

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

Adam Baybutt and Manu Navjeevan<sup>†</sup>

University of California, Los Angeles

Revised April 19, 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  $\ell_1$  (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 conditions similar to 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.

<sup>†</sup>Corresponding author: [mnavjeevan@g.ucla.edu](mailto:mnavjeevan@g.ucla.edu).

# 1 Introduction

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

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

Estimation of the CATE generally requires first fitting propensity score and/or outcome regression models. When the number of control variables  $Z$  is large ( $d_z \gg n$ ), 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.<sup>1</sup>

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. Additionally, 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 inference 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  $n$ , though the dimension of the conditioning variable  $d_x$  will remain fixed. Our approach is based on Tan (2020) wherein model-assisted inference is developed for the average treatment effect. Following Semanova 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  $\ell_1$ -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  $\ell_1$  regularized first-stage estimation incorporates

<sup>1</sup>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  $\ell_1$ -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. [Fan et al. \(2022\)](#) provides a similar analysis using a second stage kernel estimator. For estimating the conditional average treatment effect, both papers suggest the use of the aIPW signal first proposed by [Robins et al. \(1994\)](#). The resulting second stage estimators are doubly robust but do not necessarily yield model-assisted inference.

Regularization using an  $\ell_1$ -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  $\ell_1$ -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 provides evidence from a simulation study, Section 7 ties up a technical detail, and Section 8 concludes. Proofs of main results are deferred to Appendix A.

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

## 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  $\eta < \mathbb{E}[D \mid Z = z] < 1 - \eta$  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:  $\mathbb{E}[Y_1 - Y_0 \mid X = x]$ . However, for this paper we largely focus on estimation and inference for the conditional average counterfactual outcome:

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

A model-assisted inference procedure for the other conditional counterfactual outcome,  $\mathbb{E}[Y_0 \mid X = x]$ , follows a similar procedure and is described in Section 7. 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,

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

is a function of a fitted propensity score model,  $\pi(Z)$ , and a fitted outcome regression model,  $m(Z)$ . After estimating these two first-stage models, whose true values are given  $\pi^*(Z) := \Pr(D = 1 \mid Z)$  and  $m^*(Z) := \mathbb{E}[Y \mid D = 1, Z]$ , 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

$$p^k(x) := (p_1(x), \dots, p_k(x))' \in \mathbb{R}_+^k. \quad (2.3)$$

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 shifting polynomials. To ensure that the basis terms are well behaved, we will make assumptions on  $\xi_{k,\infty} := \sup_{x \in \mathcal{X}} \|p^k(x)\|_\infty$ ,  $\xi_{k,2} := \sup_{x \in \mathcal{X}} \|p^k(x)\|_2$ , and the eigenvalues of the design matrix  $Q := \mathbb{E}[p^k(x)p^k(x)']$ .

**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:

$$\begin{aligned} \pi(Z; \gamma) &= (1 + \exp(-\gamma'Z))^{-1}, \\ m(Z; \alpha) &= \alpha'Z. \end{aligned} \quad (2.4)$$

To deal with the high-dimensional  $Z$ , we implement  $\ell_1$ -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  $j = 1, \dots, k$ , the parameters of (2.4),  $\gamma, \alpha \in \mathbb{R}^{d_z}$ , are estimated by

$$\hat{\gamma}_j := \arg \min_{\gamma} \mathbb{E}_n[p_j(X)\{De^{-\gamma'Z} + (1-D)\gamma'Z\}] + \lambda_{\gamma,j}\|\gamma\|_1, \quad (2.5)$$

$$\hat{\alpha}_j := \arg \min_{\alpha} \mathbb{E}_n[p_j(Z)De^{-\hat{\gamma}_j'Z}(Y - \alpha'Z)^2]/2 + \lambda_{\alpha,j}\|\alpha\|_1. \quad (2.6)$$

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

$$\bar{\gamma}_j := \arg \min_{\gamma} \mathbb{E}[p_j(X)\{De^{-\gamma'Z} + (1-D)\gamma'Z\}], \quad (2.7)$$

$$\bar{\alpha}_j := \arg \min_{\alpha} \mathbb{E}[p_j(Z)De^{-\bar{\gamma}_j'Z}(Y - \alpha'Z)^2]. \quad (2.8)$$

We will assume that the population minimizers are sufficiently sparse. To control the sparsity, define  $\mathcal{S}_{\gamma,j} := \{j : \bar{\alpha}_j \neq 0\}$ ,  $\mathcal{S}_{\alpha,j} := \{j : \bar{\alpha}_j \neq 0\}$  and  $s_k := \max_{1 \leq j \leq k} \{|\mathcal{S}_{\gamma,j}| \vee |\mathcal{S}_{\alpha,j}|\}$ . 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

$$\hat{\pi}_j(Z) := \pi(Z; \hat{\gamma}_j) \rightarrow_p \bar{\pi}_j(Z) := \pi(Z; \bar{\gamma}_j), \quad \hat{m}_j(Z) := m(Z; \hat{\alpha}_j) \rightarrow_p \bar{m}_j(Z) := m(Z; \bar{\alpha}_j).$$

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  $j = 1, \dots, k$   $\bar{\pi}_j(Z) = \pi^*(Z)$  ( $\bar{m}_j(Z) = m^*(Z)$ ). 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  $\pi \neq \pi^*$  and  $m \neq m^*$ ,

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

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

$$\begin{aligned}Y(\bar{\pi}_j, \bar{m}_j) &= g_0(x) + \epsilon_j, & \mathbb{E}[\epsilon_j \mid X] &= 0 \\ &= g_k(x) + r_k(x) + \epsilon_j\end{aligned}\tag{2.10}$$

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  $\beta^k$  governing the projection in  $g_k(x)$  can be identified by the projection of the aIPW signal onto the basis terms  $p^k(x)$ :<sup>1</sup>

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

To estimate  $\beta^k$  we combine the  $k$  estimates according to:

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

where  $\hat{Q} := \mathbb{E}_n[p^k(x) p^k(x)']$ . 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  $\epsilon^k := (\epsilon_1, \dots, \epsilon_k)$ .

**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:

$$\begin{aligned}\tilde{\Omega} &:= Q^{-1} \mathbb{E}[\{p^k(x) \circ (\epsilon^k + r_k)\} \{p^k(x) \circ (\epsilon^k + r_k)\}' ] Q^{-1} \\ \Omega_0 &:= Q^{-1} \mathbb{E}[\{p^k(x) \circ \epsilon^k\} \{p^k(x) \circ \epsilon^k\}' ] Q^{-1}\end{aligned}\tag{2.13}$$

where  $\circ$  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

$$\hat{\Omega} := \hat{Q}^{-1} \mathbb{E}_n[\{p^k(x) \circ \hat{\epsilon}^k\} \{p^k(x) \circ \hat{\epsilon}^k\}' ] \hat{Q}^{-1}\tag{2.14}$$

<sup>1</sup>The second and third equalities of (2.11) hold under Assumption 2.1 for any  $\pi \neq \pi^*$  and  $m \neq m^*$  by double robustness of the aIPW signal (2.9).

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

$$t(x) := \frac{\hat{g}(x) - g(x)}{\hat{\sigma}(x)} \quad (2.15)$$

Denote the bootstrapped test statistic as

$$\hat{t}^b(x) := \frac{p^k(x) \hat{\Omega}^{1/2}}{\|\hat{\Omega}^{1/2} p^k(x)\|} N_k^b \quad (2.16)$$

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

$$[\underline{i}(x), \bar{i}(x)] := [\hat{g}(x) - c^*(1 - \alpha) \hat{\sigma}(x), \hat{g}(x) + c^*(1 - \alpha) \hat{\sigma}(x)]. \quad (2.17)$$

The critical value  $c^*(1 - \alpha)$  is either  $c_p^*(1 - \alpha)$ , the  $(1 - \alpha)$  quantile of a  $N(0, 1)$  distribution, for pointwise inference or  $c_u^*(1 - \alpha)$ , the  $(1 - \alpha)$ -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  $\alpha \in S^{k-1}$ ,*

$$\sqrt{n} \frac{\alpha'(\hat{\beta}^k - \beta^k)}{\|\alpha' \hat{\Omega}^{1/2}\|} \rightarrow_d N(0, 1). \quad (3.1)$$

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

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

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

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

**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 specified the uniform confidence intervals stated in (2.17) satisfy*

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

Proposition 3.1 states that our method yields model-assisted pointwise inference, that is an inference proce-

ture based on the test statistic  $t(x)$  in (2.15) remains valid when using the pointwise critical value  $c_p^*(1 - \alpha)$  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  $c_u^*(1 - \alpha)$ .

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  $\hat{\beta}^k$ . Limiting the size of  $\mathbf{B}^k$  is crucial in showing convergence of  $\hat{\beta}^k$  to the true parameter  $\beta^k$  and thus consistency of the nonparametric estimator  $\hat{g}(x)$ .

$$\mathbf{B}^k := \mathbb{E}_n \begin{bmatrix} p_1(X) \{Y(\hat{\pi}_1, \hat{m}_1) - Y(\bar{\pi}_1, \bar{m}_1)\} \\ \vdots \\ p_k(X) \{Y(\hat{\pi}_k, \hat{m}_k) - Y(\bar{\pi}_k, \bar{m}_k)\} \end{bmatrix}. \quad (3.4)$$

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  $j^{\text{th}}$  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:<sup>1</sup>

$$\mathbb{E}[\partial_\pi Y(\pi^*, m^*) \mid Z] = \mathbb{E}[\partial_m Y(\pi^*, m^*) \mid Z] = 0. \quad (3.5)$$

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  $\bar{\pi}_j = \pi^*$  and  $\bar{m}_j = m^*$  and considering the following first order Taylor expansion:

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

By orthogonality of the aIPW signal the gradient term is close to zero, which guarantees that the bias is  $\sqrt{n}$ -asymptotically negligible, i.e.  $o_p(n^{-1/2})$ , even if the nuisance parameters converge slowly to the true

<sup>1</sup>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).



values,  $\pi^*$  and  $m^*$ .<sup>2</sup> 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  $\pi^*$  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  $\mathbb{E}_n[p_j(X)\nabla_{\pi,m}Y(\pi^*, m^*)]$  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  $\hat{\pi}$  and  $\hat{m}$  could be used at each element of  $\hat{\beta}^k$  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  $\bar{\pi}_j$  and  $\bar{m}_j$ .

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

So long as either one of  $\bar{\pi}_j = \pi^*$  or  $\bar{m}_j = m^*$ , the double robustness of the aIPW signal (2.9) still allows for identification of  $\mathbb{E}[p_j(X)Y_1]$ .<sup>3</sup> However, the aIPW orthogonality tells us nothing about the expectation of the gradient away from the true parameters,  $\pi^*, m^*$ ; if either  $\bar{\pi}_j \neq \pi^*$  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 parametric rate needed to establish 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.

$$\begin{aligned} \mathbf{B}_j^k &= \mathbb{E}_n[p_j(X)Y(\hat{\pi}_j, \hat{m}_j)] - \mathbb{E}_n[p_j(X)Y(\bar{\pi}_j, \bar{m}_j)] \\ &= \mathbb{E}_n[p_j(X)\nabla_{\gamma_j, \alpha_j}Y(\bar{\pi}_j, \bar{m}_j)] \begin{bmatrix} \hat{\gamma}_j - \bar{\gamma}_j \\ \hat{\alpha}_j - \bar{\alpha}_j \end{bmatrix} + o_p(n^{-1/2}) \end{aligned} \quad (3.8)$$

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

$$Y(\bar{\pi}_j, \bar{m}_j) = DY(1 + e^{-\bar{\gamma}_j'Z}) - D(1 + e^{-\bar{\gamma}_j'Z})\bar{\alpha}'Z + \bar{\alpha}'Z.$$

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

$$\mathbb{E}[p_j(X)\nabla_{\gamma_j, \alpha_j}Y(\bar{\pi}_j, \bar{m}_j)] = \mathbb{E} \begin{bmatrix} -p_j(X)De^{-\bar{\gamma}_j'Z}(Y - \bar{\alpha}'Z)Z \\ p_j(X)\{D(1 + e^{-\bar{\gamma}_j'Z})Z + Z\} \end{bmatrix} \quad (3.9)$$

However, by definition  $\bar{\gamma}_j$  and  $\bar{\alpha}_j$  solve the minimization problems defined in (2.7)-(2.8), the population

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

<sup>3</sup>That is  $\mathbb{E}_n[p_j(X)Y(\bar{\pi}_j, \bar{m}_j)] \approx \mathbb{E}_n[p_j(X)Y_1]$

analogs of our finite sample estimating equations. The first order conditions of these minimization problems yield that

$$\mathbb{E} \left[ \underbrace{\begin{bmatrix} p_j(X)\{D(1 + e^{\bar{\gamma}'Z})Z + Z\} \\ p_j(X)De^{-\bar{\gamma}'Z}(DY - \bar{\alpha}'Z)Z \end{bmatrix}}_{\text{First order condition of } \bar{\alpha}_j} \right] = 0 \implies \mathbb{E}[p_j(X)\nabla_{\gamma_j, \alpha_j} Y(\bar{\pi}_j, \bar{m}_j)] = 0 \quad (3.10)$$

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.<sup>4</sup>

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 parametric 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  $\bar{\gamma}_l$  and  $\bar{\alpha}_l$  for  $l \neq j$ :

$$\begin{aligned} \mathbb{E}[p_j(X)\nabla_{\gamma_j, \alpha_j} Y(\bar{\pi}_j, \bar{m}_j)] &= \mathbb{E} \left[ \begin{bmatrix} -p_j(X)De^{-\bar{\gamma}'Z}(Y - \bar{\alpha}'Z)Z \\ p_j(X)\{D(1 + e^{\bar{\gamma}'Z})Z + Z\} \end{bmatrix} \right] \\ &\quad \underbrace{\hspace{10em}}_{\text{First order condition of } \bar{\gamma}_l} \\ &\neq \mathbb{E} \left[ \underbrace{\begin{bmatrix} p_l(X)\{D(1 + e^{\bar{\gamma}'Z})Z + Z\} \\ p_l(X)De^{-\bar{\gamma}'Z}(Y - \bar{\alpha}'Z)Z \end{bmatrix}}_{\text{First order condition of } \bar{\alpha}_l} \right] = 0. \end{aligned} \quad (3.11)$$

Obtaining valid inference at all points  $x \in \mathcal{X}$  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.

