

# Mean-squared-error Calculations for Average Treatment Effects

Guido W. Imbens  
UC Berkeley and NBER

Whitney Newey  
MIT

Geert Ridder  
USC

First Version: December 2003,  
Current Version: April 19, 2005

## Abstract

This paper develops

**JEL Classification:** C14, C20.

**Keywords:** *Nonparametric Estimation*

# 1 Introduction

Recently a number of estimators have been proposed for average treatment effects under the assumption of unconfoundedness or selection on observables. Many of these estimators require nonparametric estimation of an unknown function, either the regression function or the propensity score. Typically results are presented concerning the rates at which the smoothing parameters go to their limiting values, without specific recommendations regarding their values (Hahn, 1998; Hirano, Imbens and Ridder, 2000; Heckman, Ichimura and Todd, 1998; Rotnitzky and Robins, 1995; Robins, Rotnitzky and Zhao, 1995)).

In this paper we make two contributions. First, we propose a new estimator. Our estimator is a modification of an estimator introduced in an influential paper by Hahn (1998). Like Hahn, our estimator relies on consistent estimation of the two regression functions followed by averaging their difference over the empirical distribution of the covariates. Our estimator differs from Hahn's in that it directly estimates these two regression functions whereas Hahn first estimates the propensity score and the two conditional expectations of the product of the outcome and the indicators for being in the control and treatment group and then combines these to get estimates of the two regression functions. Thus our estimator completely avoids the need to estimate the propensity score<sup>1</sup>. Our second and most important contribution is that we are explicit about the choice of smoothing parameters. In our series estimation setting the smoothing parameter is the number of terms in the series. We provide a criterion for choosing this number of terms and show its asymptotic optimality in terms of expected-mean-squared-error. This criterion is related to, but differ from from the standard one for nonparametric estimation as in Li (1987) and Andrews (1991) in that it focuses explicitly on optimal estimation of the average treatment effect rather than on optimal estimation of the entire unknown function.

In the next section we discuss the basic set up and introduce the new estimator. In Section 3 we analyze the asymptotic properties of this estimator. In Section 4 we propose a method for choosing the number of terms in the series.

## 2 The Basic Framework

The basic framework is standard in this literature (e.g Rosenbaum and Rubin, 1983; Hahn, 1998; Heckman, Ichimura and Todd, 1998; Hirano, Imbens and Ridder, 2003) We have a random sample of size  $N$  from a large population. For each unit  $i$  in the sample, let  $W_i$  indicate whether the treatment of interest was received, with  $W_i = 1$  if unit  $i$  receives the treatment of interest, and  $W_i = 0$  if unit  $i$  receives the control treatment. Using the potential outcome notation popularized by Rubin (1974), let  $Y_i(0)$  denote the outcome for unit  $i$  under control and  $Y_i(1)$

---

<sup>1</sup>Independently Chen, Hong, and Tarozzi (2004) have established the efficiency of their CEP-GMM estimator that is similar to our new estimator.

the outcome under treatment. We observe  $W_i$  and  $Y_i$ , where

$$Y_i \equiv Y_i(W_i) = W_i \cdot Y_i(1) + (1 - W_i) \cdot Y_i(0).$$

In addition, we observe a vector of pre-treatment variables, or covariates, denoted by  $X_i$ . We shall focus on the population average treatment effect:

$$\tau \equiv E[Y(1) - Y(0)].$$

Similar results can be obtained for the average effect for the treated:

$$\tau_t \equiv E[Y(1) - Y(0)|W = 1].$$

The central problem of evaluation research (e.g., Holland, 1986) is that for unit  $i$  we observe  $Y_i(0)$  or  $Y_i(1)$ , but never both. Without further restrictions, the treatment effects are not consistently estimable. To solve the identification problem, we maintain throughout the paper the unconfoundedness assumption (Rubin, 1978; Rosenbaum and Rubin, 1983), which asserts that conditional on the pre-treatment variables, the treatment indicator is independent of the potential outcomes. This assumption is closely related to “selection on observables” assumptions (e.g., Barnow, Cain and Goldberger, 1980; Heckman and Robb, 1984). Formally:

**Assumption 2.1** (UNCONFOUNDEDNESS)

$$W \perp (Y(0), Y(1)) \mid X. \tag{2.1}$$

Let the propensity score be the probability of selection into the treatment group:

$$e(x) \equiv \Pr(W = 1|X = x) = \mathbb{E}[W|X = x], \tag{2.2}$$

**Assumption 2.2** (OVERLAP)

*The propensity score is bounded away from zero and one.*

Define the average effect conditional on pre-treatment variables:

$$\tau(x) \equiv \mathbb{E}[Y(1) - Y(0)|X = x]$$

Note that  $\tau(x)$  is estimable under the unconfoundedness assumption, because

$$\begin{aligned} \mathbb{E}[Y(1) - Y(0)|X = x] &= \mathbb{E}[Y(1)|W = 1, X = x] - \mathbb{E}[Y(0)|W = 0, X = x] \\ &= \mathbb{E}[Y|W = 1, X = x] - \mathbb{E}[Y|W = 0, X = x]. \end{aligned}$$

The population average treatment effect can be obtained by averaging the  $\tau(x)$  over the distribution of  $X$ :

$$\tau = \mathbb{E}[\tau(X)],$$

and therefore the average treatment effect is identified.

### 3 Efficient Estimation

In this section we review two efficient estimators previously proposed in the literature. We then discuss two new estimators, which will facilitate the mean-squared error calculations.

#### 3.1 The Hahn Estimator

Hahn (1998) studies the same model as in the current paper. He calculates the efficiency bound, and proposes an efficient estimator. His estimator also imputes the potential outcomes given covariates, followed by averaging the difference in the estimated regression functions. The difference with our estimator is that Hahn first estimates nonparametrically the three conditional expectations  $\mathbb{E}[Y \cdot W|X]$ ,  $\mathbb{E}[Y \cdot (1 - W)|X]$ , and  $e(X) = \mathbb{E}[W|X]$ . and then uses these conditional expectations to estimate the two regression function  $\mu_0(x) = \mathbb{E}[Y(0)|X = x]$  and  $\mu_1(x) = \mathbb{E}[Y(1)|X = x]$  as

$$\hat{\mu}_0(x) = \frac{\hat{\mathbb{E}}[Y(1 - W)|X = x]}{1 - \hat{e}(X)}, \quad \text{and} \quad \hat{\mu}_1(x) = \frac{\hat{\mathbb{E}}[YW|X = x]}{\hat{e}(X)}.$$

The average treatment effect is then estimated as

$$\hat{\tau}_h = \frac{1}{N} \sum_{i=1}^N (\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)).$$

Hahn shows that under regularity conditions this estimator is consistent, asymptotically normally distributed, and that it reaches the semiparametric efficiency bound.

#### 3.2 The Hirano-Imbens-Ridder Estimator

Hirano, Imbens and Ridder (2003) also study the same set up. They propose using a weighting estimator with the weights based on the estimated propensity score:

$$\hat{\tau}_{hir,1} = \frac{1}{N} \sum_{i=1}^N Y_i \cdot \left( \frac{W_i}{\hat{e}(X_i)} - \frac{1 - W_i}{1 - \hat{e}(X_i)} \right).$$

They show that under regularity conditions, and with  $\hat{e}(x)$  a nonparametric estimator for the propensity score, this estimator is consistent, asymptotically normally distributed and efficient.

It will be useful to consider a slight modification of this estimator. Consider the weights for the treated observations,  $1/\hat{e}(X_i)$ . Summing up over all treated observations and dividing by  $N$  we get  $\sum_i (W_i/\hat{e}(X_i))/N$ . This is not necessarily equal to one. We may therefore wish to modify the weights to ensure they add up to one for the treated and control units. This leads to the following estimator:

$$\hat{\tau}_{hir,2} = \sum_{i=1}^N Y_i \cdot \left( \frac{W_i}{\hat{e}(X_i)} - \frac{1 - W_i}{1 - \hat{e}(X_i)} \right) \bigg/ \sum_{i=1}^N \left( \frac{W_i}{\hat{e}(X_i)} - \frac{1 - W_i}{1 - \hat{e}(X_i)} \right).$$

### 3.3 A New Estimator

The new estimator relies on estimating the unknown regression functions  $\mu_1(x)$  and  $\mu_0(x)$  through nonparametric regression of  $Y$  on  $X$  for the two subpopulations indexed by the treatment status. It is a modification of Hahn's estimator. Like Hahn's estimator it estimates the average treatment effect as

$$\hat{\tau}_{inr} = \frac{1}{N} \sum_{i=1}^N (\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)).$$

The difference with Hahn's estimator is that the two regression functions are estimated directly without first estimating the propensity score and the two conditional expectations  $\mathbb{E}[Y \cdot W|X]$  and  $\mathbb{E}[Y \cdot (1 - W)|X]$ . This implies that we only need to estimate two regression functions nonparametrically rather than three and this will make the optimal choice of smoothing parameters easier. Note that this estimator still requires more unknown regression functions to be estimated than the HIR estimator which requires only estimation of the propensity score, but since the new estimator relies on estimation of different functions, it is not clear which is to be preferred in practice.

To provide additional insight into the structure of the problem, we introduce one final estimator, which combines features of the Hirano-Imbens-Ridder estimators and the new estimator:

$$\hat{\tau}_{mod} = \sum_{i=1}^N \left( \frac{W_i \cdot \hat{\mu}_1(X_i)}{\hat{e}(X_i)} - \frac{(1 - W_i) \cdot \hat{\mu}_0(X_i)}{1 - \hat{e}(X_i)} \right) / \sum_{i=1}^N \left( \frac{W_i}{\hat{e}(X_i)} - \frac{1 - W_i}{1 - \hat{e}(X_i)} \right).$$

In order to implement these estimators we need estimators for the two regression functions. Following Newey (1995), we use series estimators for the two regression functions  $\mu_w(x)$ , with  $K$  terms. As the basis we use power series. Let  $\lambda = (\lambda_1, \dots, \lambda_k)$  be a multi-index of dimension  $k$ , that is, a  $k$ -dimensional vector of non-negative integers, with  $|\lambda| = \sum_{i=1}^k \lambda_i$ , and let  $x^\lambda = x_1^{\lambda_1} \dots x_k^{\lambda_k}$ . Consider a series  $\{\lambda(r)\}_{r=1}^\infty$  containing all distinct such vectors such that  $|\lambda(r)|$  is nondecreasing. Let  $p_r(x) = x^{\lambda(r)}$ , where  $p^K(x) = (p_1(x), \dots, p_K(x))'$ . The nonparametric series estimator of the regression function  $\mu_w(x)$ , given  $K$  terms in the series, is given by:

$$\hat{\mu}_w(x) = p^K(x)' \left( \sum_{W_i=w} p^K(X_i) p^K(X_i)' \right)^- \sum_{W_i=w} p^K(X_i) Y_i,$$

where  $A^-$  denotes a generalized inverse of  $A$ . Given the two estimated regression functions we estimate the average treatment effect as

$$\hat{\tau} = \frac{1}{N} \sum_{i=1}^N (\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)).$$

For the propensity score we use the series logit estimator (Hirano, Imbens and Rubin, 2003) Let  $L(z) = \exp(z)/(1 + \exp(z))$  be the logistic cdf. The series logit estimator of the population

propensity score  $e(x)$  is  $\hat{e}_M(x) = L(p^M(x)' \hat{\pi}_M)$ , (for simplicity we use the same series  $p^M(x)$ , although this is not essential) where

$$\hat{\pi}_M = \arg \max_{\pi} L_{N,M}(\pi), \quad (3.3)$$

for

$$L_{N,M}(\pi) = \sum_{i=1}^N (W_i \cdot \ln L(p^M(X_i)' \pi) + (1 - W_i) \cdot \ln(1 - L(p^M(X_i)' \pi))). \quad (3.4)$$

### 3.4 First Order Equivalence of Estimators

We make the following assumptions.

**Assumption 3.1** (DISTRIBUTION OF COVARIATES)

$X \in \mathbb{X} \subset \mathbb{R}^d$ , where  $\mathbb{X}$  is the Cartesian product of intervals  $[x_{jL}, x_{jU}]$ ,  $j = 1, \dots, d$ , with  $x_{jL} < x_{jU}$ . The density of  $X$  is bounded away from zero on  $\mathbb{X}$ .

**Assumption 3.2** (PROPENSITY SCORE)

- (i) The propensity score is bounded away from zero and one.
- (ii) The propensity score is  $s$  times continuously differentiable.

**Assumption 3.3** (CONDITIONAL OUTCOME DISTRIBUTIONS)

- (i) The two regression functions  $\mu_w(x)$  are  $t$  times continuously differentiable.
- (ii) the conditional variance of  $Y_i(w)$  given  $X_i = x$  is bounded by  $\sigma_w^2$ .

**Assumption 3.4** (RATES FOR SERIES ESTIMATORS)

- (i)  $K = N^\nu$ , with  $0 < \nu < 1$ ,
- (ii)  $L = N^\xi$ , with  $0 < \xi < 1$ .

The properties of the estimators will follow from the following lemma:

**Theorem 3.1** (ASYMPTOTIC EQUIVALENCE OF  $\hat{\beta}_{inr}$ ,  $\hat{\beta}_{mod}$  AND  $\hat{\beta}_{hir}$ )

Suppose Assumptions 3.1-3.4 hold. Then

(i),

$$\sqrt{N} \cdot (\hat{\tau}_{inr} - \hat{\tau}_{mod}) = o_p(1),$$

(ii),

$$\sqrt{N} \cdot (\hat{\tau}_{mod} - \hat{\tau}_{hir,1}) = o_p(1).$$

(iii),

$$\sqrt{N} \cdot (\hat{\tau}_{hir,1} - \hat{\tau}_{hir,2}) = o_p(1).$$

and (iv),

$$\sqrt{N} \cdot (\hat{\tau}_h - \hat{\tau}_{hir,1}) = o_p(1).$$

**Proof:** See Appendix.

## 4 A feasible MSE criterion

All three estimators contain a function or functions that are estimated nonparametrically. For the Hahn estimator we need nonparametric estimates of the propensity score and the conditional expectation of the product of the outcome and the treatment indicator and of the outcome and the control indicator given the covariates. For the HIR weighting estimator an estimator of the propensity score is needed, and for the imputation estimator introduced in section 3.3 we need estimators for the conditional means in the treatment and control populations. Both Hahn and HIR use series estimators for either the propensity score or the conditional expectations. That leaves the question how to select the order of the series. For a meaningful comparison of the performance of these asymptotically efficient estimators such a selection rule is essential.

Despite its practical importance there has been little work on the selection of the nonparametric estimators. The only paper that we are aware of is Ichimura and Linton (2003) who consider bandwidth selection if the propensity score in the weighting estimator is estimated by a kernel nonparametric regression estimator. The current practice in propensity score matching, which is a nonparametric estimator that is different from the estimators considered in section 3, is that the propensity score is usually selected using the balancing score property

$$W \perp X | e(X)$$

In practice this is implemented by stratifying the sample on the propensity score and testing whether the means of the covariates are the same for treatment and controls (see e.g. Dehejia and Wahba (2000)). This method of selecting the model for the propensity score focuses exclusively on the bias in the estimation of the treatment effect. This could lead to 'overspecification' of the propensity score and inflation of the variance of the estimator of the treatment effect.

We consider both the bias and the variance associated with a choice of the nonparametric function in the treatment effect estimator. Initially we consider the missing data problem in which we observe  $X$  for all observations, but  $Y$  only if an observation indicator  $D = 1$ . If  $D = 0$  we observe  $X$  but not  $Y$ . We refer to this type of data as one-sample data. Alternatively, we could have two samples. The first sample is from the marginal population distribution of  $X$ . The second sample is a random sample from the subpopulation  $Y, X | D = 1$ . This type of data is referred as the two-sample data. We develop the theory for two-sample data, but it turns out that the results are identical for the one-sample case.

Throughout we assume that the  $Y$  is Missing At Random (MAR), i.e.

$$Y \perp D|X$$

The notation in this section will differ from that used in section 3 to reflect that we consider the missing data instead of the treatment effect problem. In particular, we use  $p(x)$  for the propensity score, and  $r_{kK}(x)$  for the orthonormal polynomials that are used in the estimation of  $\mu(x)$ .

