cemmap
centre for microdata methods and practice

Program evaluation with high-dimensional data

A. Belloni
V. Chernozhukov
I. Fernandez-Val
C. Hansen

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP33/14

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 this paper, we consider estimation of general modern moment-condition problems in econometrics in a data-rich environment where there may be many more control variables available than there are observations. The framework we consider allows for a continuum of target parameters and for Lasso-type or Post-Lasso type methods to be used as estimators of a continuum of high-dimensional nuisance functions. As an important leading example of this environment, we first provide detailed results on estimation and inference for relevant treatment effects, such as local average and quantile treatment effects. The setting we work in is designed expressly to handle many control variables, *endogenous* receipt of treatment, *heterogeneous* treatment effects, and possibly *function-valued* outcomes. To make informative inference possible, we assume that key 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 by a small number of the control variables whose identities are unknown to the researcher. This condition permits estimation and inference to proceed after data-driven selection of control variables. 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 orthogonal moment conditions. We illustrate the use of the proposed methods with an application to estimating the effect of 401(k) participation on accumulated assets.

We generalize the results from the treatment effects setting to accommodate more general moment condition models in a second part of the paper. In particular, we establish a functional central limit theorem for robust estimators of a continuum of target parameters that holds uniformly in a wide range of data generating processes (dgp's), i.e. over dgp's $P \in \mathcal{P}$ where $\mathcal{P}$ include dgp's where perfect model selection is theoretically impossible. We prove that the use of orthogonal moment conditions is key to achieving uniform validity. We also establish a functional central limit theorem for the multiplier bootstrap that resamples first-order approximations to the proposed estimators that holds uniformly over $P \in \mathcal{P}$. We propose a notion of differentiability, together with a functional delta method, that allows us to derive approximate distributions for smooth functionals of a continuum of target parameters that hold uniformly over $P \in \mathcal{P}$ and to establish the validity of the multiplier bootstrap for approximating these distributions uniformly over $P \in \mathcal{P}$. Finally, we establish rate and consistency results for continua of Lasso or Post-Lasso type estimators for continua of (nuisance) regression functions and provide practical, theoretically justified choices for the penalty parameters used in these methods. Each of these results is new and of independent interest.

Date: First version: April 2013. This version: August 8, 2014. We gratefully acknowledge research support from the NSF. We are grateful to Andres Santos, Alberto Abadie, Stephane Bonhomme, Matias Cattaneo, Jinyong Hahn, Michael Jansson, Toru Kitagawa, Roger Koenker, Simon Lee, Yuan Liao, Oliver Linton, Whitney Newey, Adam Rosen, Yuan Liao, and seminar participants at the Cornell-Princeton Conference on "Inference on Non-Standard Problems", Winter 2014 ES meeting, Semi-Plenary Lecture at of ES Summer Meeting 2014, University of Montreal, 2013 Summer NBER Institute, the UIUC, UCL, and Bristol Econometric Study Group for helpful comments. We are especially grateful to Anders Santos for many useful comments.

Key words and phrases. endogeneity, local average and quantile treatment effects, instruments, local effects of treatment on the treated, propensity score, Lasso, inference after model selection, moment condition models, moment condition models with a continuum of target parameters, Lasso and Post-Lasso with functional response data.

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 as exogenous conditional on the right set of control variables or simply controls. 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 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 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 controls 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 controls from a potentially large set 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 controls, $f(X)$, including raw variables, X, and transformations of these variables such as powers, b-splines, or interactions.[2] We allow for fully *heterogeneous* treatment effects and thus

[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.

[2]High-dimensional $f(X)$ typically occurs in either of two ways. First, the baseline set of conditioning variables itself may be large so $f(X) = X$. Second, X may be low-dimensional, but one may wish to entertain many nonlinear transformations of X in forming $f(X)$. In the second case, one might prefer to refer to X as the controls and $f(X)$

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 *endogenous* case where identification is obtained through the use of an instrumental variable, but all results carry through to the *exogenous* case where the treatment is taken as exogenous after conditioning on sufficient controls by simply replacing the instrument with the treatment variable in the estimation and inference methods and in the formal results.

The methodology for estimating treatment effects we consider allows for cases where the number of potential controls, $p := \dim f(X)$, is much larger 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 the conditional expectations $\mathrm{E}_P[D|X]$, $\mathrm{E}_P[Z|X]$, and $\mathrm{E}_P[Y|X]$ 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 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 can be 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 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 result in 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); and Belloni, Chernozhukov, and Hansen (2013; 2014).

---

as something else, such as technical regressors. For simplicity of exposition and as the formal development in the paper is agnostic about the source of high-dimensionality, we call the variables in $f(X)$ controls or control variables in either case.

A main contribution of this 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 Belloni, Chernozhukov, and Hansen (2010) and Belloni, Chen, Chernozhukov, and Hansen (2012), who were the first to demonstrate in a highly specialized context, that valid inference can proceed following model selection, allowing for model selection mistakes, under two conditions. We formulate and extend these two conditions to a rather general moment-condition framework (e.g., Hansen (1982), and Hansen and Singleton (1982)) as follows. First, estimation should be based upon "orthogonal" moment conditions that are first-order insensitive to changes in the values of nuisance parameters that will be estimated using high-dimensional methods. Specifically, if the target parameter value α_0 is identified via the moment condition

$$\mathrm{E}_P \psi(W, \alpha_0, h_0) = 0, \tag{1.1}$$

where h_0 is a function-valued nuisance 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

$$\partial_h [\mathrm{E}_P \psi(W, \alpha_0, h)]_{h=h_0} = 0, \tag{1.2}$$

where ∂_h is a functional derivative operator with respect to h restricted to directions of possible deviations of estimators of h_0 from h_0. Second, one needs to ensure that the model selection mistakes occuring in estimation of nuisance parameters are uniformly "moderately" small with respect to the underlying model.

The orthogonality condition embodied in (1.2) 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. See also Newey (1990), Andrews (1994b), Newey (1994), Robins and Rotnitzky (1995), and Linton (1996) for the use of this condition in semi-parametric problems. To the best of our knowledge, Belloni, Chernozhukov, and Hansen (2010) and Belloni, Chen, Chernozhukov, and Hansen (2012) were the first to use orthogonality (1.2) to expressly address the question of the uniform post-selection inference, either in high-dimensional settings with $p \gg n$ or in low-dimensional settings (with $p \ll n$). They applied it to a very specialized setup of the linear instrumental variables model with many instruments where the nuisance function h_0 is the optimal instrument estimated by Lasso or Post-Lasso methods and α_0 is the coefficient of the endogenous regressor. Belloni, Chernozhukov, and Hansen (2013) and Belloni, Chernozhukov, and Hansen (2014) also exploited this approach to develop a double-selection method that yields valid post-selection inference on the parameters of the linear part of a partially linear model and on average treatment effects when the treatment is binary and *exogenous* conditional on controls in both the $p \gg n$ and the $p \ll n$ setting;[3] see also Farrell (2013) who

[3]Note that these results as well as results of this paper on the uniform post-selection inference in moment-condition problems are new for either $p \ll n$ or $p \gg n$ settings. The results also apply to arbitrary model selection devices that

extended this method to estimation of average treatment effects when the treatment is multivalued and exogenous conditional on controls using group penalization for selection. In this paper, we establish that building estimators based upon moment conditions with the orthogonality condition (1.2) holding ensures that crude estimation of h_0 via post-selection or other regularization methods has an asymptotically negligible effect on the estimation of α_0 in general frameworks, which results in a regular, root-n consistent estimator of α_0, uniformly with respect to the underlying model.

In the general endogenous treatment effects setting, moment conditions satisfying (1.2) can be found as efficient influence functions for certain reduced form parameters as in Hahn (1998). We illustrate how orthogonal moment conditions coupled with methods developed for forecasting in high-dimensional approximately 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. An important feature of our main theoretical results is that they cover the use of variable selection for *functional response data* using ℓ_1-penalized methods. Functional response data arises, for example, when one is interested in the 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 or the LATE.

A second main contribution of this paper is providing a general set of results for uniformly valid estimation and inference in modern moment-condition problems in econometrics allowing for both smooth and non-smooth moment functions. We prove that the use of orthogonal moment conditions is key to achieving uniform validity. In the general framework we consider, we may have a continuum of target parameters identified via a continuum of moment conditions that involve a continuum of nuisance functions that will be estimated via modern high-dimensional methods such as Lasso or Post-Lasso or their variants. These results contain the first set of results on treatement effects relevant for program evaluation, particularly the distributional and quantile effects, as a leading special case. These results are also immediately useful in many other contexts such as nonseparable quantile models as in Chernozhukov and Hansen (2005), Chernozhukov and Hansen (2006), Chesher (2003), and Imbens and Newey (2009); semiparametric and partially identified models as in Escanciano and Zhu (2013); and many others. In our results, we first establish a functional central limit theorem for the continuum of target parameters and show that this functional central limit theorem holds uniformly in a wide range of data-generating processes P with approximately sparse continua of nuisance functions. Second, we establish a functional central limit theorem for the multiplier boostrap that resamples the first order approximations to the

---

are able to select good sparse approximating models; and "moderate" model selection errors are explicitly allowed in the paper.

standardized estimators and demonstrate its uniform-in-P validity. These uniformity results build upon and complement those given in Romano and Shaikh (2012) for the empirical bootstrap. Third, we establish a functional delta method for smooth functionals of the continuum of target parameters and a functional delta method for the multiplier bootstrap of these smooth functionals both of which hold uniformly in P using an appropriately strengthened notion of Hadamard differentiability. All of these results are new and are of independent interest outside of the application in this paper.[4]

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).[5] 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 broader set of controls than have previously been considered. We find that 401(k) participation has a moderate 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 similar to those in Chernozhukov and Hansen (2004) despite allowing for a much richer set of controls.

We organize the rest of the paper as follows. Section 2 introduces the structural parameters for policy evaluation, relates these parameters to reduced form functions, and gives conditions under which the structural parameters have a causal interpretation. Section 3 describes a three step procedure to estimate and make inference on the structural parameters and functionals of these parameters, and Section 4 provides asymptotic theory. Section 5 generalizes the setting and results to moment-condition problems with a continuum of structural parameters and a continuum of reduced form functions. Section 6 derives general asymptotic theory for the Lasso and post-Lasso estimators for functional response data used in the estimation of the reduced form functions. Section 7 presents the empirical application. We gather the notation, the proofs of all the results and additional technical results in Appendices A–G. A supplementary appendix provides implementation details for the empirical application and a Monte Carlo simulation.

[4]These results build upon the work of Belloni and Chernozhukov (2011) 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 (2011); 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.

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

2. The Setting and The Target Parameters

2.1. **Observables and Reduced Form Parameters.** The observed random variables consist of $((Y_u)_{u\in\mathcal{U}}, X, Z, D)$. 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 a binary 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, we argue that 401(k) eligibility can be considered exogenous only after conditioning on income and other individual characteristics in the empirical application. The notions of exogeneity and endogeneity we employ are standard, but we state them in Section 2.4 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

$$\alpha_V(z) := \mathrm{E}_P[g_V(z, X)] \quad \text{and} \quad \gamma_V := \mathrm{E}_P[V], \tag{2.1}$$

where $z = 0$ or $z = 1$ are the fixed values of Z.[6] The function g_V maps $\mathcal{Z}\mathcal{X}$, the support of the vector (Z, X), to the real line $\mathbb{R}$ and is defined as

$$g_V(z, x) := \mathrm{E}_P[V|Z = z, X = x]. \tag{2.2}$$

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 modern methods to deal with high-dimensional data coupled with orthogonal estimating equations. The orthogonality property is crucial for dealing with the "non-regular" nature of

[6]The expectation that defines $\alpha_V(z)$ is well-defined under the support condition $0 < \mathrm{P}_P(Z = 1 \mid X) < 1$ a.s. We impose this condition in Assumption 2.1 and Assumption 4.1.

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 (2.1) are key because the structural parameters of interest are functionals of these elementary objects. The local average structural function (LASF) defined as

$$\theta_{Y_u}(d) = \frac{\alpha_{\mathbf{1}_d(D)Y_u}(1) - \alpha_{\mathbf{1}_d(D)Y_u}(0)}{\alpha_{\mathbf{1}_d(D)}(1) - \alpha_{\mathbf{1}_d(D)}(0)}, \quad d \in \{0,1\} \tag{2.3}$$

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 by variation in the instrument in the treated and non-treated states, under standard assumptions; see, e.g. Abadie (2002; 2003). The local average treatment effect (LATE) of Imbens and Angrist (1994) corresponds to the difference of the two values of the LASF:

$$\theta_{Y_u}(1) - \theta_{Y_u}(0). \tag{2.4}$$

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 simplify to

$$\theta_{Y_u}(z) = \alpha_{Y_u}(z), \quad \theta_{Y_u}(1) - \theta_{Y_u}(0) = \alpha_{Y_u}(1) - \alpha_{Y_u}(0). \tag{2.5}$$

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

$$\alpha_{Y_u}(z), \quad \alpha_{Y_u}(1) - \alpha_{Y_u}(0).$$

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 (2.3) and (2.4) 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 $u \in \mathbb{R}$. In this case, the family of effects

$$(\theta_{Y_u}(1) - \theta_{Y_u}(0))_{u \in \mathbb{R}}, \tag{2.6}$$

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

$$\theta_Y^{\leftarrow}(\tau, d) := \inf\{u \in \mathbb{R} : \theta_{Y_u}(d) \geqslant \tau\}, \tag{2.7}$$

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

$$(\theta_Y^{\leftarrow}(\tau,1) - \theta_Y^{\leftarrow}(\tau,0))_{\tau\in(0,1)}. \tag{2.8}$$

The LQTE identify the differences of quantiles between the distribution of the outcome in the treated and non-treated states for compliers.

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 these effects is the local average structural function on the treated (LASF-T) which is defined by its two values:

$$\vartheta_{Y_u}(d) = \frac{\gamma_{\mathbf{1}_d(D)Y_u} - \alpha_{\mathbf{1}_d(D)Y_u}(0)}{\gamma_{\mathbf{1}_d(D)} - \alpha_{\mathbf{1}_d(D)(0)}}, \quad d \in \{0,1\}. \tag{2.9}$$

The LASF-T identifies 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 the difference of two values of the LASF-T:

$$\vartheta_{Y_u}(1) - \vartheta_{Y_u}(0). \tag{2.10}$$

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. The distinction between LATE-T and LATE can be relevant; for example, in our empirical application the estimated LATE-T and LATE are substantially different.

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 simplify to

$$\vartheta_{Y_u}(1) = \frac{\gamma_{\mathbf{1}_1(D)Y_u}}{\gamma_{\mathbf{1}_1(D)}}, \quad \vartheta_{Y_u}(0) = \frac{\gamma_{\mathbf{1}_0(D)Y_u} - \alpha_{Y_u}(0)}{\gamma_{\mathbf{1}_0(D)} - 1}, \quad \vartheta_{Y_u}(1) - \vartheta_{Y_u}(0); \tag{2.11}$$

and we can use our results to provide estimation and inference methods 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

$$(\vartheta_{Y_u}(1) - \vartheta_{Y_u}(0))_{u\in\mathbb{R}} \tag{2.12}$$

then describes the LDTE-T. We can also use the quantile left-inverse transform of the curve $u \mapsto \vartheta_{Y_u}(d)$, namely $\vartheta_Y^{\leftarrow}(\tau,d) := \inf\{u \in \mathbb{R} : \vartheta_{Y_u}(d) \geqslant \tau\}$, and define LQTE-T:

$$(\vartheta_Y^{\leftarrow}(\tau,1) - \vartheta_Y^{\leftarrow}(\tau,0))_{\tau\in(0,1)}. \tag{2.13}$$

Under conditional exogeneity LQTE and LQTE-T reduce to the quantile treatment effects (QTE) and quantile treatment effects on the treated (QTE-T) (Koenker, 2005, Chap. 2).

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. We briefly recall these conditions, using the 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.

Exogeneity fails when D depends 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. Let Z be a binary instrument, such as an offer of participation, that generates potential participation decisions D_1 and D_0 under the instrument states 1 and 0, respectively. 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 $D = ZD_1 + (1-Z)D_0$. We assume that Z is assigned randomly with respect to potential outcomes and participation decisions conditional on X, i.e., $(Y_{u0}, Y_{u1}, D_0, D_1) \perp\!\!\!\perp Z \mid X$ a.s.

There are many causal quantities of interest for program evaluation. Chief among these are various structural averages: $d \mapsto \mathrm{E}_P[Y_{ud}]$, the causal ASF; $d \mapsto \mathrm{E}_P[Y_{ud} \mid D = 1]$, the causal ASF-T; $d \mapsto \mathrm{E}_P[Y_{ud} \mid D_1 > D_0]$, the causal LASF; and $d \mapsto \mathrm{E}_P[Y_{ud} \mid D_1 > D_0, D = 1]$, the causal LASF-T; as well as effects derived from them such as $\mathrm{E}_P[Y_{u1} - Y_{u0}]$, the causal ATE; $\mathrm{E}_P[Y_{u1} - Y_{u0} \mid D = 1]$, the causal ATE-T; $\mathrm{E}_P[Y_{u1} - Y_{u0} \mid D_1 > D_0]$, the causal LATE; and $\mathrm{E}_P[Y_{u1} - Y_{u0} \mid D_1 > D_0, D = 1]$, the causal LATE-T. 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 2.1 (Assumptions for Causal/Structural 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)* $\mathrm{P}_P(Z = 1 \mid X) \in (0,1)$*; (Monotonicity)* $\mathrm{P}_P(D_1 \geqslant D_0 \mid X) = 1$.

This condition due to Imbens and Angrist (1994) and Abadie (2003) 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 2.1. Therefore, our referring to them as structural/causal is justified under this condition.

Lemma 2.1 (Identification of Causal Effects). *Under Assumption 2.1, 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 to define the structural parameters in Section 2 are the expectations

$$\alpha_V(z) = \mathrm{E}_P[g_V(z, X)] \text{ and } \gamma_V = \mathrm{E}_P[V], \tag{3.1}$$

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 $\mathcal{V}_u$:

$$V \in \mathcal{V}_u := \{V_{uj}\}_{j=1}^5 := \{Y_u, \mathbf{1}_0(D)Y_u, \mathbf{1}_0(D), \mathbf{1}_1(D)Y_u, \mathbf{1}_1(D)\}. \tag{3.2}$$

It is clear that $g_V(z, X)$ will play an important role in estimating $\alpha_V(z)$. A related function that will also play an important role in forming a robust estimation strategy is the propensity score $m_Z : \mathcal{Z}\mathcal{X} \mapsto \mathbb{R}$ defined by

$$m_Z(z, x) := \mathrm{P}_P[Z = z|X = x]. \tag{3.3}$$

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

$$g_V(z, x) = \sum_{d=0}^{1} e_V(d, z, x) l_D(d, z, x), \tag{3.4}$$

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

$$e_V(d, z, x) := \mathrm{E}_P[V|D = d, Z = z, X = x] \text{ and} \tag{3.5}$$

$$l_D(d, z, x) := \mathrm{P}_P[D = d|Z = z, X = x]. \tag{3.6}$$

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 the relation (3.4) to estimate g_V. Given the resulting estimate of g_V and an estimate of m_Z obtained using high-dimensional methods to model the propensity score, we will then recover an estimate of $\alpha_V(z)$.

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 binary endogenous variable model with normal disturbances.[7] This strict nesting of coherent models is less easily guaranteed 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 affine aggregation of cross-products shown in (3.4). 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 find little difference between the two approaches in our empirical example.

In the rest of this section, we describe the estimation of the reduced-form and structural parameters. 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 orthogonal 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 the plug-in rule.

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

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

$$f(X) = (f_j(X))_{j=1}^p, \tag{3.7}$$

to approximate g_V and m_Z. Specifically, we use

$$g_V(z,x) =: \Lambda_V[f(z,x)'\beta_V] + r_V(z,x), \tag{3.8}$$

$$f(z,x) := ((1-z)f(x)', zf(x)')', \quad \beta_V := (\beta_V(0)', \beta_V(1)')', \tag{3.9}$$

and

$$m_Z(1,x) =: \Lambda_Z[f(x)'\beta_Z] + r_Z(x), \quad m_Z(0,x) = 1 - \Lambda_Z[f(x)'\beta_Z] - r_Z(x). \tag{3.10}$$

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 $m_Z(1,x)$.

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

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 variable is 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, the dictionary of controls, denoted by $f(X)$, can 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, indicators, polynomials, and various interactions as in, e.g., Chen (2007), Newey (1997), Tsybakov (2009), and Wasserman (2006). The functions f 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 (3.8)-(3.10) 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 $V \in \mathcal{V} := \{\mathcal{V}_u : u \in \mathcal{U}\}$,

$$\|\beta_V\|_0 + \|\beta_Z\|_0 \leqslant s. \tag{3.11}$$

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 $V \in \mathcal{V}$,

$$\{\mathrm{E}_P[r_V^2(Z,X)]\}^{1/2} + \{\mathrm{E}_P[r_Z^2(X)]\}^{1/2} \lesssim \sqrt{s/n}. \tag{3.12}$$

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

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

These conditions ensure that the functions g_V and m_Z are estimable at $o(n^{-1/4})$ rate and are used to derive asymptotic normality results for the structural and reduced-form parameter estimators. They could be relaxed through the use of sample splitting methods as in Belloni, Chen, Chernozhukov, and Hansen (2012).

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 small relative to the sample size.[8] 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 methods 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 the 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). We provide further discussion of these methods for estimation of a continuum of functions in Section 6, and we specify detailed implementation algorithms used in the empirical example 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

$$\widehat{\beta}_V \in \arg \min_{\beta \in \mathbb{R}^{\dim(\tilde{X})}} \Big(\mathbb{E}_n[M(\tilde{Y}, \tilde{X}'\beta)] + \frac{\lambda}{n} \|\widehat{\Psi}\beta\|_1 \Big), \tag{3.13}$$

where $\widehat{\Psi} = \text{diag}(\widehat{l}_1, \ldots, \widehat{l}_{\dim(\widetilde{X})})$ 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, $\widehat{l}_j$, $j = 1, ..., \dim(\widetilde{X})$, are selected to guarantee good theoretical properties of the method. We provide theoretical choices and further detail regarding the implementation in Section 6.[9] 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,

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

[8] For example, one would select a set of basis functions, $\{f_j(X)\}_{j=1}^{\infty}$, such as power series or splines and then use only the first $s \ll n$ terms in the basis under the assumption that $s^C/n \to 0$ for some number C whose value depends on the specific context in a standard nonparametric approach using series.

[9] We also provide a detailed description of the implementation we used in the empirical example in a supplementary appendix.

The Post-Lasso estimator is then a solution to

$$\tilde{\beta}_V \in \arg \min_{\beta \in \mathbb{R}^{\dim(\tilde{X})}} \Big(\mathbb{E}_n[M(\tilde{Y}, \tilde{X}'\beta)] : \beta_j = 0, j \notin \widehat{I}_V \Big). \tag{3.14}$$

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 $\bar{\beta}_V = \tilde{\beta}_V$, achieves the near oracle rate of convergence $\sqrt{(s \log p)/n}$ and maintains 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 estimator $\tilde{\beta}_Z$ are defined analogously to $\widehat{\beta}_V$ and $\tilde{\beta}_V$ using the data $(\tilde{Y}_i, \tilde{X}_i)_{i=1}^n = (Z_i, f(X_i))_{i=1}^n$. 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 $\bar{\beta}_Z = \tilde{\beta}_Z$, also achieves the near oracle rate of convergence $\sqrt{(s \log p)/n}$ and has other good theoretic properties. The estimator of $\widehat{m}_Z(0,X)$ is then formed as $1 - \widehat{m}_Z(1,X)$.

Strategy 2. The second strategy we consider involves modeling and estimating m_Z as above via (3.10) while modeling g_V through its disaggregation into the parts e_V and l_D via (3.4). We model and estimate each of the unknown parts of e_V and l_D using the same approach as in Strategy 1.[10] Specifically, we model the conditional expectation of V given D, Z, and X by

$$e_V(d,z,x) =: \Gamma_V[f(d,z,x)'\theta_V] + \varrho_V(d,z,x), \tag{3.15}$$

$$f(d,z,x) := ((1-d)f(z,x)', df(z,x)')', \tag{3.16}$$

$$\theta_V := (\theta_V(0,0)', \theta_V(0,1)', \theta_V(1,0)', \theta_V(1,1)')'. \tag{3.17}$$

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

$$l_D(1,z,x) =: \Gamma_D[f(z,x)'\theta_D] + \varrho_D(z,x), \tag{3.18}$$

$$l_D(0,z,x) = 1 - \Gamma_D[f(z,x)'\theta_D] - \varrho_D(z,x), \tag{3.19}$$

$$f(z,x) := ((1-z)f(x)', zf(x)')', \tag{3.20}$$

$$\theta_D := (\theta_D(0)', \theta_D(1)')'. \tag{3.21}$$

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 again to be known link functions from the set $\mathcal{L} = \{\mathrm{Id}, \Phi, 1-\Phi, \Lambda_0, 1-\Lambda_0\}$ defined following equation (3.10).

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

$$\|\theta_V\|_0 + \|\theta_D\|_0 + \|\beta_Z\|_0 \leqslant s. \tag{3.22}$$

[10] 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 $d = d'$.

That is, there are at most $s = s_n \ll n$ components of θ_V, θ_D, and β_Z with nonzero values 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

$$\{\mathrm{E}_P[\varrho_V^2(D,Z,X)]\}^{1/2} + \{\mathrm{E}_P[\varrho_D^2(Z,X)]\}^{1/2} + \{\mathrm{E}_P[r_Z^2(X)]\}^{1/2} \lesssim \sqrt{s/n}. \tag{3.23}$$

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

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 estimator $\tilde{\theta}_V$ 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 $\bar{\theta}_V = \tilde{\theta}_V$, has the near oracle rate 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 $\bar{\theta}_D = \tilde{\theta}_D$, has good theoretical properties including the near oracle rate of convergence, $\sqrt{(s \log p)/n}$. The resulting estimator for g_V is then

$$\widehat{g}_V(z,x) = \sum_{d=0}^{1} \widehat{e}_V(d,z,x)\widehat{l}_D(d,z,x). \tag{3.24}$$

3.2. **Second Step: Robust Estimation of the Reduced-Form Parameters $\alpha_V(z)$ and γ_V.** Estimation of the key quantities $\alpha_V(z)$ will make heavy use of orthogonal moment functions as defined in (1.2). 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 non-regularity 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 orthogonal moment function for $\alpha_V(z)$, $z \in \mathcal{Z} = \{0,1\}$, are given respectively by

$$\psi_{V,z}^{\alpha}(W) := \psi_{V,z,g_V,m_Z}^{\alpha}(W, \alpha_V(z)) \quad \text{and} \tag{3.25}$$

$$\psi_{V,z,g,m}^{\alpha}(W,\alpha) := \frac{1(Z=z)(V-g(z,X))}{m(z,X)} + g(z,X) - \alpha. \tag{3.26}$$

This efficient influence function was derived by Hahn (1998); it was 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

$$\psi_V^{\gamma}(W) := \psi_V^{\gamma}(W, \gamma_V), \text{ and } \psi_V^{\gamma}(W,\gamma) := V - \gamma. \tag{3.27}$$

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

$$\mathbb{E}_n[\psi^{\alpha}_{V,z,\widehat{g}_V,\widehat{m}_Z}(W,\alpha)] = 0, \quad \mathbb{E}_n[\psi^{\gamma}_{V}(W,\gamma)] = 0, \tag{3.28}$$

where $\widehat{g}_V$ and $\widehat{m}_Z$ are constructed as in Section 3.1. 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[11]

$$\widehat{\rho}_u := \big(\{\widehat{\alpha}_V(0), \widehat{\alpha}_V(1), \widehat{\gamma}_V\}\big)_{V\in\mathcal{V}_u} \quad \text{of} \quad \rho_u := \big(\{\alpha_V(0), \alpha_V(1), \gamma_V\}\big)_{V\in\mathcal{V}_u}. \tag{3.29}$$

The estimator and the parameter 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

$$\sqrt{n}(\widehat{\rho}_u - \rho_u) \rightsquigarrow N(0, \mathrm{Var}_P(\psi^{\rho}_u)), \quad \psi^{\rho}_u := (\{\psi^{\alpha}_{V,0}, \psi^{\alpha}_{V,1}, \psi^{\gamma}_{V}\})_{V\in\mathcal{V}_u}, \tag{3.30}$$
$$\text{uniformly in } P \in \mathcal{P}_n,$$

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 Appendix A 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 parameters 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 function-valued parameter ρ. We establish that $\sqrt{n}(\widehat{\rho} - \rho)$ is asymptotically Gaussian: In $\ell^\infty(\mathcal{U})^{d_\rho}$,

$$\sqrt{n}(\widehat{\rho} - \rho) \rightsquigarrow Z_P := (\mathbb{G}_P \psi^{\rho}_u)_{u\in\mathcal{U}}, \text{ uniformly in } P \in \mathcal{P}_n, \tag{3.31}$$

where $\mathbb{G}_P$ denotes the P-Brownian bridge (van der Vaart and Wellner, 1996, p. 81–82). This result contains (3.30) 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 (Giné and Zinn, 1984). We define $\widehat{\rho}^* = (\widehat{\rho}^*_u)_{u\in\mathcal{U}}$, a bootstrap draw of $\widehat{\rho}$, via

$$\widehat{\rho}^*_u = \widehat{\rho}_u + n^{-1}\sum_{i=1}^{n} \xi_i \widehat{\psi}^{\rho}_u(W_i). \tag{3.32}$$

[11] By default notation, $(a_j)_{j\in\mathcal{J}}$ returns a column vector produced by stacking components together in some consistent order.

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 P_ξ does not depend on P. We also impose that

$$\mathrm{E}[\xi] = 0, \quad \mathrm{E}[\xi^2] = 1, \quad \mathrm{E}[\exp(|\xi|)] < \infty. \tag{3.33}$$

Examples of ξ include (a) $\xi = \mathcal{E} - 1$, where $\mathcal{E}$ is a standard exponential random variable, (b) $\xi = \mathcal{N}$, where $\mathcal{N}$ is a 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.[12] The choices of (a), (b), and (c) correspond respectively to the Bayesian bootstrap (e.g., Hahn (1997) and Chamberlain and Imbens (2003)), the Gaussian multiplier method (e.g, Giné and Zinn (1984) and van der Vaart and Wellner (1996, Chap. 3.6)), and the wild bootstrap method (Mammen, 1993).[13] $\widehat{\psi}_u^\rho$ in (3.32) is an estimator of the influence function ψ_u^ρ defined via the plug-in rule:

$$\widehat{\psi}_u^\rho = (\widehat{\psi}_V^\rho)_{V \in \mathcal{V}_u}, \quad \widehat{\psi}_V^\rho(W) := \{\psi^\alpha_{V,0,\widehat{g}_V,\widehat{m}_Z}(W, \widehat{\alpha}_V(0)), \psi^\alpha_{V,1,\widehat{g}_V,\widehat{m}_Z}(W, \widehat{\alpha}_V(1)), \psi_V^\gamma(W, \widehat{\gamma}_V)\}. \tag{3.34}$$

Note that this bootstrap is computationally efficient since it does not involve recomputing the influence functions $\widehat{\psi}_u^\rho$.[14] 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 prior uses of this or similar methods in low-dimensional settings such as Hansen (1996) and Kline and Santos (2012).

We establish that the bootstrap law of $\sqrt{n}(\widehat{\rho}^* - \widehat{\rho})$ is uniformly asymptotically consistent: In the metric space $\ell^\infty(\mathcal{U})^{d_\rho}$, 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 weak convergence of the bootstrap law in probability, as defined in Appendix B.

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

$$\Delta := (\Delta_q)_{q \in \mathcal{Q}}, \text{ where } \Delta_q := \phi(\rho)(q), \ q \in \mathcal{Q}. \tag{3.35}$$

The structural parameters may themselves carry an index $q \in \mathcal{Q}$ that can be different from u; for example, the LQTE is indexed by a quantile index $q \in (0, 1)$. This formulation includes as special cases all the structural functions of Section 2. 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 (extended to handle uniformity in P).

[12] 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 the conditions and analysis could be extended to cover this case.

[13] 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.

[14] Chernozhukov and Hansen (2006) and Hong and Scaillet (2006) proposed a related computationally efficient bootstrap scheme that resamples the influence functions.

For the application of the functional delta method, we require that the functional $\rho \mapsto \phi(\rho)$ be Hadamard differentiable *uniformly* in $\rho \in \mathbb{D}_\rho$, where $\mathbb{D}_\rho$ is a set that contains the true values $\rho = \rho_P$ for all $P \in \mathcal{P}_n$, tangentially to a subset that contains the realizations of Z_P for all $P \in \mathcal{P}_n$ with derivative map $h \mapsto \phi'_\rho(h) = (\phi'_\rho(h)(q))_{q\in\mathcal{Q}}$.[15] We define the estimators of the structural parameters and their bootstrap versions via the plug-in rule as

$$\widehat{\Delta} := (\widehat{\Delta}_q)_{q\in\mathcal{Q}}, \quad \widehat{\Delta}_q := \phi\left(\widehat{\rho}\right)(q), \text{ and } \widehat{\Delta}^* := (\widehat{\Delta}^*_q)_{q\in\mathcal{Q}}, \quad \widehat{\Delta}^*_q := \phi\left(\widehat{\rho}^*\right)(q). \tag{3.36}$$

We establish that these estimators are asymptotically Gaussian

$$\sqrt{n}(\widehat{\Delta} - \Delta) \rightsquigarrow \phi'_\rho(Z_P), \text{ uniformly in } P \in \mathcal{P}_n, \tag{3.37}$$

and that the bootstrap consistently estimates their large sample distribution:

$$\sqrt{n}(\widehat{\Delta}^* - \widehat{\Delta}) \rightsquigarrow_B \phi'_\rho(Z_P), \text{ uniformly in } P \in \mathcal{P}_n. \tag{3.38}$$

These results can be used to construct simultaneous confidence bands and test functional hypotheses on Δ.

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

Consider fixed sequences of numbers $\delta_n \searrow 0$, $\epsilon_n \searrow 0$, $\Delta_n \searrow 0$, at a speed at most polynomial in n (for example, $\delta_n \geqslant 1/n^c$ for some $c > 0$), $\ell_n \to \infty$, and positive constants c, C, and $c' < 1/2$. These sequences and constants will not vary with P. The probability P can vary in the set $\mathcal{P}_n$ of probability measures, termed "data-generating processes", where $\mathcal{P}_n$ is typically a set that is weakly increasing in n, i.e. $\mathcal{P}_n \subseteq \mathcal{P}_{n+1}$.

Assumption 4.1 (Basic Assumptions)**.** *(i) Consider a random element W with values in a measure space $(\mathcal{W}, \mathcal{A}_\mathcal{W})$ and law determined by a probability measure $P \in \mathcal{P}_n$. The observed data $((W_{ui})_{u\in\mathcal{U}})_{i=1}^n$ consist of n i.i.d. copies of a random element $(W_u)_{u\in\mathcal{U}} = ((Y_u)_{u\in\mathcal{U}}, D, Z, X)$, where $\mathcal{U}$ is a Polish space equipped with its Borel sigma-field and $(Y_u, D, Z, X) \in \mathbb{R}^{3+d_X}$. Each W_u is generated via a measurable transform $t(W, u)$ of W and u, namely the map $t : \mathcal{W} \times \mathcal{U} \mapsto \mathbb{R}^{3+d_X}$ is measurable, and the map can possibly depend on P. Let*

$$\mathcal{V}_u := \{V_{uj}\}_{j\in\mathcal{J}} := \{Y_u, \mathbf{1}_0(D)Y_u, \mathbf{1}_0(D), \mathbf{1}_1(D)Y_u, \mathbf{1}_1(D)\}, \quad \mathcal{V} := (\mathcal{V}_u)_{u\in\mathcal{U}},$$

where $\mathcal{J} = \{1, ..., 5\}$. (ii) For $\mathcal{P} := \cup_{n=n_0}^\infty \mathcal{P}_n$, the map $u \mapsto Y_u$ obeys the uniform continuity property:

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

where the second supremum is taken over $u, \bar{u} \in \mathcal{U}$, and $\mathcal{U}$ is a totally bounded metric space equipped with a semi-metric $d_\mathcal{U}$. The uniform covering entropy of the set $\mathcal{F}_P = \{Y_u : u \in \mathcal{U}\}$, viewed as a

[15] We give the definition of uniform Hadamard differentiability in Definition B.1 of Appendix B.

collection of maps $(\mathcal{W}, \mathcal{A}_{\mathcal{W}}) \mapsto \mathbb{R}$*, obeys*

$$\sup_Q \log N(\epsilon \|F_P\|_{Q,2}, \mathcal{F}_P, \|\cdot\|_{Q,2}) \leqslant C \log(\mathrm{e}/\epsilon) \vee 0$$

for all $P \in \mathcal{P}$*, where* $F_P(W) = \sup_{u \in \mathcal{U}} |Y_u|$*, with the supremum taken over all finitely discrete probability measures* Q *on* $(\mathcal{W}, \mathcal{A}_{\mathcal{W}})$*. (iii) For each* $P \in \mathcal{P}$*, the conditional probability of* $Z = 1$ *given* X *is bounded away from zero or one, namely* $c' \leqslant m_Z(1, X) \leqslant 1 - c'$ *P-a.s., the instrument* Z *has a non-trivial impact on* D*, namely* $c' \leqslant |l_D(1, 1, X) - l_D(1, 0, X)|$ *P-a.s, and the regression function* g_V *is bounded,* $\|g_V\|_{P,\infty} < \infty$ *for all* $V \in \mathcal{V}$*.*

Assumption 4.1 is stated to deal with the measurability issues associated with functional response data. This assumption also 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

$$Z_{n,P} \rightsquigarrow Z_P \quad \text{in } \ell^\infty(\mathcal{U})^{d_\rho}, \text{ uniformly in } P \in \mathcal{P}, \tag{4.1}$$

where

$$Z_{n,P} := (\mathbb{G}_n \psi_u^\rho)_{u \in \mathcal{U}} \text{ and } Z_P := (\mathbb{G}_P \psi_u^\rho)_{u \in \mathcal{U}}, \tag{4.2}$$

with $\mathbb{G}_P$ denoting the P-Brownian bridge (van der Vaart and Wellner, 1996, p. 81–82) and with Z_P having bounded, uniformly continuous paths uniformly in $P \in \mathcal{P}$:

$$\sup_{P \in \mathcal{P}} \mathrm{E}_P \sup_{u \in \mathcal{U}} \|Z_P(u)\| < \infty, \quad \lim_{\varepsilon \searrow 0} \sup_{P \in \mathcal{P}} \mathrm{E}_P \sup_{d_{\mathcal{U}}(u, \tilde{u}) \leqslant \varepsilon} \|Z_P(u) - Z_P(\tilde{u})\| = 0. \tag{4.3}$$

Other assumptions will be specific to the strategy adopted.

