# Complementarity and Aggregate Implications of Assortative Matching: A Nonparametric Analysis*

Bryan S. Graham† Guido W. Imbens‡ Geert Ridder§

First Draft: July 2004
This version: February 2009

**Abstract**

This paper presents methods for evaluating the effects of reallocating an indivisible input across production units. When the a production technology is nonseparable, such reallocations, although leaving the marginal distribution of the reallocated input unchanged by construction, may nonetheless alter average output. Examples include reallocations of teachers across classrooms composed of students of varying mean ability, and altering assignment mechanisms for college roommates in the presence of social interactions. We focus on the effects of reallocating one input, while holding the assignment of another, potentially complementary input, fixed. We present a class of such reallocations – correlated matching rules – that includes the status quo allocation, a random allocation, and both the perfect positive and negative assortative matching allocations as special cases. Our econometric approach involves first nonparametrically estimating the production function and then averaging this function over the distribution of inputs induced by the new assignment rule. These methods build upon the partial mean literature (e.g., Newey 1994, Linton and Nielsen 1995). We derive the large sample properties of our proposed estimators and assess their small sample properties via a limited set of Monte Carlo experiments.

**JEL Classification: C14, C21, C52**

**Keywords:** *Aggregate Redistributional Effects, Complementarity, Nonparametric Estimation, Assortative Matching*

*Financial support for this research was generously provided through NSF grant SES 0136789, SES 0452590, and SES 0820361. We thank participants in the Harvard-MIT and Brown Econometrics Seminars for comments. We thank Cristine Pinto for excellent research assistance.

†Department of Economics, University of California at Berkeley, 665 Evans Hall, Berkeley, CA 94720-3880. E-Mail: bgraham@econ.berkeley.edu, Web: http://www.econ.berkeley.edu/~bgraham/

‡Department of Economics, Harvard University, 1805 Cambridge Street, Cambridge, MA 02138, and NBER. Electronic correspondence: imbens@harvard.edu, http://www.economics.harvard.edu/faculty/imbens/imbens.html.

§Department of Economics, University of Southern California, 310A Kaprielian Hall, Los Angeles, CA 90089. E-Mail: ridder@usc.edu, Web: http://www-rcf.usc.edu/~ridder/

# 1 Introduction

Consider a production function depending on a number of inputs. We are interested in the effect of a particular input on output, and specifically in the effects of policies that change the allocation of this input to the different firms. For each firm output may be monotone in this input, but at different rates. If the input is indivisible, and its aggregate stock fixed, it will be impossible to simultaneously raise the input level for all firms. In such cases it may be of interest to consider the output effects of *reallocations* of the input across firms. Here we investigate econometric methods for assessing the effect on average output of such reallocations. We will call the causal effects of such policies Aggregate Redistributional Effects (AREs). A key feature of reallocations is that, although they potentially alter input levels for each firm, they keep the marginal distribution of the input across the population of firms fixed.

The first contribution of our paper is to introduce a framework for considering such reallocations, and to define estimands that capture their key features. These estimands include the effects of focal reallocations and a parametric class of reallocations, as well as the effect of a local reallocation. One focal reallocation redistributes the input across firms such that it has perfect rank correlation with a second input, the positive assortative matching allocation. We also consider a negative assortative matching allocation where the input is redistributed to have perfect negative rank correlation with the second input. A third allocation involves randomly assigning the input across firms. This allocation, by construction, ensures independence of the two inputs. A fourth allocation simply maintains the status quo assignment of the input. More generally we consider a two parameter family of feasible reallocations that include these four focal allocations as special cases. Reallocations in this family may depend on the distribution of a second input or firm characteristic. This characteristic may be correlated with the firm-specific return to the input to be reallocated. Our family of reallocations, called correlated matching rules, includes each of the above allocations as special cases. In particular the family traces a path from the positive to negative assortative matching allocations. Each reallocation along this path keeps the marginal distribution of the two inputs fixed, but is associated with a different level of correlation between the two inputs. Each of the reallocations we consider are members of a general class of reallocation rules that keep the marginal distributions of the two inputs fixed. We also provide a local measure of complementarity. In particular we consider whether small steps away from the status quo and toward the perfect assortative matching allocation raise average output.