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 nuisance 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.

<sup>4</sup>Correct specification is only necessary for identification.

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

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

In particular, Condition 1 implies that the bias vector  $\mathbf{B}^k$  from (3.4) satisfies  $\|\mathbf{B}^k\|_2 = o_p(n^{-1/2})$ . Establishing this condition largely follows the logic laid out in Section 3 after establishing the uniform consistency of the first-stage estimators,  $\hat{\gamma}_j$  and  $\hat{\alpha}_j$ , 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  $\ell_1$ -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  $j = 0, 1, \dots, k$ , we start with a pilot penalty parameters given by

$$\lambda_{\gamma,j}^{\text{pilot}} = c_{\gamma,j} \times \sqrt{\frac{\ln^3(d_z)}{n}} \quad \text{and} \quad \lambda_{\alpha,j}^{\text{pilot}} = c_{\alpha,j} \times \sqrt{\frac{\ln^3(d_z)}{n}} \quad (4.2)$$

for some arbitrary constants  $c_{\gamma,j}, c_{\alpha,j} > 0$ . Using  $\lambda_{\gamma,j}^{\text{pilot}}$  and  $\lambda_{\alpha,j}^{\text{pilot}}$  in lieu of  $\lambda_{\gamma,j}$  and  $\lambda_{\alpha,j}$  in (2.5)-(2.6) we generate pilot estimators  $\hat{\gamma}_j^{\text{pilot}}$  and  $\hat{\alpha}_j^{\text{pilot}}$ . These pilot estimators are used to generate plug in estimators  $\hat{U}_{\gamma,j}$  and  $\hat{U}_{\alpha,j}$  of the residuals

$$\begin{aligned} \hat{U}_{\gamma,j} &:= -p_j(X)\{De^{-\hat{\gamma}_j^{\text{pilot}'Z}} + (1-D)\} \\ \hat{U}_{\alpha,j} &:= p_j(X)De^{-\hat{\gamma}_j^{\text{pilot}'Z}}(Y - \hat{\alpha}_j^{\text{pilot}'Z}). \end{aligned} \quad (4.3)$$

We then use a multiplier bootstrap procedure to select our final penalty parameters  $\lambda_{\gamma,j}$  and  $\lambda_{\alpha,j}$ .

$$\begin{aligned} \lambda_{\gamma,j} &= c_0 \times (1-\epsilon)\text{-quantile of } \max_{1 \leq l \leq d_z} |\mathbb{E}_n[e_l \hat{U}_{\gamma,j} Z_l]| \text{ given } \{Y_i, D_i, Z_i\}_{i=1}^n, \\ \lambda_{\alpha,j} &= c_0 \times (1-\epsilon)\text{-quantile of } \max_{1 \leq l \leq d_z} |\mathbb{E}_n[e_l \hat{U}_{\alpha,j} Z_l]| \text{ given } \{Y_i, D_i, Z_i\}_{i=1}^n \end{aligned} \quad (4.4)$$

where  $e_1, \dots, e_n$  are independent standard normal random variables generated independently of the data  $\{Y_i, D_i, X_i\}_{i=1}^n$  and  $c_0 > 1$  is a fixed constant<sup>1</sup>. 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,

$$\begin{aligned} U_{\gamma,j} &:= -p_j(X)\{De^{-\bar{\gamma}_j^{\text{pilot}'Z}} + (1-D)\} \\ U_{\alpha,j} &:= p_j(X)De^{-\bar{\gamma}_j^{\text{pilot}'Z}}(Y - \bar{\alpha}_j^{\text{pilot}'Z}), \end{aligned} \quad (4.5)$$

the choice of penalty parameter in (4.4) will ensure that the penalty parameter dominates the noise with

<sup>1</sup>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 satisfies 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.

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  $c_{\gamma,j}$  and  $c_{\alpha,j}$  in (4.2). The optimal choice of these constants may depend on the underlying data degenerating process. We recommend using cross validation to pick these constants from a fixed set of possible values. As a rule of thumb, we recommend the values in the set be chosen to be on the order of the maximum value of  $\|p^k(X_i)\|_\infty \cdot \|Z_i\|_\infty$  observed in the data. So long as this set of possible values is fixed, 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,  $\max_{1 \leq l \leq d_z} |Z_l| \leq C_0$  almost surely.
- (ii) There is a constant  $B_0$  such that  $\bar{\gamma}'_j Z \geq B_0$  almost surely for all  $j$ .
- (iii) There exist fixed constants  $\xi_0 > 1$  and  $1 > \nu_0 > 0$  such that for each  $j = 1, \dots, l$  the following empirical compatability condition holds for the empirical hessian matrix  $\tilde{\Sigma}_{\gamma,j} := \mathbb{E}_n[De^{-\bar{\gamma}'_j Z} Z Z']$ . For any  $b \in \mathbb{R}^{d_z}$

$$\sum_{l \notin \mathcal{S}_{\gamma,j}} |b_l| \leq \xi_0 \sum_{l \in \mathcal{S}_{\gamma,j}} |b_l| \implies \nu_0^2 \left( \sum_{l \in \mathcal{S}_{\gamma,j}} |b_l| \right)^2 \leq |\mathcal{S}_{\gamma,j}| \left( b' \tilde{\Sigma}_{\gamma,j} b \right).$$

- (iv) There exist fixed constants  $c_u$  and  $C_U > 0$  such that for all  $j \leq k$ ,  $\mathbb{E}[U_{\gamma,j}^4] \leq (\xi_{k,\infty} C_U)^4$  and  $\min_{1 \leq l \leq d_z} \mathbb{E}[U_{\gamma,j}^2 Z_l^2] \geq c_u$ .
- (v) The following sparsity bounds hold with  $s_{k,\gamma} := \max_{1 \leq j \leq k} |\mathcal{S}_{\gamma,j}|$ .

$$\frac{\xi_{k,\infty}}{\ln(d_z n)} \rightarrow 0, \quad \frac{\xi_{k,\infty} s_{k,\gamma} \ln^5(d_z n)}{n} \rightarrow 0, \quad \text{and} \quad \frac{\xi_{k,\infty}^4 \ln^7(d_z k n)}{n} \rightarrow 0.$$

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  $\mathcal{X}$  is generally required by nonparametric estimators. The second assumption bounds all limiting propensity scores  $\bar{\pi}_j(Z)$  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 literature (Bickel et al., 2009; Belloni et al., 2012). The penultimate condition is an identifiability constraint that limits the moments of the noise and bounds it away from zero uniformly over all estimation procedures. 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,  $\xi_{k,\infty}$  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  $Y_1 - \bar{m}_j(Z)$  are uniformly subgaussian conditional on  $Z$  in the following sense. There exist fixed positive constants  $D_0$  and  $D_1$  such that for any  $j$ :

$$D_0 \mathbb{E} \left[ \exp \left( \{Y_1 - \bar{m}_j(Z)\}^2 / D_0^2 \right) - 1 \mid Z \right] \leq D_1^2$$

almost surely.

- (ii) For any  $j$  the following empirical compatibility condition holds with fixed constants  $\xi_1 > 1$  and  $1 \geq \nu_1 > 0$  for the empirical hessian matrix  $\tilde{\Sigma}_{\gamma,j}$ :

$$\sum_{l \notin \mathcal{S}_{\alpha,j}} |b_l| \leq \xi_0 \sum_{l \in \mathcal{S}_{\alpha,j}} |b_l| \implies \nu_0^2 \left( \sum_{l \in \mathcal{S}_{\alpha,j}} |b_l| \right)^2 \leq |\mathcal{S}_{\alpha,j}| \left( b' \tilde{\Sigma}_{\gamma,j} b \right).$$

- (iii) For all  $j \leq k$ ,  $\mathbb{E}[U_{\alpha,j}^4] \leq (\xi_{k,\infty} C_U)^4$  and  $\min_{1 \leq l \leq d_z} \mathbb{E}[U_{\alpha,j}^2 Z_l^2] \geq c_u$ .
- (iv) The following sparsity bound holds with  $s_{k,\alpha} = \max_{1 \leq j \leq k} |\mathcal{S}_{\alpha,j}|$ :

$$\frac{\xi_{k,\infty}^2 s_{k,\alpha}^2 \ln^5(d_z n)}{n} \rightarrow 0.$$

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 compatibility 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 bound 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 to show Condition 1.

We can now establish uniform convergence of the propensity score model and the outcome regression model parameters. The convergence will be presented in terms of the  $\ell_1$  norm of the coefficient vector as well as the symmetrized Bregman divergences,  $D_{\gamma,j}^\dagger(\hat{\gamma}_j, \bar{\gamma}_j)$  and  $D_{\alpha,j}^\dagger(\hat{\alpha}_j, \bar{\alpha}_j; \gamma_j)$ , defined by

$$\begin{aligned} D_{\gamma,j}^\dagger(\hat{\gamma}_j, \bar{\gamma}_j) &:= \mathbb{E}_n \left[ p_j(X) D \{ e^{-\hat{\gamma}_j' Z} - e^{-\bar{\gamma}_j' Z} \} \{ \hat{\gamma}_j' Z - \bar{\gamma}_j' Z \} \right], \\ D_{\alpha,j}^\dagger(\hat{\alpha}_j, \bar{\alpha}_j; \gamma_j) &:= \mathbb{E}_n \left[ p_j(X) D e^{-\hat{\gamma}_j' Z} (\bar{\alpha}_j' Z - \hat{\alpha}_j' Z)^2 \right]. \end{aligned} \tag{4.6}$$

We refer readers to discussion in Tan (2017) for details and motivation. 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-penalized 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  $\xi_0 > (c_0 + 1)/(c_0 - 1)$ ,  $k/n \rightarrow 0$ , and that  $k\epsilon \rightarrow 0$  as  $n \rightarrow \infty$ . Then the logistic propensity score model parameters are uniformly bounded:*

$$\sup_{1 \leq j \leq k} D_{\gamma,j}^{\dagger}(\hat{\gamma}_j, \bar{\gamma}_j) \lesssim_P \xi_{k,\infty}^2 \frac{s_k \ln(d_z)}{n} \quad \text{and} \quad \sup_{1 \leq j \leq k} \|\hat{\gamma}_j - \bar{\gamma}_j\|_1 \lesssim_P \xi_{k,\infty} \sqrt{\frac{s_k^2 \ln(d_z)}{n}} \quad (4.7)$$

**Theorem 4.2** (Outcome Regression Model). *Suppose that Assumptions 4.1 and 4.2 hold. In addition assume that  $\xi_0 > (c_0 + 1)/(c_0 - 1)$ ,  $k/n \rightarrow 0$ ,  $k\epsilon \rightarrow 0$ , and that there is a fixed constant  $c > 0$  such that for all  $j$ ,  $\lambda_{\alpha,j}/\lambda_{\gamma,j} \geq c$ . Then the linear outcome regression model parameters are uniformly bounded:*

$$\sup_{1 \leq j \leq k} D_{\alpha}^{\dagger}(\hat{\alpha}_j, \bar{\alpha}_j; \bar{\gamma}_j) \lesssim_P \xi_{k,\infty}^2 \frac{s_k \ln(d_z)}{n} \quad \text{and} \quad \sup_{1 \leq j \leq k} \|\hat{\alpha}_j - \bar{\alpha}_j\|_1 \lesssim_P \xi_{k,\infty} \sqrt{\frac{s_k^2 \ln(d_z)}{n}} \quad (4.8)$$