Assumption 4.2 (Approximate Sparsity for Strategy 1)**.** *Under each* $P \in \mathcal{P}_n$ *and for each* $n \geqslant n_0$*, uniformly for all* $V \in \mathcal{V}$*: (i) The approximations (3.8)-(3.10) hold with the link functions* Λ_V *and* Λ_Z *belonging to the set* $\mathcal{L}$*, the sparsity condition* $\|\beta_V\|_0 + \|\beta_Z\|_0 \leqslant s$ *holding, the approximation errors satisfying* $\|r_V\|_{P,2} + \|r_Z\|_{P,2} \leqslant \delta_n n^{-1/4}$ *and* $\|r_V\|_{P,\infty} + \|r_Z\|_{P,\infty} \leqslant \epsilon_n$*, and the sparsity index* s *and the number of terms* p *in the vector* $f(X)$ *obeying* $s^2 \log^2(p \vee n) \log^2 n \leqslant \delta_n n$*. (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* $\|f(Z, X)'(\bar{\beta}_V - \beta_V)\|_{\mathbb{P}_n,2} + \|f(X)'(\bar{\beta}_Z - \beta_Z)\|_{\mathbb{P}_n,2} \leqslant \delta_n n^{-1/4}$*,* $K_n \|\bar{\beta}_V - \beta_V\|_1 + K_n \|\bar{\beta}_Z - \beta_Z\|_1 \leqslant \epsilon_n$*; the estimators are sparse such that* $\|\bar{\beta}_V\|_0 + \|\bar{\beta}_Z\|_0 \leqslant Cs$*; and the empirical and population norms induced by the Gram matrix formed by* $(f(X_i))_{i=1}^n$ *are equivalent on sparse subsets,* $\sup_{\|\delta\|_0 \leqslant \ell_n s} |\|f(X)'\delta\|_{\mathbb{P}_n,2} / \|f(X)'\delta\|_{P,2} - 1| \leqslant \epsilon_n$*. (iii) The following boundedness conditions hold:* $\|\|f(X)\|_\infty\|_{P,\infty} \leqslant K_n$ *and* $\|V\|_{P,\infty} \leqslant C$*.*

Comment 4.1. Assumption 4.2 imposes simple intermediate-level conditions which encode both the approximate sparsity of the models as well as some reasonable behavior of the sparse estimators of m_Z and g_V. Sufficient conditions for the equivalence between empirical and population norms and primitive examples of functions admitting sparse approximations are given in Belloni, Chernozhukov, and Hansen (2014). Primitive conditions for the estimators obeying the bounds above while addressing the problem of estimating continua of approximately sparse nuisance functions

are given in Section 6. These conditions extend and generalize the conditions employed in the literature on adaptive estimation using series methods. 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 4.3 (Approximate Sparsity for Strategy 2)**.** *Under each $P \in \mathcal{P}_n$ and for each $n \geqslant n_0$, uniformly for all $V \in \mathcal{V}$: (i) The approximations (3.15)-(3.21) and (3.10) 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 in the vector $f(X)$ obeying $s^2 \log^2(p \vee n) \log^2 n \leqslant \delta_n n$. (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 population norms induced by the Gram matrix formed by $(f(X_i))_{i=1}^n$ are equivalent on sparse subsets, $\sup_{\|\delta\|_0 \leqslant \ell_n s} |\|f(X)'\delta\|_{\mathbb{P}_n,2} / \|f(X)'\delta\|_{P,2} - 1| \leqslant \epsilon_n$. (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 (3.29) obeys the following relations. We recall definitions of convergence uniformly in $P \in \mathcal{P}_n$ in Appendix A.

Theorem 4.1 (**Uniform Gaussianity of the Reduced-Form Parameter Process**)**.** *Under Assumptions 4.1 and 4.2 or 4.1 and 4.3, the reduced-form empirical process admits a linearization; namely,*

$$\widehat{Z}_{n,P} := \sqrt{n}(\widehat{\rho} - \rho) = Z_{n,P} + o_P(1) \quad \text{in } \ell^\infty(\mathcal{U})^{d_\rho}, \text{ uniformly in } P \in \mathcal{P}_n. \tag{4.4}$$

The process $\widehat{Z}_{n,P}$ is asymptotically Gaussian, namely

$$\widehat{Z}_{n,P} \rightsquigarrow Z_P \quad \text{in } \ell^\infty(\mathcal{U})^{d_\rho}, \text{ uniformly in } P \in \mathcal{P}_n, \tag{4.5}$$

where Z_P is defined in (4.2) and its paths obey the property (4.3).

Another main result of this section shows that the bootstrap law of the process

$$\widehat{Z}^*_{n,P} := \sqrt{n}(\widehat{\rho}^* - \widehat{\rho}) := \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i \widehat{\psi}^\rho_u(W_i),$$

where $\widehat{\psi}^\rho_u$ is defined in (3.34), 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 4.1 and 4.2 or 4.1 and 4.3, the bootstrap law consistently approximates the large sample law Z_P of $Z_{n,P}$ uniformly in $P \in \mathcal{P}_n$, namely,*

$$\widehat{Z}^*_{n,P} \rightsquigarrow_B Z_P \quad \text{in } \ell^\infty(\mathcal{U})^{d_\rho}, \text{ uniformly in } P \in \mathcal{P}_n. \tag{4.6}$$

Next we consider inference on the structural functionals Δ defined in (3.35). We derive the large sample distribution of the estimator $\widehat{\Delta}$ in (3.36), and show that the multiplier bootstrap law of $\widehat{\Delta}^*$ in (3.36) provides a consistent approximation to that distribution. We rely on the functional delta method in our derivations, which we modify to handle uniformity with respect to the underlying dgp P. Our argument relies on the following assumption on the structural functionals.

Assumption 4.4 (Uniform Hadamard Differentiability of Structural Functionals)**.** *Suppose that for each $P \in \mathcal{P}$, $\rho = \rho_P \in \mathbb{D}_\rho$, a compact metric space. Suppose $\varrho \mapsto \phi(\varrho)$, a functional of interest mapping $\mathbb{D}_\phi \subset \mathbb{D} = \ell^\infty(\mathcal{U})^{d_\rho}$ to $\ell^\infty(\mathcal{Q})$, where $\mathbb{D}_\rho \subset \mathbb{D}_\phi$, is Hadamard differentiable in ϱ tangentially to $\mathbb{D}_0 = UC(\mathcal{U})^{d_\rho}$ uniformly in $\varrho \in \mathbb{D}_\rho$, with the linear derivative map $\phi'_\varrho : \mathbb{D}_0 \mapsto \mathbb{D}$ such that the mapping $(\varrho, h) \mapsto \phi'_\varrho(h)$ from $\mathbb{D}_\rho \times \mathbb{D}_0$ to $\ell^\infty(\mathcal{Q})$ is continuous.*

The definition of uniform Hadamard differentiability is given in Definition B.1 of Appendix B. Assumption 4.4 holds for all examples of structural parameters listed in Section 2.

The following corollary gives the large sample law of $\sqrt{n}(\widehat{\Delta} - \Delta)$, 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 large sample law $\sqrt{n}(\widehat{\Delta} - \Delta)$. It follows from the previous theorems as well as from 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 4.1, 4.2 or 4.3, and 4.4,*

$$\sqrt{n}(\widehat{\Delta} - \Delta) \rightsquigarrow T_P := \phi'_{\rho_P}(Z_P), \quad \text{in } \ell^\infty(\mathcal{Q}), \text{ uniformly in } P \in \mathcal{P}_n, \tag{4.7}$$

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

$$\sqrt{n}(\widehat{\Delta}^* - \widehat{\Delta}) \rightsquigarrow_B T_P, \quad \text{in } \ell^\infty(\mathcal{Q}), \text{ uniformly in } P \in \mathcal{P}_n. \tag{4.8}$$

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

In this section, we consider a general setting where possibly a continuum of target parameters is of interest and Lasso-type or Post-Lasso-type methods are used to estimate a continuum of high-dimensional nuisance functions. This setting covers a rich variety of modern moment-condition problems in econometrics including the treatment effects problem. We establish a functional central limit theorem for the estimators of the continuum of target parameters that 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 derive a functional central limit theorem for the multiplier bootstrap that resamples the first order approximations to the standardized estimators of the continuum of target parameters and establish its uniform validity. Moreover, we establish the uniform validity of the functional delta method and the functional delta method for the multiplier bootstrap for

smooth functionals of the continuum of target parameters using an appropriate strengthening of Hadamard differentiability.

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

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

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

$$\mathrm{E}_P[\psi_u(W_u, \theta_u, h_u(Z_u))] = 0, \tag{5.1}$$

where W_u is a random vector that takes values in a Borel set $\mathcal{W}_u \subset \mathbb{R}^{d_w}$ and contains as a subcomponent the vector Z_u taking values in a Borel set $\mathcal{Z}_u$, the moment function

$$\psi_u : \mathcal{W}_u \times \Theta_u \times T_u \mapsto \mathbb{R}^{d_\theta}, \quad (w, \theta, t) \mapsto \psi_u(w, \theta, t) = (\psi_{uj}(w, \theta, t))_{j=1}^{d_\theta} \tag{5.2}$$

is a Borel measurable map, and the function

$$h_u : \mathcal{Z}_u \mapsto \mathbb{R}^{d_t}, \quad z \mapsto h_u(z) = (h_{um}(z))_{m=1}^{d_t} \in T_u(z), \tag{5.3}$$

is another Borel measurable map that denotes the possibly infinite-dimensional nuisance parameter. The sets $T_u(z)$ are assumed to be convex for each $u \in \mathcal{U}$ and $z \in \mathcal{Z}_u$.

We assume that the continuum of nuisance functions $(h_u)_{u \in \mathcal{U}}$ is approximately sparse and thus can be modelled and estimated using modern regularization and post-selection methods such as Lasso and Post-Lasso. We let $\widehat{h}_u = (\widehat{h}_{um})_{m=1}^{d_t}$ denote the estimator of h_u, which we assume obeys the conditions in Assumption 5.3. 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 moment condition (5.1), i.e.,

$$\|\mathbb{E}_n[\psi_u(W_u, \widehat{\theta}_u, \widehat{h}_u(Z_u))]\| \leqslant \inf_{\theta \in \Theta_u} \|\mathbb{E}_n[\psi(W_u, \theta, \widehat{h}_u(Z_u))]\| + \epsilon_n, \text{ where } \epsilon_n = o(n^{-1/2}). \tag{5.4}$$

The key condition needed for regular estimation of θ_u is an orthogonality or immunization condition. The simplest to explain, yet strongest, form of this condition can be expressed as follows:

$$\partial_t \mathrm{E}_P[\psi_u(W_u, \theta_u, h_u(Z_u))|Z_u] = 0, \text{ a.s.}, \tag{5.5}$$

subject to additional technical conditions such as continuity (5.6) and dominance (5.7) stated below, where we use the symbol ∂_t to abbreviate $\frac{\partial}{\partial t'}$.[16] This condition holds in the previous setting of inference on relevant treatment effects after interchanging the order of the derivative and expectation. The formulation here also covers certain non-smooth cases such as structural and instrumental quantile regression problems.

In the formal development, we use a more general form of the orthogonality condition.

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

Definition 5.1 (**Orthogonality for Moment Condition Models, General Form**). *For each $u \in \mathcal{U}$, suppose that (5.1)–(5.3) hold. Consider $\mathcal{H}_u$, a set of measurable functions $z \mapsto h(z) \in T_u(z)$ from $\mathcal{Z}_u$ to $\mathbb{R}^{d_t}$ such that $\|h(Z_u) - h_u(Z_u)\|_{P,2} < \infty$ for all $h \in \mathcal{H}_u$. Suppose also that the set $T_u(z)$ is a convex subset of $\mathbb{R}^{d_t}$ for each $z \in \mathcal{Z}_u$. We say that ψ_u obeys a general form of orthogonality with respect to $\mathcal{H}_u$ uniformly in $u \in \mathcal{U}$, if the following conditions hold: For each $u \in \mathcal{U}$, the derivative*

$$t \mapsto \partial_t \mathrm{E}_P[\psi_u(W_u, \theta_u, t)|Z_u] \text{ is continuous on } t \in T_u(Z_u) \text{ } P\text{-a.s.}, \tag{5.6}$$

is dominated,

$$\left\| \sup_{t \in T_u(Z_u)} \left\| \partial_t \mathrm{E}_P[\psi_u(W_u, \theta_u, t)|Z_u] \right\| \right\|_{P,2} < \infty, \tag{5.7}$$

and obeys the orthogonality condition:

$$\mathrm{E}_P\Big[\partial_t \mathrm{E}_P\big[\psi_u(W_u, \theta_u, h_u(Z_u))|Z_u\big](h(Z_u) - h_u(Z_u))\Big] = 0 \quad \text{for all } h \in \mathcal{H}_u. \tag{5.8}$$

The orthogonality condition (5.8) reduces to (5.5) when $\mathcal{H}_u$ can span all measurable functions $h : \mathcal{Z}_u \mapsto T_u$ such that $\|h\|_{P,2} < \infty$ but is more general otherwise.

Comment 5.1. It is important to use a moment function ψ_u that satisfies the orthogonality property given in (5.8). Generally, if we have a moment function $\tilde{\psi}_u$ which identifies θ_u but does not have this property, we can construct a moment function ψ_u that identifies θ_u and has the required orthogonality property by projecting the original function $\tilde{\psi}_u$ onto the orthocomplement of the tangent space for the nuisance functions h_u; see, for example, van der Vaart and Wellner (1996), van der Vaart (1998, Chap. 25), Kosorok (2008), Belloni, Chernozhukov, and Kato (2013), and Belloni, Chernozhukov, and Hansen (2014). ■

Comment 5.2 (**An alternative formulation of the orthogonality condition**). A slightly more general, though less primitive definition is as follows. For each $u \in \mathcal{U}$, suppose that (5.1)-(5.3) hold. Consider $\mathcal{H}_u$, a set of measurable functions $z \mapsto h(z) \in T_u(z)$ from $\mathcal{Z}_u$ to $\mathbb{R}^{d_t}$ such that $\|h(Z_u) - h_u(Z_u)\|_{P,2} < \infty$ for all $h \in \mathcal{H}_u$, where the set $T_u(z)$ is a convex subset of $\mathbb{R}^{d_t}$ for each $z \in \mathcal{Z}_u$. We say that ψ_u obeys a general form of orthogonality with respect to $\mathcal{H}_u$ uniformly in $u \in \mathcal{U}$, if the following conditions hold: The Gateaux derivative map

$$\mathrm{D}_{u,t}[h - h_u] := \partial_t \mathrm{E}_P\left(\psi_u\Big\{W_u, \theta_u, h_u(Z_u) + t\Big[h(Z_u) - h_u(Z_u)\Big]\Big\}\right)$$

exists for all $t \in [0, 1)$, $h \in \mathcal{H}_u$, and $u \in \mathcal{U}$ and vanishes at $t = 0$ – namely,

$$\mathrm{D}_{u,0}[h - h_u] = 0 \quad \text{for all } h \in \mathcal{H}_u. \tag{5.9}$$

Definition 5.1 implies this definition by the mean-value expansion and the dominated convergence theorem. ■

In what follows, we shall denote by δ, c_0, c, and C some positive constants. For a positive integer d, $[d]$ denotes the set $\{1, \ldots, d\}$.

Assumption 5.1 (Moment condition problem)**.** *Consider a random element W, taking values in a measure space $(\mathcal{W}, \mathcal{A}_{\mathcal{W}})$, with law determined by a probability measure $P \in \mathcal{P}_n$. The observed data $((W_{ui})_{u\in\mathcal{U}})_{i=1}^n$ consist of n i.i.d. copies of a random element $(W_u)_{u\in\mathcal{U}}$ which is generated as a suitably measurable transformation with respect to W and u. Uniformly for all $n \geqslant n_0$ and $P \in \mathcal{P}_n$, the following conditions hold: (i) The true parameter value θ_u obeys (5.1) and is interior relative to $\Theta_u \subset \Theta \subset \mathbb{R}^{d_\theta}$, namely there is a ball of radius δ centered at θ_u contained in Θ_u for all $u \in \mathcal{U}$, and Θ is compact. (ii) For $\nu := (\nu_k)_{k=1}^{d_\theta+d_t} = (\theta, t)$, each $j \in [d_\theta]$ and $u \in \mathcal{U}$, the map $\Theta_u \times T_u(Z_u) \ni \nu \mapsto \mathrm{E}_P[\psi_{uj}(W_u, \nu)|Z_u]$ is twice continuously differentiable a.s. with derivatives obeying the integrability conditions specified in Assumption 5.2. (iii) For all $u \in \mathcal{U}$, the moment function ψ_u obeys the orthogonality condition given in Definition 5.1 for the set $\mathcal{H}_u = \mathcal{H}_{un}$ specified in Assumption 5.3. (iv) The following identifiability condition holds: $\|\mathrm{E}_P[\psi_u(W_u, \theta, h_u(Z_u))]\| \geqslant 2^{-1}(\|J_u(\theta - \theta_u)\| \wedge c_0)$ for all $\theta \in \Theta_u$, where the singular values of $J_u := \partial_\theta \mathrm{E}[\psi_u(W_u, \theta_u, h_u(Z_u))]$ lie between $c > 0$ and C for all $u \in \mathcal{U}$.*

The conditions of Assumption 5.1 are mild and standard in moment condition problems. Assumption 5.1(iv) encodes sufficient global and local identifiability to obtain a rate result. The suitably measurable condition, defined in Appendix A, is a mild condition satisfied in most practical cases.

Assumption 5.2 (Entropy and smoothness)**.** *The set $(\mathcal{U}, d_{\mathcal{U}})$ is a semi-metric space such that $\log N(\epsilon, \mathcal{U}, d_{\mathcal{U}}) \leqslant C \log(\mathrm{e}/\epsilon) \vee 0$. Let $\alpha \in [1, 2]$, and let α_1 and α_2 be some positive constants. Uniformly for all $n \geqslant n_0$ and $P \in \mathcal{P}_n$, the following conditions hold: (i) The set of functions $\mathcal{F}_0 = \{\psi_{uj}(W_u, \theta_u, h_u(Z_u)) : j \in [d_\theta], u \in \mathcal{U}\}$, viewed as functions of W is suitably measurable; has an envelope function $F_0(W) = \sup_{j\in[d_\theta],u\in\mathcal{U},\nu\in\Theta_u\times T_u(Z_u)} |\psi_{uj}(W_u, \nu)|$ that is measurable with respect to W and obeys $\|F_0\|_{P,q} \leqslant C$, where $q \geqslant 4$ is a fixed constant; and has a uniform covering entropy obeying $\sup_Q \log N(\epsilon\|F_0\|_{Q,2}, \mathcal{F}_0, \|\cdot\|_{Q,2}) \leqslant C \log(\mathrm{e}/\epsilon) \vee 0$. (ii) For all $j \in [d_\theta]$ and $k, r \in [d_\theta + d_t]$, and $\psi_{uj}(W) := \psi_{uj}(W_u, \theta_u, h_u(Z_u))$,*

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

Assumption 5.2 imposes smoothness and integrability conditions on various quantities derived from ψ_u. It also imposes conditions on the complexity of the relevant function classes.

In what follows, let $\Delta_n \searrow 0$, $\delta_n \searrow 0$, and $\tau_n \searrow 0$ be sequences of constants approaching zero from above at a speed at most polynomial in n (for example, $\delta_n \geqslant 1/n^c$ for some $c > 0$).

Assumption 5.3 (Estimation of nuisance functions)**.** *The following conditions hold for each $n \geqslant n_0$ and all $P \in \mathcal{P}_n$. The estimated functions $\widehat{h}_u = (\widehat{h}_{um})_{m=1}^{d_t} \in \mathcal{H}_{un}$ with probability at least $1 - \Delta_n$,*

where $\mathcal{H}_{un}$ *is the set of measurable maps* $\mathcal{Z}_u \ni z \mapsto h = (h_m)_{m=1}^{d_t}(z) \in T_u(z)$ *such that*

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

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

$$\sup_Q \log N(\epsilon \|F_1\|_{Q,2}, \mathcal{F}_1, \|\cdot\|_{Q,2}) \leqslant s_n(\log(a_n/\epsilon)) \vee 0,$$

where $F_1(W)$ *is an envelope for* $\mathcal{F}_1$ *which is measurable with respect to* W *and satisfies* $F_1(W) \leqslant F_0(W)$*, for the* F_0 *defined in Assumption 5.2. The complexity characteristics* $a_n \geqslant \max(n, \mathrm{e})$ *and* $s_n \geqslant 1$ *obey the growth conditions:*

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

where q *and* α *are defined in Assumption 5.2.*

Assumption 5.3 imposes conditions on the estimation rate of the nuisance functions h_{um} and on the complexity of the functions sets that contain the estimators $\widehat{h}_{um}$. Within the approximately sparse framework, the index s_n corresponds to the maximum of the dimension of the approximating models and of the size of the selected models; and $a_n = p \vee n$. Under other frameworks, these parameters could be different; yet if they are well-behaved, then our results still apply. Thus, these results potentially cover other frameworks, where assumptions other than approximate sparsity are used to make the estimation problem manageable. It is important to point out that the class $\mathcal{F}_1$ need not be Donsker because its entropy is allowed to increase with n. Allowing for non-Donsker classes is crucial for accommodating modern high-dimensional estimation methods for the nuisance functions as we have seen in the previous section. This feature makes the conditions imposed here very different from the conditions imposed in various classical references on dealing with nonparametrically estimated nuisance functions; see, for example, van der Vaart and Wellner (1996), van der Vaart (1998), and Kosorok (2008).

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 5.1, 5.2, and 5.3, for an estimator* $(\widehat{\theta}_u)_{u\in\mathcal{U}}$ *that obeys equation (5.4),*

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

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

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

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

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

Comment 5.3. It is important to mention here that this result on a continuum of parameters solving a continuum of moment conditions is completely new. The prior approaches dealing with continuums of moment conditions with infinite-dimensional nuisance parameters, for example, the ones given in Chernozhukov and Hansen (2006) and Escanciano and Zhu (2013), impose the Donsker conditions on the class of functions, following Andrews (1994a), that contain the values of the estimators of these nuisance functions. This approach is completely precluded in our setting, since the resulting class of functions in our case has entropy that grows with the sample size and therefore the class is not Donsker. Hence, we develop a new approach to establishing the results which exploits the delicate interplay between the rate of growth of entropy, the biases, and the size of the estimation error. In addition, the new approach allows for obtaining results that are uniform in P. ∎

We can estimate the law of Z_P with the bootstrap law of

$$\widehat{Z}^*_{n,P} := \sqrt{n}(\widehat{\theta}^*_u - \widehat{\theta}_u)_{u\in\mathcal{U}} := \left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i \widehat{\psi}_u(W_i)\right)_{u\in\mathcal{U}}, \tag{5.10}$$

where $(\xi_i)_{i=1}^n$ are i.i.d. multipliers as defined in equation (3.33), $\widehat{\psi}_u(W_i)$ is the estimated score

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

and $\widehat{J}_u$ is a suitable estimator of J_u.[17] The bootstrap law is computed by drawing $(\xi_i)_{i=1}^n$ conditional on the data.

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 5.1, 5.2, and 5.3 hold, the estimator $(\widehat{\theta}_u)_{u\in\mathcal{U}}$ obeys equation (5.4), and that, for the constant α defined in Assumption 5.2 and some positive constant α_3, uniformly in $P \in \mathcal{P}_n$ with probability $1-\delta_n$,*

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

Then,

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

We next derive the large sample distribution and validity of the multiplier bootstrap for the estimator $\widehat{\Delta} := \phi(\widehat{\theta}) := \phi((\widehat{\theta}_u)_{u\in\mathcal{U}})$ of the functional $\Delta := \phi(\theta^0) = \phi((\theta_u)_{u\in\mathcal{U}})$ using the functional delta method. The functional $\theta^0 \mapsto \phi(\theta^0)$ is defined as a uniformly Hadamard differentiable transform of $\theta^0 = (\theta_u)_{u\in\mathcal{U}}$. The following result gives the large sample law of $\sqrt{n}(\widehat{\Delta} - \Delta)$, the properly normalized estimator. It also shows that the bootstrap law of $\sqrt{n}(\widehat{\Delta}^* - \widehat{\Delta})$, computed conditionally on the data, is consistent for the large sample law of $\sqrt{n}(\widehat{\Delta} - \Delta)$. Here $\widehat{\Delta}^* := \phi(\widehat{\theta}^*) = \phi((\widehat{\theta}^*)_{u\in\mathcal{U}})$

[17] We do not discuss the estimation of J_u since it is often a problem-specific matter. In Section 3, J_u was equal to the identity matrix, so we did not need to estimate it.

is the bootstrap version of $\widehat{\Delta}$, and $\widehat{\theta}_u^* = \widehat{\theta}_u + n^{-1}\sum_{i=1}^n \xi_i \widehat{\psi}_u(W_i)$ is the multiplier bootstrap version of $\widehat{\theta}_u$ defined via equation (5.10).

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

$$\sqrt{n}(\widehat{\Delta} - \Delta) \rightsquigarrow T_P := \phi'_{\theta_P^0}(Z_P) \quad \text{in } \ell^\infty(\mathcal{Q}), \text{ uniformly in } P \in \mathcal{P}_n, \tag{5.11}$$

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

$$\sqrt{n}(\widehat{\Delta}^* - \widehat{\Delta}) \rightsquigarrow_B T_P \quad \text{in } \ell^\infty(\mathcal{Q}), \text{ uniformly in } P \in \mathcal{P}_n. \tag{5.12}$$

To derive Theorem 5.3, we strengthen the usual notion of Hadamard differentiability to a uniform notion introduced in Definition B.1. Theorems B.3 and B.4 show that this uniform Hadamard differentiability is sufficient to guarantee the validity of the functional delta uniformly in P. These new uniform functional delta method theorems may be of independent interest.

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 $u \in \mathcal{U}$,

$$\mathrm{E}_P[Y_u \mid X] = \Lambda(f(X)'\theta_u) + r_u(X), \tag{6.1}$$

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 known link function. The notation in this section differs from the rest of the paper with Y_u and X denoting a generic response and a generic vector of covariates to facilitate the application of these results to other contexts. We only consider the linear link function, $\Lambda(t) = t$, and the logistic link function, $\Lambda(t) = \exp(t)/\{1+\exp(t)\}$, in detail.

Considering the logistic link is useful when the functional response is binary, though the linear link can be used in that case as well under some conditions. For example, it is useful for estimating a high-dimensional generalization of the distributional regression models considered in Chernozhukov, Fernández-Val, and Melly (2013) where the response variable is the continuum $(Y_u = 1(Y \leqslant u))_{u \in \mathcal{U}}$.

Even though we focus on these two cases we note that the principles discussed here apply to many other convex (or near-convex) M-estimators. In the remainder of the section, we discuss and establish results for ℓ_1-penalized and post-model selection estimators of $(\theta_u)_{u\in\mathcal{U}}$ that hold uniformly over $u \in \mathcal{U}$.

Throughout the section, we assume that $u \in \mathcal{U} \subset [0,1]^{d_u}$ and that n i.i.d. observations from dgps where (6.1) holds, $\{(Y_{ui})_{u\in\mathcal{U}}, X_i)\}_{i=1}^n$, are available to estimate $(\theta_u)_{u\in\mathcal{U}}$. For each $u \in \mathcal{U}$, a penalty level λ, and a diagonal matrix of penalty loadings $\widehat{\Psi}_u$, we define the Lasso estimator as

$$\widehat{\theta}_u \in \arg\min_{\theta\in\mathbb{R}^p} \mathbb{E}_n[M(Y_u, f(X)'\theta)] + \frac{\lambda}{n}\|\widehat{\Psi}_u\theta\|_1 \tag{6.2}$$

where $M(y,t) = \frac{1}{2}(y-\Lambda(t))^2$ 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 the logistic link function for binary response data. For each $u \in \mathcal{U}$, the Post-Lasso estimator based on a set of covariates $\widetilde{T}_u$ is then defined as

$$\widetilde{\theta}_u \in \arg\min_{\theta\in\mathbb{R}^p} \mathbb{E}_n[M(Y_u, f(X)'\theta)] \quad : \quad \mathrm{supp}(\theta) \subseteq \widetilde{T}_u \tag{6.3}$$

where the set $\widetilde{T}_u$ contains $\mathrm{supp}(\widehat{\theta}_u)$ and possibly additional variables deemed as important.[18] We will set $\widetilde{T}_u = \mathrm{supp}(\widehat{\theta}_u)$ unless otherwise noted.

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

$$\frac{\lambda}{n} \geqslant c \sup_{u\in\mathcal{U}} \left\| \widehat{\Psi}_u^{-1}\mathbb{E}_n\left[\partial_\theta M(Y_u, f(X)'\theta_u)\right] \right\|_\infty, \tag{6.4}$$

where $c > 1$ is a fixed constant. When $\mathcal{U}$ is a singleton 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 (6.4) to derive the properties of Lasso and Post-Lasso. When $\mathcal{U}$ is not a singleton, this strategy was first employed in the context of ℓ_1-penalized quantile regression processes by Belloni and Chernozhukov (2011).

To implement (6.4), we propose setting the penalty level as

$$\lambda = c\sqrt{n}\Phi^{-1}(1-\gamma/\{2pn^{d_u}\}), \tag{6.5}$$

where d_u is the dimension of $\mathcal{U}$, $1-\gamma$ with $\gamma = o(1)$ is a confidence level associated with the probability of event (6.4), and $c > 1$ is a slack constant.[19] When implementing the estimators, we set $c = 1.1.$ and $\gamma = .1/\log(n)$, though other choices are theoretically valid.

[18]The total number of additional variables $\widehat{s}_a$ should also obey the same growth conditions that s obeys. For example, if the additional variables are chosen so that $\widehat{s}_a \lesssim |\mathrm{supp}(\widehat{\theta}_u)|$ the growth condition is satisfied with probability going to one for the designs covered by Assumptions 6.1 and 6.2. See also Belloni, Chernozhukov, and Hansen (2014) for a discussion on choosing additional variables.

[19]When the set $\mathcal{U}$ is a singleton, one can use the penalty level in (6.5) with $d_u = 0$. This choice corresponds to that used in Belloni, Chernozhukov, and Hansen (2014).

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

Algorithm 1 (Estimation of Penalty Loadings)**.** Choose $\gamma \in [1/n, \min\{1/\log n, pn^{d_u - 1}\}]$ and $c > 1$ to form λ as defined in (6.5), 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.** We provide sufficient conditions for establishing good performance of the estimators discussed above when the linear link function is used. In the statement of the following assumption, $\delta_n \searrow 0$, $\ell_n \to \infty$, and $\Delta_n \searrow 0$ are fixed sequences; and c, C, κ', κ'' and $\nu \in (0, 1]$ are positive finite constants.

Assumption 6.1. *Consider a random element W taking values in a measure space $(\mathcal{W}, \mathcal{A}_\mathcal{W})$, with law determined by a probability measure $P \in \mathcal{P}_n$. The observed data $((Y_{ui})_{u\in\mathcal{U}}, X_i)_{i=1}^n$ consist of n i.i.d. copies of random element $((Y_u)_{u\in\mathcal{U}}, X)$, which is generated as a suitably measurable transformation of W and u. The model (6.1) holds with linear link $t \mapsto \Lambda(t) = t$ for all $u \in \mathcal{U} \subset [0, 1]^{d_u}$, where d_u is fixed and $\mathcal{U}$ is equipped with the semi-metric $d_\mathcal{U}$. Uniformly for all $n \geqslant n_0$ and $P \in \mathcal{P}_n$, the following conditions hold. (i) The model (6.1) is approximately sparse with sparsity index obeying $\sup_{u\in\mathcal{U}} \|\theta_u\|_0 \leqslant s$ and the growth restriction $\log(p \vee n) \leqslant \delta_n n^{1/3}$. (ii) The set $\mathcal{U}$ has uniform covering entropy obeying $\log N(\epsilon, \mathcal{U}, d_\mathcal{U}) \leqslant d_u \log(1/\epsilon) \vee 0$, and the collection $(\zeta_u = Y_u - \mathrm{E}_P[Y_u \mid X], r_u)_{u\in\mathcal{U}}$ are suitably measurable transformations of W and u. (iii) Uniformly over $u \in \mathcal{U}$, the moments of the model are boundedly heteroscedastic, namely $c \leqslant \mathrm{E}_P[\zeta_u^2 \mid X] \leqslant C$ a.s., and $\max_{j\leqslant p} \mathrm{E}_P[|f_j(X)\zeta_u|^3 + |f_j(X)Y_u|^3] \leqslant C$. (iv) For a fixed $\nu > 0$ and a sequence K_n, 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.; $K_n^2 s \log(p \vee n) \leqslant \delta_n n$. (b) With probability $1 - \Delta_n$, $\sup_{u\in\mathcal{U}} \mathbb{E}_n[r_u^2(X)] \leqslant Cs \log(p \vee n)/n$; $\sup_{u\in\mathcal{U}} \max_{j\leqslant p} |(\mathbb{E}_n - \mathrm{E}_P)[f_j^2(X)\zeta_u^2]| \vee |(\mathbb{E}_n - \mathrm{E}_P)[f_j^2(X)Y_u^2]| \leqslant \delta_n$; $\log^{1/2}(p \vee n) \sup_{d_\mathcal{U}(u,u')\leqslant 1/n} \max_{j\leqslant p} \{\mathbb{E}_n[f_j(X)^2(\zeta_u - \zeta_{u'})^2]\}^{1/2} \leqslant \delta_n$, and $\sup_{d_\mathcal{U}(u,u')\leqslant 1/n} \|\mathbb{E}_n[f(X)(\zeta_u - \zeta_{u'})]\|_\infty \leqslant \delta_n n^{-1/2}$. (c) With probability $1 - \Delta_n$, the empirical minimum and maximum sparse eigenvalues are bounded from zero and above, namely $\kappa' \leqslant \inf_{\|\delta\|_0\leqslant s\ell_n} \|f(X)'\delta\|_{\mathbb{P}_n,2} \leqslant \sup_{\|\delta\|_0\leqslant s\ell_n} \|f(X)'\delta\|_{\mathbb{P}_n,2} \leqslant \kappa''$.*

Assumption 6.1 is only a set of sufficient conditions. The finite sample results in the Appendix allow for more general conditions (for example, d_u can grow with the sample size). We verify that the more technical conditions in Assumption 6.1(iv)(b) hold in a variety of cases, see Lemma G.2

in Appendix G. Under Assumption 6.1, we establish results on the performance of the estimators (6.2) and (6.3) for the linear link function case that hold uniformly over $u \in \mathcal{U}$ and $P \in \mathcal{P}_n$.

Theorem 6.1 (Rates and Sparsity for Functional Responses under Linear Link)**.** *Under Assumption 6.1 and setting the penalty and loadings 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 4.2 and 4.3. Indeed, under the condition $s^2\log^2(p\vee n)\log^2 n \leqslant \delta_n n$, the rate of convergence established in Theorem 6.1 yields $\sqrt{s\log(p\vee n)/n} \leqslant o(n^{-1/4})$.

6.3. **Asymptotic Properties of a Continuum of Lasso and Post-Lasso Estimators for Functional Responses: Logistic Case.** 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, $Y_u \in \{0,1\}$ for all $u \in \mathcal{U}$. Consider the fixed sequences $\delta_n \searrow 0$, $\ell_n \to \infty$, and $\Delta_n \searrow 0$ and the positive finite constants c, C, κ', κ'', and $\underline{c} \leqslant 1/2$.

Assumption 6.2. *Consider a random element W taking values in a measure space $(\mathcal{W}, \mathcal{A}_\mathcal{W})$, with law determined by a probability measure $P \in \mathcal{P}_n$. The observed data $((Y_{ui})_{u\in\mathcal{U}}, X_i)_{i=1}^n$ consist of n i.i.d. copies of random element $((Y_u)_{u\in\mathcal{U}}, X)$, which is generated as a suitably measurable transformation of W and u. The model (6.1) holds with $Y_{ui} \in \{0,1\}$ with the logistic link $t \mapsto \Lambda(t) = \exp(t)/\{1+\exp(t)\}$ for each $u \in \mathcal{U} \subset [0,1]^{d_u}$, where d_u is fixed and $\mathcal{U}$ is equipped with the semi-metric $d_\mathcal{U}$. Uniformly for all $n \geqslant n_0$ and $P \in \mathcal{P}_n$, the following conditions hold. (i) The model (6.1) is approximately sparse with sparsity index obeying $\sup_{u\in\mathcal{U}} \|\theta_u\|_0 \leqslant s$ and the growth restriction $\log(p\vee n) \leqslant \delta_n n^{1/3}$. (ii) The set $\mathcal{U}$ has uniform covering entropy obeying $\log N(\epsilon, \mathcal{U}, d_\mathcal{U}) \leqslant d_u \log(1/\epsilon) \vee 0$, and the collection $(\zeta_u = Y_u - \mathrm{E}_P[Y_u \mid X], r_u)_{u\in\mathcal{U}}$ is a suitably measurable transformation of W and u. (iii) Uniformly over $u \in \mathcal{U}$ the moments of the model satisfy $\max_{j\leqslant p}\mathrm{E}_P[|f_j(X)|^3] \leqslant C$, and $\underline{c} \leqslant \mathrm{E}_P[Y_u \mid X] \leqslant 1 - \underline{c}$ a.s. (iv) For a sequence K_n, 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.; and $K_n^2 s^2 \log^2(p\vee n) \leqslant \delta_n n$. (b) With probability $1 - \Delta_n$, $\sup_{u\in\mathcal{U}} \mathbb{E}_n[r_u^2(X)] \leqslant Cs\log(p\vee n)/n$; $\sup_{u\in\mathcal{U}} \max_{j\leqslant p} |(\mathbb{E}_n - \mathrm{E}_P)[f_j^2(X)\zeta_u^2]| \leqslant$*

δ_n; $\sup_{u,u' \in \mathcal{U}, d_{\mathcal{U}}(u,u') \leqslant 1/n} \max_{j \leqslant p} \{\mathbb{E}_n[f_j(X)^2(\zeta_u - \zeta_{u'})^2]\}^{1/2} \leqslant \delta_n$, *and* $\sup_{u,u' \in \mathcal{U}, d_{\mathcal{U}}(u,u') \leqslant 1/n} \|\mathbb{E}_n[f(X)(\zeta_u - \zeta_{u'})]\|_\infty \leqslant \delta_n n^{-1/2}$. *(c) With probability* $1-\Delta_n$, *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 (6.2) and (6.3) for the logistic link function case under Assumption 6.2.