The second contribution of our paper is to derive methods for estimation and inference for the estimands considered. We derive an estimator for average output under all correlated matching allocations. Our estimator requires that the first input is exogenous conditional on the second input and additional firm characteristics. Except for the case of perfect negative and positive rank correlation the estimator has the usual parametric convergence rate. For the two extremes the rate of convergence is slower, comparable to that of estimating a regression function with a scalar covariate at a point. In all cases we drive the asymptotic distribution of the estimator.

Our focus on reallocation rules that keep the marginal distribution of the inputs fixed is appropriate in applications where the input is indivisible, such as in the allocation of teachers to classes or managers to production units. In other settings it may be more appropriate to consider allocation rules that leave the total amount of the input constant by fixing its average level. Such rules would require some modification of the methods considered in this paper.

Our methods may be useful in a variety of settings. One class of examples concerns complementarity of inputs in production functions (e.g. Athey and Stern, 1998). If the first and second inputs are everywhere complements, then the difference in average output between the positive and negative assortative matching allocations provides a nonparametric measure of the degree of complementarity. This measure is invariant to monotone transformations of the inputs. If the production function is not supermodular interpretation of this difference is not straightforward, although it still might be viewed as some sort of 'global' measure of input complementarity. With this concern in mind

A second example concerns educational production functions. Card and Krueger (1992) study the relation between educational output as measured by test scores and teacher quality. Teacher quality may improve test scores for all students, but average test scores may be higher or lower depending on whether, given a fixed supply of teachers, the best teachers are assigned to the least prepared students or vice versa. Parents concerned solely with outcomes for their own children may be most interested in the effect of raising teacher quality on expected scores. A school board, however, may be more interested in maximizing expected test scores given a fixed set of classes and teachers by optimally matching teachers to classes.

A third class of examples arises in settings with social interaction (c.f., Manski 1993; Brock and Durlauf 2001). Sacerdote (2001) studies the peer effects in college by looking at the relation between outcomes and roommate characteristics. From the perspective of the individual student or her parents it may again be of interest whether a roommate with different characteristics would, in expectation, lead to a different outcome. This is what Manski (1993) calls an exogenous or contextual effect. The college, however, may be interested in a different effect, namely the effect on average outcomes of changing the procedures for assigning roommates. While it may be very difficult for a college to change the distribution of characteristics in the incoming classes, it may be possible to change the way roommates are assigned. In Graham, Imbens and Ridder (2006b) we consider average effect of segregation policies.

If production functions are additive in inputs the questions posed above have simple answers: average outcomes are invariant to input reallocations. While reallocations may raise individual outcomes for some units, they will necessarily lower them by an offsetting amount for others. Reallocations are zero-sum games. With additive and linear functions even more general assignment rules that allow the marginal input distribution to change while keeping its average level unchanged do not affect average outcomes. In order for these questions to have interesting answers, one therefore needs to explicitly recognize and allow for non-additivity and non-linearity of a production function in its inputs. For this reason our approach is fully nonparametric.