**Remark 4.3** (Penalty Parameter Ratio). The existence of a fixed constant  $c > 0$  such  $\lambda_{\alpha,j}/\lambda_{\gamma,j} \geq c$  may seem a bit unnatural, but it can be enforced in practice by setting the penalty to instead be, say,  $\lambda_{\alpha,j}^{\text{ratio}} := \max\{\lambda_{\gamma,j}/5, \lambda_{\alpha,j}\}$ . The use of this alternative penalty parameter instead of  $\lambda_{\alpha,j}$  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  $c_0 > (\xi_0 + 1)/(\xi_0 - 1)$ ,  $k/n \rightarrow 0$ ,  $k\epsilon \rightarrow 0$ , and there is a fixed constant  $c > 0$  such that for all  $j$ ,  $\lambda_{\alpha,j}/\lambda_{\gamma,j} \geq c$ . Then the elements of the vector (3.4) converge uniformly:*

$$\max_{1 \leq j \leq k} |\mathbb{E}_n[p_j(X)Y(\hat{\pi}_j, \hat{m}_j)] - \mathbb{E}_n[p_j(X)Y(\bar{\pi}_j, \bar{m}_j)]| \lesssim_P \frac{s_k \xi_{k,\infty}^2 \ln(d_z)}{n} \quad (4.9)$$

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

$$\frac{s_k \xi_{k,\infty}^2 k^{1/2} \ln(d_z)}{\sqrt{n}} \rightarrow 0. \quad (4.10)$$

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,

$$\xi_{k,\infty} \max_{1 \leq j \leq k} \mathbb{E}_n[p_j(X)^2(Y(\hat{\pi}_j, \hat{m}_j) - Y(\bar{\pi}_j, \bar{m}_j))^2] = o_p(k^{-2}n^{-1/m}) \quad (4.11)$$

**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:*

$$\max_{1 \leq j \leq k} \mathbb{E}_n[p_j^2(X)(Y(\hat{\pi}_j, \hat{m}_j) - Y(\bar{\pi}_j, \bar{m}_j))^2] \lesssim_P \xi_{k,\infty}^4 \frac{s_k^2 \ln(d_z)}{n} \quad (4.12)$$

*In particular, Condition 2 holds if additionally*

$$\frac{\xi_{k,\infty}^5 s_k^2 k^2 \ln(d_z)}{n^{(m-1)/m}}. \quad (4.13)$$

**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:

$$\max_{1 \leq i \leq n} \max_{1 \leq j \leq k} |Y_i(\hat{\pi}_j, \hat{m}_j) - Y_i(\bar{\pi}_j, \bar{m}_j)| \quad \text{and} \quad \max_{1 \leq i \leq n} \max_{1 \leq j \leq k} |Y_i(\hat{\pi}_j, \hat{m}_j) - Y_i(\bar{\pi}_j, \bar{m}_j)|^2$$

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  $\|\hat{\pi}_j - \bar{\pi}_j\|$  and  $\|\hat{m}_j - \bar{m}_j\|$  are small. Because the difference  $Y_i(\hat{\pi}_j, \hat{m}_j) - Y_i(\bar{\pi}_j, \bar{m}_j)$  depends on  $Y$ , not knowing its expectation on the event that  $\hat{\pi}_j$  is close to  $\bar{\pi}_j$  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  $s\sqrt{\ln(d_z)/n} \rightarrow 0$ , 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 after. The constants  $c_{\gamma,j}$  and  $c_{\alpha,j}$  can be selected via cross validation from a growing set  $\Lambda_n \subseteq [\underline{c}_n, \bar{c}_n]$ . 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:

$$\frac{k|\Lambda_n|}{n^2} \rightarrow 0, \quad \frac{\underline{c}_n^{-1}\xi_{k,\infty}}{\ln(d_z n)} \rightarrow 0, \quad \frac{\bar{c}_n^2 s_k k^{1/2} \ln^3(d_z n)}{\sqrt{n}} \rightarrow 0, \quad \text{and} \quad \frac{\xi_{k,\infty}^2 \bar{c}_n^2 s_k k^2 \ln^3(d_z n)}{n^{(m-1)/m}} \rightarrow 0 \quad (4.14)$$

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 existing 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 logistite propensity score model is correctly specified and the limiting outcome regression models are uniformly bounded conditional on  $Z$ , our uniform 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  $Q = \mathbb{E}[p^k(x)p^k(x)']$  are bounded from above and away from zero.
- (ii) The conditional variance of the error terms is uniformly bounded in the following sense. There exist constants  $\underline{\sigma}^2$  and  $\bar{\sigma}^2$  such that for any  $j = 1, 2, \dots$  we have that  $\underline{\sigma}^2 \leq \text{Var}(\epsilon_j | X) \leq \bar{\sigma}^2 < \infty$ ;
- (iii) For each  $n$  and  $k$  there are finite constants  $c_k$  and  $\ell_k$  such that for each  $f \in \mathcal{G}$

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

**Assumption 5.2** (Uniform Integrability). Let  $\bar{\epsilon}_k := \max_{1 \leq j \leq k} |\epsilon_j|$ . Assume that



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

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  $j = 1, \dots, k$ . 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. See [Newey \(1997\)](#), [Chen \(2007\)](#), or [Belloni et al. \(2015\)](#) for a more in depth discussion.<sup>1</sup>

**Theorem 5.1** (Pointwise Normality). *Suppose that Condition 1, Assumption 5.1, and Assumption 5.2 holds. In addition suppose that  $\xi_k^2 \log k/n \rightarrow 0$ . Then so long as either the logistic propensity score model or linear outcome regression model is correctly specified, for any  $\alpha \in S^{k-1}$ :*

$$\sqrt{n} \frac{\alpha'(\hat{\beta}^k - \beta^k)}{\|\alpha' \Omega^{1/2}\|} \rightarrow_d N(0, 1) \quad (5.1)$$

where generally  $\Omega = \tilde{\Omega}$  but if  $\ell_k c_k \rightarrow 0$  then we can set  $\Omega = \Omega_0$ . Moreover, for any  $x \in \mathcal{X}$  and  $s(x) := \Omega^{1/2} p^k(x)$ ,

$$\sqrt{n} \frac{p^k(x)'(\hat{\beta}^k - \beta^k)}{\|s(x)\|} \rightarrow_d N(0, 1) \quad (5.2)$$

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

$$\sqrt{n} \frac{\hat{g}(x) - g(x)}{\|s(x)\|} \rightarrow_d N(0, 1) \quad (5.3)$$

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  $c_p^*(1-\alpha)$  up to procurement of a consistent estimator for the variance matrix  $\Omega$ . Theorem 5.4 will provide conditions under which the plug in estimator  $\hat{\Omega}$  defined in (2.14) is consistent.

## 5.2 Uniform Convergence

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

**Assumption 5.3** (Uniform Limit Theory). Let  $\bar{\epsilon}_k = \sup_{1 \leq j \leq k} |\epsilon_j|$ ,  $\alpha(x) := p^k(x)/\|p^k(x)\|$ , and let

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

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

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

<sup>1</sup>In practice, we recommend the use of B-splines in order 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).

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

As before, Assumption 5.3 is very similar to its analogue in Belloni et al. (2015), with the modification that the conditions are required to hold for  $\bar{\epsilon}_k$  as opposed to  $\epsilon_k$ . 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 specified:*

$$\sup_{x \in \mathcal{X}} |\alpha(x)' \mathbb{G}_n[p^k(x) \circ \epsilon^k]| \lesssim_P \sqrt{\log k} \quad (5.4)$$

Moreover, for

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

we have that

$$\sup_{x \in \mathcal{X}} |p^k(x)'(\hat{\beta}^k - \beta^k)| \lesssim_P \frac{\xi_k}{\sqrt{n}} \left( \sqrt{\log k} + \bar{R}_{1n} + \bar{R}_{2n} \right) \quad (5.5)$$

and

$$\sup_{x \in \mathcal{X}} |\hat{g}(x) - g_0(x)| \lesssim_P \frac{\xi_k}{\sqrt{n}} \left( \sqrt{\log k} + \bar{R}_{1n} + \bar{R}_{2n} \right) + \ell_k c_k \quad (5.6)$$

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  $m \geq 3$ . In addition assume that (i)  $\bar{R}_{1n} = o_p(a_n^{-1})$  and (ii)  $a_n^6 k^4 \xi_k^2 (\bar{\sigma}_k^3 + \ell_k^3 c_k^2)^2 \log^2 n/n \rightarrow 0$ . Then so long as either the propensity score model or outcome regression model is correctly specified, for some  $\mathcal{N}_k \sim N(0, I_k)$ :*

$$\sqrt{n} \frac{\alpha(x)'(\hat{\beta} - \beta)}{\|\alpha(x)'\Omega^{1/2}\|} =_d \frac{\alpha(x)'\Omega^{1/2}}{\|\alpha(x)'\Omega^{1/2}\|} \mathcal{N}_k + o_p(a_n^{-1}) \text{ in } \ell^\infty(\mathcal{X}) \quad (5.7)$$

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

$$\sqrt{n} \frac{p^k(x)'(\hat{\beta} - \beta)}{\|s(x)\|} =_d \frac{s(x)}{\|s(x)\|} \mathcal{N}_k + o_p(a_n^{-1}) \text{ in } \ell^\infty(\mathcal{X}) \quad (5.8)$$

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

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

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  $\Omega_0$  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  $\ell^\infty(X)$ . 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  $Q = \mathbb{E}[p^k(x)p^k(x)']$  and  $\Sigma = \mathbb{E}[\{p^k(x) \circ (\epsilon^k + r_k)\} \circ \{p^k(x) \circ (\epsilon^k + r_k)\}']$  are estimated using

$$\begin{aligned}\hat{Q} &= \mathbb{E}_n[p^k(x)p^k(x)'] \\ \hat{\Sigma} &= \mathbb{E}_n[\{p^k(x) \circ \hat{\epsilon}^k\}\{p^k(x) \circ \hat{\epsilon}^k\}']\end{aligned}\tag{5.10}$$

where  $\hat{\epsilon}^k := (\hat{\epsilon}_1, \dots, \hat{\epsilon}_k)$  and  $\epsilon_j = Y(\hat{\pi}_j, \hat{m}_j) - \hat{g}(x)$  for each  $j = 1, \dots, k$ . In addition define  $v_n = (\mathbb{E}[\max_{1 \leq i \leq n} |\bar{\epsilon}_k|^2])^{1/2}$ .

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

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

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

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

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 satisfied and Assumptions 5.1-5.3 hold with  $m \geq 4$ . In addition suppose (i)  $R_{1n} + R_{2n} \lesssim \log^{1/2} n$ , (ii)  $\xi_k \log^2 n / n^{1/2-1/m} = o(1)$ , (iii)  $\sup_{x \in \mathcal{X}} |r_k(x)| / \|p^k(x)\| = o(\log^{-1/2} n)$ , and (iv)  $k^4 \xi_k^2 (1 + l_k^3 r_k^3)^2 \log^5 n / n = o(1)$ . Then for  $t(x)$  defined in (2.15), so long as either the propensity score model or outcome regression model is satisfied*

$$\Pr \left( \sup_{x \in \mathcal{X}} |t(x)| \leq c^*(1 - \alpha) \right) = 1 - \alpha + o(1).$$

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

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

The width of the confidence bands  $2c^*(1 - \alpha)\hat{\sigma}(x)$  satisfy

$$2c^*(1 - \alpha)\hat{\sigma}(x) \lesssim_P \sigma(x) \sqrt{\log n} \lesssim \sqrt{\frac{\xi_k^2 \log n}{n}}.$$

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  $c_u^*(1 - \alpha)$  shrink at the uniform nonparametric rate derived in Theorem 5.2.

## 6 Simulation Study

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

### 6.1 Simulation Design

Random variables are generated i.i.d. according to the following distributions. The error term  $\epsilon$  is independently generated following  $\epsilon \sim N(0, \sigma_\epsilon^2)$  where  $\sigma_\epsilon = 0.1$ .  $Z = (X, Z_1) \in \mathbb{R}^{d_x, d_z - d_x}$  where  $d_x = 1$ ,  $d_z \in \{100, 500\}$ ,  $X$  is independently generated following  $U(0, 1)$ , and the independent regressors  $Z_1$  are jointly centered Gaussian with a covariance matrix of the Toeplitz form

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

These distributions and the following data-generating processes (DGP) follow the design of Tan (2020). Let  $Z^\dagger$  be a transformation of the regressors where  $Z_j^\dagger = Z_j + \max(0, 0.1 + Z_j)^2$ ,  $\forall j = 2, \dots, d_z$ . Let **sparsity**  $\in \{50, 100\}$  control the number of regressors in  $Z_1$ , i.e. excluding  $X$  and a constant term, entering the DGP by  $\gamma = \alpha = (\underbrace{1, \dots, 1}_{\text{sparsity}}, \underbrace{0, \dots, 0}_{d_z - \text{sparsity} - 1})$ .