Theorem 6.2 (Rates and Sparsity for Functional Response under Logistic Link)**.** *Under Assumption 6.2 and setting the penalty and loadings 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*

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

Comment 6.1. We note that the performance bounds satisfy the conditions of Assumptions 4.2 and 4.3. Moreover, since in the logistic case the link function is 1-Lipschitz and the approximation errors are assumed to be small, the results above establish the same rates of convergence for the estimators of the conditional probabilities, for example

$$\sup_{u \in \mathcal{U}} \|\mathrm{E}_P[Y_u \mid X] - \Lambda(f(X)'\widehat{\theta}_u)\|_{\mathbb{P}_n,2} \leqslant \bar{C}\sqrt{\frac{s \log(p \vee n)}{n}}.$$

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

As a practical illustration of the methods developed 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). Our goal here is to explain the practical implementation details of our methods, to illustrate how to interpret the estimation results and inference statements, and to make the following points that underscore our theoretical findings: 1) In a *low-dimensional setting*, where the number of controls is low and therefore there is no need for selection, our robust post-selection inference methods perform well. That is, the results of our methods agree with the results of standard methods that do not employ any selection. 2) In a *high-dimensional* setting, where there are (moderately) many controls, our post-selection inference methods perform well, producing well-behaved estimates and confidence intervals compared to the erratic estimates and confidence intervals produced by standard methods that do not employ selection as a means of regularization. 3) Finally, in a *very high-dimensional* setting, where the number of controls is comparable to the

sample size, the standard methods break down completely, while our methods still produce well-behaved estimates and confidence intervals. These findings are in line with our theoretical results about uniform validity of our inference methods.

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 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 high 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.[20] 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

[20]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).

relatively low-dimensional set that adequately captures the effect of confounds. 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, Y, in our analysis: net total financial assets[21] and total wealth.[22] Our treatment variable, D, is an indicator for having positive 401(k) balances; and our instrument, Z, is an indicator for being eligible to enroll in a 401(k) plan. The vector of raw covariates, 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 detailed results for five different sets of controls $f(X)$. The first set 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 (Indicator 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 second, second, fourth, and eighth order polynomials in family size, education, age, and income, respectively (Orthogonal Polynomials specification).[23] The fourth specification augments the Orthogonal Polynomials specification with all two-way interactions of the sets of variables from the Orthogonal Polynomials specification (Orthogonal Polynomials plus interactions specification). The final specification forms a larger set of potential controls by starting with all of the variables from the Orthogonal Polynomials specification and forming all two-way interactions between all of the non-income variables. The set of main effects and interactions of all non-income variables is then fully interacted with all of the income terms (Orthogonal Polynomials plus many interactions). The dimensions of the set of controls are thus 20, 167, 22, 196, and 756 for the Indicator, Indicator plus interactions, Orthogonal Polynomials, Orthogonal Polynomials plus

[21]Net total financial assets are defined as the sum of IRA balances, 401(k) balances, 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.

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

[23]The polynomials in each variable are orthogonalized using the Gram-Schmidt process. Note that the polynomials are not orthogonal across variables; e.g. the age and income polynomials may be correlated.

interactions specifications, and Orthogonal Polynomials plus many interactions, respectively. We refer to the specifications without interactions as *low-p*, to the specifications with only two-way interactions as *high-p*, and to the specification with two- and three-way interactions as *very-high-p*.

We report estimates of the LATE, LATE-T, LQTE, and LQTE-T for each specification. 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 functions 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 functions where $Y_u = 1(Y \leqslant u)$ is the outcome by logistic regression when no model selection is used or Post-ℓ_1-penalized logistic regression when selection is used. We only report selection-based estimates in the very-high-p setting.[24] We use the penalty level given in (6.5) 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 (6.5) with $d_u = 0$. This choice corresponds to that used in Belloni, Chernozhukov, and Hansen (2014). We refer to the supplementary appendix for further details on implementation.

Estimates of the LATE and LATE-T are given in Table 1. In this table, we provide point estimates for each of the five 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 with Mammen (1993) weights as multipliers. 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 six of the ten sets of estimates reported, suggesting positive and significant effects of 401(k) participation on net financial assets and total wealth with larger effects for treated compliers than for untreated compliers. This similarity is reassuring in the Indicator and Orthogonal Polynomials specifications as it illustrates that there is little impact of variable selection relative to simply including everything in a low-dimensional setting.[25] The two cases where we observe substantively different results are in the Orthogonal Polynomials plus interactions and Orthogonal Polynomials plus many interactions specifications. Both the LATE and LATE-T point estimates are of implausible magnitudes and have very large estimated standard errors in the Orthogonal Polynomials plus interactions case.

[24]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 Orthogonal Polynomials plus interactions (11 observations trimmed) and Orthogonal Polynomials plus many interactions specifications (9915 observations trimmed). We choose not to report unregularized estimates in the very-high-p specification since all observations are trimmed and, in fact, have estimated propensity scores of either 0 or 1.

[25]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.

Estimates cannot even be computed reliably in the Orthogonal Polynomials plus many interactions case due to the empirical failure of the identification condition due to the estimated propensity score hitting the boundary of 0 or 1 for every observation.

One would favor these imprecise estimates produced in the Orthogonal Polynomials plus interactions and Orthogonal Polynomials plus many interactions specifications 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 estimates and standard errors based on the both of these specifications following variable selection are sensible and similar to the other estimates. The similarity in the point 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 interactions specification, the Orthogonal Polynomials specification, the Orthogonal Polynomials plus interactions specification, and the Orthogonal Polynomials plus many interactions specification in Figures 1-5 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 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 low-p Indicator and Orthogonal Polynomials specifications. In the more flexible high-p 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 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 larger than the effect on total wealth in the upper quantiles, 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 controls. 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.[26] Finally, it is worth noting that this similarity is not mechanical or otherwise built in to the procedure. For example, applications in Belloni, Chen, Chernozhukov, and Hansen (2012) and Belloni, Chernozhukov, and Hansen (2014) use high-dimensional variable selection methods and produce sets of variables that differ substantially from intuitive baselines.

Appendix A. Notation

A.1. **Overall Notation.** We consider a random element $W = W_P$ taking values in the measure space $(\mathcal{W}, \mathcal{A}_{\mathcal{W}})$, with probability law $P \in \mathcal{P}$. Note that it is most convenient to think about P as a parameter in a parameter set $\mathcal{P}$. We shall also work with a bootstrap multiplier variable ξ taking values in $(\mathbb{R}, \mathcal{A}_{\mathbb{R}})$ that is independent of W_P, having probability law P_ξ, which is fixed throughout. We consider $(W_i)_{i=1}^\infty = (W_{i,P})_{i=1}^\infty$ and $(\xi_i)_{i=1}^\infty$ to be i.i.d. copies of W and ξ, which are also independent of each other. The data will be defined as some measurable function of W_i for $i = 1, ..., n$, where n denotes the sample size.

We require the sequences $(W_i)_{i=1}^\infty$ and $(\xi_i)_{i=1}^\infty$ to live on a probability space $(\Omega, \mathcal{A}_\Omega, \mathrm{P}_P)$ for all $P \in \mathcal{P}$; note that other variables arising in the proofs do not need to live on the same space. It is important to keep track of the dependence on P in the analysis since we want the results to hold uniformly in P in some set $\mathcal{P}_n$, which may be dependent on n, namely it will typically increase with n, i.e. $\mathcal{P}_n \subseteq \mathcal{P}_{n+1}$.

Throughout the paper we signify the dependence on P by mostly using P as a subscript in P_P, but in the proofs we sometimes use it as a subscript for variables as in W_P. The operator E denotes

[26] 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.

a generic expectation operator with respect to a generic probability measure P, while E_P denotes the expectation with respect to P_P. Note also that we use capital letters such as W to denote random elements and use the corresponding lower case letters such as w to denote fixed values that these random elements can take.

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

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

indexed by a measurable class of functions $\mathcal{F} : \mathcal{W} \mapsto \mathbb{R}$; see van der Vaart and Wellner (1996, chap. 2.3). We shall often omit the index P from $\mathbb{G}_{n,P}$ and simply write $\mathbb{G}_n$. 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, $\|v\|_1 = |v_1| + \cdots + |v_p|$, and $\|v\|$ denotes the Euclidean norm of v, that is, $\|v\| = \sqrt{v'v}$.

We say that a collection of random variables $\mathcal{F} = \{f(W,t), t \in T\}$, where $f : \mathcal{W} \times T \to \mathbb{R}$, indexed by a set T and viewed as functions of $W \in \mathcal{W}$, is *suitably measurable* with respect to W if it is image admissible Suslin class, as defined in Dudley (1999), p 186. In particular, $\mathcal{F}$ is suitably measurable if $f : \mathcal{W} \times T \to \mathbb{R}$ is measurable and T is a Polish space equipped with its Borel sigma algebra, see Dudley (1999), p 186. This condition is a mild assumption satisfied in practical cases.

For a positive integer k, $[k]$ denotes the set $\{1, \ldots, k\}$.

A.2. **Notation for Stochastic Convergence Uniformly in P.** All parameters, such as the law of the data, are indexed by P. This dependency is sometimes kept implicit. We shall allow for 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 the 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 $X_{n,P}$ and a random element Y_P, taking values in the normed space $\mathbb{D}$, defined on the probability space $(\Omega, \mathcal{A}_\Omega, \mathrm{P}_P)$. 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) $X_{n,P} = O_P(a_n)$ uniformly in $P \in \mathcal{P}_n$, if $\lim_{K\to\infty} \lim_{n\to\infty} \sup_{P\in\mathcal{P}_n} \mathrm{P}_P^*(|X_{n,P}| > Ka_n) = 0$,
(ii) $X_{n,P} = o_P(a_n)$ uniformly in $P \in \mathcal{P}_n$, if $\sup_{K>0} \lim_{n\to\infty} \sup_{P\in\mathcal{P}_n} \mathrm{P}_P^*(|X_{n,P}| > Ka_n) = 0$,
(iii) $X_{n,P} \rightsquigarrow Y_P$ uniformly in $P \in \mathcal{P}_n$, if $\sup_{P\in\mathcal{P}_n} \sup_{h\in \mathrm{BL}_1(\mathbb{D})} |\mathrm{E}_P^* h(X_{n,P}) - \mathrm{E}_P h(Y_P)| \to 0$.

Here the symbol $\rightsquigarrow$ denotes weak convergence, i.e. convergence in distribution or law, $\mathrm{BL}_1(\mathbb{D})$ denotes the space of functions mapping $\mathbb{D}$ to $[0,1]$ with Lipschitz norm at most 1, and the outer probability and expectation, P_P^* and $\mathrm{E}_{\mathrm{P}}^*$, are invoked whenever (non)-measurability arises.

Lemma A.1. *The above notions (i), (ii) and (iii) are equivalent to the following notions (a), (b), and (c), each holding for every sequence $P_n \in \mathcal{P}_n$:*

(a) $X_{n,P_n} = O_{P_n}(a_n)$, *i.e.* $\lim_{K\to\infty} \lim_{n\to\infty} \mathrm{P}^*_{P_n}(|X_{n,P_n}| > Ka_n) = 0$;
(b) $X_{n,P_n} = o_{P_n}(a_n)$, *i.e.* $\sup_{K>0} \lim_{n\to\infty} \mathrm{P}^*_{P_n}(|X_{n,P_n}| > Ka_n) = 0$;
(c) $X_{n,P_n} \rightsquigarrow Y_{P_n}$, *i.e.* $\sup_{h\in \mathrm{BL}_1(\mathbb{D})} |\mathrm{E}^*_{P_n} h(X_{n,P_n}) - \mathrm{E}_{P_n} h(Y_{P_n})| \to 0$.

The claims follow straightforwardly from the definitions, so the proof is omitted. We shall use this equivalence extensively in the proofs of the main results without explicit reference.

Appendix B. Key Tools I: Uniform in P Donsker Theorem, Multiplier Bootstrap, and Functional Delta Method

B.1. **Uniform in P Donsker Property.** Let $(W_i)_{i=1}^\infty$ be a sequence of i.i.d. copies of the random element W taking values in the measure space $(\mathcal{W}, \mathcal{A}_\mathcal{W})$ according to the probability law P on that space. Let $\mathcal{F}_P = \{f_{t,P} : t \in T\}$ be a set of suitably measurable functions $w \mapsto f_{t,P}(w)$ mapping $\mathcal{W}$ to $\mathbb{R}$, equipped with a measurable envelope $F_P : \mathcal{W} \mapsto \mathbb{R}$. The class is 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 $\|\cdot\|_{Q,2}$ for Q a finitely-discrete measure on $(\mathcal{W}, \mathcal{A}_\mathcal{W})$. We shall use the following result.

Theorem B.1 (**Uniform in P Donsker Property**). *Work with the set-up above. Suppose that for $q > 2$*

$$\sup_{P\in\mathcal{P}} \|F_P\|_{P,q} \leqslant C \text{ and } \lim_{\delta\searrow 0} \sup_{P\in\mathcal{P}} \sup_{d_T(t,\bar{t})\leqslant\delta} \|f_{t,P} - f_{\bar{t},P}\|_{P,2} = 0. \tag{B.1}$$

Furthermore, suppose that

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

Let $\mathbb{G}_P$ denote the P-Brownian Bridge, and consider

$$Z_{n,P} := (Z_{n,P}(t))_{t\in T} := (\mathbb{G}_n(f_{t,P}))_{t\in T}, \quad Z_P := (Z_P(t))_{t\in T} := (\mathbb{G}_P(f_{t,P}))_{t\in T}.$$

(a) Then, $Z_{n,P} \rightsquigarrow Z_P$ in $\ell^\infty(T)$ uniformly in $P \in \mathcal{P}$, namely

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

(b) The process $Z_{n,P}$ is stochastically equicontinuous uniformly in $P \in \mathcal{P}$, i.e., for every $\varepsilon > 0$,

$$\lim_{\delta \searrow 0} \limsup_{n \to \infty} \sup_{P \in \mathcal{P}} \mathrm{P}_P^* \left(\sup_{d_T(t,\bar{t}) \leqslant \delta} |Z_{n,P}(t) - Z_{n,P}(\bar{t})| > \varepsilon \right) = 0.$$

(c) The limit process Z_P has the following continuity properties:

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

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

Comment B.1. [Important Feature of the Theorem] This is an extension of the uniform Donsker theorem stated in Theorem 2.8.2 in van der Vaart and Wellner (1996), which allows for the function classes $\mathcal{F}$ to be **dependent on** P themselves. This generalization is crucial and is required in all of our problems.

B.2. **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 $\mathrm{E}\xi = 0$, $\mathrm{E}\xi^2 = 1$, and $\mathrm{E}|\xi|^q \leqslant C$ for $q > 2$. Consider the multiplier empirical process:

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

Here $\mathbb{G}_n$ is taken to be an extended empirical processes defined by the empirical measure that assigns mass $1/n$ to each point (W_i, ξ_i) for $i = 1, ..., n$. Let $Z_P = (Z_P(t))_{t \in T} = (\mathbb{G}_P(f_{t,P}))_{t \in T}$ as defined in Theorem B.1.

Theorem B.2 (**Uniform in P Validity of Multiplier Bootstrap**). *Assume the conditions of Theorem B.1 hold. 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*

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

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

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

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

B.3. **Uniform in P Functional Delta Method and Bootstrap.** We shall use the functional delta method, as formulated in van der Vaart and Wellner (1996, Chap. 3.9). 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 $\rho \in \mathbb{D}_\phi$ tangentially to $\mathbb{D}_0$ if there is a continuous linear map $\phi'_\rho : \mathbb{D}_0 \mapsto \mathbb{E}$ such that

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

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

We now define the following notion of the uniform Hadamard differentiability:

Definition B.1 (**Uniform Hadamard Tangential Differentiability**). *Consider a map $\phi : \mathbb{D}_\phi \mapsto \mathbb{E}$, where the domain of the map $\mathbb{D}_\phi$ is a subset of a normed space $\mathbb{D}$ and the range is a subset of the normed space $\mathbb{E}$. Let $\mathbb{D}_0$ be a normed space, with $\mathbb{D}_0 \subset \mathbb{D}$, and $\mathbb{D}_\rho$ be a compact metric space, a subset of $\mathbb{D}_\phi$. The map $\phi : \mathbb{D}_\phi \mapsto \mathbb{E}$ is called Hadamard-differentiable uniformly in $\rho \in \mathbb{D}_\rho$ tangentially to $\mathbb{D}_0$ with derivative map $h \mapsto \phi'_\rho(h)$, if*

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

for all convergent sequences $\rho_n \to \rho$ in $\mathbb{D}_\rho$, $t_n \to 0$ in $\mathbb{R}$, and $h_n \to h \in \mathbb{D}_0$ in $\mathbb{D}$ such that $\rho_n + t_n h_n \in \mathbb{D}_\phi$ for every n. As a part of the definition, we require that the derivative map $h \mapsto \phi'_\rho(h)$ from $\mathbb{D}_0$ to $\mathbb{E}$ is linear for each $\rho \in \mathbb{D}_\rho$. ■

Comment B.2. Note that the definition requires that that the derivative map $(\rho, h) \mapsto \phi'_\rho(h)$, mapping $\mathbb{D}_\rho \times \mathbb{D}_0$ to $\mathbb{E}$, is continuous at each $(\rho, h) \in \mathbb{D}_\rho \times \mathbb{D}_0$. ■

Comment B.3 (**Important Details of the Definition**). Definition B.1 is different from the definition of uniform differentiability given in van der Vaart and Wellner (1996, p. 379, eq. (3.9.12)), since our definition allows $\mathbb{D}_\rho$ to be much smaller than $\mathbb{D}_\phi$ and allows $\mathbb{D}_\rho$ to be endowed with a much stronger metric than the metric induced by the norm of $\mathbb{D}$. These differences are essential for infinite-dimensional applications. For example, the quantile/inverse map is uniformly Hadamard differentiable in the sense of Definition B.1 for a suitable choice of $\mathbb{D}_\rho$: Let $T = [\epsilon, 1-\epsilon]$, $\mathbb{D} = \ell^\infty(T)$, $\mathbb{D}_\phi$= set of cadlag functions on T, $\mathbb{D}_0 = \mathrm{UC}(T)$, and $\mathbb{D}_\rho$ be a compact subset of $C^1(T)$ such that each $\rho \in \mathbb{D}_\rho$ obeys $\partial\rho(t)/\partial t \geqslant c > 0$ on $t \in T$, where c is a positive constant. However, the quantile/inverse map is not Hadamard differentiable uniformly on $\mathbb{D}_\rho$ if we set $\mathbb{D}_\rho = \mathbb{D}_\phi$ and hence is not uniformly differentiable in the sense of the definition given in van der Vaart and Wellner (1996) which requires $\mathbb{D}_\rho = \mathbb{D}_\phi$. It is important and practical to keep the distinction between $\mathbb{D}_\rho$ and $\mathbb{D}_\phi$ since the estimated values $\widehat{\rho}$ may well be outside $\mathbb{D}_\rho$ unless explicitly imposed in estimation even though the population values of ρ are in $\mathbb{D}_\rho$ by assumption. For example, the empirical cdf is in $\mathbb{D}_\phi$, but is outside $\mathbb{D}_\rho$. ■

Theorem B.3 (**Functional delta-method uniformly in $P \in \mathcal{P}$**). *Let $\phi : \mathbb{D}_\phi \subset \mathbb{D} \mapsto \mathbb{E}$ be Hadamard-differentiable uniformly in $\rho \in \mathbb{D}_\rho \subset \mathbb{D}_\phi$ tangentially to $\mathbb{D}_0$ with derivative map ϕ'_ρ. Let $\widehat{\rho}_{n,P}$ be a sequence of stochastic processes taking values in $\mathbb{D}_\phi$, where each $\widehat{\rho}_{n,P}$ is an estimator of the parameter $\rho_P \in \mathbb{D}_\rho$. Suppose there exists a sequence of constants $r_n \to \infty$ such that $Z_{n,P} = r_n(\widehat{\rho}_{n,P} - \rho_P) \rightsquigarrow Z_P$ in $\mathbb{D}$ uniformly in $P \in \mathcal{P}_n$. The limit process Z_P is separable and takes its values in $\mathbb{D}_0$ for all $P \in \mathcal{P} = \cup_{n \geqslant n_0} \mathcal{P}_n$, where n_0 is fixed. Moreover, the set of stochastic processes $\{Z_P : P \in \mathcal{P}\}$ is relatively compact in the topology of weak convergence in $\mathbb{D}_0$, that is, every sequence in this set can be split into weakly convergent subsequences. Then, $r_n\left(\phi(\widehat{\rho}_{n,P}) - \phi(\rho_P)\right) \rightsquigarrow \phi'_{\rho_P}(Z_P)$ in $\mathbb{E}$ uniformly in $P \in \mathcal{P}_n$. If $(\rho, h) \mapsto \phi'_\rho(h)$ is defined and*

continuous on the whole of $\mathbb{D}_\rho \times \mathbb{D}$*, then the sequence* $r_n\left(\phi(\widehat{\rho}_{n,P}) - \phi(\rho_P)\right) - \phi'_{\rho_P}\left(r_n(\widehat{\rho}_{n,P} - \rho_P)\right)$ *converges to zero in outer probability uniformly in* $P \in \mathcal{P}_n$*. Moreover, the set of stochastic processes* $\{\phi'_{\rho_P}(Z_P) : P \in \mathcal{P}\}$ *is relatively compact in the topology of weak convergence in* $\mathbb{E}$*.*

The following result on the functional delta method applies to any bootstrap or other simulation method obeying certain conditions. This includes the multiplier bootstrap as a special case. Let $D_{n,P} = (W_{i,P})_{i=1}^n$ denote the data vector and $B_n = (\xi_i)_{i=1}^n$ be a vector of random variables, used to generate bootstrap or simulation draws (this may depend on the particular method). Consider sequences of stochastic processes $\widehat{\rho}_{n,P} = \widehat{\rho}_{n,P}(D_{n,P})$, where $Z_{n,P} = r_n(\widehat{\rho}_{n,P} - \rho_P) \rightsquigarrow Z_P$ in the normed space $\mathbb{D}$ uniformly in $P \in \mathcal{P}_n$. Also consider the bootstrap stochastic process $Z^*_{n,P} = Z_{n,P}(D_{n,P}, B_n)$ in $\mathbb{D}$, where $Z_{n,P}$ is a measurable function of B_n for each value of D_n. Suppose that $Z^*_{n,P}$ converges conditionally given D_n in distribution to Z_P uniformly in $P \in \mathcal{P}_n$, namely that

$$\sup_{h \in \mathrm{BL}_1(\mathbb{D})} |\mathrm{E}_{B_n}[h(Z^*_{n,P})] - \mathrm{E}_P h(Z_P)| = o^*_P(1),$$

uniformly in $P \in \mathcal{P}_n$, where E_{B_n} denotes the expectation computed with respect to the law of B_n holding the data $D_{n,P}$ fixed. This is denoted as "$Z^*_{n,P} \rightsquigarrow_B Z_P$ uniformly in $P \in \mathcal{P}_n$." Finally, let

$$\widehat{\rho}^*_{n,P} = \widehat{\rho}_{n,P} + Z^*_{n,P}/r_n$$

denote the bootstrap or simulation draw of $\widehat{\rho}_{n,P}$.

Theorem B.4 (**Uniform in** P **functional delta-method for bootstrap and other simulation methods**)**.** *Assume the conditions of Theorem B.3 hold. Let* $\widehat{\rho}_{n,P}$ *and* $\widehat{\rho}^*_{n,P}$ *be maps as indicated previously taking values in* $\mathbb{D}_\phi$ *such that* $r_n(\widehat{\rho}_{n,P} - \rho_P) \rightsquigarrow Z_P$ *and* $r_n(\widehat{\rho}^*_{n,P} - \widehat{\rho}_{n,P}) \rightsquigarrow_B Z_P$ *in* $\mathbb{D}$ *uniformly in* $P \in \mathcal{P}_n$*. Then,* $X^*_{n,P} = r_n(\phi(\widehat{\rho}^*_{n,P}) - \phi(\widehat{\rho}_{n,P})) \rightsquigarrow_B X_P = \phi'_{\rho_P}(Z_P)$ *uniformly in* $P \in \mathcal{P}_n$*.*

B.4. **Proof of Theorem B.1.** Part (a) and (b) are a direct consequence of Lemma B.2. In particular, Lemma B.2(a) implies stochastic equicontinuity under arbitrary subsequences $P_n \in \mathcal{P}$, which implies part (b). Part (a) follows from Lemma B.2(b) by splitting an arbitrary sequence $n \in \mathbb{N}$ into subsequences $n \in \mathbb{N}'$ along each of which the covariance function

$$(t, s) \mapsto c_{P_n}(t, s) := P_n f_{s,P_n} f_{t,P_n} - P_n f_{s,P_n} P_n f_{t,P_n}$$

converges uniformly and therefore also pointwise to a uniformly continuous function on (T, d_T). This is possible because $\{(t, s) \mapsto c_P(t, s) : P \in \mathcal{P}\}$ is a relatively compact set in $\ell^\infty(T \times T)$ in view of the Arzela-Ascoli Theorem, the assumptions in equation (B.1), and total boundedness of (T, d_T). By Lemma B.2(b) pointwise convergence of the covariance function implies weak convergence to a tight Gaussian process which may depend on the identity $\mathbb{N}'$ of the subsequence. Since this argument applies to each such subsequence that split the overall sequence, part (b) follows.

Part (c) is immediate from the imposed uniform covering entropy condition and Dudley's metric entropy inequality for expectations of suprema of Gaussian processes (Corollary 2.2.8 in van der

Vaart and Wellner (1996)). Claim (d) follows from claim (c) and a standard argument, based on the application of the Borel-Cantelli lemma. Indeed, let $m \in \mathbb{N}$ be a sequence and

$$\delta_m := 2^{-m} \wedge \sup \left\{ \delta > 0 : \sup_{P \in \mathcal{P}} \mathrm{E}_P \sup_{d_T(t,\bar{t}) \leqslant \delta} |Z_P(t) - Z_P(\bar{t})| < 2^{-2m} \right\},$$

then by the Markov inequality

$$\mathrm{P}_P \left(\sup_{d_T(t,\bar{t}) \leqslant \delta_m} |Z_P(t) - Z_P(\bar{t})| > 2^{-m} \right) \leqslant 2^{-2m+m} = 2^{-m}.$$

This sums to a finite number over $m \in \mathbb{N}$. Hence, by the Borel-Cantelli lemma, for almost all states $\omega \in \Omega$, $|Z_P(t)(\omega) - Z_P(\bar{t})(\omega)| \leqslant 2^{-m}$ for all $d_T(t,\bar{t}) \leqslant \delta_m \leqslant 2^{-m}$ and all m sufficiently large. Hence claim (d) follows. ∎

B.5. **Proof of Theorem B.2.** Claim (a) is verified by invoking Theorem B.1. We begin by showing that $Z_P^* = (\mathbb{G}_P \xi f_{t,P})_{t \in T}$ is equal in distribution to $Z_P = (\mathbb{G}_P f_{t,P})_{t \in T}$, in particular, Z_P^* and Z_P share identical mean and covariance function, and thus they share the continuity properties established in Theorem B.1. This claim is immediate from the fact that multiplication by ξ of each $f \in \mathcal{F}_P = \{f_{t,P} : t \in T\}$ yields a set $\xi \mathcal{F}_P$ of measurable functions $\xi f : (w, \xi) \mapsto \xi f(w)$, mapping $\mathcal{W} \times \mathbb{R}$ to $\mathbb{R}$. Each such function has mean zero under $P \times P_\xi$, i.e. $\int s f(w) dP_\xi(s) dP(w) = 0$, and the covariance function $(\xi f, \xi \tilde{f}) \mapsto P f \tilde{f} - P f P \tilde{f}$. Hence the Gaussian process $(\mathbb{G}_P(\xi f))_{\xi f \in \xi \mathcal{F}_P}$ shares the zero mean and the covariance function of $(\mathbb{G}_P(f))_{f \in \mathcal{F}_P}$.

We are claiming that $Z_{n,P}^* \rightsquigarrow Z_P^*$ in $\ell^\infty(T)$ uniformly in $P \in \mathcal{P}$, where $Z_{n,P}^* := (\mathbb{G}_n \xi f_{t,P})_{t \in T}$. We note that the function class $\mathcal{F}_P$ and the corresponding envelope F_P satisfy the conditions of Theorem B.1. The same is also true for the function class $\xi \mathcal{F}_P$ defined by $(w, \xi) \mapsto \xi f_P(w)$, which maps $\mathcal{W} \times \mathbb{R}$ to $\mathbb{R}$ and its envelope $|\xi| F_P$, since ξ is independent of W. Let Q now denote a finitely discrete measure over $\mathcal{W} \times \mathbb{R}$. By Lemma C.2 multiplication by ξ does not change qualitatively the uniform covering entropy bound:

$$\log \sup_Q N(\epsilon \| |\xi| F_P \|_{Q,2}, \xi \mathcal{F}_P, \| \cdot \|_{Q,2}) \leqslant \log \sup_Q N(2^{-1} \epsilon \| F_P \|_{Q,2}, \mathcal{F}_P, \| \cdot \|_{Q,2}).$$

Moreover, multiplication by ξ does not affect the norms, $\| \xi f_P(W) \|_{P \times P_\xi, 2} = \| f_P(W) \|_{P,2}$, since ξ is independent of W by construction and $\mathrm{E}\xi^2 = 1$. The claim then follows.

Claim (b). For each $\delta > 0$ and $t \in T$, let $\pi_\delta t$ denote a closest element in a given, finite δ-net over T. We begin by noting that

$$\begin{aligned}
\Delta_P &:= \sup_{h \in \mathrm{BL}_1} |\mathrm{E}_{B_n} h(Z_{n,P}^*) - \mathrm{E}_P h(Z_P)| \\
&\leqslant I_P + II_P + III_P := \sup_{h \in \mathrm{BL}_1} |\mathrm{E}_P h(Z_P \circ \pi_\delta) - \mathrm{E}_P h(Z_P)| \\
&\quad + \sup_{h \in \mathrm{BL}_1} |\mathrm{E}_{B_n} h(Z_{n,P}^* \circ \pi_\delta) - \mathrm{E}_P h(Z_P \circ \pi_\delta)| + \sup_{h \in \mathrm{BL}_1} |\mathrm{E}_{B_n} h(Z_{n,P}^* \circ \pi_\delta) - \mathrm{E}_{B_n} h(Z_{n,P}^*)|,
\end{aligned}$$

where here and below BL_1 abbreviates $\mathrm{BL}_1(\ell^\infty(T))$.

First, we note that

$$I_P \leqslant \mathrm{E}_P \left(\sup_{d_T(t,\bar{t}) \leqslant \delta} |Z_P(t) - Z_P(\bar{t})| \wedge 2 \right) =: \mu_P(\delta), \quad \lim_{\delta \searrow 0} \sup_{P \in \mathcal{P}} \mu_P(\delta) = 0.$$

The first assertion follows from

$$I_P \leqslant \sup_{h \in \mathrm{BL}_1} \mathrm{E}_P |h(Z^*_{n,P} \circ \pi_\delta) - h(Z^*_{n,P})| \leqslant \mathrm{E}_P \left(\sup_{t \in T} |Z_P \circ \pi_\delta(t) - Z_P(t)| \wedge 2 \right) \leqslant \mu_P(\delta),$$

and the second assertion holds by Theorem B.1 (c).

Second, we note that

$$\mathrm{E}^*_P III_P \quad \leqslant \quad \mathrm{E}^*_P \left(\sup_{d_T(t,\bar{t}) \leqslant \delta} |Z^*_{n,P}(t) - Z^*_{n,P}(\bar{t})| \wedge 2 \right) =: \mu^*_P(\delta), \quad \lim_{n \to \infty} \sup_{P \in \mathcal{P}} |\mu^*_P(\delta) - \mu_P(\delta)| = 0.$$

The first assertion follows because $\mathrm{E}^*_P III_P$ is bounded by

$$\mathrm{E}^*_P \sup_{h \in \mathrm{BL}_1} \mathrm{E}_{B_n} |h(Z^*_{n,P} \circ \pi_\delta) - h(Z^*_{n,P})| \leqslant \mathrm{E}^*_P \mathrm{E}_{B_n} \left(\sup_{t \in T} |Z^*_{n,P} \circ \pi_\delta(t) - Z^*_{n,P}(t)| \wedge 2 \right) \leqslant \mu^*_P(\delta).$$

The second assertion holds by part (a) of the present theorem. Define $\epsilon(\delta) := \delta \vee \sup_{P \in \mathcal{P}} \mu_P(\delta)$. Then, by Markov's inequality, followed by $n \to \infty$,

$$\limsup_{n \to \infty} \sup_{P \in \mathcal{P}} \mathrm{P}^*_P \left(III_P > \sqrt{\epsilon(\delta)} \right) \leqslant \limsup_{n \to \infty} \frac{\sup_{P \in \mathcal{P}} \mu^*_P(\delta)}{\sqrt{\epsilon(\delta)}} \leqslant \frac{\sup_{P \in \mathcal{P}} \mu_P(\delta)}{\sqrt{\epsilon(\delta)}} \leqslant \sqrt{\epsilon(\delta)}.$$

Finally, by Lemma B.1, for each $\varepsilon > 0$

$$\limsup_{n \to \infty} \sup_{P \in \mathcal{P}} \mathrm{P}^*_P \left(II_P > \varepsilon \right) = 0.$$

We can now conclude. Note that $\epsilon(\delta) \searrow 0$ if $\delta \searrow 0$, which holds by the definition of $\epsilon(\delta)$ and the property $\sup_{P \in \mathcal{P}} \mu_P(\delta) \searrow 0$ if $\delta \searrow 0$ noted above. Hence for each $\varepsilon > 0$ and all $0 < \delta < \delta_\varepsilon$ such that $3\sqrt{\epsilon(\delta)} < \varepsilon$,

$$\limsup_{n \to \infty} \sup_{P \in \mathcal{P}} \mathrm{P}^*_P \left(\Delta_P > \varepsilon \right) \leqslant \limsup_{n \to \infty} \sup_{P \in \mathcal{P}} \mathrm{P}^*_P \left(I_P + II_P + III_P > 3\sqrt{\epsilon(\delta)} \right) \leqslant \sqrt{\epsilon(\delta)}.$$

Sending $\delta \searrow 0$ gives the result. ∎

B.6. **Auxiliary Result: Conditional Multiplier Central Limit Theorem in $\mathbb{R}^d$ uniformly in $P \in \mathcal{P}$.** We rely on the following lemma, which is apparently new. (An analogous result can be derived for almost sure convergence from the well-known non-uniform multiplier central limit theorems, but this strategy requires us to put all the variables indexed by P on the single underlying probability space, which is much less convenient in applications.)

Lemma B.1 (Conditional Multiplier Central Limit Theorem in $\mathbb{R}^d$ uniformly in $P \in \mathcal{P}$)**.** *Let $(Z_{i,P})_{i=1}^\infty$ be i.i.d. random vectors on $\mathbb{R}^d$, indexed by a parameter $P \in \mathcal{P}$. The parameter P represents probability laws on $\mathbb{R}^d$. For each $P \in \mathcal{P}$, these vectors are assumed to be independent of the i.i.d. sequence $(\xi_i)_{i=1}^\infty$ with $\mathrm{E}\xi_1 = 0$ and $\mathrm{E}\xi_1^2 = 1$. There exist constants $2 < q < \infty$ and*

$0 < M < \infty$, such that $\mathrm{E}_P Z_{1,P} = 0$ and $(\mathrm{E}_P \|Z_{1,P}\|^q)^{1/q} \leqslant M$ uniformly for all $P \in \mathcal{P}$. Then, for every $\varepsilon > 0$

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}} \mathrm{P}_P^* \left(\sup_{h\in \mathrm{BL}_1(\mathbb{R}^d)} \left| \mathrm{E}_{B_n} h\Big(n^{-1/2} \sum_{i=1}^n \xi_i Z_{i,P}\Big) - \mathrm{E}_P h\Big(N(0, \mathrm{E}_P Z_{1,P} Z_{1,P}')\Big) \right| > \varepsilon \right) = 0,$$

where E_{B_n} denotes the expectation over $(\xi_i)_{i=1}^n$ holding $(Z_{i,P})_{i=1}^n$ fixed.

Proof of Lemma B.1. Let X and Y be random variables in $\mathbb{R}^d$, then define $d_{BL}(X,Y) := \sup_{h\in \mathrm{BL}_1(\mathbb{R}^d)} |\mathrm{E}h(X) - \mathrm{E}h(Y)|$. It suffices to show that for any sequence $P_n \in \mathcal{P}$ and $N^* \sim n^{-1/2} \sum_{i=1}^n \xi_i Z_{i,P_n} \mid (Z_{i,P_n})_{i=1}^n$, $d_{BL}\left(N^*, N(0, \mathrm{E}_{P_n} Z_{1,P_n} Z_{1,P_n}')\right) \to 0$ in probability (under P_{P_n}).

Following Bickel and Freedman (1981), we shall rely on the Mallow's metric, written m_r, which is a metric on the space of distribution functions on $\mathbb{R}^d$. For our purposes it suffices to recall that given a sequence of distribution functions $\{F_k\}$ and a distribution function F, $m_r(F_k, F) \to 0$ if and only if $\int g dF_k \to \int g dF$ for each continuous and bounded $g : \mathbb{R}^d \to \mathbb{R}$, and $\int \|z\|^r dF_k(z) \to \int \|z\|^r dF(z)$. See Bickel and Freedman (1981) for the definition of m_r.

Under the assumptions of the lemma, we can split the sequence $n \in \mathbb{N}$ into subsequences $n \in \mathbb{N}'$, along each of which the distribution function of Z_{1,P_n} converges to some distribution function F' with respect to the Mallow's metric m_r, for some $2 < r < q$. This also implies that $N(0, \mathrm{E}_{P_n} Z_{1,P_n} Z_{1,P_n}')$ converges weakly to a normal limit $N(0, Q')$ with $Q' = \int zz' dF'(z)$ such that $\|Q'\| \leqslant M$. Both Q' and F' can depend on the subsequence $\mathbb{N}'$.

Let F_k be the empirical distribution function of a sequence $(z_i)_{i=1}^k$ of constant vectors in $\mathbb{R}^d$, where $k \in \mathbb{N}$. The law of $N_{F_k}^* = k^{-1/2} \sum_{i=1}^k \xi_i z_i$ is completely determined by F_k and the law of ξ (the latter is fixed, so it does not enter as the subscript in the definition of $N_{F_k}^*$). If $m_r(F_k, F') \to 0$ as $k \to \infty$, then $d_{BL}(N_{F_k}^*, N(0, Q')) \to 0$ by Lindeberg's central limit theorem.

