Belloni, Alexandre; Chernozhukov, Victor; Fernández Val, Iván; Hansen, Christian

Working Paper

Program evaluation with high-dimensional data

cemmap working paper, No. CWP77/13

Provided in Cooperation with:
The Institute for Fiscal Studies (IFS), London

Suggested Citation: Belloni, Alexandre; Chernozhukov, Victor; Fernández Val, Iván; Hansen, Christian (2013) : Program evaluation with high-dimensional data, cemmap working paper, No. CWP77/13, Centre for Microdata Methods and Practice (cemmap), London, https://doi.org/10.1920/wp.cem.2013.7713

This Version is available at:
https://hdl.handle.net/10419/97383

cemmap
centre for microdata methods and practice

Program evaluation with high-dimensional data

Alexandre Belloni
Victor Chernozhukov
Iván Fernández Val
Christian Hansen

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP77/13

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

PROGRAM EVALUATION WITH HIGH-DIMENSIONAL DATA

A. BELLONI, V. CHERNOZHUKOV, I. FERNÁNDEZ-VAL, AND C. HANSEN

Abstract. In the first part of the paper, we consider estimation and inference on policy relevant treatment effects, such as local average and local quantile treatment effects, in a data-rich environment where there may be many more control variables available than there are observations. In addition to allowing many control variables, the setting we consider allows endogenous receipt of treatment, heterogeneous treatment effects, and function-valued outcomes. To make informative inference possible, we assume that some reduced form predictive relationships are approximately sparse. That is, we require that the relationship between the control variables and the outcome, treatment status, and instrument status can be captured up to a small approximation error using a small number of the control variables whose identities are unknown to the researcher. This condition allows estimation and inference for a wide variety of treatment parameters to proceed after selection of an appropriate set of controls formed by selecting control variables separately for each reduced form relationship and then appropriately combining these reduced form relationships. We provide conditions under which post-selection inference is uniformly valid across a wide-range of models and show that a key condition underlying the uniform validity of post-selection inference allowing for imperfect model selection is the use of approximately unbiased estimating equations. We illustrate the use of the proposed methods with an application to estimating the effect of 401(k) participation on accumulated assets.

In the second part of the paper, we present a generalization of the treatment effect framework to a much richer setting, where possibly a continuum of target parameters is of interest and the Lasso-type or post-Lasso type methods are used to estimate a continuum of high-dimensional nuisance functions. This framework encompasses the analysis of local treatment effects as a leading special case and also covers a wide variety of classical and modern moment-condition problems in econometrics. We establish a functional central limit theorem for the continuum of the target parameters, and also show that it holds uniformly in a wide range of data-generating processes P, with continua of approximately sparse nuisance functions. We also establish validity of the multiplier bootstrap for resampling the first order approximations to the standardized continuum of the estimators, and also establish uniform validity in P. We propose a notion of the functional delta method for finding limit distribution and multiplier bootstrap of the smooth functionals of the target parameters that is valid uniformly in P. Finally, we establish rate and consistency results for continua of Lasso or post-Lasso type methods for estimating continua of the (nuisance) regression functions, also providing practical, theoretically justified penalty choices. Each of these results is new and could be of independent interest.

Keywords: local average and quantile treatment effects, endogeneity, instruments, local effects of treatment on the treated, propensity score, Lasso, inference on infinite-dimensional parameters after model selection, moment conditional models with a continuum of target parameters, Lasso and Post-Lasso with functional response data.

Date: First date: April 2013. This version: December 28, 2013. This is a short version of the paper. We gratefully acknowledge research support from the NSF. We are grateful to the seminar participants at University of Montreal, 2013 Summer NBER Institute, and the University of Illinois at Urbana-Champaign for helpful comments.

1. Introduction

The goal of many empirical analyses in economics is to understand the causal effect of a treatment such as participation in a government program on economic outcomes. Such analyses are often complicated by the fact that few economic treatments or government policies are randomly assigned. The lack of true random assignment has led to the adoption of a variety of quasi-experimental approaches to estimating treatment effects that are based on observational data. Such approaches include instrumental variable (IV) methods in cases where treatment is not randomly assigned but there is some other external variable, such as eligibility for receipt of a government program or service, that is either randomly assigned or the researcher is willing to take it as exogenous conditional on the right set of control variables. Another common approach is to assume that the treatment variable itself may be taken as exogenous after conditioning on the right set of controls which leads to regression or matching based methods, among others, for estimating treatment effects.[1]

A practical problem empirical researchers must face when trying to estimate treatment effects is deciding what conditioning variables to include. When the treatment variable or instrument is not randomly assigned, a researcher must choose what needs to be conditioned on to make the argument that the instrument or treatment is exogenous plausible. Typically, economic intuition will suggest a set of variables that might be important to control for but will not identify exactly which variables are important or the functional form with which variables should enter the model. While less crucial to plausibly identifying treatment effects, the problem of selecting controls also arises in situations where the key treatment or instrumental variables are randomly assigned. In these cases, a researcher interested in obtaining precisely estimated policy effects will also typically consider including additional control variables to help absorb residual variation. As in the case where including controls is motivated by a desire to make identification of the treatment effect more plausible, one rarely knows exactly which variables will be most useful for accounting for residual variation. In either case, the lack of clear guidance about what variables to use presents the problem of selecting a set of controls from a potentially large set of control variables including raw variables available in the data as well as interactions and other transformations of these variables.

In this paper, we consider estimation of the effect of an endogenous binary treatment, D, on an outcome, Y, in the presence of a binary instrumental variable, Z, in settings with very many potential control variables, $f(X)$, including raw variables, X, and transformations of these variables such as powers, b-splines, or interactions. We allow for fully heterogeneous treatment effects and thus focus on estimation of causal quantities that are appropriate in heterogeneous effects settings such as the local average treatment effect (LATE) or the local quantile treatment effect (LQTE). We focus our discussion on the case where identification is obtained through the

[1]There is a large literature about estimation of treatment effects. See, for example, the textbook treatments in Angrist and Pischke (2008) or Wooldridge (2010) and the references therein for discussion from an economic perspective.

use of an instrumental variable, but all results carry through to the case where the treatment is taken as exogenous after conditioning on sufficient controls. Note that we can simply replace the instrument with the treatment variable in the estimation and inference methods and in the formal results.

The methodology for estimating policy-relevant effects we consider allows for cases where the number of potential control variables, $p := \dim f(X)$, is much greater than the sample size, n. Of course, informative inference about causal parameters cannot proceed allowing for $p \gg n$ without further restrictions. We impose sufficient structure through the assumption that reduced form relationships such as $\mathrm{E}[D|X]$, $\mathrm{E}[Z|X]$, and $\mathrm{E}[Y|X]$ are approximately sparse. Intuitively, approximate sparsity imposes that these reduced form relationships can be represented up to a small approximation error as a linear combination, possibly inside of a known link function such as the logistic function, of a small number $s \ll n$ of the variables in $f(X)$ whose identities are *a priori* unknown to the researcher. This assumption allows us to use methods for estimating models in high-dimensional sparse settings that are known to have good prediction properties to estimate the fundamental reduced form relationships. We may then use these estimated reduced form quantities as inputs to estimating the causal parameters of interest. Approaching the problem of estimating treatment effects within this framework allows us to accommodate the realistic scenario in which a researcher is unsure about exactly which confounding variables or transformations of these confounds are important and so must search among a broad set of controls.

Valid inference following model selection is non-trivial. Direct application of usual inference procedures following model selection does not provide valid inference about causal parameters even in low-dimensional settings, such as when there is only a single control, unless one assumes sufficient structure on the model that perfect model selection is possible. Such structure is very restrictive and seems unlikely to be satisfied in many economic applications. For example, a typical condition that allows perfect model selection in a linear model is to assume that all but a small number of coefficients are exactly zero and that the non-zero coefficients are all large enough that they can be distinguished them from zero with probability very near one in finite samples. Such a condition rules out the possibility that there may be some variables which have moderate, but non-zero, partial effects. Ignoring such variables may lead to only a small loss in predictive performance while also producing a non-ignorable omitted variables bias that has a substantive impact on estimation and inference regarding individual model parameters. (For further discussion, see Leeb and Pötscher (2008a; 2008b), Pötscher (2009), as well as Belloni, Chernozhukov, and Hansen (2011).)

A key contribution of our paper is providing inferential procedures for key parameters used in program evaluation that are theoretically valid within approximately sparse models allowing for imperfect model selection. Our procedures build upon the insights in Belloni, Chernozhukov, and Hansen (2010) and Belloni, Chen, Chernozhukov, and Hansen (2012), which demonstrate that valid inference about low-dimensional structural parameters can proceed following model

selection, allowing for model selection mistakes, under two key conditions. First, estimation should be based upon "orthogonal" moment conditions that are first-order insensitive to changes in the values of nuisance parameters. Specifically, if the target parameter value α_0 is identified via the moment condition

$$\mathrm{E}_P \psi(W, \alpha_0, h_0) = 0, \tag{1}$$

where h_0 is a nuisance function-valued parameter estimated via a post-model-selection or regularization method, one needs to use a moment function, ψ, such that the moment condition is orthogonal with respect to perturbations of h around h_0. More formally, the moment conditions should satisfy

$$\partial_h [\mathrm{E}_P \psi(W, \alpha_0, h)]_{h=h_0} = 0 \tag{2}$$

where ∂h computes the functional derivative operator with respect to h. Second, one needs to use a model selection procedure that keeps model selection errors "moderately" small.

The orthogonality condition embodied in (2) has a long history in statistics and econometrics. For example, this type of orthogonality was used by Neyman (1979) in low-dimensional settings to deal with crudely estimated parametric nuisance parameters. To the best of our knowledge, Belloni, Chernozhukov, and Hansen (2010) and Belloni, Chen, Chernozhukov, and Hansen (2012) were the first to use this property in the $p \gg n$ setting. They applied it to a linear instrumental variables model with many instruments, where the nuisance function h_0 is the optimal instrument estimated by Lasso or post-Lasso methods. Using estimators based upon moment conditions with this low-bias property insures that crude estimation of h_0 via post-selection or other regularization methods has an asymptotically negligible effect on the estimation of α_0. Belloni, Chernozhukov, and Hansen (2011) and Farrell (2013) also exploited this approach in the $p \gg n$ setting to develop a double-selection method that yields valid inference on the parameters of the linear part of the partially linear model and on average treatment effects when the treatment is exogenous conditional on observables. In the general endogenous treatment effects setting we consider in this paper, such moment conditions can be found as efficient influence functions for certain reduced form parameters as in Hahn (1998). Moreover, our analysis allows for function-valued outcomes. As a result, the parameters of interest α_0 are themselves function-valued; i.e. they can carry an index. We illustrate how these efficient influence functions coupled with methods developed for forecasting in high-dimensional sparse models can be used to estimate and obtain valid inferential statements about a variety of structural/treatment effects. We formally demonstrate the uniform validity of the resulting inference within a broad class of approximately sparse models including models where perfect model selection is theoretically impossible.

The second set of main results of the paper deals with a more general setting, where possibly a continuum of target parameters is of interest and the Lasso-type or post-Lasso type methods are used to estimate the a continuum of high-dimensional nuisance functions. This framework is quite general, and it encompasses the analysis of LATE and other effects for function-valued outcomes as a special case. It covers a very wide variety of classical and modern moment-condition problems – it covers both smooth moment conditions as well as non-smooth ones, such those arising in

the context of structural quantile analysis, for example, the nonseparable endogenous models in Chernozhukov and Hansen (2005). Here, firstly, we establish a functional central limit theorem for the continuum of the target parameters, and also show that it holds uniformly in a wide range of data-generating processes P with approximately sparse continua of nuisance functions. Secondly, we also establish validity of the multiplier bootstrap for resampling the first order approximations to the standardized continua of the estimators, and also establish uniform-in-P validity. (These uniformity results here complement those given in (Romano and Shaikh, 2012) for the empirical bootstrap.) Thirdly, we establish validity of the functional delta method uniformly in P, under an appropriately strengthened notion of Hadamard differentiability, as well as uniform-in-P validity of the functional delta method for the multiplier bootstrap for resampling the smooth functionals of the continuum of the target parameters. All of these results are new and represent the second main contribution of the paper.

In establishing our main theoretical results, we consider variable selection for functional response data using ℓ_1-penalized methods. This type of data arises, for example, when one is interested in LQTE at not just a single quantile but over a range of quantile indices or when one is interested in how $1(Y \leq u)$ relates to the treatment over a range of threshold values u. Considering such functional response data allows us to provide a unified inference procedure for interesting quantities such as the distributional effects of the treatment as well as simpler objects such as the LQTE at a single quantile. The main theoretical contribution of the paper is to demonstrate that the developed methods provide uniformly valid inference for functional response data in a high-dimensional setting allowing for model selection mistakes. Our result builds upon the work of Belloni and Chernozhukov (2011b) who provided rates of convergence for variable selection when one is interested in estimating the quantile regression process with exogenous variables. More generally, this theoretical work complements and extends the rapidly growing set of results for ℓ_1-penalized estimation methods; see, for example, Frank and Friedman (1993); Tibshirani (1996); Fan and Li (2001); Zou (2006); Candès and Tao (2007); van de Geer (2008); Huang, Horowitz, and Ma (2008); Bickel, Ritov, and Tsybakov (2009); Meinshausen and Yu (2009); Bach (2010); Huang, Horowitz, and Wei (2010); Belloni and Chernozhukov (2011a); Kato (2011); Belloni, Chen, Chernozhukov, and Hansen (2012); Belloni and Chernozhukov (2013); Belloni, Chernozhukov, and Kato (2013); Belloni, Chernozhukov, and Wei (2013); and the references therein. We also demonstrate that a simple multiplier bootstrap procedure can be used to produce asymptotically valid inferential statements on function-valued parameters, which should aid in the practical implementation of our methods.

We illustrate the use of our methods by estimating the effect of 401(k) participation on measures of accumulated assets as in Chernozhukov and Hansen (2004).[2] Similar to Chernozhukov and Hansen (2004), we provide estimates of LATE and LQTE over a range of quantiles. We differ from this previous work by using the high-dimensional methods developed in this paper to allow ourselves to consider a much broader set of control variables than have previously been

[2]See also Poterba, Venti, and Wise (1994; 1995; 1996; 2001); Benjamin (2003); and Abadie (2003) among others.

considered. We find that 401(k) participation has a small impact on accumulated financial assets at low quantiles while appearing to have a much larger impact at high quantiles. Interpreting the quantile index as "preference for savings" as in Chernozhukov and Hansen (2004), this pattern suggests that 401(k) participation has little causal impact on the accumulated financial assets of those with low desire to save but a much larger impact on those with stronger preferences for saving. It is interesting that these results are quite similar to those in Chernozhukov and Hansen (2004) despite allowing for a much richer set of control variables.

1.1. **Notation.** We have a pair of random elements (W, ξ) living on the probability space $(S, \mathcal{A}_S, P \times P_\xi)$. We have i.i.d. copies $(W_i)_{i=1}^n$ of W, the data. The variable ξ is the bootstrap multiplier variable with law determined by P_ξ, and it is independent of W. The data streams $(W_i)_{i=1}^\infty$ live on the probability space $(A, \mathcal{A}_A, \mathrm{P}_P)$, containing the infinite product of the space above as a subproduct, where notation P_P signifies the dependence on P, the "data-generating process" for W. It is important to keep track of this dependence in the analysis, if we want the results to hold uniformly in P in some set $\mathcal{P}_n$, which may be dependent on n (typically increasing in n.) The probability space $(A, \mathcal{A}_A, \mathrm{P}_P)$ will also carry i.i.d. copies of bootstrap multipliers $(\xi_i)_{i=1}^\infty$ which are independent of the data streams $(W_i)_{i=1}^\infty$. Note also that we use capital letters such as W to denote random elements and use the lower case letters such as w as fixed values that these random elements can take. The operator E_P denotes the expectation with respect to the probability measure P_P.

We denote by $\mathbb{P}_n$ the (random) empirical probability measure that assigns probability n^{-1} to each (W_i, ξ_i). $\mathbb{E}_n$ denotes the expectation with respect to the empirical measure, and $\mathbb{G}_n$ denotes the empirical process $\sqrt{n}(\mathbb{E}_n - P)$, i.e.

$$\mathbb{G}_n(f) = \mathbb{G}_n(f(W, \xi)) = n^{-1/2} \sum_{i=1}^n \{f(W_i, \xi_i) - P[f(W), \xi]\}, \quad P[f(W, \xi)] := \int f(w, \xi) dP(w) dP_\xi(\xi)$$

indexed by a measurable class of functions $\mathcal{F} : S \mapsto \mathbb{R}$; see van der Vaart and Wellner (1996), Chapter 2.3. In what follows, we use $\|\cdot\|_{\mathsf{P},q}$ to denote the $L^q(\mathsf{P})$ norm; for example, we use $\|f(W)\|_{P,q} = (\int |f(w)|^q dP(w))^{1/q}$ and $\|f(W)\|_{\mathbb{P}_n,q} = (n^{-1} \sum_{i=1}^n \|f(W_i)\|^q)^{1/q}$. For a vector $v = (v_1, \ldots, v_p) \in \mathbb{R}^p$, $\|v\|_0$ denotes the ℓ_0-"norm" of v, that is, the number of non-zero components of v, $\|v\|_1$ denotes the ℓ_1-norm of v, that is, $\|v\|_1 = |v_1| + \cdots + |v_n|$, and $\|v\|$ denotes the Euclidean norm of v, that is, $\|v\| = \sqrt{v'v}$. Given a class $\mathcal{F}$ of measurable functions from S to $\mathbb{R}$ we say that it is *suitably measurable*, if it is an image admissible Suslin class as defined in (Dudley, 1999), which is a rather mild assumption. For a positive integer k, symbol $[k]$ denotes the set $\{1, \ldots, k\}$.

2. The Setting and The Target Parameters

2.1. **Observables and Reduced Form Parameters.** The observables are a random variable $W = ((Y_u)_{u \in \mathcal{U}}, X, Z, D)$. The outcome variable of interest Y_u is indexed by $u \in \mathcal{U}$. We give examples of the index u below. The variable $D \in \mathcal{D} = \{0, 1\}$ is an indicator of the receipt of a treatment or participation in a program. It will be typically treated as endogenous; that is, we will typically view the treatment as assigned non-randomly with respect to the outcome. The

instrumental variable $Z \in \mathcal{Z} = \{0, 1\}$ is a binary indicator, such as an offer of participation, that is assumed to be randomly assigned conditional on the observable covariates X with support $\mathcal{X}$. For example, in the empirical application we argue that 401(k) eligibility can be considered exogenous only after conditioning on income and other individual characteristics. The notions of exogeneity and endogeneity we employ are standard, but we state them below for clarity and completeness. We also restate standard conditions that are sufficient for a causal interpretation of our target parameters.

The indexing of the outcome Y_u by u is useful to analyze functional data. For example, Y_u could represent an outcome falling short of a threshold, namely $Y_u = 1(Y \leqslant u)$, in the context of distributional analysis; Y_u could be a height indexed by age u in growth charts analysis; or Y_u could be a health outcome indexed by a dosage u in dosage response studies. Our framework is tailored for such functional response data. The special case with no index is included by simply considering $\mathcal{U}$ to be a singleton set.

We make use of two key types of reduced form parameters for estimating the structural parameters of interest – (local) treatment effects and related quantities. These reduced form parameters are defined as

$$\alpha_V(z) := \mathrm{E}_P[g_V(z, X)] \text{ for } z \in \{0, 1\} \text{ and } \gamma_V := \mathrm{E}_P[V], \tag{3}$$

where $z = 0$ or $z = 1$ are the fixed values of the random variable Z.[3] The function g_V, mapping the support $\mathcal{Z}\mathcal{X}$ of the vector (Z, X) to the real line $\mathbb{R}$, is defined as

$$g_V(z, x) := \mathrm{E}_P[V|Z = z, X = x]. \tag{4}$$

We use V to denote a target variable whose identity may change depending on the context such as $V = \mathbf{1}_d(D)Y_u$ or $V = \mathbf{1}_d(D)$, where $\mathbf{1}_d(D) := 1(D = d)$ is the indicator function.

All the structural parameters we consider are smooth functionals of these reduced-form parameters. In our approach to estimating treatment effects, we estimate the key reduced form parameter $\alpha_V(z)$ using recent approaches to dealing with high-dimensional data coupled with using "low-bias" estimating equations. The low-bias property is crucial for dealing with the "non-regular" nature of penalized and post-selection estimators which do not admit linearizations except under very restrictive conditions. The use of regularization by model selection or penalization is in turn motivated by the desire to accommodate high-dimensional data.

2.2. **Target Structural Parameters – Local Treatment Effects.** The reduced form parameters defined in (3) are key because the structural parameters of interest are functionals of these elementary objects. The local average structural function (LASF) defined as

$$\theta_{Y_u}(d) = \frac{\alpha_{\mathbf{1}_d(D)Y_u}(1) - \alpha_{\mathbf{1}_d(D)Y_u}(0)}{\alpha_{\mathbf{1}_d(D)}(1) - \alpha_{\mathbf{1}_d(D)}(0)}, \quad d \in \{0, 1\} \tag{5}$$

underlies the formation of many commonly used treatment effects. The LASF identifies the average outcome for the group of *compliers*, individuals whose treatment status may be influenced

[3]The expectation that defines the parameter $\alpha_V(z)$ is well-defined under the support condition $0 < P(Z = 1 \mid X) < 1$ a.s. We impose this condition in Assumption 1 and Assumption 2.

by variation in the instrument, in the treated and non-treated states under standard assumptions; see, e.g. Imbens and Angrist (1994). The local average treatment effect (LATE) is defined as the difference of the two values of the LASF:

$$\theta_{Y_u}(1) - \theta_{Y_u}(0). \tag{6}$$

The term local designates that this parameter does not measure the effect on the entire population but on the subpopulation of compliers.

When there is no endogeneity, formally when $D \equiv Z$, the LASF and LATE become the average structural function (ASF) and average treatment effect (ATE) on the entire population. Thus, our results cover this situation as a special case where the ASF and ATE are given by

$$\theta_{Y_u}(z) = \alpha_{Y_u}(z), \quad \theta_{Y_u}(1) - \theta_{Y_u}(0) = \alpha_{Y_u}(1) - \alpha_{Y_u}(0). \tag{7}$$

We also note that the impact of the instrument Z itself may be of interest since Z often encodes an offer of participation in a program. In this case, the parameters of interest are again simply the reduced form parameters $\alpha_{Y_u}(z)$ and $\alpha_{Y_u}(1)-\alpha_{Y_u}(0)$. Thus, the LASF and LATE are primary targets of interest in this paper, and the ASF and ATE are subsumed as special cases.

2.2.1. *Local Distribution and Quantile Treatment Effects.* Setting $Y_u = Y$ in (5) and (6) provides the conventional LASF and LATE. An important generalization arises by letting $Y_u = 1(Y \leqslant u)$ be the indicator of the outcome of interest falling below a threshold u. In this case, the family of effects

$$(\theta_{Y_u}(1) - \theta_{Y_u}(0))_{u\in\mathbb{R}}, \tag{8}$$

describe the local distribution treatment effects (LDTE). Similarly, we can look at the quantile left-inverse transform of the curve $u \mapsto \theta_{Y_u}(d)$,

$$\theta_Y^{\leftarrow}(\tau, d) := \inf\{u \in \mathbb{R} : \theta_{Y_u}(d) \geqslant \tau\}, \tag{9}$$

and examine the family of local quantile treatment effects (LQTE):

$$(\theta_Y^{\leftarrow}(\tau, 1) - \theta_Y^{\leftarrow}(\tau, 0))_{\tau\in(0,1)}. \tag{10}$$

2.3. **Target Structural Parameters – Local Treatment Effects on the Treated.** In addition to the local treatment effects given in Section 2.2, we may be interested in local treatment effects on the treated. The key object in defining local treatment effects on the treated is the local average structural function on the treated (LASF-T) which is defined by its two values:

$$\vartheta_{Y_u}(d) = \frac{\gamma_{\mathbf{1}_d(D)Y_u} - \alpha_{\mathbf{1}_d(D)Y_u}(0)}{\gamma_{\mathbf{1}_d(D)} - \alpha_{\mathbf{1}_d(D)(0)}}, \quad d \in \{0, 1\}. \tag{11}$$

These quantities identify the average outcome for the group of *treated compliers* in the treated and non-treated states under assumptions stated below. The local average treatment effect on the treated (LATE-T) introduced in Hong and Nekipelov (2010) is defined simply as the difference of two values of the LASF-T:

$$\vartheta_{Y_u}(1) - \vartheta_{Y_u}(0). \tag{12}$$

The LATE-T may be of interest because it measures the average treatment effect for *treated compliers*, namely the subgroup of compliers that actually receive the treatment.[4]

When the treatment is assigned randomly given controls so we can take $D = Z$, the LASF-T and LATE-T become the average structural function on the treated (ASF-T) and average treatment effect on the treated (ATE-T). In this special case, the ASF-T and ATE-T are given by

$$\vartheta_{Y_u}(1) = \frac{\gamma_{\mathbf{1}_1(D)Y_u}}{\gamma_{\mathbf{1}_1(D)}}, \quad \vartheta_{Y_u}(0) = \frac{\gamma_{\mathbf{1}_0(D)Y_u} - \alpha_{Y_u}(0)}{\gamma_{\mathbf{1}_0(D)} - 1}, \quad \vartheta_{Y_u}(1) - \vartheta_{Y_u}(0); \tag{13}$$

and we can use our results to provide estimation and inference results for these quantities.

2.3.1. *Local Distribution and Quantile Treatment Effects on the Treated.* Local distribution treatment effects on the treated (LDTE-T) and local quantile treatment effects on the treated (LQTE-T) can also be defined. As in Section 2.2.1, we let $Y_u = 1(Y \leqslant u)$ be the indicator of the outcome of interest falling below a threshold u. The family of treatment effects

$$(\vartheta_{Y_u}(1) - \vartheta_{Y_u}(0))_{u \in \mathbb{R}} \tag{14}$$

then describes the LDTE-T. We can also use the quantile left-inverse transform of the curve $u \mapsto \vartheta_{Y_u}(d)$,

$$\vartheta_Y^{\leftarrow}(\tau, d) := \inf\{u \in \mathbb{R} : \vartheta_{Y_u}(d) \geqslant \tau\}, \tag{15}$$

and define LQTE-T:

$$(\vartheta_Y^{\leftarrow}(\tau, 1) - \vartheta_Y^{\leftarrow}(\tau, 0))_{\tau \in (0,1)}. \tag{16}$$

Under conditional exogeneity LQTE and LQTE-T reduce to the quantile treatment effects (QTE) (Koenker (2005)) and quantile treatment effects on the treated (QTE-T).

2.4. **Causal Interpretations for Structural Parameters.** The quantities discussed in Sections 2.2 and 2.3 are well-defined and have causal interpretation under standard conditions. To discuss these conditions, we use potential outcomes notation. Let Y_{u1} and Y_{u0} denote the potential outcomes under the treatment states 1 and 0. These outcomes are not observed jointly, and we instead observe $Y_u = DY_{u1} + (1 - D)Y_{u0}$, where $D \in \mathcal{D} = \{0, 1\}$ is the random variable indicating program participation or treatment state. Under exogeneity, D is assigned independently of the potential outcomes conditional on covariates X, i.e. $(Y_{u1}, Y_{u0}) \perp\!\!\!\perp D \mid X$ a.s., where $\perp\!\!\!\perp$ denotes statistical independence.

When exogeneity fails, D may depend on the potential outcomes. For example, people may drop out of a program if they think the program will not benefit them. In this case, instrumental variables are useful in creating quasi-experimental fluctuations in D that may identify useful effects. To provide identification in this setting, we assume the existence of an instrument Z, such as an offer of participation, that is assigned randomly conditional on observable covariates X. We further assume the instrument is binary. Let the random variables D_1 and D_0 indicate

[4]Note that LATE-T $\neq$ LATE in general, because the distribution of X might be different for treated and non-treated compliers.

the potential participation decisions under the instrument states 1 and 0, respectively. These variables may in general depend on the potential outcomes. As with the potential outcomes, the potential participation decisions under both instrument states are not observed jointly. The realized participation decision is then given by $D = ZD_1 + (1 - Z)D_0$.

There are many causal quantities of interest for program evaluation. Chief among these are various structural averages

- average structural function (ASF): $\mathrm{E}_P[Y_{ud}]$,
- average structural function on the treated (ASF-T): $\mathrm{E}_P[Y_{ud} \mid D = 1]$,
- local average structural function (LASF): $\mathrm{E}_P[Y_{ud} \mid D_1 > D_0]$,
- local average structural function on the treated (LASF-T): $\mathrm{E}_P[Y_{ud} \mid D_1 > D_0, D = 1]$,

as well as effects derived from them such as

- average treatment effect (ATE): $\mathrm{E}_P[Y_{u1} - Y_{u0}]$,
- average treatment effect on the treated (ATE-T): $\mathrm{E}_P[Y_{u1} - Y_{u0} \mid D = 1]$,
- local average treatment effect (LATE): $\mathrm{E}_P[Y_{u1} - Y_{u0} \mid D_1 > D_0]$,
- local average treatment effect on the treated (LATE-T): $\mathrm{E}_P[Y_{u1} - Y_{u0} \mid D_1 > D_0, D = 1]$.

These causal quantities are the same as the structural parameters defined in Sections 2.2-2.3 under the following well-known sufficient condition.

Assumption 1 (Causal Interpretability)**.** *The following conditions hold P-almost surely: (Exogeneity)* $((Y_{u1}, Y_{u0})_{u \in \mathcal{U}}, D_1, D_0) \perp\!\!\!\perp Z \mid X$*; (First Stage)* $\mathrm{E}_P[D_1 \mid X] \neq \mathrm{E}_P[D_0 \mid X]$*; (Non-Degeneracy)* $P(Z = 1 \mid X) \in (0, 1)$*; (Monotonicity)* $P(D_1 \geqslant D_0 \mid X) = 1$.

This condition is much-used in the program evaluation literature. It has an equivalent formulation in terms of a simultaneous equation model with a binary endogenous variable; see Vytlacil (2002) and Heckman and Vytlacil (1999). For a thorough discussion of this assumption, we refer to Imbens and Angrist (1994). Using this assumption, we present an identification lemma which follows from results of Abadie (2003) and Hong and Nekipelov (2010) that both in turn build upon Imbens and Angrist (1994). The lemma shows that the parameters θ_{Y_u} and ϑ_{Y_u} defined earlier have a causal interpretation under Assumption 1. Therefore, our referring to them as structural/causal is justified under this condition.

Lemma 2.1 (Identification of Causal Effects)**.** *Under Assumption 1, for each $d \in \mathcal{D}$,*

$$\mathrm{E}_P[Y_{ud} \mid D_1 > D_0] = \theta_{Y_u}(d), \quad \mathrm{E}_P[Y_{ud} \mid D_1 > D_0, D = 1] = \vartheta_{Y_u}(d).$$

Furthermore, if D is exogenous, namely $D \equiv Z$ a.s., then

$$\mathrm{E}_P[Y_{ud} \mid D_1 > D_0] = \mathrm{E}_P[Y_{ud}], \quad \mathrm{E}_P[Y_{ud} \mid D_1 > D_0, D = 1] = \mathrm{E}_P[Y_{ud} \mid D = 1].$$

3. Estimation of Reduced-Form and Structural Parameters in a Data-Rich Environment

Recall that the key objects used in defining the structural parameters in Section 2 are the expectations

$$\alpha_V(z) = \mathrm{E}_P[g_V(z,X)] \text{ and } \gamma_V := \mathrm{E}[V], \tag{17}$$

where $g_V(z,X) = \mathrm{E}_P[V|Z=z,X]$ and V denotes a variable whose identity will change with the context. Specifically, we shall vary V over the set $\mathcal{V}_u$:

$$V \in \mathcal{V}_u := (V_{uj})_{j=1}^5 := \{Y_u, \mathbf{1}_0(D)Y_u, \mathbf{1}_0(D), \mathbf{1}_1(D)Y_u, \mathbf{1}_1(D)\}. \tag{18}$$

Given the definition of $\alpha_V(z) = \mathrm{E}_P[g_V(z,X)]$, it is clear that $g_V(z,X)$ will play an important role in estimating $\alpha_V(z)$. A related function that will play an important role in forming a robust estimation strategy is the propensity score $m_Z : \mathcal{Z}\mathcal{X} \to \mathbb{R}$ defined by

$$m_Z(z,x) := \mathrm{P}_P[Z=z|X=x]. \tag{19}$$

We will denote other potential values for the functions g_V and m_Z by the parameters g and m, respectively. A first approach to estimating $\alpha_V(z)$ is to try to recover g_V and m_Z directly using high-dimensional modelling and estimation methods.

As a second approach, we can further decompose g_V as

$$g_V(z,x) = \sum_{d=0}^{1} e_V(d,z,x) l_D(d,z,x), \tag{20}$$

where the regression functions e_V and l_D, mapping the support $\mathcal{D}\mathcal{Z}\mathcal{X}$ of (D,Z,X) to the real line, are defined by

$$e_V(d,z,x) := \mathrm{E}_P[V|D=d,Z=z,X=x] \text{ and} \tag{21}$$

$$l_D(d,z,x) := \mathrm{P}_P[D=d|Z=z,X=x]. \tag{22}$$

We will denote other potential values for the functions e_V and l_D by the parameters e and l. In this second approach, we can again use high-dimensional methods for modelling and estimating e_V and l_D, and we can then use relation (20) to obtain g_V. Given the resulting g_V and an estimate of m_Z obtained from using high-dimensional methods to model the propensity score, we will then recover $\alpha_V(z)$.

This second approach may be seen as a "special" case of the first. However, this approach could in fact be more principled. For example, if we use linear or generalized linear models to approximate each of the elements e_V, l_D and m_Z, then the implied approximations can strictly nest some coherent models such as the standard dummy endogenous variable model with normal disturbances.[5] This strict nesting of coherent models is more awkward in the first approach which directly approximates g_V using linear or generalized linear forms. Indeed, the "natural" functional form for g_V is not of the linear or generalized linear form but rather is given by the

[5] "Generalized linear" means "linear inside a known link function" in the context of the present paper.

affine aggregation of cross-products shown in (20). While these potential differences exist, we expect to see little quantitative difference between the estimates obtained via either approach if sufficiently flexible functional forms are used. For example, we see little difference between the two approaches in our empirical example.

In the rest of the section we describe the estimation of the reduced-form and structural parameter. The estimation method consists of 3 steps:

(1) Estimation of the predictive relationships m_Z and g_V, or m_Z, l_D and e_V, using high-dimensional nonparametric methods with model selection.
(2) Estimation of the reduced form parameters α_V and γ_V using low bias estimating equations to immunize the reduced form estimators to imperfect model selection in the first step.
(3) Estimation of the structural parameters and effects via plug-in rule.

3.1. **First Step: Modeling and Estimating Regression Function g_V, m_Z, l_D, and e_V in a Data-Rich Environment.** In this section, we elaborate the two strategies that we introduced above.

Strategy 1. We first discuss direct estimation of g_V and m_Z, which corresponds to the first strategy suggested in the previous subsection. Since the functions are unknown and potentially complicated, we use generalized linear combinations of a large number of control terms

$$f(X) = (f_j(X))_{j=1}^p, \tag{23}$$

to approximate g_V and m_Z. Specifically, we use

$$g_V(z,x) =: \Lambda_V[f(z,x)'\beta_V] + r_V(z,x), \tag{24}$$

$$f(z,x) := ((1-z)f(x)', zf(x)')', \quad \beta_V := (\beta_V(0)', \beta_V(1)')', \tag{25}$$

and

$$m_Z(1,x) =: \Lambda_Z[f(x)'\beta_Z] + r_Z(x), \quad m_Z(0,x) = 1 - \Lambda_Z[f(x)'\beta_Z] - r_Z(x). \tag{26}$$

In these equations, $r_V(z,x)$ and $r_Z(x)$ are approximation errors, and the functions $\Lambda_V(f(z,x)'\beta_V)$ and $\Lambda_Z(f(x)'\beta_Z)$ are generalized linear approximations to the target functions $g_V(z,x)$ and $m_Z(1,x)$. The functions Λ_V and Λ_Z are taken to be known link functions Λ. The most common example is the linear link $\Lambda(u) = u$. When the response variables V, Z, and D are binary, we may also use the logistic link $\Lambda(u) = \Lambda_0(u) = e^u/(1+e^u)$ and its complement $1 - \Lambda_0(u)$ or the probit link $\Lambda(u) = \Phi(u) = (2\pi)^{-1/2}\int_{-\infty}^u e^{-z^2/2}dz$ and its complement $1 - \Phi(u)$. For clarity, we use links from the finite set $\mathcal{L} = \{\mathrm{Id}, \Phi, 1-\Phi, \Lambda_0, 1-\Lambda_0\}$ where Id is the identity (linear) link.

In order to allow for a flexible specification and incorporation of pertinent confounding factors, we allow for the dictionary of controls, denoted $f(X)$, to be "rich" in the sense that its dimension $p = p_n$ may be large relative to the sample size. Specifically, our results require only that

$$\log p = o(n^{1/3})$$

along with other technical conditions. High-dimensional regressors $f(X)$ could arise for different reasons. For instance, the list of available variables could be large, i.e. $f(X) = X$ as in e.g. Koenker (1988). It could also be that many technical controls are present; i.e. the list $f(X) = (f_j(X))_{j=1}^p$ could be composed of a large number of transformations of elementary variables X such as B-splines, dummies, polynomials, and various interactions as, e.g., in Newey (1997), Tsybakov (2009), and Wasserman (2006). The functions f_j forming the dictionary can depend on n, but we suppress this dependence.

Having very many controls $f(X)$ creates a challenge for estimation and inference. A useful condition that makes it possible to perform constructive estimation and inference in such cases is termed approximate sparsity or simply sparsity. Sparsity imposes that there exist approximations of the form given in (24)-(26) that require only a small number of non-zero coefficients to render the approximation errors small relative to estimation error. More formally, sparsity relies on two conditions. First, there must exist β_V and β_Z such that, for all $V \in \mathcal{V} := \{\mathcal{V}_u : u \in \mathcal{U}\}$,

$$\|\beta_V\|_0 + \|\beta_Z\|_0 \leqslant s. \tag{27}$$

That is, there are at most $s = s_n \ll n$ components of $f(Z, X)$ and $f(X)$ with nonzero coefficient in the approximations to g_V and m_Z. Second, the sparsity condition requires that the size of the resulting approximation errors is small compared to the conjectured size of the estimation error; namely, for all $V \in \mathcal{V}$,

$$\{\mathrm{E}_P[r_V^2(Z, X)]\}^{1/2} + \{\mathrm{E}_P[r_Z^2(X)]\}^{1/2} \lesssim \sqrt{s/n}. \tag{28}$$

Note that the size of the approximating model $s = s_n$ can grow with n just as in standard series estimation, subject to the rate condition

$$s^2 \log^3(p \vee n)/n \to 0.$$

This condition ensures that the functions g_V and m_Z are estimable at the $o(n^{-1/4})$ rates and are used to derive asymptotic normality results for the structural and reduced-form parameter estimates. This condition can be substantially relaxed if sample splitting methods are used.

The high-dimensional-sparse-model framework outlined above extends the standard framework in the program evaluation literature which assumes both that the identities of the relevant controls are known and that the number of such controls s is much smaller than the sample size. Instead, we assume that there are many, p, potential controls of which at most s controls suffice to achieve a desirable approximation to the unknown functions g_V and m_Z; and we allow the identity of these controls to be unknown. Relying on this assumed sparsity, we use selection methods to choose approximately the right set of controls.

Current estimation methods that exploit approximate sparsity employ different types of regularization aimed at producing estimators that theoretically perform well in high-dimensional settings while remaining computationally tractable. Many widely used method are based on ℓ_1-penalization. The Lasso method is one such commonly used approach that adds a penalty for the weighted sum of the absolute values of model parameters to the usual objective function of

an M-estimator. A related approach is the Post-Lasso method which performs re-estimation of the model after selection of variables by Lasso. These methods are discussed at length in recent papers and review articles; see, for example, Belloni, Chernozhukov, and Hansen (2013). Rather than providing specifics of these methods here, we specify detailed implementation algorithms in a supplementary appendix.

In the following, we outline the general features of the Lasso and Post-Lasso methods focusing on estimation of g_V. Given the data $(\tilde{Y}_i, \tilde{X}_i)_{i=1}^n = (V_i, f(Z_i, X_i))_{i=1}^n$, the Lasso estimator $\widehat{\beta}_V$ solves

$$\widehat{\beta}_V \in \arg\min_{\beta \in \mathbb{R}^{2p}} \Big(\mathbb{E}_n[M(\tilde{Y}, \tilde{X}'\beta)] + \frac{\lambda}{n} \|\widehat{\Psi}\beta\|_1 \Big), \tag{29}$$

where $\widehat{\Psi} = \mathrm{diag}(\widehat{l}_1, \ldots, \widehat{l}_p)$ is a diagonal matrix of data-dependent penalty loadings, $M(y,t) = (y-t)^2/2$ in the case of linear regression, and $M(y,t) = 1(y=1)\log\Lambda(t) + 1(y=0)\log(1-\Lambda(t))$ in the case of binary regression. In the binary case, the link function Λ could be logistic or probit. The penalty level, λ, and loadings, $\widehat{l}_j$ $j = 1, ..., p$, are selected to guarantee good theoretical properties of the method. We provide theoretical choices and further detail regarding implementation in Section 5. A key consideration in this paper is that the penalty level needs to be set to account for the fact that we will be simultaneously estimating potentially a *continuum* of Lasso regressions since our V varies over the list $\mathcal{V}_u$ with u varying over the index set $\mathcal{U}$.

