Model-Assisted Inference for Conditional Average Treatment Effects with High-Dimensional Controls*

Adam Baybutt and Manu Navjeevan†

University of California, Los Angeles

Revised July 31, 2022

Abstract

Plausible identification of conditional average treatment effects (CATE's) often relies on controlling for a large number of variables to account for confounding factors. In these high-dimensional settings, estimation of the CATE requires first estimating first-stage models. The resulting second-stage CATE estimator is said to be doubly robust if it remains consistent when either one of the first-stage models is consistently estimated. Inference procedures for this estimator are termed "model-assisted" if they are asymptotically valid when either one of the first stage models is consistently estimated. These properties are especially relevant in high-dimensional settings where consistency of the first stage estimators often relies on correctly specifying their parametric forms.

While doubly robust estimators of the CATE have been studied, inference based on these estimators is typically not model-assisted. Using the popular augmented inverse propensity weighting signal, we propose an estimator for the CATE that is doubly robust and whose resulting Wald-type confidence intervals are model-assisted. 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. Our proposed estimator remains consistent at the nonparametric rate and our proposed pointwise and uniform confidence intervals remain asymptotically valid even if one of the logistic propensity score or linear outcome regression models are misspecified. These results are obtained under similar conditions to existing analyses in the high-dimensional and nonparametric literatures.

*We are grateful to Denis Chetverikov, Andres Santos, Zhipeng Liao, Jinyong Hahn, Rosa Matzkin 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 $Y \in \mathsf{R}$ and treatment $D \in \{0, 1\}$ are related to two latent potential outcomes $Y_1, Y_0 \in \mathsf{R}$ 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 $Z = (Z_1, X) \in Z_1 \times X \subseteq \mathsf{R}^{d_z - d_x} \times \mathsf{R}^{d_x}$ such that the potential outcomes are independent of the treatment decision after conditioning on the observed covariates, $(Y_1, Y_0) \perp D \mid Z$. In this setting, we are interested in estimation of and inference on the conditional average treatment effect (CATE):

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

Estimation of the CATE generally requires first fitting propensity score and/or outcome regression models. When the number of control variables Z is large ($d_z \approx n$), as is often needed to justify conditional unconfoundedness, these first stage models are typically estimated using regularized methods which converge slowly and rely on the correctness of ad-hoc parametric specifications for consistency.[1]

Fortunately, so long as 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 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 model-assisted. Under misspecification the aIPW signal orthogonality fails and the second stage CATE estimators converge slowly, rendering resulting testing procedures and confidence intervals invalid.

This paper proposes a model-assisted 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, will remain fixed in our analysis. Our approach is based on Tan (2020) wherein model-assisted inference is developed for the average treatment effect. Following Semenova and Chernozhukov (2021) we take a series approach to estimating the CATE, using a projection of 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 limited orthogonality of the aIPW signal at the limiting values of the first stage estimators. This restricted orthogonality is enough to achieve model-assisted pointwise and uniform inference; the second-stage estimator's pointwise and uniform confidence intervals are valid even if one of the logistic or linear functional forms is misspecified.

To achieve model-assisted inference at all points in the support of the conditioning variable, we must obtain this restricted orthogonality for each basis term in the series approximation. This is accomplished by employing 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 model-assisted inference, our ℓ_1 regularized first-stage estimation incorporates

[1] See Bauer and Kohler (2019); Schmidt-Hieber (2020) for some recent nonparametric results in high-dimensional settings using deep neural networks.

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.

The use of multiple pairs of nuisance parameter estimates 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 normality and strong gaussian process approximation 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. In the same setting as this paper, Tan (2020) considers estimation of the average treatment effect. After assuming a logistic form for the propensity score and a linear form for the outcome regression, Tan (2020) proposes 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 model-assisted inference. Smucler et al. (2019) extends the analysis of Tan (2020) to consider model-assisted inference for a larger class of finite dimensional target parameters with bilinear influence functions.

Inference on low-dimensional target parameters after estimation of high-dimensional nuisance parameters has its origins in the work of Belloni et al. (2012, 2013), who consider inference on causal parameters after estimation of nuisance parameters using Lasso and Post-Lasso methods. Chernozhukov et al. (2018) analyze estimation of low-dimensional target parameters using score functions that are Neyman-orthogonal with respect to nuisance parameters. They show that the resulting target parameter estimate is $\sqrt{n}$-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.

Stone (1985) provides a primary analysis of nonparametric series estimators. Newey (1997) derives convergence rates and pointwise normality results which are updated by Belloni et al. (2015). Recently, Semenova and Chernozhukov (2021) consider series estimation of functional target parameters after high-dimensional nuisance estimation using the aIPW signal. Fan et al. (2022) provides a similar analysis using a second stage kernel estimator. The resulting second stage estimators are doubly robust but do not necessarily yield model-assisted inference. Wu et al. (2021) provide model-assisted 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.

Regularization using an ℓ_1-penalty in linear regression models (Lasso) was first introduced by Tibshirani (1996) for selecting relevant features out of a high-dimensional set of controls. Near oracle rates of convergence for these estimators under sparsity are given by Bickel et al. (2009); Belloni and Chernozhukov (2013). Chetverikov et al. (2021) provide interesting new results with cross validated penalties. These works can be extended to consider rates of convergence for general ℓ_1-penalized M-estimators as seen in van der Greer (2016); Tan (2017); Chetverikov and Sørensen (2021). In particular, 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.

Paper Structure. This paper proceeds as follows. Section 2 defines the problem and introduces methods for estimation and inference. Section 3 introduces, with exposition for intuition, the main result: valid 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 4 shows uniform convergence of the first-stage estimators and describes the required conditions. Section 5 presents a full set of second-stage asymptotic results. Section 6 ties up a technical detail. Section 7 provides evidence from a simulation study while Section 8 applies our proposed estimator to examine the effect of maternal smoking on infant birth weight. Section 9 concludes. Proofs of main results are deferred to Appendix A.

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

We will assume that the researcher observes i.i.d data and that, as discussed, 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.

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 $\eta < \mathsf{E}[D \mid Z = z] < 1 - \eta$ almost surely in Z.

Our analysis allows the dimensionality of Z to grow much faster than sample size ($d_z \gg n$), while assuming the dimensionality of X remains fixed ($d_x \ll n$). The parameter of interest is the conditional average treatment effect: $\mathsf{E}[Y_1 - Y_0 \mid X = x]$. However, for this paper we largely focus on estimation and inference for the conditional average counterfactual outcome:

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

A model-assisted inference procedure for the other conditional counterfactual outcome, $\mathsf{E}[Y_0 \mid X = x]$, follows a similar procedure and is described in Section 6. The procedures can be combined for model-assisted estimation and inference of the CATE.

We utilize series estimation targeting $g_0(x)$. The augmented inverse propensity weighted (aIPW) signal we employ,

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

is a function of a fitted propensity score model, (Z); and a fitted outcome regression model, $m(Z)$: After estimating these two first-stage models, whose true values are given $^{?}(Z) := \Pr(D = 1 \mid Z)$ and $m^{?}(Z) := \mathsf{E}[Y \mid D = 1; Z]$, the target parameter $g_0(x)$ is estimated by taking a quasi-projection of the aIPW signal (2.2) onto a growing set of k, weakly positive, basis terms

$$p^k(x) := \; p_1(x); : : : ; p_k(x) \;^{0} \; 2 \; \mathsf{R}^k_+ : \tag{2.3}$$

The basis terms are required to be weakly positive as they will be used as weights within the first-stage estimators (2.5)-(2.6). Negative weights break the convexity of the objective function. Examples of weakly positive basis functions are B-splines or Fourier series basis terms[1]. To ensure that the basis terms are well behaved, we will make assumptions on $_{k;1} := \sup_{x2X} kp^k(x)k_1$, $_{k;2} := \sup_{x2X} kp^k(x)k_2$, and the eigenvalues of the design matrix $Q := \mathsf{E}[p^k(x)p^k(x)^{0}]$.

First-Stage Setup and Estimation. We estimate a separate propensity score model and outcome regression model for each basis term. Following Tan (2020) we assume a logistic regression form for the propensity score model and a linear form for the outcome regression model:

$$\begin{aligned} (Z; \;\;) &= \; 1 + \exp(\quad {}^{0}Z) \quad {}^{1}; \\ m(Z; \;\;) &= \; {}^{0}Z: \end{aligned} \tag{2.4}$$

To deal with the high-dimensional Z, we implement $`_1$-penalized optimization to estimate the parameters of the models. Importantly, these first stage estimating equations are designed as in Tan (2020) so that their first order conditions directly limit the bias passed on to the second-stage series estimator, as will be described in Section 3. For each $j = 1; : : : ; k;$ the parameters of (2.4), $\;;\;\; 2 \; \mathsf{R}^{d_z};$ are estimated by

$$\mathsf{b}_j := \arg\min \; \mathsf{E}_n[p_j(X) f De \quad {}^{0}Z + (1 \quad D) \; {}^{0}Zg] + \quad {}_{;j}k \;\; k_1; \tag{2.5}$$

$$\mathsf{b}_j := \arg\min \; \mathsf{E}_n[p_j(Z) De \;\; {}^{\mathsf{b}^{0}_j Z}(Y \quad {}^{0}Z)^2] {=} 2 + \quad {}_{;j}k \;\; k_1: \tag{2.6}$$

The penalty parameters $\;{}_{;j}$ and $\;{}_{;j}$, for each $j = 1; : : : ; k$, will be chosen via a data dependent technique described in Section 4.1. Under standard assumptions the nuisance estimators $\mathsf{b}_j; \mathsf{b}_j$ will converge uniformly over $j = 1; : : : ; k$ to the population minimizers

$$^{-}_j := \arg\min \; \mathsf{E}[p_j(X) f De \quad {}^{0}Z + (1 \quad D) \; {}^{0}Zg]; \tag{2.7}$$

$$^{-}_j := \arg\min \mathsf{E}[p_j(Z) De \quad {}^{0}_j Z (Y \quad {}^{0}Z)^2]: \tag{2.8}$$

We will assume that the population minimizers are sufficiently sparse. To control the sparsity, define $S_{;j} := fj : {}^{-}_j \; \not{6}= 0g$, $S_{;j} := fj : {}^{-}_j \; \not{6}= 0g$ and $s_k := \max_{1 \;\; j \;\; k} fjS_{;j} j \_ jS_{;j} jg$. Uniform convergence of the parameter estimators (2.5) and (2.6) to their population minimizers (2.7) and (2.8) will ensure uniform convergence of the estimated propensity score and outcome regression models to limiting values

$$\mathsf{b}_j(Z) := \quad (Z; \mathsf{b}_j) \;\; !_p \; {}^{-}_j(Z) := \quad (Z; {}^{-}_j); \qquad \not{m}_j(Z) := m(Z; \mathsf{b}_j) \;\; !_p \; \bar{m}_j(Z) := m(Z; {}^{-}_j):$$

[1] B-splines are linear transformations of regression spliens with lower multicolinearity. See De Boor (2001) for more information on B-splines.

If the logistic propensity score model (linear outcome regression model) is correctly specified, the limiting propensity score model (outcome regression model) will coincide with the true value, i.e. for all $j = 1,\dots,k$ $\bar{\gamma}_j(Z) = \gamma^*(Z)$ ($\bar{m}_j(Z) = m^*(Z)$). In general however they may differ. As discussed, our inference results will remain valid so long as one of either the logistic propensity score model or the linear outcome regression model is correctly specified.

Second Stage Setup and Estimation. Under Assumption 2.1, the aIPW signal $Y(\cdot\,;\cdot)$ is doubly robust, that is for square integrable $\gamma \neq \gamma^*$ and $m \neq m^*$,

$$\begin{aligned} \mathsf{E}[Y_1 \mid X = x] &= \mathsf{E}[Y(\gamma^*; m^*) \mid X = x] \\ &= \mathsf{E}[Y(\gamma; m^*) \mid X = x] \\ &= \mathsf{E}[Y(\gamma^*; m) \mid X = x]. \end{aligned} \tag{2.9}$$

This 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 $\bar{\gamma}_j = \gamma^*$ or $\bar{m}_j = m^*$, we can write for all $j = 1,\dots,k$:

$$\begin{aligned} Y(\bar{\gamma}_j; \bar{m}_j) &= g_0(x) + \varepsilon_j; \qquad \mathsf{E}[\varepsilon_j \mid X] = 0 \\ &= g_k(x) + r_k(x) + \varepsilon_j \end{aligned} \tag{2.10}$$

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. 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 $p^k(x)$:[2]

$$\begin{aligned} \beta^k &:= Q^{-1}\mathsf{E}[p^k(X)Y_1] \\ &= Q^{-1}\mathsf{E}[p^k(X)Y(\gamma^*; m^*)] \\ &= Q^{-1}\mathsf{E}[p^k(X)Y(\bar{\gamma}_j; \bar{m}_j)], \;\; \forall j = 1,\dots,k. \end{aligned} \tag{2.11}$$

To estimate β^k we combine the k estimates according to:

$$\hat{\beta}^k = \hat{Q}^{-1}\mathsf{E}_n\begin{bmatrix} p_1(X)Y(\hat{\gamma}_1; \hat{m}_1) \\ \vdots \\ p_k(X)Y(\hat{\gamma}_k; \hat{m}_k) \end{bmatrix}; \tag{2.12}$$

where $\hat{Q} := \mathsf{E}_n[p^k(x)p^k(x)']$. Our final estimator for $g_0(x)$ is $\hat{g}(x) := p^k(x)'\hat{\beta}^k$. Note the separate error terms for each $j = 1,\dots,k$ in (2.10), which are collected together in the vector $\varepsilon^k := (\varepsilon_1,\dots,\varepsilon_k)$.

Inference Procedure. Having defined our estimator $\hat{g}(x)$ we now describe a typical inference procedure. The variance of our second stage estimator is governed by the following variance matrices:

$$\begin{aligned} \tilde{\Omega} &:= Q^{-1}\mathsf{E}[fp^k(x)(\varepsilon^k + r_k)(\varepsilon^k + r_k)'fp^k(x)'Q^{-1}] \\ \Omega_0 &:= Q^{-1}\mathsf{E}[fp^k(x)\varepsilon^k\varepsilon^{k\prime}fp^k(x)'Q^{-1}] \end{aligned} \tag{2.13}$$

[2]The second and third equalities of (2.11) hold under Assumption 2.1 for any $\gamma \neq \gamma^*$ and $m \neq m^*$ by double robustness of the aIPW signal (2.9).

where represents the Hadamard (element-wise) product and, abusing notation somewhat, for a vector $a \in \mathbb{R}^k$ and scalar $c \in \mathbb{R}$ we let $a + c = (a_i + c)_{i=1}^k$. We will estimate these matrices using the plug-in empirical analog

$$\hat{\Omega} := \hat{Q}^{-1} \mathbb{E}_n[f p^k(x) \quad \hat{b}^k g f p^k(x) \quad \hat{b}^k g^0] \hat{Q}^{-1} \tag{2.14}$$

for $\hat{b}^k := (\hat{b}_1, \ldots, \hat{b}_k)$ and $\hat{b}_j := Y(\hat{b}_j, \hat{m}_j)$ $\hat{g}(x)$ for $j = 1, \ldots, k$. Define $\hat{s}(x) := \|\hat{\Omega}^{1/2} p^k(x)\| / \sqrt{n}$ and let the test statistic $t(x)$ be given:

$$t(x) := \frac{\hat{g}(x) \quad g(x)}{\hat{s}(x)} \tag{2.15}$$

Denote the bootstrapped test statistic as

$$\hat{t}^b(x) := \frac{p^k(x)' \hat{\Omega}^{1/2}}{\|\hat{\Omega}^{1/2} p^k(x)\|} N_k^b \tag{2.16}$$

where N_k^b is a bootstrap draw from $N(0, I_k)$. Define the confidence bands

$$\underline{I}(x), \bar{I}(x) \quad := \quad \hat{g}(x) \quad c^*(1 \quad /2)\hat{s}(x), \hat{g}(x) + c^*(1 \quad /2)\hat{s}(x) \quad . \tag{2.17}$$

The critical value $c^*(1 \quad /2)$ is either $c_p^*(1 \quad /2)$, the $(1 \quad /2)$ quantile of a $N(0, 1)$ distribution, for pointwise inference or $c_u^*(1 \quad /2)$, the $(1 \quad /2)$-quantile of $\sup_{x \in \mathcal{X}} |\hat{t}_n^b(x)|$, for uniform inference. Below, we will show that under our construction of $\hat{g}(x)$, this standard inference procedure will remain valid even under misspecification of either nuisance model.

3 Theory Overview

We now present the main results of the paper and give some intuition into how they are obtained.

3.1 Main Results

Proposition 3.1 (Pointwise Normality). *Suppose that the conditions of Theorems 4.3 and 5.4 hold. If additionally the first stage sparsity bounds of (4.10) and (4.13) hold, then so long as either the logistic propensity score model or linear outcome regression model is correctly specified we have for any $\in S^{k-1}$,*

$$\sqrt{n} \frac{'(\hat{b}^k \quad ^k)}{\| ' \hat{\Omega}^{1/2}\|} \to_d N(0, 1). \tag{3.1}$$

Moreover, for any $x \in \mathcal{X}$ and $\hat{s}(x) := \hat{\Omega}^{1/2} p^k(x)$,

$$\sqrt{n} \frac{p(x)'(\hat{b}^k \quad ^k)}{\|\hat{s}(x)\|} \to_d N(0, 1). \tag{3.2}$$

If the approximation error is negligible relative to the estimation error, namely $\sqrt{n} r_k(x) = o(\|\Omega^{1/2} p^k(x)\|)$, then

$$\sqrt{n} \frac{\hat{g}(x) \quad g(x)}{\|\hat{s}(x)\|} \to_d N(0, 1). \tag{3.3}$$

Proposition 3.2 (Uniform Inference). *Suppose that the conditions of Theorems 4.3 and 5.5 hold. If additionally the first stage sparsity bounds of (4.10) and (4.13) hold, then so long as either the logistic propensity*

score model or linear regression model is correctly speci ed the uniform con dence intervals stated in (2.17) *satisfy*

$$\Pr\ g(x)\ 2\ [\underline{l}(x);\bar{l}(x)];\ 8x\ 2\ X\ \ = 1 \qquad + o(1):$$

Proposition 3.1 states that our method yields model-assisted pointwise inference, that is an inference procedure based on the test statistic $t(x)$ in (2.15) remains valid when using the pointwise critical value $c_p^?(1\quad)$ even if either the logistic propensity score model or linear outcome regression model is misspecified, but not both. Proposition 3.2 gives conditions for model-assisted uniform inference and shows that the uniform confidence intervals described in (2.17) remain valid under misspecification when using the uniform critical value $c_u^?(1\quad)$.

In Section 5 we provide further results for the nonparametric estimator $\hat{g}(x)$, including pointwise and uniform convergence rates. These results will rely on first-stage convergence results established in Section 4. For now, we give some intuition for how model-assisted confidence bands are obtained for the nonparametric estimator.

3.2 Managing First-Stage Bias

We focus on controlling the size of the vector $\mathbf{B}^k$, defined in (3.4), which measures the bias passed on from first-stage estimation to the second-stage estimate $\hat{b}^k$. Limiting the size of $\mathbf{B}^k$ is crucial in showing convergence of $\hat{b}^k$ to the true parameter $\quad^k$ and thus consistency of the nonparametric estimator $\hat{g}(x)$.

$$\mathbf{B}^k := \mathsf{E}_n\begin{bmatrix} p_1(X) & Y(\hat{}_1;\hat{m}_1) & Y(\bar{}_1;\bar{m}_1) \\ \vdots & & \\ p_k(X) & Y(\hat{}_k;\hat{m}_k) & Y(\bar{}_k;\bar{m}_k) \end{bmatrix}: \tag{3.4}$$

For exposition, we will consider a single term of (3.4), $\mathbf{B}_j^k$, 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 (2.5)-(2.6). However, the general intuition is the same as is used in the proofs.

In addition to the doubly robust identification property (2.9), the aIPW signal is typically useful in the high-dimensional setting because it obeys an orthogonality condition:[1]

$$\mathsf{E}[@\ \ Y(\ \ ^?;m^?)\ j\ Z] = \mathsf{E}[@_m Y(\ \ ^?;m^?)\ j\ Z] = 0: \tag{3.5}$$

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 $\bar{}_j = \ \ ^?$ and $\bar{m}_j = m\ \ $ and considering the following first order

[1] Robustness and orthogonality are indeed closely related, see Theorem 6.2 in Newey and McFadden (1994) for a discussion. The orthogonality condition is sometimes called *Neyman Orthogonality* and is discussed in depth in Chernozhukov et al. (2018); Semenova and Chernozhukov (2021).

Taylor expansion:

$$\begin{aligned}\mathbf{B}_j^k &= \mathsf{E}_n[p_j(X)Y(\hat{\gamma}_j;\hat{m}_j)] \quad \mathsf{E}_n[p_j(X)Y(\gamma^*;m^*)] \\ &= \underbrace{\mathsf{E}_n[p_j(X)\, r_{;m}\, Y(\gamma^*;m^*)]}_{O_p(n^{-1/2}) \text{ by (3.5)}} \begin{bmatrix} \hat{\gamma}_j - \gamma^* \\ \hat{m}_j - m^* \end{bmatrix} + o_p(n^{-1/2}).\end{aligned} \tag{3.6}$$

By orthogonality of the aIPW signal the gradient term is close to zero, which guarantees that the bias is $p\overline{n}$-asymptotically negligible, i.e $o_p(n^{-1/2})$, even if the nuisance parameters converge slowly to the true values, γ^* and m^*.[2] Since the bias disappears faster than the rate of convergence in Proposition 3.1, the researcher can ignore first stage nuisance parameter estimation error and treat γ^* and m^* as known when analyzing the asymptotic properties of the nonparametric estimator.

Indeed since X is a subvector of Z, under mild conditions aIPW signal orthogonality yields that the gradient $\mathsf{E}_n[p_j(X)\, r_{;m} Y(\gamma^*;m^*)]$ is asymptotically bounded uniformly for all $j = 1;\dots;k$. In this case, we would not need separate nuisance estimators for each basis function. The same consistent estimators $\hat{\gamma}$ and $\hat{m}$ could be used at each element of b^k and the analysis of (3.6) would limit the size of the entire vector $\mathbf{B}^k$. This is the approach followed by Semenova and Chernozhukov (2021).

However, if the functional form of either the propensity score model or outcome regression model is misspecified, this approach does not work. Consider the first order Taylor expansion around the probability limits $\bar{\gamma}_j$ and $\bar{m}_j$.

$$\begin{aligned}\mathbf{B}_j^k &= \mathsf{E}_n[p_j(X)Y(\hat{\gamma}_j;\hat{m}_j)] \quad \mathsf{E}_n[p_j(X)Y(\bar{\gamma}_j;\bar{m}_j)] \\ &= \mathsf{E}_n[p_j(X)\, r_{;m}\, Y(\bar{\gamma}_j;\bar{m}_j)] \begin{bmatrix} \hat{\gamma}_j - \bar{\gamma}_j \\ \hat{m}_j - \bar{m}_j \end{bmatrix} + o_p(n^{-1/2})\end{aligned} \tag{3.7}$$

So long as either one of $\bar{\gamma}_j = \gamma^*$ or $\bar{m}_j = m^*$, the double robustness of the aIPW signal (2.9) still allows for identification of $\mathsf{E}[p_j(X)Y_1]$.[3] However, the aIPW orthogonality tells us nothing about the expectation of the gradient away from the true parameters, $\gamma^*;m^*$; if either $\bar{\gamma}_j \neq \gamma^*$ or $\bar{m}_j \neq m^*$ there is no reason to believe that the gradient on the right hand side of (3.7) is mean zero. 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 rate of convergence will generally be much slower than the parameteric rate needed to esablish Proposition 3.1.

To get around this, we follow a similar approach as Tan (2020) by designing the first-stage objective functions (2.5)-(2.6) such that the resulting first-order conditions control the bias passed on to the second stage. Since we have assumed specific parametric forms for our propensity score and outcome regression models we will consider expansions around the parameters themselves.

$$\begin{aligned}\mathbf{B}_j^k &= \mathsf{E}_n[p_j(X)Y(\hat{\gamma}_j;\hat{m}_j)] \quad \mathsf{E}_n[p_j(X)Y(\bar{\gamma}_j;\bar{m}_j)] \\ &= \mathsf{E}_n[p_j(X)\, r_{\hat{\gamma}_j;\,\bar{\gamma}_j}\, Y(\bar{\gamma}_j;\bar{m}_j)] \begin{bmatrix} \hat{\gamma}_j - \bar{\gamma}_j \\ \hat{\gamma}_j - \bar{\gamma}_j \end{bmatrix} + o_p(n^{-1/2})\end{aligned} \tag{3.8}$$

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

[3] That is $\mathsf{E}_n[p_j(X)Y(\bar{\gamma}_j;\bar{m}_j)] = \mathsf{E}_n[p_j(X)Y_1]$

After substituting the forms of $\bar{}_j(z) = \quad (z; \bar{}_j)$ and $\bar{m}_j(z) = m(z; \bar{}_j)$ described in (2.4) we obtain

$$Y(\bar{}_j; \bar{m}_j) = DY(1 + e^{\quad {}^0_j Z}) \quad D(1 + e^{\quad {}^0_j Z})^{-0}Z + {}^{-0}Z:$$

Differentiating with respect to the parameters $\bar{}_j$ and $\bar{}_j$ gives

$$\mathsf{E}[p_j(X)\, r_{\quad j;\ j}\, Y(\bar{}_j; \bar{m}_j)] = \mathsf{E}\begin{bmatrix} p_j(X)De^{\quad {}^0Z}(Y \quad {}^{-0}Z)Z \\ p_j(x)fD(1 + e^{\quad {}^0Z})Z + Zg \end{bmatrix}^{\#} \tag{3.9}$$

However, by definition $\bar{}_j$ and $\bar{}_j$ solve the minimization problems defined in (2.7)-(2.8), the population analogs of our finite sample estimating equations. The first order conditions of these minimization problems yield that

$$\mathsf{E}\begin{bmatrix} \overbrace{p_j(X)fD(1 + e^{\quad {}^0Z})Z + Zg}^{\text{First order condition of } \ j} \\ p_j(X)De^{\quad {}^0Z}(DY \quad {}^{-0}Z)Z \\ \underbrace{\qquad\qquad}_{\text{First order condition of } \ j} \end{bmatrix} = 0 \quad =) \quad \mathsf{E}[p_j(X)\, r_{\quad j;\ j}\, Y(\bar{}_j; \bar{m}_j)] = 0 \tag{3.10}$$

Examining the first order conditions in (3.10), we see that they exactly give us control over the gradient (3.9). This gives us that the bias examined in expansion (3.8) is $\sqrt{n}$-asymptotically negligible even under misspecification of the propensity score or outcome regression models.[4]

Notice that the orthogonality condition obtained in (3.10) is weaker than that of (3.5). Orthogonality is obtained with respect to the parameters of the logistic propensity score model and linear outcome model rather than the nuisance functions themselves. The bias is controlled only when approaching the limiting values along specific parameteric paths; hence the term model-assisted inference. Because we are assuming constant parametric models, this is exactly what is needed to recover valid inference.

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

$$\begin{aligned} \mathsf{E}[p_j(X)\, r_{\quad j;\ j}\, Y(\bar{}_j; \bar{m}_j)] = \mathsf{E}\begin{bmatrix} p_j(X)De^{\quad {}^0Z}(Y \quad {}^{-0}Z)Z \\ p_j(X)fD(1 + e^{\quad {}^0Z})Z + Zg \end{bmatrix} \\ \neq \mathsf{E}\begin{bmatrix} \overbrace{p_l(X)fD(1 + e^{\quad {}^0Z})Z + Zg}^{\text{First order condition of } \ l} \\ p_l(X)De^{\quad {}^0Z}(Y \quad {}^{-0}Z)Z \\ \underbrace{\qquad\qquad}_{\text{First order condition of } \ l} \end{bmatrix} = 0: \end{aligned} \tag{3.11}$$

Obtaining valid inference at all points $x \in X$ under misspecification as in Proposition 3.1 requires showing $\sqrt{n}$-negligible first stage estimation bias for any linear combination of the vector (3.4). 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] Correct specification is only necessary for identification.