Let $\mathbb{F}_n$ denote the empirical distribution function of $(Z_{i,P_n})_{i=1}^n$. Note that $N^* = N_{\mathbb{F}_n}^* \sim n^{-1/2} \sum_{i=1}^n \xi_i Z_{i,P_n} \mid (Z_{i,P_n})_{i=1}^n$. By the law of large numbers for arrays, $\int g d\mathbb{F}_n \to \int g dF'$ and $\int \|z\|^r d\mathbb{F}_n(z) \to \int \|z\|^r dF'(z)$ in probability along the subsequence $n \in \mathbb{N}'$. Hence $m_r(\mathbb{F}_n, F') \to 0$ in probability along the same subsequence. We can conclude that $d_{BL}(N_{\mathbb{F}_n}^*, N(0, Q')) \to 0$ in probability along the same subsequence by the extended continuous mapping theorem (van der Vaart and Wellner, 1996, Theorem 1.11.1).

The argument applies to every subsequence $\mathbb{N}'$ of the stated form. The claim in the first paragraph of the proof thus follows. ∎

B.7. **Donsker Theorems for Function Classes that depend on n.** Let $(W_i)_{i=1}^\infty$ be a sequence of i.i.d. copies of the random element W taking values in the measure space $(\mathcal{W}, \mathcal{A}_\mathcal{W})$, whose law is determined by the probability measure P, and let $w \mapsto f_{n,t}(w)$ be measurable functions $f_{n,t} : \mathcal{W} \to \mathbb{R}$ 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 a measurable envelope function F_n. It is important to note here that the dependency on n allows us to have *the class itself* be possibly dependent on the law P_n.

Lemma B.2 (**Donsker Theorem for Classes Changing with** n)**.** *Work with the set-up above. Suppose that for some fixed constant $q > 2$ and every sequence $\delta_n \searrow 0$:*

$$\|F_n\|_{P_n,q} = O(1), \sup_{d_T(s,t)\leqslant\delta_n} \|f_{n,s} - f_{n,t}\|_{P_n,2} \to 0,$$

$$\int_0^{\delta_n} \sup_Q \sqrt{\log N(\epsilon \|F_n\|_{Q,2}, \mathcal{F}_n, \|\cdot\|_{Q,2})} d\epsilon \to 0.$$

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

Proof. This is merely a restatement for subsequences of Theorem 2.11.22 in van der Vaart and Wellner (1996, p. 220-221), stated for sequences. ∎

B.8. **Proof of Theorems B.3 and B.4.** The proof consists of two parts, each proving the corresponding theorem.

Part 1. We can split $\mathbb{N}$ into subsequences $\{\mathbb{N}'\}$ along each of which

$$Z_{n,P_n} \rightsquigarrow Z' \in \mathbb{D}_0 \text{ in } \mathbb{D}, \quad \rho_{P_n} \to \rho' \text{ in } \mathbb{D}_\rho \quad (n \in \mathbb{N}'),$$

where Z' and ρ' can possibly depend on $\mathbb{N}'$. It suffices to verify that for each $\mathbb{N}'$:

$$r_n(\phi(\widehat{\rho}_{n,P_n}) - \phi(\rho_{P_n})) \rightsquigarrow \phi'_{\rho'}(Z') \quad (n \in \mathbb{N}') \tag{B.3}$$

$$r_n(\phi(\widehat{\rho}_{n,P_n}) - \phi(\rho_{P_n})) - \phi'_{\rho_{P_n}}(r_n(\widehat{\rho}_{n,P_n} - \rho_{P_n})) \rightsquigarrow 0 \quad (n \in \mathbb{N}'), \tag{B.4}$$

$$r_n(\phi(\widehat{\rho}_{n,P_n}) - \phi(\rho_{P_n})) - \phi'_{\rho'}(r_n(\widehat{\rho}_{n,P_n} - \rho_{P_n})) \rightsquigarrow 0 \quad (n \in \mathbb{N}'), \tag{B.5}$$

where the last two claims hold provided that $(\rho, h) \mapsto \phi'_\rho(h)$ is defined and continuous on the whole of $\mathbb{D}_\rho \times \mathbb{D}$. The claim (B.5) is not needed in Part 1, but we need it for the Part 2.

The map $g_n(h) = r_n(\phi(\rho_{P_n} + r_n^{-1}h) - \phi(\rho_{P_n}))$, from $\mathbb{D}_n = \{h \in \mathbb{D} : \rho_{P_n} + r_n^{-1}h \in \mathbb{D}_\phi\}$ to $\mathbb{E}$, satisfies $g_n(h_n) \to \phi'_{\rho'}(h)$ for every subsequence $h_n \to h \in \mathbb{D}_0$ (with $n \in \mathbb{N}'$). Application of the extended continuous mapping theorem (van der Vaart and Wellner, 1996, Theorem 1.11.1) yields (B.3).

Similarly, the map $m_n(h) = r_n(\phi(\rho_{P_n} + r_n^{-1}h) - \phi(\rho_{P_n})) - \phi'_{\rho_{P_n}}(h)$, from $\mathbb{D}_n = \{h \in \mathbb{D} : \rho_{P_n} + r_n^{-1}h \in \mathbb{D}_\phi\}$ to $\mathbb{E}$, satisfies $m_n(h_n) \to \phi'_{\rho'}(h) - \phi'_{\rho'}(h) = 0$ for every subsequence $h_n \to h \in \mathbb{D}_0$

(with $n \in \mathbb{N}'$). Application of the extended continuous mapping theorem (van der Vaart and Wellner, 1996, Theorem 1.11.1) yields (B.4). The proof of (B.5) is completely analogous and is omitted.

To establish relative compactness, work with each $\mathbb{N}'$. Then $\phi'_{\rho_{P_n}}(h)$ mapping $\mathbb{D}_0$ to $\mathbb{E}$ satisfies $\phi'_{\rho_{P_n}}(h_n) \to \phi'_{\rho'}(h)$ for every subsequence $h_n \to h \in \mathbb{D}_0$ (with $n \in \mathbb{N}'$). Application of the extended continuous mapping theorem (van der Vaart and Wellner, 1996, Theorem 1.11.1) yields that $\phi'_{\rho_{P_n}}(Z_P) \rightsquigarrow \phi'_{\rho'}(Z')$.

Part 2. We can split $\mathbb{N}$ into subsequences $\{\mathbb{N}'\}$ as above. Along each $\mathbb{N}'$,

$$r_n(\widehat{\rho}^*_{n,P_n} - \rho_{P_n}) \rightsquigarrow Z'' \in \mathbb{D}_0 \text{ in } \mathbb{D}, \quad r_n(\widehat{\rho}_{n,P_n} - \rho_{P_n}) \rightsquigarrow Z' \in \mathbb{D}_0 \text{ in } \mathbb{D}, \quad \rho_{P_n} \to \rho' \text{ in } \mathbb{D}_\rho \quad (n \in \mathbb{N}'),$$

where Z'' is a separable process in $\mathbb{D}_0$ (which is given by Z' plus its independent copy $\bar{Z}'$). Indeed, note that $r_n(\widehat{\rho}^*_{\rho_{n,P_n}} - \rho_{P_n}) = Z^*_{n,P_n} + Z_{n,P_n}$, and (Z^*_{n,P_n}, Z_{n,P_n}) converge weakly unconditionally to $(\bar{Z}', Z')$ by a standard argument.

Given each $\mathbb{N}'$ the proof is similar to the proof of Theorem 3.9.15 of van der Vaart and Wellner (1996). We can assume without loss of generality that the derivative $\phi'_{\rho'} : \mathbb{D} \to \mathbb{E}$ is defined and continuous on the whole of $\mathbb{D}$. Otherwise, if $\phi'_{\rho'}$ is defined and continuous only on $\mathbb{D}_0$, we can extend it to $\mathbb{D}$ by a Hahn-Banach extension such that $C = \|\phi'_{\rho'}\|_{\mathbb{D}_0 \to \mathbb{E}} = \|\phi'_{\rho'}\|_{\mathbb{D} \to \mathbb{E}} < \infty$; see van der Vaart and Wellner (1996, p. 380) for details. For each $\mathbb{N}'$, by claim (B.5), applied to $\widehat{\rho}_{n,P_n}$ and to $\widehat{\rho}^*_{n,P_n}$ replacing $\widehat{\rho}_{n,P_n}$,

$$r_n(\phi(\widehat{\rho}_{n,P_n}) - \phi(\rho_{P_n})) = \phi'_{\rho'}(r_n(\widehat{\rho}_{n,P_n} - \rho_{P_n})) + o^*_{P_n}(1),$$
$$r_n(\phi(\widehat{\rho}^*_{n,P_n}) - \phi(\rho_{P_n})) = \phi'_{\rho'}(r_n(\widehat{\rho}^*_{n,P_n} - \rho_{P_n})) + o^*_{P_n}(1).$$

Subtracting these equations conclude that for each $\varepsilon > 0$

$$\mathrm{E}_{P_n} 1\left(\left\| r_n(\phi(\widehat{\rho}^*_{n,P_n}) - \phi(\widehat{\rho}_{n,P_n})) - \phi'_{\rho'}(r_n(\widehat{\rho}^*_{n,P_n} - \widehat{\rho}_{n,P_n}))\right\|^*_{\mathbb{E}} > \varepsilon\right) \to 0 \quad (n \in \mathbb{N}'). \tag{B.6}$$

For every $h \in \mathrm{BL}_1(\mathbb{E})$, the function $h \circ \phi'_{\rho'}$ is contained in $\mathrm{BL}_C(\mathbb{D})$. Moreover, $r_n(\widehat{\rho}^*_{n,P} - \widehat{\rho}_{n,P}) \rightsquigarrow_B Z_P$ in $\mathbb{D}$ uniformly in $P \in \mathcal{P}_n$ implies $r_n(\widehat{\rho}^*_{n,P} - \widehat{\rho}_{n,P}) \rightsquigarrow_B Z'$ along the subsequence $n \in \mathbb{N}'$. These two facts imply that

$$\sup_{h \in \mathrm{BL}_1(\mathbb{E})} \left| \mathrm{E}_{B_n} h\Big(\phi'_{\rho'}(r_n(\widehat{\rho}^*_{n,P_n} - \widehat{\rho}_{n,P_n}))\Big) - \mathrm{E} h(\phi_{\rho'}(Z'))\right| = o^*_{P_n}(1) \quad (n \in \mathbb{N}').$$

Next for each $\varepsilon > 0$ and along $n \in \mathbb{N}'$

$$\sup_{h \in \mathrm{BL}_1(\mathbb{E})} \left| \mathrm{E}_{B_n} h\Big(r_n(\phi(\widehat{\rho}^*_{n,P_n}) - \phi(\widehat{\rho}_{n,P_n}))\Big) - \mathrm{E}_{B_n} h\Big(\phi'_{\rho'}(r_n(\widehat{\rho}^*_{n,P_n} - \widehat{\rho}_{n,P_n}))\Big)\right|$$
$$\leqslant \varepsilon + 2\mathrm{E}_{B_n} 1\left(\left\| r_n(\phi(\widehat{\rho}^*_{n,P_n}) - \phi(\widehat{\rho}_{n,P_n})) - \phi'_{\rho'}(r_n(\widehat{\rho}^*_{n,P_n} - \widehat{\rho}_{n,P_n}))\right\|^*_{\mathbb{E}} > \varepsilon\right) = o_{P_n}(1),$$

where the $o_{P_n}(1)$ conclusion follows by the Markov inequality and by (B.6). Conclude that

$$\sup_{h \in \mathrm{BL}_1(\mathbb{E})} \left| \mathrm{E}_{B_n} h\Big(r_n(\phi(\widehat{\rho}^*_{n,P_n}) - \phi(\widehat{\rho}_{n,P_n}))\Big) - \mathrm{E} h(\phi_{\rho'}(Z'))\right| = o^*_{P_n}(1) \quad (n \in \mathbb{N}').$$

■

Appendix C. Key Tools II: Probabilistic Inequalities

Let $(W_i)_{i=1}^n$ be a sequence of i.i.d. copies of random element W taking values in the measure space $(\mathcal{W}, \mathcal{A}_\mathcal{W})$ according to probability law P. Let $\mathcal{F}$ be a set of suitably measurable functions $f : \mathcal{W} \mapsto \mathbb{R}$, equipped with a measurable envelope $F : \mathcal{W} \mapsto \mathbb{R}$.

The following maximal inequality is due to Chernozhukov, Chetverikov, and Kato (2012).

Lemma C.1 (**A Maximal Inequality**). *Work with the setup above. Suppose that $F \geqslant \sup_{f\in\mathcal{F}} |f|$ is a measurable envelope with $\|F\|_{P,q} < \infty$ for some $q \geqslant 2$. Let $M = \max_{i\leqslant n} F(W_i)$ and $\sigma^2 > 0$ be any positive constant such that $\sup_{f\in\mathcal{F}} \|f\|_{P,2}^2 \leqslant \sigma^2 \leqslant \|F\|_{P,2}^2$. Suppose that there exist constants $a \geqslant e$ and $v \geqslant 1$ such that*

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

Then

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

where K is an absolute constant. Moreover, for every $t \geqslant 1$, with probability $> 1 - t^{-q/2}$,

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

where $K(q) > 0$ is a constant depending only on q. In particular, setting $a \geqslant n$ and $t = \log n$, with probability $> 1 - c(\log n)^{-1}$,

$$\|\mathbb{G}_n\|_\mathcal{F} \leqslant K(q,c)\left(\sigma\sqrt{v\log\left(\frac{a\|F\|_{P,2}}{\sigma}\right)} + \frac{v\|M\|_{\mathrm{P},q}}{\sqrt{n}}\log\left(\frac{a\|F\|_{P,2}}{\sigma}\right)\right), \tag{C.1}$$

where $\|M\|_{\mathrm{P},q} \leqslant n^{1/q}\|F\|_{P,q}$ and $K(q,c) > 0$ is a constant depending only on q and c.

Lemma C.2 (**Algebra for Covering Entropies**). *Work with the setup above.*
(1) Let $\mathcal{F}$ be a VC subgraph class with a finite VC index k or any other class whose entropy is bounded above by that of such a VC subgraph class, then the covering entropy of $\mathcal{F}$ obeys:

$$\sup_Q \log N(\epsilon\|F\|_{Q,2}, \mathcal{F}, \|\cdot\|_{Q,2}) \lesssim 1 + k\log(1/\epsilon) \vee 0$$

(2) For any measurable classes of functions $\mathcal{F}$ and $\mathcal{F}'$ mapping $\mathcal{W}$ to $\mathbb{R}$

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

(3) Given a measurable class $\mathcal{F}$ mapping $\mathcal{W}$ to $\mathbb{R}$ and a random variable ξ taking values in $\mathbb{R}$,

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

(4) Given measurable classes $\mathcal{F}_j$ and envelopes F_j, $j = 1,\ldots,k$, mapping $\mathcal{W}$ to $\mathbb{R}$, a function $\phi : \mathbb{R}^k \to \mathbb{R}$ such that for $f_j, g_j \in \mathcal{F}_j$, $|\phi(f_1,\ldots,f_k) - \phi(g_1,\ldots,g_k)| \leqslant \sum_{j=1}^k L_j(x)|f_j(x) - g_j(x)|$,

$L_j(x) \geqslant 0$, and fixed functions $\bar{f}_j \in \mathcal{F}_j$, the class of functions $\mathcal{L} = \{\phi(f_1, \ldots, f_k) - \phi(\bar{f}_1, \ldots, \bar{f}_k) : f_j \in \mathcal{F}_j, j = 1, \ldots, k\}$ satisfies

$$\log \sup_Q N(\epsilon \| \sum_{j=1}^k L_j F_j \|_{Q,2}, \mathcal{L}, \| \cdot \|_{Q,2}) \leqslant \sum_{j=1}^k \log \sup_Q N\left(\tfrac{\epsilon}{k} \| F_j \|_{Q,2}, \mathcal{F}_j, \| \cdot \|_{Q,2}\right).$$

Proof. For the proof (1)-(2) see, e.g., Andrews (1994a) and (3) follows from (2). To show (4) let $f = (f_1, \ldots, f_k)$ and $g = (g_1, \ldots, g_k)$ where $f_j, g_j \in \mathcal{F}_j$, $j = 1, \ldots, k$. Then, by the condition on ϕ, we have

$$\begin{array}{rl} \|\phi(f) - \phi(g)\|_{Q,2} & \leqslant \| \sum_{j=1}^k L_j |f_j - g_j| \, \|_{Q,2} \\ & \leqslant \sum_{j=1}^k \| L_j |f_j - g_j| \, \|_{Q,2} \end{array} \tag{C.2}$$

Let $\widehat{\mathcal{N}}_j$ be a (ϵ/k)-net for $\mathcal{F}_j$ with the measure $\widetilde{Q}_j$, where $d\widetilde{Q}_j(x) = L_j^2(x) dQ(x)$. Then the set $\{\phi(f_1, \ldots, f_k) - \phi(\bar{f}_1, \ldots, \bar{f}_k) : f_j \in \widehat{\mathcal{N}}_j\}$ is an ϵ-net for $\mathcal{L}$ with respect to the measure Q by (C.2). Thus, for any $\epsilon > 0$ we have that

$$\log N(\epsilon, \mathcal{L}, \| \cdot \|_{Q,2}) \leqslant \sum_{j=1}^k \log N(\epsilon/k, \mathcal{F}_j, \| \cdot \|_{\widetilde{Q}_j,2})$$

Therefore,

$$\begin{array}{rl} \log N(\epsilon \| \sum_{j=1}^k L_j F_j \|_{Q,2}, \mathcal{L}, \| \cdot \|_{Q,2}) & \leqslant \sum_{j=1}^k \log N(\frac{\epsilon}{k} \| \sum_{j=1}^k L_j F_j \|_{Q,2}, \mathcal{F}_j, \| \cdot \|_{\widetilde{Q}_j,2}) \\ & \leqslant \sum_{j=1}^k \log N(\frac{\epsilon}{k} \| L_j F_j \|_{Q,2}, \mathcal{F}_j, \| \cdot \|_{\widetilde{Q}_j,2}) \\ & = \sum_{j=1}^k \log N(\frac{\epsilon}{k} \| F_j \|_{\widetilde{Q}_j,2}, \mathcal{F}_j, \| \cdot \|_{\widetilde{Q}_j,2}) \\ & \leqslant \sum_{j=1}^k \log \sup_{\bar{Q}} N(\frac{\epsilon}{k} \| F_j \|_{\bar{Q},2}, \mathcal{F}_j, \| \cdot \|_{\bar{Q},2}) \end{array}$$

and the result follows since the right hand side no longer depends on Q. ∎

Lemma C.3 (**Covering Entropy for Classes obtained as Conditional Expectations**). *Let $\mathcal{F}$ denote a class of measurable functions $f : \mathcal{W} \times \mathcal{Y} \mapsto \mathbb{R}$ with a measurable envelope F. For a given $f \in \mathcal{F}$, let $\bar{f} : \mathcal{W} \mapsto \mathbb{R}$ be the function $\bar{f}(w) := \int f(w, y) d\mu_w(y)$ where μ_w is a regular conditional probability distribution over $y \in \mathcal{Y}$ conditional on $w \in \mathcal{W}$. Set $\bar{\mathcal{F}} = \{\bar{f} : f \in \mathcal{F}\}$ and let $\bar{F}(w) := \int F(w, y) d\mu_w(y)$ be an envelope for $\bar{\mathcal{F}}$. Then, for $r, s \geqslant 1$,*

$$\log \sup_Q N(\epsilon \| \bar{F} \|_{Q,r}, \bar{\mathcal{F}}, \| \cdot \|_{Q,r}) \leqslant \log \sup_{\widetilde{Q}} N((\epsilon/4)^r \| F \|_{\widetilde{Q},s}, \mathcal{F}, \| \cdot \|_{\widetilde{Q},s}),$$

where Q belongs to the set of finitely-discrete probability measures over $\mathcal{W}$ such that $0 < \| \bar{F} \|_{Q,r} < \infty$, and $\widetilde{Q}$ belongs to the set of finitely-discrete probability measures over $\mathcal{W} \times \mathcal{Y}$ such that $0 < \| F \|_{\widetilde{Q},s} < \infty$. In particular, for every $\epsilon > 0$ and any $k \geqslant 1$

$$\log \sup_Q N(\epsilon, \bar{\mathcal{F}}, \| \cdot \|_{Q,k}) \leqslant \log \sup_{\widetilde{Q}} N(\epsilon/2, \mathcal{F}, \| \cdot \|_{\widetilde{Q},k}).$$

Proof. The proof generalizes the proof of Lemma A.2 in Ghosal, Sen, and van der Vaart (2000). For $f, g \in \mathcal{F}$ and the corresponding $\bar{f}, \bar{g} \in \bar{\mathcal{F}}$, and any probability measure Q on $\mathcal{W}$, by Jensen's inequality, for any $k \geqslant 1$,

$$\mathrm{E}_Q[|\bar{f} - \bar{g}|^k] = \mathrm{E}_Q[|\int (f-g) d\mu_w(y)|^k] \leqslant \mathrm{E}_Q[\int |f-g|^k d\mu_w(y)] = \mathrm{E}_{\bar{Q}}[|f-g|^k]$$

where $d\bar{Q}(w,y) = dQ(w)d\mu_w(y)$. Therefore, for any $\epsilon > 0$

$$\sup_Q N(\epsilon, \bar{\mathcal{F}}, \|\cdot\|_{Q,k}) \leqslant \sup_{\bar{Q}} N(\epsilon, \mathcal{F}, \|\cdot\|_{\bar{Q},k}) \leqslant \sup_{\widetilde{Q}} N(\epsilon/2, \mathcal{F}, \|\cdot\|_{\widetilde{Q},k}),$$

where we use Problems 2.5.1-2 of van der Vaart and Wellner (1996) to replace the supremum over $\bar{Q}$ with the supremum over finitely-discrete probability measures $\widetilde{Q}$.

Moreover, $\|\bar{F}\|_{Q,1} = \mathrm{E}_Q[\bar{F}(w)] = \mathrm{E}_Q[\int F(w,y) d\mu_w(y)] = \mathrm{E}_{\bar{Q}}[F(w,y)] = \|F\|_{\bar{Q},1}$. Therefore taking $k = 1$,

$$\begin{aligned} &\sup_Q N(\epsilon \|\bar{F}\|_{Q,1}, \bar{\mathcal{F}}, \|\cdot\|_{Q,1}) \leqslant \sup_{\bar{Q}} N(\epsilon \|F\|_{\bar{Q},1}, \mathcal{F}, \|\cdot\|_{\bar{Q},1}) \\ &\leqslant \sup_{\widetilde{Q}} N((\epsilon/2)\|F\|_{\widetilde{Q},1}, \mathcal{F}, \|\cdot\|_{\widetilde{Q},1}) \leqslant \sup_{\widetilde{Q}} N((\epsilon/2)\|F\|_{\widetilde{Q},s}, \mathcal{F}, \|\cdot\|_{\widetilde{Q},s}) \end{aligned}$$

where we use Problems 2.5.1-2 of van der Vaart and Wellner (1996) to replace the supremum over $\bar{Q}$ with the supremum over finitely-discrete probability measures $\widetilde{Q}$, and then Problem 2.10.4 of van der Vaart and Wellner (1996) to argue that the last bound in weakly increasing in $s \geqslant 1$.

Also, by the second part of the proof of Theorem 2.6.7 of van der Vaart and Wellner (1996)

$$\sup_Q N(\epsilon \|F\|_{Q,r}, \mathcal{F}, \|\cdot\|_{Q,r}) \leqslant \sup_Q N((\epsilon/2)^r \|F\|_{Q,1}, \mathcal{F}, \|\cdot\|_{Q,1}).$$

■

Comment C.1. Lemma C.3 extends the result in Lemma A.2 in Ghosal, Sen, and van der Vaart (2000) and Lemma 5 in Sherman (1994) which considered integral classes with respect to a fixed measure μ on $\mathcal{Y}$. In our applications we need to allow the integration measure to vary with w, namely we allow for μ_w to be a conditional distribution. ■

Appendix D. Proofs for Section 4

D.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. (Preparation). In the proof $a \lesssim b$ means that $a \leqslant Ab$, where the constant A depends on the constants in Assumptions 4.1 and 4.2 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 assume that $n \geqslant n_0$ in what follows.

To proceed with the presentation of the 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 function and the moment function for γ_V are $\psi^{\gamma}_V(W) = \psi^{\gamma}_V(W, \gamma_V)$ and $\psi^{\gamma}_V(W,\gamma) = V - \gamma$. Recall that 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 $\widehat{g}_V(z,x) = \Lambda_V(f(z,x)'\bar{\beta}_V)$, $\widehat{m}_Z(1,x) = \Lambda_Z(f(x)'\bar{\beta}_Z)$, $\widehat{m}_Z(0,x) = 1 - \widehat{m}_Z(1,x)$, and $\bar{\beta}_V$ and $\bar{\beta}_Z$ are estimators as in Assumption 4.2. For each variable $V \in \mathcal{V}_u$,

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

we obtain the estimator $\widehat{\rho}_u = \big(\{\widehat{\alpha}_V(0), \widehat{\alpha}_V(1), \widehat{\gamma}_V\}\big)_{V\in\mathcal{V}_u}$ of $\rho_u := \big(\{\alpha_V(0), \alpha_V(1), \gamma_V\}\big)_{V\in\mathcal{V}_u}$. The estimator and the estimand are vectors in $\mathbb{R}^{d_\rho}$ with a fixed 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

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

where $Z_{n,P} = (\mathbb{G}_n \psi^{\rho}_u)_{u\in\mathcal{U}}$ and $\psi^{\rho}_u = (\{\psi^{\alpha}_{V,0}, \psi^{\alpha}_{V,1}, \psi^{\gamma}_V\})_{V\in\mathcal{V}_u}$. 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 $z \in \{0,1\}$ and each $j \in \mathcal{J}$, which we do in the steps below.

(a) We make some preliminary observations. For $t = (t_1,t_2,t_3,t_4) \in \mathbb{R}^2 \times (0,1)^2$, $v \in \mathbb{R}$, and $(z,\bar{z}) \in \{0,1\}^2$, 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 $k = (k_j)_{j=1}^4 \in \mathbb{N}^4 : 0 \leqslant |k| \leqslant 3$,

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

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

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

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

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

where

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

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

To see this, note that under Assumption 4.2 for all $n \geqslant \min\{j : \delta_j \leqslant 1/2\}$,

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

for $\beta = \bar{\beta}_Z$, 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 4.2,

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

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

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

under condition (iii) of Assumption 4.1, and

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

under condition (iii) of Assumption 4.1.

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

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

(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

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

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

(d) Note that for

$$\Delta_{V,i} := (\Delta_{1V,i}, \Delta_{2V,i}, \Delta_{3V,i}, \Delta_{4V,i}) = h(X_i) - h_V(X_i), \quad \Delta_{V,i}^k := \Delta_{1V,i}^{k_1}\Delta_{2V,i}^{k_2}\Delta_{3V,i}^{k_3}\Delta_{4V,i}^{k_4},$$

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

with h evaluated at $h = \widehat{h}$ after computing the expectations under P.

By the law of iterated expectations and the orthogonality property of the moment condition for α_V,

$$\mathrm{E}_P[\partial_t^k \varphi(V_i, Z_i, z, h_V(X_i))|X_i] = 0 \ \ \forall k \in \mathbb{N}^4 : |k| = 1, \implies III_V^a(z) = 0.$$

Moreover, uniformly for any $h \in \mathcal{H}_{V,n}$, in view of properties noted in Steps (a) and (b),

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

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

Since $\widehat{h}_V \in \mathcal{H}_{V,n}$ for all $V \in \mathcal{V} = \{V_{uj} : u \in \mathcal{U}, j \in \mathcal{J}\}$ with probability $1 - \Delta_n$, for $n \geqslant n_0$,

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

(e) Furthermore, with probability $1 - \Delta_n$

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

The classes of functions,

$$\mathcal{V} := \{V_{uj} : u \in \mathcal{U}, j \in \mathcal{J}\} \quad \text{and} \quad \mathcal{V}^* := \{g_{V_{uj}}(Z, X) : u \in \mathcal{U}, j \in \mathcal{J}\}, \tag{D.3}$$

viewed as maps from the sample space $\mathcal{W}$ to the real line, are bounded by a constant envelope and obey $\log\sup_Q N(\epsilon, \mathcal{V}, \|\cdot\|_{Q,2}) \lesssim \log(\mathrm{e}/\epsilon) \vee 0$, which holds by Assumption 4.1(ii), and $\log\sup_Q N(\epsilon, \mathcal{V}^*, \|\cdot\|_{Q,2}) \lesssim \log(\mathrm{e}/\epsilon) \vee 0$ which holds by Assumption 4.1(ii) and Lemma C.3. The uniform covering entropy of the function sets

$$\mathcal{B} = \{1(Z = z) : z \in \{0,1\}\} \text{ and } \mathcal{M}^* = \{m_Z(z, X) : z \in \{0,1\}\}$$

are 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 $\binom{2p}{Cs}$ VC-subgraph classes of functions with VC indices bounded by $C's$. Note that a fixed monotone transformations Λ preserves the VC-subgraph property (van der Vaart and Wellner, 1996, Lemma 2.6.18). Therefore

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

Similarly, the class of functions $\mathcal{M} = (\mathcal{M}(1) \cup (1 - \mathcal{M}(1)))$ has a constant envelope, 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 $C's$ since a fixed monotone transformations Λ preserves the VC-subgraph property. Therefore, $\log \sup_Q N(\epsilon, \mathcal{M}, \|\cdot\|_{Q,2}) \lesssim (s \log p + s \log(\mathrm{e}/\epsilon)) \vee 0$.

Finally, the set of functions

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

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

$$\log \sup_Q N(\epsilon, \mathcal{J}_n, \|\cdot\|_{Q,2}) \lesssim (s \log p + s \log(\mathrm{e}/\epsilon)) \vee 0.$$

Applying Lemma C.1 with $\sigma_n = C'\delta_n n^{-1/4}$ and the envelope $J_n = C'$, with probability $1 - \Delta_n$ for some constant $K > e$

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

Here we have used some simple calculations, exploiting the boundedness condition in Assumptions 4.1 and 4.2, to deduce that

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

by definition of the set $\mathcal{H}_{V,n}$, so that we can use Lemma C.1. We also note that $\log(1/\delta_n) \lesssim \log(n)$ by the assumption on δ_n and that $s^2 \log^2(p \vee n) \log^2(n)/n \leqslant \delta_n$ by Assumption 4.2(i).

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

STEP 2 (Uniform Donskerness). Here we claim that Assumption 4.1 implies that the set of vectors of functions $(\psi_u^\rho)_{u\in\mathcal{U}}$ is P-Donsker uniformly in $\mathcal{P}$, namely that

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

where $Z_{n,P} = (\mathbb{G}_n\psi_u^\rho)_{u\in\mathcal{U}}$ and $Z_P = (\mathbb{G}_P\psi_u^\rho)_{u\in\mathcal{U}}$. 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 Theorem B.1.

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

$$\mathcal{F}_z = \left\{ \frac{1\{Z = z\}(V - g_V(z,X))}{m_Z(z,X)} + g_V(z,X), V \in \mathcal{V} \right\}, \ z = 0, 1,$$

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 the conditions of Theorem B.1 apply to $\mathcal{F}_P$ uniformly in $P \in \mathcal{P}$.

Observe that $\mathcal{F}_z$ is formed as a uniform Lipschitz transform of the function sets $\mathcal{B}$, $\mathcal{V}$, $\mathcal{V}^*$ and $\mathcal{M}^*$ defined in Step 1(e), where the validity of the Lipschitz property relies on Assumption 4.1(iii) (to keep the denominator away from zero) and on the boundedness conditions in Assumption 4.1(iii) and Assumption 4.2(iii). The function sets $\mathcal{B}$, $\mathcal{V}$, $\mathcal{V}^*$ and $\mathcal{M}^*$ are uniformly bounded classes that have uniform covering entropy bounded by $\log(\mathrm{e}/\epsilon) \vee 0$ up to a multiplicative constant, and so $\mathcal{F}_z$, which is uniformly bounded under Assumption 4.1, the uniform covering 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 $\mathcal{F}_P$ has this property; namely,

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

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

We demonstrate the second condition in (B.1). Consider a sequence of positive constants ϵ approaching zero, and note that

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

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 4.1(ii). 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 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 deduce using the 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 the Holder inequality, and (3)

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

which we deduce by the definition of $g_{U_u}(z, X) = \mathrm{E}_P[U_u|X, Z = z]$ and the contraction property of the conditional expectation. ∎

D.2. **Proof of Theorem 4.2.** The proof will be similar to the proof of Theorem 4.1; and as in that proof, we only present the argument for the first strategy.

Step 0. (Preparation). In the proof $a \lesssim b$ means that $a \leqslant Ab$, where the constant A depends on the constants in Assumptions 4.1 and 4.2 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 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 $\mathbb{E}_n f = n^{-1} \sum_{i=1}^n f(\xi_i, W_i)$, and $\mathbb{G}_n$ denote the corresponding empirical process $\sqrt{n}(\mathbb{E}_n - P)$, i.e.

$$\mathbb{G}_n f = \sqrt{n}(\mathbb{E}_n f - Pf) = n^{-1/2} \sum_{i=1}^n \left(f(\xi_i, W_i) - \int f(s, w) dP_\xi(s) dP(w) \right).$$

Recall that we define the bootstrap draw as:

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

since $P[\xi\widehat{\psi}^\rho_u] = 0$ because ξ is independent of W and has zero mean. Here $\widehat{\psi}^\rho_u = (\widehat{\psi}^\rho_V)_{V\in\mathcal{V}_u}$, where $\widehat{\psi}^\rho_V(W) = \{\psi^\alpha_{V,0,\widehat{g}_V,\widehat{m}_Z}(W, \widehat{\alpha}_V(0)), \psi^\alpha_{V,1,\widehat{g}_V,\widehat{m}_Z}(W, \widehat{\alpha}_V(1)), \psi^\gamma_V(W, \widehat{\gamma}_V)\}$, is a plug-in estimator of the influence function ψ^ρ_u.

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

$$\zeta^*_{n,P} := Z^*_{n,P} - G^*_{n,P} = o_P(1), \quad \text{for } G^*_{n,P} := (\mathbb{G}_n \xi \psi^\rho_u)_{u\in\mathcal{U}}, \qquad \text{in } \mathbb{D} = \ell^\infty(\mathcal{U})^{d_\rho}, \tag{D.4}$$

where $\zeta^*_{n,P} = \zeta_{n,P}(D_n, B_n)$ is a linearization error, arising completely due to estimation of the influence function; if the influence function were known, this term would be zero.

For the components $(\sqrt{n}(\widehat{\gamma}_V^* - \widehat{\gamma}_V))_{V\in\mathcal{V}}$ of $\sqrt{n}(\widehat{\rho}^* - \widehat{\rho})$ the linearization follows by the representation,

$$\sqrt{n}(\widehat{\gamma}_V^* - \widehat{\gamma}_V) = \mathbb{G}_n \xi \psi_V^\gamma - \underbrace{(\widehat{\gamma}_V - \gamma_V)\mathbb{G}_n \xi}_{I_V^*},$$

for all $V \in \mathcal{V}$, and noting that $\sup_{V\in\mathcal{V}} |I_V^*| = \sup_{V\in\mathcal{V}} |(\widehat{\gamma}_V - \gamma_V)||\mathbb{G}_n \xi| = O_P(n^{-1/2})$, for $\mathcal{V}$ defined in (D.3) by Theorem 4.1 and by $|\mathbb{G}_n \xi| = O_P(1)$.

It remains 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 $j \in \mathcal{J}$. As in the proof of Theorem 4.1, we have that with probability at least $1 - \Delta_n$,

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

We have the representation:

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

where $\sup_{V\in\mathcal{V}, z\in\{0,1\}} (\widehat{\alpha}_V(z) - \alpha_V(z)) = O_P(n^{-1/2})$ by Theorem 4.1.

Hence to establish $\sup_{V\in\mathcal{V}} |II_V^*(z)| = o_P(1)$, it remains to show that with probability $1 - \Delta_n$

$$\sup_{z\in\{0,1\}, V\in\mathcal{V}} |\mathbb{G}_n[\xi f_{\widehat{h}_V,V,z} - \xi f_{h_V,V,z}]| \leqslant \sup_{f\in\xi\mathcal{J}_n} |\mathbb{G}_n(f)| = o_P(1),$$

where

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

By the calculations in Step 1(e) of the proof of Theorem 4.1, $\mathcal{J}_n$ obeys $\log \sup_Q N(\epsilon, \mathcal{J}_n, \|\cdot\|_{Q,2}) \lesssim (s \log p + s \log(e/\epsilon)) \vee 0$. By Lemma C.2, multiplication of this class by ξ does not change the entropy bound modulo an absolute constant, namely

$$\log \sup_Q N(\epsilon \|J_n\|_{Q,2}, \xi\mathcal{J}_n, \|\cdot\|_{Q,2}) \lesssim (s \log p + s \log(e/\epsilon)) \vee 0,$$

where the envelope J_n for $\xi\mathcal{J}_n$ is $|\xi|$ times a constant. Also, $\mathrm{E}[\exp(|\xi|)] < \infty$ implies that $(\mathrm{E}[\max_{i\leqslant n} |\xi_i|^2])^{1/2} \lesssim \log n$. Thus, applying Lemma C.1 with $\sigma = \sigma_n = C'\delta_n n^{-1/4}$ and the envelope $J_n = C'|\xi|$, for some constant $K > e$

$$\begin{aligned}
\sup_{f\in\xi\mathcal{J}_n} |\mathbb{G}_n(f)| &\lesssim \left(\sqrt{s\sigma_n^2 \log(p \vee K \vee \sigma_n^{-1})} + \frac{s \log n}{\sqrt{n}} \log(p \vee K \vee \sigma_n^{-1})\right) \\
&\lesssim \left(\sqrt{s\delta_n^2 n^{-1/2} \log(p \vee n)} + \sqrt{s^2 n^{-1} \log^2(p \vee n) \log^2(n)}\right) \\
&\lesssim \left(\delta_n \delta_n^{1/4} + \delta_n^{1/2}\right) \lesssim (\delta_n^{1/2}) = o_P(1),
\end{aligned}$$

for $\sup_{f\in\xi\mathcal{J}_n} \|f\|_{P,2} = \sup_{f\in\mathcal{J}_n} \|f\|_{P,2} \lesssim \sigma_n$; where the details of calculations are the same as in Step 1(e) of the proof of Theorem 4.1.