The post-Lasso method uses $\widehat{\beta}_V$ solely as a model selection device. Specifically, it makes use of the labels of the regressors with non-zero estimated coefficients,

$$\widehat{I}_V = \mathrm{support}(\widehat{\beta}_V).$$

The Post-Lasso estimator is then a solution to

$$\tilde{\beta}_V \in \arg\min_{\beta \in \mathbb{R}^p} \Big(\mathbb{E}_n[M(\tilde{Y}, \tilde{X}'\beta)] : \beta_j = 0, j \notin \widehat{I}_V \Big). \tag{30}$$

A main contribution of this paper is establishing that the estimator $\widehat{g}_V(Z,X) = \Lambda(f(Z,X)'\bar{\beta}_V)$ of the regression function $g_V(Z,X)$, where $\bar{\beta}_V = \widehat{\beta}_V$ or $\bar{\beta}_V = \tilde{\beta}_V$, achieve the near oracle rate of convergence $\sqrt{(s\log p)/n}$ and maintain desirable theoretic properties, while allowing for a *continuum* of response variables.

Estimation of m_Z proceeds similarly. The Lasso estimator $\widehat{\beta}_Z$ and Post-Lasso estimators $\tilde{\beta}_Z$ are defined analogously to $\widehat{\beta}_V$ and $\tilde{\beta}_V$ using the data $(\tilde{Y}_i, \tilde{X}_i)_{i=1}^n = (Z_i, f(X_i))_{i=1}^n$. As with the estimator $\widehat{g}_V(Z,X)$, the estimator $\widehat{m}_Z(1,X) = \Lambda_Z(f(X)'\bar{\beta}_Z)$ of $m_Z(X)$, with $\bar{\beta}_Z = \widehat{\beta}_Z$ or $\bar{\beta}_Z = \tilde{\beta}_Z$, achieves the near oracle rate or convergence $\sqrt{(s\log p)/n}$ and has other good theoretic properties. The estimator of $\widehat{m}_Z(0,X)$ is then given by $1 - \widehat{m}_Z(1,X)$.

Strategy 2. The second strategy we consider involves modeling and estimating m_Z as above via (26) while modeling g_V through its disaggregation into the parts e_V and l_D via (20). We model each of the unknown parts of e_V and l_D using the same approach as in Strategy 1.[6]

[6]Upon conditioning on $D = d$ some parts become known; e.g., $e_{1_d(D)Y}(d', x, z) = 0$ if $d \neq d'$ and $e_{1_d(D)}(d', x, z) = 1$ if $d = d'$.

Specifically, we model the conditional expectation of V given D, Z, and X by

$$e_V(d,z,x) =: \Gamma_V[f(d,z,x)'\theta_V] + \varrho_V(d,z,x), \tag{31}$$

$$f(d,z,x) := ((1-d)f(z,x)', df(z,x)')', \tag{32}$$

$$\theta_V := (\theta_V(0,0)', \theta_V(0,1)', \theta_V(1,0)', \theta_V(1,1)')'. \tag{33}$$

We model the conditional probability of D taking on 1 or 0, given Z and X by

$$l_D(1,z,x) =: \Gamma_D[f(z,x)'\theta_D] + \varrho_D(z,x), \tag{34}$$

$$l_D(0,z,x) = 1 - \Gamma_D[f(z,x)'\theta_D] - \varrho_D(z,x), \tag{35}$$

$$f(z,x) := ((1-z)f(x)', zf(x)')', \tag{36}$$

$$\theta_D := (\theta_D(0)', \theta_D(1)')'. \tag{37}$$

Here $\varrho_V(d,z,x)$ and $\varrho_D(z,x)$ are approximation errors, and the functions $\Gamma_V(f(d,z,x)'\theta_V)$ and $\Gamma_D(f(z,x)'\theta_D)$ are generalized linear approximations to the target functions $e_V(d,z,x)$ and $l_D(1,z,x)$. The functions Γ_V and Γ_D are taken to be known link functions $\Lambda \in \mathcal{L}$ as in the previous strategy.

As in the first strategy, we maintain approximate sparsity in the modeling framework. We assume that there exist β_Z, θ_V and θ_D such that, for all $V \in \mathcal{V}$,

$$\|\theta_V\|_0 + \|\theta_D\|_0 + \|\beta_Z\|_0 \leqslant s. \tag{38}$$

That is, there are at most $s = s_n \ll n$ components of $f(D,Z,X)$, $f(Z,X)$ and $f(X)$ with nonzero coefficient in the approximations to e_V, l_D and m_Z. The sparsity condition also requires the size of the approximation errors to be small compared to the conjectured size of the estimation error: For all $V \in \mathcal{V}$, we assume

$$\{\{\mathrm{E}_P[\varrho_V^2(D,Z,X)]\}^{1/2} + \{\mathrm{E}_P[\varrho_D^2(Z,X)] + \mathrm{E}_P[r_Z^2(X)]\}^{1/2}\}^{1/2} \lesssim \sqrt{s/n}. \tag{39}$$

Note that the size of the approximating model $s = s_n$ can grow with n just as in standard series estimation as long as $s^2 \log^3(p \vee n)/n \to 0$.

We proceed with the estimation of e_V and l_D analogously to the approach outlined in Strategy 1. The Lasso estimator $\widehat{\theta}_V$ and Post-Lasso estimators $\tilde{\theta}_V$ are defined analogously to $\widehat{\beta}_V$ and $\tilde{\beta}_V$ using the data $(\tilde{Y}_i, \tilde{X}_i)_{i=1}^n = (V_i, f(D_i,Z_i,X_i))_{i=1}^n$ and the link function $\Lambda = \Gamma_V$. The estimator $\widehat{e}_V(D,Z,X) = \Gamma_V[f(D,Z,X)'\bar{\theta}_V]$, with $\bar{\theta}_V = \widehat{\theta}_V$ or $\bar{\theta}_V = \widehat{\theta}_V$, have near oracle rates of convergence, $\sqrt{(s\log p)/n}$, and other desirable properties. The Lasso estimator $\widehat{\theta}_D$ and Post-Lasso estimators $\tilde{\theta}_D$ are also defined analogously to $\widehat{\beta}_V$ and $\tilde{\beta}_V$ using the data $(\tilde{Y}_i, \tilde{X}_i)_{i=1}^n = (D_i, f(Z_i,X_i))_{i=1}^n$ and the link function $\Lambda = \Gamma_D$. Again, the estimator $\widehat{l}_D(Z,X) = \Gamma_D[f(Z,X)'\bar{\theta}_D]$ of $l_D(Z,X)$, where $\bar{\theta}_D = \widehat{\theta}_D$ or $\bar{\theta}_D = \tilde{\theta}_D$, have good theoretical properties including the near oracle rate of convergence, $\sqrt{(s\log p)/n}$. The resulting estimator for $g_V(z,X)$ is then

$$\widehat{g}_V(z,x) = \sum_{d=0}^{1} \widehat{e}_V(d,z,x)\widehat{l}_D(d,z,x). \tag{40}$$

3.2. **Second Step: Robust Estimation of Reduced-Form Parameters $\alpha_V(z)$ and γ_V.** Estimation of the key quantities $\alpha_V(z)$ will make heavy use of "low-bias" moment functions as defined in (2). These moment functions are closely tied to efficient influence functions, where efficiency is in the sense of locally minimax semi-parametric efficiency. The use of these functions will deliver robustness with respect to the irregularity of the post-selection and penalized estimators needed to manage high-dimensional data. The use of these functions also automatically delivers semi-parametric efficiency for estimating and performing inference on the reduced-form parameters and their smooth transformations – the structural parameters.

The efficient influence function and low-bias moment function for $\alpha_V(z)$ for $z \in \mathcal{Z} = \{0, 1\}$ are given respectively by

$$\psi^{\alpha}_{V,z}(W) := \psi^{\alpha}_{V,z,g_V,m_Z}(W, \alpha_V(z)) \quad \text{and} \tag{41}$$

$$\psi^{\alpha}_{V,z,g,m}(W, \alpha) := \frac{1(Z = z)(V - g(z, X))}{m(z, X)} + g(z, X) - \alpha. \tag{42}$$

The efficient influence function was derived by Hahn (1998); they were also used by Cattaneo (2010) in the series context (with $p \ll n$) and Rothe and Firpo (2013) in the kernel context. The efficient influence function and the moment function for γ_V are trivially given by

$$\psi^{\gamma}_V(W) := \psi^{\gamma}_V(W, \gamma_V), \text{ and } \psi^{\gamma}_V(W, \gamma) := V - \gamma. \tag{43}$$

We then define the estimator of the reduced-form parameters $\alpha_V(z)$ and $\gamma_V(z)$ as solutions $\alpha = \widehat{\alpha}_V(z)$ and $\gamma = \widehat{\gamma}_V$ to the equations

$$\mathbb{E}_n[\psi^{\alpha}_{V,z,\widehat{g}_V,\widehat{m}_Z}(W, \alpha)] = 0, \quad \mathbb{E}_n[\psi^{\gamma}_V(W, \gamma)] = 0, \tag{44}$$

where $\widehat{g}_V$ and $\widehat{m}_Z$ are constructed as in the previous section. Note that $\widehat{g}_V$ may be constructed via either Strategy 1 or Strategy 2. We apply this procedure to each variable name $V \in \mathcal{V}_u$ and obtain the estimator

$$\widehat{\rho}_u := \left(\{\widehat{\alpha}_V(0), \widehat{\alpha}_V(1), \widehat{\gamma}_V\}\right)_{V \in \mathcal{V}_u} \quad \text{of} \quad \rho_u := \left(\{\alpha_V(0), \alpha_V(1), \gamma_V\}\right)_{V \in \mathcal{V}_u}.^{7} \tag{45}$$

The estimator and the estimand are vectors in $\mathbb{R}^{d_\rho}$ with dimension $d_\rho = 3 \times \dim \mathcal{V}_u = 15$.

In the next section, we formally establish a principal result which shows that

$$\sqrt{n}(\widehat{\rho}_u - \rho_u) \rightsquigarrow N(0, \mathrm{Var}_P(\psi^{\rho}_u)), \quad \psi^{\rho}_u := (\{\psi^{\alpha}_{V,0}, \psi^{\alpha}_{V,1}, \psi^{\gamma}_V\})_{V \in \mathcal{V}_u}, \tag{46}$$
$$\text{uniformly in } P \in \mathcal{P}_n,$$

where $\mathcal{P}_n$ is a rich set of data generating processes P. The notation "$Z_{n,P} \rightsquigarrow Z_P$ uniformly in $P \in \mathcal{P}_n$" is defined formally in the Appendix and can be read as "$Z_{n,P}$ is approximately distributed as Z_P uniformly in $P \in \mathcal{P}_n$." This usage corresponds to the usual notion of asymptotic distribution extended to handle uniformity in P. Here $\mathcal{P}_n$ is a "rich" set of data generating processes P which includes cases where perfect model selection is impossible theoretically.

We then stack all the reduced form estimators and the estimands over $u \in \mathcal{U}$ as

$$\widehat{\rho} = (\widehat{\rho}_u)_{u \in \mathcal{U}} \quad \text{and} \quad \rho = (\rho_u)_{u \in \mathcal{U}},$$

giving rise to the empirical reduced-form process $\widehat{\rho}$ and the reduced-form functional ρ. We establish that $\sqrt{n}(\widehat{\rho}-\rho)$ is asymptotically Gaussian: In $\ell^\infty(\mathcal{U})^{d_\rho}$,

$$\sqrt{n}(\widehat{\rho}-\rho) \rightsquigarrow Z_P := (\mathbb{G}_P \psi_u^\rho)_{u\in\mathcal{U}}, \text{ uniformly in } P \in \mathcal{P}_n \tag{47}$$

where $\mathbb{G}_P$ denotes the P-Brownian bridge (van der Vaart and Wellner, 1996). This result contains (46) as a special case and again allows $\mathcal{P}_n$ to be a "rich" set of data generating processes P that includes cases where perfect model selection is impossible theoretically. Importantly, this result verifies that the functional central limit theorem applies to the reduced-form estimators in the presence of possible model selection mistakes.

Since some of our objects of interest are complicated, inference can be facilitated by a multiplier bootstrap method. We define $\widehat{\rho}^* = (\widehat{\rho}_u^*)_{u\in\mathcal{U}}$, a bootstrap draw of $\widehat{\rho}$, via

$$\sqrt{n}(\widehat{\rho}_u^* - \widehat{\rho}_u) = n^{-1/2}\sum_{i=1}^n \xi_i \widehat{\psi}_u^\rho(W_i). \tag{48}$$

Here $(\xi_i)_{i=1}^n$ are i.i.d. copies of ξ which are independently distributed from the data $(W_i)_{i=1}^n$ and whose distribution does not depend on P. We also impose that

$$\mathrm{E}[\xi]=0, \quad \mathrm{E}[\xi^2]=1, \quad \mathrm{E}[\exp(|\xi|)]<\infty.$$

Examples of ξ include (a) $\xi=\mathcal{E}-1$, where $\mathcal{E}$ is standard exponential random variable, (b) $\xi=\mathcal{N}$, where $\mathcal{N}$ is standard normal random variable, and (c) $\xi=\mathcal{N}_1/\sqrt{2}+(\mathcal{N}_2^2-1)/2$, where $\mathcal{N}_1$ and $\mathcal{N}_2$ are mutually independent standard normal random variables.[8] Methods (a), (b), and (c) correspond respectively to the Bayesian bootstrap (e.g., Hahn (1997), Chamberlain and Imbens (2003)), the Gaussian multiplier method (e.g, van der Vaart and Wellner (1996)), and the wild bootstrap method (Mammen, 1993).[9] $\widehat{\psi}_u^\rho$ in (48) is an estimator of the influence function ψ_u^ρ defined via the plug-in rule:

$$\widehat{\psi}_u^\rho = (\widehat{\psi}_V^\rho)_{V\in\mathcal{V}_u}, \quad \widehat{\psi}_V^\rho(W) := \{\psi^\alpha_{V,0,\widehat{g}_V,\widehat{m}_Z}(W,\widehat{\alpha}_V(0)), \psi^\alpha_{V,1,\widehat{g}_V,\widehat{m}_Z}(W,\widehat{\alpha}_V(1)), \psi^\gamma_V(W,\widehat{\gamma}_V)\}. \tag{49}$$

Note that this bootstrapping is computationally efficient since it does not involve recomputing the influence functions $\widehat{\psi}_u^\rho$. Each new draw of $(\xi_i)_{i=1}^n$ generates a new draw of $\widehat{\rho}^*$ holding the data and the estimates of the influence functions fixed. This method simply amounts to resampling the first-order approximations to the estimators. Here we build upon the prior uses of this or similar methods in low-dimensional setting include Hansen (1996) and Kline and Santos (2012).

We establish that that the bootstrap law $\sqrt{n}(\widehat{\rho}^*-\widehat{\rho})$ is uniformly asymptotically valid: In the metric space $\ell^\infty(\mathcal{U})^{d_\rho}$, both unconditionally and conditionally on the data,

$$\sqrt{n}(\widehat{\rho}^*-\widehat{\rho}) \rightsquigarrow_B Z_P, \text{ uniformly in } P\in\mathcal{P}_n,$$

where $\rightsquigarrow_B$ denotes the convergence of the bootstrap law conditional on the data, as defined in the Appendix.

[8] We do not consider the nonparametric bootstrap, which corresponds to using multinomial multipliers ξ, to reduce the length of the paper; but we note that it is possible to show that it is also valid in the present setting.

[9] The motivation for method (c) is that it is able to match 3 moments since $\mathrm{E}[\xi^2]=\mathrm{E}[\xi^3]=1$. Methods (a) and (b) do not satisfy this property since $\mathrm{E}[\xi^2]=1$ but $\mathrm{E}[\xi^3]\neq 1$ for these approaches.

3.3. **Step 3: Robust Estimation of the Structural Parameters.** All structural parameters we consider take the form of smooth transformations of reduced-form parameters:

$$\Delta = (\Delta_q)_{q\in\mathcal{Q}}, \text{ where } \Delta_q := \phi(\rho)(q), \ q \in \mathcal{Q}. \tag{50}$$

The structural parameters may themselves carry an index $q \in \mathcal{Q}$ that can be different from u; for example, the structural quantile treatment effects are indexed by the quantile index $q \in (0,1)$. This formulation includes as special cases all the structural functions we previously mentioned. We estimate these quantities by the plug-in rule. We establish the asymptotic behavior of these estimators and the validity of the bootstrap as a corollary from the results outlined in Section 3.2 and the functional delta method.

For the application of the functional delta method, we require that the functionals be Hadamard differentiable – tangential to a subset that contains realizations of Z_P for all $P \in \mathcal{P}_n$ – with derivative map $h \mapsto \phi'_\rho(h) = (\phi'_{\rho,q}(h))_{q\in\mathcal{Q}}$. We define the estimators and their bootstrap versions as $\widehat{\Delta} = (\widehat{\Delta}_q)_{q\in\mathcal{Q}}$ and $\widehat{\Delta}^* = (\widehat{\Delta}^*_q)_{q\in\mathcal{Q}}$, where

$$\widehat{\Delta}_q := \phi\left(\widehat{\rho}\right)(q), \quad \widehat{\Delta}^*_q := \phi\left(\widehat{\rho}^*_u\right)(q). \tag{51}$$

We establish that these estimators are asymptotically Gaussian

$$\sqrt{n}(\widehat{\Delta} - \Delta) \rightsquigarrow \phi'_\rho(Z_P), \text{ uniformly in } P \in \mathcal{P}_n, \tag{52}$$

and that the bootstrap consistently estimates their large sample distribution:

$$\sqrt{n}(\widehat{\Delta}^* - \widehat{\Delta}) \rightsquigarrow_B \phi'_\rho(Z_P), \text{ uniformly in } P \in \mathcal{P}_n. \tag{53}$$

These results can be used to construct simultaneous confidence bands on Δ.

4. Theory of Estimation and Inference on Local Treatment Effects Functionals

Consider fixed sequences of positive numbers $\delta_n \searrow 0$, $\epsilon_n \searrow 0$, $\Delta_n \searrow 0$, $\ell_n \to \infty$, and $1 \leqslant K_n < \infty$, and positive constants c, C, and $c' < 1/2$ which will not vary with P. P is allowed to vary in the set $\mathcal{P}_n$ of probability measures, termed "data-generating processes", where $\mathcal{P}_n$ is typically a weakly increasing in n set.

Assumption 2 (Basic Assumptions)**.** *(i) For each $n \geqslant 1$, our data will consist of i.i.d. copies $(W_i)_{i=1}^n$ of the stochastic process $W = ((Y_u)_{u\in\mathcal{U}}, X, Z, D)$ defined on the probability space $(S, \mathcal{S}, P)$, where $P \in \mathcal{P}_n$, and the collection $(Y_u)_{u\in\mathcal{U}}$ is suitably measurable, namely image-admissible Suslin (Dudley, 1999, p. 186). Let*

$$V_u := (V_{uj})_{j\in\mathcal{J}} := \{Y_u, \mathbf{1}_0(D)Y_u, \mathbf{1}_0(D), \mathbf{1}_1(D)Y_u, \mathbf{1}_1(D)\}$$

where $\mathcal{J} = \{1, ..., 5\}$ and $\mathcal{V} = (V_u)_{u\in\mathcal{U}}$. (ii) For $\mathcal{P} := \cup_n \mathcal{P}_n$, the map $u \mapsto Y_u$ obeys the uniform continuity property:

$$\lim_{\epsilon\searrow 0} \sup_{P\in\mathcal{P}} \sup_{d_{\mathcal{U}}(u,\bar{u})\leqslant\epsilon} \|Y_u - Y_{\bar{u}}\|_{P,2} = 0, \ \sup_{P\in\mathcal{P}} \mathrm{E}_P \sup_{u\in\mathcal{U}} |Y_u|^{2+c} \leqslant \infty,$$

for each $j \in \mathcal{J}$, where the supremum is taken over $u, \bar{u} \in \mathcal{U}$, and $\mathcal{U}$ is a totally bounded metric space equipped with the metric $d_{\mathcal{U}}$. The uniform ϵ covering entropy of $(Y_u, u \in \mathcal{U})$ is bounded by $C \log(\mathrm{e}/\epsilon) \vee 0$. (iii) For each $P \in \mathcal{P}$, the conditional probability of $Z = 1$ given X is bounded away from zero or one: $P(c' \leqslant m_Z(z, X) \leqslant 1 - c') = 1$; the instrument Z has a non-trivial impact on D, namely $P(c' \leqslant l_D(1, 1, X) - l_D(1, 0, X)) = 1$; and the regression function g_V is bounded, $\|g_V\|_{P,\infty} < \infty$ for all $V \in \mathcal{V}$.

This assumption implies that the set of functions $(\psi_u^\rho)_{u \in \mathcal{U}}$, where $\psi_u^\rho := (\{\psi_{V,0}^\alpha, \psi_{V,1}^\alpha, \psi_V^\gamma\})_{V \in \mathcal{V}_u}$, is P-Donsker uniformly in $\mathcal{P}$. That is, it implies

$$Z_{n,P} \rightsquigarrow Z_P \quad \text{in } \ell^\infty(\mathcal{U})^{d_\rho}, \text{ uniformly in } P \in \mathcal{P}, \tag{54}$$

where

$$Z_{n,P} := (\mathbb{G}_n \psi_u^\rho)_{u \in \mathcal{U}} \text{ and } Z_P := (\mathbb{G}_P \psi_u^\rho)_{u \in \mathcal{U}}, \tag{55}$$

with $\mathbb{G}_P$ denoting the P-Brownian bridge (van der Vaart and Wellner, 1996), and Z_P having bounded, uniformly continuous paths uniformly in $P \in \mathcal{P}$:

$$\sup_{P \in \mathcal{P}} \mathrm{E}_P \sup_{u \in \mathcal{U}} \|Z_P(u)\| < \infty, \quad \lim_{\varepsilon \searrow 0} \sup_{P \in \mathcal{P}} \mathrm{E}_P \sup_{d_{\mathcal{U}}(u, \tilde{u}) \leqslant \varepsilon} \|Z_P(u) - Z_P(\tilde{u})\| = 0. \tag{56}$$

Other assumptions will be specific to the strategy taken.

Assumption 3 (Approximate Sparsity for Strategy 1)**.** *Under each $P \in \mathcal{P}_n$ and for each $n \geqslant n_0$, uniformly for all $V \in \mathcal{V}$ the following conditions hold: (i) The approximations (24)-(26) hold with the link functions Λ_V and Λ_Z belonging to the set $\mathcal{L}$, the sparsity condition holding, $\|\beta_V\|_0 + \|\beta_Z\|_0 \leqslant s$, the approximation errors satisfying $\|r_V\|_{P,2} + \|r_Z\|_{P,2} \leqslant \delta_n n^{-1/4}$, $\|r_V\|_{P,\infty} + \|r_Z\|_{P,\infty} \leqslant \epsilon_n$, and the sparsity index and the number of terms p in vector $f(X)$ obeying $s^2 \log^3(p \vee n)/n \leqslant \delta_n$. (ii) There are estimators $\bar{\beta}_V$ and $\bar{\beta}_Z$ such that, with probability no less than $1 - \Delta_n$, the estimation errors satisfy $\|f(Z, X)'(\bar{\beta}_V - \beta_V)\|_{\mathbb{P}_n,2} + \|f(X)'(\bar{\beta}_Z - \beta_Z)\|_{\mathbb{P}_n,2} \leqslant \delta_n n^{-1/4}$, $K_n \|\bar{\beta}_Z - \beta_Z\|_1 + K_n \|\bar{\beta}_Z - \beta_Z\|_1 \leqslant \delta_n$; the estimators are sparse such that $\|\bar{\beta}_V\|_0 + \|\bar{\beta}_Z\|_0 \leqslant Cs$; and the empirical and populations norms induced by the Gram matrix formed by $(f(X_i))_{i=1}^n$ are equivalent on sparse subsets, $\sup_{\|\delta\|_0 \leqslant \ell_n s} |\|f(X)'\delta\|_{\mathbb{P}_n,2} / \|f(X)'\delta\|_{P,2} - 1| \leqslant \delta_n$. (iii) The following boundedness conditions hold: $\|\|f(X)\|_\infty\|_{P,\infty} \leqslant K_n$ and $\|V\|_{P,\infty} \leqslant C$.*

Comment 4.1. These conditions are simple intermediate-level conditions which encode both the approximate sparsity of the models as well as some reasonable behavior on the sparse estimators of m_Z and g_V. Sufficient conditions for the equivalence between empirical and population norms are given in Belloni, Chernozhukov, and Hansen (2011). The boundedness conditions are made to simplify arguments, and they could be removed at the cost of more complicated proofs and more stringent side conditions. ■

Assumption 4 (Approximate Sparsity for Strategy 2)**.** *Under each $P \in \mathcal{P}_n$ and for all $n \geqslant n_0$, uniformly for all $V \in \mathcal{V}$ the following conditions hold: (i) The approximations (31)-(37) and (26) apply with the link functions Γ_V, Γ_D and Λ_Z belonging to the set $\mathcal{L}$, the sparsity condition $\|\theta_V\|_0 + \|\theta_D\|_0 + \|\beta_Z\|_0 \leqslant s$ holding, the approximation errors satisfying $\|\varrho_D\|_{P,2} + \|\varrho_V\|_{P,2} +$*

$\|r_Z\|_{P,2} \leqslant \delta_n n^{-1/4}$ *and* $\|\varrho_D\|_{P,\infty} + \|\varrho_V\|_{P,\infty} + \|r_Z\|_{P,\infty} \leqslant \epsilon_n$*, and the sparsity index* s *and the number of terms* p *obeying* $s^2 \log^3(p \vee n)/n \leqslant \delta_n$*. (ii) There are estimators* $\bar{\theta}_V$*,* $\bar{\theta}_D$*, and* $\bar{\beta}_Z$ *such that, with probability no less than* $1 - \Delta_n$*, the estimation errors satisfy* $\|f(D, Z, X)'(\bar{\theta}_V - \theta_V)\|_{\mathbb{P}_n,2} + \|f(Z, X)'(\bar{\theta}_D - \theta_D)\|_{\mathbb{P}_n,2} + \|f(X)'(\bar{\beta}_Z - \beta_Z)\|_{\mathbb{P}_n,2} \leqslant \delta_n n^{-1/4}$ *and* $K_n\|\bar{\theta}_V - \theta_V\|_1 + K_n\|\bar{\theta}_D - \theta_D\|_1 + K_n\|\bar{\beta}_Z - \beta_Z\|_1 \leqslant \epsilon_n$*; the estimators are sparse such that* $\|\bar{\theta}_V\|_0 + \|\bar{\theta}_D\|_0 + \|\bar{\beta}_Z\|_0 \leqslant Cs$*; and the empirical and populations norms induced by the Gram matrix formed by* $(f(X_i))_{i=1}^n$ *are equivalent on sparse subsets,* $\sup_{\|\delta\|_0 \leqslant \ell_n s} |\|f(X)'\delta\|_{\mathbb{P}_n,2}/\|f(X)'\delta\|_{P,2} - 1| \leqslant \delta_n$*. (iii) The following boundedness conditions hold:* $\|\|f(X)\|_\infty\|_{P,\infty} \leqslant K_n$ *and* $\|V\|_{P,\infty} \leqslant C$*.*

Under the stated assumptions, the empirical reduced form process $\widehat{Z}_{n,P} := \sqrt{n}(\widehat{\rho} - \rho)$ defined by (45) obeys the following laws.

Theorem 4.1 (**Uniform Gaussianity of the Reduced-Form Parameter Process**)**.** *Under Assumptions 2 and 3 or 2 and 4 holding, the reduced-form empirical process admits a linearization, namely*

$$\widehat{Z}_{n,P} := \sqrt{n}(\widehat{\rho} - \rho) = Z_{n,P} + o_P(1) \quad \text{in } \ell^\infty(\mathcal{U})^{d_\rho}, \text{ uniformly in } P \in \mathcal{P}_n. \tag{57}$$

The process is also asymptotically Gaussian, namely

$$\widehat{Z}_{n,P} \rightsquigarrow Z_P \quad \text{in } \ell^\infty(\mathcal{U})^{d_\rho}, \text{ uniformly in } P \in \mathcal{P}_n, \tag{58}$$

where Z_P *is defined in (55) and its paths obey the property (56) .*

Another main result of this section shows that the bootstrap law

$$\widehat{Z}^*_{n,P} = \sqrt{n}(\widehat{\rho}^* - \widehat{\rho})$$

provides a valid approximation to the large sample law of $\sqrt{n}(\widehat{\rho} - \rho)$.

Theorem 4.2 (**Validity of Multiplier Bootstrap for Inference on Reduced-Form Parameters**)**.** *Under Assumptions 2 and 3 or 2 and 4, the bootstrap law consistently approximates the large sample law* Z_P *of* $Z_{n,P}$ *uniformly in* $P \in \mathcal{P}_n$*, namely,*

$$\widehat{Z}^*_{n,P} \rightsquigarrow_B Z_P \quad \text{in } \ell^\infty(\mathcal{U})^{d_\rho}, \text{ uniformly in } P \in \mathcal{P}_n. \tag{59}$$

The notation $\rightsquigarrow_B$ is defined in the Appendix and just means the usual notion of weak convergence in probability of the bootstrap law.

We derive the large sample distribution and validity of the multiplier bootstrap for structural functionals via the functional delta method, which we modify to handle uniformity with respect to the underlying dgp P. We shall need the following assumption on the structural functionals.

Assumption 5 (Uniform Hadamard Differentiability of Structural Functionals)**.** *Suppose that for each* $P \in \mathcal{P}$*,* $\rho = \rho_P$ *is an element of a compact subset* $\mathbb{D}_1 \subset \mathbb{D} = \ell^\infty(\mathcal{U})^{d_\rho}$*. Suppose* $\varrho \mapsto \phi(\varrho)$*, a functional of interest mapping* $\mathbb{D}_\rho$ *to* $\ell^\infty(\mathcal{Q})$*, is Hadamard differentiable in* ϱ *with derivative* ϕ'_ϱ*, tangentially to* $\mathbb{D}_0 = UC(\mathcal{U})^{d_\rho}$*, uniformly in* $\rho \in \mathbb{D}_1$*, and that the mapping* $(\varrho, h) \mapsto \phi'_\varrho(h)$ *from* $\mathbb{D}_1 \times \mathbb{D}_0$ *into* $\ell^\infty(\mathcal{Q})$ *is defined and continuous.*

The definition of uniform Hadamard differentiability is given in the appendix.

This assumption holds for all examples of structural parameters listed in Section 2.

The following result gives asymptotic Gaussian law for $\sqrt{n}(\widehat{\Delta} - \Delta)$, the properly normalized structural estimator. It also shows that the bootstrap law of $\sqrt{n}(\widehat{\Delta}^* - \widehat{\Delta})$, computed conditionally on the data, approaches the asymptotic Gaussian law for $\sqrt{n}(\widehat{\Delta} - \Delta)$. The following is the corollary of the previous theorems as well as of a more general result contained in Theorem 5.3.

Corollary 4.1 (**Limit Theory and Validity of Multiplier Bootstrap for Smooth Structural Functionals**). *Under Assumptions 2, 3 or 4, and 5,*

$$\sqrt{n}(\widehat{\Delta} - \Delta) \rightsquigarrow T_P := \phi'_\rho(Z_P), \quad \text{in } \ell^\infty(\mathcal{Q}), \text{ uniformly in } P \in \mathcal{P}_n, \tag{60}$$

where T_P is a zero mean tight Gaussian process, for each $P \in \mathcal{P}$. Moreover,

$$\sqrt{n}(\widehat{\Delta}^* - \widehat{\Delta}) \rightsquigarrow_B T_P := \phi'_\rho(Z_P), \quad \text{in } \ell^\infty(\mathcal{Q}), \text{ uniformly in } P \in \mathcal{P}_n. \tag{61}$$

5. A General Problem of Inference on Function-Valued Parameters with Approximately Sparse Nuisance Functions

In this section we generalize the previous specific framework to a more general setting, where possibly a continuum of target parameters is of interest and the Lasso-type or post-Lasso type methods are used to estimate a continuum of high-dimensional nuisance function. This framework is quite general — in addition to the previous specific framework, it covers a rich variety of modern moment-condition problems in econometrics. We establish a functional central limit theorem for the estimators of the continuum of the target parameters, and also show that it holds uniformly in $P \in \mathcal{P}$, where $\mathcal{P}$ includes a wide range of data-generating processes with approximately sparse continua of nuisance functions. We also establish validity of the multiplier bootstrap for resampling the first order approximations to standardized continua of the estimators, and also establish its uniform validity. Moreover, we establish uniform validity of the functional delta method, using an appropriate strengthening of Hadamard differentiability; and we establish uniform validity of the functional delta method for the multiplier bootstrap for resampling the smooth functionals of continua of the target parameters.

We are interested in a continuum of target parameters indexed by $u \in \mathcal{U} \subset \mathbb{R}^{d_u}$. We denote the true value of the target parameter by

$$\theta^0 = (\theta_u)_{u \in \mathcal{U}}, \text{ where } \theta_u \in \Theta_u \subset \Theta \subset \mathbb{R}^{d_\theta}, \text{ for each } u \in \mathcal{U} \subset \mathbb{R}^{d_u}.$$

We assume that for each $u \in \mathcal{U}$ the true value θ_u is identified as a solution of the following moment condition:

$$\mathrm{E}_P[\psi_u(W_u, \theta_u, h_u(Z_u))] = 0, \tag{62}$$

where for each $u \in \mathcal{U}$, vector W_u is a random vector taking values in $\mathcal{W}_u \subset \mathbb{R}^{d_w}$, containing vector Z_u taking values in $\mathcal{Z}_u$ as a subcomponent; the function

$$\psi_u : \mathcal{W}_u \times \Theta_u \times T_u \to \mathbb{R}^{d_\theta}, \quad (w, \theta, t) \mapsto \psi_u(w, \theta, t) = (\psi_{uj}(w, \theta, t))_{j=1}^{d_\theta}$$

is a measurable map, and the function

$$h_u : \mathcal{Z}_u \mapsto T_u \subset \mathbb{R}^{d_t}, \quad z \mapsto h_u(z) = (h_{um}(z))_{m=1}^{d_t}$$

is another measurable map, the nuisance parameter, possibly infinite-dimensional.

We assume that the continuum of the nuisance functions $(h_u)_{u \in \mathcal{U}}$ is approximately sparse, which can be modelled and estimated using modern regularization and post-selection methods; for example, in the previous sections we described the continuum of Lasso and post-Lasso regressions. We let $\widehat{h}_u = (\widehat{h}_{um})_{m=1}^{d_t}$ denote the estimator of h_u, which obeys conditions stated below. The estimator $\widehat{\theta}_u$ of θ_u is constructed as any approximate ϵ_n-solution in Θ_u to a sample analog of the estimating equation above:

$$\|\mathbb{E}_n[\psi_u(W_u, \widehat{\theta}_u, \widehat{h}_u(Z_u))]\| \leqslant \inf_{\theta \in \Theta_u} \|\mathbb{E}_n[\psi(W_u, \theta, \widehat{h}_u(Z_u))]\| + \epsilon_n, \text{ where } \epsilon_n = o(n^{-1/2}). \tag{63}$$

The key condition needed for regular estimation of θ_u is the orthogonality or immunization condition. This condition can be expressed as follows:

$$\partial_t \mathrm{E}_P[\psi_u(W_u, \theta_u, h_u(Z_u))|Z_u] = 0, \text{ a.s.}, \tag{64}$$

where we use the symbol ∂_t to abbreviate $\frac{\partial}{\partial t}$. [10] This condition holds in the previous setting of inference on policy-relevant treatment effects. The condition above is formulated to cover certain non-smooth cases, for example, in structural and instrumental quantile regression problems.

It is important to construct the moment-functions ψ_u that have this orthogonality property. Generally, if we have a moment function $\tilde{\psi}_u$, which identifies the parameters of interest θ_u but does not have the orthogonality property, we can construct the moment-function ψ_u that has the required orthogonality property by projecting the original function $\tilde{\psi}_u$ onto the orthocomplement of the tangent space for the nuisance functions; see, for example, (van der Vaart and Wellner, 1996; van der Vaart, 1998; Kosorok, 2008).

In what follows, we shall denote by c_0, c, and C some positive constants.

Assumption S1. *For each n, we observe i.i.d. copies $(W_i)_{i=1}^n$ of $W = (W_u)_{u \in \mathcal{U}}$ with law determined by the probability measure $P \in \mathcal{P}_n$. Uniformly for all $n \geqslant n_0$ and $P \in \mathcal{P}_n$, the following conditions hold. (i) The true parameter values θ_u obeys (62) and is interior relative to $\Theta_u \subset \Theta \subset \mathbb{R}^{d_\theta}$, namely there is a ball of fixed positive radius centered at θ_u contained in Θ_u. (ii) For each j, for each $\nu = (\nu_k)_{k=1}^{d_\theta + d_t} = (\theta, t) \in \Theta_u \times T_u$ and $u \in \mathcal{U}$, the map $\nu \mapsto \mathrm{E}_P[\psi_{uj}(W_u, \nu)|Z_u]$ is twice continuously differentiable a.s. with derivatives obeying the integrability conditions specified in S2 below. (iii) The orthogonality condition (64) holds. (iv) The identifiability condition*

[10] The expression $\partial_t \mathrm{E}_P[\psi_u(W_u, \theta_u, h_u(Z_u))|Z_u]$ is understood to be $\partial_t \mathrm{E}_P[\psi_u(W_u, \theta_u, t)|Z_u]|_{t=h_u(Z_u)}$.

holds: $\|\mathrm{E}_P[\psi_u(W_u,\theta,h_u(Z_u))]\| \geqslant 2^{-1}(\|J_u(\theta-\theta_u)\| \vee c_0)$ *for all* $\theta \in \Theta_u$, *where eigenvalues of* $J_u := \partial_\theta \mathrm{E}[\psi_u(W_u,\theta_u,h_u(Z_u))]$ *lie in between* $c>0$ *and* C *for all* $u \in \mathcal{U}$.

The conditions above are mild and standard assumptions for moment condition problems. The identification condition encodes both global identifiability and local identifiability sufficient to get a rate result, not just consistency.

Assumption S2. *Let* $(\mathcal{U}, d_{\mathcal{U}})$ *be a semi-metric space, such that* $\log N(\varepsilon,\mathcal{U},d_{\mathcal{U}}) \leqslant C\log(\mathrm{e}/\varepsilon)$. *Let* $\alpha \in [1,2]$ *and* α_1 *and* α_2 *be some positive constants. Uniformly for all* $n \geqslant n_0$ *and* $P \in \mathcal{P}_n$, *the following holds: (i) The set of functions* $\mathcal{F}_0 = \{\psi_{uj}(W_u,\theta_u,h_u(Z_u)), j \in [d_\theta], u \in \mathcal{U}\}$ *is suitably measurable; the envelope* $F_0 = \sup_{j\in[d_\theta],u\in\mathcal{U},\nu\in\Theta_u\times T_u}|\psi_{uj}(W_u,\nu)|$ *is measurable and obeys* $\|F_0\|_{P,q} \leqslant C$, *were* $q \geqslant 4$ *is a fixed constant, and* $\sup_Q \log N(\epsilon\|F_0\|_{Q,2},\mathcal{F}_0,\|\cdot\|_{Q,2}) \leqslant C\log(\mathrm{e}/\varepsilon)$. *(ii) for all* $j \in [d_\theta]$ *and* $k,r \in [d_\theta+d_t]$,

(a) $\sup_{u\in\mathcal{U},(\nu,\bar\nu)\in(\Theta_u\times T_u)^2} \mathrm{E}_P[(\psi_{uj}(W_u,\nu)-\psi_{uj}(W_u,\bar\nu))^2|Z_u] \leqslant C\|\nu-\bar\nu\|^\alpha$, P-*a.s.*,
(b) $\sup_{d_{\mathcal{U}}(u,\bar u)\leqslant\delta} \mathrm{E}_P[(\psi_{uj}(W)-\psi_{uj}(W))^2] \leqslant C\delta^{\alpha_1}$, $\sup_{d_{\mathcal{U}}(u,\bar u)\leqslant\delta}\|J_u-J_{\bar u}\| \leqslant C\delta^{\alpha_2}$,
(c) $\mathrm{E}_P \sup_{u\in\mathcal{U},\nu\in\Theta_u\times T_u}|\partial_{\nu_r}\mathrm{E}_P[\psi_{uj}(W_u,\nu)\mid Z_u]|^2 \leqslant C$,
(d) $\sup_{u\in\mathcal{U},\nu\in\Theta_u\times T_u}|\partial_{\nu_k}\partial_{\nu_r}\mathrm{E}_P[\psi_{uj}(W_u,\nu))|Z_u]| \leqslant C$, P-*a.s.*

This assumption imposes various smoothness and integrability conditions on various quantities derived from ψ. It also imposes some conditions on the complexity of the relevant function classes.

In what follows, let $\delta_n \searrow 0$ and $\tau_n \searrow 0$ be a sequence of constants approaching zero from above.

Assumption AS. *The following conditions hold for each* $n \geqslant n_0$ *and all* $P \in \mathcal{P}_n$. *The function* $\widehat{h}_u = (\widehat{h}_{um})_{m=1}^{d_t} \in \mathcal{H}_{un}$ *with probability at least* $1-\delta_n$, *where* $\mathcal{H}_{un}$ *is the set of measurable maps* $h = (h_m)_{m=1}^{d_t} : \mathcal{Z}_u \to T_u$ *such that*

$$\|h_m - h_{um}\|_{P,2} \leqslant \tau_n,$$

and whose complexity does not grow two quickly in the sense that the uniform covering entropy of $\mathcal{F}_1 = \{\psi_{uj}(W_u,\theta,h(Z_u), j\in[d_\theta], u\in\mathcal{U},\theta\in\Theta_u, h\in\mathcal{H}_{un}\}$ *obeys:*

$$\sup_Q \log N(\varepsilon\|F_1\|_{Q,2},\mathcal{F}_1,\|\cdot\|_{Q,2}) \leqslant s_n(\log(a_n/\varepsilon)),$$

where $F_1 \leqslant F_0$ *is the envelope of* $\mathcal{F}_1$, *and* $a_n \geqslant \max(n,\mathrm{e})$ *and* $s_n \geqslant 1$ *are some numbers that obey the growth condition:*

$$n^{-1/2}\left(\sqrt{s_n\log(a_n)} + n^{-1/2}s_n n^{\frac{1}{q}}\log(a_n)\right) \leqslant \tau_n \text{ and } \tau_n^{\alpha/2}\sqrt{s_n\log(a_n)} \leqslant \delta_n.$$

This assumption imposes conditions on the rate of estimating nuisance functions h_{um} as well as on the complexity of functions sets to which the estimators $\widehat{h}_{um}$ belong. Under approximately sparse framework, the index s_n appearing above will correspond to the maximum of the dimension of approximating models and of the size of the selected models; and index a_n will be equal to $p \vee n$. Under other frameworks, these parameters could be different; yet if they are well behaved,

then our results apply. Thus these results potentially cover other frameworks, where assumptions other than approximate sparsity are used to get the estimation problems to be manageable. We also would like to point out that the class $\mathcal{F}_1$ need not be Donsker, in particular its entropy is allowed to increase with n. Allowing non-Donskerness is crucial for allowing modern high-dimensional estimation methods for the nuisance functions. This is one assumption that makes the conditions imposed here very different from conditions imposed in various classical references on dealing with nonparametrically estimated nuisance functions.