Of course it is not enough to analyze each term of $\mathbf{B}^k$ individually. Showing Proposition 3.1 and the rest of our nonparametric results we will require that these terms, and thus the niusance parameter estimators, converge uniformly. Section 4 provides conditions under which this uniform convergence is obtained.

4 First-Stage Convergence

In order to show nonparameteric rate convergence of the estimator described in Section 2, we seek to establish the following condition, which limits the bias passed from first-stage estimation onto the second-stage estimator.

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

$$\max_{1 \quad j \quad k} \mathsf{E}_n[p_j(X)Y(\mathsf{b}_j; \hat{m}_j)] \quad \mathsf{E}_n[p_j(X)Y(\bar{}_j; \bar{m}_j)] \; = o_p(n^{\;1=2}k^{\;1=2}): \tag{4.1}$$

In particular, Condition 1 implies that the bias vector $\mathbf{B}^k$ from (3.4) satisfies $k\mathbf{B}^k k_2 = o_p(n^{\;1=2})$. Establishing this condition largely follows the logic laid out in Section 3 after establishing the uniform consistency of the first-stage estimators, b_j and b_j, described in (2.5)-(2.6), to their corresponding probability limits. This section will establish constraints under which this uniform first-stage convergence is obtained. These provisions are similar to those existing in the $\grave{}_1$-penalized estimation literature (Bickel et al., 2009; Belloni and Chernozhukov, 2013; Chetverikov and Sørensen, 2021; Bühlmann and van de Geer, 2011).

4.1 Penalty Parameter Selection

We improve on existing work on model-assisted inference by incorporating a data dependent two step penalty parameter selection procedure based on the work of Chetverikov and Sørensen (2021). Before presenting uniform convergence results it is useful to describe this procedure.

For each $j = 0; 1 : : : ; k;$ we start with a pilot penalty parameters given by

$$^{\text{pilot}}_{;j} = c_{\;;j} \quad \sqrt{\frac{\ln^3(d_z)}{n}} \quad \text{and} \quad ^{\text{pilot}}_{;j} = c_{\;;j} \quad \sqrt{\frac{\ln^3(d_z)}{n}} \tag{4.2}$$

for some constants $c_{\;;j}; c_{\;;j}$ selected from the interval $[\underline{c}_n; \bar{c}_n]$ with $\underline{c}_n > 0$. Using $^{\text{pilot}}_{;j}$ and $^{\text{pilot}}_{;j}$ in lieu of $_{;j}$ and $_{;j}$ in (2.5)-(2.6) we generate pilot estimators $\mathsf{b}_j^{\text{pilot}}$ and $\mathsf{b}_j^{\text{pilot}}$. These pilot estimators are used to generate plug in estimators $\hat{\theta}_{\;;j}$ and $\hat{\theta}_{\;;j}$ of the residuals

$$\begin{aligned} \hat{\theta}_{\;;j} &:= \quad p_j(X)\{De^{\;\mathsf{b}_j^{\text{pilot}'}Z} + (1 \quad D)\} \\ \hat{\theta}_{\;;j} &:= p_j(X)De^{\;\mathsf{b}_j^{\text{pilot}'}Z}(Y \quad \mathsf{b}_j^{\text{pilot}'}Z): \end{aligned} \tag{4.3}$$

We then use a multiplier bootstrap procedure to select our final penalty parameters $_{;j}$ and $_{;j}$.

$$\begin{aligned} _{;j} &= c_0 \quad (1 \quad)\text{-quantile of} \max_{1 \quad l \quad d_z} |\mathsf{E}_n[e_i\hat{\theta}_{\;;j}Z_l]| \text{ given } \{Y_i; D_i; Z_i\}_{i=1}^n; \\ _{;j} &= c_0 \quad (1 \quad)\text{-quantile of} \max_{1 \quad l \quad d_z} |\mathsf{E}_n[e_i\hat{\theta}_{\;;j}Z_l]| \text{ given } \{Y_i; D_i; Z_i\}_{i=1}^n \end{aligned} \tag{4.4}$$

where $e_1; : : : ; e_n$ are independent standard normal random variables generated independently of the data

$fY_i; D_i; X_i g_{i=1}^n$ and $c_0 > 1$ is a fixed constant[1]. In line with other work we take $c_0 = 1.1$ in simulations. So long as our residual estimates converge in empirical mean square to limiting values,

$$\begin{aligned} U_{;j} &:= p_j(X)fDe^{\beta_j^{pilot'}Z} + (1 - D)g \\ U_{;j} &:= p_j(X)De^{\beta_j^{pilot'}Z}(Y - e^{\beta_j^{pilot'}Z}); \end{aligned} \tag{4.5}$$

the choice of penalty parameter in (4.4) will ensure that the penalty parameter dominates the noise with high probability. This allows for consistent variable selection and coefficient estimation.

Remark 4.1 (Pilot Penalty Selection)**.** In practice, the researcher has a fair bit of flexibility in choosing the constants $c_{;j}$ and $c_{;j}$ in (4.2). The optimal choice of these constants may depend on the underlying data degenerating process. We recommend using cross validation to pick theses constants from a fixed-cardinality set of possible values. In line with Assumption 4.1(v), the values in the set should be chosen to be on the order of the maximum value of $kp^k(X_i)k_1$ observed in the data. So long as this set of possible values is bounded and of cardinality on an order less than n, the results of this section hold for a cross-validated choice of these constants.

4.2 Uniform First-Stage Convergence

With the penalty parameter selection procedure established, we now demonstrate uniform convergence of our first-stage estimators. To do so, we will rely on the following first-stage assumptions. The first assumption will allow us to generate a uniform bound on the convergence rate of the logistic propensity score model parameters.

Assumption 4.1 (Logistic Model Convergence)**.**

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

(ii) There is a constant B_0 such that $|\beta_j' Z| \le B_0$ almost surely for all j.

(iii) There exist fixed constants $_0 > 1$ and $1 > _0 > 0$ such that for each $j = 1; \dots; l$ the following empirical compatability condition holds for the empirical hessian matrix $\tilde{\Sigma}_{;j} := \mathsf{E}_n[De^{\beta_j^{0'}Z}ZZ']$. For any $b \in \mathbb{R}^{d_z}$

$$\sum_{l \in S_{;j}} jb_l j \le \sqrt{_0 s_{;j}} \sum_{l \in S_{;j}} jb_l j =) \frac{2}{0} \sum_{l \in S_{;j}} jb_l j^2 \le s_{;j} j b' \tilde{\Sigma}_{;j} b :$$

(iv) There exist fixed constants c_u and $C_U > 0$ such that for all $j \le k$, $\mathsf{E}[U_{;j}^4] \le (_{k;1} C_U)^4$ and $\min_{1 \le l \le d_z} \mathsf{E}[U_{;j}^2 Z_l^2] \ge c_u$.

(v) The following sparsity bounds hold with $s_{k;} := \max_{1 \le j \le k} js_{;j} j$.

$$\frac{_{k;1}}{\underline{c}_n \ln(d_z n)} \to 0; \quad \frac{_{k;1} s_{k;} \bar{c}_n^2 \ln^5(d_z n)}{n} \to 0; \quad \text{and} \quad \frac{^4_{k;1} \ln^7(d_z k n)}{n} \to 0:$$

The first part of Assumption 4.1 assumes that the regressors are bounded. It can be relaxed with sufficient assumptions on the tail behavior of the controls. We should note that compactness of X is generally required

[1]The constant c_0 can actually be different for the propensity score model and for the outcome regression model, and can also vary for each $j = 1; \dots; k$. All that matters is that each constant satisfeis the requirements on c_0 in Theorem 4.3. This complicates notation, however, so the constant is assumed to be the same across estimation procedures.

by nonparametric estimators. The second assumption bounds all limiting propensity scores $\bar{}_j(Z)$ away from zero uniformly. The third assumption is an empirical compatability condition on the weighted first-stage design matrix. It is slightly weaker than the restricted eigenvalue conditions often assumed in the lterature (Bickel et al., 2009; Belloni et al., 2012). The penultimate condition is an identifiability constraint that limits the moments of the nosie and bounds it away from zero uniformly over all estimation procedures. The last condition requires 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, $_{k;1}$ which is to ensure uniform stability of the estimation procedures (2.5)-(2.6).

The next set of assumptions enable us to generate a uniform bound on the convergence rate of the linear outcome regression model parameters. The assumptions needed for this convergence are a bit stronger than those needed for the propensity score model convergence in order to manage the reliance of outcome regression estimates on the propensity model parameter estimates in (2.6).

Assumption 4.2 (Linear Model Convergence)**.**

(i) The errors $Y_1 \quad \bar{m}_j(Z)$ are uniformly subgaussian conditional on Z in the following sense. There exist fixed positive constants D_0 and D_1 such that for any j:

$$D_0\mathsf{E}\left[\exp\left(fY_1 \quad \bar{m}_j(Z)g^2=D_0^2\right) \quad 1\,j\,Z\right] \quad D_1^2$$

almost surely.

(ii) For any j the following empirical compatibility condition holds with fixed constants $_0 > 1$ and $1 \; _0 > 0$ for the empirical hessian matrix $\tilde{\Sigma}_{;j}$:

$$\sum_{l \in S_{;j}} jb_l j \quad {}_0 \sum_{l \in S_{;j}} jb_l j =) \quad {}_0^2 \sum_{l \in S_{;j}} jb_l j^2 \quad jS_{;j} j \; b^0 \tilde{\Sigma}_{;j} b \; :$$

(iii) For all $j \quad k$, $\mathsf{E}[U^4_{;j}] \quad (_{k;1} C_U)^4$ and $\min_{1 \; l \; d_z} \mathsf{E}[U^2_{;j} Z_l^2] \quad c_u$.

(iv) The following sparsity bound holds with $s_{k;} = \max_{1 \; j \; k} jS_{;j}j$:

$$\frac{{}^2_{k;1} s^2_{k;} \; \bar{c}_n^2 \ln^5(d_z n)}{n} \; ! \; 0:$$

The first condition of Assumption 4.2 assumes that tail behavior of the outcome regression errors are uniformly thin. The second and third conditions are the same empirical compatability condition and identification conditions as in Assumption 4.1 but applied to the parameters and residuals for the outcome regression. The final part of Assumption 4.2 is a sparsity bounded needed for validity of the penalty parameter selection procedure and mirrors those of Chetverikov and Sørensen (2021).

Remark 4.2 (Constants)**.** Many of the constants in Assumptions 4.1 and 4.2 are assumed to be fixed across all j. This is mainly to simplify the exposition of the results below. In practice, all constants can be allowed to grow slowly as the number of basis terms increases. However, the growth rate of these terms affects the first-stage sparsity bounds needed needed to show Condition 1.

We can now establish uniform convergence of the propensity score morel and the outcome regression model parameters. The convergence will be presented in terms of the ℓ_1 norm of the coefficient vector as well as the symmetrized Bregman divergences, $D^z_{;j}(\beta_j; \bar{}_j)$ and $D^z_{;j}(\beta_j; \bar{}_j; \ _j)$, defined by

$$\begin{aligned} D^z_{;j}(\beta_j; \bar{}_j) &:= \mathsf{E}_n\left[p_j(X) D f e^{\ \beta_j^0 Z} \quad e^{\ {}^0_j Z} g f^{-{}^0_j} Z \quad \beta_j^0 Z g \right]; \\ D^z_{;j}(\beta_j; \bar{}_j; \beta) &:= \mathsf{E}_n\left[p_j(X) D e^{\ \beta_j^0 Z} ({}^{-{}^0_j} Z \quad \beta_j^0 Z)^2 \right]: \end{aligned} \tag{4.6}$$

We refer readers to discussion in Tan (2017) for details and motiviation. For now it suffices to note that the Bregman divergence is the error resulting from approximating the non-penalized loss function at the estimated value with a first order Taylor expansion of the non-penalizd loss function at the true values. Because our loss functions are convex, these errors will always be positive. Bounds on the Bregman divergence help directly control second order terms in the remainder of (3.8).

Theorems 4.1 and 4.2 establish uniform convergence of the propensity score and outcome regression parameters at nearly the near-oracle rate from Belloni and Chernozhukov (2013). Theorem 4.2 relies on a condition on the ratio between the penalty parameters that is discussed in Remark 4.3.

Theorem 4.1 (Propensity Score Model)**.** *Suppose that Assumption 4.1 holds, that* $_0 > (c_0 + 1)/(c_0 \quad 1)$*,* $k/n \ ! \ 0$*, and that* $k \ ! \ 0$ *as* $n \ ! \ 1$*. Then the logistic propensity score model parameters are uniformly bounded:*

$$\sup_{1 \ j \ k} D^z_{;j}(\beta_j; \bar{}_j) \ . _P \ {}^2_{k;1} \frac{s_k \ln(d_z)}{n} \quad \text{and} \quad \sup_{1 \ j \ k} k\beta_j \quad \bar{}_j k_1 \ . _P \ _{k;1} \sqrt{\frac{s_k^2 \ln(d_z)}{n}} \tag{4.7}$$

Theorem 4.2 (Outcome Regression Model)**.** *Suppose that Assumptions 4.1 and 4.2 hold. In addition assume that* $_0 > (c_0 + 1)/(c_0 \quad 1)$*,* $k/n \ ! \ 0$*,* $k \ ! \ 0$*, and that there is a fixed constant* $c > 0$ *such that for all* j*,* $_{;j} / _{;j} \quad c$*. Then the linear outcome regression model parameters are uniformly bounded:*

$$\sup_{1 \ j \ k} D^z_{;j}(\beta_j; \bar{}_j; \bar{}_j) \ . _P \ {}^2_{k;1} \frac{s_k \ln(d_z)}{n} \quad \text{and} \quad \sup_{1 \ j \ k} k\beta_j \quad \bar{}_j k_1 \ . _P \ _{k;1} \sqrt{\frac{s_k^2 \ln(d_z)}{n}} \tag{4.8}$$

Remark 4.3 (Penalty Parameter Ratio)**.** The existence of a fixed constant $c > 0$ such $_{;j} / _{;j} \quad c$ may seem a bit unnatural, but it can be enforced in practice by setting the penalty to instead be, say, $^{\text{ratio}}_{;j} := \max\{ _{;j}/5; \ _{;j}\}$. The use of this alternative penalty parameter instead of $_{;j}$ does not effect any of the convergence results established in this document. In simulations, we find this constraint is rarely binding.

With the uniform convergence of the logistic regression parameters established, we can apply the logic of Section 3 to show convergence of the entire bias vector $\mathbf{B}^k$ (3.4). Again, it should be emphasized that the result of Theorem 4.3 could not be obtained using standard maximum likelihood estimating equations for the logistic propensity score parameters or on ordinary least squares loss function for the outcome regression parameters. Moreover, the rate of convergence of Theorem 4.3 would not be obtained by using the estimating equations of Tan (2020). It is the particular form of the estimating equations in (2.5)-(2.6) that allows us to show a suitable rate of convergence.

Theorem 4.3 (First-Stage Convergence)**.** *Suppose that Assumptions 4.1 and 4.2. In addition assume that* $c_0 > (_0 + 1)/(_0 \quad 1)$*,* $k/n \ ! \ 0$*,* $k \ ! \ 0$*, and there is a fixed constant* $c > 0$ *such that for all* j*,* $_{;j} / _{;j} \quad c$*.*

Then the elements of the vector (3.4) *converge uniformly:*

$$\max_{1 \le j \le k} \left| \mathsf{E}_n[p_j(X) Y(\mathsf{b}_j; \hat{m}_j)] - \mathsf{E}_n[p_j(X) Y(\bar{\beta}_j; \bar{m}_j)] \right| \cdot_P \frac{s_k \sqrt{\sum_{k=1}^{2} \ln(d_z)}}{n} \tag{4.9}$$

In particular, Condition 1 holds under the following additional sparsity bound

$$\frac{s_k \sqrt{\sum_{k=1}^{2} k^{1+2} \ln(d_z)}}{\sqrt{n}} \to 0. \tag{4.10}$$

For implementable inference using $\hat{\Omega}$ to estimate the second-stage variance as in Proposition 3.1, we additionally wish to establish the following condition uniformly bounding the rate of convergence of the second moments for elements $\mathbf{B}_j^k$ of the bias vector $\mathbf{B}^k$. Under Condition 2 and additional second-stage assumptions, Theorem 5.4 shows that the plug-in estimator for the matrices described in (2.13) is consistent.

Condition 2 (Variance Estimation)**.** Let $m > 2$ be as in Assumption 5.3. Then,

$$\max_{1 \le k \le 1} \max_{1 \le j \le k} \mathsf{E}_n[p_j(X)^2 (Y(\mathsf{b}_j; \hat{m}_j) - Y(\bar{\beta}_j; \bar{m}_j))^2] = o_p(k^{-2} n^{-1+m}) \tag{4.11}$$

Theorem 4.4 (Second Moment Convergence)**.** *Suppose that the conditions of Theorem 4.3 hold. Then the elements of the vector* (3.4) *converge uniformly in empirical mean square at the following rate:*

$$\max_{1 \le j \le k} \mathsf{E}_n[p_j^2(X) (Y(\mathsf{b}_j; \hat{m}_j) - Y(\bar{\beta}_j; \bar{m}_j))^2] \cdot_P \sum_{k=1}^{4} \frac{s_k^2 \ln(d_z)}{n} \tag{4.12}$$

In particular, Condition 2 holds if additionally

$$\frac{\sum_{k=1}^{5} s_k^2 k^2 \ln(d_z)}{n^{(m-1)/m}} \to 0. \tag{4.13}$$

Remark 4.4. The sparsity bound and conditions of Theorem 5.4 can be significantly weakened 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 we could more directly follow Semenova and Chernozhukov (2021) and use that with high probability the quantities:

$$\max_{1 \le i \le n} \max_{1 \le j \le k} |Y_i(\mathsf{b}_j; \hat{m}_j) - Y_i(\bar{\beta}_j; \bar{m}_j)| \quad \text{and} \quad \max_{1 \le i \le n} \max_{1 \le j \le k} |Y_i(\mathsf{b}_j; \hat{m}_j) - Y_i(\bar{\beta}_j; \bar{m}_j)|^2$$

are controlled by bounds that converge to zero. This could then be used in the proof of the matrix estimation result to show that the first-stage estimation error passed on to the variance estimator is negligible. Unfortunately, because we are not using cross-fitting, we cannot easily say anything about the behavior of the random variable Y on the event that $\|\mathsf{b}_j - \bar{\beta}_j\|$ and $\|\hat{m}_j - \bar{m}_j\|$ are small. Because the difference $Y_i(\mathsf{b}_j; \hat{m}_j) - Y(\bar{\beta}_j; \bar{m}_j)$ depends on Y, not knowing it's expectation on the event that b_j is close to $\bar{\beta}_j$ and $\hat{m}_j$ is close to $\bar{m}_j$ is problematic.

Remark 4.5 (Alternative Penalty Parameter Selection)**.** The additional sparsity assumptions needed by Assumptions 4.1 and 4.2 are stronger than what is required by Tan (2020) who only requires $s \sqrt{\ln(d_z)/n} \to 0$, where s denotes the maximum sparsity index. This is due to the use of the data dependent penalty parameter

selection procedure described in Section 4.1 as opposed to the theoretical penalty parameter of Tan (2020). Under a similar theoretical penalty parameter we could establish Condition 1 under only the sparsity bound in (4.10). However, in line with other work, we find that in practice the data dependent method performs better than a theoretical penalty and only requires the researcher have minimal knowledge of the underlying data generating process.

For various reasons, the researcher may not want to implement the bootstrap penalty parameter procedure described in Section 4.1. If this is the case, we note that the pilot penalty parameters of (4.2) can be used directly. The constants $c_{;j}$ and $c_{;j}$ can be selected via cross validation from a growing set Λ_n $[\underline{c}_n; \bar{c}_n]$. If the penalty parameters are selected in this fashion, Conditions 1 and 2 can be obtained under Assumption 4.1(i)-(iv), Assumption 4.2(i)-(iii) and the following modified rate conditions:

$$\frac{k\,j\Lambda_n j}{n^2} \;!\; 0; \quad \frac{k;\,1}{\underline{c}_n \ln(d_z n)} \;!\; 0; \quad \frac{\bar{c}_n^2 s_k k^{1=2} \ln^3(d_z n)}{p\overline{n}} \;!\; 0; \quad \text{and} \quad \frac{\sum_{k;\,1}^{2} \bar{c}_n^2 s_k k^2 \ln^3(d_z n)}{n^{(m\ \ 1)=m}} \;!\; 0 \tag{4.14}$$

These conditions are comparable to the rates in Assumptions 4.1 and 4.2 and equations (4.10)-(4.13). However, the rate of convergence on the bias established in Theorem 4.3 is somewhat slower. Appendix D provides details for implementation as well as formally shows the bounds in (4.14).

5 Additional Second-Stage Results

In this section, we present a set consistency and distributional results for our second-stage estimator $\hat{g}(x)$ described in Section 2. Additional results, including pointwise and uniform linearization lemmas, can be found in Appendix C. These results will be presented under the assumption that Conditions 1-2 from Section 4 hold and for the most part will follow exsting results in the nonparametric literature (Newey, 1997; Belloni et al., 2015). The main remaining difficulty is managing the k different error terms presented in (2.10). This requires some modification to the second-stage results.

In the standard setting, consistency of nonparametric estimator relies on certain conditions on the error terms. In our setting, we will 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 logistitc propensity score model is correctly specified and the limiting outcome regression models are uniformly bounded conditional on Z, 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.

5.1 Pointwise Inference

Pointwise inference will rely on the following assumptions in tandem with Condition 1. Assumption 5.1 is all that is needed for the estimator to achieve the nonparametric rate of convergence even under misspecification, while Assumption 5.2 will ensure that the second stage estimator is pointwise asymptotically linear (see Appendix C).

Assumption 5.1 (Second-Stage Consistency Assumptions)**.**

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

(ii) The conditional variance of the error terms is uniformly bounded in the following sense. There exist constants $\_^2$ and $^{-2}$ such that for any $j = 1; 2 :::$ we have that $\_^2 \quad \mathrm{Var}(\ _j \ j \ X) \quad ^{-2} < 1$;

(iii) For each n and k there are finite constants c_k and $`_k$ such that for each $f \ 2 \ G$

$$kr_k k_{L;2} = (\mathsf{E}[r_k(x)^2])^{1=2} \quad c_k \quad \text{and} \quad kr_k k_{L;1} = \sup_{x2X} jr_k(x)j \quad `_k c_k:$$

Assumption 5.2 (Uniform Integrability)**.** Let $^{-}_k := \max_{1 \ j \ k} j \ _j j$. Assume that

(i) $\sup_{x2X} \mathsf{E}[^{-2}_k \mathbf{1} f^{-}_k + `_k c_k > \ ^p\overline{n} = \ _k g \ j \ X = x] \ ! \ 0$ as $n \ ! \ 1$ for any > 0.

(ii) $\sup_{x2X} \mathsf{E}[`^2_k c^2_k \mathbf{1} f^{-}_k + `_k c_k > \ ^p\overline{n} = \ _k g \ j \ X = x] \ ! \ 0$ as $n \ ! \ 1$ for any > 0.

As mentioned, note that 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 $j = 1; :::; k$. 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]

Theorem 5.1 (Pointwise Normality)**.** *Suppose that Condition 1, Assumption 5.1, and Assumption 5.2 holds. In addition suppose that* $^2_k \log k=n \ ! \ 0$. *Then so long as either the logistic propensity score model or linear outcome regression model is correctly speci ed, for any* $\quad 2 \ S^{k \ 1}$:

$$\frac{^p\overline{n}(\mathsf{b}k \quad \ ^k)}{k \ ^0\Omega^{1=2}k} \ !_d N(0; 1) \tag{5.1}$$

where generally $\Omega = \tilde{\Omega}$ *but if* $`_k c_k \ ! \ 0$ *then we can set* $\Omega = \Omega_0$. *Moreover, for any* $x \ 2 \ X$ *and* $s(x) := \Omega^{1=2} p^k(x)$,

$$\frac{^p\overline{n}p^k(x)^0(\mathsf{b}k \quad \ ^k)}{ks(x)k} \ !_d N(0; 1) \tag{5.2}$$

and if the approximation error is negligible relative to the estimation error, namely $^p\overline{n} r_k(x) = o(ks(x)k)$, *then*

$$\frac{^p\overline{n}[\mathsf{b}g(x) \quad g(x)}{ks(x)k} \ !_d N(0; 1) \tag{5.3}$$

In conjunction with the verification of Condition 1 established in Section 4, Theorem 5.1 verifies the validity of the pointwise inference procedure defined in Section 2 based on the critical values $c^?_p(1 \quad)$ up to procurement of a consistent estimator for the variance matrix Ω: Theorem 5.4 will provide conditions under which the plug in estimator $\mathsf{b}\Omega$ defined in (2.14) is consistent.

5.2 Uniform Convergence

Next, we turn to strengthening the pointwise results to hold uniformly over all points $x \ 2 \ X$. This will require stronger conditions. We will make the following assumptions on the tail behavior of the error terms which strengthen Assumption 5.2.

[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 (2.5)-(2.6).

Assumption 5.3 (Uniform Limit Theory)**.** Let $\bar{\ell}_k = \sup_{1 \quad j \quad k} j \; j$, $(x) := p^k(x)/kp^k(x)k$, and let

$$\frac{L}{k} := \sup_{\substack{x;x^0 2X \\ x6=x^0}} \frac{k \quad (x) \qquad (x^0)k}{kx \quad x^0k}.$$

Further for any integer s let $\bar{\ell}_k^s = \sup_{x2X} \mathsf{E}[j\bar{\ell}_k j^s jX = x]$. For some $m > 2$ assume

(i) The regression errors satisfy $\sup_{x2X} \mathsf{E}[\max_{1 \quad i \quad n} j\bar{\ell}_{k;i} j^m \; j \; X = x] \; . \; n^{1=m}$

(ii) The basis functions are such that (a) $k^{2m=(m \quad 2)} \log k/n \; . \; 1$, (b) $(\bar{\ell}_k^2 \_ \bar{\ell}_k^m) \log \frac{L}{k} \; . \; \log k$, and (c) $\log \bar{\ell}_k^m \quad _k \; . \; \log k$.

As before, Assumption 5.3 is very similar to its analogue in Belloni et al. (2015), with the modification that the conditions are required to hold for $\bar{\ell}_k$ as opposed to $_k$. Under this assumption, we derive doubly-robust uniform rates of convergence and model-assisted uniform inference procedures for the conditional counterfactual outcome $g_0(x)$.

Theorem 5.2 (Uniform Rate of Convergence)**.** *Suppose that Condition 1 and Assumptions 5.1-5.3 hold. Then so long as either the propensity score model or outcome regression model is correctly speci ed:*

$$\sup_{x2X} \quad (x)^0 \mathsf{G}_n[p^k(x) \quad \quad ^k] \; . \; _P \; \sqrt{\log k} \tag{5.4}$$

Moreover, for

$$\bar{R}_{1n} := \sqrt{\frac{2 \log k}{n}} (n^{1=m} \sqrt{\log k} + p\bar{k} \grave{}_k c_k)$$
$$\bar{R}_{2n} := \sqrt{\log k} \; \grave{}_k c_k$$

we have that

$$\sup_{x2X} \; p^k(x)^0 (\mathsf{b}^k \quad \quad ^k) \; . \; _P \; p\frac{k}{n} \; \sqrt{\log k} + \bar{R}_{1n} + \bar{R}_{2n} \tag{5.5}$$

and

$$\sup_{x2X} \; \mathsf{b}g(x) \quad g_0(x) \; . \; _P \; p\frac{k}{n} \; \sqrt{\log k} + \bar{R}_{1n} + \bar{R}_{2n} \quad + \grave{}_k c_k \tag{5.6}$$

Theorem 5.2 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 (5.5)-(5.6). If these two terms go to zero, the entire estimator will converge uniformly to the true conditional expectation of interest.