Finally, we conclude that

$$\|\zeta^*_{n,P}\|_{\mathbb{D}} \lesssim \sup_{V\in\mathcal{V}} |I^*_V| + \sup_{V\in\mathcal{V},z\in\{0,1\}} |II^*_V| = o_P(1).$$

Step 2. Here we are claiming that $Z^*_{n,P} \rightsquigarrow_B Z_P$ in $\mathbb{D}$, under any sequence $P = P_n \in \mathcal{P}_n$, where $Z_P = (\mathbb{G}_P\psi^\rho_u)_{u\in\mathcal{U}}$. We have that

$$\sup_{h\in \mathrm{BL}_1(\mathbb{D})} \left|\mathrm{E}_{B_n} h(Z^*_{n,P}) - \mathrm{E}_P h(Z_P)\right| \leqslant \sup_{h\in \mathrm{BL}_1(\mathbb{D})} \left|\mathrm{E}_{B_n} h(G^*_{n,P}) - \mathrm{E}_P h(Z_P)\right| + \mathrm{E}_{B_n}(\|\zeta^*_{n,P}\|_{\mathbb{D}} \wedge 2),$$

where the first term is $o^*_P(1)$, since $G^*_{n,P} \rightsquigarrow_B Z_P$ by Theorem B.2, and the second term is $o_P(1)$ because $\|\zeta^*_{n,P}\|_{\mathbb{D}} = o_P(1)$ implies that $\mathrm{E}_P(\|\zeta^*_{n,P}\|_{\mathbb{D}} \wedge 2) = \mathrm{E}_P\mathrm{E}_{B_n}(\|\zeta^*_{n,P}\|_{\mathbb{D}} \wedge 2) \to 0$, which in turn implies that $\mathrm{E}_{B_n}(\|\zeta^*_{n,P}\|_{\mathbb{D}} \wedge 2) = o_P(1)$ by the Markov inequality. ∎

D.3. **Proof of Corollary 4.1.** This is an immediate consequence of Theorems 4.1, 4.2, B.3, and B.4. ∎

Appendix E. Proofs for Section 5

E.1. **Proof of Theorem 5.1.** In the proof $a \lesssim b$ means that $a \leqslant Ab$, where the constant A depends on the constants in Assumptions 5.1–5.3, but not on n once $n \geqslant n_0$, and not on $P \in \mathcal{P}_n$. Since the argument is asymptotic, we can assume that $n \geqslant n_0$ in what follows. 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 use P, suppressing the dependency of P_n on the sample size n.

Throughout the proof we use the notation

$$B(W) := \max_{j\in[d_\theta],k\in[d_\theta+d_t]} \sup_{\nu\in\Theta_u\times T_u(Z_u),u\in\mathcal{U}} \left|\partial_{\nu_k}\mathrm{E}_P[\psi_{uj}(W_u,\nu) \mid Z_u]\right|, \tag{E.1}$$

$$\tau_n := n^{-1/2}\left(\sqrt{s_n\log(a_n)} + n^{-1/2}s_n n^{\frac{1}{q}}\log(a_n)\right). \tag{E.2}$$

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

$$\sup_{u\in\mathcal{U}} \|\widehat{\theta}_u - \theta_u\| \lesssim \tau_n.$$

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 $1 - o(1)$

$$\left\|P[\psi_u(W_u,\widehat{\theta}_u,h_u(Z_u))]\right\| \leqslant \epsilon_n + 2I_1 + 2I_2 \lesssim \tau_n, \tag{E.3}$$

for I_1 and I_2 defined in Step 2 below. The $\lesssim$ bound in (E.3) follows from Step 2 and from the assumption $\epsilon_n = o(n^{-1/2})$. Since by Assumption 5.1(iv), $2^{-1}(\|J_u(\widehat{\theta}_u - \theta_u)\| \wedge c_0)$ does not exceed the left side of (E.3) and $\inf_{u\in\mathcal{U}} \mathrm{mineig}(J'_uJ_u)$ is bounded away from zero uniformly in n, we conclude that $\sup_{u\in\mathcal{U}} \|\widehat{\theta}_u - \theta_u\| \lesssim (\inf_{u\in\mathcal{U}} \mathrm{mineig}(J'_uJ_u))^{-1/2}\tau_n \lesssim \tau_n$.

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

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

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

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

These bounds in turn hold by the following arguments. In order to bound I_{1b} we employ Taylor's expansion and the triangle inequality. For $\bar{h}(Z,u,j,\theta)$ denoting a point on a line connecting vectors $h(Z_u)$ and $h_u(Z_u)$, and t_m denoting the mth element of the vector t,

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

where the last inequality holds by the definition of $B(W)$ given earlier and Hölder's inequality. By Assumption 5.2(ii)(c), $\|B\|_{P,2}\leqslant C$, and by Assumption 5.3, $\sup_{u\in\mathcal{U},h\in\mathcal{H}_{un},m\in[d_t]}\|h_m - h_{um}\|_{P,2}\lesssim\tau_n$, hence we conclude that $I_{1b}\lesssim\tau_n$ since d_θ and d_t are fixed.

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

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

defined in Assumption 5.3 and equipped with an envelope $F_1\leqslant F_0$, to conclude that with probability $1-o(1)$,

$$I_{1a} \lesssim n^{-1/2}\left(\sqrt{s_n\log(a_n)} + n^{-1/2}s_n n^{\frac{1}{q}}\log(a_n)\right)\lesssim\tau_n.$$

Here we use that $\log\sup_Q N(\epsilon\|F_1\|_{Q,2},\mathcal{F}_1,\|\cdot\|_{Q,2})\leqslant s_n\log(a_n/\epsilon)\vee 0$ by Assumption 5.3; $\|F_0\|_{P,q}\leqslant C$ and $\sup_{f\in\mathcal{F}_1}\|f\|_{P,2}^2\leqslant\sigma^2\leqslant\|F_0\|_{P,2}^2$ for $c\leqslant\sigma\leqslant C$ by Assumption 5.2(i); $a_n\geqslant n$ and $s_n\geqslant 1$ by Assumption 5.3; and (E.2).

Step 3. (Linearization) By definition

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

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

$$\begin{aligned} &\left\|\sqrt{n}\mathbb{E}_n\psi_u(W_u,\theta_u,h_u(Z_u)) + J_u\sqrt{n}(\widehat{\theta}_u-\theta_u) + \mathrm{D}_{u,0}(\widehat{h}_u - h_u)\right\| \\ &\leqslant \epsilon_n\sqrt{n} + \sup_{u\in\mathcal{U}}\left(\inf_{\theta\in\Theta_u}\sqrt{n}\|\mathbb{E}_n\psi_u(W_u,\theta,\widehat{h}_u(Z_u))\| + \|II_1(u)\| + \|II_2(u)\|\right) = o_P(1), \end{aligned}$$

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

$$\mathrm{D}_{u,0}(\widehat{h}_u - h_u) := \left(\sum_{m=1}^{d_t} \sqrt{n} P\Big[\partial_{t_m} P[\psi_{uj}(W_u, \theta_u, h_u(Z_u))|Z_u](\widehat{h}_m(Z_u) - h_{um}(Z_u)) \Big] \right)_{j=1}^{d_\theta} = 0$$

by the orthogonality condition. Conclude using Assumption 5.1(iv) that

$$\sup_{u\in\mathcal{U}} \left\| J_u^{-1}\sqrt{n}\mathbb{E}_n\psi_u(W_u,\theta_u,h_u(Z_u)) + \sqrt{n}(\widehat{\theta}_u - \theta_u) \right\| \leqslant o_P(1) \sup_{u\in\mathcal{U}} (\mathrm{mineg}(J_u' J_u)^{-1/2}) = o_P(1),$$

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

$$\mathcal{F}_P := \Big\{ \bar{\psi}_{uj} : j \in [d_\theta], u \in \mathcal{U} \Big\},$$

where $\bar{\psi}_{uj}$ is the j-th element of $-J_u^{-1}\psi_u(W_u,\theta_u,h_u(Z_u))$ and we make explicit the dependence of $\mathcal{F}_P$ on P. Let $\mathcal{M} = \{M_{ujk} : j,k \in [d_\theta], u \in \mathcal{U}\}$, where M_{ujk} is the (j,k) element of the matrix J_u^{-1}. $\mathcal{M}$ is a class of uniformly Hölder continuous functions on $(\mathcal{U}, d_\mathcal{U})$ with a uniform covering entropy bounded by $\log(e/\epsilon) \vee 0$ and equipped with a constant envelope C, given the stated assumptions. This result follows from the fact that by Assumption 5.2(ii)(b)

$$\begin{aligned} \max_{j,k\in[d_\theta]} |M_{ujk} - M_{\bar{u}jk}| &\leqslant \|J_u^{-1} - J_{\bar{u}}^{-1}\| = \|J_u^{-1}(J_u - J_{\bar{u}})J_{\bar{u}}^{-1}\| \\ &\leqslant \|J_u - J_{\bar{u}}\| \sup_{\tilde{u}\in\mathcal{U}} \|J_{\tilde{u}}^{-1}\|^2 \lesssim \|u - \bar{u}\|^{\alpha_2}, \end{aligned} \tag{E.4}$$

and the constant envelope follows by Assumption 5.1(iv). Since $\mathcal{F}_P$ is generated as a finite sum of products of the elements of $\mathcal{M}$ and the class $\mathcal{F}_0$ defined in Assumption 5.2, the properties of $\mathcal{M}$ and the conditions on $\mathcal{F}_0$ in Assumption 5.2(ii) imply that $\mathcal{F}_P$ has a uniformly well-behaved uniform covering entropy by Lemma C.2, namely

$$\sup_{P\in\mathcal{P}=\cup_{n\geqslant n_0}\mathcal{P}_n} \log \sup_Q N(\epsilon\|CF_0\|_{Q,2}, \mathcal{F}_P, \|\cdot\|_{Q,2}) \lesssim \log(e/\epsilon) \vee 0,$$

where $F_P = CF_0$ is an envelope for $\mathcal{F}_P$ since $\sup_{f\in\mathcal{F}_P}|f| \lesssim \sup_{u\in\mathcal{U}}\|J_u^{-1}\| \sup_{f\in\mathcal{F}_0}|f| \leqslant CF_0$ by Assumption 5.2(i). The class $\mathcal{F}_P$ is therefore Donsker uniformly in P because $\sup_{P\in\mathcal{P}}\|F_P\|_{P,q} \leqslant C\sup_{P\in\mathcal{P}}\|F_0\|_{P,q}$ is bounded by Assumption 5.2(ii), and $\sup_{P\in\mathcal{P}}\|\bar{\psi}_u - \bar{\psi}_{\bar{u}}\|_{P,2} \to 0$ as $d_\mathcal{U}(u,\bar{u}) \to 0$ by Assumption 5.2(b) and (E.4). Application of Theorem B.1 gives the results of the theorem.

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

$$\begin{aligned} II_{1j}(u) &:= \sum_{r,k=1}^{d_\nu} \sqrt{n} P\left[\partial_{\nu_k}\partial_{\nu_r} P[\psi_{uj}(W_u, \bar{\nu}_u(Z_u,j))|Z_u]\{\widehat{\nu}_{ur}(Z_u) - \nu_{ur}(Z_u)\}\{\widehat{\nu}_{uk}(Z_u) - \nu_{uk}(Z_u)\}\right], \\ II_{2j}(u) &:= \mathbb{G}_n(\psi_{uj}(W_u, \widehat{\theta}_u, \widehat{h}_u(Z_u)) - \psi_{uj}(W_u,\theta_u,h_u(Z_u))), \end{aligned}$$

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

First, by Assumptions 5.2(ii)(d) and 5.3, the claim of Step 1, and the Hölder inequality,

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

Second, 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 C.1 with an envelope $F_2\lesssim F_0$ gives that with probability $1-o(1)$

$$\sup_{f\in\mathcal{F}_2}|\mathbb{G}_n(f)|\lesssim\tau_n^{\alpha/2}\sqrt{s_n\log(a_n)}+n^{-1/2}s_nn^{\frac{1}{q}}\log(a_n),\tag{E.5}$$

since $\sup_{f\in\mathcal{F}_2}|f|\leqslant 2\sup_{f\in\mathcal{F}_1}|f|\leqslant 2F_0$ by Assumption 5.3; $\|F_0\|_{P,q}\leqslant C$ by Assumption 5.2(i); $\log\sup_Q N(\epsilon\|F_2\|_{Q,2},\mathcal{F}_2,\|\cdot\|_{Q,2})\lesssim(s_n\log a_n+s_n\log(a_n/\epsilon))\vee 0$ by Lemma C.2 because $\mathcal{F}_2=\mathcal{F}_1-\mathcal{F}_0$ for the $\mathcal{F}_0$ and $\mathcal{F}_1$ defined in Assumptions 5.2(i) and 5.3; and σ can be chosen so that $\sup_{f\in\mathcal{F}_2}\|f\|_{P,2}\leqslant\sigma\lesssim\tau_n^{\alpha/2}$. Indeed,

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

where the first inequality follows by the law of iterated expectations; the second inequality follows by Assumption 5.2(ii)(a); and the last inequality follows from $\alpha\in[1,2]$ by Assumption 5.2, the monotonicity of the norm $\|\cdot\|_{P,\alpha}$ in $\alpha\in[1,\infty]$, and Assumption 5.3.

Conclude using the growth conditions of Assumption 5.3 that with probability $1-o(1)$

$$\max_{j\in[d_\theta]}\sup_{u\in\mathcal{U}}|II_{2j}(u)|\lesssim\tau_n^{\alpha/2}\sqrt{s_n\log(a_n)}+n^{-1/2}s_nn^{\frac{1}{q}}\log(a_n)=o(1).\tag{E.6}$$

Step 5. In this step we show that

$$\sup_{u\in\mathcal{U}}\inf_{\theta\in\Theta_u}\sqrt{n}\|\mathbb{E}_n\psi_u(W_u,\theta,\widehat{h}_u(Z_u))\|=o_P(1).$$

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, $\sup_{u\in\mathcal{U}}\|\bar{\theta}_u-\theta_u\|=O_P(1/\sqrt{n})$ by the last paragraph of Step 3.

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

$$\sqrt{n}\|\mathbb{E}_n\psi_u(W_u, \bar{\theta}_u, \widehat{h}_u(Z_u))\| \leqslant \sqrt{n}\|\mathbb{E}_n\psi_u(W_u, \theta_u, h_u(Z_u)) + J_u(\bar{\theta}_u - \theta_u) + \mathrm{D}_{u,0}(\widehat{h}_u - h_u)\| + o_P(1)$$

where the first term on the right side is zero by definition of $\bar{\theta}_u$ and $\mathrm{D}_{u,0}(\widehat{h}_u - h_u) = 0$. ∎

E.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 5.1– 5.3, 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 assume that $n \geqslant n_0$ in what follows.

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

$$\mathbb{G}_n f = \sqrt{n}(\mathbb{E}_n f - Pf) = n^{-1/2}\sum_{i=1}^n \left(f(\xi_i, W_i) - \int f(s, w) dP_\xi(s) dP(w) \right).$$

Recall that we define the bootstrap draw as:

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

where $\widehat{\psi}_u(W) = -\widehat{J}_u^{-1}\psi_u(W_u, \widehat{\theta}_u, \widehat{h}_u(Z_u))$.

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

$$\zeta^*_{n,P} := Z^*_{n,P} - G^*_{n,P} = o_P(1) \quad \text{in } \mathbb{D} = \ell^\infty(\mathcal{U})^{d_\theta}, \tag{E.7}$$

where $G^*_{n,P} := (\mathbb{G}_n \xi \bar{\psi}_u)_{u \in \mathcal{U}}$, and $\bar{\psi}_u(W) = -J_u^{-1}\psi_u(W_u, \theta_u, h_u(Z_u))$.

With probability $1 - \delta_n$, $\widehat{h}_u \in \mathcal{H}_{un}$, $\widehat{\theta}_u \in \Theta_{un} = \{\theta \in \Theta_u : \|\theta - \theta_u\| \leqslant C\tau_n\}$, and $J_u \in \mathcal{J}_n$, so that $\|\zeta^*_{n,P}\|_{\mathbb{D}} \lesssim \sup_{f \in \mathcal{F}_3} |\mathbb{G}_n[\xi f]|$, where

$$\mathcal{F}_3 = \left\{ \tilde{\psi}_{uj}(\bar{\theta}_u, \bar{h}_u, \bar{J}_u) - \bar{\psi}_{uj} : j \in [d_\theta], u \in \mathcal{U}, \bar{\theta}_u \in \Theta_{un}, \bar{h}_u \in \mathcal{H}_{un}, \bar{J}_u \in \mathcal{J}_n \right\},$$

where $\tilde{\psi}_{uj}(\bar{\theta}_u, \bar{h}_u, \bar{J}_u)$ is the j-th element of $-\bar{J}_u^{-1}\psi_u(W_u, \bar{\theta}_u, \bar{h}_u(Z_u))$, and $\bar{\psi}_{uj}$ is the j-th element of $-J_u^{-1}\psi_u(W_u, \theta_u, h_u(Z_u))$. By the arguments similar to those employed in the proof of the previous theorem, under Assumption 5.3 and the additional conditions stated in the theorem, $\mathcal{F}_3$ obeys

$$\log \sup_Q N(\epsilon \|F_3\|_{Q,2}, \mathcal{F}_3, \|\cdot\|_{Q,2}) \lesssim (s_n \log a_n + s_n \log(a_n/\epsilon)) \vee 0,$$

for an envelope $F_3 \lesssim F_0$. By Lemma C.2, multiplication of this class by ξ does not change the entropy bound modulo an absolute constant, namely

$$\log \sup_Q N(\epsilon \||\xi| F_3\|_{Q,2}, \xi\mathcal{F}_3, \|\cdot\|_{Q,2}) \lesssim (s_n \log a_n + s_n \log(a_n/\epsilon)) \vee 0.$$

Also $\mathrm{E}[\exp(|\xi|)] < \infty$ implies $(\mathrm{E}[\max_{i\leqslant n}|\xi_i|^2])^{1/2} \lesssim \log n$, so that, using independence of $(\xi_i)_{i=1}^n$ from $(W_i)_{i=1}^n$ and Assumption 5.2(i),

$$\|\max_{i\leqslant n}\xi_i F_0(W_i)\|_{\mathrm{P}_P,2} \leqslant \|\max_{i\leqslant n}\xi_i\|_{\mathrm{P}_P,2}\|\max_{i\leqslant n}F_0(W_i)\|_{\mathrm{P}_P,2} \lesssim n^{1/q}\log n.$$

Applying Lemma C.1,

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

for $\sup_{f\in\xi\mathcal{F}_3}\|f\|_{P,2} = \sup_{f\in\mathcal{F}_3}\|f\|_{P,2} \lesssim \sigma_n \lesssim \tau_n^{\alpha/2}$, where the details of calculations are similar to those in the proof of Theorem 5.1. Indeed, with probability $1-o(\delta_n)$,

$$\begin{aligned}
\sup_{f\in\mathcal{F}_3}\|f\|_{P,2}^2 &\lesssim \sup_{u\in\mathcal{U}}\|J_u^{-1}\|^2 \sup_{j\in[d_\theta],u\in\mathcal{U},\nu\in\Theta_{un}\times\mathcal{H}_{un}} P\left(P[(\psi_{uj}(W_u,\nu(Z_u)) - \psi_{uj}(W_u,\nu_u(Z_u)))^2|Z_u]\right) \\
&+ \sup_{u\in\mathcal{U}}\|\bar{J}_u^{-1} - J_u^{-1}\|^2 \sup_{j\in[d_\theta],u\in\mathcal{U},\nu\in\Theta_{un}\times\mathcal{H}_{un}} P\left(P[\psi_{uj}(W_u,\nu(Z_u))^2|Z_u]\right) \\
&\lesssim \sup_{u\in\mathcal{U},\nu\in\Theta_{un}\times\mathcal{H}_{un}}\|\nu-\nu_u\|_{P,\alpha}^\alpha + \tau_n^\alpha \\
&\lesssim \sup_{u\in\mathcal{U},\nu\in\Theta_{un}\times\mathcal{H}_{un}}\|\nu-\nu_u\|_{P,2}^\alpha + \tau_n^\alpha \lesssim \tau_n^\alpha,
\end{aligned}$$

where the first inequality follows from the triangle inequality and the law of iterated expectations; the second inequality follows by Assumption 5.2(ii)(a), Assumption 5.2(i), and $\sup_{u\in\mathcal{U}}\|\bar{J}_u^{-1} - J_u^{-1}\|^2 \lesssim \tau_n^\alpha$ by the assumptions of the theorem and the continuous mapping theorem; the third inequality follows from $\alpha\in[1,2]$ by Assumption 5.2, the monotonicity of the norm $\|\cdot\|_{P,\alpha}$ in $\alpha\in[1,\infty]$, and Assumption 5.3; and the last inequality follows from $\|\nu-\nu_u\|_{P,2} \lesssim \tau_n$ by the definition of Θ_{un} and $\mathcal{H}_{un}$. The claim of Step 1 follows.

Step 2. Here we are claiming that $Z_{n,P}^* \rightsquigarrow_B Z_P$ in $\mathbb{D} = \ell^\infty(\mathcal{U})^{d_\theta}$, under any sequence $P = P_n \in \mathcal{P}_n$, were $Z_P = (\mathbb{G}_P\bar{\psi}_u)_{u\in\mathcal{U}}$. By the triangle inequality and Step 1,

$$\sup_{h\in\mathrm{BL}_1(\mathbb{D})}\left|\mathrm{E}_{B_n}h(Z_{n,P}^*) - \mathrm{E}_P h(Z_P)\right| \leqslant \sup_{h\in\mathrm{BL}_1(\mathbb{D})}\left|\mathrm{E}_{B_n}h(G_{n,P}^*) - \mathrm{E}_P h(Z_P)\right| + \mathrm{E}_{B_n}(\|\zeta_{n,P}^*\|_{\mathbb{D}}\wedge 2),$$

where the first term is $o_P^*(1)$, since $G_{n,P}^* \rightsquigarrow_B Z_P$ by Theorem B.2, and the second term is $o_P(1)$ because $\|\zeta_{n,P}^*\|_{\mathbb{D}} = o_P(1)$ implies that $\mathrm{E}_P(\|\zeta_{n,P}^*\|_{\mathbb{D}}\wedge 2) = \mathrm{E}_P\mathrm{E}_{B_n}(\|\zeta_{n,P}^*\|_{\mathbb{D}}\wedge 2) \to 0$, which in turn implies that $\mathrm{E}_{B_n}(\|\zeta_{n,P}^*\|_{\mathbb{D}}\wedge 2) = o_P(1)$ by the Markov inequality. ∎

E.3. **Proof of Theorem 5.3.** This is an immediate consequence of Theorems 5.1, 5.2, B.3, and B.4. ∎

Appendix F. Proofs for Section 6

Proof of Theorem 6.1. 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 use P, suppressing the dependency of P_n on the sample size n. To prove this result we invoke Lemmas G.3-G.5 in Appendix G. These lemmas rely on specific events (described below)

and Condition WL which is also stated in Appendix G. We will show that Assumption 6.1 implies that the required events occur with probability $1 - o(1)$ and also implies Condition WL.

Let $\widehat{\Psi}_{u0,jj} = \{\mathbb{E}_n[|f_j(X)\zeta_u|^2]\}^{1/2}$ denote the ideal penalty loadings. The three events required to occur with probability $1 - o(1)$ are the following: $E_1 := \{c_r \geqslant \sup_{u\in\mathcal{U}} \|r_u\|_{\mathbb{P}_n,2}\}$, and where $c_r := C\sqrt{s\log(p\vee n)/n}$; $E_2 := \{\lambda/n \geqslant \sqrt{c}\sup_{u\in\mathcal{U}} \|\widehat{\Psi}_{u0}^{-1}\mathbb{E}_n[\zeta_u f(X)]\|_\infty\}$, $E_3 := \{\ell\widehat{\Psi}_{u0} \leqslant \widehat{\Psi}_u \leqslant L\widehat{\Psi}_{u0}\}$, for some $1/\sqrt{c} < 1/\sqrt[4]{c} < \ell$ and L uniformly bounded for the penalty loading $\widehat{\Psi}_u$ in all iterations $k \leqslant K$ for n sufficiently large.

By Assumption 6.1(iv)(b) E_1 holds with probability $1 - o(1)$.

Next we verify that Condition WL holds. Condition WL(i) is implied by the approximate sparsity condition in Assumption 6.1(i) and the covering condition in Assumption 6.1(ii). By Assumption 6.1 we have that d_u is fixed and the Algorithm sets $\gamma \in [1/n, \min\{\log^{-1} n, pn^{d_u-1}\}]$ so that $\gamma = o(1)$ and $\Phi^{-1}(1-\gamma/\{2pn^{d_u}\}) \leqslant C\log^{1/2}(np) \leqslant C\delta_n n^{1/6}$ by Assumption 6.1(i). Since it is assumed that $\mathrm{E}_P[|f_j(X)\zeta_u|^2] \geqslant c$ and $\mathrm{E}_P[|f_j(X)\zeta_u|^3] \leqslant C$ uniformly in $j \leqslant p$ and $u \in \mathcal{U}$, Condition WL(ii) holds. Condition WL(iii) follows from Assumption 6.1(iv).

Since Condition WL holds, by Lemma G.1, the event E_2 occurs with probability $1 - o(1)$.

Next we proceed to verify occurrence of E_3. In the first iteration, the penalty loadings are defined as $\widehat{\Psi}_{ujj} = \{\mathbb{E}_n[|f_j(X)Y_u|^2]\}^{1/2}$ for $j = 1,\dots,p$, $u \in \mathcal{U}$. By Assumption 6.1, $\underline{c} \leqslant \mathrm{E}_P[|f_j(X)\zeta_u|^2] \leqslant \mathrm{E}_P[|f_j(X)Y_u|^2] \leqslant C$ uniformly over $u \in \mathcal{U}$ and $j = 1,\dots,p$. Moreover, Assumption 6.1(iv)(b) yields

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

with probability $1 - \Delta_n$. In turn this shows that for n large so that $\delta_n \leqslant \underline{c}/4$ we have[27]

$$(1 - 2\delta_n/\underline{c})\mathbb{E}_n[|f_j(X)\zeta_u|^2] \leqslant \mathbb{E}_n[|f_j(X)Y_u|^2] \leqslant (\{C+\delta_n\}/\{\underline{c}-\delta_n\})\mathbb{E}_n[|f_j(X)\zeta_u|^2]$$

with probability $1 - \Delta_n$ so that $\ell\widehat{\Psi}_{u0} \leqslant \widehat{\Psi}_u \leqslant L\widehat{\Psi}_{u0}$ for some uniformly bounded L and $\ell > 1/\sqrt[4]{c}$. Moreover, $\tilde{\mathbf{c}} = \{(L\sqrt{c}+1)/(\sqrt{c}\ell - 1)\}\sup_{u\in\mathcal{U}}\|\widehat{\Psi}_{u0}^{-1}\|_\infty\|\widehat{\Psi}_{u0}\|_\infty$ is uniformly bounded for n large enough which implies that $\kappa_{2\tilde{\mathbf{c}}}$ as defined in (G.1) in Appendix G.2 is bounded away from zero with probability $1 - \Delta_n$ by the condition on sparse eigenvalues of order $s\ell_n$ (see Bickel, Ritov, and Tsybakov (2009) Lemma 4.1(ii)).

By Lemma G.3, since $\lambda \in [cn^{1/2}\log^{1/2}(p\vee n), Cn^{1/2}\log^{1/2}(p\vee n)]$ by the choice of γ and d_u fixed, $c_r \leqslant C\sqrt{s\log(p\vee n)/n}$, $\sup_{u\in\mathcal{U}}\|\widehat{\Psi}_{u0}\|_\infty \leqslant C$, 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 G.4, by Assumption 6.1(iv)(c), we have that $\min_{m\in\mathcal{M}}\phi_{\max}(m)$ is uniformly bounded for n large enough with probability $1 - o(1)$. Thus, with probability $1 - o(1)$,

[27] Indeed, using that $\underline{c} \leqslant \mathrm{E}_P[|f_j(X)\zeta_u|^2] \leqslant \mathrm{E}_P[|f_j(X)Y_u|^2] \leqslant C$, we have $(1 - 2\delta_n/\underline{c})\mathbb{E}_n[|f_j(X)\zeta_u|^2] \leqslant (1 - 2\delta_n/\underline{c})\{\delta_n + \mathrm{E}_P[|f_j(X)\zeta_u|^2]\} \leqslant \mathrm{E}_P[|f_j(X)\zeta_u|^2] - \delta_n \leqslant \mathrm{E}_P[|f_j(X)Y_u|^2] - \delta_n \leqslant \mathbb{E}_n[|f_j(X)Y_u|^2]$. Similarly, $\mathbb{E}_n[|f_j(X)Y_u|^2] \leqslant \delta_n + \mathrm{E}_P[|f_j(X)Y_u|^2] \leqslant \delta_n + C \leqslant (\{\delta_n + C\}/\{\underline{c} - \delta_n\})\mathbb{E}_n[|f_j(X)\zeta_u|^2]$.

by Lemma G.4 we have

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

Therefore by Lemma G.5 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, since uniformly in $u\in\mathcal{U}$ we have a sparsity bound $\|(\widetilde{\theta}_u-\theta_u)\|_0 \leqslant C''s$ and that ensures that a bound on the prediction rate yields a bound on the ℓ_1-norm rate through the relations $\|v\|_1 \leqslant \sqrt{\|v\|_0}\|v\| \leqslant \sqrt{\|v\|_0}\|f(X)'v\|_{\mathbb{P}_n,2}/\sqrt{\phi_{\min}(\|v\|_0)}$.

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 $j=1,\ldots,p$, $u\in\mathcal{U}$. We assume $(\widetilde{\theta}_u^{(k)})_{u\in\mathcal{U}}$ satisfy the rates above uniformly in $u\in\mathcal{U}$. Then with probability $1-o(1)$ we have uniformly in $u\in\mathcal{U}$ and $j=1,\ldots,p$

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

where we used that $\max_{i\leqslant n, j\leqslant p}|f_j(X_i)| \leqslant K_n$ a.s., and $K_n^2 s\log(p\vee n) \leqslant \delta_n n$ by Assumption 6.1(iv)(a), and that $\inf_{u\in\mathcal{U},j\leqslant p}\widehat{\Psi}_{u0jj} \geqslant \underline{c}/2$ with probability $1-o(1)$ for n large so that $\delta_n \leqslant \underline{c}/2$. Further, for n large so that $(2\bar{C}\delta_n^{1/2}/\underline{c}) < 1-1/\sqrt[4]{c}$, this establishes that the event of the penalty loadings for the $(k+1)$th iteration also satisfy $\ell\widehat{\Psi}_{u0}^{-1} \leqslant \widehat{\Psi}_u^{-1} \leqslant L\widehat{\Psi}_{u0}^{-1}$ for a uniformly bounded L and some $\ell > 1/\sqrt[4]{c}$ with probability $1-o(1)$ uniformly in $u\in\mathcal{U}$.

This leads to the stated rates of convergence and sparsity bound. ∎

Proof of Theorem 6.2. 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 use P, suppressing the dependency of P_n on the sample size n. The proof is similar to the proof of Theorem 6.1. We invoke Lemmas G.6, G.7 and G.8 which require Condition WL and some events to occur. We show that Assumption 6.2 implies Condition WL and that the required events occur with probability at least $1-o(1)$.

Let $\widehat{\Psi}_{u0,jj} = \{\mathbb{E}_n[|f_j(X)\zeta_u|^2]\}^{1/2}$ denote the ideal penalty loadings, $w_{ui} = \mathrm{E}_P[Y_{ui} \mid X_i](1-\mathrm{E}_P[Y_{ui}\mid X_i])$ the conditional variance of Y_{ui} given X_i and $\tilde{r}_{ui} = \tilde{r}_u(X_i)$ the rescaled approximation error as defined in (G.5). The three events required to occur with probability $1-o(1)$ are as follows: $E_1 := \{c_r \geqslant \sup_{u\in\mathcal{U}}\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\}$ for $c_r := C'\sqrt{s\log(p\vee n)/n}$ where C' is large enough; $E_2 := \{\lambda/n \geqslant \sqrt{c}\sup_{u\in\mathcal{U}}\|\widehat{\Psi}_{u0}^{-1}\mathbb{E}_n[\zeta_u f(X)]\|_\infty\}$; and $E_3 := \{\ell\widehat{\Psi}_{u0} \leqslant \widehat{\Psi}_u \leqslant L\widehat{\Psi}_{u0}\}$, for $\ell > 1/\sqrt[4]{c}$ and L uniformly bounded, for the penalty loading $\widehat{\Psi}_u$ in all iterations $k\leqslant K$ for n sufficiently large.

Regarding E_1, by Assumption 6.2(iii), we have $\underline{c}(1-\underline{c}) \leqslant w_{ui} \leqslant 1/4$. Since $|r_u(X_i)| \leqslant \delta_n$ a.s. uniformly on $u\in\mathcal{U}$ for $i=1,\ldots,n$, we have that the rescaled approximation error defined in (G.5)

satisfies $|\tilde{r}_u(X_i)| \leqslant |r_u(X_i)|/\{\underline{c}(1-\underline{c}) - 2\delta_n\}_+ \leqslant \tilde{C}|r_u(X_i)|$ for n large enough so that $\delta_n \leqslant \underline{c}(1-\underline{c})/4$. Thus $\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2} \leqslant \tilde{C}\|r_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}$. Assumption 6.2(iv)(b) yields $\sup_{u\in\mathcal{U}} \|r_u/\sqrt{w_u}\|_{\mathbb{P}_n,2} \leqslant C\sqrt{s\log(p\vee n)/n}$ with probability $1-o(1)$, so E_3 occurs with probability $1-o(1)$.

To apply Lemma G.1 to show that E_2 occurs with probability $1-o(1)$ we need to verify Condition WL. Condition WL(i) is implied by the sparsity in Assumption 6.2(i) and the covering condition in Assumption 6.2(ii). By Assumption 6.2 we have that d_u is fixed and the Algorithm sets $\gamma \in [1/n, \min\{\log^{-1} n, pn^{d_u-1}\}]$ so that $\gamma = o(1)$ and $\Phi^{-1}(1-\gamma/\{2pn^{d_u}\}) \leqslant C\log^{1/2}(np) \leqslant C\delta_n n^{1/6}$ by Assumption 6.2(i). Since it is assumed that $\mathrm{E}_P[|f_j(X)\zeta_u|^2] \geqslant c$ and $\mathrm{E}_P[|f_j(X)\zeta_u|^3] \leqslant C$ uniformly in $j \leqslant p$ and $u \in \mathcal{U}$, Condition WL(ii) holds. Condition WL(iii) follows from Assumption 6.1(iv). Then, by Lemma G.1, the event E_2 occurs with probability $1-o(1)$.

Next we verify the occurrence of E_3. In the initial iteration, the penalty loadings are defined as $\widehat{\Psi}_{ujj} = \frac{1}{2}\{\mathbb{E}_n[|f_j(X)|^2]\}^{1/2}$ for $j = 1,\ldots,p$, $u \in \mathcal{U}$. Assumption 6.2(iv)(c) for the sparse eigenvalues implies that for n large enough, $c' \leqslant \mathbb{E}_n[|f_j(X)|^2] \leqslant C'$ for all $j = 1,\ldots,p$, with probability $1-o(1)$.

Moreover, Assumption 6.2(iv)(b) yields

$$\sup_{u\in\mathcal{U}} \max_{j\leqslant p} |(\mathbb{E}_n - \mathrm{E}_P)[|f_j(X)\zeta_u|^2]| \leqslant \delta_n \tag{F.1}$$

with probability $1-\Delta_n$, so that $\widehat{\Psi}_{u0jj}$ is bounded away from zero and from above uniformly over $j = 1,\ldots,p$, $u \in \mathcal{U}$, with the same probability because $\mathrm{E}_P[|f_j(X)\zeta_u|^2]$ is bounded away from zero and above. By (F.1) and $\mathrm{E}_P[|f_j(X)\zeta_u|^2] \leqslant \frac{1}{4}\mathrm{E}_P[|f_j(X)|^2]$, for n large enough, we have $\ell\widehat{\Psi}_{u0} \leqslant \widehat{\Psi}_u \leqslant L\widehat{\Psi}_{u0}$ for some uniformly bounded L and $\ell > 1/\sqrt[4]{c}$ with probability $1-\Delta_n$.

Thus, $\tilde{\mathbf{c}} = \{(L\sqrt{c}+1)/(\ell\sqrt{c}-1)\}\sup_{u\in\mathcal{U}} \|\widehat{\Psi}_{u0}^{-1}\|_\infty \|\widehat{\Psi}_{u0}\|_\infty$ is uniformly bounded. In turn, since $\inf_{u\in\mathcal{U}} \min_{i\leqslant n} w_{ui} \geqslant \underline{c}(1-\underline{c})$ is bounded away from zero, we have $\bar{\kappa}_{2\tilde{\mathbf{c}}} \geqslant \sqrt{\underline{c}(1-\underline{c})}\kappa_{2\tilde{\mathbf{c}}}$ by their definitions in (G.1) and (G.2). It follows that $\kappa_{2\tilde{\mathbf{c}}}$ is bounded away from zero by the condition on $s\ell_n$ sparse eigenvalues stated in Assumption 6.2(iv)(c), see Bickel, Ritov, and Tsybakov (2009) Lemma 4.1(ii).

