Doubly-Robust Inference for Conditional Average Treatment Effects with High-Dimensional Controls*

Adam Baybutt and Manu Navjeevan†

University of California, Los Angeles

Revised September 3, 2023

Abstract

Plausible identification of conditional average treatment effects (CATEs) can rely on controlling for a large number of variables to account for confounding factors. In these high-dimensional settings, estimation of the CATE requires estimating first-stage models whose consistency relies on correctly specifying their parametric forms. While doubly-robust estimators of the CATE exist, inference procedures based on the second-stage CATE estimator are not doubly-robust. Using the popular augmented inverse propensity weighting signal, we propose an estimator for the CATE whose resulting Wald-type confidence intervals are doubly-robust. We assume a logistic model for the propensity score and a linear model for the outcome regression, and estimate the parameters of these models using an ℓ_1 (Lasso) penalty to address the high-dimensional covariates. Inference based on this estimator remains valid even if one of the logistic propensity score or linear outcome regression models are misspecified. To our knowledge, we are the first paper to develop doubly-robust pointwise and uniform inference on an infinite dimensional target parameter after high dimensional nuisance model estimation.

Keywords: High-Dimensional, Doubly-Robust Inference, Nonparametric

JEL Codes: C01, C12, C14

*We are grateful to Denis Chetverikov, Andres Santos, Zhipeng Liao, Jinyong Hahn, Rosa Matzkin, Shuyang Sheng, Daniel Ober-Reynolds and participants in UCLA's Econometrics Proseminar for helpful comments.

†Corresponding author: mnavjeevan@g.ucla.edu.

1 Introduction

Consider a potential outcomes framework (Rubin, 1974, 1978) where an observed outcome $Y \in \mathbb{R}$ and treatment $D \in \{0,1\}$ are related to two latent potential outcomes $Y_1, Y_0 \in \mathbb{R}$ via $Y = DY_1 + (1-D)Y_0$. To account for unobserved confounding factors a common strategy is to assume the researcher has access to a vector of covariates, $Z = (Z_1', X')' \in \mathcal{Z}_1 \times \mathcal{X} \subseteq \mathbb{R}^{d_z - d_x} \times \mathbb{R}^{d_x}$, such that the potential outcomes are independent of the treatment decision after conditioning on the observed covariates, $(Y_1, Y_0) \perp D|Z$. In this setting, we are interested in estimation of and inference on the conditional average treatment effect (CATE):

$$\mathbb{E}[Y_1 - Y_0 \mid X = x]. \tag{1.1}$$

Estimation of the CATE generally requires first fitting propensity score and/or outcome regression models. When the number of control variables Z is large ($d_z \gg n$), these first-stage models must be estimated using regularized methods which converge slower than the nonparametric rate and typically rely on the correctness of parametric specifications for consistency.[1]

Fortunately, if both models are correctly specified, one can obtain a nonparametric-rate consistent estimator and valid inference procedure for the CATE by using the popular augmented inverse propensity weighted (aIPW) signal (Semenova and Chernozhukov, 2021; Fan et al., 2022). This is because the aIPW signal obeys an orthogonality condition at, crucially, the true nuisance model values that limits the first-stage estimation error passed on to the second-stage estimator. Moreover, estimators based on the aIPW signal are doubly-robust; consistency of the resulting second-stage estimators requires correct specification of only one of the first-stage propensity score or outcome regression models. However inference based on these estimators is not doubly-robust. The orthogonality of the aIPW signal fails under misspecification and the resulting testing procedures and confidence intervals are rendered invalid.

This paper proposes a doubly-robust estimator and inference procedure for the conditional average treatment effect when the number of control variables, d_z, is potentially much larger than the sample size, n. The dimensionality of the conditioning variable, d_x, remains fixed in our analysis. Our approach is based on Tan (2020) wherein doubly-robust inference is developed for the average treatment effect. We take a series approach to estimating the CATE, using a quasi-projection of the aIPW signal onto a growing set of basis functions. By assuming a logistic form for the propensity score model and a linear form for the outcome regression model, we construct novel ℓ_1-regularized first-stage estimating equations to recover a partial orthogonality of the aIPW signal at the limiting values of the first-stage estimators. So long as the limiting values of the first stage estimators have sparse representations this restricted orthogonality is enough to achieve doubly-robust pointwise and uniform inference; pointwise and uniform confidence intervals centered at the second-stage estimator are valid even if one of the logistic or linear functional forms is misspecified.

To achieve this restricted orthogonality at all points in the support of the conditioning variable, we employ distinct first-stage estimating equations for each basis term used in the second-stage series approximation. This results in the number of first-stage estimators growing with the number of basis terms. These estimators converge uniformly to limiting values under standard conditions in high-dimensional analysis. Improving on prior work in doubly-robust inference, our ℓ_1 regularized first-stage estimation incorporates a data-dependent penalty parameter based on the work of Chetverikov and Sørensen (2021). This allows practical implementation of our proposed estimation procedure with minimal knowledge of the underlying data generating process.

[1] Recent works by Bauer and Kohler (2019); Schmidt-Hieber (2020) provide some limited nonparametric results in high-dimensional settings using deep neural networks.

The use of multiple pairs of nuisance parameter estimates leaves us with multiple limiting values for the aIPW signals. So long as one of the nuisance models is correctly specified these limiting values share a conditional mean function. However, the various limiting values may all have different error terms describing their deviations from the conditional mean. This limits our ability to straightforwardly apply existing nonparametric results for series estimators (Newey, 1997; Belloni et al., 2015). Under modified conditions, we analyze the asymptotic properties of our second-stage series estimator to re-derive pointwise and uniform inference results. These modified conditions are in general slightly stronger than those of Belloni et al. (2015), though in certain special cases collapse exactly to the conditions of Belloni et al. (2015).

PRIOR LITERATURE. Chernozhukov et al. (2018) analyze the general problem of estimating finite dimensional target parameters in the presence of potentially high-dimensional nuisance functions. Using score functions that are Neyman-orthogonal with respect to nuisance parameters they show that it is possible to obtain target parameter estimates that are $\sqrt{n}$-consistent and asymptotically normal so long as the nuisance parameters are consistent at rate $n^{-1/4}$, a condition satisfied by many machine learning-based estimators. Semenova and Chernozhukov (2021) take advantage of new results for series estimation in Belloni et al. (2015) and consider series estimation of functional target parameters after high-dimensional nuisance estimation.[2] The inference results of these papers are highly dependent on the orthogonality of their second stage estimators to first stage estimation error, making it difficult to directly extend these analyses when the first stage estimators are not consistent and the orthogonality cannot be applied.

In the same setting as this paper, Tan (2020); Bradic et al. (2019) consider estimation of the average treatment effect. After assuming a logistic form for the propensity score and a linear form for the outcome regression, both papers propose ℓ_1-regularized first-stage estimators that allow for partial control of the derivative of the aIPW signal away from true nuisance values and thus allow for doubly-robust inference. Bradic et al. (2019) differs from Tan (2020) in their use of sample splitting, which allows them to achieve a "sparsity double robust" estimate of the ATE; so long as one nuisance model is sufficiently sparse the other may be more dense. Smucler et al. (2019) extends the analysis of Tan (2020) to consider doubly-robust inference for a larger class of finite dimensional target parameters with bilinear influence functions. Wu et al. (2021) provide doubly-robust inference procedures for covariate-specific treatment effects with discrete conditioning variables; their results depend on exact representation assumptions that are unlikely to hold with continuous covariates. Moreover, no uniform inference procedures are described.

These papers pioneered the approach that we will employ below, which is to directly use the first order conditions of the first stage estimators to control second stage estimation error. However, it is not a priori clear how to extend this approach to control the estimation error passed onto an infinite dimensional target parameter like the CATE. As discussed above, our analysis requires re-deriving pointwise and uniform inference results for nonparametric series estimators under modified conditions. We do not consider the sample splitting approach of Bradic et al. (2019), which may allow for relaxed sparsity conditions on our nuisance parameter estimates, but consider this an interesting future extension.

Chetverikov and Sørensen (2021) propose a data-driven "bootstrap after cross-validation" approach to penalty parameter selection that is modified for and implemented in our setting. This work is related to other work on the lasso (Tibshirani, 1996; Bickel et al., 2009; Belloni and Chernozhukov, 2013; Chetverikov et al., 2021) and ℓ_1-regularized M-estimation in high-dimensional settings (van der Greer, 2016; Tan, 2017).

[2] Fan et al. (2022) provides a similar analysis using a second-stage kernel estimator.

PAPER STRUCTURE. This paper proceeds as follows. Section 2 defines the problem and introduces our methods for estimation and inference. Section 3 provides intuition for how the first-stage estimation procedure allows for doubly-robust estimation and inference on the CATE as well as formally establishes the necessary first-stage convergence. Section 4 presents the main results: valid pointwise and uniform inference for the second-stage series estimator if either the first-stage logistic propensity score model or linear outcome regression model is correctly specified. Section 5 ties up a technical detail. Section 6 applies our proposed estimator to examine the effect of maternal smoking on infant birth weight while Section 7 provides evidence from simulation study. Section 8 concludes. Proofs of main results are deferred to Appendix A.

NOTATION. For any measure F and any function f, define the L^2 norm, $\|f\|_{F,2} = (\mathbb{E}_F[f^2])^{1/2}$ and the L^∞ norm $\|f\|_{F,\infty} = \operatorname{ess\,sup}_F |f|$. For any vector in $\mathbb{R}^p$ let $\|\cdot\|_p$ for $p \in [1,\infty]$ denote the ℓ_p norm, $\|a\|_p = (\sum_{l=1}^p a_l^p)^{1/p}$ and $\|a\|_\infty = \max_{1\le l\le\infty} |a_l|$. If the subscript is unspecified, we are using the ℓ_2 norm. For two vectors $a, b \in \mathbb{R}^p$, let $a \circ b = (a_i b_i)_{i=1}^p$ denote the Hadamard (element-wise) product. We adopt the convention that for $a \in \mathbb{R}^p$ and $c \in \mathbb{R}$, $a + c = (a_i + c)_{i=1}^p$. For a matrix $A \in \mathbb{R}^{m\times n}$ let $\|A\| = \max_{\|v\|_{\ell_2}\le 1} \|Av\|_{\ell_2}$ denote the operator norm and $\|A\|_\infty = \sup_{1\le r\le m, 1\le s\le n} |A_{rs}|$. For any real valued function f let $\mathbb{E}_n[f(X)] = \frac{1}{n}\sum_{i=1}^n f(X_i)$ denote the empirical expectation and $\mathbb{G}_n[f(X)] = \frac{1}{\sqrt{n}}\sum_{i=1}^n (f(X_i) - \mathbb{E}[X_i])$ denote the empirical process. For two sequences of random variables $\{a_n\}_\mathbb{N}$ and $\{b_n\}_\mathbb{N}$, we say $a_n \lesssim_P b_n$ or $a_n = O_p(b_n)$ if a_n/b_n is bounded in probability and say $a_n = o_p(b_n)$ if $a_n/b_n \to_p 0$.

2 SETUP

In this section, we formally define the setting and identification strategy that we consider. We then introduce our doubly-robust estimator and inference procedure. The parameter of interest is the conditional average treatment effect: $\mathbb{E}[Y_1 - Y_0 \mid X = x]$. However, for this paper we largely focus on estimation and inference for the conditional average counterfactual outcome:

$$g_0(x) \coloneqq \mathbb{E}[Y_1 \mid X = x]. \tag{2.1}$$

Doubly-robust estimation and inference on $\mathbb{E}[Y_0 \mid X = x]$ follows a similar procedure and is described in Section 5. The procedures can be combined for doubly-robust estimation and inference for the CATE.

2.1 SETTING

We assume the researcher observes i.i.d data and conditioning on Z is sufficient to control for all confounding factors affecting both the treatment decision D and the potential outcomes, Y_1 and Y_0. Our analysis allows the dimensionality of the controls, $Z = (Z_1, X)$, to grow much faster than the sample size ($d_z \gg n$), while assuming the dimensionality of the conditioning variables, X, remains fixed ($d_x \ll n$).

Assumption 2.1 (Identification)**.**

(i) $\{Y_i, D_i, Z_i\}_{i=1}^n$ *are independent and identically distributed.*

(ii) $(Y_1, Y_0) \perp D \mid Z$.

(iii) There exists a value $\eta \in (0,1)$ *such that* $\eta < \mathbb{E}[D \mid Z = z] < 1 - \eta$ *almost surely in* Z.

To obtain doubly-robust estimation and inference we use the augmented inverse propensity weighted (aIPW) signal,

$$Y(\pi, m) = \frac{DY}{\pi(Z)} - \left(\frac{D}{\pi(Z)} - 1\right) m(Z), \tag{2.2}$$

which is a function of a fitted propensity score model, $\pi(Z)$, and a fitted outcome regression model, $m(Z)$, whose true values are given $\pi^\star(Z) := \mathbb{E}[D \mid Z]$ and $m^\star(Z) := \mathbb{E}[Y \mid D = 1, Z]$. Under Assumption 2.1, the aIPW signal $Y(\cdot, \cdot)$ provides doubly-robust identification of $g_0(x)$. That is, for integrable $\pi \neq \pi^\star$ and $m \neq m^\star$,

$$\begin{aligned} \mathbb{E}[Y_1 \mid X = x] &= \mathbb{E}[Y(\pi^\star, m^\star) \mid X = x] \\ &= \mathbb{E}[Y(\pi\,, m^\star) \mid X = x] \\ &= \mathbb{E}[Y(\pi^\star, m\,) \mid X = x]. \end{aligned} \tag{2.3}$$

We use a series approach to estimate $g_0(x)$, taking a quasi-projection of the aIPW signal onto a growing set of k weakly positive basis terms:

$$p^k(x) := \big(p_1(x), \ldots, p_k(x)\big)' \in \mathbb{R}^k_+. \tag{2.4}$$

The basis terms are required to be weakly positive as they are used as weights within the convex first-stage estimators estimating equations.[1]Examples of weakly positive basis functions are B-splines or shifted polynomial series terms. To ensure that the basis terms are well behaved, we assume regularity conditions on $\xi_{k,\infty} := \sup_{x \in \mathcal{X}} \|p^k(x)\|_\infty$, $\xi_{k,2} := \sup_{x \in \mathcal{X}} \|p^k(x)\|_2$, and the eigenvalues of the design matrix $Q := \mathbb{E}[p^k(x)p^k(x)']$.

For each basis term $p_j(x)$, $j = 1, \ldots, k$, we estimate a separate propensity score model, $\widehat{\pi}_j(Z)$, and outcome regression model, $\widehat{m}_j(Z)$. Under standard moment and sparsity conditions, these converge uniformly over $j = 1, \ldots, k$ to limiting values $\bar{\pi}_j(Z)$ and $\bar{m}_j(Z)$. If the propensity score model and outcome regression models are correctly specified these limiting values coincide with the true values $\pi^\star(Z)$ and $m^\star(Z)$. However, in general the limiting and true values may differ. The double robustness of the aIPW signal allows for identification of the CATE even if only one of the nuisance models is correctly specified. If either $\bar{\pi}_j = \pi^\star$ or $\bar{m}_j = m^\star$, we can write for all $j = 1, \ldots, k$:

$$\begin{aligned} Y(\bar{\pi}_j, \bar{m}_j) &= g_0(x) + \epsilon_j, \qquad\qquad \mathbb{E}[\epsilon_j \mid X] = 0 \\ &= g_k(x) + r_k(x) + \epsilon_j \end{aligned} \tag{2.5}$$

where $g_0(x)$ is the conditional counterfactual outcome (2.1), $g_k(x) := p^k(x)'\beta^k$ is the projection of $g_0(x)$ onto the first k basis terms, and $r_k(x) := g_0(x) - g_k(x)$ denotes the approximation error from this projection. Note the separate error terms for each $j = 1, \ldots, k$ in (2.5), which are collected together in the vector $\epsilon^k := (\epsilon_1, \ldots, \epsilon_k)$. As long as one of the first-stage models is correctly specified, the least squares parameter β^k governing the projection in $g_k(x)$ can be identified by the projection of the aIPW signal onto the basis terms $p^k(x)$:

$$\begin{aligned} \beta^k &:= Q^{-1}\mathbb{E}[p^k(X)Y_1] \\ &= Q^{-1}\mathbb{E}[p^k(X)Y(\pi^\star, m^\star)] \\ &= Q^{-1}\mathbb{E}[p^k(X)Y(\bar{\pi}_j, \bar{m}_j)], \quad \forall j = 1, \ldots, k. \end{aligned} \tag{2.6}$$

[1]In case the researcher wants to use a second-stage basis that cannot be transformed to be weakly positive, we have shown a slightly modified method of constructing our doubly-robust estimator and inference procedure that does not require the first-stage weights to directly be the second-stage basis terms. This is available on request.

2.2 ESTIMATOR AND INFERENCE PROCEDURE

We assume a logistic regression form for the propensity score model and a linear form for the outcome regression model:

$$\begin{aligned} \pi(Z;\gamma) &= \left(1+\exp(-\gamma' Z)\right)^{-1}, \\ m(Z;\alpha) &= \alpha' Z. \end{aligned} \tag{2.7}$$

For each $j = 1, \ldots, k$, the parameters of (2.7), $\gamma, \alpha \in \mathbb{R}^{d_z}$, are estimated, respectively, by

$$\widehat{\gamma}_j := \arg\min_{\gamma} \mathbb{E}_n[p_j(X)\{De^{-\gamma' Z} + (1-D)\gamma' Z\}] + \lambda_{\gamma,j}\|\gamma\|_1, \tag{2.8}$$

$$\widehat{\alpha}_j := \arg\min_{\alpha} \mathbb{E}_n[p_j(X)De^{-\widehat{\gamma}_j' Z}(Y-\alpha' Z)^2]/2 + \lambda_{\alpha,j}\|\alpha\|_1. \tag{2.9}$$

The penalty parameters $\lambda_{\gamma,j}$ and $\alpha_{\gamma,j}$ are chosen via a data dependent technique described below. Under standard assumptions the parameter estimators $\widehat{\gamma}_j, \widehat{\alpha}_j$ will converge uniformly over $j = 1, \ldots, k$ to population minimizers

$$\bar{\gamma}_j := \arg\min_{\gamma} \mathbb{E}[p_j(X)\{De^{-\gamma' Z} + (1-D)\gamma' Z\}], \tag{2.10}$$

$$\bar{\alpha}_j := \arg\min_{\alpha} \mathbb{E}[p_j(Z)De^{-\bar{\gamma}_j' Z}(Y-\alpha' Z)^2]. \tag{2.11}$$

which we assume are sufficiently sparse. Our first-stage estimators are then $\widehat{\pi}_j(Z) := \pi(Z;\widehat{\gamma}_j)$ and $\widehat{m}_j(Z) := m(Z;\widehat{\alpha}_j)$ with limiting values $\bar{\pi}_j(Z) := \pi(Z;\bar{\gamma}_j)$ and $\bar{m}_j(Z) := m(Z;\bar{\alpha}_j)$, respectively.

After plugging in the functional forms of $\bar{\pi}_j(Z)$ and $\bar{m}_j(Z)$ into the aIPW signal one can verify that the derivatives of the aIPW signal with respect to the parameters γ_j and α_j are almost identical to the first order conditions of the minimization problems in (2.10)-(2.11) . Optimality of $\bar{\gamma}_j$ and $\bar{\alpha}_j$ will thus imply that the gradient of the limiting aIPW signal, weighted by $p_j(X)$, is mean zero even when the limiting values $\bar{\pi}_j(Z)$ and $\bar{m}_j(Z)$ differ from the true values $\pi^\star(Z)$ and $m^\star(Z)$. This allows us to control how sensitive the second stage CATE estimator is to first stage nuisance model estimation error even under misspecification and achieve doubly robust inference. The importance of this fact and why it is useful is discussed at greater depth in Section 3.

Our second-stage estimator is defined $\widehat{g}(x) := p^k(x)'\widehat{\beta}^k$ where $\widehat{\beta}^k$ is an estimate of the population projection parameter, β^k, obtained by combining all k pairs of first-stage estimators

$$\widehat{\beta}^k := \widehat{Q}^{-1}\mathbb{E}_n \begin{bmatrix} p_1(X)Y(\widehat{\pi}_1, \widehat{m}_1) \\ \vdots \\ p_k(X)Y(\widehat{\pi}_k, \widehat{m}_k) \end{bmatrix}, \tag{2.12}$$

and $\widehat{Q} := \mathbb{E}_n[p^k(X)p^k(X)']$. We estimate the variance of $\widehat{g}(x)$ using $\widehat{\sigma}(x) := \|\widehat{\Omega}^{1/2}p^k(x)\|/\sqrt{n}$ where

$$\widehat{\Omega} := \widehat{Q}^{-1}\mathbb{E}_n[\{p^k(X) \circ \widehat{\epsilon}^k\}\{p^k(X) \circ \widehat{\epsilon}^k\}']\widehat{Q}^{-1}, \tag{2.13}$$

and $\circ$ represents the Hadamard element-wise product. The vector $\widehat{\epsilon}^k$ collects the various estimated error terms; $\widehat{\epsilon}^k := (\widehat{\epsilon}_1, \ldots, \widehat{\epsilon}_k)$ for $\widehat{\epsilon}_j := Y(\widehat{\pi}_j, \widehat{m}_j) - \widehat{g}(x)$, $j = 1, ..., k$. Inference is based on the $100(1-\eta)\%$ confidence bands

$$\left[\underline{i}(x), \bar{i}(x)\right] := \left[\widehat{g}(x) - c^\star\left(1-\eta/2\right)\widehat{\sigma}(x),\ \widehat{g}(x) + c^\star\left(1-\eta/2\right)\widehat{\sigma}(x)\right]. \tag{2.14}$$

For pointwise inference, the critical value $c^\star(1-\eta/2)$ is taken as the $(1-\eta/2)$ quantile of a standard normal distribution. For uniform inference $c^\star(1-\eta/2)$ is taken

$$c_u^\star(1-\eta/2) := (1-\eta/2)\text{-quantile of } \sup_{x\in\mathcal{X}}\left|\frac{p^k(x)\widehat{\Omega}^{1/2}}{\widehat{\sigma}(x)}N_k^b\right|$$

where N_k^b is a bootstrap draw from $N(0, I_k)$. Sections 3 and 4 show that, under standard sparsity and moment conditions, these pointwise and uniform inference procedures remain valid even under misspecification of either first-stage model.

2.3 PENALTY PARAMETER SELECTION

To select the penalty parameters $\lambda_{\gamma,j}$ and $\lambda_{\alpha,j}$ in (2.8)-(2.9) we propose a data driven two-step procedure based on the work of Chetverikov and Sørensen (2021). For each $j = 0, 1\ldots, k$, we start with pilot penalty parameters given by

$$\lambda_{\gamma,j}^{\text{pilot}} = c_{\gamma,j}\times\sqrt{\frac{\ln^3(d_z)}{n}} \text{ and } \lambda_{\alpha,j}^{\text{pilot}} = c_{\alpha,j}\times\sqrt{\frac{\ln^3(d_z)}{n}} \tag{2.15}$$

for some constants $c_{\gamma,j}, c_{\alpha,j}$ selected from the interval $[\underline{c}_n, \bar{c}_n]$ with $\underline{c}_n > 0$. In practice, the researcher has a fair bit of flexibility in choosing these constants. The optimal choice of these constants may depend on the underlying data generating process. We recommend using cross validation to pick these constants from a fixed-cardinality set of possible values. In line with Assumption 3.1(vi), the values in the set should be chosen to be on the order of the maximum value of $\|p^k(X_i)\|_\infty$ observed in the data.

Using $\lambda_{\gamma,j}^{\text{pilot}}$ and $\lambda_{\alpha,j}^{\text{pilot}}$ in lieu of $\lambda_{\gamma,j}$ and $\lambda_{\alpha,j}$ in (2.8)-(2.9) we generate pilot estimators $\widehat{\gamma}_j^{\text{pilot}}$ and $\widehat{\alpha}_j^{\text{pilot}}$. These pilot estimators are used to generate plug in estimators $\widehat{U}_{\gamma,j}$ and $\widehat{U}_{\alpha,j}$ of the residuals

$$\begin{aligned}\widehat{U}_{\gamma,j} &:= -p_j(X)\{D(1+e^{-\widehat{\gamma}_j^{\text{pilot}\prime}Z})-1\}\\ \widehat{U}_{\alpha,j} &:= -p_j(X)De^{-\widehat{\gamma}_j^{\text{pilot}\prime}Z}(Y-\widehat{\alpha}_j^{\text{pilot}\prime}Z).\end{aligned} \tag{2.16}$$

whose true values are given

$$\begin{aligned}U_{\gamma,j} &:= -p_j(X)\{D(1+e^{-\bar{\gamma}_j'Z})-1\}\\ U_{\alpha,j} &:= -p_j(X)De^{-\bar{\gamma}'Z}(Y-\bar{\alpha}'Z)\end{aligned} \tag{2.17}$$

These true residuals are the derivatives of the minimization problems in (2.10)-(2.11) evaluated at minimizing values $\bar{\gamma}_j$ and $\bar{\alpha}_j$. After generating the residual estimates, we use a multiplier bootstrap procedure to select final penalty parameters $\lambda_{\gamma,j}$ and $\lambda_{\alpha,j}$.

$$\begin{aligned}\lambda_{\gamma,j} &= c_0\times(1-\epsilon)\text{-quantile of } \max_{1\le l\le d_z}|\mathbb{E}_n[e_i\widehat{U}_{\gamma,j}Z_l]| \text{ given } \{Y_i, D_i, Z_i\}_{i=1}^n,\\ \lambda_{\alpha,j} &= c_0\times(1-\epsilon)\text{-quantile of } \max_{1\le l\le d_z}|\mathbb{E}_n[e_i\widehat{U}_{\alpha,j}Z_l]| \text{ given } \{Y_i, D_i, Z_i\}_{i=1}^n\end{aligned} \tag{2.18}$$

where $e_1,\ldots,e_n$ are independent standard normal random variables generated independently of the data $\{Y_i, D_i, X_i\}_{i=1}^n$ and $c_0 > 1$ is a fixed constant.[1] In line with other work we find $c_0 = 1.1$

[1]The constant c_0 can be different for the propensity score and outcome regression models and can also vary for each $j = 1,\ldots,k$. All that matters is that each constant satisfies the requirements of Lemma 3.1. This complicates notation, however.

works well in simulations. So long as our residual estimates converge in empirical mean square to limiting values and $k\epsilon \to 0$, the choice of penalty parameters in (2.18) will ensure that the penalty parameters dominate the noise with probability approaching one uniformly over the k first stage estimation procedures. This allows for consistent variable selection and coefficient estimation.

3 THEORY OVERVIEW

We begin with a main technical lemma which provides a bound on rate at which first-stage estimation error is passed on to the second-stage CATE and variance estimators. This bound is comparable to others seen in the inference after model-selection literature (Belloni et al., 2013; Tan, 2020) and is achieved under standard conditions in the ℓ_1-regularized estimation literature (Bickel et al., 2009; Bühlmann and van de Geer, 2011; Belloni and Chernozhukov, 2013; Chetverikov and Sørensen, 2021). However, this bound is achieved at the limiting values of the propensity score and outcome regression models which may differ from the true values $\pi^\star$ and $m^\star$ under misspecification.

The potential misspecification of the first-stage models means we cannot directly apply orthogonality of the aIPW signal, discussed below, to show that the effect of first-stage estimation error on the second-stage is negligible. Instead, we use the first order conditions for $\widehat{\gamma}_j$ and $\widehat{\alpha}_j$ to directly control this quantity. After presenting the lemma Section 3.2 provides some intuition for how this is done. Controlling the rate at which first-stage estimation error is passed on to the second-stage estimator even at points away from the true values $\pi^\star$ and $m^\star$ is key for obtaining doubly-robust inference for the CATE.

3.1 UNIFORM FIRST-STAGE CONVERGENCE

To show uniform convergence of the first-stage estimators and thus uniform control of the bias passed on from the first-stage estimation to the second-stage estimator we rely on Assumption 3.1, below. The conditions in Assumption 3.1(v,vi) depend on the sup-norm of the basis functions, $\xi_{k,\infty} = \sup_{x\in\mathcal{X}} \|p^k(x)\|_\infty$.

Assumption 3.1 (First-Stage Convergence)**.**

(i) The regressors Z are bounded, $\max_{1\le l\le d_z} |Z_l| \le C_0$ *almost surely.*

(ii) The errors $Y_1 - \bar{m}_j(Z)$ are uniformly subgaussian conditional on Z in the following sense. There exists fixed positive constants G_0 and G_1 such that for any j:

$$G_0\mathbb{E}\left[\exp\left(\{Y_1 - \bar{m}_j(Z)\}^2/G_0^2\right) - 1 \mid Z\right] \le G_1^2$$

almost surely.

(iii) There is a constant B_0 such that $\bar{\gamma}_j' Z \ge B_0$ almost surely for all j.

(iv) There exists fixed constants $\xi_0 > 1$ and $1 > v_0 > 0$ such that for each $j = 1, \dots, k$ the following empirical compatibility condition holds for the empirical hessian matrix $\tilde{\Sigma}_{\gamma,j} := \mathbb{E}_n[De^{-\bar{\gamma}_j' Z} ZZ']$. For any $b \in \mathbb{R}^{d_z}$ and $\mathcal{S}_j = \{l : |\bar{\gamma}_{j,l}| \vee |\bar{\alpha}_{j,l}| \ne 0\}$:

$$\sum_{l\notin\mathcal{S}_j} |b_l| \le \xi_0 \sum_{l\in\mathcal{S}_j} |b_l| \implies v_0^2\Big(\sum_{l\in\mathcal{S}_j} |b_l|\Big)^2 \le |\mathcal{S}_j|\left(b'\tilde{\Sigma}_{\gamma,j} b\right).$$

(v) There exists fixed constants c_u and $C_U > 0$ such that for all $j = 1, \dots, k$, $\mathbb{E}[U_{\gamma,j}^4] \le (\xi_{k,\infty} C_U)^4$

and $\min_{1\leq l\leq d_z} \mathbb{E}[U_{\gamma,j}^2 Z_l^2] \geq c_u$.

(vi) The constant $\underline{c}_n$ *is chosen such that* $\xi_{k.\infty} \lesssim \underline{c}_n$ *and the following sparsity bounds hold for* $s_k = \max_{1\leq j\leq k} |\mathcal{S}_j|$

$$\frac{\xi_{k,\infty} s_k^2 \bar{c}_n^2 \ln^5(d_z n)}{n} \to 0, \quad \textit{and} \quad \frac{\xi_{k,\infty}^4 \ln^7(d_z k n)}{n} \to 0.$$

The first part of Assumption 3.1 assumes that the regressors are bounded while the second assumes that tail behavior of the outcome regression errors are uniformly thin. Both of these can be relaxed somewhat with sufficient moment conditions on the tail behavior of the controls and errors. We should note that compactness of $\mathcal{X}$ is generally required by nonparametric estimators. The third part of the assumption bounds all limiting propensity scores $\bar{\pi}_j(Z)$ away from zero uniformly. The fourth assumption is an empirical compatibility condition on the weighted first-stage design matrix. It is slightly weaker than the restricted eigenvalue conditions often assumed in the literature (Bickel et al., 2009; Belloni et al., 2012). The penultimate condition is an identifiability constraint that limits the moments of the noise and bounds it away from zero uniformly over all estimation procedures. Many of the constants in Assumption 3.1 are assumed to be fixed across all j. This is mainly to simplify the exposition of the results below and in practice all constants can be allowed to grow slowly with k. However, the growth rate of these terms affects the required first-stage sparsity.

The last condition is required for the validity of the bootstrap penalty parameter selection procedure and is comparable to the requirements needed for the bootstrap after cross validation technique described by Chetverikov and Sørensen (2021). The main difference is the additional assumption on the growth rate of the basis functions, $\xi_{k,\infty}$ which is to ensure uniform stability of the estimation procedures (2.8)-(2.9) as well as some assumptions on the order of the constants $c_{\gamma,j}$ and $c_{\alpha,j}$ in (2.15).

Assumptions 3.1(v,vi) depend on the sup-norm of the basis functions, $\xi_{k,\infty}$. This growth rate of this quantity will depend on the form of basis used for the second stage nonparametric estimator. In both our simulation study as well as our empirical exercise we use B-splines for which $\xi_{k,\infty} \lesssim \sqrt{k}$. Other common bases used in nonparametric estimation are polynomial series for which $\xi_k \lesssim k$, or wavelets for which $\xi_{k,\infty} \lesssim \sqrt{k}$. Belloni et al. (2015) provide a discussion for other choices of basis terms.

Lemma 3.1 (First-Stage Convergence)**.** *Suppose that Assumption 3.1 holds. In addition assume that* $c_0 > (\xi_0 + 1)/(\xi_0 - 1)$, $k/n \to 0$, $k\epsilon \to 0$, *and there is a fixed constant* $c > 0$ *such that for all* j, $\lambda_{\alpha,j}/\lambda_{\gamma,j} \geq c$.[1] *Then the following weighted means converge uniformly in absolute value at least at rate:*

$$\max_{1\leq j\leq k} \left|\mathbb{E}_n[p_j(X)Y(\widehat{\pi}_j, \widehat{m}_j)] - \mathbb{E}_n[p_j(X)Y(\bar{\pi}_j, \bar{m}_j)]\right| \lesssim_P \frac{s_k\, \xi_{k,\infty}^2 \ln(d_z)}{n} \tag{3.1}$$

and in empirical mean square at least at rate:

$$\max_{1\leq j\leq k} \mathbb{E}_n[p_j^2(X)(Y(\widehat{\pi}_j, \widehat{m}_j) - Y(\bar{\pi}_j, \bar{m}_j))^2] \lesssim_P \frac{s_k^2\, \xi_{k,\infty}^4 \ln(d_z)}{n} \tag{3.2}$$

Lemma 3.1 provides a tight bound on the first-stage estimation error passed on to the second-stage estimator even when the first-stage estimators converge to values that are not the true propensity score or outcome regression. In particular under the sparsity bound $s_k \xi_{k,\infty}^2 k^{1/2} \ln^2(d_z)/\sqrt{n} \to 0$, any linear combination of the means in both (3.1) and (3.2) is $o_p(\sqrt{n})$. This allows us to obtain

doubly-robust inference for the CATE. This sparsity bound is similar in form to others in the literature (Belloni et al., 2012; van der Greer, 2016; Chetverikov and Sørensen, 2021) however is somewhat stronger due to the additional dependence on $\xi_{k,\infty}^2 k^{1/2}$.