#### 4.1 The MSE and its estimator

As in Li (1987) we consider a population in which the joint distribution of  $Y, X$  is such that

$$Y = \mu(X) + U$$

with  $E(U|X) = 0$  and  $\text{Var}(U|X) = \sigma^2$ . Andrews (1991) has generalized Li's results to the heteroskedastic case and we could do the same. To concentrate on essentials first we maintain the assumption that  $U$  is homoskedastic.

Initially we assume that we have two independent samples. Sample 1 is a sample of size  $N_1$  from the marginal distribution of  $X$ . We denote this sample by  $\tilde{X}_i, i = 1, \dots, N_1$ . Sample 2 is a random sample of size  $N$  from the distribution of  $Y, X|D = 1$ . We denote this sample by  $Y_i, X_i, i = 1, \dots, N$ . The imputation estimator for the parameter of interest  $\mu_Y = E(Y)$  is

$$\hat{\tau}_{INR} = \hat{\mu}_{YK} = \frac{1}{N_1} \sum_{i=1}^{N_1} \hat{\mu}_K(\tilde{X}_i)$$

with

$$\hat{\mu}_K(x) = r_K(x)'(R_K' R_K)^{-1} R_K' y$$

The subscript  $K$  is the number of base functions used in the estimation of  $\mu(x)$ , and we use the notation

$$y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_N \end{pmatrix} \quad \mu = \begin{pmatrix} \mu(X_1) \\ \vdots \\ \mu(X_N) \end{pmatrix} \quad u = \begin{pmatrix} U_1 \\ \vdots \\ U_N \end{pmatrix}$$

and

$$r_K(x) = \begin{pmatrix} r_{1K}(x) \\ \vdots \\ r_{KK}(x) \end{pmatrix} \quad R_K = \begin{pmatrix} r_K(X_1)' \\ \vdots \\ r_K(X_N)' \end{pmatrix} \quad \tilde{R}_K = \begin{pmatrix} r_K(\tilde{X}_1)' \\ \vdots \\ r_K(\tilde{X}_{N_1})' \end{pmatrix}$$

The parameter of interest is  $\mu_Y = E(Y)$ . The MSE is obtained from

$$\sqrt{N}(\hat{\mu}_{YK} - \mu_Y) = \frac{\sqrt{N}}{N_1} \sum_{i=1}^{N_1} \left( \hat{\mu}_K(\tilde{X}_i) - E_{Y|X} \left[ \hat{\mu}_K(\tilde{X}_i) \right] \right) + \frac{\sqrt{N}}{N_1} \sum_{i=1}^{N_1} \left( E_{Y|X} \left[ \hat{\mu}_K(\tilde{X}_i) \right] - \mu(\tilde{X}_i) \right) +$$



(4.5)

$$+\frac{\sqrt{N}}{N_1} \sum_{i=1}^{N_1} \left( \mu(\tilde{X}_i) - \mu_Y \right)$$

We can treat  $X_i$  and  $\tilde{X}_i$  as constants. Therefore, it is reasonable to redefine the parameter of interest as

$$\mu_Y = \frac{1}{N_1} \sum_{i=1}^{N_1} \mu(\tilde{X}_i)$$

If we do so, the final term in the comparison can be omitted, and we need only consider the first two terms. This parameter is the missing data analogue of the sample treatment effect of Imbens (2003).

We now consider the first two terms separately. The first corresponds to the variance and the second to the bias term in the MSE. The first term can be written as

$$V = \frac{\sqrt{N}}{N_1} \sum_{i=1}^{N_1} \left( \hat{\mu}_K(\tilde{X}_i) - \mathbb{E}_{Y|X} \left[ \hat{\mu}_K(\tilde{X}_i) \right] \right) = \frac{\sqrt{N}}{N_1} \iota'_{N_1} \tilde{R}_K (R'_K R_K)^{-1} R'_K u$$

If we treat the covariates as constants, it is easily seen that the expected value of the cross-product of the bias and variance term is 0. Hence we can deal with the bias and variance terms separately.

The variance term can be expressed as

$$\mathbb{E}(V^2) = \sigma^2 \frac{N}{N_1^2} \iota'_{N_1} \tilde{R}_K (R'_K R_K)^{-1} \tilde{R}'_K \iota_{N_1} = \sigma^2 \left( \frac{\iota'_{N_1} \tilde{R}_K}{N_1} \right) \left( \frac{R'_K R_K}{N} \right)^{-1} \left( \frac{\iota'_{N_1} \tilde{R}_K}{N_1} \right)'$$

Because

$$f(x) = \frac{p}{p(x)} f(x|D=1) \tag{4.6}$$

with  $p = \Pr(D=1)$ , the variance term can be computed from the observations in the subpopulation  $X|D=1$  if we use the weights  $\frac{p}{p(X)}$ , so that

$$\mathbb{E}(V^2) = p^2 \frac{\sigma^2}{N} a' M_K a$$

with

$$a = \begin{pmatrix} \frac{1}{p(\tilde{X}_1)} \\ \vdots \\ \frac{1}{p(\tilde{X}_N)} \end{pmatrix} \quad M_K = R_K (R'_K R_K)^{-1} R'_K$$

The bias term is

$$\frac{\sqrt{N}}{N_1} \sum_{i=1}^{N_1} \left( \mathbb{E}_{Y|X} \left[ \hat{\mu}_K(\tilde{X}_i) \right] - \mu(\tilde{X}_i) \right) = \frac{\sqrt{N}}{N_1} \sum_{i=1}^{N_1} \left( r_K(\tilde{X}_i)' (R'_K R_K)^{-1} R'_K \mu - \mu(\tilde{X}_i) \right)$$

Again we average over the distribution of  $X|D = 1$  using the weights  $\frac{p}{p(X)}$ . Hence, the bias term is written as

$$B = \frac{1}{\sqrt{N}} \sum_{i=1}^N \frac{p}{p(X_i)} (r_K(X_i)'(R_K' R_K)^{-1} R_K' \mu - \mu(X_i)) = -\frac{1}{\sqrt{N}} p a' A_K \mu$$

with  $A_K = I - M_K$ . Note that  $\frac{a' A_K \mu}{N}$  is the covariance of the residuals of the regressions of  $\mu$  on  $R_K$  and  $a$  on  $R_K$ , respectively. By Lorentz (1986), Theorem 8, p. 90, we have that

$$|a' A_K \mu| = O\left(N K^{-s_\mu/d - s_p/d}\right)$$

so that the squared bias term is of order  $O\left(N K^{-2s_\mu/d - 2s_p/d}\right)$  with  $s_\mu$  the number of continuous derivatives of  $\mu(x)$ ,  $s_p$  the number of continuous derivatives of  $p(x)$ , and  $d$  the dimension of  $X$ .

If we add the variance and squared bias terms we obtain the population MSE

$$E_N(K) = \frac{p^2}{N} (\sigma^2 a' M_K a + (a' A_K \mu)^2) \quad (4.7)$$

To use the MSE we need to estimate the bias term. This can be done in several ways<sup>2</sup>. Let  $e_K$  be the vector of residuals of the regression of  $y$  on  $R_K$ , i.e.  $e_K = A_K y$ . Using

$$a' e_K = a' A_K \mu + a' A_K u$$

so that

$$(a' e_K)^2 = (a' A_K \mu)^2 + a' A_K u u' A_K a + 2a' A_K \mu a' A_K u$$

Hence,

$$E[(a' e_K)^2] = (a' A_K \mu)^2 + \sigma^2 a' A_K a$$

so that the bias term is estimated by

$$(a' e_K)^2 - \sigma^2 a' A_K a$$

Note that

$$(a' A_K \mu)^2 = O(N^2 K^{-2s_\mu/d - 2s_p/d})$$

and

$$\sigma^2 a' A_K a = O(N K^{-2s_p/d})$$

Upon substitution of the estimate we obtain the estimated MSE

$$C_N(K) = \frac{p^2}{N} (2\sigma^2 a' M_K a - \sigma^2 a' a + (a' e_K)^2) \quad (4.8)$$

---

<sup>2</sup>Note that  $a' A_K \hat{\mu}_K = 0$ , so that this obvious estimator cannot be used.

Both the population MSE and its estimator are proportional to  $p^2$ . Hence the value of  $K$  that minimizes the (estimated) MSE is independent of the fraction of observations for which we observe  $Y$ . In the sequel we omit  $p^2$  and redefine

$$E_N(K) = \frac{1}{N} (\sigma^2 a' M_K a + (a' A_K \mu)^2)$$

$$C_N(K) = \frac{1}{N} (2\sigma^2 a' M_K a - \sigma^2 a' a + (a' e_K)^2)$$

## 4.2 The population MSE

If  $N, N_1 \rightarrow \infty$  the variance term converges to

$$\sigma^2 \mathbf{E} \left[ r_K(\tilde{X}) \right]' \Sigma_K^{-1} \mathbf{E} \left[ r_{kK}(\tilde{X}) \right]$$

with  $\Sigma_K$  the probability limit of  $\frac{R'_K R_K}{N}$ . The variance term simplifies if we choose the base functions  $r_{kK}$  as orthonormal polynomials with respect to the density of  $X|D=1$ , so that

$$\mathbf{E} [r_{jK}(X)r_{kK}(X)|D=1] = 0, \quad j \neq k \quad \mathbf{E} [r_{kK}(X)^2|D=1] = 1$$

For this choice the variance term converges to

$$\sigma^2 \mathbf{E} \left[ r_K(\tilde{X}) \right]' \mathbf{E} \left[ r_K(\tilde{X}) \right] = \sigma^2 \sum_{k=1}^K \left( \mathbf{E} \left[ r_{kK}(\tilde{X}) \right] \right)^2$$

Because the constant function is one of the orthonormal polynomials, we have that for all non-constant base functions

$$\mathbf{E} [r_{kK}(X)] = 0$$

Using (4.6)

$$\mathbf{E} \left[ r_{kK}(\tilde{X}) \right] = \mathbf{E} \left[ \frac{p}{p(X)} r_{kK}(X) \right]$$

which is proportional to the covariance of  $\frac{1}{p(X)}$  and  $r_{kK}(X)$ . If the inverse of the probability that  $Y$  is observed is smooth in the sense that it can be well approximated by a linear combination of a relatively small number of base functions, then the variance will not increase much if  $K$  exceeds the number of base functions needed.

As an example take

$$p(x) = \frac{d}{c + \gamma'_{K_0} r_{K_0}(x)}$$

With bounded support of  $X$  we can always find a  $c$  so that  $p(x)$  is nonnegative. For this choice of  $p(x)$  we have

$$E \left[ \frac{p}{p(X)} r_{kK}(X) \right] = p\gamma_k E [r_{kK}(X)^2] = p\gamma_k$$

if the base functions are orthonormal. Hence for large  $N, N_1$  the variance term is proportional to  $p^2 \sum_{k=1}^K \gamma_k^2$ . Hence the variance term increases with  $K$  for  $K \leq K_0$ , and is constant for  $K > K_0$ .

In this example the bias decreases with  $K$  and is 0 for  $K \geq K_0$  for all  $N$ . Because the bias term increases with  $N$  we find that the value of  $K$  that minimizes the MSE is bounded by  $K_0$  if  $N$  is large. In this example  $a$  is orthogonal to the residuals  $e_{K_0}$ , and  $a' A_{K_0} a = 0$ , so that the estimate of the bias is 0 (as is the bias) for  $K \geq K_0$ .

As we will see in section 4.4 these same results apply (approximately) if  $\frac{1}{p(x)}$  is not exactly a linear combination of base functions. Unless  $\frac{1}{p(x)}$  is very unsmooth, only a few base functions are needed to minimize the MSE.

It is instructive to compare the population MSE with Li's (1986) average squared error criterion. We obtain intuitive results if we assume that

$$\frac{1}{p(x)} = \sum_{k=1}^{K_0} \gamma_k r_k(x) \quad \mu(x) = \sum_{k=1}^{K_0} \delta_k r_k(x)$$

with  $r_k(x)$  orthogonal polynomials with respect to the density of  $X|D = 1$ . Li's average squared error criterion is

$$L_N(K) = \frac{1}{N} \sum_{i=1}^N (\hat{\mu}_K(X_i) - \mu(X_i))^2$$

i.e. it is average squared deviation in the sample in which we observe both  $Y$  and  $X$ . The expected value of the variance term is

$$\sigma^2 \frac{1}{N} \sum_{i=1}^N r_K(X_i)' (R_K' R_K)^{-1} r_K(X_i) = \sigma^2 \frac{K}{N}$$

The bias term is equal to

$$\frac{1}{N} \mu' \mu - \frac{1}{N} \mu' R_K (R_K' R_K)^{-1} R_K' \mu$$

which converges to  $\sum_{k=K+1}^{K_0} \delta_k^2$ . Hence Li's criterion is

$$E [L_N(K)] = \sigma^2 \frac{K}{N} + \sum_{k=K+1}^{K_0} \delta_k^2 \tag{4.9}$$

Under these assumptions the bias in our criterion is

$$p \left( \frac{1}{N} a' \mu - \frac{1}{N} a' R_K (R_K' R_K)^{-1} R_K' \mu \right)$$

We have

$$\frac{1}{N}a'\mu = \frac{1}{N} \sum_{i=1}^N \left( \sum_{k=1}^{K_0} \gamma_k r_k(X_i) \right) \left( \sum_{l=1}^{K_0} \delta_l r_l(X_i) \right) = \frac{1}{N} \sum_{i=1}^N \sum_{k=1}^{K_0} \gamma_k \delta_k r_k(X_i)^2 + \frac{1}{N} \sum_{i=1}^N \sum_{k \neq l}^{K_0} \gamma_k \delta_l r_k(X_i) r_l(X_i)$$

Because the  $r_k$  are orthonormal, the second term on the right hand side converges to 0, and the first converges to  $\sum_{k=1}^{K_0} \gamma_k \delta_k$ . Finally, because  $\frac{1}{N}a'R_K$  converges to  $\gamma'_K I_K$  and  $\frac{1}{N}R'_K \mu$  to  $I_K \delta_K$ , the second term in the bias converges to  $\sum_{k=1}^K \gamma_k \delta_k$ . Combining the results we find

$$p^2 \left[ \sigma^2 \sum_{k=1}^K \gamma_k^2 + N \left( \sum_{k=K+1}^{K_0} \gamma_k \delta_k \right)^2 \right] \quad (4.10)$$

These two criteria are clearly different. For instance, if we take  $\gamma_k = \gamma$ , then for fixed  $N$ ,  $E[L_N(K+1)] \leq E[L_N(K)]$  if and only if  $t_{K+1}^2 \geq 1$  with

$$t_k = \frac{\delta_k}{\frac{\sigma}{\sqrt{N}}}$$

the asymptotic t-ratio for  $\delta_{K+1}$ . For our criterion we find that it decreases if and only if

$$t_{K+1} \geq \sqrt{1 + T_{K+2}^2} - T_{K+2}$$

or

$$t_{K+1} \leq -\sqrt{1 + T_{K+2}^2} - T_{K+2}$$

with  $T_{K+2} = \sum_{k=K+2}^{K_0} t_k$ .

### 4.3 Optimality of minimizer of estimated MSE

Define  $\hat{K}$  as

$$\hat{K} = \operatorname{argmin}_{K \in \mathcal{K}_N} C_N(K)$$

with  $\mathcal{K}_N$  an index set that grows with  $N$  at rate  $N^\kappa$ . We will show that  $\hat{K}$  is optimal in the sense that (Li (1987))

$$\frac{E_N(\hat{K})}{\inf_{K \in \mathcal{K}_N} E_N(K)} \xrightarrow{p} 1 \quad (4.11)$$

This does not imply that the difference between  $\hat{K}$  and the minimizer of  $E_N(K)$  converges to 0.

We make the following assumptions

**Assumption 4.1** *The smallest eigenvalue of  $E[r_K(X)r_K(X)']|D=1]$  is bounded from 0 for all  $K$ .*

**Assumption 4.2**  $E[U_1^m] < \infty$  for some integer  $m$ .

**Assumption 4.3**

$$N^{\kappa(ms_p/d-1)} \inf_{K \in \mathcal{K}_N} E_N(K)^{\frac{m}{2}} \rightarrow \infty$$

with  $d$  the dimension of  $X$ .