By the choice of γ and d_u fixed, $\lambda \in [cn^{1/2}\log^{1/2}(p\vee n), Cn^{1/2}\log^{1/2}(p\vee n)]$. By relation (G.4) and Assumption 6.2(iv)(a), $\inf_{u\in\mathcal{U}} \bar{q}_{A_u} \geqslant c'\bar{\kappa}_{2\tilde{\mathbf{c}}}/\{\sqrt{s}K_n\}$. Under the condition $K_n^2 s^2\log^2(p\vee n) \leqslant \delta_n n$, the side condition in Lemma G.6 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 Assumption 6.2(iv)(c) and $K_n^2 s^2\log^2(p\vee n) \leqslant \delta_n n$, with probability $1-o(1)$ Lemma G.7 implies

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

since $\min_{m\in\mathcal{M}}\phi_{\max}(m)$ is uniformly bounded. The rate of convergence for $\widetilde{\theta}_u$ is given by Lemma G.8, 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 by (G.20) we have uniformly in $u\in\mathcal{U}$

$$\begin{aligned}M_u(\tilde{\theta}_u)-M_u(\theta_u)&\leqslant M_u(\widehat{\theta}_u)-M_u(\theta_u)\leqslant\tfrac{\lambda}{n}\|\widehat{\Psi}_u\theta_u\|_1-\tfrac{\lambda}{n}\|\widehat{\Psi}_u\widehat{\theta}_u\|_1\leqslant\tfrac{\lambda}{n}\|\widehat{\Psi}_u(\widehat{\theta}_{uT_u}-\theta_u)\|_1\\&\leqslant\bar{C}'s\log(p\vee n)/n,\end{aligned}$$

$\sup_{u\in\mathcal{U}}\|\mathbb{E}_n[f(X)\zeta_u]\|_\infty\leqslant C\sqrt{\log(p\vee n)/n}$ by Lemma G.1, $\phi_{\min}(\widehat{s}_u+s_u)$ is bounded away from zero (by Assumption 6.2(iv)(c) and $\widehat{s}_u\leqslant C'''s$), $\inf_{u\in\mathcal{U}}\psi_u(\{\delta\in\mathbb{R}^p:\|\delta\|_0\leqslant\widehat{s}_u+s_u\})$ is bounded away from zero (because $\inf_{u\in\mathcal{U}}\min_{i\leqslant n}w_{ui}\geqslant\underline{c}(1-\underline{c})$), 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 $j=1,\ldots,p$, $u\in\mathcal{U}$. We assume $(\widetilde{\theta}_u^{(k)})_{u\in\mathcal{U}}$ satisfy the rates above uniformly in $u\in\mathcal{U}$. Then

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

and therefore, provided that $(2C\delta_n/\underline{c})<1-1/\sqrt[4]{c}$, uniformly in $u\in\mathcal{U}$, $\ell\widehat{\Psi}_{u0}\leqslant\widehat{\Psi}_u\leqslant L\widehat{\Psi}_{u0}$ for $\ell>1/\sqrt[4]{c}$ 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 G. Finite Sample Results of a Continuum of Lasso and Post-Lasso Estimators for Functional Responses

G.1. **Assumptions.** We consider the following high level conditions which are implied by the primitive Assumptions 6.1 and 6.2. For each $n\geqslant 1$, there is a sequence of independent random variables $(W_i)_{i=1}^n$, defined on the probability space $(\Omega,\mathcal{A}_\Omega,\mathrm{P}_P)$ such that model (6.1) holds with $\mathcal{U}\subset[0,1]^{d_u}$. Let $d_{\mathcal{U}}$ be a metric on $\mathcal{U}$ (and note that the results cover the case where d_u is a function of n). Throughout this section we assume that the variables $(X_i,(Y_{ui},\zeta_{ui}:=Y_{ui}-\mathrm{E}_P[Y_{ui}\mid X_i])_{u\in\mathcal{U}})$ are generated as suitably measurable transformations of W_i and $u\in\mathcal{U}$. Furthermore, this section uses the notation $\bar{\mathrm{E}}_P[\cdot]=\frac{1}{n}\sum_{i=1}^n\mathrm{E}_P[\cdot]$, because we allow for independent non-identically distributed (i.n.i.d.) data.

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 and C which will not vary with P.

Condition WL. *Let $T_u:=\mathrm{supp}(\theta_u)$, $u\in\mathcal{U}$, and suppose that: (i) for $s\geqslant 1$ we have $\sup_{u\in\mathcal{U}}\|\theta_u\|_0\leqslant s$, $\log N(\epsilon,\mathcal{U},d_{\mathcal{U}})\leqslant d_u\log(1/\epsilon)\vee 0$; (ii) uniformly over $u\in\mathcal{U}$, we have that*

$\max_{j\leqslant p}\frac{\{\bar{\mathrm{E}}_P[|f_j(X)\zeta_u|^3]\}^{1/3}}{\{\bar{\mathrm{E}}_P[|f_j(X)\zeta_u|^2]\}^{1/2}}\Phi^{-1}(1-\gamma/\{2pn^{d_u}\}) \leqslant \delta_n n^{1/6}$ *and* $0 < c \leqslant \bar{\mathrm{E}}_P[|f_j(X)\zeta_u|^2] \leqslant C$, $j = 1,\dots,p$; *(iii) with probability* $1-\Delta_n$, *we have* $\max_{i\leqslant n}\|f(X_i)\|_\infty \leqslant K_n$, $\sup_{u\in\mathcal{U}}\max_{j\leqslant p}|(\mathbb{E}_n - \bar{\mathrm{E}}_P)[f_j(X)^2\zeta_u^2]| \leqslant \delta_n$, $\log(p\vee n^{d_u+1})\sup_{d_{\mathcal{U}}(u,u')\leqslant 1/n}\max_{j\leqslant p}\mathbb{E}_n[f_j(X)^2(\zeta_u-\zeta_{u'})^2] \leqslant \delta_n$, $\sup_{d_{\mathcal{U}}(u,u')\leqslant 1/n}\|\mathbb{E}_n[f(X)(\zeta_u-\zeta_{u'})]\|_\infty \leqslant \delta_n n^{-\frac{1}{2}}$.

The following technical lemma justifies the choice of penalty level λ. It is based on self-normalized moderate deviation theory. In what follows, for $u\in\mathcal{U}$ we let $\widehat{\Psi}_{u0}$ denote a diagonal $p\times p$ matrix of "ideal loadings" with diagonal elements given by $\widehat{\Psi}_{u0jj} = \{\mathbb{E}_n[f_j^2(X)\zeta_u^2]\}^{1/2}$ for $j = 1,\dots,p$.

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

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

We note that Condition WL(iii) contains high level conditions on the process $(Y_u,\zeta_u)_{u\in\mathcal{U}}$. The following lemma provides easy to verify sufficient conditions that imply Condition WL(iii).

Lemma G.2. *Suppose the i.i.d. sequence* $((Y_{ui},\zeta_{ui})_{u\in\mathcal{U}}, X_i), i = 1,\dots,n$, *satisfies the following conditions: (i)* $c \leqslant \max_{j\leqslant p}\mathrm{E}_P[f_j(X)^2] \leqslant C$, $\max_{j\leqslant p}|f_j(X)| \leqslant K_n$, $\sup_{u\in\mathcal{U}}\max_{i\leqslant n}|Y_{ui}| \leqslant B_n$, *and* $c \leqslant \sup_{u\in\mathcal{U}}\mathrm{E}_P[\zeta_u^2 \mid X] \leqslant C$, *P-a.s.; (ii) for some random variable* Y *we have* $Y_u = G(Y,u)$ *where* $\{G(\cdot,u) : u\in\mathcal{U}\}$ *is a VC-class of functions with VC-index equal to* $C'd_u$, *(iii) For some fixed* $\nu > 0$, *we have* $\mathrm{E}_P[|Y_u - Y_{u'}|^2 \mid X] \leqslant L_n|u-u'|^\nu$ *for any* $u,u'\in\mathcal{U}$, *P-a.s. For* $\widetilde{A} := pnK_nB_nn^\nu/L_n$, *we have with probability* $1-\Delta_n$

$$\begin{aligned}
\sup_{d_{\mathcal{U}}(u,u')\leqslant 1/n}\|\mathbb{E}_n[f(X)(\zeta_u-\zeta_{u'})]\|_\infty &\lesssim \tfrac{1}{\sqrt{n}}\left\{\sqrt{\tfrac{(1+d_u)L_n\log(\widetilde{A})}{n^\nu}} + \tfrac{(1+d_u)K_nB_n\log(\widetilde{A})}{\sqrt{n}}\right\}\\
\sup_{d_{\mathcal{U}}(u,u')\leqslant 1/n}\max_{j\leqslant p}\mathbb{E}_n[f_j(X)^2(\zeta_u-\zeta_{u'})^2] &\lesssim L_nn^{-\nu}\left\{1+\sqrt{\tfrac{K_n^2\log(pnK_n^2)}{n}} + \tfrac{K_n^2}{n}\log(pnK_n^2)\right\}\\
\sup_{u\in\mathcal{U}}\max_{j\leqslant p}|(\mathbb{E}_n-\mathrm{E}_P)[f_j^2(X)\zeta_u^2]| &\lesssim \sqrt{\tfrac{(1+d_u)\log(npK_nB_n)}{n}}K_nB_n + \tfrac{(1+d_u)K_n^2B_n^2}{n}\log(npB_nK_n)
\end{aligned}$$

where Δ_n *is a fixed sequence going to zero.*

Lemma G.2 allows for several different cases including cases where Y_u is generated by a non-smooth transformation of a random variable Y. For example, if $Y_u = 1\{Y \leqslant u\}$ where Y has bounded conditional probability density function, we have $d_u = 1$, $B_n = 1$, $\nu = 1$, $L_n = \sup_y f_{Y|X}(y \mid x)$. A similar result holds for independent non-identically distributed data.

In what follows for a vector $\delta\in\mathbb{R}^p$, and a set of indices $T\subseteq\{1,\dots,p\}$, we denote by $\delta_T\in\mathbb{R}^p$ the vector such that $(\delta_T)_j = \delta_j$ if $j\in T$ and $(\delta_T)_j = 0$ if $j\notin T$. For a set T, $|T|$ denotes the cardinality of T. Moreover, let

$$\Delta_{\mathbf{c},u} := \{\delta\in\mathbb{R}^p : \|\delta_{T_u^c}\|_1 \leqslant \mathbf{c}\|\delta_{T_u}\|_1\}.$$

G.2. **Finite Sample Results: Linear Case.** For the model described in (6.1) with $\Lambda(t) = t$ and $M(y,t) = \frac{1}{2}(y-t)^2$ we will study the finite sample properties of the associated Lasso and Post-Lasso estimators of $(\theta_u)_{u\in\mathcal{U}}$ defined in relations (6.2) and (6.3).

The analysis relies on the restricted eigenvalues

$$\kappa_{\mathbf{c}} = \inf_{u\in\mathcal{U}} \min_{\delta\in\Delta_{\mathbf{c},u}} \frac{\|f(X)'\delta\|_{\mathbb{P}_n,2}}{\|\delta_{T_u}\|}, \tag{G.1}$$

and maximum and minimum sparse eigenvalues

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

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 G.3 (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 uniformly in $u\in\mathcal{U}$*

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

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 G.4 (Sparsity bound for Lasso)**.** *Consider the Lasso estimator $\widehat{\theta}_u$, its support $\widehat{T}_u = \operatorname{supp}(\widehat{\theta}_u)$, and let $\widehat{s}_u = \|\widehat{\theta}_u\|_0$. 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 $\tilde{\mathbf{c}} = c_0\sup_{u\in\mathcal{U}} \|\widehat{\Psi}_{u0}\|_\infty\|\widehat{\Psi}_{u0}^{-1}\|_\infty$ we have uniformly over $u\in\mathcal{U}$*

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

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

Lemma G.5 (Rate of Convergence of Post-Lasso)**.** *Under Conditions WL, let $\widetilde{\theta}_u$ be the Post-Lasso estimator based on the support $\widetilde{T}_u$. Then, with probability $1-o(1)$, uniformly over $u\in\mathcal{U}$, we have for $\tilde{s}_u = |\widetilde{T}_u|$*

$$\|\mathrm{E}_P[Y_u \mid X] - f(X)'\widetilde{\theta}_u\|_{\mathbb{P}_n,2} \leqslant C\frac{\sqrt{\tilde{s}_u\log(p\vee n^{d_u+1})}}{\sqrt{n\ \phi_{\min}(\tilde{s}_u)}}\|\widehat{\Psi}_{u0}\|_\infty + \min_{\operatorname{supp}(\theta)\subseteq\widetilde{T}_u}\|\mathrm{E}_P[Y_u \mid X] - f(X)'\theta\|_{\mathbb{P}_n,2}$$

Moreover, if $\operatorname{supp}(\widehat{\theta}_u) \subseteq \widetilde{T}_u$ *for every* $u \in \mathcal{U}$*, 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*

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

G.3. **Finite Sample Results: Logistic Case.** For the model described in (6.1) with $\Lambda(t) = \exp(t)/\{1+\exp(t)\}$ and $M(y,t) = -\{1\{y=1\}\log(\Lambda(t)) + 1\{y=0\}\log(1-\Lambda(t))\}$ we will study finite the sample properties of the associated Lasso and Post-Lasso estimators of $(\theta_u)_{u\in\mathcal{U}}$ defined in relations (6.2) and (6.3). In what follows we use the notation

$$M_u(\theta) = \mathbb{E}_n[M(Y_u, f(X)'\theta)].$$

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 $w_{ui} = \mathrm{E}_P[Y_{ui} \mid X_i](1 - \mathrm{E}_P[Y_{ui} \mid X_i])$, $i = 1,\ldots,n$, $u \in \mathcal{U}$, is the conditional variance of the outcome variable Y_{ui}.

For $T_u = \operatorname{supp}(\theta_u)$, $s_u = \|\theta_u\|_0 \geqslant 1$, the (logistic) restricted eigenvalue is defined as

$$\bar{\kappa}_{\mathbf{c}} := \inf_{u\in\mathcal{U}} \min_{\delta\in\Delta_{\mathbf{c},u}} \frac{\|\sqrt{w_u}f(X)'\delta\|_{\mathbb{P}_n,2}}{\|\delta_{T_u}\|}. \tag{G.2}$$

For a subset $A_u \subset \mathbb{R}^p$, $u \in \mathcal{U}$, let the non-linear impact coefficient (Belloni and Chernozhukov, 2011; Belloni, Chernozhukov, and Wei, 2013) be defined as

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

Note that $\bar{q}_{A_u}$ can be bounded 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 can lead to interesting bounds provided A_u is appropriate (like the restrictive set $\Delta_{\mathbf{c},u}$ in the definition of restricted eigenvalues). In Lemma G.6 we have $A_u = \Delta_{2\tilde{\mathbf{c}},u} \cup \{\delta \in \mathbb{R}^p : \|\delta\|_1 \leqslant \frac{6c\|\widehat{\Psi}_{u0}^{-1}\|_\infty}{\ell c-1}\frac{n}{\lambda}\|\frac{r_u}{\sqrt{w_u}}\|_{\mathbb{P}_n,2}\|\sqrt{w_u}f(X)'\delta\|_{\mathbb{P}_n,2}\}$, for $u \in \mathcal{U}$. For this choice of sets, and provided that with probability $1 - o(1)$ we have $\ell c > c' > 1$, $\sup_{u\in\mathcal{U}}\|r_u/\sqrt{w_u}\|_{\mathbb{P}_n,2} \lesssim \sqrt{s\log(p\vee n)/n}$, $\sup_{u\in\mathcal{U}}\|\widehat{\Psi}_{u0}^{-1}\|_\infty \lesssim 1$ and $\sqrt{n\log(p\vee n)} \lesssim \lambda$, we have that uniformly over $u \in \mathcal{U}$, with probability $1 - o(1)$

$$\bar{q}_{A_u} \geqslant \frac{1}{\max_{i\leqslant n}\|f(X)\|_\infty}\left(\frac{\bar{\kappa}_{2\tilde{\mathbf{c}}}}{\sqrt{s_u}(1+2\tilde{\mathbf{c}})} \wedge \frac{(\lambda/n)(\ell c-1)}{6c\|\widehat{\Psi}_{u0}^{-1}\|_\infty\|r_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}}\right) \gtrsim \frac{\bar{\kappa}_{2\tilde{\mathbf{c}}}}{\sqrt{s}\max_{i\leqslant n}\|f(X_i)\|_\infty}. \tag{G.4}$$

The definitions above differ from their counterpart in the analysis of ℓ_1-penalized least squares estimators by the weighting $0 \leqslant w_{ui} \leqslant 1$. 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

$$\tilde{r}_{ui} = \tilde{r}_u(X_i) = \Lambda^{-1}(\ \Lambda(f(X_i)'\theta_u) + r_{ui}\ ) - f(X_i)'\theta_u, \tag{G.5}$$

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 follows that $|r_{ui}| \leqslant |\tilde{r}_{ui}|$ and that[28] $|\tilde{r}_{ui}| \leqslant |r_{ui}|/\inf_{0\leqslant t\leqslant \tilde{r}_{ui}} \Lambda'(f(X_i'\theta_u) + t) \leqslant |r_{ui}|/\{w_{ui} - 2|r_{ui}|\}_+$.

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

Lemma G.6 (Rates of Convergence for ℓ_1-Logistic Estimator)**.** *Assume that*

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

for $c > 1$. Further, let $\ell\widehat{\Psi}_{u0} \leqslant \widehat{\Psi}_u \leqslant L\widehat{\Psi}_{u0}$ for $L \geqslant 1 \geqslant \ell > 1/c$, 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

$$A_u = \Delta_{2\tilde{\mathbf{c}},u} \cup \{\delta : \|\delta\|_1 \leqslant \frac{6c\|\widehat{\Psi}_{u0}^{-1}\|_\infty}{\ell c - 1}\frac{n}{\lambda}\|r_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\|\sqrt{w_u}f(X)'\delta\|_{\mathbb{P}_n,2}\}.$$

Provided that the nonlinear impact coefficient $\bar{q}_{A_u} > 3\left\{(L+\frac{1}{c})\|\widehat{\Psi}_{u0}\|_\infty\frac{\lambda\sqrt{s}}{n\bar{\kappa}_{2\tilde{\mathbf{c}}}} + 9\tilde{\mathbf{c}}\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\right\}$ for every $u \in \mathcal{U}$, we have uniformly over $u \in \mathcal{U}$

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

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

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

Lemma G.7 (Sparsity of ℓ_1-Logistic Estimator)**.** *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 $\ell\widehat{\Psi}_{u0} \leqslant \widehat{\Psi}_u \leqslant L\widehat{\Psi}_{u0}$ for $L \geqslant 1 \geqslant \ell > 1/c$, uniformly over $u \in \mathcal{U}$, $c_0 = (Lc+1)/(\ell c - 1)$, $\tilde{\mathbf{c}} = c_0\sup_{u\in\mathcal{U}} \|\widehat{\Psi}_{u0}\|_\infty\|\widehat{\Psi}_{u0}^{-1}\|_\infty$ and $A_u = \Delta_{2\tilde{\mathbf{c}},u} \cup \{\delta : \|\delta\|_1 \leqslant \frac{6c\|\widehat{\Psi}_{u0}^{-1}\|_\infty}{\ell c-1}\frac{n}{\lambda}\|r_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\|\sqrt{w_u}f(X)'\delta\|_{\mathbb{P}_n,2}\}$, and $\bar{q}_{A_u} > 3\left\{(L+\frac{1}{c})\|\widehat{\Psi}_{u0}\|_\infty\frac{\lambda\sqrt{s}}{n\bar{\kappa}_{2\tilde{\mathbf{c}}}} + 9\tilde{\mathbf{c}}\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\right\}$ for every $u \in \mathcal{U}$. Then for $\widehat{s}_u = \|\widehat{\theta}_u\|_0$, uniformly over $u \in \mathcal{U}$,*

$$\widehat{s}_u \leqslant \left(\min_{m\in\mathcal{M}} \phi_{\max}(m)\right)\left[\frac{c_0}{\psi(A_u)}\left\{3\|\widehat{\Psi}_{u0}\|_\infty\frac{\sqrt{s}}{\bar{\kappa}_{2\tilde{\mathbf{c}}}} + 28\tilde{\mathbf{c}}\frac{n\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}}{\lambda}\right\}\right]^2$$

where $\mathcal{M} = \left\{m \in \mathbb{N} : m > 2\left[\frac{c_0}{\psi(A_u)}\sup_{u\in\mathcal{U}}\left\{3\|\widehat{\Psi}_{u0}\|_\infty\frac{\sqrt{s}}{\bar{\kappa}_{2\tilde{\mathbf{c}}}} + 28\tilde{\mathbf{c}}\frac{n\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}}{\lambda}\right\}\right]^2\right\}$.

[28]The last relation follows from noting that for the logistic function we have $\inf_{0\leqslant t\leqslant\tilde{r}_{ui}} \Lambda'(f(X_i'\theta_u) + t) = \min\{\Lambda'(f(X_i'\theta_u) + \tilde{r}_{ui}), \Lambda'(f(X_i'\theta_u))\}$ since Λ' is unimodal. Moreover, $\Lambda'(f(X_i'\theta_u) + \tilde{r}_{ui}) = w_{ui}$ and $\Lambda'(f(X_i'\theta_u)) = \Lambda(f(X_i'\theta_u))[1 - \Lambda(f(X_i'\theta_u))] = [\Lambda(f(X_i'\theta_u)) + r_{ui} - r_{ui}][1 - \Lambda(f(X_i'\theta_u)) - r_{ui} + r_{ui}] \geqslant w_{ui} - 2|r_{ui}|$ since $|r_{ui}| \leqslant 1$.

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

$$\widehat{s}_u \leqslant \left(\min_{m\in\mathcal{M}}\phi_{\max}(m)\right) 4c_0^2\left\{3\|\widehat{\Psi}_{u0}\|_\infty\frac{\sqrt{s}}{\bar{\kappa}_{2\tilde{\mathbf{c}}}}+28\tilde{\mathbf{c}}\frac{n\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}}{\lambda}\right\}^2$$

where $\mathcal{M}=\left\{m\in\mathbb{N}: m>8c_0^2\sup_{u\in\mathcal{U}}\left[3\|\widehat{\Psi}_{u0}\|_\infty\frac{\sqrt{s}}{\bar{\kappa}_{2\tilde{\mathbf{c}}}}+28\tilde{\mathbf{c}}\frac{n\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}}{\lambda}\right]^2\right\}$.

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 G.8 (Rate of Convergence for Post-ℓ_1-Logistic Estimator)**.** *Consider* $\widetilde{\theta}_u$ *defined as the post model selection logistic regression with the support* $\widetilde{T}_u$ *and let* $\tilde{s}_u:=|\widetilde{T}_u|$. *Uniformly over* $u\in\mathcal{U}$ *we have*

$$\|\sqrt{w_u}f(X)'(\widetilde{\theta}_u-\theta_u)\|_{\mathbb{P}_n,2}\leqslant\sqrt{3}\sqrt{0\vee\{M_u(\widetilde{\theta}_u)-M_u(\theta_u)\}}+3\left\{\frac{\sqrt{\tilde{s}_u+s_u}\|\mathbb{E}_n[f(X)\zeta_u]\|_\infty}{\psi_u(A_u)\sqrt{\phi_{\min}(\tilde{s}_u+s_u)}}+3\left\|\frac{\tilde{r}_u}{\sqrt{w_u}}\right\|_{\mathbb{P}_n,2}\right\}$$

provided that, for every $u\in\mathcal{U}$ *and* $A_u=\{\delta\in\mathbb{R}^p:\|\delta\|_0\leqslant\tilde{s}_u+s_u\}$,

$$\bar{q}_{A_u}>6\left\{\frac{\sqrt{\tilde{s}_u+s_u}\|\mathbb{E}_n[f(X)\zeta_u]\|_\infty}{\psi_u(A_u)\sqrt{\phi_{\min}(\tilde{s}_u+s_u)}}+3\left\|\frac{\tilde{r}_u}{\sqrt{w_u}}\right\|_{\mathbb{P}_n,2}\right\}\ \text{and}\ \bar{q}_{A_u}>6\sqrt{0\vee\{M_u(\widetilde{\theta}_u)-M_u(\theta_u)\}}.$$

Comment G.1. Since for a sparse vector δ such that $\|\delta\|_0=k$ we have $\|\delta\|_1\leqslant\sqrt{k}\|\delta\|\leqslant\sqrt{k}\|f(X)'\delta\|_{\mathbb{P}_n,2}/\sqrt{\phi_{\min}(k)}$, the results above can directly establish bounds on the rate of convergence in the ℓ_1-norm.

G.4. **Proofs for Lasso with Functional Response: Penalty Level.**

Proof of Lemma G.1. By the triangle inequality

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

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

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

Lemma G.9 (**Moderate deviations for self-normalized sums**)**.** *Let* $Z_1,\ldots,Z_n$ *be independent, zero-mean random variables and* $\mu\in(0,1]$. *Let* $S_{n,n}=\sum_{i=1}^n Z_i,\ \ V_{n,n}^2=\sum_{i=1}^n Z_i^2$,

$$M_n=\left\{\frac{1}{n}\sum_{i=1}^n\mathrm{E}[Z_i^2]\right\}^{1/2}\Big/\left\{\frac{1}{n}\sum_{i=1}^n\mathrm{E}[|Z_i|^{2+\mu}]\right\}^{1/\{2+\mu\}}>0$$

and $0<\ell_n\leqslant n^{\frac{\mu}{2(2+\mu)}}M_n$. *Then for some absolute constant* A,

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

For each $j = 1, \ldots, p$, and each $u \in \mathcal{U}^\epsilon$, we will apply Lemma G.9 with $Z_i := f_j(X_i)\zeta_{ui}$, and $\mu = 1$. Then, by Lemma G.9, the union bound, and $|\mathcal{U}^\epsilon| \leqslant n^{d_u}$, we have

$$\begin{aligned} \mathrm{P}_P\left(\sup_{u\in\mathcal{U}^\epsilon}\max_{j\leqslant p}\left|\frac{\sqrt{n}\mathbb{E}_n[f_j(X)\zeta_u]}{\sqrt{\mathbb{E}_n[f_j(X)^2\zeta_u^2]}}\right| > \Phi^{-1}(1-\tfrac{\gamma}{2pn^{d_u}})\right) &\leqslant 2pn^{d_u}(\gamma/2pn^{d_u})\{1+o(1)\} \\ &\leqslant \gamma\{1+o(1)\} \end{aligned} \tag{G.6}$$

provided that $\max_{u,j}\{\bar{\mathrm{E}}_P[|f_j(X)\zeta_u|^3]^{1/3}/\bar{\mathrm{E}}_P[|f_j(X)\zeta_u|^2]^{1/2}\}\Phi^{-1}(1-\gamma/2pn^{d_u}) \leqslant \delta_n n^{1/6}$, which holds by Condition WL (under this condition there is $\ell_n \to \infty$ obeying conditions of Lemma G.9.)

Moreover, by triangle inequality we have

$$\begin{aligned} &\sup_{u\in\mathcal{U}^\epsilon, u'\in\mathcal{U}, d_{\mathcal{U}}(u,u')\leqslant\epsilon} \|\widehat{\Psi}_{u0}^{-1}\mathbb{E}_n[f(X)\zeta_u] - \widehat{\Psi}_{u'0}^{-1}\mathbb{E}_n[f(X)\zeta_{u'}]\|_\infty \\ &\leqslant \sup_{u\in\mathcal{U}^\epsilon, u'\in\mathcal{U}, d_{\mathcal{U}}(u,u')\leqslant\epsilon} \|(\widehat{\Psi}_{u0}^{-1} - \widehat{\Psi}_{u'0}^{-1})\widehat{\Psi}_{u0}\|_\infty \|\widehat{\Psi}_{u0}^{-1}\mathbb{E}_n[f(X)\zeta_u]\|_\infty \\ &+ \sup_{u,u'\in\mathcal{U}, d_{\mathcal{U}}(u,u')\leqslant\epsilon} \|\mathbb{E}_n[f(X)(\zeta_u - \zeta_{u'})]\|_\infty \|\widehat{\Psi}_{u'0}^{-1}\|_\infty. \end{aligned} \tag{G.7}$$

To control the first term in (G.7) we note that by Condition WL, $\widehat{\Psi}_{u0jj}$ is bounded away from zero with probability $1 - o(1)$ uniformly over $u \in \mathcal{U}$ and $j = 1, \ldots, p$. Thus we have uniformly over $u \in \mathcal{U}$ and $j = 1, \ldots, p$

$$|(\widehat{\Psi}_{u0jj}^{-1} - \widehat{\Psi}_{u'0jj}^{-1})\widehat{\Psi}_{u0jj}| = |\widehat{\Psi}_{u0jj} - \widehat{\Psi}_{u'0jj}|/\widehat{\Psi}_{u'0jj} \leqslant C|\widehat{\Psi}_{u0jj} - \widehat{\Psi}_{u'0jj}| \tag{G.8}$$

with the same probability. Moreover, we have

$$\begin{aligned} &\sup_{u,u'\in\mathcal{U}, d_{\mathcal{U}}(u,u')\leqslant\epsilon} \max_{j\leqslant p} |\{\mathbb{E}_n[f_j(X)^2\zeta_u^2]\}^{1/2} - \{\mathbb{E}_n[f_j(X)^2\zeta_{u'}^2]\}^{1/2}| \\ &\qquad \leqslant \sup_{u,u'\in\mathcal{U}, d_{\mathcal{U}}(u,u')\leqslant\epsilon} \max_{j\leqslant p}\{\mathbb{E}_n[f_j(X)^2(\zeta_u - \zeta_{u'})^2]\}^{1/2}. \end{aligned} \tag{G.9}$$

Thus, with $\epsilon = 1/n$, relations (G.8) and (G.9) imply that with probability $1 - o(1)$

$$\sup_{u,u'\in\mathcal{U}, d_{\mathcal{U}}(u,u')\leqslant\epsilon} \|(\widehat{\Psi}_{u0}^{-1} - \widehat{\Psi}_{u'0}^{-1})\widehat{\Psi}_{u0}\|_\infty \lesssim \sup_{u,u'\in\mathcal{U}, d_{\mathcal{U}}(u,u')\leqslant 1/n} \max_{j\leqslant p}\{\mathbb{E}_n[f_j(X)^2(\zeta_u - \zeta_{u'})^2]\}^{1/2}.$$

By (G.6)

$$\sup_{u\in\mathcal{U}^\epsilon} \|\widehat{\Psi}_{u0}^{-1}\mathbb{E}_n[f(X)\zeta_u]\|_\infty \leqslant C'\sqrt{\log(p \vee n^{d_u+1})/n}$$

with probability $1 - o(1)$, so that with the same probability

$$\begin{aligned} &\sup_{u\in\mathcal{U}^\epsilon, u'\in\mathcal{U}, d_{\mathcal{U}}(u,u')\leqslant\epsilon} \|(\widehat{\Psi}_{u0}^{-1} - \widehat{\Psi}_{u'0}^{-1})\widehat{\Psi}_{u0}\|_\infty \|\widehat{\Psi}_{u0}^{-1}\mathbb{E}_n[f(X)\zeta_u]\|_\infty \\ &\leqslant \sup_{u,u'\in\mathcal{U}, d_{\mathcal{U}}(u,u')\leqslant 1/n} \max_{j\leqslant p}\{\mathbb{E}_n[f_j(X)^2(\zeta_u - \zeta_{u'})^2]\}^{1/2} C'\sqrt{\tfrac{\log(p\vee n^{d_u+1})}{n}} \leqslant \tfrac{o(1)}{\sqrt{n}} \end{aligned}$$

where the last inequality follows by Condition WL.

Since $\epsilon = 1/n$, the last term in (G.7) is of the order $o(n^{-1/2})$ with probability $1 - o(1)$ since by Condition WL,

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

with probability $1-\Delta_n$, and noting that by Condition WL $\sup_{u\in\mathcal{U}}\|\widehat{\Psi}_{u0}^{-1}\|_\infty$ is uniformly bounded with probability at least $1-o(1)-\Delta_n$.

The results above imply that (G.7) is bounded by $o(1)/\sqrt{n}$ with probability $1-o(1)$. Since $\frac{1}{2}\sqrt{\log(2pn^{d_u}/\gamma)} \leqslant \Phi^{-1}(1-\gamma/\{2pn^{d_u}\})$ by $\gamma/\{2pn^{d_u}\} \leqslant 1/4$ and standard (lower) tail bounds, we have that with probability $1-o(1)$

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

and the result follows. ∎

Proof of Lemma G.2. We start with the last statement of the lemma since it is more difficult (others will use similar calculations). Consider the class of functions $\mathcal{F} = \{Y_u : u\in\mathcal{U}\}$, $\mathcal{F}' = \{\mathrm{E}_P[Y_u \mid X] : u\in\mathcal{U}\}$, and $\mathcal{G} = \{\zeta_u^2 = (Y_u - \mathrm{E}_P[Y_u \mid X])^2 : u\in\mathcal{U}\}$. Let F be a measurable envelope for $\mathcal{F}$ which satisfies $F \leqslant B_n$.

Because $\mathcal{F}$ is a VC-class of functions with VC index $C'd_u$, by Lemma C.2(1) we have

$$\log N(\epsilon\|F\|_{Q,2}, \mathcal{F}, \|\cdot\|_{Q,2}) \lesssim 1 + [d_u \log(e/\epsilon) \vee 0]. \tag{G.10}$$

To bound the covering number for $\mathcal{F}'$ we apply Lemma C.3, and since $\mathrm{E}[F \mid X] \leqslant F$, we have

$$\log\sup_Q N(\epsilon\|F\|_{Q,2}, \mathcal{F}', \|\cdot\|_{Q,2}) \leqslant \log\sup_Q N(\tfrac{\epsilon}{2}\|F\|_{Q,2}, \mathcal{F}, \|\cdot\|_{Q,2}). \tag{G.11}$$

Since $\mathcal{G} \subset (\mathcal{F}-\mathcal{F}')^2$, $G = 4F^2$ is an envelope for $\mathcal{G}$ and the covering number for $\mathcal{G}$ satisfies

$$\begin{aligned}
\log N(\epsilon\|4F^2\|_{Q,2}, \mathcal{G}, \|\cdot\|_{Q,2}) &\overset{(i)}{\leqslant} 2\log N(\tfrac{\epsilon}{2}\|2F\|_{Q,2}, \mathcal{F}-\mathcal{F}', \|\cdot\|_{Q,2}) \\
&\overset{(ii)}{\leqslant} 2\log N(\tfrac{\epsilon}{4}\|F\|_{Q,2}, \mathcal{F}, \|\cdot\|_{Q,2}) + 2\log N(\tfrac{\epsilon}{4}\|F\|_{Q,2}, \mathcal{F}', \|\cdot\|_{Q,2}) \\
&\overset{(iii)}{\leqslant} 4\log\sup_Q N(\tfrac{\epsilon}{8}\|F\|_{Q,2}, \mathcal{F}, \|\cdot\|_{Q,2}),
\end{aligned} \tag{G.12}$$

where (i) and (ii) follow by Lemma C.2(2), and (iii) follows from (G.11).

Hence, the entropy bound for the class $\mathcal{M} = \cup_{j\in[p]}\mathcal{M}_j$, where $\mathcal{M}_j = \{f_j^2(X)\mathcal{G}\}$, $j\in[p]$ and envelope $M = 4K_n^2F^2$, satisfies

$$\begin{aligned}
\log N(\epsilon\|M\|_{Q,2}, \mathcal{M}, \|\cdot\|_{Q,2}) &\overset{(a)}{\leqslant} \log p + \max_{j\in[p]}\log N(\epsilon\|4K_n^2F^2\|_{Q,2}, \mathcal{M}_j, \|\cdot\|_{Q,2}) \\
&\overset{(b)}{\leqslant} \log p + \log N(\epsilon\|4F^2\|_{Q,2}, \mathcal{G}, \|\cdot\|_{Q,2}) \\
&\overset{(c)}{\leqslant} \log p + 4\log\sup_Q N(\tfrac{\epsilon}{8}\|F\|_{Q,2}, \mathcal{F}, \|\cdot\|_{Q,2}) \\
&\overset{(d)}{\lesssim} \log p + [(1+d_u)\log(e/\epsilon) \vee 0],
\end{aligned}$$

where (a) follows by Lemma C.2(2) for union of classes, (b) holds by Lemma C.2(2) when one class has only a single function, (c) by (G.12) and (d) follows from (G.10) and $\epsilon \leqslant 1$. Therefore, since

$\sup_{u\in\mathcal{U}}\max_{j\leqslant p}\mathrm{E}_P[f_j^2(X)\zeta_u^2]$ is bounded away from zero and from above, by Lemma C.1 we have with probability $1-O(1/\log n)$ that

$$\sup_{u\in\mathcal{U}}\max_{j\leqslant p}|(\mathbb{E}_n-\mathrm{E}_P)[f_j^2(X)\zeta_u^2]| \lesssim \sqrt{\tfrac{(1+d_u)\log(npK_n^2B_n^2)}{n}}+\tfrac{(1+d_u)K_n^2B_n^2}{n}\log(npB_n^2K_n^2).$$

using the envelope $M=4K_n^2B_n^2$, $v=C'$, $a=pn$ and a constant σ.

Consider the first term. By Lemma C.1 we have with probability $1-O(1/\log n)$ that

$$\begin{aligned}\sup_{d_{\mathcal{U}}(u,u')\leqslant 1/n}\|\mathbb{E}_n[f(X)(\zeta_u-\zeta_{u'})]\|_\infty &= \sup_{d_{\mathcal{U}}(u,u')\leqslant 1/n}\frac{1}{\sqrt{n}}\max_{j\leqslant p}|\mathbb{G}_n(f_j(X)(\zeta_u-\zeta_{u'}))|\\ &\lesssim \tfrac{1}{\sqrt{n}}\sqrt{\tfrac{(1+d_u)L_n\log(pnK_nB_n\frac{n^\nu}{L_n})}{n^\nu}}+\tfrac{(1+d_u)K_nB_n\log(pnK_nB_n\frac{n^\nu}{L_n})}{n}\end{aligned}$$

using the envelope $F=2K_nB_n$, $v=C'$, $a=pn$, the entropy bound in Lemma C.3, and $\sigma^2\propto L_nn^{-\nu}\leqslant F^2$ for all n sufficiently large, because $L_nn^{-\nu}\searrow 0$ and

$$\begin{aligned}\sup_{d_{\mathcal{U}}(u,u')\leqslant 1/n}\max_{j\leqslant p}\mathrm{E}_P[f_j(X)^2(\zeta_u-\zeta_{u'})^2] &\leqslant \sup_{d_{\mathcal{U}}(u,u')\leqslant 1/n}\max_{j\leqslant p}\mathrm{E}_P[f_j(X)^2(Y_u-Y_{u'})^2]\\ &\leqslant \sup_{d_{\mathcal{U}}(u,u')\leqslant 1/n}L_n|u-u'|^\nu\max_{j\leqslant p}\mathrm{E}_P[f_j(X)^2]\leqslant CL_nn^{-\nu}.\end{aligned}$$

To bound the second term in the statement of the lemma, it follows that

$$\begin{aligned}\sup_{d_{\mathcal{U}}(u,u')\leqslant 1/n}\max_{j\leqslant p}\mathbb{E}_n[f_j(X)^2(\zeta_u-\zeta_{u'})^2] &= \sup_{d_{\mathcal{U}}(u,u')\leqslant 1/n}\max_{j\leqslant p}\mathbb{E}_n[f_j(X)^2(\mathrm{E}_P[Y_u-Y_{u'}\mid X])^2]\\ &\leqslant \sup_{d_{\mathcal{U}}(u,u')\leqslant 1/n}\max_{j\leqslant p}\mathbb{E}_n[f_j(X)^2\mathrm{E}_P[|Y_u-Y_{u'}|^2\mid X]]\\ &\leqslant \max_{j\leqslant p}\mathbb{E}_n[f_j(X)^2]\sup_{d_{\mathcal{U}}(u,u')\leqslant 1/n}L_n|u-u'|^\nu\end{aligned}\tag{G.13}$$

where the first inequality holds by Jensen's inequality, and the second inequality holds by assumption. Since $c\leqslant\max_{j\leqslant p}\{\mathrm{E}_P[f_j(X)^2]\}^{1/2}\leqslant C$, the result follows by Lemma C.1 which yields with probability $1-O(1/\log n)$

$$\max_{j\leqslant p}|(\mathbb{E}_n-\mathrm{E}_P)[f_j(X)^2]| \lesssim \sqrt{\frac{\log(pnK_n^2)}{n}}+\frac{K_n^2}{n}\log(pnK_n^2),\tag{G.14}$$

where we used the choice $C\leqslant\sigma=C'\leqslant F=K_n^2$, $v=C$, $a=pn$. ∎

G.5. **Proofs for Lasso with Functional Response: Linear Case.**

Proof of Lemma G.3. Let $\widehat{\delta}_u=\widehat{\theta}_u-\theta_u$. Throughout the proof we assume that the events $c_r^2\geqslant\sup_{u\in\mathcal{U}}\mathbb{E}_n[r_u^2]$, $\lambda/n\geqslant c\sup_{u\in\mathcal{U}}\|\widehat{\Psi}_{u0}^{-1}\mathbb{E}_n[\zeta_uf(X)]\|_\infty$ and $\ell\widehat{\Psi}_{u0}\leqslant\widehat{\Psi}_u\leqslant L\widehat{\Psi}_{u0}$ occur.