Theorem 5.3 (Strong Approximation by a Gaussian Process)**.** *Assume that Condition 1 holds and that Assumptions 5.1-5.3 hold with $m \quad 3$. In addition assume that (i) $\bar{R}_{1n} = o_p(a_n^{\;1})$ and (ii) $a_n^6 k^4 \; k(\bar{\ell}_k^3 + \grave{}_k^3 c_k^2)^2 \log^2 n/n \; ! \; 0$. Then so long as either the propensity score model or outcome regression model is correctly speci ed, for some $N_k \quad N(0; I_k)$:*

$$p\overline{n} \frac{(x)^0(\mathsf{b} \quad)}{k \; (x)^0 \Omega^{1=2} k} =_d \frac{(x)^0 \Omega^{1=2}}{k \; (x)^0 \Omega^{1=2} k} N_k + o_p(a_n^{\;1}) \;\; in \;\; \grave{}^{\;1}(X) \tag{5.7}$$

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

$$\sqrt{n}\frac{p^k(x)'(\hat{b} - b)}{\|s(x)\|} =_d \frac{s(x)}{\|s(x)\|} N_k + o_p(a_n^{-1}) \text{ in } \ell^1(X) \tag{5.8}$$

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

$$\sqrt{n}\frac{\hat{g}(x) - g(x)}{\|s(x)\|} =_d \frac{s(x)'}{\|s(x)\|} N_k + o_p(a_n^{-1}) \text{ in } \ell^1(X) \tag{5.9}$$

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* (2.13).

Theorem 5.3 establishes conditions under which uniform confidence intervals can be created for the function $g_0(x)$. As noted by Belloni et al. (2015), this is distinctly different from a Donsker type convergence result for the estimator $\hat{g}(x)$ as viewed as a random element of $\ell^1(X)$. In particular, the covariance kernel is left completely unspecified and in general need not be well behaved.

5.3 Matrix Estimation and Uniform Inference

The final second-stage results presented establish the consistency of various matrix estimations and validity of the uniform confidence intervals described in Section 2. The matrix estimation results will allow for consistent estimation of the asymptotic variances in Theorem 5.1 and Theorem 5.3 and thus allow for practically implementable, asymptotically valid confidence intervals for $g_0(x)$, even under misspecification of one of the nuisance parameter models.

The matrices $Q = E[p^k(x)p^k(x)']$ and $\Sigma = E[\{p^k(x)(\hat{\gamma}_j - \gamma_j) + r_k\}\{p^k(x)(\hat{\gamma}_l - \gamma_l) + r_k\}]$ are estimated using

$$\begin{aligned} \hat{Q} &= E_n[p^k(x)p^k(x)'] \\ \hat{\Sigma} &= E_n[\{p^k(x)'(\hat{b}^k - b^k) + \hat{r}_k\}\{p^k(x)'(\hat{b}^k - b^k) + \hat{r}_k\}] \end{aligned} \tag{5.10}$$

where $b^k := (b_1, \ldots, b_k)$ and $\hat{\gamma}_j = Y(b_j, \hat{m}_j) - \hat{g}(x)$ for each $j = 1, \ldots, k$. In addition define $v_n = (E[\max_{1 \le i \le n} \|\hat{\gamma}_i - \gamma_i\|^2])^{1=2}$.

Theorem 5.4 (Matrix Estimation)**.** *Suppose that Conditions 1 and 2 and Assumptions 5.1-5.3 hold. In addition, assume that* $\bar{R}_{1n} + \bar{R}_{2n} \cdot (\log k)^{1=2}$. *Then, so long as either the propensity score model or outcome regression model is correctly specified:*

$$\begin{aligned} \|\hat{Q} - Q\| &\cdot_P \sqrt{\frac{\zeta_k^2 \log k}{n}} = o(1) \text{ and} \\ \|\hat{\Sigma} - \Sigma\| &\cdot_P (v_n \vee \zeta_k c_k) \sqrt{\frac{\zeta_k^2 \log k}{n}} = o(1) \end{aligned}$$

Moreover, for $\hat{\Omega} = \hat{Q}^{-1}\hat{\Sigma}\hat{Q}^{-1}$:

$$\|\hat{\Omega} - \Omega\| \cdot_P (v_n \vee \zeta_k c_k) \sqrt{\frac{\zeta_k^2 \log k}{n}} = o(1)$$

We now establish the validity of uniform inference based on the gaussian bootstrap critical values $c_u^*(1 - \alpha)$

defined in Section 2.

Theorem 5.5 (Validity of Uniform Confidence Bands)**.** *Suppose Conditions 1 and 2 are satis ed and Assumptions 5.1{5.3 hold with m 4. In addition suppose (i) $R_{1n} + R_{2n} \cdot \log^{1=2} n$, (ii) $_k \log^2 n=n^{1=2}$ $^{1=m} = o(1)$, (iii) $\sup_{x2X} jr_k(x)j=kp^k(x)k = o(\log^{1=2} n)$, and (iv) $k^4 \frac{2}{k}(1 + l_k^3 r_k^3)^2 \log^5 n=n = o(1)$. Then for $t(x)$ de ned in (2.15), so long as either the propensity score model or outcome regression model is satis ed*

$$\Pr\left(\sup_{x2X} jt(x)j \quad c^?(1 \quad) \right) = 1 \quad + o(1):$$

As a result, uniform con dence intervals formed in (2.17) satisfy

$$\Pr(g(x) \ 2 \ [\underline{i}(x); \bar{i}(x)]; \ 8x \ 2 \ X) = 1 \quad + o(1):$$

The width of the con dence bands $2c^?(1 \quad)\mathsf{b}(x)$ satisfy

$$2c^?(1 \quad)\mathsf{b}(x) \cdot_P (x) \sqrt{\log n} \cdot \sqrt{\frac{\frac{2}{k}\log n}{n}}:$$

In conjunction with theorems 4.3, 4.4 and 5.4, this shows Proposition 3.2. It also shows that the confidence intervals based on the uniform critical values $c_u^?(1 \quad)$ shrink at the uniform nonparametric rate derived in Theorem 5.2.

6 Estimation of the Conditional Average Treatment E ect

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

$$g_0(x) = \mathsf{E}[Y_1 \ j \ X = x]:$$

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

$$\tilde{g}_0(x) = \mathsf{E}[Y_0 \ j \ X = x]:$$

To do so, we use the alternate aIPW signal

$$Y_0(\ _0; m_0) = \frac{(1 \quad D)Y}{1 \quad _0(Z)} + \frac{1 \quad D}{1 \quad _0(Z)} \quad 1 \quad m_0(Z)$$

where as before the true value for $_0^?(z) = \Pr(D = 1 \ j \ Z = z)$ but now $m_0^?(z) = \mathsf{E}[Y \ j \ D = 0; Z = z]$. To estimate these nuisance models we again assume a logistic form for the propensity score model $_0(z) = (z; \ ^0)$ and a linear form for the outcome regression model $m_0(z) = m(z; \ ^0)$ as in (2.4) 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 β_j^0 differ from those in (2.5)-(2.6) however, and are instead given

$$\mathsf{b}_j^0 := \arg\min \mathsf{E}_n[p_j(X)f(1 \quad D)e^{\ {}^0Z} \quad D\ {}^0Zg] + \quad {}_{;j}k\ k_1$$

$$\mathsf{b}_j^0 := \arg\min \mathsf{E}_n[p_j(Z)(1 \quad D)e^{\mathsf{b}_j^{0\,0}Z}(Y \quad {}^0Z)^2]{=}2 + \quad {}_{;j}k\ k_1$$

which under the natural analogs of Assumptions 4.1 and 4.2 converge uniformly to population minimizers:

$$\bar{}_j^0 := \arg\min \mathsf{E}[p_j(X)f(1 \quad D)e^{\ {}^0Z} \quad D\ {}^0Zg]$$

$$\bar{}_j^0 := \arg\min \mathsf{E}[p_j(Z)(1 \quad D)e^{\ {}_j^{0\,0}Z}(Y \quad {}^0Z)^2]$$

Letting $\bar{}_{0;j}(z) = \quad (z; \bar{}_j^0)$, and $\bar{m}_{0;j}(z) = m(z; \bar{}_j^0)$ we can repeat the decomposition of Section 3, expressing $\tilde{Y}(\bar{}_{0;j}; \bar{m}_{0;j})$ as functions of the parameters $\bar{}_j^0$ and $\bar{}_j^0$ and show that the first order conditions for $\bar{}_j^0$ and $\bar{}_j^0$ 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 4 and 5. Combining estimation and inference of the two conditional counterfactual outcomes then gives a doubly-robust estimator and model-assisted inference procedure for the CATE. To perform inference on the CATE we can use the variance matrix

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

where Ω_0 is as in (2.13) but Ω_1 and Ω_2 are given

$$\begin{aligned} \Omega_1 &= Q\ {}^1\mathsf{E}[fp^k(x) \quad {}_0^k gfp^k(x) \quad {}_0^k g^0]Q\ {}^1 \\ \Omega_2 &= Q\ {}^1\mathsf{E}[fp^k(x) \quad {}^k gfp^k(x) \quad {}_0^k g^0]Q\ {}^1 \end{aligned} \tag{6.1}$$

where ${}_{0;j}^k = Y_0(\bar{}_{0;j}; \bar{m}_{0;j}) \quad \tilde{g}_0(x)$ and ${}_0^k = ({}_{0;1}^k; \ldots; {}_{0;k}^k)^0$. These matrices can be consistently estimated using their natural empirical analogs as in (5.10).

7 Simulation Study

The following simulations investigate the finite-sample estimation error and coverage properties, under misspecification, of the second stage estimator $\hat{\boldsymbol{\theta}}(x)$ described in Section 2.

7.1 Simulation Design

Random variables are generated i.i.d. according to the following distributions. The error term $\quad$ is independently generated following $\quad N(0; \ {}^2)$ where $\quad = 1$: $Z = (X; \frac{1}{2}X^2; Z_1)\ 2\ \mathsf{R}^{2d_x + d_z}$ where $d_x = 1$, $d_z = 100$, X is independently generated following $U(1; 2)$, and the independent regressors Z_1 are jointly centered Gaussian with a covariance matrix of the Toeplitz form

$$\mathrm{Cov}(Z_{1;j}; Z_{1;k}) = \mathsf{E}[Z_{1;j}Z_{1;k}] = 2\ {}^{jj\ \ kj}; \quad 1 \quad j; k \quad d_z:$$

These distributions and the following data-generating processes (DGP) follow the design of Tan (2020). Let Z^y be a transformation of the regressors in Z_1 where $Z_j^y = Z_j + \max(0; 1 + Z_j)^2; \ 8\ j = 3; \ldots; d_z$. Let `sparsity` control the number of regressors in Z_1, i.e. excluding X terms and a constant term, entering the

DGP by $\quad = \quad = (\underbrace{1;\ldots;1}_{\text{sparsity}};\ \underbrace{0;\ldots;0}_{d_z\ \text{sparsity}})$:

(S1) *Correct speci cation*: Generate D given Z from a Bernoulli distribution with $\Pr(D = 1jZ) = f1 + \exp(p_1 \quad X \quad 0{:}5X^2 \quad {}^{0}Z)g^{\ 1}$ and $Y = 1 + D + X + 0{:}5X^2 + \ {}^{0}Z_1 1 + \ :$

(S2) *Propensity score model correctly speci ed, but outcome regression model misspeci ed*: Generate D given Z as in (S1), but $Y = 1 + D + X + 0{:}5X^2 + \ {}^{0}Z_1^y + \ :$

(S3) *Propensity score model misspeci ed, but outcome regression model correctly speci ed*: Generate Y according to (S1), but generate D given Z from a Bernoulli distribution with $\Pr(D = 1jZ) = f1 + \exp(p_2 \quad X \quad 0{:}5X^2 = \ {}^{0}Z_1^y)g^{\ 1}$.

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 $N \ 2 \ f500; 1; 000g$ 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

In our first-stage estimation, we need to set values for the arbitrary constants $c_{;j}; c_{;j}$ for $j \ 2 \ f1;\ldots;kg$ that appear in the pilot penalty parameters specified in (4.2). Following Tan (2020), we cross-validate to find suitable constants. These pilot penalty parameters are used to estimate $\mathsf{b}_j^{\text{pilot}}$ and $\mathsf{b}_j^{\text{pilot}}$ for $j \ 2 \ f1;\ldots;kg$. Following the multiplier bootstrap procedure outlined in (4.3) and (4.4) with $B = 10{;}000$ bootstrap samples and $c_0 = 1{:}1$, the final penalty parameters $\quad_{;j}$ and $\quad_{;j}$ are obtained for use in the first-stage estimators (2.5) and (2.6). These $2 \quad k$ estimations for b_j and b_j are implemented in parallel using the `optim` function in R's `stats` package–which uses Nelder Mead by default–with `maxit`= 500 and `ndeps`= 0:01 and the remaining settings using default values.

In our second-stage estimation, we follow the series estimation outlined in (2.12) to obtain b^k which is used to estimate the target parameter $g_0(x)$ with $\hat{\mathsf{g}}(x) := p^k(x)^{0}\mathsf{b}^k$. Results are reported for $k = 3$: For an approximating basis, B-splines are used from the `bSpline` function in the R package `splines2`, which uses the specification detailed in Perperoglou et al. (2019).

This Monte Carlo study aims to compare the proposed estimation and inference procedures to a state-of-the-art benchmark method targeting the CATE. Specifically, we use a benchmark method from Semenova and Chernozhukov (2021), which uses DML series estimation. This *DML* method follows Belloni et al. (2015) but uses the aIPW signal, which requires estimating the outcome regression in the first-stage. We follow the publically available code to use sample splitting in both first-stages; estimate the propensity score via logistic regression with an $`_1$-penalty; and, estimate the outcome regression via lasso.[1] Let *MA-DML* refer to our method, which aims to demonstrate, in relation to *DML*, the benefits of model assisted inference under misspecification.

To estimate the variance of $\hat{\mathsf{g}}(x)$ and form confidence intervals, we follow (2.14) to obtain the estimated covariance matrix $\hat{\Omega}$. Confidence bands are formed using the estimated standard error $\mathsf{b}(x) := k\hat{\Omega}^{1=2}p^k(x)k=\sqrt{n}$ according to (2.17), which uses different bootstrapped critical values for pointwise and uniform confidence intervals.

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

Estimation error is studied for each method's $\hat{b}(x)$ for the target parameter $g_0(x) = \mathsf{E}[Y|D = 1; X = x]$ over a grid of 100 points spaced across $x \in [1; 2]$, i.e. the support of X. We study coverage across S simulations of each method's pointwise (i.e. $\mathsf{b}(x)$ at $x = 1.5$) and uniform confidence intervals.

To compare the estimation error for the target parameter $g(x)$ across the two different estimators $\hat{g}_s(x)$ for each simulation $s = 1; \ldots; S$, we utilize integrated bias, variance, and mean-squared error where $\bar{g}(x) = S^{-1} \sum_{s=1}^{S} \hat{g}_s(x)$;

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

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

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

7.3 Simulation Results

We aim to study the estimation error and coverage properties of this paper's proposed method under correctly and misspecified regimes according to S1-S3 as compared to a state-of-the-art benchmark method. Specifically, do the model-assisted inference methods improve the estimator's coverage properties under misspecification? How does estimation error change under misspecification? How does performance respond to different degrees of high-dimensionality?

Table 7.1 presents the simulation results. We present all three specifications, i.e. (S1)-(S3), in two different regimes, i.e. $N = 500$ in the top panel and $N = 1000$ in the lower panel, for both estimators *DML* and *MA-DML*. Estimation errors measures integrated squared bias, variance, and mean squared error are presented in columns (1)-(3). Pointwise and uniform coverage results are presented in columns (4)-(7).

First, 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.

Next, 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.

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.

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 $N = 500$.

Our findings should not be interpreted as a critique of the authors of this benchmark method, whose work we rely on and were inspired by.

8 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. A full summary of the data used as well as additional details/analysis from our empirical analysis can be found in Appendix F.

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

[1]The dataset can be downloaded here.

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).

8.1 Empirical Results

Figure 8.1 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 (4.4) and and select the first stage penalty parameters using the 99^{th}, 95^{th}, and 90^{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).

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 paramater, though the model assisted procedure displays a bit more sensitivity to the level of regularization introduced.[1]

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, compared to the kernel first 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 using first degree b-splines instead of second degree splines. These results are displayed in Figure 8.2. 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 using a standard MLE/OLS first stage is very stable to penalty choice here while the shape of the model assisted CATE function displays a bit more instability here at the two lower levels of regularization.

Finally, Table 8.1 reports the smoothed average treatment effect estimates taken from averaging the model assisted CATE estimates from Figure 8.1 across observations. Again, these estimates are generally in line with prior work

[1] Numerically solving the minimization problems in (2.5)-(2.6) also typically requires more iterations to converge than solving the standard MLE/OLS minimzaitn problems.

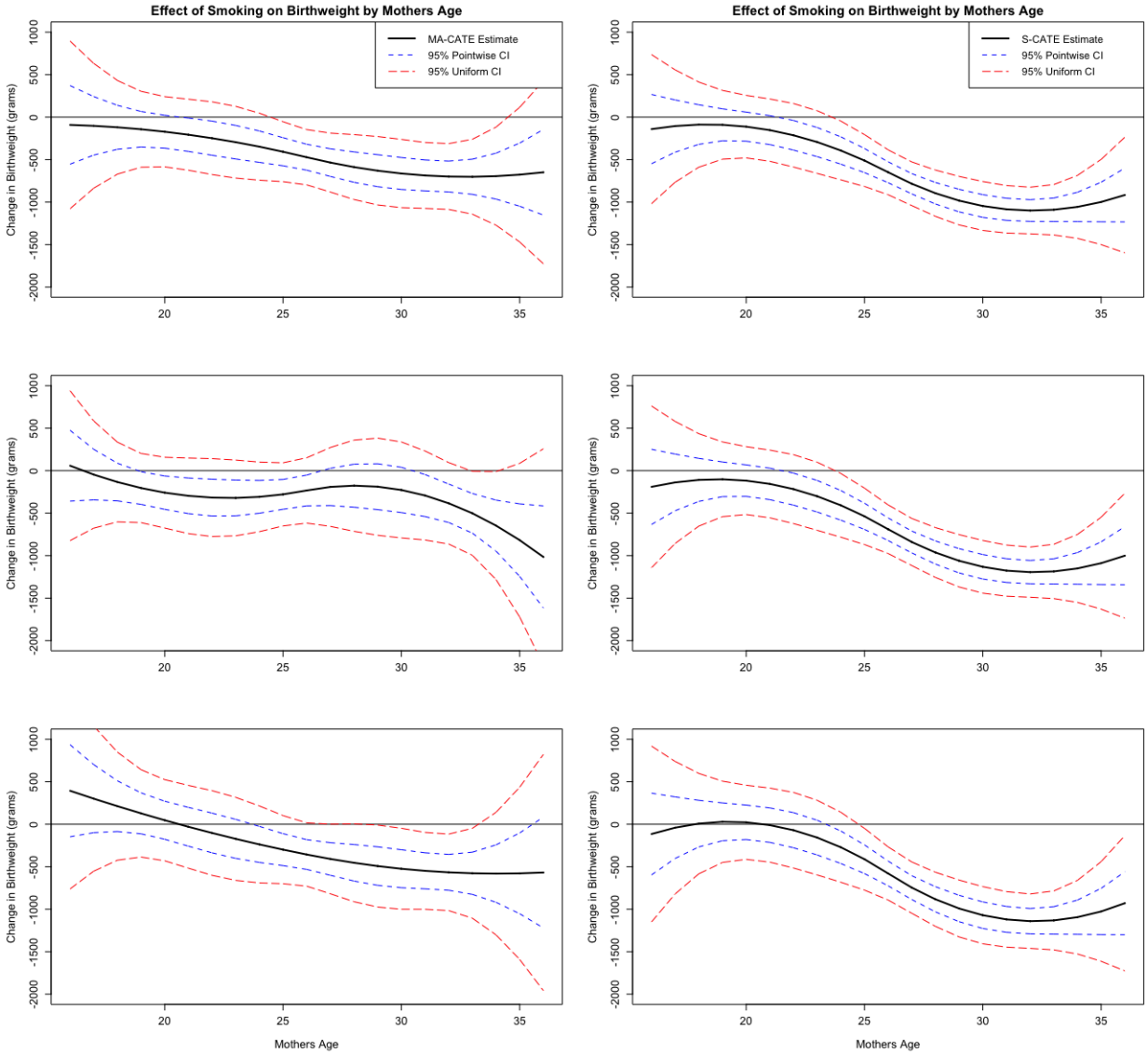

Figure 8.1: CATE of maternal smoking estimated using model assisted estimating equations (left) and standard MLE/OLS estimating equations (right). Top row uses the 99^{th} quantile of the bootstrap distribution to select the penalty parameters, second row uses 95^{th} quantile, and final row uses the 90^{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.

Table 8.1: Smoothed Model Assisted ATE Estimates

Bootstrap Penalty Qt.	99^{th}	95^{th}	90^{th}
Implied ATE	-295.221	-292.9086	-453.2242

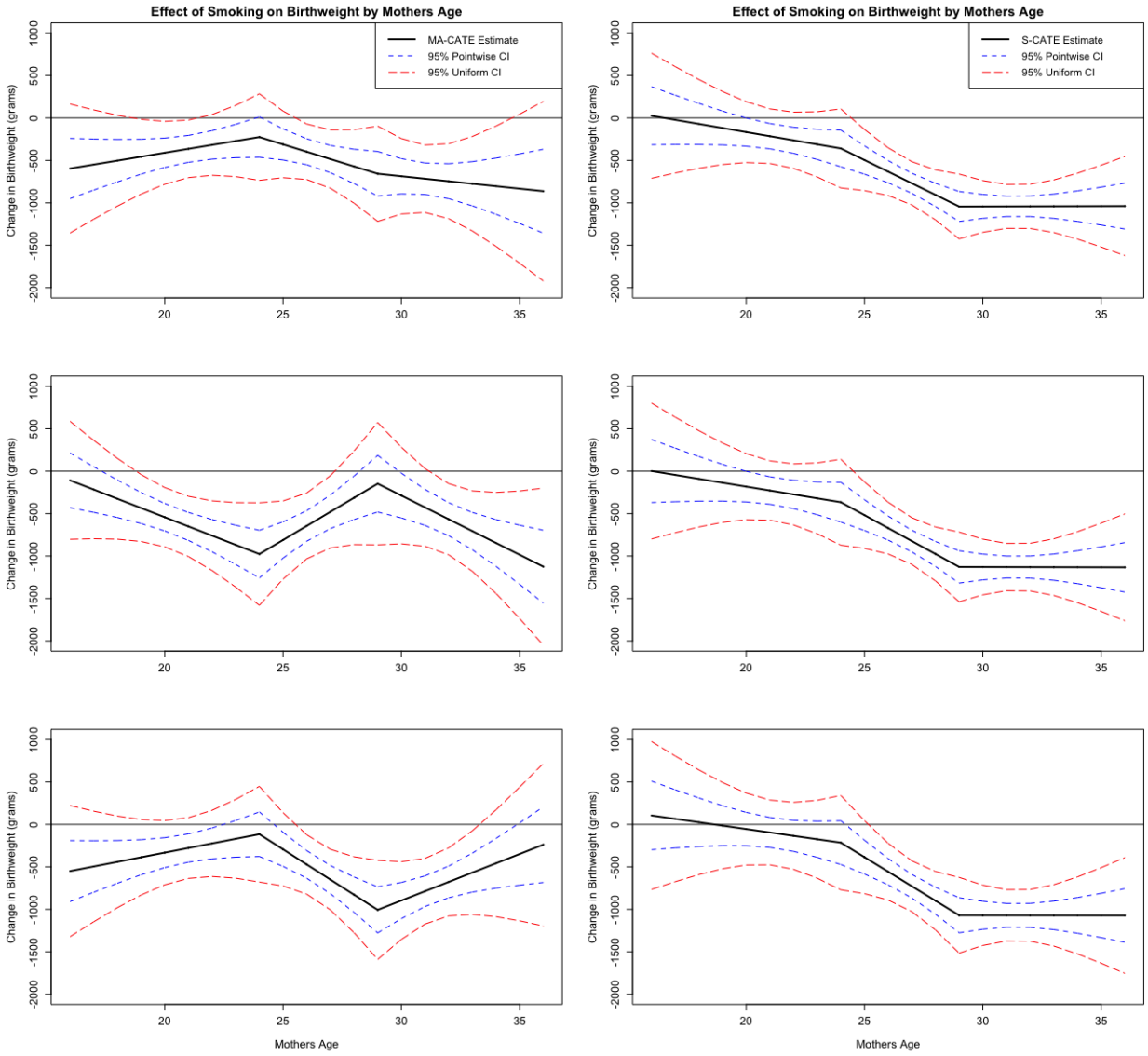

Figure 8.2: CATE of maternal smoking estimated using model assisted estimating equations (left) and standard MLE/OLS estimating equations (right). Top row uses the 99^{th} quantile of the bootstrap distribution to select the penalty parameters, second row uses 95^{th} quantile, and final row uses the 90^{th} quantile. Second stage is computed using b-splines of the first 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.

9 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 resuling 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., C. Klaassen, Y. Ritov, and J. Wellner (1993). *E cient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins series in the mathematical sciences. Johns Hopkins University Press.

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.

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 `$_1$ regularized estimation of causal contrasts. *ArXiv NA*, 1–125.

Stone, C. J. (1985). Additive Regression and Other Nonparametric Models. *The Annals of Statistics 13* (2), 689 – 705.

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 3-5. The proofs for Section 4 rely on an assortment of supporting lemmas proved in Appendix B.

A.1 Proof of Proposition 3.1

Immediately follows from Theorems 4.3, 4.4, 5.1 and 5.4 under the sparsity bounds in (4.10) and (4.13).

A.2 Proof of Proposition 3.2

Immediately follows from Theorems 4.3, 4.4, 5.4 and 5.5 under the sparsity bounds in (4.10) and (4.13).

A.3 Proofs of Results in Section 4

Proof of Theorem 4.1

Lemma B.7 shows that there is a constant M that does not depend on k such that under the conditions of Assumption 4.1 and Theorem 4.1, $\Pr(\Omega_{k;1} \setminus \Omega_{k;2}) \quad 1 \quad o(1)$ where

$$\bar{}_k = M \; {}_{k;1} \; \sqrt{\frac{\log(d_z =)}{n}}:$$

Lemma B.1 provides the bound under this event.

Proof of Theorem 4.2

Lemmas B.7-B.10 show that there is a constant M that does not depend on k such that under the conditions of Assumptions 4.1 and 4.2 and Theorem 4.2, $\Pr(\bigcap_{m=1}^{6} \Omega_{k;m}) \quad 1 \quad o(1)$ where

$$\bar{}_k = M \; {}_{k;1} \; \sqrt{\frac{\log(d_z =)}{n}}:$$

Lemma B.2 provides the bound under this event.

Proof of Theorem 4.3

Lemmas B.7-B.10 show that there is a constant M that does not depend on k such that under the conditions of Assumptions 4.1 and 4.2 and Theorem 4.3, $\Pr(\bigcap_{m=1}^{6} \Omega_{k;m}) \quad 1 \quad o(1)$ where

$$\bar{}_k = M \; {}_{k;1} \; \sqrt{\frac{\log(d_z =)}{n}}:$$

Lemma B.3 provides the bound under this event.

