol{\mu}} \pm z_{\alpha/2} \cdot \widehat{\mathrm{SE}}(\hat{\boldsymbol{\mu}})$$

where $z_{\alpha/2}$ is the critical value from the standard normal distribution.

The simulation results provide strong empirical support for our proposed semiparametric framework. Notably, our estimator consistently maintains a stable coverage rate around the nominal 95%, under different levels of correlation ρ_{UV} , highlighting its robustness in handling endogeneity. In contrast, traditional methods like the two-stage least squares (2SLS) and the estimator by Farrell et al. (2021) exhibit significant bias and lower coverage rates, particularly in high-endogeneity scenarios.

Overall, the simulation results confirm that our framework outperforms existing methods, particularly in scenarios involving generated regressors. This validates the practical relevance and adaptability of our approach, making it a valuable tool for modern econometric analysis.

3 Model Setting and Examples

In this section, we introduce the model framework to capture strutrual individual heterogeneity with generated regressors. Suppose the observed variables are $W = (Y, T, \boldsymbol{X}, Z)$ with cumulative distribution function F^*. For exhibition, let's assume Y and T are univariate[1].

In this framework, the existed relation between T and Z deliver the generated regressor in the first step.

[1]Model setting notations refer to Escanciano and Pérez-Izquierdo (2023)

$$V \equiv \varphi(T, Z, g^*) \tag{3}$$

where $g^*(\cdot)$ is the true population function between T and Z in the Hilbert space $L_2(Z)$ [2] and obtained by solving the following the orthogonal moments

$$\mathbb{E}[\gamma_1(Z)(T - g^*(Z))] = 0 \text{ for any } \gamma_1 \in L_2(Z), \tag{4}$$

and typically, first-step unknown functions consist of reduced form functions of exogenous variables such as conditional mean (or quantile) regressions or conditional choice probabilities(Olley and Pakes, 1992; Newey et al., 1999; Newey and Powell, 2003). Specifically, we allow for the generated regressors to be estimated from various machine learning(ML) methods, such as lasso-type models, tree-based models, neural network models and mixtures among ML methods.

The second step of this setting is the relation between Y and $(T, \boldsymbol{X}, Z)$ by satisfying the moment restrictions

$$\mathbb{E}[\gamma_2(T, \boldsymbol{X}, V)(Y - h^*(\boldsymbol{X}, T, V))] = 0 \text{ for any } \gamma_2 \in L_2^{res}(T, \boldsymbol{X}, V|g^*), \tag{5}$$

where the regression model h^* is located in a Hilbert space of semiparametric setting $L_2^{res}(T, X, Z|g_0) = \{h(\cdot)|\int h(X,T,V)^2 < \infty, h(X,T,V) = \theta(\boldsymbol{X})T + m(V)\}$. $\theta(\mathbf{X})$ denotes the structural heterogeneous parameter function which captures the individual treatment effect of treatment variable T with respect to individual characteristics $\mathbf{X}$.

With the structural heterogeneous parameter functions $\theta^*(\boldsymbol{X})$ and control function $m^*(V)$ from h^*, we can construct the structural parameters of inferential interest denoted by $\mu^* \in \mathbb{R}^{d_{\boldsymbol{\mu}}}$, and its smooth functions are represented by $\boldsymbol{H} : \{\mathbb{R}^{d_{\boldsymbol{X}}} \times \mathbb{R}^{d_{\boldsymbol{\theta}}} \times \mathbb{R}^{d_m}\} \to \mathbb{R}^{d_{\boldsymbol{\mu}}}$,

$$\boldsymbol{\mu}^* = \mathbb{E}\big[\boldsymbol{H}(\boldsymbol{X}, m^*(V), \theta^*(\boldsymbol{X}); \boldsymbol{t}^*)\big] \tag{6}$$

where $\boldsymbol{t}_i^*$ fixed values. The smooth function $\mathbf{H}$, as specified by the researcher, adheres to economically interesting statistics mostly derived from structural economic models.

The following example is used to illustrate usages of this framework and explain the notation and concepts.

EXAMPLE 1 (*Control Function Approach*) Suppose the observed variables, $W = (Y, \boldsymbol{X}, T, Z)$, follows with the output equation

$$Y = \theta^*(\boldsymbol{X})T + U \tag{7}$$

[2] $L_2(Z) = \{g(\cdot)|\int g(Z)^2 dF_0 < \infty\}$

and

$$T = g^*(Z) + V \tag{8}$$

where T denotes endogenous variable and Z denotes instrumental variable. Define generated regressor

$$V \equiv \varphi(T, Z, g^*) = T - g^*(Z),$$

Under assumptions of control function approach, $\mathbb{E}[U|T, \mathbf{X}, V] = \mathbb{E}[U|\mathbf{X}, V] = \mathbb{E}[U|V] = \mathrm{m(V)}$, the control function approach derives the following

$$\mathbb{E}[Y|T, \boldsymbol{X}, V] = \theta^*(\boldsymbol{X})T + m(V).$$

Therefore, a straightforward inferential parameter, average treatment effect(ATE) is

$$\mu = \mathbb{E}[\theta^*(\mathbf{X})]$$

More relevant details of control function methods refer to Newey et al. (1999); Blundell and Powell (2004); Wooldridge (2015).

EXAMPLE 2 (*Sample Selection Models*) Suppose the observed variables $W = (Y, T, Z)$ follow the model $Y^* = \theta^*(\mathbf{X})T + \epsilon$ and $Y = DY^*$, which is observed. Here, Z denotes a vector of variables that determine the selection, and D denotes the binary selection indicator defined as $D = \mathbb{1}[g_0(Z) - U \geq 0]$, where U is uniformly distributed in $[0, 1]$. In this example, the generated regressor is:

$$V \equiv \varphi(T, Z, g_0) = g_0(Z) = \mathbb{E}[D \mid Z]$$

Thus, the conditional expectation of Y given Z is:

$$\mathbb{E}[Y \mid Z] = \theta^*(\mathbf{X})T + \lambda_0(V)$$

Consider the Average Partial Effects (APE) as the parameter of interest, defined as:

$$\mu = \mathbb{E}[\theta^*(\mathbf{X})]$$

For more details on the origin and semiparametric extension of sample selection models, refer to Heckman (1979); Ahn and Powell (1993).

4 Structural Deep Learning with generated regressors by adversarial training

In this section, we delve into the advantages of utilizing deep neural networks (DNN) for estimating structural parameter functions and control functions using generated regressors through adversarial training. This method harnesses the flexibility and power of DNNs to handle complex, high-dimensional data while providing robust estimations even in the presence of endogeneity.

The foremost advantage of employing deep neural networks in this framework is their capability to estimate structural parameter functions $\theta^*(X)$ and control functions $m(V)$ concurrently. This dual capability is illustrated in Figure 1 on the parameter layer, demonstrating the network's ability to integrate these components into a unified estimation process. Traditional econometric methods often treat these components separately, but the DNN architecture's simultaneous approach ensures more accurate and efficient estimations.

Deep learning models excel at capturing complex, high-dimensional relationships within data. Unlike classical nonparametric methods, which can struggle with high dimensionality, DNNs maintain performance and deliver outstanding results. This capability is crucial for econometric models where individual heterogeneity is a highly complex function of observable characteristics X. The architecture proposed effectively reveals structural parameter functions that capture this heterogeneity, enabling more precise and meaningful policy targeting and inference.

Our framework significantly differs from Farrell et al. (2021) approach by incorporating control functions with generated regressors directly within the DNN architecture. This design allows for the correction of endogeneity, a frequent challenge in econometric analysis. By addressing endogeneity directly within the network, our model can accurately estimate structural parameter functions and provide more reliable causal inferences. This integration ensures that the parameter functions retain their economic interpretability and structural validity, essential for meaningful economic analysis.

4.1 Sup-norm Convergence of deep neural network estimator for nonparametric regression by adversarial training

To achieve robust and accurate estimations, we employ a preprocessed adversarial training scheme. This method helps the model achieve sup-norm convergence, a property crucial for ensuring that the estimated functions uniformly converge to their true values at a fast rate.

Adversarial training involves training the model using inputs slightly perturbed to test its robustness. Originally designed to defend against adversarial attacks in classification tasks, this technique is adapted here for regression problems to enhance the model's resilience and accuracy. The preprocessed adversarial training framework smooths the output variable initially, reducing bias introduced by adversarial perturbations and ensuring the estimator remains consistent and efficient(Imaizumi, 2023).

Let's consider a nonparametric regression context, assume that there exist a identical and independent distributed sequence of random variables $(\boldsymbol{X}_1, Y_1), (\boldsymbol{X}_2, Y_2), ..., (\boldsymbol{X}_n, Y_n) \in [0,1]^d \times \mathbb{R}$ follows the regression model [3]

$$Y = f^*(\boldsymbol{X}) + \varepsilon. \tag{9}$$

In mathematics, the adversarial training framework defines its loss function by finding an input point within the neighborhood of a data observation that maximizes the loss. The neighborhood of $x \in [0,1]^d$ is specified by the proximity distance $h \in (\underline{h}, 1)$, where $\underline{h} > 0$:

$$\Delta_h^p(x) = \left\{ x' \in [0,1]^d \mid \|x - x'\|_p \leq h \right\} \subset [0,1]^d.$$

To illustrate this, we use a quadratic loss function as our empirical risk. With surrogate outputs $\hat{Y}$, the empirical preprocessed adversarial risk is defined as:

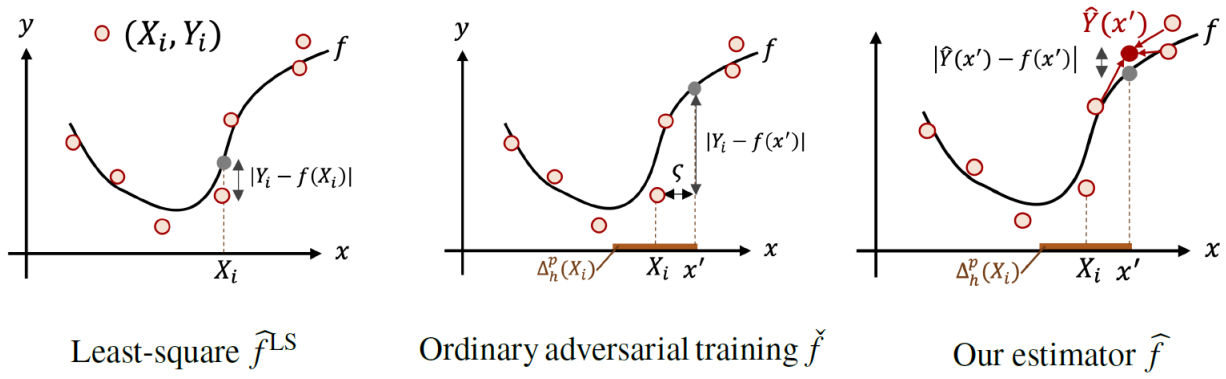
$$R_n(f) := \frac{1}{n} \sum_{i=1}^{n} \sup_{x' \in \Delta_h^p(X_i)} \left(\widehat{Y}(x') - f(x') \right)^2,$$

for a function $f \in L^2([0,1]^d)$. The preprocessed output $\hat{Y}$ is developed by a random map $\hat{Y} : [0,1]^d \to \mathbb{R}$, which can be created using methods like the k-nearest neighbor approach. This loss function extends the standard adversarial risk by incorporating the preprocessed $\widehat{Y}$.

The main purpose of preprocessing is to adapt the output variable in response to perturbations in the input X_i caused by adversarial training. When the input point X_i is shifted by $\xi = x' - X_i$, the output is also adjusted to accommodate this shift. This is achieved using

[3] We slightly have abuse of notation for $\boldsymbol{X}$, which is input of regression model here.

Figure 2: Structured deep neural networks to estimate parameters functions



preprocessing methods such as the nearest neighbor approach. Figure 2[4] illustrates the differences among the least-square estimator $\hat{f}^{LS}$, ordinary adversarial training $\check{f}$, and the proposed estimator $\hat{f}$ with adversarial training and preprocessing. Specifically, the figure shows how preprocessing helps the model better approximate the true function f by reducing the bias introduced by adversarial perturbations. This ensures that the estimator remains consistent and robust, leading to more accurate and reliable predictions.

By implementing this particular concept, we define an estimator as the element that works to minimize the empirical risk as

$$\widehat{f} \in \underset{f \in \mathcal{F}(L,H)}{\operatorname{argmin}} R_n(f) \tag{10}$$

where we define a set of functions using deep neural networks with L layers and H maximum width. Given a tuple $(L, H) \in \mathbb{N}^2$ and an upper bound $B \geq 1$, the set of functions by deep neural networks is defined as:

$$\mathcal{F}(L, H) := \{f_\theta \mid \|f_\theta\|_{L^\infty} \leq B, \theta \in \Theta_{L,\mathcal{W}}, \|\mathcal{H}\|_\infty \leq H\}.$$

The upper bound condition B can be met by applying a clipping operation with the ReLU activation function.

In this adversarial training framework with preprocessing, we derive a non-asymptotic unpper bound for the L^∞-risk of the estimator $\hat{f}$. Before presenting the theorem, the required assumptions on the theorem 1 are standard and common in nonparametric regresssion analysis. Except assumption 6, it requires certain extend of fast sup-norm convergent rate for

[4]The source of figure is from Imaizumi (2023)

prepocessing $\hat{Y}$ in this framework. More detailed descriptions of these common assumptions for our convergent theoreoms are presented in section 4.2. These assumptions are the following:

Assumption 1. *Suppose marginal measure $P_{\boldsymbol{X}}$ of $\boldsymbol{X}$ has a density function which is uniformly lower bounded by $C_{P_{\boldsymbol{X}}} > 0$ on $[0,1]^d$.*

Assumption 2. *There exists $\beta > 0$ such that $f^*(\boldsymbol{x})$ lies in $\mathcal{W}^{\beta',\infty}\left([0,1]^d\right)$ with any smoothness $\beta' \in (0, \beta]$,*

$$f_*(x) \in \mathcal{W}^{\beta',\infty}\left([0,1]^d\right) := \left\{ f : \max_{\boldsymbol{\alpha}, |\boldsymbol{\alpha}| \leq \beta'} \mathrm{ess}_{x \in [0,1]^d} \left| D^{\alpha} f(x) \right| \leq B \right\}$$

with its radius $B \geq 1$, where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d), |\boldsymbol{\alpha}| = \alpha_1 + \cdots + \alpha_d$ and $D^{\boldsymbol{\alpha}} f$ is the weak derivative.

Assumption 3 (Preprocessing)**.** *$\widehat{Y}(x)$ is continuous and $\mathbb{E}\left[\|\widehat{Y}\|_{L^\infty}^2\right] \leq V^2$ with some $V > 0$. Also, there exists a non-negative sequence $\{\zeta_n\}_{n \in \mathbb{N}}$ such that $\zeta_n \to 0$ as $n \to \infty$ such that the following holds for all $n \in \mathbb{N}$:*

$$\mathbb{E}\left[\left\| \widehat{Y} - f^* \right\|_{L^\infty}^2 \right] \leq \zeta_n^2$$

where the preprocessed output is developed by a random map $\hat{Y} : [0,1]^d \to \mathbb{R}$ which can be completed by several methods like the k-nearest neighbor method.

Hence, the asymptotic property of estimator $\hat{f}$ is given in the following

Theorem 1. *(General Feedforward Architecture) Suppose that Assumption 1, 2, and 6 hold for some $\beta > 0$. Let $f^*(\boldsymbol{x})$ be regression function 9 and $\widehat{f}$ be corrected adversarial estimator 10 under the function class $\mathcal{F}(L, H)$ of deep neural networks. Then*

$$\mathbb{E}\left[\left\| \widehat{f} - f^* \right\|_{L^\infty}^2 \right] \leq C_{P_X, p, B, d, \beta} h^{-d} \left(\frac{(1 + h^{-d})(HL)^2 \log(HL) \log n}{n} + (LH)^{-4\beta/d} + \mathbb{E}\left[\|\Xi\|_{L^\infty}^2 \right] \right)$$

for every $n \geq \bar{n}$ with some $\bar{n} \in \mathbb{N}$.

The upper bound includes three terms: the first term $O((HL)^2 \log(HL) \log n)$ represents the complexity error and the second term $O((LH)^{-4\beta/d})$ describes the approximation error. Both of these two errors typically exist in risk bounds on an L^2- risk of deep neural networks(Farrell et al., 2020; Shen et al., 2021). More relevant details can refer to Imaizumi (2023).

With dedicate selection of width and depth of deep neural networks, we can further derive the convergence rate of the $L^\infty-$risk with respect to sample size n, as following:

Corollary 1. *Suppose that Assumption 1, 2, and 6 hold for some $\beta > 0$. Let $f^*(\boldsymbol{x})$ be regression function 9 and $\widehat{f}$ be corrected adversarial estimator 10 under the function class $\mathcal{F}(L, H)$ of deep neural networks with depth L and width H setted as $HL \asymp n^{d/(4\beta+2d)}$. Assume that $\zeta_n^2 = O\left(n^{-2\beta/(2\beta+d)} \log^{\beta^*} n\right)$ for some $\beta^* > 0$. Then*

$$\mathbb{E}\left[\left\|\widehat{f} - f^*\right\|_{L^\infty}^2\right] = O\left(n^{-2\beta/(2\beta+d)} \log^{2\vee\beta^*} n\right).$$

4.2 Sup-norm Convergence of Structural Deep Neural Network with Generated Regressors by Adversarial Training

In this section, we thoroughly explore the sup-norm convergence properties of structural deep neural networks (DNNs) when integrated with generated regressors via adversarial training. Our objective is to establish a comprehensive theoretical framework and practical results demonstrating the sup-norm convergence rates, with detailed comparison to existing methodologies.

Our analysis hinges on several key assumptions, which can be categorized into three main groups: (1) regularity conditions on the random vector $W = (Y, \boldsymbol{X}', \boldsymbol{T}', \boldsymbol{Z}')$, (2) conditions on the parameter functions $\boldsymbol{\theta}^*(\boldsymbol{X})$ and $m^*(\boldsymbol{V})$, and (3) requirements concerning the preprocessing step $\hat{Y}(\boldsymbol{x}, \boldsymbol{t}, \boldsymbol{v})$ and the model governed by the loss function $\ell(y, \boldsymbol{t}, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}))$.

For the regularity conditions on $W = (Y, \boldsymbol{X}', \boldsymbol{T}', \boldsymbol{Z}')$, we assume that all continuous random variables involved have compact supports and that their corresponding density functions are uniformly bounded below by a positive constant.

Assumption 4. *Assume that the marginal distributions $P_{\boldsymbol{X}}, P_{\boldsymbol{V}}, P_{\boldsymbol{T}}$ for the variables $\boldsymbol{X}, \boldsymbol{V}, \boldsymbol{T}$ have density functions that are uniformly bounded below by positive constants $C_{P_{\boldsymbol{X}}} > 0$ on $[0,1]^{d_x}$, $C_{P_{\boldsymbol{V}}} > 0$ on $[0,1]^{d_v}$, and $C_{P_{\boldsymbol{T}}} > 0$ on $[0,1]^{d_t}$, respectively.*

These conditions are standard in the context of nonparametric regression involving neural networks and have been employed in previous studies (e.g., Farrell et al. (2020), Schmidt-Hieber (2020), Bauer and Kohler (2019)). The inclusion of binary random variables is straightforward and does not alter the validity of the results. For simplicity, we do not explicitly include binary variables in the notation.

We also assume that the parameter functions $\boldsymbol{\theta}^*(\boldsymbol{X})$ and the control function $m^*(\boldsymbol{V})$ belong to a Hölder space characterized by a smoothness index β. Our focus is on deep neural networks with fully-connected layers and ReLU activation functions, which are well-suited for capturing the smoothness properties of these functions.

Assumption 5. *Assume that all components of* $\mathbf{W} = (\boldsymbol{Y}', \boldsymbol{T}', \boldsymbol{X}', \boldsymbol{Z}')$ *are bounded random variables and that* $\boldsymbol{b}^*(\boldsymbol{X}, V) = (m^*(V), \theta^*(\boldsymbol{X})')'$ *in equation (1) are nonparametrically identified. Furthermore, there exists a constant* $\beta > 0$ *such that each component* $\theta_k(\boldsymbol{x}) \in \mathcal{W}^{\beta',\infty}([0,1]^{d_x})$ *for* $k = 1, \ldots, d_\theta$ *and* $m(\boldsymbol{v}) \in \mathcal{W}^{\beta',\infty}([0,1]^{d_v})$ *with any smoothness* $\beta' \in (0, \beta]$. *The Hölder ball* $\mathcal{W}^{p,\infty}([0,1]^q)$ *is defined for* $p, q \in \mathbb{N}^+$ *as:*

$$\mathcal{W}^{p,\infty}([0,1]^q) := \left\{ f : \max_{\boldsymbol{\alpha}, |\boldsymbol{\alpha}| \leq p} \operatorname{ess}_{x \in [0,1]^q} |D^\alpha f(x)| \leq B \right\},$$

where $B \geq 1$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)$, $|\boldsymbol{\alpha}| = \alpha_1 + \cdots + \alpha_d$, *and* $D^{\boldsymbol{\alpha}} f$ *denotes the weak derivative.*

Imposing a smoothness condition via the Hölder ball is a classical approach in nonparametric regression models (e.g., Farrell et al. (2020), Schmidt-Hieber (2020), Imaizumi (2023)). The convergence rate is closely tied to the Hölder smoothness parameter β, which is a critical factor in achieving a sufficiently fast rate of convergence for subsequent inference.

We require the preprocessing $\hat{Y}$ to satisfy a certain level of convergence as established in Imaizumi (2023). The following non-negative sequence $\{\zeta_n\}_{n \in \mathbb{N}}$ represents the convergence rate of $\hat{Y}$ to f^*, as well as the finiteness of the second moment, which will be important in the general loss function setting.

Assumption 6 (Preprocessing)**.** $\widehat{Y}(x)$ *is continuous and* $\mathbb{E}\left[\|\widehat{Y}\|_{L^\infty}^2\right] \leq V^2$ *for some* $V > 0$. *Additionally, there exists a non-negative sequence* $\{\zeta_n\}_{n \in \mathbb{N}}$ *such that* $\zeta_n \to 0$ *as* $n \to \infty$, *with the following holding for all* $n \in \mathbb{N}$*:*

$$\mathbb{E}\left[\left\|\widehat{Y} - f^*\right\|_{L^\infty}^2\right] \leq \zeta_n^2,$$

where the preprocessed output is developed by a random map $\hat{Y} : [0,1]^d \to \mathbb{R}$, *which can be implemented using several methods including the k-nearest neighbor method and posterior mean by Bayesian methods as mentioned in Imaizumi (2023).*

In the k-nearest neighbor method, we have $\xi_n^2 = O(n^{-2\beta/(2\beta+d)})$ with the setting of $k \asymp n^{2\beta/(2\beta+d)}$ as established in Theorem 1 of Jiang (2019). The key requirement for preprocessing is that the error in estimating a smooth function converges in the sup-norm sense at a sufficiently fast rate, as the preprocessing error will impact the convergence rate of the nuisance functions and subsequent inference.

For the loss functions, we impose Lipschitz continuity and a degree of curvature to ensure proper estimation of $\boldsymbol{\theta}^*(\boldsymbol{X})$ and $m^*(\boldsymbol{V})$. Additional assumptions on the loss functions may

be required during the inference stage. Similar conditions have been considered in prior work (Farrell et al. (2020, 2021); Imaizumi (2023); Shen et al. (2021)).

Assumption 7. *Assume the loss function $\ell(y, \boldsymbol{t}, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}))$ is symmetric with respect to y and $\boldsymbol{t}'\boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v})$, and that it is Lipschitz continuous in each argument with a Lipschitz constant $C_\ell > 0$. Additionally, $\ell(y, \boldsymbol{t}, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v})) = 0$ if and only if $y = \boldsymbol{t}'\boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v})$, and there exists a constant $c_\ell > 0$ and $q \geq 1$ such that:*

$$\ell(y, \boldsymbol{t}, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v})) \geq c_\ell |y - \boldsymbol{t}'\boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v})|^q, \quad \forall y, \boldsymbol{t}'\boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}) \in \mathbb{R}.$$

A range of common loss functions satisfy Assumption 7, including the absolute loss $\ell(y, x) = |y - x|$, the quantile loss $\ell(y, x) = (\mathbf{1}\{y \geq x\}\tau + 1\{y \leq x\}(\tau - 1))(y - x)$ for $\tau \in (0, 1)$, and the Cauchy loss $\ell(y, x) = \log(1 + \kappa^2(y - x)^2)$ for $\kappa > 0$.

Now we will formally present our theoretical results for sup-norm convergent rate. Since we focus on structured deep neural network, we define the set of deep neural networks, $\mathcal{F}_{L_h, H_h, L_c, H_c}$ as shown in Figure 5, composed of upper and lower modules with an upper bound B. The structure of the upper neural network is defined by the depth L_h and width H_h, while the lower one is determined by the depth L_c and width H_c. Given a tuple $(L_h, H_h, L_c, H_c) \in \mathbb{N}^4$ and an upper bound $B \geq 1$, the set of functions by deep neural networks is defined as:

$$\mathcal{F}(L_h, H_h, L_c, L_c) := \left\{ \boldsymbol{b}_\theta \mid \|m_\theta\|_{L^\infty} \leq B, \|\theta_{k,\theta}\|_{L^\infty} \leq B \text{ for } k = 1, \ldots, d_\theta \text{ and } \theta \in \Theta_{L_h, \mathcal{H}_h, L_c, H_c} \right\}.$$

Let's we first consider a quadratic loss function as our empirical risk for illustration. With surrogate outputs $\hat{Y}$, the empirical preprocessed adversarial risk is defined as

$$R_n(\boldsymbol{b}) := \frac{1}{n} \sum_{i=1}^{n} \max_{(\boldsymbol{x}, \boldsymbol{v}) \in \Delta_h^p(X_i, V_i)} \left(\widehat{Y}\left(\boldsymbol{x}, \boldsymbol{v}, \boldsymbol{t}_i\right) - m(\boldsymbol{v}) - \boldsymbol{t}_i' \boldsymbol{\theta}(\boldsymbol{x}) \right)^2$$

for $m(\mathbf{v}), \{\theta_k(x)\}_{k=1}^{d_\theta} \in L^2\left([0, 1]^d\right)$. This loss function is an extended form of the standard adversarial risk with preprocessing $\widehat{Y}$. By implementing this particular concept, we define an estimator as the element that works to minimize the empirical risk as

$$\widehat{\boldsymbol{b}} \in \underset{\boldsymbol{b} \in \mathcal{F}(L_h, H_h, L_c, H_c)}{\operatorname{argmin}} R_n(\boldsymbol{b}) \tag{11}$$

Theorem 2. *(General Structure Feedforward Architecture) Suppose that Assumption 4, 5, and 6 hold for some $\beta > 0$. Let $\boldsymbol{b}^*(\boldsymbol{x}, \boldsymbol{v})$ be regression function 1 and $\widehat{\boldsymbol{b}}$ be corrected adversarial*

estimator 11 under the function class $\mathcal{F}(L_h, H_h, L_c, H_c)$ *of deep neural networks. Then*

$$\mathbb{E}\left[\|m - m^*\|_{L^\infty}^2 + \sum_{k=1}^{d_\theta} C_k \|\theta_k - \theta_k^*\|_{L^\infty}^2\right]$$
$$\leq C_{P_V, P_X, p, d_v, d_x, h, B}\left(\frac{H^2 L^2 \log(HL)\log(n)}{n} + \mathbb{E}\left[\|\Xi\|_{L^\infty}^2\right] + (LH)^{-4\beta/d_v} + (LH)^{-4\beta/d_x}\right)$$

for every $n \geq \bar{n}$ *with some* $\bar{n} \in \mathbb{N}$.

Theorem 2 deliver the generalized results to accommodate a broader range of neural network architectures, allowing for flexibility in design. This includes fixed-width, very deep networks as well as fixed-depth, very wide networks, making the results applicable to a variety of user-specified configurations. With dedicate selection of depth and width, we can further obtain sup-norm convergent rate with respect to n.

Corollary 2. *Suppose that Assumption 4, 5, and 6 hold for some* $\beta > 0$. *Let* $\boldsymbol{b}^*(\boldsymbol{x}, \boldsymbol{v})$ *be regression function 1 and* $\widehat{\boldsymbol{b}}$ *be corrected adversarial estimator 11 under the function class* $\mathcal{F}(L_h, H_h, L_c, H_c)$ *of deep neural networks with depth* $L_h = L_c = L$ *and width* $H_h = L_h = H$ *setted as* $HL \asymp n^{(d_x+d_v)/(4\beta+2d_x+2d_v)}$. *Assume that* $\zeta_n^2 = O\left(n^{-2\beta/(2\beta+d_v+d_x)} \log^{\beta^*} n\right)$ *for some* $\beta^* > 0$. *Then*

$$\mathbb{E}\left[\|m - m^*\|_{L^\infty}^2 + \sum_{k=1}^{d_\theta} C_k \|\theta_k - \theta_k^*\|_{L^\infty}^2\right] = O\left(n^{-2\beta/(2\beta+d_x+d_v)} \log^{2\vee\beta^*} n\right)$$

and for $\boldsymbol{b}^*(\boldsymbol{X}, V) = (m^*(V), \theta^*(\boldsymbol{X})')'$, *we have*

$$\mathbb{E}\left[\|\boldsymbol{b} - \boldsymbol{b}^*\|_{L^\infty}^2\right] = O\left(n^{-2\beta/(2\beta+d_x+d_v)} \log^{2\vee\beta^*} n\right).$$

The result of Corollary 2 shows robust convergent property in terms of sup-norm perspective compared other deep neural networks(Farrell et al., 2020, 2021) and complement the convergent properties in nonparametric M estimation literature. The derived convergent rate is identical to the minimax optimal rate of risk measured in the sup-norm in the problem of estimating a function from $\mathcal{H}^\beta\left([0,1]^d\right)$ (Stone, 1980, 1982). Sup-norm convergent property of individual parameter functions and control function can better deal with deriving the convergent rate with the influence of generated regressors. The estimators $\hat{\boldsymbol{b}}(\boldsymbol{X}, \hat{\boldsymbol{V}})$ have two source of sample variations. One typical source is to estimate the $\boldsymbol{b}$ if we truthly know generated regressors $\boldsymbol{v}$ where this case is usual sample variation for normal estimators. The other source of sample variation is coming from estimation of generated regressors $\hat{V}$ and its variation passed

to estimation of $\hat{\boldsymbol{b}}(\boldsymbol{X}, \hat{\boldsymbol{V}})$. With the sup-norm convergent property, the input distribution of generated regressors can be greatly disregarded since the sup-norm convergence property ensures that the maximum deviation between the estimated and true functions over all possible inputs is minimized.

The following Theorem demonstrate the sup-norm property in the general loss functions. As mentioned above, we provide sup-norm convergent property in more general loss function setting. It makes this framework more adaptable to various estimators in terms of M estimation with generated regressors, which can allow us to deal with endogeneity in terms of various econometric models.

We consider a general loss function which satisfies assumption 7. Similarly, the empirical preprocessed adversarial risk is generalized as

$$\widetilde{R}_n(\boldsymbol{b}) := \frac{1}{n}\sum_{i=1}^{n} \max_{(\boldsymbol{x},\boldsymbol{v})\in\Delta_h^p\left(X_i,\hat{V}_i\right)} \ell(\widehat{Y}\left(\boldsymbol{x},\boldsymbol{v},\boldsymbol{t}_i\right),\boldsymbol{t}_i,\boldsymbol{b}(\boldsymbol{x},\boldsymbol{v}))$$

for $m(\mathbf{v}), \{\theta_k(x)\}_{k=1}^{d_\theta} \in L^2\left([0,1]^d\right)$. With general loss function setting, we obtain an generalized estimator as the element that works to minimize the empirical risk as

$$\widetilde{\boldsymbol{b}} \in \underset{\boldsymbol{b}\in\mathcal{F}(L_h,H_h,L_c,H_c)}{\operatorname{argmin}} \widetilde{R}_n(\boldsymbol{b}) \tag{12}$$

Theorem 3. *Consider the regression model (1) and the adversarial estimator $\tilde{\boldsymbol{b}}$ in (12) with the function class by deep neural networks with a tuple (L_h, H_h, L_c, H_c) and $h \in (0,1)$. Suppose Assumption 4 and 5 for $\beta > 0$, Assumption 6 holds with $\zeta_n^2 = O\left(n^{-2\beta/(2\beta+d)}\log^{\beta^*} n\right)$ for some $\beta^* > 0$ and $\widehat{Y}$ is independent of $\{(\boldsymbol{X}_i, \boldsymbol{T}_i, \boldsymbol{Z}_i, Y_i)_{i=1}^n\}$, and Assumption 7 holds with $q \in [1,\infty)$. Then, we have the following as $n \to \infty$:*

$$\mathbb{E}\left[\left\|\tilde{\boldsymbol{b}} - \boldsymbol{b}^*\right\|_{L^\infty}^2\right] =\leq C_{P_X,B,p,d,\ell,q,V} h^{-2d/q}\left\{n^{-\beta/(q(\beta+d))}\log^{4/q} n + n^{-2\beta/(2\beta+d)}\log^{\beta^*} n\right\}. \tag{13}$$

With certain prior smoothness indexed by β and q, our adversarial training scheme achieves a convergence rate of $O(N^{-1/4})$, which is fast enough for later inference. This convergence rate is critical for wide inferential applications of M estimation requiring precise function estimation. The sup-norm convergent rate in this section contributes to the literature on statistical properties of deep learning for M estimation with generated regressors.