**Theorem 4.1** If assumptions 4.1-4.3 hold, then

$$\frac{E_N(\hat{K})}{\inf_{K \in \mathcal{K}_N} E_N(K)} \xrightarrow{p} 1$$

**Proof** See Appendix.

#### 4.4 Simulation results

The finite sample performance of our estimator is investigated in a number of sampling experiments. The population model is

$$Y = \mu(X) + U$$

with

$$\mu(x) = 5 \sum_{k=1}^8 x^k$$

$X$  a scalar variable and  $U$  standard normal. We choose a simple logit model for  $p(x)$

$$p(x) = \frac{e^{1.5+3x}}{1 + e^{1.5+3x}}$$

As base functions we choose the Legendre polynomials. These polynomials are defined on  $[-1, 1]$  and are orthogonal with weight function equal to 1 on this interval. For that reason we choose the marginal distribution of  $X$  such that  $X|D = 1$  is a uniform distribution on  $[-1, 1]$ , i.e. the distribution of  $X$  has density

$$f(x) = \frac{p}{2} \frac{1}{p(x)} \quad -1 \leq x \leq 1$$

with

$$p = \frac{6}{6 + (e^{1.5} - e^{-4.5})}$$

Of course, in practice it may not be feasible to choose polynomials that are orthogonal with a weight function that is equal to the distribution  $X|D = 1$ .

Table 1 gives some statistics for the data generating process. The fraction with a missing  $Y$  is .429. The mean of both  $X$  and  $Y$  is larger in the subpopulation with observed  $Y$ .

Table 1: Statistics sampling experiment

	Mean	Std dev
$Y$	-.222	1.914
$X$	-.287	.587
$Y D = 1$	.504	2.032
$D$	.571	

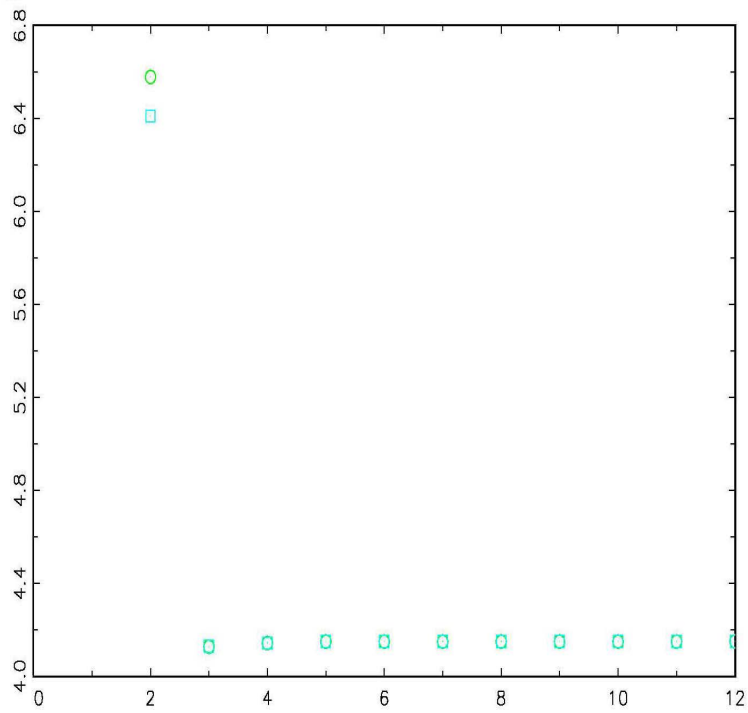
Table 2: Optimality of minimizer estimated GMM

	Fraction $\hat{K} = \tilde{K}$	$\frac{E_N(\hat{K})}{E_N(\tilde{K})}$
$N_1 = 100$	.621	1.044
$N_1 = 1000$	.650	1.011
$N_1 = 10000$	.740	1.0007

We consider three sample sizes  $N_1 = 100, 1000, 10000$ . The number of repetitions is 1000. The results are in Table 2 and the figures.

The second column is in line with the asymptotic result in section 4.3. That result does not imply that the minimizer of the population and estimated population MSE are closer if the sample size increases. However, this is what we find in the experiments.

Figure 1: Population and estimated MSE,  $N_1 = 100$



In the figures 1-3 we report the population MSE and its estimator. Note that with the simple logit model the MSE is flat after the inclusion of a few basis functions. To avoid scaling problems we omit some population and estimated MSE values for small  $K$ .



Figure 2: Population and estimated MSE,  $N_1 = 1000$

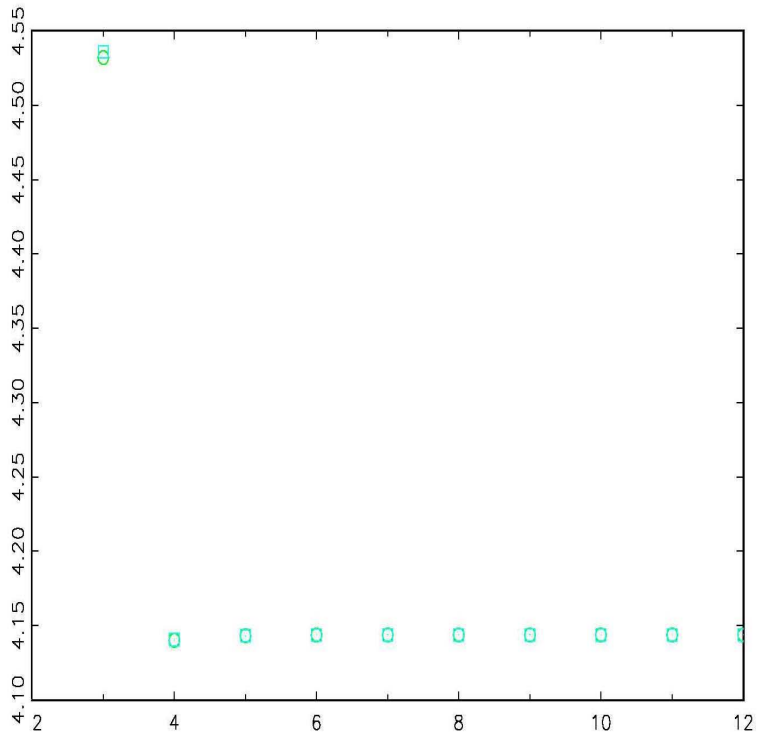
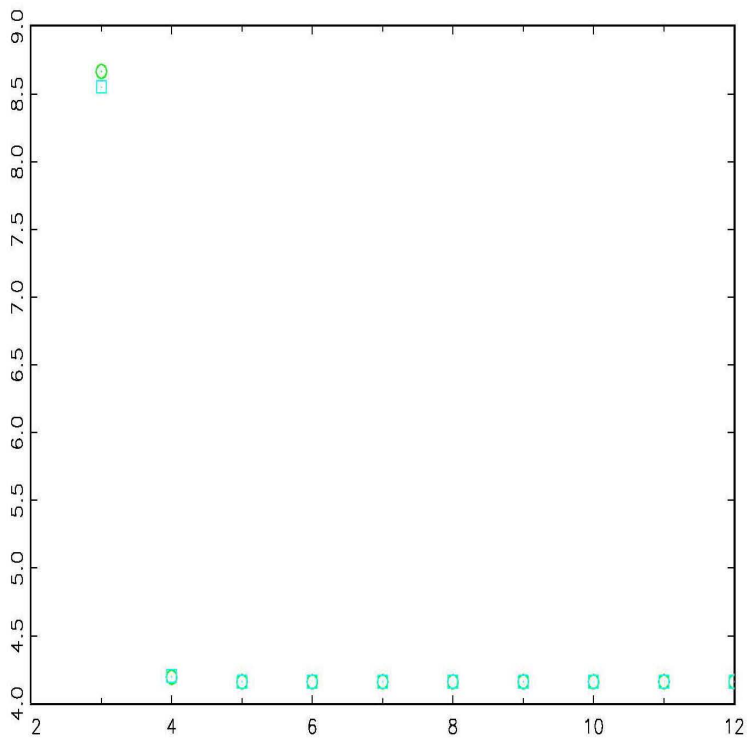


Figure 3: Population and estimated MSE,  $N_1 = 10000$



## 5 Appendix

For matrices  $A$  we use the norm  $\|A\| = (\text{tr}(A'A))^{1/2}$ .

**Lemma A.1** (PROPERTIES OF NORM) *For conformable matrices  $A$  and  $B$ ,*

$$(i), \|AB\| \leq \|A\| \cdot \|B\|,$$

$$(ii), \|A'BA\| \leq \|B\| \cdot \|A'A\|.$$

*If  $B$  is positive semi-definite and symmetric, then, for  $\lambda_{\max}(B)$  equal to the maximum eigenvalue of  $B$ ,*

$$(iii), \text{tr}(A'BA) \leq \|A\|^2 \cdot \lambda_{\max}(B),$$

$$(iv), \|AB\| \leq \|A\| \cdot \lambda_{\max}(B),$$

$$(v), \|BA\| \leq \|A\| \cdot \lambda_{\max}(B).$$

**Proof:** Let  $A$  be of dimension  $K \times L$ , and  $B$  of dimension  $L \times M$ . First we prove (i):  $AB$  is of dimension  $K \times M$ , with its  $(k, m)$  element equal to  $\sum_{l=1}^L a_{kl}b_{lm}$ . Hence the  $(m, m)$  element of the  $M \times M$  matrix  $B'A'AB$  is equal to  $\sum_{k=1}^K \left( \sum_{l=1}^L a_{kl}b_{lm} \right)^2$ , and thus

$$\|AB\|^2 = \text{tr}(B'A'AB) = \sum_{m=1}^M \sum_{k=1}^K \left( \sum_{l=1}^L a_{kl}b_{lm} \right)^2 \leq \sum_{m=1}^M \sum_{k=1}^K \sum_{l=1}^L a_{kl}^2 \sum_{l=1}^L b_{lm}^2 = \|A\|^2 \cdot \|B\|^2,$$

where the last inequality follows from the Cauchy-Schwartz inequality which implies that  $\left( \sum_{l=1}^L a_{kl}b_{lm} \right)^2 \leq \sum_{l=1}^L a_{kl}^2 \sum_{l=1}^L b_{lm}^2$ . Next, consider (ii):

$$\|A'BA\|^2 = \text{tr}(A'B'AA'BA) = \text{tr}(B'AA'BAA') = \text{tr}(D'E),$$

where  $D = AA'B$  and  $E = BAA'$ . Let  $\delta = \text{tr}(D'E)/\text{tr}(E'E)$ , and  $F = D - \delta E$ , so that  $\text{tr}(F'E) = 0$ . Then  $\text{tr}(D'E) = \text{tr}((\delta E + F)'E) = \delta \text{tr}(E'E)$ , and  $\text{tr}(D'D) = \delta^2 \cdot \text{tr}(E'E) + \text{tr}(F'F)$ , so that

$$\text{tr}(D'E)^2 = \delta^2 \cdot \text{tr}(E'E)^2 \leq (\delta^2 \cdot \text{tr}(E'E) + \text{tr}(F'F)) \cdot \text{tr}(E'E) = \text{tr}(D'D) \cdot \text{tr}(E'E).$$

Thus

$$\|A'BA\|^2 = \text{tr}(B'AA'BAA') \leq (\text{tr}(B'AA'AA'B))^{1/2} \cdot (\text{tr}(AA'B'BAA'))^{1/2} = \|AA'B\| \cdot \|BAA'\|.$$

By part (i) of the lemma,  $\|AA'B\| \leq \|AA'\| \cdot \|B\|$ , and  $\|BAA'\| \leq \|B\| \cdot \|AA'\|$ , so that the conclusion follows.

Next, consider (iii). Because  $B$  is symmetric  $L \times L$ , and positive definite, we have  $B = SAS'$ , where  $S'S = I_L$ , and  $\Lambda$  is diagonal with the eigenvalues of  $B$  on its diagonal, so that  $\lambda_{\max}(B) = \lambda_{\max}(\Lambda)$ . Let  $C = A'S$ , with  $c_{ij}$  the  $(i, j)$  element of  $A'S$ . With  $A$  of dimension  $L \times K$ ,  $C$  is of dimension  $K \times L$ . Then

$$\begin{aligned} \text{tr}(A'BA) &= \text{tr}(A'S\Lambda S'A) = \text{tr}(\Lambda S'AA'S) \\ &= \text{tr}(\Lambda C'C) = \sum_{j=1}^L \lambda_j \sum_{i=1}^K c_{ij}^2 \leq \lambda_{\max}(B) \sum_{j=1}^L \sum_{i=1}^K c_{ij}^2 = \lambda_{\max}(B) \cdot \text{tr}(C'C) \\ &= \lambda_{\max}(B) \cdot \text{tr}(S'AA'S) = \lambda_{\max}(B) \cdot \text{tr}(AA'SS') = \lambda_{\max}(B) \cdot \text{tr}(AA') \\ &= \lambda_{\max}(B) \cdot \text{tr}(A'A) = \lambda_{\max}(B) \cdot \|A\|^2. \end{aligned}$$

Next, consider (iv). Again write  $B = S\Lambda S'$ , with  $\Lambda$  diagonal and  $S'S = I_L$ . Then:

$$\begin{aligned}\|AB\|^2 &= \text{tr}(B'A'AB) = \text{tr}(A'ABB') = \text{tr}(A'AS\Lambda S'S\Lambda S') = \text{tr}(A'AS\Lambda^2 S') \\ &\leq \lambda_{\max}^2(B) \cdot \text{tr}(S'A'AS) = \lambda_{\max}^2(B) \cdot \text{tr}(A'ASS') = \lambda_{\max}^2(B) \cdot \text{tr}(A'A) = \lambda_{\max}^2(B) \cdot \|A\|^2.\end{aligned}$$

Finally, (v) can be proven the same way.  $\square$

The data consist of two random samples  $(X_i, Y_i), i = 1, \dots, N$ , and  $(Z_i), i = 1, \dots, M$ . Let  $\mu(x) = \mathbb{E}[Y|X = x]$  be the conditional expectation of the outcome given the covariates.

We use a series estimator for the regression function  $\mu(x)$ . Let  $K$  denote the number of terms in the series. As the basis we use power series. Let  $\lambda = (\lambda_1, \dots, \lambda_d)$  be a multi-index of dimension  $d$ , that is, a  $d$ -dimensional vector of non-negative integers, with  $|\lambda| = \sum_{k=1}^d \lambda_k$ , and let  $x^\lambda = x_1^{\lambda_1} \dots x_d^{\lambda_d}$ . Consider a series  $\{\lambda(r)\}_{r=1}^\infty$  containing all distinct vectors such that  $|\lambda(r)|$  is nondecreasing. Let  $p_r(x) = x^{\lambda(r)}$ , where  $P_r(x) = (p_1(x), \dots, p_r(x))'$ .

**Assumption A.1**  $X \in \mathbb{X}$ , where  $\mathbb{X}$  is the Cartesian product of intervals  $[x_{jL}, x_{jU}]$ ,  $j = 1, \dots, d$ , with  $x_{jL} < x_{jU}$ . The density of  $X$  is bounded away from zero on  $\mathbb{X}$ .

**Assumption A.2**  $Z \in \mathbb{X}$ , and the density of  $Z$  is bounded away from zero on  $\mathbb{X}$ .

Given Assumption A.1 the expectation  $\Omega_K = \mathbb{E}[P_K(X)P_K(X)']$  is nonsingular for all  $K$ . Hence we can construct a sequence  $R_K(x) = \Omega_K^{-1/2}P_K(x)$  with  $\mathbb{E}[R_K(X)R_K(X)'] = I_K$ . Let  $R_{kK}(x)$  be the  $k$ th element of the vector  $R_K(x)$ . It will be convenient to work with this sequence of basis function  $R_K(x)$ . Define

$$\zeta(K) = \sup_{x \in \mathbb{X}} \|R_K(x)\|.$$

**Lemma A.2** (NEWHEY, 1994)

$$\zeta(K) = O(K).$$

**Proof:** (see other notes)

Let  $R_K$  be the  $N \times K$  matrix with  $i$ th row equal to  $R'_K(X_i)$ , and let  $\hat{\Omega}_{K,N} = R'_K R_K / N$ .