Proof of Theorem 4.4

Lemmas B.7-B.11 show that there is a constant M that does not depend on k such that under the conditions of Assumptions 4.1 and 4.2 and Theorem 4.4, $\Pr(\bigcap_{m=1}^{7} \Omega_{k;m}) \quad 1 \quad o(1)$ where

$$\bar{}_k = M \; {}_{k;1} \; \sqrt{\frac{\log(d_z =)}{n}}:$$

Lemma B.4 provides the bound under this event.

A.4 Proofs of Results in Section 5

The proofs for Section 5 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 Appendix C .

Proof of Theorem 5.1

Equation (3.2) follows from applying (3.1) with $= p(x)=kp(x)k$ and (3.3) follows from (3.2). So it suffices to prove (3.1).

For any $2 S^{k\ 1}$, $1\ .\ k\ {}^{0}\Omega^{1=2}k$ because of the conditional variance of $^{-2}_{j}$ is bounded from below and from above and under the positive semidefinite ranking

$$\Omega \quad \Omega_0 \quad _^{2}Q\ ^{1}:$$

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

$$\begin{aligned} p_{\overline{n}\ {}^{0}}(\mathsf{b} \qquad) &= \frac{p_{\overline{n}\ {}^{0}}}{k\ {}^{0}\Omega^{1=2}k}\mathsf{G}_n[p^k(x) \quad (\ ^k + r_k)] + o_p(1) \\ &= \sum_{i=1} \frac{p}{\overline{n}k\ {}^{0}\Omega^{1=2}k} {}^{0} fp^k(x) \quad (\ ^k + r_k)g: \end{aligned}$$

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 $a = (a_1;:::;a_k)^0\ 2\ S^{k\ 1}$, let $jaj = (ja_1j;:::;ja_kj)^0$ and note that $jaj\ 2\ S^{k\ 1}$ as well:

$$^{\sim 0}_{n} = \frac{p}{\overline{n}k\ {}^{0}_{n}\Omega^{1=2}k}; \quad !_n := j\sim j^{0}p^k(x); \quad \text{and} \quad \bar{}_k := \sup_{1\ j\ k} j\ _j j$$

Now, by the definition of Ω we have that

$$\mathrm{Var}\left(\sum_{i=1} \frac{p}{\overline{n}k\ {}^{0}\Omega^{1=2}k} {}^{0} fp^k(x) \quad (\ ^k + r_k)g\right) = 1:$$

Second for each > 0

$$\begin{aligned} &\sum_{i=1} \mathsf{E}\ (\sim^{0}_{n} fp^k(x) \quad (\ ^k + r_k)g)^2 \mathbf{1}\left\{ j\sim^{0}_{n} fp^k(x) \quad (\ ^k + r_k)gj > \right\} \\ &\sum_{i=1} \mathsf{E}\ !^2_n \mathsf{E}\left[{}^{-2}_{k}\mathbf{1}\{j!_n jj\bar{}_k + `_k c_k j > \} \ g\ j\ X = x\right] \end{aligned} \tag{A.1}$$

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

$$f\ {}^{0}(a\quad b)g = \sum_{j=1}^{k} {}_{j}a_j b_j \quad kbk_1 \sum_{j=1}^{k} j\ {}_{j}ja_j = (j\ j^{0}a)kbk_1 : \tag{A.2}$$

To bound the right hand side of (A.1) use the fact that $1\ .\ k\ {}^{0}\Omega^{1=2}k$ because $1\ .\ \_^2$ and

$$\Omega \quad \Omega_0 \quad \_^2 Q\ {}^{1}$$

in the positive semidefinite sense. Using these two we have

$$n\mathsf{E}j\,!_n j^2 \quad \mathsf{E}[(j\ j^{0}p^k(x))^2]=(\ {}^{0}\Omega\ )\ .\ 1:$$

By the bounded eigenvalue condition and using the trace operator:

$$\mathsf{E}[(j\ jp^k(x))^2] = \mathrm{trace}(\mathsf{E}[j\ j^{0}p^k(x)^{0}p^k(x)j\ j]) = j\ j^{0}Qjj\ j\ .\ k\ k = 1$$

Further note, $j\,!_n j\ .\ \frac{pk}{n}$. Using $(a+b)^2 \quad 2a^2+2b^2$, the right hand side of (A.1) is bounded by

$$2n\mathsf{E}[j\,!_n j^{2-2}_k \mathbf{1}fj^{-}_k j + \grave{}_k c_k > \ =j\,!_n jg] + 2n\mathsf{E}[j\,!_n j^2 \grave{}^2_k c^2_k \mathbf{1}fj^{-}_k j + \grave{}_k c_k > \ =j\,!_n jg]$$

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

$$\begin{aligned} 2n\mathsf{E}[j\,!_n j^{2-2}_k \mathbf{1}fj^{-}_k j + \grave{}_k c_k > \ =j\,!_n jg]\ .\ n\mathsf{E}[j\,!_n j^2] \sup_{x2X} \mathsf{E}[{}^{-2}_k \mathbf{1}f^{-2}_k + \grave{}_k c_k > c\ {}^{p}\overline{n}= {}_k g\ j\ X = x] \\ = o(1) \end{aligned}$$

where here we use the first part of Assumption 5.2. To show the second term converges to zero, follow the same steps as for the first term, but apply the second part of Assumption 5.2.

Proof of Theorem 5.2

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 5.2

Theorem (Gine and Koltchinskii, 2006). *Let $X_1, \ldots, X_n$ 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 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 $\sup_{f \in F} \mathrm{Var}(f) \le \sigma^2 \le \|F\|^2_{L^2(P)}$. Suppose there exist constats $A > e^2$ and $V \ge 2$ such that $\sup_Q N(F, L^2(Q), \varepsilon \|F\|_{L^2(Q)}) \le (A/\varepsilon)^V$ for all $0 < \varepsilon \le 1$. Then*

$$\mathbb{E}\left[\sup_{f \in F} \left| \sum_{i=1}^n f(X_i) - \mathbb{E}[f(X_1)] \right|\right] \le C \left[\sqrt{n \sigma^2 V \log \frac{A\|F\|_{L^2(P)}}{\sigma}} + V \|F\|_{L^2(P)} \log \frac{A\|F\|_{L^2(P)}}{\sigma} \right]. \tag{GK}$$

where C is a universal constant.

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

$$\|\mathbb{G}_n\|_{F, P;p} \lesssim \|J(1, F)\|_{P;p} \cdot \|F\|_{P;p} \cdot \|J(\delta_n, F)\|_{P;2p} \tag{VW}$$

where $\delta_n = \|f\|_{n,F}/\|F\|_{n,F}$, where $\|\cdot\|_{n,F}$ 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

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

We would like to apply these theorems to bound $\sup_{x \in X} |(v)^0 \mathbb{G}_n[p^k(x) - \mu^k]|$ and thus show (5.4). The other two statements of Theorem 5.2 follow from this. To this end, let's consider the class of functions

$$G := \{(x^k, x) \mapsto (v)^0(p^k(x) - \mu^k); v \in X\}.$$

Let's note that $|(v)^0 p^k(x)| \le \sigma_k$, $\mathrm{Var}((v)^0 p^k(x)) = 1$, and for any $v, \tilde{v} \in X$

$$|(v)^0(p^k(x) - \mu^k) - (\tilde{v})^0(p^k(x) - \mu^k)| \le |v - \tilde{v}|_k \sigma_k \le \|v - \tilde{v}\|,$$

where $\sigma_k = \max_k \sigma_{1k}$. Then, taking $G(x^k, x) \le \sigma_k \bar{\sigma}_k$ we have that

$$\sup_Q N(G, L^2(Q), \varepsilon \|G\|_{L^2(Q)}) \le \left(\frac{C \sigma_k}{\varepsilon} \right)^d. \tag{A.3}$$

Now, for a $\tau > 0$ specified later define $\sigma_k = \max_k \mathbf{1}\{|v - \mu|_k \le \tau\}$ and $\sigma_k^+ = \max_k \mathbf{1}\{|v - \mu|_k > \tau\}$. Since $\mathbb{E}[\sigma_k^+ | X] = 0$ we have that $\sigma_k = \sigma_k + \sigma_k^+$. Using this decompose:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n (v)^0(p^k(x) - \mu^k) = \sum_{i=1}^n (v)^0(p^k(x) - \mu_k)\sigma_k / \sqrt{n} + \sum_{i=1}^n (v)^0(p^k(x) - \mu^k)\sigma_k^+ / \sqrt{n}.$$

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 (A.3) gives us the constants

$A = C\ \frac{L}{k}$ and $V = d_x \_ 2$. To get $\ ^2$ note that for any $v\ 2\ X$,

$$\begin{aligned}\mathrm{Var}(\ \ (v)^0(p^k(x)\quad\ _k)=^p\overline{n}) \quad & \mathsf{E}[(\ \ (v)^0(p^k(x)\quad\ _k)=^p\overline{n})^2] \\ & \frac{1}{n}\mathsf{E}[(\ \ (v)^0p^k(x))^2]\ \sup_{x\,2\,X}\mathsf{E}[k\ _k k_1^2\ \ j\ X = x] \\ & \frac{^{-2}_k\ \wedge\ \ ^2}{n}\end{aligned}$$

Finally note that we can take the envelope $G = k\ _k k_1\ \ _k=^p\overline{n}$ where $kGk_{L^2(P)}\quad \frac{\wedge}{\overline{n}}$ and $kGk_1\quad\ _k=^p\overline{n}$.

We can now apply (GK) to get that

$$\mathsf{E}[\sup_{x\,2\,X} j\ \ (x)^0\mathsf{G}_n[p^k(x)\quad\ _k]j]\ .\ \ ^q\overline{^{-2}_k\ \wedge\ \ ^2\log(\ \frac{L}{k})} + \frac{k\log(\ \frac{L}{k})}{p\overline{n}}:$$

Step 2: For the second term, we set up for an application of (VW) with the envelope function $G = k\ ^+_k k_1\ \ _k=^p\overline{n}$ and note that

$$\mathsf{E}[k\ ^+_k k_1^2\,] \quad \mathsf{E}[^{-2}_k\mathbf{1}fj^-_kj > \ \ g] \qquad ^{m+2}\mathsf{E}[j^-_kj^m]$$

We can now use (VW) to bound

$$\begin{aligned}\mathsf{E}\ \sup_{x\,2\,X} j\ \ (x)^0\mathsf{G}_n[p^k(x)\quad\ ^+_k]j\ \ . & \ ^p\overline{\mathsf{E}[j^-_kj^m]}\quad ^{m=2+1}_k\int_0^1\ ^q\overline{\log(\ \tfrac{L}{k}=")}\,d" \\ . & \ ^p\overline{\tfrac{m}{k}}\quad ^{m=2+1}_k\ ^q\overline{\log(\ \tfrac{L}{k})}:\end{aligned}$$

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

$$\begin{aligned}\sup_{x\,2\,X} j\ \ (x)^0\mathsf{G}_n[p^k(x)\quad\ _k]j\ .\ _P & \ ^q\overline{^{-2}_k\log(\ \tfrac{L}{k})} + \frac{^{2m=(m\ \ 2)}_k\log(\ \frac{L}{k})}{p\overline{n}} \\ \sup_{x\,2\,X} j\ \ (x)^0\mathsf{G}_n[p^k(x)\quad\ ^+_k]j\ .\ _P & \ ^q\overline{^{-m}_k\log(\ \tfrac{L}{k})}\end{aligned}$$

Applying Assumption 5.3 along with the inequality

$$\frac{^{m=(m\ \ 2)}_k\log k}{p\overline{n}} = \ ^p\overline{\log k}\ \ \frac{^{2m=(m\ \ 2)}_k\log k}{n}\ .\ \log k$$

completes the proof.

Proof of Theorem 5.3

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

Yurinskii's Coupling Lemma

Let $_1;:::;\ _n$ be independent random k-vectors with $\mathsf{E}[\ _i] = 0$ and $:= \sum_{i=1}^{n} \mathsf{E}[k\ _i k^3]$ finite. Let $S := \ _1 + \quad + \ _n$. For each > 0 there exists a random vector T with a $N(0; \mathrm{var}(S))$ distribution such that

$$\mathsf{P}(jS \quad Tj > 3\ ) \quad C_0 B \quad 1 + \frac{j \log(1=B) j}{k} \qquad \text{where } B := \quad k \quad {}^3 \tag{YC}$$

for some universal constant C_0.

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

$$\frac{1}{p\overline{n}} \sum_{i=1} \ _i; \qquad _i = \Omega \ {}^{1=2} p^k(x) \quad (\ {}^k + r_k):$$

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

$$\begin{aligned} \mathsf{E}k\ _i k^3 &. \mathsf{E}[kp^k(x) \quad (\ {}^k(x) + r_k)k^3] \\ &. \mathsf{E}[kp^k(x)k^3(j^- {}_k j^3 + j r_k j^3)] \\ &. \mathsf{E}[kp^k(x)k^3](^-{}_k^3 + \ `{}_k^3 c_k^3) \\ &. \mathsf{E}[kp^k(x)k^3]\ {}_k(^-{}_k^3 + \ `{}_k^3 c_k^3) \\ &. k\ {}_k(^-{}_k^3 + \ `{}_k^3 c_k^3) \end{aligned}$$

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

$$\begin{aligned} \Pr\left(\left\{ k \sum_{i=1} \ _i =^p \overline{n} \quad N_k k > 3\ a_n{}^1 \right\}; \right) &. \frac{nk^2\ {}_k(^-{}_m^3 + \ `{}_k^3 c_k^3)}{(\ a_n{}^1 p\overline{n})^3} \quad 1 + \frac{\log(k^3\ {}_k(^-{}_k^3 + \ `{}_k^3 c_k^3))}{k} \\ &. \frac{a_n^3 k^2\ {}_k(^-{}_k^3 + \ `{}_k^3 c_k^3)}{{}^3 n^{1=2}} \quad 1 + \frac{\log n}{k} \quad ! \ 0: \end{aligned}$$

because $a_n^6 k^2\ {}_k(^-{}_m^3 + \ `{}_k^3 c_k^3) \log^2 n = n \ ! \ 0$. Using the first two results from Lemma C.3, (C.6)-(C.7), we obtain that

$$k^p \overline{n} \ (x)^0(\mathsf{b}k \quad {}^k) \quad (x)^0 \Omega^{1=2} N_k k \quad k1 =^p \overline{n} \sum_{i=1} \ (x)^0 \Omega^{1=2}\ _i \quad (x)^0 \Omega^{1=2} N_k k + \bar{R}_{1n} = o_p(a_n{}^1):$$

uniformly over $x \ 2 \ X$. Since $k \ (x)^0 \Omega^{1=2} k$ is bounded from below uniformly over $x \ 2 \ X$ we obtain the first statetment of Theorem 5.2 from which the second statement directly follows.

Finally, under the assumption that $\sup_{x2X} n^{1=2} jr(x) j = ks(x) k = o_p(a_n{}^1)$,

$$\frac{\sqrt{np(x)^0(\mathsf{b}k \quad {}^k)}}{ks(x)k} \quad \frac{p\overline{n}(\mathsf{b}g(x) \quad g_0(x))}{ks(x)k} = o_p(a_n{}^1)$$

so that the third statement, (5.9) holds.

Proof of Theorem 5.4

Preliminaries for Proof of Theorem 5.4

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*

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

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

For $p \ge 1$ 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 $p=1$ 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, $i=1,\ldots,n$, $2 \le p < \infty$, and an i.i.d sequence of Rademacher random variables $\epsilon_1,\ldots,\epsilon_n$ we have*

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

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

$$\mathsf{E}\left[\|\mathbb{G}_n[\epsilon_i Q_i]\|\right] \le C\sqrt{\log k}\,\|(\mathsf{E}_n[Q_i^2])^{1/2}\| \tag{KI-2}$$

for some (possibly different) absolute constant C.

We will establish consistent estimation of

$$\Sigma = \mathsf{E}[f p^k(x)\,(\gamma^k + r_k)(\gamma^k + r_k)' f p^k(x)'\,(\gamma^k + r_k)(\gamma^k + r_k)']$$

using

$$\hat{\Sigma} = \mathsf{E}_n[\hat{f} p^k(x)\,\hat{b}^k \hat{f} p^k(x)'\,\hat{b}^k \hat{f} p^k(x)']$$

Consistency of $\hat{\Omega}$ will then follow from the consistency of $\hat{Q}$ established by Lemma C.1. To save notation, define the vectors

$$\hat{p} := \begin{bmatrix} \hat{Y}(\hat{b}_1; \hat{m}_1) \\ \vdots \\ \hat{Y}(\hat{b}_k; \hat{m}_k) \end{bmatrix} \quad \text{and} \quad p := \begin{bmatrix} Y(b_1; m_1) \\ \vdots \\ Y(b_k; m_k) \end{bmatrix} \tag{A.4}$$

Also define $\gamma^k := (\gamma_1^k,\ldots,\gamma_k^k)$ so that $\gamma_j^k := Y(\bar{\beta}_j; \bar{m}_j)'\,\hat{b}(x)$. Ideally, we would like to use γ^k to estimate $\hat{\Sigma}$, but we don't observe γ^k. Define $\Delta := \hat{b}^k\,\gamma^k - \hat{p}^k\,\bar{Y}^k \in \mathbb{R}^k$.

Using this, we can decompose

$$\begin{aligned}\Sigma &= \mathsf{E}_n[fp^k(x) \quad (\Delta + {}^{\cdot k})gfp^k(x) \quad (\Delta + {}^{\cdot k})g] \\ &= \underbrace{\mathsf{E}_n[fp^k(x) \quad \Delta gfp^k(x) \quad \Delta g^0]}_{1} + \underbrace{\mathsf{E}_n[fp^k(x) \quad {}^{\cdot k}gfp^k(x) \quad \Delta g^0]}_{2} \\ &\quad + \underbrace{\mathsf{E}_n[fp^k(x) \quad \Delta gfp^k(x) \quad {}^{\cdot k}g^0]}_{3} + \underbrace{\mathsf{E}_n[fp^k(x) \quad {}^{\cdot k}gfp^k(x) \quad {}^{\cdot k}g]}_{4}\end{aligned} \tag{A.5}$$

We first show that $k\Sigma_4 \quad \Sigma k \; ! \; _p \; 0$. This is nonstandard because of the Hadamard product.

Lemma A.1 (Psuedo-Variance Estimator Consistency)**.** *Suppose Assumption 5.1 and Assumption 5.3 hold. Further, de ne* $v_n = \mathsf{E}[\max_{1 \; i \; n} j^{-}{}_k j^2]^{1=2}$*. In addition, assume that* $\bar{R}_{1n} + \bar{R}_{2n} \; . \; (\log k)^{1=2}$*. Then,*

$$k\hat{\Sigma} \quad Qk \; ._P \; \sqrt{\frac{{}^2_k \log k}{n}} = o(1)$$

$$\text{and} \quad k\Sigma_4 \quad \Sigma k \; ._P \; (v_n \_ 1 + \grave{}_k c_k) \sqrt{\frac{{}^2_k \log k}{n}}$$

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.5) we can define $\dot{\Delta} \quad {}^{\cdot k} \quad {}^{k} = g_0(x) \quad \hat{g}(x)$[1] and decompose

$$\begin{aligned}\Sigma_4 &= \underbrace{\mathsf{E}_n[p^k(x)p^k(x)^0\dot{\Delta}^2]}_{41} + \underbrace{\mathsf{E}_n[fp^k(x) \quad ({}^{k} + r_k)gfp^k(x) \quad \dot{\Delta}g^0]}_{42} \\ &\quad + \underbrace{\mathsf{E}_n[fp^k(x) \quad \dot{\Delta}gfp^k(x) \quad ({}^{k} + r_k)g^0]}_{43} + \underbrace{\mathsf{E}_n[fp^k(x) \quad ({}^{k} + r_k)gfp^k(x) \quad ({}^{k} + r_k)g]}_{44}\end{aligned}$$

The terms $\Sigma_{41}; \Sigma_{42}$ and Σ_{43} are simple to show are negligible.

$$\begin{aligned}&k\Sigma_{41} + \Sigma_{42} + \Sigma_{43}k \\ &\quad k\mathsf{E}_n[fp^k(x)^0(\hat{} \quad {}^{k})gp^k(x)p^k(x)^0]k + k\mathsf{E}_n[fp^k(x) \quad ({}^{k} + r_k)gp^k(x)^0fp^k(x)^0(\hat{} \quad {}^{k})g]k \\ &\quad + k\mathsf{E}_n[p^k(x)fp^k(x)^0(\hat{} \quad {}^{k})gfp^k(x) \quad ({}^{k} + r_k)g^0]k \\ &\quad \max_{1 \; i \; n} jp^k(x)(\hat{} \quad {}^{k})j^2 k\mathsf{E}_n[p^k(x)p^k(x)^0]k \\ &\quad + 2 \max_{1 \; i \; n} j^{-}{}_{k;i}j + jr_{k;i}j \max_{1 \; i \; n} jp^k(x)^0(\hat{} \quad)jk\mathsf{E}_n[p^k(x)p^k(x)^0]k\end{aligned}$$

By Theorem 5.2 $j \max_{1 \; i \; n} jp^k(x)^0(\hat{} \quad {}^{k})j \; ._P \; {}^2_k(\sqrt{\log k} + \bar{R}_{1n} + \bar{R}_{2n})^2 = n$, by Assumption 5.1 the approximation error is bounded $\max_{1 \; i \; n} jr_{k;i}j \quad \grave{}_k c_k$, by Assumption 5.3 and Markov's inequality the errors are bounded $\max_{1 \; i \; n} j^{-}{}_{k;i}j \; ._p \; v_n^2$. Finally, by the first part of Lemma A.1 $k\hat{Q}k \; ._P \; kQk \; . \; 1$. Putting this all together with $\bar{R}_{1n} + \bar{R}_{2n} \; . \; (\log k)^{1=2}$ and ${}^2_k \log k = n \; ! \; 0$ gives

$$k\Sigma_{41} + \Sigma_{42} + \Sigma_{43}k \; ._P \; (v_n \_ 1 + \grave{}_k c_k) \sqrt{\frac{{}^2_k \log k}{n}}:$$

Next, we want to control $\Sigma_{44} \quad \Sigma$. To do this, let $_1; : : : ; \; _n$ be independent Rademacher random variables

[1] It is useful to recall that ${}^{\cdot k} = \bar{Y}^k \quad \hat{g}(x)$ and ${}^{k} = \bar{Y}^k \quad g_0(x)$

generated independently from the data. Then for $= (\ _1; \dots;\ _n)$

$$\begin{aligned}
&\mathsf{E}[k\mathsf{E}_n[fp^k(x) \quad (\ ^k + r_k)gfp^k(x) \quad (\ ^k + r_k)g^0] \quad \Sigma k] \\
&\quad . \mathsf{E}[\mathsf{E}\ [\mathsf{E}_n k\ \ fp^k(x) \quad (\ ^k + r_k)gfp^k(x) \quad (\ ^k + r_k)g^0 k]] \\
&\quad . \sqrt{\frac{\log k}{n}}\mathsf{E}[(k\mathsf{E}_n[kp^k(x)k^2(\bar{}_k + r_k)^2 fp^k(x) \quad (\ ^k + r_k)gfp^k(x) \quad (\ ^k + r_k)g^0]k)^{1=2}] \\
&\quad . \sqrt{\frac{^2_k \log k}{n}}\mathsf{E}[\max_{1\ i\ n} j\bar{}_{k;i} + r_k j(k\mathsf{E}_n[fp^k(x) \quad (\ ^k + r_k)gfp^k(x) \quad (\ ^k + r_k)g^0]k)^{1=2}] \\
&\quad \sqrt{\frac{^2_k \log k}{n}}(\mathsf{E}[\max_{1\ i\ n} j\bar{}_{k;i} + r_k j^2])^{1=2} \quad (\mathsf{E}[k\mathsf{E}_n[fp^k(x) \quad (\ ^k + r_k)gfp^k(x) \quad (\ ^k + r_k)g^0]k])^{1=2}
\end{aligned}$$

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

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

$$\begin{aligned}
&\mathsf{E}[k\mathsf{E}_n[fp^k(x) \quad (\ ^k + r_k)gfp^k(x) \quad (\ ^k + r_k)g^0] \quad \Sigma k] \\
&\qquad . \frac{^2_k \log k}{n}(v_n^2 + \grave{}^2_k c_k^2) + \quad \frac{^2_k \log k}{n} f v_n^2 + \grave{}^2_k c_k^2 g^{\ !\ 1=2} \quad k\Sigma k^{1=2}
\end{aligned}$$

and so, because $k\Sigma k$. 1 and $(v_n^2 + \grave{}^2_k c_k^2)\ ^2_k \log k=n\ !\ 0$ we have

$$\mathsf{E}[k\mathsf{E}_n[fp^k(x) \quad (\ ^k + r_k)gfp^k(x) \quad (\ ^k + r_k)g^0] \quad \Sigma k] \ . \ (v_n \_ 1 + \grave{}_k c_k) \sqrt{\frac{^2_k \log k}{n}}:$$

The second result of Lemma A.1 follows from Markov's inequality. □

Now, we need to take care of the terms

$$\begin{aligned}
\Sigma_1 &= \mathsf{E}_n[fp^k(x) \quad \Delta gfp^k(x) \quad \Delta g^0] \\
\Sigma_2 &= \mathsf{E}_n[fp^k(x) \quad \cdot^k gfp^k(x) \quad \Delta g^0] \\
\Sigma_3 &= \mathsf{E}_n[fp^k(x) \quad \Delta gfp^k(x) \quad \cdot^k g^0]
\end{aligned}$$

where $\Delta = \hat{p}^k \quad \bar{Y}^k$ and $\cdot^k = \bar{Y}^k \quad \hat{g}(x) = \hat{g}(x) \quad g^k(x) + \ ^k$. To do so we will use Condition 2.

Lemma A.2 (Negligible Variance Bias). *Suppose that Condition 2, Assumption 5.1 and Assumption 5.3 hold. Then*

$$k\Sigma_1 + \Sigma_2 + \Sigma_3 k = o_p(1):$$

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

that

$$k\Sigma_1 k \quad \sup_{\substack{1 \le l \le k \\ 1 \le j \le k}} |\mathsf{E}_n[p_l(X)(Y(\hat{\ }_l; \hat{m}_l) \quad Y(\bar{}_j; \bar{m}_j))p_l(X)(Y(\mathsf{b}_l; \hat{m}_l) \quad Y(\bar{}_l; \bar{m}_l))]|$$
$$\sup_{1 \le l \le k} (\mathsf{E}_n[p_j(X)^2(Y(\mathsf{b}_j; \hat{m}_j) \quad Y(\bar{}_j; \bar{m}_j))^2])^{1/2} \sup_{1 \le j \le k} (\mathsf{E}_n[p_j(X)^2(Y(\mathsf{b}_j; \hat{m}_j) \quad Y(\bar{}_j; \bar{m}_j))^2])^{1/2}$$
$$= o_p(1):$$

To see that Σ_2 is negligible notice that

$$k\Sigma_2 k \quad \sup_{\substack{1 \le l \le k \\ 1 \le j \le k}} \mathsf{E}_n[p_l(X)(\ _l + p^k(x)^0(\mathsf{b}^k \quad ^k))p_j(X)(Y(\mathsf{b}_j; \hat{m}_j) \quad Y(\bar{}_j; \bar{m}_j))]$$
$$\sup_{1 \le l \le k} \mathsf{E}_n[p_l(X)^2(\ _l + p^k(x)^0(\mathsf{b} \quad))^2]^{1/2} \mathsf{E}_n[p_j(X)^2(Y(\mathsf{b}_j; \hat{m}_j) \quad Y(\bar{}_j; \bar{m}_j))^2]^{1/2}$$
$$_{k;1}(\max_{1 \le i \le n} |\bar{}_k| + \max_{1 \le i \le n} p^k(x)^0(\mathsf{b} \quad))\mathsf{E}_n[p_j(X)^2(Y(\mathsf{b}_j; \hat{m}_j) \quad Y(\bar{}_j; \bar{m}_j))^2]^{1/2}$$

Applying Assumption 5.3 and Theorem 5.2 gives

$$\cdot_P k \ _{k;1} n^{1/m} \mathsf{E}[p_j(X)^2 Y(\mathsf{b}_j; \hat{m}_j) \quad Y(\bar{}_j; \bar{m}_j))^2]^{1/2} = o_p(1)$$

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

Proof of Theorem 5.5

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 5.3 and consistent variance estimation as in Theorem 5.4.