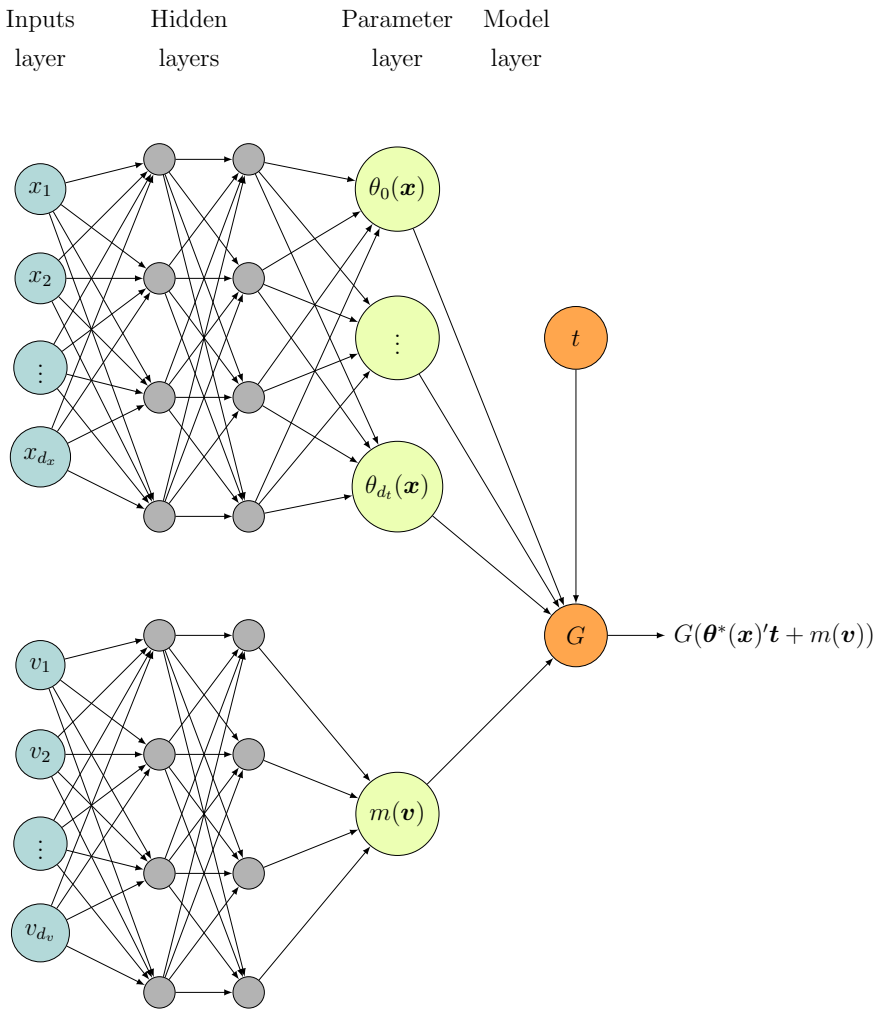


Figure 3: Illustration of the structured deep neural network estimation of the parameter functions $\boldsymbol{\theta}(\boldsymbol{x})$ and control function m($\boldsymbol{v}$) for generic model (14)

The above assumptions are high level in the context of general loss functions. The usual verification is necessary but straightforward in most of cases. To illustrate this, we start with a prevalent empirical case where loss function is built around conditional mean restriction and the parameters are the intercept and the slopes as in Farrell et al. (2021). Let's assume a know function $G(u), u \in \mathbb{R}$,

$$\mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{T} = \boldsymbol{t}, \boldsymbol{V} = \boldsymbol{v}] = G\left(\theta^*(\boldsymbol{x})'\boldsymbol{t} + m^*(\boldsymbol{v})\right). \tag{14}$$

where $\boldsymbol{t}$ includes a constant term, $\boldsymbol{v}$ are generated regressors and $m^*(\boldsymbol{v})$ is the control function used to deal with endogeneity or sample selection bias. The G function is linked to the dedicate selection of loss function. For example, The typical squared loss function relate G function to

identity function and nonlinear least squares or likelihood link G function with logistic function. It's also straightforward to adapt to generalized additive models with this architectural design with generated regressors.

The required assumptions are following

Assumption 8. *(i) The conditional expectation $G\left(\boldsymbol{\theta}^*(\boldsymbol{x})'\boldsymbol{t} + m^*(\boldsymbol{v})\right)$ is incorporated into the loss function via a known, real-valued transformation $g(\cdot)$. Both g and G are continuously invertible, and the normalized functions $g/\|g\|_\infty$ and $G/\|G\|_\infty$ belong to the Hölder space $\mathcal{W}^{\beta',\infty}([0,1])$ for some smoothness parameter $\beta' \in (0,\beta)$. (ii) Assumption 7 holds, with the loss function $\ell(\boldsymbol{y},\boldsymbol{t},\boldsymbol{\theta}(\boldsymbol{x}))$ replaced by $\ell(\boldsymbol{y},g)$, where the conditions are applied to the scalar argument g. (iii) The eigenvalues of the matrix $\mathbb{E}\left[\boldsymbol{T}\boldsymbol{T}' \mid \boldsymbol{X}=\boldsymbol{x}, \boldsymbol{V}=\boldsymbol{v}\right]$ are uniformly bounded and strictly positive across all $\boldsymbol{x}$ and $\boldsymbol{v}$.*

Our assumption of 8 requires certain smoothness and curvature assumptions. Differently, the standard positive variance condition conditioning on not only charatertistics $\boldsymbol{x}$ but also generated regressors $\boldsymbol{v}$. These assumptions are common and standard in M estimation. The main difference is to include generated regressors and control function in these assumptions, which are kind of being acceptable and not harming the generality of adopting this structured deep neural network framework as in Farrell et al. (2021). Then by applying Theorem 2 to this generic model, we can obtain the following theoretical results.

Corollary 3. *Consider the regression model (1) and the adversarial estimator $\tilde{\boldsymbol{b}}$ in (12) with the function class by deep neural networks with a tuple (L_h, H_h, L_c, H_c) and $h \in (0,1)$. Suppose Assumption 4 and 5 for $\beta > 0$, Assumption 6 holds with $\zeta_n^2 = O\left(n^{-2\beta/(2\beta+d)} \log^{\beta^*} n\right)$ for some $\beta^* > 0$ and $\widehat{Y}$ is independent of $\{(\boldsymbol{X}_i, \boldsymbol{T}_i, \boldsymbol{Z}_i, Y_i)_{i=1}^n\}$, and Assumption 8 holds with $q \in [1,\infty)$. Then, we have the following as $n \to \infty$:*

$$\mathbb{E}\left[\left\|\tilde{\boldsymbol{b}} - \boldsymbol{b}^*\right\|_{L^\infty}^2\right] = O(n^{-\beta/(q(\beta+d))} \log^{4/q} n + n^{-2\beta/(2\beta+d)} \log^{\beta^*} n) \tag{15}$$

and

$$\mathbb{E}\left[\left\|G(\widetilde{\boldsymbol{\theta}}(\boldsymbol{x})'\boldsymbol{t} + \widetilde{m}(\hat{\boldsymbol{v}})) - G(\boldsymbol{\theta}^*(\boldsymbol{x})'\boldsymbol{t} + m^*(\boldsymbol{v}))\right\|_{L^\infty}^2\right] = O(n^{-\beta/(q(\beta+d))} \log^{4/q} n + n^{-2\beta/(2\beta+d)} \log^{\beta^*} n)$$

The above results demonstrate the good convergent rate on parameter functions and control function in a sup-norm perspective, which is a direct application of Theorem 3. The cost of our proposed structured deep neural network with advasarial training is convergent rate depends

on $d = d_{\boldsymbol{x}} + d_{\boldsymbol{t}} + d_{\boldsymbol{v}}$ since we include a nonparametric preposessing technique and advasarial trianing scheme. However, with the rapid development of computation technique in deep neural network, computation cost may not be a worrisome problem for our proposed framework.

5 Semiparametric Inference and Influence Function

In econometrics, semiparametric estimation has become an essential tool for analyzing models that involve both parametric and nonparametric components. One particular challenge arises when generated regressors are introduced into the model. These generated regressors, often derived from a first-stage estimation, add complexity to the inference process, requiring the development of specialized methods to ensure valid and robust estimators. This section elaborates on the semiparametric inference framework developed in this paper, focusing on the integration of influence functions with generated regressors in the double machine learning frame work(Chernozhukov et al. (2018, 2021, 2022)), and situating this contribution within the broader econometric literature.

One main relevant strand of literature is that the studies on semiparametric estimators with generated regressors can be traced back to Ahn and Powell (1993); Newey et al. (1999); Imbens and Newey (2009); Rothe (2009); Mammen et al. (2012, 2016). These studies addressed the inherent challenges of dealing with generated regressors, particularly in the context of endogeneity and sample selection models. When regressors are generated, they introduce additional variability into the estimation process, which can lead to biased and inconsistent estimates if not properly accounted for. Our framework extends this literature by incorporating advanced machine learning techniques into the generation of regressors, allowing for a more flexible and robust modeling of individual heterogeneity under endogeneity or sample selection. Specifically, we develop an influence function that accounts for the generated regressors, ensuring that the asymptotic properties of the estimator—such as consistency and asymptotic normality—are preserved in the presence of high-dimensional nuisance parameter. This advancement is critical for applications where the structural parameters of interest are influenced by high-dimensional or complex observable characteristics with usage of Chernozhukov et al. (2018) method.

5.1 Influence Function

The influence function (Neyman orthogonal, double robust or locally robust) is a powerful tool for conducting inference in semiparametric models, as it captures the sensitivity of an estimator to perturbations in the data. Mathematically, it can be seen as the pathwise derivative of the estimator with respect to a perturbation in the empirical distribution. This strand of literature can refer to Newey (1994); Ichimura and Newey (2015, 2022) for more background details including regularity conditions of existence of influence functions and derivation recipe. Chernozhukov et al. (2018) introduced the double machine learning (DML) framework, which combines the flexibility of machine learning with the rigor of econometric inference. The DML framework is particularly powerful in high-dimensional settings, where it allows for the estimation of causal parameters while controlling for high-dimensional nuisance parameters. The cornerstone of this approach is the use of Neyman orthogonal scores, which ensure that the second-stage estimation is robust to errors in the first-stage machine learning predictions. However, their methodology primarily addresses cases without generated regressors.

In this paper, we generalize Chernozhukov et al. (2018); Farrell et al. (2021)'s approach to accommodate generated regressors. The key innovation lies in the decomposition of the debiasing correction into components that address both the first-stage generation of regressors and the second-stage estimation. This generalized influence function mitigates the bias introduced by the generated regressors. Our approach involves deriving an influence function that adapts to the structure of the generated regressors. This allows us to maintain the robustness of the inference process, even when the generated regressors are estimated through complex machine learning models. The derived influence function can be applied directly to various economic models, making it a versatile tool for conducting semiparametric inference in the presence of generated regressors.

The development of influence functions in the presence of generated regressors builds on a rich body of literature. Earlier studies by Hahn and Ridder (2013, 2019); Mammen et al. (2012, 2016) explored the asymptotic properties of semiparametric estimators with generated regressors. These contributions laid the groundwork for the derivation of influence functions with generated regressors, which is particularly relevant in the context of double machine learning framework. Our work contributes to this literature by providing a more general and robust framework in second stage for semiparametric estimation as we do not focus on regressions and pivot around interpretable parameter functions and control function, which is similar to Farrell et al. (2021) but still differ from it in the existence of generated regressors.

Let's introduce the formal results. We can present our assumption and then discuss on the form of influence function. Before we illustrate each main outcome, we need to define gradient and hessian matrix of loss function first. Let's define gradient $\ell_{\boldsymbol{b}}$ as an d_θ-vector of first derivatives of $\ell(y, \boldsymbol{t}, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}))$ with respect to $\boldsymbol{b}$ as following

$$\ell_{\boldsymbol{b}}(y, \boldsymbol{t}, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v})) = \left. \frac{\partial \ell(w, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right|_{\beta = \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v})} \tag{16}$$

and hessian matrix $\ell_{\boldsymbol{bb}}(y, \boldsymbol{t}, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}))$ as a $d_{\boldsymbol{b}} \times d_{\boldsymbol{b}}$ squared matrix of second-order partial derivatives of $\ell(y, \boldsymbol{t}, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}))$ with respect to $\boldsymbol{b}$ where the i-th row and j-th column element is defined as

$$\left[\ell_{\boldsymbol{\theta\theta}}(y, \boldsymbol{t}, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}))\right]_{i,j} = \left. \frac{\partial^2 \ell(\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{b})}{\partial b_i \partial b_j} \right|_{\boldsymbol{b} = \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v})}, \tag{17}$$

where b_i and b_j denote the respective i-th and j-th elements of $\boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v})$.

Similar to Farrell et al. (2021), the key required assumptions are first order condition of loss function and structural parameters $\boldsymbol{\mu}^*$ is pathwise differentiable. Both of conditions are guarantee to derive influence function for later inference.

Assumption 9. *The following conditions hold on the distribution of* $\boldsymbol{W} = (Y, \boldsymbol{X}', \boldsymbol{T}', \boldsymbol{Z}')'$*, uniformly in the given conditioning elements. (i) Equation (6) holds and identifies* $\boldsymbol{b}^*(\boldsymbol{x})$*, where* $\boldsymbol{\ell}(\boldsymbol{w}, \boldsymbol{b})$ *is thrice continuously differentiable with respect to* $\boldsymbol{b}$*. (ii)* $\mathbb{E}\left[\boldsymbol{J}(\boldsymbol{x}, \boldsymbol{v})\ell_{\boldsymbol{b}}(W, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v})) \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{V} = \boldsymbol{v}\right]$ $\boldsymbol{0}$*. (iii) For* $\ell_{\boldsymbol{bb}}$ *of (17),* $\boldsymbol{\Lambda}(\boldsymbol{x}, \boldsymbol{v}) := \mathbb{E}\left[\ell_{\boldsymbol{bb}}(\boldsymbol{Y}, \boldsymbol{T}, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v})) \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{V} = \boldsymbol{v}\right]$ *is invertible with bounded inverse. (iv) The parameter* $\boldsymbol{\mu}_0$ *of Equation (11) is identified and pathwise differentiable and* $\boldsymbol{H}$ *is thrice continuously differentiable in* $\boldsymbol{b}$*. (v)* $\boldsymbol{H}\left(\boldsymbol{x}, \boldsymbol{b}^*(\boldsymbol{x}, \boldsymbol{v}); \boldsymbol{t}^*\right)$ *and* $\boldsymbol{\ell}_{\boldsymbol{b}}(\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}))$ *possess* $q > 4$ *finite absolute moments and positive variance.*

Differently, both the gradient and Hessian matrix of loss function condition on characteristics $\boldsymbol{X}$ and generated regrssors $\boldsymbol{V}$ simutaneously. This part will bring stronger assumption requirement for introducing control function with generated regressors.

We give a formal introduction of the influence function result as following

Theorem 4. *Suppose assumption 9 holds, we can have the following influence function. Define* $\boldsymbol{H}_{\boldsymbol{b}}\left(\boldsymbol{x}, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}); \boldsymbol{t}^*\right)$ *as the* $d_{\boldsymbol{\mu}} \times d_{\boldsymbol{\theta}}$ *Jacobian of* $\boldsymbol{H}$ *with respect to* $\boldsymbol{b}$*, that is, the matrix with* $\{i, j\}$ *element, for* $i = 1, \ldots, d_{\boldsymbol{\mu}}, j = 1, \ldots, d_{\boldsymbol{b}}$*, given by*

$$\left[\boldsymbol{H}_{\boldsymbol{b}}\left(\boldsymbol{x}, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}); \boldsymbol{t}^*\right)\right]_{i,j} = \left. \frac{\partial H_i\left(\boldsymbol{x}, \boldsymbol{b}; \boldsymbol{t}^*\right)}{\partial b_j} \right|_{b = \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v})}$$

with H_i the i^{th} element of $\boldsymbol{H}$ and b_j the j^{th} element of $\boldsymbol{b}$. Then for $\boldsymbol{\mu}^*$ of Equation (6), a valid and Neyman orthogonal score is $\boldsymbol{\psi}(\boldsymbol{w}, \boldsymbol{\theta}, \boldsymbol{\Lambda}) - \boldsymbol{\mu}_0$, where

$$\boldsymbol{\psi}(\boldsymbol{w}, \boldsymbol{b}^*, \boldsymbol{\Lambda}) = \boldsymbol{H}(\boldsymbol{x}, \boldsymbol{b}^*(\boldsymbol{x}, \boldsymbol{v}^*); \boldsymbol{t}^*) - \boldsymbol{H_b}(\boldsymbol{x}, \boldsymbol{b}^*(\boldsymbol{x}, \boldsymbol{v}^*); \boldsymbol{t}^*)\boldsymbol{\Lambda}(\boldsymbol{x}, \boldsymbol{v}^*)^{-1}\boldsymbol{\ell_b}(\boldsymbol{w}, \boldsymbol{b}^*(\boldsymbol{x}, \boldsymbol{v}^*))) \tag{18}$$

Our influence function is more general and robust with taking care of generated regressors compared to Farrell et al. (2021) while keeping concise and standard function form. Our influence function can handle various applied economics under the existing of endogenity and sample selection context with this form of generality. For example, the typical econometrics models as average treatment effect, average partial effects and partially linear model lie in the coverage of our general influence function. In particular, our setting don't limit loss function to squared loss and don't require nonparametric form in second stage as Hahn and Ridder (2013, 2019); Mammen et al. (2012), which demonstrate the comprehensive usages and yield more new contexts for inference after machine learning through our setting.

Our influence function includes two terms: the leading main term which is the usual formula for influence function of an m-estimator when we truthly know nuisances function and generated regressros; The second term is the bias correction term. This term measure the effect that include estimation of $\boldsymbol{v}$ in first step and the estimation of $\boldsymbol{b}^*(\boldsymbol{x}, \boldsymbol{v}^*)$ in second step. To illustrate this idea, we define

$$\boldsymbol{\mu}^* = \bar{\boldsymbol{H}}(\boldsymbol{x}, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}^*); t^*) = \mathbb{E}[\boldsymbol{H}(\boldsymbol{X}, \boldsymbol{b}(\boldsymbol{X}, \boldsymbol{V}^*); t^*)]$$

Then these two step effects can be expressed by chain rule as in equation (19)when we derive the influence function as following (More details see the appendix).

$$\frac{\partial \bar{\boldsymbol{H}}(\boldsymbol{x}, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}_\eta; \eta, \boldsymbol{v}_\eta); t^*)}{\partial \eta} = \underbrace{\frac{\partial \bar{\boldsymbol{H}}(\boldsymbol{x}, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}_\eta; \eta = 0, \boldsymbol{v}_\eta); t^*)}{\partial \eta}}_{\text{Effect by estimation of } \boldsymbol{v} \text{ in first step}} + \underbrace{\frac{\partial \bar{\boldsymbol{H}}(\boldsymbol{x}, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}^*; \eta, \boldsymbol{v}^*); t^*)}{\partial \eta}}_{\text{Effect by estimation of } \boldsymbol{b} \text{ in second step}} \tag{19}$$

$$= \underbrace{\frac{\partial \bar{\boldsymbol{H}}(\boldsymbol{x}, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}_\eta; \eta = 0, \boldsymbol{v}^*); t^*)}{\partial \eta}}_{\text{Direct effect through evaluating in first step}} + \underbrace{\frac{\partial \bar{\boldsymbol{H}}(\boldsymbol{x}, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}^*; \eta = 0, \boldsymbol{v}_\eta); t^*)}{\partial \eta}}_{\text{Indirect effect through condictioning in first step}} \tag{20}$$

$$+ \underbrace{\frac{\partial \bar{\boldsymbol{H}}(\boldsymbol{x}, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}^*; \eta, \boldsymbol{v}^*); t^*)}{\partial \eta}}_{\text{Effect by estimation of } \boldsymbol{b} \text{ in second step}} \tag{21}$$

The first step impact by estimation of $\boldsymbol{v}$ can be splited into two sources as in equation (20). The direct source come from emerge from the effect of evaluating $\boldsymbol{b}$ through the generated regressors while the indirect source of impact on $\boldsymbol{\mu}^*$ stem from generated regressors $\boldsymbol{v}$ affecting estimation of $\boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v})$ through conditioning as in Hahn and Ridder (2013). The second partial derivative on the right hand side of equation (19) describe the bias impacts arise from the estimation of $\boldsymbol{b}^*(\boldsymbol{x}, \boldsymbol{v}^*)$. This impact is described by assuming the generated regressors are known, $\boldsymbol{v}_\eta = \boldsymbol{v}^*$.

Under the setting in this paper, these biases incurred by estimation of $\boldsymbol{v}$ in first step and $\boldsymbol{b}^*(\boldsymbol{x}, \boldsymbol{v}^*)$, these two effects are merged and simultaneously taken care in the second bias correction term of influence function as in (18). This unexpected but plausible derivation results differ our work from Hahn and Ridder (2013); Hahn et al. (2021); Escanciano and Pérez-Izquierdo (2023) where they separately study the correction bias terms in influence function for first and second step estimation, which usually develop complex analytical forms . One highlight of Escanciano and Pérez-Izquierdo (2023) deserve to be pointed out is their innovative automatic estimation idea to construct estimators of nuisance functions without resorting to plug-in non-parametric estimators of nuisance functions in bias correction terms. That can benefit to avoid cumbersome computations because the analytical form of nuisance functions are typically complex in the existence of generated regressors. This kind of auto-DML ideas(Chernozhukov et al., 2021, 2022; Escanciano and Pérez-Izquierdo, 2023), where the biases correction terms in influence function are estimated from data instead of deriving their analytical forms explicitly, are also implicitly inherited in this paper.[5] With knowing smooth function $\boldsymbol{H}(\boldsymbol{x}, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}^*); t^*)$ and three pieces of bias correction term including $\boldsymbol{H}_{\boldsymbol{b}}(\boldsymbol{x}, \boldsymbol{b}^*(\boldsymbol{x}, \boldsymbol{v}^*); \boldsymbol{t}^*)$, $\boldsymbol{\Lambda}(\boldsymbol{x}, \boldsymbol{v}^*)$, $\boldsymbol{\ell}_{\boldsymbol{b}}(\boldsymbol{w}, \boldsymbol{b}^*(\boldsymbol{x}, \boldsymbol{v}^*)))$, we can simply calculate each piece value through automatic differentiation without full derivation of analytical forms as in Farrell et al. (2021). With this general influence function under the existence of generated regressors, our influence function maintain standard form and still provide general and comprehensive usages through automatic differentiation even if we don't know their derived analytical forms.

Our conditional expectation of hessian matrix $\Lambda(\boldsymbol{x}, \boldsymbol{v}^*)$ contain generated regressors $\boldsymbol{v}^*$ as inputs. Since generated regressors $\boldsymbol{v}^*$ are related to $\boldsymbol{t}$ in loss function, we would always include generated regressors $\boldsymbol{v}^*$ in this function even if randomization case exists. In our method, $\Lambda(\boldsymbol{x}, \boldsymbol{v}^*)$ will degenerate back to $\Lambda(\boldsymbol{v}^*)$ under full randomization scenario. Our methodology

[5]Incorporating automatic locally robust estimation in this paper provide alternative methodology to derive influence function and achieve asymptotic normality. We will leave this work for future.

bring some certain loss of generality of our methodology compared to remark 4 in Farrell et al. (2021) but still endurable for introducing generated regressors to deal with endogeneity or sample selection. Since $\Lambda(\boldsymbol{x}, \boldsymbol{v}^*)$ can be obtained through regression, we project loss function into characteristics covariates $\boldsymbol{X}$ and generated regressors $\boldsymbol{V}$ and then calculate inverse matrix to be plugged into influence function for achieving empirical analog. To ensure well-behaved inverse of $\Lambda(\boldsymbol{x}, \boldsymbol{v}^*)$, our requirement of the "demonimator" is bounded away from zero is stronger than that in Farrell et al. (2021). Apparently, the trimming technique or other regularization is mostly required to ensure stability and valibility of the inverse of $\Lambda(\boldsymbol{X}, \boldsymbol{V})$ in practice. For example, $(\Lambda(\boldsymbol{x}, \boldsymbol{v}) + \alpha I_{d_{\boldsymbol{b}}})^{-1}$ is a typical regularization solution to this issue with small value of α.

5.2 Asymptotic Normality

In this section, we establish the asymptotic normality of the estimators derived within our semiparametric framework, which incorporates generated regressors. We build on the double machine learning (DML) framework introduced by Chernozhukov et al. (2018), which provides a general and robust approach for inference in high-dimensional settings. The core result of asymptotic normality in our context is a direct application and extension of the asymptotic theory presented by Chernozhukov et al. (2018), adapted to accommodate the complexities introduced by generated regressors.

Chernozhukov et al. (2018) developed a comprehensive framework for establishing asymptotic normality in models involving high-dimensional nuisance parameters. Their methodology hinges on two critical components: the Neyman-orthogonal score and cross-fitting.

The Neyman-orthogonal score is designed to ensure robustness against small errors in the estimation of nuisance parameters. Orthogonality reduces the sensitivity of the estimator to these errors, which is crucial for maintaining the validity of inference when dealing with complex models. This orthogonality condition is achieved by constructing the score such that the first-order derivative of the score with respect to the nuisance parameters is zero at the true value of the parameter of interest.

In our study, we extend Chernozhukov et al. (2018) asymptotic normality results to a semiparametric framework that includes generated regressors. The primary challenge in this extension lies in ensuring that the influence function we developed maintains the Neyman-orthogonality property even when generated regressors are present. The introduction of generated regressors necessitates adjustments to account for the biases introduced in the first stage

of the estimation process, where these regressors are constructed.

To address this challenge, our developed influence function explicitly accounts for the impact of generated regressors. By carefully incorporating the estimation error from the first-stage generation of regressors, our influence function adjusts the orthogonal score to preserve its Neyman-orthogonality. This ensures that the orthogonality property holds robustly, even in the presence of generated regressors, thereby enabling valid and reliable inference. The meticulous design of this influence function is critical, as it mitigates the potential biases that could otherwise compromise the asymptotic properties of the estimator. Consequently, we are able to maintain the robustness and accuracy of the Neyman-orthogonal score, ensuring that the final estimator achieves the desired asymptotic normality.

Cross-fitting is a sample-splitting technique that further enhances the estimator's robustness. By dividing the sample into multiple subsets and using different subsets to estimate the nuisance parameters and the target parameter, cross-fitting mitigates the risk of overfitting and improves the reliability of the asymptotic variance estimation. Together, these techniques form the backbone of the asymptotic framework in the DML methodology, resulting in estimators that are asymptotically normal under mild regularity conditions.

In addition, we employ a deep learning-based adversarial training scheme in the first stage. This scheme is designed to estimate both the parameter functions and the control functions associated with the generated regressors. By achieving a fast sup-norm convergence rate—specifically, one that is sufficiently fast for the purposes of asymptotic inference—we ensure that the Neyman-orthogonality of the score is preserved even in the presence of generated regressors. This enables us to apply the asymptotic normality results derived by Chernozhukov et al. (2018) directly to our context.

We here introduce the whole estimation procedure of DML in our setting as in Chernozhukov et al. (2018); Farrell et al. (2021). Let's assume we have n data observations and take a K-fold random partition $I_k \subset \{1, 2, \cdots, N\}$ for $k = 1, 2, \cdots, K$ with size of each fold I_k equal to $n = \lfloor N/K \rfloor$. We define complement set of I_k as $I_k^c := \{1, ..., N\} \backslash I_k$ for each k. Our estimation procedure include three stages: we use I_k^c to estimate generated regressors $\boldsymbol{V}$, preposeeced $\hat{Y}$ and parameter functions $\boldsymbol{\theta}^*(\boldsymbol{X})$ with control function $m^*(\boldsymbol{V})$. Further division of I_k^c into two halfs. One of halfs is used to estimate $\boldsymbol{V}$ and the other one is used to estimate preprocessed $\hat{Y}$, hessian matrix function $\boldsymbol{\Lambda}(\boldsymbol{X}, \boldsymbol{V})$ and parameter functions $\boldsymbol{\theta}^*(\boldsymbol{X})$ with control function $m^*(\boldsymbol{V})$. If necessary, I_k^c need to be splitted into three folds including the first to estimate $\boldsymbol{V}$, the second to estimate $\hat{Y}$ and the third to estimate hessian matrix

function $\boldsymbol{\Lambda}(\boldsymbol{X}, \boldsymbol{V})$ and parameter functions $\boldsymbol{\theta}^*(\boldsymbol{X})$ with control function $m^*(\boldsymbol{V})$. Therefore, we calculate the empirical analog of influence function to obtain the estimator of inferential parameters $\boldsymbol{\mu}^*$ as following

$$\hat{\boldsymbol{\mu}}_k = \frac{1}{|I_k|} \sum_{i \in I_k} \boldsymbol{\psi}(\boldsymbol{w}_i, \hat{\boldsymbol{b}}_k(\boldsymbol{x}_i, \hat{\boldsymbol{v}}_i), \hat{\Lambda}_k(\boldsymbol{x}_i, \hat{\boldsymbol{v}}_i))$$

and aggregate the estimators to get the final estimator

$$\hat{\boldsymbol{\mu}} = \frac{1}{K} \sum_{k=1}^{K} \hat{\mu}_k$$

where $|I_k|$ denotes the data size of fold I_k. Besides estimator $\hat{\boldsymbol{\mu}}$, we need to estimate empirical variance of the asymptotic variance of $\boldsymbol{\Psi} = \mathbb{V}[\boldsymbol{\psi}(\boldsymbol{W}, \boldsymbol{\theta}(\boldsymbol{X}, \boldsymbol{V}), \boldsymbol{\Lambda}(\boldsymbol{X}, \boldsymbol{V}))]$ and its empirical analog is

$$\hat{\boldsymbol{\Psi}}_k = \frac{1}{|I_k|} \sum_{i \in |I_k|} \left(\boldsymbol{\psi}(\boldsymbol{w}_i, \boldsymbol{b}_k(\boldsymbol{x}_i, \hat{\boldsymbol{v}}_i), \boldsymbol{\Lambda}_k(\boldsymbol{x}_i, \boldsymbol{v}_i) - \hat{\boldsymbol{\mu}})\right)^2$$

and aggregate them as the estimator of asymptotic variance

$$\hat{\boldsymbol{\Psi}} = \sum_{k=1}^{K} \hat{\boldsymbol{\Psi}}_k \tag{22}$$

by the application of theorem 3.2 in Chernozhukov et al. (2018).

The key requirement of asymptotic normality is fast enough convergent rate of $\hat{\boldsymbol{b}}(\boldsymbol{x}, \hat{\boldsymbol{v}})$ and $\hat{\boldsymbol{\Lambda}}(\boldsymbol{x}, \hat{\boldsymbol{v}})$ guaranteed by $o(N^{-1/4})$ with application of our estimation scheme of structured deep neural network by adversarial training. Other machine learning estimators or traditional nonparametrics may also satisfy rate requirement by sup-norm convergent property in the existence of generated regressors.

We formally introduce our asymptotic theory here which is a direct result from Chernozhukov et al. (2018).