The following theorem is one of the main results of the paper:

Theorem 5.1 (**Uniform Functional Central Limit Theorem for a Continuum of Target Parameters**). *Under Assumptions S1, S2, and AS, an estimator $(\widehat{\theta}_u)_{u\in\mathcal{U}}$ that obeys equation (63), satisfies uniformly in $P \in \mathcal{P}_n$:*

$$\sqrt{n}(\widehat{\theta}_u - \theta_u)_{u\in\mathcal{U}} = (\mathbb{G}_n \bar{\psi}_u)_{u\in\mathcal{U}} + o_P(1) \text{ in } \ell^\infty(\mathcal{U})^{d_\theta},$$

where $\bar{\psi}_u(W) := J_u^{-1}\psi_u(W_u, \theta_u, h_u(Z_u))$, and uniformly in $P \in \mathcal{P}_n$

$$(\mathbb{G}_n \bar{\psi}_u)_{u\in\mathcal{U}} \rightsquigarrow (\mathbb{G}_P \bar{\psi}_u)_{u\in\mathcal{U}} \text{ in } \ell^\infty(\mathcal{U})^{d_\theta},$$

where the paths of $u \mapsto \mathbb{G}_P\bar{\psi}_u$ are a.s. uniformly continuous on $(\mathcal{U}, d_\mathcal{U})$ and

$$\sup_{P\in\mathcal{P}_n} \mathrm{E}_P \sup_{u\in\mathcal{U}} \|\mathbb{G}_P\bar{\psi}_u\| < \infty \text{ and } \lim_{\delta\to 0} \sup_{P\in\mathcal{P}_n} \mathrm{E}_P \sup_{d_\mathcal{U}(u,\bar{u})\leqslant\delta} \|\mathbb{G}_P\bar{\psi}_u - \mathbb{G}_P\bar{\psi}_{\bar{u}}\| = 0.$$

We can estimate the law of Z_P by using the the bootstrap law of

$$\widehat{Z}^*_{n,P} := \sqrt{n}(\widehat{\theta}^*_u - \widehat{\theta}_u)_{u\in\mathcal{U}} := \frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i \widehat{\psi}_u(W_i), \tag{65}$$

where $(\xi_i)_{i=1}^n$ are i.i.d. multiplier variables defined in Section 3. The bootstrap law is computed by drawing $(\xi_i)_{i=1}^n$ conditional on the data. The estimated score above is

$$\widehat{\psi}_u(W_i) := \widehat{J}_u^{-1}\psi_u(W_{ui}, \widehat{\theta}_u, \widehat{h}_u(Z_{ui})),$$

where $\widehat{J}_u^{-1}$ is a suitable estimator of J_u^{-1}. Here we do not discuss the estimation of J_u, since it is often a problem-specific matter. In Section 3, we had $J_u = I$, so we did not need to estimate it.

The following theorem shows that the multiplier bootstrap provides a valid approximation to the large sample law of $\sqrt{n}(\widehat{\theta}_u - \theta_u)_{u\in\mathcal{U}}$.

Theorem 5.2 (**Uniform Validity of Multiplier Bootstrap**). *Suppose Assumptions S1, S2, and AS hold and that, for some positive constant α_3, uniformly in $P \in \mathcal{P}_n$ with probability $1-\delta_n$,*

$$(u \mapsto \widehat{J}_u) \in \mathcal{J}_n = \{u \mapsto \bar{J}_u : \|\bar{J}_u - \bar{J}_{\bar{u}}\| \leqslant C\|u - u\|^{\alpha_3}, \|\bar{J}_u - J_u\| \leqslant \tau_n, \text{ for all } (u,\bar{u}) \in \mathcal{U}^2\}.$$

Then $\widehat{Z}^*_{n,P} \rightsquigarrow_B Z_P$ in $\ell^\infty(\mathcal{U})^{d_\theta}$, uniformly in $P \in \mathcal{P}_n$.

Using a functional delta methods, we next derive the large sample distribution and validity of the multiplier bootstrap for estimators $\widehat{\Delta} := \phi(\widehat{\theta}) := \phi((\widehat{\theta}_u)_{u\in\mathcal{U}})$ of structural functionals $\Delta := \phi(\theta^0) = \phi((\theta_u)_{u\in\mathcal{U}})$. The latter functionals defined as "suitably differentiable" transforms of

$\theta^0 = (\theta_u)_{u\in\mathcal{U}}$. We suitably modify the functional delta method to handle uniformity with respect to the underlying dgp P. The following result gives asymptotic Gaussian law for $\sqrt{n}(\widehat{\Delta} - \Delta)$, the properly normalized structural estimator. It also shows that the bootstrap law of $\sqrt{n}(\widehat{\Delta}^* - \widehat{\Delta})$, computed conditionally on the data, approaches the asymptotic Gaussian law for $\sqrt{n}(\widehat{\Delta} - \Delta)$. Here $\widehat{\Delta}^* := \phi(\widehat{\theta}^*) := \phi((\widehat{\theta}^*)_{u\in\mathcal{U}})$ is the boostrap version of $\widehat{\Delta}$, and $\widehat{\theta}^*_u = \widehat{\theta}_u + \frac{1}{n}\sum_{i=1}^n \xi_i \widehat{\psi}_u(W_i)$ is the multiplier bootstrap version of $\widehat{\theta}_u$ defined via equation (65).

Theorem 5.3 (**Uniform Limit Theory and Validity of Multiplier Bootstrap for Smooth Functionals of θ.**). *Suppose that for each $P \in \mathcal{P} := \cup_{n\geqslant n_0}\mathcal{P}_n$, $\theta^0 = \theta^0_P$ is an element of a compact subset $\mathbb{D}_1 \subset \mathbb{D} = \ell^\infty(\mathcal{U})^{d_\theta}$. Suppose $\vartheta \mapsto \phi(\vartheta)$, a functional of interest mapping $\mathbb{D}_1$ to $\ell^\infty(\mathcal{Q})$, is Hadamard differentiable in ϑ with derivative ϕ'_ϑ, tangentially to $\mathbb{D}_0 = UC(\mathcal{U})^{d_\theta}$, uniformly in $\vartheta \in \mathbb{D}_1$, and that the mapping $(\vartheta, g) \mapsto \phi'_\vartheta(g)$ from $\mathbb{D}_1 \times \mathbb{D}_0$ into $\ell^\infty(\mathcal{Q})$ is defined and continuous. Then,*

$$\sqrt{n}(\widehat{\Delta} - \Delta) \rightsquigarrow T_P := \phi'_{\theta^0_P}(Z_P), \quad \text{in } \ell^\infty(\mathcal{Q}), \text{ uniformly in } P \in \mathcal{P}_n, \tag{66}$$

where T_P is a zero mean tight Gaussian process, for each $P \in \mathcal{P}$. Moreover,

$$\sqrt{n}(\widehat{\Delta}^* - \widehat{\Delta}) \rightsquigarrow_B T_P := \phi'_{\theta^0_P}(Z_P), \quad \text{in } \ell^\infty(\mathcal{Q}), \text{ uniformly in } P \in \mathcal{P}_n. \tag{67}$$

Here the usual notion of Hadamard differentiability is strengthened to a uniform notion of Hadamard differentiability as defined in the Appendix. The strengthening is sufficient to guarantee the uniform validity with respect to P. This result may be of independent interest in other problems.

6. Generic Lasso and Post-Lasso Methods for Functional Response Data

In this section, we provide estimation and inference results for Lasso and Post-Lasso estimators with function-valued outcomes and linear or logistic links. These results are of interest beyond the context of treatment effects estimation, and thus we present this section in a way that leaves it autonomous with respect to the rest of the paper.

6.1. **The generic setting with function-valued outcomes.**

Consider a data generating process with a functional response variable $(Y_u)_{u\in\mathcal{U}}$ and observable covariates X satisfying for each $u \in \mathcal{U}$

$$\mathrm{E}[Y_u \mid X] = \Lambda(f(X)'\theta_u) + r_u(X), \tag{68}$$

where $f : \mathcal{X} \to \mathbb{R}^p$ is a set of p measurable transformations of the initial controls X, θ_u is a p-dimensional vector, r_u is an approximation error, and Λ is a fixed link function. We note that the notation in this section differs from the rest of the paper with Y_u and X denoting a generic response and generic covariates to facilitate the application of these results in other contexts. We only consider the cases of linear link function, $\Lambda(t) = t$, and the logistic link[11] function $\Lambda(t) = \exp(t)/\{1 + \exp(t)\}$, in detail; but we note that the principles discussed here apply to any

[11]Considering the logistic link is useful for binary response data where $Y_u \in \{0, 1\}$ for each $u \in \mathcal{U}$, though the linear link can be used in this case as well.

M-estimator. In the remainder of the section, we discuss and establish results for ℓ_1-penalized and post-model selection estimators for $\theta_u, u \in \mathcal{U}$, that hold uniformly over $u \in \mathcal{U}$.

Throughout the section, we assume that $u \in \mathcal{U} \subset [0,1]^\iota$ and that i.i.d. observations from a dgp where (68) holds, $\{(Y_{ui}, u \in \mathcal{U}, X_i, f(X_i)) : i = 1, \ldots, n\}$, are available to estimate $(\theta_u)_{u \in \mathcal{U}}$. For $u \in \mathcal{U}$, a penalty level λ, and a diagonal matrix of penalty loadings $\widehat{\Psi}_u$, we define the Lasso estimator as

$$\widehat{\theta}_u \in \arg\min_\theta \mathbb{E}_n[M(Y_u, f(X)'\theta)] + \frac{\lambda}{n}\|\widehat{\Psi}_u \theta\|_1 \tag{69}$$

where $M(y,t) = \frac{1}{2}(y - \Lambda(t))^2$ for the case of linear regression, and $M(y,t) = 1(y = 1)\log \Lambda(t) + 1(y = 0)\log(1 - \Lambda(t))$ in the case of the logistic link function for binary response data. The corresponding Post-Lasso estimator is then defined as

$$\widetilde{\theta}_u \in \arg\min_\theta \mathbb{E}_n[M(Y_u, f(X)'\theta)] \quad : \quad \text{supp}(\theta) \subseteq \text{supp}(\widehat{\theta}_u). \tag{70}$$

The chief departure between analysis of Lasso and Post-Lasso when $\mathcal{U}$ is a singleton and the functional response case is that the penalty parameter needs to be set to control selection errors uniformly over $u \in \mathcal{U}$. To uniformly control these errors, we will essentially set the penalty parameter λ so that with high probability

$$\frac{\lambda}{n} \geqslant c \sup_{u \in \mathcal{U}} \left\| \widehat{\Psi}_u^{-1} \mathbb{E}_n \left[\nabla_\theta M(Y_u, f(X)'\theta_u) \right] \right\|_\infty . \tag{71}$$

The strategy above is similar to Bickel, Ritov, and Tsybakov (2009); Belloni and Chernozhukov (2013); and Belloni, Chernozhukov, and Wang (2011) who use an analog of (71) that derive properties of Lasso and Post-Lasso when $\mathcal{U}$ is a singleton. In the context of quantile regression a related uniform choice of penalty parameter was used in Belloni and Chernozhukov (2011a). In the functional outcome case guaranteeing that the "regularization event" (71) holds with high probability also plays a key role in establishing desirable properties of Lasso and Post-Lasso estimators uniformly over $u \in \mathcal{U}$.

To implement (71), we propose setting the penalty level as

$$\lambda = c\sqrt{n}\Phi^{-1}(1 - \gamma/\{2pn^\iota\}), \tag{72}$$

where ι is the dimension of $\mathcal{U}$, $1 - \gamma$ with $\gamma = o(1)$ is a confidence level associated with the probability of event (71), and $c > 1$ is a slack constant similar to that of Bickel, Ritov, and Tsybakov (2009). In practice, we set $c = 1.1$ and $\gamma = .1/\log(n)$ though many other choices are theoretically valid.

In addition to the penalty parameter λ, we also need to construct a penalty loading matrix $\widehat{\Psi}_u = \text{diag}(\{\widehat{l}_{uj,k}, j = 1, \ldots, p\})$. This loading matrix can be formed according to the following iterative algorithm.

Algorithm 1 (Estimation of Penalty Loadings)**.** Choose $\gamma \in [1/n, 1/\log n]$ can $c > 1$ to form λ as defined in (72), and choose a constant $K \geqslant 1$ as an upper bound on the number of iterations. (0) Set $k = 0$, and initialize $\widehat{l}_{uj,0}$ for each $j = 1, \ldots, p$. For the linear link function, set $\widehat{l}_{uj,0} = \{\mathbb{E}_n[f_j^2(X)(Y_u - \bar{Y}_u)^2]\}^{1/2}$ with $\bar{Y}_u = \mathbb{E}_n[Y_u]$. For the logistic link function, set

$\widehat{l}_{uj,0} = \frac{1}{2}\{\mathbb{E}_n[f_j^2(X)]\}^{1/2}$. (1) Compute the Lasso and Post-Lasso estimators, $\widehat{\theta}_u$ and $\widetilde{\theta}_u$, based on $\widehat{\Psi}_u = \mathrm{diag}(\{\widehat{l}_{uj,k}, j = 1, \ldots, p\})$. (2) Set $\widehat{l}_{uj,k+1} := \{\mathbb{E}_n[f_j^2(X)(Y_u - \Lambda(f(X)'\widetilde{\theta}_u))^2]\}^{1/2}$. (3) If $k > K$, stop; otherwise set $k \leftarrow k + 1$ and go to step (1).

6.2. **Asymptotic Properties of a Continuum of Lasso and Post-Lasso Estimators for Functional Responses: Linear Case.** In the following, we provide sufficient conditions for establishing good performance of the estimators discussed in Section 5.1 when the linear link function is used. In the statement of the following assumption, $\delta_n \searrow 0$, $\ell_n \nearrow \infty$, and $\Delta_n \searrow 0$ are fixed sequences; and c, C, κ', κ'' and $\nu \in (0, 1]$ are positive finite constants.

Assumption 6. *For each $n \geqslant 1$, our data consist of i.i.d. copies $(W_i)_{i=1}^n$ of the stochastic process $W = ((Y_u)_{u\in\mathcal{U}}, X)$ defined on the probability space $(S, \mathcal{S}, P)$ such that model (68) holds with $\mathcal{U} \subset [0, 1]^\iota$. Consider $\Lambda(t) = t$ and $\zeta_u = Y_u - \mathrm{E}[Y_u \mid X]$. Suppose the following conditions hold uniformly for all $P \in \mathcal{P}_n$: (i) the model (68) is approximately sparse with sparsity index obeying $\sup_{u\in\mathcal{U}} \|\theta_u\|_0 \leqslant s$ and the growth restriction $\log(pn/\gamma) \leqslant \delta_n n^{1/3}$. (ii) The set $\mathcal{U}$ has covering entropy bounded as $\log N(\epsilon, \mathcal{U}, d_\mathcal{U}) \leqslant \iota \log(1/\epsilon) \vee 0$, and the collection $(Y_u, \zeta_u, r_u)_{u\in\mathcal{U}}$ is suitably measurable. (iii) Uniformly over $u \in \mathcal{U}$, the model's moments are boundedly heteroscedastic, namely $c \leqslant \mathrm{E}_P[\zeta_u^2 \mid X] \leqslant C$ and $\max_{j\leqslant p}\mathrm{E}_P[|f_j(X)\zeta_u|^3 + |f_j(X)Y_u|^3] \leqslant C$. (iv) We have that the dictionary functions, approximation errors, and empirical errors obey the following boundedness and empirical regularity conditions: (a) $c \leqslant \mathrm{E}_P[f_j^2(X)] \leqslant C$, $j = 1, \ldots, p$; $\max_{j\leqslant p} |f_j(X)| \leqslant K_n$ a.s.; $K_n \log(p \vee n) \leqslant \delta_n n^{\{\nu\wedge\frac{1}{2}\}}$. (b) With probability $1 - \Delta_n$, $\sup_{u\in\mathcal{U}} \mathbb{E}_n[r_u^2(X)] \leqslant Cs \log(p \vee n)/n$; $\sup_{u\in\mathcal{U}} \max_{j\leqslant p} |(\mathbb{E}_n - \mathrm{E}_P)[f_j^2(X)\zeta_u^2]| \vee |(\mathbb{E}_n - \mathrm{E}_P)[f_j^2(X)Y_u^2]| \leqslant \delta_n$; $\sup_{d_\mathcal{U}(u,u')\leqslant\epsilon}\{(\mathbb{E}_n + \mathrm{E}_P)[(\zeta_u - \zeta_{u'})^2]\}^{1/2} \leqslant C\{\epsilon^\nu + n^{-1/2}\}$. (c) The empirical minimum and maximum sparse eigenvalues are bounded from zero and above, $\kappa' \leqslant \inf_{\|\delta\|_0\leqslant s\ell_n} \|f(X)'\delta\|_{\mathbb{P}_n,2} \leqslant \sup_{\|\delta\|_0\leqslant s\ell_n} \|f(X)'\delta\|_{\mathbb{P}_n,2} \leqslant \kappa''$.*

Under Assumption 6, we establish results on the performance of the estimators (69) and (70) for the linear link function case that hold uniformly over $u \in \mathcal{U}$.

Theorem 6.1 (Rates and Sparsity for Functional Responses under Linear Link)**.** *Under Assumption 6 and setting penalties as in Algorithm 1, for all n large enough, uniformly for all $P \in \mathcal{P}_n$ with P_P probability $1 - o(1)$, for some constant $\bar{C}$, the Lasso estimator $\widehat{\theta}_u$ is uniformly sparse, $\sup_{u\in\mathcal{U}} \|\widehat{\theta}_u\|_0 \leqslant \bar{C}s$, and the following performance bounds hold:*

$$\sup_{u\in\mathcal{U}} \|f(X)'(\widehat{\theta}_u - \theta_u)\|_{\mathbb{P}_n,2} \leqslant \bar{C}\sqrt{\frac{s \log(p \vee n)}{n}} \text{ and } \sup_{u\in\mathcal{U}} \|\widehat{\theta}_u - \theta_u\|_1 \leqslant \bar{C}\sqrt{\frac{s^2 \log(p \vee n)}{n}}.$$

For all n large enough, uniformly for all $P \in \mathcal{P}_n$, with P_P probability $1 - o(1)$, the Post-Lasso estimator corresponding to $\widehat{\theta}_u$ obeys

$$\sup_{u\in\mathcal{U}} \|f(X)'(\widetilde{\theta}_u - \theta_u)\|_{\mathbb{P}_n,2} \leqslant \bar{C}\sqrt{\frac{s \log(p \vee n)}{n}}, \text{ and } \sup_{u\in\mathcal{U}} \|\widetilde{\theta}_u - \theta_u\|_1 \leqslant \bar{C}\sqrt{\frac{s^2 \log(p \vee n)}{n}}.$$

We note that the performance bounds are exactly of the type used in Assumptions 3 and 4.

6.3. **Asymptotic Properties of a Continuum of Lasso and Post-Lasso Estimators for Functional Responses: Logistic Case.** Next we provide sufficient conditions to state results on the performance of the estimators discussed above for the logistic link function. This case corresponds to $M(y,t) = 1(y=1)\log\Lambda(t) + 1(y=0)\log(1-\Lambda(t))$ with $\Lambda(t) = \exp(t)/\{1+\exp(t)\}$ where the response variable is assumed to be binary, $Y_{ui} \in \{0,1\}$ for all $u \in \mathcal{U}$ and $i = 1,\ldots,n$.

Consider fixed sequences $\delta_n \to 0$, $\ell_n \nearrow \infty$, $\Delta_n \to 0$ and positive finite constants c, C, κ', κ'' and $\nu \in (0,1]$.

Assumption 7. *For each $n \geqslant 1$, our data consist of i.i.d. copies $(W_i)_{i=1}^n$ of the stochastic process $W = ((Y_u)_{u\in\mathcal{U}}, X)$ defined on the probability space $(S, \mathcal{S}, P)$ such that model (68) holds with $\mathcal{U} \subset [0,1]^\iota$. Consider $\Lambda(t) = \exp(t)/\{1+\exp(t)\}$, $Y_u \in \{0,1\}$, and $\zeta_u = Y_u - \mathrm{E}[Y_u \mid X]$. Suppose the following conditions hold uniformly for all $P \in \mathcal{P}_n$: (i) the model (68) is approximately sparse form with sparsity index obeying $\sup_{u\in\mathcal{U}} \|\theta_u\|_0 \leqslant s$ and the growth restriction $\log(pn/\gamma) \leqslant \delta_n n^{1/3}$. (ii) The set $\mathcal{U}$ has covering entropy bounded as $\log N(\epsilon, \mathcal{U}, d_{\mathcal{U}}) \leqslant \iota \log(1/\epsilon) \vee 0$, and the collection $(Y_u, \zeta_u, r_u)_{u\in\mathcal{U}}$ is suitably measurable. (iii) Uniformly over $u \in \mathcal{U}$ the model's moments satisfy $\max_{j\leqslant p} \mathrm{E}_P[|f_j(X)|^3] \leqslant C$, and $\underline{c} \leqslant \mathrm{E}_P[Y_u \mid X] \leqslant 1 - \underline{c}$. (iv) We have that the dictionary functions, approximation errors, and empirical errors obey the following boundedness and empirical regularity conditions: (a) $\sup_{u\in\mathcal{U}} |r_u(X)| \leqslant \delta_n$ a.s.; $c \leqslant \mathrm{E}_P[f_j^2(X)] \leqslant C$, $j = 1,\ldots,p$; $\max_{j\leqslant p} |f_j(X)| \leqslant K_n$ a.s.; $K_n \log(p \vee n) \leqslant \delta_n n^{\{\nu \wedge \frac{1}{2}\}}$ and $K_n^2 s^2 \log^2(p\vee n) \leqslant \delta_n n$. (b) With probability $1-\Delta_n$, $\sup_{u\in\mathcal{U}} \mathbb{E}_n[r_u^2(X)] \leqslant Cs\log(p\vee n)/n$; $\sup_{u\in\mathcal{U}} \max_{j\leqslant p} |(\mathbb{E}_n - \mathrm{E}_P)[f_j^2(X)\zeta_u^2]| \leqslant \delta_n$; $\sup_{d_{\mathcal{U}}(u,u')\leqslant\epsilon} \{(\mathbb{E}_n + \mathrm{E}_P)[(\zeta_u - \zeta_{u'})^2]\}^{1/2} \leqslant C\{\epsilon^\nu + n^{-1/2}\}$. (c) The empirical minimum and maximum sparse eigenvalues are bounded from zero and above: $\kappa' \leqslant \inf_{\|\delta\|_0\leqslant s\ell_n} \|f(X)'\delta\|_{\mathbb{P}_n,2} \leqslant \sup_{\|\delta\|_0\leqslant s\ell_n} \|f(X)'\delta\|_{\mathbb{P}_n,2} \leqslant \kappa''$.*

The following result characterizes the performance of the estimators (69) and (70) for the logistic link function case under Assumption 7.

Theorem 6.2 (Rates and Sparsity for Functional Response under Logistic Link)**.** *Under Assumption 7 and setting penalties as in Algorithm 1, for all n large enough, uniformly for all $P \in \mathcal{P}_n$ with P_P probability $1 - o(1)$, the following performance bounds hold for some constant $\bar{C}$:*

$$\sup_{u\in\mathcal{U}} \|f(X)'(\widehat{\theta}_u - \theta_u)\|_{\mathbb{P}_n,2} \leqslant \bar{C}\sqrt{\frac{s\log(p\vee n)}{n}} \text{ and } \sup_{u\in\mathcal{U}} \|\widehat{\theta}_u - \theta_u\|_1 \leqslant \bar{C}\sqrt{\frac{s^2\log(p\vee n)}{n}}.$$

and the estimator is uniformly sparse: $\sup_{u\in\mathcal{U}} \|\widehat{\theta}_u\|_0 \leqslant \bar{C}s$. For all n large enough, uniformly for all $P \in \mathcal{P}_n$, with P_P probability $1 - o(1)$, the Post-Lasso estimator corresponding to $\widehat{\theta}_u$ obeys

$$\sup_{u\in\mathcal{U}} \|f(X)'(\widetilde{\theta}_u - \theta_u)\|_{\mathbb{P}_n,2} \leqslant \bar{C}\sqrt{\frac{s\log(p\vee n)}{n}}, \text{ and } \sup_{u\in\mathcal{U}} \|\widetilde{\theta}_u - \theta_u\|_1 \leqslant \bar{C}\sqrt{\frac{s^2\log(p\vee n)}{n}}.$$

We note that the performance bounds satisfy the conditions of Assumptions 3 and 4.

7. Estimating the Effect of 401(k) Participation on Financial Asset Holdings

As an illustration of the methods in this paper, we consider the estimation of the effect of 401(k) participation on accumulated assets as in Abadie (2003) and Chernozhukov and Hansen (2004). The key problem in determining the effect of participation in 401(k) plans on accumulated assets is saver heterogeneity coupled with the fact that the decision of whether to enroll in a 401(k) is non-random. It is generally recognized that some people have a higher preference for saving than others. It also seems likely that those individuals with the highest unobserved preference for saving would be most likely to choose to participate in tax-advantaged retirement savings plans and would tend to have otherwise high amounts of accumulated assets. The presence of unobserved savings preferences with these properties then implies that conventional estimates that do not account for saver heterogeneity and endogeneity of participation will be biased upward, tending to overstate the savings effects of 401(k) participation.

To overcome the endogeneity of 401(k) participation, Abadie (2003) and Chernozhukov and Hansen (2004) adopt the strategy detailed in Poterba, Venti, and Wise (1994; 1995; 1996; 2001) and Benjamin (2003), who used data from the 1991 Survey of Income and Program Participation and argue that eligibility for enrolling in 401(k) plan in this data can be taken as exogenous after conditioning on a few observables of which the most important for their argument is income. The basic idea of their argument is that, at least around the time 401(k)'s initially became available, people were unlikely to be basing their employment decisions on whether an employer offered a 401(k) but would instead focus on income. Thus, eligibility for a 401(k) could be taken as exogenous conditional on income, and the causal effect of 401(k) eligibility could be directly estimated by appropriate comparison across eligible and ineligible individuals.[12] Abadie (2003) and Chernozhukov and Hansen (2004) use this argument for the exogeneity of eligibility conditional on controls to argue that 401(k) eligibility provides a valid instrument for 401(k) participation and employ IV methods to estimate the effect of 401(k) participation on accumulated assets.

As a complement to the work cited above, we estimate various treatment effects of 401(k) participation on holdings of financial assets using high-dimensional methods. A key component of the argument underlying the exogeneity of 401(k) eligibility is that eligibility may only be taken as exogenous after conditioning on income. Both Abadie (2003) and Chernozhukov and Hansen (2004) adopt this argument but control only for a small number of terms. One might wonder whether the small number of terms considered is sufficient to adequately control for income and other related confounds. At the same time, the power to learn anything about the effect of 401(k) participation decreases as one controls more flexibly for confounds. The methods developed in this paper offer one resolution to this tension by allowing us to consider a very broad set of controls and functional forms under the assumption that among the set of variables we consider there is a relatively low-dimensional set that adequately captures the effect of confounds.

[12]Poterba, Venti, and Wise (1994; 1995; 1996; 2001) and Benjamin (2003) all focus on estimating the effect of 401(k) eligibility, the intention to treat parameter. Also note that there are arguments that eligibility should not be taken as exogenous given income; see, for example, Engen, Gale, and Scholz (1996) and Engen and Gale (2000).

This approach is more general than that pursued in Chernozhukov and Hansen (2004) or Abadie (2003) which both implicitly assume that confounding effects can adequately be controlled for by a small number of variables chosen *ex ante* by the researcher.

We use the same data as Abadie (2003), Benjamin (2003), and Chernozhukov and Hansen (2004). The data consist of 9,915 observations at the household level drawn from the 1991 SIPP. We consider two different outcome variables in our analysis: net total financial assets[13] and total wealth.[14] Our treatment variable, D, is an indicator for having positive 401(k) balances; and our instruments, Z, is an indicator for working at a firm that offers a 401(k) plan. The vector of controls, X, consists of age, income, family size, years of education, a married indicator, a two-earner status indicator, a defined benefit pension status indicator, an IRA participation indicator, and a home ownership indicator. Further details about the sample and variables used can be found in Chernozhukov and Hansen (2004).

We present results for four different sets of control variables $f(X)$. The first set of control variables uses the indicators of marital status, two-earner status, defined benefit pension status, IRA participation status, and home ownership status, a linear term for family size, five categories for age, four categories for education, and seven categories for income (Indicator specification). We use the same definitions of categories as in Chernozhukov and Hansen (2004) and note that this is identical to the specification in Chernozhukov and Hansen (2004) and Benjamin (2003). The second specification augments the Indicator specification with all two-way interactions between the variables from the Indicator specification (Indicators plus interactions specification). The third specification uses the indicators of marital status, two-earner status, defined benefit pension status, IRA participation status, and home ownership status, and cubic b-splines with one, one, three, and five interior knots for family size, education, age, and income, respectively (B-Spline specification). The fourth specification augments the B-Spline specification with all two-way interactions of the sets of variables from the B-Spline specification (B-Spline plus interactions specification). The dimensions of the set of control variables are thus 20, 167, 27, and 323 for the Indicator, Indicator plus interactions, B-Spline, and B-Spline plus interactions specifications, respectively.

We report estimates of the LATE, LATE-T, LQTE, and LQTE-T for each of the four sets of control variables. Estimation of all of the treatment effects depends on first-stage estimation of reduced form functions as detailed in Section 3. We estimate reduced form quantities where $Y_u = Y$ is the outcome using least squares when no model selection is used or Post-Lasso when selection is used. We estimate propensity scores and reduced form quantities where $Y_u = 1(Y \leqslant u)$ is the outcome by logistic regression when no model selection is used or Post-ℓ_1-penalized

[13]Net total financial assets are defined as the sum of checking accounts, U.S. saving bonds, other interest-earning accounts in banks and other financial institutions, other interest-earning assets (such as bonds held personally), stocks and mutual funds less nonmortgage debt, IRA balances, and 401(k) balances.

[14]Total wealth is net financial assets plus housing equity, housing value minus mortgage, and the value of business, property, and motor vehicles.

logistic regression when selection is used.[15] We use the penalty level given in 72 and construct penalty loadings using the method detailed in Algorithm 1. For the LATE and LATE-T where the set $\mathcal{U}$ is a singleton, we use the penalty level in 72 with $\iota = 0$. This choice corresponds to that used in Belloni, Chernozhukov, and Hansen (2011).

Estimates of the LATE and LATE-T are given in Table 1. In this table, we provide point estimates for each of the four sets of controls with and without variable selection. We also report both analytic and multiplier bootstrap standard errors. The bootstrap standard errors are based on 500 bootstrap replications and wild bootstrap weights. Looking first at the two sets of standard error estimates, we see that the bootstrap and analytic standard are quite similar and that one would not draw substantively different conclusions from one versus the other.

It is interesting that the estimated LATE and LATE-T are similar in seven of the eight sets of estimates reported, suggesting positive and significant effects of 401(k) participation on net financial assets and total wealth. This similarity is unsurprising but reassuring in the Indicator and B-Spline specifications as it illustrates that there is little impact of variable selection relative to simply including everything in a low-dimensional setting.[16] The one case where we observe substantively different results is in the B-Spline specification with interactions when we do not use variable selection. In this case, both the LATE and LATE-T point estimates are large with associated very large estimated standard errors. One would favor these imprecise estimates from the B-spline plus interactions specification if there were important nonlinearity that is missed by the simpler specifications. The concern that there is important nonlinearity missed by the other specifications that renders the estimated treatment effects too imprecise to be useful is alleviated by noting that the point estimate and standard error based on the B-spline plus interactions specification following variable selection are sensible and similar to the other estimates. The fact that estimates following variable selection are similar to the other estimates suggests the bulk of the reduced form predictive power is contained in a set of variables similar to those used in the other specifications and that there is not a small number of the added variables that pick out important sources of nonlinearity neglected by the other specifications. Thus, the large point estimates and standard errors in this case seem to be driven by including many variables which have little to no predictive power in the reduced form relationships but result in overfitting.

We provide estimates of the LQTE and LQTE-T based on the Indicator specification, the Indicator plus interaction specification, the B-Spline specification, and the B-Spline plus interaction specification in Figures 1, 2, 3, and 4, respectively. The left column in each figure gives results for the LQTE, and the right column displays the results for the LQTE-T. In the top row of each

[15]The estimated propensity score shows up in the denominator of the efficient moment conditions. As is conventional, we use trimming to keep the denominator bounded away from zero with trimming set to 10^{-12}. Trimming only occurs when selection is not done in the B-spline plus interaction specification.

[16]In the low-dimensional setting, using all available controls is semi-parametrically efficient and allows uniformly valid inference. Thus, the similarity between the results in this case is an important feature of our method which results from our reliance on low-bias moment functions and sensible variable selection devices to produce semi-parametrically efficient estimators and uniformly valid inference statements *following* model selection.

figure, we display the results with net financial assets as the dependent variable, and we give the results based on total wealth as the dependent variable in the middle row. The bottom row of each figure displays the selection-based estimate of the treatment effect on net total financial assets along with the selection-based estimate of the treatment effect on total wealth. In each graphic, we use solid lines for point estimates and report uniform 95% confidence intervals with dashed lines.

Looking across the figures, we see a similar pattern to that seen for the LATE and LATE-T in that the selection-based estimates are stable across all specifications and are similar to the estimates obtained without selection from the baseline Indicators specification and the B-Spline specification. In the more flexible specifications that include interactions, the estimates that do not make use of selection start to behave erratically. This erratic behavior is especially apparent in the estimated LQTE of 401(k) participation on total wealth where we observe that small changes in the quantile index may result in large swings in the point estimate of the LQTE and estimated standard errors are large enough that meaningful conclusions cannot be drawn. Again, this erratic behavior is likely due to overfitting as the variable selection methods select a roughly common low-dimensional set of variables that are useful for reduced form prediction in all cases.

If we focus on the LQTE and LQTE-T estimated from variable selection methods, we find that 401(k) participation has a small impact on accumulated net total financial assets at low quantiles while appearing to have a much larger impact at high quantiles. Looking at the uniform confidence intervals, we can see that this pattern is statistically significant at the 5% level and that we would reject the hypothesis that 401(k) participation has no effect and reject the hypothesis of a constant treatment effect more generally. For total wealth, we can also reject the hypothesis of zero treatment effect and the hypothesis of a constant treatment effect, though the uniform confidence bands are much wider. Interestingly, the only evidence of a statistically significant impact on total wealth occurs for low and intermediate quantiles; one cannot rule out the hypothesis of no effect of 401(k) participation on total wealth in the upper quantiles. This pattern is especially interesting when coupled with the evidence of essentially a uniformly positive effect of participation on net total financial assets which suggests that some of the effect on financial assets may be attributed to substitution from non-financial assets into the tax-advantaged 401(k) assets.

It is interesting that our results are similar to those in Chernozhukov and Hansen (2004) despite allowing for a much richer set of control variables. The similarity is due to the fact that the variable selection methods consistently pick a set of variables similar to those used in previous work. The fact that we allow for a rich set of controls but produce similar results to those previously available lends further credibility to the claim that previous work controlled adequately for the available observables.[17] Finally, it is worth noting that this similarity is not mechanical or otherwise built in to the procedure. For example, applications in Belloni, Chen,

[17]Of course, the estimates are still not valid causal estimates if one does not believe that 401(k) eligibility can be taken as exogenous after controlling for income and the other included variables.

Chernozhukov, and Hansen (2012) and Belloni, Chernozhukov, and Hansen (2011) use high-dimensional variable selection methods and produce sets of variables that differ substantially from intuitive baselines.

Appendix A. Some Tools

A.1. **Stochastic Convergence Uniformly in P.** All parameters, such as the law of the data, are indexed by P, sometimes referred to as the the data generating process. This dependency, which is well understood, is kept implicit throughout. We shall allow the possibility that the probability measure $P = P_n$ can depend on n. We shall conduct our stochastic convergence analysis uniformly in P, where P can vary within some set $\mathcal{P}_n$, which itself may vary with n.

The convergence analysis, namely stochastic order relations and convergence in distribution, uniformly in $P \in \mathcal{P}_n$ and the analysis under all sequences $P_n \in \mathcal{P}_n$ are equivalent. Specifically, consider a sequence of stochastic processes X_n and a random element Y, taking values in the metric space $\mathbb{D}$, defined on the probability space $(A, \mathcal{A}_A, \mathrm{P}_P)$. Through most of the Appendix $\mathbb{D} = \ell^\infty(\mathcal{U})$, the space of uniformly bounded functions mapping an arbitrary index set $\mathcal{U}$ to the real line. Consider also a sequence of deterministic positive constants a_n. We shall say that

(i) $X_n = O_P(a_n)$ uniformly in $P \in \mathcal{P}_n$, if $\lim_{K \nearrow \infty} \lim_{n\to\infty} \sup_{P\in\mathcal{P}_n} \mathrm{P}^*_P(|X_n| > Ka_n) = 0$,
(ii) $X_n = o_P(a_n)$ uniformly in $P \in \mathcal{P}_n$, if $\sup_{K>0} \lim_{n\to\infty} \sup_{P\in\mathcal{P}_n} \mathrm{P}^*_P(|X_n| > Ka_n) = 0$,
(iii) $X_n \rightsquigarrow Y$ (with law dependent on P) uniformly in $P \in \mathcal{P}_n$, if

$$\sup_{P\in\mathcal{P}_n} \sup_{h\in \mathrm{BL}_1(\mathbb{D},\mathbb{R})} |\mathrm{E}^*_P h(X_n) - \mathrm{E}_P h(Y)| \to 0.$$

Here the symbol $\rightsquigarrow$ denotes weak convergence, i.e. convergence in distribution or law, $\mathrm{BL}_1(\mathbb{D}, \mathbb{R})$ denotes the space of functions mapping $\mathbb{D}$ to $\mathbb{R}$ with Lipschitz norm at most 1, and the outer probability and expectations, P^*_P and E^*_P, are invoked whenever (non)-measurability arises.

Lemma A.1. *The above notions are equivalent to the following notions:*

(i) *for every sequence $P_n \in \mathcal{P}_n$, $X_n = O_{P_n}(a_n)$, i.e. $\lim_{K \nearrow \infty} \lim_{n\to\infty} \mathrm{P}^*_{P_n}(|X_n| > Ka_n) = 0$,*
(ii) *for every sequence $P_n \in \mathcal{P}_n$, $X_n = o_{P_n}(a_n)$, i.e. $\sup_{K>0} \lim_{n\to\infty} \mathrm{P}^*_{P_n}(|X_n| > Ka_n) = 0$,*
(iii) *for every sequence $P_n \in \mathcal{P}_n$, $X_n \rightsquigarrow Y$, i.e.*

$$\sup_{h\in \mathrm{BL}_1(\mathbb{D},\mathbb{R})} |\mathrm{E}^*_{P_n} h(X_n) - \mathrm{E}_{P_n} h(Y)| \to 0.$$

Proof of Lemma A.1. The claims follow straightforwardly from the definitions, and so the proof is omitted.

A.2. **Uniform in P Donsker Property.** Let $(W_i)_{i=1}^\infty$ be a sequence of i.i.d. copies of random element $W : S \mapsto \mathcal{W}$, taking values in the sample space $(\mathcal{W}, \mathcal{A}_\mathcal{W})$, with law determined by a probability measure $P \in \mathcal{P}$ defined on a measurable space $(S, \mathcal{A}_S)$. Let $\mathcal{F}_P$ be a set of measurable functions $w \mapsto f_{P,t}(w)$ mapping $\mathcal{W}$ to $\mathbb{R}$ indexed by $P \in \mathcal{P}$ and $t \in T$, where T is a fixed, totally bounded semi-metric space equipped with a semi-metric d_T. Let $N(\epsilon, \mathcal{F}_P, \|\cdot\|_{Q,2})$ denote the

ϵ-covering number of the class of functions $\mathcal{F}_P$ with respect to the $L^2(Q)$ seminorm $\|\cdot\|_{Q,2}$, where Q is finitely discrete. We shall invoke the following lemma.