The current paper builds on the larger treatment effect and program evaluation literature.[1] More directly, it is complementary to the small literature on the effect of treatment assignment rules (Manski, 2004; Dehejia, 2004; Hirano and Porter, 2005). Our focus is different from that in the Manski, Dehejia, and Hirano-Porter papers. First, we allow for continuous rather than discrete or binary treatments. Second, our assignment policies to do not change the marginal distribution of the treatment, whereas in the previous papers treatment assignment for one unit is not restricted by assignment for other units. Our policies are fundamentally redistributions. In the current paper we focus on estimation and inference for specific assignment rules. It is also interesting to consider optimal rules as in Manski, Dehejia and Hirano-Porter. The class of feasible reallocations/redistributions includes all joint distributions of the two inputs with

[1] For recent surveys see Angrist and Krueger (2001), Heckman, Lalonde and Smith (2001), and Imbens (2004).

fixed marginal distributions. When the inputs are continuously-valued, as we assume in the current paper, the class potential rules is very large. Characterizing the optimal allocation within this class is therefore a non-trivial problem. When both inputs are discrete-valued the problem with finding the optimal allocation is tractable as the joint distribution of the inputs is characterized by a finite number of parameters. In Graham, Imbens and Ridder (2006a) we consider optimal allocation rules when both inputs are binary, allowing for general complementarity or substitutability of the inputs.

Our paper is also related to recent work on identification and estimation of models of social interactions (e.g., Manski 1993, Brock and Durlauf 2001). We do not focus on directly characterizing the within-group structure of social interactions, an important theme of this literature. Rather our goal is simply to estimate the average relationship between group composition and outcomes. The average we estimate may reflect endogenous behavioral responses by agents to changes in group composition, or even equal an average over multiple equilibria. Viewed in this light our approach is reduced form in nature. However it is sufficient for, say, an university administrator to characterize the outcome effects of alternative roommate assignment procedures.

The econometric approach taken here builds on the partial mean literature (e.g., Newey, 1994; Linton and Nielsen, 1995). In this literature one first estimates a regression function nonparametrically. In the second stage the regression function is averaged, possibly after some weighting with a known or estimable weight function, over some of the regressors. Similarly here we first estimate a nonparametric regression function of the outcome on the input and other characteristics. In the second stage the averaging is over the distribution of the regressors induced by the new assignment rule. This typically involves the original marginal distribution of some of the regressors, but a different conditional distribution for others. Complications arise because this conditional covariate distribution may be degenerate, which will affect the rate of convergence for the estimator. In addition the conditional covariate distribution itself may require nonparametric estimation through its dependence on the assignment rule. For the policies we consider the assignment rule will involve distribution functions and their inverses similar to the way these enter in the changes-in-changes model of Athey and Imbens (2005).

The next section lays out our basic model and approach to identification. Section 3 then defines and motivates the estimands we seek to estimate. Section 4 presents of our estimators, and derives their large-sample properties, for the case where inputs are continuously-valued. 5 presents an application and the results of a small Monte Carlo exercise.

# 2 Model

In this section we present the basic set up and identifying assumptions. For clarity of exposition we use the production function terminology; although our methods are appropriate for a wide range of applications, as emphasized in the introduction. For firm $i$, for $i = 1, \ldots, N$, the production function relates a triple of observed inputs, $(W_i, X_i, Z_i)$, and an unobserved input $\varepsilon_i$, to an output $Y_i$:

$$Y_i = h(W_i, X_i, Z_i, \varepsilon_i). \tag{2.1}$$

The inputs $W_i$ and $X_i$, and the output $Y_i$ are scalars. The third observed input $Z_i$ and the unobserved input $\varepsilon_i$ can both be vectors. We are interested in reallocating the input $W$ across firms. We focus upon reallocations which hold the marginal distribution of $W$ fixed. As such they are appropriate for settings where $W$ is a plausibly indivisible input, such as a manager or teacher with a certain level of experience and expertise. The presumption is also that the aggregate stock of $W$ is difficult to augment. In addition to $W$ there are two other (observed) firm characteristics that may affect output: $X$ and $Z$, where $X$ is a scalar and $Z$ is a vector of dimension $K$. The first characteristic $X$ could be a measure of, say, the quality of the long-run capital stock, with $Z$ being other characteristics of the firm such as location and age. These characteristics may themselves be inputs that can be varied, but this is not necessary for the arguments that follow. In particular the exogeneity assumption that we make for the first input need not hold for these characteristics.