Theorem 5. *Assume a random data set $\boldsymbol{w}_i$ follows assumption 9 and convergent rate requirements for all subsamples $k = 1, \cdots, K$ as following $||\hat{\theta}_i - \theta_i^*||_{L^\infty} = o(n^{-1/4})$, $||\hat{m} - m^*||_{L^\infty} = o(n^{-1/4})$ and $\left\|[\widehat{\boldsymbol{\Lambda}}]_{i,j} - [\boldsymbol{\Lambda}]_{i,j}\right\|_{L^\infty} = o\left(n^{-1/4}\right)$ for all $i, j \in \{1, \ldots, d_{\boldsymbol{\theta}}\}$ with uniformly invertible $\hat{\Lambda}_k(\boldsymbol{x}, \hat{\boldsymbol{v}})$ for each k in $1, \cdots, K$. Hence, the asymptotic normality is following*

$$\sqrt{n}\widehat{\Psi}^{-1/2}\left(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{\Psi}^{-1/2} \psi\left(\boldsymbol{w}_i, \boldsymbol{\theta}_0\left(\boldsymbol{x}_i\right), \boldsymbol{\Lambda}\left(\boldsymbol{x}_i\right)\right) + o_p(1) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}_{d_\mu}, \boldsymbol{I}_{d_\mu}\right) \tag{23}$$

Our framework with influence function accounting for generated regressors allow us to test various inferential structure parameters in the context of endogeneity and sample selection even when the closed form of inferential parameters is not available. In addition, our framework can deal with not only typical squared loss function but handle nonlinear models where classification-based loss or quantile regression may be covered. It allows us to test widespread different economics models while studying individual heterogeneity according to individual characteristics.

6 Estimation Procedure

In this section, we illustrate a deep neural network architecture to estimate structural parameter functions $\boldsymbol{\theta}^*(\boldsymbol{X})$ and control function $m^*(\boldsymbol{V})$ with generated regressors $\hat{\boldsymbol{V}}$ by adversarial training. Our estimation procedure follows a robust three-stage approach, integrating advanced deep learning architectures and adversarial training to ensure robustness and accuracy in the estimation of structural parameters.

We consider triangular system in the context of endogeneity with a typical linear parametric setting as a simple exhibition for estimation proccesure of our framework, where the parameters are reforged into non-linear functions of individual heterogeneity variables, $\boldsymbol{\theta}(\boldsymbol{x}) = (\theta_0(\boldsymbol{x}), \boldsymbol{\theta_1}(\boldsymbol{x})')$. Then the model specification can be straightforwardly given by structural form equation

$$y = \theta_0(\boldsymbol{x}) + \boldsymbol{\theta}_1(\boldsymbol{x})'\boldsymbol{t} + u, \tag{24}$$

and reduced form equation

$$\boldsymbol{t} = \boldsymbol{g}(\boldsymbol{z}) + \boldsymbol{v}, \tag{25}$$

where $\boldsymbol{g(z)} = (g_1(\boldsymbol{z}), ..., g_{d_t}(\boldsymbol{z}))'$ and $\boldsymbol{v} = (v_{i1}, v_{i2}, ..., v_{id_t})'$.

In the following, we will describe the control function method for the individual heterogeneity model with endogeneity, estimated by deep learning model. In the method of control function, we assume

$$\mathbb{E}(u|\boldsymbol{t}, \boldsymbol{x}, \boldsymbol{v}) = \mathbb{E}(u|\boldsymbol{v}) = m(\boldsymbol{v}),$$

$$\mathbb{E}(\boldsymbol{v}|\boldsymbol{z}) = \boldsymbol{0}.$$

Then we can rearrange the model to incorporate control function with generated regressors into 24 as below

$$\begin{aligned} y_i &= \theta_0(\boldsymbol{x}_i) + \boldsymbol{\theta}_1(\boldsymbol{x}_i)'\boldsymbol{t}_i + u_i \\ &= \theta_0(\boldsymbol{x}_i) + \boldsymbol{\theta}_1(\boldsymbol{x}_i)'\boldsymbol{t}_i + \mathbb{E}(u_i|\boldsymbol{v}_i) + u_i - \mathbb{E}(u_i|\boldsymbol{v}_i) \\ &= \theta_0(\boldsymbol{x}_i) + m(\boldsymbol{v}_i) + \boldsymbol{\theta}_1(\boldsymbol{x}_i)'\boldsymbol{t}_i + u^* \end{aligned} \tag{26}$$

where $u_i^* = u_i - \mathbb{E}(u_i|\boldsymbol{v}_i)$.

6.1 First-Stage Estimation

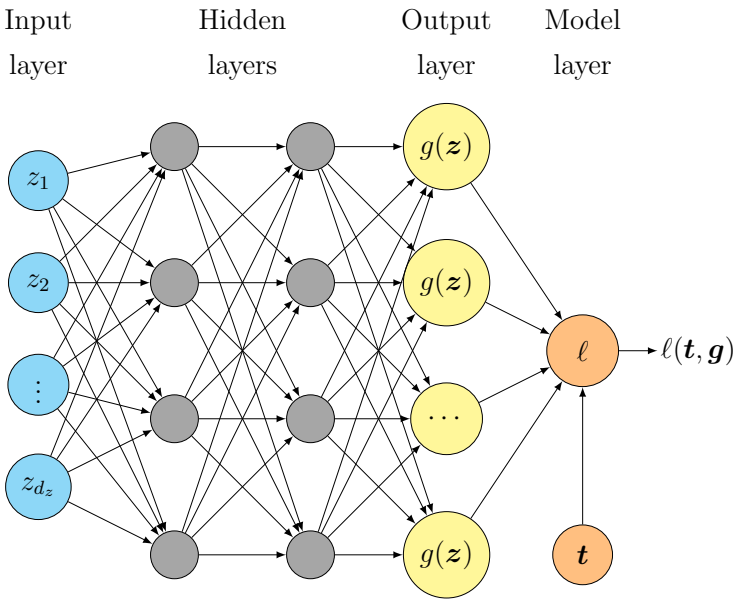


Figure 4: Unstructured deep neural network to estimate $\hat{g}_{DNN}(\cdot)$

In order to estimate our structural parameter functions $\boldsymbol{\theta}(\boldsymbol{X})$ and control function $m(\boldsymbol{V})$ in equation (26), we need to estimate $\boldsymbol{V}$ in our first stage. In the first stage estimation, the functional form $\boldsymbol{g}$ in reduced form equation can be presented in parametric, semiparametric and nonparametric ways with great generality. To deal with possible misspecification problems, we focus on nonparametric setting in reduced form equation. Not only traditional nonparametric methods are allowed as kernel regression, local polynomial regression or sieve regression in estimating generated regressors but also various machine learning methods including tree-based models, deep neural network moedel, or ensemble models among different machine learning

methods can be employed. Here we employ a unstructured deep neural networks to estimate $\boldsymbol{g}(\boldsymbol{z})$. That is

$$\hat{\boldsymbol{g}}_{\text{DNN}}(\cdot) = \underset{\boldsymbol{g}\in\mathcal{F}_{\text{DNN}}}{\operatorname{argmin}} \sum_{i=1}^{n}\sum_{j=1}^{d_{\boldsymbol{t}}} \ell(t_{ij}, g_j(\boldsymbol{z}_i)), \tag{27}$$

where $\mathcal{F}_{DNN}$ denotes the functional class of ReLU-DNN and t_{ij} represent i-th data observation for jth element of endogenous variables $\boldsymbol{t}$ and loss function ℓ will adopt squared loss function. With $\hat{\boldsymbol{g}}_{\text{DNN}}(\boldsymbol{z})$ functions estimator by deep neural network, we can derive the estimators of generated regressors as following

$$\hat{\boldsymbol{v}}_i = \boldsymbol{t}_i - \hat{\boldsymbol{g}}_{DNN}(\boldsymbol{z}_i).$$

6.2 Second-stage estimation

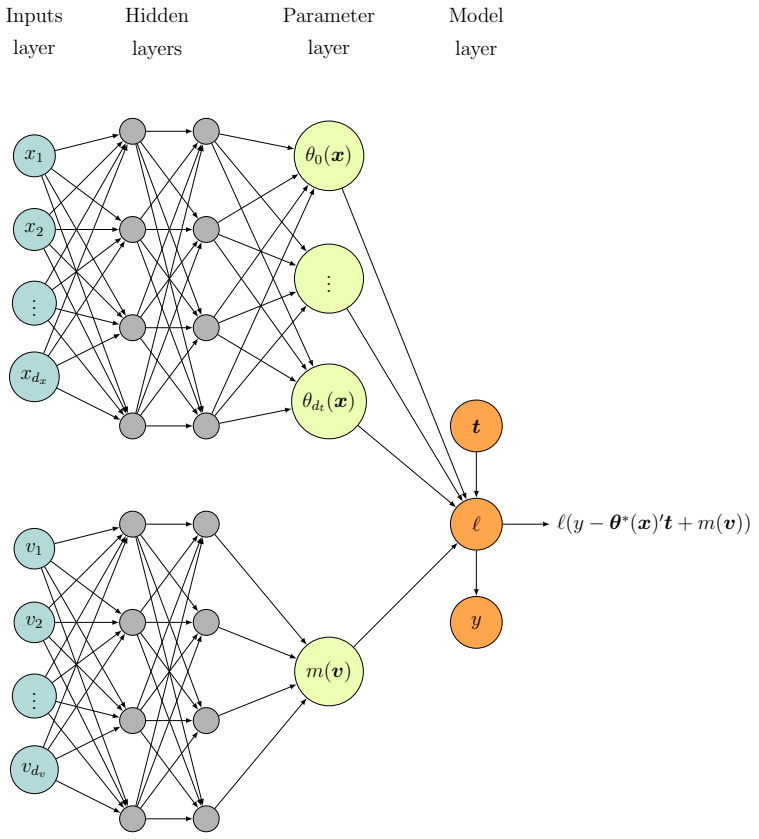


Figure 5: Illustration of the structured deep neural network estimation of the parameter functions $\boldsymbol{\theta}(\boldsymbol{x})$ and control function m($\boldsymbol{v}$)

Next, we construct the structured deep neural network to estimate parameters function $\boldsymbol{\theta}(\boldsymbol{X})$ and $m^*(\boldsymbol{V})$ in equation (26) with adversarial training. The architecture of structured deep neural network comprise two modules as in (5). The upper module of structured deep neural network aims at estimating parameters functions $\boldsymbol{\theta}(\boldsymbol{x}) = (\theta_0(\boldsymbol{x}), \boldsymbol{\theta}_1(\boldsymbol{x})')'$ with individual characteristics variables $\boldsymbol{x}$ where the architecture is configured by width H_h and depth L_h while the lower module of deep neural networks focuses on estimating control function $m(\hat{\boldsymbol{v}})$ with inputs $\hat{\boldsymbol{v}}$ from first-stage estimation where the architecture is similarly determined by width H_c and depth L_c. With the outcomes of parameter functions and control function in parameter layer together with y, $\boldsymbol{t}$, the structural deep neural network is encapsulated and completed by a per-observation loss function $\ell(y, \boldsymbol{t}, \boldsymbol{\theta}(\boldsymbol{x}), m(\boldsymbol{v}))$, which is squared loss function for the example of triangular system.

Regarding estimation, we apply the advasarial training scheme to obtain estimation of parameter functions and control function in second stage of estimation. Adversarial training is a robust optimization technique that involves solving a min-max optimization problem, where the goal is to minimize the worst-case loss across all possible adversarial perturbations of the input data. This approach enhances the model's robustness against small, deliberate perturbations that could otherwise lead to significant estimation errors. The optimization problem can be formalized as follows:

In the context of triangular system, we adopt a quadratic loss function as our empirical risk for illustration. With surrogate outputs $\hat{Y}$, the empirical preprocessed adversarial risk is defined as

$$R_n(\boldsymbol{b}) := \frac{1}{n}\sum_{i=1}^{n} \max_{(\boldsymbol{x},\boldsymbol{v})\in\Delta_h^p\left(X_i,\widehat{V}_i\right)} \left(\widehat{Y}\left(\boldsymbol{x}, \boldsymbol{v}, \boldsymbol{t}_i\right) - m(\boldsymbol{v}) - \boldsymbol{t}_i'\boldsymbol{\theta}(\boldsymbol{x})\right)^2 \tag{28}$$

for $m(\mathbf{v}), \{\theta_k(\boldsymbol{x})\}_{k=1}^{d_\theta} \in L^2\left([0,1]^d\right)$. This loss function is an extended form of the standard adversarial risk with preprocessing $\widehat{Y}$. It is an robustified empirical risk by an inner maximization which is the first step in adversarial training. By implementing this particular concept, we obtain our goal estimators as the element that works to minimize the empirical risk as

$$\widehat{\boldsymbol{b}} \in \underset{\boldsymbol{b}\in\mathcal{F}(L_h,H_h,L_c,H_c)}{\operatorname{argmin}} R_n(\boldsymbol{b}) \tag{29}$$

where the set of deep neural networks, $\mathcal{F}_{L_h,H_h,L_c,H_c}$ as shown in Figure 5, composed of upper and lower modules with an upper bound B.

The optimization process consists of two key steps: generating adversarial examples through inner maximization and refining model parameters via outer minimization to ensure robustness against these adversarial examples.

In the inner maximization step, given a set of input data $(\mathbf{X}_i, \mathbf{V}_i)$, the goal is to identify the worst-case adversarial data(or perturbations) that maximize the loss function as in (28). This typically involves using projected gradient descent (PGD), where small, calculated perturbations are iteratively applied to the input data to increase the loss (Madry, 2017).

Once the adversarial examples are generated, the outer minimization step adjusts the model parameters θ and m to minimize the loss across these worst-case scenarios as in (29). This is done using stochastic gradient descent (SGD) or similar optimization algorithms, which iteratively refine the model parameters based on the gradients computed from the adversarially perturbed inputs.

In this context, adversarial training adjusts both $\theta(\mathbf{X})$ and $m(\mathbf{V})$ to ensure that the estimated structural parameters are robust against adversarial perturbations in the covariates and generated regressors. The goal is to ensure that small changes in the inputs $\mathbf{X}$ and $\mathbf{V}$, which may result from unobserved confounders or measurement errors, do not lead to significant biases in the estimated parameters. In certain extend, it also provide robustness of estimated parameter functions and control function against sample variation of estimating generated regression in the first stage.

With the optimization procedure, the parameter functions and control function can simultaneously estimated and readily used for the next semiparametric inference.

6.3 Third-Stage Estimation for Inferential Parameters

In this stage, we employ the double machine learning (DML) framework to estimate the inferential parameters $\boldsymbol{\mu}^*$ and construct their corresponding confidence intervals. This procedure builds on the parameter functions $\hat{\boldsymbol{\theta}}(\boldsymbol{X})$ and control function $\hat{m}(\widehat{\boldsymbol{V}})$ estimated in the previous stages, utilizing cross-fitting to ensure robustness and valid inference.

We assume a dataset with N observations, which is divided into K folds. Let $I_k \subset \{1, 2, \ldots, N\}$ denote the indices of the observations in the kth fold, with $|I_k| = n = \lfloor N/K \rfloor$, and let $I_k^c = \{1, \ldots, N\} \setminus I_k$ denote the complement of I_k. The estimation procedure proceeds in three stages:

Firstly, for each fold k, the complement set I_k^c is used to estimate the generated regressors $\boldsymbol{V}$, the preprocessed $\hat{Y}$, and the parameter functions $\hat{\boldsymbol{\theta}}(\boldsymbol{X})$ along with the control function $\hat{m}(\boldsymbol{V})$. This process may involve further splitting I_k^c into sub-folds to sequentially estimate these components. Specifically, one half of I_k^c is used to estimate $\boldsymbol{V}$, while the other half is used to estimate $\hat{Y}$, the Hessian matrix function $\boldsymbol{\Lambda}(\boldsymbol{X}, \widehat{\boldsymbol{V}})$, and the parameter functions $\hat{\boldsymbol{\theta}}(\boldsymbol{X})$

and $\hat{m}(\widehat{\boldsymbol{V}})$.[6].

Using the estimates from the first and second stages, we calculate the empirical analog of the influence function to obtain the estimator of the inferential parameters $\boldsymbol{\mu}^*$:

$$\hat{\boldsymbol{\mu}} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{|I_k|}\sum_{i\in I_k}\boldsymbol{\psi}\left(\boldsymbol{w}_i, \hat{\boldsymbol{b}}_k(\boldsymbol{x}_i, \hat{\boldsymbol{v}}_i), \hat{\boldsymbol{\Lambda}}_k(\boldsymbol{x}_i, \hat{\boldsymbol{v}}_i)\right)$$

where $\boldsymbol{\psi}$ is the influence function, $\hat{\boldsymbol{b}}_k$ denotes the estimates of structural parameters for fold k, and $\hat{\boldsymbol{\Lambda}}_k$ is the estimated Hessian matrix function and $|I_k|$ is the size of the kth fold.

The empirical variance of the influence function is used to estimate the asymptotic variance $\boldsymbol{\Psi}$ of $\boldsymbol{\mu}^*$. The overall estimator for the asymptotic variance is calculated as:

$$\hat{\boldsymbol{\Psi}} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{|I_k|}\sum_{i\in I_k}\left(\boldsymbol{\psi}\left(\boldsymbol{w}_i, \hat{\boldsymbol{b}}_k(\boldsymbol{x}_i, \hat{\boldsymbol{v}}_i), \hat{\boldsymbol{\Lambda}}_k(\boldsymbol{x}_i, \hat{\boldsymbol{v}}_i)\right) - \hat{\boldsymbol{\mu}}\right)^2$$

Finally, the standard error $\widehat{\mathrm{SE}}(\hat{\boldsymbol{\mu}})$ is derived from $\hat{\boldsymbol{\Psi}}$, and the confidence intervals for $\boldsymbol{\mu}^*$ are constructed as:

$$\hat{\boldsymbol{\mu}} \pm z_{\alpha/2}\cdot\widehat{\mathrm{SE}}(\hat{\boldsymbol{\mu}})$$

where $z_{\alpha/2}$ is the critical value from the standard normal distribution.

This comprehensive approach ensures that the final estimator $\hat{\boldsymbol{\mu}}$ is both robust to errors in the estimation of nuisance parameters and consistent, providing reliable inference in high-dimensional settings.

7 Simulations

7.1 Data Generating Process

To evaluate the performance of our proposed method, we conduct simulations using two distinct data-generating processes (DGPs). These DGPs simulate econometric challenges of endogeneity problems, providing a robust framework for testing the estimator's capabilities under different structural conditions. The key difference between the two DGPs lies in the correlation structure. While the first DGP assumes that the treatment variable is correlated with the error term but not directly with the characteristics, the second DGP incorporates correlations between the treatment variable and the individual characteristics. This distinction is crucial

[6]More divisions may be required if necessary

for evaluating the robustness of our estimator in handling endogeneity, particularly in scenarios where the treatment and characteristics are not independent.

In both DGPs, the parameter functions $\theta_1(X_i)$ and $\theta_2(X_i)$ are modeled as complex, non-linear functions of the individual characteristics, ensuring that the heterogeneity of the population is appropriately captured. In addition, the introduction of correlated characteristics in DGPs is expected to increase the challenge of estimation, providing a more rigorous test of our estimator's performance.

7.1.1 First Data Generating Process

The first DGP models the outcome variable Y_i as a function of individual characteristics X_i, an endogenous treatment variable T_i, and an error term U_i:

$$Y_i = \theta_1(X_i) + \theta_2(X_i) \cdot T_i + U_i, \quad i = 1, \ldots, N$$

where Y_i represents the dependent variable, while T_i is the endogenous treatment variable, whose causal effect on Y_i is the primary focus of our estimation. The individual characteristics X_i influence the parameter functions $\theta_1(X_i)$ and $\theta_2(X_i)$, both of which vary with X_i, reflecting the heterogeneity present within the population. The error term U_i captures unobserved factors that influence Y_i.

Endogeneity is introduced through the correlation between T_i and U_i, as the errors U_i and V_i are jointly normally distributed:

$$\begin{pmatrix} U_i \\ V_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_U^2 & \sigma_{UV} \\ \sigma_{UV} & \sigma_V^2 \end{pmatrix} \right)$$

where $\sigma_U^2 = 1.0$, $\sigma_V^2 = 1.0$, and $\sigma_{UV} = 0.9$, ensuring that V_i and T_i are correlated, introducing endogeneity.

The individual characteristics $X_i = (X_{i1}, X_{i2}, \ldots, X_{i20})$ are generated from a Beta distribution:

$$X_{ij} \sim \text{Beta}(\alpha = 2, \beta = 5), \quad j = 1, \ldots, 20$$

These variables are standardized to have zero mean and unit variance, with correlations introduced via an exponential decay structure in the correlation matrix Σ_X:

$$\Sigma_X[j, k] = \rho^{|j-k|}, \quad \text{where } \rho = 0.5$$

Next, we generate the instrumental variables Z_i from a uniform distribution:

$$Z_{il} \sim \mathcal{U}(0,1), \quad l = 1, \ldots, 3$$

The endogenous treatment variable T_i is generated as a nonlinear function of the instrumental variables Z_i and the error term V_i, which is transformed using the logistic sigmoid function:

$$T_i = \frac{1}{2}\left(\sin(\pi Z_{i1} Z_{i3}) \sin(\pi Z_{i2}^2) + \cos(2\pi e^{Z_{i3}} Z_{i1} Z_{i2})^2 + \frac{1}{1+e^{-V_i}}\right)$$

This process models the treatment variable T_i as dependent on both the instrumental variables and the error V_i, creating a source of endogeneity.

7.1.2 Second Data Generating Process

The second DGP introduces a more complex structure by incorporating correlations between the endogenous treatment variable T_i and the individual characteristics X_i, reflecting a more realistic economic setting. Here, individual characteristics X_i are generated as in the first DGP but combined with correlated logistic-transformed errors that affect both the characteristics and the treatment variable.

First, we generate errors for both the characteristics and treatment using a multivariate normal distribution:

$$\begin{pmatrix} \text{Errors for } X_i \\ V_{\text{err},i} \end{pmatrix} \sim \mathcal{N}(0, \Sigma_{X,V})$$

where the covariance matrix $\Sigma_{X,V}$ governs the correlation between the individual characteristics and the treatment errors:

$$\Sigma_{X,V}[i,j] = \rho \cdot \frac{1}{1+|i-j|}, \quad \text{where } \rho = 0.5$$

The last row and column of the covariance matrix correspond to the treatment error $V_{\text{err},i}$, which is correlated with the individual characteristics X_i.

Next, we generate individual characteristics by adding logistic-transformed errors:

$$X_{ij} = \frac{1}{2}\left(\text{Beta}(2,2) + \text{Logistic-Transformed Errors for } X_i\right)$$

This ensures the characteristics lie within a reasonable range while being influenced by the correlated errors.

The endogenous treatment variable T_i is generated as:

$$T_i = \frac{1}{3}\left(f_Z(Z_i) + V_i + V_{\text{err},i}\right)$$

where $f_Z(Z_i)$ is a complex non-linear function of the instrumental variables $Z_i = (Z_{i1}, Z_{i2}, Z_{i3})$:

$$f_Z(Z_i) = \frac{1}{2}\sin(\pi Z_{i1} Z_{i3})\sin(\pi Z_{i2}^2) + \frac{1}{2}\cos(2\pi e^{Z_{i3}} Z_{i1} Z_{i2})^2$$

This generation process reflects a more challenging econometric setting, where the characteristics and the endogenous treatment variable are correlated through shared errors.

In both DGPs, the parameter functions $\theta_1(X_i)$ and $\theta_2(X_i)$ are modeled as complex non-linear functions of the individual characteristics:

$$\begin{aligned}\theta_1(X_i) = &\sin(2\pi X_{i1}) + \cos(2\pi X_{i2}) + X_{i3}^2 + X_{i4}X_{i5} + X_{i6}^3 - \sin(2\pi X_{i7}) + \cos(2\pi X_{i8}) + X_{i9}X_{i10} - X_{i11}^2 \\ &+ X_{i12}X_{i13} + X_{i14}^3 - \sin(2\pi X_{i15}) + \cos(2\pi X_{i16}) + X_{i17}X_{i18} + X_{i19}^2 + X_{i20}^3\end{aligned}$$

$$\begin{aligned}\theta_2(X_i) = &\cos(2\pi X_{i1}X_{i2}) + X_{i3}^2 + X_{i4}X_{i5} + \sin(X_{i6}X_{i7}) - \cos(X_{i8} + X_{i9}) + X_{i10}^2 + X_{i11}X_{i12} \\ &+ \sin(2\pi X_{i13}X_{i14}) - X_{i15}^3 + X_{i16}X_{i17} + \cos(X_{i18}X_{i19}) + X_{i20}^2 + X_{i1}X_{i2} + X_{i3}^3 - \sin(2\pi X_{i4}X_{i5})\end{aligned}$$

These functions introduce nonlinearity and interactions among the individual characteristics, reflecting realistic economic relationships.

In both DGPs, the dependent variable Y_i is generated as:

$$Y_i = \theta_1(X_i) + \theta_2(X_i) \cdot T_i + U_i$$

where U_i is the error term for the outcome.

By comparing the results of these two DGPs, we assess the robustness and flexibility of our proposed estimator under different levels of endogeneity and structural complexity.

7.2 Simulation Results

In our simulations, we assess the performance of four estimators: (i) our proposed estimator (DLIHWG); (ii) the partially linear IV regression model (PLMIV) from Chernozhukov et al. (2018); (iii) the estimator proposed by Farrell et al. (2021) (DLIH); and (iv) the two-stage least squares (2SLS) estimator. The target parameter for estimation is $\mu = \mathbb{E}[\theta_2(X)]$. We conduct 1000 simulation iterations, varying the correlation between the error terms, ρ_{UV}, at levels 0, 0.2, 0.5, and 0.9.

For estimation, we utilize fully connected neural networks with ReLU activation functions to estimate the generated regressors and Hessian matrix functions. Structured fully connected neural networks are employed for estimating the parameter functions and control functions, as

shown in Figure (5). The depth and width of these architectures are optimized via grid search, with depths (number of layers) ranging from 2 to 4 and widths (number of neurons) between 30 and 150. For adversarial training, we follow the projected gradient descent algorithm (Madry, 2017), using a neighborhood distance $h = 0.02$. In the preprocessing step, we apply the K-nearest neighbor method, selecting the tuning parameter K via cross-validation, with values ranging from 2 to 100.

Table 1: Simulation results comparison with $\rho_{UV} = 0.9$ for DGP1

Sample Size	Mean Absolute Bias				Standard Deviation				Coverage Rate			
	DLIHWG	**DLIH**	**PLMIV**	**2SLS**	**DLIHWG**	**DLIH**	**PLMIV**	**2SLS**	**DLIHWG**	**DLIH**	**PLMIV**	**2SLS**
5000	1.687	4.274	1.907	9.129	5.929	0.353	2.430	0.219	75.40%	0.00%	96.40%	77.80%
10000	0.931	4.261	1.263	9.132	2.926	0.355	1.595	0.162	92.20%	0.00%	97.80%	65.00%
20000	0.409	4.124	0.826	9.126	1.563	0.214	1.068	0.107	94.20%	0.00%	97.80%	39.20%
50000	0.405	4.224	0.532	9.134	1.380	0.291	0.656	0.067	92.40%	0.00%	99.40%	8.00%
100000	0.338	4.077	0.357	9.134	1.367	0.311	0.455	0.049	95.50%	0.00%	99.40%	0.40%

The simulation results in Table 1 demonstrate that our proposed estimator (DLIHWG) effectively infers the true parameter (5.63), achieving the appropriate coverage rate at the 5% significance level as the sample size increases. In contrast, the estimator proposed by Farrell et al. (2021) (DLIH) exhibits significant absolute bias, which increases with the sample size. This bias arises primarily due to the exclusion of generated regressors in their framework. As a result, the coverage rate for DLIH remains at 0%, reflecting the persistent non-zero bias. These findings highlight the robustness and generality of our framework in handling generated regressors and addressing endogeneity and sample selection bias.

The 2SLS estimator performs poorly across the board, with increasing bias as the sample size grows and a rapid decline in coverage rate to nearly 0%. This decline is mainly due to the bias induced by the linearity assumption in both stages, which fails to capture individual heterogeneity. Finally, our proposed method demonstrates comparable performance to the partially linear IV regression model (PLMIV) from Chernozhukov et al. (2018) in terms of mean absolute bias and standard deviation. However, the coverage rate of our method is more stable, consistently hovering around 95%, whereas other models show amplified rates. Overall, our proposed estimator exhibits robustness and consistently infers the true parameter at the 5% significance level.

Table 2: Simulation results comparison under different endogenity levels for DGP1

Panel (a) : $\rho_{UV} = 0.9$												
Sample Size	Mean Absolute Bias				Standard Deviation				Coverage Rate			
	DLIHWG	DLIH	PLMIV	2SLS	DLIHWG	DLIH	PLMIV	2SLS	DLIHWG	DLIH	PLMIV	2SLS
5000	1.978	4.428	0.740	6.830	3.928	1.588	0.944	1.076	93.30%	5.70%	96.90%	0.00%
10000	1.517	4.389	0.534	6.813	3.790	1.829	0.669	0.780	91.70%	5.00%	97.70%	0.00%
Panel (b) : $\rho_{UV} = 0.5$												
Sample Size	Mean Absolute Bias				Standard Deviation				Coverage Rate			
	DLIHWG	DLIH	PLMIV	2SLS	DLIHWG	DLIH	PLMIV	2SLS	DLIHWG	DLIH	PLMIV	2SLS
5000	2.045	2.463	0.738	6.838	4.550	0.446	0.943	1.013	92.90%	3.10%	96.80%	0.00%
10000	1.259	2.417	0.533	6.816	3.779	0.299	0.668	0.737	91.70%	3.70%	97.60%	0.00%
Panel (c) : $\rho_{UV} = 0.2$												
Sample Size	Mean Absolute Bias				Standard Deviation				Coverage Rate			
	DLIHWG	DLIH	PLMIV	2SLS	DLIHWG	DLIH	PLMIV	2SLS	DLIHWG	DLIH	PLMIV	2SLS
5000	2.090	1.038	0.736	6.844	5.260	1.432	0.942	0.965	98.20%	4.80%	96.30%	0.00%
10000	1.359	0.936	0.532	6.819	4.859	0.329	0.667	0.702	91.70%	1.20%	97.20%	0.00%
Panel (d) : $\rho_{UV} = 0$												
Sample Size	Mean Absolute Bias				Standard Deviation				Coverage Rate			
	DLIHWG	DLIH	PLMIV	2SLS	DLIHWG	DLIH	PLMIV	2SLS	DLIHWG	DLIH	PLMIV	2SLS
5000	2.350	0.354	0.735	6.848	6.320	3.578	0.940	0.932	96.70%	96.90%	96.40%	0.00%
10000	1.003	0.174	0.532	6.821	3.230	0.935	0.667	0.678	91.60%	95.70%	97.20%	0.00%

We also conduct simulations under varying levels of endogeneity to test the robustness of our proposed methodology. In Table 2, we observe that the mean absolute bias of DLIHWG is lower than that of the other models at most endogeneity levels, except when the endogeneity level is zero. At $\rho_{UV} = 0$, the DLIH model restores its accuracy in estimating causal parameters as benchmark model, with a correct coverage rate under the 5% significance level. Conversely, DLIH consistently fails to infer the true parameters across various endogeneity levels, as its coverage rates remain close to 0%. This is primarily due to the endogeneity issue, which DLIH cannot adequately address. As the endogeneity level decreases, the mean absolute bias of DLIH also decreases, returning to unbiased estimation when the endogeneity level reaches zero.

Since our data-generating process generalizes the partially linear model, the PLMIV estimator serves as a benchmark. Our proposed estimator consistently outperforms PLMIV across various levels of endogeneity, exhibiting lower mean absolute bias and slightly higher standard deviation. When the endogeneity level is small, our estimator dominates PLMIV in terms of both bias and standard deviation.

The poor performance of the 2SLS estimator across different endogeneity levels aligns with previous analyses, as it cannot account for nonlinearities in either stage of the estimation, resulting in its failure to capture individual heterogeneity.

These results are similarly robust in the simulations for DGP2, as demonstrated in Table 3, further supporting the robustness and consistency of our proposed estimator across more complex data-generating processes.

After conducting these simulations, we confirm that our estimator achieves consistency and asymptotic normality, even in the presence of generated regressors and endogeneity. The integration of adversarial training significantly enhances the estimator's robustness, making it suitable for a broad range of econometric applications.