Theorem A.1. *For each $P \in \mathcal{P}$, let $\mathcal{F}_P$ be a suitably measurable class of functions mapping $\mathcal{W}$ to $\mathbb{R}$, equipped with a measurable envelope $F_P : \mathcal{W} \mapsto \mathbb{R}$. Suppose that for $q > 2$*

$$\sup_{P\in\mathcal{P}} \|F_P\|_{P,q} \leqslant C \text{ and } \lim_{\delta\searrow 0} \sup_{P\in\mathcal{P}} \sup_{(f_t,f_{\bar{t}})\in\mathcal{F}_P^2 : d_T(t,\bar{t})\leqslant\delta} \|f_t - f_{\bar{t}}\|_{P,2} = 0.$$

Furthermore, suppose that

$$\lim_{\delta\searrow 0} \sup_{P\in\mathcal{P}} \int_0^\delta \sup_Q \sqrt{\log N(\epsilon\|F_P\|_{Q,2}, \mathcal{F}_P, \|\cdot\|_{Q,2})}d\epsilon = 0.$$

Consider $Z_{n,P} := (Z_n(t))_{t\in T} := (\mathbb{G}_n(f_t))_{t\in T}$ and $Z_P := (Z_P(t))_{t\in T} := (\mathbb{G}_P(f_t))_{t\in T}$.
(a) Then for $Z_{n,P} \rightsquigarrow Z_P$ in $\ell^\infty(T)$ uniformly in $P \in \mathcal{P}$, namely

$$\sup_{P\in\mathcal{P}} \sup_{h\in BL_1(\ell^\infty(T),\mathbb{R})} |\mathrm{E}_P^* h(Z_{n,P}) - \mathrm{E}_P h(Z_P)| \to 0.$$

(b) Moreover, the limit process hat the following continuity properties:

$$\sup_{P\in\mathcal{P}} \mathrm{E}_P \sup_{t\in T} |Z_P(t)| < \infty, \quad \lim_{\delta\searrow 0} \sup_{P\in\mathcal{P}} \mathrm{E}_P \sup_{d_T(t,\bar{t})\leqslant\delta} |Z_P(t) - Z_P(\bar{t})| = 0.$$

(c) The paths $t \mapsto Z_P(t)$ are a.s. uniformly continuous on (T, d_T) under each $P \in \mathcal{P}$.

This is a version of uniform Donsker theorem stated in Theorem 2.8.2 in van der Vaart and Wellner (1996), which allows for the function classes to be depend on P themselves. The latter case is critically needed in all of our problems.

Proof. Part (a) is a direct consequence of Lemma A.2 stated below, and part (b) can be demonstrated similarly to the proof of Theorem 2.8.2 in van der Vaart and Wellner (1996). Claim (c) follows from claim (b) and a standard argument, based on application of Borell-Canteli lemma and reasoning as in Van der Vaart (1998). ■

A.3. **Uniform in P Validity of Multiplier Bootstrap.** Consider the setting of the preceding subsection. Let $(\xi_{i=1}^n)$ be i.i.d multipliers whose distribution does not depend on P, such that $\mathrm{E}(|\xi|^q) \leqslant C$ for $q > 2$. Consider the multiplier empirical process:

$$Z_{n,P}^* := (Z_n^*(t))_{t\in T} := (\mathbb{G}_n(\xi f_t))_{t\in T} = \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i f_t\right)_{t\in T},$$

and $Z_P := (Z_P(t))_{t\in T} := (\mathbb{G}_P(f_t))_{t\in\mathcal{F}_P}$ as defined before.

Theorem A.2. *Consider the conditions of Theorem A.1. Then (a) the following unconditional convergence takes place, $Z_{n,P}^* \rightsquigarrow Z_P$ in $\ell^\infty(T)$ uniformly in $P \in \mathcal{P}$, namely*

$$\sup_{P\in\mathcal{P}} \sup_{h\in BL_1(\ell^\infty(T),\mathbb{R})} |\mathrm{E}_P^* h(Z_{n,P}^*) - \mathrm{E}_P h(Z_P)| \to 0.$$

(b) and the following conditional convergence takes place, $Z^*_{n,P} \rightsquigarrow_B Z_P$ *in* $\ell^\infty(T)$ *uniformly in* $P \in \mathcal{P}$*, namely*

$$\sup_{P\in\mathcal{P}} \sup_{h\in BL_1(\ell^\infty(T),\mathbb{R})} |\mathrm{E}_M h(Z^*_{n,P}) - \mathrm{E}_P h(Z_P)| = o_P(1),$$

where E_M *denotes expectation over multiplier weights* $(\xi_i)_{i=1}^n$ *holding the data* $(W_i)_{i=1}^n$ *fixed.*

Proof. We begin by claiming that (i) $Z^*_{n,P} \rightsquigarrow Z^*_P$ in $\ell^\infty(T)$, where $Z^*_{n,P} := (\mathbb{G}_n \xi f_u)_{t\in T}$; and (ii) $Z^*_P := (\mathbb{G}_P \xi f_t)_{t\in T}$ is equal in distribution to $Z_P := (\mathbb{G}_P f_t)_{t\in T}$, in particular, Z^*_P and Z_P share the identical covariance function (and so they also share the continuity properties, which we have established in the preceding theorem).

Claim (ii) is immediate, since multiplication by ξ to create $\mathbb{G}_P(\xi f)$ does not change the covariance function of $\mathbb{G}_P(f)$, that is, the P-Gaussian processes indexed by $\xi\mathcal{F}$ and by $\mathcal{F}$ are equal in distribution.

Claim (a) is verified by invoking Lemma A.1. To demonstrate the claim, we note that $\mathcal{F}$ and F_P satisfy conditions Lemma A.1. The same is also true of $\xi\mathcal{F}_P$ and its envelope $|\xi|F_P$, since ξ is independent of W. Indeed by Lemma A.5 multiplication by ξ does not change qualitatively the uniform entropy bound: $\log \sup_Q N(\varepsilon \||\xi|F_P\|_{Q,2}, \xi\mathcal{F}_P, \|\cdot\|_{Q,2}) \leqslant C \log \sup_Q N(\varepsilon \||\xi|F_P\|_{Q,2}/C, \xi\mathcal{F}_P, \|\cdot\|_{Q,2})$. Moreover, multiplication by ξ does not affect the $\|\cdot\|_{P,2}$ norm of the functions, since ξ is independent of W by construction. The claim then follows.

Claim (b). The previous argument implies unconditional convergence in distribution under any sequence $P = P_n \in \mathcal{P}_n$. Using the same argument as in the first part of the proof of Theorem 2.9.6 in van der Vaart and Wellner (1996), we can claim that the conditional convergence takes place under any sequence $P = P_n \in \mathcal{P}_n$, using the unconditional convergence to establish that the stochastic equicontinuity holds conditionally. The marginal convergence holds by the central limit theorem for triangular array and by tightness of Z_P along any sequence $P \in \mathcal{P}$. ∎

A.4. **Donsker Theorems for Function Classes that depend on** n**.** Let $(W_i)_{i=1}^\infty$ be a sequence of i.i.d. copies of random element W with law $P = P_n$ defined on a measurable space $(S, \mathcal{A}_S)$, and let $w \mapsto f_{n,t}(w)$ be measurable functions from $\mathcal{W}$ to $\mathbb{R}$ indexed by $n \in \mathbb{N}$ and a fixed, totally bounded semi-metric space (T, d_T). Consider the stochastic process

$$(\mathbb{G}_n f_{n,t})_{t\in T} := \left\{ n^{-1/2} \sum_{i=1}^n (f_{n,t}(W_i) - P f_{n,t}) \right\}_{t\in T}.$$

This empirical process is indexed by a class of functions $\mathcal{F}_n = \{f_{n,t} : t \in T\}$ with an envelope function F_n. It is important to note here that the dependency on n allow us to have *the class itself* be possibly dependent on the law $P = P_n$.

Lemma A.2 (**Donsker Theorem for Classes Changing with** n**.**) *For each* n*, let* $\mathcal{F}_n = \{f_{n,t}\}_{t\in T}$ *be a class of suitably measurable functions, and measurable envelope* F_n*, indexed by a fixed, totally bounded semimetric space* (T, d_T)*. Suppose that for some fixed constants* $q > 2$ *and*

that for every sequence $\delta_n \searrow 0$*:*

$$\|F_n\|_{P_n,q} = O(1), \sup_{\rho(s,t)\leqslant\delta_n} \|f_{n,s} - f_{n,t}\|_{P,2} \to 0,$$
$$\int_0^{\delta_n} \sup_Q \sqrt{\log N(\epsilon\|F_n\|_{Q,2}, \mathcal{F}_n, L_2(Q))}d\epsilon \to 0.$$

Then the empirical process $(\mathbb{G}_n f_{n,t})_{t\in T}$ *is asymptotically tight in* $\ell^\infty(T)$*. For any subsequence such that the covariance function* $Pf_{n,s}f_{n,t} - Pf_{n,s}Pf_{n,t}$ *converges pointwise on* $T\times T$*, it converges to a Gaussian process with covariance function given by the limit of the covariance function along that subsequence.*

Proof. This is an immediate consequence of Theorem 2.11.12 in (van der Vaart and Wellner, 1996), p. 220-221. ∎

A.5. **Probabilistic Inequalities.** Let $\sigma^2 > 0$ be any positive constant such that $\sup_{f\in\mathcal{F}} \mathrm{P}f^2 \leqslant \sigma^2 \leqslant \|F\|^2_{\mathrm{P},2}$.

Lemma A.3 (**A Maximal Inequality**)**.** *Let* $\mathcal{F}$ *be an image admissible Suslin set of functions with a measurable envelope* F*. Suppose that* $F = \sup_{f\in\mathcal{F}}|f|$ *with* $\|F\|_{Q,q} < \infty$ *for some* $q \geqslant 2$*. Let* $M = \max_{i\leqslant n} F(W_i)$*. Suppose that there exist constants* $a \geqslant e$ *and* $v \geqslant 1$ *such that*

$$\log \sup_Q N(\varepsilon\|F\|_{Q,2}, \mathcal{F}, \|\cdot\|_{Q,2}) \leqslant v(\log a + \log(1/\varepsilon)),\ 0 < \forall\varepsilon \leqslant 1.$$

Then

$$\mathrm{E}_P[\|\mathbb{G}_n\|_{\mathcal{F}}] \lesssim \sqrt{v\sigma^2 \log\left(\frac{a\|F\|_{P,2}}{\sigma}\right)} + \frac{v\|M\|_{\mathrm{P}_P,2}}{\sqrt{n}} \log\left(\frac{a\|F\|_{P,2}}{\sigma}\right).$$

Moreover, for every $t \geqslant 1$*, with probability* $> 1 - t^{-q/2}$*,*

$$\|\mathbb{G}_n\|_{\mathcal{F}} \leqslant (1+\alpha)\mathrm{E}_P[\|\mathbb{G}_n\|_{\mathcal{F}}] + K(q)\Big[(\sigma + n^{-1/2}\|M\|_{\mathrm{P}_P,q})\sqrt{t} + \alpha^{-1}n^{-1/2}\|M\|_{\mathrm{P}_P,2}t\Big],\ \forall\alpha > 0,$$

where $K(q) > 0$ *is a constant depending only on* q*.*

Proof. See Chernozhukov, Chetverikov, and Kato (2012). ∎

Lemma A.4 (**A Self-Normalized Maximal Inequality**)**.** *Let* $\mathcal{F}$ *be an image-admissible Suslin set of functions with a measurable envelope* F*. Suppose that* $F \geqslant \sup_{f\in\mathcal{F}}|f| \geqslant 1$*, and suppose that there exist some constants* $p > 1$*,* $m \geqslant 1$*, and* $\kappa \geqslant 3 \vee n$ *such that*

$$\log N(\epsilon\|F\|_{\mathbb{P}_n,2}, \mathcal{F}, \|\cdot\|_{\mathbb{P}_n,2}) \leqslant (\kappa/\epsilon)^m,\ 0 < \epsilon < 1.$$

Then for every $\delta \in (0, 1/6)$*, with probability at least* $1 - \delta$*,*

$$\|\mathbb{G}_n\|_{\mathcal{F}} \leqslant (C'/\sqrt{\delta})\sqrt{m\log(\kappa\|F\|_{\mathbb{P}_n,2})}\max\left\{\sup_{f\in\mathcal{F}}\|f\|_{\mathrm{P},2}, \sup_{f\in\mathcal{F}}\|f\|_{\mathbb{P}_n,2}\right\},$$

where the constant C' *is universal.*

Proof. The inequality can be deduced from Belloni and Chernozhukov (2011b), with the exception that the envelope is allowed to be larger than $\sup_{f\in\mathcal{F}}|f|$. ■

Lemma A.5 (**Algebra for Covering Entropies**).

(1) *Let $\mathcal{F}$ be a measurable VC class with a finite VC index k or any other class whose entropy is bounded above by that of such a VC class, then the covering entropy of $\mathcal{F}$ obeys:*

$$\sup_Q \log N(\epsilon\|F\|_{Q,2},\mathcal{F},\|\cdot\|_{Q,2}) \lesssim 1+k\log(1/\epsilon)$$

Examples include $\mathcal{F}=\{\alpha' z,\alpha\in\mathbb{R}^k,\|\alpha\|\leqslant C\}$ and $\mathcal{F}=\{1\{\alpha' z>0\},\alpha\in\mathbb{R}^k,\|\alpha\|\leqslant C\}$.

(2) *For any measurable classes of functions $\mathcal{F}$ and $\mathcal{F}'$:*

$$\log N(\epsilon\|F+F'\|_{Q,2},\mathcal{F}+\mathcal{F}',\|\cdot\|_{Q,2})\leqslant B$$
$$\log N(\epsilon\|F\cdot F'\|_{Q,2},\mathcal{F}\cdot\mathcal{F}',\|\cdot\|_{Q,2})\leqslant B$$
$$\log N(\epsilon\|F\vee F'\|_{Q,2},\mathcal{F}\cup\mathcal{F}',\|\cdot\|_{Q,2})\leqslant B$$
$$B=\log N\left(\frac{\epsilon}{2}\|F\|_{Q,2},\mathcal{F},\|\cdot\|_{Q,2}\right)+\log N\left(\frac{\epsilon}{2}\|F'\|_{Q,2},\mathcal{F}',\|\cdot\|_{Q,2}\right).$$

(3) *Given a measurable class $\mathcal{F}$ and a random variable ξ:*

$$\log\sup_Q N(\epsilon\||\xi|F\|_{Q,2},\xi\mathcal{F},\|\cdot\|_{Q,2})\lesssim\log\sup_Q N\left(\epsilon/2\|F\|_{Q,2},\mathcal{F},\|\cdot\|_{Q,2}\right)$$

(4) *For the class $\mathcal{F}^*$ created by integrating $\mathcal{F}$, i.e.*

$$\mathcal{F}=\left\{f^*:f^*(x):=\int f(x,y)d\mu(y),\ \textit{for some } \mu \textit{ a probability measure}\right\}$$

, we have that

$$\log N(\epsilon\|F\|_{Q,2},\mathcal{F}^*,\|\cdot\|_{Q,2})\leqslant\log N\left(\epsilon\|F\|_{Q,2},\mathcal{F},\|\cdot\|_{Q,2}\right)$$

Proof. For the proof of assertions (1)-(3) see, e.g., Andrews (1994). The fact (4) was noted in Chandraksekhar et al (2011), though it is rather elementary and follows from convexity of the norm and Jensen's inequality: $\|f^*-\tilde{f}^*\|_{Q,2}\leqslant\int\|f-\tilde{f}\|_{Q,2}d\mu=\|f-\tilde{f}\|_{Q,2}$, from which the stated bound follows immediately. In other words, any averaging done over components of the function contracts distances between functions and therefore does not expand the covering entropy. A related, slightly different bound is stated in Ghosal and Van der Vaart (2009), but we need the bound above. ■

Lemma A.6 (**Contractivity of Conditional Expectation**). *Let (V,X) and (V',X) be random vectors in $\mathbb{R}\times\mathbb{R}^k$ defined on the probability space $(S,\mathcal{A}_S,Q)$, with the first components being scalar, then for any $1\leqslant q\leqslant\infty$,*

$$\|\mathrm{E}_Q(V|X)-\mathrm{E}_Q(V'|X)\|_{Q,q}\leqslant\|V-V'\|_{Q,q}.$$

This is an instance of a well known result on the contractive property of the conditional expectation. We recall it here since we shall use it frequently.

A.6. **Hadamard Differentiability for Sequences and Delta Method for Sequences.** We shall use the functional delta method, as formulated in van der Vaart and Wellner (1996). Let $\mathbb{D}_0$, $\mathbb{D}$, and $\mathbb{E}$ be normed spaces, with $\mathbb{D}_0 \subset \mathbb{D}$. A map $\phi : \mathbb{D}_\phi \subset \mathbb{D} \mapsto \mathbb{E}$ is called *Hadamard-differentiable* at $\theta \in \mathbb{D}_\phi$ tangentially to $\mathbb{D}_0$ if there is a continuous linear map $\phi'_\theta : \mathbb{D}_0 \mapsto \mathbb{E}$ such that

$$\frac{\phi(\theta + t_n h_n) - \phi(\theta)}{t_n} \to \phi'_\theta(h), \quad n \to \infty,$$

for all sequences $t_n \to 0$ and $h_n \to h \in \mathbb{D}_0$ such that $\theta + t_n h_n \in \mathbb{D}_\phi$ for every n.

A map $\phi : \mathbb{D}_\phi \subset \mathbb{D} \mapsto \mathbb{E}$ is called *Hadamard-differentiable uniformly* in $\theta \in \mathbb{D}_\phi$, a compact subset of $\mathbb{D}$, tangentially to $\mathbb{D}_0$, if

$$\left| \frac{\phi(\theta + t_n h_n) - \phi(\theta)}{t_n} - \phi'_\theta(h) \right| \to 0, \quad n \to \infty,$$

for all sequences $\theta_n \to \theta$, $t_n \to 0$, and $h_n \to h \in \mathbb{D}_0$ such that $\theta + t_n h_n \in \mathbb{D}_\phi$ for every n. As a part of the definition, we require that the map $h \mapsto \phi'_\theta(h)$ from $\mathbb{D}_0$ to $\mathbb{E}$ is continuous and linear, and that the map $(\theta, h) \mapsto \phi'_\theta(h)$ from $\mathbb{D}_\phi \times \mathbb{D}_0$ to $\mathbb{E}$ is continuous.

Lemma A.7 (Functional delta-method for sequences)**.** *Let $\mathbb{D}_0$, $\mathbb{D}$, and $\mathbb{E}$ be normed spaces. Let $\phi : \mathbb{D}_\phi \subset \mathbb{D} \mapsto \mathbb{E}$ be Hadamard-differentiable uniformly in $\theta \in \mathbb{D}_\phi$ tangentially to $\mathbb{D}_0$, with derivative map ϕ'_θ. Let X_n be a sub-sequence of stochastic processes taking values in $\mathbb{D}_\phi$ such that $r_n(X_n - \theta_n) \rightsquigarrow X$ and $\theta_n \to \theta$ in $\mathbb{D}$ along a subsequence $n \in \mathbb{Z}' \subset \mathbb{Z}$, where X possibly depends on $\mathbb{Z}'$ and is separable and takes its values in $\mathbb{D}_0$, for some sequence of constants $r_n \to \infty$. Then $r_n(\phi(X_n) - \phi(\theta_n)) \rightsquigarrow \phi'_\theta(X)$ in $\mathbb{E}$ along the same subsequence. If ϕ'_θ is defined and continuous on the whole of $\mathbb{D}$, then the sequences $r_n(\phi(X_n) - \phi(\theta_n)) - \phi'_{\theta_n}(r_n(X_n - \theta_n))$ and $\phi'_{\theta_n}(r_n(X_n - \theta_n)) - \phi'_\theta(r_n(X_n - \theta_n))$ converge to zero in outer probability along the same subsequence.*

Let $D_n = (W_i)_{i=1}^n$ denote the data vector and $M_n = (\xi_i)_{i=1}^n$ be a vector of random variables, used to generate bootstrap draws or simulation draws (this may depend on particular method). Consider sequences of stochastic processes $X_n = X_n(D_n)$, where the sequence $G_n = \sqrt{n}(X_n - \theta_n)$ weakly converges unconditionally to the tight random element G in the normed space $\mathbb{D}$ along a subsequence, and $\theta_n \to \theta$. This means that

$$\sup_{h \in \mathrm{BL}_1(\mathbb{D}, \mathbb{R})} |\mathrm{E}^*_{P_n} h(G_n) - \mathrm{E}_G h(G)| \to 0,$$

along $n \in \mathbb{Z}'$, where E_G denotes the expectation computed with respect to the law of G. This is denoted as $G_n \rightsquigarrow G$ along $n \in \mathbb{Z}'$. Also consider the bootstrap stochastic process $G^*_n = G_n(D_n, M_n)$ in $\mathbb{D}$, where G_n is a measurable function of M_n for each value of D_n. Suppose that G^*_n converges conditionally given D_n in distribution to G, in probability, that is

$$\sup_{h \in \mathrm{BL}_1(\mathbb{D}, \mathbb{R})} |\mathrm{E}_{M_n}[h(G^*_n)] - \mathrm{E}_G h(G)| \to 0,$$

in outer probability along $n \in \mathbb{Z}'$, where E_{M_n} denotes the expectation computed with respect to the law of M_n holding the data D_n fixed. This is denoted as $G^*_n \rightsquigarrow_B G$ along $n \in \mathbb{Z}'$, respectively.

Let $X_n^* = X_n + G_n^*/\sqrt{n}$ denote the bootstrap or simulation draw of X_n.

Lemma A.8 (Delta-method for bootstrap and other simulation methods for sequences)**.** *Let $\mathbb{D}_0$, $\mathbb{D}$, and $\mathbb{E}$ be normed spaces, with $\mathbb{D}_0 \subset \mathbb{D}$. Let $\phi : \mathbb{D}_\phi \subset \mathbb{D} \mapsto \mathbb{E}$ be Hadamard-differentiable uniformly in $\theta \in \mathbb{D}_\phi$ tangentially to $\mathbb{D}_0$, with the derivative map ϕ'_θ. Let X_n and X_n^* be maps as indicated previously with values in $\mathbb{D}_\phi$ such that $\sqrt{n}(X_n - \theta_n) \rightsquigarrow G$, $\theta_n \to \theta$, and $\sqrt{n}(X_n^* - X_n) \rightsquigarrow_B G$ in $\mathbb{D}$ along a subsequence of integers $n \in \mathbb{Z}' \subset \mathbb{Z}$, where G is separable and takes its values in $\mathbb{D}_0$. Then $\sqrt{n}(\phi(X_n^*) - \phi(\theta_n)) \rightsquigarrow_B \phi'_\theta(G)$ in $\mathbb{E}$ along the same subsequence.*

Another technical result that we use in the sequel concerns the equivalence of continuous and uniform convergence.

Lemma A.9 (Uniform convergence via continuous convergence)**.** *Let $\mathbb{D}$ and $\mathbb{E}$ be complete separable metric spaces, with $\mathbb{D}$ compact. Suppose $f : \mathbb{D} \mapsto \mathbb{E}$ is continuous. Then a sequence of functions $f_n : \mathbb{D} \mapsto \mathbb{E}$ converges to f uniformly on $\mathbb{D}$ if and only if for any convergent sequence $x_n \to x$ in $\mathbb{D}$ we have that $f_n(x_n) \to f(x)$.*

Proofs of Lemmas A.7 and A.8. The result follows from the proofs in VW, Chap. 3.9, where proofs (pointwise in θ) are given for a sequence of integers $n \in \{1, 2, ...\}$. The claim extends to subsequences trivially. ■

Proof of Lemma A.9. See, for example, Resnick (1987, p. 2).

Appendix B. Proofs for Section 4

B.1. **Proof of Theorem 4.1.** The results for the two strategies have similar structure, so we only give the proof for Strategy 1.

Step 0. (A Preamble). In the proof $a \lesssim b$ means that $a \leqslant Ab$, where the constant A depends on the constants in Assumptions only, but not on n once $n \geqslant n_0 = \min\{j : \delta_j \leqslant 1/2\}$, and not on $P \in \mathcal{P}_n$. We consider a sequence P_n in $\mathcal{P}_n$, but for simplicity, we write $P = P_n$ throughout the proof, *suppressing* the index n. Since the argument is asymptotic, we can just assume that $n \geqslant n_0$ in what follows.

To proceed with the presentation of proofs, it might be convenient for the reader to have the notation collected in one place. The influence function and low-bias moment functions for $\alpha_V(z)$ for $z \in \mathcal{Z} = \{0, 1\}$ are given respectively by:

$$\psi^\alpha_{V,z}(W) := \psi^\alpha_{V,z,g_V,m_Z}(W, \alpha_V(z)), \quad \psi^\alpha_{V,z,g,m}(W, \alpha) := \frac{1(Z = z)(V - g(z, X))}{m(z, X)} + g(z, X) - \alpha.$$

The influence functions and the moment functions for γ_V are given by $\psi^\gamma_V(W) := \psi^\gamma_V(W, \gamma_V)$ and $\psi^\gamma_V(W, \gamma) := V - \gamma$. Recall that the the estimator of the reduced-form parameters $\alpha_V(z)$ and γ_V are solutions $\alpha = \widehat{\alpha}_V(z)$ and $\gamma = \widehat{\gamma}_V$ to the equations:

$$\mathbb{E}_n[\psi^\alpha_{V,z,\widehat{g}_V,\widehat{m}_Z}(W, \alpha)] = 0, \quad \mathbb{E}_n[\psi^\gamma_V(W, \gamma)] = 0,$$

where $\widehat{g}_V(z,x) = \Lambda_V(f(z,x)'\bar{\beta}_V)$ and $\widehat{m}_Z(z,x) = \Lambda_Z(f(x)'\bar{\beta}_Z)$, where $\bar{\beta}_V$ and $\bar{\beta}_Z$ are estimators as in Assumption 3. For each variable name $V \in V_u$,

$$V_u := (V_{uj})_{j=1}^5 := (Y_u, \mathbf{1}_0(D)Y_u, \mathbf{1}_0(D), \mathbf{1}_1(D)Y_u, \mathbf{1}_1(D)),$$

we obtain the estimator $\widehat{\rho}_u := \left(\{\widehat{\alpha}_V(0), \widehat{\alpha}_V(1), \widehat{\gamma}_V\}\right)_{V\in V_u}$ of $\rho_u := \left(\{\alpha_V(0), \alpha_V(1), \gamma_V\}\right)_{V\in V_u}$. The estimator and the estimand are vectors in $\mathbb{R}^{d_\rho}$ with a finite dimension. We stack these vectors into the processes $\widehat{\rho} = (\widehat{\rho}_u)_{u\in\mathcal{U}}$ and $\rho = (\rho_u)_{u\in\mathcal{U}}$.

Step 1.(Linearization) In this step we establish the first claim, namely that

$$\sqrt{n}(\widehat{\rho} - \rho) = Z_{n,P} + o_P(1) \quad \text{in } \ell^\infty(\mathcal{U})^{d_\rho}, \tag{73}$$

where $Z_{n,P} := (\mathbb{G}_n \psi_u^\rho)_{u\in\mathcal{U}}$. The components $(\sqrt{n}(\widehat{\gamma}_{V_{uj}} - \gamma_{V_{uj}}))_{u\in\mathcal{U}}$ of $\sqrt{n}(\widehat{\rho} - \rho)$ trivially have the linear representation (with no error) for each $j \in \mathcal{J}$. We only need to establish the claim for the empirical process $(\sqrt{n}(\widehat{\alpha}_{V_{uj}}(z) - \alpha_{V_{uj}}(z)))_{u\in\mathcal{U}}$ for $z \in \{0,1\}$, which we do in the steps below.

(a) We make some preliminary observations. For $t = (t_1, t_2, t_3, t_4) \in \mathbb{R}^2 \times (0,1)^2$ and $v \in \mathbb{R}$, $(z, \bar{z}) \in \{0,1\}^2$, we define the function $(v, z, \bar{z}, t) \mapsto \varphi(v, z, \bar{z}, t)$ via:

$$\varphi(v,z,1,t) = \frac{1(z=1)(v-t_2)}{t_4} + t_2, \quad \varphi(v,z,0,t) = \frac{1(z=0)(v-t_1)}{t_3} + t_1.$$

The derivatives of this function with respect to t obey for all $k = (k_j)_{j=1}^4 \in \mathbb{N}^4 : 0 \leqslant |k| \leqslant 4$,

$$|\partial_t^k \varphi(v,z,\bar{z},t)| \leqslant L, \quad \forall (v,\bar{z},z,t) : |v| \leqslant C, |t_1|, |t_2| \leqslant C, c'/2 \leqslant |t_3|, |t_4| \leqslant 1 - c'/2, \tag{74}$$

where L depends only on c' and C, $|k| = \sum_{j=1}^4 k_j$, and $\partial_t^k := \partial_{t_1}^{k_1}\partial_{t_2}^{k_2}\partial_{t_3}^{k_3}\partial_{t_4}^{k_4}$.

(b) Let

$$\begin{aligned}
&\widehat{h}_V(X_i) := (\widehat{g}_V(0,X_i), \widehat{g}_V(1,X_i), 1 - \widehat{m}(1,X_i), \widehat{m}(1,X_i))', \\
&h_V(X_i) := (g_V(0,X_i), g_V(1,X_i), m(0,X_i), m(1,X_i))', \\
&f_{\widehat{h}_V,V,z}(W) := \varphi(V,Z,z,\widehat{h}_V(X_i)), \\
&f_{h_V,V,z}(W) := \varphi(V,Z,z,h_V(X_i)).
\end{aligned}$$

We observe that with probability no less than $1 - \Delta_n$,

$$\widehat{g}_V(0,\cdot) \in \mathcal{G}_V(0), \quad \widehat{g}_V(1,\cdot) \in \mathcal{G}_V(1), \quad \widehat{m}(1,\cdot) \in \mathcal{M}(1), \quad \widehat{m}(0,\cdot) \in \mathcal{M}(0) = 1 - \mathcal{M}(1),$$

where

$$\mathcal{G}_V(z) := \left\{ \begin{array}{l} (x,z) \mapsto \Lambda_V(f(z,x)'\beta) : \|\beta\|_0 \leqslant sC \\ \|\Lambda_V(f(z,X)'\beta) - g_V(z,X)\|_{\mathrm{P},2} \lesssim \delta_n n^{-1/4} \\ \|\Lambda_V(f(z,X)'\beta) - g_V(z,X)\|_{\mathrm{P},\infty} \lesssim \epsilon_n \end{array} \right\},$$

$$\mathcal{M}(1) := \left\{ \begin{array}{l} x \mapsto \Lambda_Z(f(x)'\beta) : \|\beta\|_0 \leqslant sC \\ \|\Lambda_Z(f(X)'\beta) - m_Z(1,X)\|_{\mathrm{P},2} \lesssim \delta_n n^{-1/4} \\ \|\Lambda_Z(f(X)'\beta) - m_Z(1,X)\|_{\mathrm{P},\infty} \lesssim \epsilon_n \end{array} \right\}.$$

To see this note, that under Assumption 3, under conditions (i)-(ii), under the event occurring under condition (ii) of that assumption: for all $n \geqslant \min\{j : \delta_j \leqslant 1/2\}$,

$$\begin{aligned}
\|\Lambda_Z(f(X)'\beta) - m_Z(1,X)\|_{\mathrm{P},2} &\leqslant \|\Lambda_Z(f(X)'\beta) - \Lambda_Z(f(X)'\beta_Z)\|_{\mathrm{P},2} + \|r_Z(X)\|_{\mathrm{P},2} \\
&\lesssim \|\partial\Lambda_Z\|_\infty \|f(X)'(\beta-\beta_Z)\|_{\mathrm{P},2} + \|r_Z(X)\|_{\mathrm{P},2} \\
&\lesssim \|\partial\Lambda_Z\|_\infty \|f(X)'(\beta-\beta_Z)\|_{\mathbb{P}_n,2} + \|r_Z(X)\|_{\mathrm{P},2} \lesssim \delta_n n^{-1/4} \\
\|\Lambda_Z(f(X)'\beta) - m_Z(1,X)\|_{\mathrm{P},\infty} &\leqslant \|\Lambda_Z(f(X)'\beta) - \Lambda_Z(f(X)'\beta_Z)\|_{\mathrm{P},\infty} + \|r_Z(X)\|_{\mathrm{P},\infty} \\
&\leqslant \|\partial\Lambda_Z\|_\infty \|f(X)'(\beta-\beta_Z)\|_{\mathrm{P},\infty} + \|r_Z(X)\|_{\mathrm{P},\infty} \\
&\lesssim K_n\|\beta-\beta_Z\|_1 + \epsilon_n \leqslant 2\epsilon_n,
\end{aligned}$$

for $\beta = \widehat{\beta}_Z$, with evaluation after computing the norms, and for $\|\partial\Lambda\|_\infty$ denoting $\sup_{l\in\mathbb{R}}|\partial\Lambda(l)|$ here and below. Similarly, under Assumption 3,

$$\begin{aligned}
\|\Lambda_V(f(Z,X)'\beta) - g_V(Z,X)\|_{\mathrm{P},2} &\lesssim \|\partial\Lambda_V\|_\infty \|f(Z,X)'(\beta-\beta_V)\|_{\mathbb{P}_n,2} + \|r_V(Z,X)\|_{\mathrm{P},2} \lesssim \delta_n n^{-1/4} \\
\|\Lambda_V(f(Z,X)'\beta) - g_V(Z,X)\|_{\mathrm{P},\infty} &\lesssim K_n\|\beta-\beta_V\|_1 + \epsilon_n \leqslant 2\epsilon_n,
\end{aligned}$$

for $\beta = \widehat{\beta}_V$, with evaluation after computing the norms, and noting that for any β

$$\|\Lambda_V(f(0,X)'\beta) - g_V(0,X)\|_{\mathrm{P},2} \vee \|\Lambda_V(f(1,X)'\beta) - g_V(1,X)\|_{\mathrm{P},2} \lesssim \|\Lambda_V(f(Z,X)'\beta) - g_V(Z,X)\|_{\mathrm{P},2}$$

under condition (iii) of Assumption 2, and trivially

$$\|\Lambda_V(f(0,X)'\beta) - g_V(0,X)\|_{\mathrm{P},\infty} \vee \|\Lambda_V(f(1,X)'\beta) - g_V(1,X)\|_{\mathrm{P},\infty} \leqslant \|\Lambda_V(f(Z,X)'\beta) - g_V(Z,X)\|_{\mathrm{P},\infty}$$

under condition (iii) of Assumption 2.

Hence with probability at least $1-\Delta_n$,

$$\widehat{h}_V \in \mathcal{H}_{V,n} := \{h = (\bar{g}(0,\cdot), \bar{g}(1,\cdot), \bar{m}(0,\cdot), \bar{m}(1,\cdot)) \in \mathcal{G}_V(0) \times \mathcal{G}_V(1) \times \mathcal{M}(0) \times \mathcal{M}(1)\}.$$

(c) We have that

$$\alpha_V(z) = \mathrm{E}_P[f_{h_V,V,z}] \text{ and } \widehat{\alpha}(z) = \mathbb{E}_n[f_{\widehat{h}_V,V,z}],$$

so that

$$\sqrt{n}(\widehat{\alpha}_V(z) - \alpha_V(z)) = \underbrace{\mathbb{G}_n[f_{h_V,V,z}]}_{I_V(z)} + \underbrace{(\mathbb{G}_n[f_{h,V,z}] - \mathbb{G}_n[f_{h_V,V,z}])}_{II_V(z)} + \underbrace{\sqrt{n}(\mathrm{E}_P[f_{h,V,z} - f_{h_V,h,z}])}_{III_V(z)},$$

with h evaluated at $h = \widehat{h}_V$.

(d) Note that for $\Delta_{V,i} = h(Z_i,X_i) - h_V(Z_i,X_i)$ and $\Delta_{V,i}^k = \Delta_{1V,i}^{k_1}\Delta_{2V,i}^{k_2}\Delta_{3V,i}^{k_3}\Delta_{4V,i}^{k_4}$,

$$\begin{aligned}
III_V(z) &= \sqrt{n}\sum_{|k|=1} \mathrm{E}_P[\partial_t^k \varphi(V_i,Z_i,z,h_V(Z_i,X_i))\Delta_{V,i}^k] \\
&+ \sqrt{n}\sum_{|k|=2} 2^{-1}\mathrm{E}_P[\partial_t^k \varphi(V_i,Z_i,z,h_V(Z_i,X_i))\Delta_{V,i}^k] \\
&+ \sqrt{n}\sum_{|k|=3} \int_0^1 6^{-1}\mathrm{E}_P[\partial_t^k \varphi(V_i,Z_i,z,h_V(Z_i,X_i)+\lambda\Delta_{V,i})\Delta_{V,i}^k]d\lambda, \\
&=: III_V^a(z) + III_V^b(z) + III_V^b(z),
\end{aligned}$$

(with h evaluated at $h = \widehat{h}$ after computing expectations).

By the law of iterated expectations and the low-bias property of the estimation equation for α_V,

$$\mathrm{E}_P[\partial_t^k \varphi(V_i, Z_i, z, h_V(Z_i, X_i))|Z_i, X_i] = 0 \quad \forall k \in \mathbb{N}^3 : |k| = 1,$$

so we have that $III_V^a(z) = 0$.

Moreover, uniformly for any $h \in \mathcal{H}_{V,n}$ we have that, in view of properties noted in step (a),

$$|III_V^b(z)| \lesssim \sqrt{n}\|h - h_V\|_{\mathrm{P},2}^2 \lesssim \sqrt{n}(\delta_n n^{-1/4})^2 \leqslant \delta_n^2,$$

$$|III_V^c(z)| \lesssim \sqrt{n}\|h - h_V\|_{\mathrm{P},2}^2 \|h - h_V\|_{\mathrm{P},\infty} \lesssim \sqrt{n}(\delta_n n^{-1/4})^2 \epsilon_n \leqslant \delta_n^2 \epsilon_n.$$

Since $\widehat{h}_V \in \mathcal{H}_{V,n}$ for all $V \in \mathcal{V}$ with probability $1 - \Delta_n$, we have that once $n \geqslant n_0$,

$$\mathrm{P}_P\Big(|III_V(z)| \lesssim \delta_n^2, \forall z \in \{0,1\}, \forall V \in \mathcal{V}\Big) \geqslant 1 - \Delta_n.$$

(e) Furthermore, we have that

$$\sup_{V \in \mathcal{V}} \max_{z \in \{0,1\}} |II_V(z)| \leqslant \sup_{h \in \mathcal{H}_{V,n}, z \in \{0,1\}, V \in \mathcal{V}} |\mathbb{G}_n[f_{h,V,z}] - \mathbb{G}_n[f_{h_V,V,z}]|.$$

The classes of functions, viewed as maps from the sample space S to the real line,

$$\mathcal{V} := \{V_{uj}, u \in \mathcal{U}, j \in \mathcal{J}\} \quad \text{and} \quad \mathcal{V}^* := \{g_{V_{uj}}(Z, X), u \in \mathcal{U}, j \in \mathcal{J}\}$$

are bounded by a constant envelope and have the uniform covering ϵ-entropy bounded by a multiple of $\log(\mathrm{e}/\epsilon) \vee 0$, that is $\log \sup_Q N(\epsilon, \mathcal{V}, \|\cdot\|_{Q,2}) \lesssim \log(\mathrm{e}/\epsilon) \vee 0$, which holds by Assumption 2, and $\log \sup_Q N(\epsilon, \mathcal{V}^*, \|\cdot\|_{Q,2}) \lesssim \log(\mathrm{e}/\epsilon) \vee 0$ which holds by contractivity of conditions expectations noted in Lemma A.6 (or by Lemma A.5, item (iv)). The uniform covering ϵ-entropy of the function set $\mathcal{B} = \{1(Z = z), z \in \{0,1\}\}$ is trivially bounded by $\log(\mathrm{e}/\epsilon) \vee 0$.

The class of functions

$$\mathcal{G} := \{\mathcal{G}_V(z), V \in \mathcal{V}, z \in \{0,1\}\}$$

has a constant envelope and is a subset of

$$\{(x,z) \mapsto \Lambda(f(z,x)'\beta) : \|\beta\|_0 \leqslant sC, \Lambda \in \mathcal{L} = \{\mathrm{Id}, \Phi, 1 - \Phi, \Lambda_0, 1 - \Lambda_0\}\},$$

which is a union of 5 sets of the form

$$\{(x,z) \mapsto \Lambda(f(z,x)'\beta) : \|\beta\|_0 \leqslant sC\}$$

with $\Lambda \in \mathcal{L}$ a fixed monotone function for each of the 5 sets; each of these sets are the unions of at most $\binom{p}{Cs}$ VC-subgraph classes of functions with VC indices bounded by $C's$ (note that a fixed monotone transformations Λ preserves the VC-subgraph property). Therefore

$$\log \sup_Q N(\varepsilon, \mathcal{G}, \|\cdot\|_{Q,2}) \lesssim (s \log p + s \log(1/\varepsilon)) \vee 0.$$

Similarly, the class of functions $\mathcal{M} = (\mathcal{M}(1) \cup (1 - \mathcal{M}(1)))$ has a constant envelope, which is a union of at most 5 sets, which are themselves the unions of at most $\binom{p}{Cs}$ VC-subgraph classes

of functions with VC indices bounded by $C's$ (a fixed monotone transformations Λ preserves the VC-subgraph property). Therefore, $\log \sup_Q N(\varepsilon, \mathcal{M}, \|\cdot\|_{Q,2}) \lesssim (s \log p + s \log(1/\varepsilon)) \vee 0$.

Finally, the set of functions

$$\mathcal{F}_n = \{f_{h,V,z} - f_{h_V,V,z} : z \in \{0,1\}, V \in \mathcal{V}, h \in \mathcal{H}_{V,n}\},$$

is a Lipschitz transform of function sets $\mathcal{V}$, $\mathcal{V}^*$, $\mathcal{B}$, $\mathcal{G}$, $\mathcal{M}$, with bounded Lipschitz coefficients and with a constant envelope. Therefore, we have that

$$\log \sup_Q N(\varepsilon, \mathcal{F}_n, \|\cdot\|_{Q,2}) \lesssim (s \log p + s \log(1/\varepsilon)) \vee 0.$$

Applying Lemma A.3 and Markov inequality, we have for some constant $K > e$

$$\begin{aligned}
&\sup_{V \in \mathcal{V}} \max_{z \in \{0,1\}} |II_V(z)| \leqslant \sup_{f \in \mathcal{F}_n} |\mathbb{G}_n(f)| \\
&= O_{\mathrm{P}}(1) \left(\sqrt{s\sigma_n^2 \log(p \vee K \vee \sigma_n^{-1})} + \frac{s}{\sqrt{n}} \log(p \vee K \vee \sigma_n^{-1}) \right) \\
&= O_{\mathrm{P}}(1) \left(\sqrt{s\delta_n^2 n^{-1/2} \log(p \vee n)} + \sqrt{s^2 n^{-1} \log^2(p \vee n)} \right) \\
&= O_{\mathrm{P}}(1) \left(\delta_n \delta_n^{1/4} + \delta_n^{1/2} \right) = O_{\mathrm{P}}(\delta_n^{1/2}),
\end{aligned}$$

for $\sigma_n = \sup_{f \in \mathcal{F}_n} \|f\|_{P,2}$; and we used some simple calculations, exploiting the boundedness conditions in Assumptions 2 and 3, to deduce that,

$$\sigma_n = \sup_{f \in \mathcal{F}_n} \|f\|_{P,2} \lesssim \sup_{h \in \mathcal{H}_{V,n}, V \in \mathcal{V}} \|h - h_V\|_{P,2} \lesssim \delta_n n^{-1/4},$$

since $\sup_{h \in \mathcal{H}_{V,n}, V \in \mathcal{V}} \|h - h_V\|_{P,2} \lesssim \delta_n n^{-1/4}$ by definition of the set $\mathcal{H}_{V,n}$; and then we used that $s^2 \log^3(p \vee n)/n \leqslant \delta_n$ by Assumption 3.