**Lemma A.3** (NEWHEY, 1997)

$$\|\hat{\Omega}_{K,N} - I_K\| = O_p\left(\zeta(K)K^{1/2}N^{-1/2}\right)$$

**Proof:** We will show that

$$\mathbb{E}[\|R'_K R_K / N - I_K\|^2] \leq \zeta(K)^2 K / N, \tag{A.1}$$

so that the result follows by Markov's inequality.

$$\begin{aligned}\mathbb{E}[\|R'_K R_K / N - I_K\|^2] &= \mathbb{E}[\text{tr}(R'_K R_K R'_K R_K / N^2 - 2R'_K R_K / N + I_K)] \\ &= \text{tr}(\mathbb{E}[R'_K R_K R'_K R_K / N^2] - 2\mathbb{E}[R'_K R_K / N] + I_K) \\ &= \text{tr}(\mathbb{E}[R'_K R_K R'_K R_K / N^2]) - K\end{aligned}$$

$$= \sum_{k=1}^K \sum_{l=1}^K \sum_{i=1}^N \sum_{j=1}^N \mathbb{E} [R_{kK}(X_i)R_{lK}(X_i)R_{lK}(X_j)R_{kK}(X_j)/N^2] - K.$$

The terms with  $i \neq j$  and  $k \neq l$  have expectation zero. The terms with  $k = l$  and  $i \neq j$  have expectation  $\mathbb{E}[R_{kK}(X_i)^2 R_{kK}(X_j)^2] = 1$ . There are  $K \times N \times (N - 1)$  of those terms, each divided by  $N^2$ , leading to a total of  $K(N - 1)/N = K - K/N$ . Hence

$$\begin{aligned} & \left| \sum_{k=1}^K \sum_{l=1}^K \sum_{i=1}^N \sum_{j=1}^N \mathbb{E} [R_{kK}(X_i)R_{lK}(X_i)R_{lK}(X_j)R_{kK}(X_j)/N^2] - K \right| \\ & \leq \sum_{i=1}^N \sum_{k=1}^K \sum_{l=1}^K \mathbb{E} [R_{kK}(X_i)R_{lK}(X_i)R_{lK}(X_i)R_{kK}(X_i)/N^2] \\ & = \frac{1}{N^2} \sum_{i=1}^N \mathbb{E} \left[ \sum_{k=1}^K R_{kK}(X_i)^2 \sum_{l=1}^K R_{lK}(X_i)^2 \right]. \end{aligned} \quad (\text{A.2})$$

By definition,  $\zeta(K) = \sup_x \|R_K(x)\| = \sup_x (\sum_k R_{kK}^2(x))^{1/2}$ . Hence it follows that  $\sum_k R_{kK}^2(x) \leq \zeta^2(K)$ , and thus (A.2) can be bounded by

$$\frac{1}{N^2} \sum_{i=1}^N \zeta^2(K) \cdot \mathbb{E} \left[ \sum_{l=1}^K R_{lK}(X_i)^2 \right] = K\zeta^2(K)/N.$$

This shows that (A.1) holds, and thus finishes the proof.  $\square$

Let  $U_i = Y_i - \mu(X_i)$ . Let  $\mathbf{U}$ ,  $\mathbf{Y}$ , and  $\mathbf{X}$  be the  $N$  vectors and  $N \times d$  matrix with  $i$ th row equal to  $U_i$ ,  $Y_i$ , and  $X_i'$ . Furthermore, let  $1_N$  be an indicator for the event  $\lambda_{\min}(\hat{\Omega}_{K,N}) > 1/2$ . By Lemma A.3, it follows that  $1_N \xrightarrow{p} 1$ . For a symmetric matrix  $A$ , we can write  $A$  as  $F\Lambda F$ , where  $F$  is symmetric and  $\Lambda$  diagonal. Let  $\Lambda^{-1/2}$  be the diagonal matrix with each diagonal element equal to the reciprocal of the square root of the corresponding diagonal element of  $\Lambda$ , and let  $A^{-1/2} = F\Lambda^{-1/2}$ , so that  $A^{-1/2}$  is symmetric and satisfies  $A^{-1/2}A^{-1/2} = A^{-1}$ .

### Assumption A.3

$$\sup_{x \in \mathcal{X}} \text{Var}(Y|X) \leq \bar{\sigma}^2.$$

**Lemma A.4** (i),

$$1_N \cdot \left\| \hat{\Omega}_{K,N}^{-1/2} R_K' \mathbf{U} / N \right\| = O_p(K^{1/2} N^{-1/2}),$$

and (ii),

$$1_N \cdot \left\| \hat{\Omega}_{K,N}^{-1} R_K' \mathbf{U} / N \right\| = O_p(K^{1/2} N^{-1/2}),$$

**Proof:** First we prove (i).

$$\begin{aligned} & \mathbb{E} \left[ 1_N \cdot \left\| \hat{\Omega}_{K,N}^{-1/2} R_K' \mathbf{U} / N \right\|^2 \middle| X \right] \\ & = \mathbb{E} \left[ 1_N \cdot \mathbf{U}' R_K \hat{\Omega}_{K,N}^{-1} R_K' \mathbf{U} / N^2 \middle| X \right] \end{aligned}$$

$$\begin{aligned}
&= \mathbb{E} \left[ 1_N \cdot \mathbf{U}' R_K (R_K' R_K)^{-1} R_K' \mathbf{U} / N \mid X \right] \\
&= 1_N \cdot \mathbb{E} \left[ \text{tr} \left( \mathbf{U}' R_K (R_K' R_K)^{-1} R_K' \mathbf{U} \right) \mid X \right] / N \\
&= 1_N \cdot \mathbb{E} \left[ \text{tr} \left( R_K (R_K' R_K)^{-1} R_K' \mathbf{U} \mathbf{U}' \right) \mid X \right] / N \\
&= 1_N \cdot \text{tr} \left( R_K (R_K' R_K)^{-1} R_K' \mathbb{E} [\mathbf{U} \mathbf{U}' \mid X] \right) / N \\
&\leq \bar{\sigma}^2 \cdot 1_N \cdot \text{tr} \left( R_K (R_K' R_K)^{-1} R_K' \right) / N \\
&= \bar{\sigma}^2 \cdot 1_N \cdot K / N \\
&\leq \bar{\sigma}^2 K / N.
\end{aligned}$$

Then by the Markov inequality  $1_N \|\hat{\Omega}_{K,N}^{-1/2} R_K' \mathbf{U} / N\| = O_p(K^{1/2} / N^{-1/2})$ .

Next, consider part (ii). Using Lemma A.1(v),

$$1_N \cdot \left\| \hat{\Omega}_{K,N}^{-1} R_K' \mathbf{U} / N \right\| \leq 1_N \cdot \lambda_{\max}(\hat{\Omega}_{K,N}^{-1/2}) \cdot \left\| \hat{\Omega}_{K,N}^{-1/2} R_K' \mathbf{U} / N \right\|.$$

Since  $\lambda_{\max}(\hat{\Omega}_{K,N}^{-1/2}) = O_p(1)$ , the conclusion follows.  $\square$

This conditional expectation  $\mu(x)$  is estimated by the series estimator  $\hat{\mu}_K(x) = R_K'(x) \hat{\gamma}_K$  with  $\hat{\gamma}_K$  the least squares estimator. Formally  $R_K' R_K$  may be singular, although Lemma A.3 shows that this happens with probability going to zero. To deal with this case we define  $\hat{\gamma}_K$  as

$$\hat{\gamma}_K = \begin{cases} 0_K & \text{if } \lambda_{\min}(R_K' R_K / N) \leq 1/2, \\ \left( \sum_{i=1}^N R_K(X_i) R_K'(X_i) \right)^{-1} \sum_{i=1}^N R_K(X_i) Y_i & \text{otherwise,} \end{cases} \quad (\text{A.3})$$

where  $0_K$  is a  $K$ -dimensional vector of zeros, and  $\lambda_{\min}(A)$  is the minimum eigenvalue of the matrix  $A$ .

Define  $\gamma_K^*$  to be the pseudo true value defined as

$$\gamma_K^* = \arg \min_{\gamma} \mathbb{E} \left[ (\mu(X) - R_K(X)' \gamma)^2 \mid W = 1 \right], \quad (\text{A.4})$$

with the corresponding pseudo true value of the regression function denoted by  $\mu_K^*(x) = R_K(x)' \gamma_K^*$ .

First we state some of the properties of the estimator for the regression function. In order to do so it is useful to first give some approximation properties for  $\mu(x)$ .

**Assumption A.4**  $\mu(x)$  is  $s$  times continuously differentiable on  $\mathbb{X}$ .

**Lemma A.5** (LORENTZ) Suppose Assumptions A.1-A.2 hold. Then there is a sequence  $\gamma_K^0$  such that

$$\sup_x |\mu(x) - R_K(x)' \gamma_K^0| = O\left(K^{-s/d}\right).$$

**Proof:** (see notes)

For the sequence  $\gamma_K^0$  in Lemma A.5, define the corresponding sequence of regression functions,  $\mu_K^0(x) = R_K(x)' \gamma_K^0$ .

**Lemma A.6** (CONVERGENCE RATE FOR REGRESSION FUNCTION ESTIMATORS)

Suppose Assumptions A.1-A.2 hold. Then (i):

$$\|\gamma_K^* - \gamma_K^0\| = O\left(\zeta(K)K^{-s/d}\right), \quad (\text{A.5})$$

(ii):

$$\sup_x |\mu_K^*(x) - \mu_K^0(x)| = O(\zeta^2(K)K^{-s/d}). \quad (\text{A.6})$$

(iii):

$$\|\hat{\gamma}_K - \gamma_K^*\| = O_p(K^{1/2}N^{-1/2} + K^{-s/d}), \quad (\text{A.7})$$

(iv):

$$\|\hat{\gamma}_K - \gamma_K^0\| = O_p\left(K^{1/2}N^{-1/2} + K^{-s/d}\right), \quad (\text{A.8})$$

(v):

$$\sup_x |\hat{\mu}_K(x) - \mu_K^*(x)| = O_p(\zeta(K)K^{1/2}N^{-1/2} + \zeta(K)K^{-s/d}), \quad (\text{A.9})$$

and (vi):

$$\sup_x |\hat{\mu}_K(x) - \mu(x)| = O_p(\zeta(K)K^{1/2}N^{-1/2} + \zeta^2(K)K^{-s/d}), \quad (\text{A.10})$$

**Proof:** First, consider (i):

$$\begin{aligned} \gamma_K^* &= (\mathbb{E}[R_K(X)R_K'(X)])^{-1} \mathbb{E}[R_K(X)Y] = (\mathbb{E}[R_K(X)R_K'(X)])^{-1} \mathbb{E}[R_K(X)\mu(X)] \\ &= \mathbb{E}[R_K(X)\mu(X)], \end{aligned}$$

where we use both  $[R_K(X)U] = 0$  and  $\mathbb{E}[R_K(X)R_K'(X)] = I_K$ . Also,  $\gamma_K^0 = \mathbb{E}[R_K(X)R_K'(X)\gamma_K^0]$ , so that

$$\begin{aligned} \|\gamma_K^* - \gamma_K^0\| &= \|\mathbb{E}[R_K(X)\mu(X)] - \mathbb{E}[R_K(X)R_K'(X)\gamma_K^0]\| \\ &= \|\mathbb{E}[R_K(X)(\mu(X) - R_K(X)'\gamma_K^0)]\| \\ &\leq \sup_x \|R_K(x)\| \cdot \sup_x |\mu(x) - R_K(x)'\gamma_K^0| \leq C\zeta(K)K^{-s/d}. \end{aligned}$$

Next, consider (ii):

$$\begin{aligned} \sup_x |\mu_K^*(x) - \mu_K^0(x)| &= \sup_x |R_K(x)'(\gamma_K^* - \gamma_K^0)| \\ &\leq \sup_x \|R_K(x)\| \cdot \|\gamma_K^* - \gamma_K^0\| \leq C\zeta^2(K)K^{-s/d}, \end{aligned}$$

using the result in Lemma A.6(i).

Next, consider (iii):

$$\|\hat{\gamma}_K - \gamma_K^*\| = 1_N \cdot \|\hat{\gamma}_K - \gamma_K^*\| + (1 - 1_N) \cdot \|\hat{\gamma}_K - \gamma_K^*\|,$$

since  $\Pr(1_N = 1) \rightarrow 1$ . The second term is

$$(1 - 1_N) \cdot \|\hat{\gamma}_K - \gamma_K^*\| = (1 - 1_N) \cdot \|\gamma_K^*\| = o_p(1).$$

The first term is

$$\begin{aligned} 1_N \cdot \|\hat{\gamma}_K - \gamma_K^*\| &= 1_N \cdot \|\hat{\Omega}_{K,N}^{-1}(R'_K \mathbf{Y}/N) - \hat{\Omega}_{K,N}^{-1}(R'_K R_K \gamma_K^*/N)\| \\ &= 1_N \cdot \|\hat{\Omega}_{K,N}^{-1}(R'_K \mathbf{U}/N) + \hat{\Omega}_{K,N}^{-1}(R'_K(\mu(\mathbf{X}) - R_K \gamma_K^*)/N)\| \\ &\leq 1_N \cdot \|\hat{\Omega}_{K,N}^{-1}(R'_K \mathbf{U}/N)\| + 1_N \cdot \|\hat{\Omega}_{K,N}^{-1}(R'_K(\mu(\mathbf{X}) - R_K \gamma_K^*)/N)\|. \end{aligned}$$

By Lemma A.4(ii) the first term is  $O_p(K^{1/2}N^{-1/2})$ . For the second term, we have, using Lemma A.1(v), and using the fact that if  $1_N$  is equal to 1, then  $\lambda_{\max}(\hat{\Omega}_{K,N}^{-1/2}) \leq \sqrt{2}$ ,

$$\begin{aligned} 1_N \cdot \|\hat{\Omega}_{K,N}^{-1}(R'_K(\mu(\mathbf{X}) - R_K \gamma_K^*)/N)\| &\leq 1_N \cdot \lambda_{\max}(\hat{\Omega}_{K,N}^{-1/2}) \cdot \|\hat{\Omega}_{K,N}^{-1/2}(R'_K(\mu(\mathbf{X}) - R_K \gamma_K^*)/N)\| \\ &\leq 1_N \cdot \sqrt{2} \cdot \left( \frac{1}{N^2} (\mu(\mathbf{X}) - R_K \gamma_K^*)' R'_K \hat{\Omega}_{K,N}^{-1/2} \hat{\Omega}_{K,N}^{-1/2} R'_K (\mu(\mathbf{X}) - R_K \gamma_K^*) \right)^{1/2} \\ &= 1_N \cdot \sqrt{2} \cdot \left( \frac{1}{N} (\mu(\mathbf{X}) - R_K \gamma_K^*)' R'_K (R'_K R_K)^{-1} R'_K (\mu(\mathbf{X}) - R_K \gamma_K^*) \right)^{1/2} \\ &\leq 1_N \cdot \sqrt{2} \cdot \left( \frac{1}{N} (\mu(\mathbf{X}) - R_K \gamma_K^*)' (\mu(\mathbf{X}) - R_K \gamma_K^*) \right)^{1/2} \end{aligned} \quad (\text{A.11})$$

where we use the fact that because  $R_K (R'_K R_K)^{-1} R'_K$  is a projection matrix, it follows that  $I_N - R_K (R'_K R_K)^{-1} R'_K$  is positive semi-definite. Since by the definition of  $\gamma_K^*$

$$\begin{aligned} \mathbb{E} \left[ \frac{1}{N} (\mu(\mathbf{X}) - R_K \gamma_K^*)' (\mu(\mathbf{X}) - R_K \gamma_K^*) \right] &\leq \mathbb{E} \left[ \frac{1}{N} (\mu(\mathbf{X}) - R_K \gamma_K^0)' (\mu(\mathbf{X}) - R_K \gamma_K^0) \right] \\ &\leq \sup_x |\mu(x) - R_K(x)' \gamma_K^0|^2 \leq CK^{-2s/d}, \end{aligned}$$

it follows by the Markov inequality that (A.11) is  $O_p(K^{-s/d})$ . Hence  $\|\hat{\gamma}_K - \gamma_K^*\| = O_p(K^{-s/d}) + o_p(1) + O_p(K^{1/2}N^{-1/2}) = O_p(K^{-s/d} + K^{1/2}N^{-1/2})$ .