3.2 MANAGING FIRST-STAGE BIAS

We now provide some intuition for how this result is obtained and the role our particular estimating equations play in establishing this fact. We focus on control of the vector $\mathbf{B}^k$, defined in (3.3), which measures the bias passed on from first-stage estimation to the second-stage estimate $\widehat{\beta}^k$. Limiting the size of $\mathbf{B}^k$ is crucial in showing convergence of $\widehat{\beta}^k$ to the true parameter β^k and thus consistency of the nonparametric estimator $\widehat{g}(x)$.

$$\mathbf{B}^k := \mathbb{E}_n \begin{bmatrix} p_1(X) \left\{ Y(\widehat{\pi}_1, \widehat{m}_1) - Y(\bar{\pi}_1, \bar{m}_1) \right\} \\ \vdots \\ p_k(X) \left\{ Y(\widehat{\pi}_k, \widehat{m}_k) - Y(\bar{\pi}_k, \bar{m}_k) \right\} \end{bmatrix}. \tag{3.3}$$

For exposition, we consider a single term of (3.3), $\mathbf{B}_j^k$, which roughly measures the first-stage estimation bias taken on from adding the j^{th} basis term to our series approximation of $g_0(x)$. The discussion that follows is a bit informal, instead of considering the derivatives with respect to the true parameters below our proof strategy will directly use the Kuhn-Tucker conditions of the optimization routines in (2.8)-(2.9). However, the general intuition is the same as is used in the proofs.

In addition to the doubly-robust identification property (2.3), the aIPW signal is typically useful in the high-dimensional setting because it obeys an orthogonality condition at the true values $(\pi^\star, m^\star)$:[1]

$$\mathbb{E}[\nabla_{\pi,m} Y(\pi^\star, m^\star) \mid Z] = 0. \tag{3.4}$$

When both the propensity score model and outcome regression model are correctly specified we can (loosely speaking) examine the bias $\mathbf{B}_j^k$ by replacing $\bar{\pi}_j = \pi^\star$ and $\bar{m}_j = m^*$ and considering the following first order expansion:

$$\begin{aligned} \mathbf{B}_j^k &= \mathbb{E}_n[p_j(X) Y(\widehat{\pi}_j, \widehat{m}_j)] - \mathbb{E}_n[p_j(X) Y(\pi^\star, m^\star)] \\ &= \underbrace{\mathbb{E}_n[p_j(X) \nabla_{\pi,m} Y(\pi^\star, m^\star)]}_{O_p(n^{-1/2}) \text{ by (3.4)}} \begin{bmatrix} \widehat{\pi}_j - \pi^\star \\ \widehat{m}_j - m^\star \end{bmatrix} + o_p(n^{-1/2}). \end{aligned} \tag{3.5}$$

By orthogonality of the aIPW signal the gradient term is close to zero, which guarantees that the bias is asymptotically negligible even if the nuisance parameters converge slowly to the true values, $\pi^\star$ and $m^\star$.[2] This allows the researcher to ignore first-stage nuisance parameter estimation error and treat $\pi^\star$ and $m^\star$ as known when analyzing the asymptotic properties of the second-stage series estimator. Indeed, since the aIPW signal orthogonality holds conditional on $Z = (Z_1, X)$, if both models are correctly specified only a single pair of first-stage estimators would be needed to provide control over all the elements in $\mathbf{B}^k$. This is the approach followed

[1] The requirement $\lambda_{\alpha,j}/\lambda_{\gamma,j} \geq c$ may seem a bit unnatural, but it can be enforced in practice without upsetting any assumptions by using the alternative linear penalty $\lambda_{\alpha,j}^{\text{ratio}} := \max\{\lambda_{\gamma,j} \geq c, \lambda_{\alpha,j}\}$. In simulations, we find this constraint is rarely binding. The constant c here is arbitrary, it is only important that the ratio $\lambda_{\gamma,j}/\lambda_{\alpha,j}$ is bounded from above.

[1] Robustness and orthogonality are indeed closely related, see Theorem 6.2 in Newey and McFadden (1994) for a discussion.

[2] Typically all that is required is that $\|\hat{\pi}_j - \pi^\star\| = o_p(n^{-1/4})$ and $\|\hat{m}_j - m^\star\| = o_p(n^{-1/4})$ in order to make the second order remainder term $\sqrt{n}$-negligible

by Semenova and Chernozhukov (2021).

So long as either one of $\bar{\pi}_j = \pi^\star$ or $\bar{m}_j = m^\star$ we still have that $\mathbb{E}[p_j(X)Y_1] \approx \mathbb{E}_n[p_j(X)Y(\bar{\pi}_j, \bar{m}_j)]$ by double-robustness of the aIPW signal (2.3). However, the aIPW orthogonality tells us nothing about the expectation of the gradient away from the true parameters, $\pi^\star, m^\star$; if either $\bar{\pi}_j \neq \pi^\star$ or $\bar{m}_j \neq m^\star$ there is no reason to believe that the gradient on the right hand side of (3.5) is mean zero when evaluated instead at $Y(\bar{\pi}_j, \bar{m}_j)$. In general, the bias $\mathbf{B}_j^k$ will then diminish at the rate of convergence of our nuisance parameters. Because we have high dimensional controls, this convergence rate will generally be much slower than the standard nonparametric rate (Newey, 1997; Belloni et al., 2015).

To get around this, we design the first-stage objective functions (2.8)-(2.9) such that the resulting first-order conditions control the bias passed on to the second-stage. Consider the following expansion instead around the limiting parameters $\bar{\gamma}_j$ and $\bar{\alpha}_k$.

$$\begin{aligned} \mathbf{B}_j^k &= \mathbb{E}_n[p_j(X)Y(\widehat{\pi}_j, \widehat{m}_j)] - \mathbb{E}_n[p_j(X)Y(\bar{\pi}_j, \bar{m}_j)] \\ &= \mathbb{E}_n[p_j(X)\nabla_{\gamma_j,\alpha_j} Y(\bar{\pi}_j, \bar{m}_j)] \begin{bmatrix} \widehat{\gamma}_j - \bar{\gamma}_j \\ \widehat{\alpha}_j - \bar{\alpha}_j \end{bmatrix} + o_p(n^{-1/2}) \end{aligned} \tag{3.6}$$

After substituting the forms of $\bar{\pi}_j(z) = \pi(z; \bar{\gamma}_j)$ and $\bar{m}_j(z) = m(z; \bar{\alpha}_j)$ described in (2.7) and differentiating with respect to γ_j and α_j we obtain

$$\mathbb{E}[p_j(X)\nabla_{\gamma_j,\alpha_j} Y(\bar{\pi}_j, \bar{m}_j)] = \mathbb{E}\begin{bmatrix} -p_j(X)De^{-\bar{\gamma}_j'Z}(Y - \bar{\alpha}_j'Z)Z \\ -p_j(x)\{D(1 + e^{-\bar{\gamma}_j'Z})Z - Z\} \end{bmatrix} \tag{3.7}$$

However, by definition $\bar{\gamma}_j$ and $\bar{\alpha}_j$ solve the minimization problems defined in (2.10)-(2.11), the population analogs of our finite sample estimating equations. The first order conditions of these minimization problems yield

$$\mathbb{E}\overbrace{\underbrace{\begin{bmatrix} -p_j(X)\{D(1 + e^{\bar{\gamma}'Z})Z - Z\} \\ -p_j(X)De^{-\bar{\gamma}'Z}(DY - \bar{\alpha}'Z)Z \end{bmatrix}}_{\text{First order condition of } \bar{\alpha}_j}}^{\text{First order condition of } \bar{\gamma}_j} = 0 \implies \mathbb{E}[p_j(X)\nabla_{\gamma_j,\alpha_j} Y(\bar{\pi}_j, \bar{m}_j)] = 0 \tag{3.8}$$

Examining the first order conditions in (3.8), we see that they exactly give us control over the gradient (3.7). Under suitable convergence of the first-stage parameter estimates, this guarantees the bias examined in expansion (3.6) is negligible even under misspecification of the propensity score or outcome regression models.

Control of this gradient under misspecification is not provided using other estimating equations, such as maximum likelihood for the logistic propensity score model or ordinary least squares for the linear outcome regression model. Moreover, control over the gradient of $\mathbf{B}_j^k$

from (3.3) is not provided by the first-order conditions for $\bar{\gamma}_l$ and $\bar{\alpha}_l$ for $l \neq j$:

$$\begin{aligned}\mathbb{E}[p_j(X)\nabla_{\gamma_j,\alpha_j} Y(\bar{\pi}_j, \bar{m}_j)] &= \mathbb{E}\overbrace{\begin{bmatrix} -p_j(X)De^{-\bar{\gamma}'Z}(Y - \bar{\alpha}'Z)Z \\ -p_j(X)\{D(1 + e^{\bar{\gamma}'Z})Z - Z\}\end{bmatrix}}^{\text{First order condition of } \bar{\gamma}_l} \\ &\neq \mathbb{E}\underbrace{\begin{bmatrix} -p_l(X)\{D(1 + e^{\bar{\gamma}'Z})Z - Z\} \\ -p_l(X)De^{-\bar{\gamma}'Z}(Y - \bar{\alpha}'Z)Z\end{bmatrix}}_{\text{First order condition of } \bar{\alpha}_l}.\end{aligned} \tag{3.9}$$

Showing that the inference procedure of Section 2 remains valid at all points $x \in \mathcal{X}$ under misspecification requires showing negligible first-stage estimation bias for any linear transformation of the vector (3.3). As outlined above, this requires using k separate pairs of nuisance parameter estimator to obtain k separate pairs of first order conditions, one for each term of the vector.

4 Main Results

In this section, we present the main consistency and distributional results for our second-stage estimator $\widehat{g}(x)$ described in Section 2. A full set of second-stage results, including pointwise and uniform linearization lemmas and uniform convergence rates, can be found in the Online Appendix. The first set of results is established under the following condition, which limits the bias passed from first-stage estimation onto the second-stage estimator. In particular, Condition 1 implies that the bias vector $\mathbf{B}^k$ from (3.3) satisfies $\|\mathbf{B}^k\| = o_p(n^{-1/2})$.

Condition 1 (No Effect of First-Stage Bias)**.**

$$\max_{1\leq j\leq k} \left|\mathbb{E}_n[p_j(X)Y(\widehat{\pi}_j, \widehat{m}_j)] - \mathbb{E}_n[p_j(X)Y(\bar{\pi}_j, \bar{m}_j)]\right| = o_p(n^{-1/2}k^{-1/2}). \tag{4.1}$$

Via Lemma 3.1 we can see that is a logistic propensity score model and a linear outcome regression model and estimating the first-stage models using the estimating equations (2.8)-(2.9), Condition 1 can be achieved under Assumption 3.1 and the sparsity bound

$$\frac{s_k\, \xi_{k,\infty}^2 k^{1/2} \ln(d_z)}{\sqrt{n}} \to 0. \tag{4.2}$$

If the researcher were to assume different parametric forms for the first-stage model, different first estimating equations would have to be used to obtain doubly-robust estimation and inference. However, so long as the Condition 1 can be established at the limiting values of the first-stage models, the results of this section hold.

Having dealt with the first-stage estimation error, the main complication remaining is that under misspecification the aIPW signals $Y(\hat{\pi}_j, \hat{m}_j)$ for $j = 1, \ldots, k$ do not all converge to the same limiting values. However, so long as at least one of the first-stage models is correctly specified, all of the limiting aIPW signals have the same conditional mean, $g_0(x)$. In the standard setting, consistency of nonparametric estimator relies on certain conditions on the error terms. In our setting, we require that these assumptions hold uniformly over k the error terms. We note though that there is a non-trivial dependence structure between that limiting aIPW signals. This strong dependence gives plausibility to our uniform conditions. For example, if the logistic propensity score model is correctly specified and the difference between

the limiting outcome regression models is bounded, $|\max_{1\leq j\leq k} \bar{m}_j(Z) - \min_{1\leq j\leq k} m_j(Z)| \leq C$ almost surely, our conditions reduce exactly to the conditions of Belloni et al. (2015). In general, however, the uniform conditions suggest that a degree of undersmoothing is optimal when implementing our estimation procedure; the optimal choice of k may be smaller than in standard nonparametric regression.

4.1 Pointwise Inference

Pointwise inference relies on the following assumption in tandem with Condition 1.

Assumption 4.1 (Second-Stage Pointwise Assumption). *Let $\bar{\epsilon}_k := \max_{1\leq j\leq k} |\epsilon_j|$. Assume that*

(i) Uniformly over all n, the eigenvalues of $Q = \mathbb{E}[p^k(x)p^k(x)']$ are bounded from above and away from zero.

(ii) The conditional variance of the error terms is uniformly bounded in the following sense. There exists constants $\underline{\sigma}^2$ and $\bar{\sigma}^2$ such that for any $j = 1, 2 \ldots$ we have that $\underline{\sigma}^2 \leq \mathrm{Var}(\epsilon_j \mid X) \leq \bar{\sigma}^2 < \infty$;

(iii) For each n and k there are finite constants c_k and ℓ_k such that for each $f \in \mathcal{G}$

$$\|r_k\|_{L,2} = (\mathbb{E}[r_k(x)^2])^{1/2} \leq c_k \;\; \text{and} \;\; \|r_k\|_{L,\infty} = \sup_{x\in\mathcal{X}} |r_k(x)| \leq \ell_k c_k.$$

(iv) $\sup_{x\in\mathcal{X}} \mathbb{E}[\bar{\epsilon}_k^2 \mathbf{1}\{\bar{\epsilon}_k + \ell_k c_k > \delta\sqrt{n}/\xi_k\} \mid X = x] \to 0$ as $n \to \infty$ and $\sup_{x\in\mathcal{X}} \mathbb{E}[\ell_k^2 c_k^2 \mathbf{1}\{\bar{\epsilon}_k + \ell_k c_k > \delta\sqrt{n}/\xi_k\} \mid X = x] \to 0$ as $n \to \infty$ for any $\delta > 0$.

As mentioned, these are exactly the conditions required by Belloni et al. (2015), with the modification that the bounds on conditional variance and other moment conditions on the error term hold uniformly over $j = 1, \ldots, k$. The assumptions on the series terms being used in the approximation can be shown to be satisfied by a number of commonly used functional bases, such as polynomial bases or splines, under adequate normalizations and smoothness of the underlying regression function. Readers should refer to Newey (1997), Chen (2007), or Belloni et al. (2015) for a more in depth discussion of these assumptions.[1]

Under these assumptions, the variance of our second-stage estimator is governed by one of the following variance matrices:

$$\begin{aligned} \tilde{\Omega} &:= Q^{-1}\mathbb{E}[\{p^k(x) \circ (\epsilon^k + r_k)\}\{p^k(x) \circ (\epsilon^k + r_k)\}']Q^{-1} \\ \Omega_0 &:= Q^{-1}\mathbb{E}[\{p^k(x) \circ \epsilon^k\}\{p^k(x) \circ \epsilon^k\}']Q^{-1} \end{aligned} \tag{4.3}$$

where $\circ$ represents the Hadamard (element-wise) product and, abusing notation, for a vector $a \in \mathbb{R}^k$ and scalar $c \in \mathbb{R}$ we let $a + c = (a_i + c)_{i=1}^k$. Later on, we establish the validity of the plug-in analog $\widehat{\Omega}$ (2.13), as an estimator of these matrices.

Theorem 4.1 (Pointwise Normality). *Suppose that Condition 1 and Assumption 4.1 hold. In addition suppose that $\xi_k^2 \log k/n \to 0$. Then so long as either the logistic propensity score model or linear outcome regression model is correctly specified, for any $\alpha \in S^{k-1}$:*

$$\sqrt{n}\frac{\alpha'(\widehat{\beta}^k - \beta^k)}{\|\alpha'\Omega^{1/2}\|} \to_d N(0,1) \tag{4.4}$$

where generally $\Omega = \tilde{\Omega}$ but if $\ell_k c_k \to 0$ then we can set $\Omega = \Omega_0$. Moreover, for any $x \in \mathcal{X}$ and

[1] In practice, we recommend the use of B-splines in order to to satisfy the first requirement that the basis functions are weakly positive and to reduce instability of the convex optimization programs described in (2.8)-(2.9).

$s(x) := \Omega^{1/2} p^k(x)$,

$$\sqrt{n}\frac{p^k(x)'(\widehat{\beta}^k - \beta^k)}{\|s(x)\|} \to_d N(0,1) \tag{4.5}$$

and if the approximation error is negligible relative to the estimation error, namely $\sqrt{n} r_k(x) = o(\|s(x)\|)$, *then*

$$\sqrt{n}\frac{\widehat{g}(x) - g(x)}{\|s(x)\|} \to_d N(0,1) \tag{4.6}$$

Theorem 4.1 shows that the estimator proposed in Section 2 has a limiting gaussian distribution even under misspecification of either first-stage model. This allows for doubly-robust pointwise inference after establishing a consistent variance estimator.

4.2 UNIFORM CONVERGENCE

Next, we turn to strengthening the pointwise results to hold uniformly over all points $x \in \mathcal{X}$. This requires stronger conditions. We make the following assumptions on the tail behavior of the error terms which strengthens Assumption 4.1.

Assumption 4.2 (Uniform Limit Theory). *Let* $\bar{\epsilon}_k = \sup_{1\le j\le k} |\epsilon_j|$, $\alpha(x) := p^k(x)/\|p^k(x)\|$, *and let*

$$\xi_k^L := \sup_{\substack{x,x'\in\mathcal{X}\\ x\neq x'}} \frac{\|\alpha(x) - \alpha(x')\|}{\|x - x'\|}.$$

Further for any integer s *let* $\bar{\sigma}_k^s = \sup_{x\in\mathcal{X}} \mathbb{E}[|\bar{\epsilon}_k|^s | X = x]$. *For some* $m > 2$ *assume*

(i) The regression errors satisfy $\sup_{x\in\mathcal{X}} \mathbb{E}[\max_{1\le i\le n} |\bar{\epsilon}_{k,i}|^m \mid X = x] \lesssim_P n^{1/m}$

(ii) The basis functions are such that (a) $\xi_k^{2m/(m-2)} \log k/n \lesssim 1$, *(b)* $(\bar{\sigma}_k^2 \vee \bar{\sigma}_k^m) \log \xi_k^L \lesssim \log k$, *and (c)* $\log \bar{\sigma}_k^m \xi_k \lesssim \log k$.

As before, Assumption 4.2 is very similar to its analogue in Belloni et al. (2015), with the modification that the conditions are required to hold for $\bar{\epsilon}_k$ as opposed to ϵ_k. Under this assumption, we derive doubly-robust uniform rates of convergence uniform inference procedures for the conditional counterfactual outcome $g_0(x)$.

Theorem 4.2 (Strong Approximation by a Gaussian Process). *Assume that Condition 1 holds and that Assumptions 4.1-4.2 hold with* $m \ge 3$. *In addition assume that (i)* $\bar{R}_{1n} = o_p(a_n^{-1})$ *and (ii)* $a_n^6 k^4 \xi_k^2 (\bar{\sigma}_k^3 + \ell_k^3 c_k^2)^2 \log^2 n/n \to 0$ *where*

$$\bar{R}_{1n} := \sqrt{\frac{\xi_k^2 \log k}{n}} (n^{1/m}\sqrt{\log k} + \sqrt{k}\ell_k c_k) \;\text{ and }\; \bar{R}_{2n} := \sqrt{\log k}\cdot \ell_k c_k$$

Then so long as either the propensity score model or outcome regression model is correctly specified, for some $\mathcal{N}_k \sim N(0, I_k)$:

$$\sqrt{n}\frac{\alpha(x)'(\widehat{\beta} - \beta)}{\|\alpha(x)'\Omega^{1/2}\|} =_d \frac{\alpha(x)'\Omega^{1/2}}{\|\alpha(x)'\Omega^{1/2}\|} N_k + o_p(a_n^{-1}) \;\text{ in } \ell^\infty(\mathcal{X}) \tag{4.7}$$

so that for $s(x) := \Omega^{1/2} p^k(x)$

$$\sqrt{n}\frac{p^k(x)'(\widehat{\beta} - \beta)}{\|s(x)\|} =_d \frac{s(x)}{\|s(x)\|} N_k + o_p(a_n^{-1}) \;\text{ in } \ell^\infty(\mathcal{X}) \tag{4.8}$$

and if $\sup_{x\in\mathcal{X}} \sqrt{n}|r_k(x)|/\|s(x)\| = o(a_n^{-1})$, *then*

$$\sqrt{n}\frac{\widehat{g}(x) - g(x)}{\|s(x)\|} =_d \frac{s(x)'}{\|s(x)\|}N_k + o_p(a_n^{-1}) \;\; in\; \ell^\infty(\mathcal{X}) \tag{4.9}$$

where in general we take $\Omega = \tilde{\Omega}$ *but if* $\bar{R}_{2n} = o_p(a_n^{-1})$ *then we can set* $\Omega = \Omega_0$ *where* $\tilde{\Omega}$ *and* Ω_0 *are as in* (4.3).

Theorem 4.2 establishes conditions under which we obtain a doubly-robust strong approximation of the empirical process $x \mapsto \sqrt{n}(\widehat{g}(x) - g_0(x))$ by a Gaussian process. After establishing consistent estimation of the matrix Ω, this strong approximation result allows us to show validity of the uniform confidence bands described in Section 2. As noted by Belloni et al. (2015), this is distinctly different from a Donsker type weak convergence result for the estimator $\widehat{g}(x)$ as viewed as a random element of $\ell^\infty(X)$. In particular, the covariance kernel is left completely unspecified and in general need not be well behaved.

4.3 MATRIX ESTIMATION AND UNIFORM INFERENCE

We establish that the estimator $\widehat{\Omega}$ proposed in (2.13) is a consistent estimator of the true limiting variance Ω, where $\Omega = \tilde{\Omega}$ in general but if $\bar{R}_{2n} = o_p(a_n^{-1})$ then $\Omega = \Omega_0$. To do so, we rely on the second-stage assumptions Assumptions 4.1 and 4.2 as well as the following condition limiting the first-stage estimation error passed on to the variance estimator $\widehat{\Omega}$.

Condition 2 (Variance Estimation)**.** Let $m > 2$ be as in Assumption 4.2. Then,

$$\xi_{k,\infty} \max_{1\le j\le k} \mathbb{E}_n[p_j(X)^2(Y(\widehat{\pi}_j, \widehat{m}_j) - Y(\bar{\pi}_j, \bar{m}_j))^2] = o_p(k^{-2}n^{-1/m}) \tag{4.10}$$

Via Lemma 3.1 we can establish Condition 2 under Assumption 3.1 as well as the additional sparsity bound[1]

$$\frac{\xi_{k,\infty}^5 s_k^2 k^2 \ln(d_z)}{n^{(m-1)/m}}. \tag{4.11}$$

Theorem 4.3 (Matrix Estimation)**.** *Suppose that Conditions 1 and 2 and Assumptions 4.1-4.2 hold. In addition, assume that* $\bar{R}_{1n} + \bar{R}_{2n} \lesssim (\log k)^{1/2}$. *Then, so long as either the propensity score model or outcome regression model is correctly specified then for* $\widehat{\Omega} = \widehat{Q}^{-1}\widehat{\Sigma}\widehat{Q}^{-1}$:

$$\|\widehat{\Omega} - \Omega\| \lesssim_P (v_n \vee \ell_k c_k)\sqrt{\frac{\xi_k^2 \log k}{n}} = o(1)$$

Theorem 4.3 establishes that pointwise inference based on the test statistic described in Section 2, obtained by replacing Ω in Theorem 4.1 with the consistent estimator $\widehat{\Omega}$, is doubly-robust. Hypothesis tests based on the test statistic as well as pointwise confidence intervals for $g_0(x)$ remain valid even if one of the first-stage parameters is misspecified.

We now establish the validity of uniform inference based on the gaussian bootstrap critical values $c_u^\star(1-\alpha)$ defined in Section 2.

[1]The sparsity bound (4.11) required for consistent variance estimation can be significantly sharpened if the researcher is willing to use a cross fitting procedure, using one sample to estimate the nuisance parameters and another to evaluate the aIPW signal. This is because one could more directly follow Semenova and Chernozhukov (2021) and control alternate quantities with bounds that converge more quickly to zero.

Theorem 4.4 (Validity of Uniform Confidence Bands). *Suppose Conditions 1 and 2 are satisfied and Assumptions 4.1–4.2 hold with $m \geq 4$. In addition suppose (i) $R_{1n} + R_{2n} \lesssim \log^{1/2} n$, (ii) $\xi_k \log^2 n / n^{1/2-1/m} = o(1)$, (iii) $\sup_{x\in\mathcal{X}} |r_k(x)|/\|p^k(x)\| = o(\log^{-1/2} n)$, and (iv) $k^4\xi_k^2(1 + l_k^3 r_k^3)^2 \log^5 n/n = o(1)$. Then, so long as either the propensity score model or outcome regression model is satisfied*

$$\Pr\left(\sup_{x\in\mathcal{X}} \left|\frac{\widehat{g}(x) - g(x)}{\widehat{\sigma}(x)}\right| \leq c^\star(1-\alpha)\right) = 1 - \alpha + o(1).$$

As a result, uniform confidence intervals formed in (2.14) satisfy

$$\Pr(g(x) \in [\underline{i}(x), \bar{i}(x)],\ \forall x \in \mathcal{X}) = 1 - \alpha + o(1).$$

In conjunction with Lemma 3.1, Theorem 4.1 and Theorem 4.3, Theorem 4.4 shows the validity of the uniform inference procedure described in Section 2.

5 Estimation of the Conditional Average Treatment Effect

Up to now, we have mainly focused on doubly-robust estimation and model-assisted inference for the function

$$g_0(x) = \mathbb{E}[Y_1 \mid X = x].$$

We conclude by noting that we can use a symmetric procedure to obtain model-assisted inference for the additional conditional counterfactual outcome

$$\tilde{g}_0(x) = \mathbb{E}[Y_0 \mid X = x].$$

To do so, we use the alternate aIPW signal

$$Y_0(\pi_0, m_0) = \frac{(1-D)Y}{1-\pi_0(Z)} + \left(\frac{1-D}{1-\pi_0(Z)} - 1\right) m_0(Z)$$

where as before the true value for $\pi_0^\star(z) = \Pr(D = 1 \mid Z = z)$ but now $m_0^\star(z) = \mathbb{E}[Y \mid D = 0, Z = z]$. To estimate these nuisance models we again assume a logistic form for the propensity score model $\pi_0(z) = \pi(z; \gamma^0)$ and a linear form for the outcome regression model $m_0(z) = m(z, \alpha^0)$ as in (2.7) and use a separate estimation procedure for each basis term in our series approximation of $\tilde{g}_0(x)$. The estimating equations we use to estimate each γ_j^0 and α_j^0 differ from those in (2.8)-(2.9) however, and are instead given

$$\widehat{\gamma}_j^0 := \arg\min_\gamma \mathbb{E}_n[p_j(X)\{(1-D)e^{\gamma' Z} - D\gamma' Z\}] + \lambda_{\gamma,j}\|\gamma\|_1$$

$$\widehat{\alpha}_j^0 := \arg\min_\alpha \mathbb{E}_n[p_j(Z)(1-D)e^{\widehat{\gamma}_j^{0\prime} Z}(Y - \alpha' Z)^2]/2 + \lambda_{\alpha,j}\|\alpha\|_1$$

which under the natural analog of Assumption 3.1 converge uniformly to population minimizers:

$$\bar{\gamma}_j^0 := \arg\min_\gamma \mathbb{E}[p_j(X)\{(1-D)e^{\gamma' Z} - D\gamma' Z\}]$$

$$\bar{\alpha}_j^0 := \arg\min_\alpha \mathbb{E}[p_j(Z)(1-D)e^{\bar{\gamma}_j^{0\prime} Z}(Y - \alpha' Z)^2]$$

Letting $\bar{\pi}_{0,j}(z) = \pi(z, \bar{\gamma}_j^0)$, and $\bar{m}_{0,j}(z) = m(z, \bar{\alpha}_j^0)$ we can repeat the decomposition of Section 3, expressing $\tilde{Y}(\bar{\pi}_{0,j}, \bar{m}_{0,j})$ as functions of the parameters $\bar{\gamma}_j^0$ and $\bar{\alpha}_j^0$ and show that the first order

conditions for $\bar{\gamma}_j^0$ and $\bar{\alpha}_j^0$ directly control the bias passed on to the second stage nonparametric estimator for $\tilde{g}_0(x)$. Convergence rates and validity of inference then follow from symmetric analysis of the results in Sections 3 and 4. Combining estimation and inference of the two conditional counterfactual outcomes then gives a doubly-robust estimator and inference procedure for the CATE. To perform inference on the CATE we can use the variance matrix

$$\bar{\Omega} = \Omega_0 + \Omega_1 - 2\Omega_2$$

where Ω_0 is as in (4.3) but Ω_1 and Ω_2 are given

$$\begin{aligned} \Omega_1 &= Q^{-1}\mathbb{E}[\{p^k(x) \circ \epsilon_0^k\}\{p^k(x) \circ \epsilon_0^k\}']Q^{-1} \\ \Omega_2 &= Q^{-1}\mathbb{E}[\{p^k(x) \circ \epsilon^k\}\{p^k(x) \circ \epsilon_0^k\}']Q^{-1} \end{aligned} \tag{5.1}$$

where $\epsilon_{0,j}^k = Y_0(\bar{\pi}_{0,j}, \bar{m}_{0,j}) - \tilde{g}_0(x)$ and $\epsilon_0^k = (\epsilon_{0,1}^k, \ldots, \epsilon_{0,k}^k)'$. These matrices can be consistently estimated using their natural empirical analogs as in (2.13).

6 Empirical Application

We apply the model assisted estimator to estimate the effect of maternal smoking on infant birthweight conditional on the age of the mother. We use the Cattaneo (2010) dataset which can be found online on the Stata website.[1] The dataset describes each infant's birthweight in grams, Y, whether or not the mother smoked during pregnancy, $D = 1$ indicating smoking, and a number of covariates containing information on the mother's health and socioeconomic background, $Z = (X, Z_1)$, where X represents the conditioning variable, maternal age. The dataset includes a base of 21 control variables. We additionally construct quadratic powers and interactions of continuous control variables to generate an additional 29 control variables so that in total $d_z = 50$. A full summary of the data used as well as additional details/analysis from our empirical analysis can be found in Appendix E.

We compare the model assisted estimator of the CATE against one where standard MLE and OLS loss functions are used to estimate the first stage propensity score and outcome regression models. We also qualitatively compare our results to Zimmert and Lechner (2019), who use a kernel based approach to estimate the CATE in this setting. While this sort of comparison is not perfect since we do not know the true DGP, this setting is advantageous for analysis since we strongly expect that (i) the effect of smoking on birthweight will be negative and (ii) this effect should grow stronger in magnitude as the age of the mother increases. These hypotheses have been corroborated by other work that examines the conditional average treatment effect in this setting (Zimmert and Lechner, 2019; Abrevaya, 2006; Lee et al., 2017).

6.1 Empirical Results

Figure 6.1 displays our main results from implementing both the model assisted and standard MLE/OLS estimation procedures. After removing the top 3% and bottom 3% of smoker and non-smoker birthweights by maternal age, we select the penalty parameters for the first stage models via the bootstrap procedure described in Section 4. The pilot penalty parameters are uniformly taken to be equal to zero, so that the residuals used in the bootstrap procedure are generated from non-regularized estimations. We take $c_0 = 2$ in (2.18) and select the first stage penalty parameters using the 90^{th}, 85^{th}, and 80^{th} quantiles of the bootstrap distribution. For the second stage basis functions we implement second degree b-splines with 3 knots via the splines2 package in R (Wang and Yan, 2021).

[1] The dataset can be downloaded here.