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 Assumptions 4.1 and 4.2 these events happen with probability approaching one. Define the scores

$$\begin{aligned} S \ _{;j} &:= \mathsf{E}_n[U \ _{;j} Z] \\ S \ _{;j} &:= \mathsf{E}_n[U \ _{;j} Z] \end{aligned} \tag{B.1}$$

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

$$\begin{aligned} \Omega_{k;1} &:= f \quad _{;j} \quad c_0 \quad kS \ _{;j} k \ _1 ; 8j \quad kg \\ \Omega_{k;2} &:= f \quad _{;j} \quad ^{-}_k ; 8j \quad kg \end{aligned} \tag{B.2}$$

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

Lemma B.1 (Nonasymptotic Bounds for Logistic Model). *Suppose that Assumption 4.1 holds with* $_0 > (c_0+1)/(c_0 \quad 1)$ *and* $2C_0 \ _0{}^2 s_k \bar{} \ _k \quad < 1$. *Then, under the events* $\Omega_{k;1} \setminus \Omega_{k;2}$ *defined in* (B.2), *there exists a finite constant* M_0 *that does not depend on* k *such that*

$$\max_{1 \ j \ k} D^z(\bar{g}; \mathbf{b}) \quad M_0 s_k \bar{}^2_k \quad \text{and} \quad \max_{1 \ j \ k} k\mathbf{b}_j \quad \bar{}_j k_1 \quad M_0 s_k \bar{}_k \tag{B.3}$$

Proof. We show that the bound of (B.3) holds for each $j = 1; : : : ; k$. For any $\quad 2 \ \mathsf{R}^d$ define $\tilde{}_j(\ ) := \mathsf{E}_n[p_j(X) f D e \quad {}^0 Z + (1 \quad D) \ {}^0 Z g]$. By optimality of $\mathbf{b}_j$ we must have, for any $u \ 2 \ (0; 1]$:

$$\tilde{}_j \ \mathbf{b}_j \ + \quad {}_{;j} k\mathbf{b}_j k_1 \quad \tilde{} \ (1 \quad u)\mathbf{b}_j + u \bar{}_j \ + \quad {}_{;j} k(1 \quad u)\mathbf{b}_j + u \bar{}_j k_1 :$$

Using convexity of the $\grave{}_1$ norm $k \ k_1$, this gives after rearrangment

$$\tilde{}_j \ \mathbf{b}_j \quad \tilde{} \ (1 \quad u)\mathbf{b}_j + u \bar{}_j \ + \quad {}_{;j} uk\mathbf{b}_j k_1 \quad {}_{;j} uk \bar{}_j k_1 :$$

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

$$\mathsf{E}_n[p_j(X) D f e \ \ {}^{\mathrm{b}^0 Z} + (1 \quad D) g f \mathrm{b}_j^0 Z \quad {}^{-0}_j Z g] + \quad {}_{;j} k\mathbf{b}_j k_1 \quad {}_{;j} k \bar{}_j k_1 :$$

By direct calculation, we have that $D^z{}_{;j}(\mathbf{b}_j; \bar{}_j)$ from (4.6) can be expressed

$$D^z{}_{;j}(\mathbf{b}_j; \bar{}_j) = \mathsf{E}_n[p_j(X) D f e \ \ {}^{\mathrm{b}^0 Z} + (1 \quad D) g f \mathrm{b}_j^0 Z \quad {}^{-0}_j Z g] \quad \mathsf{E}_n[p_j(X) D f e \ \ {}^{0 Z} + (1 \quad D) g f \mathrm{b}_j^0 Z \quad {}^{-0}_j Z g] :$$

Combining the last two displays yields

$$D^z{}_{;j}(\mathbf{b}_j; \bar{}_j) + \mathsf{E}_n[p_j(X) D f e \ \ {}^{0 Z} + (1 \quad D) g f \mathrm{b}_j^0 Z \quad {}^{-0}_j Z g] + \quad {}_{;j} k\mathbf{b}_j k_1 \quad {}_{;j} k \bar{}_j k_1 \tag{L.1}$$

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

$$j \mathsf{E}_n[p_j(X) D f e \ \ {}^{0 Z} + (1 \quad D) f \mathrm{b}^0 Z \quad {}^{-0} Z g g] \quad c_0 \ {}^1 \quad {}_{;j} k\mathbf{b}_j \quad \bar{}_j k_1 \tag{L.2}$$

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

$$D^z{}_{;j}(\mathbf{b}_j; \bar{}_j) + \quad {}_{;j} k\mathbf{b}_j k_1 \quad {}_{;j} k \bar{}_j k + c_0 \ {}^1 \quad {}_{;j} k\mathbf{b}_j \quad \bar{}_j k_1 :$$

Expanding $k \ _j k_1 = \sum_{l 2 S \ ;j} j \ _l j + \sum_{l 2 S \ ;j} j \ _l j$ for $\quad = \mathbf{b}_j; \bar{}_j$ and applying the triangle inequalities $j \mathbf{b}_{j;l} j$ $j \bar{}_{j;l} j \quad j \mathbf{b}_{j;l} \quad \bar{}_{j;l} j$ for $l \ 2 \ S \ _{;j}$ and the equality $\mathbf{b}_{j;l} = \mathbf{b}_{j;l} \quad \bar{}_{j;l}$ gives

$$\begin{aligned} D^z{}_{;j}(\mathbf{b}_j; \bar{}_j) + \quad {}_{;j} & \sum_{l 2 S \ ;j} j \bar{}_{j;l} j \quad \sum_{l 2 S \ ;j} j \mathbf{b}_{j;l} \quad \bar{}_{j;l} j + \sum_{j \in S \ ;j} j \mathbf{b}_{j;l} \quad \bar{}_{j;l} j \\ & {}_{;j} \sum_{l 2 S \ ;j} j \bar{}_{j;l} j + c_0 \ {}^1 \sum_{l 2 S \ ;j} j \mathbf{b}_{j;l} \quad \bar{}_{j;l} j + c_0 \ {}^1 \sum_{j \in S \ ;j} j \mathbf{b}_{j;l} \quad \bar{}_{j;l} j \end{aligned}$$

Rearrange to get

$$D^z_{;j}(\mathsf{b}_j;\bar{}_j)+(1\quad c_0{}^1)\quad {}_{;j}\sum_{l\in S}\quad j\mathsf{b}_{j;l}\quad \bar{}_{j;l}j\quad (1+c_0)\quad {}^1\quad {}_{;j}\sum_{l2S\ ;j}\quad j\mathsf{b}_{j;l}\quad \bar{}_{j;l}j:$$

Adding $(1\quad c_0{}^1)\quad {}_{;j}\sum_{l2S\ ;j} j\mathsf{b}_{j;l}\quad \bar{}_{j;l}j$ gives

$$D^z_{;j}(\mathsf{b}_j;\bar{}_j)+(1\quad c_0{}^1)k\mathsf{b}_j\quad \bar{}_j k_1\quad 2\quad {}_{;j}\sum_{l2S\ ;j}\quad j\mathsf{b}_{j;l}\quad \bar{}_{j;l}j \tag{L.3}$$

By Lemma 4 in Appendix V.3 of Tan (2017) we have that for $\quad {}_j := \mathsf{b}_j\quad \bar{}_j$

$$D^z_{;j}(\mathsf{b}_j;\bar{}_j)\quad \frac{1\quad e^{\quad C_0 k\ {}_j k_1}}{C_0 k\mathsf{b}_j k}\quad {}^0_j\tilde{\Sigma}_{\ ;j\ \ j} \tag{L.4}$$

By (L.3) and $\quad {}_0 > (c_0+1)/(c_0\quad 1)$ we have that $\sum_{l\in S\ ;j} j\ {}_{j;l}j\quad {}_0\sum_{l2S\ ;j} j\ {}_{j;l}j$. Applying the empirical compatability condition from Assumption 4.1 to (L.3) then yields

$$D^z_{;j}(\mathsf{b}_j;\bar{}_j)+(1\quad c_0{}^1)\quad {}_{;j}k\ {}_j k_1\quad 2\quad {}_{;j}\quad {}_0{}^1 jS_{\ ;j}j^{1=2}(\ {}^0_j\tilde{\Sigma}_{\ ;j\ \ j})^{1=2} \tag{L.5}$$

Combining (L.4) and (L.5) to get an upper bound on $(\ {}^0_j\tilde{\Sigma}_{\ j})^{1=2}$ gives

$${}_0 k\ {}_j k_2\quad (\ {}^0_j\tilde{\Sigma}_{\ ;j\ \ j})^{1=2}\quad 2\quad {}_{;j}\quad {}_0{}^1 jS_{\ ;j}j^{1=2}\frac{C_0 k\ {}_j k_1}{1\quad e^{\quad C_0 k\ {}_j k_1}}:$$

Plugging the second bound into (L.5) gives

$$D^z_{;j}(\mathsf{b}_j;\bar{}_j)+(1\quad c_0{}^1)\quad {}_{;j}k\ {}_j k_1\quad 2\sum_{l2S\ ;j} j\ {}_{j;l}j\quad 4\ {}^2_{;j}\quad {}_0{}^2 jS_{\ ;j}j\frac{C_0 k\ {}_j k_1}{1\quad e^{\quad C_0 k\ {}_j k_1}}:$$

The second inequality and $\sum_{l\in S\ ;j} j\ {}_{j;l}j\quad {}_0\sum_{l2S\ ;j} j\ {}_{j;l}j$ imply $1\quad e^{\quad C_0 k\ {}_j k_1}\quad 2C_0\quad {}_{;j}\quad {}_0{}^2 jS_{\ ;j}j\quad$ so,

$$\frac{1\quad e^{\quad C_0 k\ {}_j k_1}}{C_0 k\ {}_j k_1}=\int_0^1 e^{\quad C_0 k\ {}_j k_1 u}\,du\quad e^{\quad C_0 k\ {}_j k_1}\quad 1\quad :$$

Combining the last two displays gives

$$D^z_{;j}(\mathsf{b}_j;\bar{}_j)+(1\quad c_0{}^1)\quad {}_{;j}k\mathsf{b}_j\quad \bar{}_j k_1\quad 4\ {}^2_{;j}\quad {}_0{}^2(1\quad)jS_{\ ;j}j \tag{L.6}$$

Applying $\Omega_{k;2}$ to bound $\quad {}_{;j}\quad \bar{}_k$ and noting that $jS_{\ ;j}j\quad s_k$ by definition gives (B.3) with $M_0=\frac{4\ {}_0{}^1(1\quad)}{1\quad c_0{}^1}$. □

For each j, consider the matrices,

$$\begin{aligned}\tilde{\Sigma}_{\ ;j}&:=\mathsf{E}_n[p_j(X)De\quad {}^0_j Z(Y\quad \bar{}^0_j Z)^2 ZZ^0]\\ \tilde{\Sigma}_{\ ;j}&:=\mathsf{E}_n[p_j(X)De\quad {}^0_j ZZZ^0]\end{aligned} \tag{B.4}$$

In addition define $\Sigma_{\ ;j}:=\mathsf{E}\tilde{\Sigma}_{\ ;j}$ and $\Sigma_{\ ;j}:=\mathsf{E}\tilde{\Sigma}_{\ ;j}$. For the outcome regression model, we will consider

nonasymptotic bounds under the following additional events:

$$\begin{aligned}\Omega_{k;3} &:= f \quad _{;j} \quad c_0 k S \ _{;j} k_1 ; 8j \quad kg \\ \Omega_{k;4} &:= f \quad _{;j} \quad \bar{} _k ; 8j \quad kg \\ \Omega_{k;5} &:= fk\tilde{\Sigma} \ _{;j} \quad \Sigma \ _{;j} k_1 \quad \bar{} _k ; 8j \quad kg \\ \Omega_{k;6} &:= fk\tilde{\Sigma} \ _{;j} \quad \Sigma \ _{;j} k_1 \quad \bar{} _k ; 8j \quad kg\end{aligned} \tag{B.5}$$

Lemma B.2 (Nonasymptotic Bounds for Linear Model)**.** *Suppose that Assumptions 4.1 and 4.2 hold,* $_0 > (c_0 + 1)/(c_0 \quad 1)$*, and* $2C_0 \ _0 \ ^2 s_k \bar{} _k \quad < 1$*. In addition, assume there is a constant* $c > 0$ *such that* $_{;j} = \ _{;j} \quad c$ *for all* $j \quad k$*. 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*

$$\max_{1 \ j \ k} D^z \ _{;j}(\mathrm{b}_j ; \bar{} _j ; \bar{} _j) \quad M_1 s_k \bar{} _k^2 \quad \text{and} \quad \max_{1 \ j \ k} k\mathrm{b}_j \quad \bar{} _j k_1 \quad M_1 s_k \bar{} _k \tag{B.6}$$

Proof. We show that the bound of (B.6) holds for each $j = 1; : : : ; k$. We proceed in a few steps.

Step 1: Optimization Step. Let $\tilde{} _j(\ ; \mathrm{b}_j) := \mathsf{E}_n[p_j(X) De \ ^{\mathrm{b}_j^0 Z} fY \quad \ ^0 Zg^2]/2$. Optimality of b_j implies that for any $u \in (0; 1]$:

$$\tilde{} _j \quad \mathrm{b}_j ; \mathrm{b}_j \quad \tilde{} _j \quad (1 \quad u)\mathrm{b}_j + u \bar{} _j ; \mathrm{b}_j \quad + \quad _{;j} k\mathrm{b}_j k_1 \quad _{;j} k(1 \quad u)\mathrm{b}_j + u \bar{} _j k_1 :$$

Convexity of the ℓ_1 norm $k \ k_1$ gives

$$\tilde{} _j \quad \mathrm{b}_j ; \mathrm{b}_j \quad \tilde{} _j \quad (1 \quad u)\mathrm{b}_j + u \bar{} _j ; \mathrm{b}_j \quad + \quad _{;j} uk\mathrm{b}_j k_1 \quad _{;j} uk \bar{} _j k_1 :$$

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

$$\mathsf{E}_n[p_j(X) De \ ^{\mathrm{b}_j^0 Z} fY \quad \mathrm{b}_j^0 Z g f \mathrm{b}_j^0 Z \quad \bar{} _j^0 Z g] + \quad _{;j} k\mathrm{b}_j k_1 \quad _{;j} k \bar{} _j k_1 :$$

Rearranging using the form of $D^z \ _{;j}$ in (4.6) yields:

$$D^z \ _{;j}(\mathrm{b}_j ; \bar{} _j ; \mathrm{b}_j) + \quad _{;j} k\mathrm{b}_j k_1 \quad (\mathrm{b}_j \quad \bar{} _j^0)\mathsf{E}_n[p_j(X) De \ ^{\mathrm{b}^0 Z} fY \quad \bar{} _j^0 Z g Z] + \quad _{;j} k \bar{} _j k_1 \tag{O.1}$$

Step 2: Quasi-Score Domination and relating $\bar{} _j$ *to* b_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 $D^z \ _{;j}(\mathrm{b}_j ; \bar{} _j)$ from (4.6) we find that for some $u \in (0; 1)$:

$$\begin{aligned} D^z \ _{;j}(\mathrm{b}_j ; \bar{} _j) = \quad & \mathsf{E}_n[p_j(X) D f e \ ^{\mathrm{b}_j^0 Z} \quad e \ ^{\ _j^0 Z} g f \mathrm{b}_j^0 Z \quad \bar{} _j^0 Z g] \\ = \quad & \mathsf{E}_n[p_j(X) De \ ^{u(\mathrm{b}_j \quad \ _j)^0 Z} e \ ^{\ _j^0 Z} f \mathrm{b}_j^0 Z \quad \bar{} _j^0 Z g^2] \end{aligned}$$

where the second step uses the mean value theorem:

$$e \ ^{\mathrm{b}_j^0 Z} \quad e \ ^{\ _j^0 Z} = e \ ^{u\mathrm{b}_j^0 Z \quad (1 \quad u) \ _j^0 Z}(\mathrm{b}_j \quad \bar{} _j)^0 Z \tag{O.2}$$

In the event $\Omega_{k;1} \setminus \Omega_{k;2}$ using the bound in (B.3) and the fact that $C_0 \ {}_0{}^2 s_k \bar{}_k \quad < 1$ gives us that

$$C_0 k \mathrm{b}_j \quad \bar{}_j k_1 \quad C_0 M_0 s_k \bar{}_k \quad M_0 \ : \tag{O.3}$$

In the event $\Omega_{k;1} \setminus \Omega_{k;2}$ the bound in (L.6) also gives us that $D^z{}_{;j}(\mathrm{b}_j; \bar{}_j) \quad M_0 s_k \ {}^2{}_{;j}$. Combining the above displays then yields

$$M_0 s_k \ {}^2{}_{;j} \quad D^z{}_{;j}(\mathrm{b}_j; \bar{}_j) \quad e^{M_0} \ \mathsf{E}_n[p_j(X) D e \ {}^{0}_{j} Z f \mathrm{b}^0_j Z \quad \bar{}^0_j Z g^2]: \tag{O.4}$$

Again applying the bound on $C_0 k \mathrm{b}_j \quad \bar{}_j k_1$ (O.3) gives

$$\begin{aligned} D^z{}_{;j}(\mathrm{b}_j; \bar{}_j; \mathrm{b}_j) &= \mathsf{E}_n[p_j(X) D e \ {}^{\mathrm{b}^0_j Z}(\mathrm{b}^0_j Z \quad \bar{}^0_j Z)^2] \\ &= \mathsf{E}_n[p_j(X) D e \ {}^{(\mathrm{b}_j \quad {}_j)^0 Z} e \ {}^{0}_{j} Z (\mathrm{b}^0_j Z \quad \bar{}^0_j Z)^2] \\ &\quad e \ {}^{M_0} \ D^z{}_{;j}(\mathrm{b}_j; \bar{}_j; \bar{}_j) \end{aligned} \tag{O.5}$$

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

$$(\mathrm{b}_j \quad \bar{}_j)^0 \mathsf{E}_n[p_j(X) D e \ {}^{\mathrm{b}^0_j Z} f Y \quad \bar{}^0_j Z g Z] = \underbrace{(\mathrm{b}_j \quad \bar{}_j)^0 \mathsf{E}_n[p_j(X) D e \ {}^{0}_{j} Z f Y \quad \bar{}^0_j Z g Z]}_{\{z \ {}_{1;j}} + \underbrace{\mathsf{E}_n[p_j(X) D f e \ {}^{\mathrm{b}^0_j Z} \quad e \ {}^{0}_{j} Z g f Y \quad \bar{}^0_j Z g f \mathrm{b}^0_j Z \quad \bar{}^0_j Z g]}_{\{z \ {}_{2;j}}$$

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

$${}_{1;j} \quad c_0 \ {}^1 k \mathrm{b}_j \quad \bar{}_j k_1 \quad {}_{;j} \tag{O.6}$$

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

$$\begin{aligned} {}_{2;j} \quad e^{C_0 k \mathrm{b}_j \quad {}_j k_1} \quad & \mathsf{E}_n^{1=2}[p_j(X) D e \ {}^{0}_{j} Z f \mathrm{b}^0 Z \quad \bar{}^0 Z g^2] \\ & \mathsf{E}_n^{1=2}[p_j(X) D e \ {}^{0}_{j} Z f Y \quad \bar{}^0_j Z g^2 f \mathrm{b}^0_j Z \quad \bar{}^0_j Z g^2] \end{aligned} \tag{O.7}$$

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

$$(\mathsf{E}_n \quad \mathsf{E})[p_j(X) D e \ {}^{0}_{j} Z f Y \quad \bar{}^0_j Z g^2 f \mathrm{b}^0_j Z \quad \bar{}^0_j Z g] \quad \bar{}_k k \mathrm{b}_j \quad \bar{}_j k_1^2:$$

By Assumption 4.2 and Lemma E.6 we have that $\mathsf{E}[D f Y \quad \bar{}^0_j Z g^2] \quad D_0^2 + D_1^2$ so that:

$$\mathsf{E}[p_j(X) D e \ {}^{0}_{j} Z f Y \quad \bar{}^0_j Z g^2 f \mathrm{b}^0_j Z \quad \bar{}^0_j Z g^2] \quad (D_0^2 + D_1^2) \mathsf{E}[p_j(X) D e \ {}^{0}_{j} Z f \mathrm{b}^0_j Z \quad \bar{}^0_j Z g^2]:$$

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

$$(\mathsf{E}_n \quad \mathsf{E})[p_j(X) D e \ {}^{0}_{j} Z f \mathrm{b}^0_j Z \quad \bar{}^0_j Z g^2] \quad \bar{}_k k \mathrm{b}_j \quad \bar{}_j k_1:$$

and we can bound $\mathsf{E}_n[p_j(X)De^{\ \ {}^{0}_{j}Z}fb^{0}_{j}Z \quad {}^{-0}_{j}Zg^2]$ using (O.4). Putting this together gives

$$\begin{aligned}\mathsf{E}_n[p_j(X)De^{\ \ {}^{0}_{j}Z}fY \quad {}^{-0}_{j}Zg^2fb^{0}_{j}Z \quad {}^{-0}_{j}Zg^2] \quad {}^{-}_{k}kb_j \quad {}^{-}_{j}k_1^2 \\ +(D_0^2+D_1^2){}^{-}_{k}kb_j \quad {}^{-}_{j}k_1^2 \\ +(D_0^2+D_1^2)e^{\ M_0\ M_0s_k\ ^2}_{;j}\end{aligned} \tag{O.8}$$

Applying convexity of p_- and the bounds on $kb_j \quad {}^{-}_{j}k_1^2$ in the event $\Omega_{k;1} \setminus \Omega_{k;2}$ from (L.6) gives

$$\begin{aligned}{}_{2;j} \quad & fe^{M_0}\ (1+(D_0^2+D_1^2)^{1=2})(M_0{}^{-}_{k} \quad {}_{;j}s_k)^{1=2}+(D_0^2+D_1)^2(M_0s_k\ ^2_{;j})^{1=2}gD^{z}_{;j}(b_j;{}^{-}_{j};{}^{-}_{j})^{1=2} \\ & \tilde{C}f({}^{-}_{k} \quad {}_{;j}s_k)^{1=2}+(s_k \quad {}_{;j})^{1=2}gD^{z}_{;j}(b_j;{}^{-}_{j};{}^{-}_{j})^{1=2}\end{aligned} \tag{O.9}$$

where $\tilde{C}=\max fe^{M_0}\ M_0^{1=2}(1+D_0+D_1);(D_0^2+D_1^2)M_0^{1=2}g$. Combining (O.6) and (O.9) gives a bound on the empirical expectation on the RHS of (O.1).

$$\begin{aligned}(b_j \quad {}^{-}_{j})^{0}\mathsf{E}_n[p_j(X)De^{\ b^{0}_{j}Z}fY \quad {}^{-0}_{j}ZgZ] \quad & \underbrace{c_0{}^{1}kb_j \quad {}^{-}_{j}k_1 \quad {}_{;j}}_{\text{Bound on } {}_{1;j} \text{ from (O.6)}} \\ & +\underbrace{\tilde{C}f({}^{-}_{k} \quad {}_{;j}s_k)^{1=2}+(s_k\ ^2_{;j})^{1=2}gD^{z}_{;j}(b_j;{}^{-}_{j};{}^{-}_{j})^{1=2}}_{\text{Bound on } {}_{2;j} \text{ from (O.9)}}\end{aligned} \tag{O.10}$$

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

Step 3: Express Minimization Constraint in Terms of ${}^{-}_{j}$ 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

$$e^{\ M_0}\ D^{z}_{;j}(b_j;{}^{-}_{j};{}^{-}_{j})+ \quad {}_{;j}kb_jk_1 \quad c_0{}^{1} \quad {}_{;j}kb_j \quad {}^{-}_{j}k_1+ \quad {}_{;j}k{}^{-}_{j}k_1+ \quad {}_{2;j} \tag{O.11}$$

Apply the triangle inequality $jb_{j;l}j \quad j{}^{-}_{j;l}j \quad jb_{j;l} \quad {}^{-}_{j;l}j$ for $l \in S_{;j}$ and $jb_{j;l}j=jb_{j;l} \quad {}^{-}_{j;l}j$ for $l \notin S_{;j}$ to the above to obtain

$$e^{\ M_0}\ D^{z}_{;j}(b_j;{}^{-}_{j};{}^{-}_{j})+(1 \quad c_0{}^{1})kb_j \quad {}^{-}_{j}k_1 \quad 2 \quad {}_{;j}\sum_{l\in S_{;j}} jb_{j;l} \quad {}^{-}_{j;l}j+ \quad {}_{2;j}:$$

Let ${}_j=b_j \quad {}^{-}_{j}$. We use the form $D^{z}_{;j}(b_j;{}^{-}_{j})=\mathsf{E}_n[p_j(X)De^{\ {}^{0}_{j}Z}fb^{0}_{j}Z \quad {}^{-0}_{j}Zg^2]= \quad {}^{0}_{j}\tilde{\Sigma}_{;j\ j}$ to expand out

$$\begin{aligned}e^{\ M_0}\ (\ {}^{0}_{j}\tilde{\Sigma}_{;j\ j})+(1 \quad c_0{}^{1}) \quad {}_{;j}k\ k_1 \quad 2 \quad {}_{;j}\sum_{l\in S_{;j}} j\ {}_{j;l}j \\ +\tilde{C}f(s_k{}^{-}_{k} \quad {}_{;j})^{1=2}+(s_k \quad {}_{;j})^{1=2}g(\ {}^{0}_{j}\tilde{\Sigma}_{;j\ j})^{1=2}\end{aligned} \tag{O.12}$$

Step 4: Apply Empirical Compatability Condition. Let ${}_{3;j}:=\tilde{C}f(s_k{}^{-}_{k} \quad {}_{;j})^{1=2}+(s_k \quad {}_{;j})^{1=2}g$ and $D^{?}_{;j}:=e^{\ M_0}\ (\ {}^{0}_{j}\tilde{\Sigma}_{;j\ j})+(1 \quad c_0{}^{1}) \quad {}_{;j}k\ {}_{j}k_1$. In the even $\Omega_{k;1} \setminus \Omega_{k;2} \setminus \Omega_{k;3} \setminus \Omega_{k;5} \setminus \Omega_{k;6}$ that (O.12) holds, there are two possibilities. For ${}_2=1 \quad 2c_0=f(\ {}_1+1)(c_0 \quad 1)g \in (0;1]$ either

$${}_2D^{?}_{;j} \quad {}_{3;j}(\ {}^{0}_{j}\tilde{\Sigma}_{;j\ j})^{1=2} \tag{O.13}$$

or $(1 \quad {}_2)D^{?}{}_{;j} \quad 2 \quad {}_{;j} \sum_{l2S\ {}_{;j}} j\ {}_{j;l}j$, that is

$$D^{?}{}_{j} \quad (\ {}_1+1)(c_0 \quad 1)c_0{}^{1} \quad {}_{;j} \sum_{l2S\ {}_{;j}} j\ {}_{j;l}j \tag{O.14}$$

We deal with these two cases separately. First, if (O.14) holds, then $\sum_{l\in S\ {}_{;j}} j\ {}_{j;l}j \quad {}_1 \sum_{l2S_{j;l}} j\ {}_{j;l}j$. We can apply the empirical compatability of Assumption 4.2 to (O.14) to obtain.