- (S1) *Correct specification:* Generate  $D$  given  $Z$  from a Bernoulli distribution with  $\Pr(D = 1|Z) = \{1 + \exp(0.5 - \gamma'Z)\}^{-1}$  and  $Y = 1 + D + \alpha'Z + \epsilon$ .
- (S2) *Propensity score model correctly specified, but outcome regression model misspecified:* Generate  $D$  given  $Z$  as in (S1), but  $Y = 1 + D + \alpha'Z^\dagger + \epsilon$ .
- (S3) *Propensity score model misspecified, but outcome regression model correctly specified:* Generate  $Y$  according to (S1), but generate  $D$  given  $Z$  from a Bernoulli distribution with  $\Pr(D = 1|Z) = \{1 + \exp(0.5 - \gamma'Z^\dagger)\}^{-1}$ .

When  $d_z = 100$ , **sparsity** = 10 and when  $d_z = 500$ , **sparsity** = 50. To consider various degrees of high-dimensionality,  $N \in \{500, 1000\}$  to allow  $d_z = N$  as well as  $d_z = \frac{N}{2}$  and  $d_z = \frac{N}{5}$ . Results are reported for  $S = 1000$  repeated simulations.

## 6.2 Estimators and Implementation

In our first-stage estimation, we need to set values for the arbitrary constants  $c_{\gamma,j}, c_{\alpha,j}$  for  $j \in \{1, \dots, k\}$  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  $\hat{\gamma}_j^{\text{pilot}}$  and  $\hat{\alpha}_j^{\text{pilot}}$  for  $j \in \{1, \dots, k\}$ . Following the multiplier bootstrap procedure outlined in (4.3) and (4.4) with  $B = 1,000$  bootstrap samples and  $c_0 = 1.1$ , the final penalty parameters  $\lambda_{\gamma,j}$  and  $\lambda_{\alpha,j}$  are obtained for use in the first-stage estimators (2.5) and (2.6). These  $2 \times k$  estimations for  $\hat{\gamma}_j$  and  $\hat{\alpha}_j$  are implemented in parallel using the `optimize.minimize` function in Python's `scipy` package with the `SLSQP` method and the remaining settings using default values.

In our second-stage estimation, we follow the series estimation outlined in (2.12) to obtain  $\hat{\beta}^k$  which is used to estimate the target parameter  $g_0(x)$  with  $\hat{g}(x) := p^k(x)' \hat{\beta}^k$ . Results are reported for  $k \in \{3, 10\}$ . For an approximating basis, B-splines are used according to Powell et al. (1981). Specifically, we take

$$p_j(x) = \sum_{i=j}^{j+q+1} \left( \max(0, x - t_i)^q \prod_{l=i+1}^{j+q+1} (t_l - t_i) \right) \quad (6.1)$$

where  $j = 1, \dots, k$ ;  $q, m$  are positive integers such that  $m > q + 1$  and  $k = m - q - 1$ ;  $\{t_i; i = 1, \dots, m\}$  are equally spaced knots within  $(a, b)$ ; and,  $a, b$  are set to the minimum and maximum value of the support of  $X$ , respectively. When  $k = 3$ ,  $m = 5$  and  $q = 1$ ; similarly, when  $k = 10$ ,  $m = 12$  and  $q = 1$ .

To estimate the variance of  $\hat{g}(x)$  and form confidence intervals, we follow (5.10) to obtain the estimated covariance matrix  $\hat{\Omega}$ . Pointwise confidence intervals are then formed using the estimated standard error  $\hat{s}(x) = \sqrt{p^k(x)' \hat{\Omega} p^k(x)}$  according to Theorem 5.1.

This Monte Carlo study aims to compare the proposed estimation and inference procedures to commonly used methods targeting the CATE. First, for the base method, we follow Newey (1997) to regress the IPW signal for  $Y_1$  of  $Y_{IPW} = \frac{Y \cdot D}{\pi(Z)}$  on the series terms defined in (6.1). These series terms are common across all benchmarks. This first method is denoted as *Newey*. The propensity score  $\pi(Z)$  is estimated using logistic regression. Next, to examine the benefits of regularization for high-dimensional controls, we repeat the first method, but add an  $\ell_1$ -penalty to the first-stage logistic regression. This second method is denoted as *Newey-Reg*. To demonstrate the benefits of the aIPW in handling regularization bias, we add a third and final benchmark method following Semenova and Chernozhukov (2021) to use 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  $\ell_1$ -penalty; and, estimate the outcome regression via lasso.<sup>1</sup> Let *MA-DML* refer to our method, which aims to demonstrate, in relation to *DML*, the benefits of model assisted inference under misspecification.

Across all three benchmark methods, we cross-validate for optimal penalty parameter over a grid of values, akin to our estimation procedure. Each method's logistic regression is performed using Python's `sklearn` package with the `saga` solver. *DML*'s lasso for the outcome regression is implemented using Python's `sklearn`

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

package. Finally, the variance estimator is also common across the three methods:

$$\widehat{\Omega} = \left( \mathbb{E}_n \left[ p^k(X) p^k(X) \right] \right)^{-1} \mathbb{E}_n \left[ p^k(X) p^k(X) (Y(\widehat{\eta}) - \widehat{g}(X))^2 \right] \left( \mathbb{E}_n \left[ p^k(X) p^k(X) \right] \right)^{-1}$$

where  $\widehat{\eta} = \widehat{\pi}$  for the classic methods and  $\widehat{\eta} = \{\widehat{\pi}, \widehat{m}\}$  for *DML*.

We study the estimation error of each method's  $\widehat{g}(x)$  for the target parameter  $g_0(x) = \mathbb{E}[Y|D = 1, X = x]$  over a grid of 100 points equally spaced across  $x \in [0, 1]$ , i.e. the support of  $X$ . We study the coverage across the  $S$  simulations of each method's pointwise confidence interval formed using  $\widehat{s}(x)$  at  $x = 0.5$ . To be precise, the DGP gives  $g_0(x) = 2 + x + \mathbb{E}[\alpha'_{2:d_z} \tilde{Z}_1]$  where  $\alpha_{2:d_z}$  is the vector  $\alpha$  without the first component and  $\tilde{Z} \in \{Z_1, Z_1^\dagger\}$  according to specifications S1-3. Furthermore, the series terms  $p^k(x)$  are formed using the grid values of  $x$  while both  $\widehat{\beta}^k$  and  $\widehat{\Omega}$  are estimated using the underlying data across all four methods.

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

$$\begin{aligned} \text{IBias}^2 &= \int_0^1 (\bar{g}(x) - g_0(x))^2 dx, \\ \text{IVar} &= S^{-1} \sum_{s=1}^S \int_0^1 (\widehat{g}_s(x) - \bar{g}(x))^2 dx, \\ \text{IMSE} &= S^{-1} \sum_{s=1}^S \int_0^1 (\widehat{g}_s(x) - g_0(x))^2 dx. \end{aligned}$$

### 6.3 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 commonly used benchmark methods. Specifically, do the DML methods improve coverage under regularization to handle high-dimensionality? Do the model assisted inference methods improve the estimator's coverage properties under misspecification? How does performance respond to different degrees of high-dimensionality and the number of series terms specified?

**TODO: Add paragraphs describing the results in Table 1.**

- Describe structure of the table.
- Describe the key results in relation to how they answer our research questions.
- State what the take-aways are for our method in relation to benchmarks and how this answers research questions.

**FIGURE(S) (A LA VIRA FIG 2): FIGURES WITH THREE PANELS FOR S1, S2, and S3 WITH ALL/SOME OF THE FOUR METHODS REPORTING OUT THE ESITMATED G(X) NORMALIZED TO HORIZONATAL LINE WITH THEIR CI'S. SHOW REGIMES WHERE HIGH-DIM BREAKS CONVENTIONAL METHODS; MISSPECIFICATION +**

Table 1. Simulation study.

	(S1) cor PS, cor OR				(S2) cor PS, mis OR				(S3) mis PS, cor OR			
	Newey	Newey-Reg	DML	MA-DML	Newey	Newey-Reg	DML	MA-DML	Newey	Newey-Reg	DML	MA-DML
K=3, N=500, and $d_z=10$												
IBias <sup>2</sup>	0.96	0.05	0.05	0.05	3.26	0.08	0.08	0.08	1.09	0.03	0.27	0.32
Ivar	0.06	0.11	0.11	0.11	0.13	0.37	0.37	0.37	0.33	0.08	0.11	0.25
IMSE	1.02	0.15	0.15	0.15	3.39	0.39	0.39	0.39	1.42	0.11	0.38	0.55
Cov90	0.00	0.89	0.89	0.89	0.00	0.93	0.93	0.93	0.06	0.85	0.49	0.69
Cov95	0.01	0.95	0.95	0.95	0.00	0.97	0.97	0.97	0.11	0.87	0.56	0.77
K=3, N=500, and $d_z=100$												
IBias <sup>2</sup>	5.73	0.03	0.03	0.03	15.16	0.08	0.08	0.14	2040.19	0.09	0.21	1.42
Ivar	49.81	0.18	0.18	0.19	59.43	0.86	0.86	0.99	238542.49	0.12	1.70	0.60
IMSE	55.82	0.20	0.20	0.21	75.49	0.81	0.81	1.01	240574.19	0.22	1.89	2.00
Cov90	0.18	0.93	0.93	0.91	0.15	0.95	0.95	0.92	0.19	0.79	0.76	0.70
Cov95	0.24	0.96	0.96	0.97	0.21	0.97	0.97	0.95	0.19	0.85	0.80	0.80
K=3, N=500, and $d_z=500$												
IBias <sup>2</sup>												
Ivar												
IMSE												
Cov90												
Cov95												

Note: IBias<sup>2</sup>, IVar, IMSE, refer to integrated squared bias, variance, and mean-squared error, respectively. Cov90 and Cov95 are the coverage proportion of the 90% and 95% 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  $Z$ .

## HIGH-DIM BREAKS ALL BUT OURS.

**TODO: Add paragraphs describing the results in the Figures.**

- Describe structure of the figures.
- Describe the key results in relation to how they answer our research questions.
- State what the take-aways are for our method in relation to benchmarks and how this answers research questions.

**TODO: Final paragraph to back off critique of benchmark methods by praising them but close with sentence on what MA-DML offers. See Denis WP for example.**

## 7 Estimation of the Conditional Average Treatment Effect

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

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

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

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

To do so, we use the alternate aIPW signal

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

where as before the true value for  $\pi_0^*(z) = \Pr(D = 1 \mid Z = z)$  but now  $m_0^*(z) = \mathbb{E}[Y \mid D = 0, Z = z]$ . To estimate these nuisance models we again assume a logistic form for the propensity score model  $\pi_0(z) = \pi(z; \gamma^0)$  and a linear form for the outcome regression model  $m_0(z) = m(z, \alpha^0)$  as in (2.4) 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  $\gamma_j^0$  and  $\alpha_j^0$  differ from those in (2.5)-(2.6) however, and are instead given

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

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

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

Letting  $\bar{\pi}_{0,j}(z) = \pi(z, \bar{\gamma}_j^0)$ , and  $\bar{m}_{0,j}(z) = m(z, \bar{\alpha}_j^0)$  we can repeat the decomposition of Section 3, expressing  $\tilde{Y}(\bar{\pi}_{0,j}, \bar{m}_{0,j})$  as functions of the parameters  $\bar{\gamma}_j^0$  and  $\bar{\alpha}_j^0$  and show that the first order conditions for  $\bar{\gamma}_j^0$  and  $\bar{\alpha}_j^0$  directly control the bias passed on to the second stage nonparametric estimator for  $\tilde{g}_0(x)$ . Convergence rates



and validity of inference then follow from symmetric analysis of the results in Sections 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.

## 8 Conclusion

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

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

## References

- 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). *Efficient 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.
- 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.
- 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.
- 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.
- Pollard, D. (2001). *A User’s Guide to Measure Theoretic Probability*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.
- Powell, M. J. D. et al. (1981). *Approximation theory and methods*. Cambridge university press.
- Robins, J. M., A. Rotnitzky, and L. P. Zhao (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* 89(427), 846–866.
- 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 approach for doubly-robust  $\ell_1$  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.

## 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,  $\Pr(\Omega_{k,1} \cap \Omega_{k,2}) \geq 1 - o(1)$  where

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

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,  $\Pr(\bigcap_{m=1}^6 \Omega_{k,m}) \geq 1 - o(1)$  where

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

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,  $\Pr(\bigcap_{m=1}^6 \Omega_{k,m}) \geq 1 - o(1)$  where

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

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,  $\Pr(\bigcap_{m=1}^7 \Omega_{k,m}) \geq 1 - o(1)$  where

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

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  $\alpha = p(x)/\|p(x)\|$  and (3.3) follows from (3.2). So it suffices to prove (3.1).