Table 3: Simulation results comparison under different endogeneity levels for DGP2

	Panel (a) : $\rho_{UV} = 0.9$											
Sample Size	Mean Absolute Bias				Standard Deviation				Coverage Rate			
	DLIHWG	DLIH	PLMIV	2SLS	DLIHWG	DLIH	PLMIV	2SLS	DLIHWG	DLIH	PLMIV	2SLS
5000	1.614	4.153	1.847	9.132	6.103	0.338	2.364	0.220	92.70%	0.00%	97.40%	0.00%
10000	1.040	4.137	1.224	9.131	4.683	0.281	1.554	0.161	91.70%	0.00%	98.20%	0.00%
	Panel (b) : $\rho_{UV} = 0.5$											
Sample Size	Mean Absolute Bias				Standard Deviation				Coverage Rate			
	DLIHWG	DLIH	PLMIV	2SLS	DLIHWG	DLIH	PLMIV	2SLS	DLIHWG	DLIH	PLMIV	2SLS
5000	0.969	2.334	1.816	9.132	4.301	0.341	2.324	0.222	97.20%	0.00%	97.73%	0.00%
10000	0.834	2.267	1.195	9.131	3.548	0.265	1.537	0.163	95.90%	0.00%	98.00%	0.00%
	Panel (c) : $\rho_{UV} = 0.2$											
Sample Size	Mean Absolute Bias				Standard Deviation				Coverage Rate			
	DLIHWG	DLIH	PLMIV	2SLS	DLIHWG	DLIH	PLMIV	2SLS	DLIHWG	DLIH	PLMIV	2SLS
5000	0.515	0.940	1.842	9.133	2.995	0.360	2.346	0.223	93.40%	4.40%	97.40%	0.00%
10000	0.424	0.850	1.227	9.131	1.486	0.256	1.564	0.164	91.70%	2.30%	98.10%	0.00%
	Panel (d) : $\rho_{UV} = 0$											
Sample Size	Mean Absolute Bias				Standard Deviation				Coverage Rate			
	DLIHWG	DLIH	PLMIV	2SLS	DLIHWG	DLIH	PLMIV	2SLS	DLIHWG	DLIH	PLMIV	2SLS
5000	0.426	0.179	1.839	9.133	1.841	0.265	2.337	0.225	99.10%	93.80%	97.40%	0.00%
10000	0.240	0.107	1.232	9.131	1.835	0.165	1.232	0.165	95.70%	96.60%	98.00%	0.00%

8 Conclusion

This paper introduces significant methodological advancements that integrates machine learning, including the development of a structural deep learning architecture with adversarial

training, which ensures sup-norm convergence of estimated parameter functions. This innovation allows for the simultaneous estimation of structural parameters and control functions, effectively addressing the challenges of endogeneity and sample selection bias. Additionally, the research contributes a generalized influence function that explicitly accounts for the estimation process of generated regressors, extending the double machine learning framework. This advancement enhances the framework's applicability to models with endogeneity or sample selection concerns, while preserving the ability to capture individual heterogeneity through high-dimensional characteristics, thus maintaining the interpretability and validity of causal inferences in econometric analysis.

The simulation results validate the effectiveness of our proposed method, demonstrating that it consistently achieves low bias and stable coverage rates, even in scenarios with high levels of endogeneity. This robustness is particularly evident when compared to traditional methods, such as DLIH and two-stage least squares (2SLS), which suffer from significant biases under similar conditions.

Looking forward, this framework paves the way for further research in multiple directions. First, automatic selection of individual variables is essential, as including irrelevant variables in parameter functions can diminish model performance. Integrating such variable selection into this framework would enhance both its practicality and robustness. Additionally, future research could extend this framework to applications such as personalized policy evaluations and structural economic estimations, where high-dimensional individual characteristics and endogeneity, or sample selection bias, are significant concerns. Specifically, we aim to explore how deep neural networks can model heterogeneous treatment effects in econometrics under endogeneity, building on recent advances in semiparametric inference and influence functions in structural models, especially within industrial organization contexts. Lastly, incorporating automatic locally robust estimation offers an alternative method to derive influence functions and achieve asymptotic normality, as noted in Escanciano and Pérez-Izquierdo (2023). Their innovative approach constructs estimators for nuisance functions without relying on nonparametric plug-in estimators for bias correction, thus minimizing complex computations typically required for generated regressors. Although this concept of auto-DML, where bias correction terms are empirically estimated rather than derived analytically, is implicitly present in our framework, automatic locally robust estimation could further enhance the practical application of this method across empirical settings, maximizing its impact and potential.

References

Ahn H, Powell JL (1993) Semiparametric estimation of censored selection models with a nonparametric selection mechanism. *Journal of Econometrics* 58(1-2):3–29.

Anthony M, Bartlett PL, Bartlett PL, et al. (1999) *Neural network learning: Theoretical foundations*, volume 9 (cambridge university press Cambridge).

Bartlett PL, Harvey N, Liaw C, Mehrabian A (2019) Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research* 20(63):1–17.

Bauer B, Kohler M (2019) On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *The Annals of Statistics* 47(4):2261–2285.

Blundell RW, Powell JL (2004) Endogeneity in semiparametric binary response models. *The Review of Economic Studies* 71(3):655–679.

Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W, Robins J (2018) Double/debiased machine learning for treatment and structural parameters.

Chernozhukov V, Newey WK, Quintas-Martinez V, Syrgkanis V (2021) Automatic debiased machine learning via neural nets for generalized linear regression. *arXiv preprint arXiv:2104.14737* .

Chernozhukov V, Newey WK, Singh R (2022) Automatic debiased machine learning of causal and structural effects. *Econometrica* 90(3):967–1027.

Escanciano JC, Pérez-Izquierdo T (2023) Automatic locally robust estimation with generated regressors. *arXiv preprint arXiv:2301.10643* .

Farrell MH, Liang T, Misra S (2020) Deep learning for individual heterogeneity: an automatic inference framework. *arXiv preprint arXiv:2010.14694* .

Farrell MH, Liang T, Misra S (2021) Deep neural networks for estimation and inference. *Econometrica* 89(1):181–213.

Giné E, Nickl R (2021) *Mathematical foundations of infinite-dimensional statistical models* (Cambridge university press).

Hahn J, Liao Z, Ridder G, Shi R, et al. (2021) The influence function of semiparametric two-step estimators with estimated control variables. Technical report.

Hahn J, Ridder G (2013) Asymptotic variance of semiparametric estimators with generated regressors. *Econometrica* 81(1):315–340.

Hahn J, Ridder G (2019) Three-stage semi-parametric inference: Control variables and differentiability. *Journal of econometrics* 211(1):262–293.

Heckman JJ (1979) Sample selection bias as a specification error. *Econometrica: Journal of the econometric society* 153–161.

Ichimura H, Newey WK (2015) The influence function of semiparametric estimators .

Ichimura H, Newey WK (2022) The influence function of semiparametric estimators. *Quantitative Economics* 13(1):29–61.

Imaizumi M (2023) Sup-norm convergence of deep neural network estimator for nonparametric regression by adversarial training. *arXiv preprint arXiv:2307.04042* .

Imbens GW, Newey WK (2009) Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica* 77(5):1481–1512.

Jiang H (2019) Non-asymptotic uniform rates of consistency for k-nn regression. *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 3999–4006.

Lu J, Shen Z, Yang H, Zhang S (2021) Deep network approximation for smooth functions. *SIAM Journal on Mathematical Analysis* 53(5):5465–5506.

Madry A (2017) Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083* .

Mammen E, Rothe C, Schienle M (2012) Nonparametric regression with nonparametrically generated covariates. *The Annals of Statistics* 1132–1170.

Mammen E, Rothe C, Schienle M (2016) Semiparametric estimation with generated covariates. *Econometric Theory* 32(5):1140–1177.

Newey WK (1994) The asymptotic variance of semiparametric estimators. *Econometrica: Journal of the Econometric Society* 1349–1382.

Newey WK, Powell JL (2003) Instrumental variable estimation of nonparametric models. *Econometrica* 71(5):1565–1578.

Newey WK, Powell JL, Vella F (1999) Nonparametric estimation of triangular simultaneous equations models. *Econometrica* 67(3):565–603.

Olley S, Pakes A (1992) The dynamics of productivity in the telecommunications equipment industry.

Rothe C (2009) Semiparametric estimation of binary response models with endogenous regressors. *Journal of Econometrics* 153(1):51–64.

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

Shen G, Jiao Y, Lin Y, Huang J (2021) Robust nonparametric regression with deep neural networks. *arXiv preprint arXiv:2107.10343* .

Stone CJ (1980) Optimal rates of convergence for nonparametric estimators. *The annals of Statistics* 1348–1360.

Stone CJ (1982) Optimal global rates of convergence for nonparametric regression. *The annals of statistics* 1040–1053.

Van Der Vaart AW, Wellner JA, van der Vaart AW, Wellner JA (1996) *Weak convergence* (Springer).

Wooldridge JM (2015) Control function methods in applied econometrics. *Journal of Human Resources* 50(2):420–445.

APPENDIX: PROOFS OF RESULTS

Appendix A: Convergent Properties for Deep Neural Networks

A.1 Proofs of Sup-norm Convergence for Deep Neural Network

The proof ideas mainly follows with Imaizumi (2023) and adapt to our context.

Before we introduce the main proofs, we need to define several notations related to adversarial training. Let's define an adversarial pseudo-norm of $f : [0,1]^d \to \mathbb{R}$

$$\|f\|_{P,\Delta}^2 := \mathbb{E}_{X\sim P}\left[\max_{x'\in\Delta_h^p(X)} |f(x')|^2\right],$$

and its empirical analogue

$$\|f\|_{n,\Delta}^2 := n^{-1}\sum_{i=1}^{n} \max_{x'\in\Delta_h^p(X_i)} |f(x')|^2,$$

where $\Delta_h^p(x) = \left\{x' \in [0,1]^d \mid \|x - x'\|_p \le h\right\}$ is a neighbourhood of $x \in [0,1]^d$ with a scale multiplier $h \in (\underline{h}, 1)$ with $\underline{h} > 0$. We introduce a uniform covering number for $\mathcal{F}(L,H)$ as defined in empirical process theory (Van Der Vaart et al., 1996)

$$N_{L,H}(\delta) := \sup_{Q_n} N\left(\delta, \mathcal{F}(L,H), \|\cdot\|_{L^2(Q_n)}\right),$$

where Q_n is an empirical measure with n samples and $N\left(\delta, \mathcal{F}(L,H), \|\cdot\|_{L^2(Q_n)}\right)$ is minimum cardinality of a δ-covering set, denoted by $\{f_1, \ldots, f_N\} \subset \mathcal{F}$, of $\mathcal{F}(L,H)$ for $\delta \in (0,1]$.

We further define an approximation error of deep neural networks in $\mathcal{F}(L,H)$ as

$$\Phi_{L,H} := \inf_{f\in\mathcal{F}(L,H)} \|f - f^*\|_{L^\infty}. \tag{30}$$

This term illustrates the expressive capabilities of neural networks in $\mathcal{F}(L,H)$, a characteristic that diminishes with the increase of either L or H as demonstrated in Lu et al. (2021) for an example. For notation simplication, we define residual of preprocessed output $\hat{Y}$ as following

$$\Xi(x) := \widehat{Y}(x) - f^*(x).$$

A.1.1 Proofs of Theorem 1

Theorem 1.(General Feedforward Architecture) Suppose that Assumption 1, 2, and 6 hold for some $\beta > 0$. Let $f^*(\boldsymbol{x})$ be regression function 9 and $\widehat{f}$ be corrected adversarial estimator

10 under the function class $\mathcal{F}(L, H)$ of deep neural networks. Then

$$\mathbb{E}\left[\left\|\widehat{f} - f^*\right\|_{L^\infty}^2\right] \leq C_{P_X,p,B,d,\beta} h^{-d} \left(\frac{(1+h^{-d})(HL)^2 \log(HL) \log n}{n} + (LH)^{-4\beta/d} + \mathbb{E}\left[\|\Xi\|_{L^\infty}^2\right]\right) \tag{31}$$

for every $n \geq \bar{n}$ with some $\bar{n} \in \mathbb{N}$.

Proof. In the process of preparation, Lemma 6 establishes the following bound:

$$\Phi_{L,W} \leq C_{d,\beta}(LH)^{-2\beta/d}.$$

By leveraging this bound for $\Phi_{L,W}$, we employ Proposition 1 to deduce the following inequality:

$$\begin{aligned}
&\mathbb{E}\left[\left\|\widehat{f} - f^*\right\|_{L^\infty}^2\right] \\
&\leq C_{P_X,p,B,d,\beta} h^{-d} \left(\frac{(1+h^{-d})(HL)^2 \log(HL) \log n}{n} + (LH)^{-4\beta/d} + \mathbb{E}\left[\|\Xi\|_{L^\infty}\right](LH)^{-2\beta/d} + \mathbb{E}\left[\|\Xi\|_{L^\infty}^2\right]\right).
\end{aligned}$$

In addition, it is feasible to represent the following expression:

$$(LH)^{-4\beta/d} + \mathbb{E}\left[\|\Xi\|_{L^\infty}\right](LH)^{-2\beta/d} + \mathbb{E}\left[\|\Xi\|_{L^\infty}^2\right] \leq \left\{(LH)^{-2\beta/d} + \mathbb{E}\left[\|\Xi\|_{L^\infty}^2\right]^{1/2}\right\}^2,$$

Therefore, we can derive the statement as following

$$\mathbb{E}\left[\left\|\widehat{f} - f^*\right\|_{L^\infty}^2\right] \leq C_{P_X,p,B,d,\beta} h^{-d} \left(\frac{(1+h^{-d})(HL)^2 \log(HL) \log n}{n} + (LH)^{-4\beta/d} + \mathbb{E}\left[\|\Xi\|_{L^\infty}^2\right]\right).$$

□

A.1.2 Proofs of corollary 1

Corollary 1. Suppose that Assumption 1, 2, and 6 hold for some $\beta > 0$. Let $f^*(\boldsymbol{x})$ be regression function 9 and $\widehat{f}$ be corrected adversarial estimator 10 under the function class $\mathcal{F}(L, H)$ of deep neural networks with depth L and width H setted as $HL \asymp n^{d/(4\beta+2d)}$. Assume that $\zeta_n^2 = O\left(n^{-2\beta/(2\beta+d)} \log^{\beta^*} n\right)$ for some $\beta^* > 0$. Then

$$\mathbb{E}\left[\left\|\widehat{f} - f^*\right\|_{L^\infty}^2\right] = O\left(n^{-2\beta/(2\beta+d)} \log^{2\vee\beta^*} n\right).$$

Proof. Based on theorem 1, we place $HL = Cn^{d/(4\beta+2d)}$ and $\zeta_n^2 = C'n^{-2\beta/(2\beta+d)} \log^{\beta^*} n$, for $n > \bar{n}'$ with some $\bar{n}' \in \mathbb{N}$, into the right side of inequality 31. Then we obtain

$$\begin{aligned}
&\mathbb{E}\left[\left\|\widehat{f} - f^*\right\|_{L^\infty}^2\right] \\
&\leq C_{P_X,p,d,B,d,\beta} h^{-d} \left(n^{-2\beta/(2\beta+d)} \left((1+h^{-d})\log^2 n + 1\right) + \mathbb{E}\left[\|\Xi\|_{L^\infty}^2\right]\right) \\
&\leq C_{P_X,p,d,B,d,\beta} h^{-d}(1+h^{-d}) \left(n^{-2\beta/(2\beta+d)} \left(\log^2 n + 1\right) + n^{-2\beta/(2\beta+d)} \log^{\beta^*} n\right) \\
&= O\left(n^{-2\beta/(2\beta+d)} \log^{2\vee\beta^*} n\right)
\end{aligned}$$

□

A.1.3 Proofs of corollary 1

Proposition 1. *Suppose that Assumption 1 and 2 hold. Let $f^*(\boldsymbol{x})$ be regression function (9) and $\hat{f}$ be the corrected adversarial estimator(10) restricted to the function class $\mathcal{F}(L, H)$ of deep neural networks. Then*

$$\begin{aligned}
&\mathbb{E}\left[\left\|\widehat{f} - f^*\right\|_{L^\infty}^2\right] \\
&\leq C_{P_X,p,d,B} h^{-d} \left((1+h^{-d})\frac{H^2L^2\log(HL)\log(n)}{n} + \Phi_{L,H}^2 + \mathbb{E}\left[\|\Xi\|_{L^\infty}\right]\Phi_{L,H} + \mathbb{E}\left[\|\Xi\|_{L^\infty}^2\right]\right).
\end{aligned}$$

Proof. Let's start with Lemma 1 to bound the sup-norm as

$$\left\|\widehat{f} - f^*\right\|_{L^\infty}^2 \leq 2\left(C_{P_X,p,d}h^d\right)^{-1} \left\|\widehat{f} - f^*\right\|_{P_X,\Delta}^2. \tag{32}$$

Then with lemma 2, we have the following inequality

$$\begin{aligned}
&\mathbb{E}\left[\left\|\widehat{f} - f^*\right\|_{P_X,\Delta}^2\right] \\
&\leq \mathbb{E}\left[\left\|\widehat{f} - f^*\right\|_{n,\Delta}^2\right] + \frac{(18B^2+6B)\log N_{L,H}(\delta) + 6B + 32B^2}{n} + \frac{22B^2}{n^2} + 20B\delta + 2\delta^2.
\end{aligned}$$

Based on lemma 2, we further apply Proposition 12 of Imaizumi (2023)(3) to obtain further inequality

$$\begin{aligned}
&\mathbb{E}\left[\left\|\widehat{f} - f^*\right\|_{P_X,\Delta}^2\right] \\
&\leq \mathbb{E}\left[\left\|\widehat{f} - f^*\right\|_{n,\Delta}^2\right] + \frac{(18B^2+6B)\log N_{L,H}(\delta) + 6B + 32B^2}{n} + \frac{22B^2}{n^2} + 20B\delta + 2\delta^2 \\
&\leq \left(2\mathbb{E}\left[\left\|\widehat{f} - f^*\right\|_{L^\infty}^2\right]^{1/2} + 2\delta\right)\left(\frac{\log N_{L,W}(\delta) + \mathbb{E}\left[\|\Xi\|_{L^\infty}^2\right]}{n}\right)^{1/2} + 2\mathbb{E}\left[\|\Xi\|_{L^\infty}^2\right]^{1/2}\delta \\
&+ \frac{(18B^2+6B)\log N_{L,H}(\delta) + 6B + 32B^2}{n} + \frac{22B^2}{n^2} + 20B\delta + 2\delta^2 + \Phi_{L,W}^2 + 2\mathbb{E}\left[\|\Xi\|_{L^\infty}\right]\Phi_{L,W} + 2\mathbb{E}\left[\|\Xi\|_{L^\infty}^2\right],
\end{aligned}$$

Then we combine the inequality with expectation of lemma 1 inequality and derive following

$$\mathbb{E}\left[\left\|\widehat{f}-f^*\right\|_{L^\infty}^2\right]$$
$$\leq C_{P_X,p,d}h^{-d}\left(\mathbb{E}\left[\left\|\widehat{f}-f^*\right\|_{L^\infty}^2\right]^{1/2}+\delta\right)\left(\frac{\log N_{L,W}(\delta)+\mathbb{E}\left[\|\Xi\|_{L^\infty}^2\right]}{n}\right)^{1/2}+$$
$$C_{P_X,p,d}h^{-d}\left(\frac{(B^2+B)\log N_{L,H}(\delta)+B+B^2}{n}+\frac{B^2}{n^2}+B\delta+\delta^2+\Phi_{L,H}^2+\mathbb{E}\left[\|\Xi\|_{L^\infty}\right]\Phi_{L,H}+\mathbb{E}\left[\|\Xi\|_{L^\infty}^2\right]+\mathbb{E}\left[\|\Xi\|_{L^\infty}^2\right]^{1/2}\delta\right)$$

With a quadratic inequality $z^2\leq az+b$ for $a,b\geq 0$ and $z\in\mathbb{R}$, we can derive $z^2\leq a^2+2b$. Then, we can further obtain the following inequality with $z=\mathbb{E}\left[\left\|\widehat{f}-f^*\right\|_{L^\infty}^2\right]^{1/2}$

$$\mathbb{E}\left[\left\|\widehat{f}-f^*\right\|_{L^\infty}^2\right]$$
$$\leq C_{P_X,p,d,B}h^{-d}\left\{(1+h^{-d})\frac{\log N_{L,H}(\delta)}{n}+\frac{1}{n^2}+\delta+\delta^2+\Phi_{L,H}^2+\mathbb{E}\left[\|\Xi\|_{L^\infty}\right]\Phi_{L,H}+\left(1+\frac{h^{-d}}{n}\right)\mathbb{E}\left[\|\Xi\|_{L^\infty}^2\right]\right.$$
$$\left.+\mathbb{E}\left[\|\Xi\|_{L^\infty}^2\right]^{1/2}\delta+\delta\left(\frac{\log N_{L,H}(\delta)+\mathbb{E}\left[\|\Xi\|_{L^\infty}^2\right]}{n}\right)^{1/2}\right\}. \tag{33}$$

In addition, we can calculate the bound of metric entropy by lemma 4 with $\delta=1/n$, which is

$$\log N_{L,H}(1/n)=\log\sup_{Q_n}N\left(1/n,\mathcal{F}(L,H),\|\cdot\|_{L^2(Q_n)}\right)\leq CH^2L^2\log(HL)\log\left(Bn^2\right).$$

By replace the bound of metric entropy into the inequality 33, we may obtain

$$\mathbb{E}\left[\left\|\widehat{f}-f^*\right\|_{L^\infty}^2\right]$$
$$\leq C_{P_X,p,d,B}h^{-d}\left((1+h^{-d})\frac{H^2L^2\log(HL)\log(n)}{n}+\Phi_{L,H}^2+\mathbb{E}\left[\|\Xi\|_{L^\infty}\right]\Phi_{L,H}+\mathbb{E}\left[\|\Xi\|_{L^\infty}^2\right]\right).$$

□

Supportive Results

Lemma 1. *Suppose P_X satisfies Assumption 1 and f^* is continuous. For any bounded and continuous $f:[0,1]^d\to\mathbb{R}$, we have*

$$\|f-f^*\|_{L^\infty}^2\leq\left(C_{P_X,p,d}h^d\right)^{-1}\|f-f^*\|_{P_X,\Delta}^2.$$

Proof. Since both $f(x)$ and f^* are bounded and continuous by compact domain of $[0,1]^d$, it implies $g(\cdot) = (f(\cdot) - f_*(\cdot))^2$ is also bounded and continuous. With Lemma 15 of Imaizumi (2023), we derive that

$$\|f - f^*\|_{P_X,\Delta}^2 \geq C_{P_X,p,d} h^d \sup_{x' \in [0,1]^d} \left(f\left(x'\right) - f^*\left(x'\right)\right)^2 = C_{P_X,p,d} h^d \left\|f - f^*\right\|_{L^\infty}^2 .$$

Dividing the constant factor $C_{P_X,p,d}h^d$ on both sides of the above inequality completes the proof. □

Lemma 2. *Consider $f^*(\boldsymbol{x})$ is continuous with $\|f^*\|_{L^\infty} \leq B$ and function class $\mathcal{F}$ with all continuous functions. Then given any $\delta > 0$, we have*

$$\mathbb{E}\left[\left\|\widehat{f} - f^*\right\|_{P_X,\Delta}^2\right] \leq \mathbb{E}\left[\left\|\widehat{f} - f^*\right\|_{n,\Delta}^2\right] + \frac{(18B^2 + 6B) log N_{L,H}(\delta) + 6B + 32B^2}{n} + \frac{22B^2}{n^2} + 20B\delta + 2\delta^2,$$

Proof. Without loss of generality, we assume that $N_{L,H}(\delta) \geq 3$ and $\log N_{L,H}(\delta) \leq n$. Let X_i' be an i.i.d. samples from P_X for $i = 1, \ldots, n$. Note that $\widehat{Y}$ depends on $X_1, \ldots, X_n$. Let's define the closest element of δ-cover of $\mathcal{F}$ to $\hat{f}$ as $\widehat{j} \coloneqq \mathrm{argmin}_{j'=1,\ldots,N} \sup_{Q_n} \left\|f_{j'} - \widehat{f}\right\|_{L^2(Q_n)}$. Suppose there are two sequences of independent and identically distributed data samples denoted as (X_i, X_i') from the probability distribution P_X for i = 1, 2,...,n. The sequence of random variables X_i is designated for training the function $\hat{f}$, while the other sequence of random variables X_i' is allocated for the symmetrization technique employed in the empirical process.

We start with bounding the below difference

$$
\begin{aligned}
&\left|\mathbb{E}\left[\left\|\widehat{f}-f^*\right\|_{P_X,\Delta}^2\right]-\mathbb{E}\left[\left\|\widehat{f}-f^*\right\|_{n,\Delta}^2\right]\right| \\
&=\left|\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\sup_{x'\in\Delta_h^p\left(X_i'\right)}\left(\widehat{f}\left(x'\right)-f^*\left(x'\right)\right)^2-\sup_{x'\in\Delta_h^p(X_i)}\left(\widehat{f}\left(x'\right)-f^*\left(x'\right)\right)^2\right]\right| \\
&=\left|\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\sup_{x'\in\Delta_h^p\left(X_i'\right)}\left(\widehat{f}\left(x'\right)-f_{\widehat{j}}\left(x'\right)+f_{\widehat{j}}\left(x'\right)-f^*\left(x'\right)\right)^2-\sup_{x'\in\Delta_h^p(X_i)}\left(\widehat{f}\left(x'\right)-f_{\widehat{j}}\left(x'\right)+f_{\widehat{j}}\left(x'\right)-f^*\left(x'\right)\right)^2\right]\right| \\
&\leq\left|\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\underbrace{\sup_{x'\in\Delta_h^p\left(X_i'\right)}\left(f_{\widehat{j}}\left(x'\right)-f^*\left(x'\right)\right)^2-\sup_{x'\in\Delta_h^p(X_i)}\left(f_{\widehat{j}}\left(x'\right)-f^*\left(x'\right)\right)^2}_{=:g_{\widehat{j}}\left(X_i,X_i'\right)}\right]\right| \\
&+2\left|\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\sup_{x'\in\Delta_h^p\left(X_i'\right)}\left(\widehat{f}\left(x'\right)-f_{\widehat{j}}\left(x'\right)\right)\left(f_{\widehat{j}}\left(x'\right)-f^*\left(x'\right)\right)\right)-\sup_{x'\in\Delta_h^p((X_i)}\left(\widehat{f}\left(x'\right)-f_{\widehat{j}}\left(x'\right)\right)\left(f_{\widehat{j}}\left(x'\right)-f^*\left(x'\right)\right)\right)\right] \\
&+\left|\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\sup_{x'\in\Delta_h^p\left(X_i'\right)}\left(\widehat{f}\left(x'\right)-f_{\widehat{j}}\left(x'\right)\right)^2-\sup_{x'\in\Delta_h^p(X_i)}\left(\widehat{f}\left(x'\right)-f_{\widehat{j}}\left(x'\right)\right)^2\right]\right| \\
&\leq\left|\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}g_{\widehat{j}}\left(X_i,X_i'\right)\right]\right|+4\mathbb{E}\left[\sup_{Q_n}\left\|\widehat{f}-f_{\widehat{j}}\right\|_{L^2(Q_n)}^2\right]^{1/2}\mathbb{E}\left[\sup_{Q_n}\left\|f_{\widehat{j}}-f^*\right\|_{L^2(Q_n)}^2\right]^{1/2}+2\mathbb{E}\left[\sup_{Q_n}\left\|\widehat{f}-f_{\widehat{j}}\right\|_{L^2(Q_n)}^2\right] \\
&\leq\left|\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}g_{\widehat{j}}\left(X_i,X_i'\right)\right]\right|+4\delta\mathbb{E}\left[\left\|f_{\widehat{j}}-f^*\right\|_{L^\infty}^2\right]^{1/2}+2\delta^2 \\
&\leq\left|\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}g_{\widehat{j}}\left(X_i,X_i'\right)\right]\right|+8\delta B+2\delta^2. \qquad (34)
\end{aligned}
$$

The final inequality is derived from the definition of $\widehat{j}$ and the constraints on $f\in\mathcal{F}$ and f^*, bounded by B. Moving on to analyzing the first term in the bound above, we introduce the following: $r_j=B\max\left\{\left\|f_j-f^*\right\|_{P_X,\Delta},\left(n^{-1}\log N_{L,H}(\delta)\right)^{1/2}\right\}$, for $j=1,\ldots,N$, which leads to

$$
\begin{aligned}
r_{\widehat{j}} &\leq B\mathbb{E}_{X|X_{1:n},X_{1:n}'}\left[\sup_{x'\in\Delta_h^p(X)}\left(f_{\widehat{j}}\left(x'\right)-f^*\left(x'\right)\right)^2\right]^{1/2}+B\left(n^{-1}\log N_{L,H}(\delta)\right)^{1/2} \\
&\leq B\mathbb{E}_{X|X_{1:n},X_{1:n}'}\left[\sup_{x'\in\Delta_h^p(X)}\left(\widehat{f}\left(x'\right)-f^*\left(x'\right)\right)^2\right]^{1/2}+B\left(n^{-1}\log N_{L,H}(\delta)\right)^{1/2}+B\delta,
\end{aligned}
$$

where $\mathbb{E}_{X|X_{1:n},X'_{1:n}}[\cdot]$ denotes a conditional expectation with given $X_1,\ldots,X_n$ and $X'_1,\ldots,X'_n$. By the law of iterated expectation, the first term of the bound is decomposed as

$$
\begin{aligned}
&\left|\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} g_{\widehat{j}}\left(X_i,X'_i\right)\right]\right| \\
&= \frac{1}{n}\left|\mathbb{E}\left[\sum_{i=1}^{n}\underbrace{\frac{g_{\widehat{j}}\left(X_i,X'_i\right)}{r_{\widehat{j}}}}_{=:\tilde{g}_{\widehat{j}}\left(X_i,X'_i\right)} r_{\widehat{j}}\right]\right| \\
&\leq \frac{1}{n}\left|\mathbb{E}\left[\sum_{i=1}^{n}\widetilde{g}_{\widehat{j}}\left(X_i,X'_i\right)\left(B\mathbb{E}_{X|X_{1:n},X'_{1:n}}\left[\sup_{x'\in\Delta_h^p(X)}\left(\widehat{f}\left(x'\right)-f^*\left(x'\right)\right)^2\right]^{1/2}+B\left(n^{-1}\log N_{L,H}(\delta)\right)^{1/2}+B\delta\right)\right]\right| \\
&\leq \frac{1}{n}\left|\mathbb{E}\left[\max_{j=1,\ldots,N_{L,H}(\delta)}\sum_{i=1}^{n}\widetilde{g}_j\left(X_i,X'_i\right)\left(B\mathbb{E}_{X|X_{1:n},X'_{1:n}}\left[\sup_{x'\in\Delta_h^p(X)}\left(\widehat{f}\left(x'\right)-f^*\left(x'\right)\right)^2\right]^{1/2}\right)\right]\right| \\
&+\frac{B}{n}\left|\mathbb{E}\max_{j=1,\ldots,N_{L,H}(\delta)}\sum_{i=1}^{n}\widetilde{g}_j\left(X_i,X'_i\right)\left(\left(n^{-1}\log N_{L,H}(\delta)\right)^{1/2}+\delta\right)\right| \\
&\leq \frac{B}{n}\left|\mathbb{E}\left[\left(\max_{j=1,\ldots,N_{L,H}(\delta)}\sum_{i=1}^{n}\widetilde{g}_j\left(X_i,X'_i\right)\right)^2\right]^{1/2}\mathbb{E}\left[\left\|\widehat{f}-f^*\right\|_{P_X,\Delta}^2\right]^{1/2}\right| \\
&+\frac{B}{n}\mathbb{E}\left[\max_{j=1,\ldots,N_{L,H}(\delta)}\sum_{i=1}^{n}\widetilde{g}_j\left(X_i,X'_i\right)\right]\left(\left(n^{-1}\log N_{L,H}(\delta)\right)^{1/2}+\delta\right) \\
&\leq \frac{B}{n}\left(36n\log N_{L,H}(\delta)+64n+\frac{392n}{9\log N_{L,H}(\delta)}\right)^{1/2}\mathbb{E}\left[\left\|\widehat{f}-f^*\right\|_{P_X,\Delta}^2\right]^{1/2} \\
&+\frac{B}{n}\left(6\log N_{L,H}(\delta)+\frac{16}{3}+6\delta\left(n\log N_{L,H}(\delta)\right)^{1/2}+\frac{16\delta n^{1/2}}{3\left(\log N_{L,H}(\delta)\right)^{1/2}}\right). \qquad (35)
\end{aligned}
$$

The penultimate inequality arises as a result of the Cauchy-Schwarz inequality. Through the application of Lemma 4 and the condition $1 \leq \log N_{L,H}(\delta) \leq n$, we derive the ultimate inequality. Upon substituting the result from 35 into the constraint 34, we obtain the subsequent

inequality:

$$\begin{aligned}
&\left|\mathbb{E}\left[\left\|\widehat{f}-f^*\right\|_{P_X,\Delta}^2\right]-\mathbb{E}\left[\left\|\widehat{f}-f^*\right\|_{n,\Delta}^2\right]\right| \\
&\leq \frac{B}{n}\left(36n\log N_{L,H}(\delta)+64n+\frac{392n}{9\log N_{L,H}(\delta)}\right)^{1/2}\mathbb{E}\left[\left\|\widehat{f}-f^*\right\|_{P_X,\Delta}^2\right]^{1/2} \\
&+\frac{B}{n}\left(6\log N_{L,H}(\delta)+\frac{16}{3}+6\delta\left(n\log N_{L,H}(\delta)\right)^{1/2}+\frac{16\delta n^{1/2}}{3\left(\log N_{L,H}(\delta)\right)^{1/2}}\right)+8\delta B+2\delta^2
\end{aligned}$$

We rearrange the term and obtain that

$$\begin{aligned}
\mathbb{E}\left[\left\|\widehat{f}-f^*\right\|_{P_X,\Delta}^2\right] &\leq \mathbb{E}\left[\left\|\widehat{f}-f^*\right\|_{n,\Delta}^2\right] \\
&+\frac{B}{n}\left(6\log N_{L,H}(\delta)+\frac{16}{3}+6\delta\left(n\log N_{L,H}(\delta)\right)^{1/2}+\frac{16\delta n^{1/2}}{3\left(\log N_{L,H}(\delta)\right)^{1/2}}\right) \\
&+\frac{B^2\left(36n\log N_{L,H}(\delta)+64n+\frac{392n}{9\log N_{L,H}(\delta)}\right)}{2n^2}+8\delta B+2\delta^2,
\end{aligned}$$

where δ will be chosen as $\mathcal{O}(n^{-1})$ and simplify the inequality to obtain the statement.