Next, consider (iv). We use the same argument as for (iii):

$$\begin{aligned} 1_N \cdot \|\hat{\gamma}_K - \gamma_K^0\| &= 1_N \cdot \|\hat{\Omega}_{K,N}^{-1}(R'_K \mathbf{Y}/N) - \hat{\Omega}_{K,N}^{-1}(R'_K R_K \gamma_K^0/N)\| \\ &\leq 1_N \cdot \|\hat{\Omega}_{K,N}^{-1}(R'_K \mathbf{U}/N)\| + 1_N \cdot \|\hat{\Omega}_{K,N}^{-1}(R'_K(\mu(\mathbf{X}) - R'_K \gamma_K^0)/N)\|. \end{aligned}$$

The first term is again  $O_p(K^{1/2}N^{-1/2})$ . For the second term we now have:

$$\begin{aligned} 1_N \cdot \|\hat{\Omega}_{K,N}^{-1}(R'_K(\mu(\mathbf{X}) - R_K \gamma_K^0)/N)\| &\leq 1_N \cdot \lambda_{\max}(\hat{\Omega}_{K,N}^{-1/2}) \cdot \|\hat{\Omega}_{K,N}^{-1/2}(R'_K(\mu(\mathbf{X}) - R_K \gamma_K^0)/N)\| \\ &\leq 1_N \cdot \sqrt{2} \cdot \sup_x |\mu(x) - R_K(x)' \gamma_K^0| = O(K^{-s/d}). \end{aligned}$$

Next, consider (v):

$$\begin{aligned} \sup_x |\hat{\mu}_K(x) - \mu_K^*(x)| &= \sup_x |R_K(x)'(\hat{\gamma}_K - \gamma_K^*)| \\ &\leq \sup_x \|R_K(x)\| \cdot \|\hat{\gamma}_K - \gamma_K^*\| = O_p(\zeta(K)K^{1/2}N^{-1/2} + \zeta(K)K^{-s/d}). \end{aligned}$$



Finally, consider (vi).

$$\begin{aligned} \sup_x |\hat{\mu}_K(x) - \mu(x)| &\leq \sup_x |\hat{\mu}_K(x) - \mu_K^*(x)| + \sup_x |\mu_K^*(x) - \mu_K^0(x)| + \sup_x |\mu_K^0(x) - \mu(x)| \\ &= O_p(\zeta(K)K^{1/2}N^{-1/2} + \zeta(K)K^{-s/d} + \zeta^2(K)K^{-s/d} + K^{-s/d}) = O_p(\zeta(K)K^{1/2}N^{-1/2} + \zeta^2(K)K^{-s/d}). \end{aligned}$$

□

The new imputation estimator developed in this paper is

$$\hat{\beta}_{inr} = \frac{1}{N} \sum_{i=1}^N \hat{\mu}_K(X_i) = \frac{1}{N} \sum_{i=1}^N R_K(X_i)' \hat{\gamma}_K \quad (\text{A.12})$$

It is useful to consider in this discussion three additional estimators that are based on the propensity score. In all cases we estimate the propensity score using a logistic series estimator, as suggested in Hirano, Imbens and Ridder (2003). Here we briefly summarize the relevant results. Let  $L(z) = \exp(z)/(1 + \exp(z))$  be the logistic cdf and  $L'(z) = L(z) \cdot (1 - L(z))$ . The series logit estimator of the population propensity score  $e^*(x)$  is  $\hat{e}_K(x) = L(R_K(x)' \hat{\pi}_K)$ , (for simplicity we use the same series  $R_K(x)$ , although this is not essential) where

$$\hat{\pi}_K = \arg \max_{\pi} L_N(\pi), \quad (\text{A.13})$$

for

$$L_N(\pi) = \sum_{i=1}^N (W_i \cdot \ln L(R_K(X_i)' \pi) + (1 - W_i) \cdot \ln(1 - L(R_K(X_i)' \pi))). \quad (\text{A.14})$$

For  $N \rightarrow \infty$  and  $K$  fixed we have  $\hat{\pi}_K \xrightarrow{P} \pi_K^*$ , with  $\pi_K^*$  the pseudo true value:

$$\pi_K^* = \arg \max_{\pi} \mathbb{E} [e^*(X) \cdot \ln L(R_K(X)' \pi) + (1 - e^*(X)) \cdot \ln(1 - L(R_K(X)' \pi))]. \quad (\text{A.15})$$

We also define the pseudo true propensity score:  $e_K^*(x) = L(R_K(x)' \pi_K^*)$ .

**Assumption A.5**  $e(x)$  is  $s$  times continuously differentiable on  $\mathbb{X}$ .

**Lemma A.7** Suppose Assumptions A.1, A.3 hold. Then there is a sequence  $\pi_L^0$  such that

$$\sup_x |e(x) - L(R_L(x)' \pi_K^0)| = O(L^{-s/d}).$$

**Assumption A.6**  $\inf_x e(x) > 0$  and  $\sup_x e(x) < 1$ .

**Lemma A.8** (CONVERGENCE RATE FOR PROPENSITY SCORE ESTIMATORS)

Suppose Assumptions ??, A.1, and A.4 hold. Then (i):

$$\|\hat{\pi}_L - \pi_L^*\| = O_p(L^{1/2}N^{-1/2}), \quad (\text{A.16})$$

(ii):

$$\|\pi_L^* - \pi_L^0\| = O(L^{-s/(2d)}), \quad (\text{A.17})$$

(iii):

$$\sup_x |\hat{e}_L(x) - e_L^*(x)| = O_p(\zeta(L)L^{1/2}N^{-1/2}). \quad (\text{A.18})$$

(iv):

$$\sup_x |e_L^*(x) - e_L^0(x)| = O\left(\zeta(L)L^{-s/(2d)}\right). \quad (\text{A.19})$$

(v):

$$\sup_x |e_L^0(x) - e(x)| = O_p\left(L^{-s/d}\right). \quad (\text{A.20})$$

(vi),

$$\|\hat{\pi}_L - \pi_L^0\| = O_p(L^{1/2}N^{-1/2} + L^{-s/(2d)}), \quad (\text{A.21})$$

(vii),

$$\sup_x |\hat{e}_L(x) - e_L^0(x)| = O_p(\zeta(L)(L^{1/2}N^{-1/2} + L^{-s(2d)})), \quad (\text{A.22})$$

and (viii),

$$\sup_x |\hat{e}_L(x) - e(x)| = O_p(\zeta(L)(L^{1/2}N^{-1/2} + L^{-s(2d)})). \quad (\text{A.23})$$

**Proof:** See Hirano, Imbens and Ridder (2003).

The second estimator is a modified imputation estimator that only averages over the observations with  $W_i = 1$ :

$$\hat{\beta}_{mod} = \frac{1}{N} \sum_{i=1}^N \frac{W_i \cdot \hat{\mu}_K(X_i)}{\hat{e}_L(X_i)}. \quad (\text{A.24})$$

The third estimator is the weighting estimator proposed by Hirano, Imbens and Ridder (2003):

$$\hat{\beta}_{hir,1} = \frac{1}{N} \sum_{i=1}^N \frac{W_i \cdot Y_i}{\hat{e}_L(X_i)}. \quad (\text{A.25})$$

Hirano, Imbens and Ridder (2003) show that  $\hat{\beta}_{hir,1}$  is consistent, asymptotically normal and efficient.

The fourth estimator is a modified version of the Hirano-Imbens-Ridder estimator where the weights are normalized to add up to unity:

$$\hat{\beta}_{hir,2} = \sum_{i=1}^N \frac{W_i \cdot Y_i}{\hat{e}_L(X_i)} \bigg/ \sum_{j=1}^N \frac{W_j}{\hat{e}_L(X_j)}. \quad (\text{A.26})$$

The properties of  $\hat{\beta}_{inr}$  will follow from the following lemma:

**Proof of Theorem 3.1:** First we prove (i). We have

$$\begin{aligned} \sqrt{N}|\hat{\beta}_{inr} - \hat{\beta}_{mod}| &= \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \hat{\mu}_K(X_i) - \frac{1}{\sqrt{N}} \sum_{i=1}^N \frac{W_i \cdot \hat{\mu}_K(X_i)}{\hat{e}_L(X_i)} \right| \\ &= \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \left( \frac{\hat{\mu}_K(X_i) \cdot \hat{e}_L(X_i)}{\hat{e}_L(X_i)} - \frac{W_i \cdot \hat{\mu}_K(X_i)}{\hat{e}_L(X_i)} \right) \right| \\ &= \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \left( \frac{\hat{\mu}_K(X_i) \cdot \hat{e}_L(X_i)}{\hat{e}_L(X_i)} - \frac{W_i \cdot \hat{\mu}_K(X_i)}{\hat{e}_L(X_i)} \right) \right| \end{aligned} \quad (\text{A.27})$$

$$\begin{aligned}
& + \frac{\hat{\mu}_K(X_i) \cdot \hat{e}_L(X_i)}{e_L(X_i)} - \frac{\hat{\mu}_K(X_i) \cdot \hat{e}_L(X_i)}{e_L(X_i)} + \frac{\hat{\mu}_K(X_i) \cdot W_i}{e_L^0(X_i)} - \frac{\hat{\mu}_K(X_i) \cdot W_i}{e_L^0(X_i)} \\
& + \frac{\mu_K^0(X_i) \cdot \hat{e}_L(X_i)}{e_L^0(X_i)} - \frac{\mu_K^0(X_i) \cdot \hat{e}_L(X_i)}{e_L^0(X_i)} + \frac{\hat{\mu}_K(X_i) \cdot e_L^0(X_i)}{e_L^0(X_i)} - \frac{\hat{\mu}_K(X_i) \cdot e_L^0(X_i)}{e_L^0(X_i)} \\
& + \frac{\mu_K^0(X_i) \cdot e_L^0(X_i)}{e_L^0(X_i)} - \frac{\mu_K^0(X_i) \cdot e_L^0(X_i)}{e_L^0(X_i)} + \frac{\hat{\mu}_K(X_i) \cdot e(X_i)}{e_L^0(X_i)} - \frac{\hat{\mu}_K(X_i) \cdot e(X_i)}{e_L^0(X_i)} \\
& + \frac{\mu_K^0(X_i) \cdot e(X_i)}{e_L^0(X_i)} - \frac{\mu_K^0(X_i) \cdot e(X_i)}{e_L^0(X_i)} + \frac{\hat{\mu}_K^0(X_i) \cdot e(X_i)}{e(X_i)} - \frac{\hat{\mu}_K^0(X_i) \cdot e(X_i)}{e(X_i)} \\
& + \frac{\mu_K^0(X_i) \cdot e(X_i)}{e(X_i)} - \frac{\mu_K^0(X_i) \cdot e(X_i)}{e(X_i)} + \frac{\hat{\mu}_K(X_i) \cdot W_i}{e(X_i)} - \frac{\hat{\mu}_K(X_i) \cdot W_i}{e(X_i)} \\
& + \frac{\mu_K^0(X_i) \cdot W_i}{e_L^0(X_i)} - \frac{\mu_K^0(X_i) \cdot W_i}{e_L^0(X_i)} + \frac{\mu_K^0(X_i) \cdot W_i}{e(X_i)} - \frac{\mu_K^0(X_i) \cdot W_i}{e(X_i)} \\
& + \left. \frac{\mu(X_i) \cdot \hat{e}_L(X_i)}{e(X_i)} - \frac{\mu(X_i) \cdot \hat{e}_L(X_i)}{e(X_i)} + \frac{\mu(X_i) \cdot W_i}{e(X_i)} - \frac{\mu(X_i) \cdot W_i}{e(X_i)} \right) \Big| \\
= & \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \left( \frac{\hat{\mu}_K(X_i) \cdot \hat{e}_L(X_i)}{\hat{e}_L(X_i)} - \frac{\hat{\mu}_K(X_i) \cdot \hat{e}_L(X_i)}{e_L^0(X_i)} - \frac{W_i \cdot \hat{\mu}_K(X_i)}{\hat{e}_L^0(X_i)} + \frac{\hat{\mu}_K(X_i) \cdot W_i}{e_L^0(X_i)} \right. \right. \\
& + \frac{\hat{\mu}_K(X_i) \cdot \hat{e}_L(X_i)}{e_L^0(X_i)} - \frac{\mu_K^0(X_i) \cdot \hat{e}_L(X_i)}{e_L^0(X_i)} - \frac{\hat{\mu}_K(X_i) \cdot e_L^0(X_i)}{e_L^0(X_i)} + \frac{\mu_K^0(X_i) \cdot e_L^0(X_i)}{e_L^0(X_i)} \\
& + \frac{\hat{\mu}_K(X_i) \cdot e_L^0(X_i)}{e_L^0(X_i)} - \frac{\mu_K^0(X_i) \cdot e_L(X_i)}{e_L^0(X_i)} - \frac{\hat{\mu}_K(X_i) \cdot e(X_i)}{e_L^0(X_i)} + \frac{\mu_K^0(X_i) \cdot e(X_i)}{e_L^0(X_i)} \\
& + \frac{\hat{\mu}_K(X_i) \cdot e(X_i)}{e_L^0(X_i)} - \frac{\mu_K(X_i) \cdot e(X_i)}{e_L^0(X_i)} - \frac{\hat{\mu}_K(X_i) \cdot e(X_i)}{e(X_i)} + \frac{\mu_K^0(X_i) \cdot e(X_i)}{e(X_i)} \\
& - \frac{\hat{\mu}_K(X_i) \cdot W_i}{e_L^0(X_i)} + \frac{\mu_K(X_i) \cdot W_i}{e_L^0(X_i)} + \frac{\hat{\mu}_K(X_i) \cdot W_i}{e(X_i)} - \frac{\mu_K^0(X_i) \cdot W_i}{e(X_i)} \\
& + \frac{\hat{\mu}_K(X_i) \cdot e(X_i)}{e(X_i)} - \frac{\mu_K^0(X_i) \cdot e(X_i)}{e(X_i)} - \frac{\hat{\mu}_K(X_i) \cdot W_i}{e(X_i)} + \frac{\mu_K^0(X_i) \cdot W_i}{e(X_i)} \\
& + \frac{\mu_K^0(X_i) \cdot \hat{e}_L^0(X_i)}{e_L^0(X_i)} - \frac{\mu(X_i) \cdot \hat{e}_L(X_i)}{e(X_i)} - \frac{\mu_K^0(X_i) \cdot W_i}{e_L^0(X_i)} + \frac{\mu(X_i) \cdot W_i}{e(X_i)} \\
& \left. + \frac{\mu(X_i) \cdot \hat{e}_L(X_i)}{e(X_i)} - \frac{\mu(X_i) \cdot W_i}{e(X_i)} \right) \Big| \\
\leq & \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \left( \frac{\hat{\mu}_K(X_i)}{\hat{e}_L(X_i)} - \frac{\hat{\mu}_K^0(X_i)}{e_L^0(X_i)} \right) (\hat{e}_L(X_i) - W_i) \right| \tag{A.28} \\
& + \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \frac{\hat{\mu}_K(X_i) - \mu_K^0(X_i)}{e_L(X_i)} (\hat{e}_L(X_i) - e_L^0(X_i)) \right| \tag{A.29} \\
& + \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \frac{\hat{\mu}_K(X_i) - \mu_K^0(X_i)}{e_L^0(X_i)} (e_L^0(X_i) - e(X_i)) \right| \tag{A.30} \\
& + \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \left( \frac{\hat{\mu}_K(X_i) - \mu_K^0(X_i)}{e_L(X_i)} - \frac{\hat{\mu}_K(X_i) - \mu_K^0(X_i)}{e(X_i)} \right) \cdot (e(X_i) - W_i) \right| \tag{A.31}
\end{aligned}$$

$$+ \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \frac{\hat{\mu}_K(X_i) - \mu_K^0(X_i)}{e(X_i)} \cdot (e(X_i) - W_i) \right| \quad (\text{A.32})$$

$$+ \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \left( \frac{\mu_K(X_i)}{e_L^0(X_i)} - \frac{\mu(X_i)}{e(X_i)} \right) \cdot (\hat{e}_L(X_i) - W_i) \right| \quad (\text{A.33})$$

$$+ \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \frac{\mu(X_i)}{e(X_i)} (\hat{e}_L(X_i) - W_i) \right|. \quad (\text{A.34})$$