$$e^{\ M_0}\ (\ {}^{0}_{j}\tilde{\Sigma}\ {}_{;j}\ {}_j)+(1 \quad c_0{}^{1}) \quad {}_{;j}k\ {}_{j;l}k \quad {}_1(\ {}_1+1)(\ {}_1 \quad 1) \quad {}_{;j}(s_j\ {}_j\tilde{\Sigma}\ {}_{;j}\ {}_j)^{1=2}:$$

Inverting for $(\ {}_j\tilde{\Sigma}\ {}_{;j}\ {}_j)^{1=2}$ and plugging in gives

$$e^{\ M_0}\ D^{z}{}_{;j}(\mathrm{b};{}^{-}{}_j;{}^{-}{}_j)+(1 \quad c_0{}^{1}) \quad {}_{;j}k\mathrm{b}_j \quad {}^{-}{}_jk_1 \quad \tilde{M}s_k\ {}^{2}{}_{;j} \tag{O.15}$$

where $\tilde{M}=e^{M_0}\ (\ {}_1+1)(c_0 \quad 1)c_0{}^{1}$. Next, assume that (O.13) holds. In this case, we can directly invert for $(\ {}_j\tilde{\Sigma}\ {}_{;j}\ {}_j)^{1=2}$ to get that

$$e^{\ M_0}\ D^{z}{}_{;j}(\mathrm{b}_j;{}^{-}{}_j;{}^{-}{}_j)+(1 \quad c_0{}^{1}) \quad {}_{;j}k\mathrm{b}_j \quad {}^{-}{}_jk_1 \quad {}_2{}^{1}\tilde{C}f(s_k{}^{-}{}_k \quad {}_{;j})^{1=2}+(s_k\ {}^{2}{}_{;j})^{1=2}g^2 \tag{O.16}$$

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

$$\begin{aligned} e^{\ M_0}\ D^{z}{}_{;j}(\mathrm{b}_j;{}^{-}{}_j;{}^{-}{}_j)+(1 \quad c_0{}^{1}) \quad {}_{;j}k\mathrm{b}_j \quad {}^{-}{}_jk_1 \quad \tilde{M}s_k\ {}^{2}{}_{;j} \\ + \ {}_2{}^{1}\tilde{C}f(s_k{}^{-}{}_k \quad {}_{;j})^{1=2}+(s_k\ {}^{2}{}_{;j})^{1=2}g^2 \end{aligned} \tag{O.17}$$

Step 5: Apply Penalty Majorization and Bounded Penalty Ratio. Use the fact that ${}_{;j}= \quad {}_{;j} \quad c\ {}^{1}$ to express (O.17) as

$$\begin{aligned} D^{z}{}_{;j}(\mathrm{b}_j;{}^{-}{}_j;{}^{-}{}_j) \quad e^{M_0}\ \tilde{M}s_k\ {}^{2}{}_{;j}+e^{M_0} \quad {}_2{}^{1}\tilde{C}f(s_k{}^{-}{}_k \quad {}_{;j})^{1=2}+(s_k\ {}^{2}{}_{;j})^{1=2}g^2 \\ k\mathrm{b}_j \quad {}^{-}{}_jk_1 \quad (1 \quad c_0{}^{1})\ {}^{1}\tilde{M}s_k \quad {}_{;j}+(1 \quad c_0{}^{1})\ {}^{1}c\ {}^{1}\tilde{C}f(s_k{}^{-}{}_k)^{1=2}+(s_k \quad {}_{;j})^{1=2}g^2 \end{aligned}$$

In the event $\Omega_{k;2} \setminus \Omega_{k;3}$ we have that ${}_{;j}\_ \quad {}_{;j} \quad {}^{-}{}_k$, so that the above simplifies to

$$\begin{aligned} D^{z}{}_{;j}(\mathrm{b}_j;{}^{-}{}_j;{}^{-}{}_j) \quad M_1s_k{}^{-}{}^{2}_{k} \\ k\mathrm{b}_j \quad {}^{-}{}_jk_1 \quad M_1s_k{}^{-}{}_k \end{aligned} \tag{O.18}$$

for $M_1=\max\{e^{M_0}\ ;c\ {}^{1}(1 \quad c_0{}^{1})\ {}^{1}g(\tilde{M}+2e^{M_0} \quad {}_2{}^{1}\tilde{C})$. This completes the result (B.6). □

Let $\tilde{\Sigma}^{1}{}_{;j}:=\mathsf{E}_n[p_j(X)De \quad {}^{0}_{j}ZjY \quad {}^{-0}_{j}ZjZZ^{0}]$ and $\Sigma^{1}{}_{;j}:=\mathsf{E}\tilde{\Sigma}^{1}{}_{;j}$. Define the event

$$\Omega_{k;7}:=\{k\tilde{\Sigma}^{1}{}_{;j} \quad \Sigma^{1}{}_{;j}k_1 \quad {}^{-}{}_k;8j \quad kg \tag{B.7}$$

Lemma B.3 (Nonasymptotic Bounds for Weighted Means)**.** *Suppose that Assumptions 4.1 and 4.2 hold,* ${}_0 > (c_0+1)=(c_0 \quad 1)$*, and* $2C_0\ {}_0{}^{2}s_k{}^{-}{}_k \quad < 1$*. In addition, assume there is a constant* $c>0$ *such that* ${}_{;j}= \quad {}_{;j} \quad c$ *for all* $j \quad k$*. 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

$$\max_{1 \le j \le k} |\mathsf{E}_n[p_j(X)Y(\beta_j; \hat{m}_j)] - \mathsf{E}_n[p_j(X)Y(\bar{\beta}_j; \bar{m}_j)]| \le M_2 s_k^{-2} \tag{B.8}$$

Proof. We show that the bound of (B.8) holds for any $j = 1, \ldots, k$ in a couple steps. To save notation, define

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

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

$$\begin{aligned} \psi_j(\beta_j; \hat{m}_j) - \psi_j(\bar{\beta}_j; \bar{m}_j) &= \mathsf{E}\left[p_j(X)\{\hat{m}_j(Z) - \bar{m}_j(Z)\}\left(1 - \frac{D}{\bar{e}_j(X)}\right)\right] \\ &\quad + \mathsf{E}_n\left[p_j(X)D\{Y - \bar{m}_j(Z)\}\left(\frac{1}{\hat{e}_j(Z)} - \frac{1}{\bar{e}_j(Z)}\right)\right] \\ &\quad + \mathsf{E}_n\left[p_j(X)\{\hat{m}_j(Z) - \bar{m}_j(Z)\}\left(\frac{D}{\bar{e}_j(Z)} - \frac{D}{\hat{e}_j(Z)}\right)\right] \\ &:= \Delta_{1;j} + \Delta_{2;j} + \Delta_{3;j} \end{aligned}$$

Notice that $\Delta_{1;j} + \Delta_{3;j} = (\hat{\beta}_j - \bar{\beta}_j)'\mathsf{E}_n[p_j(X)(1 - D/\bar{e}_j(Z))Z]$. By the first order conditions for $\hat{\beta}_j$ we have that

$$|\mathsf{E}_n[p_j(X)\{Z_l - D\bar{Z}_l/\bar{e}_j(Z)\}]| \le \varepsilon_{j,n} \;\; \forall l = 1, \ldots, d_z \;\Rightarrow\; \|\mathsf{E}_n[p_j(X)\{Z_l - D\bar{Z}_l/\bar{e}_j(Z)\}]\|_1 \le \varepsilon_{j,n}.$$

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

$$|\Delta_{1;j} + \Delta_{3;j}| \le \|\hat{\beta}_j - \bar{\beta}_j\|_1 \varepsilon_{j,n} \le \|\hat{\beta}_j - \bar{\beta}_j\| \varepsilon_{j,n}.$$

By Lemma B.2 on the event $\bigcap_{m=1}^{6} \Omega_{k;m}$ and under the conditions of Lemma B.3, $\|\hat{\beta}_j - \bar{\beta}_j\| \le M_1 s_k \bar{\kappa}_k$ where M_1 is a constant that does not depend on k. So

$$|\Delta_{1;j} + \Delta_{3;j}| \le M_1 s_k \bar{\kappa}_k^{-2} \tag{M.1}$$

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 $u \in (0,1)$

$$\begin{aligned} \Delta_{2;j} &= (\hat{\beta}_j - \bar{\beta}_j)'\mathsf{E}_n[p_j(X)D\{Y - \bar{m}_j(Z)\}e^{\bar{\beta}_j'Z}Z] \\ &\quad + (\hat{\beta}_j - \bar{\beta}_j)'\mathsf{E}_n[p_j(X)D\{Y - \bar{m}_j(Z)\}e^{u\bar{\beta}_j'Z + (1-u)\hat{\beta}_j'Z}ZZ']'(\hat{\beta}_j - \bar{\beta}_j)/2 \\ &:= \Delta_{21;j} + \Delta_{22;j} \end{aligned}$$

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 $\Delta_{21} \le M_0 s_k \bar{\kappa}_k^{-2}$.

The term $\Delta_{22;j}$ is second order. On the event $\Omega_{k;0} \cap \Omega_{k;1}$ where $\|\hat{\beta}_j - \bar{\beta}_j\|_1 \le M_0 s_k \bar{\kappa}_k \le M_0 / C_0$ it can be

bounded with

$$\begin{aligned} {}_{22;j} \quad & e^{C_0 k b_j} \quad {}_j k_1 \mathsf{E}_n[p_j(X) D e \quad {}^{0}_{j} Z j Y \quad \bar{m}_j(Z) j f b^{0}_{j} Z \quad {}^{-0}_{j} Z g^2] \\ & e^{M_0} \; \mathsf{E}_n[p_j(X) D e \quad {}^{0}_{j} Z j Y \quad \bar{m}_j(Z) j f b^{0}_{j} Z \quad {}^{-0}_{j} Z g^2]: \end{aligned}$$

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

$$(\mathsf{E}_n \quad \mathsf{E})[p_j(X) D e \quad {}^{0}_{j} Z j Y \quad \bar{m}_j(Z) j f b^{0}_{j} Z \quad {}^{-0}_{j} Z g^2] \quad {}^{-}_{k} k b_j \quad {}^{-}_{j} k_1^2:$$

By Assumption 4.2 and Lemma E.8 we have that $D_0^2 E[D j Y \quad \bar{m}_j(Z) j \, j \, Z] \quad D_1^2 = D_0 + D_0$ so that,

$$\mathsf{E}[p_j(X) D e \quad {}^{0}_{j} Z j Y \quad \bar{m}_j(Z) j f b^{0}_{j} Z \quad {}^{-0}_{j} Z g^2] \quad (D_1^2 = D_0 + D_0) \mathsf{E}[p_j(X) D e \quad {}^{0}_{j} Z f b^{0}_{j} Z \quad {}^{-0}_{j} Z g^2]:$$

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

$$(\mathsf{E}_n \quad \mathsf{E})[p_j(X) D e \quad {}^{0}_{j} Z f b^{0}_{j} Z \quad {}^{-0}_{j} Z g^2] \quad {}^{-}_{k} k b_j \quad {}^{-}_{j} k_1:$$

Putting these all together gives

$$\begin{aligned} \mathsf{E}_n[p_j(X) D e \quad {}^{0}_{j} Z j Y \quad \bar{m}_j(Z) j f b^{0}_{j} Z \quad {}^{-0}_{j} Z g^2] \quad {}^{-}_{k} k b_j \quad {}^{-}_{j} k_1^2 + (D_1^2 = D_0 + D_0) {}^{-}_{k} k b_j \quad {}^{-}_{j} k_1^2 \\ + (D_1^2 = D_0 + D_0) \mathsf{E}_n[p_j(X) D e \quad {}^{0}_{Z} f b^{0}_{j} Z \quad {}^{-0}_{j} Z g] \end{aligned} \tag{M.2}$$

To bound (M.2) note again that in the event $\Omega_{k;1} \setminus \Omega_{k;2}$, $k b_j \quad {}^{-}_{j} k_1 \quad M_0 s_k {}^{-}_{k}$ and that using (O.4)

$$\mathsf{E}_n[p_j(X) D e \quad {}^{0}_{j} Z f b^{0}_{j} Z \quad {}^{-0}_{j} Z g^2] \quad e \quad {}^{M_0} \; M_0 s_k {}^{-2}_{k}:$$

Plugging these into (M.2) gives

$${}_{22;j} \quad e^{M_0} \; M_0^2 s_k^{2} {}^{-3}_{k} + e^{M_0} \; (D_1^2 = D_0 + D_0) M_0^2 s_k^{2} {}^{-3}_{k} + (D_1^2 = D_0 + D_0) M_0 s_k {}^{-2}_{k} \tag{M.3}$$

so that in total ${}_{2;j} = {}_{21;j} + {}_{22;j}$ is bouned

$${}_{2;j} \quad M_0 s_k (D_1^2 = D_0 + D_0 + 1) {}^{-2}_{k} + e^{M_0} \; M_0^2 s_k^2 (D_1^2 = D_0 + D_0 + 1) {}^{-3}_{k} \tag{M.4}$$

Step 3: Combine Terms. Putting this together yields

$$\begin{aligned} j \, {}_{1;j} + {}_{2;j} + {}_{3;j} j \quad & f M_1 + M_0 (D_1^2 = D_0 + D_0 + 1) g s_k {}^{-2}_{k} \\ & + e^{M_0} \; (D_1^2 = D_0 + D_0) M_0^2 s_k^{2} {}^{-3}_{k} \end{aligned} \tag{M.5}$$

Use the fact that $s_k {}^{-}_{k} \quad < 1$ to simplify the last term of this expression

$$\begin{aligned} j \, {}_{1;j} + {}_{2;j} + {}_{3;j} j \quad & f M_1 + M_0 (D_1^2 = D_0 + D_0 + 1) g s_k {}^{-2}_{k} \\ & + e^{M_0} \; (D_1^2 = D_0 + D_0) M_0^2 s_k {}^{-}_{k} \end{aligned} \tag{M.6}$$

This gives the result (B.8) after taking $M_2 = M_1 + M_0 (D_1^{=} D_0 + D_0 + 1) + e^{M_0} \; (D_1^2 = D_0 + D_0) M_0^2$.

□

Lemma B.4 (Nonasymptotic Bounds for Variance Estimation)**.** *Suppose that Assumptions 4.1 and 4.2 hold,* $_0 > (c_0 + 1)/(c_0 \quad 1)$*, and* $2C_0 \ _0{}^2 s_k^{\ -}{}_k \quad < 1$*. In addition, assume there is a constant* $c > 0$ *such that* $_{;j} = \ _{;j} \quad c$ *for all* $j \quad k$*. 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*

$$\max_{1 \ j \ k} \mathsf{E}_n[p_j^2(X)(Y(\mathsf{b}_j; \hat{m}_j) \quad Y(\bar{}_j; \bar{m}_j))^2] \quad M_3 \ _{k;1} s_k^{2\ -}{}_k^2 \tag{B.9}$$

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

$$\begin{aligned} p_j(X)(Y(\mathsf{b}_j; \hat{m}_j) \quad Y(\bar{}_j; \bar{m}_j)) &= p_j(X)\{\hat{m}_j(Z) \quad \bar{m}_j(Z)\} \quad 1 \quad \frac{D}{\bar{}_j(X)} \\ &\quad + p_j(X)D\{Y \quad \bar{m}_j(Z)\} \ \frac{1}{\mathsf{b}_j(Z)} \quad \frac{1}{\bar{}_j(Z)} \\ &\quad + p_j(X)\{\hat{m}_j(Z) \quad \bar{m}_j(Z)\} \ \frac{D}{\bar{}_j(Z)} \quad \frac{D}{\mathsf{b}_j(Z)} \\ &:= \tilde{}_{1;j} + \tilde{}_{2;j} + \tilde{}_{3;j} \end{aligned}$$

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

$$\mathsf{E}_n[p_j^2(X)(Y(\mathsf{b}_j; \hat{m}_j) \quad Y(\bar{}_j; \bar{m}_j))^2] \quad 4\mathsf{E}_n[\tilde{}^2_{1;j}] + 4\mathsf{E}_n[\tilde{}^2_{2;j}] + 4\mathsf{E}_n[\tilde{}^2_{3;j}]: \tag{V.1}$$

To bound $\mathsf{E}_n[\tilde{}^2_{2;j}]$ use the mean value equation (O.2) and the lower bound on $\bar{g}_j(z)$ from Assumption 4.1

$$\begin{aligned} \mathsf{E}_n[\tilde{}^2_{2;j}] &= \mathsf{E}_n[p_j^2(X)D\{Y \quad \bar{m}_j(Z)\}^2\{\mathsf{b}_j \ ^1(Z) \quad \bar{}_j \ ^1(Z)\}^2] \\ &\quad _{k;1} e \ ^{B_0} \ 1 + e^{C_0 k \mathsf{b}_j} \quad _j{}_{k_1} \ ^2 \mathsf{E}_n[p_j(X)De \ ^{{}^0_j Z} \{Y \quad \bar{m}_j(Z)\}^2\{\hat{\mathsf{b}}_j(Z) \quad \bar{g}_j(Z)\}^2] \end{aligned}$$

Applying (O.8), Lemma B.1, and $s_k \ ^{-}{}_k \quad < 1$ 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

$$\tilde{M}_1 \ _{k;1} s_k \ ^{-2}_k \tag{V.2}$$

To bound $\mathsf{E}_n[\tilde{}^2_{3;j}]$ write $\mathsf{b}_j \ ^1(Z) \quad \bar{}_j \ ^1(Z) = e \ ^{{}^0_j Z} \{e \ ^{\mathsf{b}^0_j Z + \ {}^0_j Z} \quad 1\}$ and use the lower bound on $\bar{g}_j(z)$ from Assumption 4.1:

$$\begin{aligned} \mathsf{E}_n[\tilde{}^2_{3;j}] &= \mathsf{E}_n[p_j^2(X)D\{\hat{m}_j(Z) \quad \bar{m}_j(Z)\}^2\{\mathsf{b}_j \ ^1(Z) \quad \bar{}_j \ ^1(Z)\}^2] \\ &\quad _{k;1} e \ ^{B_0} \ 1 + e^{C_0 k \mathsf{b}_j} \quad _j{}_{k_1} \ ^2 \mathsf{E}_n[p_j(X)e \ ^{{}^0_j Z} \{\hat{m}_j(Z) \quad \bar{m}_j(Z)\}^2] \end{aligned}$$

Noting that this last term is exactly $D^z{}_{;j}(\mathsf{b}_j; \bar{}_j; \bar{}_j)$ and applying 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

$$\tilde{M}_2 \ _{k;1} s_k \ ^{-2}_k \tag{V.3}$$

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

$$\begin{aligned}\mathsf{E}_n[\tilde{\ }^2_{1;j}] = \mathsf{E}_n[p_j^2(X)\,fm(z) \quad \bar{m}(z)g^2 f1 \quad D=^-{}_j(Z)g^2] \\ {}^2_{k;1}(1+e^{\ B_0})^2\mathsf{E}_n[fm_j(Z) \quad \bar{m}_j(Z)g^2] \\ {}^2_{k;1}(1+e^{\ B_0})^2 C_0^2 k\mathsf{b}_j \quad {}^-{}_j k_1^2\end{aligned}$$

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

$$\tilde{M}_3\ {}^2_{k;1}\, s_k^{2\,-2}_{\ k} \tag{V.4}$$

The result (B.9) follows by collecting (V.1)-(V.4). □

B.2 Nonasymptotic Bounds for Residual Estimation

We now provide nonasymptotic bounds that are useful for showing consistent residual estimation. These bounds will be shown under the events in (B.2), (B.5), and (B.7) using the results in Lemmas B.1 and B.2.

Lemma B.5 (Nonasymptotic Logistic Residual Bound). *Suppose that Assumption 4.1 and the conditions of Lemma B.1 hold. Then, in the event $\Omega_{k;1} \setminus \Omega_{k;2}$ described on (B.2) there is a constant $M_{\ ;r}$ that does not depend on k such that:*

$$\max_{1\ j\ k} \mathsf{E}_n[(\hat{\theta}_{\ ;j} \quad U_{\ ;j})^2] \quad M_{\ ;r\ k;1}\, s_k^{\ -2}{}_k; \tag{B.10}$$

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

$$\begin{aligned}(\hat{\theta}_{\ ;j} \quad U_{\ ;j})^2 = p_j(X)^2 D f e^{\ \mathsf{b}_j^{0\prime} Z} \quad e^{\ {}^0_j Z} g f e^{\ \mathsf{b}_j^{0\prime} Z} \quad e^{\ {}^0_j Z} g \\ {}_{k;1}\, p_j(X) D f e^{\ \mathsf{b}_j^{0\prime} Z} \quad e^{\ {}^0_j Z} g e^{\ {}^0_j Z\ \ u(\mathsf{b}_j \quad {}_j)^{0\prime} Z} f^{-0}_j Z \quad \mathsf{b}_j^{0\prime} Z g \\ {}_{k;1}\, e^{\ B_0+M_0} D f e^{\ \mathsf{b}_j^{0\prime} Z} \quad e^{\ {}^0_j Z} g f^{-0}_j Z \quad \mathsf{b}_j^{0\prime} Z g\end{aligned}$$

So that

$$\begin{aligned}\mathsf{E}_n[(\hat{\theta}_{\ ;j} \quad U_{\ ;j})^2] \quad e^{\ B_0+M_0}\ {}_{k;1}\ \underbrace{\mathsf{E}_n[p_j(X) D f e^{\ \mathsf{b}_j^{0\prime} Z} g f \mathsf{b}_j^{0\prime} Z \quad {}^{-0}_j Z g]}_{=D^z{}_{;j}(\mathsf{b}_j;\ {}_j)} \\ e^{\ B_0+M_0}\ {}_{k;1}\, s_k^{\ -2}{}_k\end{aligned}$$

□

Lemma B.6 (Nonasymptotic Linear Residual Bound). *Suppose that Assumptions 4.1 and 4.2 and the conditions of Lemma B.2 hold. Then, in the event $\mathsf{T}^6_{m=1}\Omega_{k;m}$, there is a constant $M_{\ ;r}$ that does not depend on k such that*

$$\max_{1\ j\ k} \mathsf{E}_n[(\hat{\theta}_{\ ;j} \quad U_{\ ;j})^2] \quad M_{\ ;r\ {}^2_{k;1}}\, s_k^{2\,-2}{}_k \tag{B.11}$$

Proof. Recall that $\hat{\theta}_{\ ;j} = p_j(X) D e^{\ \mathsf{b}_j^{0\prime} Z}(Y \quad \mathsf{b}_j^{0\prime} Z)$ and $U_{\ ;j} = p_j(X) D e^{\ {}^0_j Z}(Y \quad {}^{-0}_j Z)$. As an intermediary, define $\dot{U}_{\ ;j} = p_j(X) D e^{\ \mathsf{b}_j^{0\prime} Z}(Y \quad {}^{-0}_j Z)$. We will show a bound on the empirical mean square error between

$\hat{\theta}_{;j}$ and $\dot{U}_{;j}$ as well as on the empirical mean square error between $\dot{U}_{;j}$ and $U_{;j}$. The bound in (B.11) will then follow from $(a+b)^2 \quad 2a^2+2b^2$.

First consider $(\hat{\theta}_{;j} \quad \dot{U}_{;j})^2$:

$$\begin{aligned}
\mathsf{E}_n[(\hat{\theta}_{;j} \quad \bar{U}_{;j})^2] &= \mathsf{E}_n p_j^2(X) De^{\ 2b_j^0 Z}(b_j^0 Z \quad {}_j^0 Z)^2] \\
&= \mathsf{E}_n[p_j^2(X) De^{\ 2({}_j^0 Z \quad (b_j \quad {}_j)^0 Z)}(b_j^0 Z \quad {}_j^0 Z)^2] \\
&\quad {}_{k\,1} e^{\ B_0} e^{2M_0} \underbrace{\mathsf{E}_n[p_j(X) De^{\ {}_j^0 Z}(b_j^0 Z \quad {}_j^0 Z)]}_{= D^z{}_{;j}(b_j;\ {}_j;\ {}_j)} \\
&\quad e^{2M_0} \quad {}^{B_0} M_{1\ k;1} s_k^{\ -2}{}_k
\end{aligned}$$

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

$$\begin{aligned}
\mathsf{E}_n[(\dot{U}_{;j} \quad U_{;j})^2] &= \mathsf{E}_n[p_j^2(X) D f e^{\ b^0 Z} \quad e^{\ {}^0 Z} g^2 f Y \quad {}_j^0 Z g^2] \\
&= \mathsf{E}_n[p_j^2(X) D f e^{\ {}^0 Z \ u(b \quad)^0 Z}(-{}_j^0 Z \quad b_j^0 Z) g^2 (Y \quad {}_j^0 Z)^2] \\
&\quad 2e^{M_0} \quad {}^{B_0} C_0^2{}_{k;1}(M_1 s_k^{\ -}{}_k)^2 \mathsf{E}_n[p_j(X) De^{\ {}_j^0 Z}(Y \quad {}_j^0 Z)^2]
\end{aligned}$$

To proceed we assume that Z contains a constant. That is $Z = (1; Z_2; :::; Z_{d_z})$. 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

$$\begin{aligned}
\mathsf{E}_n[p_j(X) De^{\ {}_j^0 Z}(Y \quad {}_j^0 Z)^2] &= (\mathsf{E}_n \quad \mathsf{E})[p_j(X) De^{\ {}_j^0 Z}(Y \quad {}_j^0 Z)^2] + \mathsf{E}[p_j(X) De^{\ {}_j^0 Z}(Y \quad \bar{m}_j(X))^2] \\
&\quad {}^-{}_k + \quad {}_{k;1} e^{\ B_0}(D_0 + D_1)^2:
\end{aligned}$$

Combining the above, and using the fact that $s_k^{\ -}{}_k \quad < 1$ 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.5), and (B.7) occurs under Assumptions 4.1 and 4.2 with probability approaching one.

Lemma B.7 (Logistic Score Domination and Penalty Majorization)**.** *Suppose Assumption 4.1 holds and that the penalty parameter ${}_{;j}$ is chosen as described in Section 4.1. Then, for n su ciently large, the event $\Omega_{k;1}$ holds with probability $1 \quad {}_{;n}$ where*

$$_{;n} = C \max\left(\frac{4kn+4k}{n^2};\ \frac{\tilde{M}_{\ k;1} s_{k;}\ \bar{c}_n^2 \ln^5(d_z n)}{n}^{\ 1=2};\ \frac{\tilde{M}^{\ 4}_{\ k;1} \ln^7(d_z kn)}{n}^{\ 1=6};\ \frac{1}{\ln^2(d_z kn)} \right) : \tag{B.12}$$

where $C; \tilde{M}$ are absolute constants that do not depend on k. In particular so long as $\ ! \ 0$ as $n \ ! \ 1$, this shows that $\Pr(\Omega_{k;1}) = 1 \quad o(1)$ under the rate conditions of Assumption 4.1.