□

Lemma 3. (*Proposition 12 in Imaizumi* (2023)). *Consider the setting in Theorem 3 of Imaizumi (2023). Then, for any $\delta \in (0,1]$, we have*

$$\begin{aligned}
\mathbb{E}\left[\left\|\widehat{f}-f^*\right\|_{n,\Delta}^2\right] &\leq \left(2\mathbb{E}\left[\left\|\widehat{f}-f^*\right\|_{L^\infty}^2\right]^{1/2}+2\delta\right)\left(\frac{\log N_{L,W}(\delta)+\mathbb{E}\left[\|\Xi\|_{L^\infty}^2\right]}{n}\right)^{1/2}+2\mathbb{E}\left[\|\Xi\|_{L^\infty}^2\right]^{1/2}\delta \\
&+\Phi_{L,W}^2+2\mathbb{E}\left[\|\Xi\|_{L^\infty}\right]\Phi_{L,W}+2\mathbb{E}\left[\|\Xi\|_{L^\infty}^2\right].
\end{aligned}$$

Proof. According to the definition of the minimization problem, $L_n(\widehat{f}) \leq L_n(f)$ for any $f \in \mathcal{F}(L,H)$. Therefore, we derive the following fundamental inequality:

$$\frac{1}{n}\sum_{i=1}^n \max_{x'\in\Delta_h^p(X_i)}\left(\widehat{Y}(x')-\widehat{f}(x')\right)^2 \leq \frac{1}{n}\sum_{i=1}^n \max_{x'\in\Delta_h^p(X_i)}\left(\widehat{Y}(x')-f(x')\right)^2,$$

which can be expressed as

$$\frac{1}{n}\sum_{i=1}^n \max_{x'\in\Delta_h^p(X_i)}\left(f^*(x')+\Xi(x')-\widehat{f}(x')\right)^2 \leq \frac{1}{n}\sum_{i=1}^n \max_{x'\in\Delta_h^p(X_i)}\left(f^*(x')+\Xi(x')-f(x')\right)^2. \quad (36)$$

We bound both sides of 36. The left-hand side (LHS) of 36 is lower bounded as

$$\begin{aligned}\text{LHS of (36)} &= \frac{1}{n}\sum_{i=1}^{n}\max_{x'\in\Delta_h^p(X_i)}\left\{\left(f^*\left(x'\right)-\widehat{f}\left(x'\right)\right)^2+\Xi\left(x'\right)^2+2\Xi\left(x'\right)\left(f^*\left(x'\right)-\widehat{f}\left(x'\right)\right)\right\}\\ &\geq \left\|f^*-\widehat{f}\right\|_{n,\Delta}^2-\|\Xi\|_{n,\Delta}^2-\frac{2}{n}\sum_{i=1}^{n}\max_{x'\in\Delta_h^p(X_i)}\left|\Xi\left(x'\right)\left(f^*\left(x'\right)-\widehat{f}\left(x'\right)\right)\right|, \qquad (37)\end{aligned}$$

Similarly, we bound the right-hand side of (36) as

$$\begin{aligned}\text{RHS of (36)} &= \frac{1}{n}\sum_{i=1}^{n}\max_{x'\in\Delta_h^p(X_i)}\left\{\left(f^*\left(x'\right)-f\left(x'\right)\right)^2+\Xi\left(x'\right)^2+2\Xi\left(x'\right)\left(f^*\left(x'\right)-f\left(x'\right)\right)\right\}\\ &\leq \|f^*-f\|_{n,\Delta}^2+\|\Xi\|_{n,\Delta}^2+\frac{2}{n}\sum_{i=1}^{n}\max_{x'\in\Delta_h^p(X_i)}\left|\Xi\left(x'\right)\left(f^*\left(x'\right)-f\left(x'\right)\right)\right|. \qquad (38)\end{aligned}$$

By combining equations (37) and (38) with (36), we obtain the following inequality:

$$\begin{aligned}\left\|f^*-\widehat{f}\right\|_{n,\Delta}^2 \leq \|f^*-f\|_{n,\Delta}^2+2\|\Xi\|_{n,\Delta}^2+\underbrace{\frac{2}{n}\sum_{i=1}^{n}\max_{x'\in\Lambda_h^p(X_i)}\left|\Xi\left(x'\right)\left(f^*\left(x'\right)-\widehat{f}\left(x'\right)\right)\right|}_{=:S_1}\\ +\frac{2}{n}\sum_{i=1}^{n}\max_{x'\in\Lambda_h^p(X_i)}\left|\Xi\left(x'\right)\left(f^*\left(x'\right)-f\left(x'\right)\right)\right|\\ \leq \Phi_{L,W}^2+2\|\Xi\|_{L^\infty}^2+S_1+2\|\Xi\|_{L^\infty}\Phi_{L,W}, \qquad (39)\end{aligned}$$

by the definition of $\Phi_{L,W}$ in (30).

To bound $\mathbb{E}\left[S_1\right]$, we define the nearest element of the covering set to $\widehat{f}$, denoted as $\widehat{j} := \operatorname{argmin}_{j'=1,\ldots,N}\sup_{Q_n}\left\|f_{j'}-\widehat{f}\right\|_{L^2(Q_n)}$. Then, we can bound $\mathbb{E}\left[S_1\right]$ as follows:

$$
\begin{aligned}
\mathbb{E}\left[S_1\right] &= \mathbb{E}\left[\frac{2}{n}\sum_{i=1}^{n}\max_{x'\in\Delta_h(X_i)}\left|\Xi\left(x'\right)\left(f^*\left(x'\right)-f_{\widehat{j}}\left(x'\right)+f_{\widehat{j}}\left(x'\right)-\widehat{f}\left(x'\right)\right)\right|\right] \\
&\leq \mathbb{E}\left[\frac{2}{n}\sum_{i=1}^{n}\max_{x'\in\Delta_h(X_i)}\left|\Xi\left(x'\right)\left(f^*\left(x'\right)-f_{\hat{j}}\left(x'\right)\right)\right|\right]+\mathbb{E}\left[\frac{2}{n}\sum_{i=1}^{n}\max_{x'\in\Delta_h(X_i)}\left|\Xi\left(x'\right)\left(f_{\hat{j}}\left(x'\right)-\widehat{f}\left(x'\right)\right)\right|\right] \\
&\leq \mathbb{E}\left[\frac{2}{n}\sum_{i=1}^{n}\max_{x'\in\Delta_h(X_i)}\left|\Xi\left(x'\right)\left(f^*\left(x'\right)-f_{\hat{j}}\left(x'\right)\right)\right|\frac{\left\|\widehat{f}-f^*\right\|_{L^\infty}+\delta}{\left\|f_{\widehat{j}}-f^*\right\|_{L^\infty}}\right] \\
&\quad + 2\mathbb{E}\left[\sup_{Q_n}\|\Xi\|^2_{L^2(Q_n)}\right]^{1/2}\mathbb{E}\left[\sup_{Q_n}\left\|f_{\widehat{j}}-\widehat{f}\right\|^2_{L^2(Q_n)}\right]^{1/2} \\
&\leq \mathbb{E}\left[\left(\left\|\widehat{f}-f^*\right\|_{L^\infty}+\delta\right)\underbrace{\frac{2}{n}\sum_{i=1}^{n}\frac{\max_{x'\in\Delta_h(X_i)}\left|\Xi\left(x'\right)\left(f^*\left(x'\right)-f_{\hat{j}}\left(x'\right)\right)\right|}{\left\|f_{\widehat{j}}-f^*\right\|_{L^\infty}}}_{=:Z_{\widehat{j}}}\right]+2\mathbb{E}\left[\|\Xi\|^2_{L^\infty}\right]^{1/2}\delta.
\end{aligned}
\tag{40}
$$

We can further bound the first term of (40) using the Cauchy-Schwarz inequality, which results in

$$
\begin{aligned}
\mathbb{E}\left[\left(\left\|\widehat{f}-f^*\right\|_{L^\infty}+\delta\right)Z_{\widehat{j}}\right] &\leq \mathbb{E}\left[\left(\left\|\widehat{f}-f^*\right\|_{L^\infty}+\delta\right)^2\right]^{1/2}\mathbb{E}\left[Z^2_{\widehat{j}}\right]^{1/2} \\
&\leq \left(\mathbb{E}\left[\left\|\widehat{f}-f^*\right\|^2_{L^\infty}\right]^{1/2}+\delta\right)\mathbb{E}\left[\max_{j=1,\ldots,N_{L,W}(\delta)}Z_j^2\right]^{1/2} \\
&\leq 2\left(\mathbb{E}\left[\left\|\widehat{f}-f^*\right\|^2_{L^\infty}\right]^{1/2}+\delta\right)\left(\frac{\log N_{L,W}(\delta)+\mathbb{E}\left[\|\Xi\|^2_{L^\infty}\right]}{n}\right)^{1/2},
\end{aligned}
$$

where the last inequality is guaranteed by the bounded random process of Z_j^2,

$$
Z_j \leq \frac{2}{n}\sum_{i=1}^{n}\left|\frac{\max_{x'\in A_h(X_i)}\left\{\left|\Xi\left(x'\right)\|\left(f^*\left(x'\right)-f_j\left(x'\right)\right)\right|\right\}}{\|f_j-f^*\|_{L^\infty}}\right| \leq 2\|\Xi\|_{L^\infty} \text{ and } Z_j^2 \leq 4\|\Xi\|^2_{L^\infty},
$$

for any $j=1,\ldots,N$, and the application of the maximal inequality in theorem 3.1.10 of Giné and Nickl (2021). Therefore, we have

$$
\mathbb{E}\left[S_1\right] \leq \left(2\mathbb{E}\left[\left\|\widehat{f}-f^*\right\|^2_{L^\infty}\right]^{1/2}+2\delta\right)\left(\frac{\log N_{L,W}(\delta)+\mathbb{E}\left[\|\Xi\|^2_{L^\infty}\right]}{n}\right)^{1/2}+2\mathbb{E}\left[\|\Xi\|^2_{L^\infty}\right]^{1/2}\delta \tag{41}
$$

By inserting bound 41 into 39, we can calculate the expectation of inequality (39) to establish the lemma. □

Lemma 4. *Suppose that $N_{L,H}(\delta) \geq 3$. For the function $\widetilde{g}_j(X_i, X_i')$ defined in the proof of Lemma 2. we have*

$$\mathbb{E}\left[\max_{j=1,\ldots,N_{L,H}(\delta)} \sum_{i=1}^{n} \tilde{g}_j(X_i, X_i')\right] \leq 6\left(n \log N_{L,H}(\delta)\right)^{1/2} + \frac{16n^{1/2}}{3\left(\log N_{L,H}(\delta)\right)^{1/2}}, \tag{42}$$

and

$$\mathbb{E}\left[\left(\max_{j=1,\ldots,N_{L,H}(\delta)} \sum_{i=1}^{n} \widetilde{g}_j(X_i, X_i')\right)^2\right] \leq 36n \log N_{L,H}(\delta) + 64n + \frac{392n}{9 \log N_{L,H}(\delta)}. \tag{43}$$

Proof. We begin by deriving a tail bound for the sums of $\sum_{i=1}^{n} \widetilde{g}_j(X_i, X_i')$ for any $j = 1, \ldots, N_{L,H}(\delta)$. To prepare for the Bernstein inequality, it is known that $\mathbb{E}\left[\widetilde{g}_j(X_i, X_i')\right] = 0$ and $|\widetilde{g}_j(X_i, X_i')| \leq 4B^2/r_j \leq 4n^{1/2}/\left(\log N_{L,H}(\delta)\right)^{1/2} \coloneqq M$ as per the definition of r_j. Additionally,

$$\begin{aligned}
\operatorname{Var}\left(\widetilde{g}_j(X_i, X_i')\right) &= 2r_j^{-2} \operatorname{Var}\left(\sup_{x' \in \Delta_h^p(X_1)} \left(f_j(x') - f^*(x')\right)^2\right) \\
&\leq 2r_j^{-2} \mathbb{E}\left[\left(\sup_{x' \in \Delta_h^p(X_1)} \left(f_j(x') - f^*(x')\right)^2\right)^2\right] \\
&\leq 8r_j^{-2} \mathbb{E}\left[\|f_j - f^*\|_{P_X,\Delta}^2\right] B^2 \\
&\leq 8,
\end{aligned}$$

where the penultimate inequality is derived from Hölder's inequality. By employing the bounds above, we apply the Bernstein inequality as given $t \geq 6\left(n \log N_{L,H}(\delta)\right)^{1/2}$,

$$\begin{aligned}
\mathbb{P}\left(\sum_{i=1}^{n} \widetilde{g}_j(X_i, X_i') \geq t\right) &\leq \exp\left(-\frac{\frac{1}{2}t^2}{\sum_{i=1}^{n} \mathrm{E}\left[\widetilde{g}_j(X_i, X_i')^2\right] + tM/3}\right) \\
&= \exp\left(-\frac{t^2}{16n + 8tn^{1/2}/\left(\log N_{L,H}(\delta)\right)^{1/2}/3}\right) \\
&\leq \exp\left(-\frac{3t\left(\log N_{L,H}(\delta)\right)^{1/2}}{16n^{1/2}}\right),
\end{aligned} \tag{44}$$

where the final inequality is derived by $8tn^{1/2}\left(\log N_{L,H}(\delta)\right)^{-1/2}/3 \geq 16n$ when t exceeding the threshold of $6(n \log N)^{1/2}$.

With 44 associated with $t \geq 6\left(n \log N_{L,H}(\delta)\right)^{1/2}$, we derive the maximum inequality as following

$$
\begin{aligned}
&\mathbb{E}\left[\max_{j=1,\ldots,N_{L,H}(\delta)} \sum_{i=1}^{n} \tilde{g}_j\left(X_i, X_i'\right)\right] \\
&= \int_0^{\infty} \mathbb{P}\left(\max_{j=1,\ldots,N_{L,H}(\delta)} \sum_{i=1}^{n} \widetilde{g}_j\left(X_i, X_i'\right) \geq t\right) dt \\
&\leq 6\left(n \log N_{L,H}(\delta)\right)^{1/2} + N_{L,H}(\delta) \int_{6\left(n \log N_{L,H}(\delta)\right)^{1/2}}^{\infty} \max_{j=1,\ldots,N_{L,H}(\delta)} \mathbb{P}\left(\sum_{i=1}^{n} \widetilde{g}_j\left(X_i, X_i'\right) \geq t\right) dt \\
&\leq 6\left(n \log N_{L,H}(\delta)\right)^{1/2} + N_{L,H}(\delta) \int_{6\left(n \log N_{L,H}(\delta)\right)^{1/2}}^{\infty} \exp\left(-\frac{3t\left(\log N_{L,H}(\delta)\right)^{1/2}}{16 n^{1/2}}\right) dt \\
&\leq 6\left(n \log N_{L,H}(\delta)\right)^{1/2} + \frac{16 n^{1/2}}{3\left(\log N_{L,H}(\delta)\right)^{1/2}} \exp\left(-6(nlogN_{L,W})^{\frac{1}{2}}\right) \\
&\leq 6\left(n \log N_{L,H}(\delta)\right)^{1/2} + \frac{16 n^{1/2}}{3\left(\log N_{L,H}(\delta)\right)^{1/2}},
\end{aligned}
$$

when $N_{L,H} > 3$ by assumption. Then, the first statement 42 is proved.

Similarly, we can prove the second statment 43 associated with $t \geq 6\left(n \log N_{L,H}(\delta)\right)^{1/2}$ by deriving maximum inequality as following

$$
\begin{aligned}
&\mathbb{E}\left[\left(\max_{j=1,\ldots,N_{L,H}(\delta)} \sum_{i=1}^{n} \widetilde{g}_j\left(X_i, X_i'\right)\right)^2\right] \\
&= \int_0^{\infty} \mathbb{P}\left(\max_{j=1,\ldots,N_{L,H}(\delta)} \sum_{i=1}^{n} \widetilde{g}_j\left(X_i, X_i'\right) \geq t^{1/2}\right) dt \\
&\leq 36 n \log N_{L,H}(\delta) + N_{L,H}(\delta) \int_{36 n \log N_{L,H}(\delta)}^{\infty} \max_{j=1,\ldots,N_{L,H}(\delta)} \mathbb{P}\left(\sum_{i=1}^{n} \widetilde{g}_j\left(X_i, X_i'\right) \geq t^{1/2}\right) \mathrm{d}t \\
&\leq 36 n \log N_{L,H}(\delta) + N_{L,H}(\delta) \int_{36 n \log N_{L,H}(\delta)}^{\infty} \exp\left(-\frac{3 t^{1/2}\left(\log N_{L,H}(\delta)\right)^{1/2}}{16 n^{1/2}}\right) dt \\
&\leq 36 n \log N_{L,H}(\delta) + \left(64 n + \frac{392 n}{9 \log N_{L,H}(\delta)}\right) \exp\left(-\frac{9 \log N_{L,H}(\delta)}{8}\right) \\
&\leq 36 n \log N_{L,H}(\delta) + 64 n + \frac{392 n}{9 \log N_{L,H}(\delta)}.
\end{aligned}
$$

□

Lemma 5. *Consider the set of deep neural networks as (7) with the depth L, the width H, and the upper bound B. For any $\delta > 0$ and sufficiently large n, we have*

$$\log N\left(\delta, \mathcal{F}(L,H), \|\cdot\|_{L^2(P_n)}\right) \leq CH^2L^2\log(HL)\log(Bn/\delta).$$

Proof. Let D be the VC-dimension of $\mathcal{F}$, and $W\,(\leq H^2L)$ be a number of parameters in $\mathcal{F}$. By Theorem 3 and 7 in Bartlett et al. (2019), we bound the VC-dimension as $D = O(HL\log(W)) \leq O\left(H^2L^2\log(HL)\right)$. Using this inequality and Theorem 12.2 in Anthony et al. (1999), we have

$$\log N\left(\delta, \mathcal{F}(L,H), \|\cdot\|_{L^2(P_n)}\right) \leq D\log\left(\frac{enB}{\delta D}\right) \leq CH^2L^2\log(HL)\log(Bn/\delta).$$

for $n = \Omega\left(H^2L^2\log(HL)\right)$. □

Lemma 6. (*Theorem 1.1 in Luet al.* (2021)). *For any $N, M \in \mathbb{N}^+$, $\mathcal{F}(L,H)$ is a set of functions with width $H = C_d(N+2)\log_2(8N)$ and depth $L = C_s(M+2)\log_2(4M) + 2d$ such that*

$$\inf_{f\in\mathcal{F}} \sup_{f^*\in C_1^s\left([0,1]^d\right)} \|f - f^*\|_{L^\infty} \leq C_{d,s}N^{-2s/d}M^{-2s/d}.$$

Appendix B: Proofs of Sup-Norm Convergence of Nuisance Functions

B.1 Squared Loss Function Setting

B.1.1 Proofs of Theorem 2

Theorem 2. (General Structure Feedforward Architecture) Suppose that Assumption 4, 5, and 6 hold for some $\beta > 0$. Let $\boldsymbol{b}^*(\boldsymbol{x}, \boldsymbol{v})$ be regression function 1 and $\widehat{\boldsymbol{b}}$ be corrected adversarial estimator 11 under the function class $\mathcal{F}(L_h, H_h, L_c, H_c)$ of deep neural networks. Then

$$\begin{aligned}
&\mathbb{E}\left[\|m - m^*\|_{L^\infty}^2 + \sum_{k=1}^{d_\theta} C_k \|\theta_k - \theta_k^*\|_{L^\infty}^2\right] \\
&\leq C_{P_V, P_X, p, d_v, d_x, h, B}\left(\frac{H^2 L^2 \log(HL)\log(n)}{n} + \mathbb{E}\left[\|\Xi\|_{L^\infty}^2\right] + (LH)^{-4\beta/d_v} + (LH)^{-4\beta/d_x}\right)
\end{aligned}$$

for every $n \geq \bar{n}$ with some $\bar{n} \in \mathbb{N}$.

Proof. Since we set $L_h = L_c = L$ and $H_h = H_c = H$, we further simplify proposition 2 into

$$\begin{aligned}
&\mathbb{E}\left[\|m - m^*\|_{L^\infty}^2 + \sum_{k=1}^{d_\theta} C_k \|\theta_k - \theta_k^*\|_{L^\infty}^2\right] \\
&\leq C_{P_V, P_X, p, d_v, d_x, h, B}\left(\frac{H^2 L^2 \log(HL)\log(n)}{n} + \mathbb{E}\left[\|\Xi\|_{L^\infty}^2\right] + \mathbb{E}\left[\|\Xi\|_{L^\infty}\right]\Phi_{L_c, H_c}\right. \\
&\left.+\mathbb{E}\left[\|\Xi\|_{L^\infty}\right]\Phi_{L_h, H_h} + \Phi_{L_c, H_c}^2 + \Phi_{L_c, H_c}\Phi_{L_h, H_h} + \Phi_{L_h, H_h}^2\right)
\end{aligned}$$

In the process of preparation, Lemma 6 establishes the following bound:

$$\Phi_{L,W} \leq C_{d,\beta}(LH)^{-2\beta/d}.$$

By leveraging this bound for $\Phi_{L,W}$, we employ Proposition 1 to deduce the following inequality:

$$\begin{aligned}
&\mathbb{E}\left[\|m - m^*\|_{L^\infty}^2 + \sum_{k=1}^{d_\theta} C_k \|\theta_k - \theta_k^*\|_{L^\infty}^2\right] \\
&\leq C_{P_V, P_X, p, d_v, d_x, h, B}\left(\frac{H^2 L^2 \log(HL)\log(n)}{n} + \mathbb{E}\left[\|\Xi\|_{L^\infty}^2\right] + \mathbb{E}\left[\|\Xi\|_{L^\infty}\right](LH)^{-2\beta/d_v}\right. \\
&\left.+\mathbb{E}\left[\|\Xi\|_{L^\infty}\right](LH)^{-2\beta/d_x} + (LH)^{-4\beta/d_v} + (LH)^{-2\beta/d_v}(LH)^{-2\beta/d_h} + (LH)^{-4\beta/d_x}\right)
\end{aligned}$$

In addition, it is feasible to represent the following expressions:

$$\frac{1}{2}(LH)^{-4\beta/d_i} + \mathbb{E}\left[\|\Xi\|_{L^\infty}\right](LH)^{-2\beta/d_i} + \frac{1}{2}\mathbb{E}\left[\|\Xi\|_{L^\infty}^2\right] \leq \frac{1}{2}\left\{(LH)^{-2\beta/d_i} + \mathbb{E}\left[\|\Xi\|_{L^\infty}^2\right]^{1/2}\right\}^2 \quad \text{for } i \in \{c, h\}$$

and

$$\frac{1}{2}(LH)^{-4\beta/d_v} + (LH)^{-2\beta/d_v}(LH)^{-2\beta/d_h} + \frac{1}{2}(LH)^{-4\beta/d_x} \leq \frac{1}{2}\left\{(LH)^{-2\beta/d_v} + (LH)^{-2\beta/d_x}\right\}^2$$

Therefore, we can derive the statement as following

$$\begin{aligned}
&\mathbb{E}\left[\|m - m^*\|_{L^\infty}^2 + \sum_{k=1}^{d_\theta} C_k \|\theta_k - \theta_k^*\|_{L^\infty}^2\right] \\
&\leq C_{P_V,P_X,p,d_v,d_x,h,B}\left(\frac{H^2L^2\log(HL)\log(n)}{n} + \mathbb{E}\left[\|\Xi\|_{L^\infty}^2\right] + (LH)^{-4\beta/d_v} + (LH)^{-4\beta/d_x}\right)
\end{aligned}$$

□

B.1.2 Proofs of Corollary 2

Corollary 2. Suppose that Assumption 4, 5, and 6 hold for some $\beta > 0$. Let $\boldsymbol{b}^*(\boldsymbol{x}, \boldsymbol{v})$ be regression function 1 and $\widehat{\boldsymbol{b}}$ be corrected adversarial estimator 11 under the function class $\mathcal{F}(L_h, H_h, L_c, H_c)$ of deep neural networks with depth $L_h = L_c = L$ and width $H_h = L_h = H$ setted as $HL \asymp n^{(d_x+d_v)/(4\beta+2d_x+2d_v)}$. Assume that $\zeta_n^2 = O\left(n^{-2\beta/(2\beta+d_v+d_x)}\log^{\beta^*} n\right)$ for some $\beta^* > 0$. Then

$$\mathbb{E}\left[\|m - m^*\|_{L^\infty}^2 + \sum_{k=1}^{d_\theta} C_k \|\theta_k - \theta_k^*\|_{L^\infty}^2\right] = O\left(n^{-2\beta/(2\beta+d_x+d_v)}\log^{2\vee\beta^*} n\right)$$

and for $\boldsymbol{b}^*(\boldsymbol{X}, V) = (m^*(V), \theta^*(\boldsymbol{X})')'$, we have

$$\mathbb{E}\left[\|\boldsymbol{b} - \boldsymbol{b}^*\|_{L^\infty}^2\right] = O\left(n^{-2\beta/(2\beta+d_x+d_v)}\log^{2\vee\beta^*} n\right).$$

Proof. Based on theorem 2, we place $HL = Cn^{(d_x+d_v)/(4\beta+2d_x+2d_v)}$ and $\zeta_n^2 = C'n^{-2\beta/(2\beta+d_v+d_x)}\log^{\beta^*} n$, for $n > \bar{n}'$ with some $\bar{n}' \in \mathbb{N}$, into the right side of inequality 31. Then we obtain

$$\begin{aligned}
&\mathbb{E}\left[\|m - m^*\|_{L^\infty}^2 + \sum_{k=1}^{d_\theta} C_k \|\theta_k - \theta_k^*\|_{L^\infty}^2\right] \\
&\leq C_{P_V,P_X,p,d_v,d_x,h,B}\left(n^{-2\beta/(2\beta+d_x+d_v)}\log^2 n + n^{-2\beta/(2\beta+d_x+d_v)}\log^{\beta*} n + n^{-\frac{2\beta}{2\beta+d_x+d_v}\frac{d_x+d_v}{d_v}} + n^{-\frac{2\beta}{2\beta+d_x+d_v}\frac{d_x+d_v}{d_x}}\right) \\
&= O\left(n^{-2\beta/(2\beta+d_x+d_v)}\log^{2\vee\beta^*} n\right)
\end{aligned}$$

□

B.1.3 Proofs of Proposition 2

Proposition 2. *Suppose that Assumption 4 and 5 hold. Let $\boldsymbol{b}^*(\boldsymbol{x}, \boldsymbol{v})$ be regression function (1) and $\hat{\boldsymbol{b}}$ be the corrected adversarial estimator(11) restricted to the function class $\mathcal{F}(L_h, H_h, L_c, H_c)$ of deep neural networks. Then*

$$\begin{aligned}
&\mathbb{E}\left[\|m - m^*\|_{L^\infty}^2 + \sum_{k=1}^{d_\theta} C_k \|\theta_k - \theta_k^*\|_{L^\infty}^2\right] \\
&\leq C_{P_V, P_X, p, d_v, d_x, h, B}\left(\frac{(H_h^2 + H_c^2)\, L^2 \log(H_h^2 L + H_c^2 L)\log(n)}{n} + \mathbb{E}\left[\|\Xi\|_{L^\infty}^2\right] + \mathbb{E}\left[\|\Xi\|_{L^\infty}\right]\Phi_{L_c, H_c}\right. \\
&\left.+\mathbb{E}\left[\|\Xi\|_{L^\infty}\right]\Phi_{L_h, H_h} + \Phi_{L_c, H_c}^2 + \Phi_{L_c, H_c}\Phi_{L_h, H_h} + \Phi_{L_h, H_h}^2\right)
\end{aligned}$$

Proof. Let's start with Lemma 7 to bound the sup-norm as

$$\|m - m^*\|_{L^\infty}^2 + \sum_{k=1}^{d_\theta} C_k \|\theta_k - \theta_k^*\|_{L^\infty}^2 \leq C_{P_V, P_X, p, d_v, d_x, h}\left(\|m - m^*\|_{P_V, \Delta}^2 + \sum_{k=1}^{d_\theta} C_k \|\theta_k - \theta_k^*\|_{P_X, \Delta}^2\right), \tag{45}$$

Then with lemma 8, we have the following inequality

$$\begin{aligned}
\mathbb{E}\left[\|\widehat{m} - m^*\|_{P_V, \Delta}^2 + \sum_{k=1}^{d_\theta} C_k \left\|\widehat{\theta}_k - \theta_k^*\right\|_{P_X, \Delta}^2\right] &\leq \mathbb{E}\left[\|\widehat{m} - m^*\|_{n, \Delta}^2 + \sum_{k=1}^{d_\theta} C_k \left\|\widehat{\theta}_k - \theta_k^*\right\|_{n, \Delta}^2\right] \\
&+ \left(1 + \sum_{k=1}^{d_\theta} C_k\right)\left(\frac{(18B^2 + 6B)\log N_{L_h, H_h, L_c, H_c}(\delta) + 6B + 32B^2}{n} + \frac{22B^2}{n^2} + 20B\delta + 2\delta^2\right),
\end{aligned}$$

Based on lemma 8, we further apply lemma 9 to obtain further inequality

$$
\begin{aligned}
&\mathbb{E}\left[\|\widehat{m}-m^*\|_{P_V,\Delta}^2+\sum_{k=1}^{d_\theta}C_k\left\|\widehat{\theta}_k-\theta_k^*\right\|_{P_X,\Delta}^2\right]\\
&\leq \mathbb{E}\left[\|\widehat{m}-m^*\|_{n,\Delta}^2+\sum_{k=1}^{d_\theta}C_k\left\|\widehat{\theta}_k-\theta_k^*\right\|_{n,\Delta}^2\right]\\
&\quad+\left(1+\sum_{k=1}^{d_\theta}C_k\right)\left(\frac{(18B^2+6B)\log N_{L_h,H_h,L_c,H_c}(\delta)+6B+32B^2}{n}+\frac{22B^2}{n^2}+20B\delta+2\delta^2\right)\\
&\leq 2\left(1+d_\theta\right)\left(\mathbb{E}\left[\|\widehat{m}-m^*\|_{L^\infty}^2+\sum_{k=1}^{d_\theta}C_k\left\|\widehat{\theta}_k-\theta_k^*\right\|_{L^\infty}^2\right]^{1/2}+\delta\left(1+\sum_{k=1}^{d_\theta}C_k\right)\right)\left(\frac{\log N_{L_h,H_h,L_c,H_c}(\delta)+\mathbb{E}\left[\|\Xi\|_{L^\infty}^2\right]}{n}\right)^{1/2}\\
&\quad+2\left(1+\sum_{k=1}^{d_\theta}C_k\right)\mathbb{E}\left[\|\Xi\|_{L^\infty}^2\right]^{1/2}\delta+\Phi_{L_c,H_c}^2+2\mathbb{E}\left[\|\Xi\|_{L^\infty}^2\right]+2\mathbb{E}\left[\|\Xi\|_{L^\infty}\right]\Phi_{L_c,H_c}\\
&\quad+2\sum_{j=1}^{d_\theta}C_j\mathbb{E}\left[\|\Xi\|_{L^\infty}\right]\Phi_{L_h,H_h}+2\sum_{j=1}^{d_\theta}C_j\Phi_{L_c,H_c}\Phi_{L_h,H_h}+\left(\sum_{j=1}^{d_\theta}C_j+\sum_{j=1}^{d_\theta}\sum_{k=1}^{d_\theta}C_{jk}\right)\Phi_{L_h,H_h}^2\\
&\quad+\left(1+\sum_{k=1}^{d_\theta}C_k\right)\left(\frac{(18B^2+6B)\log N_{L_h,H_h,L_c,H_c}(\delta)+6B+32B^2}{n}+\frac{22B^2}{n^2}+20B\delta+2\delta^2\right)
\end{aligned}
$$

.