We will deal with equations (A.28)-(A.34) separately. First consider (A.28). This itself will be broken up into seven parts.

$$\begin{aligned} & \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \left( \frac{R_K(X_i)' \hat{\gamma}_K}{L(R_L(X_i)' \hat{\pi}_L)} - \frac{R_K(X_i)' \hat{\gamma}_K}{L(R_L(X_i)' \pi_L^0)} \right) (\hat{e}_L(X_i) - W_i) \right| \\ &= \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \left( \frac{\hat{\mu}_K(X_i)}{\hat{e}_L(X_i)} - \frac{\hat{\mu}_K(X_i)}{e_L^0(X_i)} \right) \cdot (\hat{e}_L(X_i) - W_i) \right| \\ &= \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \mu_K(X_i) \cdot \left( \frac{1}{\hat{e}_L(X_i)} - \frac{1}{e_L^0(X_i)} \right) \cdot (\hat{e}_L(X_i) - W_i) \right| \\ &\leq \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N (\hat{\mu}_K(X_i) - \mu_K(X_i)) \cdot \left( \frac{1}{\hat{e}_L(X_i)} - \frac{1}{e_L(X_i)} \right) \cdot (\hat{e}_L(X_i) - W_i) \right| \end{aligned} \quad (\text{A.35})$$

$$+ \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \mu_K(X_i) \cdot \left( \frac{\hat{e}_L(X_i) - e_L(X_i)}{e_L^2(X_i)} - \left( \frac{1}{\hat{e}_L(X_i)} - \frac{1}{e_L(X_i)} \right) \right) \cdot (\hat{e}_L(X_i) - W_i) \right| \quad (\text{A.36})$$

$$+ \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \left( \frac{\mu_K(X_i)}{e_L^2(X_i)} - \frac{\mu(X_i)}{e^2(X_i)} \right) \cdot (\hat{e}_L(X_i) - e_L(X_i)) \cdot (\hat{e}_L(X_i) - W_i) \right| \quad (\text{A.37})$$

$$+ \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \frac{\mu(X_i)}{e^2(X_i)} \cdot \left( \hat{e}_L(X_i) - e_L(X_i) - e(X_i)(1 - e(X_i)) R'_L(X_i) (\hat{\pi}_L - \pi_L^0) \right) \cdot (\hat{e}_L(X_i) - W_i) \right| \quad (\text{A.38})$$

$$+ \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \frac{\mu(X_i) e(X_i) (1 - e(X_i))}{e^2(X_i)} \cdot R'_L(X_i) (\hat{\pi}_L - \pi_L^0) (\hat{e}_L(X_i) - e_L(X_i)) \right| \quad (\text{A.39})$$

$$+ \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \frac{\mu(X_i) e(X_i) (1 - e(X_i))}{e^2(X_i)} \cdot R'_L(X_i) (\hat{\pi}_L - \pi_L^0) (e_L(X_i) - e(X_i)) \right| \quad (\text{A.40})$$

$$+ \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \frac{\mu(X_i) e(X_i) (1 - e(X_i))}{e^2(X_i)} \cdot R'_L(X_i) (\hat{\pi}_L - \pi_L^0) (e(X_i) - W_i) \right| \quad (\text{A.41})$$

First consider (A.35). Since  $\inf_x e(x) > c$  for some  $c > 0$ , it follows that for large enough  $N$ ,  $\inf_x e_L^0(x) > c/2$ , and therefore for large enough  $N$  we have  $\inf_x e_L^*(x) > c/4$ . Thus for large enough  $N$ , with arbitrarily high probability,  $\inf_x \hat{e}_L(x) > c/8$ . Thus, since  $\sup_x |\hat{e}_L(x) - e_L^0(x)| = O_p(\zeta(L)L^{-s/(2d)} + \zeta(L)L^{1/2}N^{-1/2})$ , it follows that

$$\left| \frac{1}{\hat{e}_L(X_i)} - \frac{1}{e_L(X_i)} \right| = O_p(\zeta(L)L^{-s/(2d)} + \zeta(L)L^{1/2}N^{-1/2}).$$

Thus

$$\begin{aligned}
& \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N (\hat{\mu}_K(X_i) - \mu_K(X_i)) \cdot \left( \frac{1}{\hat{e}_L(X_i)} - \frac{1}{e_L(X_i)} \right) \cdot (\hat{e}_L(X_i) - W_i) \right| \\
& \leq N^{1/2} \cdot \sup_x |\hat{\mu}_K(x) - \mu_K(x)| \cdot \sup_x \left| \frac{1}{\hat{e}_L(X_i)} - \frac{1}{e_L(X_i)} \right| \cdot 2 \\
& = O_p(N^{1/2} \cdot (\zeta(K)K^{-s/d} + \zeta(K)K^{1/2}N^{-1/2}) \cdot (\zeta(L)L^{-s/(2d)} + \zeta(L)L^{1/2}N^{-1/2})).
\end{aligned}$$

Next, consider (A.36).

$$\begin{aligned}
& \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \mu_K(X_i) \cdot \left( \left( \frac{1}{\hat{e}_L(X_i)} - \frac{1}{e_L(X_i)} \right) + \frac{\hat{e}_L(X_i) - e_L(X_i)}{e_L^2(X_i)} \right) \cdot (\hat{e}_L(X_i) - W_i) \right| \\
& = \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \mu_K(X_i) \cdot \left( \left( \frac{e_L(X_i) - \hat{e}_L(X_i)}{e_L(X_i)\hat{e}_L(X_i)} \right) + \frac{\hat{e}_L(X_i) - e_L(X_i)}{e_L^2(X_i)} \right) \cdot (\hat{e}_L(X_i) - W_i) \right| \\
& = \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \frac{\mu_K(X_i)}{e_L(X_i)} \cdot (e_L(X_i) - \hat{e}_L(X_i)) \cdot \left( \frac{1}{\hat{e}_L(X_i)} - \frac{1}{e_L(X_i)} \right) \cdot (\hat{e}_L(X_i) - W_i) \right| \\
& \leq N^{1/2} \cdot \sup_x |\hat{e}_L(x) - e_L^0(x)| \cdot \sup_x \left| \frac{1}{\hat{e}_L(X_i)} - \frac{1}{e_L(X_i)} \right| \cdot 2 \\
& = O_p(N^{1/2}(\zeta(K)K^{-s/d} + \zeta(K)K^{1/2}/N^{1/2})^2).
\end{aligned}$$

Next, consider (A.37).

$$\begin{aligned}
& \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \left( \frac{\mu_K(X_i)}{e_L^2(X_i)} - \frac{\mu(X_i)}{e^2(X_i)} \right) \cdot (\hat{e}_L(X_i) - e_L(X_i)) \cdot (\hat{e}_L(X_i) - W_i) \right| \\
& \leq \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \left( \frac{\mu_K(X_i)}{e_L^2(X_i)} - \frac{\mu(X_i)}{e_L^2(X_i)} \right) \cdot (\hat{e}_L(X_i) - e_L(X_i)) \cdot (\hat{e}_L(X_i) - W_i) \right| \\
& \quad + \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \left( \frac{\mu(X_i)}{e_L^2(X_i)} - \frac{\mu(X_i)}{e^2(X_i)} \right) \cdot (\hat{e}_L(X_i) - e_L(X_i)) \cdot (\hat{e}_L(X_i) - W_i) \right| \\
& \leq N^{1/2} \sup_x \left| \frac{\mu_K(X_i)}{e_L^2(X_i)} - \frac{\mu(X_i)}{e_L^2(X_i)} \right| \cdot \sup_x |\hat{e}_L(x) - e_L(x)| \cdot 2 \\
& \quad + N^{1/2} \sup_x \left| \frac{\mu(X_i)}{e_L^2(X_i)} - \frac{\mu(X_i)}{e^2(X_i)} \right| \cdot \sup_x |\hat{e}_L(x) - e_L(x)| \cdot 2 \\
& = O_p(N^{1/2}\zeta^2(K)K^{-s/d}(\zeta(L)L^{1/2}N^{-1/2} + \zeta(L)L^{-s/(2d)})) \\
& \quad + O_p(N^{1/2}\zeta(L)L^{-s/(2d)}(\zeta(L)L^{1/2}N^{-1/2} + \zeta(L)L^{-s/(2d)})).
\end{aligned}$$

Next, consider (A.38). Using a mean value theorem, we can write

$$\begin{aligned}
\hat{e}_L(X_i) - e_L^0(X_i) &= L(R_L(X_i)\hat{\pi}_L) - L(R_L(X_i)\pi_L^0) \\
&= L(R_L(X_i)\tilde{\pi}_L) \cdot (1 - L(R_L(X_i)\tilde{\pi}_L))R_L(X_i)'(\hat{\pi}_L - \pi_L^0),
\end{aligned}$$

for some intermediate value  $\tilde{\pi}_L$ . Since  $\tilde{\pi}_L$  is between  $\hat{\pi}_L$  and  $\pi_L^0$ , it follows that

$$\|\tilde{\pi}_L - \pi_L^0\| = LO_p(L^{1/2}N^{-1/2}),$$

and

$$\sup_x |L(R_L(x)\tilde{\pi}_L) - L(R_L(x)\pi_L^0)| = O_p(\zeta(L)L^{1/2}N^{-1/2}).$$

Hence

$$\begin{aligned} & \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \frac{\mu(X_i)}{e^2(X_i)} \cdot (\hat{e}_L(X_i) - e_L(X_i) - e(X_i)(1 - e(X_i)))R_L(X_i)(\hat{\pi}_L - \pi_L^0) \cdot (\hat{e}_L(X_i) - W_i) \right| \\ &= \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \frac{\mu(X_i)}{e^2(X_i)} \cdot (L(R_K(X_i)\tilde{\pi}_L)(1 - L(R_K(X_i)\tilde{\pi}_L)) - e(X_i)(1 - e(X_i)))R_L(X_i)(\hat{\pi}_L - \pi_L^0) \cdot (\hat{e}_L(X_i) - W_i) \right| \\ &\leq N^{1/2} \sup_x \left| \frac{\mu(X_i)}{e^2(X_i)} \right| \sup_x |L(R_K(X_i)\tilde{\pi}_L)(1 - L(R_K(X_i)\tilde{\pi}_L)) - e(X_i)(1 - e(X_i))| \sup_x \|R_L(x)\| \cdot \|\hat{\pi}_L - \pi_L^0\| \cdot 2 \\ &\quad = O_p(N^{1/2}\zeta(L)L^{1/2}N^{-1/2}\zeta(L)\zeta(L)L^{1/2}N^{-1/2}). \end{aligned}$$

Next, consider (A.39).

$$\begin{aligned} & \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \frac{\mu(X_i)e(X_i)(1 - e(X_i))}{e^2(X_i)} \cdot R_L(X_i)(\hat{\pi}_L - \pi_L^0) (\hat{e}_L(X_i) - e_L(X_i)) \right| \\ &\leq N^{1/2} \cdot \sup_x \frac{\mu(x)e(x)(1 - e(x))}{e^2(x)} \cdot \sup_x \|R_K(x)\| \cdot \|\hat{\pi}_L - \pi_L^0\| \cdot \sup_x |\hat{e}_L(x) - e_L^0(x)| \\ &= O_p(N^{1/2}\zeta(L)(L^{1/2}N^{-1/2} + L^{-s/(2d)})(\zeta(L)L^{1/2}N^{-1/2} + \zeta(L)L^{-s/(2d)}). \end{aligned}$$

Next, consider (A.40).

$$\begin{aligned} & \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \frac{\mu(X_i)e(X_i)(1 - e(X_i))}{e^2(X_i)} \cdot R_L(X_i)(\hat{\pi}_L - \pi_L^0) (e_L(X_i) - e(X_i)) \right| \\ &\leq N^{1/2} \cdot \sup_x \frac{\mu(x)e(x)(1 - e(x))}{e^2(x)} \cdot \sup_x \|R_K(x)\| \cdot \|\hat{\pi}_L - \pi_L^0\| \cdot \sup_x |e_L^0(x) - e(x)| \\ &= O_p(N^{1/2}\zeta(L)(L^{1/2}N^{-1/2} + L^{-s/(2d)})L^{-s/d}). \end{aligned}$$

Next, consider (A.41).

$$\begin{aligned} & \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \frac{\mu(X_i)e(X_i)(1 - e(X_i))}{e^2(X_i)} \cdot R_L(X_i)(\hat{\pi}_L - \pi_L^0) (e(X_i) - W_i) \right| \\ &\leq \|\hat{\pi}_L - \pi_L^0\| \cdot \left\| \frac{1}{\sqrt{N}} \sum_{i=1}^N \frac{\mu(X_i)e(X_i)(1 - e(X_i))}{e^2(X_i)} R_L(X_i)(e(X_i) - W_i) \right\|. \end{aligned}$$

Since

$$\mathbb{E} \left[ \left\| \frac{1}{\sqrt{N}} \sum_{i=1}^N \frac{\mu(X_i)e(X_i)(1 - e(X_i))}{e^2(X_i)} R_L(X_i)(e(X_i) - W_i) \right\|^2 \right]$$

$$\leq \sup_x \frac{\mu(x)e(x)(1-e(x))}{e^2(x)} \mathbb{E}[R_L(X_i)'R_L(X_i)] = O(L),$$

it follows that (A.41) is  $O_p(L(L^{1/2}N^{-1/2} + L^{-s/(2d)}))$ . Expression (A.29) is bounded by

$$\begin{aligned} & \|\hat{\gamma}_K - \gamma_K\| \cdot \sup_x \|R_K(x)\| \cdot N^{1/2} \cdot \sup_x \frac{1}{e_L(x)} \cdot \sup_x |\hat{e}_L(x) - e_L(x)| \\ &= O_p(N_{-1/2}\zeta(K)) \cdot O_p(\zeta(K)) \cdot N^{1/2} \cdot O_p(N^{-1/2}\xi^2(L)) = O_p(N^{-1/2}\xi^2(K)\xi^2(L)). \end{aligned}$$

(Note that  $\inf_x e_L(x) > \sup_x e(x)/2$  for  $L$  large enough, so that  $\sup_x (1/e_L(x)) = O(1)$ .) Expression (A.30) is bounded by

$$\begin{aligned} & \|\hat{\gamma}_K - \gamma_K\| \cdot \sup_x \|R_K(x)\| \cdot N^{1/2} \cdot \sup_x \frac{1}{e_L(x)} \cdot \sup_x |e_L(x) - e(x)| \\ &= O_p(N_{-1/2}\zeta(K)) \cdot O_p(\zeta(K)) \cdot N^{1/2} \cdot O_p(L^{-t/d}) = O_p(L^{-t/d}\xi^2(K)). \end{aligned}$$

Expression (A.31) is bounded by

$$\begin{aligned} & \|\hat{\gamma}_K - \gamma_K\| \cdot \sup_x \|R_K(x)\| \cdot N^{1/2} \cdot \sup_x \left| \frac{1}{e_L(x)} - \frac{1}{e(x)} \right| \\ &= O_p(N_{-1/2}\zeta(K)) \cdot O_p(\zeta(K)) \cdot N^{1/2} \cdot O_p(L^{-t/d}) = O_p(L^{-t/d}\xi^2(K)). \end{aligned}$$

Expression (A.32) is bounded by

$$\|\hat{\gamma}_K - \gamma_K\| \cdot \left\| \frac{1}{\sqrt{N}} \sum_{i=1}^N \frac{R_K(X_i)}{e(X_i)} \cdot (e(X_i) - W_i) \right\|$$

The first factor is  $O_p(N^{-1/2}\zeta(K))$ . Because

$$\mathbb{E} \left[ \left\| \frac{1}{\sqrt{N}} \sum_{i=1}^N \frac{R_K(X_i)}{e(X_i)} \cdot (e(X_i) - W_i) \right\|^2 \right] \leq C \cdot \xi^2(K),$$

it follows that the second factor is  $O_p(\zeta(K))$ . Thus (A.32) is  $O_p(N^{-1/2}\xi^2(K))$ .