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

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

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

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

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

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

Now, by the definition of  $\Omega$  we have that

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

Second for each  $\delta > 0$

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

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

$$\{\alpha'(a \circ b)\} = \sum_{j=1}^k \alpha_j a_j b_j \leq \|b\|_\infty \sum_{j=1}^k |\alpha_j| a_j = (|\alpha'|a) \|b\|_\infty. \quad (\text{A.2})$$

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

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

in the positive semidefinite sense. Using these two we have

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

By the bounded eigenvalue condition and using the trace operator:

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

Further note,  $|\omega_{ni}| \lesssim \frac{\xi_k}{\sqrt{n}}$ . Using  $(a+b)^2 \leq 2a^2 + 2b^2$ , the right hand side of (A.1) is bounded by

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

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

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

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

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

where  $C$  is a universal constant.

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

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

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

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

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

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

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

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

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

$$\sup_Q N(\mathcal{G}, L^2(Q), \varepsilon \|G\|_{L^2(Q)}) \leq \left( \frac{C \xi_k^L}{\varepsilon} \right)^d. \quad (\text{A.3})$$

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

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

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

**Step 1:** For the first term, we set up for an application of (GK). Equation (A.3) gives us the constants

$A = C\xi_k^L$  and  $V = d_x \vee 2$ . To get  $\sigma^2$  note that for any  $v \in \mathcal{X}$ ,

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

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

We can now apply (GK) to get that

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

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

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

We can now use (VW) to bound

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

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

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

Applying Assumption 5.3 along with the inequality

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

completes the proof.

### Proof of Theorem 5.3

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



### Yurinskii's Coupling Lemma

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

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

for some universal constant  $C_0$ .

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

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

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

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

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

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

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

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

uniformly over  $x \in \mathcal{X}$ . Since  $\|\alpha(x)' \Omega^{1/2}\|$  is bounded from below uniformly over  $x \in \mathcal{X}$  we obtain the first statement of Theorem 5.2 from which the second statement directly follows.

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

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

so that the third statement, (5.9) holds.

### Proof of Theorem 5.4

#### Preliminaries for Proof of Theorem 5.4

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

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

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

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

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

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

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

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

for some (possibly different) absolute constant  $C$ .

We will establish consistent estimation of

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

using

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

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

$$\widehat{Y} := \begin{bmatrix} Y(\widehat{\pi}_1, \widehat{m}_1) \\ \vdots \\ Y(\widehat{\pi}_k, \widehat{m}_k) \end{bmatrix} \quad \text{and} \quad \widehat{Y} := \begin{bmatrix} Y(\widehat{\pi}_1, \widehat{m}_1) \\ \vdots \\ Y(\widehat{\pi}_k, \widehat{m}_k) \end{bmatrix} \quad (\text{A.4})$$

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

Using this, we can decompose

$$\begin{aligned}
\widehat{\Sigma} &= \mathbb{E}_n[\{p^k(x) \circ (\Delta + \epsilon^k)\} \{p^k(x) \circ (\Delta + \epsilon^k)\}] \\
&= \underbrace{\mathbb{E}_n[\{p^k(x) \circ \Delta\} \{p^k(x) \circ \Delta\}']}_{\Sigma_1} + \underbrace{\mathbb{E}_n[\{p^k(x) \circ \epsilon^k\} \{p^k(x) \circ \Delta\}']}_{\Sigma_2} \\
&\quad + \underbrace{\mathbb{E}_n[\{p^k(x) \circ \Delta\} \{p^k(x) \circ \epsilon^k\}']}_{\Sigma_3} + \underbrace{\mathbb{E}_n[\{p^k(x) \circ \epsilon^k\} \{p^k(x) \circ \epsilon^k\}']}_{\Sigma_4}
\end{aligned} \tag{A.5}$$

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

**Lemma A.1** (Pseudo-Variance Estimator Consistency). *Suppose Assumption 5.1 and Assumption 5.3 hold. Further, define  $v_n = \mathbb{E}[\max_{1 \leq i \leq n} |\bar{\epsilon}_k|^2]^{1/2}$ . In addition, assume that  $\bar{R}_{1n} + \bar{R}_{2n} \lesssim (\log k)^{1/2}$ . Then,*

$$\begin{aligned}
\|\widehat{Q} - Q\| &\lesssim_P \sqrt{\frac{\xi_k^2 \log k}{n}} = o(1) \\
\text{and } \|\Sigma_4 - \Sigma\| &\lesssim_P (v_n \vee 1 + \ell_k c_k) \sqrt{\frac{\xi_k^2 \log k}{n}}
\end{aligned}$$

*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  $\dot{\Delta} \equiv \epsilon^k - \epsilon^k = g_0(x) - \widehat{g}(x)$ <sup>1</sup> and decompose

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

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

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

By Theorem 5.2  $|\max_{1 \leq i \leq n} |p^k(x)'(\widehat{\beta}^k - \beta^k)|| \lesssim_P \xi_k^2 (\sqrt{\log k} + \bar{R}_{1n} + \bar{R}_{2n})^2/n$ , by Assumption 5.1 the approximation error is bounded  $\max_{1 \leq i \leq n} |r_{k,i}| \leq \ell_k c_k$ , by Assumption 5.3 and Markov's inequality the errors are bounded  $\max_{1 \leq i \leq n} |\bar{\epsilon}_{k,i}| \lesssim_p v_n^2$ . Finally, by the first part of Lemma A.1  $\|\widehat{Q}\| \lesssim_P \|Q\| \lesssim 1$ . Putting this all together with  $\bar{R}_{1n} + \bar{R}_{2n} \lesssim (\log k)^{1/2}$  and  $\xi_k^2 \log k/n \rightarrow 0$  gives

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

Next, we want to control  $\Sigma_{44} - \Sigma$ . To do this, let  $\eta_1, \dots, \eta_n$  be independent Rademacher random variables

<sup>1</sup>It is useful to recall that  $\epsilon^k = \bar{Y}^k - \widehat{g}(x)$  and  $\epsilon^k = \bar{Y}^k - g_0(x)$

generated independently from the data. Then for  $\eta = (\eta_1, \dots, \eta_n)$

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

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

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

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

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

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

The second result of Lemma A.1 follows from Markov's inequality.  $\square$

Now, we need to take care of the terms

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

where  $\Delta = \hat{Y}^k - \bar{Y}^k$  and  $\epsilon^k = \bar{Y}^k - \hat{g}(x) = \hat{g}(x) - g^k(x) + \epsilon^k$ . 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*

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

*Proof.* From Condition 2, the term  $\Sigma_1$  being negligible immediately follows from Cauchy-Schwarz. Notice

that

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

To see that  $\Sigma_2$  is negligible notice that

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

Applying Assumption 5.3 and Theorem 5.2 gives

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

where the final line is via Condition 2. Showing negligibility of  $\Sigma_3$  follows the same steps.  $\square$

### 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

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

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

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

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

**Lemma B.1** (Nonasymptotic Bounds for Logistic Model). *Suppose that Assumption 4.1 holds with  $\xi_0 > (c_0 + 1)/(c_0 - 1)$  and  $2C_0\nu_0^{-2}s_k\bar{\lambda}_k \leq \eta < 1$ . Then, under the events  $\Omega_{k,1} \cap \Omega_{k,2}$  defined in (B.2), there exists a finite constant  $M_0$  that does not depend on  $k$  such that*

$$\max_{1 \leq j \leq k} D^\dagger(\bar{g}, \hat{g}) \leq M_0 s_k \bar{\lambda}_k^2 \quad \text{and} \quad \max_{1 \leq j \leq k} \|\hat{\gamma}_j - \bar{\gamma}_j\|_1 \leq M_0 s_k \bar{\lambda}_k \quad (\text{B.3})$$

*Proof.* We show that the bound of (B.3) holds for each  $j = 1, \dots, k$ . For any  $\gamma \in \mathbb{R}^d$  define  $\tilde{\ell}_j(\gamma) := \mathbb{E}_n[p_j(X)\{De^{-\gamma'Z} + (1-D)\gamma'Z\}]$ . By optimality of  $\hat{\gamma}_j$  we must have, for any  $u \in (0, 1]$ :

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

Using convexity of the  $\ell_1$  norm  $\|\cdot\|_1$ , this gives after rearrangement

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

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

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

By direct calculation, we have that  $D_{\gamma,j}^\dagger(\hat{\gamma}_j, \bar{\gamma}_j)$  from (4.6) can be expressed

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

Combining the last two displays yields

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

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

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

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

$$D_{\gamma,j}^\dagger(\hat{\gamma}_j, \bar{\gamma}_j) + \lambda_{\gamma,j} \|\hat{\gamma}_j\|_1 \leq \lambda_{\gamma,j} \|\bar{\gamma}_j\|_1 + c_0^{-1} \lambda_{\gamma,j} \|\hat{\gamma}_j - \bar{\gamma}_j\|_1.$$

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

$$\begin{aligned} D_{\gamma,j}^\dagger(\hat{\gamma}_j, \bar{\gamma}_j) + \lambda_{\gamma,j} \left\{ \sum_{l \in \mathcal{S}_{\gamma,j}} |\bar{\gamma}_{j,l}| - \sum_{l \in \mathcal{S}_{\gamma,j}} |\hat{\gamma}_{j,l} - \bar{\gamma}_{j,l}| + \sum_{j \notin \mathcal{S}_{\gamma,j}} |\hat{\gamma}_{j,l} - \bar{\gamma}_{j,l}| \right\} \\ \leq \lambda_{\gamma,j} \left\{ \sum_{l \in \mathcal{S}_{\gamma,j}} |\bar{\gamma}_{j,l}| + c_0^{-1} \sum_{l \in \mathcal{S}_{\gamma,j}} |\hat{\gamma}_{j,l} - \bar{\gamma}_{j,l}| + c_0^{-1} \sum_{j \notin \mathcal{S}_{\gamma,j}} |\hat{\gamma}_{j,l} - \bar{\gamma}_{j,l}| \right\} \end{aligned}$$

Rearrange to get

$$D_{\gamma,j}^\dagger(\hat{\gamma}_j, \bar{\gamma}_j) + (1 - c_0^{-1})\lambda_{\gamma,j} \sum_{l \notin \mathcal{S}_{\gamma,j}} |\hat{\gamma}_{j,l} - \bar{\gamma}_{j,l}| \leq (1 + c_0)^{-1}\lambda_{\gamma,j} \sum_{l \in \mathcal{S}_{\gamma,j}} |\hat{\gamma}_{j,l} - \bar{\gamma}_{j,l}|.$$

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

$$D_{\gamma,j}^\dagger(\hat{\gamma}_j, \bar{\gamma}_j) + (1 - c_0^{-1})\|\hat{\gamma}_j - \bar{\gamma}_j\|_1 \leq 2\lambda_{\gamma,j} \sum_{l \in \mathcal{S}_{\gamma,j}} |\hat{\gamma}_{j,l} - \bar{\gamma}_{j,l}| \quad (\text{L.3})$$

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

$$D_{\gamma,j}^\dagger(\hat{\gamma}_j, \bar{\gamma}_j) \geq \frac{1 - e^{-C_0\|\delta_j\|_1}}{C_0\|\delta_j\|} \left( \delta_j' \tilde{\Sigma}_{\gamma,j} \delta_j \right) \quad (\text{L.4})$$

By (L.3) and  $\xi_0 > (c_0 + 1)/(c_0 - 1)$  we have that  $\sum_{l \notin \mathcal{S}_{\gamma,j}} |\delta_{j,l}| \leq \xi_0 \sum_{l \in \mathcal{S}_{\gamma,j}} |\delta_{j,l}|$ . Applying the empirical compatibility condition from Assumption 4.1 to (L.3) then yields

$$D_{\gamma,j}^\dagger(\hat{\gamma}_j, \bar{\gamma}_j) + (1 - c_0^{-1})\lambda_{\gamma,j}\|\delta_j\|_1 \leq 2\lambda_{\gamma,j}\nu_0^{-1}|\mathcal{S}_{\gamma,j}|^{1/2}(\delta_j' \tilde{\Sigma}_{\gamma,j} \delta_j)^{1/2} \quad (\text{L.5})$$

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

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

Plugging the second bound into (L.5) gives

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

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

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

Combining the last two displays gives

$$D_{\gamma,j}^\dagger(\hat{\gamma}_j, \bar{\gamma}_j) + (1 - c_0^{-1})\lambda_{\gamma,j}\|\hat{\gamma}_j - \bar{\gamma}_j\|_1 \leq 4\lambda_{\gamma,j}^2\nu_0^{-2}(1 - \eta)|\mathcal{S}_{\gamma,j}| \quad (\text{L.6})$$