(f) The claim of Step 1 follows by collecting steps (a)-(e).

Step 2 (Uniform Donskerness). Here we claim that Assumption 2 implies two assertions:

(a) The set of vector functions $(\psi_u^\rho)_{u \in \mathcal{U}}$, where $\psi_u^\rho := (\{\psi_{V,0}^\alpha, \psi_{V,1}^\alpha, \psi_V^\gamma\})_{V \in V_u}$, is P-Donsker uniformly in $\mathcal{P}$, namely that

$$Z_{n,P} \rightsquigarrow Z_P \quad \text{in } \ell^\infty(\mathcal{U})^{d_\rho}, \text{ uniformly in } P \in \mathcal{P},$$

where $Z_{n,P} := (\mathbb{G}_n \psi_u^\rho)_{u \in \mathcal{U}}$ and $Z_P := (\mathbb{G}_P \psi_u^\rho)_{u \in \mathcal{U}}$.

(b) Moreover, Z_P has bounded, uniformly continuous paths uniformly in $P \in \mathcal{P}$:

$$\sup_{P \in \mathcal{P}} \mathrm{E}_P \sup_{u \in \mathcal{U}} \|Z_P(u)\| < \infty, \quad \lim_{\varepsilon \searrow 0} \sup_{P \in \mathcal{P}} \mathrm{E}_P \sup_{d_{\mathcal{U}}(u,\tilde{u}) \leqslant \varepsilon} \|Z_P(u) - Z_P(\tilde{u})\| = 0.$$

To verify these claims we shall invoke Lemma A.1.

To demonstrate the claim, it will suffice to consider the set of $\mathbb{R}$-valued functions $\Psi = (\psi_{uk}, u \in \mathcal{U}, k \in 1, ..., d_\rho)$. Further, we notice that $\mathbb{G}_n \psi_{V,z}^\alpha = \mathbb{G}_n f$, for $f \in \mathcal{F}_z$,

$$\mathcal{F}_z = \left\{ \frac{1\{Z = z\}(V - g_V(z,X))}{m_Z(z,X)} + g_V(z,X), V \in \mathcal{V} \right\},$$

and that $\mathbb{G}_n\psi_V^\gamma = \mathbb{G}_n f$, for $f = V \in \mathcal{V}$. Hence $\mathbb{G}_n(\psi_{uk}) = \mathbb{G}_n(f)$ for $f \in \mathcal{F}_P = \mathcal{F}_0 \cup \mathcal{F}_1 \cup \mathcal{V}$. We thus need to check that conditions of Lemma A.1 apply $\mathcal{F}$ uniformly in $P \in \mathcal{P}$.

Observe that $\mathcal{F}_z$ is formed as a uniform Lipschitz transform of function sets $\mathcal{B}, \mathcal{V}, \mathcal{V}^*, \mathcal{M}$ where the validity of the Lipschitz property relies on Assumption 2 (to keep the denominator away from zero) and on boundedness conditions in Assumption 3. The latter function sets are uniformly bounded classes that have the uniform covering ϵ-entropy bounded by $\log(\mathrm{e}/\epsilon) \vee 0$ up to a multiplicative constant, and so this class, which is uniformly bounded under Assumption 2, has the uniform ϵ-entropy bounded by $\log(\mathrm{e}/\epsilon) \vee 0$ up to a multiplicative constant (e.g. van der Vaart and Wellner (1996)). Since $\mathcal{F}_P$ is uniformly bounded and is a finite union of function sets with the uniform entropies obeying the said properties, it also follows that it has this property, namely:

$$\sup_{P\in\mathcal{P}} \log \sup_{Q} N(\varepsilon, \mathcal{F}_P, \|\cdot\|_{Q,2}) \lesssim \log(\mathrm{e}/\epsilon) \vee 0.$$

Since $\int_0^\infty \sqrt{\log(\mathrm{e}/\epsilon) \vee 0} d\epsilon = \mathrm{e}\sqrt{\pi}/2 < \infty$ and $\mathcal{F}$ is uniformly bounded, the entropy and bounded moments condition in Lemma A.1 holds.

We demonstrate other conditions. Consider a sequence of positive constants ε approaching zero, and note that

$$\sup_{d_{\mathcal{U}}(u,\tilde{u})\leqslant\varepsilon} \max_{k\leqslant d_\rho} \|\psi_{uk} - \psi_{\tilde{u}k}\|_{P,2} \lesssim \sup_{d_{\mathcal{U}}(u,\tilde{u})\leqslant\varepsilon} \|f_u - f_{\tilde{u}}\|_{P,2}$$

where f_u and $f_{\tilde{u}}$ must be of the form:

$$\frac{1\{Z=z\}(U_u - g_{U_u}(z,X))}{m_Z(z,X)} + g_{U_u}(z,X), \frac{1\{Z=z\}(U_{\tilde{u}} - g_{U_{\tilde{u}}}(z,X))}{m_Z(z,X)} + g_{U_{\tilde{u}}}(z,X),$$

with $(U_u, U_{\tilde{u}})$ equal to either $(Y_u, Y_{\tilde{u}})$ or $(1_d(D)Y_u, 1_d(D)Y_{\tilde{u}})$, for $d = 0$ or 1, and $z = 0$ or 1. Then

$$\sup_{P\in\mathcal{P}} \|f_u - f_{\tilde{u}}\|_{P,2} \lesssim \sup_{P\in\mathcal{P}} \|Y_u - Y_{\tilde{u}}\|_{P,2} \to 0,$$

as $d_{\mathcal{U}}(u,\tilde{u}) \to 0$ by Assumption 2. Indeed, $\sup_{P\in\mathcal{P}} \|f_u - f_{\tilde{u}}\|_{P,2} \lesssim \sup_{P\in\mathcal{P}} \|Y_u - Y_{\tilde{u}}\|_{P,2}$ follows follows from a sequence of inequalities holding uniformly in $P \in \mathcal{P}$: (1)

$$\|f_u - f_{\tilde{u}}\|_{P,2} \lesssim \|U_u - U_{\tilde{u}}\|_{P,2} + \|g_{U_u}(z,X) - g_{U_{\tilde{u}}}(z,X)\|_{P,2},$$

which we deduced using triangle inequality and the fact that $m_Z(z,X)$ is bounded away from zero, (2) $\|U_u - U_{\tilde{u}}\|_{P,2} \leqslant \|Y_u - Y_{\tilde{u}}\|_{P,2}$, which we deduced using a Holder inequality, (3)

$$\|g_{U_u}(z,X) - g_{U_{\tilde{u}}}(z,X)\|_{P,2} \leqslant \|U_u - U_{\tilde{u}}\|_{P,2},$$

which we deduced by the definition of $g_{U_u}(z,X) = \mathrm{E}_P(U_u|X, Z=z)$ and the contraction property of conditional expectation recalled in Lemma A.6. ■

B.2. **Proof of Theorem 4.2.** The proof will be similar to the previous proof, and as in that proof we only focus the presentation on the first strategy.

STEP 0. (A Preamble). In the proof $a \lesssim b$ means that $a \leqslant Ab$, where the constant A depends on the constants in Assumptions only, but not on n once $n \geqslant n_0 = \min\{j : \delta_j \leqslant 1/2\}$, and not on $P \in \mathcal{P}_n$. We consider a sequence P_n in $\mathcal{P}_n$, but for simplicity, we write $P = P_n$ throughout the proof, suppressing the index n. Since the argument is asymptotic, we can just assume that $n \geqslant n_0$ in what follows.

Let $\mathbb{P}_n$ denote the measure that puts mass n^{-1} on points (ξ_i, W_i) for $i = 1, ..., n$. Let $\mathbb{E}_n$ denote the expectation with respect to this measure, so that $\mathbb{E}_n f = n^{-1} \sum_{i=1}^n f(\xi_i, W_i)$.

Recall that we define the bootstrap draw as:

$$\sqrt{n}(\widehat{\rho}^* - \widehat{\rho}) = \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i \widehat{\psi}_u^\rho(W_i) \right)_{u \in \mathcal{U}} = \left(\mathbb{G}_n(\xi \widehat{\psi}_u^\rho) \right)_{u \in \mathcal{U}}.$$

STEP 1.(Linearization) In this step we establish that

$$\sqrt{n}(\widehat{\rho}^* - \widehat{\rho}) = Z_{n,P}^* + o_P(1) \quad \text{in } \ell^\infty(\mathcal{U})^{d_\rho}, \tag{75}$$

where $Z_{n,P}^* := (\mathbb{G}_n \xi \psi_u^\rho)_{u \in \mathcal{U}}$. The components $(\sqrt{n}(\widehat{\gamma}_{V_{uj}}^* - \widehat{\gamma}_{V_{uj}}))_{u \in \mathcal{U}}$ of $\sqrt{n}(\widehat{\rho}^* - \widehat{\rho})$ trivially have the linear representation (with no error) for each $j \in \mathcal{J}$. We only need to establish the claim for the empirical process $(\sqrt{n}(\widehat{\alpha}_{V_{uj}}^*(z) - \widehat{\alpha}_{V_{uj}}(z)))_{u \in \mathcal{U}}$ for $z \in \{0, 1\}$ and $j \in \mathcal{J}$, which we do in the steps below.

(a) As in the previous proof, we have that with probability at least $1 - \Delta_n$,

$$\widehat{h}_V \in \mathcal{H}_{V,n} := \{h = (\bar{g}(0, \cdot), \bar{g}(1, \cdot), \bar{m}(0, \cdot), \bar{m}(1, \cdot)) \in \mathcal{G}_V(0) \times \mathcal{G}_V(1) \times \mathcal{M}(0) \times \mathcal{M}(1)\}.$$

(b) We have that

$$\sqrt{n}(\widehat{\alpha}_V(z) - \alpha_V(z)) = \underbrace{\mathbb{G}_n[\xi f_{h_V,V,z}]}_{I_V^*(z)} + \underbrace{(\mathbb{G}_n[\xi f_{h,V,z}] - \mathbb{G}_n[\xi f_{h_V,V,z}])}_{II_V^*(z)} + \underbrace{\sqrt{n}(\mathrm{E}_P[\xi f_{h,V,z} - \xi f_{h_V,h,z}])}_{III_V^*(z)},$$

with h evaluated at $h = \widehat{h}_V$.

(c) Note that $III_V^*(z) = 0$ since ξ is independent of W and has zero mean.

(d) Furthermore, we have that

$$\sup_{V \in \mathcal{V}} \max_{z \in \{0,1\}} |II_V^*(z)| \leqslant \sup_{h \in \mathcal{H}_{V,n}, z \in \{0,1\}, V \in \mathcal{V}} |\mathbb{G}_n[\xi f_{h,V,z}] - \mathbb{G}_n[\xi f_{h_V,V,z}]|.$$

By the previous proof the class of functions, $\mathcal{F}_n = \{f_{h,V,z} - f_{h_V,V,z} : z \in \{0, 1\}, V \in \mathcal{V}, h \in \mathcal{H}_{V,n}\}$, obeys $\log \sup_Q N(\varepsilon, \mathcal{F}_n, \|\cdot\|_{Q,2}) \lesssim (s \log p + s \log(1/\varepsilon)) \vee 0$. By Lemma A.5, multiplication of this class with ξ does not change the entropy bound modulo an absolute constant, namely

$$\log \sup_Q N(\varepsilon, \xi \mathcal{F}_n, \|\cdot\|_{Q,2}) \lesssim (s \log p + s \log(1/\varepsilon)) \vee 0,$$

since the envelope for $\xi\mathcal{F}_n$ is $|\xi|$ times a constant, and $\mathrm{E}[\xi^2] = 1$. We also have that, by standard calculations using that $\mathrm{E}[\exp(|\xi|)] < \infty$,

$$(\mathrm{E}[\max_{i\leqslant n} |\xi|^2])^{1/2} \lesssim \log n.$$

Applying Lemma A.3 and Markov inequality, we have for some constant $K > e$

$$\begin{aligned}
&\sup_{V\in\mathcal{V}} \max_{z\in\{0,1\}} |II_V^*(z)| \leqslant \sup_{f\in\xi\mathcal{F}_n} |\mathbb{G}_n(f)| \\
&= O_{\mathrm{P}}(1)\left(\sqrt{s\sigma_n^2 \log(p\vee K\vee \sigma_n^{-1})} + \frac{s\log n}{\sqrt{n}}\log(p\vee K\vee\sigma_n^{-1})\right) \\
&= O_{\mathrm{P}}(1)\left(\sqrt{s\delta_n^2 n^{-1/2}\log(p\vee n)} + \sqrt{s^2 n^{-1}\log^3(p\vee n)}\right) \\
&= O_{\mathrm{P}}(1)\left(\delta_n\delta_n^{1/4} + \delta_n^{1/2}\right) = O_{\mathrm{P}}(\delta_n^{1/2}),
\end{aligned}$$

for $\sigma_n = \sup_{f\in\xi\mathcal{F}_n} \|f\|_{P,2} = \sup_{f\in\mathcal{F}_n} \|f\|_{P,2}$; where the details of calculations are the same as in the previous proof.

(e) The claim of Step 1 follows by collecting steps (a)-(d).

Step 2 Here by Lemma A.2, we have the conditional convergence:

$$Z_{n,P}^* \rightsquigarrow_B Z_P \quad \text{in } \ell^\infty(\mathcal{U})^{d_\rho}, \text{ uniformly in } P\in\mathcal{P},$$

where $Z_{n,P}^* := (\mathbb{G}_n \xi\psi_u^\rho)_{u\in\mathcal{U}}$; and (b) $Z_P := (\mathbb{G}_P\psi_u^\rho)_{u\in\mathcal{U}}$.

Moreover, the linearization error in Step 1 converges to zero in unconditional probability. It is known that this is stronger than the conditional convergence. The final claim follows by combining the steps. ∎

Appendix C. Proofs for Section 5

C.1. **Proof of Theorem 5.1.** In order to establish the result uniformly in $P\in\mathcal{P}_n$, it suffices to establish the result under the probability measure induced by any sequence $P = P_n \in \mathcal{P}_n$. In the proof we shall suppress the dependency of P on the sample size n.

Throughout the proof we use the notation

$$\begin{aligned}
B(W) &:= \max_{j,k} \sup_{\nu_u\in\Theta_u\times T_u, u\in\mathcal{U}} \left|\partial_{\nu_k}\mathrm{E}[\psi_{uj}(W_u,\nu)\mid Z_u]\right|, \\
\tau_n &:= n^{-1/2}\left(\sqrt{s\log(a_n)} + n^{-1/2}sn^{\frac{1}{q}}\log(a_n)\right).
\end{aligned}$$

Step 1. (A Preliminary Rate Result). In this step we claim that with probability $1-o(1)$,

$$\sup_{u\in\mathcal{U}} \|\widehat{\theta}_u - \theta_u\| \leqslant \bar{C}\tau_n,$$

for some finite constant $\bar{C}$. By definition

$$\|\mathbb{E}_n\psi_u(W_u,\widehat{\theta}_u,\widehat{h}_u(Z_u))\| \leqslant \inf_{\theta\in\Theta_u} \|\mathbb{E}_n\psi_u(W_u,\theta,\widehat{h}_u(Z_u))\| + \epsilon_n \text{ for each } u\in\mathcal{U},$$

which implies via triangle inequality that uniformly in $u \in \mathcal{U}$ with probability $1 - o(1)$

$$\left\| P[\psi_u(W_u, \widehat{\theta}_u, h_u(Z_u))] \right\| \leqslant \epsilon_n + 2I_1 + 2I_2 \lesssim \tau_n, \tag{76}$$

where we define and bound I_1 and I_2 in Step 2 below. The second bound in (76) follows from Step 2 and from the assumption $\epsilon_n = o(n^{-1/2})$. Since by Assumption S1, $2^{-1}(\|J_u(\widehat{\theta}_u - \theta_u)\| \vee c_0)$ does not exceed the left side of (76) and $\inf_{u\in\mathcal{U}} \text{mineig}(J_u)$ is bounded away from zero uniformly in n, we conclude that

$$\sup_{u\in\mathcal{U}} \|\widehat{\theta}_u - \theta_u\| \lesssim \sup_{u\in\mathcal{U}} (\text{mineig}(J_u))^{-1} \tau_n \lesssim \tau_n.$$

Step 2. (Define and bound I_1 and I_2) We claim that with probability $1 - o(1)$:

$$\begin{aligned} I_1 &:= \sup_{\theta\in\Theta_u, u\in\mathcal{U}} \left\| \mathbb{E}_n \psi_u(W_u, \theta, \widehat{h}_u(Z_u)) - \mathbb{E}_n \psi_u(W_u, \theta, h_u(Z_u)) \right\| \lesssim \tau_n, \\ I_2 &:= \sup_{\theta\in\Theta_u, u\in\mathcal{U}} \left\| \mathbb{E}_n \psi_u(W_u, \theta, h_u(Z_u)) - \mathrm{E}_P \psi_u(W_u, \theta, h_u(Z_u)) \right\| \lesssim \tau_n. \end{aligned}$$

To establish the bound, we can bound $I_1 \leqslant I_{1a} + I_{1b}$ and $I_2 \leqslant I_{1a}$, where with probability $1-o(1)$,

$$\begin{aligned} I_{1a} &:= \sup_{\theta\in\Theta_u, u\in\mathcal{U}, h\in\mathcal{H}_{un}\cup\{h_u\}} \left\| \mathbb{E}_n \psi_u(W_u, \theta, h(Z_u)) - \mathrm{E}_P \psi_u(W_u, \theta, h(Z_u)) \right\| \lesssim \tau_n, \\ I_{1b} &:= \sup_{\theta\in\Theta_u, u\in\mathcal{U}, h\in\mathcal{H}_{un}\cup\{h_u\}} \left\| \mathrm{E}_P \psi_u(W_u, \theta, h(Z_u)) - \mathrm{E}_P \psi_u(W_u, \theta, h_u(Z_u)) \right\| \lesssim \tau_n. \end{aligned}$$

The latter bounds hold by the following arguments.

In order to bound I_{1b} we employ Taylor's expansion and triangle inequalities. For $\bar{h}(Z, u, j, \theta)$ denoting a point on a line connecting vectors $h(Z_u)$ and $h_u(Z_u)$,

$$\begin{aligned} I_{1b} &\leqslant \sum_{j=1}^{d_t} \sum_{m=1}^{d_\theta} \sup_{\theta\in\Theta_u, u\in\mathcal{U}, h\in\mathcal{H}_{un}} \left| P\left[\partial_{t_m} P\left[\psi_{uj}(W_u, \theta, \bar{h}(Z, u, j, \theta)) | Z_u \right] (h_m(Z_u) - h_{um}(Z_u)) \right] \right| \\ &\leqslant d_\theta d_t \|B\|_{P,2} \max_{m\in[d_t]} \|h_m - h_{um}\|_{P,2}, \end{aligned}$$

where the last inequality holds by definition of $B(w)$ given earlier and the Holder's inequality. By Assumption S2 $\|B\|_{P,2} \leqslant C$ and by Assumption AS d_θ and d_t are fixed and $\sup_{h\in\mathcal{H}_{un}} \max_m \|h_m - h_{um}\|_{P,2} \lesssim \tau_n$, conclude that $I_{1b} \lesssim \tau_n$.

In order to bound I_{1a} we employ the maximal inequality of Lemma A.3 to the class

$$\mathcal{F}_1 := \{\psi_{uj}(W_u, \theta, h(Z_u)), j \in [d_\theta], u \in \mathcal{U}, \theta \in \Theta_u, h \in \mathcal{H}_{un} \cup \{h_u\}\},$$

equipped with the envelope $F_1 \leqslant F_0$, to conclude that there exists a constant $C > 0$ such that with probability $1 - o(1)$,

$$\begin{aligned} I_{1a} &\leqslant Cn^{-1/2}\left(\|F_0\|_{P,2}\sqrt{s\log(a_n)} + n^{-1/2}s\left\|\max_{i\leqslant n} F_0(W_i)\right\|_{\mathrm{P}_P,2}\log(a_n)\right) \\ &+ C\left[\left(\|F_0\|_{P,2} + n^{-1/2}\left\|\max_{i\leqslant n} F_0(W_i)\right\|_{\mathrm{P}_P,2}\right)\sqrt{\log n} + n^{-1/2}\left\|\max_{i\leqslant n} F_0(W_i)\right\|_{\mathrm{P}_P,2}\log n\right], \\ &\lesssim n^{-1/2}\left(\|F_0\|_{P,2}\sqrt{s\log(a_n)} + n^{-1/2}sn^{\frac{1}{q}}\|F_0\|_{P,q}\log(a_n)\right) \lesssim \tau_n, \end{aligned}$$

using the assumptions on the entropy bound of $\mathcal{F}_1$, and that $\|F_0\|_{P,q} \leqslant C$ and using the condition that $a_n \geqslant n$ and $s \geqslant 1$, and using the elementary inequality $\|\max_{i\leqslant n} F_0(W_i)\|_{\mathrm{P}_P,2} \leqslant n^{\frac{1}{q}}\|F_0\|_{P,q}$.

Step 3. (Linearization) We have that

$$\sqrt{n}\|\mathbb{E}_n\psi_u(W_u,\widehat{\theta}_u,\widehat{h}_u(Z_u))\| \leqslant \inf_{\theta\in\Theta_u}\sqrt{n}\|\mathbb{E}_n\psi_u(W_u,\theta,\widehat{h}_u(Z_u))\| + \epsilon_n n^{1/2}.$$

Application of Taylor's theorem and the triangle inequality gives that for all $u \in \mathcal{U}$

$$\begin{aligned} &\left\|\sqrt{n}\mathbb{E}_n\psi_u(W_u,\theta_u,h_u(Z_u)) + J_u\sqrt{n}(\widehat{\theta}_u - \theta_u) + D_u(\widehat{h}_u - h_u)\right\| \\ &\leqslant \epsilon_n\sqrt{n} + \sup_{u\in\mathcal{U}}\Big(\inf_{\theta\in\Theta_u}\sqrt{n}\|\mathbb{E}_n\psi_u(W_u,\theta,\widehat{h}_u(Z_u))\| + \|II_1(u)\| + \|II_2(u)\|\Big) = o_P(1), \end{aligned}$$

where the terms II_1 and II_2 are defined and bounded below in Step 4; the $o_P(1)$ bound follows from Step 4 and from $\epsilon_n\sqrt{n} = o(1)$ holding by assumption and from Step 5; and

$$D_u(\widehat{h}_u - h_u) := \Big(\sum_{m=1}^{d_t}\sqrt{n}P\big[\underbrace{\partial_{t_m}P[\psi_u(W_u,\theta_u,h_u(Z_u))|Z_u]}_{=0}(\widehat{h}_m(Z_u) - h_{um}(Z_u))\big]\Big)_{j=1}^{d_\theta} = 0,$$

where evanishment occurs because of the orthogonality condition. Conclude using Assumption S1 that

$$\sup_{u\in\mathcal{U}}\left\|J_u^{-1}\sqrt{n}\mathbb{E}_n\psi_u(W_u,\theta_u,h_u(Z_u)) + \sqrt{n}(\widehat{\theta}_u - \theta_u)\right\| \leqslant o_P(1)\sup_{u\in\mathcal{U}}(\mathrm{mineg}(J_u))^{-1} = o_P(1),$$

Furthermore, the empirical process $(\sqrt{n}\mathbb{E}_n J_u^{-1}\psi_u(W_u,\theta_u,h_u(Z_u)), u \in \mathcal{U})$ is equivalent to an empirical process $\mathbb{G}_n$ indexed by

$$\mathcal{F}_0^* := \Big(\bar{\psi}_{uj} : j \in [d_\theta], u \in \mathcal{U}\Big),$$

where $\bar{\psi}_u(W) = -J_u^{-1}\psi_u(W_u,\theta_u,h_u(Z_u))$. We can show that that $u \mapsto J_u^{-1}$ is uniformly Holder on $(\mathcal{U}, d_{\mathcal{U}})$, given the stated assumptions. Indeed, $\|J_u - J_{\bar{u}}\| = \|J_u(J_u^{-1} - J_{\bar{u}}^{-1})J_{\bar{u}}\| \leqslant \sup_{u\in\mathcal{U}}\|J_u^{-1}\|^2\|J_u - J_{\bar{u}}\| \lesssim \|u - \bar{u}\|^{\alpha_2}$. This property and assumptions on $\mathcal{F}_0$ imply by a standard argument that $\mathcal{F}_0^*$, which depends on P, has a uniformly well-behaved uniform covering entropy, namely

$$\sup_{P\in\mathcal{P}=\cup_{n\geqslant n_0}\mathcal{P}_n}\log\sup_Q N(\varepsilon\|CF_0\|_{Q,2},\mathcal{F}_0^*,\|\cdot\|_{Q,2}) \lesssim \log(e/\varepsilon),$$

where CF_0 is the envelope of $\mathcal{F}_0^*$ for some constant C. Application of Lemma A.1 gives the required result.

Step 4. (Define and Bound II_1 and II_2). Let $II_1(u) := (II_{1j}(u))_{j=1}^{d_\theta}$ and $II_2(u) = (II_{2j}(u))_{j=1}^{d_\theta}$, where

$$II_{1j}(u) := \sum_{r,k=1}^{K} \sqrt{n} P\left[\partial_{\nu_k}\partial_{\nu_r} P[\psi_{uj}(W_u, \bar{\nu}_u(Z_u, j))|Z_u]\{\widehat{\nu}_{ur}(Z_u) - \nu_{ur}(Z_u)\}\{\widehat{\nu}_{uk}(Z_u) - \nu_{uk}(Z_u)\}\right],$$

$$II_{2j}(u) := \mathbb{G}_n(\psi_{uj}(W_u, \widehat{\theta}_u, \widehat{h}_u(Z_u)) - \psi_{ju}(W_u, \theta_u, h_u(Z_u))),$$

where $\nu_u(Z_u) := (\nu_{uk}(Z_u))_{k=1}^K := (\theta_u', h_u(Z_u)')'$; $K = d_\theta + d_t$; $\widehat{\nu}_u(Z_u) := (\widehat{\nu}_{uk}(Z_u))_{k=1}^K := (\widehat{\theta}_u', \widehat{h}_u(Z_u)')'$, and $\bar{\nu}_u(Z_u, j)$ is a vector on the line connecting $\nu_u(W)$ and $\widehat{\nu}_u(W)$.

Using Assumption AS, the claim of Step 1, and Holder inequalities, we conclude that

$$\begin{aligned}
\max_{j\in[d_\theta]} \sup_{u\in\mathcal{U}} |II_{1j}(u)| &\leqslant \sum_{r,k=1}^{K} \sqrt{n} P\left[C|\widehat{\nu}_r(Z_u) - \nu_{ur}(Z_u)||\widehat{\nu}_k(Z_u) - \nu_{uk}(Z_u)|\right] \\
&\leqslant C\sqrt{n}K^2 \max_{k\in[K]} \|\widehat{\nu}_k - \nu_{uk}\|_{P,2}^2 \lesssim_P \sqrt{n}\tau_n^2 = o(1).
\end{aligned}$$

We have that with probability $1 - o(1)$,

$$\max_{j\in[d_\theta]} \sup_{u\in\mathcal{U}} |II_{2j}(u)| \lesssim \sup_{f\in\mathcal{F}_2} |\mathbb{G}_n(f)|$$

where, for $\Theta_{un} := \{\theta \in \Theta_u : \|\theta - \theta_u\| \leqslant C\tau_n\}$,

$$\mathcal{F}_2 = \Big(\psi_{uj}(W_u, \theta, h(Z_u)) - \psi_{uj}(W_u, \theta_u, h_u(Z_u)) : j \in [d_\theta], u \in \mathcal{U}, h \in \mathcal{H}_{un}, \theta \in \Theta_{un}\Big).$$

Application of Lemma A.3 gives:

$$\sup_{f\in\mathcal{F}_2} |\mathbb{G}_n(f)| \lesssim \tau_n^{\alpha/2}\sqrt{s\log(a_n)} + n^{-1/2}sn^{\frac{1}{q}}\|F_1\|_{P,q}\log(a_n) + \tau_n, \tag{77}$$

since σ in Lemma A.3 can be chosen so that $\sup_{f\in\mathcal{F}_2}\|f\|_{P,2} \leqslant \sigma \lesssim \tau_n^{\alpha/2}$. Indeed,

$$\begin{aligned}
\sup_{f\in\mathcal{F}_2} \|f\|_{P,2}^2 &\leqslant \sup_{j\in[d_\theta], u\in\mathcal{U}, \nu\in\Theta_{un}\times\mathcal{H}_{un}} \mathrm{E}_P\mathrm{E}_P[(\psi_{uj}(W_u, \nu(Z_u)) - \psi_{uj}(W_u, \nu_u(Z_u)))^2|Z_u], \\
&\leqslant \sup_{u\in\mathcal{U}, \nu\in\Theta_{un}\times\mathcal{H}_{un}} \mathrm{E}_P C\|\nu(Z_u) - \nu_u(Z_u)\|^\alpha, \\
&= \sup_{u\in\mathcal{U}, \nu\in\Theta_{un}\times\mathcal{H}_{un}} C\|\nu - \nu_u\|_{P,\alpha}^\alpha \leqslant \sup_{u\in\mathcal{U}, \nu\in\Theta_{un}\times\mathcal{H}_{un}} C\|\nu - \nu_u\|_{P,2}^\alpha \lesssim \tau_n^\alpha,
\end{aligned}$$

where $\nu_u(Z_u) := (\nu_{uk}(Z_u))_{k=1}^K := (\theta_u', h_u(Z_u)')'$; $K = d_\theta + d_t$; $\nu(Z_u) := (\nu_k(Z_u))_{k=1}^K := (\theta', h(Z_u)')'$; where the first inequality follows by the law of iterated expectations; the second inequality follows by Assumption AS; and the last inequality follows from $\alpha \in [1, 2]$ by Assumption AS and the monotonicity of the norm $\|\cdot\|_{P,q}$ in $q \in [1, \infty]$.

Conclude that using the growth conditions of Assumption AS:

$$\max_{j\in[d_\theta]} \sup_{u\in\mathcal{U}} |II_{2j}(u)| \lesssim_P \tau_n^{\alpha/2}\sqrt{s\log(a_n)} + n^{-1/2}sn^{\frac{1}{q}}\log(a_n) = o(1). \tag{78}$$

Step 5. In this step we show that $\inf_{\theta\in\Theta_u}\sqrt{n}\|\mathbb{E}_n\psi_u(W_u,\theta,\widehat{h}_u(Z_u))\| = o_P(1)$. We have that with probability $1-o(1)$

$$\inf_{\theta\in\Theta_u}\sqrt{n}\|\mathbb{E}_n\psi_u(W_u,\theta,\widehat{h}_u(Z_u))\| \leqslant \sqrt{n}\|\mathbb{E}_n\psi_u(W_u,\bar{\theta}_u,\widehat{h}_u(Z_u))\|$$

where $\bar{\theta}_u = \theta_u - J_u^{-1}\mathbb{E}_n\psi_u(W_u,\theta_u,h_u(Z_u))$, since $\bar{\theta}_u\in\Theta_u$ for all $u\in\mathcal{U}$ with probability $1-o(1)$, and in fact $\sup_{u\in\mathcal{U}}\|\bar{\theta}_u-\theta_u\| \lesssim_P 1/\sqrt{n}$ by the last paragraph of Step 3.

Then, arguing similarly to Step 3 and 4, we can conclude that uniformly in $u\in\mathcal{U}$:

$$\sqrt{n}\|\mathbb{E}_n\psi_u(W_u,\bar{\theta}_u,\widehat{h}_u(Z_u))\| \leqslant \sqrt{n}\|\mathbb{E}_n\psi_u(W_u,\theta_u,h_u(Z_u)) + J_u(\bar{\theta}_u-\theta_u) + D_u(\widehat{h}_u-h_u)\| + o_P(1)$$

where the first term on the right side vanishes by definition of $\bar{\theta}_u$ and by $D_u(\widehat{h}_u-h_u)=0$. ∎

C.2. **Proof of Theorem 5.2.** Step 0. In the proof $a\lesssim b$ means that $a\leqslant Ab$, where the constant A depends on the constants in assumptions only, but not on n once $n\geqslant n_0$, and not on $P\in\mathcal{P}_n$. In Step 1, we consider a sequence P_n in $\mathcal{P}_n$, but for simplicity, we write $P=P_n$ throughout the proof, suppressing the index n. Since the argument is asymptotic, we can just assume that $n\geqslant n_0$ in what follows.

Let $\mathbb{P}_n$ denote the measure that puts mass n^{-1} on points (ξ_i,W_i) for $i=1,...,n$. Let $\mathbb{E}_n$ denote the expectation with respect to this measure, so that $\mathbb{E}_n f = n^{-1}\sum_{i=1}^n f(\xi_i,W_i)$.

Recall that we define the bootstrap draw as:

$$\sqrt{n}(\widehat{\theta}^*-\widehat{\theta}) = \left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i\widehat{\psi}_u(W_i)\right)_{u\in\mathcal{U}} = \left(\mathbb{G}_n\xi\widehat{\psi}_u\right)_{u\in\mathcal{U}},$$

where

$$\bar{\psi}_u(W) = -J_u^{-1}\psi_u(W_u,\theta_u,h_u(Z_u)),\quad \widehat{\psi}_u(W) = -\widehat{J}_u^{-1}\psi_u(W_u,\widehat{\theta}_u,\widehat{h}_u(Z_u)).$$

Step 1.(Linearization) In this step we establish that

$$\sqrt{n}(\widehat{\theta}^*-\widehat{\theta}) = Z^*_{n,P} + o_P(1)\quad \text{in } \ell^\infty(\mathcal{U})^{d_\theta}, \tag{79}$$

where $Z^*_{n,P} := (\mathbb{G}_n\xi\psi_u^\rho)_{u\in\mathcal{U}}$.

(a) We have that $h_u\in\mathcal{H}_{un}$ with probability $1-\delta_n$.

(b) We have that

$$\sqrt{n}(\widehat{\theta}^*_u-\widehat{\theta}_u) = \underbrace{\mathbb{G}_n[\xi\bar{\psi}_u(W)]}_{I^*(u)} + \underbrace{(\mathbb{G}_n[\xi\bar{\psi}_u(W)]-\mathbb{G}_n[\xi\widehat{\psi}_u(W)])}_{II^*(u)} + \underbrace{\sqrt{n}(P[\widehat{\psi}_u(W)-\xi\bar{\psi}_u(W)])}_{III^*(u)} + R_n(u),$$

where notation $Pf(\xi,W)$ means $\int f(\xi,w)d\mathrm{P}(\xi)dP(w)$, and where $R_n(u) := \sqrt{n}(\widehat{\theta}-\theta_u) - \mathbb{G}_n\bar{\psi}_u$ obeys by the preceding theorem:

$$\sup_{u\in\mathcal{U}}\|R_n(u)\| = o_P(1).$$

(c) Note that $III^*(u)=0$ since ξ is independent of W and has zero mean.

(d) Furthermore, we have that with probability at least $1-\delta_n$:

$$\sup_{u\in\mathcal{U}}|II^*(u)| \leqslant \sup_{f\in\mathcal{F}_3}|\mathbb{G}_n[\xi f]|.$$

where

$$\mathcal{F}_3 = \{\bar{J}_u^{-1}\psi_u(W_u, \bar{\theta}_u, \bar{h}_u(Z_u)) - J_u^{-1}\psi_u(W_u, \theta_u, h_u(Z_u)), u \in \mathcal{U}, \bar{\theta}_u \in \Theta_{un}, \bar{h}_u \in \mathcal{H}_{un}, \bar{J} \in \mathcal{J}_{un}\},$$

By the standard reasoning, under Assumption AS and additional conditions stated in the theorem, we can conclude that this class obeys

$$\log \sup_Q N(\varepsilon \|F_3\|_{Q,2}, \mathcal{F}_3, \|\cdot\|_{Q,2}) \lesssim (s \log a_n + s \log(a_n/\varepsilon)),$$

with the envelope $F_3 \lesssim F_0$. By Lemma A.5, multiplication of this class with ξ does not change the entropy bound modulo an absolute constant, namely

$$\log \sup_Q N(\varepsilon \||\xi| F_3\|_{Q,2}, \xi\mathcal{F}_3, \|\cdot\|_{Q,2}) \lesssim (s \log a_n + s \log(a_n/\varepsilon)).$$

We also have that $(\mathrm{E}[\max_{i \leqslant n} |\xi|^2])^{1/2} \lesssim \log n$ by standard calculations, using that $\mathrm{E}[\exp(|\xi|)] < \infty$. Applying Lemma A.3 and Markov inequality, we have that

$$\sup_{f \in \xi\mathcal{F}_3} |\mathbb{G}_n(f)| = O_P(1)\left(\sqrt{s\sigma_n^2 \log(a_n)} + \frac{sn^{1/q}\|F_0\|_{P,q} \log n}{\sqrt{n}} \log(a_n)\right)$$
$$= O_P(1)\left(\tau_n^{\alpha/2}\sqrt{s \log(a_n)} + \frac{sn^{1/q}\|F_0\|_{P,q} \log n}{\sqrt{n}} \log(a_n)\right) = o_P(1),$$

for $\sigma_n = \sup_{f \in \xi\mathcal{F}_3} \|f\|_{P,2} = \sup_{f \in \mathcal{F}_3} \|f\|_{P,2} \lesssim \tau_n^{\alpha/2}$; where the details of calculations are the same as in the previous proof and therefore omitted in this version.

(e) The claim of Step 1 follows by collecting steps (a)-(d).

Step 2. Here by Lemma A.2, we have the conditional convergence of the bootstrap law:

$$Z_{n,P}^* \rightsquigarrow_B Z_P \quad \text{in } \ell^\infty(\mathcal{U})^{d_\rho}, \text{ uniformly in } P \in \mathcal{P},$$

where $Z_{n,P}^* := (\mathbb{G}_n \xi \bar{\psi}_u)_{u \in \mathcal{U}}$; and $Z_P := (\mathbb{G}_P \bar{\psi}_u)_{u \in \mathcal{U}}$. Moreover, the linearization error R_n in Step 1 converges to zero in probability unconditionally on the data. It is known that this is stronger than the convergence in probability conditional on the data. The final claim follows by combining the steps. ∎

C.3. **Proof of Theorem 5.3.** We have that under any sequence $P_n \in \mathcal{P}_n$

$$Z_{P_n,n} \rightsquigarrow Z_{P_n},$$

which means that

$$\sup_{h \in BL_1(\ell^\infty(\mathcal{U})^{d_\theta}, \mathbb{R})} |\mathrm{E}_{P_n}^* h(Z_{P_n,n}) - \mathrm{E}_{P_n} h(Z_{P_n})| \to_{n \in \mathbb{Z}} 0,$$

where $\mathbb{Z} = \{1, 2, ..., \}$. By the uniform in $P \in \mathcal{P}$ tightness of Z_P and compactness of $\mathbb{D}_1$ we can split $\mathbb{Z} = \{1, 2, ..., \}$ into a collection of subsequences $\{\mathbb{Z}'\}$, along each of which

$$Z_{P_n} \rightsquigarrow Z', \quad \theta_{P_n}^0 \to \theta^{0\prime},$$

where the former means that

$$\sup_{h\in BL_1(\ell^\infty(\mathcal{U})^{d_\theta},\mathbb{R})} |\mathrm{E}_{P_n} h(Z_{P_n}) - \mathrm{E}h(Z')| \to_{n\in\mathbb{Z}'} 0,$$

where Z' is a tight Gaussian process, which depends on a subsequence $\mathbb{Z}'$ with paths that are uniformly continuous on $(\mathcal{U}, d_\mathcal{U})$, with covariance function equal to the limit of the covariance function Z_{P_n} along the subsequence, which may depend on the subsequence $\mathbb{Z}'$, and $\theta^{0\prime}$ is some value in $\mathbb{D}_1$ that also may depend on the subsequence $\mathbb{Z}'$. We can conclude by the triangle inequality that along that same subsequence,

$$Z_{P_n,n} \rightsquigarrow Z' \text{ , that is, } \sup_{h\in BL_1(\ell^\infty(\mathcal{U})^{d_\theta},\mathbb{R})} |\mathrm{E}_{P_n} h(Z_{P_n,n}) - \mathrm{E}h(Z')| \to_{n\in\mathbb{Z}'} 0.$$

For each such subsequence, application of the functional delta method for subsequences, Lemma A.7, yields $\sqrt{n}(\widehat{\Delta} - \Delta) \rightsquigarrow \phi'_{\theta^{0\prime}}(Z')$ and, furthermore, by the continuity of the map $(\vartheta, g) \mapsto \phi'_\vartheta(g)$ on the domain $\mathbb{D}_1 \times \mathbb{D}_0$ and the Extended Continuous Mapping Theorem, $\phi'_{\theta^0_{P_n}}(Z_{P_n}) \rightsquigarrow \phi'_{\theta^{0\prime}}(Z')$, which gives that, via the triangle inequality, $\sqrt{n}(\widehat{\Delta} - \Delta) \rightsquigarrow \phi'_{\theta^0_{P_n}}(Z_{P_n})$, that is,

$$\sup_{h\in BL_1(\ell^\infty(\mathcal{Q}),\mathbb{R})} |\mathrm{E}_P h(\sqrt{n}(\widehat{\Delta} - \Delta)) - \mathrm{E}_P h(\phi'_{\theta^0_{P_n}}(Z_{P_n}))| \to_{n\in\mathbb{Z}'} 0.$$

Since the argument above works for all subsequences as defined above, we conclude that

$$\sup_{P\in\mathcal{P}_n} \sup_{h\in BL_1(\ell^\infty(\mathcal{Q}),\mathbb{R})} |\mathrm{E}_P h(\sqrt{n}(\widehat{\Delta} - \Delta)) - \mathrm{E}_P h(\phi'_{\theta^0_P}(Z_{P_n}))| \to 0,$$

or, using more compact notation,

$$\sqrt{n}(\widehat{\Delta} - \Delta) \rightsquigarrow \phi'_{\theta^0}(Z_P), \quad \text{in } \ell^\infty(\mathcal{Q}), \text{ uniformly in } P \in \mathcal{P}_n.$$

The argument for bootstrap follows similarly, except now we apply Lemma A.8. ∎

Appendix D. Proofs for Section 6

Proof of Theorem 6.1. Condition WL is implied by Assumption 6. Recall that the Algorithm sets $\gamma \in [1/n, 1/\log n]$ so that $\gamma = o(1)$ and $\log(np/\gamma) \leqslant 3\log(n \vee p)$. We will establish that the events $\{\lambda/n \geqslant \sqrt{c}\sup_{u\in\mathcal{U}} \|\widehat{\Psi}_{u0}^{-1}\mathbb{E}_n[\zeta_u f(X)]\|_\infty\}$, $\{\ell\widehat{\Psi}_{u0} \leqslant \widehat{\Psi}_u \leqslant L\widehat{\Psi}_{u0}\}$, for $\ell > 1/\sqrt{c}$ and L uniformly bounded, hold with probability $1 - o(1)$ for all iterations K for n sufficiently large. Since $c_r \geqslant \sup_{u\in\mathcal{U}} \|r_u\|_{\mathbb{P}_n,2}$ is assumed, we will be able to invoke the conclusions of Lemmas F.2, F.3 and F.4.