Then we combine the inequality with expectation of lemma 7 inequality and derive following

$$
\begin{aligned}
&\mathbb{E}\left[\|m-m^*\|_{L^\infty}^2+\sum_{k=1}^{d_\theta}C_k\|\theta_k-\theta_k^*\|_{L^\infty}^2\right]\\
&\leq C_{P_V,P_X,p,d_v,d_x,h}\left(\mathbb{E}\left[\|\widehat{m}-m^*\|_{L^\infty}^2+\sum_{k=1}^{d_\theta}C_k\left\|\widehat{\theta}_k-\theta_k^*\right\|_{L^\infty}^2\right]^{1/2}+\delta\right)\left(\frac{\log N_{L_h,H_h,L_c,H_c}(\delta)+\mathbb{E}\left[\|\Xi\|_{L^\infty}^2\right]}{n}\right)^{1/2}\\
&+C_{P_V,P_X,p,d_v,d_x,h}\left(\mathbb{E}\left[\|\Xi\|_{L^\infty}^2\right]^{1/2}\delta+\Phi_{L_c,H_c}^2+\mathbb{E}\left[\|\Xi\|_{L^\infty}^2\right]+\mathbb{E}\left[\|\Xi\|_{L^\infty}\right]\Phi_{L_c,H_c}+\mathbb{E}\left[\|\Xi\|_{L^\infty}\right]\Phi_{L_h,H_h}\right.\\
&\left.+\Phi_{L_c,H_c}\Phi_{L_h,H_h}+\Phi_{L_h,H_h}^2+\frac{(B^2+B)\log N_{L_h,H_h,L_c,H_c}(\delta)+B+B^2}{n}+\frac{B^2}{n^2}+B\delta+\delta^2\right)
\end{aligned}
$$

.

With a quadratic inequality $z^2\leq az+b$ for $a,b\geq 0$ and $z\in\mathbb{R}$, we can derive $z^2\leq a^2+2b$. Then, we can further obtain the following inequality with $z=\mathbb{E}\left[\|\widehat{m}-m^*\|_{L^\infty}^2+\sum_{k=1}^{d_\theta}C_k\left\|\widehat{\theta}_k-\theta_k^*\right\|_{L^\infty}^2\right]^{1/2}$

$$\mathbb{E}\left[\|m-m^*\|_{L^\infty}^2+\sum_{k=1}^{d_\theta}C_k\|\theta_k-\theta_k^*\|_{L^\infty}^2\right]$$
$$\leq C_{P_V,P_X,p,d_v,d_x,h,B}\left(\frac{\log N_{L_h,H_h,L_c,H_c}(\delta)}{n}+\left(1+\frac{1}{n}\right)\mathbb{E}\left[\|\Xi\|_{L^\infty}^2\right]+\mathbb{E}\left[\|\Xi\|_{L^\infty}\right]\Phi_{L_c,H_c}+\mathbb{E}\left[\|\Xi\|_{L^\infty}\right]\Phi_{L_h,H_h}\right.$$
$$\left.+\Phi_{L_c,H_c}^2+\Phi_{L_c,H_c}\Phi_{L_h,H_h}+\Phi_{L_h,H_h}^2+\frac{1}{n^2}+\delta+\delta^2+\mathbb{E}\left[\|\Xi\|_{L^\infty}^2\right]^{1/2}\delta+\delta\left(\frac{\log N_{L_h,H_h,L_c,H_c}(\delta)+\mathbb{E}\left[\|\Xi\|_{L^\infty}^2\right]}{n}\right)^{1/2}\right)$$

In addition, we can calculate the bound of metric entropy by lemma 10 with $\delta=1/n$, which is

$$\log N_{L_h,H_h,L_c,H_c}(1/n)=\log\sup_{Q_n}N\left(\delta,\mathcal{F}(L_h,H_h,L_c,H_c),\|\cdot\|_{L^2(Q_n)}\right)\leq C\left(H_h^2+H_c^2\right)L^2\log(H_h^2L+H_c^2L)\log(Bn/\delta).$$

By replace the bound of metric entropy into the inequality 33, we may obtain

$$\mathbb{E}\left[\|m-m^*\|_{L^\infty}^2+\sum_{k=1}^{d_\theta}C_k\|\theta_k-\theta_k^*\|_{L^\infty}^2\right]$$
$$\leq C_{P_V,P_X,p,d_v,d_x,h,B}\left(\frac{(H_h^2+H_c^2)L^2\log(H_h^2L+H_c^2L)\log(n)}{n}+\mathbb{E}\left[\|\Xi\|_{L^\infty}^2\right]+\mathbb{E}\left[\|\Xi\|_{L^\infty}\right]\Phi_{L_c,H_c}\right.$$
$$\left.+\mathbb{E}\left[\|\Xi\|_{L^\infty}\right]\Phi_{L_h,H_h}+\Phi_{L_c,H_c}^2+\Phi_{L_c,H_c}\Phi_{L_h,H_h}+\Phi_{L_h,H_h}^2\right)$$

□

Supportive results for squared loss function setting

Lemma 7. *Suppose $P_{\boldsymbol{X}},P_{\boldsymbol{V}}$ satisfies Assumption 1 and $m^*(\boldsymbol{V}),\{\theta_k^*(\boldsymbol{X})\}_{k=1}^{d_\theta}$ are continuous. For any bounded and continuous $m:[0,1]^{d_v}\to\mathbb{R}$ and $\theta_k:[0,1]^{d_x}\to\mathbb{R}$ for $k=1,\ldots,d_\theta$, we have*

$$\|m-m^*\|_{L^\infty}^2\leq\left(C_{P_V,p,d_v}h^{d_v}\right)^{-1}\|m-m^*\|_{P_V,\Delta}^2,\tag{46}$$

and

$$\|\theta_k-\theta_k^*\|_{L^\infty}^2\leq\left(C_{P_X,p,d_x}h^{d_x}\right)^{-1}\|\theta_k-\theta_k^*\|_{P_X,\Delta}^2\quad\text{for}\ \ k=1,\ldots,d_\theta.\tag{47}$$

Proof. This lemma follows directly from Lemma 1 and is automatically obtained. □

Lemma 8. *Suppose $m^*(\boldsymbol{V}), \{\theta_k^*(\boldsymbol{X})\}_{k=1}^{d_\theta}$ are continuous with $\|m^*\|_{L^\infty} \le B$, $\|\theta_k^*\|_{L^\infty} \le B$ and function class $\mathcal{F}$ with all continuous functions. Then given any $\delta > 0$, we have*

$$\mathbb{E}\left[\|\widehat{m} - m^*\|_{P_V,\Delta}^2\right] \le \mathbb{E}\left[\|\widehat{m} - m^*\|_{n,\Delta}^2\right] + \frac{(18B^2+6B)\log N_{L_h,H_h,L_c,H_c}(\delta) + 6B + 32B^2}{n} + \frac{22B^2}{n^2} + 20B\delta + 2\delta^2,$$

and for $k = 1, \ldots, d_\theta$,

$$\mathbb{E}\left[\left\|\widehat{\theta}_k - \theta_k^*\right\|_{P_X,\Delta}^2\right] \le \mathbb{E}\left[\left\|\widehat{\theta}_k - \theta_k^*\right\|_{n,\Delta}^2\right] + \frac{(18B^2+6B)\log N_{L_h,H_h,L_c,H_c}(\delta) + 6B + 32B^2}{n} + \frac{22B^2}{n^2} + 20B\delta + 2\delta^2.$$

Proof. This lemma is also a direct result of Lemma 2 for each nuisance function. □

Lemma 9. *Consider the setting in Theorem 2. Then, for any $\delta \in (0, 1]$, we have*

$$\begin{aligned}
&\mathbb{E}\left[\|m^*(\boldsymbol{v}) - \widehat{m}(\boldsymbol{v})\|_{n,\Delta}^2 + \sum_{j=1}^{d_\theta} C_j \left\|\theta_j^*(\boldsymbol{x}) - \widehat{\theta}_j(\boldsymbol{x})\right\|_{n,\Delta}^2\right] \\
&\le 2\left(1 + d_\theta\right)\left(\mathbb{E}\left[\|\widehat{m} - m^*\|_{L^\infty}^2 + \sum_{k=1}^{d_\theta} C_k \left\|\widehat{\theta}_k - \theta_k^*\right\|_{L^\infty}^2\right]^{1/2} + \delta\left(1 + \sum_{k=1}^{d_\theta} C_k\right)\right)\left(\frac{\log N_{L_h,H_h,L_c,H_c}(\delta) + \mathbb{E}\left[\|\Xi\|_{L^\infty}^2\right]}{n}\right)^{1/2} \\
&+ 2\left(1 + \sum_{k=1}^{d_\theta} C_k\right)\mathbb{E}\left[\|\Xi\|_{L^\infty}^2\right]^{1/2}\delta + \Phi_{L_c,H_c}^2 + 2\mathbb{E}\left[\|\Xi\|_{L^\infty}^2\right] + 2\mathbb{E}\left[\|\Xi\|_{L^\infty}\right]\Phi_{L_c,H_c} \\
&+ 2\sum_{j=1}^{d_\theta} C_j \mathbb{E}\left[\|\Xi\|_{L^\infty}\right]\Phi_{L_h,H_h} + 2\sum_{j=1}^{d_\theta} C_j \Phi_{L_c,H_c}\Phi_{L_h,H_h} + \left(\sum_{j=1}^{d_\theta} C_j + \sum_{j=1}^{d_\theta}\sum_{k=1}^{d_\theta} C_{jk}\right)\Phi_{L_h,H_h}^2. \qquad (48)
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}\left[\|m^*(\boldsymbol{v}) - \widehat{m}(\boldsymbol{v})\|_{n,\Delta}^2 + \sum_{j=1}^{d_\theta} C_j \left\|\theta_j^*(\boldsymbol{x}) - \widehat{\theta}_j(\boldsymbol{x})\right\|_{n,\Delta}^2\right] \\
&\le 2\left(1 + d_\theta\right)\left(\mathbb{E}\left[\|\widehat{m} - m^*\|_{L^\infty}^2 + \sum_{k=1}^{d_\theta} C_k \left\|\widehat{\theta}_k - \theta_k^*\right\|_{L^\infty}^2\right]^{1/2} + \delta\left(1 + \sum_{k=1}^{d_\theta} C_k\right)\right)\left(\frac{\log N_{L_h,H_h,L_c,H_c}(\delta) + \mathbb{E}\left[\|\Xi\|_{L^\infty}^2\right]}{n}\right)^{1/2} \\
&+ 2\left(1 + \sum_{k=1}^{d_\theta} C_k\right)\mathbb{E}\left[\|\Xi\|_{L^\infty}^2\right]^{1/2}\delta + 2\mathbb{E}\left[\|\Xi\|_{L^\infty}^2\right] + 2\left(1 + \sum_{j=1}^{d_\theta} C_j\right)\mathbb{E}\left[\|\Xi\|_{L^\infty}\right]\Phi_{L,H} \\
&+ (1 + 3\sum_{j=1}^{d_\theta} C_j + \sum_{j=1}^{d_\theta}\sum_{k=1}^{d_\theta} C_{jk})\Phi_{L,H}^2. \qquad (49)
\end{aligned}$$

Proof. By the definition of the minimization problem, $L_n(\widehat{f}) \leq L_n(f)$ holds for any $f \in \mathcal{F}(L_h, H_h, L_c, H_c)$, hence we have the following basic inequality as

$$\frac{1}{n}\sum_{i=1}^{n} \max_{(\boldsymbol{x},\boldsymbol{v})\in\Delta_h^p(X_i,V_i)} \left(\widehat{Y}\left(\boldsymbol{x},\boldsymbol{v},\boldsymbol{t}_i\right) - \widehat{m}(\boldsymbol{v}) - \boldsymbol{t}_i'\widehat{\boldsymbol{\theta}}(\boldsymbol{x})\right)^2 \leq \frac{1}{n}\sum_{i=1}^{n} \max_{(\boldsymbol{x},\boldsymbol{v})\in\Delta_h^p(X_i,V_i)} \left(\widehat{Y}\left(\boldsymbol{x},\boldsymbol{v},\boldsymbol{t}_i\right) - m(\boldsymbol{v}) - \boldsymbol{t}_i'\boldsymbol{\theta}(\boldsymbol{x})\right)^2,$$

which can be rewritten as

$$\frac{1}{n}\sum_{i=1}^{n} \max_{(\boldsymbol{x},\boldsymbol{v})\in\Delta_h^p(X_i,V_i)} \left(m^*(\boldsymbol{v}) + \boldsymbol{t}_i'\boldsymbol{\theta}^*(\boldsymbol{x}) + \Xi\left(\boldsymbol{x},\boldsymbol{v},\boldsymbol{t}_i\right) - \widehat{m}(\boldsymbol{v}) - \boldsymbol{t}_i'\widehat{\boldsymbol{\theta}}(\boldsymbol{x})\right)^2 \tag{50}$$

$$\leq \frac{1}{n}\sum_{i=1}^{n} \max_{(\boldsymbol{x},\boldsymbol{v})\in\Delta_h^p(X_i,V_i)} \left(m^*(\boldsymbol{v}) + \boldsymbol{t}_i'\boldsymbol{\theta}^*(\boldsymbol{x}) + \Xi\left(\boldsymbol{x},\boldsymbol{v},\boldsymbol{t}_i\right) - m(\boldsymbol{v}) - \boldsymbol{t}_i'\boldsymbol{\theta}(\boldsymbol{x})\right)^2. \tag{51}$$

We bound the both-hand side of inequality. The left-hand side (LHS) of (50) is lower bounded as

$$\begin{aligned}
&\text{LHS of (50)} = \frac{1}{n}\sum_{i=1}^{n} \max_{(x,v)\in\Delta_h^p(X_i,V_i)} \left\{ \left(m^*(\boldsymbol{v}) - \widehat{m}(\boldsymbol{v})\right)^2 + \sum_{j=1}^{d_\theta} t_{ij}^2 \left(\theta_j^*(\boldsymbol{x}) - \widehat{\theta}_j(\boldsymbol{x})\right)^2 + \Xi\left(\boldsymbol{x},\boldsymbol{v},\boldsymbol{t}_i\right)^2 \right. \\
&+2\Xi\left(\boldsymbol{x},\boldsymbol{v},\boldsymbol{t}_i\right)\left(m^*(\boldsymbol{v}) - \widehat{m}(\boldsymbol{v})\right) + 2\sum_{j=1}^{d_\theta} t_{ij}\Xi\left(\boldsymbol{x},\boldsymbol{v},\boldsymbol{t}_i\right)\left(\theta_j^*(\boldsymbol{x}) - \widehat{\theta}_j(\boldsymbol{x})\right) \\
&\left.+2\sum_{j=1}^{d_\theta} t_{ij}\left(m^*(\boldsymbol{v}) - \widehat{m}(\boldsymbol{v})\right)\left(\theta_j^*(\boldsymbol{x}) - \widehat{\theta}_j(\boldsymbol{x})\right) + \sum_{j\neq k} t_{ij}t_{ik}\left(\theta_j^*(\boldsymbol{x}) - \widehat{\theta}_j(\boldsymbol{x})\right)\left(\theta_k^*(\boldsymbol{x}) - \widehat{\theta}_k(\boldsymbol{x})\right) \right\} \\
&\geq \left\|m^*(\boldsymbol{v}) - \widehat{m}(\boldsymbol{v})\right\|_{n,\Delta}^2 + \sum_{j=1}^{d_\theta} C_j \left\|\theta_j^*(\boldsymbol{x}) - \widehat{\theta}_j(\boldsymbol{x})\right\|_{n,\Delta}^2 - \left\|\Xi\right\|_{n,\Delta}^2 - \\
&\frac{2}{n}\sum_{i=1}^{n} \max_{(x',v')\in\Delta_h^p(X_i,V_i)} \left|\Xi\left(\boldsymbol{x},\boldsymbol{v},\boldsymbol{t}_i\right)\left(m^*(\boldsymbol{v}) - \widehat{m}(\boldsymbol{v})\right)\right| - \sum_{j=1}^{d_\theta} C_j \frac{2}{n}\sum_{i=1}^{n} \max_{(x',v')\in\Delta_h^p(X_i,V_i)} \left|\Xi\left(\boldsymbol{x},\boldsymbol{v},\boldsymbol{t}_i\right)\left(\theta_j^*(\boldsymbol{v}) - \theta_j(\boldsymbol{v})\right)\right|,
\end{aligned} \tag{52}$$

by bounded assumption for $\boldsymbol{T}$ and orthogonal products. Similarly, we bound the right-hand side of (51) as

$$\text{RHS of (51)} = \frac{1}{n}\sum_{i=1}^{n} \max_{(x,v)\in\Delta_h^p(X_i,V_i)} \Bigg\{ (m^*(\boldsymbol{v}) - m(\boldsymbol{v}))^2 + \sum_{j=1}^{d_\theta} t_{ij}^2 \left(\theta_j^*(\boldsymbol{x}) - \theta_j(\boldsymbol{x})\right)^2 + \Xi\left(\boldsymbol{x}, \boldsymbol{v}, \boldsymbol{t}_i\right)^2 +$$

$$2\Xi\left(\boldsymbol{x}, \boldsymbol{v}, \boldsymbol{t}_i\right)\left(m^*(\boldsymbol{v}) - m(\boldsymbol{v})\right) + 2\sum_{j=1}^{d_\theta} t_{ij}\Xi\left(\boldsymbol{x}, \boldsymbol{v}, \boldsymbol{t}_i\right)\left(\theta_j^*(\boldsymbol{x}) - \theta_j(\boldsymbol{x})\right) +$$

$$2\sum_{j=1}^{d_\theta} t_{ij}\left(m^*(\boldsymbol{v}) - m(\boldsymbol{v})\right)\left(\theta_j^*(\boldsymbol{x}) - \theta_j(\boldsymbol{x})\right) + \sum_{j\neq k} t_{ij}t_{ik}\left(\theta_j^*(\boldsymbol{x}) - \theta_j(\boldsymbol{x})\right)\left(\theta_k^*(\boldsymbol{x}) - \theta_k(\boldsymbol{x})\right) \Bigg\}$$

$$\leq \left\|m^*(\boldsymbol{v}) - m(\boldsymbol{v})\right\|_{n,\Delta}^2 + \sum_{j=1}^{d_\theta} C_j \left\|\theta_j^*(\boldsymbol{x}) - \theta_j(\boldsymbol{x})\right\|_{n,\Delta}^2 + \|\Xi\|_{n,\Delta}^2 +$$

$$\frac{2}{n}\sum_{i=1}^{n} \max_{(x',v')\in\Delta_h^p(X_i,V_i)} \left|\Xi\left(\boldsymbol{x}, \boldsymbol{v}, \boldsymbol{t}_i\right)\left(m^*(\boldsymbol{v}) - m(\boldsymbol{v})\right)\right| + \sum_{j=1}^{d_\theta} C_j \frac{2}{n}\sum_{i=1}^{n} \max_{(x',v')\in\Delta_h^p(X_i,V_i)} \left|\Xi\left(\boldsymbol{x}, \boldsymbol{v}, \boldsymbol{t}_i\right)\left(\theta_j^*(\boldsymbol{v}) - \theta_j(\boldsymbol{v})\right)\right|$$

$$+ \sum_{j=1}^{d_\theta} C_j \frac{2}{n}\sum_{i=1}^{n} \max_{(x',v')\in\Delta_h^p(X_i,V_i)} \left|\left(m^*(\boldsymbol{v}) - m(\boldsymbol{v})\right)\left(\theta_j^*(\boldsymbol{v}) - \theta_j(\boldsymbol{v})\right)\right| +$$

$$\sum_{j\neq k} C_{jk} \frac{2}{n}\sum_{i=1}^{n} \max_{(x',v')\in\Delta_h^p(X_i,V_i)} \left|\left(\theta_j^*(\boldsymbol{v}) - \theta_j(\boldsymbol{v})\right)\left(\theta_k^*(\boldsymbol{v}) - \theta_k(\boldsymbol{v})\right)\right|$$

$$\leq \Phi_{L_m,H_m}^2 + \sum_{j=1}^{d_\theta} C_j \Phi_{L_{\theta_j},H_{\theta_j}}^2 + \|\Xi\|_{L^\infty}^2 + 2\|\Xi\|_{L^\infty}\Phi_{L_m,H_m} + 2\sum_{j=1}^{d_\theta} C_j \|\Xi\|_{L^\infty}\Phi_{L_{\theta_j},H_{\theta_j}} +$$

$$2\sum_{j=1}^{d_\theta} C_j \Phi_{L_m,H_m}\Phi_{L_{\theta_j},H_{\theta_j}} + \sum_{j\neq k} C_{jk}\Phi_{L_{\theta_j},H_{\theta_j}}\Phi_{L_{\theta_k},H_{\theta_k}}. \tag{53}$$

Combining (53) and (52) with (50,51), we obtain

$$
\begin{aligned}
\|m^*(\boldsymbol{v}) - \widehat{m}(\boldsymbol{v})\|_{n,\Delta}^2 + \sum_{j=1}^{d_\theta} C_j \left\|\theta_j^*(\boldsymbol{x}) - \widehat{\theta}_j(\boldsymbol{x})\right\|_{n,\Delta}^2 \leq & \underbrace{\frac{2}{n}\sum_{i=1}^{n} \max_{(x',v')\in\Delta_h^p(X_i,V_i)} \left|\Xi\left(\boldsymbol{x},\boldsymbol{v},\boldsymbol{t}_i\right)\left(m^*(\boldsymbol{v}) - \widehat{m}(\boldsymbol{v})\right)\right|}_{=:S_m} \\
& + \sum_{j=1}^{d_\theta} C_j \underbrace{\frac{2}{n}\sum_{i=1}^{n} \max_{(x',v')\in\Delta_h^p(X_i,V_i)} \left|\Xi\left(\boldsymbol{x},\boldsymbol{v},\boldsymbol{t}_i\right)\left(\theta_j^*(\boldsymbol{v}) - \widehat{\theta}_j(\boldsymbol{v})\right)\right|}_{=:S_{\theta_j}} \\
& + \Phi_{L_m,H_m}^2 + \sum_{j=1}^{d_\theta} C_j \Phi_{L_{\theta_j},H_{\theta_j}}^2 + 2\|\Xi\|_{L^\infty}^2 + 2\|\Xi\|_{L^\infty}\Phi_{L_m,H_m} \\
& + 2\sum_{j=1}^{d_\theta} C_j \|\Xi\|_{L^\infty}\Phi_{L_{\theta_j},H_{\theta_j}} + 2\sum_{j=1}^{d_\theta} C_j \Phi_{L_m,H_m}\Phi_{L_{\theta_j},H_{\theta_j}} \\
& + \sum_{j\neq k}^{d_\theta} C_{jk}\Phi_{L_{\theta_j},H_{\theta_j}}\Phi_{L_{\theta_k},H_{\theta_k}} \\
\leq & S_m + \sum_{j=1}^{d_\theta} C_j S_{\theta_j} + \Phi_{L_m,H_m}^2 + \sum_{j=1}^{d_\theta} C_j \Phi_{L_{\theta_j},H_{\theta_j}}^2 + 2\|\Xi\|_{L^\infty}^2 \\
& + 2\|\Xi\|_{L^\infty}\Phi_{L_m,H_m} + 2\sum_{j=1}^{d_\theta} C_j \|\Xi\|_{L^\infty}\Phi_{L_{\theta_j},H_{\theta_j}} \\
& + 2\sum_{j=1}^{d_\theta} C_j \Phi_{L_m,H_m}\Phi_{L_{\theta_j},H_{\theta_j}} + \sum_{j\neq k}^{d_\theta} C_{jk}\Phi_{L_{\theta_j},H_{\theta_j}}\Phi_{L_{\theta_k},H_{\theta_k}}
\end{aligned} \tag{54}
$$

by the definition of $\Phi_{L,W}$ in (12). We will bound an expectation the terms. Note that the expectations of the terms are guaranteed to exist, by the boundedness of f^* and $\widehat{f}, f \in \mathcal{F}(L,W)$, and $\widehat{Y}$.

We bound $\mathbb{E}\left[S_m + \sum_{j=1}^{d_\theta} C_j S_{\theta_j}\right]$. We define the nearest element of the covering set to nuisance functions $\left\{\widehat{m}, \{\widehat{\theta}_i\}_{i=1}^{d_\theta}\right\}$ that is, we define $\widehat{j}_m := \operatorname{argmin}_{j'=1,\ldots,N} \sup_{Q_n} \|m_{j'} - \widehat{m}\|_{L^2(Q_n)}$ and $\widehat{j}_{\theta_i} := \operatorname{argmin}_{j'=1,\ldots,N} \sup_{Q_n} \left\|\theta_{i,j'} - \widehat{\theta}_i\right\|_{L^2(Q_n)}$ for $i = 1,\ldots,d_\theta$. Then, we bound $\mathbb{E}\left[S_m + \sum_{j=1}^{d_\theta} C_j S_{\theta_j}\right]$ as

$$
\mathbb{E}\left[S_m + \sum_{j=1}^{d_\theta} C_j S_{\theta_j}\right] = \mathbb{E}\left[\frac{2}{n}\sum_{i=1}^{n} \max_{(x',v')\in\Delta_h^p(X_i,V_i)} \left|\Xi\left(\boldsymbol{x},\boldsymbol{v},\boldsymbol{t}_i\right)\left(m^*(\boldsymbol{v}) - m_{\widehat{j}}(\boldsymbol{v}) + m_{\widehat{j}}(\boldsymbol{v}) - \widehat{m}(\boldsymbol{v})\right)\right|\right]
$$

$$
\begin{aligned}
&+\sum_{k=1}^{d_\theta} C_k \mathbb{E}\left[\frac{2}{n}\sum_{i=1}^{n}\max_{(x,v)\in\Delta_h^p(X_i,V_i)}\left|\Xi\left(\boldsymbol{x},\boldsymbol{v},\boldsymbol{t}_i\right)\left(\theta_k^*(\boldsymbol{v})-\theta_{k,\widehat{j}}(\boldsymbol{v})+\theta_{k,\widehat{j}}(\boldsymbol{v})-\widehat{\theta}_k(\boldsymbol{v})\right)\right|\right]\\
&\leq \mathbb{E}\left[\frac{2}{n}\sum_{i=1}^{n}\max_{(x',v')\in\Delta_h^p(X_i,V_i)}\left|\Xi\left(\boldsymbol{x},\boldsymbol{v},\boldsymbol{t}_i\right)\left(m^*(\boldsymbol{v})-m_{\widehat{j}}(\boldsymbol{v})\right)\right|\right]\\
&+\mathbb{E}\left[\frac{2}{n}\sum_{i=1}^{n}\max_{(x',v')\in\Delta_h^p(X_i,V_i)}\left|\Xi\left(\boldsymbol{x},\boldsymbol{v},\boldsymbol{t}_i\right)\left(m_{\widehat{j}}(\boldsymbol{v})-\widehat{m}(\boldsymbol{v})\right)\right|\right]\\
&+\sum_{k=1}^{d_\theta} C_k \mathbb{E}\left[\frac{2}{n}\sum_{i=1}^{n}\max_{(x,v)\in\Delta_h^p(X_i,V_i)}\left|\Xi\left(\boldsymbol{x},\boldsymbol{v},\boldsymbol{t}_i\right)\left(\theta_k^*(\boldsymbol{v})-\theta_{k,\widehat{j}}(\boldsymbol{v})\right)\right|\right]\\
&+\sum_{k=1}^{d_\theta} C_k \mathbb{E}\left[\frac{2}{n}\sum_{i=1}^{n}\max_{(x,v)\in\Delta_h^p(X_i,V_i)}\left|\Xi\left(\boldsymbol{x},\boldsymbol{v},\boldsymbol{t}_i\right)\left(\theta_{k,\widehat{j}}(\boldsymbol{v})-\widehat{\theta}_k(\boldsymbol{v})\right)\right|\right]\\
&\leq \mathbb{E}\left[\frac{2}{n}\sum_{i=1}^{n}\max_{(x',v')\in\Delta_h^p(X_i,V_i)}\left|\Xi\left(\boldsymbol{x},\boldsymbol{v},\boldsymbol{t}_i\right)\left(m^*(\boldsymbol{v})-m_{\widehat{j}}(\boldsymbol{v})\right)\right|\frac{\left\|\widehat{m}-m^*\right\|_{L^\infty}+\delta}{\left\|m_{\widehat{j}}-m^*\right\|_{L^\infty}}\right]\\
&+2\mathbb{E}\left[\sup_{Q_n}\|\Xi\|_{L^2(Q_n)}^2\right]^{1/2}\mathbb{E}\left[\sup_{Q_n}\left\|m_{\widehat{j}}-\widehat{m}\right\|_{L^2(Q_n)}^2\right]^{1/2}\\
&+\sum_{k=1}^{d_\theta} C_k \mathbb{E}\left[\frac{2}{n}\sum_{i=1}^{n}\max_{(x,v)\in\Delta_h^p(X_i,V_i)}\left|\Xi\left(\boldsymbol{x},\boldsymbol{v},\boldsymbol{t}_i\right)\left(\theta_k^*(\boldsymbol{v})-\theta_{k,\widehat{j}}(\boldsymbol{v})\right)\right|\frac{\left\|\widehat{\theta}_k-\theta_k^*\right\|_{L^\infty}+\delta}{\left\|\theta_{k,\widehat{j}}-\theta_k^*\right\|_{L^\infty}}\right]\\
&+2\sum_{k=1}^{d_\theta} C_k \mathbb{E}\left[\sup_{Q_n}\|\Xi\|_{L^2(Q_n)}^2\right]^{1/2}\mathbb{E}\left[\sup_{Q_n}\left\|\theta_{k,\widehat{j}}-\widehat{\theta}_k\right\|_{L^2(Q_n)}^2\right]^{1/2}\\
&\leq \mathbb{E}\left[\left(\|\widehat{m}-m^*\|_{L^\infty}+\delta\right)\underbrace{\frac{2}{n}\sum_{i=1}^{n}\frac{\max_{(x',v')\in\Delta_h^p(X_i,V_i)}\left|\Xi\left(\boldsymbol{x},\boldsymbol{v},\boldsymbol{t}_i\right)\left(m^*(\boldsymbol{v})-m_{\widehat{j}}(\boldsymbol{v})\right)\right|}{\left\|m_{\widehat{j}}-m^*\right\|_{L^\infty}}}_{Z_{\widehat{j}m}}\right]\\
&+\mathbb{E}\left[\left(\sum_{k=1}^{d_\theta} C_k\left\|\widehat{\theta}_k-\theta_k^*\right\|_{L^\infty}+\delta\sum_{k=1}^{d_\theta} C_k\right)\sum_{k=1}^{d_\theta}\underbrace{\frac{2}{n}\sum_{i=1}^{n}\frac{\max_{(x,v)\in\Delta_h^p(X_i,V_i)}\left|\Xi\left(\boldsymbol{x},\boldsymbol{v},\boldsymbol{t}_i\right)\left(\theta_k^*(\boldsymbol{v})-\theta_{k,\widehat{j}}(\boldsymbol{v})\right)\right|}{\left\|\theta_{k,\widehat{j}}-\theta_k^*\right\|_{L^\infty}}}_{=:Z_{\widehat{j}\theta_k}}\right]\\
&+2\left(1+\sum_{k=1}^{d_\theta} C_k\right)\mathbb{E}\left[\|\Xi\|_{L^\infty}^2\right]^{1/2}\delta
\end{aligned}
$$

$$\leq \mathbb{E}\left[\left(\left\|\widehat{m}-m^*\right\|_{L^\infty}+\sum_{k=1}^{d_\theta}C_k\left\|\widehat{\theta}_k-\theta_k^*\right\|_{L^\infty}+\delta\left(1+\sum_{k=1}^{d_\theta}C_k\right)\right)\left(\underbrace{Z_{\widehat{j}_m}+\sum_{k=1}^{d_\theta}Z_{\widehat{j}_{\theta_k}}}_{Z_{\widehat{j}}}\right)\right] \tag{55}$$

$$+2\left(1+\sum_{k=1}^{d_\theta}C_k\right)\mathbb{E}\left[\|\Xi\|_{L^\infty}^2\right]^{1/2}\delta$$