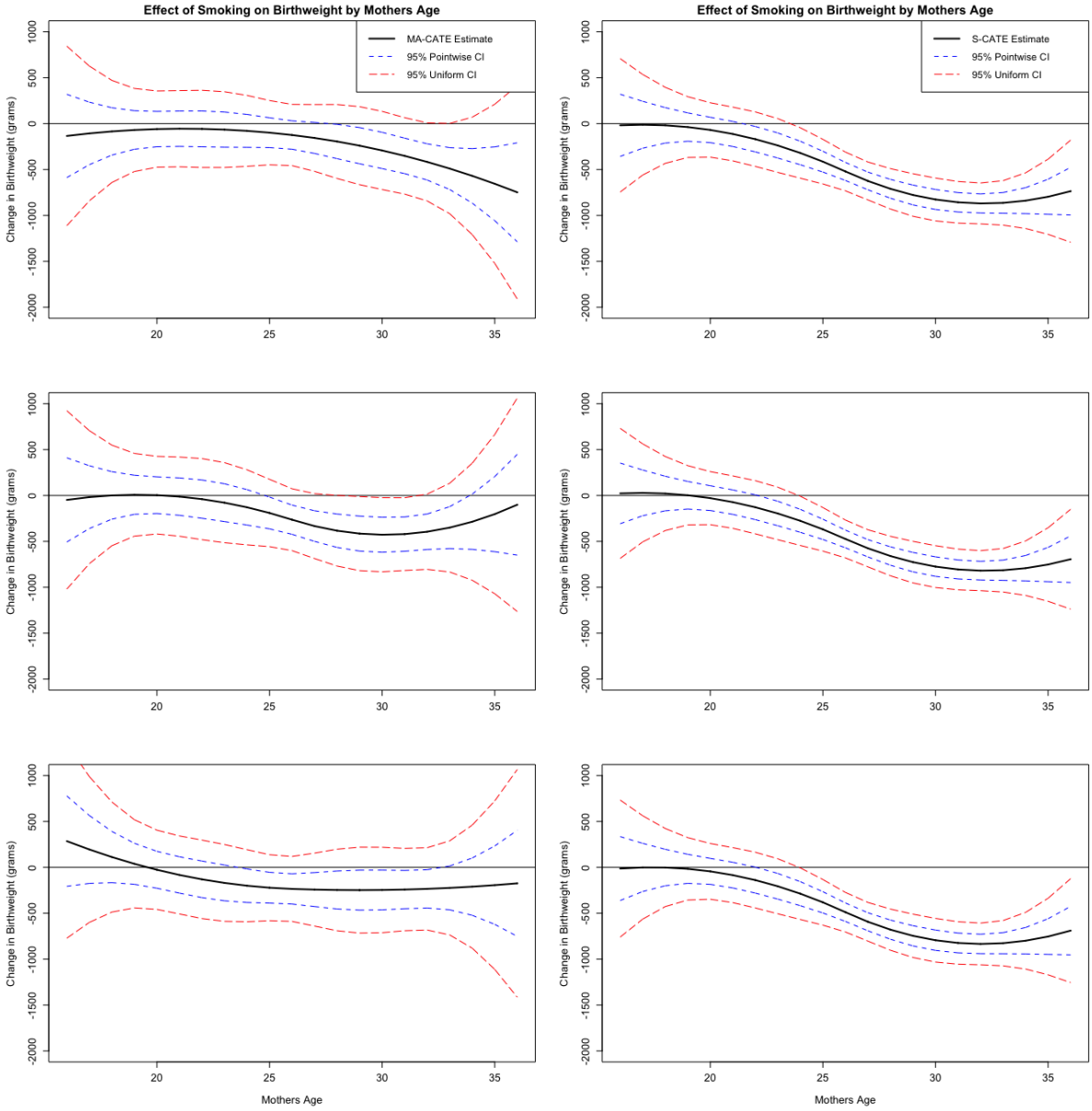

Figure 6.1: CATE of maternal smoking estimated using model assisted estimating equations (left) and standard MLE/OLS estimating equations (right). Top row uses the 90^{th} quantile of the bootstrap distribution to select the penalty parameters, second row uses 85^{th} quantile, and final row uses the 80^{th} quantile. Second stage is computed using b-splines of the second degree with 3 knots. 95% pointwise confidence intervals are displayed in blue short dashes and 95% uniform confidence bands are displayed in long red dashes.

Consistent with prior work, both estimators of the CATE suggest that the effect of smoking on birthweight becomes more negative with age. Both estimation procedures also generally produces negative estimates for the CATE, but it should be noted that for the lowest levels of penalization the model assisted CATE estimate suggests a slightly positive effect of smoking for particularly young mothers, though this difference is not significantly different from zero. The shapes of the estimated functions remain relatively stable under various sizes of the penalty parameter, though the model assisted procedure is more sensitive to the level of regularization introduced.[1] Overall, the magnitude of the CATE estimates produced by the model assisted estimator seem to be more reasonable those produced by the standard estimator.

For the most part, the effects found here are similar to those found in Zimmert and Lechner (2019), though the effects estimated using standard first stage loss functions have somewhat larger magnitudes and in general both series estimation procedures seem to give less reasonable results on the boundaries. An advantage of using a series second stage however, in contrast to the kernel second stage of Zimmert and Lechner (2019), is the existence of the uniform confidence bands displayed. Reassuringly, the estimates of Zimmert and Lechner (2019) seem to be within the 95% uniform confidence bands generated by the model assisted estimator.

As a robustness check, we also try estimating the treatment effect via second degree splines with five knots and first degree splines with seven knots. These results are displayed in Figures 6.2 and 6.3, respectively. Again, we find that the effect of smoking on child birthweight is almost uniformly negative regardless of estimation procedure used or choice of penalty parameter. The shape of the estimated CATE function remains fairly stable under both alternative specifications. Again, the confidence bands from the model assisted procedure remain larger than the confidence bands from the standard procedure. However, the in the first degree spline specification the uniform confidence bands for the standard procedure suggest a significantly positive CATE for some values of maternal age; an implausible result.

Finally, Table 6.1 reports the smoothed average treatment effect estimates taken from averaging the model assisted CATE estimates from Figure 6.1 across observations. Again, these estimates are in line with prior work

Table 6.1: Smoothed Model Assisted ATE Estimates

Bootstrap Penalty Qt.	90^{th}	85^{th}	80^{th}
Implied ATE	-163.257	-222.431	-207.827

7 Simulation Study

We investigate the finite-sample performance of the doubly-robust estimator and inference procedure via simulation study. We find that our proposed estimation procedure retains good coverage properties even under misspecification.

7.1 Simulation Design

Observations are generated i.i.d. according to the following distributions The error term is generated following $\epsilon \sim N(0,1)$. The controls are set $Z_i = (Z_{1i}, X_i) \in \mathbb{R}^{d_z}$ where $d_z = 100$, $X \sim U(1,2)$, and the independent regressors Z_1 are jointly centered Gaussian with a covariance matrix of the Toeplitz form

$$\text{Cov}(Z_{1,j}, Z_{1,k}) = \mathbb{E}[Z_{1,j}Z_{1,k}] = 2^{-|j-k|}, \quad 3 \leq j,k \leq d_z.$$

[1] Numerically solving the minimization problems in (2.8)-(2.9) also typically requires more iterations to converge than solving the standard MLE/OLS minimization problems.

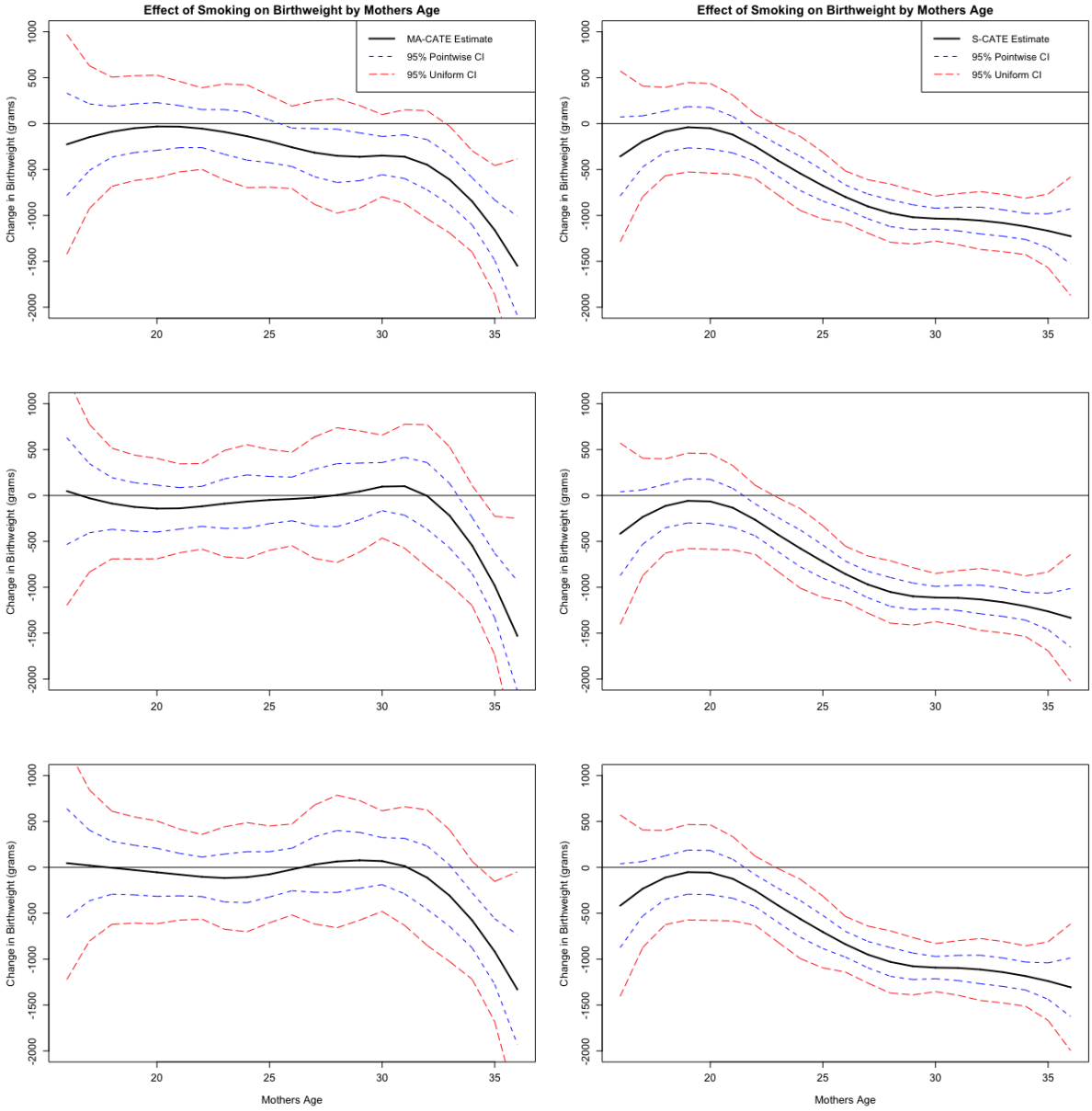

Figure 6.2: CATE of maternal smoking estimated using model assisted estimating equations (left) and standard MLE/OLS estimating equations (right). Top row uses the 95th quantile of the bootstrap distribution to select the penalty parameters, second row uses 90th quantile, and final row uses the 85th quantile. Second stage is computed using b-splines of the second degree with 5 knots. 95% pointwise confidence intervals are displayed in blue short dashes and 95% uniform confidence bands are displayed in long red dashes.

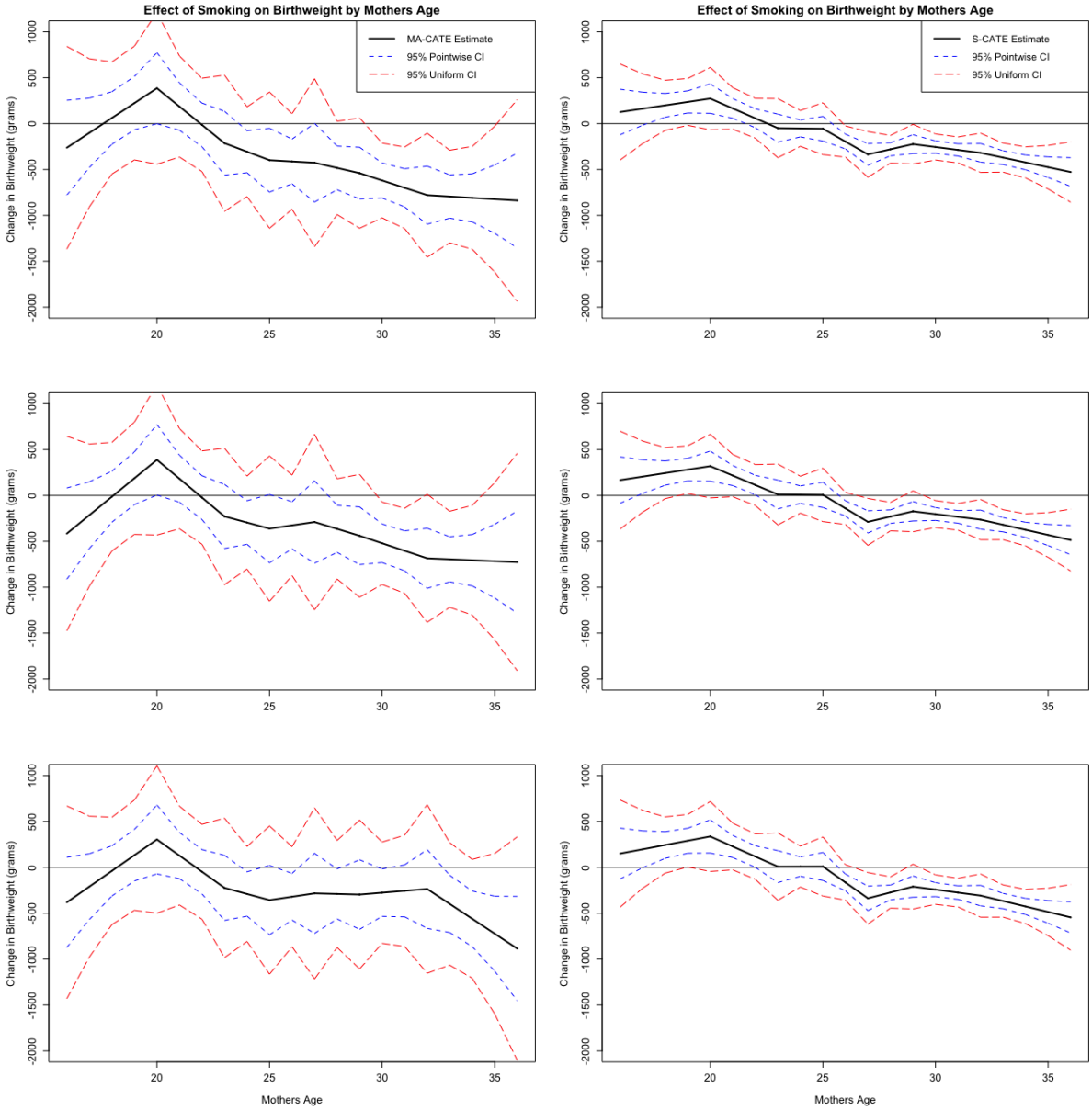

Figure 6.3: CATE of maternal smoking estimated using model assisted estimating equations (left) and standard MLE/OLS estimating equations (right). Top row uses the 95th quantile of the bootstrap distribution to select the penalty parameters, second row uses 90th quantile, and final row uses the 85th quantile. Second stage is computed using b-splines of the second degree with 5 knots. 95% pointwise confidence intervals are displayed in blue short dashes and 95% uniform confidence bands are displayed in long red dashes.

To capture misspecification, we let $Z^\dagger$ be a transformation of the regressors in Z_1 where $Z_j^\dagger = Z_j + \max(0, 1+Z_j)^2$, $\forall\, j = 3, \ldots, d_z$. Let `sparsity` control the number of regressors in $Z = (Z_1, X)$ entering the DGP.

(S1) *Correct specification*: Generate D given Z from a Bernoulli distribution with $\Pr(D = 1|Z) = \{1 + \exp(p_1 - X - 0.5X^2 - \gamma' Z_1)\}^{-1}$ and $Y = D(1 + X + 0.5X^2 + \gamma' Z_1) + \epsilon$.

(S2) *Propensity score model correctly specified, but outcome regression model misspecified*: Generate D given Z as in (S1), but $Y = D(1 + X + 0.5X^2 + \gamma' Z_1^\dagger) + \epsilon$.

(S3) *Propensity score model misspecified, but outcome regression model correctly specified*: Generate Y according to (S1), but generate D given Z from a Bernoulli distribution with $\Pr(D = 1|Z) = \{1 + \exp(p_2 - X - 0.5X^2 + \gamma' Z_1^\dagger)\}^{-1}$.

where the constants p_1 and p_2 differ in various simulation setups but are always set so that the average probability of treatment is about one half. To consider various degrees of high-dimensionality, we implement $N \in \{500, 1000\}$ with $d_z = 100$. For (S1), `sparsity`$= 6$; for (S2), `sparsity`$= 4$; and, for (S3), `sparsity`$= 5$. Results are reported for $S = 1,000$ repeated simulations.

7.2 ESTIMATORS AND IMPLEMENTATION

To select the first stage penalty parameters, we implement the multiplier bootstrap procedure described in Section 2.3. The constants $c_{\gamma,j}$ and $c_{\alpha,j}$ in the pilot penalty parameters (2.15) are selected via cross validation from a set of size 5. To select the final bootstrap penalty parameter we set $c_0 = 1.1$ and select the 95^{th} quantile of $B = 10000$ bootstrap replications. In our second-stage estimation, we use a b-spline basis of size $k = 3$. B-splines are implemented from the R package `splines2` (Wang and Yan, 2021), which uses the specification detailed in Perperoglou et al. (2019). In the tables below, we refer to our method as *MA-DML* (model assisted double machine learning).

We compare our proposed estimator and inference procedure to that of Semenova and Chernozhukov (2021), which projects a single aIPW signal onto a growing series of basis terms. In implementing this *DML* method, we use the standard ℓ_1-penalized maximum likelihood (MLE) and ordinary least squares (OLS) loss functions to estimate the first stage propensity score and outcome regression models, respectively.[1]

Estimation error is studied for the target parameter $g_0(x) = \mathbb{E}[Y|D = 1, X = x]$ over a grid of 100 points spaced across $x \in [1, 2]$, i.e. the support of X. We study average coverage across simulations of each method's pointwise (at $x = 1.5$) and uniform confidence intervals. To compare the estimation error for the target parameter $g(x)$ across the two different estimators $\widehat{g}_s(x)$ for each simulation $s = 1, \ldots, S$, we utilize integrated bias, variance, and mean-squared

[1] Vira Semenova provides several example R scripts implementing *DML*: https://sites.google.com/view/semenovavira/research.

Table 7.1: Simulation study.

DGP	Estimator	IBias^2 (1)	IVar (2)	IMSE (3)	Cov90 (4)	Cov95 (5)	UCov90 (6)	UCov95 (7)
		K=3, n=500, $d_z = 100$						
(S1)	DML	0.04	0.31	0.35	0.92	0.96	1.00	1.00
	MA-DML	~0.0	0.34	0.34	0.93	0.97	1.00	1.00
(S2)	DML	0.16	2.17	2.33	0.92	0.97	0.83	0.86
	MA-DML	0.03	2.12	2.15	0.90	0.94	0.88	0.91
(S3)	DML	0.03	0.55	0.59	0.87	0.93	0.95	0.97
	MA-DML	0.01	0.79	0.80	0.91	0.95	0.99	0.99
		K=3, n=1000, $d_z = 100$						
(S1)	DML	0.12	0.20	0.32	0.83	0.90	0.96	0.96
	MA-DML	0.01	0.22	0.23	0.83	0.90	0.99	0.99
(S2)	DML	0.40	2.1	2.5	0.84	0.91	0.33	0.39
	MA-DML	0.19	2.07	2.26	0.83	0.89	0.50	0.55
(S3)	DML	0.11	0.34	0.46	0.74	0.82	0.80	0.84
	MA-DML	0.01	0.53	0.54	0.84	0.89	0.89	0.91

Note: DGP refers to the three various data generating processes introduced above. IBias^2, IVar, and IMSE refer to integrated squared bias, variance, and mean squared error, respectively. Cov90, Cov95, UCov90, and UCov95 refer to the coverage proportion of the 90% and 95% pointwise and uniform confidence intervals across simulations. K refers to the number of series terms, N to the sample size, and d_z to the dimensionality of the random variable Z_1.

error where $\bar{g}(x) = S^{-1}\sum_{s=1}^{S}\widehat{g}_s(x)$,

$$\text{IBias}^2 = \int_0^1 (\bar{g}(x) - g_0(x))^2 dx,$$

$$\text{IVar} = S^{-1}\sum_{s=1}^{S}\int_0^1 (\widehat{g}_s(x) - \bar{g}(x))^2 dx,$$

$$\text{IMSE} = S^{-1}\sum_{s=1}^{S}\int_0^1 (\widehat{g}_s(x) - g_0(x))^2 dx.$$

7.3 SIMULATION RESULTS

Table 7.1 presents the simulation results for all three specifications (S1)-(S3) for $n = 500$ and $n = 1000$. Integrated squared bias, variance, and mean squared error are presented in columns (1)-(3), respectively. Pointwise and uniform coverage results are presented in columns (4)-(7).

For pointwise and uniform coverage under correct specification regime (S1), *MA-DML* has some slight improvements. Under misspecification DGPs (S2) and (S3), the pointwise coverage of *MA-DML* is closer to the targets except in the $N = 1000$ and (S2) case where it slightly underperforms. However, *MA-DML* has a notable improvement over *DML* in the (S3) case when $N = 1000$. Similarly, *MA-DML* outperforms *DML* in three of the four misspecified regimes, i.e. all but (S3) when $N = 500$ where *MA-DML* has over-coverage. Under (S2) when

$N = 1000$, both methods are markedly deterioated uniform coverage, although *MA-DML* is noticably closer to target.

In regards to estimation error, in four of the six settings, *MA-DML* has a lower MSE than *DML* where regardless of sample size *MA-DML* underperforms in (S3). Notably, it does appear *MA-DML* has substantially smaller IBias^2 across the DGPs.

Finally, we were surprised to find for both estimators that coverage properties, in general, improve under the higher-dimensional regime of $N = 500$ with $d_z = 100$ compared to $N = 1,000$ and $d_z = 100$. In particular, with a higher ratio of covariates to observations, the uniform coverage properties under regime (S2) were substantially better. The estimation error results were in line with our priors as the higher-dimensional regime sees in general higher estimation errors for both methods.

For coverage under correct specification, we did anticipate the underperformance of *MA-DML* given it is designed to handle misspecification with the cost of other estimators outperforming under correct specification. Additionally, we attribute the poor uniform coverage in DGP (S2) for both estimators under $N = 1,000$ to a lack of a rich enough cross-validation given the performance was improved under a more difficult regime when the number of observations drops to $N = 500$. The integrated bias of *MA-DML* is lower across the various DGPs compared to *DML*. Following the discussion in Section 3 this is expected since the first stage estimating equations for the model assisted procedure are specifically designed to minimize the bias passed on to the second stage estimator. However, the model assisted procedure has higher values of integrated variance compared to the standard procedure, which could be attributable to the use of k distinct first-stage estimations.

Our findings should not be interpreted as a critique of the Semenova and Chernozhukov (2021) benchmark method, whose work we rely on and were inspired by.

8 CONCLUSION

Estimation of conditional average treatment effects with high dimensional controls typically relies on first estimating two nuisance parameters: a propensity score model and an outcome regression model. In a high-dimensional setting, consistency of the nuisance parameter estimators typically relies on correctly specifying their functional forms. While the resulting second-stage estimator for the conditional average treatment effect typically remains consistent even if one of the nuisance parameters is inconsistent, the confidence intervals may no longer be valid.

In this paper, we consider estimation and valid inference on the conditional average treatment effect in the presence of high dimensional controls and nuisance parameter misspecification. We present a nonparametric estimator for the CATE that remains consistent at the nonparametric rate, under slightly modified conditions, even under misspecification of either the logistic propensity score model or linear outcome regression model. The resulting Wald-type confidence intervals based on this estimator also provide valid asymptotic coverage under nuisance parameter misspecification.

REFERENCES

Abrevaya, J. (2006). Estimating the effect of smoking on birth outcomes using a matched panel data approach. *Journal of Applied Econometrics 21*(4), 489–519.

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

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*(6), 2369–2429.

Belloni, A. and V. Chernozhukov (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli 19*(2), 521 – 547.

Belloni, A., V. Chernozhukov, D. Chetverikov, C. Hansen, and K. Kato (2018). High-dimensional econometrics and regularized gmm.

Belloni, A., V. Chernozhukov, D. Chetverikov, and K. Kato (2015). Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics 186*(2), 345–366. High Dimensional Problems in Econometrics.

Belloni, A., V. Chernozhukov, and C. Hansen (2013, 11). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies 81*(2), 608–650.

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics 37*(4), 1705 – 1732.

Bradic, J., S. Wager, and Y. Zhu (2019, May). Sparsity Double Robust Inference of Average Treatment Effects. Papers 1905.00744, arXiv.org.

Bühlmann, P. and S. van de Geer (2011). *Statistics for high-dimensional data*. Springer Series in Statistics. Springer, Heidelberg. Methods, theory and applications.

Cattaneo, M. (2010). Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *Journal of Econometrics 155*(2), 138–154.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. In J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics* (1 ed.), Volume 6B, Chapter 76, pp. 5549–5632. Elsevier.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018, 01). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal 21*(1), C1–C68.

Chernozhukov, V., D. Chetverikov, and K. Kato (2017). Central limit theorems and bootstrap in high dimensions. *The Annals of Probability 45*(4), 2309–2352.

Chetverikov, D., Z. Liao, and V. Chernozhukov (2021). On cross-validated Lasso in high dimensions. *The Annals of Statistics 49*(3), 1300 – 1317.

Chetverikov, D. and J. R.-V. Sørensen (2021). Analytic and bootstrap-after-cross-validation methods for selecting penalty parameters of high-dimensional m-estimators. *ArXiv NA*, 1–50.

De Boor, C. (2001). *A practical guide to splines; rev. ed.* Applied mathematical sciences. Berlin: Springer.

der Vaart, A. V. and J. Wellner (1996). *Weak Convergence and Empirical Processes* (1 ed.). Springer Series in Statistics. Springer, New York, NY.

Dudley, R. (1967). The sizes of compact subsets of hilbert space and continuity of gaussian processes. *Journal of Functional Analysis 1*(3), 290–330.

Fan, Q., Y.-C. Hsu, R. P. Lieli, and Y. Zhang (2022). Estimation of conditional average treatment effects with high-dimensional data. *Journal of Business & Economic Statistics 40*(1), 313–327.

Giné, E. and V. Koltchinskii (2006). Concentration inequalities and asymptotic results for ratio type empirical processes. *The Annals of Probability 34*(3), 1143 – 1216.

Hlavac, M. (2022). *stargazer: Well-Formatted Regression and Summary Statistics Tables*. Bratislava, Slovakia: Social Policy Institute. R package version 5.2.3.

Lee, S., R. Okui, and Y.-J. Whang (2017). Doubly robust uniform confidence band for the conditional average treatment effect function. *Journal of Applied Econometrics 32*(7), 1207–1225.

Newey, W. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics 79*(1), 147–168.

Newey, W. K. and D. McFadden (1994). Chapter 36 large sample estimation and hypothesis testing. *Handbook of Econometrics 4*, 2111–2245.

Perperoglou, A., W. Sauerbrei, M. Abrahamowicz, and M. Schmid (2019). A review of spline function procedures in r. *BMC medical research methodology 19*(1), 1–16.

Pollard, D. (2001). *A User's Guide to Measure Theoretic Probability*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology 66*, 688–701.

Rubin, D. B. (1978). Bayesian inference for causal effects. *The Annals of Statistics 6*(1), 34–58.

Rudelson, M. (1999). Random vectors in the isotropic position. *J. Funct. Anal 164*, 60–72.

Schmidt-Hieber, J. (2020, 08). Nonparametric regression using deep neural networks with relu activation function. *Annals of Statistics 48*, 1875–1897.

Semenova, V. and V. Chernozhukov (2021, 08). Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal 24*, 264–289. utaa027.

Smucler, E., A. Rotnitzky, and J. M. Robins (2019). A unifying apptoach for doubly-robust ℓ_1 regularized estimation of causal contrasts. *ArXiv NA*, 1–125.

Tan, Z. (2017). Regularized calibrated estimation of propensity scores with model misspecification and high-dimensional data. *ArXiv NA*, 1–60.

Tan, Z. (2020). Model-assisted inference for treatment effects using regularized calibrated estimation with high-dimensional data. *The Annals of Statistics 48*(2), 811 – 837.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological) 58*(1), 267–288.

van der Greer, S. (2016). *Estimation and Testing under Sparsity*. Lecture Notes in Mathematics. Springer, New York, NY.

Wang, W. and J. Yan (2021). Shape-restricted regression splines with R package splines2. *Journal of Data Science 19*(3), 498–517.

Wu, P., Z. Tan, W. Hu, and X.-H. Zhou (2021). Model-assisted inference for covariate-specific treatment effects with high-dimensional data.

Zimmert, M. and M. Lechner (2019). Nonparametric estimation of causal heterogeneity under high-dimensional confounding.

A Proofs for Results in Main Text

Here we provide proofs of the main results in Sections 3-4. The proofs for Section 4 rely on an assortment of supporting lemmas proved in Appendix B.

A.1 Proofs for Main First Stage Results

Proof of Lemma 3.1

The proof of Lemma 3.1 relies on a series of non-asymptotic bounds that are established in Online Appendix Lemmas B.1 and B.2 that hold on $\bigcap_{m=1}^{6} \Omega_{k,m}$ and depend on the quantity

$$\bar{\lambda}_k = M\xi_{k,\infty}\sqrt{\frac{\log(d_z/\epsilon)}{n}}$$

where M is a fixed constant. In addition let $\tilde{\Sigma}^1_{\alpha,j} := \mathbb{E}_n[p_j(X)De^{-\bar{\gamma}_j'Z}|Y - \bar{\alpha}_j'Z|ZZ']$ and $\Sigma^1_{\alpha,j} := \mathbb{E}\tilde{\Sigma}^1_{\alpha,j}$ and define the event

$$\Omega_{k,7} := \{\|\tilde{\Sigma}^1_{\alpha,j} - \Sigma^1_{\alpha,j}\|_\infty \leq \bar{\lambda}_k, \forall j \leq k\} \tag{A.1}$$

In Online Appendix B.3 we show that $\Pr(\bigcap_{m=1}^{7}) \geq 1 - o(1)$. Under these events, Lemma A.1, below provides the bound needed for first statement of Lemma 3.1 while Lemma A.2 provides the bound needed for the second statement.

Lemma A.1 (Nonasymptotic Bounds for Weighted Means)**.** *Suppose that Assumption 3.1 holds, $\xi_0 > (c_0+1)/(c_0-1)$, and $2C_0\nu_0^{-2}s_k\bar{\lambda}_k \leq \eta < 1$. In addition, assume there is a constant $c > 0$ such that $\lambda_{\alpha,j}/\lambda_{\gamma,j} \geq c$ for all $j \leq k$. Then, under the event $\bigcap_{m=1}^{7}\Omega_{k,m}$, there is a constant M_2 that does not depend on k such that*

$$\max_{1\leq j\leq k} |\mathbb{E}_n[p_j(X)Y(\widehat{\pi}_j, \widehat{m}_j)] - \mathbb{E}_n[p_j(X)Y(\bar{\pi}_j, \bar{m}_j)]| \leq M_2 s_k \bar{\lambda}_k^2 \tag{A.2}$$

Proof. We show that the bound of (A.2) holds for any $j = 1, \ldots, k$ in a couple steps. To save notation, define

$$\begin{aligned}\mu_j(\pi, m) &:= \mathbb{E}_n\left[p_j(X)Y(\pi, m)\right] \\ &= \mathbb{E}_n\left[p_j(X)\left\{\frac{DY}{\pi(Z)} + \left(\frac{D}{\pi(Z)} - 1\right)m(Z)\right\}\right]\end{aligned}$$

Step 1: Decompose Difference and Use Logistic FOCs. Consider the following decomposition

$$\begin{aligned}\mu_j(\widehat{\pi}_j, \widehat{m}_j) - \mu_j(\bar{\pi}_j, \bar{m}_j) = \mathbb{E}\left[p_j(X)\{\widehat{m}_j(Z) - \bar{m}_j(Z)\}\left(1 - \frac{D}{\bar{\pi}_j(X)}\right)\right] \\ + \mathbb{E}_n\left[p_j(X)D\{Y - \bar{m}_j(Z)\}\left(\frac{1}{\widehat{\pi}_j(Z)} - \frac{1}{\bar{\pi}_j(Z)}\right)\right] \\ + \mathbb{E}_n\left[p_j(X)\{\widehat{m}_j(Z) - \bar{m}_j(Z)\}\left(\frac{D}{\bar{\pi}_j(Z)} - \frac{D}{\widehat{\pi}_j(Z)}\right)\right] \\ := \delta_{1,j} + \delta_{2,j} + \delta_{3,j}\end{aligned}$$

Notice that $\delta_{1,j} + \delta_{3,j} = (\widehat{\alpha}_j - \bar{\alpha}_j)'\mathbb{E}_n[p_j(X)(1 - D/\widehat{\pi}_j(Z))Z]$. By the first order conditions for $\widehat{\gamma}_j$ we have that

$$|\mathbb{E}_n[p_j(X)\{Z_l - DZ_l/\widehat{\pi}_j(Z)\}]| \leq \lambda_{\gamma,j}\ \forall l = 1, \ldots, d_z \implies \|\mathbb{E}_n[p_j(X)\{Z_l - DZ_l/\widehat{\pi}_j(Z)\}]\|_\infty \leq \lambda_{\gamma,j}.$$

Applying Hölder's inequality to $\delta_{1,j} + \delta_{3,j}$ then gives us that on the event $\Omega_{k,2}$

$$|\delta_{1,j} + \delta_{3,j}| \leq \|\widehat{\alpha}_j - \bar{\alpha}_j\|_1 \lambda_{\gamma,j} \leq \|\widehat{\alpha}_j - \bar{\alpha}_j\|\bar{\lambda}_k.$$

By Lemma B.2 on the event $\bigcap_{m=1}^{6}\Omega_{k,m}$ and under the conditions of Lemma A.1, $\|\widehat{\alpha}_j - \bar{\alpha}_j\| \leq M_1 s_k \bar{\lambda}_k$ where M_1 is a constant that does not depend on k. So

$$|\delta_{1,j} + \delta_{3,j}| \leq M_1 s_k \bar{\lambda}_k^2 \tag{M.1}$$

Step 2: Use Outcome Regression Score Domination to Bound $\delta_{2,j}$. Now deal with the term $\delta_{2,j}$. By first order Taylor expansion, for some $u \in (0,1)$

$$\begin{aligned}\delta_{2,j} &= -(\widehat{\gamma}_j - \bar{\gamma}_j)'\mathbb{E}_n[p_j(X)D\{Y - \bar{m}_j(Z)\}e^{-\bar{\gamma}_j'Z}Z] \\ &\quad + (\widehat{\gamma}_j - \bar{\gamma}_j)'\mathbb{E}_n[p_j(X)D\{Y - \bar{m}_j(Z)\}e^{-u\widehat{\gamma}_j'Z - (1-u)\bar{\gamma}_j'Z}ZZ'](\widehat{\gamma}_j - \bar{\gamma}_j)/2 \\ &:= \delta_{21,j} + \delta_{22,j}\end{aligned}$$

In the event $\Omega_{k,1} \cap \Omega_{k,2} \cap \Omega_{k,3} \cap \Omega_{k,4}$ we have by score domination of the linear outcome regression model and Lemma B.1 that $\delta_{21} \leq M_0 s_k \bar{\lambda}_k^2$.