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

$$^-{}_k = \max\{M_2; M_4; M_5; M_6; M_7\}_{\ k;1} \sqrt{\frac{\ln(d_z n)}{n}} \tag{B.13}$$

where M_4, M_5, M_6 and M_7 are all constants that also do not depend on k described in Lemma B.8 and Lemmas B.9-B.11. 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.5 and the probability bounds from Lemmas B.9–B.12 we find that, (eventually) with probability at least $1 - \frac{4k}{n} - \frac{4k}{n^2}$:

$$\max_{1 \le j \le k} \mathsf{E}_n[(\hat{\theta}_{\cdot,j} Z_l - U_{\cdot,j} Z_l)^2] \le M_{\cdot,r} C_0^2 \frac{\kappa_{k,1} s_{k,} \bar{c}_n^2 \ln^3(d_z n)}{n}. \tag{P.1}$$

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

$$\begin{aligned} W_k &:= (U_{\cdot,1} Z^0, \ldots, U_{\cdot,k} Z^0)' \in \mathsf{R}^{k d_z} \\ &:= (W_{k,1}^0, \ldots, W_{k,k}^0)' \\ \widehat{W}_k &:= (\hat{\theta}_{\cdot,1} Z^0, \ldots, \hat{\theta}_{\cdot,k} Z^0)' \in \mathsf{R}^{k d_z} \\ &:= (\widehat{W}_{k,1}^0, \ldots, \widehat{W}_{k,k}^0)'. \end{aligned}$$

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 $\Sigma_k = \frac{1}{n} \sum_{i=1}^n \mathsf{E}[W_k W_k']$ exists and is finite. Define the sequences of constants

$$\begin{aligned} \delta_{\cdot,n}^2 &:= M_{\cdot,r} C_0^2 \kappa_{k,1} s_{k,} \bar{c}_n^2 \ln^5(d_z n)/n \\ \varepsilon_{\cdot,n} &:= \frac{4k}{n} + \frac{4k}{n^2} \end{aligned}$$

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

$$\Pr\left[k \mathsf{E}_n[(\widehat{W}_k - W_k)^2] k_1 > \delta_n^2 / \ln^2(d_z n) \right] \le \varepsilon_n. \tag{P.2}$$

Let $e_1, \ldots, e_n$ be i.i.d normal random variables generated independently of the data as in Section 4.1. Define the scaled random variables and the multiplier bootstrap process

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

and let $\Pr_e$ denote the probability measure with respect to the e_i's conditional on the observed data. Assumption 4.1 implies that the conditions of (E.1) hold for $Z = W_k$ with b replaced by c_u and B_n replaced by $B_k = (\kappa_{k,1} C_0 C_U)^3 \vee 1$. Further, via (P.2) the residual estimation requirement of with δ_n and ε_n replaced by $\delta_{\cdot,n}$ and $\varepsilon_{\cdot,n}$.

Let $\hat{\phi}_{\cdot,j}(\cdot)$ be the $\cdot$ quantile of $k\hat{S}_{\cdot,j}^{e'} k$ *conditional* on the data Z_i and the estimates $\hat{\theta}_i$. Theorem E.4 then shows that there is a finite constant depending only on c_u such that

$$\max_{1 \le j \le k} \sup_{\cdot \in (0,1)} \Pr(k S_{\cdot,j} k \le \hat{\phi}_{\cdot,j}(\cdot)) - \cdot \le C \max\left(\delta_{\cdot,n}, \varepsilon_{\cdot,n}, \frac{B_k^4 \ln^7(k d_z n)}{n} \right)^{1/6}, \frac{1}{\ln^2(k d_z n)} \Bigg).$$

This gives the first claim of Lemma B.7 by construction of $\hat{c}_{\cdot,j}$.

The second claim follows Lemma E.1. For this second claim we will consider the marginal convergence of each $U_{;j}Z$ as opposed to their joint convergence (the convergence of W_k). First, notice that condiitonal on the data, the random vector $\mathsf{E}_n[e\hat{\theta}_{;j}Z]$ is centered gaussian in R^{d_z}. Lemma E.1 then shows that

$$\phi_{;j}(\ ) \quad (2+\overline{p2})\ \sqrt{\frac{\ln(d_z=\ )}{n}}\ \max_{1\ l\ d_z}\mathsf{E}_n[\hat{\theta}^2_{;j}Z_l^2]:$$

Furthermore, with probability at least $1 \quad {}_{;n} \quad \frac{1}{n}$ we have that, for all $j = 1;:::;k$:

$$\max_{1\ l\ d_z}\mathsf{E}_n[\hat{\theta}^2_{;j}Z_l^2] \quad C_0^2\mathsf{E}_n[\hat{\theta}^2_{;j}] \quad 2C_0^2(\mathsf{E}_n[U^2_{;j}]+\mathsf{E}_n[(\hat{\theta}^2_{;j} \quad U_{;j})^2]) \quad 4C_0^2\ {}^2_{k;1}C_U^2 + {}^2_{;n}=\ln^2(d_z n))$$

Under the rate conditions of Assumption 4.1, ${}^2_{;n}=\ln^2(d_z n)$ will eventually be smaller than 1 and so the claim in (B.13) holds with $M_2 = 8C_0^2C_U^2 \_ 1$. □

Lemma B.8 (Linear Score Domination and Penalty Majorization)**.** *Suppose Assumptions 4.1 and 4.2 holds and that the penalty parameters* ${}_{;j}$ *and* ${}_{;j}$ *are chosen as described in Section 4.1. Then, for n su ciently large, the event* $\Omega_{k;3}$ *holds with probability* $1 \quad {}_{;n}$ *where:*

$${}_{;n} = C\max\left(\frac{4kn+4k}{n^2};\ \frac{\tilde{M}\ {}^2_{k;1}s^2_{k;}\ \bar{c}_n^2\ln^5(d_z n)}{n}\ {}^{1=2};\ \frac{\tilde{M}\ {}^4_{k;1}\ln^7(d_z kn)}{n}\ {}^{1=6};\frac{1}{\ln^2(d_z kn)}\right): \tag{B.14}$$

where $C;\tilde{M}$ *are absolute constants that do not depend on k. In particular so long as* $\ !\ 0$ *as* $n\ !\ 1$ *, this shows that* $\Pr(\Omega_{k;3}) = 1 \quad o(1)$ *under Assumption 4.1.*

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

$$\bar{}_k = \max\{M_2;M_4;M_5;M_6;M_7\}\ {}_{k;1}\sqrt{\frac{\ln(d_z n)}{n}} \tag{B.15}$$

where $M_2;M_5;M_6$ *and* M_7 *are all constants that also do not depend on k described in Lemma B.7 and Lemmas B.9-B.11. In particular, so long as* $k=n\ !\ 0$, $\Pr(\Omega_{k;4}) = 1 \quad o(1)$.

Proof. Apply the same steps as the proof of Lemma B.7 with

$${}^2_{;n} = M\ {}_{;r}C_0^2\ {}^2_{k;1}s_k^2\bar{c}_n^2\ln^5(d_z n)=n$$
$${}_{;n} = \frac{4}{n}+\frac{4}{n^2}$$

□

Lemma B.9 (Probabilistic Bound on $\Omega_{k;5}$)**.** *Let* $\tilde{\Sigma}_{;j}$ *and* $\Sigma_{;j} = \mathsf{E}\tilde{\Sigma}_{;j}$ *be as in* (B.4)*. Under Assumptions 4.1 and 4.2 if*

$$\bar{}_k \quad 4\ {}_{k;1}(D_0^2+D_0D_1)C_0^2\left[D_0^2\log(d_z=\ )=n+D_0D_1\sqrt{\log(d_z=\ )=n}\right]$$

Then $\Pr(\Omega_{k;5}) \quad 1 \quad 2k^{\ 2}$. *In particular, there is a constant* M_5 *that does not depend on k, such that if* $\bar{}_k \quad {}_{k;1}M_5\sqrt{\log(d_z=\ )=n}$ *and* $k^{\ 2}\ !\ 0$ *as* $n\ !\ 1$ *then under the conditions of Assumption 4.2,* $\Pr(\Omega_{k;5}) = 1 \quad o(1)$.

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

$$p_j(X)e^{\hat{\beta}_j' Z} D_j Y - \bar{m}_j(Z) g^2 Z_l Z_h$$

is the product of $p_k(X)e^{\hat{\beta}_j' Z} Z_l Z_h$, which is bounded in absolute value by $\bar{\kappa}_{k,1} C_0^2 e^{B_0}$, and $D_j Y - \bar{m}_j(Z)$, which is uniformly sub-gaussian conditional on Z. By Lemma E.7 we have:

$$\mathsf{E}\left[\left(\hat{\Sigma}_{j,lh} - \Sigma_{j,lh}\right)^k\right] \le \frac{k!}{2}(2\bar{\kappa}_{k,1} C_0^2 e^{B_0} D_0^2)^{k-2}(2\bar{\kappa}_{k,1} C_0^2 e^{B_0} D_0 D_1)^2, \quad k = 2,3,\dots$$

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

$$\Pr\left[\left|\hat{\Sigma}_{j,lh} - \Sigma_{j,lh}\right| > 2e^{B_0}\bar{\kappa}_{k,1} C_0^2 D_0^2 t + 2e^{B_0}\bar{\kappa}_{k,1} C_0^2 D_0 D_1 \sqrt{2t}\right] \le 2\delta^2/d_z^2.$$

A union bound completes the argument. □

Lemma B.10 (Probabilistic Bound on $\Omega_{k,6}$)**.** *Let $\hat{\Sigma}_{j,lh}$ and $\Sigma_{j,lh} = \mathsf{E}\hat{\Sigma}_{j,lh}$ be as in (B.4). Under Assumptions 4.1 and 4.2 if*

$$\bar{\kappa}_k \le \bar{\kappa}_{k,1}\sqrt{2}(e^{B_0} + 1)C_0\sqrt{\log(d_z^2/\delta^2)/n},$$

then $\Pr(\Omega_{k,6}) \ge 1 - 2k\delta^2$. In particular, there is a constant M_6 that does not depend on k, such that if $\bar{\kappa}_k \le \bar{\kappa}_{k,1} M_6\sqrt{\log(d_z^2/\delta^2)/n}$ and $k\delta^2 \to 0$ as $n \to \infty$ then under the conditions of Assumption 4.2, $\Pr(\Omega_{k,6}) = 1 - o(1)$.

Proof. Consider each j separately. For any $l,h = 1,\dots,d_z$, note $|\hat{\Sigma}_{j,lh}| = |p_j(X)De^{\hat{\beta}_j' Z} Z_l Z_h| \le \bar{\kappa}_{k,1} C_0^2 e^{B_0}$ so that $(\hat{\Sigma}_{j,lh} - \Sigma_{j,lh})$ is mean zero and bounded in abosulte values by $2\bar{\kappa}_{k,1} C_0^2 e^{B_0}$. Applying Lemma E.3 with $\bar{\kappa}_k \ge 4\bar{\kappa}_{k,1} C_0^2 e^{B_0}\sqrt{\log(d_z^2/\delta^2)/n}$ yields:

$$\Pr\left[\left|\hat{\Sigma}_{j,lh} - \Sigma_{j,lh}\right| \ge \bar{\kappa}_k\right] \le 2\delta^2/d_z^2.$$

A union bound completes the argument. □

Lemma B.11 (Probabilitstic Bound on $\Omega_{k,7}$)**.** *Let $\hat{\Sigma}^1_{j,l}$ and $\Sigma^1_{j,l} = \mathsf{E}\hat{\Sigma}^1_{j,l}$ be as in (B.7). Under Assumptions 4.1 and 4.2 if*

$$\bar{\kappa}_k \le \bar{\kappa}_{k,1} 4(D_0^2 + D_1^2)^{1/2} e^{B_0} C_0^2\sqrt{\log(d_z^2/\delta^2)/n},$$

then $\Pr(\Omega_{k,7}) \ge 1 - 2k\delta^2$. In particular, there is a constant M_7 that does not depend on k such that if $\bar{\kappa}_k \le \bar{\kappa}_{k,1} M_7\sqrt{\log(d_z^2/\delta^2)/n}$ and $k\delta^2 \to 0$ as $n \to \infty$ then, under the conditions of Assumptions 4.1 and 4.2, $\Pr(\Omega_{k,7}) \ge 1 - o(1)$.

Proof. We deal with each j term separately. The variables $p_j(X)e^{\hat{\beta}_j' Z} Y - \bar{m}_j(Z)$ are uniformly sub-gaussian conditional on Z because $|p_j(X)e^{\hat{\beta}_j' Z} Z_l Z_h| \le \bar{\kappa}_{k,1} e^{B_0} C_0^2$ and $D_j Y - \bar{m}_j(Z)$ is uniformly sub-gaussian. Applying Lemma E.4 for $\bar{\kappa}_k \ge e^{B_0}\bar{\kappa}_{k,1} C_0^2\sqrt{8(D_0^2 + D_1)^2}\sqrt{\log(d_z^2/\delta^2)/n}$ yields

$$\Pr\left[\left|\hat{\Sigma}_{j,lh} - \Sigma_{j,lh}\right| \ge \bar{\kappa}_k\right] \le 2\delta^2/d_z^2.$$

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.12 (Deterministic Logistic Score Domination)**.** *Under Assumption 4.1 let*

$$\bar{}_k \quad {}_{k;1} \frac{p}{2}(e^{B_0}+1)C_0 \sqrt{\ln(d_z=)=n}:$$

Then if for all $j = 1;:::;k$ *we let* $_{;j} \quad \bar{}_k$, $\Pr(\Omega_{k;1} \setminus \Omega_{k;2}) \quad 1 \quad 2k$ *. In particular, there is a constant* M_8^p *that does not depend on k such that if* $\bar{}_k \quad M_8^p {}_{k;1} \sqrt{\ln(d_z n)=n}$ $\Pr(\Omega_{k;1} \setminus \Omega_{k;2}) \quad 1 \quad 2k=n^p$.

Proof. Let us recall that

$$kS_j k_1 = \max_{1 \quad l \quad d_z} jE_n[p_j(X)f \quad De^{\ {}^0_j Z} + (1 \quad D)gZ_j]j:$$

Notice for each $1 \quad l \quad d_z$, $S_{j;l} = p_j(X)f \quad De^{\ {}^0_j Z} + (1 \quad D)gZ_l$ is bounded in absolute value by $C_0 \ {}_{k;1}(e^{B_0}+1)$ and is mean zero by optimality of $\bar{}_j$. For $\bar{}_k \quad 2(e^{B_0}+1)C_0 \sqrt{\ln(d_z=)=n}$ apply Lemma E.3 to see the result. □

Lemma B.13 (Deterministic Linear Score Domination)**.** *Under Assumptions 4.1 and 4.2 let*

$$\bar{}_k \quad {}_{k;1}(e^{B_0}C_0) \sqrt{8(D_0^2+D_1^2)} \sqrt{\ln(d_z=)=n}:$$

Then if for all $j = 1;:::;k$ *we let* $_{;j} \quad \bar{}_k$, $\Pr(\Omega_{k;3} \setminus \Omega_{k;4}) \quad 1 \quad 2k$ *. In particular, there is a constant* M_9^p *that does not depend on k such that if* $\bar{}_k \quad M_9^p {}_{k;1} \sqrt{\ln(d_z n)=n}$, $\Pr(\Omega_{k;3} \setminus \Omega_{k;4}) \quad 1 \quad 2k=n^p$.

Proof. Notice $S_{j;l} = p_j(X)De^{\ {}^0_j Z} fY \quad \bar{m}_j(Z)gZ_l$ for $l = 1;:::;p$. By optimality of $\bar{}_j$, $S_{j;l}$ is mean zero. Under Assumption 4.1, $jS_{j;l}j \quad e^{B_0}C_0 jD fY \quad \bar{m}_j(Z)gj$ so by Assumption 4.2 the variables $S_{j;l}$ are uniformly sub-gaussian conditional on Z in the following sense:

$$\max_{l=1;:::;p} \tilde{D}_0^2 E[\exp(S_{j;l}^2=\tilde{D}_0^2) \quad 1] \quad \tilde{D}_1^2$$

for $\tilde{D}_0 = {}_{k;1} C_0 D_0 e^{B_0}$ and $\tilde{D}_1 = {}_{k;1} C_0 D_1 e^{B_0}$. Apply Lemma E.4 for $\bar{}_k$ defined above in the statement of Lemma B.13 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 5.1 hold. In addition suppose that* $\frac{2}{k}\log k=n \ ! \ 0$ *and* $c_k \ ! \ 0$*. Then if either the propensity score our outcome regression model are correctly speci ed:*

$$k\hat{g}_k \quad g_0 k_{L;2} = (E[(\hat{g}(x) \quad g_0(x))^2])^{1=2} \ . \ _p \sqrt{k=n} + c_k \tag{C.1}$$

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

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 $Q = \mathsf{E}[\mathsf{E}_n Q_i]$ and $\|Q_i\| \leq M$ a.s. Then for $\hat{Q} = \mathsf{E}_n[Q_i]$,*

$$\Delta := \mathsf{E}\|\hat{Q} - Q\| \leq \frac{M \log k}{n} + \sqrt{\frac{M \|Q\| \log k}{n}}.$$

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

$$\Delta := \mathsf{E}\|\hat{Q} - Q\| \leq \frac{\zeta_k^2 \log k}{n} + \sqrt{\frac{\zeta_k^2 \|Q\| \log k}{n}}.$$

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

$$\begin{aligned} \|\hat{\beta}_k - \beta_k\| \leq g_0 \|L_{:2}\| \, & \|\hat{p}^k(x)' \hat{\beta}^k - p^k(x)' \beta^k\| \leq \|p^k(x)' - \hat{p}^k(x)'\| \|\beta^k L_{:2}\| + \|\hat{p}^k(x)'\| \|\beta^k - \hat{\beta}^k\| \|L_{:2}\| \\ & \|\hat{p}^k(x)' \hat{\beta}^k - p^k(x)' \beta^k\| \leq \|p^k(x)' - \hat{p}^k(x)'\| \|\beta^k L_{:2}\| + c_k \end{aligned}$$

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

$$\|p^k(x)'\| \leq \|p^k(x)\|, \quad \|L_{:2}\| = \|\beta_{:2}\| \leq \|\beta\|$$

Further,

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

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

$$\leq \|\mathsf{E}[p^k(x)(y - \bar{Y})]\| + \|\mathsf{E}_n[p^k(x)\beta^k]\| + \|\mathsf{E}_n[p^k(x) r_k]\|$$

Under Condition 1 the first term is $o_p(1/\sqrt{k/n})$. 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 5.1 to see

$$\mathsf{E}\|\mathsf{E}_n[p^k(x)\beta^k]\|^2 = \mathsf{E}\sum_{j=1}^{k} \zeta_j^2 p_j(x)^2/n \beta^{-2} \mathsf{E}_n[p^k(x)p^k(x)'/n] \beta \leq \mathsf{E}[p^k(x)p^k(x)'/n] = k/n.$$

This gives $\|\mathsf{E}_n[p^k(x)\beta^k]\| \to_p \sqrt{k/n}$ and thus shows (C.1). □

The following lemma is a building block for asymptotic pointwise normality. It establishes conditions under which the coefficient estimator $\hat{\beta}$ is asymptotically linear in the sense of Bickel et al. (1993).

Lemma C.2 (Pointwise Linearization). *Suppose that Condition 1 and Assumption 5.1, hold. In addition assume that $\zeta_k^2 \log k / n \to 0$. Then for any $\gamma \in S^{k-1}$,*

$$\sqrt{n}\, \gamma'(\hat{\beta}^k - \beta^k) = \gamma' \mathsf{G}_n[p^k(x)(\beta^k + r_k)] + R_{1n}(\gamma) \tag{C.2}$$

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

$$R_{1n}(\ ) \ .\ _p \ \sqrt{\frac{\ ^2_k \log k}{n}}(1 + \sqrt{\ _k c_k}) \tag{C.3}$$

Moreover,

$$\sqrt{n}\ ^0(\hat{b}^k \quad \ ^k) = \ ^0G_n[p^k(x) \quad \ ^k] + R_{1n}(\ ) + R_{2n}(\ ) \tag{C.4}$$

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

$$R_{2n}(\ ) \ .\ _p \ \sqrt{\ _k c_k} \tag{C.5}$$

Proof. Decompose as before,

$$\begin{aligned}\sqrt{n}\ ^0(\hat{b}^k \quad \ ^k) = &\sqrt{n}\ ^0[\hat{Q}^{\ 1}\mathbb{E}_n[p^k(x) \quad (\hat{Y} \quad \bar{Y})]\\ &+ \ ^0G_n[p^k(x) \quad (\ ^k + r_k)]\\ &+ \ ^0[\hat{Q}^{\ 1} \quad I]G_n[p^k(x) \quad (\ ^k + r_k)]:\end{aligned}$$

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

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

$$^0[\hat{Q}^{\ 1} \quad I]G_n[p^k(x) \quad \ ^k]:$$

has mean zero and variance bounded by $\ ^{-2}\ ^0[\hat{Q}^{\ 1} \quad I]\hat{Q}^{\ 1}[\hat{Q}^{\ 1} \quad I]$. Next, by Lemma C.1, with probability approaching one, all eigenvalues of $\hat{Q}^{\ 1}$ are bounded from above and away zero. So,

$$\ ^{-2}\ ^0[\hat{Q}^{\ 1} \quad I_k]\hat{Q}^{\ 1}[\hat{Q}^{\ 1} \quad I_k] \ . \ \ ^{-2}k\hat{Q}kk\hat{Q}^{\ 1}k^2k\hat{Q}^{\ 1} \quad I_k k^2 \ .\ _p \ \frac{\ ^2_k \log k}{n}:$$

so by Chebyshev's inequality,

$$^0[\hat{Q}^{\ 1} \quad I]G_n[p^k(x) \quad \ ^k] \ .\ _p \ \sqrt{\frac{\ ^2_k \log k}{n}}:$$

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

$$^0(\hat{Q}^{\ 1} \quad I_k)G_n[p^k(x)r_k] \ .\ _p \ \sqrt{\frac{\ ^2_k \log k}{n}}\ _k c_k \sqrt{\ _k}$$

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

$$R_{2n}(\ ) = \ ^0G_n[p^k(x)r_k] \ .\ _p \ \sqrt{\ _k c_k};$$

which gives (C.4)-(C.5). □

The following lemma shows that, after adding Assumption 5.3 the linearization of our coefficient estimator $\hat{b}^k$ established in Lemma C.2 holds uniformly over all points $x \ 2 \ X$. 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 5.2 and 5.3.

Lemma C.3 (Uniform Linearization). *Suppose that Condition 1 and Assumption 5.1-5.3 hold. Then if either the propensity score model our outcome regression model is correctly specified:*

$$\sqrt{n}\,p(x)'(\hat{b}_k - b_k) = p(x)'\mathbb{G}_n[p^k(x)(\varepsilon^k + r_k)] + R_{1n}(\mathcal{X}(x)) \tag{C.6}$$

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

$$R_{1n}(\mathcal{X}(x)) \cdot_p \sqrt{\frac{k\log k}{n}}\left(n^{1/m}\sqrt{\log k} + \sqrt{k}\bar{\iota}_k c_k\right) := \bar{R}_{1n} \tag{C.7}$$

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

$$\sqrt{n}\,p(x)'(\hat{b}_k - b_k) = p(x)'\mathbb{G}_n[p^k(x)\varepsilon^k] + R_{1n}(\mathcal{X}(x)) + R_{2n}(\mathcal{X}(x)) \tag{C.8}$$

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

$$R_{2n}(\mathcal{X}(x)) \cdot_p \sqrt{\log k}\,\bar{\iota}_k c_k := \bar{R}_{2n} \tag{C.9}$$

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

$$\begin{aligned}\sqrt{n}\,p(x)'(\hat{b}_k - b_k) = {} & \sqrt{n}\,p(x)'\hat{Q}^{-1}\mathbb{E}_n[p^k(x)(\hat{Y} - \bar{Y})] \\ & + p(x)'\mathbb{G}_n[p^k(x)(\varepsilon^k + r_k)] \\ & + p(x)'[\hat{Q}^{-1} - I]\mathbb{G}_n[p^k(x)(\varepsilon^k + r_k)].\end{aligned} \tag{C.10}$$

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

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

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

Step 1: Conditional on the data let

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

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 $(\mathcal{T}, \|\cdot\|_{n,2})$ is a subset $\mathcal{T}_\varepsilon$ of $\mathcal{T}$ such that for every $t \in \mathcal{T}$ there is a point $t_\varepsilon \in \mathcal{T}_\varepsilon$ such that $\|t - t_\varepsilon\|_{n,2} < \varepsilon$. The covering number $\mathcal{N}(\mathcal{T}, \|\cdot\|_{n,2}, \varepsilon)$ of $\mathcal{T}$ is the infimum of the cardinality of ε-nets of $\mathcal{T}$.

Let $\gamma_1, \ldots, \gamma_n$ be independent Rademacher random variables that are independent of the data. Let $\gamma = (\gamma_1, \ldots, \gamma_n)$. Let $\mathbb{E}_\gamma[\cdot]$ denote the expectation with respect to the distribution of γ. By Dudley's inequality

(Dudley, 1967),

$$\mathsf{E}\left[\sup_{x\in X}\left|(x)^{0}[\hat{Q}^{-1}\quad I]G_n[\ {}_i p^k(x)\quad {}^k]\right|\right]\ .\ \int_0^{\infty}\sqrt{\log N(T;k\quad k_{n;2};\varepsilon)}\,d\varepsilon:$$

where

$$\begin{aligned}&:= 2\sup_{t\in T}\|t\|_{n;2}\\ &= 2\sup_{x\in X}\ \mathsf{E}_n[(\ (x)^{0}(\hat{Q}^{-1}\quad I)p^k(x)\quad {}^k)^2]^{1=2}\\ &\quad 2\max_{1\ i\ n}|j^{-}{}_{k;i}\|k\hat{Q}^{-1}\quad I\|k\hat{Q}\|^{1=2};\end{aligned}$$

by (A.2). Now, for any $x\ 2\ X$,

$$\begin{aligned}&\mathsf{E}_n[(\ (x)^{0}(\hat{Q}^{-1}\quad I)p^k(x)\quad {}^k\quad(\tilde{x})^{0}(\hat{Q}^{-1}\quad I)p^k(x)\quad {}^k)^2]^{1=2}\\ &\max_{1\ i\ n}|j^{-}{}_{k;i}\|k\ \ (x)\quad(\tilde{x})\|k\hat{Q}^{-1}\quad I\|k\hat{Q}\|^{1=2}\\ &\tfrac{L}{k}\max_{1\ i\ n}|j^{-}{}_{k;i}\|k\hat{Q}^{-1}\quad I\|k\hat{Q}\|^{1=2}\|x\quad\tilde{x}\|\end{aligned}$$

So, for some $C > 0$,

$$N(T;k\quad k_{n;2};\varepsilon)\qquad\frac{C\ \frac{L}{k}\max_{1\ i\ n}|j^{-}{}_{k;i}\|k\hat{Q}^{-1}\quad I\|k\hat{Q}\|^{1=2}}{\varepsilon}\Bigg|^{d_x}:$$

This gives us that

$$\int_0^{\infty}\sqrt{\log(N(T;k\quad k_{2;n};\varepsilon))}\,d\varepsilon\quad\max_{1\ i\ n}|j^{-}{}_{k;i}\|k\hat{Q}^{-1}\quad I\|k\hat{Q}\|^{1=2}\int_0^{2}\sqrt{d_x\log(C\ \tfrac{L}{k}=\varepsilon)}\,d\varepsilon:$$

By Assumption 5.3 we have that $\mathsf{E}[\max_{1\ i\ n}|j^{-}{}_{k;i}|j\ j\ X] \ .\ P\ n^{1=m}$ where $X = (x_1;:::;x_n)$. In addition $\sum_k^{2m=(m\ \ 2)}\log k=n\ .\ 1$ for $m > 2$ gives that $\frac{2}{k}=\log k=n\ !\ 0$. So, $\|k\hat{Q}^{-1}\quad I\|k\ ._P\ (\ \frac{2}{k}\log k=n)^{1=2}$ and $\|k\hat{Q}^{-1}\|k\ ._P\ 1$. Combining this all with $\log\ \frac{L}{k}\ .\ \log k$ implies

$$\begin{aligned}&\mathsf{E}\left[\sup_{x\in X}\left|(x)^{0}[\hat{Q}^{-1}\quad I]G_n[p^k(x)\quad {}^k]\right|\ |\ X\right]\quad 2\mathsf{E}\left[\mathsf{E}\left[\sup_{x\in X}\left|(x)^{0}[\hat{Q}^{-1}\quad I]G_n[\ {}_i p^k(x)\quad {}^k]\right|\ |\ X\right]\right]\\ &\ ._P\ n^{1=m}\sqrt{\frac{\frac{2}{k}\log^2 k}{n}}\end{aligned}$$

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

$$\sup_{x\in X}\left|(x)^{0}[\hat{Q}^{-1}\quad I]G_n[p^k(x)\quad {}^k]\right|\ ._p\ n^{1=m}\sqrt{\frac{\frac{2}{k}\log^2 k}{n}}\tag{C.11}$$

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 $Y(\bar{}_k; \bar{m}_k)$, there is no Hadamard product to deal with.

$$\sup_{x2X} (x)^0[\text{Q}^{\ 1} \quad I]\mathsf{G}_n[p^k(x)r_k] \quad .\ _P \quad \sqrt{\frac{^2_k \log k}{n}}\ {}_k c_k \ \sqrt{p}\ \bar{k} \tag{C.12}$$

$$\sup_{x2X} (x)^0 \mathsf{G}_n[p^k(x)r_k] \quad .\ _P \ `_k c_k \ \sqrt{p \log k} \tag{C.13}$$

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

Theorem C.2 (Validity of Gaussian Bootstrap). *Suppose that the assumptions of Theorem 5.3 hold with $a_n = \log n$ and the assumptions of Theorem 5.4 hold with $a_n = O(n^{\ b})$ for some $b > 0$. In addition, suppose that there exists a sequence ^0_n obeying $1 \ . \ ^0_n \ . \ kp^k(x)k$ uniformly for all $x \ 2 \ X$ such that $kp^k(x) \quad p^k(x^0)k = ^0_n \quad L_n kx \quad x^0k$, where $\log L_n \ . \ \log n$. Let N^b_k be a bootstrap draw from $N(0; I_k)$ and $P^?$ be the distribution conditional on the observed data $fY_i; D_i; Z_i g^n_{i=1}$. Then the following approximation holds uniformly in $`^{\ 1}(X)$:*

$$\frac{p^k(x)^0 \text{Q}^{1=2}}{\text{Q}^{1=2} p^k(x)} N^b_k =^d \frac{p^k(x)^0 \Omega^{1=2}}{k\Omega^{1=2} p^k(x)k} + o_{P^?}(\log^{\ 1} N) \tag{C.14}$$

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

D 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 (4.2) directly, after choosing constants $c_{;j}$ and $c_{;j}$ from a (finite) set via cross validation. For each j we will assume that

$$c_{;j}; c_{;j} \ 2 \ \Lambda_n \quad [\underline{c}_n; \bar{c}_n] \tag{D.1}$$

where $j\Lambda_n j$ can be fairly large (on the order of $n^2=k$).