We can further bound the first term (55) of inequality using the Cauchy-Schwarz inequality and obtain

$$\begin{aligned}
&\mathbb{E}\left[\left(\left\|\widehat{m}-m^*\right\|_{L^\infty}+\sum_{k=1}^{d_\theta}C_k\left\|\widehat{\theta}_k-\theta_k^*\right\|_{L^\infty}+\delta\left(1+\sum_{k=1}^{d_\theta}C_k\right)\right)Z_{\widehat{j}}\right]\\
&\leq \mathbb{E}\left[\left(\left\|\widehat{m}-m^*\right\|_{L^\infty}+\sum_{k=1}^{d_\theta}C_k\left\|\widehat{\theta}_k-\theta_k^*\right\|_{L^\infty}+\delta\left(1+\sum_{k=1}^{d_\theta}C_k\right)\right)^2\right]^{1/2}\mathbb{E}\left[Z_{\widehat{j}}^2\right]^{1/2}\\
&\leq \left(\mathbb{E}\left[\left\|\widehat{m}-m^*\right\|_{L^\infty}^2+\sum_{k=1}^{d_\theta}C_k\left\|\widehat{\theta}_k-\theta_k^*\right\|_{L^\infty}^2\right]^{1/2}+\delta\left(1+\sum_{k=1}^{d_\theta}C_k\right)\right)\mathbb{E}\left[\max_{j=1,\ldots,N_{L_h,H_h,L_c,H_c}(\delta)}Z_j^2\right]^{1/2}\\
&\leq 2\left(1+d_\theta\right)\left(\mathbb{E}\left[\left\|\widehat{m}-m^*\right\|_{L^\infty}^2+\sum_{k=1}^{d_\theta}C_k\left\|\widehat{\theta}_k-\theta_k^*\right\|_{L^\infty}^2\right]^{1/2}+\delta\left(1+\sum_{k=1}^{d_\theta}C_k\right)\right)\left(\frac{\log N_{L_h,H_h,L_c,H_c}(\delta)+\mathbb{E}\left[\|\Xi\|_{L^\infty}^2\right]}{n}\right)^{1/2},
\end{aligned}$$

where the last inequality is guaranteed by the bounded random process of Z_j^2,

$$Z_j\leq 2\left(1+d_\theta\right)\|\Xi\|_{L^\infty}\text{ and }Z_j^2\leq 4\left(1+d_\theta\right)^2\|\Xi\|_{L^\infty}^2,$$

for any $j=1,\ldots,N$, and the application of the maximal inequality in theorem 3.1.10 of Giné and Nickl (2021). Therefore, we have

$$\begin{aligned}
&\mathbb{E}\left[S_m+\sum_{j=1}^{d_\theta}C_jS_{\theta_j}\right]\\
&\leq 2\left(1+d_\theta\right)\left(\mathbb{E}\left[\left\|\widehat{m}-m^*\right\|_{L^\infty}^2+\sum_{k=1}^{d_\theta}C_k\left\|\widehat{\theta}_k-\theta_k^*\right\|_{L^\infty}^2\right]^{1/2}+\delta\left(1+\sum_{k=1}^{d_\theta}C_k\right)\right)\left(\frac{\log N_{L_h,H_h,L_c,H_c}(\delta)+\mathbb{E}\left[\|\Xi\|_{L^\infty}^2\right]}{n}\right)^{1/2}\\
&+2\left(1+\sum_{k=1}^{d_\theta}C_k\right)\mathbb{E}\left[\|\Xi\|_{L^\infty}^2\right]^{1/2}\delta
\end{aligned} \tag{56}$$

By inserting bound 56 into 54, we can calculate the expectation of inequality (54) to establish the lemma:

$$\mathbb{E}\left[\|m^*(\boldsymbol{v})-\widehat{m}(\boldsymbol{v})\|_{n,\Delta}^2+\sum_{j=1}^{d_\theta}C_j\left\|\theta_j^*(\boldsymbol{x})-\widehat{\theta}_j(\boldsymbol{x})\right\|_{n,\Delta}^2\right]$$
$$\leq 2\left(1+d_\theta\right)\left(\mathbb{E}\left[\|\widehat{m}-m^*\|_{L^\infty}^2+\sum_{k=1}^{d_\theta}C_k\left\|\widehat{\theta}_k-\theta_k^*\right\|_{L^\infty}^2\right]^{1/2}+\delta\left(1+\sum_{k=1}^{d_\theta}C_k\right)\right)\left(\frac{\log N_{L_h,H_h,L_c,H_c}(\delta)+\mathbb{E}\left[\|\Xi\|_{L^\infty}^2\right]}{n}\right)^{1/2}$$
$$+2\left(1+\sum_{k=1}^{d_\theta}C_k\right)\mathbb{E}\left[\|\Xi\|_{L^\infty}^2\right]^{1/2}\delta+\Phi_{L_c,H_c}^2+\sum_{j=1}^{d_\theta}C_j\Phi_{L_h,H_h}^2+2\mathbb{E}\left[\|\Xi\|_{L^\infty}^2\right]+2\mathbb{E}\left[\|\Xi\|_{L^\infty}\right]\Phi_{L_c,H_c}$$
$$+2\sum_{j=1}^{d_\theta}C_j\mathbb{E}\left[\|\Xi\|_{L^\infty}\right]\Phi_{L_h,H_h}+2\sum_{j=1}^{d_\theta}C_j\Phi_{L_c,H_c}\Phi_{L_h,H_h}+\sum_{j=1}^{d_\theta}\sum_{k=1}^{d_\theta}C_{jk}\Phi_{L_h,H_h}^2. \tag{57}$$

Note that our deep neural network architecture $\mathcal{F}_{L_h,H_h,L_c,H_c}$ includes upper and lower modules as figure 5. The upper module is configured with depth L_h and width H_h, aiming at estimating parameter functions, $\{\theta_k^*(\boldsymbol{X})\}_{k=1}^{d_\theta}$, while the lower module focuses on estimating control function, $m^*(\boldsymbol{V})$, with configuration determined by depth L_c and width H_c. The first statement is proved.

For the second statement, we set $L_h = L_c = L$ and $H_h = H_c = H$ to further simplify inequality 57 into

$$\mathbb{E}\left[\|m^*(\boldsymbol{v})-\widehat{m}(\boldsymbol{v})\|_{n,\Delta}^2+\sum_{j=1}^{d_\theta}C_j\left\|\theta_j^*(\boldsymbol{x})-\widehat{\theta}_j(\boldsymbol{x})\right\|_{n,\Delta}^2\right]$$
$$\leq 2\left(1+d_\theta\right)\left(\mathbb{E}\left[\|\widehat{m}-m^*\|_{L^\infty}^2+\sum_{k=1}^{d_\theta}C_k\left\|\widehat{\theta}_k-\theta_k^*\right\|_{L^\infty}^2\right]^{1/2}+\delta\left(1+\sum_{k=1}^{d_\theta}C_k\right)\right)\left(\frac{\log N_{L_h,H_h,L_c,H_c}(\delta)+\mathbb{E}\left[\|\Xi\|_{L^\infty}^2\right]}{n}\right)^{1/2}$$
$$+2\left(1+\sum_{k=1}^{d_\theta}C_k\right)\mathbb{E}\left[\|\Xi\|_{L^\infty}^2\right]^{1/2}\delta+2\mathbb{E}\left[\|\Xi\|_{L^\infty}^2\right]+2\left(1+\sum_{j=1}^{d_\theta}C_j\right)\mathbb{E}\left[\|\Xi\|_{L^\infty}\right]\Phi_{L,H}+(1+3\sum_{j=1}^{d_\theta}C_j+\sum_{j=1}^{d_\theta}\sum_{k=1}^{d_\theta}C_{jk})\Phi_{L,H}^2.$$

□

Lemma 10. *Consider the set of deep neural networks, $\mathcal{F}_{L_h,H_h,L_c,H_c}$ as shown in Figure **??**, composed of upper and lower modules with an upper bound B. The structure of the upper neural network is defined by the depth L_h and width H_h, while the lower one is determined by the depth L_c and width H_c. For any $\delta > 0$ and sufficiently large n, we have*

$$\log N\left(\delta,\mathcal{F}(L_h,H_h,L_c,H_c),\|\cdot\|_{L^2(P_n)}\right)\leq C\left(H_h^2+H_c^2\right)L^2\log(H_h^2L+H_c^2L)\log(Bn/\delta).$$

Proof. The proof procedure is same with lemma 5. Similarly, let D be the VC-dimension of $\mathcal{F}_{L_h,H_h,L_c,H_c}$, and $W\,(\leq H_h^2 L_h + H_c^2 L_c)$ be the number of parameters in $\mathcal{F}$. By Theorems 3 and 7 in Bartlett et al. (2019), we bound the VC-dimension as $D = O(WL\log(W)) \leq O\left(\left(H_h^2 + H_c^2\right) L^2 \log(H_h^2 L + H_c^2 L)\right)$, where $L = \max\{L_h, L_c\}$. Utilizing this inequality and Theorem 12.2 in Anthony et al. (1999), we obtain

$$\log N\left(\delta, \mathcal{F}(L_h, H_h, L_c, H_c), \|\cdot\|_{L^2(P_n)}\right) \leq D \log\left(\frac{enB}{\delta D}\right) \leq C\left(H_h^2 + H_c^2\right) L^2 \log(H_h^2 L + H_c^2 L) \log(Bn/\delta).$$

for $n = \Omega\left(\left(H_h^2 + H_c^2\right) L^2 \log(H_h^2 L + H_c^2 L)\right)$. □

B.2 General Loss Function Setting

B.2.1 Proofs of Theorem 3

Theorem 3. Consider the regression model (1) and the adversarial estimator $\tilde{\boldsymbol{b}}$ in (12) with the function class by deep neural networks with a tuple (L_h, H_h, L_c, H_c) and $h \in (0, 1)$. Suppose Assumption 4 and 5 for $\beta > 0$, Assumption 6 holds with $\zeta_n^2 = O\left(n^{-2\beta/(2\beta+d)} \log^{\beta^*} n\right)$ for some $\beta^* > 0$ and $\widehat{Y}$ is independent of $\{(\boldsymbol{X}_i, \boldsymbol{T}_i, \boldsymbol{Z}_i, Y_i)_{i=1}^n\}$, and Assumption 7 holds with $q \in [1, \infty)$. Then, we have the following as $n \to \infty$:

$$\mathbb{E}\left[\left\|\tilde{\boldsymbol{b}} - \boldsymbol{b}^*\right\|_{L^\infty}^2\right] =\leq C_{P_X,B,p,d,\ell,q,V} h^{-2d/q} \left\{n^{-\beta/(q(\beta+d))} \log^{4/q} n + n^{-2\beta/(2\beta+d)} \log^{\beta^*} n\right\}. \quad (58)$$

Proof. By Lemma 11 and Lemma 12, we have

$$
\begin{aligned}
\mathbb{E}\left[\left\|\boldsymbol{b}^* - \widetilde{\boldsymbol{b}}\right\|_{L^\infty}^2\right] &\leq C_{P_X,p,d,\ell,q} h^{-2d/q}\left(\mathbb{E}\left[\left(\widetilde{R}(\widetilde{\boldsymbol{b}}) - \widetilde{R}\left(\boldsymbol{b}^*\right)\right)^{2/q}\right] + \mathbb{E}\left[\|\Xi\|_{L^\infty}^2\right]\right) \\
&\leq C_{P_X,B,p,d,\ell,q,} h^{-2d/q}\left\{\left(\frac{\log N_{L_h,H_h,L_c,H_c}(1/n)}{n^{1/2}}\right)^{2/q} + \Phi_{L_h,H_h}^{2/q} + \Phi_{L_c,H_c}^{2/q} + \zeta_n^2\right\} \\
&\leq C_{P_X,B,p,d,\ell,q,V} h^{-2d/q}\left\{\left(\frac{(H_h^2+H_c^2)L^2\log((H_h^2+H_c^2)L)\log n}{n^{1/2}}\right)^{2/q} + \Phi_{L_h,H_h}^{2/q} + \Phi_{L_c,H_c}^{2/q} + \zeta_n^2\right\} \\
&\leq C_{P_X,B,p,d,\ell,q,V} h^{-2d/q}\left\{\left(\frac{(H^2L^2\log(HL)\log n}{n^{1/2}}\right)^{2/q} + \Phi_{L_h,H_h}^{2/q} + \Phi_{L_c,H_c}^{2/q} + \zeta_n^2\right\} \\
&\leq C_{P_X,B,p,d,\ell,q,V} h^{-2d/q}\left\{\left(\frac{(H^2L^2\log(HL)\log n}{n^{1/2}}\right)^{2/q} + (LH)^{-4\beta/q} + \zeta_n^2\right\} \\
&\leq C_{P_X,B,p,d,\ell,q,V} h^{-2d/q}\left\{\left(\frac{(n^{d/(2\beta+2d)}\log^2 n}{n^{1/2}}\right)^{2/q} + n^{-d\beta/(\beta+d)q} + \zeta_n^2\right\} \\
&\leq C_{P_X,B,p,d,\ell,q,V} h^{-2d/q}\left\{n^{-\beta/(q(\beta+d))}\log^{4/q} n + n^{-2\beta/(2\beta+d)}\log^{\beta^*} n\right\}
\end{aligned}
$$

The third inequality follows Lemma 10. We set $H_h = H_c = H$ and $L_h = L_c = L$ and obtain the fourth inequality. Furthermore, by lemma 6 we derive the following bound $\Phi_{L,W} \leq C_{d,\beta}(LH)^{-2\beta/d}$ and obtain the fifth inequality. Lastly, we set $HL \asymp n^{d/(4\beta+4d)}$ to derive the penult inequality where $d = d_x + d_v + d_t$.

□

Lemma 11. *Consider the setting in Theorem 3. Then, we have for n such that $\log N(1/n) \geq 1$:*

$$
\mathbb{E}\left[\widetilde{R}(\widetilde{\boldsymbol{b}}) - \widetilde{R}\left(\boldsymbol{b}^*\right)\right] \leq \frac{C_{\ell,B}\left(\log N_{L_h,H_h,L_c,H_c}(1/n) + V^2\right)}{n^{1/2}} + C_\ell\left(\Phi_{L_h,H_h} + \Phi_{L_c,H_c} + \mathbb{E}\left[\left\|\Xi_n\right\|_{L^\infty}\right]\right).
$$

Proof. This proof is similar to Lemma 3.1 in Shen et al. (2021). A difference between Shen et al. (2021)and our result is that a property of the loss depends on f in our setting, so we have to modify it. Hence, we write down the proof.

We develop the proof in the following four steps: (i) a basic decomposition, (ii) bounding a variance, (iii) bounding a bias, and (iv) combining every bound.

Step 1: Basic decomposition. We define i.i.d. copies of the observations $D := \{(\boldsymbol{X}_i, \boldsymbol{T}_i, \boldsymbol{Z}_i, Y_i)_{i=1}^n\}$

as $D' := \{(\boldsymbol{X}_i', \boldsymbol{T}_i', \boldsymbol{Z}_i', Y_i')_{i=1}^n\}$, and also define an excess loss as

$$g(\boldsymbol{X}_i, \widehat{\boldsymbol{V}}_i, \boldsymbol{T}_i, \widehat{Y}, \boldsymbol{b}) = \sup_{(\boldsymbol{x},\boldsymbol{v},\boldsymbol{t})\in\Delta_h^p\left(X_i,\widehat{V}_i,T_i\right)} \ell(\widehat{Y}\left(\boldsymbol{x},\boldsymbol{v},\boldsymbol{t}\right), \boldsymbol{t}, \boldsymbol{b}(\boldsymbol{x},\boldsymbol{v})) - \sup_{(\boldsymbol{x},\boldsymbol{v},\boldsymbol{t})\in\Delta_h^p\left(\boldsymbol{X}_i,\widehat{\boldsymbol{V}}_i,T_i\right)} \ell(\widehat{Y}\left(\boldsymbol{x},\boldsymbol{v},\boldsymbol{t}\right), \boldsymbol{t}, \boldsymbol{b}^*(\boldsymbol{x},\boldsymbol{v})) \tag{59}$$

We further define empirical means of the excess loss as $G_n(\boldsymbol{b}) := n^{-1}\sum_{i=1}^n g(\boldsymbol{X}_i, \widehat{\boldsymbol{V}}_i, \boldsymbol{T}_i, \widehat{Y}, \boldsymbol{b})$ with the observations D, and $G_n'(\boldsymbol{b}) := n^{-1}\sum_{i=1}^n g(\boldsymbol{X}_i', \widehat{\boldsymbol{V}}_i', \boldsymbol{T}_i', \widehat{Y}', \boldsymbol{b})$ with the copies D'. Since $\widehat{\boldsymbol{b}}$ is independent to D', we can rewrite the expected risk as

$$\mathbb{E}\left[\widetilde{R}(\widetilde{\boldsymbol{b}}) - \widetilde{R}\left(\boldsymbol{b}^*\right)\right] = \mathbb{E}\left[\mathbb{E}_{D'}\left[G_n'(\widetilde{\boldsymbol{b}})\right]\right].$$

Since $\widetilde{\boldsymbol{b}}$ is the minimizer of the empirical risk and the loss is bounded, we obtain the following inequality of expectations:

$$\mathbb{E}\left[G_n(\widetilde{\boldsymbol{b}})\right] \leq \mathbb{E}\left[G_n(\boldsymbol{b})\right], \tag{60}$$

for any $\boldsymbol{b} \in \mathcal{F}(L, W)$. We set set $\bar{\boldsymbol{b}}$ such that $\left\|\bar{\boldsymbol{b}} - \boldsymbol{b}^*\right\|_{L^\infty} = \inf_{\boldsymbol{b}\in\mathcal{F}(L,W)} \left\|\boldsymbol{b} - \boldsymbol{b}^*\right\|_{L^\infty}$ Using this fact, we decompose the excess risk as

$$\mathbb{E}[\widetilde{R}(\widetilde{\boldsymbol{b}}) - \widetilde{R}(\boldsymbol{b}^*)] = \mathbb{E}\left[\mathbb{E}_{D'}\left[G_n'(\widetilde{\boldsymbol{b}})\right]\right] \leq \mathbb{E}[\underbrace{\mathbb{E}_{D'}\left[G_n'(\widetilde{\boldsymbol{b}}) - 2G_n(\widetilde{\boldsymbol{b}})\right]}_{=:\mathcal{A}}] + 2\mathbb{E}[\underbrace{G_n(\bar{\boldsymbol{b}})}_{=:\mathcal{B}}]. \tag{61}$$

The inequality follows (60).

Step 2: Bound the variance $\mathbb{E}[\mathcal{A}]$. We bound an expectation of the term $\mathcal{A}$. By the boundedness of both $\widetilde{\boldsymbol{b}}$ and $\widehat{Y}$ by Assumption 5 and 6, the expectation $\mathbb{E}[\mathcal{A}]$ exists.

We prepare additional notations. Fix $\delta \in (0, 1]$. We consider a covering set $\{\boldsymbol{b}_j\}_{j=1}^{N_{L_h,H_h,L_c,H_c}(\delta)} \subset \mathcal{F}$, then we pick $\boldsymbol{b}_j$ from the set such that $\sup_{Q_n}\left\|\boldsymbol{b}_j - \widetilde{\boldsymbol{b}}\right\|_{L^2(Q_n)} \leq \delta$. We define a term $\widetilde{g}(\boldsymbol{X}_i, \widehat{\boldsymbol{V}}_i, \boldsymbol{T}_i, \widehat{Y}, \widetilde{\boldsymbol{b}})$ by the following reform of $\mathcal{A}$ as

$$\mathcal{A} = \frac{1}{n}\sum_{i=1}^n \left\{\mathbb{E}_{D'}\left[G_n'(\widetilde{\boldsymbol{b}})\right] - 2g(\boldsymbol{X}_i, \widehat{\boldsymbol{V}}_i, \boldsymbol{T}_i, \widehat{Y}, \widetilde{\boldsymbol{b}})\right\} =: \frac{1}{n}\sum_{i=1}^n \widetilde{g}(\boldsymbol{X}_i, \widehat{\boldsymbol{V}}_i, \boldsymbol{T}_i, \widehat{Y}, \widetilde{\boldsymbol{b}})$$

which yields the following form

$$\begin{aligned}
\mathbb{E}[\mathcal{A}] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \widetilde{g}(\boldsymbol{X}_i, \widehat{\boldsymbol{V}}_i, \boldsymbol{T}_i, \widehat{Y}, \widetilde{\boldsymbol{b}})\right] \\
&= \underbrace{\mathbb{E}[\frac{1}{n}\sum_{i=1}^n \widetilde{g}(\boldsymbol{X}_i, \widehat{\boldsymbol{V}}_i, \boldsymbol{T}_i, \widehat{Y}, \boldsymbol{b}_j)]}_{:=\mathcal{A}_1} + \underbrace{\mathbb{E}[\frac{1}{n}\sum_{i=1}^n \widetilde{g}(\boldsymbol{X}_i, \widehat{\boldsymbol{V}}_i, \boldsymbol{T}_i, \widehat{Y}, \widetilde{\boldsymbol{b}}) - \frac{1}{n}\sum_{i=1}^n \widetilde{g}(\boldsymbol{X}_i, \widehat{\boldsymbol{V}}_i, \boldsymbol{T}_i, \widehat{Y}, \boldsymbol{b}_j)]}_{=:\mathcal{A}_2}.
\end{aligned} \tag{62}$$

We will bound both $\mathbb{E}[\mathcal{A}_1]$ and $\mathbb{E}[\mathcal{A}_2]$, separately. We bound the term $\mathbb{E}[\mathcal{A}_2]$. Since g in (59) is Lipschitz continuous in f with its Lipschitz constant C_ℓ, we easily see that $\widetilde{g}$ is Lipschitz continuous in f with its Lipschitz constant $6C_\ell$.

Thus, we obtain that

$$\mathbb{E}[\mathcal{A}_2] \leq \left| \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \widetilde{g}(\boldsymbol{X}_i, \widehat{\boldsymbol{V}}_i, \boldsymbol{T}_i, \widehat{Y}, \widetilde{\boldsymbol{b}})\right] - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \widetilde{g}(\boldsymbol{X}_i, \widehat{\boldsymbol{V}}_i, \boldsymbol{T}_i, \widehat{Y}, \boldsymbol{b}_j)\right] \right| \leq 6C_\ell\delta. \tag{63}$$

Next, we bound the term $\mathbb{E}[\mathcal{A}_1]$. Here, we need to consider a uniformly bounded function $y : [0,1]^d \to [-B, B]$ For each $\boldsymbol{b}_j$ in the covering set, $t > 0$, and the bounded function y, we use the Bernstein inequality to derive its stochastic upper bound. As preparation, we consider a threshold $B_n \geq 1$ depending on n and define a clipped preprocessing $\widehat{Y}_{B_n}(\cdot) := \max\left\{\min\left\{\widehat{Y}(\cdot), B_n\right\}, -B_n\right\}$. We firstly approximate $\mathbb{E}[\mathcal{A}_1]$ by the Lipschitz continuity of ℓ as

$$\mathbb{E}[\mathcal{A}_1] \leq \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \widetilde{g}\left(\boldsymbol{X}_i, \widehat{\boldsymbol{V}}_i, \boldsymbol{T}_i, \widehat{Y}_{B_n}, \boldsymbol{b}_j\right)\right] + 6C_\ell\mathbb{E}\left[\left\|\widehat{Y} - \widehat{Y}_{B_n}\right\|_{L^\infty}\right] \tag{64}$$

Since $\left|\widehat{Y}(\boldsymbol{x}, \boldsymbol{v}, \boldsymbol{t}) - \widehat{Y}_{B_n}(\boldsymbol{x}, \boldsymbol{v}, \boldsymbol{t})\right| = |\widehat{Y}(\boldsymbol{x}, \boldsymbol{v}, \boldsymbol{t})|\mathbf{1}\left\{|\widehat{Y}(\boldsymbol{x}, \boldsymbol{v}, \boldsymbol{t})| \geq B_n\right\}$ holds, we can bound the expectation in the second term of the right-hand side as

$$\begin{aligned}
\mathbb{E}\left[\left\|\widehat{Y} - \widehat{Y}_{B_n}\right\|_{L^\infty}\right] &= \mathbb{E}\left[\sup_{(x,v,t)\in[0,1]^{d_x+d_v+d_t}} |\widehat{Y}(\boldsymbol{x}, \boldsymbol{v}, \boldsymbol{t})|\mathbf{1}\left\{|\widehat{Y}(\boldsymbol{x}, \boldsymbol{v}, \boldsymbol{t})| \geq B_n\right\} |\right] \\
&\leq \mathbb{E}\left[\sup_{(x,v,t)\in[0,1]^{d_x+d_v+d_t}} |\widehat{Y}(\boldsymbol{x}, \boldsymbol{v}, \boldsymbol{t})| \sup_{x\in[0,1]^d} \mathbf{1}\left\{|\widehat{Y}(\boldsymbol{x}, \boldsymbol{v}, \boldsymbol{t})| \geq B_n\right\} |\right] \\
&\leq \mathbb{E}\left[\|\widehat{Y}\|_{L^\infty} 1\left\{\|\widehat{Y}\|_{L^\infty} \geq B_n\right\}\right] \\
&\leq \mathbb{E}\left[\|\widehat{Y}\|_{L^\infty}^2 / B_n\right].
\end{aligned}$$

The last inequality follows $\mathbf{1}\{x \geq 1\} \leq x$ for any $x \geq 0$. The existence of the second moment is guaranteed by Assumption 6. We put this result to 63 and obtain

$$\mathbb{E}[\mathcal{A}_1] \leq \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \widetilde{g}\left(\boldsymbol{X}_i, \widehat{\boldsymbol{V}}_i, \boldsymbol{T}_i, \widehat{Y}_{B_n}, \boldsymbol{b}_j\right)\right] + 6C_\ell\mathbb{E}\left[\|\widehat{Y}\|_{L^\infty}^2 / B_n\right]. \tag{65}$$

Then, we bound the first term $\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \widetilde{g}\left(\boldsymbol{X}_i, \widehat{\boldsymbol{V}}_i, \boldsymbol{T}_i, \widehat{Y}_{B_n}, \boldsymbol{b}_j\right)\right]$. Since we have $\left|g\left(\boldsymbol{X}_i, \widehat{\boldsymbol{V}}_i, \boldsymbol{T}_i, \widehat{Y}_{B_n}, \boldsymbol{b}_j\right)\right| \leq C_\ell\left(B_n \vee B\right)$ for any $x \in [0,1]^d$ and $\boldsymbol{b} : \|\boldsymbol{b}\|_{L^\infty} \leq B$, we obtain the following inequality with

fixed $\widehat{Y}_{B_n}$:

$$
\begin{aligned}
&\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\widetilde{g}\left(\boldsymbol{X}_i,\widehat{\boldsymbol{V}}_i,\boldsymbol{T}_i,\widehat{Y}_{B_n},\boldsymbol{b}_j\right)>t\right)\\
&=\mathbb{P}\left(\mathbb{E}_{D'}\left[g\left(\boldsymbol{X}_i',\widehat{\boldsymbol{V}}_i',\boldsymbol{T}_i',\widehat{Y}_{B_n},\boldsymbol{b}_j\right)\right]-\frac{2}{n}\sum_{i=1}^{n}g\left(\boldsymbol{X}_i,\widehat{\boldsymbol{V}}_i,\boldsymbol{T}_i,\widehat{Y}_{B_n},\boldsymbol{b}_j\right)>t\right)\\
&=\mathbb{P}\left(\mathbb{E}_{D'}\left[g\left(\boldsymbol{X}_i',\widehat{\boldsymbol{V}}_i',\boldsymbol{T}_i',\widehat{Y}_{B_n},\boldsymbol{b}_j\right)\right]-\frac{1}{n}\sum_{i=1}^{n}g\left(\boldsymbol{X}_i,\widehat{\boldsymbol{V}}_i,\boldsymbol{T}_i,\widehat{Y}_{B_n},\boldsymbol{b}_j\right)>\frac{t}{2}+\frac{1}{2}\mathbb{E}_{D'}\left[g\left(\boldsymbol{X}_i',\widehat{\boldsymbol{V}}_i',\boldsymbol{T}_i',\widehat{Y}_{B_n},\boldsymbol{b}_j\right)\right]\right)\\
&\leq\mathbb{P}\left(\mathbb{E}_{D'}\left[g\left(\boldsymbol{X}_i',\widehat{\boldsymbol{V}}_i',\boldsymbol{T}_i',\widehat{Y}_{B_n},\boldsymbol{b}_j\right)\right]-\frac{1}{n}\sum_{i=1}^{n}g\left(\boldsymbol{X}_i,\widehat{\boldsymbol{V}}_i,\boldsymbol{T}_i,\widehat{Y}_{B_n},\boldsymbol{b}_j\right)>\frac{t}{2}+\frac{1}{2}\frac{\mathrm{Var}_{D'}\left(g\left(\boldsymbol{X}_i,\widehat{\boldsymbol{V}}_i,\boldsymbol{T}_i,\widehat{Y}_{B_n},\boldsymbol{b}_j\right)\right)}{4C_\ell B_n}\right)\\
&\leq\exp\left(-\frac{n\left(t'\right)^2}{2\,\mathrm{Var}_{D'}\left(g\left(\boldsymbol{X}_i,\widehat{\boldsymbol{V}}_i,\boldsymbol{T}_i,\widehat{Y}_{B_n},\boldsymbol{b}_j\right)\right)+16C_\ell\left(B_n\vee B\right)t'/3}\right)\\
&\leq\exp\left(-\frac{n\left(t'\right)^2}{2t'C_\ell\left(B_n\vee B\right)+C_\ell\left(B_n\vee B\right)t'/3}\right)\\
&\leq\exp\left(-\frac{n\left(t'\right)^2}{16t'C_\ell\left(B_n\vee B\right)+16C_\ell\left(B_n\vee B\right)t'/3}\right)\\
&\leq\exp\left(-\frac{3nt'}{64C_\ell\left(B_n\vee B\right)}\right)\\
&\leq\exp\left(-\frac{3nt}{128C_\ell\left(B_n\vee B\right)}\right).
\end{aligned}
$$

The first and third inequalities follow $\mathrm{Var}_{D'}\left(g\left(\boldsymbol{X}_i,\widehat{\boldsymbol{V}}_i,\boldsymbol{T}_i,\widehat{Y}_{B_n},\boldsymbol{b}_j\right)\right)\leq 4C_\ell B_n\mathbb{E}_{D'}\left[g\left(\boldsymbol{X}_i,\widehat{\boldsymbol{V}}_i,\boldsymbol{T}_i,\widehat{Y}_{B_n},\boldsymbol{b}_j\right)\right]$, and the second and last inequalities follows a setting $t'=t/2+\mathrm{Var}_{D'}\left(g\left(\boldsymbol{X}_i,\widehat{\boldsymbol{V}}_i,\boldsymbol{T}_i,\widehat{Y}_{B_n},\boldsymbol{b}_j\right)\right)/\left(8C_\ell\left(B\vee B_n\right.\right.$ Using this inequality for a uniform bound in terms of the covering set $\{\boldsymbol{b}_j\}_{j=1}^{N_{L_h,H_h,L_c,H_c}(\delta)}$ and the independent random functions $\widehat{Y}$ and $\widehat{Y}_{B_n}$, we obtain

$$
\mathbb{P}\left(\max_{j=1,\ldots,N_{L_h,H_h,L_c,H_c}(\delta)}\frac{1}{n}\sum_{i=1}^{n}\widetilde{g}\left(\boldsymbol{X}_i,\widehat{\boldsymbol{V}}_i,\boldsymbol{T}_i,\widehat{Y}_{B_n},\boldsymbol{b}_j\right)>t\right)\leq N_{L_h,H_h,L_c,H_c}(\delta)\exp\left(-\frac{3nt}{128C_\ell\left(B_n\vee B\right)}\right)
$$