The term $\delta_{22,j}$ is second order. On the event $\Omega_{k,0} \cap \Omega_{k,1}$ where $\|\widehat{\gamma}_j - \bar{\gamma}_j\|_1 \leq M_0 s_k \bar{\lambda}_k \leq M_0\eta/C_0$ it can be bounded with

$$\begin{aligned}\delta_{22,j} &\leq e^{C_0\|\widehat{\gamma}_j - \bar{\gamma}_j\|_1}\mathbb{E}_n[p_j(X)De^{-\bar{\gamma}_j'Z}|Y - \bar{m}_j(Z)|\{\widehat{\gamma}_j'Z - \bar{\gamma}_j'Z\}^2] \\ &\leq e^{M_0\eta}\mathbb{E}_n[p_j(X)De^{-\bar{\gamma}_j'Z}|Y - \bar{m}_j(Z)|\{\widehat{\gamma}_j'Z - \bar{\gamma}_j'Z\}^2].\end{aligned}$$

This in turn is bounded in a few steps. First note on the event $\Omega_{k,7}$

$$(\mathbb{E}_n - \mathbb{E})[p_j(X)De^{-\bar{\gamma}_j'Z}|Y - \bar{m}_j(Z)|\{\widehat{\gamma}_j'Z - \bar{\gamma}_j'Z\}^2] \leq \bar{\lambda}_k\|\widehat{\gamma}_j - \bar{\lambda}_j\|_1^2.$$

By Assumption 3.1 we have that $G_0^2 E[D|Y - \bar{m}_j(Z)| \mid Z] \leq G_1^2/G_0 + G_0$ so that,

$$\mathbb{E}[p_j(X)De^{-\bar{\gamma}_j'Z}|Y - \bar{m}_j(Z)|\{\widehat{\gamma}_j'Z - \bar{\gamma}_j'Z\}^2] \leq (G_1^2/G_0 + G_0)\mathbb{E}[p_j(X)De^{-\bar{\gamma}_j'Z}\{\widehat{\gamma}_j'Z - \bar{\gamma}_j'Z\}^2].$$

On the event $\Omega_{k,6}$ we have that

$$(\mathbb{E}_n - \mathbb{E})[p_j(X)De^{-\bar{\gamma}_j'Z}\{\widehat{\gamma}_j'Z - \bar{\gamma}_j'Z\}^2] \leq \bar{\lambda}_k\|\widehat{\gamma}_j - \bar{\gamma}_j\|_1.$$

Putting these all together gives

$$\begin{aligned}\mathbb{E}_n[p_j(X)De^{-\bar{\gamma}_j'Z}|Y - \bar{m}_j(Z)|\{\widehat{\gamma}_j'Z - \bar{\gamma}_j'Z\}^2] \\ \leq \bar{\lambda}_k\|\widehat{\gamma}_j - \bar{\gamma}_j\|_1^2 + (G_1^2/G_0 + G_0)\bar{\lambda}_k\|\widehat{\gamma}_j - \bar{\gamma}_j\|_1^2 \\ + (G_1^2/G_0 + G_0)\mathbb{E}_n[p_j(X)De^{-\bar{\gamma}'Z}\{\widehat{\gamma}_j'Z - \bar{\gamma}_j'Z\}]\end{aligned} \tag{M.2}$$

To bound (M.2) note again that in the event $\Omega_{k,1} \cap \Omega_{k,2}$, $\|\widehat{\gamma}_j - \bar{\gamma}_j\|_1 \leq M_0 s_k \bar{\lambda}_k$ and that using by (O.4) in Online Appendix Lemma B.2:

$$\mathbb{E}_n[p_j(X)De^{-\bar{\gamma}_j'Z}\{\widehat{\gamma}_j'Z - \bar{\gamma}_j'Z\}^2] \leq e^{-M_0\eta}M_0 s_k\bar{\lambda}_k^2.$$

Plugging these into (M.2) gives

$$\delta_{22,j} \leq e^{M_0\eta}M_0^2 s_k^2\bar{\lambda}_k^3 + e^{M_0\eta}(G_1^2/G_0 + G_0)M_0^2 s_k^2\bar{\lambda}_k^3 + (G_1^2/G_0 + G_0)M_0 s_k\bar{\lambda}_k^2 \tag{M.3}$$

so that in total $\delta_{2,j} = \delta_{21,j} + \delta_{22,j}$ is bouned

$$\delta_{2,j} \leq M_0 s_k(G_1^2/G_0 + G_0 + 1)\bar{\lambda}_k^2 + e^{M_0\eta}M_0^2 s_k^2(G_1^2/G_0 + G_0 + 1)\bar{\lambda}_k^3 \tag{M.4}$$

Step 3: Combine Terms. Putting this together yields

$$\begin{aligned}|\delta_{1,j} + \delta_{2,j} + \delta_{3,j}| \leq \{M_1 + M_0(G_1^2/G_0 + G_0 + 1)\}s_k\bar{\lambda}_k^2 \\ + e^{M_0\eta}(G_1^2/G_0 + G_0)M_0^2 s_k^2\bar{\lambda}_k^3\end{aligned} \tag{M.5}$$

Use the fact that $s_k\bar{\lambda}_k \leq \eta < 1$ to simplify the last term of this expression

$$\begin{aligned}|\delta_{1,j} + \delta_{2,j} + \delta_{3,j}| \leq \{M_1 + M_0(G_1^2/G_0 + G_0 + 1)\}s_k\bar{\lambda}_k^2 \\ + e^{M_0\eta}(G_1^2/G_0 + G_0)M_0^2 s_k\bar{\lambda}_k\end{aligned} \tag{M.6}$$

This gives the result (A.2) after taking $M_2 = M_1 + M_0(G_1^{/}G_0 + G_0 + 1) + e^{M_0\eta}(G_1^2/G_0 + G_0)M_0^2$.

□

Lemma A.2 (Nonasymptotic Bounds for Variance Estimation). *Suppose that Assumption 3.1 hold, $\xi_0 > (c_0 + 1)/(c_0 - 1)$, and $2C_0\nu_0^{-2}s_k\bar{\lambda}_k \leq \eta < 1$. In addition, assume there is a constant $c > 0$ such that $\lambda_{\alpha,j}/\lambda_{\gamma,j} \geq c$ for all $j \leq k$. Then, under the event $\bigcap_{m=1}^7 \Omega_{k,m}$, there is a constant M_3 that does not*

depend on k such that

$$\max_{1\le j\le k} \mathbb{E}_n[p_j^2(X)(Y(\widehat{\pi}_j,\widehat{m}_j)-Y(\bar{\pi}_j,\bar{m}_j))^2] \le M_3 \xi_{k,\infty}^2 s_k^2 \bar{\lambda}_k^2 \tag{A.3}$$

Proof. We show the bound holds for each $j = 1,\dots,k$. We start by decomposing

$$\begin{aligned} p_j(X)(Y(\widehat{\pi}_j,\widehat{m}_j)-Y(\bar{\pi}_j,\bar{m}_j)) &= p_j(X)\{\widehat{m}_j(Z)-\bar{m}_j(Z)\}\left(1-\frac{D}{\bar{\pi}_j(X)}\right) \\ &\quad + p_j(X)D\{Y-\bar{m}_j(Z)\}\left(\frac{1}{\widehat{\pi}_j(Z)}-\frac{1}{\bar{\pi}_j(Z)}\right) \\ &\quad + p_j(X)\{\widehat{m}_j(Z)-\bar{m}_j(Z)\}\left(\frac{D}{\bar{\pi}_j(Z)}-\frac{D}{\widehat{\pi}_j(Z)}\right) \\ &:= \tilde{\delta}_{1,j}+\tilde{\delta}_{2,j}+\tilde{\delta}_{3,j} \end{aligned}$$

We will use the fact that $(a+b+c)^2 \le 4a^2+4b^2+4c^2$ to bound

$$\mathbb{E}_n[p_j^2(X)(Y(\widehat{\pi}_j,\widehat{m}_j)-Y(\bar{\pi}_j,\bar{m}_j))^2] \le 4\mathbb{E}_n[\tilde{\delta}_{1,j}^2]+4\mathbb{E}_n[\tilde{\delta}_{2,j}^2]+4\mathbb{E}_n[\tilde{\delta}_{3,j}^2]. \tag{V.1}$$

To bound $\mathbb{E}_n[\tilde{\delta}_{2,j}]$ use the mean value equation (O.2) in Online Appendix Lemma B.2 and the lower bound on $\bar{g}_j(z)$ from Assumption 3.1

$$\begin{aligned} \mathbb{E}_n[\tilde{\delta}_{2,j}^2] &= \mathbb{E}_n[p_j^2(X)D\{Y-\bar{m}_j(Z)\}^2\{\widehat{\pi}_j^{-1}(Z)-\bar{\pi}_j^{-1}(Z)\}^2] \\ &\le \xi_{k,\infty}e^{-B_0}\left(1+e^{C_0\|\widehat{\gamma}_j-\bar{\gamma}_j\|_1}\right)^2 \mathbb{E}_n[p_j(X)De^{-\bar{\gamma}_j'Z}\{Y-\bar{m}_j(Z)\}^2\{\widehat{g}_j(Z)-\bar{g}_j(Z)\}^2] \end{aligned}$$

Applying (O.8) in Online Appendix Lemma B.2, Online Appendix Lemma B.1, and $s_k\bar{\lambda}_k \le \eta < 1$ there is a constant $\tilde{M}_1$ that does not depend on k such that in the event $\bigcap_{m=1}^7 \Omega_{k,m}$ this is bounded

$$\le \tilde{M}_1 \xi_{k,\infty} s_k \bar{\lambda}_k^2 \tag{V.2}$$

To bound $\mathbb{E}_n[\tilde{\delta}_{3,j}]$ write $\widehat{\pi}_j^{-1}(Z)-\bar{\pi}_j^{-1}(Z) = e^{-\bar{\gamma}_j'Z}\{e^{-\widehat{\gamma}_j'Z+\bar{\gamma}_j'Z}-1\}$ and use the lower bound on $\bar{g}_j(z)$ from Assumption 3.1:

$$\begin{aligned} \mathbb{E}_n[\tilde{\delta}_{3,j}^2] &= \mathbb{E}_n[p_j^2(X)D\{\widehat{m}_j(Z)-\bar{m}_j(Z)\}^2\{\widehat{\pi}_j^{-1}(Z)-\bar{\pi}_j^{-1}(Z)\}^2] \\ &\le \xi_{k,\infty}e^{-B_0}\left(1+e^{C_0\|\widehat{\gamma}_j-\bar{\gamma}_j\|_1}\right)^2 \mathbb{E}_n[p_j(X)e^{-\bar{\gamma}_j'Z}\{\widehat{m}_j(Z)-\bar{m}_j(Z)\}^2] \end{aligned}$$

Applying Online Appendix Lemma B.2, there is a constant $\tilde{M}_2$ that does not depend on k such that on the event $\bigcap_{m=1}^6 \Omega_{k,m}$ this is bounded

$$\le \tilde{M}_2 \xi_{k,\infty} s_k \bar{\lambda}_k^2 \tag{V.3}$$

Finally, to bound $\mathbb{E}_n[\tilde{\delta}_{1,j}^2]$ again use the lower bound on $\bar{g}_j(z)$ and decompose

$$\begin{aligned}\mathbb{E}_n[\tilde{\delta}_{1,j}^2] &= \mathbb{E}_n[p_j^2(X)\{\widehat{m}(z) - \bar{m}(z)\}^2\{1 - D/\bar{\pi}_j(Z)\}^2] \\ &\leq \xi_{k,\infty}^2(1 + e^{-B_0})^2\mathbb{E}_n[\{\widehat{m}_j(Z) - \bar{m}_j(Z)\}^2] \\ &\leq \xi_{k,\infty}^2(1 + e^{-B_0})^2 C_0^2\|\widehat{\alpha}_j - \bar{\alpha}_j\|_1^2\end{aligned}$$

Again on the event $\bigcap_{m=1}^6 \Omega_{k,m}$ apply Online Appendix Lemma B.2 this is bounded, for some constant $\tilde{M}_3$ that does not depend on k by

$$\leq \tilde{M}_3\xi_{k,\infty}^2 s_k^2\bar{\lambda}_k^2 \tag{V.4}$$

The result (A.3) follows by collecting (V.1)-(V.4). □

A.2 PROOFS OF MAIN SECOND STAGE RESULTS

The proofs for Section 4 closely follow those of Belloni et al. (2015) with some modifications to deal with the various error terms. They also rely on some additional second stage results proved in Online Appendix C .

PROOF OF THEOREM 4.1

Equation (4.5) follows from applying (4.4) with $\alpha = p(x)/\|p(x)\|$ and (4.6) follows from (4.5). So it suffices to prove (4.4).

For any $\alpha \in S^{k-1}$, $1 \lesssim \|\alpha'\Omega^{1/2}\|$ because of the conditional variance of $\bar{\epsilon}_j^2$ is bounded from below and from above and under the positive semidefinite ranking

$$\Omega \geq \Omega_0 \geq \underline{\sigma}^2 Q^{-1}.$$

Moreover, by condition (ii) of the theorem and Lemma C.2, $R_{1n}(\alpha) = o_p(1)$. So we can write

$$\begin{aligned}\sqrt{n}\alpha'(\widehat{\beta} - \beta) &= \frac{\sqrt{n}\alpha'}{\|\alpha'\Omega^{1/2}\|}\mathbb{G}_n[p^k(x) \circ (\epsilon^k + r_k)] + o_p(1) \\ &= \sum_{i=1}^n \frac{\alpha'}{\sqrt{n}\|\alpha'\Omega^{1/2}\|}\{p^k(x) \circ (\epsilon^k + r_k)\}.\end{aligned}$$

Goal will be to verify Lindberg's condition for the CLT. Throughout the rest of the proof, it will be helpful to make the following notations. First, for any vector $a = (a_1, \ldots, a_k)' \in S^{k-1}$, let $|a| = (|a_1|, \ldots, |a_k|)'$ and note that $|a| \in S^{k-1}$ as well:

$$\tilde{\alpha}_n' = \frac{\alpha'}{\sqrt{n}\|\alpha_n'\Omega^{1/2}\|}, \quad \omega_n := |\tilde{\alpha}|'p^k(x), \text{ and } \bar{\epsilon}_k := \sup_{1\leq j\leq k}|\epsilon_j|$$

Now, by the definition of Ω we have that

$$\text{Var}\left(\sum_{i=1}^{n} \frac{\alpha'}{\sqrt{n}\|\alpha'\Omega^{1/2}\|}\{p^k(x) \circ (\epsilon^k + r_k)\}\right) = 1.$$

Second for each $\delta > 0$

$$\sum_{i=1}^{n} \mathbb{E}\left[(\tilde{\alpha}_n'\{p^k(x) \circ (\epsilon^k + r_k)\})^2 \mathbf{1}\left\{|\tilde{\alpha}_n'\{p^k(x) \circ (\epsilon^k + r_k)\}| > \delta\right\}\right]$$
$$\leq \sum_{i=1}^{n} \mathbb{E}\left[\omega_n^2 \mathbb{E}\left[\bar{\epsilon}_k^2 \mathbf{1}\{|\omega_n||\bar{\epsilon}_k + \ell_k c_k| > \delta\} \mid X = x\right]\right] \tag{A.4}$$

What we are using here is the following. Suppose α is a nonrandom vector in $\mathbb{R}^k$, a is a (positive) random vector in $\mathbb{R}^k$ and b is a random vector in $\mathbb{R}^k$. Then,

$$\{\alpha'(a \circ b)\} = \sum_{j=1}^{k} \alpha_j a_j b_j \leq \|b\|_\infty \sum_{j=1}^{k} |\alpha_j| a_j = (|\alpha|'a)\|b\|_\infty. \tag{A.5}$$

To bound the right hand side of (A.4) use the fact that $1 \lesssim \|\alpha'\Omega^{1/2}\|$ because $1 \lesssim \underline{\sigma}^2$ and

$$\Omega \geq \Omega_0 \geq \underline{\sigma}^2 Q^{-1}$$

in the positive semidefinite sense. Using these two we have

$$n\mathbb{E}|\omega_n|^2 \leq \mathbb{E}[(|\alpha|'p^k(x))^2]/(\alpha'\Omega\alpha) \lesssim 1.$$

By the bounded eigenvalue condition and using the trace operator:

$$\mathbb{E}[(|\alpha|p^k(x))^2] = \text{trace}(\mathbb{E}[|\alpha|'p^k(x)'p^k(x)|\alpha|]) = |\alpha|'Q|\alpha| \lesssim \|\alpha\| = 1$$

Further note, $|\omega_{ni}| \lesssim \frac{\xi_k}{\sqrt{n}}$. Using $(a+b)^2 \leq 2a^2 + 2b^2$, the right hand side of (A.4) is bounded by

$$2n\mathbb{E}[|\omega_n|^2\bar{\epsilon}_k^2 \mathbf{1}\{|\bar{\epsilon}_k| + \ell_k c_k > \delta/|\omega_n|\}] + 2n\mathbb{E}[|\omega_n|^2\ell_k^2 c_k^2 \mathbf{1}\{|\bar{\epsilon}_k| + \ell_k c_k > \delta/|\omega_n|\}]$$

and both terms converge to zero. Indeed, to bound the first term note that, for some $c > 0$:

$$2n\mathbb{E}[|\omega_n|^2\bar{\epsilon}_k^2 \mathbf{1}\{|\bar{\epsilon}_k| + \ell_k c_k > \delta/|\omega_n|\}] \lesssim n\mathbb{E}[|\omega_n|^2] \sup_{x \in \mathcal{X}} \mathbb{E}[\bar{\epsilon}_k^2 \mathbf{1}\{\bar{\epsilon}_k^2 + \ell_k c_k > c\delta\sqrt{n}/\xi_k\} \mid X = x]$$
$$= o(1)$$

where here we use the first part of Assumption 4.1(iv). To show the second term converges to zero, follow the same steps as for the first term, but apply the second part of Assumption 4.1(iv).

PROOF OF THEOREM 4.2

We apply Yurinskii's coupling lemma (Pollard, 2001)

Yurinskii's Coupling Lemma

Let $\xi_1, \dots, \xi_n$ be independent random k-vectors with $\mathbb{E}[\xi_i] = 0$ and $\beta := \sum_{i=1}^n \mathbb{E}[\|\xi_i\|^3]$ finite. Let $S := \xi_1 + \cdots + \xi_n$. For each $\delta > 0$ there exists a random vector T with a $N(0, \mathrm{var}(S))$ distribution such that

$$\mathbb{P}(|S - T| > 3\delta) \leq C_0 B \left(1 + \frac{|\log(1/B)|}{k}\right) \quad \text{where } B := \beta k \delta^{-3} \tag{YC}$$

for some universal constant C_0.

In order to apply the coupling, we want to consider a first order approximation to the estimator

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i, \quad \zeta_i = \Omega^{-1/2} p^k(x) \circ (\epsilon^k + r_k).$$

When $\bar{R}_{2n} = o_p(a_n^{-1})$ a similar argument can be used with $\zeta_i = \Omega^{-1/2} p^k(x) \circ (\epsilon^k + r_k)$ replaced with $\Omega^{-1/2} p^k(x) \circ \epsilon^k$. As before, the eigenvalues of Ω are bounded away from zero, therefore

$$\begin{aligned}
\mathbb{E}\|\zeta_i\|^3 &\lesssim \mathbb{E}[\|p^k(x) \circ (\epsilon^k(x) + r_k)\|^3] \\
&\lesssim \mathbb{E}[\|p^k(x)\|^3 (|\bar{\epsilon}_k|^3 + |r_k|^3)] \\
&\lesssim \mathbb{E}[\|p^k(x)\|^3](\bar{\sigma}_k^3 + \ell_k^3 c_k^3) \\
&\lesssim \mathbb{E}[\|p^k(x)\|^3]\xi_k(\bar{\sigma}_k^3 + \ell_k^3 c_k^3) \\
&\lesssim k\xi_k(\bar{\sigma}_k^3 + \ell_k^3 c_k^3)
\end{aligned}$$

Therefore, by Yurinskii's coupling lemma (YC), for each $\delta > 0$,

$$\begin{aligned}
\Pr\left\{\|\sum_{i=1}^n \zeta_i/\sqrt{n} - \mathcal{N}_k\| > 3\delta a_n^{-1}\right\} &\lesssim \frac{nk^2\xi_k(\bar{\sigma}_m^3 + \ell_k^3 c_k^3)}{(\delta a_n^{-1}\sqrt{n})^3}\left(1 + \frac{\log(k^3\xi_k(\bar{\sigma}_k^3 + \ell_k^3 c_k^3))}{k}\right) \\
&\lesssim \frac{a_n^3 k^2 \xi_k(\bar{\sigma}_k^3 + \ell_k^3 c_k^3)}{\delta^3 n^{1/2}}\left(1 + \frac{\log n}{k}\right) \to 0.
\end{aligned}$$

because $a_n^6 k^2 \xi_k(\bar{\sigma}_m^3 + \ell_k^3 c_k^3) \log^2 n/n \to 0$. Using the first two results from Lemma C.3, (C.6)-(C.7), we obtain that

$$\|\sqrt{n}\alpha(x)'(\widehat{\beta}^k - \beta^k) - \alpha(x)'\Omega^{1/2}\mathcal{N}_k\| \leq \|1/\sqrt{n}\sum_{i=1}^n \alpha(x)'\Omega^{1/2}\zeta_i - \alpha(x)'\Omega^{1/2}\mathcal{N}_k\| + \bar{R}_{1n} = o_p(a_n^{-1}).$$

uniformly over $x \in \mathcal{X}$. Since $\|\alpha(x)'\Omega^{1/2}\|$ is bounded from below uniformly over $x \in \mathcal{X}$ we obtain the first statetment of Theorem C.2 from which the second statement directly follows.

Finally, under the assumption that $\sup_{x\in\mathcal{X}} n^{1/2}|r(x)|/\|s(x)\| = o_p(a_n^{-1})$,

$$\frac{\sqrt{np(x)'(\widehat{\beta}^k - \beta^k)}}{\|s(x)\|} - \frac{\sqrt{n}(\widehat{g}(x) - g_0(x))}{\|s(x)\|} = o_p(a_n^{-1})$$

so that the third statement, (4.9) holds.

PROOF OF THEOREM 4.3

Preliminaries for Proof of Theorem 4.3

Lemma (Symmetrization). *Let $Z_1, \ldots, Z_n$ be independent stochastic processes with mean zero and let $\epsilon_1, \ldots, \epsilon_n$ be independent Rademacher random variables generated independetly of the data. Then*

$$\mathbb{E}^*\Phi\left(\frac{1}{2}\Big\|\sum_{i=1}^n \epsilon_i Z_i\Big\|_{\mathcal{F}}\right) \leq \mathbb{E}^*\Phi\left(\Big\|\sum_{i=1}^n Z_i\Big\|_{\mathcal{F}}\right) \leq \mathbb{E}^*\Phi\left(2\Big\|\epsilon_i(Z_i - \mu_i)\Big\|_{\mathcal{F}}\right), \tag{SI}$$

for every nondecreasing, convex $\Phi : \mathbb{R} \to \mathbb{R}$ and arbitrary functions $\mu_i : \mathcal{F} \to \mathbb{R}$.

For $p \geq 1$ consider the Shatten norm S_p on symmetrix $k \times k$ matrices Q defined by $\|Q\|_{S_p} = (\sum_{j=1}^k |\lambda_j(Q)|^p)^{1/p}$ where $\lambda_1(Q), \ldots, \lambda_k(Q)$ are the eigenvalues of Q. The case $p = \infty$ recovers the operator norm and $p = 2$ recovers the Frobenius norm.

Lemma (Khinchin's Inequality for Matrices). *For symmetric $k \times k$ matrices Q_i, $i = 1, \ldots, n$, $2 \leq p \leq \infty$, and an i.i.d sequence of Rademacher random variables $\epsilon_1, \ldots, \epsilon_n$ we have*

$$\left\|\left(\mathbb{E}_n[Q_i^2]\right)^{1/2}\right\|_{S_p} \leq \left(\mathbb{E}_\epsilon\|\mathbb{G}_n[\epsilon_i Q_i]\|_{S_p}^p\right)^{1/p} \leq C\sqrt{p}\left\|\left(\mathbb{E}_n[Q_i^2]\right)^{1/2}\right\|_{S_p} \tag{KI-1}$$

where C is an absolute constant. So, for $k \geq 2$,

$$\mathbb{E}_\epsilon[\|\mathbb{G}_n[\epsilon_i Q_i]\|] \leq C\sqrt{\log k}\|(\mathbb{E}_n[Q_i^2])^{1/2}\| \tag{KI-2}$$

for some (possibly different) absolute constant C.

We will establish consistent estimation of

$$\Sigma = \mathbb{E}[\{p^k(x) \circ (\epsilon^k + r_k)\}\{p^k(x) \circ (\epsilon^k + r_k)\}']$$

using

$$\widehat{\Sigma} = \mathbb{E}_n[\{p^k(x) \circ \widehat{\epsilon}^k\}\{p^k(x) \circ \widehat{\epsilon}^k\}']$$

Consistency of $\widehat{\Omega}$ will then follow from the consistency of $\widehat{Q}$ established by Lemma C.1. To

save notation, define the vectors

$$\widehat{Y} := \begin{bmatrix} Y(\widehat{\pi}_1, \widehat{m}_1) \\ \vdots \\ Y(\widehat{\pi}_k, \widehat{m}_k) \end{bmatrix} \quad \text{and} \quad \widehat{Y} := \begin{bmatrix} Y(\widehat{\pi}_1, \widehat{m}_1) \\ \vdots \\ Y(\widehat{\pi}_k, \widehat{m}_k) \end{bmatrix} \tag{A.6}$$

Also define $\dot{\epsilon}^k := (\dot{\epsilon}_1^k, \ldots, \dot{\epsilon}_k^k)$ so that $\dot{\epsilon}_j^k := Y(\bar{\pi}_j, \bar{m}_j) - \widehat{g}(x)$. Ideally, we would like to use $\dot{\epsilon}^k$ to estimate $\widehat{\Sigma}$, but we don't observe $\dot{\epsilon}^k$. Define $\Delta := \widehat{\epsilon}^k - \dot{\epsilon}^k = \widehat{Y}^k - \bar{Y}^k \in \mathbb{R}^k$.

Using this, we can decompose

$$\begin{aligned} \widehat{\Sigma} &= \mathbb{E}_n[\{p^k(x) \circ (\Delta + \dot{\epsilon}^k)\}\{p^k(x) \circ (\Delta + \dot{\epsilon}^k)\}] \\ &= \underbrace{\mathbb{E}_n[\{p^k(x) \circ \Delta\}\{p^k(x) \circ \Delta\}']}_{\Sigma_1} + \underbrace{\mathbb{E}_n[\{p^k(x) \circ \dot{\epsilon}^k\}\{p^k(x) \circ \Delta\}']}_{\Sigma_2} \\ &\quad + \underbrace{\mathbb{E}_n[\{p^k(x) \circ \Delta\}\{p^k(x) \circ \dot{\epsilon}^k\}']}_{\Sigma_3} + \underbrace{\mathbb{E}_n[\{p^k(x) \circ \dot{\epsilon}^k\}\{p^k(x) \circ \dot{\epsilon}^k\}]}_{\Sigma_4} \end{aligned} \tag{A.7}$$

We first show that $\|\Sigma_4 - \Sigma\| \to_p 0$. This is nonstandard because of the Hadamard product.

Lemma A.3 (Psuedo-Variance Estimator Consistency)**.** *Suppose Assumption 4.1 and Assumption 4.2 hold. Further, define* $v_n = \mathbb{E}[\max_{1 \le i \le n} |\bar{\epsilon}_k|^2]^{1/2}$. *In addition, assume that* $\bar{R}_{1n} + \bar{R}_{2n} \lesssim (\log k)^{1/2}$. *Then,*

$$\|\widehat{Q} - Q\| \lesssim_P \sqrt{\frac{\xi_k^2 \log k}{n}} = o(1)$$

$$\textit{and} \quad \|\Sigma_4 - \Sigma\| \lesssim_P (v_n \vee 1 + \ell_k c_k)\sqrt{\frac{\xi_k^2 \log k}{n}}$$

Proof. The first result is established by Lemma C.1 (Matrix LLN). Rest of proof will follow proof of Theorem 4.6 in Belloni et al. (2015). Like in (A.7) we can define $\dot{\Delta} \equiv \dot{\epsilon}^k - \epsilon^k = g_0(x) - \widehat{g}(x)$[1] and decompose

$$\begin{aligned} \Sigma_4 &= \underbrace{\mathbb{E}_n[p^k(x)p^k(x)'\dot{\Delta}^2]}_{\Sigma_{41}} + \underbrace{\mathbb{E}_n[\{p^k(x) \circ (\epsilon^k + r_k)\}\{p^k(x) \cdot \dot{\Delta}\}']}_{\Sigma_{42}} \\ &\quad + \underbrace{\mathbb{E}_n[\{p^k(x) \cdot \dot{\Delta}\}\{p^k(x) \circ (\epsilon^k + r_k)\}']}_{\Sigma_{43}} + \underbrace{\mathbb{E}_n[\{p^k(x) \circ (\epsilon^k + r_k)\}\{p^k(x) \circ (\epsilon^k + r_k)\}]}_{\Sigma_{44}} \end{aligned}$$

[1] It is useful to recall that $\dot{\epsilon}^k = \bar{Y}^k - \widehat{g}(x)$ and $\epsilon^k = \bar{Y}^k - g_0(x)$

The terms Σ_{41}, Σ_{42} and Σ_{43} are simple to show are negligible.

$$\begin{aligned}
&\|\Sigma_{41} + \Sigma_{42} + \Sigma_{43}\| \\
&\quad \leq \|\mathbb{E}_n[\{p^k(x)'(\widehat{\beta}^k - \beta^k)\}p^k(x)p^k(x)']\| + \|\mathbb{E}_n[\{p^k(x) \circ (\epsilon^k + r_k)\}p^k(x)'\{p^k(x)'(\widehat{\beta}^k - \beta^k)\}]\| \\
&\quad + \|\mathbb{E}_n[p^k(x)\{p^k(x)'(\widehat{\beta}^k - \beta^k)\}\{p^k(x) \circ (\epsilon^k + r_k)\}']\| \\
&\quad \leq \max_{1\leq i\leq n} |p^k(x)(\widehat{\beta}^k - \beta^k)|^2 \|\mathbb{E}_n[p^k(x)p^k(x)']\| \\
&\quad + 2 \max_{1\leq i\leq n} |\bar{\epsilon}_{k,i}| + |r_{k,i}| \max_{1\leq i\leq n} |p^k(x)'(\widehat{\beta} - \beta)| \|\mathbb{E}_n[p^k(x)p^k(x)']\|
\end{aligned}$$

By Theorem C.2 $|\max_{1\leq i\leq n} |p^k(x)'(\widehat{\beta}^k - \beta^k)| \lesssim_P \xi_k^2(\sqrt{\log k} + \bar{R}_{1n} + \bar{R}_{2n})^2/n$, by Assumption 4.1 the approximation error is bounded $\max_{1\leq i\leq n} |r_{k,i}| \leq \ell_k c_k$, by Assumption 4.2 and Markov's inequality the errors are bounded $\max_{1\leq i\leq n} |\bar{\epsilon}_{k,i}| \lesssim_p v_n^2$. Finally, by the first part of Lemma A.3 $\|\widehat{Q}\| \lesssim_P \|Q\| \lesssim 1$. Putting this all together with $\bar{R}_{1n} + \bar{R}_{2n} \lesssim (\log k)^{1/2}$ and $\xi_k^2 \log k/n \to 0$ gives

$$\|\Sigma_{41} + \Sigma_{42} + \Sigma_{43}\| \lesssim_P (v_n \vee 1 + \ell_k c_k)\sqrt{\frac{\xi_k^2 \log k}{n}}.$$

Next, we want to control $\Sigma_{44} - \Sigma$. To do this, let $\eta_1, \ldots, \eta_n$ be independent Rademacher random variables generated independently from the data. Then for $\eta = (\eta_1, \ldots, \eta_n)$

$$\begin{aligned}
&\mathbb{E}[\|\mathbb{E}_n[\{p^k(x) \circ (\epsilon^k + r_k)\}\{p^k(x) \circ (\epsilon^k + r_k)\}'] - \Sigma\|] \\
&\qquad \lesssim \mathbb{E}[\mathbb{E}_\eta[\mathbb{E}_n\|\eta\{p^k(x) \circ (\epsilon^k + r_k)\}\{p^k(x) \circ (\epsilon^k + r_k)\}'\|]] \\
&\qquad \lesssim \sqrt{\frac{\log k}{n}}\mathbb{E}[(\|\mathbb{E}_n[\|p^k(x)\|^2(\bar{\epsilon}_k + r_k)^2\{p^k(x) \circ (\epsilon^k + r_k)\}\{p^k(x) \circ (\epsilon^k + r_k)\}']\|)^{1/2}] \\
&\qquad \lesssim \sqrt{\frac{\xi_k^2 \log k}{n}}\mathbb{E}[\max_{1\leq i\leq n} |\bar{\epsilon}_{k,i} + r_k|(\|\mathbb{E}_n[\{p^k(x) \circ (\epsilon^k + r_k)\}\{p^k(x) \circ (\epsilon^k + r_k)\}']\|)^{1/2}] \\
&\qquad \lesssim \sqrt{\frac{\xi_k^2 \log k}{n}}(\mathbb{E}[\max_{1\leq i\leq n} |\bar{\epsilon}_{k,i} + r_k|^2])^{1/2} \times (\mathbb{E}[\|\mathbb{E}_n[\{p^k(x) \circ (\epsilon^k + r_k)\}\{p^k(x) \circ (\epsilon^k + r_k)\}']\|])^{1/2}
\end{aligned}$$

where the first inequality holds from Symmetrization (SI), the second from Khinchin's inequality (KI-1), the third by $\max_{1\leq i\leq n} \|p^k(x)\| \leq \xi_k$ and the fourth by Cauchy-Schwarz inequality.