Next, consider (A.33):

$$\begin{aligned} & \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \left( \frac{\mu_K(X_i)}{e_L(X_i)} - \frac{\mu(X_i)}{e(X_i)} \right) \cdot (\hat{e}(X_i) - W_i) \right| \\ &= \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \frac{\mu_K(X_i) \cdot e(X_i) - \mu(X_i) \cdot e_L(X_i)}{e_L(X_i) \cdot e(X_i)} \cdot (\hat{e}(X_i) - W_i) \right| \\ &\leq \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \left( \frac{\mu_K(X_i) \cdot e(X_i) - \mu(X_i) \cdot e_L(X_i)}{e^2(X_i)} - \frac{\mu_K(X_i) \cdot e(X_i) - \mu(X_i) \cdot e_L(X_i)}{e_L(X_i) \cdot e(X_i)} \right) \cdot (\hat{e}(X_i) - W_i) \right| \\ &\quad + \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \frac{\mu_K(X_i) \cdot e(X_i) - \mu(X_i) \cdot e_L(X_i)}{e^2(X_i)} \cdot (\hat{e}(X_i) - W_i) \right| \\ &\leq \sup_x \left| \frac{1}{e_L(x)} - \frac{1}{e(x)} \right| \cdot \frac{1}{\sqrt{N}} \sum_{i=1}^N \left| \frac{\mu_K(X_i) \cdot e(X_i) - \mu(X_i) \cdot e_L(X_i)}{e(X_i)} \cdot (\hat{e}(X_i) - W_i) \right| \end{aligned}$$

$$\begin{aligned}
& + \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \frac{\mu_K(X_i) \cdot e(X_i) - \mu(X_i) \cdot e(X_i)}{e^2(X_i)} \cdot (\hat{e}(X_i) - W_i) \right| \\
& + \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \frac{\mu(X_i) \cdot e(X_i) - \mu(X_i) \cdot e_L(X_i)}{e^2(X_i)} \cdot (\hat{e}(X_i) - W_i) \right| \\
& = O_p(L^{-t/d}) + O_p(K^{-s/d}) + O_p(L^{-t/d}).
\end{aligned}$$

so that expression (A.33) is  $O_p(L^{-t/d} + K^{-s/d})$ .

Next, consider (A.34). Because  $e(x)$  is bounded away from 0 on  $\mathbb{X}$ ,  $\mathbb{X}$  is a compact subset of  $\mathbb{R}^d$ , and  $\mu(x)$  and  $e(x)$  are  $s$  and  $t$  times continuously differentiable, it follows that  $\frac{\mu(x)}{e(x)}$  is  $\min(s, t)$  times continuously differentiable, has a finite second moment, and the projection of this random variable on  $R_L(X)$  exists. Define

$$\delta_L = \arg \min_{\delta} \mathbb{E} \left[ \left( \frac{\mu(X)}{e(X)} - R_L(X)' \delta \right)^2 \right] \quad (\text{A.42})$$

By Lorentz (), the uniform approximation error is

$$\sup_{x \in \mathbb{X}} \left| \frac{\mu(x)}{e(x)} - R_L(x)' \delta_L \right| = O \left( L^{-\frac{\min(s,t)}{d}} \right). \quad (\text{A.43})$$

Hence,

$$\left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \frac{\mu(X_i)}{e(X_i)} (\hat{e}_L(X_i) - W_i) \right| \quad (\text{A.44})$$

$$\begin{aligned}
& \leq \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \left( \frac{\mu(X_i)}{e(X_i)} - R_L(X_i)' \delta_L \right) (\hat{e}_L(X_i) - W_i) \right| + \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N R_L(X_i)' \delta_L (\hat{e}_L(X_i) - W_i) \right| \\
& = \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \left( \frac{\mu(X_i)}{e(X_i)} - R_L(X_i)' \delta_L \right) (\hat{e}_L(X_i) - W_i) \right|, \quad (\text{A.45})
\end{aligned}$$

because the second term vanishes as a result of the first order conditions for the series logit estimator which imply that  $\sum_i R_L(X_i) (\hat{e}(X_i) - W_i) = 0$ . The last expression, (A.45) can be bounded by

$$\sup_{x \in \mathbb{X}} \left| \frac{\mu(x)}{e(x)} - R_L(x)' \delta_L \right| \cdot \sum_{i=1}^N |\hat{e}(X_i) - W_i| \leq 2 \cdot \sup_{x \in \mathbb{X}} \left| \frac{\mu(x)}{e(x)} - R_L(x)' \delta_L \right| = O \left( L^{-\frac{\min(s,t)}{d}} \right),$$

because of (A.43). This proves that (A.34) is  $O(L^{-\min(s,t)/d})$ . Thus combined with (), this finishes the proof of the first assertion in Theorem 3.1.

Next, consider part (ii) of Theorem 3.1. By the triangle inequality we have

$$\begin{aligned}
\sqrt{N} \cdot \left( \hat{\beta}_{mod} - \hat{\beta}_{hir} \right) & = \frac{1}{\sqrt{N}} \sum_{i=1}^N \frac{W_i (Y_i - \hat{\mu}_K(X_i))}{\hat{e}_L(X_i)} \\
& \leq \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \frac{W_i (Y_i - \mu(X_i))}{e(X_i)} \right| \quad (\text{A.46})
\end{aligned}$$

$$+ \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N W_i (\mu(X_i) - \hat{\mu}_K(X_i)) \cdot \left( \frac{1}{\hat{e}_L(X_i)} - \frac{1}{e(X_i)} \right) \right| \quad (\text{A.47})$$



$$+ \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N W_i(Y_i - \mu_K(X_i)) \cdot \left( \frac{1}{e_L^0(X_i)} - \frac{1}{e(X_i)} \right) \right| \quad (\text{A.48})$$

$$+ \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N W_i(Y_i - \mu_K(X_i)) \cdot \left( \frac{1}{\hat{e}_L(X_i)} - \frac{1}{e_L^0(X_i)} \right) \right|. \quad (\text{A.49})$$

First consider (A.46). Because  $e(x)$  is  $s$  times continuously differentiable and bounded away from zero, it follows that there is a  $\delta_K$  such that for some finite  $C$  we can approximate  $1/e(x)$  by  $R_K(x)\delta_K$ , with

$$\sup_x \left| \frac{1}{e(x)} - R_K(x)\delta_K \right| \leq CK^{-s/d}.$$

By the first order conditions for  $\hat{\gamma}_K$  it follows that

$$\sum_{i=1}^N W_i(Y_i - \hat{\mu}_K(X_i))R_K(X_i) = 0.$$

Hence

$$\begin{aligned} & \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \frac{W_i(Y_i - \hat{\mu}_K(X_i))}{e(X_i)} \right| \\ & \leq \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N W_i(Y_i - \hat{\mu}_K(X_i)) \cdot \left( \frac{1}{e(X_i)} - R_K(X_i)\delta_K \right) \right| \\ & \quad + \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N W_i(Y_i - \hat{\mu}_K(X_i)) \cdot R_K(X_i)\delta_K \right| \\ & \leq \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N W_i(Y_i - \hat{\mu}_K(X_i)) \cdot \left( \frac{1}{e(X_i)} - R_K(X_i)\delta_K \right) \right| \\ & \leq \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N W_i(Y_i - \mu(X_i)) \cdot \left( \frac{1}{e(X_i)} - R_K(X_i)\delta_K \right) \right| \\ & \quad + \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N W_i(\mu(X_i) - \hat{\mu}_K(X_i)) \cdot \left( \frac{1}{e(X_i)} - R_K(X_i)\delta_K \right) \right| \\ & \leq N^{1/2} \frac{1}{N} \sum_{i=1}^N |W_i(Y_i - \mu(X_i))| \cdot \sup_x \left| \frac{1}{e(x)} - R_K(x)\delta_K \right| \\ & \quad + N^{1/2} \sup_x |\mu(x) - \hat{\mu}_K(x)| \cdot \sup_x \left| \frac{1}{e(x)} - R_K(x)\delta_K \right| \\ & = O_p(N^{1/2}K^{-s/d}) + O_p(N^{1/2}(\zeta(K)K^{1/2}/N^{1/2} + \zeta^2(K)K^{-s/d})K^{-s/d}) = O_p(N^{1/2}K^{-s/d}). \end{aligned}$$

Next, consider (A.47).

$$\begin{aligned} & \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N W_i(\mu(X_i) - \hat{\mu}_K(X_i)) \cdot \left( \frac{1}{\hat{e}_L(X_i)} - \frac{1}{e(X_i)} \right) \right| \\ & \leq \sqrt{N} \sup_x |\mu(x) - \hat{\mu}_K(x)| \cdot \sup_x \left| \frac{1}{\hat{e}_L(x)} - \frac{1}{e(x)} \right| \end{aligned}$$

$$= O_p(N^{1/2}(\zeta(K)K^{1/2}/N^{1/2} + \zeta^2(K)K^{-s/d})(\zeta(L)(L^{1/2}N^{-1/2} + L^{-s(2d)})).$$

Next, consider (A.48).

$$\begin{aligned} & \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N W_i(Y_i - \mu(X_i)) \cdot \left( \frac{1}{e_L^0(X_i)} - \frac{1}{e(X_i)} \right) \right| \\ & \leq \sqrt{N} \cdot \left( \frac{1}{N} \sum_{i=1}^N |W_i(Y_i - \mu(X_i))| \right) \cdot \sup_x \left| \frac{1}{e(X_i)} - \frac{1}{e_L^0(X_i)} \right| \\ & = O_p(N^{1/2}L^{-s/d}). \end{aligned}$$

Finally, consider (A.49).

$$\begin{aligned} & \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N W_i(Y_i - \mu(X_i)) \cdot \left( \frac{1}{\hat{e}_L(X_i)} - \frac{1}{e_L^0(X_i)} \right) \right| \\ & = \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N W_i(Y_i - \mu(X_i)) \cdot \frac{\hat{e}_L(X_i) - e_L^0(X_i)}{\hat{e}_L(X_i)e_L^0(X_i)} \right| \\ & \leq \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N W_i(Y_i - \mu(X_i)) \cdot \left( \frac{\hat{e}_L(X_i) - e_L^0(X_i)}{\hat{e}_L(X_i)e_L^0(X_i)} - \frac{\hat{e}_L(X_i) - e_L^0(X_i)}{e_L^0(X_i)e_L^0(X_i)} \right) \right| \end{aligned} \quad (\text{A.50})$$

$$+ \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N W_i(Y_i - \mu(X_i)) \cdot \left( \frac{\hat{e}_L(X_i) - e_L^0(X_i)}{e_L^0(X_i)e_L^0(X_i)} - \frac{1 - e(X_i)}{e(X_i)} R_L(X_i)(\hat{\pi}_L - \pi_L^0) \right) \right| \quad (\text{A.51})$$

$$+ \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N W_i(Y_i - \mu(X_i)) \cdot \frac{1 - e(X_i)}{e(X_i)} R_L(X_i)(\hat{\pi}_L - \pi_L^0) \right|. \quad (\text{A.52})$$

First, consider (A.50).

$$\begin{aligned} & \leq \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N W_i(Y_i - \mu(X_i)) \cdot \left( \frac{\hat{e}_L(X_i) - e_L^0(X_i)}{\hat{e}_L(X_i)e_L^0(X_i)} - \frac{\hat{e}_L(X_i) - e_L^0(X_i)}{e_L^0(X_i)e_L^0(X_i)} \right) \right| \\ & \leq N^{1/2} \cdot \left( \frac{1}{N} \sum_{i=1}^N |W_i(Y_i - \mu(X_i))| \right) \cdot \sup_x |\hat{e}_L(x) - e_L^0(x)| \cdot \sup_x \left| \frac{1}{e_L^0(x) \cdot \hat{e}_L(x)} - \frac{1}{e_L^0(x) \cdot e_L^0(x)} \right| \\ & = O_p(N^{1/2}(\zeta(L)(L^{1/2}N^{1/2} + L^{-s/(2d)}))(\zeta(L)(L^{1/2}N^{1/2} + L^{-s/(2d)})). \end{aligned}$$

Next, consider (A.51). Note that

$$\hat{e}_L(x) - e_L^0(x) = L(R_L(x)\hat{\pi}_L) - L(R_L(x)\pi_L^0) = L(R_L(x)\tilde{\pi}_L)(1 - L(R_L(x)\tilde{\pi}_L))R_L(x)(\hat{\pi}_L - \pi_L^0),$$

with  $\tilde{\pi}$  in between  $\hat{\pi}$  and  $\pi_L^0$ . Hence  $\|\tilde{\pi} - \pi_L^0\| = O_p(L^{1/2}N^{-1/2} + L^{-s/(2d)})$ , and  $\sup_x |L(R_L(x)\tilde{\pi}_L) - e(x)| = O_p(\zeta(L)(L^{1/2}N^{-1/2} + L^{-s/(2d)}))$ . Thus

$$\begin{aligned} & \sup |\hat{e}_L(x) - e_L^0(x) - e(x)(1 - e(x))R_L(x)(\hat{\pi}_L - \pi_L^0)| \\ & = O_p(\zeta(L)(L^{1/2}N^{-1/2} + L^{-s/(2d)})) \cdot \sup_x \|R_L(x)\| \cdot \|\hat{\pi}_L - \pi_L^0\| \\ & = O_p(\zeta^2(L)(L^{1/2}N^{-1/2} + L^{-s/(2d)})(L^{1/2}N^{-1/2} + L^{-s/(2d)})). \end{aligned}$$

Then we have

$$\begin{aligned} & \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N W_i (Y_i - \mu(X_i)) \cdot \left( \frac{\hat{e}_L(X_i) - e_L^0(X_i)}{e_L^0(X_i)e_L^0(X_i)} - \frac{1 - e(X_i)}{e(X_i)} R_L(X_i) (\hat{\pi}_L - \pi_L^0) \right) \right| \\ & \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N W_i (Y_i - \mu(X_i)) \cdot \left( \frac{\hat{e}_L(X_i) - e_L^0(X_i)}{e(X_i)e(X_i)} - \frac{1 - e(X_i)}{e(X_i)} R_L(X_i) (\hat{\pi}_L - \pi_L^0) \right) \right| \\ & + \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N W_i (Y_i - \mu(X_i)) \cdot \left( \frac{\hat{e}_L(X_i) - e_L^0(X_i)}{e_L^0(X_i)e_L^0(X_i)} - \frac{\hat{e}_L(X_i) - e_L^0(X_i)}{e(X_i)e(X_i)} \right) \right|. \end{aligned}$$

The first term is  $O_p(\zeta^2(L)(L^{1/2}N^{-1/2} + L^{-s/(2d)})(L^{1/2}N^{-1/2} + L^{-s/(2d)}))$ . The second term is bounded by

$$\begin{aligned} & \sqrt{N} \frac{1}{N} \sum_{i=1}^N |W_i(Y_i - \mu(X_i))| \cdot \sup_x |\hat{e}_L(X_i) - e_L^0(X_i)| \cdot \sup_x \left| \frac{1}{e_L^0(X_i)e_L^0(X_i)} - \frac{1}{e(X_i)e(X_i)} \right| \\ & = O_p(N^{1/2}(\zeta(L)(L^{1/2}/N^{1/2} + L^{-s/(2d)})(\zeta(L)(L^{1/2}/N^{1/2} + L^{-s/(2d)}))) \end{aligned}$$

Finally, consider (A.52).

$$\begin{aligned} & \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N W_i (Y_i - \mu(X_i)) \cdot \frac{1 - e(X_i)}{e(X_i)} R_L(X_i) (\hat{\pi}_L - \pi_L^0) \right| \\ & \leq \|\hat{\pi}_L - \pi_L^0\| \cdot \left\| \frac{1}{\sqrt{N}} \sum_{i=1}^N W_i (Y_i - \mu(X_i)) \cdot \frac{1 - e(X_i)}{e(X_i)} R_L(X_i) \right\| \end{aligned}$$

The first factor is  $O_p(\zeta(L)(L^{1/2}N^{-1/2} + L^{-s/(2d)}))$ . The expectation of the square of the second factor is

$$\begin{aligned} & \mathbb{E} \left[ \left\| \frac{1}{\sqrt{N}} \sum_{i=1}^N W_i (Y_i - \mu(X_i)) \cdot \frac{1 - e(X_i)}{e(X_i)} R_L(X_i) \right\|^2 \right] \\ & = \mathbb{E} \left[ T^2 (Y - \mu(X))^2 \left( \frac{1 - e(X)}{e(X)} \right)^2 R_L(X)' R_L(X) \right] \\ & \leq \sup_x \sigma_Y^2(x) \cdot \sup_x \left( \frac{1 - e(x)}{e(x)} \right)^2 \mathbb{E}[R_L(X)' R_L(X)] = O(L). \end{aligned}$$

Hence (A.52) is  $O_p(L\zeta(L)(L^{1/2}N^{-1/2} + L^{-s/(2d)}))$ . This finishes the proof of part (ii) of the Theorem.