By definition of $\widehat{\theta}_u$,

$$\widehat{\theta}_u\in\arg\min_{\theta\in\mathbb{R}^p}\mathbb{E}_n[(Y_u-f(X)'\theta)^2]+\frac{2\lambda}{n}\|\widehat{\Psi}_u\theta\|_1,$$

and $\ell\widehat{\Psi}_{u0} \leqslant \widehat{\Psi}_u \leqslant L\widehat{\Psi}_{u0}$, we have

$$\begin{aligned}
&\mathbb{E}_n[(f(X)'\widehat{\delta}_u)^2] - 2\mathbb{E}_n[(Y_u - f(X)'\theta_u)f(X)]'\widehat{\delta}_u \\
&= \mathbb{E}_n[(Y_u - f(X)'\widehat{\theta}_u)^2] - \mathbb{E}_n[(Y_u - f(X)'\theta_u)^2] \\
&\leqslant \tfrac{2\lambda}{n}\|\widehat{\Psi}_u\theta_u\|_1 - \tfrac{2\lambda}{n}\|\widehat{\Psi}_u\widehat{\theta}_u\|_1 \\
&\leqslant \tfrac{2\lambda}{n}\|\widehat{\Psi}_u\widehat{\delta}_{uT_u}\|_1 - \tfrac{2\lambda}{n}\|\widehat{\Psi}_u\widehat{\delta}_{uT_u^c}\|_1 \\
&\leqslant \tfrac{2\lambda}{n}L\|\widehat{\Psi}_{u0}\widehat{\delta}_{uT_u}\|_1 - \tfrac{2\lambda}{n}\ell\|\widehat{\Psi}_{u0}\widehat{\delta}_{uT_u^c}\|_1.
\end{aligned} \tag{G.15}$$

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

$$\begin{aligned}
&\mathbb{E}_n[(f(X)'\widehat{\delta}_u)^2] \\
&\leqslant 2\mathbb{E}_n[r_u f(X)]'\widehat{\delta}_u + 2(\widehat{\Psi}_{u0}^{-1}\mathbb{E}_n[\zeta_u f(X)])'(\widehat{\Psi}_{u0}\widehat{\delta}_u) + \tfrac{2\lambda}{n}L\|\widehat{\Psi}_{u0}\widehat{\delta}_{uT_u}\|_1 - \tfrac{2\lambda}{n}\ell\|\widehat{\Psi}_{u0}\widehat{\delta}_{uT_u^c}\|_1 \\
&\leqslant 2c_r\{\mathbb{E}_n[(f(X)'\widehat{\delta}_u)^2]\}^{1/2} + 2\|\widehat{\Psi}_{u0}^{-1}\mathbb{E}_n[\zeta_u f(X)]\|_\infty\|\widehat{\Psi}_{u0}\widehat{\delta}_u\|_1 + \tfrac{2\lambda}{n}L\|\widehat{\Psi}_{u0}\widehat{\delta}_{uT_u}\|_1 - \tfrac{2\lambda}{n}\ell\|\widehat{\Psi}_{u0}\widehat{\delta}_{uT_u^c}\|_1 \\
&\leqslant 2c_r\{\mathbb{E}_n[(f(X)'\widehat{\delta}_u)^2]\}^{1/2} + \tfrac{2\lambda}{cn}\|\widehat{\Psi}_{u0}\widehat{\delta}_u\|_1 + \tfrac{2\lambda}{n}L\|\widehat{\Psi}_{u0}\widehat{\delta}_{uT_u}\|_1 - \tfrac{2\lambda}{n}\ell\|\widehat{\Psi}_{u0}\widehat{\delta}_{uT_u^c}\|_1 \\
&\leqslant 2c_r\{\mathbb{E}_n[(f(X)'\widehat{\delta}_u)^2]\}^{1/2} + \tfrac{2\lambda}{n}\left(L + \tfrac{1}{c}\right)\|\widehat{\Psi}_{u0}\widehat{\delta}_{uT_u}\|_1 - \tfrac{2\lambda}{n}\left(\ell - \tfrac{1}{c}\right)\|\widehat{\Psi}_{u0}\widehat{\delta}_{uT_u^c}\|_1.
\end{aligned} \tag{G.16}$$

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 $\widehat{\delta}_u \notin \Delta_{\tilde{\mathbf{c}},u} = \{\delta \in \mathbb{R}^p : \|\delta_{T_u^c}\|_1 \leqslant \tilde{\mathbf{c}}\|\delta_{T_u}\|_1\}$, 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 $\widehat{\delta}_u \in \Delta_{\tilde{\mathbf{c}},u}$. In this case (G.16), the definition of $\kappa_{\tilde{\mathbf{c}}}$, and $\|\widehat{\delta}_{uT_u}\|_1 \leqslant \sqrt{s}\|\widehat{\delta}_{uT_u}\|$, we have

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

which implies

$$\{\mathbb{E}_n[(f(X)'\widehat{\delta}_u)^2]\}^{1/2} \leqslant 2c_r + \frac{2\lambda\sqrt{s}}{n\kappa_{\tilde{\mathbf{c}}}}\left(L + \frac{1}{c}\right)\|\widehat{\Psi}_{u0}\|_\infty. \tag{G.17}$$

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

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

where we used that $\|\widehat{\delta}_{uT_u}\|_1 \leqslant \sqrt{s}\|\widehat{\delta}_{uT_u}\|$, the definition of the restricted eigenvalue, and the prediction rate derived in (G.17).

Otherwise note that $\widehat{\delta}_u \notin \Delta_{2\tilde{\mathbf{c}},u}$ 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 (G.16) yields

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

where we used that $\max_t t(2c_r - t) \leqslant c_r^2$. 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 G.4. Step 1. 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 Step 2 below and the definition of L_u we have

$$\widehat{s}_u \leqslant \phi_{\max}(\widehat{s}_u) L_u^2. \tag{G.18}$$

Consider any $M \in \mathcal{M} = \{m \in \mathbb{N} : m > 2\phi_{\max}(m) \sup_{u \in \mathcal{U}} L_u^2\}$, and suppose $\widehat{s}_u > M$.

Next recall the sublinearity of the maximum sparse eigenvalue (for a proof see Lemma 3 in Belloni and Chernozhukov (2013)), namely, for any integer $k \geqslant 0$ and constant $\ell \geqslant 1$ we have $\phi_{\max}(\ell k) \leqslant \lceil \ell \rceil \phi_{\max}(k)$, where $\lceil \ell \rceil$ denotes the ceiling of ℓ. Therefore

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

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 (G.18) 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 $M \in \mathcal{M}$.

Step 2. In this step we establish that uniformly over $u \in \mathcal{U}$

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

Let $R_u = (r_{u1}, \ldots, r_{un})'$, $\mathbf{Y}_u = (Y_{u1}, \ldots, Y_{un})'$, $\bar{\zeta}_u = (\zeta_{u1}, \ldots, \zeta_{un})'$, and $F = [f(X_1); \ldots; f(X_n)]'$. We have from the optimality conditions that the Lasso estimator $\widehat{\theta}_u$ satisfies

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

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

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

where we used that $\|v\| \leqslant \|v\|_0^{1/2}\|v\|_\infty$ and

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

Since $\lambda/c \geqslant \sup_{u\in\mathcal{U}} \|\widehat{\Psi}_{u0}^{-1} F'\bar{\zeta}_u\|_\infty$, and by Lemma G.3, we have that the estimate $\widehat{\theta}_u$ satisfies $\|f(X)'(\widehat{\theta}_u - \theta_u)\|_{\mathbb{P}_n,2} \leqslant 2c_r + 2\left(L + \frac{1}{c}\right)\frac{\lambda\sqrt{s}}{n\kappa_{\tilde{\mathbf{c}}}}\|\widehat{\Psi}_{u0}\|_\infty$ so that

$$\begin{aligned}
\sqrt{\widehat{s}_u} &\leqslant \frac{\sqrt{\phi_{\max}(\widehat{s}_u)}\frac{\|\widehat{\Psi}_{u0}^{-1}\|_\infty}{\ell}\left[\frac{3nc_r}{\lambda} + 3\left(L + \frac{1}{c}\right)\frac{\sqrt{s}}{\kappa_{\tilde{\mathbf{c}}}}\|\widehat{\Psi}_{u0}\|_\infty\right]}{\left(1 - \frac{1}{c\ell}\right)} \\
&\leqslant 4\frac{\left(L + \frac{1}{c}\right)}{\left(1 - \frac{1}{c\ell}\right)}\frac{1}{\ell}\sqrt{\phi_{\max}(\widehat{s}_u)}\|\widehat{\Psi}_{u0}^{-1}\|_\infty\left[\frac{nc_r}{\lambda} + \frac{\sqrt{s}}{\kappa_{\tilde{\mathbf{c}}}}\|\widehat{\Psi}_{u0}\|_\infty\right].
\end{aligned}$$

The result follows by noting that $(L + [1/c])/(1 - 1/[\ell c]) = c_0\ell$ by definition of c_0.

■

Proof of Lemma G.5. Define $m_u := (\mathrm{E}[Y_{u1} \mid X_1], \ldots, \mathrm{E}[Y_{un} \mid X_n])'$, $\bar{\zeta}_u := (\zeta_{u1}, \ldots, \zeta_{un})'$, and the $n \times p$ matrix $F := [f(X_1); \ldots; f(X_n)]'$. For a set of indices $S \subset \{1, \ldots, p\}$ we define $\widehat{P}_S = F[S](F[S]'F[S])^{-1}F[S]'$ denote the projection matrix on the columns associated with the indices in S where we interpret $\widehat{P}_S$ as a null operator if S is empty.

Since $Y_{ui} = m_{ui} + \zeta_{ui}$ we have

$$m_u - F\widetilde{\theta}_u = (I - \widehat{P}_{\widetilde{T}_u})m_u - \widehat{P}_{\widetilde{T}_u}\bar{\zeta}_u$$

where I is the identity operator. Therefore

$$\|m_u - F\widetilde{\theta}_u\| \leqslant \|(I - \widehat{P}_{\widetilde{T}_u})m_u\| + \|\widehat{P}_{\widetilde{T}_u}\bar{\zeta}_u\|. \tag{G.19}$$

Since $\|F[\widetilde{T}_u]/\sqrt{n}(F[\widetilde{T}_u]'F[\widetilde{T}_u]/n)^{-1}\| \leqslant \sqrt{1/\phi_{\min}(\tilde{s}_u)}$, the last term in (G.19) satisfies

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

By Lemma G.1 with $\gamma = 1/n$, we have that with probability $1 - o(1)$, uniformly in $u \in \mathcal{U}$

$$\|F'\bar{\zeta}_u/\sqrt{n}\|_\infty \leqslant C\sqrt{\log(p \vee n^{d_u+1})}\max_{1\leqslant j\leqslant p}\sqrt{\mathbb{E}_n[f_j(X)^2\zeta_u^2]} = C\sqrt{\log(p \vee n^{d_u+1})}\|\widehat{\Psi}_{u0}\|_\infty.$$

The result follows.

The last statement follows from noting that the mean square approximation error provides an upper bound to the best mean square approximation error based on the model $\widetilde{T}_u$ provided that

the model include the Lasso's mode, i.e. $\widehat{T}_u \subseteq \widetilde{T}_u$. Indeed, we have

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

where we invoked Lemma G.3 to bound $\|f(X)'(\widehat{\theta}_u - \theta_u)\|_{\mathbb{P}_n,2}$. ∎

G.6. **Proofs for Lasso with Functional Response: Logistic Case.**

Proof of Lemma G.6. 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 $M_u(\widehat{\theta}_u) + \frac{\lambda}{n}\|\widehat{\Psi}_u\widehat{\theta}_u\|_1 \leqslant M_u(\theta_u) + \frac{\lambda}{n}\|\widehat{\Psi}_u\theta_u\|_1$. Thus,

$$\begin{aligned}M_u(\widehat{\theta}_u) - M_u(\theta_u) &\leqslant \tfrac{\lambda}{n}\|\widehat{\Psi}_u\theta_u\|_1 - \tfrac{\lambda}{n}\|\widehat{\Psi}_u\widehat{\theta}_u\|_1 \\ &\leqslant \tfrac{\lambda}{n}\|\widehat{\Psi}_u\delta_{u,T_u}\|_1 - \tfrac{\lambda}{n}\|\widehat{\Psi}_u\delta_{u,T_u^c}\|_1 \leqslant \tfrac{\lambda L}{n}\|\widehat{\Psi}_{u0}\delta_{u,T_u}\|_1 - \tfrac{\lambda\ell}{n}\|\widehat{\Psi}_{u0}\delta_{u,T_u^c}\|_1.\end{aligned} \tag{G.20}$$

Moreover, by convexity of $M_u(\cdot)$ and Hölder's inequality we have

$$M_u(\widehat{\theta}_u) - M_u(\theta_u) \geqslant \partial_\theta M_u(\theta_u) \geqslant -\tfrac{\lambda}{n}\tfrac{1}{c}\|\widehat{\Psi}_{u0}\delta_u\|_1 - \|r_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\|\sqrt{w_u}f(X)'\delta_u\|_{\mathbb{P}_n,2} \tag{G.21}$$

because

$$\begin{aligned}|\partial_\theta M_u(\theta_u)'\delta_u| &= |S_u'\delta_u + \{\partial_\theta M_u(\theta_u) - S_u\}'\delta_u| \leqslant |S_u'\delta_u| + |\{\partial_\theta M_u(\theta_u) - S_u\}'\delta_u| \\ &\leqslant \|\widehat{\Psi}_{u0}^{-1}S_u\|_\infty\|\widehat{\Psi}_{u0}\delta_u\|_1 + \|r_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\|\sqrt{w_u}f(X)'\delta_u\|_{\mathbb{P}_n,2} \\ &\leqslant \frac{\lambda}{n}\frac{1}{c}\|\widehat{\Psi}_{u0}\delta_u\|_1 + \|r_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\|\sqrt{w_u}f(X)'\delta_u\|_{\mathbb{P}_n,2},\end{aligned} \tag{G.22}$$

where we used that $\lambda/n \geqslant c\sup_{u\in\mathcal{U}}\|\widehat{\Psi}_{u0}^{-1}S_u\|_\infty$ and that $\partial_\theta M_u(\theta_u) = \mathbb{E}_n[\{\zeta_u + r_u\}f(X)]$ so that

$$|\{\partial_\theta M_u(\theta_u) - S_u\}'\delta_u| = |\mathbb{E}_n[r_u f(X)'\delta_u]| \leqslant \|r_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\|\sqrt{w_u}f(X)'\delta_u\|_{\mathbb{P}_n,2}. \tag{G.23}$$

Combining (G.20) and (G.21) we have

$$\frac{\lambda}{n}\frac{c\ell - 1}{c}\|\widehat{\Psi}_{u0}\delta_{u,T_u^c}\|_1 \leqslant \frac{\lambda}{n}\frac{Lc+1}{c}\|\widehat{\Psi}_{u0}\delta_{u,T_u}\|_1 + \|r_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\|\sqrt{w_u}f(X)'\delta_u\|_{\mathbb{P}_n,2} \tag{G.24}$$

and for $\tilde{\mathbf{c}} = \frac{Lc+1}{\ell c-1}\sup_{u\in\mathcal{U}}\|\widehat{\Psi}_{u0}\|_\infty\|\widehat{\Psi}_{u0}^{-1}\|_\infty \geqslant 1$ we have

$$\|\delta_{u,T_u^c}\|_1 \leqslant \tilde{\mathbf{c}}\|\delta_{u,T_u}\|_1 + \frac{n}{\lambda}\frac{c\|\widehat{\Psi}_{u0}^{-1}\|_\infty}{\ell c - 1}\|r_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\|\sqrt{w_u}f(X)'\delta_u\|_{\mathbb{P}_n,2}.$$

Suppose $\delta_u \notin \Delta_{2\tilde{\mathbf{c}},u}$, namely $\|\delta_{u,T_u^c}\|_1 \geqslant 2\tilde{\mathbf{c}}\|\delta_{u,T_u}\|_1$. Thus,

$$\begin{aligned}\|\delta_u\|_1 &\leqslant (1 + \{2\tilde{\mathbf{c}}\}^{-1})\|\delta_{u,T_u^c}\|_1 \\ &\leqslant (1 + \{2\tilde{\mathbf{c}}\}^{-1})\tilde{\mathbf{c}}\|\delta_{u,T_u}\|_1 + (1 + \{2\tilde{\mathbf{c}}\}^{-1})\tfrac{n}{\lambda}\tfrac{c\|\widehat{\Psi}_{u0}^{-1}\|_\infty}{\ell c-1}\|r_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\|\sqrt{w_u}f(X)'\delta_u\|_{\mathbb{P}_n,2} \\ &\leqslant (1 + \{2\tilde{\mathbf{c}}\}^{-1})\tfrac{1}{2}\|\delta_{u,T_u^c}\|_1 + (1 + \{2\tilde{\mathbf{c}}\}^{-1})\tfrac{n}{\lambda}\tfrac{c\|\widehat{\Psi}_{u0}^{-1}\|_\infty}{\ell c-1}\|r_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\|\sqrt{w_u}f(X)'\delta_u\|_{\mathbb{P}_n,2}.\end{aligned}$$

The relation above implies that if $\delta_u \notin \Delta_{2\tilde{\mathbf{c}},u}$

$$\begin{aligned}\|\delta_u\|_1 &\leqslant \tfrac{4\tilde{\mathbf{c}}}{2\tilde{\mathbf{c}}-1}(1+\{2\tilde{\mathbf{c}}\}^{-1})\tfrac{n}{\lambda}\tfrac{c\|\widehat{\Psi}_{u0}^{-1}\|_\infty}{\ell c-1}\|r_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\|\sqrt{w_u}f(X)'\delta_u\|_{\mathbb{P}_n,2} \\ &\leqslant \tfrac{6c\|\widehat{\Psi}_{u0}^{-1}\|_\infty}{\ell c-1}\tfrac{n}{\lambda}\|r_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\|\sqrt{w_u}f(X)'\delta_u\|_{\mathbb{P}_n,2} =: I_u,\end{aligned} \tag{G.25}$$

where we used that $\frac{4\tilde{\mathbf{c}}}{2\tilde{\mathbf{c}}-1}(1+\{2\tilde{\mathbf{c}}\}^{-1}) \leqslant 6$ since $\tilde{\mathbf{c}} \geqslant 1$. Combining the bound with the bound

$$\|\delta_{u,T_u}\|_1 \leqslant \frac{\sqrt{s}}{\bar{\kappa}_{2\tilde{\mathbf{c}}}}\|\sqrt{w_u}f(X)'\delta_u\|_{\mathbb{P}_n,2} =: II_u, \quad \text{if } \delta_u \in \Delta_{2\tilde{\mathbf{c}},u},$$

we have that δ_u satisfies

$$\|\delta_{u,T_u}\|_1 \leqslant I_u + II_u. \tag{G.26}$$

For every $u \in \mathcal{U}$, since $A_u = \Delta_{2\tilde{\mathbf{c}},u} \cup \{\delta : \|\delta\|_1 \leqslant \frac{6c\|\widehat{\Psi}_{u0}^{-1}\|_\infty}{\ell c-1}\frac{n}{\lambda}\|r_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\|\sqrt{w_u}f(X)'\delta\|_{\mathbb{P}_n,2}\}$, it follows that $\delta_u \in A_u$, and we have

$$\begin{aligned}&\tfrac{1}{3}\|\sqrt{w_u}f(X)'\delta_u\|_{\mathbb{P}_n,2}^2 \wedge \left\{\tfrac{\bar{q}_{A_u}}{3}\|\sqrt{w_u}f(X)'\delta_u\|_{\mathbb{P}_n,2}\right\} \\ &\leqslant_{(1)} M_u(\widehat{\theta}_u) - M_u(\theta_u) - \partial_\theta M_u(\theta_u)'\delta_u + 2\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\|\sqrt{w_u}f(X)'\delta_u\|_{\mathbb{P}_n,2} \\ &\leqslant_{(2)} (L+\tfrac{1}{c})\tfrac{\lambda}{n}\|\widehat{\Psi}_{u0}\delta_{u,T_u}\|_1 + 3\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\|\sqrt{w_u}f(X)'\delta_u\|_{\mathbb{P}_n,2} \\ &\leqslant_{(3)} (L+\tfrac{1}{c})\|\widehat{\Psi}_{u0}\|_\infty\tfrac{\lambda}{n}\{I_u + II_u\} + 3\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\|\sqrt{w_u}f(X)'\delta_u\|_{\mathbb{P}_n,2} \\ &\leqslant_{(4)} \left\{(L+\tfrac{1}{c})\|\widehat{\Psi}_{u0}\|_\infty\tfrac{\lambda\sqrt{s}}{n\bar{\kappa}_{2\tilde{\mathbf{c}}}} + 9\tilde{\mathbf{c}}\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\right\}\|\sqrt{w_u}f(X)'\delta_u\|_{\mathbb{P}_n,2},\end{aligned}$$

where (1) follows by Lemma G.10 with A_u, (2) follows from (G.22) and $|r_{ui}| \leqslant |\tilde{r}_{ui}|$, (3) follows by $\|\widehat{\Psi}_{u0}\delta_{u,T_u}\|_1 \leqslant \|\widehat{\Psi}_{u0}\|_\infty\|\delta_{u,T_u}\|_1$ and (G.26), (4) follows from simplifications and $|r_{ui}| \leqslant |\tilde{r}_{ui}|$. Since the inequality $(x^2 \wedge ax) \leqslant bx$ holding for $x > 0$ and $b < a < 0$ implies $x \leqslant b$, the above system of the inequalities, provided that for every $u \in \mathcal{U}$

$$\bar{q}_{A_u} > 3\left\{(L+\frac{1}{c})\|\widehat{\Psi}_{u0}\|_\infty\frac{\lambda\sqrt{s}}{n\bar{\kappa}_{2\tilde{\mathbf{c}}}} + 9\tilde{\mathbf{c}}\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\right\},$$

implies that

$$\|\sqrt{w_u}f(X)'\delta_u\|_{\mathbb{P}_n,2} \leqslant 3\left\{(L+\frac{1}{c})\|\widehat{\Psi}_{u0}\|_\infty\frac{\lambda\sqrt{s}}{n\bar{\kappa}_{2\tilde{\mathbf{c}}}} + 9\tilde{\mathbf{c}}\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\right\} =: III_u \quad \text{for every} \quad u \in \mathcal{U}.$$

The second result follows from the definition of $\bar{\kappa}_{2\tilde{\mathbf{c}}}$, (G.25) and the bound on $\|\sqrt{w_u}f(X)'\delta_u\|_{\mathbb{P}_n,2}$ just derived, namely for every $u \in \mathcal{U}$ we have

$$\begin{aligned}\|\delta_u\|_1 &\leqslant 1\{\delta_u \in \Delta_{2\tilde{\mathbf{c}},u}\}\|\delta_u\|_1 + 1\{\delta_u \notin \Delta_{2\tilde{\mathbf{c}},u}\}\|\delta_u\|_1 \\ &\leqslant (1+2\tilde{\mathbf{c}})II_u + I_u \leqslant 3\left\{\tfrac{(1+2\tilde{\mathbf{c}})\sqrt{s}}{\bar{\kappa}_{2\tilde{\mathbf{c}}}} + \tfrac{6c\|\widehat{\Psi}_{u0}^{-1}\|_\infty}{\ell c-1}\tfrac{n}{\lambda}\left\|\tfrac{r_u}{\sqrt{w_u}}\right\|_{\mathbb{P}_n,2}\right\} III_u\end{aligned}$$

■

Proof of Lemma G.7. The proof of both bounds are similar to the proof of sparsity for the linear case (Lemma G.4) differing only on the definition of L_u which are a consequence of pre-sparsity bounds established in Step 2 and Step 3.

Step 1. To establish the first bound by Step 2 below, triangle inequality and the definition of $\psi(A_u)$ we have

$$\begin{aligned}\sqrt{\widehat{s}_u} &\leqslant \tfrac{c(n/\lambda)}{(c\ell-1)}\sqrt{\phi_{\max}(\widehat{s}_u)}\|f(X)'(\widehat{\theta}_u-\theta_u)-r_u\|_{\mathbb{P}_n,2}\\ &\leqslant \tfrac{c(n/\lambda)}{(c\ell-1)}\sqrt{\phi_{\max}(\widehat{s}_u)}\left\{\tfrac{\|\sqrt{w_u}f(X)'(\widehat{\theta}_u-\theta_u)\|_{\mathbb{P}_n,2}}{\psi(A_u)}+\|r_u\|_{\mathbb{P}_n,2}\right\}\end{aligned}$$

uniformly in $u\in\mathcal{U}$. By Lemma G.6, $\psi(A_u)\leqslant 1$ and $\|r_u\|_{\mathbb{P}_n,2}\leqslant\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}$ we have

$$\begin{aligned}\sqrt{\widehat{s}_u} &\leqslant \sqrt{\phi_{\max}(\widehat{s}_u)}\tfrac{c(n/\lambda)}{(c\ell-1)\psi(A_u)}\left\{3(L+\tfrac{1}{c})\|\widehat{\Psi}_{u0}\|_\infty\tfrac{(\lambda/n)\sqrt{s}}{\bar{\kappa}_{2\tilde{\mathbf{c}}}}+28\tilde{\mathbf{c}}\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\right\}\\ &\leqslant \sqrt{\phi_{\max}(\widehat{s}_u)}\tfrac{c_0}{\psi(A_u)}\left\{3\|\widehat{\Psi}_{u0}\|_\infty\tfrac{\sqrt{s}}{\bar{\kappa}_{2\tilde{\mathbf{c}}}}+28\tilde{\mathbf{c}}\tfrac{n\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}}{\lambda}\right\}\end{aligned}$$

Let $L_u=\frac{c_0}{\psi(A_u)}\left\{3\|\widehat{\Psi}_{u0}\|_\infty\frac{\sqrt{s}}{\bar{\kappa}_{2\tilde{\mathbf{c}}}}+28\tilde{\mathbf{c}}\frac{n\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}}{\lambda}\right\}$. Thus we have

$$\widehat{s}_u\leqslant\phi_{\max}(\widehat{s}_u)L_u^2. \tag{G.27}$$

which has the same structure as (G.18) in the Step 1 of the proof of Lemma G.4.

Consider any $M\in\mathcal{M}=\{m\in\mathbb{N}: m>2\phi_{\max}(m)\sup_{u\in\mathcal{U}}L_u^2\}$, and suppose $\widehat{s}_u>M$. By the sublinearity of the maximum sparse eigenvalue (Lemma 3 in Belloni and Chernozhukov (2013)), for any integer $k\geqslant 0$ and constant $\ell\geqslant 1$ we have $\phi_{\max}(\ell k)\leqslant\lceil\ell\rceil\phi_{\max}(k)$, where $\lceil\ell\rceil$ denotes the ceiling of ℓ. Therefore

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

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 (G.27) 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 $M\in\mathcal{M}$.

Next we establish the second bound. By Step 3 below we have

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

By Lemma G.6 and that $\|\sqrt{w_u}\tilde{r}_u\|_{\mathbb{P}_n,2}\leqslant\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}$ we have

$$\begin{aligned}\sqrt{\widehat{s}_u} &\leqslant \sqrt{\phi_{\max}(\widehat{s}_u)}\tfrac{2c(n/\lambda)}{(c\ell-1)}\left\{3(L+\tfrac{1}{c})\|\widehat{\Psi}_{u0}\|_\infty\tfrac{(\lambda/n)\sqrt{s}}{\bar{\kappa}_{2\tilde{\mathbf{c}}}}+28\tilde{\mathbf{c}}\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\right\}\\ &\leqslant \sqrt{\phi_{\max}(\widehat{s}_u)}2c_0\left\{3\|\widehat{\Psi}_{u0}\|_\infty\tfrac{\sqrt{s}}{\bar{\kappa}_{2\tilde{\mathbf{c}}}}+28\tilde{\mathbf{c}}\tfrac{n\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}}{\lambda}\right\}\end{aligned}$$

Let $L_u=2c_0\left\{3\|\widehat{\Psi}_{u0}\|_\infty\frac{\sqrt{s}}{\bar{\kappa}_{2\tilde{\mathbf{c}}}}+28\tilde{\mathbf{c}}\frac{n\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}}{\lambda}\right\}$. Thus again we obtained the relation (G.18) and the proof follows similarly to the Step 1 in the proof of Lemma G.4.

Step 2. In this step we show that uniformly over $u\in\mathcal{U}$,

$$\sqrt{\widehat{s}_u}\leqslant\frac{c(n/\lambda)}{(c\ell-1)}\sqrt{\phi_{\max}(\widehat{s}_u)}\|f(X)'(\widehat{\theta}_u-\theta_u)-r_u\|_{\mathbb{P}_n,2}. \tag{G.28}$$

Let $\Lambda_{ui}:=\mathrm{E}_P[Y_{ui}\mid X_i]$ and $S_u=-\mathbb{E}_n[f(X)\zeta_u]=-\mathbb{E}_n[(Y_u-\Lambda_u)f(X)]$. Let $\widehat{T}_u=\mathrm{supp}(\widehat{\theta}_u)$, $\widehat{s}_u=\|\widehat{\theta}_u\|_0$, $\delta_u=\widehat{\theta}_u-\theta_u$, and $\widehat{\Lambda}_{ui}=\exp(f(X_i)'\widehat{\theta}_u)/\{1+\exp(f(X_i)'\widehat{\theta}_u)\}$. 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

$$\begin{aligned}
\tfrac{\lambda}{n}\sqrt{\widehat{s}_u} &= \|(\widehat{\Psi}_u^{-1}\mathbb{E}_n[(Y_u - \widehat{\Lambda}_u)f(X))_{\widehat{T}_u}]\| \\
&\leqslant \|\widehat{\Psi}_u^{-1}\widehat{\Psi}_{u0}\|_\infty \|\widehat{\Psi}_{u0}^{-1}\mathbb{E}_n[(Y_u - \Lambda_u)f_{\widehat{T}_u}(X)]\| + \|\widehat{\Psi}_u^{-1}\widehat{\Psi}_{u0}\|_\infty \|\widehat{\Psi}_{u0}^{-1}\|_\infty \|\mathbb{E}_n[(\widehat{\Lambda}_u - \Lambda_u)f_{\widehat{T}_u}(X)]\| \\
&\leqslant \sqrt{\widehat{s}_u}(1/\ell)\|\widehat{\Psi}_{u0}^{-1}\mathbb{E}_n[\zeta_u f(X)]\|_\infty + (1/\ell)\|\widehat{\Psi}_{u0}^{-1}\|_\infty \sup_{\|\theta\|_0 \leqslant \widehat{s}_u, \|\theta\|=1} \mathbb{E}_n[|\widehat{\Lambda}_u - \Lambda_u|\ |f(X)'\theta|] \\
&\leqslant \tfrac{\lambda}{\ell cn}\sqrt{\widehat{s}_u} + \sqrt{\phi_{\max}(\widehat{s}_u)}(1/\ell)\|\widehat{\Psi}_{u0}^{-1}\|_\infty \|f(X)'\delta_u - r_u\|_{\mathbb{P}_n,2}
\end{aligned}$$

uniformly in $u \in \mathcal{U}$, where we used that Λ is 1-Lipschitz. This relation implies (G.28).

Step 3. In this step we show that if $\max_{i\leqslant n}|f(X_i)'(\widehat{\theta}_u - \theta_u) - \tilde{r}_{ui}| \leqslant 1$ we have

$$\sqrt{\widehat{s}_u} \leqslant \frac{2c(n/\lambda)}{(c\ell - 1)}\sqrt{\phi_{\max}(\widehat{s}_u)}\|\sqrt{w_u}\{f(X)'(\widehat{\theta}_u - \theta_u) - \tilde{r}_u\}\|_{\mathbb{P}_n,2} \tag{G.29}$$

Note that uniformly in $u \in \mathcal{U}$, Lemma G.13 establishes 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. Thus, combining this bound with the calculations performed in Step 2 we obtain

$$\tfrac{\lambda}{n}\sqrt{\widehat{s}_u} \leqslant \tfrac{\lambda}{\ell cn}\sqrt{\widehat{s}_u} + (2/\ell)\|\widehat{\Psi}_{u0}^{-1}\|_\infty\sqrt{\phi_{\max}(\widehat{s}_u)}\|\sqrt{w_u}\{f(X)'\delta_u - \tilde{r}_u\}\|_{\mathbb{P}_n,2}$$

which implies (G.29). ∎

Proof of Lemma G.8. 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 G.10 with $A_u = \{\delta \in \mathbb{R}^p : \|\delta\|_0 \leqslant \tilde{s}_u + s_u\}$, we have

$$\begin{aligned}
\tfrac{1}{3}\tilde{t}_u^2 \wedge \left\{\tfrac{\bar{q}_{A_u}}{3}\tilde{t}_u\right\} &\leqslant M_u(\tilde{\theta}_u) - M_u(\theta_u) - \partial_\theta M_u(\theta_u)'\tilde{\delta}_u + 2\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\tilde{t}_u \\
&\leqslant M_u(\tilde{\theta}_u) - M_u(\theta_u) + \|S_u\|_\infty\|\tilde{\delta}_u\|_1 + 3\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\tilde{t}_u \\
&\leqslant M_u(\tilde{\theta}_u) - M_u(\theta_u) + \tilde{t}_u\left\{\frac{\sqrt{\tilde{s}_u + s_u}\|S_u\|_\infty}{\psi_u(A_u)\sqrt{\phi_{\min}(\tilde{s}_u + s_u)}} + 3\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\right\}.
\end{aligned}$$

where the second inequality holds by calculations as in (G.22) and Hölder's inequality, and the last inequality follows from

$$\|\tilde{\delta}_u\|_1 \leqslant \sqrt{\tilde{s}_u + s_u}\|\tilde{\delta}_u\|_1 \leqslant \frac{\sqrt{\tilde{s}_u + s_u}}{\sqrt{\phi_{\min}(\tilde{s}_u + s_u)}}\|f(X)'\tilde{\delta}_u\|_{\mathbb{P}_n,2} \leqslant \frac{\sqrt{\tilde{s}_u + s_u}}{\sqrt{\phi_{\min}(\tilde{s}_u + s_u)}}\frac{\|\sqrt{w_u}f(X)'\tilde{\delta}_u\|_{\mathbb{P}_n,2}}{\psi_u(A_u)}$$

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

Recall the assumed conditions $\bar{q}_{A_u}/6 > \left\{\frac{\sqrt{\tilde{s}_u + s_u}\|S_u\|_\infty}{\psi_u(A_u)\sqrt{\phi_{\min}(\tilde{s}_u + s_u)}} + 3\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\right\}$ and $\bar{q}_{A_u}/6 > \sqrt{M_u(\tilde{\theta}_u) - M_u(\theta_u)}$. If $\frac{1}{3}\tilde{t}_u^2 > \left\{\frac{\bar{q}_{A_u}}{3}\tilde{t}_u\right\}$, then

$$\frac{\bar{q}_{A_u}}{3}\tilde{t}_u \leqslant \frac{\bar{q}_{A_u}}{6}\sqrt{M_u(\tilde{\theta}_u) - M_u(\theta_u)} + \frac{\bar{q}_{A_u}}{6}\tilde{t}_u,$$

so that $\tilde{t}_u \leqslant \sqrt{0 \vee \{M_u(\tilde{\theta}_u) - M_u(\theta_u)\}}$ which implies the result. Otherwise, we have

$$\frac{1}{3}\tilde{t}_u^2 \leqslant \{M_u(\tilde{\theta}_u) - M_u(\theta_u)\} + \tilde{t}_u\left\{\frac{\sqrt{\tilde{s}_u + s_u}\|S_u\|_\infty}{\psi_u(A_u)\sqrt{\phi_{\min}(\tilde{s}_u + s_u)}} + 3\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\right\},$$

since for positive numbers a, b, c, inequality $a^2 \leqslant b + ac$ implies $a \leqslant \sqrt{b} + c$, we have

$$\tilde{t}_u \leqslant \sqrt{3}\sqrt{0 \vee \{M_u(\tilde{\theta}_u) - M_u(\theta_u)\}} + 3\left\{\frac{\sqrt{\tilde{s}_u + s_u}\|S_u\|_\infty}{\psi_u(A_u)\sqrt{\phi_{\min}(\tilde{s}_u + s_u)}} + 3\|\tilde{r}_{ui}/\sqrt{w_{ui}}\|_{\mathbb{P}_n,2}\right\}.$$

■

G.7. **Technical Lemmas: Logistic Case.** The proof of the following lower bound builds upon ideas developed in Belloni and Chernozhukov (2011) for high-dimensional quantile regressions.