Since for any positive numbers a, b and R, $a \leq R(a+b)^{1/2}$ implies $a \leq R^2 + R\sqrt{b}$, the expression above and the triangle inequality yields

$$\begin{aligned}
&\mathbb{E}[\|\mathbb{E}_n[\{p^k(x) \circ (\epsilon^k + r_k)\}\{p^k(x) \circ (\epsilon^k + r_k)\}'] - \Sigma\|] \\
&\qquad \lesssim \frac{\xi_k^2 \log k}{n}(v_n^2 + \ell_k^2 c_k^2) + \left(\frac{\xi_k^2 \log k}{n}\{v_n^2 + \ell_k^2 c_k^2\}\right)^{1/2} \|\Sigma\|^{1/2}
\end{aligned}$$

and so, because $\|\Sigma\| \lesssim 1$ and $(v_n^2 + \ell_k^2 c_k^2)\xi_k^2 \log k/n \to 0$ we have

$$\mathbb{E}[\|\mathbb{E}_n[\{p^k(x) \circ (\epsilon^k + r_k)\}\{p^k(x) \circ (\epsilon^k + r_k)\}'] - \Sigma\|] \lesssim (v_n \vee 1 + \ell_k c_k)\sqrt{\frac{\xi_k^2 \log k}{n}}.$$

The second result of Lemma A.3 follows from Markov's inequality. □

Now, we need to take care of the terms

$$\Sigma_1 = \mathbb{E}_n[\{p^k(x) \circ \Delta\}\{p^k(x) \circ \Delta\}']$$
$$\Sigma_2 = \mathbb{E}_n[\{p^k(x) \circ \dot{\epsilon}^k\}\{p^k(x) \circ \Delta\}']$$
$$\Sigma_3 = \mathbb{E}_n[\{p^k(x) \circ \Delta\}\{p^k(x) \circ \dot{\epsilon}^k\}']$$

where $\Delta = \widehat{Y}^k - \bar{Y}^k$ and $\dot{\epsilon}^k = \bar{Y}^k - \widehat{g}(x) = \widehat{g}(x) - g^k(x) + \epsilon^k$. To do so we will use Condition 2.

Lemma A.4 (Negligible Variance Bias)**.** *Suppose that Condition 2, Assumption 4.1 and Assumption 4.2 hold. Then*

$$\|\Sigma_1 + \Sigma_2 + \Sigma_3\| = o_p(1).$$

Proof. From Condition 2, the term Σ_1 being negligible immediately follows from Cauchy-Schwarz. Notice that

$$\begin{aligned}
\|\Sigma_1\| &\leq k \sup_{\substack{1\leq l\leq k\\ 1\leq j\leq k}} |\mathbb{E}_n[p_l(X)(Y(\hat{\pi}_l, \hat{m}_l) - Y(\bar{\pi}_j, \bar{m}_j))p_l(X)(Y(\widehat{\pi}_l, \widehat{m}_l) - Y(\bar{\pi}_l, \bar{m}_l))]| \\
&\leq k \sup_{1\leq l\leq k} (\mathbb{E}_n[p_j(X)^2(Y(\widehat{\pi}_j, \widehat{m}_j) - Y(\bar{\pi}_j, \bar{m}_j))^2])^{1/2} \sup_{1\leq j\leq k} (\mathbb{E}_n[p_j(X)^2(Y(\widehat{\pi}_j, \widehat{m}_j) - Y(\bar{\pi}_j, \bar{m}_j))^2])^{1/2} \\
&= o_p(1).
\end{aligned}$$

To see that Σ_2 is negligible notice that

$$\begin{aligned}
\|\Sigma_2\| &\leq k \sup_{\substack{1\leq l\leq k\\ 1\leq j\leq k}} \mathbb{E}_n[p_l(X)(\epsilon_l + p^k(x)'(\widehat{\beta}^k - \beta^k))p_j(X)(Y(\widehat{\pi}_j, \widehat{m}_j) - Y(\bar{\pi}_j, \bar{m}_j))] \\
&\leq k \sup_{1\leq l\leq k} \mathbb{E}_n[p_l(X)^2(\epsilon_l + p^k(x)'(\widehat{\beta} - \beta))^2]^{1/2}\mathbb{E}_n[p_j(X)^2(Y(\widehat{\pi}_j, \widehat{m}_j) - Y(\bar{\pi}_j, \bar{m}_j))^2]^{1/2} \\
&\leq \xi_{k,\infty}(\max_{1\leq i\leq n} |\bar{\epsilon}_k| + \max_{1\leq i\leq n} p^k(x)'(\widehat{\beta} - \beta))\mathbb{E}_n[p_j(X)^2(Y(\widehat{\pi}_j, \widehat{m}_j) - Y(\bar{\pi}_j, \bar{m}_j))^2]^{1/2}
\end{aligned}$$

Applying Assumption 4.2 and Theorem C.2 gives

$$\lesssim_P k\xi_{k,\infty}n^{1/m}\mathbb{E}_[p_j(X)^2Y(\widehat{\pi}_j, \widehat{m}_j) - Y(\bar{\pi}_j, \bar{m}_j))^2]^{1/2} = o_p(1)$$

where the final line is via Condition 2. Showing negligibility of Σ_3 follows the same steps. □

Proof of Theorem 4.4

Follows from the exact same steps as Theorem 3.5 in Semenova and Chernozhukov (2021) after establishing strong approximation by a gaussian process as in Theorem 4.2 and consistent variance estimation as in Theorem 4.3.

Online Appendix

B SUPPORTING LEMMAS FOR FIRST STAGE

Here we provide supporting lemmas and their proofs. We start off with non-asymptotic bounds for first stage parameters and means.

B.1 NONASYMPTOTIC BOUNDS FOR THE FIRST STAGE

The nonasymptotic bounds for the first stage will depend on certain events. In Appendix B.3 we will show that under Assumption 3.1 these events happen with probability approaching one. To control sparsity, define $\mathcal{S}_{\gamma,j} := \{j : \bar{\alpha}_j \neq 0\}$, $\mathcal{S}_{\alpha,j} := \{j : \bar{\alpha}_j \neq 0\}$. Recall $s_k := \max_{1\leq j\leq k}\{|\mathcal{S}_{\gamma,j}| \vee |\mathcal{S}_{\alpha,j}|\}$. Define the scores

$$\begin{aligned} S_{\gamma,j} &:= \mathbb{E}_n[U_{\gamma,j}Z] \\ S_{\alpha,j} &:= \mathbb{E}_n[U_{\alpha,j}Z] \end{aligned} \tag{B.1}$$

With these in mind, we will consider nonasymptotic bounds under the events:

$$\begin{aligned} \Omega_{k,1} &:= \{\lambda_{\gamma,j} \geq c_0 \cdot \|S_{\gamma,j}\|_\infty, \forall j \leq k\} \\ \Omega_{k,2} &:= \{\lambda_{\gamma,j} \leq \bar{\lambda}_k, \forall j \leq k\} \end{aligned} \tag{B.2}$$

Following Chetverikov and Sørensen (2021), the first event is referred to as "score domination" while the second event is referred to as "penalty majorization".

Bounds will be established on the ℓ_1 convergence rate of the estimated coefficient vector as well as on the symmetrized Bregman divergences, $D^\ddagger_{\gamma,j}(\widehat{\gamma}_j, \bar{\gamma}_j)$ and $D^\ddagger_{\alpha,j}(\widehat{\alpha}_j, \bar{\alpha}_j; \gamma_j)$, defined by

$$\begin{aligned} D^\ddagger_{\gamma,j}(\widehat{\gamma}_j, \bar{\gamma}_j) &:= \mathbb{E}_n\left[p_j(X)D\{e^{-\widehat{\gamma}_j'Z} - e^{-\bar{\gamma}_j'Z}\}\{\bar{\gamma}_j'Z - \widehat{\gamma}_j'Z\}\right], \\ D^\ddagger_{\alpha,j}(\widehat{\alpha}_j, \bar{\alpha}_j; \widehat{\gamma}) &:= \mathbb{E}_n\left[p_j(X)De^{-\widehat{\gamma}_j'Z}(\bar{\alpha}_j'Z - \widehat{\alpha}_j'Z)^2\right]. \end{aligned} \tag{B.3}$$

Lemma B.1 (Nonasymptotic Bounds for Logistic Model)**.** *Suppose that Assumption 3.1 holds with $\xi_0 > (c_0+1)/(c_0-1)$ and $2C_0\nu_0^{-2}s_k\bar{\lambda}_k \leq \eta < 1$. Then, under the events $\Omega_{k,1} \cap \Omega_{k,2}$ defined in (B.2), there exists a finite constant M_0 that does not depend on k such that*

$$\max_{1\leq j\leq k} D^\ddagger(\bar{g}, \widehat{g}) \leq M_0 s_k \bar{\lambda}_k^2 \text{ and } \max_{1\leq j\leq k} \|\widehat{\gamma}_j - \bar{\gamma}_j\|_1 \leq M_0 s_k \bar{\lambda}_k \tag{B.4}$$

Proof. We show that the bound of (B.4) holds for each $j = 1, \ldots, k$. For any $\gamma \in \mathbb{R}^d$ define $\tilde{\ell}_j(\gamma) := \mathbb{E}_n[p_j(X)\{De^{-\gamma'Z} + (1-D)\gamma'Z\}]$. By optimality of $\widehat{\gamma}_j$ we must have, for any $u \in (0,1]$:

$$\tilde{\ell}_j\left(\widehat{\gamma}_j\right) + \lambda_{\gamma,j}\|\widehat{\gamma}_j\|_1 \leq \tilde{\ell}\left((1-u)\widehat{\gamma}_j + u\bar{\gamma}_j\right) + \lambda_{\gamma,j}\|(1-u)\widehat{\gamma}_j + u\bar{\gamma}_j\|_1.$$

Using convexity of the ℓ_1 norm $\|\cdot\|_1$, this gives after rearrangment

$$\tilde{\ell}_j\left(\widehat{\gamma}_j\right) - \tilde{\ell}\left((1-u)\widehat{\gamma}_j + u\bar{\gamma}_j\right) + \lambda_{\gamma,j}u\|\widehat{\gamma}_j\|_1 \leq \lambda_{\gamma,j}u\|\bar{\gamma}_j\|_1.$$

Divide both sides by u and let $u \to^+ 0$

$$\mathbb{E}_n[p_j(X)D\{e^{-\widehat{\gamma}'Z} + (1-D)\}\{\widehat{\gamma}_j'Z - \bar{\gamma}_j'Z\}] + \lambda_{\gamma,j}\|\widehat{\gamma}_j\|_1 \leq \lambda_{\gamma,j}\|\bar{\gamma}_j\|_1.$$

By direct calculation, we have that $D^{\ddagger}_{\gamma,j}(\widehat{\gamma}_j, \bar{\gamma}_j)$ from (B.3) can be expressed

$$D^{\ddagger}_{\gamma,j}(\widehat{\gamma}_j, \bar{\gamma}_j) = \mathbb{E}_n[p_j(X)D\{e^{-\widehat{\gamma}'Z}+(1-D)\}\{\widehat{\gamma}_j'Z-\bar{\gamma}_j'Z\}]-\mathbb{E}_n[p_j(X)D\{e^{-\bar{\gamma}'Z}+(1-D)\}\{\widehat{\gamma}_j'Z-\bar{\gamma}_j'Z\}].$$

Combining the last two displays yields

$$D^{\ddagger}_{\gamma,j}(\widehat{\gamma}_j, \bar{\gamma}_j) + \mathbb{E}_n[p_j(X)D\{e^{-\bar{\gamma}'Z} + (1-D)\}\{\widehat{\gamma}_j'Z - \bar{\gamma}_j'Z\}] + \lambda_{\gamma,j}\|\widehat{\gamma}_j\|_1 \leq \lambda_{\gamma,j}\|\bar{\gamma}_j\|_1 \tag{L.1}$$

In the event $\Omega_{k,1}$ we have that

$$|\mathbb{E}_n[p_j(X)D\{e^{-\bar{\gamma}'Z} + (1-D)\{\widehat{\gamma}'Z - \bar{\gamma}'Z\}\}] \leq c_0^{-1}\lambda_{\gamma,j}\|\widehat{\gamma}_j - \bar{\gamma}_j\|_1 \tag{L.2}$$

Combining (L.1) and (L.2) yields

$$D^{\ddagger}_{\gamma,j}(\widehat{\gamma}_j, \bar{\gamma}_j) + \lambda_{\gamma,j}\|\widehat{\gamma}_j\|_1 \leq \lambda_{\gamma,j}\|\bar{\gamma}_j\| + c_0^{-1}\lambda_{\gamma,j}\|\widehat{\gamma}_j - \bar{\gamma}_j\|_1.$$

Expanding $\|\gamma_j\|_1 = \sum_{l\in\mathcal{S}_{\gamma,j}}|\gamma_l| + \sum_{l\notin\mathcal{S}_{\gamma,j}}|\gamma_l|$ for $\gamma = \widehat{\gamma}_j, \bar{\gamma}_j$ and applying the triangle inequalities $|\widehat{\gamma}_{j,l}| \geq |\bar{\gamma}_{j,l}| - |\widehat{\gamma}_{j,l} - \bar{\gamma}_{j,l}|$ for $l \in \mathcal{S}_{\gamma,j}$ and the equality $\widehat{\gamma}_{j,l} = \widehat{\gamma}_{j,l} - \bar{\gamma}_{j,l}$ gives

$$\begin{aligned} D^{\ddagger}_{\gamma,j}(\widehat{\gamma}_j, \bar{\gamma}_j) + \lambda_{\gamma,j}\Bigg\{\sum_{l\in\mathcal{S}_{\gamma,j}}|\bar{\gamma}_{j,l}| - \sum_{l\in\mathcal{S}_{\gamma,j}}|\widehat{\gamma}_{j,l} - \bar{\gamma}_{j,l}| + \sum_{j\notin\mathcal{S}_{\gamma,j}}|\widehat{\gamma}_{j,l} - \bar{\gamma}_{j,l}|\Bigg\} \\ \leq \lambda_{\gamma,j}\Bigg\{\sum_{l\in\mathcal{S}_{\gamma,j}}|\bar{\gamma}_{j,l}| + c_0^{-1}\sum_{l\in\mathcal{S}_{\gamma,j}}|\widehat{\gamma}_{j,l} - \bar{\gamma}_{j,l}| + c_0^{-1}\sum_{j\notin\mathcal{S}_{\gamma,j}}|\widehat{\gamma}_{j,l} - \bar{\gamma}_{j,l}|\Bigg\} \end{aligned}$$

Rearrange to get

$$D^{\ddagger}_{\gamma,j}(\widehat{\gamma}_j, \bar{\gamma}_j) + (1-c_0^{-1})\lambda_{\gamma,j}\sum_{l\notin\mathcal{S}_\beta}|\widehat{\gamma}_{j,l} - \bar{\gamma}_{j,l}| \leq (1+c_0)^{-1}\lambda_{\gamma,j}\sum_{l\in\mathcal{S}_{\gamma,j}}|\widehat{\gamma}_{j,l} - \bar{\gamma}_{j,l}|.$$

Adding $(1-c_0^{-1})\lambda_{\gamma,j}\sum_{l\in\mathcal{S}_{\gamma,j}}|\widehat{\gamma}_{j,l} - \bar{\gamma}_{j,l}|$ gives

$$D^{\ddagger}_{\gamma,j}(\widehat{\gamma}_j, \bar{\gamma}_j) + (1-c_0^{-1})\|\widehat{\gamma}_j - \bar{\gamma}_j\|_1 \leq 2\lambda_{\gamma,j}\sum_{l\in\mathcal{S}_{\gamma,j}}|\widehat{\gamma}_{j,l} - \bar{\gamma}_{j,l}| \tag{L.3}$$

By Lemma 4 in Appendix V.3 of Tan (2017) we have that for $\delta_j := \widehat{\gamma}_j - \bar{\gamma}_j$

$$D^{\ddagger}_{\gamma,j}(\widehat{\gamma}_j, \bar{\gamma}_j) \geq \frac{1-e^{-C_0\|\delta_j\|_1}}{C_0\|\widehat{\delta}_j\|}\left(\delta_j'\tilde{\Sigma}_{\gamma,j}\delta_j\right) \tag{L.4}$$

By (L.3) and $\xi_0 > (c_0+1)/(c_0-1)$ we have that $\sum_{l\notin\mathcal{S}_{\gamma,j}}|\delta_{j,l}| \leq \xi_0\sum_{l\in\mathcal{S}_{\gamma,j}}|\delta_{j,l}|$. Applying the

empirical compatability condition from Assumption 3.1 to (L.3) then yields

$$D^{\ddagger}_{\gamma,j}(\widehat{\gamma}_j, \bar{\gamma}_j) + (1 - c_0^{-1})\lambda_{\gamma,j}\|\delta_j\|_1 \leq 2\lambda_{\gamma,j}\nu_0^{-1}|\mathcal{S}_{\gamma,j}|^{1/2}(\delta_j'\tilde{\Sigma}_{\gamma,j}\delta_j)^{1/2} \tag{L.5}$$

Combining (L.4) and (L.5) to get an upper bound on $(\delta_j'\tilde{\Sigma}\delta_j)^{1/2}$ gives

$$\nu_0\|\delta_j\|_2 \leq (\delta_j'\tilde{\Sigma}_{\gamma,j}\delta_j)^{1/2} \leq 2\lambda_{\gamma,j}\nu_0^{-1}|\mathcal{S}_{\gamma,j}|^{1/2}\frac{C_0\|\delta_j\|_1}{1 - e^{-C_0\|\delta_j\|_1}}.$$

Plugging the second bound into (L.5) gives

$$D^{\ddagger}_{\gamma,j}(\widehat{\gamma}_j, \bar{\gamma}_j) + (1 - c_0^{-1})\lambda_{\gamma,j}\|\delta_j\|_1 \leq 2\lambda \sum_{l\in\mathcal{S}_{\gamma,j}} |\delta_{j,l}| \leq 4\lambda_{\gamma,j}^2\nu_0^{-2}|\mathcal{S}_{\gamma,j}|\frac{C_0\|\delta_j\|_1}{1 - e^{-C_0\|\delta_j\|_1}}.$$

The second inequality and $\sum_{l\notin\mathcal{S}_{\gamma,j}} |\delta_{j,l}| \leq \xi_0 \sum_{l\in\mathcal{S}_{\gamma,j}} |\delta_{j,l}|$ imply $1 - e^{-C_0\|\delta_j\|_1} \leq 2C_0\lambda_{\gamma,j}\nu_0^{-2}|\mathcal{S}_{\gamma,j}| \leq \eta$ so,

$$\frac{1 - e^{-C_0\|\delta_j\|_1}}{C_0\|\delta_j\|_1} = \int_0^1 e^{-C_0\|\delta_j\|_1 u}\, du \geq e^{-C_0\|\delta_j\|_1} \geq 1 - \eta.$$

Combining the last two displays gives

$$D^{\ddagger}_{\gamma,j}(\widehat{\gamma}_j, \bar{\gamma}_j) + (1 - c_0^{-1})\lambda_{\gamma,j}\|\widehat{\gamma}_j - \bar{\gamma}_j\|_1 \leq 4\lambda_{\gamma,j}^2\nu_0^{-2}(1 - \eta)|\mathcal{S}_{\gamma,j}| \tag{L.6}$$

Applying $\Omega_{k,2}$ to bound $\lambda_{\gamma,j} \leq \bar{\lambda}_k$ and noting that $|\mathcal{S}_{\gamma,j}| \leq s_k$ by definition gives (B.4) with $M_0 = \frac{4\nu_0^{-1}(1-\eta)}{1-c_0^{-1}}$. □

For each j, consider the matrices,

$$\begin{aligned}
\tilde{\Sigma}_{\alpha,j} &:= \mathbb{E}_n[p_j(X)De^{-\bar{\gamma}_j'Z}(Y - \bar{\alpha}_j'Z)^2ZZ'] \\
\tilde{\Sigma}_{\gamma,j} &:= \mathbb{E}_n[p_j(X)De^{-\bar{\gamma}_j'Z}ZZ']
\end{aligned} \tag{B.5}$$

In addition define $\Sigma_{\alpha,j} := \mathbb{E}\tilde{\Sigma}_{\alpha,j}$ and $\Sigma_{\gamma,j} := \mathbb{E}\tilde{\Sigma}_{\gamma,j}$. For the outcome regression model, we will consider nonasymptotic bounds under the following additional events:

$$\begin{aligned}
\Omega_{k,3} &:= \{\lambda_{\alpha,j} \geq c_0\|S_{\alpha,j}\|_\infty, \forall j \leq k\} \\
\Omega_{k,4} &:= \{\lambda_{\alpha,j} \leq \bar{\lambda}_k, \forall j \leq k\} \\
\Omega_{k,5} &:= \{\|\tilde{\Sigma}_{\alpha,j} - \Sigma_{\alpha,j}\|_\infty \leq \bar{\lambda}_k, \forall j \leq k\} \\
\Omega_{k,6} &:= \{\|\tilde{\Sigma}_{\gamma,j} - \Sigma_{\gamma,j}\|_\infty \leq \bar{\lambda}_k, \forall j \leq k\}
\end{aligned} \tag{B.6}$$

Lemma B.2 (Nonasymptotic Bounds for Linear Model). *Suppose that Assumption 3.1 holds, $\xi_0 > (c_0 + 1)/(c_0 - 1)$, and $2C_0\nu_0^{-2}s_k\bar{\lambda}_k \leq \eta < 1$. In addition, assume there is a constant $c > 0$ such that $\lambda_{\alpha,j}/\lambda_{\gamma,j} \geq c$ for all $j \leq k$. Then, under the event $\bigcap_{m=1}^6 \Omega_{k,m}$ there is a constant M_1 that does not depend on k such that*

$$\max_{1\leq j\leq k} D^{\ddagger}_{\alpha,j}(\widehat{\alpha}_j, \bar{\alpha}_j; \bar{\gamma}_j) \leq M_1 s_k\bar{\lambda}_k^2 \text{ and } \max_{1\leq j\leq k} \|\widehat{\alpha}_j - \bar{\alpha}_j\|_1 \leq M_1 s_k\bar{\lambda}_k \tag{B.7}$$

Proof. We show that the bound of (B.7) holds for each $j = 1, \ldots, k$. We proceed in a few steps.

Step 1: Optimization Step. Let $\tilde{\ell}_j(\alpha; \widehat{\gamma}_j) := \mathbb{E}_n[p_j(X)De^{-\widehat{\gamma}_j'Z}\{Y - \alpha'Z\}^2]/2$. Optimality of $\widehat{\alpha}_j$ implies that for any $u \in (0, 1]$:

$$\tilde{\ell}_j\left(\widehat{\alpha}_j; \widehat{\gamma}_j\right) - \tilde{\ell}_j\left((1-u)\widehat{\alpha}_j + u\bar{\alpha}_j; \widehat{\gamma}_j\right) + \lambda_{\alpha,j}\|\widehat{\alpha}_j\|_1 \leq \lambda_{\alpha,j}\|(1-u)\widehat{\alpha}_j + u\bar{\alpha}_j\|_1.$$

Convexity of the ℓ_1 norm $\|\cdot\|_1$ gives

$$\tilde{\ell}_j\left(\widehat{\alpha}_j; \widehat{\gamma}_j\right) - \tilde{\ell}_j\left((1-u)\widehat{\alpha}_j + u\bar{\alpha}_j; \widehat{\gamma}_j\right) + \lambda_{\alpha,j}u\|\widehat{\alpha}_j\|_1 \leq \lambda_{\alpha,j}u\|\bar{\alpha}_j\|_1.$$

Dividing both sides by u and letting $u \to 0^+$ gives:

$$-\mathbb{E}_n[p_j(X)De^{-\widehat{\gamma}_j'Z}\{Y - \widehat{\alpha}_j'Z\}\{\widehat{\alpha}_j'Z - \bar{\alpha}_j'Z\}] + \lambda_{\alpha,j}\|\widehat{\alpha}_j\|_1 \leq \lambda_{\alpha,j}\|\bar{\alpha}_j\|_1.$$

Rearranging using the form of $D^{\ddagger}_{\alpha,j}$ in (B.3) yields:

$$D^{\ddagger}_{\alpha,j}(\widehat{\alpha}_j, \bar{\alpha}_j; \widehat{\gamma}_j) + \lambda_{\alpha,j}\|\widehat{\alpha}_j\|_1 \leq (\widehat{\alpha}_j - \bar{\alpha}_j')\mathbb{E}_n[p_j(X)De^{-\widehat{\gamma}'Z}\{Y - \bar{\alpha}_j'Z\}Z] + \lambda_{\alpha,j}\|\bar{\alpha}_j\|_1 \tag{O.1}$$

Step 2: Quasi-Score Domination and relating $\bar{\gamma}_j$ to $\widehat{\gamma}_j$. For this step, we will use the fact that we are in the event $\Omega_{k,1} \cap \Omega_{k,2} \cap \Omega_{k,3} \cap \Omega_{k,5} \cap \Omega_{k,6}$. Using the expression for $D^{\ddagger}_{\gamma,j}(\widehat{\gamma}_j, \bar{\gamma}_j)$ from (B.3) we find that for some $u \in (0, 1)$:

$$\begin{aligned} D^{\ddagger}_{\gamma,j}(\widehat{\gamma}_j, \bar{\gamma}_j) &= -\mathbb{E}_n[p_j(X)D\{e^{-\widehat{\gamma}_j'Z} - e^{-\bar{\gamma}_j'Z}\}\{\widehat{\gamma}_j'Z - \bar{\gamma}_j'Z\}] \\ &= \mathbb{E}_n[p_j(X)De^{-u(\widehat{\gamma}_j - \bar{\gamma}_j)'Z}e^{-\bar{\gamma}_j'Z}\{\widehat{\gamma}_j'Z - \bar{\gamma}_j'Z\}^2] \end{aligned}$$

where the second step uses the mean value theorem:

$$e^{-\widehat{\gamma}_j'Z} - e^{-\bar{\gamma}_j'Z} = e^{-u\widehat{\gamma}_j'Z - (1-u)\bar{\gamma}_j'Z}(\widehat{\gamma}_j - \bar{\gamma}_j)'Z \tag{O.2}$$

In the event $\Omega_{k,1} \cap \Omega_{k,2}$ using the bound in Online Appendix Lemma B.1 and the fact that $C_0\nu_0^{-2}s_k\bar{\lambda}_k \leq \eta < 1$ gives us that

$$C_0\|\widehat{\gamma}_j - \bar{\gamma}_j\|_1 \leq C_0M_0s_k\bar{\lambda}_k \leq M_0\eta. \tag{O.3}$$

In the event $\Omega_{k,1} \cap \Omega_{k,2}$ the bound in (L.6) also gives us that $D^{\ddagger}_{\gamma,j}(\widehat{\gamma}_j, \bar{\gamma}_j) \leq M_0s_k\lambda^2_{\gamma,j}$. Combining the above displays then yields

$$M_0s_k\lambda^2_{\gamma,j} \geq D^{\ddagger}_{\gamma,j}(\widehat{\gamma}_j, \bar{\gamma}_j) \geq e^{M_0\eta}\mathbb{E}_n[p_j(X)De^{-\bar{\gamma}_j'Z}\{\widehat{\gamma}_j'Z - \bar{\gamma}_j'Z\}^2]. \tag{O.4}$$

Again applying the bound on $C_0\|\widehat{\gamma}_j - \bar{\gamma}_j\|_1$ (O.3) gives

$$\begin{aligned} D^{\ddagger}_{\alpha,j}(\widehat{\alpha}_j, \bar{\alpha}_j; \widehat{\gamma}_j) &= \mathbb{E}_n[p_j(X)De^{-\widehat{\gamma}_j'Z}(\widehat{\alpha}_j'Z - \bar{\alpha}_j'Z)^2] \\ &= \mathbb{E}_n[p_j(X)De^{-(\widehat{\gamma}_j-\bar{\gamma}_j)'Z}e^{-\bar{\gamma}_j'Z}(\widehat{\alpha}_j'Z - \bar{\alpha}_j'Z)^2] \\ &\geq e^{-M_0\eta}D^{\ddagger}_{\alpha,j}(\widehat{\alpha}_j, \bar{\alpha}_j; \bar{\gamma}_j) \end{aligned} \tag{O.5}$$

Decomposing the empirical expectation on the RHS of (O.1) gives

$$\begin{aligned} (\widehat{\alpha}_j - \bar{\alpha}_j)'\mathbb{E}_n[p_j(X)De^{-\widehat{\gamma}_j'Z}\{Y - \bar{\alpha}_j'Z\}Z] = &\underbrace{(\widehat{\alpha}_j - \bar{\alpha}_j)'\mathbb{E}_n[p_j(X)De^{-\bar{\gamma}_j'Z}\{Y - \bar{\alpha}_j'Z\}Z]}_{\delta_{1,j}} \\ &+ \underbrace{\mathbb{E}_n[p_j(X)D\{e^{-\widehat{\gamma}_j'Z} - e^{-\bar{\gamma}_j'Z}\}\{Y - \bar{\alpha}_j'Z\}\{\widehat{\alpha}_j'Z - \bar{\alpha}_j'Z\}]}_{\delta_{2,j}} \end{aligned}$$

By Hölder's inequality, in the event $\Omega_{k,3}$, $\delta_{1,j}$ is bounded

$$\delta_{1,j} \leq c_0^{-1}\|\widehat{\alpha}_j - \bar{\alpha}_j\|_1\lambda_{\alpha,j} \tag{O.6}$$

By the mean value equation (O.2) and the Cauchy-Schwarz inequality, $\delta_{2,j}$ can be bounded from above by

$$\begin{aligned} \delta_{2,j} \leq e^{C_0\|\widehat{\gamma}_j-\bar{\gamma}_j\|_1} &\times \mathbb{E}_n^{1/2}[p_j(X)De^{-\bar{\gamma}_j'Z}\{\widehat{\alpha}'Z - \bar{\alpha}'Z\}^2] \\ &\times \mathbb{E}_n^{1/2}[p_j(X)De^{-\bar{\gamma}_j'Z}\{Y - \bar{\alpha}_j'Z\}^2\{\widehat{\gamma}_j'Z - \bar{\gamma}_j'Z\}^2] \end{aligned} \tag{O.7}$$

Using (O.3) the first term in (O.7) can be bounded by $e^{M_0\eta}$. The second term is exactly the square root of $D^{\ddagger}_{\alpha,j}(\widehat{\alpha}_j, \bar{\alpha}_j; \bar{\gamma}_j)$. The third term is bounded in a few steps. First, in the event $\Omega_{k,5}$ we have that

$$(\mathbb{E}_n - \mathbb{E})[p_j(X)De^{-\bar{\gamma}_j'Z}\{Y - \bar{\alpha}_j'Z\}^2\{\widehat{\gamma}_j'Z - \bar{\gamma}_j'Z\}] \leq \bar{\lambda}_k\|\widehat{\gamma}_j - \bar{\gamma}_j\|_1^2.$$

By Assumption 3.1 and Lemma D.6 we have that $\mathbb{E}[D\{Y - \bar{\alpha}_j'Z\}^2] \leq G_0^2 + G_1^2$ so that:

$$\mathbb{E}[p_j(X)De^{-\bar{\gamma}_j'Z}\{Y - \bar{\alpha}_j'Z\}^2\{\widehat{\gamma}_j'Z - \bar{\gamma}_j'Z\}^2] \leq (G_0^2 + G_1^2)\mathbb{E}[p_j(X)De^{-\bar{\gamma}_j'Z}\{\widehat{\gamma}_j'Z - \bar{\gamma}_j'Z\}^2].$$

In the event $\Omega_{k,6}$ we have that

$$(\mathbb{E}_n - \mathbb{E})[p_j(X)De^{-\bar{\gamma}_j'Z}\{\widehat{\gamma}_j'Z - \bar{\gamma}_j'Z\}^2] \leq \bar{\lambda}_k\|\widehat{\gamma}_j - \bar{\gamma}_j\|_1.$$

and we can bound $\mathbb{E}_n[p_j(X)De^{-\bar{\gamma}_j'Z}\{\widehat{\gamma}_j'Z - \bar{\gamma}_j'Z\}^2]$ using (O.4). Putting this together gives

$$\begin{aligned}\mathbb{E}_n[p_j(X)De^{-\bar{\gamma}_j'Z}\{Y - \bar{\alpha}_j'Z\}^2\{\widehat{\gamma}_j'Z - \bar{\gamma}_j'Z\}^2] \leq \bar{\lambda}_k\|\widehat{\gamma}_j - \bar{\gamma}_j\|_1^2 \\ +(G_0^2 + G_1^2)\bar{\lambda}_k\|\widehat{\gamma}_j - \bar{\gamma}_j\|_1^2 \\ + (G_0^2 + G_1^2)e^{-M_0\eta}M_0 s_k\lambda_{\gamma,j}^2\end{aligned} \tag{O.8}$$

Applying convexity of $\sqrt{\cdot}$ and the bounds on $\|\widehat{\gamma}_j - \bar{\gamma}_j\|_1^2$ in the event $\Omega_{k,1} \cap \Omega_{k,2}$ from (L.6) gives

$$\begin{aligned}\delta_{2,j} &\leq \{e^{M_0\eta}(1 + (G_0^2 + G_1^2)^{1/2})(M_0\bar{\lambda}_k\lambda_{\gamma,j}s_k)^{1/2} + (G_0^2 + G_1)^2(M_0 s_k\lambda_{\gamma,j}^2)^{1/2}\}D_{\alpha,j}^{\ddagger}(\widehat{\alpha}_j, \bar{\alpha}_j; \bar{\gamma}_j)^{1/2} \\ &\leq \tilde{C}\{(\bar{\lambda}_k\lambda_{\gamma,j}s_k)^{1/2} + (s_k\lambda_{\gamma,j})^{1/2}\}D_{\alpha,j}^{\ddagger}(\widehat{\alpha}_j, \bar{\alpha}_j; \bar{\gamma}_j)^{1/2}\end{aligned} \tag{O.9}$$

where $\tilde{C} = \max\{e^{M_0\eta}M_0^{1/2}(1 + G_0 + G_1), (G_0^2 + G_1^2)M_0^{1/2}\}$. Combining (O.6) and (O.9) gives a bound on the empirical expectation on the RHS of (O.1).

$$\begin{aligned}(\widehat{\alpha}_j - \bar{\alpha}_j)'\mathbb{E}_n[p_j(X)De^{-\bar{\gamma}_j'Z}\{Y - \bar{\alpha}_j'Z\}Z] \leq \underbrace{c_0^{-1}\|\widehat{\alpha}_j - \bar{\alpha}_j\|_1\lambda_{\alpha,j}}_{\text{Bound on } \delta_{1,j} \text{ from (O.6)}} \\ + \underbrace{\tilde{C}\{(\bar{\lambda}_k\lambda_{\gamma,j}s_k)^{1/2} + (s_k\lambda_{\gamma,j}^2)^{1/2}\}D_{\alpha,j}^{\ddagger}(\widehat{\alpha}_j, \bar{\alpha}_j; \bar{\gamma}_j)^{1/2}}_{\text{Bound on } \delta_{2,j} \text{ from (O.9)}}\end{aligned} \tag{O.10}$$

For convenience, we will sometimes continue to refer to the bound on $\delta_{2,j}$ from (O.9) as simply $\delta_{2,j}$.