Finally, consider part (iii) of the Theorem. First we prove

$$N^{1/2} \left( \sum_{i=1}^N \frac{W_i}{\hat{e}_L(X_i)} / N - 1 \right) = o_p(1). \quad (\text{A.53})$$

$$\begin{aligned} & \left| N^{1/2} \left( \sum_{i=1}^N \frac{W_i}{\hat{e}_L(X_i)} / N - 1 \right) \right| = \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \frac{W_i - \hat{e}_L(X_i)}{\hat{e}_L(X_i)} \right| \\ & \leq \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \frac{W_i - \hat{e}_L(X_i)}{e(X_i)} \right| \quad (\text{A.54}) \end{aligned}$$

$$+ \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N (e(X_i) - \hat{e}_L(X_i)) \cdot \left( \frac{1}{\hat{e}(X_i)} - \frac{1}{e(X_i)} \right) \right| \quad (\text{A.55})$$

$$+ \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N (W_i - e(X_i)) \cdot \left( \frac{1}{e_L^0(X_i)} - \frac{1}{e(X_i)} \right) \right| \quad (\text{A.56})$$

$$+ \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N (W_i - e(X_i)) \cdot \left( \frac{1}{\hat{e}_L(X_i)} - \frac{1}{e_L^0(X_i)} \right) \right|. \quad (\text{A.57})$$

We will deal with (A.54)-(A.57) one by one. First consider (A.54). There is a sequence of  $\delta_L$  such that for some finite  $C$  we have

$$\sup_x \left| \frac{1}{e(x)} - R_L(x)' \delta_L \right| \leq CL^{-s/d}.$$

Then

$$\begin{aligned} & \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \frac{W_i - \hat{e}_L(X_i)}{e(X_i)} \right| \\ & \leq \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N (W_i - \hat{e}_L(X_i)) R_L(X_i)' \delta \right| + \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N (W_i - \hat{e}_L(X_i)) \cdot \left( R_L(X_i)' \delta_L - \frac{1}{e(X_i)} \right) \right|. \end{aligned}$$

Because of the first order conditions for  $\hat{\pi}_L$  the first term vanishes. Then The second term is bounded by

$$C \cdot N^{1/2} \cdot \sup_x \left| R_L(x)' \delta_L - \frac{1}{e(x)} \right| = O(N^{1/2} L^{-s/d}) = o_p(1).$$

Next, consider (A.55).

$$\begin{aligned} & \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N (e(X_i) - \hat{e}_L(X_i)) \cdot \left( \frac{1}{\hat{e}(X_i)} - \frac{1}{e(X_i)} \right) \right| \\ & \leq N^{1/2} \cdot \sup_x |e(x) - \hat{e}_L(x)| \cdot \sup_x \left| \frac{1}{e(x)} - \frac{1}{\hat{e}_L(x)} \right| \\ & = O_p(N^{1/2} \zeta(L) (L^{1/2} N^{-1/2} + L^{-s/d}) \zeta(L) (L^{1/2} N^{-1/2} + L^{-s/d})) = o_p(1). \end{aligned}$$

Next, consider (A.56).

$$\begin{aligned} & \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N (W_i - e(X_i)) \cdot \left( \frac{1}{e_L^0(X_i)} - \frac{1}{e(X_i)} \right) \right| \\ & \leq 2 \cdot \sqrt{N} \sup_x \left| \frac{1}{e_L^0(x)} - \frac{1}{e(x)} \right| = O_p(N^{1/2} L^{-s/d}) = o_p(1). \end{aligned}$$

Finally, consider (A.57).

$$\begin{aligned} & \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N (W_i - e(X_i)) \cdot \left( \frac{1}{\hat{e}_L(X_i)} - \frac{1}{e_L^0(X_i)} \right) \right| \\ & = \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N (W_i - e(X_i)) \cdot \frac{e_L^0(X_i) - \hat{e}_L(X_i)}{e_L^0(X_i) \hat{e}_L(X_i)} \right| \end{aligned}$$

$$\begin{aligned}
&\leq \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N (W_i - e(X_i)) \cdot (e_L^0(X_i) - \hat{e}_L(X_i)) \cdot \left( \frac{1}{e_L^0(X_i) \hat{e}_L(X_i)} - \frac{1}{e^2(X_i)} \right) \right| \\
&\quad + \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N (W_i - e(X_i)) \cdot \left( \frac{e_L^0(X_i) - \hat{e}_L(X_i)}{e^2(X_i)} - \frac{1 - e(X_i)}{e(X_i)} R_L(X_i)' (\hat{\pi}_L - \pi_L^0) \right) \right| \\
&\quad + \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N (W_i - e(X_i)) \cdot \frac{1 - e(X_i)}{e(X_i)} R_L(X_i)' (\hat{\pi}_L - \pi_L^0) \right|
\end{aligned}$$

The first of these terms is of order  $O_p(N^{1/2} \zeta^2(L) (L^{1/2} N^{-1/2} + L^{-s/(2d)})^2) = o_p(1)$ . The second term is of order  $O_p(1) = o_p(1)$ . For the third term, note that

$$\begin{aligned}
&\mathbb{E} \left[ \left\| \frac{1}{\sqrt{N}} \sum_{i=1}^N (W_i - e(X_i)) \cdot \frac{1 - e(X_i)}{e(X_i)} R_L(X_i) \right\|^2 \right] \\
&= \mathbb{E} \left[ \frac{1 - e(X)}{e(X)} R_L(X)' R_L(X) \right] \leq CL
\end{aligned}$$

so that

$$\begin{aligned}
&\left| \frac{1}{\sqrt{N}} \sum_{i=1}^N (W_i - e(X_i)) \cdot \frac{1 - e(X_i)}{e(X_i)} R_L(X_i)' (\hat{\pi}_L - \pi_L^0) \right| \\
&\leq CL^{1/2} \cdot \|\hat{\pi}_L - \pi_L^0\| = O_p(L^{1/2} (L^{1/2} N^{-1/2} + L^{-s/(2d)})) = o_p(1).
\end{aligned}$$

This finishes the proof of (A.53). Next, define

$$C_N = \frac{1}{N} \sum_{i=1}^N \frac{W_i \cdot Y_i}{\hat{e}_L(X_i)} - 1,$$

so that  $\sqrt{N} \cdot C_N = o_p(1)$ , and

$$\hat{\beta}_{hir,1} - \hat{\beta}_{hir,2} = \frac{1}{N} \sum_{i=1}^N \frac{W_i \cdot Y_i}{\hat{e}_L(X_i)} \cdot \left( 1 - \frac{1}{1 + C_N} \right).$$

Hence

$$\begin{aligned}
&\sqrt{N} \cdot \left| \hat{\beta}_{hir,1} - \hat{\beta}_{hir,2} \right| \\
&\leq \left| \frac{1}{N} \sum_{i=1}^N \frac{W_i \cdot Y_i}{\hat{e}_L(X_i)} \right| \cdot \left| \frac{\sqrt{N} \cdot C_N}{1 + C_N} \right| = O_p(1) \cdot o_p(1) = o_p(1).
\end{aligned}$$

□

**Lemma A.9** *If*

$$\sup_{K \in \mathcal{K}_N} \frac{|E_N(K) - C_N(K)|}{E_N(K)} \xrightarrow{p} 0$$

*then*

$$\frac{E_N(\hat{K})}{\inf_{K \in \mathcal{K}_N} E_N(K)} \xrightarrow{p} 1$$

**Proof:** Define

$$\tilde{K} = \operatorname{argmin}_{K \in \mathcal{K}_N} E_N(K)$$

For  $N$  large enough we have that for any  $\delta, \eta > 0$

$$\Pr \left( \sup_{K \in \mathcal{K}_N} \left| \frac{C_N(K)}{E_N(K)} - 1 \right| < \delta \right) > 1 - \eta$$

Hence if  $N$  is sufficiently large then with probability of at least  $1 - \eta$

$$\frac{1 + \delta}{1 - \delta} \geq \frac{C_N(\tilde{K})}{(1 - \delta)E_N(\tilde{K})} \geq \frac{C_N(\hat{K})}{(1 - \delta)E_N(\hat{K})} \geq \frac{E_N(\hat{K})}{E_N(\tilde{K})} \geq 1$$

and the conclusion follows because  $\delta, \eta >$  are arbitrary.  $\square$

**Proof of Theorem 4.1:** We have

$$C_N(K) = E_N(K) + \frac{2}{N} \varepsilon' A_K a a' A_K \mu + \frac{1}{N} (a' A_K \varepsilon \varepsilon' A_K a - \sigma^2 a' A_K a)$$

We need to show

$$\sup_{K \in \mathcal{K}_N} \frac{\frac{1}{N} |u' A_K a a' A_K \mu|}{E_N(K)} \xrightarrow{p} 0 \tag{A.58}$$

and

$$\sup_{K \in \mathcal{K}_N} \frac{\frac{1}{N} |a' A_K u u' A_K a - \sigma^2 a' A_K a|}{E_N(K)} \xrightarrow{p} 0 \tag{A.59}$$

I consider the first statement.

We have for  $\delta > 0$

$$\begin{aligned} \Pr \left( \sup_{K \in \mathcal{K}_N} \frac{1}{N} \frac{|a' A_K \mu| |u' A_K \mu|}{E_N(K)} > \delta \right) &\leq \sum_{K \in \mathcal{K}_N} \Pr \left( \frac{1}{N} \frac{|a' A_K \mu| |u' A_K \mu|}{E_N(K)} > \delta \right) \leq \\ &\leq \sum_{K \in \mathcal{K}_N} \frac{|a' A_K \mu|^m E(|u' A_K a|^m)}{\delta^m N^m E_N(K)^m} \end{aligned}$$

By Whittle's theorem

$$E(|u' A_K a|^m) \leq C (a' A_K a)^{\frac{m}{2}}$$

We have

$$a' A_K a \leq C N K^{-2s_p/d}$$

Also

$$(a' A_K \mu)^2 \leq N E_N(K)$$

Combining this we find

$$\sum_{K \in \mathcal{K}_N} \frac{|a' A_K \mu|^m E(|u' A_K a|^m)}{\delta^m N^m E_N(K)^m} \leq \sum_{K \in \mathcal{K}_N} C \frac{K^{-ms_p/d}}{E_N(K)^{\frac{m}{2}}}$$

so that we require that (compare with Li)

$$\sum_{K \in \mathcal{K}_N} \frac{1}{K^{ms_p/d} E_N(K)^{\frac{m}{2}}} \rightarrow 0$$

We have

$$\sum_{K \in \mathcal{K}_N} \frac{1}{K^{ms_p/d} E_N(K)^{\frac{m}{2}}} \leq \frac{1}{\inf_{K \in \mathcal{K}_N} E_N(K)^{\frac{m}{2}}} \sum_{K \in \mathcal{K}_N} K^{-ms_p/d}$$

If we take the dimension of  $\mathcal{K}_N$  as  $N^\kappa$ , this holds if (compare with Li)

$$N^{\kappa(ms_p/d-1)} \inf_{K \in \mathcal{K}_N} E_N(K)^{\frac{m}{2}} \rightarrow \infty \tag{A.60}$$

so that optimal MSE cannot decrease too fast with  $N$ .

We now consider (A.59). We have

$$\begin{aligned} & \Pr \left( \sup_{K \in \mathcal{K}_N} \frac{|u' A_K a a' A_K u - \sigma^2 a' A_K a|}{N E_N(K)} > \delta \right) \leq \\ & \leq \sum_{K \in \mathcal{K}_N} \Pr \left( \frac{|u' A_K a a' A_K u - \sigma^2 a' A_K a|}{N E_N(K)} > \delta \right) \leq \\ & \leq \sum_{K \in \mathcal{K}_N} \frac{E(|u' A_K a a' A_K u - \sigma^2 a' A_K a|^m)}{\delta^m N^m E_N(K)^m} \leq \\ & \leq \sum_{K \in \mathcal{K}_N} \frac{C \text{tr}((A_K a a' A_K)^2)^{\frac{m}{2}}}{\delta^m N^m E_N(K)^m} \leq \sum_{K \in \mathcal{K}_N} \frac{C(a' A_K a)^m}{\delta^m N^m E_N(K)^m} \end{aligned}$$

so that we require

$$\sum_{K \in \mathcal{K}_N} \frac{1}{K^{2ms_p/d} E_N(K)^m} \rightarrow 0$$

or

$$N^{\kappa(2ms_p/d-1)} \inf_{K \in \mathcal{K}_N} E_N(K)^m \rightarrow \infty$$

which is implied by (A.60).  $\square$

## REFERENCES

- ANDREWS, D. (1991): "Asymptotic Optimality of Generalized  $C_L$ , Cross-validation, and Generalized Cross-validation in Regression with Heteroskedastic Errors," *Journal of Econometrics*, 47, 359-377.
- BARNOW, B.S., G.G. CAIN AND A.S. GOLDBERGER (1980), "Issues in the Analysis of Selectivity Bias," in *Evaluation Studies*, vol. 5, ed. by E. Stromsdorfer and G. Farkas. San Francisco: Sage.
- CHEN, X., HONG, H., AND TAROZZI, A., (2004), "Semiparametric Efficiency in GMM Models of Nonclassical Measurement Error, Missing Data and Treatment Effects." Working Paper.

- HAHN, J., (1998), "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica* 66 (2), 315-331.
- HECKMAN, J., AND R. ROBB, (1984), "Alternative Methods for Evaluating the Impact of Interventions," in Heckman and Singer (eds.), *Longitudinal Analysis of Labor Market Data*, Cambridge, Cambridge University Press.
- HECKMAN, J., H. ICHIMURA, AND P. TODD, (1998), "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies* 65, 261-294.
- HECKMAN, J., H. ICHIMURA, J. SMITH, AND P. TODD, (1998), "Characterizing Selection Bias Using Experimental Data," *Econometrica* 66, 1017-1098.
- HIRANO, K., G. IMBENS, AND G. RIDDER (2000), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score. " NBER Working Paper.
- HOLLAND, P. (1986), "Statistics and Causal Inference," *Journal of the American Statistical Association*, 81, 945-970.
- ICHIMURA, H., AND O. LINTON, (2001), "Asymptotic Expansions for some Semiparametric Program Evaluation Estimators." Institute for Fiscal Studies, cemmap working paper cwp04/01.
- LI, K. (1987), "Asymptotic Optimality for  $C_p$ ,  $C_L$ , Cross-validation and Generalized Cross-validation: Discrete Index Set," *Annals of Statistics*, 15(3), 958-975.
- NEWBY, W., (1994), "The Asymptotic Variance of Semiparametric Estimators," *Econometrica* 62, 1349-1382.
- NEWBY, W. (1995), "Convergence Rates for Series Estimators," in *Advances in Econometrics and Quantitative Economics: Essays in Honor of Professor C. R. Rao*, Maddala, Phillips, and Srinivasan (eds.), Cambridge, Basil Blackwell.
- NEWBY, W., AND D. MCFADDEN (1994), "Large Sample Estimation," in *Handbook of Econometrics*, Vol. 4, Engle and McFadden (eds.), North Holland.
- ROBINS, J., A. ROTNITZKY, AND L. ZHAO (1995), "Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data," *Journal of the American Statistical Association*, 90 (429), 106-121.
- ROSENBAUM, P., AND D. RUBIN, (1983a), "The Central Role of the Propensity Score in Observational Studies for Causal Effects", *Biometrika*, 70, 41-55.
- ROTNITZKY, A., AND J. ROBINS (1995), "Semiparametric Regression Estimation in the Presence of Dependent Censoring," *Biometrika* 82 (4), 805-820.
- RUBIN, D. (1974), "Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies," *Journal of Educational Psychology*, 66, 688-701.
- RUBIN, D. B., (1978), "Bayesian inference for causal effects: The Role of Randomization", *Annals of Statistics*, 6:34-58.