D.1 Theory Overview

Let $M_5; M_6; M_7; M^2_8; M^2_9$ be constants that do not depend on k as in Lemmas B.9–B.13. Whenever

$$\underline{c}_n \sqrt{\frac{\ln^3(d_z n)}{n}} \quad {}_{k;\,1} \max \left\{ M_5; M_6; M_7; M^2_8; M^2_9 \right\} \sqrt{\frac{\ln(d_z n)}{n}}. \tag{D.2}$$

we will have that, under Assumption 4.1(i)-(iv) and Assumption 4.2(i)-(iii) the event $\bigcap_{k=1}^{7} \Omega_{k;7}$ occurs with probability at least $1 \quad 10k=n^2$ for the $2k$ pilot penalty parameters chosen with any values $c_{;j}; c_{;j} \ 2 \ \Lambda_n$ and

$$\bar{}_k := \bar{c}_n \sqrt{\frac{\ln^3(d_z n)}{n}}.$$

In this event, apply Lemmas B.1 and B.2 to obtain the following finte sample bounds for the parameter estimates

$$\max_{1 \le j \le k} D^z_{;j}(\mathsf{b}_j; \bar{}_j) \quad M_0 \frac{s_k \bar{c}_n^2 \ln^3(d_z n)}{n} \quad \text{and} \quad \max_{1 \le j \le k} k\mathsf{b}_j \quad \bar{}_j k_1 \quad M_0 s_k \bar{c}_n \sqrt{\frac{\ln^3(d_z n)}{n}}$$

$$\max_{1 \le j \le k} D^z_{;j}(\mathsf{b}_j; \bar{}_j; \bar{}_j) \quad M_1 \frac{s_k \bar{c}_n^2 \ln^3(d_z n)}{n} \quad \text{and} \quad \max_{1 \le j \le l} k\mathsf{b}_j \quad \bar{}_j k_1 \quad M_1 s_k \bar{c}_n \sqrt{\frac{\ln^3(d_z n)}{n}}$$

and Lemma B.3 to obtain the following finite sample bound for the weighted means:

$$\max_{1 \le j \le k} j\mathsf{E}_n[p_j(X)(Y(\mathsf{b}_j; \hat{m}_j) \quad Y(\bar{}_j; \bar{m}_j))]j \quad M_2 \frac{\bar{c}_n^2 s_k \ln^3(d_z n)}{n} \tag{D.3}$$

$$\max_{1 \le j \le k} j\mathsf{E}_n[p_j^2(X)(Y(\mathsf{b}_j; \hat{m}_j) \quad Y(\bar{}_j; \bar{m}_j))^2] \quad M_3 \frac{\sum_{k;1}^2 \bar{c}_n^2 s_k^2 \ln^3(d_z n)}{n} \tag{D.4}$$

Combining (D.2) and (D.3) we can see that Condition 1 can be obtained under Assumption 4.1(i)-(iv), Assumption 4.2(i)-(iii), and the following modified sparsity bounds

$$\frac{kj\Lambda_n j}{n^2} \; ! \; 0; \; \frac{\underline{c}_n^{\,1} \; {}_{k;1}}{\ln(d_z n)} \; ! \; 0 \quad \text{and} \quad \frac{\bar{c}_n^2 s_k k^{1=2} \ln^3(d_z n)}{p\bar{n}} \; ! \; 0: \tag{D.5}$$

Simlarly combining (D.2) and (D.4), Condition 2 can additionally be obtained by strengthening the rates in (D.5) to include

$$\frac{\sum_{k;1}^2 \bar{c}_n^2 s_k k^2 \ln^3(d_z n)}{n^{(m \; 1)=m}} \; ! \; 0 \tag{D.6}$$

for $m > 2$ as in Assumption 5.3. These rates are comparable and in certain cases may be more palatable than those presented in the main text, Assumptions 4.1 and 4.2. They come at the cost of slower rates of convergence for the weighted means as seen by comparing eqs. (D.3)–(D.4) to eqs. (4.9) and (4.12).

D.2 Practical Implementation

In practice, the constants $M_5; M_6; M_7; M_8^2; M_9^2$ from Lemmas B.9–B.13 are roughly on the order of kZk_1. We therefore reccomend setting

$$\underline{c}_n = \frac{1}{2\log^{1=2}(d_z n)} \max_{1 \le i \le n} kp^k(X_i)k_1 \max_{1 \le i \le n} kZ_i k_1$$

$$\bar{c}_n = \frac{3\log^{1=2}(d_z n)}{2} \max_{1 \le i \le n} kp^k(X_i)k_1 \max_{1 \le i \le n} kZ_i k_1$$

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 ;c and designate a fold to hold out.
3. Estimate nuisance model parameters using $\hat{}^{\text{pilot}}_{;j}$ and $\hat{}^{\text{pilot}}_{;j}$ 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 $c_{;j}$ and $c_{;j}$ with the lowest average loss.

In practice we find this procedure works well with small K, around $K = 5$ and with $j\Lambda_n j$ on the order of about 10-20.

E High Dimensional Probability Results

E.1 High Dimensional Central Limit and Bootstrap Theorems

Lemma E.1 (Gaussian Quantile Bound). *Let $Y = (Y_1; : : : ; Y_p)$ be centered Gaussian in $\mathbb{R}^p$ with ^{2} $\max_{1 \ j \ p} \mathbb{E}[Y_j^2]$ and $p \overline{2}$ $\mathrm{p}\ 2$. Let $q^Y(1 \quad)$ denote the $(1 \quad)$-quantile of $kY k_1$ for $2 (0; 1)$. Then $q^Y(1 \quad) \quad (2 + p \overline{2})\ \mathrm{p} \overline{\ln(p=)}$.*

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

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

$$S_n := p\frac{1}{n} \sum_{i=1}^{n} Z_i \quad \text{and} \quad \Sigma := \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[Z_i Z_i^{\prime}]:$$

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

$$(U; V) := \sup_{A 2 A_p} \Pr(U \ 2 \ A) \quad \Pr(V \ 2 \ A)$$

where A_p denotes the collection of all hyperrectangles in $\mathbb{R}^p$. For any symmetric positive matrix $M \ 2 \ \mathbb{R}^{p \ p}$, write $N_M := N(\mathbf{0}; M)$.

Theorem E.1 (High-Dimensional CLT). *If, for some nite constants $b > 0$ and $B_n \quad 1$,*

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[Z_{ij}^2] \quad b; \ \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[jZ_{ij}j^{2+k}] \quad B_n^k \ \text{and} \ \mathbb{E}\left[\max_{1 \ j \ p} Z_{ij}^4\right] \quad B_n^4: \tag{E.1}$$

for all i 2 $f1; : : : ; ng$; j 2 $f1; : : : ; pg$ and k 2 $f1; 2g$, then there exists a nite constant C_b, depending only on b, such that:

$$(S_n; N \ ) \quad C_b \left(\frac{B_n^4 \ln^7(pn)}{n}\right)^{1=6} :$$

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

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

$$\tilde{}(S_n^e; N \ ) := \sup_{A 2 A_p} \Pr_e \left[S_n^e \ 2 \ A\right] \quad \Pr(N \quad 2 \ A) :$$

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

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

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

$$\tilde{\rho}(\hat{S}_n^{be}, N) \leq C_b \max\left\{\varepsilon_n, \left(\frac{B_n \ln^6(pn)}{n}\right)^{1/6}\right\}.$$

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 $\hat{Z}$ into k subvectors.

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

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 $q_{M;j}^N : \mathbb{R} \to \bar{\mathbb{R}}$ as the (extended) quantile function of $\|N_{M;j}\|_1$,

$$q_{M;j}^N(\alpha) := \inf\{t \in \mathbb{R} : \Pr(\|N_{M;j}\|_1 \leq t) \geq \alpha\}.$$

Define $q_{M;j}^N(\alpha) = +\infty$ if $\alpha > 1$ and $= -\infty$ if $\alpha < 0$ so that $q_{M;j}^N$ is always montone (strictly) increasing.

Lemma E.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* $j = 1, \ldots, k$ *let* q_j *denote the quantile function of* $\|U_j\|_1$*. Then for any* $j = 1, \ldots, k$*,*

$$q_{M;j}^N(\alpha - 2\Delta(U; N_M)) - q_j(\alpha) \leq q_{M;j}^N(\alpha + \Delta(U; N_M)) \text{ for all } \alpha \in (0, 1).$$

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 $\|N_M\|_1$ 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.

$$\begin{aligned}
\Pr(\|U_1\|_1 \leq t) &= \Pr(\|N_{M;1}\|_1 \leq t) + \Pr(\|U_1\|_1 \leq t) - \Pr(\|N_{M;1}\|_1 \leq t) \\
&= \Pr(\|N_{M;1}\|_1 \leq t) + \Pr(U \in [-t, t]^p \times \mathbb{R}^{p-d_1}) - \Pr(N_M \in [-t, t]^p \times \mathbb{R}^{p-d_1}) \\
&\leq \Pr(\|N_{M;1}\|_1 \leq t) + \Delta(U; N_M)
\end{aligned}$$

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

$$\Pr(\|U_1\|_1 \leq t) \geq \Pr(\|N_{M;1}\|_1 \leq t) - \Delta(U; N_M).$$

Substituting $t = q^N_{M;1}(\quad 2\ (U;N_M))$ into the upper bound on $\Pr(kU_1k_1 \quad t)$ gives the lower bound statement, while $t = q^N_{M;1}(\ +\ (U;N_M))$ and using the lower bound on $\Pr(kU_1k_1 \quad t)$ gives the upper bound statement. □

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

$$S_n = (S^0_{n;1};:::;S^0_{n;k})^0 \ 2\ \mathsf{R}^{d_1;:::;d_k} \quad \text{and} \quad \tilde{S}^e_n = (\tilde{S}^{e^0}_{n;1};:::;\tilde{S}^{e^0}_{n;k})^0 \ 2\ \mathsf{R}^{d_1;:::;d_k}:$$

For each $j = 1;:::;k$ define $q_{n;j}(\ )$ as the -quantile of $kS_{n;j}k_1$

$$q_{n;j}(\ ) := \inf\{t\ 2\ \mathsf{R} : \Pr(kS_{n;j}k_1 \quad t) \quad g \quad \text{for} \quad 2\ (0;1):$$

Let $\tilde{q}_{n;j}(\ )$ be the -quantile of $k\tilde{S}^e_{n;j}k_1$, computed conditionally on X_i and $\tilde{X}_i$'s,

$$\tilde{q}_{n;j}(\ ) := \inf\{t\ 2\ \mathsf{R} : \Pr_e(k\tilde{S}^e_{n;j}k_1 \quad t) \quad g \ \text{for} \quad 2\ (0;1):$$

Theorem E.3 (Quantile Comparasion)**.** *If* (E.1) *holds for some nite constants* $b > 0$ *and* $B_n \quad 1$*, and*

$$_n := 2C_b \quad \frac{B^4_n \ln^7(pn)}{n}^{!\ 1=6}$$

denotes the upper bound in Theorem E.1 multiplied by two, then for all $j = 1;:::;k$

$$q^N_{;j}(1 \qquad _n) \quad q_{n;j}(1 \quad) \quad q^N_{;j}(1 \quad + \ _n) \ \textit{for all} \quad 2\ (0;1):$$

If, in addition, (E.2) *holds for some sequences* $\{\ _n\}_{\mathsf{N}}$ *and* $\{\ _n\}_{\mathsf{N}}$ *converging to zero, and*

$$^0_n \quad 2C^0_b \max\left\{ ; \ \frac{B^4_n \ln^6(pn)}{n}^{!\ 1=6} \right\};$$

denotes the upper bound in Theorem E.2 multiplied by two, then with probability at least $1 \quad _n \quad 1=\ln^2(pn)$ *we have for all* $j = 1;:::;k$*,*

$$q^N_{;j}(1 \qquad ^0_n) \quad \tilde{q}_{n;j}(1 \quad) \quad q^N_{;j}(1 \quad + \ ^0_n) \ \textit{for all} \quad 2\ (0;1):$$

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

$$q^N_{;j}(1 \qquad 2\ (S_n;N\ )) \quad q_{n;j}(1 \quad) \quad q^N_{;j}(1 \quad + \ (S_n;N\ )):$$

The first chain of inequalities then follows from $2\ (S_n;N\ ) \quad _n$ by Theorem E.1.

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

$$q^N_{;j}(1 \qquad 2\tilde{}(\tilde{S}^e_n;N\ )) \quad \tilde{q}_n(1 \quad) \quad q^N_{;j}(1 \quad + \tilde{}(\tilde{S}^e_n;N\ )):$$

The second chain of inequalities then follows on the event $2\tilde{}(\tilde{S}^e_n;N\ ) \quad ^0_n$, which by Theorem E.2 happens

with probability at least 1 $\quad_n \quad 1=\ln^2(pn)$. □

Theorem E.4 (Multiplier Bootstrap Consistency)**.** *Let* (E.1) *and* (E.2) *hold for some constants* $b > 0$ *and* $B_n \quad 1$ *and some sequences* $f\ _n g_N$ *and* $f\ _n g_N$ *in* R_{++} *converging to zero. Then, there exists a nite constant* C_b*, depending only on* b*, such that*

$$\max_{1\ j\ k}\ \sup_{2(0;1)} \Pr(kS_{n;j}k_1 \quad \phi_{n;j}(1 \quad)) \qquad C_b \max\left\{ {}_n;\ {}_n;\ \frac{B_n^4 \ln^7(pn)}{n}\right)^{1=6};\ \frac{1}{\ln^2(pn)}\right\}:$$

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

$$\begin{aligned}\Pr(kS_{n;j}k_1 \quad \phi_{n;j}(1 \quad)) \quad & \Pr(kS_{n;j}k_1 \quad q^N_{;j}(1 \quad + \ {}^0_n)) + \ {}_n + \frac{1}{\ln^2(pn)} \\ & \Pr(kN\ _{;j}k_1 \quad q^N_{;j}(1 \quad + \ {}^0_n)) + \ {}_n + \ {}_n + \frac{1}{\ln^2(pn)} \\ & 1 \quad + \ {}^0_n + \ {}_n + \ {}_n + \frac{1}{\ln^2(pn)}\end{aligned}$$

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

$$\Pr(kS_{n;j}k_1 \quad \phi_{n;j}(1 \quad)) \quad 1 \qquad {}^0_n + \ {}_n + \ {}_n + \frac{1}{\ln^2(pn)}\ :$$

Combining these two inequalities gives the result. □

E.2 Concentration and Tail Bounds

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

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

$$\Pr\left(\frac{1}{n}\sum_{i=1} Y_i > t\right) \qquad 2\exp\left(\frac{nt^2}{2c_0^2}\right):$$

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

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

Lemma E.5. *Let* $(Y_1;:::;Y_n)$ *be independent variables such that* $E[Y_i] = 0$ *for* $i = 1;:::;n$ *and*

$$\frac{1}{n}\sum_{i=1} E[jY_ij^k] \quad \frac{k!}{2}c_3^{k\ 2}c_4^2;\quad k = 2;3;:::;$$

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

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

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

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

Lemma E.7. *Suppose that* X *is bounded,* $|X| \leq c_0$, *and* Y *is sub-gaussian,* $c_2^2 \mathrm{E}[\exp(Y^2/c_1^2) - 1] \leq c_2^2$ *for some constants* $(c_1; c_2)$. *Then* $Z = XY^2$ *satisfies*

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

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

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

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

Proof. Using the fact that $e^{x^2} \geq |x|$ gives

$$\begin{aligned} c_0^2 \mathrm{E}[\exp(Y^2/c_0^2) - 1] \leq c_1^2 &\Longrightarrow \mathrm{E}[\exp(Y^2/c_0^2)] \leq c_1^2/c_0^2 + 1 \\ &\Longrightarrow \mathrm{E}[|Y|/c_0] \leq c_1^2/c_0^2 + 1 \\ &\Longrightarrow \mathrm{E}[|Y|] \leq c_1^2/c_0 + c_0 \end{aligned}$$

□

F 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 11 interaction/higher order variables that we believe may be useful in controlling for confounding. Table F.1 provides a final summary of the data after our cleaning process. The generated variables represent the bottom 11 variables in Table F.1.[1]

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.

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

Table F.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
mbsmoke ∗alcohol	4,107	0.017	0.128	0	1
medu ∗fedu	4,107	161.518	64.291	0	289
mage ∗fage	4,107	730.422	328.522	0	2,088
msmoke^2	4,107	0.944	2.422	0	9
msmoke ∗alcohol	4,107	0.037	0.302	0	3
mage^2	4,107	707.741	262.383	256	1,296
mage ∗mmarried	4,107	19.570	13.090	0	36
mage ∗medu	4,107	336.911	108.588	0	612
mage ∗fedu	4,107	328.405	128.438	0	612
monthslb^2	4,107	1,396.431	3,509.883	0	42,849
msmoke ∗monthslb^2	4,107	750.703	4,407.398	0	112,908

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.

In addition to the results presented in Section 8 we present some additional specifications below. Figure F.1 presents results from using a local constant regression with 3 knots in the first stage while Figure F.2. Both show that the results of analysis generally hold up in a variety of specifications, namely that the effect of maternal smoking on birthweight is negative and increasing in magnitude with age. The model-assisted estimated typically produce values that are more in line with previous work, though the shape of CATE estimates that use standard loss functions in the first stage are more stable to second stage basis.

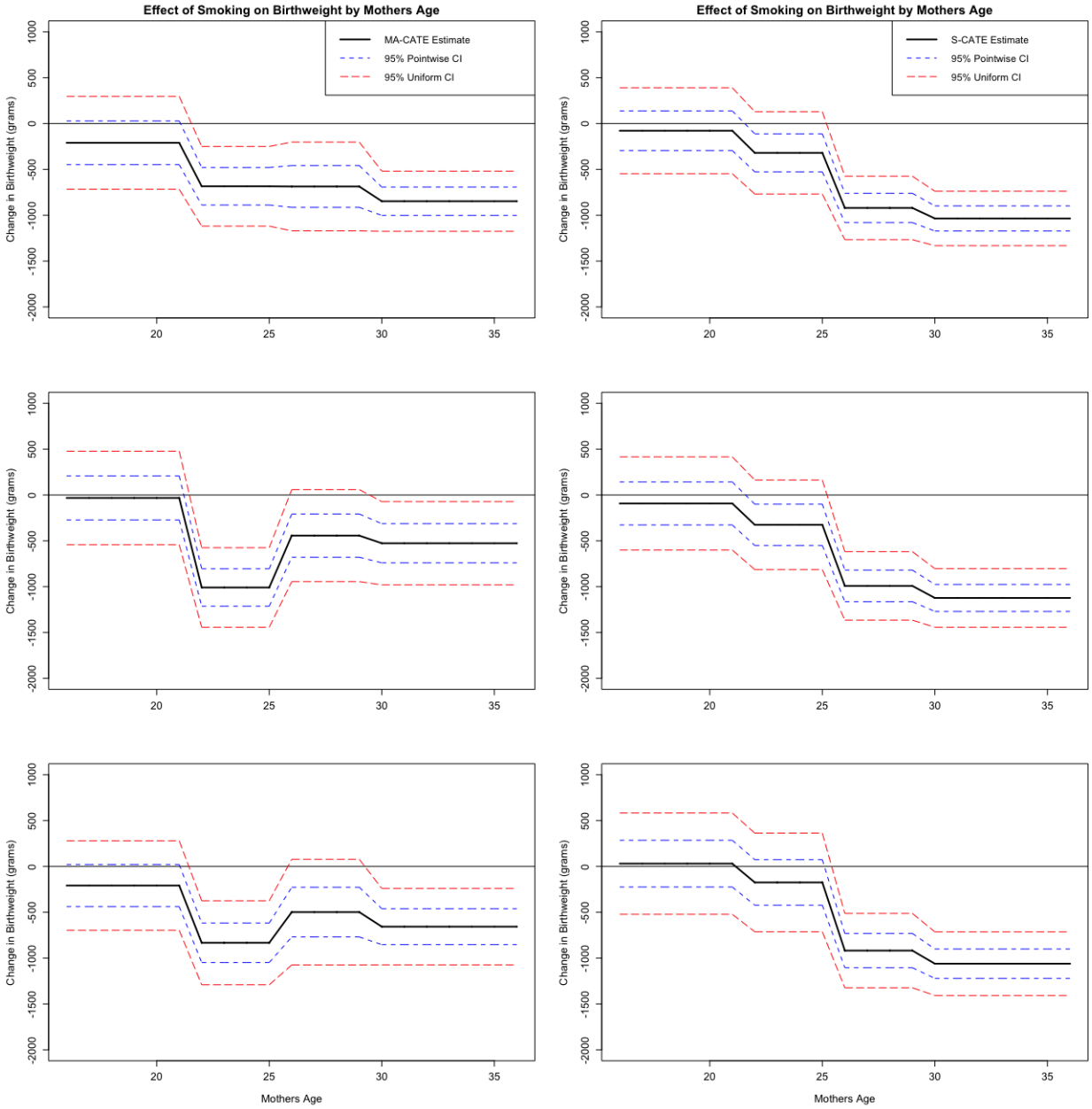

Figure F.1: CATE of maternal smoking estimated using model assisted estimating equations (left) and standard MLE/OLS estimating equations (right). Top row uses the 99^{th} quantile of the bootstrap distribution to select the penalty parameters, second row uses 95^{th} quantile, and final row uses the 90^{th} quantile. Second stage is computed using a local constant with 3 knots. 95% pointwise confidence intervals are displayed in blue short dashes and 95% uniform confidence bands are displayed in long red dashes.

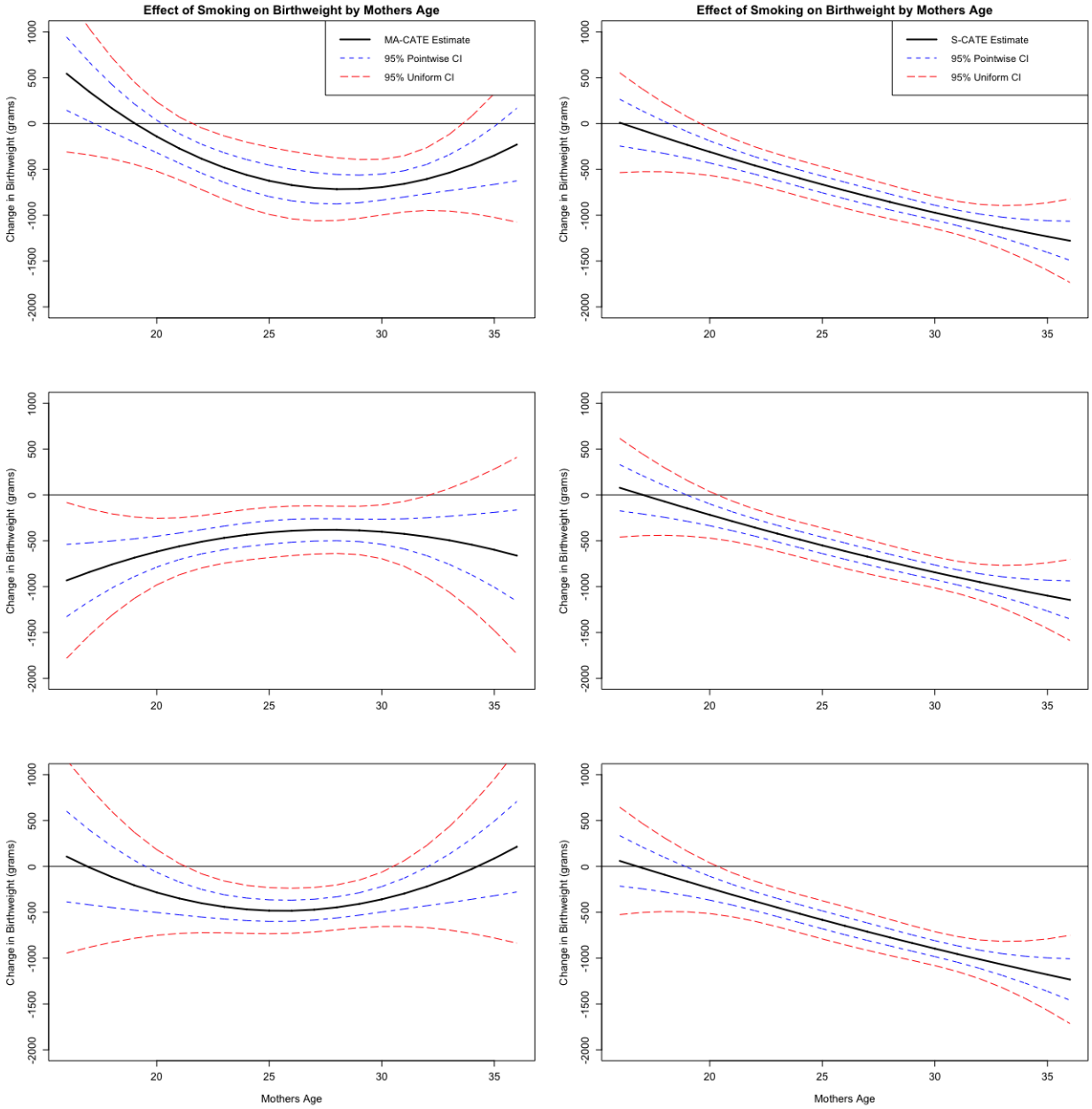

Figure F.2: CATE of maternal smoking estimated using model assisted estimating equations (left) and standard MLE/OLS estimating equations (right). Top row uses the 99^{th} quantile of the bootstrap distribution to select the penalty parameters, second row uses 95^{th} quantile, and final row uses the 90^{th} quantile. Second stage is computed using second degree b-splines with 2 knots. 95% pointwise confidence intervals are displayed in blue short dashes and 95% uniform confidence bands are displayed in long red dashes.