Applying  $\Omega_{k,2}$  to bound  $\lambda_{\gamma,j} \leq \bar{\lambda}_k$  and noting that  $|\mathcal{S}_{\gamma,j}| \leq s_k$  by definition gives (B.3) with  $M_0 = \frac{4\nu_0^{-1}(1-\eta)}{1-c_0^{-1}}$ .  $\square$

For each  $j$ , consider the matrices,

$$\begin{aligned} \tilde{\Sigma}_{\alpha,j} &:= \mathbb{E}_n[p_j(X)De^{-\tilde{\gamma}_j'Z}(Y - \bar{\alpha}_j'Z)^2ZZ'] \\ \tilde{\Sigma}_{\gamma,j} &:= \mathbb{E}_n[p_j(X)De^{-\tilde{\gamma}_j'Z}ZZ'] \end{aligned} \quad (\text{B.4})$$

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

nonasymptotic bounds under the following additional events:

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

**Lemma B.2** (Nonasymptotic Bounds for Linear Model). *Suppose that Assumptions 4.1 and 4.2 hold,  $\xi_0 > (c_0 + 1)/(c_0 - 1)$ , and  $2C_0\nu_0^{-2}s_k\bar{\lambda}_k \leq \eta < 1$ . In addition, assume there is a constant  $c > 0$  such that  $\lambda_{\alpha,j}/\lambda_{\gamma,j} \geq c$  for all  $j \leq k$ . Then, under the event  $\bigcap_{m=1}^6 \Omega_{k,m}$  there is a constant  $M_1$  that does not depend on  $k$  such that*

$$\max_{1 \leq j \leq k} D_{\alpha,j}^\dagger(\hat{\alpha}_j, \bar{\alpha}_j; \hat{\gamma}_j) \leq M_1 s_k \bar{\lambda}_k^2 \quad \text{and} \quad \max_{1 \leq j \leq k} \|\hat{\alpha}_j - \bar{\alpha}_j\|_1 \leq M_1 s_k \bar{\lambda}_k \tag{B.6}$$

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

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

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

Convexity of the  $\ell_1$  norm  $\|\cdot\|_1$  gives

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

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

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

Rearranging using the form of  $D_{\alpha,j}^\dagger$  in (4.6) yields:

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

*Step 2: Quasi-Score Domination and relating  $\bar{\gamma}_j$  to  $\hat{\gamma}_j$ .* For this step, we will use the fact that we are in the event  $\Omega_{k,1} \cap \Omega_{k,2} \cap \Omega_{k,3} \cap \Omega_{k,5} \cap \Omega_{k,6}$ . Using the expression for  $D_{\gamma,j}^\dagger(\hat{\gamma}_j, \bar{\gamma}_j)$  from (4.6) we find that for some  $u \in (0, 1)$ :

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

where the second step uses the mean value theorem:

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



In the event  $\Omega_{k,1} \cap \Omega_{k,2}$  using the bound in (B.3) and the fact that  $C_0 \nu_0^{-2} s_k \bar{\lambda}_k \leq \eta < 1$  gives us that

$$C_0 \|\hat{\gamma}_j - \bar{\gamma}_j\|_1 \leq C_0 M_0 s_k \bar{\lambda}_k \leq M_0 \eta. \quad (\text{O.3})$$

In the event  $\Omega_{k,1} \cap \Omega_{k,2}$  the bound in (L.6) also gives us that  $D_{\gamma,j}^\dagger(\hat{\gamma}_j, \bar{\gamma}_j) \leq M_0 s_k \lambda_{\gamma,j}^2$ . Combining the above displays then yields

$$M_0 s_k \lambda_{\gamma,j}^2 \geq D_{\gamma,j}^\dagger(\hat{\gamma}_j, \bar{\gamma}_j) \geq e^{M_0 \eta} \mathbb{E}_n[p_j(X) D e^{-\bar{\gamma}_j' Z} \{\hat{\gamma}_j' Z - \bar{\gamma}_j' Z\}^2]. \quad (\text{O.4})$$

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

$$\begin{aligned} D_{\alpha,j}^\dagger(\hat{\alpha}_j, \bar{\alpha}_j; \hat{\gamma}_j) &= \mathbb{E}_n[p_j(X) D e^{-\hat{\gamma}_j' Z} (\hat{\alpha}_j' Z - \bar{\alpha}_j' Z)^2] \\ &= \mathbb{E}_n[p_j(X) D e^{-(\hat{\gamma}_j - \bar{\gamma}_j)' Z} e^{-\bar{\gamma}_j' Z} (\hat{\alpha}_j' Z - \bar{\alpha}_j' Z)^2] \\ &\geq e^{-M_0 \eta} D_{\alpha,j}^\dagger(\hat{\alpha}_j, \bar{\alpha}_j; \bar{\gamma}_j) \end{aligned} \quad (\text{O.5})$$

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

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

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

$$\delta_{1,j} \leq c_0^{-1} \|\hat{\alpha}_j - \bar{\alpha}_j\|_1 \lambda_{\alpha,j} \quad (\text{O.6})$$

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

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

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

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

By Assumption 4.2 and Lemma E.6 we have that  $\mathbb{E}[D\{Y - \bar{\alpha}_j' Z\}^2] \leq D_0^2 + D_1^2$  so that:

$$\mathbb{E}[p_j(X) D e^{-\bar{\gamma}_j' Z} \{Y - \bar{\alpha}_j' Z\}^2 \{\hat{\gamma}_j' Z - \bar{\gamma}_j' Z\}^2] \leq (D_0^2 + D_1^2) \mathbb{E}[p_j(X) D e^{-\bar{\gamma}_j' Z} \{\hat{\gamma}_j' Z - \bar{\gamma}_j' Z\}^2].$$

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

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

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

$$\begin{aligned} \mathbb{E}_n[p_j(X)De^{-\tilde{\gamma}'_j Z}\{Y - \bar{\alpha}'_j Z\}^2\{\hat{\gamma}'_j Z - \bar{\gamma}'_j Z\}^2] &\leq \bar{\lambda}_k \|\hat{\gamma}_j - \bar{\gamma}_j\|_1^2 \\ &\quad + (D_0^2 + D_1^2) \bar{\lambda}_k \|\hat{\gamma}_j - \bar{\gamma}_j\|_1^2 \\ &\quad + (D_0^2 + D_1^2) e^{-M_0 \eta} M_0 s_k \lambda_{\gamma,j}^2 \end{aligned} \quad (\text{O.8})$$

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

$$\begin{aligned} \delta_{2,j} &\leq \{e^{M_0 \eta} (1 + (D_0^2 + D_1^2)^{1/2}) (M_0 \bar{\lambda}_k \lambda_{\gamma,j} s_k)^{1/2} + (D_0^2 + D_1^2) (M_0 s_k \lambda_{\gamma,j}^2)^{1/2}\} D_{\alpha,j}^\dagger(\hat{\alpha}_j, \bar{\alpha}_j; \bar{\gamma}_j)^{1/2} \\ &\leq \tilde{C} \{(\bar{\lambda}_k \lambda_{\gamma,j} s_k)^{1/2} + (s_k \lambda_{\gamma,j})^{1/2}\} D_{\alpha,j}^\dagger(\hat{\alpha}_j, \bar{\alpha}_j; \bar{\gamma}_j)^{1/2} \end{aligned} \quad (\text{O.9})$$

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

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

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

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

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

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

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

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

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

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

$$\xi_2 D_{\alpha,j}^\star \leq \delta_{3,j} (\delta'_j \tilde{\Sigma}_{\gamma,j} \delta_j)^{1/2} \quad (\text{O.13})$$

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

$$D_{\alpha,j}^* \leq (\xi_1 + 1)(c_0 - 1)c_0^{-1}\lambda_{\alpha,j} \sum_{l \in \mathcal{S}_{\alpha,j}} |\delta_{j,l}| \quad (\text{O.14})$$

We deal with these two cases separately. First, if (O.14) holds, then  $\sum_{l \notin \mathcal{S}_{\alpha,j}} |\delta_{j,l}| \leq \xi_1 \sum_{l \in \mathcal{S}_{\alpha,j}} |\delta_{j,l}|$ . We can apply the empirical compatability of Assumption 4.2 to (O.14) to obtain.

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

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

$$e^{-M_0\eta}D_{\alpha,j}^\dagger(\hat{\alpha}_j, \bar{\alpha}_j; \bar{\gamma}_j) + (1 - c_0^{-1})\lambda_{\alpha,j} \|\hat{\alpha}_j - \bar{\alpha}_j\|_1 \leq \tilde{M}s_k\lambda_{\alpha,j}^2 \quad (\text{O.15})$$

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

$$e^{-M_0\eta}D_{\alpha,j}^\dagger(\hat{\alpha}_j, \bar{\alpha}_j; \bar{\gamma}_j) + (1 - c_0^{-1})\lambda_{\alpha,j} \|\hat{\alpha}_j - \bar{\alpha}_j\|_1 \leq \xi_2^{-1}\tilde{C}\{(s_k\bar{\lambda}_k\lambda_{\gamma,j})^{1/2} + (s_k\lambda_{\gamma,j}^2)^{1/2}\}^2 \quad (\text{O.16})$$

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

$$e^{-M_0\eta}D_{\alpha,j}^\dagger(\hat{\alpha}_j, \bar{\alpha}_j; \bar{\gamma}_j) + (1 - c_0^{-1})\lambda_{\alpha,j} \|\hat{\alpha}_j - \bar{\alpha}_j\|_1 \leq \tilde{M}s_k\lambda_{\alpha,j}^2 + \xi_2^{-1}\tilde{C}\{(s_k\bar{\lambda}_k\lambda_{\gamma,j})^{1/2} + (s_k\lambda_{\gamma,j}^2)^{1/2}\}^2 \quad (\text{O.17})$$

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

$$D_{\alpha,j}^\dagger(\hat{\alpha}_j, \bar{\alpha}_j; \bar{\gamma}_j) \leq e^{M_0\eta}\tilde{M}s_k\lambda_{\alpha,j}^2 + e^{M_0\eta}\xi_2^{-1}\tilde{C}\{(s_k\bar{\lambda}_k\lambda_{\gamma,j})^{1/2} + (s_k\lambda_{\gamma,j}^2)^{1/2}\}^2$$

$$\|\hat{\alpha}_j - \bar{\alpha}_j\|_1 \leq (1 - c_0^{-1})^{-1}\tilde{M}s_k\lambda_{\alpha,j} + (1 - c_0^{-1})^{-1}c^{-1}\tilde{C}\{(s_k\bar{\lambda}_k)^{1/2} + (s_k\lambda_{\gamma,j})^{1/2}\}^2$$

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

$$D_{\alpha,j}^\dagger(\hat{\alpha}_j, \bar{\alpha}_j; \bar{\gamma}_j) \leq M_1s_k\bar{\lambda}_k^2$$

$$\|\hat{\alpha}_j - \bar{\alpha}_j\|_1 \leq M_1s_k\bar{\lambda}_k \quad (\text{O.18})$$

for  $M_1 = \max\{e^{M_0\eta}, c^{-1}(1 - c_0^{-1})^{-1}\}(\tilde{M} + 2e^{M_0\eta}\xi_2^{-1}\tilde{C})$ . This completes the result (B.6).  $\square$

Let  $\tilde{\Sigma}_{\alpha,j}^1 := \mathbb{E}_n[p_j(X)De^{-\bar{\gamma}_j'Z}|Y - \bar{\alpha}_j'Z|ZZ']$  and  $\Sigma_{\alpha,j}^1 := \mathbb{E}\tilde{\Sigma}_{\alpha,j}^1$ . Define the event

$$\Omega_{k,7} := \{\|\tilde{\Sigma}_{\alpha,j}^1 - \Sigma_{\alpha,j}^1\|_\infty \leq \bar{\lambda}_k, \forall j \leq k\} \quad (\text{B.7})$$

**Lemma B.3** (Nonasymptotic Bounds for Weighted Means). *Suppose that Assumptions 4.1 and 4.2 hold,  $\xi_0 > (c_0 + 1)/(c_0 - 1)$ , and  $2C_0\nu_0^{-2}s_k\bar{\lambda}_k \leq \eta < 1$ . In addition, assume there is a constant  $c > 0$  such that  $\lambda_{\alpha,j}/\lambda_{\gamma,j} \geq c$  for all  $j \leq k$ . Then, under the event  $\bigcap_{m=1}^7 \Omega_{k,m}$ , there is a constant  $M_2$  that does not depend*

on  $k$  such that

$$\max_{1 \leq j \leq k} |\mathbb{E}_n[p_j(X)Y(\hat{\pi}_j, \hat{m}_j)] - \mathbb{E}_n[p_j(X)Y(\bar{\pi}_j, \bar{m}_j)]| \leq M_2 s_k \bar{\lambda}_k^2 \quad (\text{B.8})$$