By Assumption 6 it follows that $\underline{c} \leqslant \mathrm{E}[|f_j(X)\zeta_u|^2] \leqslant \mathrm{E}[|f_j(X)Y_u|^2] \leqslant C$ uniformly over $u \in \mathcal{U}$ and $j = 1, \ldots, p$. Moreover, Assumption 6 yields

$$\sup_{u\in\mathcal{U}} \max_{j\leqslant p} |(\mathbb{E}_n - \bar{\mathrm{E}})[|f_j(X)Y_u|^2]| \leqslant \delta_n \quad \text{and} \quad \sup_{u\in\mathcal{U}} \max_{j\leqslant p} |(\mathbb{E}_n - \bar{\mathrm{E}})[|f_j(X)\zeta_u|^2]| \leqslant \delta_n$$

with probability $1 - \Delta_n$. In turn this shows that $\mathbb{E}_n[|f_j(X)Y_u|^2] \leqslant C + \delta_n \leqslant L^2\mathbb{E}_n[|f_j(X)\zeta_u|^2]$ for some uniformly bounded L with probability $1 - \Delta_n$ so that $\widehat{\Psi}_u \leqslant L\widehat{\Psi}_{u0}$. With the same probability we have $\mathbb{E}_n[|f_j(X)Y_u|^2] \geqslant \mathbb{E}_n[|f_j(X)\zeta_u|^2](1 - 2\delta_n/c)$ for $n \geqslant n_0 := \min\{n : \delta_n \leqslant \underline{c}/2\}$

sufficiently large so that $\ell\widehat{\Psi}_{u0} \leqslant \widehat{\Psi}_u$ for $\ell = (1 - 2\delta_n/\underline{c}) \to_P 1$. This shows $\ell\widehat{\Psi}_{u0} \leqslant \widehat{\Psi}_u \leqslant L\widehat{\Psi}_{u0}$ with probability $1 - \Delta_n$ where $\ell \to_P 1$ and $L \leqslant C$ uniformly in $u \in \mathcal{U}$ for $n \geqslant n_0$. Moreover $\tilde{\mathbf{c}}$ is uniformly bounded which implies that $\kappa_{\tilde{\mathbf{c}}}$ is bounded away from zero by the condition on sparse eigenvalues of order $s\ell_n$.

By Lemma F.1, the event $\lambda/n \geqslant \sqrt{c}\sup_{u\in\mathcal{U}} \|\widehat{\Psi}_{u0}^{-1}\mathbb{E}_n[f(X)\zeta_u]\|_\infty$ occurs with probability $1 - o(1)$. Finally, by assumption $c_r^2 \leqslant Cs\log(p \vee n)/n$ with probability $1 - \Delta_n$. By Lemma F.2 we have

$$\sup_{u\in\mathcal{U}} \|f(X)'(\widehat{\theta}_u - \theta_u)\|_{\mathbb{P}_n,2} \leqslant C\sqrt{\frac{s\log(p \vee n)}{n}} \quad \text{and} \quad \sup_{u\in\mathcal{U}} \|\widehat{\theta}_u - \theta_u\|_1 \leqslant C\sqrt{\frac{s^2\log(p \vee n)}{n}}$$

In the application of Lemma F.3, since $s\log(p \vee n) \leqslant \delta_n n$ and the conditions on the sparse eigenvalues, we have that $\min_{m\in\mathcal{M}} \phi_{\max}(m)$ is uniformly bounded. Thus, with probability $1 - o(1)$, by Lemma F.3 we have

$$\sup_{u\in\mathcal{U}} \widehat{s}_u \leqslant C\left[\frac{nc_r}{\lambda} + \sqrt{s}\right]^2 \leqslant Cs.$$

Therefore by Lemma F.4 the Post-Lasso estimators $(\widetilde{\theta}_u)_{u\in\mathcal{U}}$ satisfy with probability $1 - o(1)$

$$\sup_{u\in\mathcal{U}} \|f(X)'(\widetilde{\theta}_u - \theta_u)\|_{\mathbb{P}_n,2} \leqslant \bar{C}\sqrt{\frac{s\log(p \vee n)}{n}} \quad \text{and} \quad \sup_{u\in\mathcal{U}} \|\widetilde{\theta}_u - \theta_u\|_1 \leqslant \bar{C}\sqrt{\frac{s^2\log(p \vee n)}{n}}$$

for some $\bar{C}$ independent of n.

In the kth iteration, the penalty loadings are constructed based on $(\widetilde{\theta}_u^{(k)})_{u\in\mathcal{U}}$, defined as $\widehat{\Psi}_{ujj} = \mathbb{E}_n[|f_j(X)\{Y_u - f(X)'\widetilde{\theta}_u^{(k)}\}|^2]\}^{1/2}$ for $j = 1, \ldots, p$. We assume $(\widetilde{\theta}_u^{(k)})_{u\in\mathcal{U}}$ satisfy the rates above. Then with probability $1 - o(1)$ we have

$$\begin{aligned}|\widehat{\Psi}_{ujj} - \widehat{\Psi}_{u0jj}| &\leqslant \{\mathbb{E}_n[|f_j(X)\{f(X)'(\widetilde{\theta}_u - \theta_u)\}|^2]\}^{1/2} + \{\mathbb{E}_n[|f_j(X)r_u|^2]\}^{1/2} \\ &\leqslant K_n\|f(X)'(\widetilde{\theta}_u - \theta_u)\|_{\mathbb{P}_n,2} + K_n\|r_u\|_{\mathbb{P}_n,2} \leqslant \bar{C}K_n\sqrt{\tfrac{s\log(p\vee n)}{n}} \leqslant \bar{C}\delta_n.\end{aligned}$$

This establish the event of the penalty loadings for the $(k+1)$th iteration which leads to the stated rates of convergence and sparsity bound. ∎

Proof of Theorem 6.2. The proof is similar to the proof of Theorem 6.1. Condition WL is implied by Assumption 7. Recall that the Algorithm sets $\gamma \in [1/n, 1/\log n]$ so that $\gamma = o(1)$ and $\log(np/\gamma) \leqslant 3\log(n \vee p)$. Moreover, by Assumption 7 $w_{ui} = \mathrm{E}[Y_{ui} \mid X_i](1 - \mathrm{E}[Y_{ui} \mid X_i]) \leqslant 1$ is bounded away from zero. Since $\underline{c}(1 - \underline{c}) \leqslant w_{ui} \leqslant 1$ we have $\ell\widehat{\Psi}_{u0} \leqslant \widehat{\Psi}_u \leqslant L\widehat{\Psi}_{u0}$ for some uniformly bounded L and $\ell = 1$. Moreover, since $|r_u(X)| \leqslant \delta_n$ a.s. uniformly on $u \in \mathcal{U}$, we have $|\tilde{r}_u(X)| \leqslant |r_u(X)|/\{\underline{c}(1 - \underline{c}) - \delta_n\} \leqslant \tilde{C}|r_u(X)|$. Thus $\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2} \leqslant \tilde{C}\|r_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}$.

By Assumption 7 it follows that $\underline{c}(1 - \underline{c})c \leqslant \mathrm{E}[|f_j(X)\zeta_u|^2] \leqslant \mathrm{E}[|f_j(X)|^2] \leqslant C$ uniformly over $u \in \mathcal{U}$ and $j = 1, \ldots, p$. Moreover, Assumption 7 yields

$$\sup_{u\in\mathcal{U}} \max_{j\leqslant p} |(\mathbb{E}_n - \bar{\mathrm{E}})[|f_j(X)\zeta_u|^2]| \leqslant \delta_n$$

with probability $1-\Delta_n$. Note that by Lemma A.4 we have with probability $1-\delta_n^{1/2}$

$$\begin{aligned}\max_{j\leqslant p}|(\mathbb{E}_n-\bar{\mathrm{E}})[|f_j(X)|^2]| &\leqslant \tfrac{C}{\delta_n^{1/4}}\sqrt{\tfrac{\log(p\vee n)}{n}}\max_{j\leqslant p}\sqrt{(\mathbb{E}_n+\bar{\mathrm{E}})[|f_j(X)|^4]}\\ &\leqslant \tfrac{C}{\delta_n^{1/4}}\sqrt{\tfrac{\log(p\vee n)}{n}}K_n\max_{j\leqslant p}\sqrt{(\mathbb{E}_n+\bar{\mathrm{E}})[|f_j(X)|^2]}\\ &\leqslant \delta_n^{1/4}\max_{j\leqslant p}\sqrt{\mathbb{E}_n[|f_j(X)|^2]}+\delta_n^{1/4}\sqrt{C}\end{aligned}$$

under the condition $K_n^2 s^2\log^2(p\vee n)\leqslant\delta_n n$. Since for positive numbers $a\leqslant b+\sqrt{a}$ implies $a\leqslant(b+1)^2$, we have $\max_{j\leqslant p}\mathbb{E}_n[|f_j(X)|^2]\leqslant C'$ with probability $1-o(1)$. Thus $\tilde{\mathbf{c}}$ is uniformly bounded which implies that $\kappa_{\tilde{\mathbf{c}}}$ is bounded away from zero by the condition on sparse eigenvalues of order $s\ell_n$.

By Lemma F.1, the event $\lambda/n\geqslant\sqrt{c}\sup_{u\in\mathcal{U}}\|\widehat{\Psi}_{u0}^{-1}\mathbb{E}_n[f(X)\zeta_u]\|_\infty$ occurs with probability $1-o(1)$. Finally, since $w_{ui}\geqslant\underline{c}(1-\underline{c})$, by assumption $\|r_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\leqslant c_r/\underline{c}(1-\underline{c})\leqslant C\sqrt{s\log(p\vee n)/n}$ with probability $1-\Delta_n$.

We have that $q_{A_u}\geqslant b\kappa_{\mathbf{c}}/\sqrt{s}K_n$. Under the condition $K_n^2s^2\log^2(p\vee n)\leqslant\delta_n n$, the side condition in Lemma F.5 holds with probability $1-o(1)$, and the lemma yields

$$\sup_{u\in\mathcal{U}}\|f(X)'(\widehat{\theta}_u-\theta_u)\|_{\mathbb{P}_n,2}\leqslant C\sqrt{\frac{s\log(p\vee n)}{n}}\quad\text{and}\quad\sup_{u\in\mathcal{U}}\|\widehat{\theta}_u-\theta_u\|_1\leqslant C\sqrt{\frac{s^2\log(p\vee n)}{n}}$$

In turn, under our conditions on sparse eigenvalues and $K_n^2s^2\log^2(p\vee n)\leqslant\delta_n n$, with probability $1-o(1)$ Lemma F.6 implies

$$\sup_{u\in\mathcal{U}}\widehat{s}_u\leqslant C\left[\frac{nc_r}{\lambda}+\sqrt{s}\right]^2\leqslant Cs$$

since $\min_{m\in\mathcal{M}}\phi_{\max}(m)$ is uniformly bounded. The rate of convergence for $\widetilde{\theta}_u$ is given by Lemma F.7, namely with probability $1-o(1)$

$$\sup_{u\in\mathcal{U}}\|f(X)'(\widetilde{\theta}_u-\theta_u)\|_{\mathbb{P}_n,2}\leqslant\bar{C}\sqrt{\frac{s\log(p\vee n)}{n}}\quad\text{and}\quad\sup_{u\in\mathcal{U}}\|\widetilde{\theta}_u-\theta_u\|_1\leqslant\bar{C}\sqrt{\frac{s^2\log(p\vee n)}{n}}$$

for some $\bar{C}$ independent of n, since $M_u(\tilde{\theta}_u)-M_u(\theta_u)\leqslant Cs\log(p\vee n)/n$ and $\|\mathbb{E}_n[f(X)\zeta_u]\|_\infty\leqslant C\sqrt{\log(p\vee n)/n}$ by Lemma F.1 and $\sup_{u\in\mathcal{U}}\|\widehat{\Psi}_{u0}\|_\infty\leqslant C$ with probability $1-o(1)$.

In the kth iteration, the penalty loadings are constructed based on $(\widetilde{\theta}_u^{(k)})_{u\in\mathcal{U}}$, defined as $\widehat{\Psi}_{ujj}=\mathbb{E}_n[|f_j(X)\{Y_u-\Lambda(f(X)'\widetilde{\theta}_u^{(k)})\}|^2]\}^{1/2}$ for $j=1,\ldots,p$. We assume $(\widetilde{\theta}_u^{(k)})_{u\in\mathcal{U}}$ satisfy the rates above. Then

$$\begin{aligned}|\widehat{\Psi}_{ujj}-\widehat{\Psi}_{u0jj}| &\leqslant\{\mathbb{E}_n[|f_j(X)\{f(X)'(\widetilde{\theta}_u-\theta_u)\}|^2]\}^{1/2}+\{\mathbb{E}_n[|f_j(X)r_u|^2]\}^{1/2}\\ &\leqslant K_n\|f(X)'(\widetilde{\theta}_u-\theta_u)\|_{\mathbb{P}_n,2}+K_n\|r_u\|_{\mathbb{P}_n,2}\lesssim_P K_n\sqrt{\tfrac{s\log(p\vee n)}{n}}\lesssim\delta_n\end{aligned}$$

and we have $\ell\widehat{\Psi}_{u0}\leqslant\widehat{\Psi}_u\leqslant L\widehat{\Psi}_{u0}$ for $\ell\to_P 1$ and L uniformly bounded with probability $1-o(1)$. Then the same proof for the initial penalty loading choice applies to the iterate $(k+1)$. ∎

Appendix E. Finite Sample Results of a Continuum of Lasso and Post-Lasso Estimators for Functional Responses

This section uses notation $\bar{\mathrm{E}}[\cdot]=\frac{1}{n}\sum_{i=1}^n\mathrm{E}[\cdot]$, because it allows for i.n.i.d. data.

Appendix F. Assumptions and Results

We consider the following high level conditions which are implied by the primitive Assumptions 6 and 7. For each $n \geqslant 1$, our data consist of fixed regressors $(X_i)_{i=1}^n$ and independent $(W_i)_{i=1}^n$ stochastic process $W_i = (\zeta_{ui} := Y_{ui} - \mathrm{E}[Y_{ui} \mid X_i])_{u\in\mathcal{U}}$ defined on the probability space $(S, \mathcal{S}, P)$ such that model (68) holds with $\mathcal{U} \subset [0,1]^\iota$.

Condition WL. *Let ι be fixed, normalize $\mathbb{E}_n[f_j(X_i)^2] = 1$, $j = 1, \ldots, p$, and suppose that:*
(i) for $s \geqslant 1$ we have $\sup_{u\in\mathcal{U}} \|\theta_u\|_0 \leqslant s$, $\Phi^{-1}(1-\gamma/\{2pn^\iota\}) \leqslant \delta_n n^{1/6}$, $N(\epsilon, \mathcal{U}, d_{\mathcal{U}}) \leqslant (1/\epsilon)^\iota$.
(ii) Uniformly over $u \in \mathcal{U}$, $0 < \underline{c} \leqslant \mathrm{E}[\zeta_{ui}^2 \mid X_i] \leqslant \bar{c} < \infty$, a.s., $\max_{j\leqslant p} \frac{\{\bar{\mathrm{E}}[|f_j(X)\zeta_u|^3]\}^{1/3}}{\{\bar{\mathrm{E}}[|f_j(X)\zeta_u|^2]\}^{1/2}} \leqslant C$.
(iii) with probability $1-\Delta_n$ we have $\max_{i\leqslant n} \|f(X_i)\|_\infty \leqslant K_n$, and for a fixed parameter $\nu \in (0,1]$, $K_n \log(p \vee n) \leqslant \delta_n n^{\{\nu\wedge\frac{1}{2}\}}$, and

$$\sup_{u\in\mathcal{U}} \max_{j\leqslant p} |(\mathbb{E}_n - \bar{\mathrm{E}})[f_j(X)^2\zeta_u^2]| \leqslant \delta_n, \quad \sup_{u\in\mathcal{U}} \mathbb{E}_n[r_u^2] \leqslant c_r^2,$$
$$\sup_{u,u'\in\mathcal{U}, d_{\mathcal{U}}\leqslant\epsilon} \{(\mathbb{E}_n + \bar{\mathrm{E}})[(\zeta_u - \zeta_{u'})^2]\}^{1/2} \leqslant C\{\epsilon^\nu + n^{-1/2}\}.$$

The following important technical lemma formally justify the choice of penalty level λ. It is based on self-normalized moderate deviation theory.

Lemma F.1 (Choice of λ)**.** *Suppose Condition WL holds, let $c' > c > 1$, $\gamma \in [1/n, 1/\log n]$, and $\lambda = c'\sqrt{n}\Phi^{-1}(1-\gamma/\{2pn^\iota\})$. Then for $n \geqslant n_0$ large enough depending only on Condition WL,*

$$P\left(\lambda/n \geqslant c \sup_{u\in\mathcal{U}} \|\widehat{\Psi}_{u0}^{-1}\mathbb{E}_n[f(X)\zeta_u]\|_\infty\right) \geqslant 1 - \gamma - o(1).$$

F.1. **Finite Sample Results: Linear Case.** For the model described in (68) with $\Lambda(t) = t$ and $M(y,t) = \frac{1}{2}(y-t)^2$ we will study finite sample properties of the associated Lasso and Post-Lasso estimators of $(\theta_u)_{u\in\mathcal{U}}$.

The analysis relies on restricted eigenvalues

$$\kappa_{\mathbf{c}} = \inf_{u\in\mathcal{U}} \min_{\|\delta_{T_u^c}\|_1 \leqslant \mathbf{c}\|\delta_{T_u}\|_1} \frac{\|f(X)'\delta\|_{\mathbb{P}_n,2}}{\|\delta_{T_u}\|},$$

maximal and minimum sparse eigenvalues

$$\phi_{\min}(m) = \min_{1\leqslant\|\delta\|_0\leqslant m} \frac{\|f(X)'\delta\|_{\mathbb{P}_n,2}}{\|\delta\|} \quad \text{and} \quad \phi_{\max}(m) = \max_{1\leqslant\|\delta\|_0\leqslant m} \frac{\|f(X)'\delta\|_{\mathbb{P}_n,2}}{\|\delta\|}$$

and also the "ideal loadings" $\widehat{\Psi}_{u0jj} = \{\mathbb{E}_n[f_j(X)^2\zeta_u^2]\}^{1/2}$.

Next we present technical results on the performance of the estimators generated by Lasso that are used in the proof of Theorem 6.1.

Lemma F.2 (Rates of Convergence for Lasso). *The events $c_r \geqslant \sup_{u\in\mathcal{U}} \|r_u\|_{\mathbb{P}_n,2}$, $\ell\widehat{\Psi}_{u0} \leqslant \widehat{\Psi}_u \leqslant L\widehat{\Psi}_{u0}$, $u \in \mathcal{U}$, and $\lambda/n \geqslant c\sup_{u\in\mathcal{U}} \|\widehat{\Psi}_{u0}^{-1}\mathbb{E}_n[f(X)\zeta_u]\|_\infty$, for $c > 1/\ell$, imply that*

$$\sup_{u\in\mathcal{U}} \|f(X)'(\widehat{\theta}_u - \theta_u)\|_{\mathbb{P}_n,2} \leqslant 2c_r + \frac{2\lambda\sqrt{s}}{n\kappa_{\tilde{\mathbf{c}}}}\left(L + \frac{1}{c}\right)\|\widehat{\Psi}_{u0}\|_\infty$$
$$\sup_{u\in\mathcal{U}} \|\widehat{\theta}_u - \theta_u\|_1 \leqslant 2\frac{\sqrt{s}c_r}{\widehat{\kappa}_{2\tilde{\mathbf{c}}}} + \frac{2\lambda s}{n\kappa_{\tilde{\mathbf{c}}}\kappa_{2\tilde{\mathbf{c}}}}\left(L + \frac{1}{c}\right)\|\widehat{\Psi}_{u0}\|_\infty + \left(1 + \frac{1}{2\tilde{\mathbf{c}}}\right)\frac{c\sup_{u\in\mathcal{U}}\|\widehat{\Psi}_{u0}^{-1}\|_\infty}{\ell c - 1}\frac{n}{\lambda}c_r^2$$

where $\tilde{\mathbf{c}} = \sup_{u\in\mathcal{U}} \|\widehat{\Psi}_{u0}^{-1}\|_\infty\|\widehat{\Psi}_{u0}\|_\infty(Lc+1)/(\ell c - 1)$

The following lemma summarizes sparsity properties of $(\widehat{\theta}_u)_{u\in\mathcal{U}}$.

Lemma F.3 (Sparsity bound for Estimated Lasso under data-driven penalty). *Consider the Lasso estimator $\widehat{\theta}_u$, its support $\widehat{T}_u = \mathrm{supp}(\widehat{\theta}_u)$, and let $\widehat{s}_u = |\widehat{T}_u|$. Assume that $c_r \geqslant \sup_{u\in\mathcal{U}} \|r_u\|_{\mathbb{P}_n,2}$, $\lambda/n \geqslant c\sup_{u\in\mathcal{U}} \|\widehat{\Psi}_{u0}^{-1}\mathbb{E}_n[f(X)\zeta_u]\|_\infty$ and $\ell\widehat{\Psi}_{u0} \leqslant \widehat{\Psi}_u \leqslant L\widehat{\Psi}_{u0}$ for all $u \in \mathcal{U}$, with $L \geqslant 1 \geqslant \ell > 1/c$. Then, for $c_0 = (Lc+1)/(\ell c - 1)$ and $\tilde{\mathbf{c}} = (Lc+1)/(\ell c - 1)\sup_{u\in\mathcal{U}} \|\widehat{\Psi}_{u0}\|_\infty\|\widehat{\Psi}_{u0}^{-1}\|_\infty$ we have*

$$\sup_{u\in\mathcal{U}} \widehat{s}_u \leqslant 16\left(\min_{m\in\mathcal{M}} \phi_{\max}(m)\right)c_0^2 \sup_{u\in\mathcal{U}}\left[\frac{nc_r}{\lambda} + \frac{\sqrt{s}}{\kappa_{\tilde{\mathbf{c}}}}\|\widehat{\Psi}_{u0}\|_\infty\right]^2 \|\widehat{\Psi}_{u0}^{-1}\|_\infty^2$$

where $\mathcal{M} = \left\{m \in \mathbb{N} : m > \sup_{u\in\mathcal{U}} 32\|\widehat{\Psi}_{u0}^{-1}\|_\infty^2 c_0^2\left[\frac{nc_r}{\lambda} + \frac{\sqrt{s}}{\kappa_{\tilde{\mathbf{c}}}}\|\widehat{\Psi}_{u0}\|_\infty\right]^2\right\}$.

Lemma F.4 (Performance of the Post-Lasso). *Under Conditions WL, let $\widehat{T}_u$ denote the support selected by $\widehat{\theta}_u$, and $\widetilde{\theta}_u$ be the Post-Lasso estimator based on $\widehat{T}_u$. Then, with probability $1 - o(1)$, uniformly over $u \in \mathcal{U}$, we have for $\widehat{s}_u = |\widehat{T}_u|$*

$$\|\mathrm{E}[Y_u \mid X] - f(X)'\widetilde{\theta}_u\|_{\mathbb{P}_n,2} \leqslant \frac{\sqrt{\widehat{s}_u}\sqrt{\log(p\vee n)}}{\sqrt{n}\ \phi_{\min}(\widehat{s}_u)} + \min_{\mathrm{supp}(\theta)\subseteq\widehat{T}_u} \|\mathrm{E}[Y_u \mid X] - f(X)'\theta\|_{\mathbb{P}_n,2}$$

Moreover, the following events $c_r \geqslant \sup_{u\in\mathcal{U}} \|r_u\|_{\mathbb{P}_n,2}$, $\ell\widehat{\Psi}_{u0} \leqslant \widehat{\Psi}_u \leqslant L\widehat{\Psi}_{u0}$, $u \in \mathcal{U}$, and $\lambda/n \geqslant c\sup_{u\in\mathcal{U}} \|\widehat{\Psi}_{u0}^{-1}\mathbb{E}_n[f(X)\zeta_u]\|_\infty$, for $c > 1/\ell$, imply that

$$\sup_{u\in\mathcal{U}} \min_{\mathrm{supp}(\theta)\subseteq\widehat{T}_u} \|\mathrm{E}[Y_u \mid X] - f(X)'\theta\|_{\mathbb{P}_n,2} \leqslant 3c_r + \left(L + \frac{1}{c}\right)\frac{2\lambda\sqrt{s}}{n\kappa_{\tilde{\mathbf{c}}}}\sup_{u\in\mathcal{U}}\|\widehat{\Psi}_{u0}\|_\infty.$$

F.2. **Finite Sample Results: Logistic Case.** For the model described in (68) with $\Lambda(t) = \exp(t)/\{1+\exp(t)\}$ and $M(y,t) = 1\{y=0\}\log(\Lambda(t)) + 1\{y=1\}\log(1-\Lambda(t))$ we will study finite sample properties of the associated Lasso and Post-Lasso estimators of $(\theta_u)_{u\in\mathcal{U}}$. In what follows we use the notation $M_u(\theta) = \mathbb{E}_n[M(Y_u, f(X)'\theta)]$ for convenience.

In the finite sample analysis we will consider not only the design matrix $\mathbb{E}_n[f(X)f(X)']$ but also a weighted counterpart $\mathbb{E}_n[w_u f(X)f(X)']$ where $w_{ui} = \mathrm{E}[Y_{ui} \mid X_i](1 - \mathrm{E}[Y_{ui} \mid X_i])$, $i = 1,\ldots,n$, $u \in \mathcal{U}$, is the conditional variance of the outcome variable Y_{ui}.

For $T_u = \mathrm{supp}(\theta_u)$, $|T_u| \geqslant 1$, the (logistic) restricted eigenvalue is defined as

$$\kappa_{\mathbf{c}} := \inf_{u\in\mathcal{U}} \min_{\|\delta_{T_u^c}\|_1 \leqslant \mathbf{c}\|\delta_{T_u}\|_1} \frac{\|\sqrt{w_u}f(X)'\delta\|_{\mathbb{P}_n,2}}{\|\delta_{T_u}\|}.$$

For a subset $A_u \subset \mathbb{R}^p$, $u \in \mathcal{U}$, let the non-linear impact coefficient be defined as

$$\bar{q}_{A_u} = \inf_{\delta \in A_u} \mathbb{E}_n\left[w_u |f(X)'\delta|^2\right]^{3/2} / \mathbb{E}_n\left[w_u |f(X)'\delta|^3\right].$$

Note that $\bar{q}_{A_u}$ can be bounded based as

$$\bar{q}_{A_u} = \inf_{\delta \in A_u} \frac{\mathbb{E}_n\left[w_u |f(X)'\delta|^2\right]^{3/2}}{\mathbb{E}_n\left[w_u |f(X)'\delta|^3\right]} \geqslant \inf_{\delta \in A_u} \frac{\mathbb{E}_n\left[w_u |f(X)'\delta|^2\right]^{1/2}}{\max_{i \leqslant n} \|f(X_i)\|_\infty \|\delta\|_1}$$

which induces good behavior provided A_u is appropriate (like the restrictive set in the definition of restricted eigenvalues). In Lemma F.5 we have $A_u = \Delta_{2\mathbf{c}} \cup \{\delta \in \mathbb{R}^p : \|\delta\|_1 \leqslant \frac{6c\|\widehat{\Psi}_{u0}^{-1}\|_\infty}{\ell c - 1} \frac{n}{\lambda} \|\frac{r_u}{\sqrt{w_u}}\|_{\mathbb{P}_n,2} \|\sqrt{w_u} f(X)'\delta_u\|_{\mathbb{P}_n,2}\}$, which leads to

$$\bar{q}_{A_u} \geqslant \frac{1}{\max_{i \leqslant n} \|f(X)\|_\infty} \left(\frac{\kappa_{2\mathbf{c}}}{\sqrt{s_u}(1+2\mathbf{c})} \wedge \frac{\lambda(\ell c - 1)}{6cn \|\widehat{\Psi}_{u0}^{-1}\|_\infty \|r_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}} \right) \gtrsim_P \frac{\kappa_{2\mathbf{c}}}{\sqrt{s_u} \max_{i \leqslant n} \|f(X_i)\|_\infty}$$

under $\|r_u/\sqrt{w_u}\|_{\mathbb{P}_n,2} \lesssim_P \sqrt{s_u/n}$ and $\lambda \gtrsim_P \sqrt{n \log p}$.

The definitions above differ from their counterpart in the analysis of ℓ_1-penalized least squares estimators by the weighting $0 \leqslant w_i \leqslant 1$. Thus it is relevant to understand their relations through the quantities

$$\psi_u(A) := \min_{\delta \in A} \frac{\|\sqrt{w_u} f(X)'\delta\|_{\mathbb{P}_n,2}}{\|f(X)'\delta\|_{\mathbb{P}_n,2}}.$$

Many primitive conditions on the data generating process will imply $\psi_u(A)$ to be bounded away from zero for the relevant choices of A. We refer to (Belloni, Chernozhukov, and Wei, 2013) for bounds on ψ_u. For notational convenience we will also work with a rescaling of the approximation errors $\tilde{r}_u(X)$ defined as

$$\tilde{r}_{ui} = \tilde{r}_u(X_i) = \Lambda^{-1}(\, \Lambda(f(X_i)'\theta_u) + r_{ui} \,) - f(X_i)'\theta_u,$$

which is the unique solution to $\Lambda(f(X_i)'\theta_u + \tilde{r}_u(X_i)) = \Lambda(f(X_i)'\theta_u) + r_u(X_i)$. It trivially follows that $|r_{ui}| \leqslant |\tilde{r}_{ui}|$ and that $|\tilde{r}_{ui}| \leqslant |r_{ui}| / \inf_{0 \leqslant t \leqslant \tilde{r}_{ui}} \Lambda'(f(X_i'\theta_u) + t) \leqslant |r_{ui}| / \{w_{ui} - 2|r_{ui}|\}$.

Next we derive finite sample bounds provided some crucial events occur.

Lemma F.5. *Assume $\lambda/n \geqslant c \sup_{u \in \mathcal{U}} \|\widehat{\Psi}_{u0}^{-1} \mathbb{E}_n[f(X)\zeta_u]\|_\infty$ for $c > 1$. Further, let $\ell \widehat{\Psi}_{u0} \leqslant \widehat{\Psi}_u \leqslant L\widehat{\Psi}_{u0}$ uniformly over $u \in \mathcal{U}$, $\tilde{\mathbf{c}} = (Lc+1)/(\ell c - 1) \sup_{u \in \mathcal{U}} \|\widehat{\Psi}_{u0}\|_\infty \|\widehat{\Psi}_{u0}^{-1}\|_\infty$. and $A_u = \Delta_{2\tilde{\mathbf{c}}} \cup \{\delta : \|\delta\|_1 \leqslant \frac{6c\|\widehat{\Psi}_{u0}^{-1}\|_\infty}{\ell c - 1} \frac{n}{\lambda} \|r_u/\sqrt{w_u}\|_{\mathbb{P}_n,2} \|\sqrt{w_u} f(X)'\delta_u\|_{\mathbb{P}_n,2}\}$. Provided that the nonlinear impact coefficient $\bar{q}_{A_u} > 3\left\{(L + \frac{1}{c})\|\widehat{\Psi}_{u0}\|_\infty \frac{\lambda\sqrt{s}}{n\kappa_{2\tilde{\mathbf{c}}}} + 9\tilde{\mathbf{c}} \|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\right\}$ for every $u \in \mathcal{U}$, we have uniformly over $u \in \mathcal{U}$*

$$\|\sqrt{w_u} f(X)'(\widehat{\theta}_u - \theta_u)\|_{\mathbb{P}_n,2} \leqslant 3\left\{(L + \frac{1}{c})\|\widehat{\Psi}_{u0}\|_\infty \frac{\lambda\sqrt{s}}{n\kappa_{2\tilde{\mathbf{c}}}} + 9\tilde{\mathbf{c}} \|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\right\} \quad \text{and}$$

$$\|\widehat{\theta}_u - \theta_u\|_1 \leqslant 3\left\{\frac{(1+\tilde{\mathbf{c}})\sqrt{s}}{\kappa_{2\tilde{\mathbf{c}}}} + \frac{6c\|\widehat{\Psi}_{u0}^{-1}\|_\infty}{\ell c - 1} \frac{n}{\lambda} \left\|\frac{r_u}{\sqrt{w_u}}\right\|_{\mathbb{P}_n,2}\right\} \left\{(L + \frac{1}{c})\|\widehat{\Psi}_{u0}\|_\infty \frac{\lambda\sqrt{s}}{n\kappa_{2\tilde{\mathbf{c}}}} + 9\tilde{\mathbf{c}} \left\|\frac{\tilde{r}_u}{\sqrt{w_u}}\right\|_{\mathbb{P}_n,2}\right\}$$

The following provides a bounds on the number of non-zero coefficients in the ℓ_1-penalized estimator $\widehat{\theta}_u$, uniformly over $u \in \mathcal{U}$.

Lemma F.6 (Sparsity of ℓ_1-Logistic Estimator). *Suppose* $\lambda/n \geqslant c\sup_{u\in\mathcal{U}} \|\widehat{\Psi}_{u0}^{-1}\mathbb{E}_n[f(X)\zeta_u]\|_\infty$ *then for* $\widehat{s}_u = |\mathrm{supp}(\widehat{\theta}_u)|$

$$\widehat{s}_u \leqslant \frac{\ell^2 c^2 (n/\lambda)^2}{(c\ell-1)^2}\phi_{\max}(\widehat{s}_u)\|f(X)'(\widehat{\theta}_u - \theta_u) + r_u\|_{\mathbb{P}_n,2}^2.$$

Moreover, if $\max_{i\leqslant n}|f(X_i)'(\widehat{\theta}_u - \theta_u) + \tilde{r}_{ui}| \leqslant 1$ *we have*

$$\widehat{s}_u \leqslant \frac{\ell^2 c^2 (n/\lambda)^2}{(c\ell-1)^2}\phi_{\max}(\widehat{s}_u)\|\sqrt{w_u}\{f(X)'(\widehat{\theta}_u - \theta_u) + \tilde{r}_u\}\|_{\mathbb{P}_n,2}^2.$$

Next we turn to finite sample bounds for the logistic regression estimator where the support was selected based on ℓ_1-penalized logistic regression. The results will hold uniformly over $u \in \mathcal{U}$ provided the side conditions also hold uniformly over $\mathcal{U}$.

Lemma F.7 (Post model selection Logistic regression rate). *Consider* $\widetilde{\theta}_u$ *defined by the support* $\widehat{T}_u^*$ *and let* $\widehat{s}_u^* := |\widehat{T}_u^*|$. *For each* $u \in \mathcal{U}$ *we have*

$$\|\sqrt{w_u}f(X)'(\widetilde{\theta}_u - \theta_u)\|_{\mathbb{P}_n,2} \leqslant \sqrt{3}\sqrt{M_u(\widetilde{\theta}_u) - M_u(\theta_u)} + 3\left\{\frac{\sqrt{\widehat{s}_u^* + s_u}\|\mathbb{E}_n[f(X)\zeta_u]\|_\infty}{\psi_u(A_u)\sqrt{\phi_{\min}(\widehat{s}_u^* + s_u)}} + 3\left\|\frac{\tilde{r}_u}{\sqrt{w_u}}\right\|_{\mathbb{P}_n,2}\right\}$$

provided that $\bar{q}_{A_u} > 6\left\{\frac{\sqrt{\widehat{s}_u^*+s_u}\|\mathbb{E}_n[f(X)\zeta_u]\|_\infty}{\psi_u(A_u)\sqrt{\phi_{\min}(\widehat{s}_u^*+s_u)}} + 3\left\|\frac{\tilde{r}_u}{\sqrt{w_u}}\right\|_{\mathbb{P}_n,2}\right\}$ *and* $\bar{q}_{A_u} > 6\sqrt{M_u(\widetilde{\theta}_u) - M_u(\theta_u)}$ *for* $A_u = \{\delta \in \mathbb{R}^p : \|\delta\|_0 \leqslant \widehat{s}_u^* + s_u\}$.

F.3. **Proofs for Lasso with Functional Response: Penalty Level.**

Proof of Lemma F.1. Condition WL implies that $\widehat{\Psi}_{u0jj}$ is bounded away from zero and from above uniformly in $j = 1,\ldots,p$ and n with probability at least $1 - \Delta_n$ for $n \geqslant n_0$. Also,

By the triangle inequality

$$\begin{aligned}\sup_{u\in\mathcal{U}}\|\widehat{\Psi}_{u0}^{-1}\mathbb{E}_n[f(X)\zeta_u]\|_\infty &\leqslant \sup_{u\in\mathcal{U}^\epsilon}\|\widehat{\Psi}_{u0}^{-1}\mathbb{E}_n[f(X)\zeta_u]\|_\infty \\ &\quad + \sup_{u,u'\in\mathcal{U},d_{\mathcal{U}}(u,u')\leqslant\epsilon}\|\widehat{\Psi}_{u0}^{-1}\mathbb{E}_n[f(X)\zeta_u] - \widehat{\Psi}_{u'0}^{-1}\mathbb{E}_n[f(X)\zeta_{u'}]\|_\infty\end{aligned}$$

where $\mathcal{U}^\epsilon$ is a minimal ϵ-cover of $\mathcal{U}$. We will set $\epsilon = 1/n$ so that $|\mathcal{U}^\epsilon| \leqslant n^\iota$.

The proofs in this section rely on the following result due to (Jing, Shao, and Wang, 2003).

Lemma F.8 (**Moderate deviations for self-normalized sums**). *Let* $X_1,\ldots, X_n$ *be independent, zero-mean random variables and* $\delta \in (0,1]$. *Let* $S_{n,n} = n\mathbb{E}_n[X_i]$, $V_{n,n}^2 = n\mathbb{E}_n[X_i^2]$ *and* $M_n = \{\bar{\mathrm{E}}[X_i^2]\}^{1/2}/\{\bar{\mathrm{E}}|X_i|^{2+\delta}]\}^{1/\{2+\delta\}} > 0$. *Suppose that for some* $\ell_n \to \infty$ *such that* $n^{\frac{\delta}{2(2+\delta)}}M_n/\ell_n \geqslant 1$. *Then for some absolute constant* A, *uniformly on* $0 \leqslant x \leqslant n^{\frac{\delta}{2(2+\delta)}}M_n/\ell_n - 1$, *we have*

$$\left|\frac{\mathrm{P}(|S_{n,n}/V_{n,n}| \geqslant x)}{2(1-\Phi(x))} - 1\right| \leqslant \frac{A}{\ell_n^{2+\delta}} \to 0.$$

Using Lemma F.8 with $\delta = 1$, $|\mathcal{U}^\epsilon| \leqslant n^\iota$ and the union bound, we have

$$\mathrm{P}\left(\sup_{u\in\mathcal{U}^\epsilon}\max_{j\leqslant p}\left|\frac{\sqrt{n}\mathbb{E}_n[f_j(X)\zeta_u]}{\sqrt{\mathbb{E}_n[f_j(X)^2\zeta_u^2]}}\right| > \Phi^{-1}(1-\gamma/2pn^\iota)\right) \leqslant \gamma\{1+o(1)\}$$

provided that $\Phi^{-1}(1-\gamma/2pn^{\iota}) \leqslant \delta_n n^{1/6}$ which holds by by Condition WL.

Moreover, by triangle inequality we have

$$\begin{aligned}
&\sup_{u,u' \in \mathcal{U}, d_{\mathcal{U}}(u,u') \leqslant \epsilon} \|\widehat{\Psi}_{u0}^{-1}\mathbb{E}_n[f(X)\zeta_u] - \widehat{\Psi}_{u'0}^{-1}\mathbb{E}_n[f(X)\zeta_{u'}]\|_\infty \\
&\quad \leqslant \sup_{u,u' \in \mathcal{U}, d_{\mathcal{U}}(u,u') \leqslant \epsilon} \|\widehat{\Psi}_{u0}^{-1} - \widehat{\Psi}_{u'0}^{-1}\|_\infty \|\mathbb{E}_n[f(X)\zeta_u]\|_\infty \\
&\quad + \sup_{u,u' \in \mathcal{U}, d_{\mathcal{U}}(u,u') \leqslant \epsilon} \|\mathbb{E}_n[f(X)(\zeta_u - \zeta_{u'})]\|_\infty \|\widehat{\Psi}_{u'0}^{-1}\|_\infty.
\end{aligned} \tag{80}$$

The last term in (80) is of the order $o(n^{-1/2})$ by Lemma F.9 and noting that $\sup_{u\in\mathcal{U}} \|\widehat{\Psi}_{u0}^{-1}\|_\infty$ is uniformly bounded with probability at least $1 - o(1) - \Delta_n$.