Lemma G.10 (Minoration Lemma)**.** *For any $u \in \mathcal{U}$ and $\delta \in A_u \subset \mathbb{R}^p$, we have*

$$\begin{aligned} &M_u(\theta_u + \delta) - M_u(\theta_u) - \partial_\theta M_u(\theta_u)'\delta + 2\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\|\sqrt{w_u}f(X)'\delta\|_{\mathbb{P}_n,2} \\ &\geqslant \left\{\tfrac{1}{3}\|\sqrt{w_u}f(X)'\delta\|^2_{\mathbb{P}_n,2}\right\} \wedge \left\{\tfrac{\bar{q}_{A_u}}{3}\|\sqrt{w_u}f(X)'\delta\|_{\mathbb{P}_n,2}\right\} \end{aligned}$$

where

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

Proof. Step 1. (Minoration). Consider the following non-negative convex function

$$F_u(\delta) = M_u(\theta_u + \delta) - M_u(\theta_u) - \partial_\theta M_u(\theta_u)'\delta + 2\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\|\sqrt{w_u}f(X)'\delta\|_{\mathbb{P}_n,2}.$$

Note that if $\bar{q}_{A_u} = 0$ the statement is trivial since $F_u(\delta) \geqslant 0$. Thus we can assume $\bar{q}_{A_u} > 0$.

Step 2 below shows that for any $\delta = t\tilde{\delta} \in \mathbb{R}^p$ where $t \in \mathbb{R}$ and $\tilde{\delta} \in A_u$ such that $\|\sqrt{w_u}f(X)'\delta\|_{\mathbb{P}_n,2} \leqslant \bar{q}_{A_u}$ we have

$$F_u(\delta) \geqslant \frac{1}{3}\|\sqrt{w_u}f(X)'\delta\|^2_{\mathbb{P}_n,2}. \tag{G.30}$$

Thus (G.30) covers the case that $\delta \in A_u$ and $\|\sqrt{w_u}f(X)'\delta\|_{\mathbb{P}_n,2} \leqslant \bar{q}_{A_u}$.

In the case that $\delta \in A_u$ and $\|\sqrt{w_u}f(X)'\delta\|_{\mathbb{P}_n,2} > \bar{q}_{A_u}$, by convexity[29] of F_u and $F_u(0) = 0$ we have

$$F_u(\delta) \geqslant \frac{\|\sqrt{w_u}f(X)'\delta\|_{\mathbb{P}_n,2}}{\bar{q}_{A_u}} F_u\left(\delta\frac{\bar{q}_{A_u}}{\|\sqrt{w_u}f(X)'\delta\|_{\mathbb{P}_n,2}}\right) \geqslant \frac{\bar{q}_{A_u}\|\sqrt{w_u}f(X)'\delta\|_{\mathbb{P}_n,2}}{3}, \tag{G.31}$$

where the last step follows by (G.30) since

$$\|\sqrt{w_u}f(X)'\bar{\delta}\|_{\mathbb{P}_n,2} = \bar{q}_{A_u} \text{ for } \bar{\delta} = \delta\frac{\bar{q}_{A_u}}{\|\sqrt{w_u}f(X)'\delta\|_{\mathbb{P}_n,2}}.$$

Combining (G.30) and (G.31) we have

$$F_u(\delta) \geqslant \left\{\frac{1}{3}\|\sqrt{w_u}f(X)'\delta\|^2_{\mathbb{P}_n,2}\right\} \wedge \left\{\frac{\bar{q}_{A_u}}{3}\|\sqrt{w_u}f(X)'\delta\|_{\mathbb{P}_n,2}\right\}.$$

[29] If ϕ is a convex function with $\phi(0) = 0$, for $\alpha \in (0,1)$ we have $\phi(t) \geqslant \phi(\alpha t)/\alpha$. Indeed, by convexity, $\phi(\alpha t + (1-\alpha)0) \leqslant (1-\alpha)\phi(0) + \alpha\phi(t) = \alpha\phi(t)$.

Step 2. (Proof of (G.30)) Let $\tilde{r}_{ui}$ be such that $\Lambda(f(X_i)'\theta_u + \tilde{r}_{ui}) = \Lambda(f(X_i)'\theta_u) + r_{ui} = \mathrm{E}_P[Y_{ui} \mid X_i]$. Defining $g_{ui}(t) = \log\{1 + \exp(f(X_i)'\theta_u + \tilde{r}_{ui} + tf(X_i)'\delta)\}$, $\tilde{g}_{ui}(t) = \log\{1 + \exp(f(X_i)'\theta_u + tf(X_i)'\delta)\}$, $\Lambda_{ui} := \mathrm{E}_P[Y_{ui} \mid X_i]$, $\tilde{\Lambda}_{ui} := \exp(f(X_i)'\theta_u)/\{1 + \exp(f(X_i)'\theta_u)\}$, we have

$$
\begin{aligned}
&M_u(\theta_u + \delta) - M_u(\theta_u) - \partial_\theta M_u(\theta_u)'\delta = \\
&= \mathbb{E}_n\left[\log\{1 + \exp(f(X)'\{\theta_u + \delta\})\} - Y_u f(X)'(\theta_u + \delta)\right] \\
&\quad -\mathbb{E}_n\left[\log\{1 + \exp(f(X)'\theta_u)\} - Y_u f(X)'\theta_u\right] - \mathbb{E}_n\left[(\tilde{\Lambda}_u - Y_u)f(X)'\delta\right] \\
&= \mathbb{E}_n\left[\log\{1 + \exp(f(X)'\{\theta_u + \delta\})\} - \log\{1 + \exp(f(X)'\theta_u)\} - \tilde{\Lambda}_u f(X)'\delta\right] \\
&= \mathbb{E}_n[\tilde{g}_u(1) - \tilde{g}_u(0) - \tilde{g}_u'(0)] \\
&= \mathbb{E}_n[g_u(1) - g_u(0) - g_u'(0)] + \mathbb{E}_n[\{\tilde{g}_u(1) - g_u(1)\} - \{\tilde{g}_u(0) - g_u(0)\} - \{\tilde{g}_u'(0) - g_u'(0)\}]
\end{aligned}
\tag{G.32}
$$

Note that the function g_{ui} is three times differentiable and satisfies,

$$
g_{ui}'(t) = (f(X_i)'\delta)\Lambda_{ui}(t), \quad g_{ui}''(t) = (f(X_i)'\delta)^2\Lambda_{ui}(t)[1 - \Lambda_{ui}(t)], \quad \text{and}
$$
$$
g_{ui}'''(t) = (f(X_i)'\delta)^3\Lambda_{ui}(t)[1 - \Lambda_{ui}(t)][1 - 2\Lambda_{ui}(t)]
$$

where $\Lambda_{ui}(t) := \exp(f(X_i)'\theta_u + \tilde{r}_{ui} + tf(X_i)'\delta)/\{1 + \exp(f(X_i)'\theta_u + \tilde{r}_{ui} + tf(X)'\delta)\}$. Thus we have $|g_{ui}'''(t)| \leqslant |f(X)'\delta| g_{ui}''(t)$. Therefore, by Lemmas G.11 and G.12 given following the conclusion of this proof, we have

$$
\begin{aligned}
g_{ui}(1) - g_{ui}(0) - g_{ui}'(0) &\geqslant \tfrac{(f(X_i)'\delta)^2 w_{ui}}{(f(X_i)'\delta)^2}\left\{\exp(-|f(X_i)'\delta|) + |f(X_i)'\delta| - 1\right\} \\
&\geqslant w_{ui}\left\{\tfrac{|f(X_i)'\delta|^2}{2} - \tfrac{|f(X_i)'\delta|^3}{6}\right\}
\end{aligned}
\tag{G.33}
$$

Moreover, letting $\Upsilon_{ui}(t) = \tilde{g}_{ui}(t) - g_{ui}(t)$ we have

$$
|\Upsilon_{ui}'(t)| = |(f(X_i)'\delta)\{\Lambda_{ui}(t) - \tilde{\Lambda}_{ui}(t)\}| \leqslant |f(X_i)'\delta|\ |\tilde{r}_{ui}|
$$

where $\tilde{\Lambda}_{ui}(t) := \exp(f(X_i)'\theta_u + tf(X_i)'\delta)/\{1 + \exp(f(X_i)'\theta_u + tf(X_i)'\delta)\}$. Thus

$$
\begin{aligned}
&|\mathbb{E}_n[\{\tilde{g}_u(1) - g_u(1)\} - \{\tilde{g}_u(0) - g_u(0)\} - \{\tilde{g}_u'(0) - g_u'(0)\}]| = \\
&= |\mathbb{E}_n[\Upsilon_u(1) - \Upsilon_u(0) - \{\tilde{\Lambda}_u - \Lambda_u\}f(X)'\delta]| \\
&\leqslant 2\mathbb{E}_n[|\tilde{r}_u|\ |f(X)'\delta|].
\end{aligned}
\tag{G.34}
$$

Therefore, combining (G.32) with the bounds (G.33) and (G.34) we have

$$
\begin{aligned}
M_u(\theta_u + \delta) - M_u(\theta_u) - \partial_\theta M_u(\theta_u)'\delta &\geqslant \tfrac{1}{2}\mathbb{E}_n\left[w_u|f(X)'\delta|^2\right] - \tfrac{1}{6}\mathbb{E}_n\left[w_u|f(X)'\delta|^3\right] \\
&\quad -2\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\|\sqrt{w_u}f(X)'\delta\|_{\mathbb{P}_n,2},
\end{aligned}
$$

which holds for any $\delta \in \mathbb{R}^p$.

Take any $\delta = t\tilde{\delta}$, $t \in \mathbb{R} \setminus \{0\}$, $\tilde{\delta} \in A_u$ such that $\|\sqrt{w_u}f(X)'\delta\|_{\mathbb{P}_n,2} \leqslant \bar{q}_{A_u}$. (Note that the case of $\delta = 0$ is trivial.) We have

$$
\begin{aligned}
\mathbb{E}_n[w_u|f(X)'\delta|^2]^{1/2} = \|\sqrt{w_u}f(X)'\delta\|_{\mathbb{P}_n,2} \leqslant \bar{q}_{A_u} &\leqslant \mathbb{E}_n\left[w_u|f(X)'\tilde{\delta}|^2\right]^{3/2}/\mathbb{E}_n\left[w_u|f(X)'\tilde{\delta}|^3\right] \\
&= \mathbb{E}_n\left[w_u|f(X)'\delta|^2\right]^{3/2}/\mathbb{E}_n\left[w_u|f(X)'\delta|^3\right],
\end{aligned}
$$

since the scalar t cancels out. Thus, $\mathbb{E}_n[w_u|f(X)'\delta|^3] \leqslant \mathbb{E}_n[w_u|f(X)'\delta|^2]$. Therefore we have

$$\tfrac{1}{2}\mathbb{E}_n\left[w_u|f(X)'\delta|^2\right] - \tfrac{1}{6}\mathbb{E}_n\left[w_u|f(X)'\delta|^3\right] \geqslant \frac{1}{3}\mathbb{E}_n\left[w_u|f(X)'\delta|^2\right] \quad \text{and}$$

$$M_u(\theta_u+\delta) - M_u(\theta_u) - \partial_\theta M_u(\theta_u)'\delta \quad \geqslant \tfrac{1}{3}\mathbb{E}_n\left[w_u|f(X)'\delta|^2\right] - 2\|\tfrac{\tilde{r}_u}{\sqrt{w_u}}\|_{\mathbb{P}_n,2}\|\sqrt{w_u}f(X)'\delta\|_{\mathbb{P}_n,2},$$

which establishes that $F_u(\delta) := M_u(\theta_u+\delta) - M_u(\theta_u) - \partial_\theta M_u(\theta_u)'\delta + 2\|\frac{\tilde{r}_u}{\sqrt{w_u}}\|_{\mathbb{P}_n,2}\|\sqrt{w_u}f(X)'\delta\|_{\mathbb{P}_n,2}$ is larger than $\frac{1}{3}\mathbb{E}_n\left[w_u|f(X)'\delta|^2\right]$ for any $\delta = t\tilde{\delta}$, $t \in \mathbb{R}$, $\tilde{\delta} \in A_u$ and $\|\sqrt{w_u}f(X)'\delta\|_{\mathbb{P}_n,2} \leqslant \bar{q}_{A_u}$. ∎

Lemma G.11 (Lemma 1 from Bach (2010))**.** *Let $g : \mathbb{R} \to \mathbb{R}$ be a three times differentiable convex function such that for all $t \in \mathbb{R}$, $|g'''(t)| \leqslant Mg''(t)$ 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 G.12. *For $t \geqslant 0$ we have $\exp(-t) + t - 1 \geqslant \frac{1}{2}t^2 - \frac{1}{6}t^3$.*

Proof of Lemma G.12. 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 G.13. *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 $\Lambda'(t_0)\exp(-s) \leqslant \Lambda'(s+t_0) \leqslant \Lambda'(t_0)\exp(s)$. For $t > 0$, integrating one more time from 0 to t,

$$\Lambda'(t_0)\{1 - \exp(-t)\} \leqslant \Lambda(t+t_0) - \Lambda(t_0) \leqslant \Lambda'(t_0)\{\exp(t) - 1\}.$$

Similarly, for $t < 0$, integrating from t to 0, we have

$$\Lambda'(t_0)\{1 - \exp(t)\} \leqslant \Lambda(t+t_0) - \Lambda(t_0) \leqslant \Lambda'(t_0)\{\exp(-t) - 1\}.$$

The first result follows by noting that $1 - \exp(-|t|) \leqslant \exp(|t|) - 1$. The second follows by verification. ∎

References

Abadie, A. (2002): "Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models," *Journal of the American Statistical Association*, 97, 284–292.

——— (2003): "Semiparametric Instrumental Variable Estimation of Treatment Response Models," *Journal of Econometrics*, 113, 231–263.

Andrews, D. W. (1994a): "Empirical process methods in econometrics," *Handbook of Econometrics*, 4, 2247–2294.

Andrews, D. W. K. (1994b): "Asymptotics for semiparametric econometric models via stochastic equicontinuity," *Econometrica*, 62(1), 43–72.

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 (2011): "ℓ_1-Penalized Quantile Regression for High Dimensional Sparse Models," *Annals of Statistics*, 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.

——— (2013): "Inference for High-Dimensional Sparse Econometric Models," *Advances in Economics and Econometrics. 10th World Congress of Econometric Society. August 2010*, III, 245–295.

——— (2014): "Inference on Treatment Effects After Selection Amongst High-Dimensional Controls," *Review of Economic Studies*, 81, 608–650.

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., and D. A. Freedman (1981): "Some asymptotic theory for the bootstrap," *The Annals of Statistics*, pp. 1196–1217.

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.

Chen, X. (2007): "Large Sample Sieve Estimatin of Semi-Nonparametric Models," *Handbook of Econometrics*, 6, 5559–5632.

Chernozhukov, V., D. Chetverikov, and K. Kato (2012): "Gaussian approximation of suprema of empirical processes," *ArXiv e-prints*.

Chernozhukov, V., I. Fernández-Val, and B. Melly (2013): "Inference on counterfactual distributions," *Econometrica*, 81(6), 2205–2268.

Chernozhukov, V., and C. Hansen (2004): "The impact of 401(k) participation on the wealth distribution: An instrumental quantile regression analysis," *Review of Economics and Statistics*, 86(3), 735–751.

——— (2005): "An IV Model of Quantile Treatment Effects," *Econometrica*, 73(1), 245–262.

——— (2006): "Instrumental quantile regression inference for structural and treatment effect models," *J. Econometrics*, 132(2), 491–525.

Chesher, A. (2003): "Identification in nonseparable models," *Econometrica*, 71(5), 1405–1441.

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.

Escanciano, J. C., and L. Zhu (2013): "Set inferences and sensitivity analysis in semiparametric conditionally identified models," CeMMAP working papers CWP55/13, Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

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," *Working Paper*.

Frank, I. E., and J. H. Friedman (1993): "A Statistical View of Some Chemometrics Regression Tools," *Technometrics*, 35(2), 109–135.

Ghosal, S., A. Sen, and A. W. van der Vaart (2000): "Testing Monotonicity of Regression," *Ann. Statist.*, 28(4), 1054–1082.

Giné, E., and J. Zinn (1984): "Some limit theorems for empirical processes," *Ann. Probab.*, 12(4), 929–998, With discussion.

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.

Hansen, L. P. (1982): "Large sample properties of generalized method of moments estimators," *Econometrica*, 50(4), 1029–1054.

Hansen, L. P., and K. J. Singleton (1982): "Generalized instrumental variables estimation of nonlinear rational expectations models," *Econometrica*, 50(5), 1269–1286.

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.

Hong, H., and O. Scaillet (2006): "A fast subsampling method for nonlinear dynamic models," *J. Econometrics*, 133(2), 557–578.

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.

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

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.

LINTON, O. (1996): "Edgeworth approximation for MINPIN estimators in semiparametric regression models," *Econometric Theory*, 12(1), 30–60.

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. (1990): "Semiparametric efficiency bounds," *Journal of Applied Econometrics*, 5(2), 99–135.

——— (1994): "The asymptotic variance of semiparametric estimators," *Econometrica*, 62(6), 1349–1382.

——— (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.

ROBINS, J. M., AND A. ROTNITZKY (1995): "Semiparametric efficiency in multivariate regression models with missing data," *J. Amer. Statist. Assoc.*, 90(429), 122–129.

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.

SHERMAN, R. (1994): "Maximal inequalities for degenerate U-processes with applications to optimization estimators," *Ann. Statist.*, 22, 439–459.

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
Indicator	20	N	11833	16120	8972	12500
			(1638)	(2224)	(2692)	(3572)
			{1598}	{2199}	{2598}	{3360}
Indicator	20	Y	13960	16727	12211	13729
			(1690)	(2267)	(2750)	(3672)
			{1693}	{2272}	{2813}	{3868}
Indicator plus interactions	167	N	11856	16216	9996	12131
			(1632)	(2224)	(2675)	(3428)
			{1614}	{2189}	{2767}	{3385}
Indicator plus interactions	167	Y	14295	17011	13907	13476
			(1705)	(2331)	(2749)	(3759)
			{1687}	{2329}	{2741}	{3748}
Orthogonal Polynomials	22	N	9314	16089	6897	11807
			(2916)	(2155)	(3122)	(3504)
			{3000}	{2079}	{3029}	{3695}
Orthogonal Polynomials	22	Y	11578	16683	8057	11993
			(1645)	(2193)	(2704)	(3539)
			{1638}	{2254}	{2695}	{3571}
Orthogonal Polynomials plus interactions	196	N	-324200	-89042	-132900	-38723
			(282770)	(64321)	(129570)	(50593)
			{266300}	{66211}	{134380}	{49132}
Orthogonal Polynomials plus interactions	196	Y	10335	13062	8989	9656
			(2352)	(4536)	(2780)	(4002)
			{2485}	{4307}	{2840}	{4167}
Orthogonal Polynomials plus many interactions	756	N	-	-	-	-
			-	-	-	-
			-	-	-	-
Orthogonal Polynomials plus many interactions	756	Y	10118	12476	9382	10692
			(2465)	(4923)	(3014)	(4849)
			{2381}	{4881}	{3237}	{4781}

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. Orthogonal Polynomials uses a fourth order polynomial in age, an eigth order polynomial in income, and quadratic polynomials in education and family size. Polynomials in each variable are orthogonalized via the Gram-Schmidt process. Marital status, two-earner status, defined benefit pension status, IRA participation status, and home ownership status are included as indicators in all the specifications. Specifications denoted with "plus interactions" include all first-order interactions. The specifications denoted "plus many interactions" take all first-order interactions between all non-income variables and then fully interact these interactions as well as the main effects with all income variables. Analytic standard errors are given in parentheses. Bootstrap standard errors based on 500 repetitions with Mammen (1993) multipliers are given in braces.

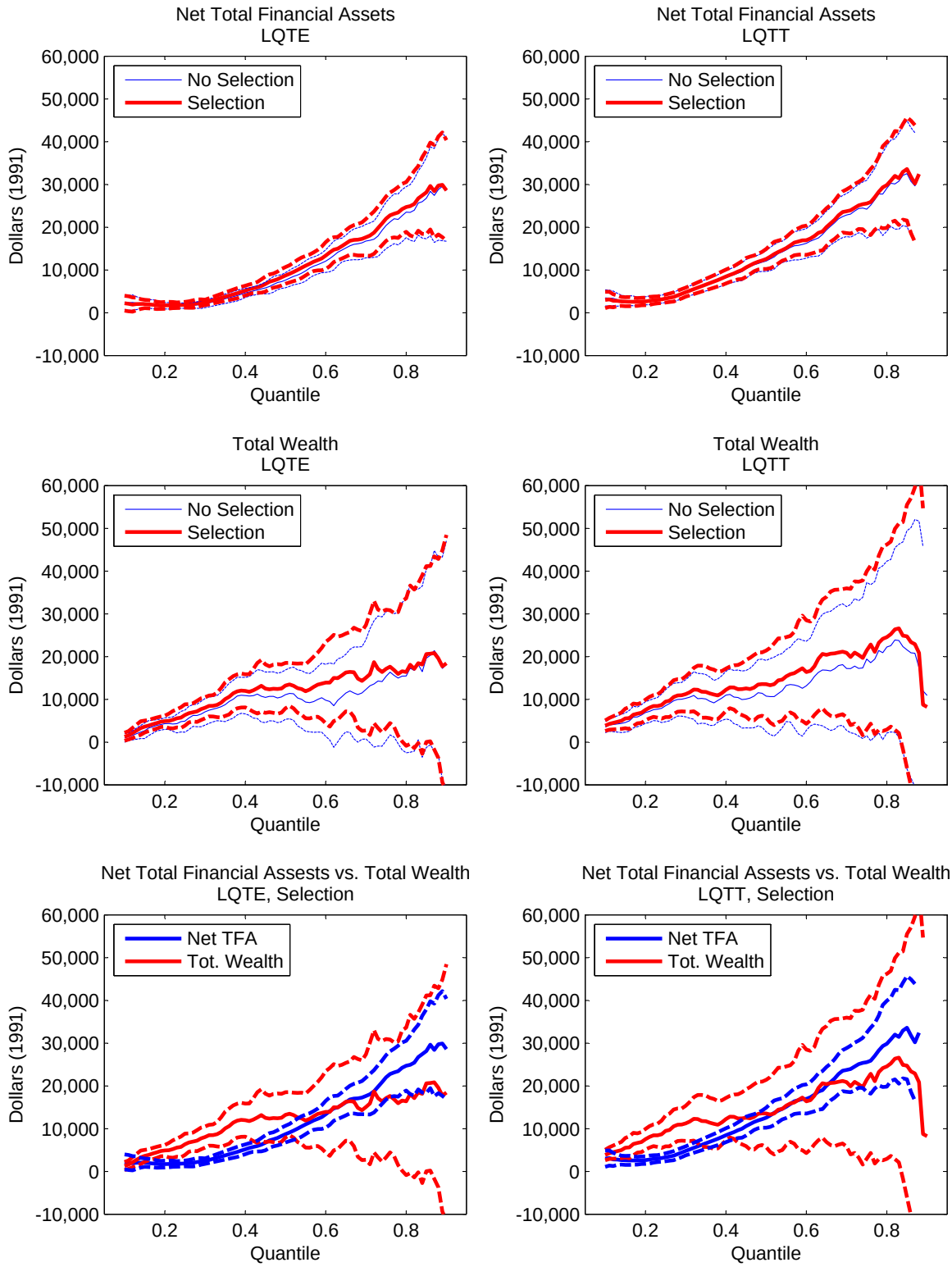

FIGURE 1. LQTE and LQTE-T estimates based on low-p indicators specification.

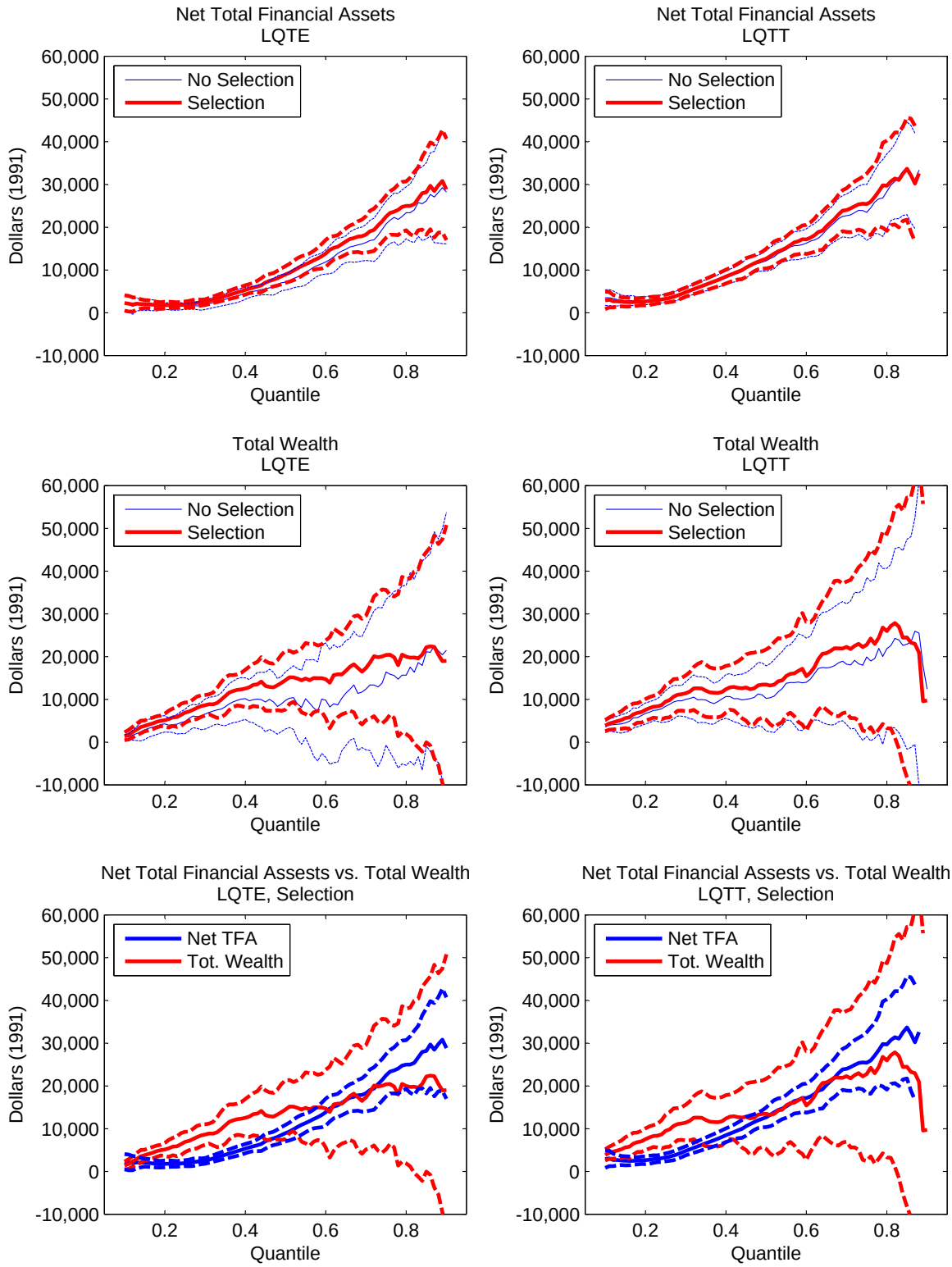

FIGURE 2. LQTE and LQTE-T estimates based on high-p indicators plus interactions specification.

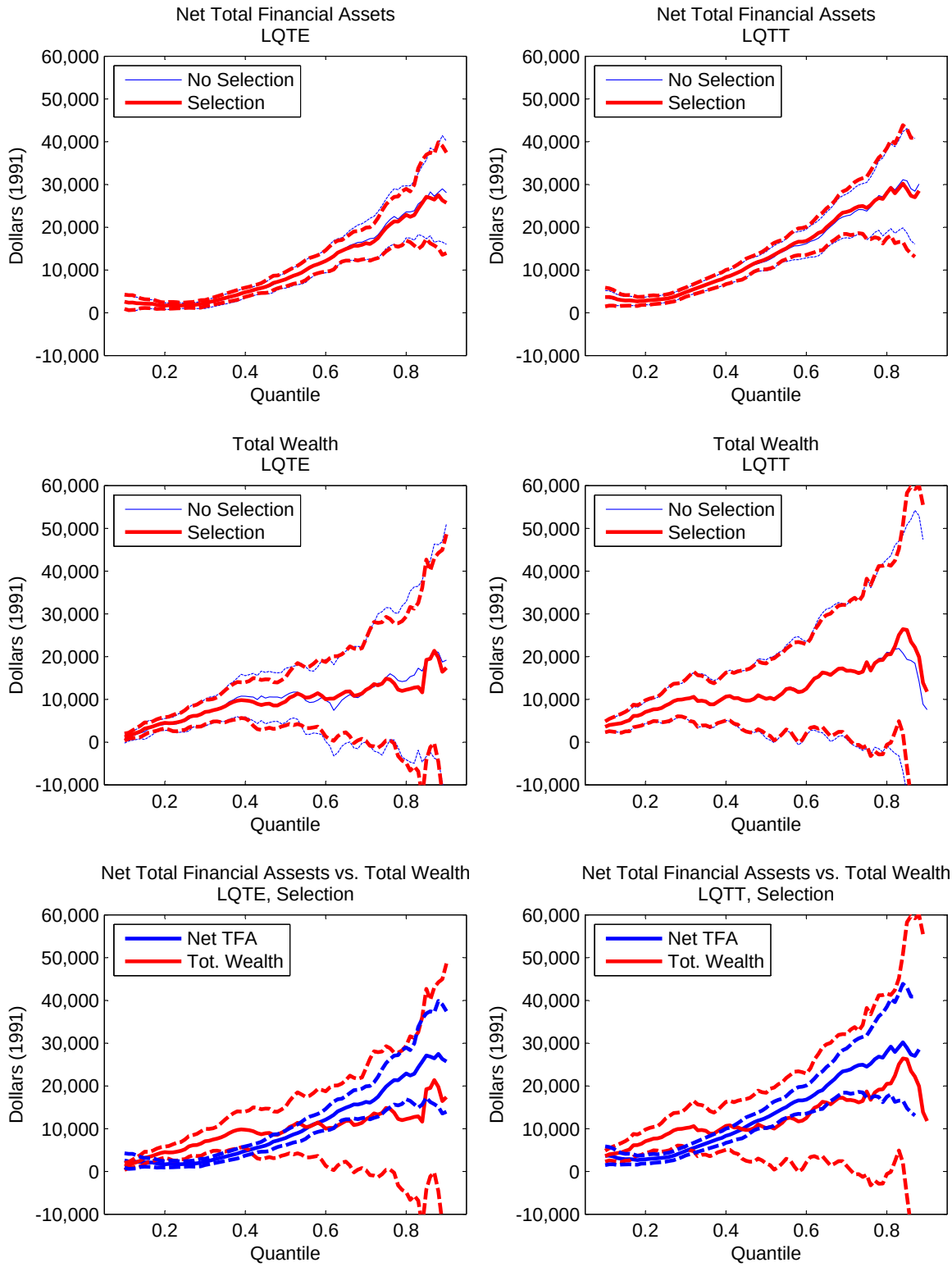

FIGURE 3. LQTE and LQTE-T estimates based on low-p orthogonal polynomial specification.

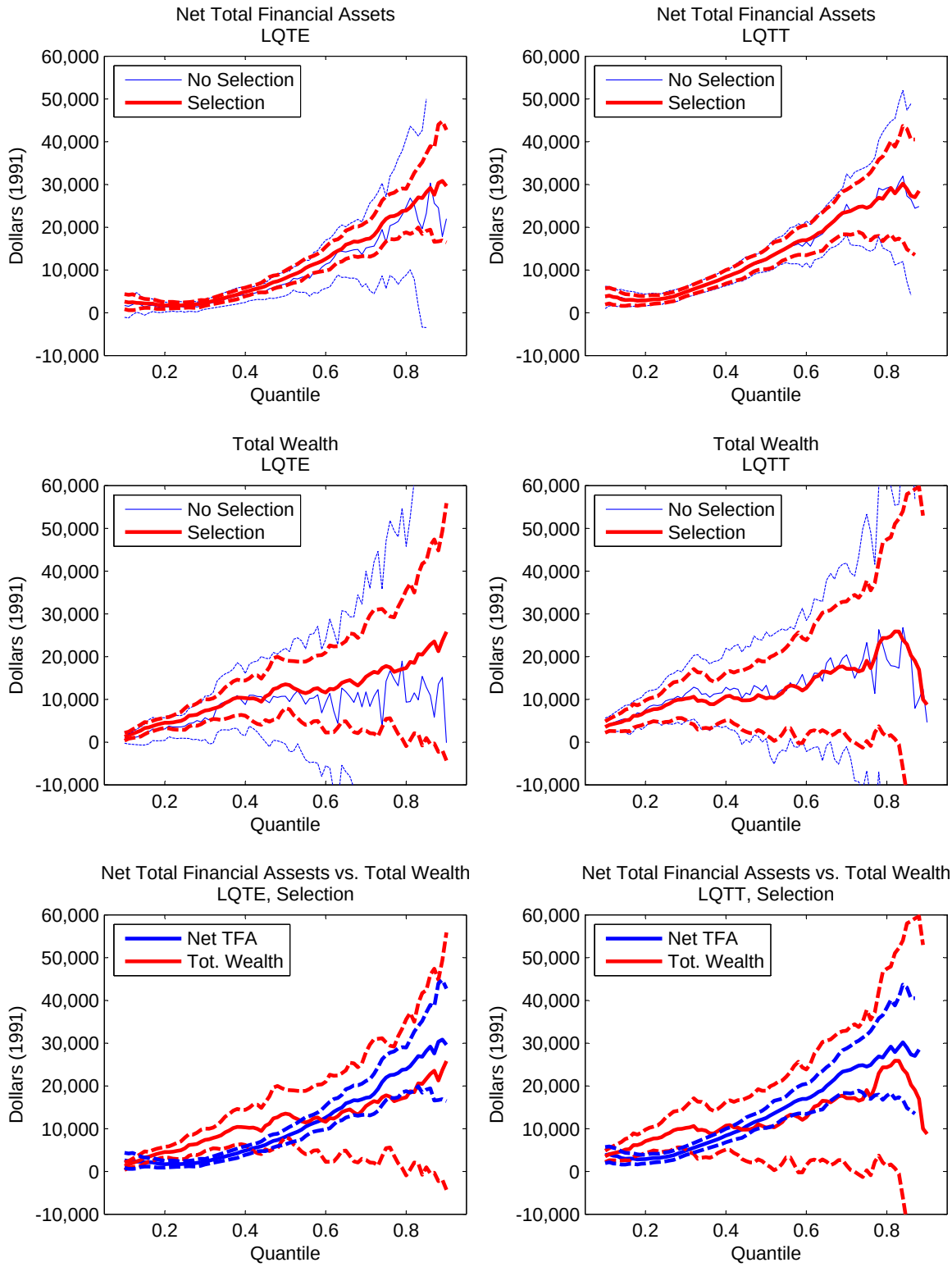

FIGURE 4. LQTE and LQTE-T estimates based on high-p orthogonal polynomial plus interactions specification.

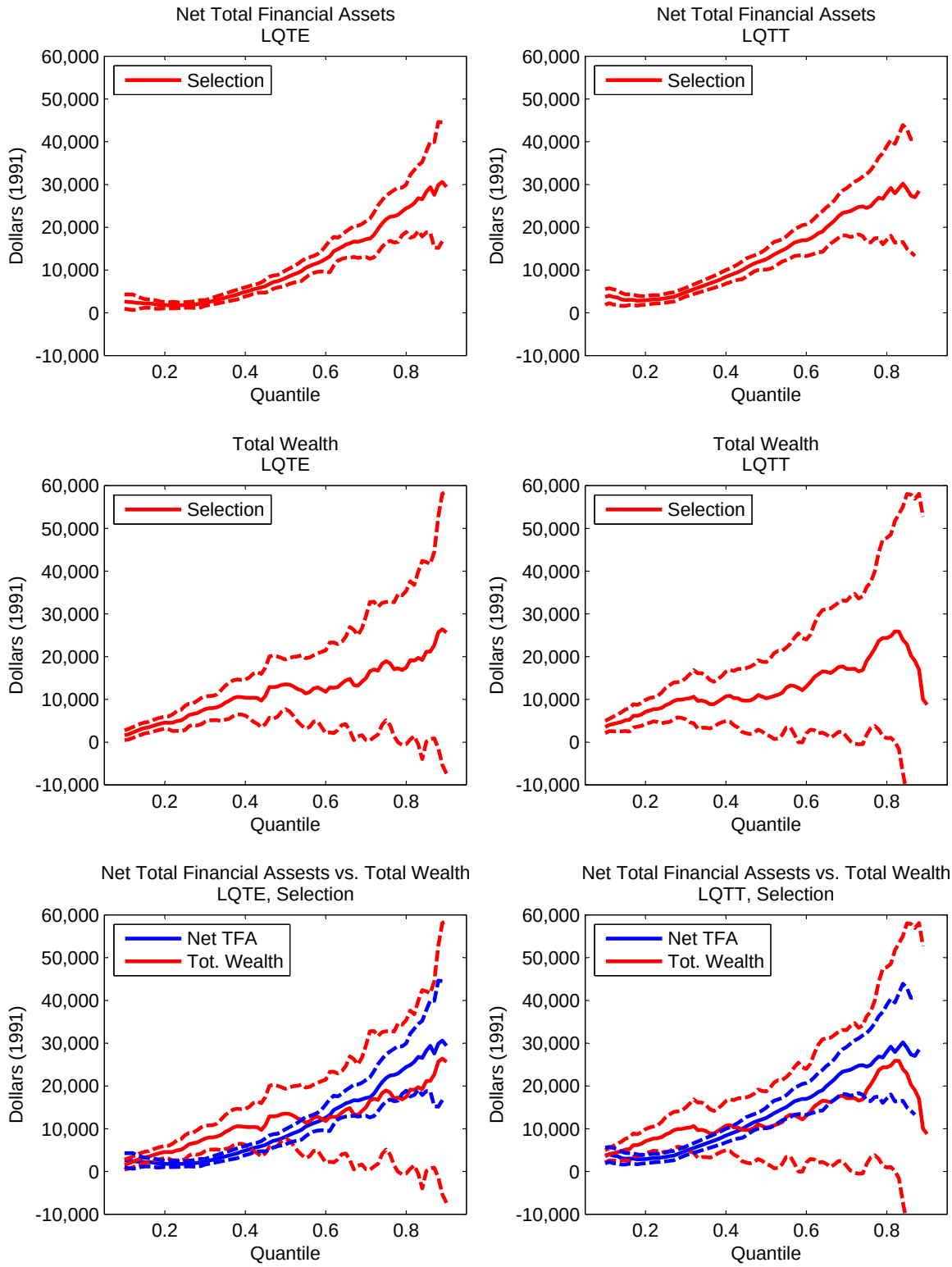

FIGURE 5. LQTE and LQTE-T estimates based on very-high-p orthogonal polynomial plus many interactions specification.