Then, by the maximal inequality (Corollary 2.2.8 in Van Der Vaart et al. (1996)), for any

$\eta > 0$, we have

$$\begin{aligned}
&\mathbb{E}\left[\max_{j=1,\ldots,N_{L_h,H_h,L_c,H_c}(\delta)} \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \widetilde{g}\left(\boldsymbol{X}_i, \widehat{\boldsymbol{V}}_i, \boldsymbol{T}_i, \widehat{Y}_{B_n}, \boldsymbol{b}_j\right)\right]\right] \\
&\leq \eta + \int_{\eta}^{\infty} \mathbb{P}\left(\max_{j=1,\ldots,N_{L_h,H_h,L_c,H_c}(\delta)} \frac{1}{n}\sum_{i=1}^{n} \widetilde{g}\left(\boldsymbol{X}_i, \widehat{\boldsymbol{V}}_i, \boldsymbol{T}_i, \widehat{Y}_{B_n}, \boldsymbol{b}_j\right) > t\right) \mathrm{d}t \\
&\leq \eta + \int_{\eta}^{\infty} N_{L_h,H_h,L_c,H_c}(\delta) \exp\left(-\frac{3nt}{128C_\ell\left(B_n \vee B\right)}\right) \mathrm{d}t \\
&\leq \eta + \frac{N_{L_h,H_h,L_c,H_c}(\delta)\left(128C_\ell\left(B_n \vee B\right)\right)}{3n} \exp\left(-\frac{3n\eta}{128C_\ell\left(B_n \vee B\right)}\right).
\end{aligned}$$

We set $B_n = n^{1/2}$, hence we have $(B \vee B_n) \leq C_B n^{1/2}$. Also, we set $\eta = \left(128C_{B,\ell}n^{1/2}\right)\log N_{L_h,H_h,L_c,H_c}(\delta)/(3n)$ and put this result into (65), we obtain

$$\begin{aligned}
\mathbb{E}\left[\mathcal{A}_1\right] &\leq \mathbb{E}\left[\max_{j=1,\ldots,N} \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \widetilde{g}\left(\boldsymbol{X}_i, \widehat{\boldsymbol{V}}_i, \boldsymbol{T}_i, \widehat{Y}_{B_n}, \boldsymbol{b}_j\right)\right]\right] + 6C_\ell \mathbb{E}\left[\|\widehat{Y}\|_{L^\infty}^2/B_n\right] \\
&\leq \frac{C_{\ell,B}\left(\log N_{L_h,H_h,L_c,H_c}(\delta) + \mathbb{E}\left[\|\widehat{Y}\|_{L^\infty}^2\right]\right)}{n^{1/2}}. \qquad (66)
\end{aligned}$$

Combining the inequalities (63) and (66) into (62) and set $\delta = 1/n$, we obtain

$$\mathbb{E}[\mathcal{A}] \leq \frac{(2C_\ell^2 B_2 + C_\ell B/3)\left(\log N_{L_h,H_h,L_c,H_c}(1/n) + \mathbb{E}\left[\|\widehat{Y}\|_{L^\infty}^2\right]\right)}{n^{1/2}}$$

Step 3: Bound the bias $\mathbb{E}[\mathcal{B}]$. By the Lipschitz continuity of the loss ℓ by Assumption 4, we have

$$\begin{aligned}
\mathbb{E}[\mathcal{B}] &= \mathbb{E}\left[n^{-1}\sum_{i=1}^{n} g(\boldsymbol{X}_i, \widehat{\boldsymbol{V}}_i, \boldsymbol{T}_i, \widehat{Y}, \bar{\boldsymbol{b}})\right] \\
&= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \sup_{(\boldsymbol{x},\boldsymbol{v},\boldsymbol{t})\in\Delta_h^p(X_i,V_i,T_i)} \ell(\widehat{Y}\left(\boldsymbol{x},\boldsymbol{v},\boldsymbol{t}\right), \boldsymbol{t}, \bar{\boldsymbol{b}}(\boldsymbol{x},\boldsymbol{v}))\right] \\
&\leq \mathbb{E}\left[\sup_{(x,v,t)\in[0,1]^{d_X+d_V+dT}} \ell(\widehat{Y}\left(\boldsymbol{x},\boldsymbol{v},\boldsymbol{t}\right), \boldsymbol{t}, \bar{\boldsymbol{b}}(\boldsymbol{x},\boldsymbol{v}))\right] \\
&\leq \mathbb{E}\left[\sup_{(x,v,t)\in[0,1]^{d_X+d_V+d_T}} C_\ell|\bar{\boldsymbol{\theta}}(\boldsymbol{x})'\boldsymbol{t} + \bar{m}(\boldsymbol{v}) - \widehat{Y}\left(\boldsymbol{x},\boldsymbol{v},\boldsymbol{t}\right)| + \ell(\widehat{Y}\left(\boldsymbol{x},\boldsymbol{v},\boldsymbol{t}\right), \widehat{Y}\left(\boldsymbol{x},\boldsymbol{v},\boldsymbol{t}\right))\right] \\
&\leq C_\ell \mathbb{E}\left[\|\boldsymbol{t}\bar{\boldsymbol{\theta}}(\boldsymbol{x})' + \bar{m}(\boldsymbol{v}) - \widehat{Y}\left(\boldsymbol{x},\boldsymbol{v},\boldsymbol{t}\right)\|_{L^\infty}\right] \\
&\leq C_\ell\left(\left\|\boldsymbol{t}\left(\bar{\boldsymbol{\theta}}(\boldsymbol{x}) - \boldsymbol{\theta}^*(\boldsymbol{x})\right)' + \bar{m}(\boldsymbol{v}) - m^*(\boldsymbol{v})\right\|_{L^\infty} + \mathbb{E}\left[\left\|\boldsymbol{t}\boldsymbol{\theta}^*(\boldsymbol{x})' + m^*(\boldsymbol{v}) - \widehat{Y}\right\|_{L^\infty}\right]\right) \\
&\leq C_\ell\left(\Phi_{L_h,H_h} + \Phi_{L_c,H_c} + \mathbb{E}\left[\|\Xi_n\|_{L^\infty}\right]\right). \qquad (67)
\end{aligned}$$

The last inequality holds by setting $\bar{\theta}_j$ such that $\left\|\bar{\theta}_j - \theta^*\right\|_{L^\infty} = \inf_{\theta_j \in \mathcal{F}(L_h, H_h)} \left\|\theta_j - \theta_j^*\right\|_{L^\infty}$ for $j \in \{1, 2, \cdots, d_\theta\}$ and $\bar{m}$ such that $\|\bar{m} - m^*\|_{L^\infty} = \inf_{\theta_j \in \mathcal{F}(L_c, H_c)} \|m - m^*\|_{L^\infty}$. Step 4: Combining the bounds. We combine the result in Step 3 and Step 4 into the decomposition (30), then obtain the statement. □

Lemma 12. *Consider the expected adversarial risk $\widetilde{R}(\cdot)$ with general losses as (??). Then, for the estimator $\widetilde{\boldsymbol{b}}$ as (12) and $q \in [1, \infty)$, we have*

$$\left\|\boldsymbol{b}^* - \widetilde{\boldsymbol{b}}\right\|_{L^\infty}^q \leq C_{P_X, p, d, \ell, q} h^{-d} \left(\widetilde{R}(\widetilde{\boldsymbol{b}}) - \widetilde{R}\left(\boldsymbol{b}^*\right) + \|\Xi\|_{L^\infty}^q \vee \|\Xi\|_{L^\infty}\right).$$

Proof. We develop a lower bound of $\widetilde{R}(\widetilde{\boldsymbol{b}}) - \widetilde{R}\left(\boldsymbol{b}^*\right)$ as

$$\begin{aligned}
\widetilde{R}(\widetilde{\boldsymbol{b}}) - \widetilde{R}\left(\boldsymbol{b}^*\right) &= \mathbb{E}\left[\sup_{(\boldsymbol{x}, \boldsymbol{v}, \boldsymbol{t}) \in \Delta_h^p(X_i, V_i, T_i)} \ell(\widehat{Y}\left(\boldsymbol{x}, \boldsymbol{v}, \boldsymbol{t}\right), \boldsymbol{t}, \widetilde{\boldsymbol{b}}(\boldsymbol{x}, \boldsymbol{v})) - \sup_{(\boldsymbol{x}, \boldsymbol{v}, \boldsymbol{t}) \in \Delta_h^p(X_i, V_i, T_i)} \ell(\widehat{Y}\left(\boldsymbol{x}, \boldsymbol{v}, \boldsymbol{t}\right), \boldsymbol{t}, \boldsymbol{b}^*(\boldsymbol{x}, \boldsymbol{v}))\right] \\
&\geq C_{P_X, p, d} h^{d_x + d_v + d_t} \sup_{(x, v, t) \in [0,1]^{d_x + d_v + d_t}} \left|\ell(\widehat{Y}\left(\boldsymbol{x}, \boldsymbol{v}, \boldsymbol{t}\right), \boldsymbol{t}, \widetilde{\boldsymbol{b}}(\boldsymbol{x}, \boldsymbol{v}))\right| - C_\ell \left\|\widehat{Y} - \boldsymbol{t}'\boldsymbol{\theta}^*(\boldsymbol{x}) - m^*(\boldsymbol{v})\right\|_{L^\infty} \\
&\geq C_{P_X, p, d, \ell} h^d \|\widehat{Y} - \boldsymbol{t}'\widetilde{\theta}(\boldsymbol{x}) - \widetilde{m}(\boldsymbol{v})\|_{L^\infty}^q - C_\ell \|\Xi\|_{L^\infty} \\
&\geq C_{P_X, p, d, \ell, q} h^d \left(\left\|\boldsymbol{t}'(\theta^*(\boldsymbol{x}) - \widetilde{\theta}(\boldsymbol{x})) + m^*(\boldsymbol{v}) - \widetilde{m}(\boldsymbol{v})\right\|_{L^\infty}^q - \|\Xi\|_{L^\infty}^q\right) - C_\ell \|\Xi\|_{L^\infty} \\
&\geq C_{P_X, p, d, \ell, q} h^d \left(\left\|\boldsymbol{b}^* - \widetilde{\boldsymbol{b}}\right\|_{L^\infty}^q - \|\Xi\|_{L^\infty}^q\right) - C_\ell \|\Xi\|_{L^\infty}.
\end{aligned}$$

Here, the first inequality follows Lemma 15 and the Lipschitz continuity of ℓ by Assumption 4, and the second inequality follows $(a+b)^q \leq C_q\left(a^q + b^q\right)$ for $q \in [1, \infty)$ and $a, b \geq 0$ and the last inequality followed by boundness of $\boldsymbol{t}$. Then we reorganize the inequality and achieve the inequality of lemma. □

Appendix C: Influence Function and Asmptotic Properties

C.1 Influence Function: Proof of Theorem 4

Let's calculate the influence function of the structural parameters.

The starting point is a parametric submodel, indexed by a parameter η. Distributions and other nonparametric objects are indexed by η, and thus we define $\boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}_\eta; \eta, \boldsymbol{v}_\eta)$ and $\boldsymbol{\mu}_0(\eta)$ as

$$\boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}_\eta; \eta, \boldsymbol{v}_\eta) = \arg\min_{\beta} \int \ell(\boldsymbol{w}, \boldsymbol{\beta}(\boldsymbol{x}, \boldsymbol{v}_\eta)) f_{\boldsymbol{w}}(\boldsymbol{w}; \eta) d\boldsymbol{w} \tag{68}$$

and

$$\boldsymbol{\mu}(\eta) = \int \boldsymbol{H}\left(\boldsymbol{x}, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}_\eta; \eta, \boldsymbol{v}_\eta); \boldsymbol{t}^*\right) f_{\boldsymbol{x}, \boldsymbol{v}_\eta}(\boldsymbol{x}, \boldsymbol{v}_\eta; \eta) d\boldsymbol{x} d\boldsymbol{v}_\eta \tag{69}$$

where $f_{\boldsymbol{w}}$ and $f_{\boldsymbol{x}}$ are the distributions of $\boldsymbol{w} = (\boldsymbol{y}', \boldsymbol{t}', \boldsymbol{x}')'$ and $\boldsymbol{x}$ respectively.

The true data generating process is obtained at $\eta = 0$. When evaluating at $\eta = 0$ we will often omit the dependence on η, such as $f_{\boldsymbol{x}}(\boldsymbol{x}; \eta) = f_{\boldsymbol{x}}(\boldsymbol{x}), \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}_\eta; \eta, \boldsymbol{v}_\eta) = \boldsymbol{b}_0(\boldsymbol{x})$, or $\mathbb{E}[\cdot]$ for expectations with respect to the true distribution.

According to Newey (1994)'s recipe of the pathwise derivative approach, the goal is to derive influence function $\boldsymbol{\psi}(\boldsymbol{w})$ such that

$$\left.\frac{\partial \boldsymbol{\mu}(\eta)}{\partial \eta}\right|_{\eta=0} = \mathbb{E}[\boldsymbol{\psi}(\boldsymbol{W}) S(\boldsymbol{W})], \tag{70}$$

for the (true) score $S(\boldsymbol{w}) = S(\boldsymbol{w}; \eta)|_{\eta=0}$.

Then we can have the following deduction

$$\begin{aligned}
&\left.\frac{\partial \boldsymbol{\mu}(\eta)}{\partial \eta}\right|_{\eta=0} = \frac{\partial}{\partial \eta}\left\{\int \boldsymbol{H}\left(\boldsymbol{x}, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}_\eta; \eta, \boldsymbol{v}_\eta); \boldsymbol{t}^*\right) f_{\boldsymbol{x}, \boldsymbol{v}_\eta}(\boldsymbol{x}, \boldsymbol{v}_\eta; \eta) d\boldsymbol{x} d\boldsymbol{v}_\eta\right\}\Big|_{\eta=0} \\
&= \int \boldsymbol{H}\left(\boldsymbol{x}, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}_{\eta=0}; \eta = 0, \boldsymbol{v}_{\eta=0}); \boldsymbol{t}^*\right) \left.\frac{\partial f_{\boldsymbol{x}, \boldsymbol{v}_\eta}(\boldsymbol{x}, \boldsymbol{v}_\eta; \eta)}{\partial \eta}\right|_{\eta=0} d\boldsymbol{x} d\boldsymbol{v}^* && (71) \\
&+ \int \left.\frac{\partial \boldsymbol{H}\left(\boldsymbol{x}, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}_\eta; \eta, \boldsymbol{v}_\eta); \boldsymbol{t}^*\right)}{\partial \eta}\right|_{\eta=0} f_{\boldsymbol{x}, \boldsymbol{v}_*}(\boldsymbol{x}, \boldsymbol{v}_*; \eta) d\boldsymbol{x} d\boldsymbol{v}^* && (72) \\
&= \int \boldsymbol{H}\left(\boldsymbol{x}, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}_*; \eta = 0, \boldsymbol{v}_*); \boldsymbol{t}^*\right) \left.\frac{\partial f_{\boldsymbol{x}, \boldsymbol{v}_\eta}(\boldsymbol{x}, \boldsymbol{v}_\eta; \eta)}{\partial \eta}\right|_{\eta=0} d\boldsymbol{x} d\boldsymbol{v}^* && (73) \\
&+ \int \boldsymbol{H}_{\boldsymbol{b}}\left(\boldsymbol{x}, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}_*; \eta = 0, \boldsymbol{v}_*); \boldsymbol{t}^*\right) \left.\frac{\partial \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}_\eta; \eta, \boldsymbol{v}_\eta)}{\partial \eta}\right|_{\eta=0} f_{\boldsymbol{x}, \boldsymbol{v}_*}(\boldsymbol{x}, \boldsymbol{v}_*; \eta = 0) d\boldsymbol{x} d\boldsymbol{v}^*, && (74)
\end{aligned}$$

To derive $\boldsymbol{\alpha}(\boldsymbol{x}, \boldsymbol{z})$ debiasness term, we follow with the proof by Farrell et al. (2021) and some idea of Hahn and Ridder (2013). Let's begin with the fact that the first order condition

holds as an identity in η and conditional on $\boldsymbol{X}$. That is, as an identity in η,

$$\mathbb{E}_\eta \left[\boldsymbol{J}(\boldsymbol{x}, \boldsymbol{v}_\eta) \ell_{\boldsymbol{b}}(W, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}_\eta; \eta, \boldsymbol{v}_\eta)) \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{V} = \boldsymbol{v} \right] \equiv \boldsymbol{0}$$

where $\ell_{\boldsymbol{b}}$ is the d_θ-vector gradient of ℓ with respect to $\boldsymbol{b}$, given by

$$\ell_{\boldsymbol{b}}(W, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}_\eta; \eta, \boldsymbol{v}_\eta)) = \left. \frac{\partial \ell(w, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right|_{\beta = \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}_\eta; \eta, \boldsymbol{v}_\eta)}$$

The expectation is also indexed by η in the submodel, as the density depends on η. To be explicit, as an identity in η we have

$$\int \left. \frac{\partial \ell(\boldsymbol{w}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right|_{\boldsymbol{\beta} = \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}_\eta; \eta, \boldsymbol{v}_\eta))} f_{y, \boldsymbol{t} | x, z}(y, \boldsymbol{t}; \eta \mid \boldsymbol{x}, \boldsymbol{z}) dy d\boldsymbol{t} \equiv 0.$$

Define $\ell_{\boldsymbol{bb}}(\boldsymbol{w}, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}_\eta; \eta, \boldsymbol{v}_\eta); \eta))$ as the $d_\theta \times d_\theta$ matrix of second derivatives of $\ell(\boldsymbol{w}, \boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$, evaluated at $\boldsymbol{\beta} = \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}_\eta; \eta, \boldsymbol{v}_\eta)$. That is, $\ell_{\boldsymbol{bb}}(\boldsymbol{w}, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}_\eta; \eta, \boldsymbol{v}_\eta); \eta))$ has $\{k_1, k_2\}$ element given by

$$\left[\ell_{\boldsymbol{bb}}(\boldsymbol{w}, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}_\eta; \eta, \boldsymbol{v}_\eta); \eta)) \right]_{k_1, k_2} = \left. \frac{\partial^2 \ell(\boldsymbol{w}, \boldsymbol{\beta})}{\partial \beta_{k_1} \partial \beta_{k_2}} \right|_{\boldsymbol{\beta} = \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}_\eta; \eta, \boldsymbol{v}_\eta); \eta)},$$

where β_{k_1} and β_{k_2} are the respective elements of $\boldsymbol{b}$. With this notation, differentiating the above identity with respect to η and applying the chain rule we find

$$\begin{aligned}
\int \boldsymbol{J}(\boldsymbol{x}, \boldsymbol{v}_\eta) \ell_{\boldsymbol{b}}(\boldsymbol{w}, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}_\eta; \eta, \boldsymbol{v}_\eta)) \frac{\partial f_{y, \boldsymbol{t} | x, v}(y, \boldsymbol{t}; \eta \mid \boldsymbol{x}, \boldsymbol{v})}{\partial \eta} dy d\boldsymbol{t} \\
+ \int \boldsymbol{J}(\boldsymbol{x}, \boldsymbol{v}_\eta) \ell_{\boldsymbol{bb}}(\boldsymbol{w}, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}_\eta; \eta, \boldsymbol{v}_\eta)) \boldsymbol{b}_\eta(\boldsymbol{x}, \boldsymbol{v}_\eta; \eta, \boldsymbol{v}_\eta) f_{y, \boldsymbol{t} | x, v(\boldsymbol{y}, \boldsymbol{t}; \eta | \boldsymbol{x}, \boldsymbol{v})} d\boldsymbol{y} d\boldsymbol{t} \\
+ \int \boldsymbol{J}_\eta(\boldsymbol{x}, \boldsymbol{v}_\eta)) \ell_{\boldsymbol{b}}(\boldsymbol{w}, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}_\eta; \eta, \boldsymbol{v})) f_{y, t | x, v_\eta}(\boldsymbol{y}, \boldsymbol{t}; \eta \mid \boldsymbol{x}, \boldsymbol{v}) d\boldsymbol{y} d\boldsymbol{t} = 0
\end{aligned}$$

Evaluating this result at $\eta = 0$, we obtain

$$\begin{aligned}
&\mathbb{E} \left[\boldsymbol{J}(\boldsymbol{x}, \boldsymbol{v}^*) \ell_{\boldsymbol{b}}(\boldsymbol{w}, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}^*; \eta = 0, \boldsymbol{v}^*))) S(\boldsymbol{Y}, \boldsymbol{T} \mid \boldsymbol{X}, \boldsymbol{V}) \mid \boldsymbol{X}, \boldsymbol{V} \right] \\
&+ \mathbb{E} \left[\boldsymbol{J}(\boldsymbol{x}, \boldsymbol{v}^*) \ell_{\boldsymbol{bb}}(\boldsymbol{w}, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}^*; \eta = 0, \boldsymbol{v}^*)) \boldsymbol{b}_\eta(\boldsymbol{x}, \boldsymbol{v}_\eta; \eta, \boldsymbol{v}_\eta) \mid \boldsymbol{X}, \boldsymbol{V} \right] \\
&+ \mathbb{E} \left[\boldsymbol{J}_\eta(\boldsymbol{x}, \boldsymbol{v}_\eta) \mid_{\eta = 0} \ell_{\boldsymbol{b}}(\boldsymbol{w}, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}^*; \eta, \boldsymbol{v}^*)) \mid \boldsymbol{X}, \boldsymbol{V} \right] = 0,
\end{aligned} \tag{75}$$

Rearranging 75, and using that $\boldsymbol{b}$ is only a function of $\boldsymbol{X}$ and $\boldsymbol{V}^*$, gives

$$\begin{aligned}
&\boldsymbol{J}(\boldsymbol{x}, \boldsymbol{v}^*) \mathbb{E} \left[\ell_{\boldsymbol{bb}}(\boldsymbol{w}, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}^*; \eta = 0, \boldsymbol{v}^*)) \mid \boldsymbol{X}, \boldsymbol{V} \right] \boldsymbol{b}_\eta(\boldsymbol{x}, \boldsymbol{v}^*; \eta = 0, \boldsymbol{v}^*) = \\
&- \mathbb{E} \left[\boldsymbol{J}(\boldsymbol{x}, \boldsymbol{v}^*)) \ell_{\boldsymbol{b}}(\boldsymbol{w}, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}^*; \eta = 0, \boldsymbol{v}^*))) S(\boldsymbol{Y}, \boldsymbol{T} \mid \boldsymbol{X}, \boldsymbol{V}) \mid \boldsymbol{X}, \boldsymbol{V} \right] \\
&- \mathbb{E} \left[\boldsymbol{J}_\eta(\boldsymbol{x}, \boldsymbol{v}_\eta)) \mid_{\eta = 0} \ell_{\boldsymbol{b}}(\boldsymbol{w}, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}^*; \eta, \boldsymbol{v}^*)) \mid \boldsymbol{X}, \boldsymbol{V} \right]
\end{aligned} \tag{76}$$

Then, because $\Lambda(\boldsymbol{x}, \boldsymbol{v}^*) := \mathbb{E}\left[\ell_{\boldsymbol{bb}}(\boldsymbol{w}, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}^*; \eta = 0, \boldsymbol{v}^*)) \mid \boldsymbol{X}, \boldsymbol{V}\right]$ is invertible and we set $\boldsymbol{J}(\boldsymbol{x}, \boldsymbol{v}^*)) = \Lambda(\boldsymbol{x}, \boldsymbol{v}^*)^{-1}$, we have

$$\begin{aligned}
\boldsymbol{b}_\eta(\boldsymbol{x}, \boldsymbol{v}^*; \eta = 0, \boldsymbol{v}^*) &= -\mathbb{E}\left[\Lambda(\boldsymbol{x}, \boldsymbol{v}^*)^{-1}\ell_{\boldsymbol{b}}(\boldsymbol{w}, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}^*; \eta = 0, \boldsymbol{v}^*)))S(\boldsymbol{Y}, \boldsymbol{T} \mid \boldsymbol{X}, \boldsymbol{V}) \mid \boldsymbol{X}, \boldsymbol{V}\right] \\
&\quad - \mathbb{E}\left[\Lambda_\eta(\boldsymbol{x}, \boldsymbol{v}_\eta)^{-1} \mid_{\eta=0} \ell_{\boldsymbol{b}}(\boldsymbol{w}, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}_\eta; \eta, \boldsymbol{v}_\eta)) \mid \boldsymbol{X}, \boldsymbol{V}\right] \\
&= -\mathbb{E}\left[\Lambda(\boldsymbol{x}, \boldsymbol{v}^*)^{-1}\ell_{\boldsymbol{b}}(\boldsymbol{w}, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}^*; \eta = 0, \boldsymbol{v}^*)))S(\boldsymbol{Y}, \boldsymbol{T} \mid \boldsymbol{X}, \boldsymbol{V}) \mid \boldsymbol{X}, \boldsymbol{V}\right]. \qquad (77)
\end{aligned}$$

Substituting this into the second term of Equation 74 and applying iterated expectations, we have

$$\begin{aligned}
&\int \boldsymbol{H}_{\boldsymbol{b}}\left(\boldsymbol{x}, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}^*; \eta = 0, \boldsymbol{v}^*); \boldsymbol{t}^*\right) \boldsymbol{b}_\eta(\boldsymbol{x}, \boldsymbol{v}^*; \eta = 0, \boldsymbol{v}^*) f_{\boldsymbol{x}, \boldsymbol{v}^*}(\boldsymbol{x}, \boldsymbol{v}^*; \eta = 0) d\boldsymbol{x} d\boldsymbol{v}^* \\
&= -\mathbb{E}\left[\boldsymbol{H}_{\boldsymbol{b}}\left(\boldsymbol{x}, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}^*; \eta = 0, \boldsymbol{v}^*); \boldsymbol{t}^*\right) \mathbb{E}\left[\Lambda(\boldsymbol{X}, \boldsymbol{V}^*)^{-1}\ell_{\boldsymbol{b}}(W, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}^*; \eta = 0, \boldsymbol{v}^*))S(\boldsymbol{Y}, \boldsymbol{T} \mid \boldsymbol{X}, \boldsymbol{V}) \mid \boldsymbol{X}, \boldsymbol{V}\right]\right] \\
&= -\mathbb{E}\left[\mathbb{E}\left[\boldsymbol{H}_{\boldsymbol{b}}\left(\boldsymbol{x}, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}^*; \eta = 0, \boldsymbol{v}^*); \boldsymbol{t}^*\right) \Lambda(\boldsymbol{X}, \boldsymbol{V}^*)^{-1}\ell_{\boldsymbol{b}}(W, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}^*; \eta = 0, \boldsymbol{v}^*))S(\boldsymbol{Y}, \boldsymbol{T} \mid \boldsymbol{X}, \boldsymbol{V}) \mid \boldsymbol{X}, \boldsymbol{V}\right]\right] \\
&= -\mathbb{E}\left[\boldsymbol{H}_{\boldsymbol{b}}\left(\boldsymbol{x}, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}^*; \eta = 0, \boldsymbol{v}^*); \boldsymbol{t}^*\right) \Lambda(\boldsymbol{X}, \boldsymbol{V}^*)^{-1}\ell_{\boldsymbol{b}}(W, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}^*; \eta = 0, \boldsymbol{v}^*))S(\boldsymbol{Y}, \boldsymbol{T} \mid \boldsymbol{X}, \boldsymbol{V})\right]
\end{aligned}$$

Next, because the first order condition holds conditionally,

$$\begin{aligned}
&\mathbb{E}\left[\boldsymbol{H}_{\boldsymbol{b}}\left(\boldsymbol{x}, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}^*; \eta = 0, \boldsymbol{v}^*); \boldsymbol{t}^*\right) \Lambda(\boldsymbol{X}, \boldsymbol{V}^*)^{-1}\ell_{\boldsymbol{b}}(W, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}^*; \eta = 0, \boldsymbol{v}^*))S(\boldsymbol{X}, \boldsymbol{V}^*)\right] = \\
&\mathbb{E}\left[\boldsymbol{H}_{\boldsymbol{b}}\left(\boldsymbol{x}, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}^*; \eta = 0, \boldsymbol{v}^*); \boldsymbol{t}^*\right) \Lambda(\boldsymbol{X}, \boldsymbol{V}^*)^{-1}\mathbb{E}\left[\ell_{\boldsymbol{b}}(W, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}^*; \eta = 0, \boldsymbol{v}^*)) \mid \boldsymbol{X}, \boldsymbol{V}^*\right] S(\boldsymbol{X}, \boldsymbol{V}^*)\right] = 0.
\end{aligned}$$

Therefore,the second term of Equation 74 is of the required form:

$$-\mathbb{E}\left[\boldsymbol{H}_{\boldsymbol{b}}\left(\boldsymbol{x}, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}_*; \eta = 0, \boldsymbol{v}_*); \boldsymbol{t}^*\right) \Lambda(\boldsymbol{X}, \boldsymbol{V}^*)^{-1}\ell_{\boldsymbol{b}}(W, \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{v}_*; \eta = 0, \boldsymbol{v}_*))S(\boldsymbol{Y}, \boldsymbol{T}, \boldsymbol{X}, \boldsymbol{V}^*)\right]$$

So we can derive the influence function as below:

$$\boldsymbol{\psi}(\boldsymbol{w}, \boldsymbol{b}_*(\boldsymbol{x}, \boldsymbol{v}_*), \boldsymbol{\Lambda}(\boldsymbol{x}, \boldsymbol{z})) = \boldsymbol{H}(\boldsymbol{x}, \boldsymbol{b}_*(\boldsymbol{x}, \boldsymbol{v}_*); \boldsymbol{t}^*) - \boldsymbol{H}_{\boldsymbol{b}}(\boldsymbol{x}, \boldsymbol{b}_*(\boldsymbol{x}, \boldsymbol{v}_*); \boldsymbol{t}^*)\boldsymbol{\Lambda}(\boldsymbol{x}, \boldsymbol{v}^*)^{-1}\boldsymbol{\ell}_{\boldsymbol{b}}(\boldsymbol{w}, \boldsymbol{b}_*(\boldsymbol{x}, \boldsymbol{v}_*))) \qquad (78)$$

C.2 Asymptotic Normality: Proof of Theorem 4

The asymptotic normality result is a direct application of Theorems 3.1 and 3.2 from Chernozhukov et al. (2018), which are satisfied under Assumptions 3.1 and 3.2 as outlined in Farrell et al. (2021).

For Assumption 3.1(a), the condition is satisfied by the term $\boldsymbol{\psi}(\boldsymbol{w}, \boldsymbol{b}^*, \boldsymbol{\Lambda}) - \boldsymbol{\mu}$, as indicated by Theorem 4. Specifically, the first term of the influence function, $\boldsymbol{\psi}(\boldsymbol{w}, \boldsymbol{b}^*, \boldsymbol{\Lambda})$, has expectation $\boldsymbol{\mu}^*$, as constructed for the inferential parameters in equation (6). Moreover, the conditional expectation of the second term in $\boldsymbol{\psi}(\boldsymbol{w}, \boldsymbol{b}^*, \boldsymbol{\Lambda})$ equals zero, which is guaranteed by Assumption 9.

The linearity condition of Assumption 3.1(b) is satisfied by the definition of the inferential parameters in equation (6) and the corresponding score in Theorem 4. Assumption 3.1(c), which requires the score to be twice continuously Gateaux-differentiable with respect to the nuisance functions on $\boldsymbol{T}$, is also guaranteed by Assumption 9. Furthermore, the Neyman orthogonality property for the score $\boldsymbol{\psi}$ is established through the derivation of the influence function in Theorem 4. Finally, the identification condition in Assumption 3.1(e) is automatically satisfied, as the matrix J_0 is the identity matrix in this setting.

Moving on to Assumption 3.2, both Assumptions 3.2(b) and 3.2(d) are satisfied by the moment conditions provided in Assumption 9. Assumptions 3.2(a) and 3.2(c) can be derived from Equations (3.7) and (3.8) in Chernozhukov et al. (2018), with the necessary rate condition $\epsilon_N = o(N^{-\frac{1}{4}})$. This rate condition for Equations (3.7) and (3.8) in Chernozhukov et al. (2018) is satisfied by the sup-norm convergence rate of the first-stage estimates, as established in Corollary 2 or Theorem 2. Therefore, all the required conditions hold, leading to the asymptotic normality of the proposed estimator.