To control the first term in (80) we note that by Condition WL with probability $1 - \Delta_n$

$$\begin{aligned}
&\sup_{u,u' \in \mathcal{U}, d_{\mathcal{U}}(u,u') \leqslant \epsilon} \max_{j \leqslant p} |\{\mathbb{E}_n[f_j(X)^2\zeta_u^2]\}^{1/2} - \{\mathbb{E}_n[f_j(X)^2\zeta_{u'}^2]\}^{1/2}| \\
&\quad \leqslant \sup_{u,u' \in \mathcal{U}, d_{\mathcal{U}}(u,u') \leqslant \epsilon} K_n\{\mathbb{E}_n[(\zeta_u - \zeta_{u'})^2]\}^{1/2} \leqslant K_n C\{\epsilon^\nu + n^{-1/2}\}
\end{aligned} \tag{81}$$

Since $\widehat{\Psi}_{u0jj}$ is bounded away from zero with probability $1 - o(1)$ uniformly over $u \in \mathcal{U}$ and $j = 1, \ldots, p$, we have $|\widehat{\Psi}_{u0jj}^{-1} - \widehat{\Psi}_{u'0jj}^{-1}| = |\widehat{\Psi}_{u0jj} - \widehat{\Psi}_{u0jj}|/\{\widehat{\Psi}_{u0jj}\widehat{\Psi}_{u'0jj}\} \leqslant C|\widehat{\Psi}_{u0jj} - \widehat{\Psi}_{u0jj}|$ with the same probability. Thus, with $\epsilon = 1/n$ the relation (81) implies with probability $1 - o(1)$

$$\sup_{u,u' \in \mathcal{U}, d_{\mathcal{U}}(u,u') \leqslant \epsilon} \|\widehat{\Psi}_{u0}^{-1} - \widehat{\Psi}_{u'0}^{-1}\|_\infty \lesssim K_n n^{-(\nu \wedge \frac{1}{2})} \lesssim \delta_n / \log(p \vee n).$$

By Lemma A.4 with probability $1 - \delta$ we have

$$\sup_{u\in\mathcal{U}} \|\mathbb{E}_n[f(X)\zeta_u]\|_\infty \leqslant \tfrac{C'}{\sqrt{\delta}} \sqrt{\tfrac{\iota \log(p\vee n)}{n}} \sup_{u\in\mathcal{U}} \{(\mathbb{E}_n + \bar{\mathrm{E}})[f_j(X)^2\zeta_u^2]\}^{1/2}.$$

By Condition WL, with probability $1 - \Delta_n$ we have $\sup_{u\in\mathcal{U}} \{(\mathbb{E}_n + \bar{\mathrm{E}})[f_j(X)^2\zeta_u^2]\}^{1/2} \leqslant C$. Therefore, setting $\delta = \delta_n^{1/2}$ we have with probability $1 - \Delta_n - \delta_n^{1/2} = 1 - o(1)$

$$\sup_{u,u' \in \mathcal{U}, d_{\mathcal{U}}(u,u') \leqslant \epsilon} \|\widehat{\Psi}_{u0}^{-1} - \widehat{\Psi}_{u'0}^{-1}\|_\infty \|\mathbb{E}_n[f(X)\zeta_u]\|_\infty \leqslant \frac{\delta_n}{\log(p \vee n)} \frac{C''}{\delta_n^{1/4}} \sqrt{\frac{\log(p \vee n)}{n}} \leqslant o(1)/\sqrt{n}.$$

The results above imply that (80) is bounded by $o(1)/\sqrt{n}$ with probability $1 - o(1)$. Since $\sqrt{\log(2pn^{\iota}/\gamma)} \leqslant \Phi^{-1}(1 - \gamma/\{2pn^{\iota}\})$ we have that

$$P\left(\frac{(c' - c)}{\sqrt{n}} \Phi^{-1}(1 - \gamma/\{2pn^{\iota}\}) \geqslant \sup_{u,u' \in \mathcal{U}, d_{\mathcal{U}}(u,u') \leqslant \epsilon} \|\widehat{\Psi}_{u0}^{-1}\mathbb{E}_n[f(X)\zeta_u] - \widehat{\Psi}_{u'0}^{-1}\mathbb{E}_n[f(X)\zeta_{u'}]\|_\infty \right) = 1 - o(1)$$

and the results follows. ∎

Lemma F.9. *Under Condition WL we have that with probability* $1 - o(1)$

$$\sup_{u,u' \in \mathcal{U}, d_{\mathcal{U}}(u,u') \leqslant 1/n} \max_{j \leqslant p} \|\mathbb{E}_n[f_j(X_i)(\zeta_{ui} - \zeta_{u'i})]\|_\infty \leqslant o(1)/\sqrt{n}.$$

Proof. By Lemma A.4 we have that with probability $1-\delta$

$$\begin{aligned}
&\sup_{u,u'\in\mathcal{U},d_{\mathcal{U}}(u,u')\leqslant\epsilon}\|\mathbb{E}_n[f(X)(\zeta_u-\zeta_{u'})]\|_\infty \\
&\leqslant \tfrac{C'}{\sqrt{\delta}}\sqrt{\tfrac{2\iota\log(p\vee n)}{n}}\sup_{u,u'\in\mathcal{U},d_{\mathcal{U}}(u,u')\leqslant\epsilon}\{(\mathbb{E}_n+\bar{\mathrm{E}})[f_j(X_i)^2(\zeta_{ui}-\zeta_{u'i})^2]\}^{1/2} \\
&\leqslant \tfrac{C'}{\sqrt{\delta}}\sqrt{\tfrac{2\iota\log(p\vee n)}{n}}K_n\sup_{u,u'\in\mathcal{U},d_{\mathcal{U}}(u,u')\leqslant\epsilon}\{(\mathbb{E}_n+\bar{\mathrm{E}})[(\zeta_u-\zeta_{u'})^2]\}^{1/2}.
\end{aligned} \tag{82}$$

By Condition WL, ι is fixed, $K_n\log^{1/2}(p\vee n)n^{-(\nu\wedge\frac{1}{2})}\leqslant\delta_n$, and with probability $1-\Delta_n$

$$\sup_{u,u'\in\mathcal{U},d_{\mathcal{U}}(u,u')\leqslant\epsilon}\{(\mathbb{E}_n+\bar{\mathrm{E}})[(\zeta_u-\zeta_{u'})^2]\}^{1/2}\leqslant C\{\epsilon^\nu+n^{-1/2}\}$$

Taking $\epsilon=1/n$, with probability $1-\delta_n^{1/2}-\Delta_n=1-o(1)$ we have

$$\sup_{u,u'\in\mathcal{U},d_{\mathcal{U}}(u,u')\leqslant\epsilon}\|\mathbb{E}_n[f(X)(\zeta_u-\zeta_{u'})]\|_\infty\leqslant C''\delta_n^{1/4}n^{-1/2}.$$

■

F.4. **Proofs for Lasso with Functional Response: Linear Case.**

Proof of Lemma F.2. Let $\widehat{\delta}_u=\widehat{\theta}_u-\theta_u$. Throughout the proof we consider the events $c_r^2\geqslant\sup_{u\in\mathcal{U}}\mathbb{E}_n[r_u^2]$, $\lambda/n\geqslant c\sup_{u\in\mathcal{U}}\|\widehat{\Psi}_{u0}^{-1}\mathbb{E}_n[\zeta_u f(X)]\|_\infty$ and $\ell\widehat{\Psi}_{u0}\leqslant\widehat{\Psi}_u\leqslant L\widehat{\Psi}_{u0}$.

By definition of $\widehat{\theta}_u$ we have

$$\begin{aligned}
&\mathbb{E}_n[(f(X)'\widehat{\delta}_u)^2]-2\mathbb{E}_n[(Y_u-f(X)'\theta_u)f(X)]'\widehat{\delta}_u \\
&=\mathbb{E}_n[(Y_u-f(X)'\widehat{\theta}_u)^2]-\mathbb{E}_n[(Y_u-f(X)'\theta_u)^2] \\
&\leqslant\tfrac{2\lambda}{n}\|\widehat{\Psi}_u\theta_u\|_1-\tfrac{2\lambda}{n}\|\widehat{\Psi}_u\widehat{\theta}_u\|_1 \\
&\leqslant\tfrac{2\lambda}{n}\|\widehat{\Psi}_u\widehat{\delta}_{uT_u}\|_1-\tfrac{2\lambda}{n}\|\widehat{\Psi}_u\widehat{\delta}_{uT_u^c}\|_1 \\
&\leqslant\tfrac{2\lambda}{n}L\|\widehat{\Psi}_{u0}\widehat{\delta}_{uT_u}\|_1-\tfrac{2\lambda}{n}\ell\|\widehat{\Psi}_{u0}\widehat{\delta}_{uT_u^c}\|_1
\end{aligned} \tag{83}$$

Therefore, by $c_r^2\geqslant\sup_{u\in\mathcal{U}}\mathbb{E}_n[r_u^2]$ and $\lambda/n\geqslant c\sup_{u\in\mathcal{U}}\|\widehat{\Psi}_{u0}^{-1}\mathbb{E}_n[\zeta_u f(X)]\|_\infty$, we have

$$\begin{aligned}
&\mathbb{E}_n[(f(X)'\widehat{\delta}_u)^2] \\
&\leqslant 2\mathbb{E}_n[r_u f(X)]'\widehat{\delta}_u+2(\widehat{\Psi}_{u0}^{-1}\mathbb{E}_n[\zeta_u f(X)])'(\widehat{\Psi}_{u0}\widehat{\delta}_u)+\tfrac{2\lambda}{n}L\|\widehat{\Psi}_{u0}\widehat{\delta}_{uT_u}\|_1-\tfrac{2\lambda}{n}\ell\|\widehat{\Psi}_{u0}\widehat{\delta}_{uT_u^c}\|_1 \\
&\leqslant 2c_r\{\mathbb{E}_n[(f(X)'\widehat{\delta}_u)^2]\}^{1/2}+2\|\widehat{\Psi}_{u0}^{-1}\mathbb{E}_n[\zeta_u f(X)]\|_\infty\|\widehat{\Psi}_{u0}\widehat{\delta}_u\|_1 \\
&+\tfrac{2\lambda}{n}L\|\widehat{\Psi}_{u0}\widehat{\delta}_{uT_u}\|_1-\tfrac{2\lambda}{n}\ell\|\widehat{\Psi}_{u0}\widehat{\delta}_{uT_u^c}\|_1 \\
&\leqslant 2c_r\{\mathbb{E}_n[(f(X)'\widehat{\delta}_u)^2]\}^{1/2}+\tfrac{2\lambda}{cn}\|\widehat{\Psi}_{u0}\widehat{\delta}_u\|_1+\tfrac{2\lambda}{n}L\|\widehat{\Psi}_{u0}\widehat{\delta}_{uT_u}\|_1-\tfrac{2\lambda}{n}\ell\|\widehat{\Psi}_{u0}\widehat{\delta}_{uT_u^c}\|_1 \\
&\leqslant 2c_r\{\mathbb{E}_n[(f(X)'\widehat{\delta}_u)^2]\}^{1/2}+\tfrac{2\lambda}{n}\left(L+\tfrac{1}{c}\right)\|\widehat{\Psi}_{u0}\widehat{\delta}_{uT_u}\|_1-\tfrac{2\lambda}{n}\left(\ell-\tfrac{1}{c}\right)\|\widehat{\Psi}_{u0}\widehat{\delta}_{uT_u^c}\|_1
\end{aligned} \tag{84}$$

Let $\tilde{\mathbf{c}}=\frac{cL+1}{c\ell-1}\sup_{u\in\mathcal{U}}\|\widehat{\Psi}_{u0}\|_\infty\|\widehat{\Psi}_{u0}^{-1}\|_\infty$. Therefore if $\widehat{\delta}_u\notin\Delta_{\tilde{\mathbf{c}}}$ we have that $\left(L+\frac{1}{c}\right)\|\widehat{\Psi}_{u0}\widehat{\delta}_{uT_u}\|_1\leqslant\left(\ell-\frac{1}{c}\right)\|\widehat{\Psi}_{u0}\widehat{\delta}_{uT_u^c}\|_1$ so that

$$\{\mathbb{E}_n[(f(X)'\widehat{\delta}_u)^2]\}^{1/2}\leqslant 2c_r.$$

Otherwise assume $\widehat{\delta}_u\in\Delta_{\tilde{\mathbf{c}}}$. In this case (84) and the definition of $\kappa_{\tilde{\mathbf{c}}}$ yields

$$\begin{aligned}
\mathbb{E}_n[(f(X)'\widehat{\delta}_u)^2] &\leqslant 2c_r\{\mathbb{E}_n[(f(X)'\widehat{\delta})^2]\}^{1/2}+\tfrac{2\lambda}{n}\left(L+\tfrac{1}{c}\right)\|\widehat{\Psi}_{u0}\widehat{\delta}_{uT_u}\|_1-\tfrac{2\lambda}{n}\left(\ell-\tfrac{1}{c}\right)\|\widehat{\Psi}_{u0}\widehat{\delta}_{uT_u^c}\|_1 \\
&\leqslant 2c_r\{\mathbb{E}_n[(f(X)'\widehat{\delta}_u)^2]\}^{1/2}+\tfrac{2\lambda}{n}\left(L+\tfrac{1}{c}\right)\|\widehat{\Psi}_{u0}\|_\infty\sqrt{s}\{\mathbb{E}_n[(f(X)'\widehat{\delta}_u)^2]\}^{1/2}/\kappa_{\tilde{\mathbf{c}}}
\end{aligned}$$

which implies

$$\{\mathbb{E}_n[(f(X)'\widehat{\delta}_u)^2]\}^{1/2} \leqslant 2c_r + \frac{2\lambda\sqrt{s}}{n\kappa_{\tilde{\mathbf{c}}}}\left(L + \frac{1}{c}\right)\|\widehat{\Psi}_{u0}\|_\infty$$

To establish the ℓ_1-bound, first assume that $\widehat{\delta}_u \in \Delta_{2\tilde{\mathbf{c}}}$. In that case

$$\|\widehat{\delta}_u\|_1 \leqslant (1+2\tilde{\mathbf{c}})\|\widehat{\delta}_{uT_u}\|_1 \leqslant \sqrt{s}\{\mathbb{E}_n[(f(X)'\widehat{\delta}_u)^2]\}^{1/2}/\kappa_{2\tilde{\mathbf{c}}} \leqslant 2\frac{\sqrt{s}\ c_r}{\kappa_{2\tilde{\mathbf{c}}}} + \frac{2\lambda s}{n\kappa_{\tilde{\mathbf{c}}}\kappa_{2\tilde{\mathbf{c}}}}\left(L+\frac{1}{c}\right)\|\widehat{\Psi}_{u0}\|_\infty.$$

Otherwise note that $\widehat{\delta}_u \notin \Delta_{2\tilde{\mathbf{c}}}$ implies that $\left(L+\frac{1}{c}\right)\|\widehat{\Psi}_{u0}\widehat{\delta}_{uT_u}\|_1 \leqslant \frac{1}{2}\left(\ell - \frac{1}{c}\right)\|\widehat{\Psi}_{u0}\widehat{\delta}_{uT_u^c}\|_1$ so that (84) yields

$$\frac{1}{2}\frac{2\lambda}{n}\left(\ell - \frac{1}{c}\right)\|\widehat{\Psi}_{u0}\widehat{\delta}_{uT_u^c}\|_1 \leqslant \{\mathbb{E}_n[(f(X)'\widehat{\delta}_u)^2]\}^{1/2}\left(2c_r - \{\mathbb{E}_n[(f(X)'\widehat{\delta}_u)^2]\}^{1/2}\right) \leqslant c_r^2.$$

Therefore

$$\|\widehat{\delta}_u\|_1 \leqslant \left(1+\frac{1}{2\tilde{\mathbf{c}}}\right)\|\widehat{\delta}_{uT_u^c}\|_1 \leqslant \left(1+\frac{1}{2\tilde{\mathbf{c}}}\right)\|\widehat{\Psi}_{u0}^{-1}\|_\infty\|\widehat{\Psi}_{u0}\widehat{\delta}_{uT_u^c}\|_1 \leqslant \left(1+\frac{1}{2\tilde{\mathbf{c}}}\right)\frac{c\|\widehat{\Psi}_{u0}^{-1}\|_\infty}{\ell c - 1}\frac{n}{\lambda}c_r^2$$

■

Proof of Lemma F.3. Let $L_u = 4c_0\|\widehat{\Psi}_{u0}^{-1}\|_\infty\left[\frac{nc_r}{\lambda} + \frac{\sqrt{s}}{\kappa_{\tilde{\mathbf{c}}}}\|\widehat{\Psi}_{u0}\|_\infty\right]$. By Lemma F.10 and the definition of L_u we have

$$\widehat{s}_u \leqslant \phi_{\max}(\widehat{s}_u)L_u^2. \tag{85}$$

Consider any $M \in \mathcal{M}$, and suppose $\widehat{s}_u > M$. Therefore by the sublinearity of the maximum sparse eigenvalue (see Lemma F.11)

$$\widehat{s}_u \leqslant \left\lceil \frac{\widehat{s}_u}{M} \right\rceil \phi_{\max}(M)L_u^2.$$

Thus, since $\lceil k \rceil \leqslant 2k$ for any $k \geqslant 1$ we have

$$M \leqslant 2\phi_{\max}(M)L_u^2$$

which violates the condition that $M \in \mathcal{M}$. Therefore, we have $\widehat{s}_u \leqslant M$.

In turn, applying (85) once more with $\widehat{s}_u \leqslant M$ we obtain

$$\widehat{s}_u \leqslant \phi_{\max}(M)L_u^2.$$

The result follows by minimizing the bound over $M \in \mathcal{M}$. ■

Proof of Lemma F.4. Let $m_{ui} = \mathrm{E}[Y_{ui} \mid X_i]$, $F = [f(X_1); \ldots; f(X_n)]'$ and for a set of indices $S \subset \{1, \ldots, p\}$ we define $P_S = F[S](F[S]'F[S])^{-1}F[S]'$ and $\widehat{P}_S = F[S](F[S]'F[S])^{-1}F[S]'$ denote the projection matrix on the columns associated with the indices in S. Since $Y_{ui} = m_{ui} + \zeta_{ui}$ we have

$$m_u - f(X)\widetilde{\theta}_u = (I - \widehat{P}_{\widehat{T}_u})m_u - \widehat{P}_{\widehat{T}_u}\zeta_u$$

where I is the identity operator. Therefore

$$\|m_u - f(X)\widetilde{\theta}_u\| \leqslant \|(I - \widehat{P}_{\widehat{T}_u})m_u\| + \|\widehat{P}_{\widehat{T}_u}\zeta_u\|. \tag{86}$$

Since $\|X[\widehat{T}_u]/\sqrt{n}(F[\widehat{T}_u]'F[\widehat{T}_u]/n)^{-1}\| \leqslant \sqrt{1/\phi_{\min}(\widehat{s}_u)}$, the last term in (86) satisfies

$$\begin{aligned}\|\widehat{P}_{\widehat{T}_u}\zeta_u\| &\leqslant \sqrt{1/\phi_{\min}(\widehat{s}_u)}\ \|F[\widehat{T}_u]'\zeta_u/\sqrt{n}\| \\ &\leqslant \sqrt{1/\phi_{\min}(\widehat{s}_u)}\|F[\widehat{T}_u]'\zeta_u/\sqrt{n}\| \\ &\leqslant \sqrt{1/\phi_{\min}(\widehat{s}_u)}\sqrt{\widehat{s}_u}\|F'\zeta_u/\sqrt{n}\|_\infty.\end{aligned}$$

By Lemma F.1 with $\gamma = 1/n$, we have that with probability $1 - o(1)$

$$\sup_{u\in\mathcal{U}}\|F'\zeta_u/\sqrt{n}\|_\infty \leqslant C\sqrt{\iota\log(p\vee n)}\sup_{u\in\mathcal{U}}\max_{1\leqslant j\leqslant p}\sqrt{\mathbb{E}_n[f_j(X)^2\zeta_u^2]}.$$

By Condition WL, $\sup_{u\in\mathcal{U}}\max_{1\leqslant j\leqslant p}|(\mathbb{E}_n - \bar{\mathrm{E}})[f_j(X)^2\zeta_u^2]| \leqslant \delta_n$ with probability $1-\Delta_n$, and $\sup_{u\in\mathcal{U}}\max_{1\leqslant j\leqslant p}\bar{\mathrm{E}}[f_j(X)^2\zeta_u^2] \leqslant \bar{c}^2\mathbb{E}_n[f_j(X)^2] \leqslant \bar{c}^2$. Thus with probability $1-\Delta_n$,

$$\sup_{u\in\mathcal{U}}\max_{1\leqslant j\leqslant p}\sqrt{\mathbb{E}_n[f_j(X)^2\zeta_u^2]} \leqslant C.$$

The result follows.

The last statement follows from noting that the Lasso solution provides an upper bound to the approximation of the best model based on $\widehat{T}_u$, and the application of Lemma F.2. ∎

F.5. **Auxiliary Technical Lemmas for Lasso with Functional Response.**

Lemma F.10 (Empirical pre-sparsity for Lasso)**.** *Let $\widehat{T}_u$ denote the support selected by the Lasso estimator, $\widehat{s}_u = |\widehat{T}_u|$, assume $\lambda/n \geqslant c\sup_{u\in\mathcal{U}}\|\mathbb{E}_n[\widehat{\Psi}_{u0}^{-1}f(X)\zeta_u]\|_\infty$, and $\ell\widehat{\Psi}_{u0} \leqslant \widehat{\Psi}_u \leqslant L\widehat{\Psi}_{u0}$ for all $u\in\mathcal{U}$, with $L \geqslant 1 \geqslant \ell > 1/c$. Then, for $c_0 = (Lc+1)/(\ell c - 1)$ and $\tilde{\mathbf{c}} = (Lc+1)/(\ell c - 1)\sup_{u\in\mathcal{U}}\|\widehat{\Psi}_{u0}\|_\infty\|\widehat{\Psi}_{u0}^{-1}\|_\infty$ we have uniformly over $u\in\mathcal{U}$*

$$\sqrt{\widehat{s}_u} \leqslant 4\sqrt{\phi_{\max}(\widehat{s}_u)}\|\widehat{\Psi}_{u0}^{-1}\|_\infty c_0\left[\frac{nc_r}{\lambda} + \frac{\sqrt{s}}{\kappa_{\tilde{\mathbf{c}}}}\|\widehat{\Psi}_{u0}\|_\infty\right].$$

Proof of Lemma F.10. Let $R_u = (r_{u1},\ldots,r_{un})'$, and $F = [f(X_1);\ldots;f(X_n)]'$. We have from the optimality conditions that the Lasso estimator $\widehat{\theta}_u$ satisfies

$$\mathbb{E}_n[\widehat{\Psi}_{ujj}^{-1}f_j(X)(Y_u - f(X)'\widehat{\theta}_u)] = \mathrm{sign}(\widehat{\theta}_{uj})\lambda/n \quad \text{for each} \quad j\in\widehat{T}_u.$$

Therefore, noting that $\|\widehat{\Psi}_u^{-1}\widehat{\Psi}_{u0}\|_\infty \leqslant 1/\ell$, we have

$$\begin{aligned}&\sqrt{\widehat{s}_u}\lambda = \|(\widehat{\Psi}_u^{-1}F'(Y_u - f(X)\widehat{\theta}_u))_{\widehat{T}_u}\| \\ &\leqslant \|(\widehat{\Psi}_u^{-1}F'\zeta_u)_{\widehat{T}_u}\| + \|(\widehat{\Psi}_u^{-1}F'R_u)_{\widehat{T}_u}\| + \|(\widehat{\Psi}_u^{-1}F'F(\theta_u - \widehat{\theta}_u))_{\widehat{T}_u}\| \\ &\leqslant \sqrt{\widehat{s}_u}\ \|\widehat{\Psi}_u^{-1}\widehat{\Psi}_{u0}\|_\infty\|\widehat{\Psi}_{u0}^{-1}F'\zeta_u\|_\infty + n\sqrt{\phi_{\max}(\widehat{s}_u)}\|\widehat{\Psi}_u^{-1}\|_\infty c_r + \\ &n\sqrt{\phi_{\max}(\widehat{s}_u)}\|\widehat{\Psi}_u^{-1}\|_\infty\|F'(\widehat{\theta}_u - \theta_u)\|_{\mathbb{P}_n,2}, \\ &\leqslant \sqrt{\widehat{s}_u}\ (1/\ell)\ n\|\widehat{\Psi}_{u0}^{-1}F'\zeta_u\|_\infty + n\sqrt{\phi_{\max}(\widehat{s}_u)}\frac{\|\widehat{\Psi}_{u0}^{-1}\|_\infty}{\ell}\{c_r + \|F'(\widehat{\theta}_u - \theta_u)\|_{\mathbb{P}_n,2}\},\end{aligned}$$

where we used that

$$\begin{aligned}&\|(F'(\theta_u - \widehat{\theta}_u))_{\widehat{T}_u}\| \\ &\leqslant \sup_{\|\delta\|_0\leqslant\widehat{s}_u,\|\delta\|\leqslant 1}|\delta'F'F(\theta_u - \widehat{\theta}_u)| \leqslant \sup_{\|\delta\|_0\leqslant\widehat{s}_u,\|\delta\|\leqslant 1}\|\delta'F'\|\|F(\theta_u - \widehat{\theta}_u)\| \\ &\leqslant \sup_{\|\delta\|_0\leqslant\widehat{s}_u,\|\delta\|\leqslant 1}\{\delta'F'F\delta\}^{1/2}\|F(\theta_u - \widehat{\theta}_u)\| \leqslant n\sqrt{\phi_{\max}(\widehat{s}_u)}\|f(X)'(\theta_u - \widehat{\theta}_u)\|_{\mathbb{P}_n,2}.\end{aligned}$$

Since $\lambda/c \geqslant \sup_{u\in\mathcal{U}} \|\widehat{\Psi}_{u0}^{-1}F'\zeta_u\|_\infty$, and by Lemma F.2, we have that the estimate $\widehat{\theta}_u$ satisfies $\|f(X)'(\widehat{\theta}_u - \theta_u)\|_{\mathbb{P}_n,2} \leqslant 2c_r + \left(L + \frac{1}{c}\right)\frac{\lambda\sqrt{s_u}}{n\kappa_{\tilde{\mathbf{c}}}}\|\widehat{\Psi}_{u0}\|_\infty$ so that

$$\sqrt{\widehat{s}_u} \leqslant \frac{\sqrt{\phi_{\max}(\widehat{s}_u)}\frac{\|\widehat{\Psi}_{u0}^{-1}\|_\infty}{\ell}\left[\frac{3nc_r}{\lambda} + \left(L+\frac{1}{c}\right)\frac{\sqrt{s}}{\kappa_{\tilde{\mathbf{c}}}}\|\widehat{\Psi}_{u0}\|_\infty\right]}{\left(1-\frac{1}{c\ell}\right)}.$$

The result follows by noting that $(L + [1/c])/(1 - 1/[\ell c]) = c_0\ell$ by definition of c_0. ∎

Lemma F.11 (Sub-linearity of maximal sparse eigenvalues)**.** *Let M be a semi-definite positive matrix. For any integer $k \geqslant 0$ and constant $\ell \geqslant 1$ we have $\phi_{\max}(\lceil \ell k\rceil)(M) \leqslant \lceil \ell\rceil\phi_{\max}(k)(M)$.*

F.6. **Proofs for Lasso with Functional Response: Logistic Case.**

Proof of Lemma F.5. Let $\delta_u = \widehat{\theta}_u - \theta_u$ and $S_u = \mathbb{E}_n[f(X)\zeta_u]$. By definition of $\widehat{\theta}_u$ we have $M_u(\widehat{\theta}_u) + \frac{\lambda}{n}\|\widehat{\Psi}_u\widehat{\theta}_u\|_1 \leqslant M_u(\theta_u) + \frac{\lambda}{n}\|\widehat{\Psi}\theta_u\|_1$. Thus,

$$\begin{aligned} M_u(\widehat{\theta}_u) - M_u(\theta_u) &\leqslant \tfrac{\lambda}{n}\|\widehat{\Psi}_u\theta_u\|_1 - \tfrac{\lambda}{n}\|\widehat{\Psi}_u\widehat{\theta}_u\|_1 \\ &\leqslant \tfrac{\lambda}{n}\|\widehat{\Psi}_u\delta_{T_u}\|_1 - \tfrac{\lambda}{n}\|\widehat{\Psi}_u\delta_{T_u^c}\|_1 \\ &\leqslant \tfrac{\lambda L}{n}\|\widehat{\Psi}_{u0}\delta_{T_u}\|_1 - \tfrac{\lambda\ell}{n}\|\widehat{\Psi}_{u0}\delta_{T_u^c}\|_1 \end{aligned} \tag{87}$$

However, by convexity of $M_u(\cdot)$ and Hölder's inequality we have

$$\begin{aligned} M_u(\widehat{\theta}_u) - M_u(\theta_u) &\geqslant \nabla M_u(\theta_u)'\delta_u \\ &= \{\nabla M_u(\theta_u) - S_u\}'\delta_u + S_u'\delta_u \\ &\geqslant -\|\widehat{\Psi}_{u0}^{-1}S_u\|_\infty\|\widehat{\Psi}_{u0}\delta_u\|_1 - \|r_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\|\sqrt{w_u}f(X)'\delta_u\|_{\mathbb{P}_n,2} \\ &\geqslant -\tfrac{\lambda}{n}\tfrac{1}{c}\|\widehat{\Psi}_{u0}\delta_{u,T_u}\|_1 - \tfrac{\lambda}{n}\tfrac{1}{c}\|\widehat{\Psi}_{u0}\delta_{u,T_u^c}\|_1 \\ &\quad -\|r_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\|\sqrt{w_u}f(X)'\delta_u\|_{\mathbb{P}_n,2} \end{aligned} \tag{88}$$

Combining (87) and (88) we have

$$\frac{\lambda}{n}\frac{c\ell-1}{c}\|\widehat{\Psi}_{u0}\delta_{u,T_u^c}\|_1 \leqslant \frac{\lambda}{n}\frac{Lc+1}{c}\|\widehat{\Psi}_{u0}\delta_{u,T_u}\|_1 + \|r_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\|\sqrt{w_u}f(X)'\delta_u\|_{\mathbb{P}_n,2}$$

and for $\tilde{\mathbf{c}} = \frac{Lc+1}{\ell c-1}\sup_{u\in\mathcal{U}}\|\widehat{\Psi}_{u0}\|_\infty\|\widehat{\Psi}_{u0}^{-1}\|_\infty$ we have

$$\|\delta_{uT_u^c}\|_1 \leqslant \tilde{\mathbf{c}}\|\delta_{uT_u}\|_1 + \frac{n}{\lambda}\frac{c\|\widehat{\Psi}_{u0}^{-1}\|_\infty}{\ell c-1}\|r_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\|\sqrt{w_u}f(X)'\delta_u\|_{\mathbb{P}_n,2}$$

Suppose $\|\delta_{u,T_u^c}\|_1 \geqslant 2\tilde{\mathbf{c}}\|\delta_{u,T_u}\|_1$. Thus,

$$\begin{aligned} \|\delta_u\|_1 &\leqslant (1+\{2\tilde{\mathbf{c}}\}^{-1})\|\delta_{u,T_u^c}\|_1 \\ &\leqslant (1+\{2\tilde{\mathbf{c}}\}^{-1})\tilde{\mathbf{c}}\|\delta_{u,T_u}\|_1 + (1+\{2\tilde{\mathbf{c}}\}^{-1})\tfrac{n}{\lambda}\tfrac{c\|\widehat{\Psi}_{u0}^{-1}\|_\infty}{\ell c-1}\|r_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\|\sqrt{w_u}f(X)'\delta_u\|_{\mathbb{P}_n,2} \\ &\leqslant (1+\{2\tilde{\mathbf{c}}\}^{-1})\tfrac{1}{2}\|\delta_{u,T_u^c}\|_1 + (1+\{2\tilde{\mathbf{c}}\}^{-1})\tfrac{n}{\lambda}\tfrac{c\|\widehat{\Psi}_{u0}^{-1}\|_\infty}{\ell c-1}\|r_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\|\sqrt{w_u}f(X)'\delta_u\|_{\mathbb{P}_n,2} \\ &\leqslant \tfrac{4\tilde{\mathbf{c}}}{2\tilde{\mathbf{c}}-1}(1+\{2\tilde{\mathbf{c}}\}^{-1})\tfrac{n}{\lambda}\tfrac{c\|\widehat{\Psi}_{u0}^{-1}\|_\infty}{\ell c-1}\|r_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\|\sqrt{w_u}f(X)'\delta_u\|_{\mathbb{P}_n,2} \\ &\leqslant \tfrac{6c\|\widehat{\Psi}_{u0}^{-1}\|_\infty}{\ell c-1}\tfrac{n}{\lambda}\|r_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\|\sqrt{w_u}f(X)'\delta_u\|_{\mathbb{P}_n,2} \end{aligned}$$

Note that the relation above combined with the definition of $\kappa_{2\tilde{\mathbf{c}}}$ yields for δ_u that

$$\|\delta_{u,T_u}\|_1 \leqslant \sqrt{s}\frac{\|\sqrt{w_u}f(X)'\delta_u\|_{\mathbb{P}_n,2}}{\kappa_{2\tilde{\mathbf{c}}}} + \frac{6c\|\widehat{\Psi}_{u0}^{-1}\|_\infty}{\ell c-1}\frac{n}{\lambda}\|r_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\|\sqrt{w_u}f(X)'\delta_u\|_{\mathbb{P}_n,2}. \tag{89}$$

Since $A_u := \Delta_{2\tilde{\mathbf{c}}} \cup \{\delta : \|\delta\|_1 \leqslant \frac{6c\|\widehat{\Psi}_{u0}^{-1}\|_\infty}{\ell c-1}\frac{n}{\lambda}\|r_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\|\sqrt{w_u}f(X)'\delta_u\|_{\mathbb{P}_n,2}\}$, we have

$$
\begin{aligned}
&\tfrac{1}{3}\|\sqrt{w_u}f(X)'\delta_u\|_{\mathbb{P}_n,2}^2 \wedge \left\{\tfrac{\bar{q}_A}{3}\|\sqrt{w_u}f(X)'\delta_u\|_{\mathbb{P}_n,2}\right\} \\
&\leqslant_{(1)} M_u(\widehat{\theta}_u) - M_u(\theta_u) - \nabla M_u(\theta_u)'\delta_u + 2\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\|\sqrt{w_u}f(X)'\delta_u\|_{\mathbb{P}_n,2} \\
&\leqslant_{(2)} (L+\tfrac{1}{c})\tfrac{\lambda}{n}\|\widehat{\Psi}_{u0}\delta_{u,T_u}\|_1 + 3\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\|\sqrt{w_u}f(X)'\delta_u\|_{\mathbb{P}_n,2} \\
&\leqslant_{(3)} (L+\tfrac{1}{c})\|\widehat{\Psi}_{u0}\|_\infty\tfrac{\lambda}{n}\left\{\sqrt{s}\tfrac{\|\sqrt{w_u}f(X)'\delta_u\|_{\mathbb{P}_n,2}}{\kappa_{2\tilde{\mathbf{c}}}} + \tfrac{6c\|\widehat{\Psi}_{u0}^{-1}\|_\infty}{\ell c-1}\tfrac{n}{\lambda}\|r_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\|\sqrt{w_u}f(X)'\delta_u\|_{\mathbb{P}_n,2}\right\} \\
&+3\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\|\sqrt{w_u}f(X)'\delta_u\|_{\mathbb{P}_n,2} \\
&\leqslant \left\{(L+\tfrac{1}{c})\|\widehat{\Psi}_{u0}\|_\infty\tfrac{\lambda\sqrt{s}}{n\kappa_{2\tilde{\mathbf{c}}}} + 9\tilde{\mathbf{c}}\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\right\}\|\sqrt{w_u}f(X)'\delta_u\|_{\mathbb{P}_n,2}
\end{aligned}
$$

where (1) follows by by Lemma F.12 with A_u, (2) follows from $|r_{ui}| \leqslant |\tilde{r}_{ui}|$, (87) and (88), (3) follows by (89), and (4) follows from simplifications. Provided that

$$
\bar{q}_{A_u} > 3\left\{(L+\frac{1}{c})\|\widehat{\Psi}_{u0}\|_\infty\frac{\lambda\sqrt{s}}{n\kappa_{2\tilde{\mathbf{c}}}} + 9\tilde{\mathbf{c}}\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\right\},
$$

so that the minimum on the LHS needs to be the quadratic term for each $u \in \mathcal{U}$, we have

$$
\|\sqrt{w_u}f(X)'\delta_u\|_{\mathbb{P}_n,2} \leqslant 3\left\{(L+\frac{1}{c})\|\widehat{\Psi}_{u0}\|_\infty\frac{\lambda\sqrt{s}}{n\kappa_{2\tilde{\mathbf{c}}}} + 9\tilde{\mathbf{c}}\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\right\} \quad \text{for every} \quad u \in \mathcal{U}.
$$

■

Proof of Lemma F.6. Recall that $\Lambda_{ui} = \mathrm{E}[Y_{ui} \mid X]$ and $S_u = \mathbb{E}_n[f(X)\zeta_u] = \mathbb{E}_n[(Y_u - \Lambda_u)f(X)]$. Let $\widehat{T}_u = \mathrm{supp}(\widehat{\theta}_u)$, $\widehat{s}_u = |\widehat{T}_u|$, $\delta_u = \widehat{\theta}_u - \theta_u$, and $\widehat{\Lambda}_{ui} = \exp(f(X)'\widehat{\theta}_u)/\{1+\exp(f(X)'\widehat{\theta}_u)\}$. For any $j \in \widehat{T}_u$ we have $|\mathbb{E}_n[(Y_u - \widehat{\Lambda}_u)f_j(X)]| = \widehat{\Psi}_{ujj}\lambda/n$.