Step 3: Express Minimization Constraint in Terms of $\bar{\gamma}_j$ and Simplify. We use the results from *Step 2* to rewrite the minimization bound (O.1) from *Step 1*. Using (O.5) and (O.10) together with the minimization bound (O.1) yields

$$e^{-M_0\eta}D_{\alpha,j}^{\ddagger}(\widehat{\alpha}_j, \bar{\alpha}_j; \bar{\gamma}_j) + \lambda_{\alpha,j}\|\widehat{\alpha}_j\|_1 \leq c_0^{-1}\lambda_{\alpha,j}\|\widehat{\alpha}_j - \bar{\alpha}_j\|_1 + \lambda_{\alpha,j}\|\bar{\alpha}_j\|_1 + \delta_{2,j} \tag{O.11}$$

Apply the triangle inequality $|\widehat{\alpha}_{j,l}| \geq |\bar{\alpha}_{j,l}| - |\widehat{\alpha}_{j,l} - \bar{\alpha}_{j,l}|$ for $l \in \mathcal{S}_{\alpha,j}$ and $|\widehat{\alpha}_{j,l}| = |\widehat{\alpha}_{j,l} - \bar{\alpha}_{j,l}|$ for $l \notin \mathcal{S}_{\alpha,j}$ to the above to obtain

$$e^{-M_0\eta}D_{\alpha,j}^{\ddagger}(\widehat{\alpha}_j, \bar{\alpha}_j; \bar{\gamma}_j) + (1 - c_0^{-1})\|\widehat{\alpha}_j - \bar{\alpha}_j\|_1 \leq 2\lambda_{\alpha,j}\sum_{l \in \mathcal{S}_{\alpha,j}}|\widehat{\alpha}_{j,l} - \bar{\alpha}_{j,l}| + \delta_{2,j}.$$

Let $\delta_j = \widehat{\alpha}_j - \bar{\alpha}_j$. We use the form $D_{\alpha,j}^{\ddagger}(\widehat{\alpha}_j, \bar{\alpha}_j) = \mathbb{E}_n[p_j(X)De^{-\bar{\gamma}_j'Z}\{\widehat{\alpha}_j'Z - \bar{\alpha}_j'Z\}^2] = \delta_j'\tilde{\Sigma}_{\gamma,j}\delta_j$ to expand out

$$\begin{aligned}e^{-M_0\eta}(\delta_j'\tilde{\Sigma}_{\gamma,j}\delta_j) + (1 - c_0^{-1})\lambda_{\alpha,j}\|\delta\|_1 \leq 2\lambda_{\alpha,j}\sum_{l \in \mathcal{S}_{\alpha,j}}|\delta_{j,l}| \\ + \tilde{C}\{(s_k\bar{\lambda}_k\lambda_{\gamma,j})^{1/2} + (s_k\lambda_{\gamma,j})^{1/2}\}(\delta_j'\tilde{\Sigma}_{\gamma,j}\delta_j)^{1/2}\end{aligned} \tag{O.12}$$

Step 4: Apply Empirical Compatability Condition. Let $\delta_{3,j} := \tilde{C}\{(s_k\bar{\lambda}_k\lambda_{\gamma,j})^{1/2} + (s_k\lambda_{\gamma,j})^{1/2}\}$ and $D^{\star}_{\alpha,j} := e^{-M_0\eta}(\delta_j'\tilde{\Sigma}_{\gamma,j}\delta_j) + (1-c_0^{-1})\lambda_{\alpha,j}\|\delta_j\|_1$. In the even $\Omega_{k,1}\cap\Omega_{k,2}\cap\Omega_{k,3}\cap\Omega_{k,5}\cap\Omega_{k,6}$ that (O.12) holds, there are two possibilities. For $\xi_2 = 1 - 2c_0/\{(\xi_1+1)(c_0-1)\} \in (0,1]$ either

$$\xi_2 D^{\star}_{\alpha,j} \le \delta_{3,j}(\delta_j'\tilde{\Sigma}_{\gamma,j}\delta_j)^{1/2} \tag{O.13}$$

or $(1-\xi_2)D^{\star}_{\alpha,j} \le 2\lambda_{\alpha,j}\sum_{l\in\mathcal{S}_{\alpha,j}}|\delta_{j,l}|$, that is

$$D^{\star}_{\alpha_j} \le (\xi_1+1)(c_0-1)c_0^{-1}\lambda_{\alpha,j}\sum_{l\in\mathcal{S}_{\alpha,j}}|\delta_{j,l}| \tag{O.14}$$

We deal with these two cases separately. First, if (O.14) holds, then $\sum_{l\notin\mathcal{S}_{\alpha,j}}|\delta_{j,l}| \le \xi_1\sum_{l\in\mathcal{S}_{j,l}}|\delta_{j,l}|$. We can apply the empirical compatability of Assumption 3.1 to (O.14) to obtain.

$$e^{-M_0\eta}(\delta_j'\tilde{\Sigma}_{\gamma,j}\delta_j) + (1-c_0^{-1})\lambda_{\alpha,j}\|\delta_{j,l}\| \le \nu_1(\xi_1+1)(\xi_1-1)\lambda_{\alpha,j}(s_j\delta_j\tilde{\Sigma}_{\gamma,j}\delta_j)^{1/2}.$$

Inverting for $(\delta_j\tilde{\Sigma}_{\gamma,j}\delta_j)^{1/2}$ and plugging in gives

$$e^{-M_0\eta}D^{\ddagger}_{\alpha,j}(\widehat{\alpha},\bar{\alpha}_j;\bar{\gamma}_j) + (1-c_0^{-1})\lambda_{\alpha,j}\|\widehat{\alpha}_j-\bar{\alpha}_j\|_1 \le \tilde{M}s_k\lambda^2_{\alpha,j} \tag{O.15}$$

where $\tilde{M} = e^{M_0\eta}(\xi_1+1)(c_0-1)c_0^{-1}$. Next, assume that (O.13) holds. In this case, we can directly invert for $(\delta_j\tilde{\Sigma}_{\gamma,j}\delta_j)^{1/2}$ to get that

$$e^{-M_0\eta}D^{\ddagger}_{\alpha,j}(\widehat{\alpha}_j,\bar{\alpha}_j;\bar{\gamma}_j) + (1-c_0^{-1})\lambda_{\alpha,j}\|\widehat{\alpha}_j-\bar{\alpha}_j\|_1 \le \xi_2^{-1}\tilde{C}\{(s_k\bar{\lambda}_k\lambda_{\gamma,j})^{1/2} + (s_k\lambda^2_{\gamma,j})^{1/2}\}^2 \tag{O.16}$$

Combining (O.15) and (O.16) gives

$$\begin{aligned} e^{-M_0\eta}D^{\ddagger}_{\alpha,j}(\widehat{\alpha}_j,\bar{\alpha}_j;\bar{\gamma}_j) + (1-c_0^{-1})\lambda_{\alpha,j}\|\widehat{\alpha}_j-\bar{\alpha}_j\|_1 &\le \tilde{M}s_k\lambda^2_{\alpha,j} \\ &\quad + \xi_2^{-1}\tilde{C}\{(s_k\bar{\lambda}_k\lambda_{\gamma,j})^{1/2} + (s_k\lambda^2_{\gamma,j})^{1/2}\}^2 \end{aligned} \tag{O.17}$$

Step 5: Apply Penalty Majorization and Bounded Penalty Ratio. Use the fact that $\lambda_{\gamma,j}/\lambda_{\alpha,j} \le c^{-1}$ to express (O.17) as

$$\begin{aligned} D^{\ddagger}_{\alpha,j}(\widehat{\alpha}_j,\bar{\alpha}_j;\bar{\gamma}_j) &\le e^{M_0\eta}\tilde{M}s_k\lambda^2_{\alpha,j} + e^{M_0\eta}\xi_2^{-1}\tilde{C}\{(s_k\bar{\lambda}_k\lambda_{\gamma,j})^{1/2} + (s_k\lambda^2_{\gamma,j})^{1/2}\}^2 \\ \|\widehat{\alpha}_j-\bar{\alpha}_j\|_1 &\le (1-c_0^{-1})^{-1}\tilde{M}s_k\lambda_{\alpha,j} + (1-c_0^{-1})^{-1}c^{-1}\tilde{C}\{(s_k\bar{\lambda}_k)^{1/2} + (s_k\lambda_{\gamma,j})^{1/2}\}^2 \end{aligned}$$

In the event $\Omega_{k,2}\cap\Omega_{k,3}$ we have that $\lambda_{\gamma,j}\vee\lambda_{\alpha,j} \le \bar{\lambda}_k$, so that the above simplifies to

$$\begin{aligned} D^{\ddagger}_{\alpha,j}(\widehat{\alpha}_j,\bar{\alpha}_j;\bar{\gamma}_j) &\le M_1 s_k\bar{\lambda}_k^2 \\ \|\widehat{\alpha}_j-\bar{\alpha}_j\|_1 &\le M_1 s_k\bar{\lambda}_k \end{aligned} \tag{O.18}$$

for $M_1 = \max\{e^{M_0\eta}, c^{-1}(1-c_0^{-1})^{-1}\}(\tilde{M} + 2e^{M_0\eta}\xi_2^{-1}\tilde{C})$. This completes the result (B.7). □

B.2 NONASYMPTOTIC BOUNDS FOR RESIDUAL ESTIMATION

We now provide nonasymptotic bounds on the empirical mean square error between the estimated residuals $\widehat{U}_{\gamma,j}$ and $\widehat{U}_{\alpha,j}$ and the true residuals

$$\begin{aligned} U_{\gamma,j} &:= -p_j(X)\{De^{-\bar{\gamma}_j'Z} + (1-D)\} \\ U_{\alpha,j} &:= p_j(X)De^{-\bar{\gamma}_j'Z}(Y - \bar{\alpha}_j^{\text{pilot}'}Z), \end{aligned} \tag{B.8}$$

These bounds will be shown under the events in (B.2), (B.6), and (A.1) using the results in Lemmas B.1 and B.2.

Lemma B.3 (Nonasymptotic Logistic Residual Bound)**.** *Suppose that Assumption 3.1 and the conditions of Lemma B.1 hold. Then, in the event $\Omega_{k,1} \cap \Omega_{k,2}$ described on (B.2) there is a constant $M_{\gamma,r}$ that does not depend on k such that:*

$$\max_{1\le j\le k} \mathbb{E}_n[(\widehat{U}_{\gamma,j} - U_{\gamma,j})^2] \le M_{\gamma,r}\xi_{k,\infty}s_k\bar{\lambda}_k^2. \tag{B.9}$$

Proof. Consider each j separately. By applying the mean value theorem (O.2) and Lemma B.1, we can write

$$\begin{aligned} (\widehat{U}_{\gamma,j} - U_{\gamma,j})^2 &= p_j(X)^2D\{e^{-\widehat{\gamma}_j'Z} - e^{-\bar{\gamma}_j'Z}\}\{e^{-\widehat{\gamma}_j'Z} - e^{-\bar{\gamma}_j'Z}\} \\ &\le \xi_{k,\infty}p_j(X)D\{e^{-\widehat{\gamma}_j'Z} - e^{-\bar{\gamma}_j'Z}\}e^{-\bar{\gamma}_j'Z - u(\widehat{\gamma}_j-\bar{\gamma}_j)'Z}\{\bar{\gamma}_j'Z - \widehat{\gamma}_j'Z\} \\ &\le \xi_{k,\infty}e^{-B_0+M_0\eta}D\{e^{-\widehat{\gamma}_j'Z} - e^{-\bar{\gamma}_j'Z}\}\{\bar{\gamma}_j'Z - \widehat{\gamma}_j'Z\} \end{aligned}$$

So that

$$\begin{aligned} \mathbb{E}_n[(\widehat{U}_{\gamma,j} - U_{\gamma,j})^2] &\le e^{-B_0+M_0\eta}\xi_{k,\infty}\underbrace{\mathbb{E}_n[p_j(X)D\{e^{-\widehat{\gamma}_j'Z}\}\{\widehat{\gamma}_j'Z - \bar{\gamma}_j'Z\}]}_{=D^{\ddagger}_{\gamma,j}(\widehat{\gamma}_j,\bar{\gamma}_j)} \\ &\le e^{-B_0+M_0\eta}\xi_{k,\infty}s_k\bar{\lambda}_k^2 \end{aligned}$$

□

Lemma B.4 (Nonasymptotic Linear Residual Bound)**.** *Suppose that Assumption 3.1 and the conditions of Lemma B.2 hold. Then, in the event $\bigcap_{m=1}^{6}\Omega_{k,m}$, there is a constant $M_{\alpha,r}$ that does not depend on k such that*

$$\max_{1\le j\le k} \mathbb{E}_n[(\widehat{U}_{\alpha,j} - U_{\alpha,j})^2] \le M_{\alpha,r}\xi_{k,\infty}^2s_k^2\bar{\lambda}_k^2 \tag{B.10}$$

Proof. Recall that $\widehat{U}_{\alpha,j} = p_j(X)De^{-\widehat{\gamma}_j'Z}(Y - \widehat{\alpha}_j'Z)$ and $U_{\alpha,j} = p_j(X)De^{-\bar{\gamma}_j'Z}(Y - \bar{\alpha}_j'Z)$. As an intermediary, define $\dot{U}_{\gamma,j} = p_j(X)De^{-\widehat{\gamma}_j'Z}(Y - \bar{\alpha}_j'Z)$. We will show a bound on the empirical mean square error between $\widehat{U}_{\alpha,j}$ and $\dot{U}_{\alpha,j}$ as well as on the empirical mean square error between $\dot{U}_{\alpha,j}$ and $U_{\alpha,j}$. The bound in (B.10) will then follow from $(a+b)^2 \le 2a^2 + 2b^2$.

First consider $(\widehat{U}_{\alpha,j} - \dot{U}_{\alpha,j})^2$:

$$\begin{aligned}
\mathbb{E}_n[(\widehat{U}_{\alpha,j} - \bar{U}_{\alpha,j})^2] &= \mathbb{E}_n p_j^2(X) D e^{-2\widehat{\gamma}_j' Z}(\widehat{\alpha}_j' Z - \bar{\alpha}_j' Z)^2] \\
&= \mathbb{E}_n[p_j^2(X) D e^{-2(\bar{\gamma}_j' Z - (\widehat{\gamma}_j - \bar{\gamma}_j)' Z)}(\widehat{\alpha}_j' Z - \bar{\alpha}_j' Z)^2] \\
&\leq \xi_{k\infty} e^{-B_0} e^{2M_0\eta} \underbrace{\mathbb{E}_n[p_j(X) D e^{-\bar{\gamma}_j' Z}(\widehat{\alpha}_j' Z - \bar{\alpha}_j' Z)]}_{= D^{\ddagger}_{\alpha,j}(\widehat{\alpha}_j, \bar{\alpha}_j; \bar{\gamma}_j)} \\
&\leq e^{2M_0\eta - B_0} M_1 \xi_{k,\infty} s_k \bar{\lambda}_k^2
\end{aligned}$$

Where the last empirical expectation is bounded by Lemma B.2. Next, consider $(\dot{U}_{\alpha,j} - U_{\alpha,j})^2$:

$$\begin{aligned}
\mathbb{E}_n[(\dot{U}_{\alpha,j} - U_{\alpha,j})^2] &= \mathbb{E}_n[p_j^2(X) D\{e^{-\widehat{\gamma}' Z} - e^{-\bar{\gamma}' Z}\}^2\{Y - \bar{\alpha}_j' Z\}^2] \\
&= \mathbb{E}_n[p_j^2(X) D\{e^{-\bar{\gamma}' Z - u(\widehat{\gamma} - \bar{\gamma})' Z}(\bar{\gamma}_j' Z - \widehat{\gamma}_j' Z)\}^2 (Y - \bar{\alpha}_j' Z)^2] \\
&\leq 2 e^{M_0\eta - B_0} C_0^2 \xi_{k,\infty} (M_1 s_k \bar{\lambda}_k)^2 \mathbb{E}_n[p_j(X) D e^{-\bar{\gamma}_j' Z}(Y - \bar{\alpha}_j' Z)^2]
\end{aligned}$$

To proceed we assume that Z contains a constant. That is $Z = (1, Z_2, \ldots, Z_{d_z})$. However, this is not necessary it just simplifies the proof a bit. We bound the final empirical expectation in the event $\Omega_{k,5}$. In this event we can bound

$$\begin{aligned}
\mathbb{E}_n[p_j(X) D e^{-\bar{\gamma}_j' Z}(Y - \bar{\alpha}_j' Z)^2] &= (\mathbb{E}_n - \mathbb{E})[p_j(X) D e^{-\bar{\gamma}_j' Z}(Y - \bar{\alpha}_j' Z)^2] + \mathbb{E}[p_j(X) D e^{-\bar{\gamma}_j' Z}(Y - \bar{m}_j(X))^2] \\
&\leq \bar{\lambda}_k + \xi_{k,\infty} e^{-B_0}(D_0 + D_1)^2.
\end{aligned}$$

Combining the above, and using the fact that $s_k \bar{\lambda}_k \leq \eta < 1$ completes the reult.

□

B.3 PROBABILITY BOUNDS FOR THE FIRST STAGE

In this section we establish that each of the events in (B.2), (B.6), and (A.1) occurs under Assumption 3.1 with probability approaching one.

Lemma B.5 (Logistic Score Domination and Penalty Majorization)**.** *Suppose Assumption 3.1 holds and that the penalty parameter $\lambda_{\gamma,j}$ is chosen as described in Section 2. Then, for n sufficiently large, the event $\Omega_{k,1}$ holds with probability $1 - \epsilon - \rho_{\gamma,n}$ where*

$$\rho_{\gamma,n} = C \max\left\{ \frac{4kn + 4k}{n^2}, \left(\frac{\tilde{M}\xi_{k,\infty} s_{k,\gamma} \bar{c}_n^2 \ln^5(d_z n)}{n}\right)^{1/2}, \left(\frac{\tilde{M}\xi_{k,\infty}^4 \ln^7(d_z kn)}{n}\right)^{1/6}, \frac{1}{\ln^2(d_z kn)} \right\}. \tag{B.11}$$

where $C, \tilde{M}$ are absolute constants that do not depend on k. In particular so long as $\epsilon \to 0$ as $n \to \infty$, this shows that $\Pr(\Omega_{k,1}) = 1 - o(1)$ under the rate conditions of Assumption 3.1.

Moreover, with probability at least $1 - \frac{5k}{n} - \frac{4k}{n^2}$ there is a constant M_2 that does not depend on k such

that $\Omega_{k,2}$ holds with

$$\bar{\lambda}_k = \max\{M_2, M_4, M_5, M_6, M_7\}\xi_{k,\infty}\sqrt{\frac{\ln(d_z n)}{n}} \tag{B.12}$$

where M_4, M_5, M_6 and M_7 are all constants that also do not depend on k described in Lemma B.6 and Lemmas B.7-B.9. In particular, so long as $k/n \to 0$, $\Pr(\Omega_{k,2}) = 1 - o(1)$.

Proof. Collecting the logistic nonasymptotic residual bound from Lemma B.3 and the probability bounds from Lemmas B.7–B.10 we find that, (eventually) with probability at least $1 - \frac{4k}{n} - \frac{4k}{n^2}$:

$$\max_{\substack{1\le j\le k\\ 1\le l\le d_z}} \mathbb{E}_n[(\widehat{U}_{\gamma,j}Z_l - U_{\gamma,j}Z_l)^2] \le M_{\gamma,r}C_0^2\frac{\xi_{k,\infty}s_{k,\gamma}\bar{c}_n^2\ln^3(d_z n)}{n}. \tag{P.1}$$

where $M_{\gamma,r}$ is a constant that does not depend on k. Define the vectors

$$\begin{aligned} W_k &:= (U_{\gamma,1}Z', \ldots, U_{\gamma,k}Z')' \in \mathbb{R}^{kd_z} \\ &:= (W_{k,1}', \ldots, W_{k,k}')' \\ \widehat{W}_k &:= (\widehat{U}_{\gamma,1}Z', \ldots, \widehat{U}_{\gamma,k}Z')' \in \mathbb{R}^{kd_z} \\ &:= (\widehat{W}_{k,1}', \ldots, \widehat{W}_{k,k}')'. \end{aligned}$$

Notice by optimality of $\bar{\gamma}_1, \ldots, \bar{\gamma}_k$ that W_k is a mean zero vector. Under our assumptions the covariance matrix $\Sigma_k = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[W_k W_k']$ exists and is finite. Define the sequences of constants

$$\begin{aligned} \delta_{\gamma,n}^2 &:= M_{\gamma,r}C_0^2\xi_{k,\infty}s_{k,\gamma}\bar{c}_n^2\ln^5(d_z n)/n \\ \beta_{\gamma,n} &:= \frac{4k}{n} + \frac{4k}{n^2} \end{aligned}$$

Then, by (P.1) we have that with probability at least $1 - \beta_{\gamma,n}$

$$\Pr\left(\|\mathbb{E}_n[(\widehat{W}_k - W_k)^2]\|_\infty > \delta_n^2/\ln^2(d_z n)\right) \le \beta_n. \tag{P.2}$$

Let $e_1, \ldots, e_n$ be i.i.d normal random variables generated independently of the data. Define the scaled random variables and the multiplier bootstrap process

$$\begin{aligned} \widehat{S}_{\gamma,n}^e &:= n^{-1/2}\sum_{i=1}^n e_i\widehat{W}_{k,i} \\ &:= (\widehat{S}_{\gamma,1}^{e\prime}, \ldots, \widehat{S}_{\gamma,k}^{e\prime})' \end{aligned}$$

and let $\Pr_e$ denote the probability measure with respect to the $e_i's$ conditional on the observed data. Assumption 3.1 implies that the conditions of (D.1) hold for $Z = W_k$ with b replaced by c_u and B_n replaced by $B_k = (\xi_{k,\infty}C_0C_U)^3 \vee 1$. Further, via (P.2) the residual estimation requirement of with δ_n and β_n replaced by $\delta_{\gamma,n}$ and $\beta_{\gamma,n}$.

Let $\widehat{q}_{\gamma,j}(\alpha)$ be the α quantile of $\|\widehat{S}_{\gamma,j}^{e\prime}\|$ *conditional* on the data Z_i and the estimates $\widehat{Z}_i$. Theo-

rem D.4 then shows that there is a finite constant depending only on c_u such that

$$\max_{1\le j\le k}\sup_{\alpha\in(0,1)}\left|\Pr(\|S_{\gamma,j}\|\ge \widehat{q}_{\gamma,j}(\alpha))-\alpha\right| \le C\max\left\{\beta_{\gamma,n},\delta_{\gamma,n},\left(\frac{B_k^4\ln^7(kd_zn)}{n}\right)^{1/6},\frac{1}{\ln^2(kd_zn)}\right\}.$$

This gives the first claim of Lemma B.5 by construction of $\lambda_{\gamma,j}$. The second claim follows Lemma D.1. For this second claim we will consider the marginal convergence of each $U_{\gamma,j}Z$ as opposed to their joint convergence (the convergence of W_k). First, notice that condiitonal on the data, the random vector $\mathbb{E}_n[e\widehat{U}_{\gamma,j}Z]$ is centered gaussian in $\mathbb{R}^{d_z}$. Lemma D.1 then shows that

$$\widehat{q}_{\gamma,j}(\epsilon)\le(2+\sqrt{2})\sqrt{\frac{\ln(d_z/\epsilon)}{n}\max_{1\le l\le d_z}\mathbb{E}_n[\widehat{U}_{\gamma,j}^2Z_l^2]}.$$

Furthermore, with probability at least $1-\beta_{\gamma,n}-\frac{1}{n}$ we have that, for all $j=1,\dots,k$:

$$\max_{1\le l\le d_z}\mathbb{E}_n[\widehat{U}_{\gamma,j}^2Z_l^2]\le C_0^2\mathbb{E}_n[\widehat{U}_{\gamma,j}^2]\le 2C_0^2(\mathbb{E}_n[U_{\gamma,j}^2]+\mathbb{E}_n[(\widehat{U}_{\gamma,j}^2-U_{\gamma,j})^2])\le 4C_0^2\xi_{k,\infty}^2C_U^2+\delta_{\gamma,n}^2/\ln^2(d_zn))$$

Under the rate conditions of Assumption 3.1, $\delta_{\gamma,n}^2/\ln^2(d_zn)$ will eventually be smaller than 1 and so the claim in (B.12) holds with $M_2=8C_0^2C_U^2\vee 1$. □

Lemma B.6 (Linear Score Domination and Penalty Majorization). *Suppose Assumption 3.1 holds and that the penalty parameters $\lambda_{\gamma,j}$ and $\lambda_{\alpha,j}$ are chosen as described in Section 2. Then, for n sufficiently large, the event $\Omega_{k,3}$ holds with probability $1-\epsilon-\rho_{\alpha,n}$ where:*

$$\rho_{\alpha,n}=C\max\left\{\frac{4kn+4k}{n^2},\left(\frac{\tilde{M}\xi_{k,\infty}^2s_{k,\alpha}^2\bar{c}_n^2\ln^5(d_zn)}{n}\right)^{1/2},\left(\frac{\tilde{M}\xi_{k,\infty}^4\ln^7(d_zkn)}{n}\right)^{1/6},\frac{1}{\ln^2(d_zkn)}\right\}. \tag{B.13}$$

where $C,\tilde{M}$ are absolute constants that do not depend on k. In particular so long as $\epsilon\to 0$ as $n\to\infty$, this shows that $\Pr(\Omega_{k,3})=1-o(1)$ under Assumption 3.1.

Moreover, with probability at least $1-\frac{5k}{n}-\frac{4k}{n^2}$ there is a constant M_4 that does not depend on k such that $\Omega_{k,4}$ holds with

$$\bar{\lambda}_k=\max\{M_2,M_4,M_5,M_6,M_7\}\xi_{k,\infty}\sqrt{\frac{\ln(d_zn)}{n}} \tag{B.14}$$

where M_2,M_5,M_6 and M_7 are all constants that also do not depend on k described in Lemma B.5 and Lemmas B.7-B.9. In particular, so long as $k/n\to 0$, $\Pr(\Omega_{k,4})=1-o(1)$.

Proof. Apply the same steps as the proof of Lemma B.5 with

$$\delta_{\alpha,n}^2=M_{\alpha,r}C_0^2\xi_{k,\infty}^2s_k^2\bar{c}_n^2\ln^5(d_zn)/n$$
$$\beta_{\alpha,n}=\frac{4}{n}+\frac{4}{n^2}$$

□

Lemma B.7 (Probabilistic Bound on $\Omega_{k,5}$). *Let $\tilde{\Sigma}_{\alpha,j}$ and $\Sigma_{\alpha,j} = \mathbb{E}\tilde{\Sigma}_{\alpha,j}$ be as in (B.5). Under Assumption 3.1 if*

$$\bar{\lambda}_k \geq 4\xi_{k,\infty}(G_0^2 + G_0G_1)C_0^2\left[G_0^2\log(d_z/\epsilon)/n + G_0G_1\sqrt{\log(d_z/\epsilon)/n}\right]$$

Then $\Pr(\Omega_{k,5}) \geq 1 - 2k\epsilon^2$. In particular, there is a constant M_5 that does not depend on k, such that if $\bar{\lambda}_k \geq \xi_{k,\infty}M_5\sqrt{\log(d_z/\epsilon)/n}$ and $k\epsilon^2 \to 0$ as $n \to \infty$ then under the conditions of Assumption 3.1, $\Pr(\Omega_{k,5}) = 1 - o(1)$.

Proof. We show that this happens with probability $1 - 2\epsilon^2$ for each $j = 1, \ldots, k$. For any $l, h = 1, \ldots, d_z$, the variable

$$p_j(X)e^{-\bar{\gamma}'Z}D\{Y - \bar{m}_j(Z)\}^2Z_lZ_h$$

is the product of $p_k(X)e^{-\bar{\gamma}_j'Z}Z_lZ_h$, which is bounded in absolute value by $\xi_{k,\infty}C_0^2e^{-B_0}$, and $D\{Y - \bar{m}_j(Z)\}$, which is uniformly sub-gaussian conditional on Z. By Lemma D.7 we have:

$$\mathbb{E}\left[|(\tilde{\Sigma}_{\alpha,j})_{lh} - (\tilde{\Sigma}_{\alpha,j})_{lh}|^k\right] \leq \frac{k!}{2}(2\xi_{k,\infty}C_0^{-2}e^{-B_0}G_0^2)^{k-2}(2\xi_{k,\infty}C_0^2e^{-B_0}G_0G_1)^2, \quad k = 2, 3, \ldots.$$

Apply the above and Lemma D.5 with $t = \log(d_z^2/\epsilon^2)/n$ to obtain

$$\Pr\left(|(\tilde{\Sigma}_{\alpha,j})_{lh} - (\tilde{\Sigma}_{\alpha,j})_{lh}| > 2e^{-B_0}\xi_{k,\infty}C_0^2G_0^2t + 2e^{-B_0}\xi_{k,\infty}C_0^2G_0G_1\sqrt{2t}\right) \leq 2\epsilon^2/d_z^2.$$

A union bound completes the argument. □

Lemma B.8 (Probabilistic Bound on $\Omega_{k,6}$). *Let $\tilde{\Sigma}_{\gamma,j}$ and $\Sigma_{\gamma,j} = \mathbb{E}\tilde{\Sigma}_{\gamma,j}$ be as in (B.5). Under Assumption 3.1 if*

$$\bar{\lambda}_k \geq \xi_{k,\infty}\sqrt{2}(e^{-B_0} + 1)C_0\sqrt{\log(d_z/\epsilon)/n},$$

then $\Pr(\Omega_{k,6}) \leq 1 - 2k\epsilon^2$. In particular, there is a constant M_6 that does not depend on k, such that if $\bar{\lambda}_k \geq \xi_{k,\infty}M_6\sqrt{\log(d_z/\epsilon)/n}$ and $k\epsilon^2 \to 0$ as $n \to \infty$ then under the conditions of Assumption 3.1, $\Pr(\Omega_{k,6}) = 1 - o(1)$.

Proof. Consider each j separately. For any $l, h = 1, \ldots, d_z$, note $|(\tilde{\Sigma}_{\gamma,j})_{lh}| = |p_j(X)De^{-\bar{\gamma}_j'Z}Z_lZ_h| \leq \xi_{k,\infty}C_0^2e^{-B_0}$ so that $(\tilde{\Sigma}_{\gamma,j})_{lh} - (\Sigma_{\gamma,j})_{lh}$ is mean zero and bounded in abosulte values by $2\xi_{k,\infty}C_0^2e^{-B_0}$. Applying Lemma D.3 with $\bar{\lambda}_k \geq 4\xi_{k,\infty}C_0^2e^{-B_0}\sqrt{\log(d_z/\epsilon)/n}$ yields:

$$\Pr\left(|(\tilde{\Sigma}_{\gamma,j})_{lh} - (\Sigma_{\gamma,j})_{lh}| \geq \bar{\lambda}_k\right) \leq 2\epsilon^2/d_z^2.$$

A union bound completes the argument. □

Lemma B.9 (Probabilitstic Bound on $\Omega_{k,7}$). *Let $\tilde{\Sigma}^1_{\alpha,j}$ and $\Sigma^1_{\alpha,j} = \mathbb{E}\tilde{\Sigma}^1_{\alpha,j}$ be as in (A.1). Under Assumption 3.1 if*

$$\bar{\lambda}_k \geq \xi_{k\infty}4(G_0^2 + G_1^2)^{1/2}e^{-B_0}C_0^2\sqrt{\log(d_z/\epsilon)/n},$$

then $\Pr(\Omega_{k,7}) \geq 1 - 2k\epsilon^2$. *In particular, there is a constant* M_7 *that does not depend on* k *such that if* $\bar{\lambda}_k \geq \xi_{k,\infty} M_7 \sqrt{\log(d_z/\epsilon)/n}$ *and* $k\epsilon^2 \to 0$ *as* $n \to \infty$ *then, under the conditions of Assumption 3.1,* $\Pr(\Omega_{k,7}) \geq 1 - o(1)$.