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

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

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

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

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

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

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

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

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

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

*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  $u \in (0, 1)$

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

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

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

bounded with

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

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

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

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

$$\mathbb{E}[p_j(X)De^{-\hat{\gamma}'_j Z}|Y - \bar{m}_j(Z)|\{\hat{\gamma}'_j Z - \bar{\gamma}'_j Z\}^2] \leq (D_1^2/D_0 + D_0)\mathbb{E}[p_j(X)De^{-\hat{\gamma}'_j Z}\{\hat{\gamma}'_j Z - \bar{\gamma}'_j Z\}^2].$$

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

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

Putting these all together gives

$$\begin{aligned}\mathbb{E}_n[p_j(X)De^{-\hat{\gamma}'_j Z}|Y - \bar{m}_j(Z)|\{\hat{\gamma}'_j Z - \bar{\gamma}'_j Z\}^2] &\leq \bar{\lambda}_k \|\hat{\gamma}_j - \bar{\gamma}_j\|_1^2 + (D_1^2/D_0 + D_0)\bar{\lambda}_k \|\hat{\gamma}_j - \bar{\gamma}_j\|_1^2 \\ &\quad + (D_1^2/D_0 + D_0)\mathbb{E}_n[p_j(X)De^{-\hat{\gamma}'_j Z}\{\hat{\gamma}'_j Z - \bar{\gamma}'_j Z\}^2]\end{aligned}\tag{M.2}$$

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

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

Plugging these into (M.2) gives

$$\delta_{22,j} \leq e^{M_0\eta} M_0^2 s_k^2 \bar{\lambda}_k^3 + e^{M_0\eta} (D_1^2/D_0 + D_0) M_0^2 s_k^2 \bar{\lambda}_k^3 + (D_1^2/D_0 + D_0) M_0 s_k \bar{\lambda}_k^2\tag{M.3}$$

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

$$\delta_{2,j} \leq M_0 s_k (D_1^2/D_0 + D_0 + 1) \bar{\lambda}_k^2 + e^{M_0\eta} M_0^2 s_k^2 (D_1^2/D_0 + D_0 + 1) \bar{\lambda}_k^3\tag{M.4}$$

*Step 3: Combine Terms.* Putting this together yields

$$\begin{aligned}|\delta_{1,j} + \delta_{2,j} + \delta_{3,j}| &\leq \{M_1 + M_0(D_1^2/D_0 + D_0 + 1)\} s_k \bar{\lambda}_k^2 \\ &\quad + e^{M_0\eta} (D_1^2/D_0 + D_0) M_0^2 s_k^2 \bar{\lambda}_k^3\end{aligned}\tag{M.5}$$

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

$$\begin{aligned}|\delta_{1,j} + \delta_{2,j} + \delta_{3,j}| &\leq \{M_1 + M_0(D_1^2/D_0 + D_0 + 1)\} s_k \bar{\lambda}_k^2 \\ &\quad + e^{M_0\eta} (D_1^2/D_0 + D_0) M_0^2 s_k \bar{\lambda}_k\end{aligned}\tag{M.6}$$

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

□

**Lemma B.4** (Nonasymptotic Bounds for Variance Estimation). *Suppose that Assumptions 4.1 and 4.2 hold,  $\xi_0 > (c_0 + 1)/(c_0 - 1)$ , and  $2C_0\nu_0^{-2}s_k\bar{\lambda}_k \leq \eta < 1$ . In addition, assume there is a constant  $c > 0$  such that  $\lambda_{\alpha,j}/\lambda_{\gamma,j} \geq c$  for all  $j \leq k$ . Then, under the event  $\bigcap_{m=1}^7 \Omega_{k,m}$ , there is a constant  $M_3$  that does not depend on  $k$  such that*

$$\max_{1 \leq j \leq k} \mathbb{E}_n[p_j^2(X)(Y(\hat{\pi}_j, \hat{m}_j) - Y(\bar{\pi}_j, \bar{m}_j))^2] \leq M_3 \xi_{k,\infty}^2 s_k^2 \bar{\lambda}_k^2 \quad (\text{B.9})$$

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

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

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

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

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

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

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

$$\leq \tilde{M}_1 \xi_{k,\infty} s_k \bar{\lambda}_k^2 \quad (\text{V.2})$$

To bound  $\mathbb{E}_n[\tilde{\delta}_{3,j}^2]$  write  $\hat{\pi}_j^{-1}(Z) - \bar{\pi}_j^{-1}(Z) = e^{-\bar{\gamma}_j'Z}\{e^{-\hat{\gamma}_j'Z + \bar{\gamma}_j'Z} - 1\}$  and use the lower bound on  $\bar{g}_j(z)$  from Assumption 4.1:

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

Noting that this last term is exactly  $D_{\alpha,j}^\dagger(\hat{\alpha}_j, \bar{\alpha}_j; \bar{\gamma}_j)$  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

$$\leq \tilde{M}_2 \xi_{k,\infty} s_k \bar{\lambda}_k^2 \quad (\text{V.3})$$

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

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

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

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

The result (B.9) follows by collecting (V.1)-(V.4).  $\square$

## 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  $\Omega_{k,1} \cap \Omega_{k,2}$  described on (B.2) there is a constant  $M_{\gamma,r}$  that does not depend on  $k$  such that:*

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

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

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

So that

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

$\square$

**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  $\bigcap_{m=1}^6 \Omega_{k,m}$ , there is a constant  $M_{\alpha,r}$  that does not depend on  $k$  such that*

$$\max_{1 \leq j \leq k} \mathbb{E}_n[(\widehat{U}_{\alpha,j} - U_{\alpha,j})^2] \leq M_{\alpha,r} \xi_{k,\infty}^2 s_k^2 \bar{\lambda}_k^2 \quad (\text{B.11})$$

*Proof.* Recall that  $\widehat{U}_{\alpha,j} = p_j(X) D e^{-\widehat{\gamma}'_j Z} (Y - \widehat{\alpha}'_j Z)$  and  $U_{\alpha,j} = p_j(X) D e^{-\bar{\gamma}'_j Z} (Y - \bar{\alpha}'_j Z)$ . As an intermediary, define  $\dot{U}_{\gamma,j} = p_j(X) D e^{-\widehat{\gamma}'_j Z} (Y - \bar{\alpha}'_j Z)$ . We will show a bound on the empirical mean square error between

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

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

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

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

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

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

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

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

### 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  $\lambda_{\gamma,j}$  is chosen as described in Section 4.1. Then, for  $n$  sufficiently large, the event  $\Omega_{k,1}$  holds with probability  $1 - \epsilon - \rho_{\gamma,n}$  where*

$$\rho_{\gamma,n} = C \max \left\{ \frac{4kn + 4k}{n^2}, \left( \frac{\tilde{M} \xi_{k,\infty} s_{k,\gamma} \ln^5(d_z n)}{n} \right)^{1/2}, \left( \frac{\tilde{M} \xi_{k,\infty}^4 \ln^7(d_z kn)}{n} \right)^{1/6}, \frac{1}{\ln^2(d_z kn)} \right\}. \quad (\text{B.12})$$

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

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

$$\bar{\lambda}_k = \max\{M_2, M_4, M_5, M_6, M_7\} \xi_{k,\infty} \sqrt{\frac{\ln(d_z n)}{n}} \quad (\text{B.13})$$



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  $k/n \rightarrow 0$ ,  $\Pr(\Omega_{k,2}) = 1 - o(1)$ .

*Proof.* Collecting the logistic nonasymptotic residual bound from Lemma B.5 and the probability bounds from Lemmas B.9-B.12 we find that, we probability at least  $1 - \frac{4k}{n} - \frac{4k}{n^2}$ :

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

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

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

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

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

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

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

Let  $e_1, \dots, e_n$  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

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

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  $B_k = (\xi_{k,\infty} C_0 C_U)^3 \vee 1$ . Further, via (P.2) the residual estimation requirement of with  $\delta_n$  and  $\beta_n$  replaced by  $\delta_{\gamma,n}$  and  $\beta_{\gamma,n}$ .

Let  $\widehat{q}_{\gamma,j}(\alpha)$  be the  $\alpha$  quantile of  $\|\widehat{S}_{\gamma,j}^{e'}\|$  conditional on the data  $Z_i$  and the estimates  $\widehat{Z}_i$ . Theorem E.4 then shows that there is a finite constant depending only on  $c_u$  such that

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

This gives the first claim of Lemma B.7 by construction of  $\lambda_{\gamma,j}$ .

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

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

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

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

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

**Lemma B.8** (Linear Score Domination and Penalty Majorization). *Suppose Assumptions 4.1 and 4.2 holds and that the penalty parameters  $\lambda_{\gamma,j}$  and  $\lambda_{\alpha,j}$  are chosen as described in Section 4.1. Then, for  $n$  sufficiently large, the event  $\Omega_{k,3}$  holds with probability  $1 - \epsilon - \rho_{\alpha,n}$  where:*

$$\rho_{\alpha,n} = C \max \left\{ \frac{4kn + 4k}{n^2}, \left( \frac{\tilde{M} \xi_{k,\infty}^2 s_{k,\alpha}^2 \ln^5(d_z n)}{n} \right)^{1/2}, \left( \frac{\tilde{M} \xi_{k,\infty}^4 \ln^7(d_z kn)}{n} \right)^{1/6}, \frac{1}{\ln^2(d_z kn)} \right\}. \quad (\text{B.14})$$

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

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

$$\bar{\lambda}_k = \max\{M_2, M_4, M_5, M_6, M_7\} \xi_{k,\infty} \sqrt{\frac{\ln(d_z n)}{n}} \quad (\text{B.15})$$

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  $k/n \rightarrow 0$ ,  $\Pr(\Omega_{k,4}) = 1 - o(1)$ .

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

$$\begin{aligned} \delta_{\alpha,n}^2 &= M_{\alpha,r} C_0^2 \xi_{k,\infty}^2 s_k^2 \ln^5(d_z n) / n \\ \beta_{\alpha,n} &= \frac{4}{n} + \frac{4}{n^2} \end{aligned}$$

$\square$

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

$$\bar{\lambda}_k \geq 4\xi_{k,\infty} (D_0^2 + D_0 D_1) C_0^2 \left[ D_0^2 \log(d_z/\epsilon) / n + D_0 D_1 \sqrt{\log(d_z/\epsilon) / n} \right]$$

Then  $\Pr(\Omega_{k,5}) \geq 1 - 2k\epsilon^2$ . In particular, there is a constant  $M_5$  that does not depend on  $k$ , such that if  $\bar{\lambda}_k \geq \xi_{k,\infty} M_5 \sqrt{\log(d_z/\epsilon) / n}$  and  $k\epsilon^2 \rightarrow 0$  as  $n \rightarrow \infty$  then under the conditions of Assumption 4.2,  $\Pr(\Omega_{k,5}) = 1 - o(1)$ .

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

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

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

$$\mathbb{E} \left[ |(\tilde{\Sigma}_{\alpha,j})_{lh} - (\tilde{\Sigma}_{\alpha,j})_{lh}|^k \right] \leq \frac{k!}{2} (2\xi_{k,\infty} C_0^{-2} e^{-B_0} D_0^2)^{k-2} (2\xi_{k,\infty} C_0^2 e^{-B_0} D_0 D_1)^2, \quad k = 2, 3, \dots$$

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

$$\Pr \left( |(\tilde{\Sigma}_{\alpha,j})_{lh} - (\tilde{\Sigma}_{\alpha,j})_{lh}| > 2e^{-B_0} \xi_{k,\infty} C_0^2 D_0^2 t + 2e^{-B_0} \xi_{k,\infty} C_0^2 D_0 D_1 \sqrt{2t} \right) \leq 2\epsilon^2/d_z^2.$$

A union bound completes the argument.  $\square$

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

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

*then  $\Pr(\Omega_{k,6}) \leq 1 - 2k\epsilon^2$ . In particular, there is a constant  $M_6$  that does not depend on  $k$ , such that if  $\bar{\lambda}_k \geq \xi_{k,\infty} M_6 \sqrt{\log(d_z/\epsilon)/n}$  and  $k\epsilon^2 \rightarrow 0$  as  $n \rightarrow \infty$  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  $|(\tilde{\Sigma}_{\gamma,j})_{lh}| = |p_j(X) D e^{-\tilde{\gamma}'_j Z} Z_l Z_h| \leq \xi_{k,\infty} C_0^2 e^{-B_0}$  so that  $(\tilde{\Sigma}_{\gamma,j})_{lh} - (\Sigma_{\gamma,j})_{lh}$  is mean zero and bounded in absolute values by  $2\xi_{k,\infty} C_0^2 e^{-B_0}$ . Applying Lemma E.3 with  $\bar{\lambda}_k \geq 4\xi_{k,\infty} C_0^2 e^{-B_0} \sqrt{\log(d_z/\epsilon)/n}$  yields:

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

A union bound completes the argument.  $\square$

**Lemma B.11** (Probabilistic Bound on  $\Omega_{k,7}$ ). *Let  $\tilde{\Sigma}_{\alpha,j}^1$  and  $\Sigma_{\alpha,j}^1 = \mathbb{E}\tilde{\Sigma}_{\alpha,j}^1$  be as in (B.7). Under Assumptions 4.1 and 4.2 if*

$$\bar{\lambda}_k \geq \xi_{k,\infty} 4(D_0^2 + D_1^2)^{1/2} e^{-B_0} C_0^2 \sqrt{\log(d_z/\epsilon)/n},$$

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

*Proof.* We deal with each  $j$  term separately. The variables  $p_j(X)e^{-\tilde{\gamma}'_j Z} |Y - \bar{m}_j(Z)| Z_l Z_h$  are uniformly sub-gaussian conditional on  $Z$  because  $|p_j(X)e^{-\tilde{\gamma}'_j Z} Z_l Z_h| \leq \xi_{k,\infty} e^{-B_0} C_0^2$  and  $D|Y - \bar{m}_j(Z)|$  is uniformly sub-gaussian. Applying Lemma E.4 for  $\bar{\lambda}_k \geq e^{-B_0} \xi_{k,\infty} C_0^2 \sqrt{8(D_0^2 + D_1^2)^2 \log(d_z/\epsilon)/n}$  yields

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

A union bound completes the argument.  $\square$

## 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*

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

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

*Proof.* Let us recall that

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

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

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

$$\bar{\lambda}_k \geq \xi_{k,\infty}(e^{-B_0} C_0) \sqrt{8(D_0^2 + D_1^2)} \sqrt{\ln(d_z/\epsilon)/n}.$$

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

*Proof.* Notice  $S_{j,l} = p_j(X)De^{-\bar{\gamma}_j' Z}\{Y - \bar{m}_j(Z)\}Z_l$  for  $l = 1, \dots, p$ . By optimality of  $\bar{\alpha}_j$ ,  $S_{j,l}$  is mean zero. Under Assumption 4.1,  $|S_{j,l}| \leq e^{-B_0} C_0 |D\{Y - \bar{m}_j(Z)\}|$  so by Assumption 4.2 the variables  $S_{j,l}$  are uniformly sub-gaussian conditional on  $Z$  in the following sense:

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

for  $\tilde{D}_0 = \xi_{k,\infty} C_0 D_0 e^{-B_0}$  and  $\tilde{D}_1 = \xi_{k,\infty} C_0 D_1 e^{-B_0}$ . Apply Lemma E.4 for  $\bar{\lambda}_k$  defined above in the statement of Lemma B.13 and union bound to obtain the result.  $\square$

## 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  $\xi_k^2 \log k/n \rightarrow 0$  and  $c_k \rightarrow 0$ . Then if either the propensity score or outcome regression model are correctly specified:*

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

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

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

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

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

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

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

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

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

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

Further,

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

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

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

Under Condition 1 the first term is  $o_p(\sqrt{k/n})$ . By equation (A.48) in Belloni et al. (2015) the third term is bounded in probability by  $c_k$ . For the second term apply the third condition in Assumption 5.1 to see

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

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

The following lemma is a building block for asymptotic pointwise normality. It establishes conditions under which the coefficient estimator  $\widehat{\beta}$  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  $\xi_k^2 \log k/n \rightarrow 0$ . Then for any  $\alpha \in S^{k-1}$ ,*

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

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

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

Moreover,

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

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

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

*Proof.* Decompose as before,

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

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

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

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

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

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

so by Chebyshev's inequality,

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

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

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

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

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

which gives (C.4)-(C.5).  $\square$

The following lemma shows that, after adding Assumption 5.3 the linearization of our coefficient estimator  $\hat{\beta}^k$  established in Lemma C.2 holds uniformly over all points  $x \in \mathcal{X}$ . That is to say the error from linearization

is bounded in probability uniformly over all  $x \in \mathcal{X}$ . It will form an important building block in uniform consistency and strong approximation results presented in Theorems 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 or outcome regression model is correctly specified:*

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

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

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

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

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

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

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

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

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

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

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

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

**Step 1:** Conditional on the data let

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

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

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

(Dudley, 1967),

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

where

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

by (A.2). Now, for any  $x \in \mathcal{X}$ ,

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

So, for some  $C > 0$ ,

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

This gives us that

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

By Assumption 5.3 we have that  $\mathbb{E}[\max_{1 \leq i \leq n} |\bar{\epsilon}_{k,i}| \mid X] \lesssim_P n^{1/m}$  where  $X = (x_1, \dots, x_n)$ . In addition  $\xi_k^{2m/(m-2)} \log k/n \lesssim 1$  for  $m > 2$  gives that  $\xi_k^2 / \log k/n \rightarrow 0$ . So,  $\|\hat{Q}^{-1} - I\| \lesssim_P (\xi_k^2 \log k/n)^{1/2}$  and  $\|\hat{Q}^{-1}\| \lesssim_P 1$ . Combining this all with  $\log \xi_k^L \lesssim \log k$  implies

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

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

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

**Step 2:** Now simply report the results on approximation error from Belloni et al. (2015). Since the



approximation error is the same for all signals  $Y(\bar{\pi}_k, \bar{m}_k)$ , there is no Hadamard product to deal with.

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

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

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

**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  $\xi'_n$  obeying  $1 \lesssim \xi'_n \lesssim \|p^k(x)\|$  uniformly for all  $x \in \mathcal{X}$  such that  $\|p^k(x) - p^k(x')\|/\xi'_n \leq L_n \|x - x'\|$ , where  $\log L_n \lesssim \log n$ . Let  $N_k^b$  be a bootstrap draw from  $N(0, I_k)$  and  $P^*$  be the distribution conditional on the observed data  $\{Y_i, D_i, Z_i\}_{i=1}^n$ . Then the following approximation holds uniformly in  $\ell^\infty(\mathcal{X})$ :*

$$\frac{p^k(x)' \hat{\Omega}^{1/2}}{\hat{\Omega}^{1/2} p^k(x)} N_k^b =^d \frac{p^k(x)' \Omega^{1/2}}{\|\Omega^{1/2} p^k(x)\|} + o_{P^*}(\log^{-1} N) \quad (\text{C.14})$$

*Proof.* See Theorem 3.4 in Semanova and Chernozhukov (2021).  $\square$

## 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  $c_{\gamma,j}$  and  $c_{\alpha,j}$  from a (finite) set via cross validation. For each  $j$  we will assume that

$$c_{\gamma,j}, c_{\alpha,j} \in \Lambda_n \subseteq [\underline{c}_n, \bar{c}_n] \quad (\text{D.1})$$

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

### D.1 Theory Overview

Let  $M_5, M_6, M_7, M_8^2, M_9^2$  be constants that do not depend on  $k$  as in Lemmas B.9–B.13. Whenever

$$\underline{c}_n \sqrt{\frac{\ln^3(d_z n)}{n}} \geq \xi_{k,\infty} \max \{M_5, M_6, M_7, M_8^2, M_9^2\} \sqrt{\frac{\ln(d_z n)}{n}}, \quad (\text{D.2})$$

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  $1 - 10k/n^2$  for the  $2k$  pilot penalty parameters chosen with any values  $c_{\gamma,j}, c_{\alpha,j} \in \Lambda_n$  and

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

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

$$\begin{aligned} \max_{1 \leq j \leq k} D_{\gamma,j}^\dagger(\hat{\gamma}_j, \bar{\gamma}_j) &\leq M_0 \frac{s_k \bar{c}_n^2 \ln^3(d_z n)}{n} \quad \text{and} \quad \max_{1 \leq j \leq k} \|\hat{\gamma}_j - \bar{\gamma}_j\|_1 \leq M_0 s_k \bar{c}_n \sqrt{\frac{\ln^3(d_z n)}{n}} \\ \max_{1 \leq j \leq k} D_{\alpha,j}^\dagger(\hat{\alpha}_j, \bar{\alpha}_j; \bar{\gamma}_j) &\leq M_1 \frac{s_k \bar{c}_n^2 \ln^3(d_z n)}{n} \quad \text{and} \quad \max_{1 \leq j \leq l} \|\hat{\alpha}_j - \bar{\alpha}_j\|_1 \leq M_1 s_k \bar{c}_n \sqrt{\frac{\ln^3(d_z n)}{n}} \end{aligned}$$

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

$$\max_{1 \leq j \leq k} |\mathbb{E}_n[p_j(X)(Y(\hat{\pi}_j, \hat{m}_j) - Y(\bar{\pi}_j, \bar{m}_j))]| \leq M_2 \frac{\bar{c}_n^2 s_k \ln^3(d_z n)}{n} \quad (\text{D.3})$$

$$\max_{1 \leq j \leq k} |\mathbb{E}_n[p_j^2(X)(Y(\hat{\pi}_j, \hat{m}_j) - Y(\bar{\pi}_j, \bar{m}_j))^2]| \leq M_3 \frac{\xi_{k,\infty}^2 \bar{c}_n^2 s_k^2 \ln^3(d_z n)}{n} \quad (\text{D.4})$$

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

$$\frac{k|\Lambda_n|}{n^2} \rightarrow 0, \quad \frac{\underline{c}_n^{-1} \xi_{k,\infty}}{\ln(d_z n)} \rightarrow 0 \quad \text{and} \quad \frac{\bar{c}_n^2 s_k k^{1/2} \ln^3(d_z n)}{\sqrt{n}} \rightarrow 0. \quad (\text{D.5})$$

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

$$\frac{\xi_{k,\infty}^2 \bar{c}_n^2 s_k k^2 \ln^3(d_z n)}{n^{(m-1)/m}} \rightarrow 0 \quad (\text{D.6})$$

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  $\|Z\|_\infty$ . We therefore recommend setting

$$\begin{aligned} \underline{c}_n &= \frac{1}{2 \log^{1/2}(d_z n)} \max_{1 \leq i \leq n} \|p^k(X_i)\|_\infty \max_{1 \leq i \leq n} \|Z_i\|_\infty \\ \bar{c}_n &= \frac{3 \log^{1/2}(d_z n)}{2} \max_{1 \leq i \leq n} \|p^k(X_i)\|_\infty \max_{1 \leq i \leq n} \|Z_i\|_\infty \end{aligned}$$

and letting  $\Lambda_n$  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  $c_\alpha, c_\gamma$  and designate a fold to hold out.
3. Estimate nuisance model parameters using  $\lambda_{\gamma,j}^{\text{pilot}}$  and  $\lambda_{\alpha,j}^{\text{pilot}}$  on the remaining folds.

4. Evaluate the resulting models on held out fold using non-penalized loss functions.
5. Repeat  $K$  times and record average loss over all folds.
6. Choose values of  $c_{\gamma,j}$  and  $c_{\alpha,j}$  with the lowest average loss.

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

## E High Dimensional Probability Results

### E.1 High Dimensional Central Limit and Bootstrap Theorems

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

*Proof.* See [Chetverikov and Sørensen \(2021\)](#), Lemma D.2. □

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

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

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

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

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

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

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

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

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

*Proof.* See [Chernozhukov et al. \(2017\)](#), Proposition 2.1. □

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

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

**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  $\{\beta_n\}_{\mathbb{N}}$  and  $\{\delta_n\}_{\mathbb{N}}$  be sequences in  $\mathbb{R}_{++}$  converging to zero such that*

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

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

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

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

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

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

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 to 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  $j = 1, \dots, k$  define  $q_{M,j}^N : \mathbb{R} \rightarrow \bar{\mathbb{R}}$  as the (extended) quantile function of  $\|N_{M,j}\|_\infty$ ,

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

Define  $q_{M,j}^N(\epsilon) = +\infty$  if  $\epsilon \geq 1$  and  $-\infty$  if  $\epsilon \leq 0$  so that  $q_{M,j}^N$  is always monotone (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,  $U = (U'_1, \dots, U'_k)' \in \mathbb{R}^{d_1, \dots, d_k}$  where  $d_1 + \dots + d_k = p$ . For each  $j = 1, \dots, k$  let  $q_j$  denote the quantile function of  $\|U_j\|_\infty$ . Then for any  $j = 1, \dots, k$ ,*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

The second chain of inequalities then follows on the event  $2\tilde{\rho}(\widehat{S}_n^e, N_\Sigma) \leq \rho'_n$ , which by Theorem E.2 happens

with probability at least  $1 - \beta_n - 1/\ln^2(pn)$ .  $\square$

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

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

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

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

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

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

Combining these two inequalities gives the result.  $\square$

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

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

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

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

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

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

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

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

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

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

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

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

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  $c_0^2 \mathbb{E}[\exp(Y^2/c_0^2) - 1] \leq c_1^2$ . Then

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

*Proof.* Using the fact that  $e^{x^2} > |x|$  gives

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

□