Since $\ell\widehat{\Psi}_{u0} \leqslant \widehat{\Psi}_u$ implies $\|\widehat{\Psi}_u^{-1}\widehat{\Psi}_{u0}\|_\infty \leqslant 1/\ell$, the first relation follows from

$$
\begin{aligned}
\tfrac{\lambda}{n}\sqrt{\widehat{s}_u} &= \|\widehat{\Psi}_{u\widehat{T}_u}^{-1}\mathbb{E}_n[(Y_u - \widehat{\Lambda}_u)f_{\widehat{T}_u}(X)]\|_2 \\
&\leqslant \|\widehat{\Psi}_u^{-1}\widehat{\Psi}_{u0}\|_\infty\|\widehat{\Psi}_{u0}^{-1}\mathbb{E}_n[(Y_u - \Lambda_u)f_{\widehat{T}_u}(X)]\|_2 + \|\widehat{\Psi}_u^{-1}\widehat{\Psi}_{u0}\|_\infty\|\widehat{\Psi}_{u0}^{-1}\|_\infty\|\mathbb{E}_n[(\widehat{\Lambda}_u - \Lambda_u)f_{\widehat{T}_u}(X)]\|_2 \\
&\leqslant \sqrt{\widehat{s}_u}(1/\ell)\|\widehat{\Psi}_{u0}^{-1}\mathbb{E}_n[\zeta_u f(X)]\|_\infty + (1/\ell)\|\widehat{\Psi}_{u0}^{-1}\|_\infty \sup_{\|\theta\|_0\leqslant\widehat{s}_u,\|\theta\|=1}\mathbb{E}_n[|\widehat{\Lambda}_u - \Lambda_u|\cdot|f(X)'\theta|] \\
&\leqslant \tfrac{\lambda}{\ell cn}\sqrt{\widehat{s}_u} + \sqrt{\phi_{\max}(\widehat{s}_u)}(1/\ell)\|\widehat{\Psi}_{u0}^{-1}\|_\infty\|f(X)'\delta_u + r_u\|_{\mathbb{P}_n,2}
\end{aligned}
$$

The other relation follows from

$$
\begin{aligned}
\tfrac{\lambda}{n}\sqrt{\widehat{s}_u} &= \|\widehat{\Psi}_{u\widehat{T}_u}^{-1}\mathbb{E}_n[(Y_u - \widehat{\Lambda}_u)f_{\widehat{T}_u}(X)]\|_2 \\
&\leqslant (1/\ell)\|\widehat{\Psi}_{u0}^{-1}\mathbb{E}_n[(Y_u - \Lambda_u)f_{\widehat{T}_u}(X)]\|_2 + (1/\ell)\|\widehat{\Psi}_{u0}^{-1}\|_\infty\|\mathbb{E}_n[(\widehat{\Lambda}_u - \Lambda_u)f_{\widehat{T}_u}(X)]\|_2 \\
&\leqslant \sqrt{\widehat{s}_u}(1/\ell)\|\mathbb{E}_n[\zeta_u f(X)]\|_\infty + (1/\ell)\|\widehat{\Psi}_{u0}^{-1}\|_\infty \sup_{\|\theta\|_0\leqslant\widehat{s}_u,\|\theta\|=1}\mathbb{E}_n[|\widehat{\Lambda}_u - \Lambda_u|\cdot|f(X)'\theta|] \\
&\leqslant \tfrac{\lambda}{\ell cn}\sqrt{\widehat{s}_u} + (2/\ell)\|\widehat{\Psi}_{u0}^{-1}\|_\infty\sqrt{\phi_{\max}(\widehat{s}_u)}\|\sqrt{w_u}f(X)'\delta_u + \tilde{r}_u\|_{\mathbb{P}_n,2}
\end{aligned}
$$

where we used Lemma F.15 so that $|\widehat{\Lambda}_{ui} - \Lambda_{ui}| \leqslant w_{ui}2|f(X)'\delta_u + \tilde{r}_{ui}|$ since $\max_{i\leqslant n}|f(X_i)'\delta_u + \tilde{r}_{ui}| \leqslant 1$ is assumed. ■

Proof of Lemma F.7. Let $\tilde{\delta}_u = \tilde{\theta}_u - \theta_u$ and $\tilde{t}_u = \|\sqrt{w_u} f(X)'\tilde{\delta}_u\|_{\mathbb{P}_n,2}$ and $S_u = \mathbb{E}_n[f(X)\zeta_u]$. By Lemma F.12 with $A_u = \{\delta \in \mathbb{R}^p : \|\delta\|_0 \leqslant \widehat{s}_u^* + s_u\}$, we have

$$\begin{aligned}
\tfrac{1}{3}\tilde{t}_u^2 \wedge \left\{\tfrac{\bar{q}_{A_u}}{3}\tilde{t}_u\right\} &\leqslant M_u(\tilde{\theta}_u) - M_u(\theta_u) - \nabla M_u(\theta_u)'\tilde{\delta}_u + 2\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\tilde{t}_u \\
&\leqslant M_u(\tilde{\theta}_u) - M_u(\theta_u) + \|S_u\|_\infty \|\tilde{\delta}_u\|_1 + 3\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\tilde{t}_{\mathbb{P}_n,2} \\
&\leqslant M_u(\tilde{\theta}_u) - M_u(\theta_u) + \tilde{t}_{\mathbb{P}_n,2}\left\{\frac{\sqrt{\widehat{s}_u^* + s_u}\|S_u\|_\infty}{\psi_u(A_u)\sqrt{\phi_{\min}(\widehat{s}_u^* + s_u)}} + 3\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\right\}
\end{aligned}$$

Provided that $\bar{q}_{A_u}/6 > \left\{\frac{\sqrt{\widehat{s}_u^* + s_u}\|S_u\|_\infty}{\psi_u(A_u)\sqrt{\phi_{\min}(\widehat{s}_u^* + s_u)}} + 3\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\right\}$ and $\bar{q}_{A_u}/6 > \sqrt{M_u(\tilde{\theta}_u) - M_u(\theta_u)}$, if the minimum on the LHS is the linear term, we have $\tilde{t}_u \leqslant \sqrt{M_u(\tilde{\theta}_u) - \Lambda(\theta_u)}$ which implies the result. Otherwise, since for positive numbers $a^2 \leqslant b + ac$ implies $a \leqslant \sqrt{b} + c$, we have

$$\tilde{t}_{\mathbb{P}_n,2} \leqslant \sqrt{3}\sqrt{M_u(\tilde{\theta}_u) - M_u(\theta_u)} + 3\left\{\frac{\sqrt{\widehat{s}_u^* + s_u}\|S_u\|_\infty}{\psi_u(A_u)\sqrt{\phi_{\min}(\widehat{s}_u^* + s_u)}} + 3\|\tilde{r}_{ui}/\sqrt{w_{ui}}\|_{\mathbb{P}_n,2}\right\}.$$

■

F.7. **Technical Lemmas: Logistic Case.**

Lemma F.12 (Minoration Lemma)**.** *For any $u \in \mathcal{U}$ and $\delta \in A$, we have*

$$\begin{aligned}
&M_u(\theta_u + \delta) - M_u(\theta_u) - \nabla M_u(\theta_u)'\delta + 2\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\|\sqrt{w_u} f(X)'\delta\|_{\mathbb{P}_n,2} \\
&\quad \geqslant \left\{\tfrac{1}{3}\|\sqrt{w_u} f(X)'\delta\|_{\mathbb{P}_n,2}^2\right\} \wedge \left\{\tfrac{\bar{q}_A}{3}\|\sqrt{w_u} f(X)'\delta\|_{\mathbb{P}_n,2}\right\}
\end{aligned}$$

Proof. Step 1. (Minoration). Consider the following convex function

$$F_u(\delta) = M_u(\theta_u + \delta) - M_u(\theta_u) - \nabla M_u(\theta_u)'\delta + 2\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\|\sqrt{w_u} f(X)'\delta\|_{\mathbb{P}_n,2}.$$

Define the maximal radius over which F can be minorated by a quadratic function

$$r_{A_u} = \sup_r \left\{ r \ : \quad F(\delta) \geqslant \tfrac{1}{3}\|\sqrt{w_u} f(X)'\delta\|_{\mathbb{P}_n,2}^2, \text{for all } \delta \in A,\ \|\sqrt{w_u} f(X)'\delta\|_{\mathbb{P}_n,2} \leqslant r \ \right\}.$$

Step 2 below shows that $r_{A_u} \geqslant \bar{q}_{A_u}$. By construction of r_{A_u}, the convexity of F_u and $F_u(0) = 0$,

$$\begin{aligned}
F_u(\delta) &\geqslant \tfrac{\|\sqrt{w_u} f(X)'\delta\|_{\mathbb{P}_n,2}^2}{3} \wedge \left\{\tfrac{\|\sqrt{w_u} f(X)'\delta\|_{\mathbb{P}_n,2}}{r_{A_u}} \cdot \inf_{\tilde{\delta} \in A_u, \|\sqrt{w_u} f(X)'\tilde{\delta}\|_{\mathbb{P}_n,2} \geqslant r_{A_u}} F_u(\tilde{\delta})\right\} \\
&\geqslant \tfrac{\|\sqrt{w_u} f(X)'\delta\|_{\mathbb{P}_n,2}^2}{3} \wedge \left\{\tfrac{\|\sqrt{w_u} f(X)'\delta\|_{\mathbb{P}_n,2}}{r_{A_u}} \tfrac{r_{A_u}^2}{3}\right\} \geqslant \tfrac{\|\sqrt{w_u} f(X)'\delta\|_{\mathbb{P}_n,2}^2}{3} \wedge \left\{\tfrac{\bar{q}_{A_u}}{3}\|\sqrt{w_u} f(X)'\delta\|_{\mathbb{P}_n,2}\right\}.
\end{aligned}$$

Step 2. ($r_{A_u} \geqslant \bar{q}_{A_u}$) Let $\tilde{r}_{ui}$ be such that $\Lambda(f(X_i'\theta_u) + \tilde{r}_{ui}) = \Lambda(f(X_i'\theta_u)) + r_{ui} = \mathrm{E}[Y_{ui} \mid X]$. Defining $g_{ui}(t) = \log\{1 + \exp(f(X_i)'\theta_u + \tilde{r}_{ui} + tf(X_i)'\delta)\}$, $\tilde{g}_{ui}(t) = \log\{1 + \exp(f(X)'\theta_u + tf(X)'\delta)\}$,

$\Lambda_{ui} := \mathrm{E}[Y_{ui} \mid X]$, $\tilde{\Lambda}_{ui} := \exp(f(X)'\theta_u)/\{1+\exp(f(X)'\theta_u)\}$, we have

$$\begin{aligned}
&M_u(\theta_u+\delta) - M_u(\theta_u) - \nabla M_u(\theta_u)'\delta = \\
&= \mathbb{E}_n\left[\log\{1+\exp(f(X)'\{\theta_u+\delta\})\} - Y_u f(X)'(\theta_u+\delta)\right] \\
&\quad -\mathbb{E}_n\left[\log\{1+\exp(f(X)'\theta_u) - Y_u f(X)'\theta_u\}\right] - \mathbb{E}_n\left[(\tilde{\Lambda}_u - Y_u)f(X)'\delta\right] \\
&= \mathbb{E}_n\left[\log\{1+\exp(f(X)'\{\theta_u+\delta\})\} - \log\{1+\exp(f(X)'\theta_u)\} - \tilde{\Lambda}_u f(X)'\delta\right] \\
&= \mathbb{E}_n[\tilde{g}_u(1) - \tilde{g}_u(0) - 1\cdot\tilde{g}_u'(0)] \\
&= \mathbb{E}_n[g_u(1) - g_u(0) - 1\cdot g_u'(0)] + \mathbb{E}_n[\{\tilde{g}_u(1) - g_u(1)\} - \{\tilde{g}_u(0) - g_u(0)\} - \{\tilde{g}_u'(0) - g_u'(0)\}]
\end{aligned}$$

Note that the function g_{ui} is three times differentiable and satisfies, for

$$\begin{gathered}
\Lambda_{ui}(t) := \exp(f(X_i)'\theta_u + \tilde{r}_{ui} + t f(X_i)'\delta)/\{1+\exp(f(X_i)'\theta_u + \tilde{r}_{ui} + t f(X)'\delta)\}, \\
g_{ui}'(t) = (f(X_i)'\delta)\Lambda_{ui}(t), \quad g_{ui}''(t) = (f(X_i)'\delta)^2\Lambda_{ui}(t)[1-\Lambda_{ui}(t)], \\
g_{ui}'''(t) = (f(X_i)'\delta)^3\Lambda_{ui}(t)[1-\Lambda_{ui}(t)][1-2\Lambda_{ui}(t)].
\end{gathered}$$

Thus $|g_{ui}'''(t)| \leqslant |f(X)'\delta| g_{ui}''(t)$. Therefore, by Lemmas F.13 and F.14 we have

$$\begin{aligned}
g_{ui}(1) - g_{ui}(0) - 1\cdot g_{ui}'(0) &\geqslant \tfrac{(f(X_i)'\delta)^2 w_{ui}}{(f(X_i)'\delta)^2}\left\{\exp(-|f(X_i)'\delta|) + |f(X_i)'\delta| - 1\right\} \\
&\geqslant w_{ui}\left\{\tfrac{|f(X_i)'\delta|^2}{2} - \tfrac{|f(X_i)'\delta|^3}{6}\right\}
\end{aligned}$$

Moreover, letting $\Upsilon_{ui}(t) = \tilde{g}_{ui}(t) - g_{ui}(t)$ we have $|\Upsilon_{ui}'(t)| = |(f(X_i)'\delta)\{\Lambda_{ui}(t) - \tilde{\Lambda}_{ui}(t)\}| \leqslant |f(X_i)'\delta|\cdot|r_{ui}|$ where $\tilde{\Lambda}_{ui}(t) := \exp(f(X_i)'\theta_u + t f(X_i)'\delta)/\{1+\exp(f(X_i)'\theta_u + t f(X_i)'\delta)\}$. Thus

$$\begin{aligned}
&|\mathbb{E}_n[\{\tilde{g}_u(1) - g_u(1)\} - \{\tilde{g}_u(0) - g_u(0)\} - \{\tilde{g}_u'(0) - g_u'(0)\}]| = \\
&= |\mathbb{E}_n[\Upsilon_u(1) - \Upsilon_u(0) - \{\tilde{\Lambda}_u - \Lambda_u\}f(X)'\delta]| \\
&\leqslant 2\mathbb{E}_n[|\tilde{r}_u|\ |f(X)'\delta|]
\end{aligned}$$

Therefore we have

$$\begin{aligned}
M_u(\theta_u+\delta) - M_u(\theta_u) - \nabla M_u(\theta_u)'\delta &\geqslant \tfrac{1}{2}\mathbb{E}_n\left[w_u|f(X)'\delta|^2\right] - \tfrac{1}{6}\mathbb{E}_n\left[w_u|f(X)'\delta|^3\right] \\
&\quad -2\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\|\sqrt{w_u}f(X)'\delta\|_{\mathbb{P}_n,2}
\end{aligned}$$

Note that for any $\delta \in A_u$ such that $\|\sqrt{w_u}f(X)'\delta\|_{\mathbb{P}_n,2} \leqslant \bar{q}_{A_u}$ we have

$$\|\sqrt{w_u}f(X)'\delta\|_{\mathbb{P}_n,2} \leqslant \bar{q}_{A_u} \leqslant \|\sqrt{w_u}f(X)'\delta\|_{\mathbb{P}_n,2}^3/\mathbb{E}_n\left[w_u|f(X)'\delta|^3\right],$$

so that $\mathbb{E}_n[w_u|f(X)'\delta|^3] \leqslant \mathbb{E}_n[w_u|f(X)'\delta|^2]$. Therefore we have

$$\tfrac{1}{2}\mathbb{E}_n\left[w_u|f(X)'\delta|^2\right] - \tfrac{1}{6}\mathbb{E}_n\left[w_u|f(X)'\delta|^3\right] \geqslant \frac{1}{3}\mathbb{E}_n\left[w_u|f(X)'\delta|^2\right] \quad \text{and}$$

$$M_u(\theta_u+\delta) - M_u(\theta_u) - \nabla M_u(\theta_u)'\delta \geqslant \tfrac{1}{3}\mathbb{E}_n\left[w_u|f(X)'\delta|^2\right] - 2\|\tfrac{\tilde{r}_u}{\sqrt{w_u}}\|_{\mathbb{P}_n,2}\|\sqrt{w_u}f(X)'\delta\|_{\mathbb{P}_n,2}$$

■

Lemma F.13 (Lemma 1 from (Bach, 2010))**.** *Let $g : \mathbb{R} \to \mathbb{R}$ be a convex three times differentiable function such that for all $t \in \mathbb{R}$, $|g'''(t)| \leqslant M g''(t)$ for some $M \geqslant 0$. Then, for all $t \geqslant 0$ we have*

$$\frac{g''(0)}{M^2}\left\{\exp(-Mt) + Mt - 1\right\} \leqslant g(t) - g(0) - g'(0)t \leqslant \frac{g''(0)}{M^2}\left\{\exp(Mt) + Mt - 1\right\}.$$

Lemma F.14. *For $t \geqslant 0$ we have $\exp(-t) + t - 1 \geqslant \frac{1}{2}t^2 - \frac{1}{6}t^3$.*

Proof of Lemma F.14. For $t \geqslant 0$, consider the function $f(t) = \exp(-t) + t^3/6 - t^2/2 + t - 1$. The statement is equivalent to $f(t) \geqslant 0$ for $t \geqslant 0$. It follows that $f(0) = 0$, $f'(0) = 0$, and $f''(t) = \exp(-t) + t - 1 \geqslant 0$ so that f is convex. Therefore $f(t) \geqslant f(0) + tf'(0) = 0$. ■

Lemma F.15. *The logistic link function satisfies* $|\Lambda(t + t_0) - \Lambda(t_0)| \leqslant \Lambda'(t_0)\{\exp(|t|) - 1\}$. *If* $|t| \leqslant 1$ *we have* $\exp(|t|) - 1 \leqslant 2|t|$.

Proof. Note that $|\Lambda''(s)| \leqslant \Lambda'(s)$ for all $s \in \mathbb{R}$. So that $-1 \leqslant \frac{d}{ds}\log(\Lambda'(s)) = \frac{\Lambda''(s)}{\Lambda'(s)} \leqslant 1$. Suppose $s \geqslant 0$. Therefore

$$-s \leqslant \log(\Lambda'(s + t_0)) - \log(\Lambda'(t_0)) \leqslant s.$$

In turn this implies $\Lambda'(t_0)\exp(-s) \leqslant \Lambda'(s + t_0) \leqslant \Lambda'(t_0)\exp(s)$. Integrating one more time from 0 to t,

$$\Lambda'(t_0)\{1 - \exp(-t)\} \leqslant \Lambda(t + t_0) - \Lambda(t_0) \leqslant \Lambda'(t_0)\{\exp(t) - 1\}.$$

The first result follows by noting that $1-\exp(-t) \leqslant \exp(t)-1$. The second follows by verification. ■

References

Abadie, A. (2003): "Semiparametric Instrumental Variable Estimation of Treatment Response Models," *Journal of Econometrics*, 113, 231–263.

Andrews, D. W. (1994): "Empirical process methods in econometrics," *Handbook of Econometrics*, 4, 2247–2294.

Angrist, J. D., and J.-S. Pischke (2008): *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Bach, F. (2010): "Self-concordant analysis for logistic regression," *Electronic Journal of Statistics*, 4, 384–414.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012): "Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain," *Econometrica*, 80, 2369–2429, Arxiv, 2010.

Belloni, A., and V. Chernozhukov (2011a): "ℓ_1-Penalized Quantile Regression for High Dimensional Sparse Models," *Annals of Statistics*, 39(1), 82–130.

——— (2011b): "ℓ_1-penalized quantile regression in high-dimensional sparse models," *Ann. Statist.*, 39(1), 82–130.

——— (2013): "Least Squares After Model Selection in High-dimensional Sparse Models," *Bernoulli*, 19(2), 521–547, ArXiv, 2009.

Belloni, A., V. Chernozhukov, and C. Hansen (2010): "LASSO Methods for Gaussian Instrumental Variables Models," 2010 arXiv:[math.ST], http://arxiv.org/abs/1012.1297.

——— (2011): "Inference on Treatment Effects After Selection Amongst High-Dimensional Controls," *ArXiv*, forthcoming, The Review of Economic Studies.

——— (2013): "Inference for High-Dimensional Sparse Econometric Models," *Advances in Economics and Econometrics. 10th World Congress of Econometric Society. August 2010*, III, 245–295.

Belloni, A., V. Chernozhukov, and K. Kato (2013): "Uniform Post Selection Inference for LAD Regression Models," *arXiv preprint arXiv:1304.0282*.

Belloni, A., V. Chernozhukov, and L. Wang (2011): "Square-Root-LASSO: Pivotal Recovery of Sparse Signals via Conic Programming," *Biometrika*, 98(4), 791–806, Arxiv, 2010.

Belloni, A., V. Chernozhukov, and Y. Wei (2013): "Honest Confidence Regions for Logistic Regression with a Large Number of Controls," *arXiv preprint arXiv:1304.3969*.

Benjamin, D. J. (2003): "Does 401(k) eligibility increase saving? Evidence from propensity score subclassification," *Journal of Public Economics*, 87, 1259–1290.

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009): "Simultaneous analysis of Lasso and Dantzig selector," *Annals of Statistics*, 37(4), 1705–1732.

Candès, E., and T. Tao (2007): "The Dantzig selector: statistical estimation when p is much larger than n," *Ann. Statist.*, 35(6), 2313–2351.

Cattaneo, M. D. (2010): "Efficient semiparametric estimation of multi-valued treatment effects under ignorability," *Journal of Econometrics*, 155(2), 138–154.

Chamberlain, G., and G. W. Imbens (2003): "Nonparametric applications of Bayesian inference," *Journal of Business & Economic Statistics*, 21(1), 12–18.

Chernozhukov, V., D. Chetverikov, and K. Kato (2012): "Gaussian approximation of suprema of empirical processes," *ArXiv e-prints*.

Chernozhukov, V., and C. Hansen (2004): "The impace of 401(k) participation on the wealth distribution: An instrumental quantile regression analysis," *Review of Economics and Statistics*, 86(3), 735–751.

Dudley, R. M. (1999): *Uniform central limit theorems*, vol. 63 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge.

Engen, E. M., and W. G. Gale (2000): "The Effects of 401(k) Plans on Household Wealth: Differences Across Earnings Groups," Working Paper 8032, National Bureau of Economic Research.

Engen, E. M., W. G. Gale, and J. K. Scholz (1996): "The Illusory Effects of Saving Incentives on Saving," *Journal of Economic Perspectives*, 10, 113–138.

Fan, J., and R. Li (2001): "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of American Statistical Association*, 96(456), 1348–1360.

Farrell, M. (2013): "Robust Inference on Average Treatment Effects with Possibly More Covariates than Observations," .

Frank, I. E., and J. H. Friedman (1993): "A Statistical View of Some Chemometrics Regression Tools," *Technometrics*, 35(2), 109–135.

Hahn, J. (1997): "Bayesian bootstrap of the quantile regression estimator: a large sample study," *Internat. Econom. Rev.*, 38(4), 795–808.

——— (1998): "On the role of the propensity score in efficient semiparametric estimation of average treatment effects," *Econometrica*, pp. 315–331.

Hansen, B. E. (1996): "Inference when a nuisance parameter is not identified under the null hypothesis," *Econometrica*, 64(2), 413–430.

Heckman, J., and E. J. Vytlacil (1999): "Local instrumental variables and latent variable models for identifying and bounding treatment effects," *Proc. Natl. Acad. Sci. USA*, 96(8), 4730–4734 (electronic).

Hong, H., and D. Nekipelov (2010): "Semiparametric efficiency in nonlinear LATE models," *Quantitative Economics*, 1, 279–304.

Huang, J., J. L. Horowitz, and S. Ma (2008): "Asymptotic properties of bridge estimators in sparse high-dimensional regression models," *The Annals of Statistics*, 36(2), 587613.

Huang, J., J. L. Horowitz, and F. Wei (2010): "Variable selection in nonparametric additive models," *Ann. Statist.*, 38(4), 2282–2313.

Imbens, G. W., and J. D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–475.

Jing, B.-Y., Q.-M. Shao, and Q. Wang (2003): "Self-normalized Cramr-type large deviations for independent random variables," *Ann. Probab.*, 31(4), 2167–2215.

Kato, K. (2011): "Group Lasso for high dimensional sparse quantile regression models," Preprint, ArXiv.

Kline, P., and A. Santos (2012): "A Score Based Approach to Wild Bootstrap Inference," *Journal of Econometric Methods*, 1(1), 23–41.

Koenker, R. (1988): "Asymptotic Theory and Econometric Practice," *Journal of Aplpied Econometrics*, 3, 139–147.

——— (2005): *Quantile regression*. Cambridge university press.

Kosorok, M. R. (2008): *Introduction to Empirical Processes and Semiparametric Inference*, Series in Statistics. Springer, Berlin.

Leeb, H., and B. M. Pötscher (2008a): "Can one estimate the unconditional distribution of post-model-selection estimators?," *Econometric Theory*, 24(2), 338–376.

——— (2008b): "Recent developments in model selection and related areas," *Econometric Theory*, 24(2), 319–322.

Mammen, E. (1993): "Bootstrap and wild bootstrap for high dimensional linear models," *The Annals of Statistics*, pp. 255–285.

Meinshausen, N., and B. Yu (2009): "Lasso-type recovery of sparse representations for high-dimensional data," *Annals of Statistics*, 37(1), 2246–2270.

Newey, W. K. (1997): "Convergence Rates and Asymptotic Normality for Series Estimators," *Journal of Econometrics*, 79, 147–168.

Neyman, J. (1979): "$C(\alpha)$ tests and their use," *Sankhya*, 41, 1–21.

Poterba, J. M., S. F. Venti, and D. A. Wise (1994): "401(k) Plans and Tax-Deferred savings," in *Studies in the Economics of Aging*, ed. by D. A. Wise. Chicago, IL: University of Chicago Press.

——— (1995): "Do 401(k) Contributions Crowd Out Other Personal Saving?," *Journal of Public Economics*, 58, 1–32.

——— (1996): "Personal Retirement Saving Programs and Asset Accumulation: Reconciling the Evidence," Working Paper 5599, National Bureau of Economic Research.

——— (2001): "The Transition to Personal Accounts and Increasing Retirement Wealth: Macro and Micro Evidence," Working Paper 8610, National Bureau of Economic Research.

Pötscher, B. (2009): "Confidence Sets Based on Sparse Estimators Are Necessarily Large," *Sankhya*, 71-A, 1–18.

Resnick, S. I. (1987): *Extreme values, regular variation, and point processes*, vol. 4 of *Applied Probability. A Series of the Applied Probability Trust*. Springer-Verlag, New York.

Romano, J. P., and A. M. Shaikh (2012): "On the uniform asymptotic validity of subsampling and the bootstrap," *The Annals of Statistics*, 40(6), 2798–2822.

Rothe, C., and S. Firpo (2013): "Semiparametric Estimation and Inference Using Doubly Robust Moment Conditions," Discussion paper, NYU preprint.

Tibshirani, R. (1996): "Regression shrinkage and selection via the Lasso," *J. Roy. Statist. Soc. Ser. B*, 58, 267–288.

Tsybakov, A. B. (2009): *Introduction to nonparametric estimation*. Springer.

van de Geer, S. A. (2008): "High-dimensional generalized linear models and the lasso," *Annals of Statistics*, 36(2), 614–645.

van der Vaart, A. W. (1998): *Asymptotic Statistics*. Cambridge University Press.

van der Vaart, A. W., and J. A. Wellner (1996): *Weak Convergence and Empirical Processes*. Springer Series in Statistics.

Vytlacil, E. J. (2002): "Independence, Monotonicity, and Latent Index Models: An Equivalence Result," *Econometrica*, 70, 331–341.

Wasserman, L. (2006): *All of nonparametric statistics*. Springer New York.

Wooldridge, J. M. (2010): *Econometric Analysis of Cross Section and Panel Data*. Cambridge, Massachusetts: The MIT Press, second edn.

Zou, H. (2006): "The Adaptive Lasso And Its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429.

Table 1: Estimates and standard errors of average effects

Specification			Net Total Financial Assets		Total Wealth	
Series approximation	Dimension	Selection	LATE	LATE-T	LATE	LATE-T
Indicators	20	N	11833	16120	8972	12500
			(1638)	(2224)	(2692)	(3572)
			{1685}	{2238}	{2597}	{3376}
Indicators	20	Y	14658	16895	13717	13711
			(1676)	(2265)	(2726)	(3645)
			{1685}	{2306}	{2640}	{3471}
Indicators plus interactions	167	N	11856	16216	9996	12131
			(1632)	(2224)	(2675)	(3428)
			{1683}	{2198}	{2767}	{3385}
Indicators plus interactions	167	Y	14653	16969	12926	13391
			(1693)	(2316)	(2709)	(3715)
			{1680}	{2307}	{2711}	{3700}
B-splines	27	N	11558	15572	8537	11431
			(1573)	(2140)	(2625)	(3502)
			{1516}	{2194}	{2499}	{3347}
B-splines	27	Y	11795	15956	8826	12016
			(1632)	(2172)	(2674)	(3520)
			{1513}	{2086}	{2751}	{3636}
B-splines plus interactions	323	N	40333	86032	31021	58152
			(17526)	(46158)	(12449)	(32123)
			{17092}	{47065}	{11692}	{33342}
B-splines plus interactions	323	Y	12337	16099	9706	10042
			(1629)	(2227)	(2649)	(3586)
			{1618}	{2120}	{2627}	{3468}

Notes: The sample is drawn from the 1991 SIPP and consists of 9,915 observations. All the specifications control for age, income, family size, education, marital status, two-earner status, defined benefit pension status, IRA participation status, and home ownership status. Indicators specification uses a linear term for family size, 5 categories for age, 4 categories for education, and 7 categories for income. B-splines specification uses cubic b-splines with 1, 1, 3, and 5 interior knots for family size, education, age, and income, respectively. Marital status, two-earner status, defined benefit pension status, IRA participation status, and home ownership status are included as indicators in all the specifications. Analytical standard errors are given in parentheses. Wild bootstrap standard errors based on 500 repetitions with Mammen (1993) weights are given in braces.

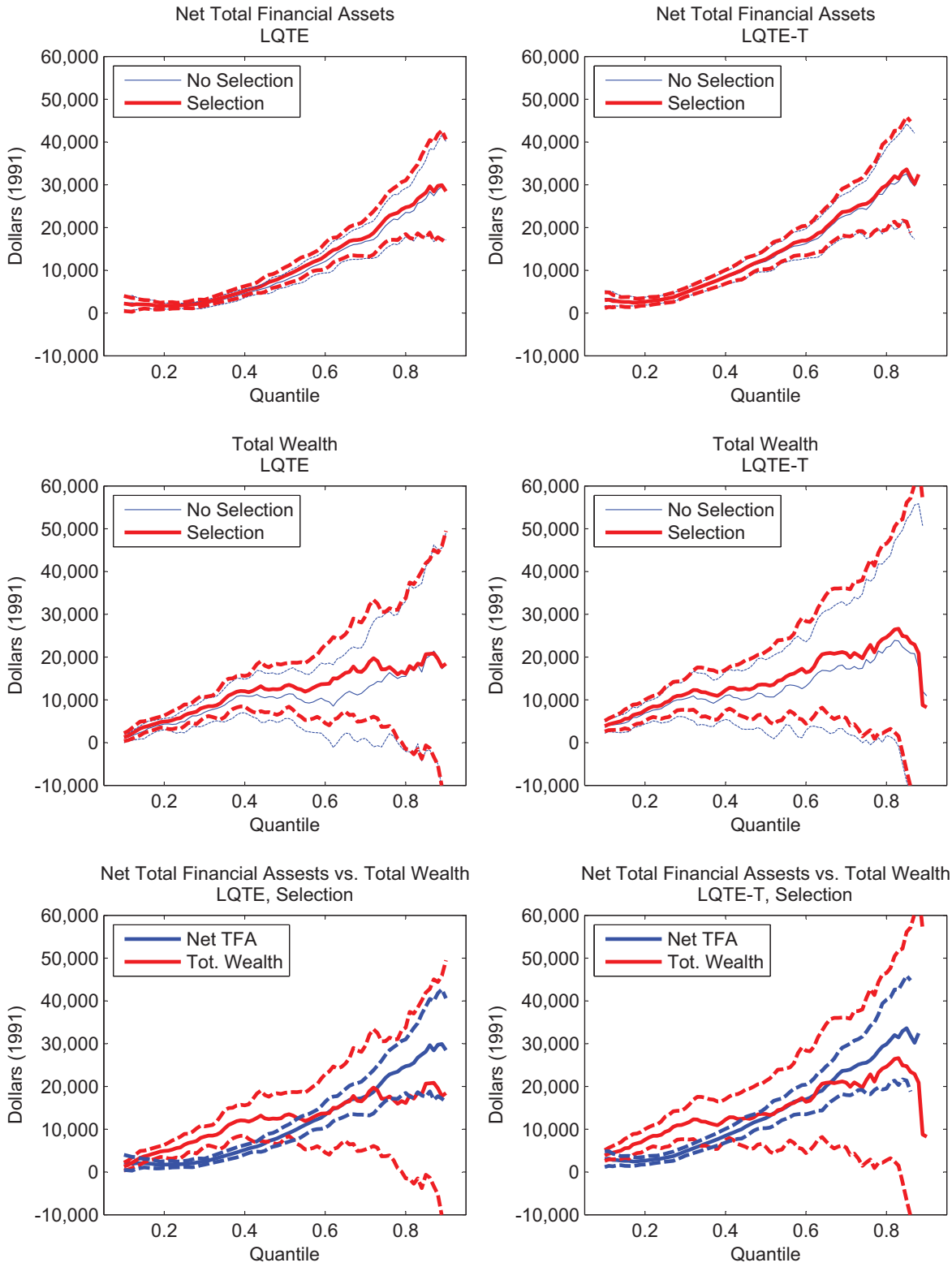

FIGURE 1. LQTE and LQTE-T estimates based on indicators specification.

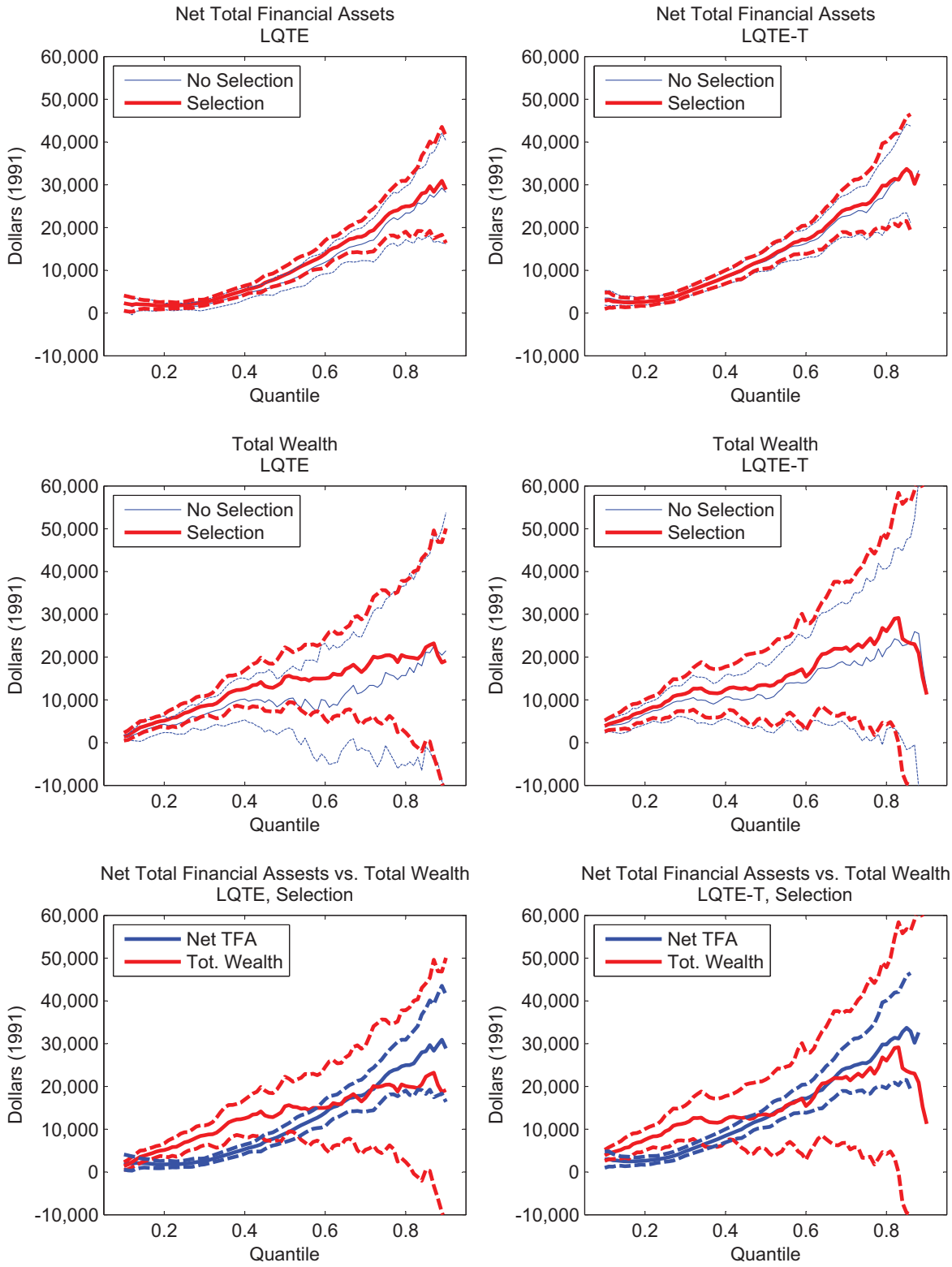

FIGURE 2. LQTE and LQTE-T estimates based on indicators plus interactions specification.

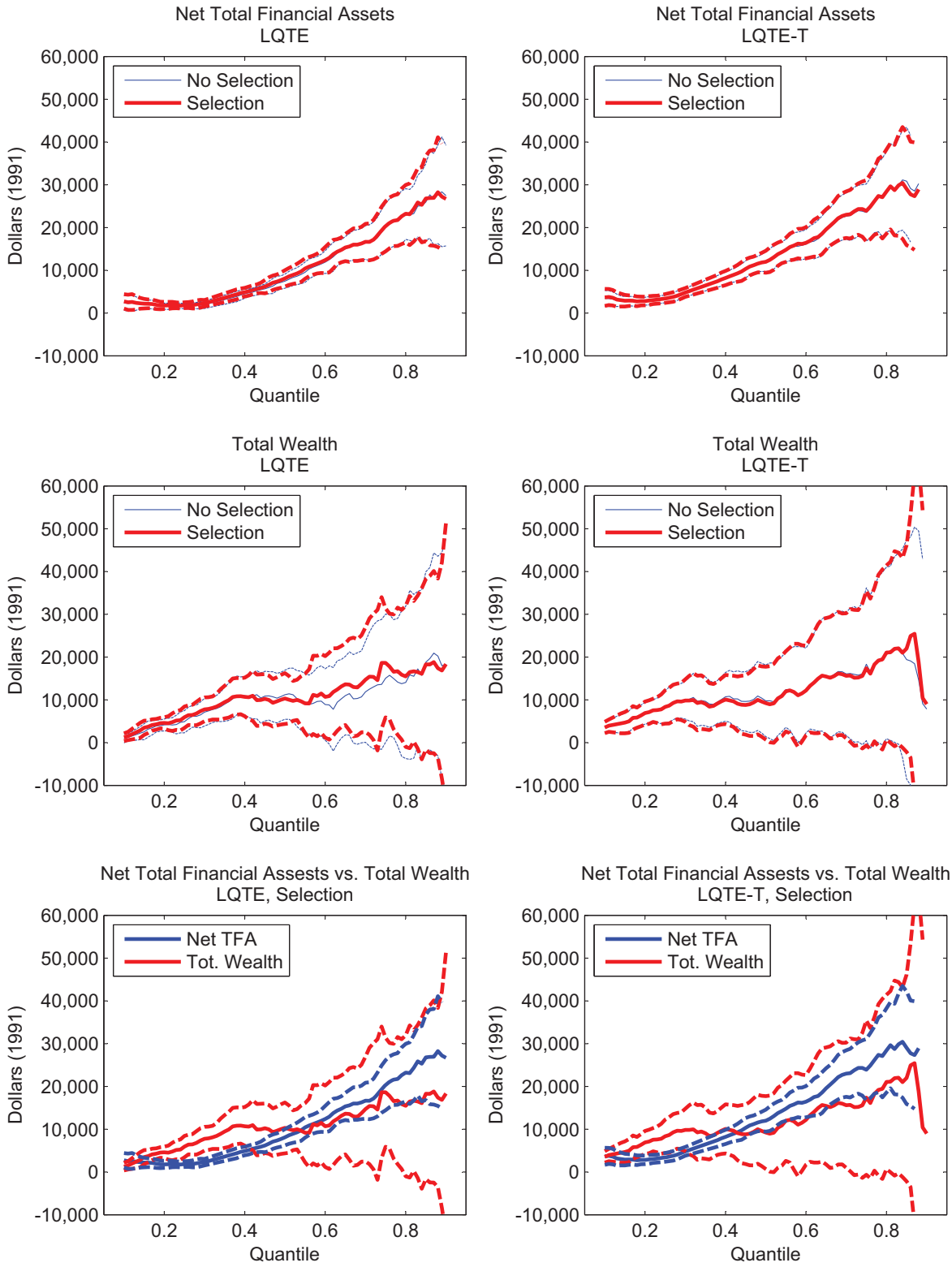

FIGURE 3. LQTE and LQTE-T estimates based on b-spline specification.

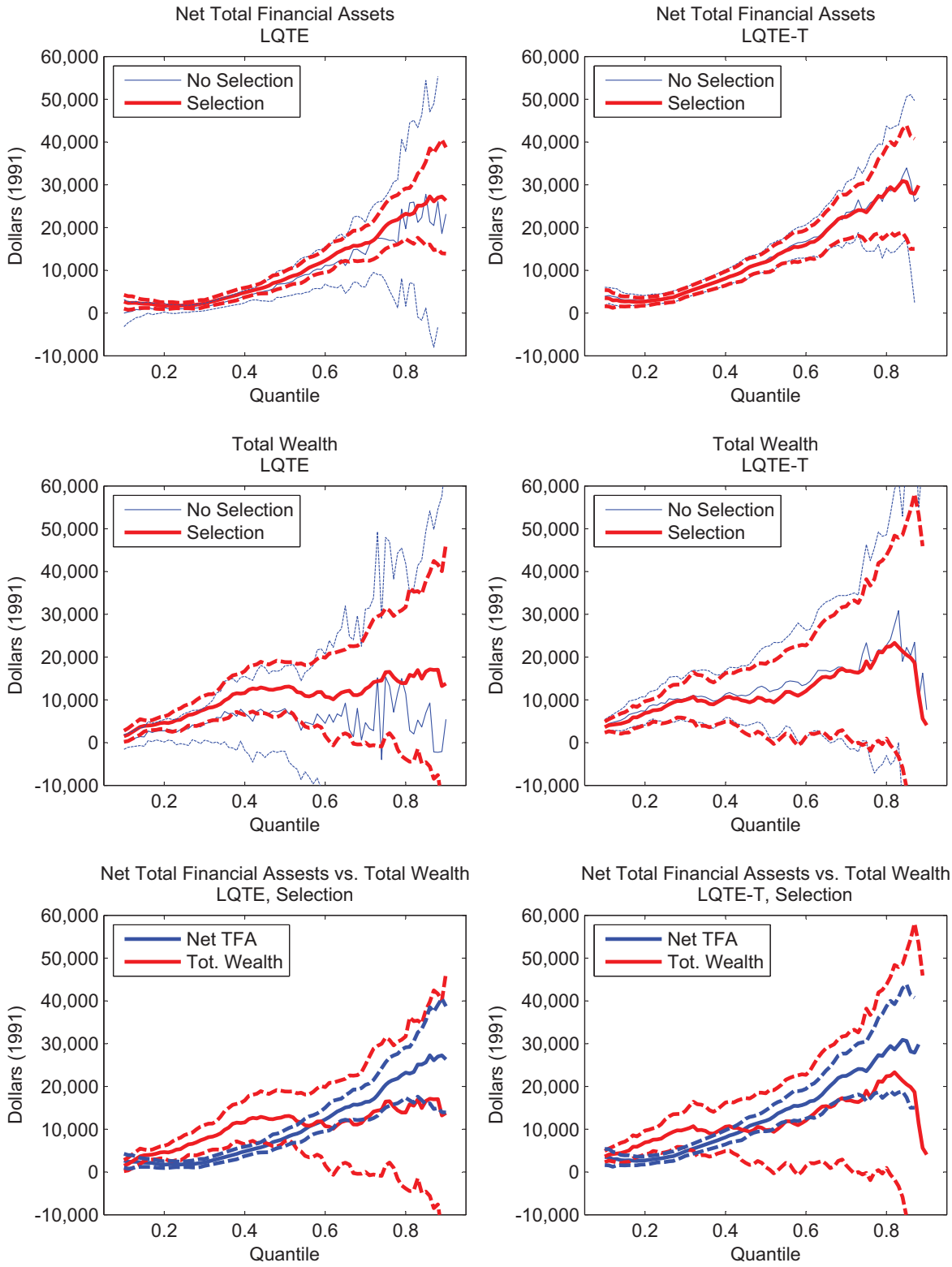

FIGURE 4. LQTE and LQTE-T estimates based on b-spline plus interactions specification.