Proof. We deal with each j term separately. The variables $p_j(X)e^{-\bar{\gamma}_j' Z}|Y - \bar{m}_j(Z)|Z_l Z_h$ are uniformly sub-gaussian conditional on Z because $|p_j(X)e^{-\bar{\gamma}_j' Z} Z_l Z_h| \leq \xi_{k,\infty} e^{-B_0} C_0^2$ and $D|Y - \bar{m}_j(Z)|$ is uniformly sub-gaussian. Applying Lemma D.4 for $\bar{\lambda}_k \geq e^{-B_0}\xi_{k,\infty} C_0^2 \sqrt{8(G_0^2 + G_1)^2}\sqrt{\log(d_z/\epsilon)/n}$ yields

$$\Pr\left(|(\tilde{\Sigma}_{\gamma,j})_{lh} - (\Sigma_{\gamma,j})_{lh}| \geq \bar{\lambda}_k\right) \leq 2\epsilon^2/d_z^2.$$

A union bound completes the argument. □

B.4 Probability Bounds for Residual Estimation

For showing consistent residual estimation, we employ the following two lemmas.

Lemma B.10 (Deterministic Logistic Score Domination)**.** *Under Assumption 3.1 let*

$$\bar{\lambda}_k \geq \xi_{k,\infty}\sqrt{2}(e^{-B_0} + 1)C_0\sqrt{\ln(d_z/\epsilon)/n}.$$

Then if for all $j = 1, \ldots, k$ *we let* $\lambda_{\gamma,j} \equiv \bar{\lambda}_k$, $\Pr(\Omega_{k,1} \cap \Omega_{k,2}) \geq 1 - 2k\epsilon$. *In particular, there is a constant* M_8^p *that does not depend on* k *such that if* $\bar{\lambda}_k \geq M_8^p \xi_{k,\infty}\sqrt{\ln(d_z n)/n}$ $\Pr(\Omega_{k,1} \cap \Omega_{k,2}) \geq 1 - 2k/n^p$.

Proof. Let us recall that

$$\|S_j\|_\infty = \max_{1 \leq l \leq d_z} |\mathbb{E}_n[p_j(X)\{-De^{-\bar{\gamma}_j' Z} + (1 - D)\}Z_j]|.$$

Notice for each $1 \leq l \leq d_z$, $S_{j,l} = p_j(X)\{-De^{-\bar{\gamma}_j' Z} + (1 - D)\}Z_l$ is bounded in absolute value by $C_0\xi_{k,\infty}(e^{-B_0} + 1)$ and is mean zero by optimality of $\bar{\gamma}_j$. For $\bar{\lambda}_k \geq 2(e^{-B_0} + 1)C_0\sqrt{\ln(d_z/\epsilon)/n}$ apply Lemma D.3 to see the result. □

Lemma B.11 (Deterministic Linear Score Domination)**.** *Under Assumption 3.1 let*

$$\bar{\lambda}_k \geq \xi_{k,\infty}(e^{-B_0}C_0)\sqrt{8(G_0^2 + G_1^2)}\sqrt{\ln(d_z/\epsilon)/n}.$$

Then if for all $j = 1, \ldots, k$ *we let* $\lambda_{\gamma,j} \equiv \bar{\lambda}_k$, $\Pr(\Omega_{k,3} \cap \Omega_{k,4}) \geq 1 - 2k\epsilon$. *In particular, there is a constant* M_9^p *that does not depend on* k *such that if* $\bar{\lambda}_k \geq M_9^p \xi_{k,\infty}\sqrt{\ln(d_z n)/n}$, $\Pr(\Omega_{k,3} \cap \Omega_{k,4}) \geq 1 - 2k/n^p$.

Proof. Notice $S_{j,l} = p_j(X)De^{-\bar{\gamma}_j' Z}\{Y - \bar{m}_j(Z)\}Z_l$ for $l = 1, \ldots, p$. By optimality of $\bar{\alpha}_j$, $S_{j,l}$ is mean zero. Under Assumption 3.1, $|S_{j,l}| \leq e^{-B_0}C_0|D\{Y - \bar{m}_j(Z)\}|$ so by Assumption 3.1 the variables $S_{j,l}$ are uniformly sub-gaussian conditional on Z in the following sense:

$$\max_{l=1,\ldots,p} \tilde{G}_0^2 \mathbb{E}[\exp(S_{j,l}^2/\tilde{G}_0^2) - 1] \leq \tilde{G}_1^2$$

for $\tilde{G}_0 = \xi_{k,\infty} C_0 G_0 e^{-B_0}$ and $\tilde{G}_1 = \xi_{k,\infty} C_0 G_1 e^{-B_0}$. Apply Lemma D.4 for $\bar{\lambda}_k$ defined above in the statement of Lemma B.11 and union bound to obtain the result. □

C Additional Second Stage Results

Theorem C.1 (Integrated Rate of Convergence). *Assume that Condition 1 and Assumption 4.1 hold. In addition suppose that $\xi_k^2 \log k/n \to 0$ and $c_k \to 0$. Then if either the propensity score our outcome regression model are correctly specified:*

$$\|\widehat{g}_k - g_0\|_{L,2} = (\mathbb{E}[(\widehat{g}(x) - g_0(x))^2])^{1/2} \lesssim_p \sqrt{k/n} + c_k \tag{C.1}$$

Proof. We begin with a matrix law of large numbers from Rudelson (1999), which is used to show $\widehat{Q} \to_p Q$.

Lemma C.1 (Rudelson's LLN for Matrices). *Let $Q_1, \ldots, Q_n$ be a sequence of independent, symmetric, non-negative $k \times k$ matrix valued random variables with $k \geq 2$ such that $Q = \mathbb{E}[\mathbb{E}_n Q_i]$ and $\|Q_i\| \leq M$ a.s. Then for $\widehat{Q} = \mathbb{E}_n[Q_i]$,*

$$\Delta := \mathbb{E}\|\widehat{Q} - Q\| \lesssim \frac{M \log k}{n} + \sqrt{\frac{M\|Q\| \log k}{n}}.$$

In particular if $Q_i = p_i p_i'$ with $\|p_i\| \leq \xi_k$ almost surely, then

$$\Delta := \mathbb{E}\|\widehat{Q} - Q\| \lesssim \frac{\xi_k^2 \log k}{n} + \sqrt{\frac{\xi_k^2 \|Q\| \log k}{n}}.$$

Now, to prove Theorem C.1 we have that:

$$\begin{aligned}\|\widehat{g}_k - g_0\|_{L,2} &\leq \|p^k(x)'\widehat{\beta}^k - p^k(x)'\beta^k\|_{L,2} + \|p^k(x)'\beta^k - g\|_{L,2} \\ &\leq \|p^k(x)'\widehat{\beta}^k - p^k(x)'\beta^k\|_{L,2} + c_k\end{aligned}$$

where under the normalization $Q = I_k$ we have that

$$\|p'\widehat{\beta} - p'\beta\|_{L,2} = \|\widehat{\beta} - \beta\|$$

Further,

$$\begin{aligned}\|\widehat{\beta}^k - \beta^k\| &= \|\widehat{Q}^{-1}\mathbb{E}[p^k(x) \circ (\widehat{Y} - \bar{Y})]\| + \|\widehat{Q}^{-1}\mathbb{E}_n[p^k(x) \circ (\epsilon^k + r_k)]\| \\ &\leq \|\widehat{Q}^{-1}\mathbb{E}[p^k(x) \circ (\widehat{Y} - \bar{Y})]\| + \|\widehat{Q}^{-1}\mathbb{E}_n[p^k(x) \circ \epsilon^k]\| + \|\widehat{Q}^{-1}\mathbb{E}_n[p^k(x) r_k]\|\end{aligned}$$

By the matrix LLN (Lemma C.1) we have that since $\xi_k^2 \log k/n \to 0$, $\|\widehat{Q} - Q\| \to_p 0$. This means that with probability approaching one all eigenvalues of $\widehat{Q}$ are boundedaway from zero, in particular they are larger than $1/2$. So w.p.a 1

$$\lesssim \|\mathbb{E}[p^k(x) \circ (\widehat{Y} - \bar{Y})]\| + \|\mathbb{E}_n[p^k(x) \circ \epsilon^k]\| + \|\mathbb{E}_n[p^k(x) r_k]\|$$

Under Condition 1 the first term is $o_p(\sqrt{k/n})$. By equation (A.48) in Belloni et al. (2015) the third term is bounded in probability by c_k. For the second term apply the third condition in Assumption 4.1 to see

$$\mathbb{E}\|\mathbb{E}_n[p^k(x) \circ \epsilon^k]\|^2 = \mathbb{E}\sum_{j=1}^{k} \epsilon_j^2 p_j(x)^2/n \leq \bar{\sigma}^2 \mathbb{E}_n[p^k(x)p^k(x)'/n] \lesssim \mathbb{E}[p^k(x)p^k(x)'/n] = k/n.$$

This gives $\|\mathbb{E}_n[p^k(x) \circ \epsilon^k]\| \lesssim_p \sqrt{k/n}$ and thus shows (C.1). □

Lemma C.2 (Pointwise Linearization)**.** *Suppose that Condition 1 and Assumption 4.1, hold. In addition assume that $\xi_k^2 \log k/n \to 0$. Then for any $\alpha \in S^{k-1}$,*

$$\sqrt{n}\alpha'(\widehat{\beta}^k - \beta^k) = \alpha'\mathbb{G}_n[p^k(x) \circ (\epsilon^k + r_k)] + R_{1n}(\alpha) \tag{C.2}$$

where the term $R_{1n}(\alpha)$, summarizing the impact of unknown design, obeys

$$R_{1n}(\alpha) \lesssim_p \sqrt{\frac{\xi_k^2 \log k}{n}}(1 + \sqrt{k}\ell_k c_k) \tag{C.3}$$

Moreover,

$$\sqrt{n}\alpha'(\widehat{\beta}^k - \beta^k) = \alpha'\mathbb{G}_n[p^k(x) \circ \epsilon^k] + R_{1n}(\alpha) + R_{2n}(\alpha) \tag{C.4}$$

where the term R_{2n}, summarizing the impact of approximation error on the sampling error of the estimator, obeys

$$R_{2n}(\alpha) \lesssim_p \ell_k c_k \tag{C.5}$$

Proof. Decompose as before,

$$\begin{aligned}\sqrt{n}\alpha'(\widehat{\beta}^k - \beta^k) = \sqrt{n}\alpha'\widehat{Q}^{-1}\mathbb{E}_n[p^k(x) \circ (\widehat{Y} - \bar{Y})] \\ + \alpha'\mathbb{G}_n[p^k(x) \circ (\epsilon^k + r_k)] \\ + \alpha'[\widehat{Q}^{-1} - I]\mathbb{G}_n[p^k(x) \circ (\epsilon^k + r_k)].\end{aligned}$$

The first term is $o_p(1)$ under Condition 1, we can just include this term in $R_{1n}(\alpha)$. Now bound $R_{1n}(\alpha)$ and $R_{2n}(\alpha)$.

Step 1. Conditional $X = [x_1, \ldots, x_n]$, the term

$$\alpha'[\widehat{Q}^{-1} - I]\mathbb{G}_n[p^k(x) \circ \epsilon^k].$$

has mean zero and variance bounded by $\bar{\sigma}^2\alpha'[\widehat{Q}^{-1} - I]\widehat{Q}^{-1}[\widehat{Q}^{-1} - I]\alpha$. Next, by Lemma C.1, with probability approaching one, all eigenvalues of $\widehat{Q}^{-1}$ are bounded from above and away zero. So,

$$\bar{\sigma}^2\alpha'[\widehat{Q}^{-1} - I_k]\widehat{Q}^{-1}[\widehat{Q}^{-1} - I_k]\alpha \lesssim \bar{\sigma}^2\|\widehat{Q}\|\|\widehat{Q}^{-1}\|^2\|\widehat{Q}^{-1} - I_k\|^2 \lesssim_p \frac{\xi_k^2 \log k}{n}.$$

so by Chebyshev's inequality,

$$\alpha'[\widehat{Q}^{-1} - I]\mathbb{G}_n[p^k(x) \circ \epsilon^k] \lesssim_p \sqrt{\frac{\xi_k^2 \log k}{n}}.$$

Step 2. From the proof of Lemma 4.1 in Belloni et al. (2015), we get that

$$\alpha'(\widehat{Q}^{-1} - I_k)\mathbb{G}_n[p^k(x)r_k] \lesssim_p \sqrt{\frac{\xi_k^2 \log k}{n}} \ell_k c_k \sqrt{k}$$

This completes the bound on $R_{1n}(\alpha)$ and gives (C.2)-(C.3). Next, also from the proof of Lemma 4.1 from Belloni et al. (2015),

$$R_{2n}(\alpha) = \alpha'\mathbb{G}_n[p^k(x)r_k] \lesssim_p \ell_k c_k,$$

which gives (C.4)-(C.5). □

The following lemma shows that, after adding Assumption 4.2 the linearization of our coefficient estimator $\widehat{\beta}^k$ established in Lemma C.2 holds uniformly over all points $x \in \mathcal{X}$. That is to say the error from linearization is bounded in probability uniformly over all $x \in \mathcal{X}$. It will form an important building block in uniform consistency and strong approximation results presented in Theorems C.2 and 4.2.

Lemma C.3 (Uniform Linearization)**.** *Suppose that Condition 1 and Assumption 4.1-4.2 hold. Then if either the propensity score model our outcome regression model is correctly specified:*

$$\sqrt{n}\alpha(x)'(\widehat{\beta}^k - \beta^k) = \alpha(x)'\mathbb{G}_n[p^k(x) \circ (\epsilon^k + r_k)] + R_{1n}(\alpha(x)) \tag{C.6}$$

where $R_{1n}(\alpha(x))$ describes the design error and satisfies

$$R_{1n}(\alpha(x)) \lesssim_p \sqrt{\frac{\xi_k^2 \log k}{n}}(n^{1/m}\sqrt{\log k} + \sqrt{k}\ell_k c_k) := \bar{R}_{1n} \tag{C.7}$$

uniformly over $x \in \mathcal{X}$. Moreover,

$$\sqrt{n}\alpha(x)'(\widehat{\beta}^k - \beta^k) = \alpha(x)'\mathbb{G}_n[p^k(x) \circ \epsilon^k] + R_{1n}(\alpha(x)) + R_{2n}(\alpha(x)) \tag{C.8}$$

where $R_{2n}(\alpha(x))$ describes the sampling error and satisfies, uniformly over $x \in \mathcal{X}$:

$$R_{2n}(\alpha(x)) \lesssim_P \sqrt{\log k} \cdot \ell_k c_k := \bar{R}_{2n} \tag{C.9}$$

Proof. As in the proof of Lemma C.2, we decompose

$$\begin{aligned}\sqrt{n}\alpha(x)'(\widehat{\beta}^k - \beta^k) = \sqrt{n}\alpha(x)'\widehat{Q}^{-1}\mathbb{E}_n[p^k(x) \circ (\widehat{Y} - \bar{Y})] \\ + \alpha(x)'\mathbb{G}_n[p^k(x) \circ (\epsilon^k + r_k)] \\ + \alpha(x)'[\widehat{Q}^{-1} - I]\mathbb{G}_n[p^k(x) \circ (\epsilon^k + r_k)].\end{aligned} \tag{C.10}$$

Using Condition 1, the matrix LLN (Lemma C.1), and bounded eigenvalues of the design matix, we have that:

$$\sup_{x\in\mathcal{X}} \sqrt{n}\alpha(x)'\widehat{Q}^{-1}\mathbb{E}_n[p^k(x)\circ(\widehat{Y}-\bar{Y})] = o_p(1).$$

Since this is $o_p(1)$, we can simply include this term in $R_{1n}(\alpha(x))$. Now derive bounds on $R_{1n}(\alpha(x))$ and $R_{2n}(\alpha(x))$.

Step 1: Conditional on the data let

$$T := \left\{t = (t_1,\ldots,t_n)\in\mathbb{R}^n : t_i = \alpha(x)'(\widehat{Q}^{-1}-I)p^k(x)\circ\epsilon^k, x\in\mathcal{X}\right\}.$$

Define the norm $\|\cdot\|_{n,2}$ on $\mathbb{R}^n$ by $\|t\|_{n,2}^2 = n^{-1}\sum_{i=1}^n t_i^2$. For an $\varepsilon > 0$ an ε-net of the normed space $(T, \|\cdot\|_{n,2})$ is a subset T_ε of T such that for every $t\in T$ there is a point $t_\varepsilon\in T_\varepsilon$ such that $\|t-t_\varepsilon\|_{n,2} < \varepsilon$. The covering number $N(T,\|\cdot\|_{n,2},\varepsilon)$ of T is the infimum of the cardinality of ε-nets of T.

Let $\eta_1,\ldots,\eta_n$ be independent Rademacher random variables that are independent of the data. Let $\eta = (\eta_1,\ldots,\eta_n)$. Let $\mathbb{E}_\eta[\cdot]$ denote the expectation with respect to the distribution of η. By Dudley's inequality (Dudley, 1967),

$$\mathbb{E}_\eta\left[\sup_{x\in\mathcal{X}}\left|\alpha(x)'[\widehat{Q}^{-1}-I]\mathbb{G}_n[\eta_i p^k(x)\circ\epsilon^k]\right|\right] \lesssim \int_0^\theta \sqrt{\log N(T,\|\cdot\|_{n,2},\varepsilon)}\,d\varepsilon.$$

where

$$\begin{aligned}\theta &:= 2\sup_{t\in T}\|t\|_{n,2}\\ &= 2\sup_{x\in\mathcal{X}}\left(\mathbb{E}_n[(\alpha(x)'(\widehat{Q}^{-1}-I)p^k(x)\circ\epsilon^k)^2]\right)^{1/2}\\ &\le 2\max_{1\le i\le n}|\bar{\epsilon}_{k,i}|\|\widehat{Q}^{-1}-I\|\|\widehat{Q}\|^{1/2},\end{aligned}$$

by (A.5). Now, for any $x\in\mathcal{X}$,

$$\begin{aligned}&\left(\mathbb{E}_n[(\alpha(x)'(\widehat{Q}^{-1}-I)p^k(x)\circ\epsilon^k - \alpha(\tilde{x})'(\widehat{Q}^{-1}-I)p^k(x)\circ\epsilon^k)^2]\right)^{1/2}\\ &\quad\le \max_{1\le i\le n}|\bar{\epsilon}_{k,i}|\|\alpha(x)-\alpha(\tilde{x})\|\|\widehat{Q}^{-1}-I\|\|\widehat{Q}\|^{1/2}\\ &\quad\le \xi_k^L\max_{1\le i\le n}|\bar{\epsilon}_{k,i}|\|\widehat{Q}^{-1}-I\|\|\widehat{Q}\|^{1/2}\|x-\tilde{x}\|\end{aligned}$$

So, for some $C > 0$,

$$N(T,\|\cdot\|_{n,2},\varepsilon) \le \left(\frac{C\xi_k^L\max_{1\le i\le n}|\bar{\epsilon}_{k,i}|\|\widehat{Q}^{-1}-I\|\|\widehat{Q}\|^{1/2}}{\varepsilon}\right)^{d_x}.$$

This gives us that

$$\int_0^\theta \sqrt{\log(N(T, \|\cdot\|_{2,n}, \varepsilon))}\, d\varepsilon \leq \max_{1\leq i\leq n} |\bar{\epsilon}_{k,i}| \|\widehat{Q}^{-1} - I\| \|\widehat{Q}\|^{1/2} \int_0^2 \sqrt{d_x \log(C\xi_k^L/\varepsilon)}\, d\varepsilon.$$

By Assumption 4.2 we have that $\mathbb{E}[\max_{1\leq i\leq n} |\bar{\epsilon}_{k,i}| \mid X] \lesssim_P n^{1/m}$ where $X = (x_1, \ldots, x_n)$. In addition $\xi_k^{2m/(m-2)} \log k/n \lesssim 1$ for $m > 2$ gives that $\xi_k^2/\log k/n \to 0$. So, $\|\widehat{Q}^{-1} - I\| \lesssim_P (\xi_k^2 \log k/n)^{1/2}$ and $\|\widehat{Q}^{-1}\| \lesssim_P 1$. Combining this all with $\log \xi_k^L \lesssim \log k$ implies

$$\begin{aligned}\mathbb{E}\left[\sup_{x\in\mathcal{X}} \left|\alpha(x)'[\widehat{Q}^{-1} - I]\mathbb{G}_n[p^k(x) \circ \epsilon^k]\right| \mid X\right] &\leq 2\mathbb{E}\left[\mathbb{E}_\eta \sup_{x\in\mathcal{X}} \left|\alpha(x)'[\widehat{Q}^{-1} - I]\mathbb{G}_n[\eta_i p^k(x) \circ \epsilon^k]\right| \mid X\right] \\ &\lesssim_P n^{1/m}\sqrt{\frac{\xi_k^2 \log^2 k}{n}}\end{aligned}$$

where the first line is due to symmetrization inequality. This gives us

$$\sup_{x\in\mathcal{X}} \left|\alpha(x)'[\widehat{Q}^{-1} - I]\mathbb{G}_n[p^k(x) \circ \epsilon^k]\right| \lesssim_p n^{1/m}\sqrt{\frac{\xi_k^2 \log^2 k}{n}} \tag{C.11}$$

Step 2: Now simply report the results on approximation error from Belloni et al. (2015) . Since the approximation error is the same for all signals $Y(\bar{\pi}_k, \bar{m}_k)$, there is no Hadamard product to deal with.

$$\sup_{x\in\mathcal{X}} \left|\alpha(x)'[\widehat{Q}^{-1} - I]\mathbb{G}_n[p^k(x) r_k]\right| \lesssim_P \sqrt{\frac{\xi_k^2 \log k}{n}} \ell_k c_k \sqrt{k} \tag{C.12}$$

$$\sup_{x\in\mathcal{X}} \left|\alpha(x)'\mathbb{G}_n[p^k(x) r_k]\right| \lesssim_P \ell_k c_k \sqrt{\log k} \tag{C.13}$$

Looking at (C.10) and combining (C.11)-(C.12) gives the bound on $R_{1n}(\alpha(x))$ while (C.13) gives the bound on $R_{2n}(\alpha(x))$. □

Theorem C.2 gives conditions under which our estimator converges in probability to the true conditional counterfactual outcome $g_0(x)$. In particular, this convergence happens uniformly at the rates defined in (C.15)-(C.16). If these two terms go to zero, the entire estimator will converge uniformly to the true conditional expectation of interest.

Theorem C.2 (Uniform Rate of Convergence)**.** *Suppose that Condition 1 and Assumptions 4.1-4.2 hold. Then so long as either the propensity score model or outcome regression model is correctly specified:*

$$\sup_{x\in\mathcal{X}} \left|\alpha(x)'\mathbb{G}_n[p^k(x) \circ \epsilon^k]\right| \lesssim_P \sqrt{\log k} \tag{C.14}$$

Moreover, for

$$\bar{R}_{1n} := \sqrt{\frac{\xi_k^2 \log k}{n}}(n^{1/m}\sqrt{\log k} + \sqrt{k}\ell_k c_k)$$
$$\bar{R}_{2n} := \sqrt{\log k} \cdot \ell_k c_k$$

we have that

$$\sup_{x\in\mathcal{X}} \left|p^k(x)'(\widehat{\beta}^k - \beta^k)\right| \lesssim_P \frac{\xi_k}{\sqrt{n}}\left(\sqrt{\log k} + \bar{R}_{1n} + \bar{R}_{2n}\right) \tag{C.15}$$

and

$$\sup_{x\in\mathcal{X}} \left|\widehat{g}(x) - g_0(x)\right| \lesssim_P \frac{\xi_k}{\sqrt{n}}\left(\sqrt{\log k} + \bar{R}_{1n} + \bar{R}_{2n}\right) + \ell_k c_k \tag{C.16}$$

Proof. The goal will be to apply the following two theorems from Giné and Koltchinskii (2006) and der Vaart and Wellner (1996).

Preliminaries for Proof of Theorem C.2

Theorem (Gine and Koltchinskii, 2006)**.** *Let $\xi_1, \ldots, \xi_n$ be i.i.d random variables taking values in a measurable space $(S, \mathcal{S})$ with a common distribution P defined on the underlying n-fold product space. Let $\mathcal{F}$ be a measurable class of functions mapping $S \to \mathbb{R}$ with a measurable envelope F. Let σ^2 be a constant such that $\sup_{f\in\mathcal{F}} \mathrm{Var}(f) \le \sigma^2 \le \|F\|^2_{L^2(P)}$. Suppose there exist constats $A > e^2$ and $V \ge 2$ such that $\sup_Q N(\mathcal{F}, L^2(Q), \varepsilon\|F\|_{L^2(Q)}) \le (A/\varepsilon)^V$ for all $0 < \varepsilon \le 1$. Then*

$$\mathbb{E}\left[\left\|\sum_{i=1}^n \{f(\xi_i) - \mathbb{E}[f(\xi_1)]\}\right\|_{\mathcal{F}}\right] \le C\left[\sqrt{n\sigma^2 V \log\frac{A\|F\|_{L^2(P)}}{\sigma}} + V\|F\|_\infty \log\frac{A\|F\|_{L^2(P)}}{\sigma}\right]. \tag{GK}$$

where C is a universal constant.

Theorem (VdV&W 2.14.1)**.** *Let $\mathcal{F}$ be a P-measurable class of measurable functions with a measurable envelope function F. Then for any $p \ge 1$,*

$$\left\|\|\mathbb{G}_n\|^*_{\mathcal{F}}\right\|_{P,p} \lesssim \|J(\theta_n, \mathcal{F})\|F\|_n\|_{P,p} \lesssim J(1, \mathcal{F})\|F\|_{P,2\vee p} \tag{VW}$$

where $\theta_n = \left\|\|f\|_n\right\|^*_{\mathcal{F}} / \|F\|_n$, where $\|\cdot\|_n$ is the $L_2(\mathbb{P}_n)$ seminorm and the inequalities are valid up to constants depending only on the p in the statement. The term $J(\cdot,\cdot)$ is given

$$J(\delta, \mathcal{F}) = \sup_Q \int_0^\delta \sqrt{1 + \log N(\mathcal{F}, \|\cdot\|_{L^2(Q)}, \varepsilon\|F\|_{L^2(Q)})}\, d\varepsilon.$$

We would like to apply these theorems to bound $\sup_{x\in\mathcal{X}} |\alpha(x)'\mathbb{G}_n[p^k(x) \circ \epsilon^k]|$ and thus show (C.14). The other two statements of Theorem C.2 follow from this. To this end, let's consider the class of functions

$$\mathcal{G} := \{(\epsilon^k, x) \mapsto \alpha(v)'(p^k(x) \circ \epsilon^k), v \in \mathcal{X}\}.$$

Let's note that $|\alpha(v)'p^k(x)| \leq \xi_k$, $\text{Var}(\alpha(v)'p^k(x)) = 1$, and for any $v, \tilde{v} \in \mathcal{X}$

$$|\alpha(v)'(p^k(x) \circ \epsilon^k) - \alpha(\tilde{v})'(p^k(x) \circ \epsilon^k)| \leq |\bar{\epsilon}_k| \xi_k^L \xi_k \|v - \tilde{v}\|,$$

where $\bar{\epsilon}_k = \|\epsilon^k\|_\infty$. Then, taking $G(\epsilon^k, x) \leq \bar{\epsilon}_k \xi_k$ we have that

$$\sup_Q N(\mathcal{G}, L^2(Q), \varepsilon \|G\|_{L^2(Q)}) \leq \left(\frac{C \xi_k^L}{\varepsilon} \right)^d. \tag{C.17}$$

Now, for a $\tau \geq 0$ specified later define $\epsilon_k^- = \epsilon^k \mathbf{1}\{|\bar{\epsilon}_k| \leq \tau\} - \mathbb{E}[\epsilon^k \mathbf{1}\{|\bar{\epsilon}_k| \leq \tau\} \mid X]$ and $\epsilon_k^+ = \epsilon^k \mathbf{1}\{|\bar{\epsilon}_k| > \tau\} - \mathbb{E}[\epsilon^k \mathbf{1}\{|\bar{\epsilon}_k| > \tau\} \mid X]$. Since $\mathbb{E}[\epsilon^k \mid X] = 0$ we have that $\epsilon^k = \epsilon_k^- + \epsilon_k^+$. Using this decompose:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \alpha(v)'(p^k(x) \circ \epsilon^k) = \sum_{i=1}^n \alpha(v)'(p^k(x) \circ \epsilon_k^-)/\sqrt{n} + \sum_{i=1}^n \alpha(v)'(p^k(x) \circ \epsilon_k^+)/\sqrt{n}.$$

We deal with each of these terms individually, in two steps.

Step 1: For the first term, we set up for an application of (GK). Equation (C.17) gives us the constants $A = C\xi_k^L$ and $V = d_x \vee 2$. To get σ^2 note that for any $v \in \mathcal{X}$,

$$\begin{aligned}
\text{Var}(\alpha(v)'(p^k(x) \circ \epsilon_k^-)/\sqrt{n}) &\leq \mathbb{E}[(\alpha(v)'(p^k(x) \circ \epsilon_k^-)/\sqrt{n})^2] \\
&\leq \frac{1}{n} \mathbb{E}[(\alpha(v)'p^k(x))^2] \sup_{x \in \mathcal{X}} \mathbb{E}[\|\epsilon_k^-\|_\infty^2 \mid X = x] \\
&\leq \frac{\bar{\sigma}_k^2 \wedge \tau^2}{n}
\end{aligned}$$

Finally note that we can take the envelope $G = \|\epsilon_k^-\|_\infty \xi_k/\sqrt{n}$ where $\|G\|_{L^2(P)} \leq \frac{\bar{\sigma}_k \wedge \tau}{\sqrt{n}}$ and $\|G\|_\infty \leq \tau \xi_k/\sqrt{n}$.

We can now apply (GK) to get that

$$\mathbb{E}[\sup_{x \in \mathcal{X}} |\alpha(x)'\mathbb{G}_n[p^k(x) \circ \epsilon_k^-]|] \lesssim \sqrt{\bar{\sigma}_k^2 \wedge \tau^2 \log(\xi_k^L)} + \frac{\tau \xi_k \log(\xi_k^L)}{\sqrt{n}}.$$

Step 2: For the second term, we set up for an application of (VW) with the envelope function $G = \|\epsilon_k^+\|_\infty \xi_k/\sqrt{n}$ and note that

$$\mathbb{E}[\|\epsilon_k^+\|_\infty^2] \leq \mathbb{E}[\bar{\epsilon}_k^2 \mathbf{1}\{|\bar{\epsilon}_k| > \tau\}] \leq \tau^{-m+2} \mathbb{E}[|\bar{\epsilon}_k|^m]$$

We can now use (VW) to bound

$$\begin{aligned}
\mathbb{E}\left[\left\|\sup_{x \in \mathcal{X}} |\alpha(x)'\mathbb{G}_n[p^k(x) \circ \epsilon_k^+]|\right\|\right] &\lesssim \sqrt{\mathbb{E}[|\bar{\epsilon}_k|^m]} \tau^{-m/2+1} \xi_k \int_0^1 \sqrt{\log(\xi_k^L/\varepsilon)}\, d\varepsilon \\
&\lesssim \sqrt{\sigma_k^m} \tau^{-m/2+1} \xi_k \sqrt{\log(\xi_k^L)}.
\end{aligned}$$

Step 3: Let $\tau = \xi_k^{2/(m-2)}$ and apply Markov's inequality. The bounds from step one and two become

$$\sup_{x\in\mathcal{X}} |\alpha(x)'\mathbb{G}_n[p^k(x)\circ\epsilon_k^-]| \lesssim_P \sqrt{\bar{\sigma}_k^2 \log(\xi_k^L)} + \frac{\xi_k^{2m/(m-2)}\log(\xi_k^L)}{\sqrt{n}}$$
$$\sup_{x\in\mathcal{X}} |\alpha(x)'\mathbb{G}_n[p^k(x)\circ\epsilon_k^+]| \lesssim_P \sqrt{\bar{\sigma}_k^m \log(\xi_k^L)}$$

Applying Assumption 4.2 along with the inequality

$$\frac{\xi_k^{m/(m-2)}\log k}{\sqrt{n}} = \sqrt{\log k}\sqrt{\frac{\xi_k^{2m/(m-2)}\log k}{n}} \lesssim \log k$$

completes the proof. □

Theorem C.3 (Validity of Gaussian Bootstrap)**.** *Suppose that the assumptions of Theorem 4.2 hold with $a_n = \log n$ and the assumptions of Theorem 4.3 hold with $a_n = O(n^{-b})$ for some $b > 0$. In addition, suppose that there exists a sequence ξ_n' obeying $1 \lesssim \xi_n' \lesssim \|p^k(x)\|$ uniformly for all $x \in \mathcal{X}$ such that $\|p^k(x) - p^k(x')\|/\xi_n' \le L_n\|x - x'\|$, where $\log L_n \lesssim \log n$. Let N_k^b be a bootstrap draw from $N(0, I_k)$ and $P^\star$ be the distribution conditional on the observed data $\{Y_i, D_i, Z_i\}_{i=1}^n$. Then the following approximation holds uniformly in $\ell^\infty(\mathcal{X})$:*

$$\frac{p^k(x)'\widehat{\Omega}^{1/2}}{\widehat{\Omega}^{1/2}p^k(x)} N_k^b =^d \frac{p^k(x)'\Omega^{1/2}}{\|\Omega^{1/2}p^k(x)\|} + o_{P^\star}(\log^{-1} N) \tag{C.18}$$

Proof. See Theorem 3.4 in Semenova and Chernozhukov (2021). □

D High Dimensional Probability Results

D.1 High Dimensional Central Limit and Bootstrap Theorems

Lemma D.1 (Gaussian Quantile Bound)**.** *Let $Y = (Y_1, \ldots, Y_p)$ be centered Gaussian in $\mathbb{R}^p$ with $\sigma^2 \le \max_{1\le j\le p}\mathbb{E}[Y_j^2]$ and $\rho \ge 2$. Let $q^Y(1-\epsilon)$ denote the $(1-\epsilon)$-quantile of $\|Y\|_\infty$ for $\epsilon \in (0,1)$. Then $q^Y(1-\epsilon) \le (2+\sqrt{2})\sigma\sqrt{\ln(p/\epsilon)}$.*

Proof. See Chetverikov and Sørensen (2021), Lemma D.2. □

Now let $Z_1, \ldots, Z_n$ be independent, mean zero random variables in $\mathbb{R}^p$, and denote their scaled average and variance by

$$S_n := \frac{1}{\sqrt{n}}\sum_{i=1}^n Z_i \text{ and } \Sigma := \frac{1}{n}\sum_{i=1}^n \mathbb{E}[Z_i Z_i'].$$

For $\mathbb{R}^p$ values random variables U and V, define the distributional measure of distance

$$\rho(U, V) := \sup_{A\in\mathcal{A}_p} \left|\Pr(U \in A) - \Pr(V \in A)\right|$$

where $\mathcal{A}_p$ denotes the collection of all hyperrectangles in $\mathbb{R}^p$. For any symmetric positive matrix $M \in \mathbb{R}^{p\times p}$, write $N_M := N(\mathbf{0}, M)$.

Theorem D.1 (High-Dimensional CLT)**.** *If, for some finite constants $b > 0$ and $B_n \geq 1$,*

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[Z_{ij}^2] \geq b,\ \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[|Z_{ij}|^{2+k}] \leq B_n^k \text{ and } \mathbb{E}\left[\max_{1\leq j\leq p} Z_{ij}^4\right] \leq B_n^4. \tag{D.1}$$

for all $i \in \{1,\ldots,n\}, j \in \{1,\ldots,p\}$ and $k \in \{1,2\}$, then there exists a finite constant C_b, depending only on b, such that:

$$\rho(S_n, N_\Sigma) \leq C_b\left(\frac{B_n^4 \ln^7(pn)}{n}\right)^{1/6}.$$

Proof. See Chernozhukov et al. (2017), Proposition 2.1. □

Let $\widehat{Z}_i$ be an estimator of Z_i and let $e_1,\ldots,e_n$ be i.i.d $N(0,1)$ and independent of both the Z_i's and $\widehat{Z}_i$'s. Define $\widehat{S}_n^e := \frac{1}{\sqrt{n}}\sum_{i=1}^n e_i\widehat{Z}_i$ and let $\Pr_e$ denote the conditional probability measure computed with respect to the $e_i's$ for fixed Z_i's and $\widehat{Z}_i$'s. Also abbreviate

$$\tilde{\rho}(\widehat{S}_n^e, N_\Sigma) := \sup_{A\in\mathcal{A}_p}\left|\Pr_e\left(\widehat{S}_n^e \in A\right) - \Pr\left(N_\Sigma \in A\right)\right|.$$

Theorem D.2 (Multiplier Bootstrap for Many Approximate Means)**.** *Let (D.1) hold for some finite constants $b > 0$ and $B_n \geq 1$, and let $\{\beta_n\}_\mathbb{N}$ and $\{\delta_n\}_\mathbb{N}$ be sequences in $\mathbb{R}_{++}$ converging to zero such that*

$$\Pr\left(\max_{1\leq j\leq p}\frac{1}{n}\sum_{i=1}^{n}(\widehat{Z}_{ij} - Z_{ij})^2 > \frac{\delta_n^2}{\ln^2(pn)}\right) \leq \beta_n \tag{D.2}$$

Then, there exists a finite constant C_b depending only on b such that with probability at least $1 - \beta_n - 1/\ln^2(pn)$,

$$\tilde{\rho}(\widehat{S}_n^e, N_\Sigma) \leq C_b \max\left\{\delta_n, \left(\frac{B_n \ln^6(pn)}{n}\right)^{1/6}\right\}.$$

Proof. See Belloni et al. (2018), Theorem 2.2 or Chetverikov and Sørensen (2021) Theorem D.2. □

We now consider a partition of Z and $\widehat{Z}$ into k subvectors.

$$Z := (Z_1',\ldots,Z_k')' \in \mathbb{R}^{d_1,\ldots,d_k} \text{ and } \widehat{Z} := (\widehat{Z}_1',\ldots,Z_k')' \in \mathbb{R}^{d_1,\ldots,d_k}$$

where $\sum_{j=1}^k d_j = p$. Given such a partition, for any symmetric, positive definite $M \in \mathbb{R}^{p\times p}$ let $N_{M,j}$ denote the marginal distribution of the subvector of N_M corresponding the the indices of partition j. In other words, N_{M_1} would denote the marginal distribution of the first d_1 elements of an $\mathbb{R}^p$ vector with distribution N_M, N_2 would denote the marginal distribution of the next

d_2 elements and so on. For each $j = 1, \ldots, k$ define $q^N_{M,j} : \mathbb{R} \to \bar{\mathbb{R}}$ as the (extended) quantile function of $\|N_{M,j}\|_\infty$,

$$q^N_{M,j}(\epsilon) := \inf\left\{t \in \mathbb{R} : \Pr(\|N_{M,j}\|_\infty \leq t) \geq \epsilon\right\}.$$

Define $q^N_{M,j}(\epsilon) = +\infty$ if $\epsilon \geq 1$ and $-\infty$ if $\epsilon \leq 0$ so that $q^N_{M,j}$ is always montone (strictly) increasing.

Lemma D.2. *Let $M \in \mathbb{R}^{p\times p}$ be symmetric positive definite, let U be a random variable in $\mathbb{R}^p$. Partition U into k subvectors, $U = (U_1', \ldots, U_k')' \in \mathbb{R}^{d_1,\ldots,d_k}$ where $d_1 + \cdots + d_k = p$. For each $j = 1, .., k$ let q_j denote the quantile function of $\|U_j\|_\infty$. Then for any $j = 1, \ldots, k$,*

$$q^N_{M,j}(\epsilon - 2\rho(U, N_M)) \leq q_j(\epsilon) \leq q^N_{M,j}(\epsilon + \rho(U, N_M)) \text{ for all } \epsilon \in (0, 1).$$

Proof. Proof is a slight modification of that of Lemma D.3 in Chetverikov and Sørensen (2021). Main idea is to add and substract a $\|N_M\|_\infty$ term and use the fact that the approximation is achieved over all hyperrectangles. We show the bound holds for each $j = 1, \ldots, k$. Without loss of generality, consider U_1. Let $N_{M,1}$ denote the maginal distribution of the first d_1 elements of a $\mathbb{R}^p$ vector with distribution N_M.

$$\begin{aligned}
\Pr(\|U_1\|_\infty \leq t) &= \Pr(\|N_{M,1}\|_\infty \leq t) + \Pr(\|U_1\|_\infty \leq t) - \Pr(\|N_{M,1}\|_\infty \leq t) \\
&= \Pr(\|N_{M,1}\|_\infty \leq t) + \left(\Pr(U \in [-t, t]^p \times \mathbb{R}^{p-d_1}) - \Pr(N_M \in [-t, t]^p \times \mathbb{R}^{p-d_1})\right) \\
&\leq \Pr(\|N_{M,1}\|_\infty \leq t) + \rho(U, N_M)
\end{aligned}$$

for any $t \in \mathbb{R}$. A similar construction will give that

$$\Pr(\|U_1\|_\infty \leq t) \geq \Pr(\|N_{M,1}\|_\infty \leq t) - \rho(U, N_M).$$

Substituting $t = q^N_{M,1}(\epsilon - 2\rho(U, N_M))$ into the upper bound on $\Pr(\|U_1\|_\infty \leq t)$ gives the lower bound statement, while $t = q^N_{M,1}(\epsilon + \rho(U, N_M))$ and using the lower bound on $\Pr(\|U_1\|_\infty \leq t)$ gives the upper bound statement. □

As with Z partition S_n and $\widehat{S}^e_n$ into

$$S_n = (S'_{n,1}, \ldots, S'_{n,k})' \in \mathbb{R}^{d_1,\ldots,d_k} \text{ and } \widehat{S}^e_n = (\widehat{S}^{e'}_{n,1}, \ldots, \widehat{S}^{e'}_{n,k})' \in \mathbb{R}^{d_1,\ldots,d_k}.$$

For each $j = 1, \ldots, k$ define $q_{n,j}(\epsilon)$ as the ϵ-quantile of $\|S_{n,j}\|_\infty$

$$q_{n,j}(\epsilon) := \inf\{t \in \mathbb{R} : \Pr(\|S_{n,j}\|_\infty \leq t) \geq \epsilon\} \text{ for } \epsilon \in (0, 1).$$

Let $\widehat{q}_{n,j}(\epsilon)$ be the ϵ-quantile of $\|\widehat{S}^e_{n,j}\|_\infty$, computed conditionally on X_i and $\widehat{X}_i$'s,

$$\widehat{q}_{n,j}(\epsilon) := \inf\{t \in \mathbb{R} : \Pr_e(\|\widehat{S}^e_{n,j}\|_\infty \leq t) \geq \epsilon\} \text{ for } \epsilon \in (0, 1).$$

Theorem D.3 (Quantile Comparasion)**.** *If* (D.1) *holds for some finite constants $b > 0$ and $B_n \geq 1$,*

and

$$\rho_n := 2C_b \left(\frac{B_n^4 \ln^7(pn)}{n} \right)^{1/6}$$

denotes the upper bound in Theorem D.1 multiplied by two, then for all $j = 1, \dots, k$

$$q^N_{\Sigma,j}(1 - \epsilon - \rho_n) \leq q_{n,j}(1 - \epsilon) \leq q^N_{\Sigma,j}(1 - \epsilon + \rho_n) \text{ for all } \epsilon \in (0, 1).$$

If, in addition, (D.2) holds for some sequences $\{\delta_n\}_{\mathbb{N}}$ *and* $\{\beta_n\}_{\mathbb{N}}$ *converging to zero, and*

$$\rho'_n \leq 2C'_b \max \left\{ \delta, \left(\frac{B_n^4 \ln^6(pn)}{n} \right)^{1/6} \right\}$$

denotes the upper bound in Theorem D.2 multiplied by two, then with probability at least $1 - \beta_n - 1/\ln^2(pn)$ *we have for all* $j = 1, \dots, k$,

$$q^N_{\Sigma,j}(1 - \epsilon - \rho'_n) \leq \widehat{q}_{n,j}(1 - \epsilon) \leq q^N_{\Sigma,j}(1 - \epsilon + \rho'_n) \text{ for all } \epsilon \in (0, 1).$$

Proof. From Lemma D.2 with $U = S_n$ we obtain

$$q^N_{\Sigma,j}(1 - \epsilon - 2\rho(S_n, N_\Sigma)) \leq q_{n,j}(1 - \epsilon) \leq q^N_{\Sigma,j}(1 - \epsilon + \rho(S_n, N_\Sigma)).$$

The first chain of inequalities then follows from $2\rho(S_n, N_\Sigma) \leq \rho_n$ by Theorem D.1.

For the second claim, apply Lemma D.2 with $U = \widehat{S}^e_n$ and condition on the Z_i's and $\widehat{Z}_i$'s obtain

$$q^N_{\Sigma,j}(1 - \epsilon - 2\tilde{\rho}(\widehat{S}^e_n, N_\Sigma)) \leq \widehat{q}_n(1 - \epsilon) \leq q^N_{\Sigma,j}(1 - \epsilon + \tilde{\rho}(\widehat{S}^e_n, N_\Sigma)).$$

The second chain of inequalities then follows on the event $2\tilde{\rho}(\widehat{S}^e_n, N_\Sigma) \leq \rho'_n$, which by Theorem D.2 happens with probability at least $1 - \beta_n - 1/\ln^2(pn)$. □

Theorem D.4 (Multiplier Bootstrap Consistency)**.** *Let (D.1) and (D.2) hold for some constants* $b > 0$ *and* $B_n \geq 1$ *and some sequences* $\{\delta_n\}_{\mathbb{N}}$ *and* $\{\beta_n\}_{\mathbb{N}}$ *in* $\mathbb{R}_{++}$ *converging to zero. Then, there exists a finite constant* C_b*, depending only on* b*, such that*

$$\max_{1 \leq j \leq k} \sup_{\epsilon \in (0,1)} \left| \Pr(\|S_{n,j}\|_\infty \geq \widehat{q}_{n,j}(1 - \alpha)) - \alpha \right| \leq C_b \max \left\{ \beta_n, \delta_n, \left(\frac{B_n^4 \ln^7(pn)}{n} \right)^{1/6}, \frac{1}{\ln^2(pn)} \right\}.$$

Proof. By Theorem D.1 and Theorem D.3,

$$\begin{aligned}
\Pr(\|S_{n,j}\|_\infty \leq \widehat{q}_{n,j}(1 - \epsilon)) &\leq \Pr(\|S_{n,j}\|_\infty \leq q^N_{\Sigma,j}(1 - \epsilon + \rho'_n)) + \beta_n + \frac{1}{\ln^2(pn)} \\
&\leq \Pr(\|N_{\Sigma,j}\|_\infty \leq q^N_{\Sigma,j}(1 - \epsilon + \rho'_n)) + \rho_n + \beta_n + \frac{1}{\ln^2(pn)} \\
&\leq 1 - \epsilon + \rho'_n + \rho_n + \beta_n + \frac{1}{\ln^2(pn)}
\end{aligned}$$

Where the second inequality is making use of the same rectangle argument as before. A parallel argument shows that

$$\Pr(\|S_{n,j}\|_\infty \leq \widehat{q}_{n,j}(1-\epsilon)) \geq 1 - \epsilon - \left(\rho'_n + \rho_n + \beta_n + \frac{1}{\ln^2(pn)}\right).$$

Combining these two inequalities gives the result.

□

D.2 Concentration and Tail Bounds

We make use of the following concentration and tail bounds. Lemmas D.3–D.7 can be found in Bühlmann and van de Geer (2011). The proof of Lemma D.8 is trivial but provided here.

Lemma D.3. *Let $(Y_1, \ldots, Y_n)$ be independent random variables such that $\mathbb{E}[Y_i] = 0$ for $i = 1, \ldots, n$ and $\max_{i=1,\ldots,m} |Y_i| \leq c_0$ for some constant c_0. Then, for any $t > 0$,*

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^n Y_i\right| > t\right) \leq 2\exp\left(-\frac{nt^2}{2c_0^2}\right).$$

Lemma D.4. *Let $(Y_1, \ldots, Y_n)$ be independent random variables such that $\mathbb{E}[Y_i] = 0$ for $i = 1, \ldots,$ and $(Y_1, \ldots, Y_n)$ are uniformly sub-gaussian: $\max_{1\leq i\leq n} c_1^2\mathbb{E}[\exp(Y_i^2/c_1^2) - 1] \leq c_2^2$ for some constants (c_1, c_2). Then for any $t > 0$,*

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^n Y_i\right| > t\right) \leq 2\exp\left(-\frac{nt^2}{8(c_1^2 + c_2^2)}\right).$$

Lemma D.5. *Let $(Y_1, \ldots, Y_n)$ be independent variables such that $\mathbb{E}[Y_i] = 0$ for $i = 1, \ldots, n$ and*

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}[|Y_i|^k] \leq \frac{k!}{2}c_3^{k-2}c_4^2, \quad k = 2, 3, \ldots,$$

for some constants (c_3, c_4). Then, for any $t > 0$,

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^n Y_i\right| > c_3 t + c_4\sqrt{2t}\right) \leq 2\exp(-nt).$$

Lemma D.6. *Suppose that Y is sub-gaussian: $c_1^2\mathbb{E}[\exp(Y^2/c_1^2) - 1] \leq c_2^2$ for some constants (c_1, c_2). Then*

$$\mathbb{E}[|Y|^k] \leq \Gamma\left(\frac{k}{2} + 1\right)(c_1^2 + c_2^2)c_1^{k-2}, \quad k = 2, 3, \ldots.$$

Lemma D.7. *Suppose that X is bounded, $|X| \leq c_0$, and Y is sub-gaussian, $c_2^2\mathbb{E}[\exp(Y^2/c_1^2) - 1] \leq c_2^2$*

for some constants (c_1, c_2)*. Then* $Z = XY^2$ *satisfies*

$$\mathbb{E}\left[|Z - \mathbb{E}[Z]|^k\right] \leq \frac{k!}{2} c_3^{k-2} c_4^2, \quad k = 2, 3, \ldots,$$

for $c_3 = 2c_0 c_1^2$ *and* $c_4 = 2c_0 c_1 c_2$*.*

Lemma D.8. *Suppose that* Y *is sub-gaussian in the following sense, there exist positive constants* $c_0, c_1 > 0$ *such that* $c_0^2 \mathbb{E}[\exp(Y^2/c_0^2) - 1] \leq c_1^2$*. Then*

$$\mathbb{E}[|Y|] \leq c_1^2/c_0 + c_0.$$

Proof. Use the fact that $e^{x^2} > |x|$ and the characterization of sub-gaussian. □

E Additional Details on Empirical Application

As mentioned in the setup, to avoid outlier contamination we drop the top 3% and bottom 3% of birthweights by maternal age. We also drop ages for which there are fewer than 10 smoker or non smoker observations. The result is a dataset with 4107 (of an initial 4602) observations on the outcome variable, birthweight. In addition to the 21 control variables (Z) available in the dataset, we further generate an additional 29 interaction/higher order variables that we believe may be useful in controlling for confounding as well as a constant. Table E.1 provides a summary of the initial 21 control variables.[1]

In addition to these 21 control variables, we include the folowing interactions: mbsmoke × alcohol, medu × fedu, mage × fage, msmoke^2, msmoke × alcohol, mage^2, mage × mmarried, mage × medu, mage × fedu, monthslb^2, msmoke ×monthslb^2, monthslb^2 × msmoke^2, msmoke^2 × prenatal^2, msmoke^2 × mage^2, mage^2 × monthslb^2, mage^2 × fage, fage^2 × mage^2, fage^2 × mage, mage^2 × mrace, fage^2 × frace, msmoke^2 × alcohol, mage^2 × alcohol, fage^2 × alcohol, monthslb^2 × alcohol, mage^2 × mhisp, fage^2 × fhisp, medu × mage^2. We also include indicators for the month of birth.

In conducting analysis, we found it quite helpful to the stability of the final model assisted estimator to do some light trimming of the estimated propensity score and outcome regression models. In particular we trim the estimated propensity score(s) to be between 0.01 and 0.99 and trim the estimated mean regression models so that they take a value no more than roughly 12.5% higher or lower than the maximum or minimum value of Y observed in the data.

Because the control variables are all of different magnitudes, it is common to do some normalization before estimating the ℓ_1-regularized propensity score and outcome regression models so that all variables are "punished" equally by the penalty. We normalize our data by scaling each variable to take on values between zero and one.

[1] This table is generated using the wonderful stargazer package in R (Hlavac, 2022).

Table E.1: Summary of Data used in Emprical Exercise

Statistic	N	Mean	St. Dev.	Min	Max
bweight	4,107	3,384.354	447.616	1,544	4,668
mmarried	4,107	0.708	0.455	0	1
mhisp	4,107	0.034	0.181	0	1
fhisp	4,107	0.038	0.192	0	1
foreign	4,107	0.054	0.226	0	1
alcohol	4,107	0.031	0.174	0	1
deadkids	4,107	0.252	0.434	0	1
mage	4,107	26.125	5.025	16	36
medu	4,107	12.703	2.470	0	17
fage	4,107	27.000	9.022	0	60
fedu	4,107	12.324	3.624	0	17
nprenatal	4,107	10.822	3.613	0	40
monthslb	4,107	21.938	30.255	0	207
order	4,107	1.858	1.056	0	12
msmoke	4,107	0.390	0.890	0	3
mbsmoke	4,107	0.183	0.386	0	1
mrace	4,107	0.847	0.360	0	1
frace	4,107	0.822	0.382	0	1
prenatal	4,107	1.204	0.507	0	3
birthmonth	4,107	6.556	3.352	1	12
lbweight	4,107	0.025	0.155	0	1
fbaby	4,107	0.443	0.497	0	1
prenatal1	4,107	0.803	0.398	0	1

F CONSISTENCY BETWEEN FIRST STAGE AND SECOND STAGE ASSUMPTIONS

In this section, we examine the consistency between the first stage and second stage assumptions on the basis terms $p^k(x)$. In particular, we are interested in finding a positive basis that also satisfies the bounded eigenvalue condition on the design matrix in Assumption 4.1. We also discuss how to construct the model assisted estimator with weights in (2.8)-(2.9) that are not directly the second stage basis terms in case the researcher is worried about their choice of basis terms satisfying the first stage and second stage stage assumptions simultaneously.

Suppose that $\mathcal{X} = [0, 1]$. First, note that the first stage non-negativity and second stage design assumptions can be trivially satisfied by using a locally constant basis; that is by taking

$$p_j(x) = \mathbf{1}_{[\ell_{j-1}, \ell_j)}(x) \tag{F.1}$$

for some $0 = \ell_0 < \ell_1 < \cdots < \ell_t = 1$. While this basis may have poor approximation qualities, the general principle can be extended to any basis whose elements have disjoint (or limitedly overlapping) supports. Higher order piecewise polynomial approximations can often be implemented using *B-splines* which are orthonormalized regression splines. See De Boor (2001) for an in-depth discussion or Newey (1997) for an application of B-splines to nonparametric series regression.

These higher order splines can be defined recursively. For a given (weakly increasing) knot sequence $\ell := (\ell_j)_{j=1}^t$ we define the "first-order" B-splines denoted $B_{1,1}(x), \ldots, B_{t,1}(x)$ using (F.1), that is $B_{j,1}(x) = p_j(x)$. On top of these functions, we can define higher order B-splines via the recursive relation (De Boor (2001), p.90)

$$B_{j,d+1} := \omega_{j,d}(x)B_{j,d}(x) + [1 - \omega_{j+1,d}(x)]B_{j+1,d}(x). \tag{F.2}$$

where

$$\omega_{j,d}(x) := \begin{cases} \frac{x-\ell_j}{\ell_{j+d}-\ell_j} & \text{if } \ell_{j+d} \neq \ell_j \\ 0 & \text{otherwise} \end{cases}.$$

If X is continuously distributed on an open set containing the knots (ℓ_j), De Boor (2001) shows that the B-spline basis is almost surely positive. Moreover, B-splines is locally supported in the sense each $B_{j,d}$ is positive on (ℓ_j, ℓ_{j+d}), zero off this support and for each d:

$$\sum_{j=1}^{t} B_{j,d} = 1 \quad \text{on } [0,1].$$

where the summation is taken pointwise (see De Boor (2001), p.36). From the final property we can see the B-spline basis using $k = td$ basis terms, $p^k(x) = (B_{j,l}(x))_{\substack{j=1,..,t \\ l=1,\ldots,d}}$ are totally bounded so that.

B-splines used directly in this manner, however, do not lead to a design matrix $Q = \mathbb{E}[p^k(x)p^k(x)']$ with eigenvalues which are bounded away from zero. To achieve this, the basis fucntions must be divided by their ℓ_2 norm. In practice, this leads to b-spline terms who are grown at rate $\xi_{k,\infty} \lesssim \sqrt{k}$. The pilot penalty constants can be chosen from a set whose bounds are on the order of $\sqrt{k}$ and the sparsity bounds of Assumption 3.1 reduce to

$$\frac{s_k\, k^{3/2} \ln^5(d_z n)}{n} \to 0 \text{ and } \frac{k^2 \ln^7(d_z k n)}{n} \to 0$$

while the bounds in (4.2) and (4.11) reduce respectively to

$$\frac{s_k\, k^{3/2} \ln(d_z)}{\sqrt{n}} \to 0 \text{ and } \frac{s_k^2\, k^{7/2} \ln(d_z)}{n^{(m-1)/m}} \to 0.$$

F.1 ALTERNATE WEIGHTING

So long as the second stage basis $p^k(x)$ contains a constant term, it is possible to weight the estimating equations (2.8)-(2.9) by some $p^k(x) = p^k(x) + c_k$ with minimal modification to the model assisted estimator. The constants $c_k \in \mathbb{R}$ can be allowed to grow with k so long as we replace $\xi_{k,\infty}$ with the maximum of $\tilde{\xi}_{k,\infty} := \sup_{x\in\mathcal{X}} \|\tilde{p}^k(x)\|_\infty$ and $\xi_{k,\infty}$ in the sparsity bounds of Section 4. Without loss of generality we will assume that the first basis term is a constant so that $p_1(x) \equiv 1$

After estimating the models $(\hat{\pi}_1, \hat{m}_1), \ldots, (\hat{\pi}_k, \hat{m}_k)$ using $(\tilde{p}_1(x), \ldots, \tilde{p}_k(x))$ in place of $(p_1(x), \ldots, p_k(x))$

in (2.8)-(2.9) we would construct the second stage estimate $\hat{\beta}^k$

$$\tilde{\beta}^k = \widehat{Q}^{-1}\mathbb{E}_n \begin{bmatrix} \tilde{p}_1(x)Y(\hat{\pi}_1, \hat{m}_1) - c_k Y(\hat{\pi}_1, \hat{m}_1) \\ \tilde{p}_2(x)Y(\hat{\pi}_2, \hat{m}_2) - c_k Y(\hat{\pi}_1, \hat{m}_1) \\ \vdots \\ \tilde{p}_k(x)Y(\hat{\pi}_k, \hat{m}_k) - c_k Y(\hat{\pi}_1, \hat{m}_1) \end{bmatrix}.$$

Via the same analysis of Sections 3 and 4 we will still be able to show that the bias passed on from first stage estimation to the second stage parameter $\tilde{\beta}^k$ remains negligible even under misspecification of either first stage model. This is because Lemma 3.1 will establish that

$$\begin{aligned} &\max_{1\le j\le k} |\mathbb{E}_n[\tilde{p}_j(x)Y(\hat{\pi}_j, \hat{m}_j)] - \mathbb{E}_n[\tilde{p}_j(x)Y(\bar{\pi}_j, \bar{m}_j)]| = o_p(n^{-1/2}k^{-1/2}) \text{ and} \\ &\max_{1\le j\le k} \tilde{\xi}_{k,\infty} \max_{1\le j\le k} \mathbb{E}_n[\tilde{p}_j(x)^2(Y(\widehat{\pi}_j, \widehat{m}_j) - Y(\bar{\pi}_j, \bar{m}_j))^2] = o_p(k^{-2}n^{-1/m}) \end{aligned}.$$

Using the first statement, we can immediately establish via the triangle inequality that

$$\max_{1\le j\le k} |\mathbb{E}_n[\tilde{p}_j(x)Y(\hat{\pi}_j, \hat{m}_j) - c_k Y(\hat{\pi}_1, \hat{m}_1)] - \mathbb{E}_n[\tilde{p}_j(x)Y(\bar{\pi}_j, \bar{m}_j) - c_k Y(\bar{\pi}_1, \bar{m}_1)]| = o_p(n^{-1/2}k^{-1/2})$$

which is the exact analog of Condition 1 needed to establish consistency at the nonparameteric rate of the modified model assisted estimator. Similarly, using the second statement and $(a+b)^2 \le 2a^2 + 2b^2$ we can immediately establish that

$$\max_{1\le j\le k} \mathbb{E}_n[(\tilde{p}_j(x)Y(\hat{\pi}_j, \hat{m}_j) - cY(\hat{\pi}_1, \hat{m}_1) - \tilde{p}_j(x)Y(\bar{\pi}_j, \bar{m}_j) + cY(\bar{\pi}_j, \bar{m}_j))^2] = o_p(k^{-2}n^{-1/m})$$

which is the exact analog of Condition 2 needed to establish a consistent variance estimator when $\tilde{\beta}^k$ is used instead of the $\hat{\beta}^k$ from (2.12).

This logic can be extended slightly if the researcher would like to weight the estimating equations (2.8)-(2.9) by some $\tilde{p}^k(x) = G^k p^k(x)$ for an invertible and bounded sequence of linear operators $G^k : \mathbb{R}^k \to \mathbb{R}^k$. In this case, one would again use $\tilde{p}^k(x)$ in place of $p^k(x)$ in (2.8)-(2.9) and construct the second stage coeffecients via

$$\tilde{\beta}^k := \widehat{Q}^{-1}G^{k,-1}\mathbb{E}_n \begin{bmatrix} \tilde{p}_1(x)Y(\hat{\pi}_1, \hat{m}_1) \\ \vdots \\ \tilde{p}_k(x)Y(\hat{\pi}_k, \hat{m}_k) \end{bmatrix}$$

After constructing the second stage estimator using $\tilde{\beta}^k$, inference procedures would proceed normally as described in Section 2.

G ALTERNATIVE CV-TYPE METHOD FOR PENALTY PARAMETER SELECTION

In this section we consider a procedure for penalty parameter selection where we use the pilot penalty parameters described in (2.15) directly, after choosing constants $c_{\gamma,j}$ and $c_{\alpha,j}$ from a

(finite) set via cross validation. For each j we will assume that

$$c_{\gamma,j}, c_{\alpha,j} \in \Lambda_n \subseteq [\underline{c}_n, \bar{c}_n] \tag{G.1}$$

where $|\Lambda_n|$ can be fairly large (on the order of n^2/k).

G.1 THEORY OVERVIEW

Let $M_5, M_6, M_7, M_8^2, M_9^2$ be constants that do not depend on k as in Lemmas B.7–B.11. Whenever

$$\underline{c}_n \sqrt{\frac{\ln^3(d_z n)}{n}} \geq \xi_{k,\infty} \max\left\{M_5, M_6, M_7, M_8^2, M_9^2\right\} \sqrt{\frac{\ln(d_z n)}{n}}. \tag{G.2}$$

we will have that, under Assumption 3.1(i)-(iv) the event $\bigcap_{k=1}^7 \Omega_{k,7}$ occurs with probability at least $1 - 10k/n^2$ for the $2k$ pilot penalty parameters chosen with any values $c_{\gamma,j}, c_{\alpha,j} \in \Lambda_n$ and

$$\bar{\lambda}_k := \bar{c}_n \sqrt{\frac{\ln^3(d_z n)}{n}}.$$

In this event, apply Lemmas B.1 and B.2 to obtain the following finte sample bounds for the parameter estimates

$$\max_{1\leq j\leq k} D^{\ddagger}_{\gamma,j}(\widehat{\gamma}_j, \bar{\gamma}_j) \leq M_0 \frac{s_k \bar{c}_n^2 \ln^3(d_z n)}{n} \quad \text{and} \quad \max_{1\leq j\leq k} \|\widehat{\gamma}_j - \bar{\gamma}_j\|_1 \leq M_0 s_k \bar{c}_n \sqrt{\frac{\ln^3(d_z n)}{n}}$$

$$\max_{1\leq j\leq k} D^{\ddagger}_{\alpha,j}(\widehat{\alpha}_j, \bar{\alpha}_j; \bar{\gamma}_j) \leq M_1 \frac{s_k \bar{c}_n^2 \ln^3(d_z n)}{n} \quad \text{and} \quad \max_{1\leq j\leq l} \|\widehat{\alpha}_j - \bar{\alpha}_j\|_1 \leq M_1 s_k \bar{c}_n \sqrt{\frac{\ln^3(d_z n)}{n}}$$

and Lemma A.1 to obtain the following finite sample bound for the weighted means:

$$\max_{1\leq j\leq k} |\mathbb{E}_n[p_j(X)(Y(\widehat{\pi}_j, \widehat{m}_j) - Y(\bar{\pi}_j, \bar{m}_j))]| \leq M_2 \frac{\bar{c}_n^2 s_k \ln^3(d_z n)}{n} \tag{G.3}$$

$$\max_{1\leq j\leq k} |\mathbb{E}_n[p_j^2(X)(Y(\widehat{\pi}_j, \widehat{m}_j) - Y(\bar{\pi}_j, \bar{m}_j))^2] \leq M_3 \frac{\xi_{k,\infty}^2 \bar{c}_n^2 s_k^2 \ln^3(d_z n)}{n} \tag{G.4}$$

Combining (G.2) and (G.3) we can see that Condition 1 can be obtained under Assumption 3.1(i)-(iv) and the following modified sparsity bounds

$$\frac{k|\Lambda_n|}{n^2} \to 0, \quad \frac{\underline{c}_n^{-1} \xi_{k,\infty}}{\ln(d_z n)} \to 0 \quad \text{and} \quad \frac{\bar{c}_n^2 s_k k^{1/2} \ln^3(d_z n)}{\sqrt{n}} \to 0. \tag{G.5}$$

Simlarly combining (G.2) and (G.4), Condition 2 can additionally be obtained by strengthening the rates in (G.5) to include

$$\frac{\xi_{k,\infty}^2 \bar{c}_n^2 s_k k^2 \ln^3(d_z n)}{n^{(m-1)/m}} \to 0 \tag{G.6}$$

for $m > 2$ as in Assumption 4.2. These rates are comparable and in certain cases may be more palatable than those presented in the main text, Assumption 3.1(vi). They come at the cost of slower rates of convergence for the weighted means as seen by comparing eqs. (G.3)–(G.4) to eqs. (3.1) and (3.2).

G.2 PRACTICAL IMPLEMENTATION

In practice, the constants $M_5, M_6, M_7, M_8^2, M_9^2$ from Lemmas B.7–B.11 are roughly on the order of $\|Z\|_\infty$. We therefore reccomend setting

$$\underline{c}_n = \frac{1}{2\log^{1/2}(d_z n)} \max_{1\le i\le n} \|p^k(X_i)\|_\infty \max_{1\le i\le n} \|Z_i\|_\infty$$
$$\bar{c}_n = \frac{3\log^{1/2}(d_z n)}{2} \max_{1\le i\le n} \|p^k(X_i)\|_\infty \max_{1\le i\le n} \|Z_i\|_\infty$$

and letting Λ_n be a set of points evenly spaced between $\underline{c}_n$ and $\bar{c}_n$. The cross validation procedure then can be implemented in the following steps.

1. Split the sample into K folds.
2. Consider a single pair of values for c_α, c_γ and designate a fold to hold out.
3. Estimate nuisance model parameters using $\lambda_{\gamma,j}^{\text{pilot}}$ and $\lambda_{\alpha,j}^{\text{pilot}}$ on the remaining folds.
4. Evaluate the resulting models on held out fold using non-penalized loss functions.
5. Repeat K times and record average loss over all folds.
6. Choose values of $c_{\gamma,j}$ and $c_{\alpha,j}$ with the lowest average loss.

In practice we find this procedure works well with small K, around $K = 5$ and with $|\Lambda_n|$ on the order of about 10-20.