We observe for each firm $i = 1, \ldots, N$ the level of the input, $W_i$, the characteristics $X_i$ and $Z_i$, and the realized output level, $Y_i = Y_i(W_i)$. In the educational example the unit of observation would be a classroom. The variable input $W$ would be teacher quality, and $X$ would be a measure of quality of the class, e.g., average test scores in prior years. The second characteristic $Z$ could include other measures of the class, e.g., its age or gender composition, as elements. In the roommate example the unit would be the individual, with $W$ the quality of the roommate (measured by, for example, a high school test score), and the characteristic $X$ would be own quality. The second set of characteristics $Z$ could be other characteristics of the dorm or of either of the two roommates such as smoking habits (which may be used by university administrators in the assignment of roommates).

Our identifying assumption is that conditional on firm characteristics $(X, Z')'$ the assignment of $W$, the level of the input to be reallocated, is exogenous.

**Assumption 2.1** (Exogeneity)

$$\varepsilon \;\perp\; W \;\Big|\; X, Z.$$

Define

$$g(w, x, z) = \mathbb{E}[Y|W = w, X = x, Z = z],$$

denote the average output associated with input level $w$ and characteristics $x$ and $z$. Under unconfoundedness we have – among firms with identical values of $X$ and $Z$ – an equality between the counterfactual average output that we would observe if all firms in this subpopulation were assigned $W = w$, and the average output we observe for the subset of firms within this subpopulation that are in fact assigned $W = w$. Alternatively, the exogeneity assumption implies that the difference in $g(w, x, z)$ evaluated at two values of $w$, $w_0$ and $w_1$, has a causal interpretation as the average effect of assigning $W = w_1$ rather than $W = w_0$:

$$g(w_1, x, z) - g(w_0, x, z) = \mathbb{E}\left[h(w_1, X, Z, \varepsilon) - h(w_0, X, Z, \varepsilon) \middle| X = x, Z = z\right].$$

Assumption 2.1 has proved controversial (c.f., Imbens 2004). It holds under conditional random assignment of $W$ to units; as would occur in a randomized experiment. However randomized allocation mechanisms are also used by administrators in some institutional settings. For example some universities match freshman roommates randomly conditional on responses in a housing questionnaire (e.g., Sacerdote 2001). This assignment mechanism is consistent with Assumption 2.1. In other settings, particularly where assignment is bureaucratic, as may be true in some educational settings, a plausible set of conditioning variables may be available. In this paper we focus upon identification and estimation under Assumption 2.1. In principle, however, the methods could be extended to accommodate other approaches to identification based upon, for example, instrumental variables.

Much of the treatment effect literature (e.g., Angrist and Krueger, 2000; Heckman, Lalonde and Smith, 2000; Manski, 1990; Imbens, 2004) has focused on the average effect of an increase in the value of the treatment. In particular, in the binary treatment case ($w \in \{0, 1\}$) interest has centered on the average treatment effect

$$\mathbb{E}[g(1, X, Z) - g(0, X, Z)].$$

With continuous inputs one may be interested in the full average output function $g(w, x, z)$ (Imbens, 2000; Flores, 2005) or in its derivative with respect to the input,

$$\frac{\partial g}{\partial w}(w, x, z), \qquad \text{or} \quad \mathbb{E}\left[\frac{\partial g}{\partial w}(W, X, Z)\right],$$

either at a point or averaged over some distribution of inputs and characteristics (e.g., Powell, Stock and Stoker, 1989; Hardle and Stoker, 1989).

Here we are interested in a fundamentally different estimand, that has appeared to have received no attention in the econometrics literature. We focus on policies that redistribute the input $W$ according to a rule based on the $X$ characteristic of the unit. For example upon assignment mechanisms that match teachers of varying experience to classes of students based on their mean ability. One might assign those teachers with the most experience (highest values of $W$) to those classrooms with the highest ability students (highest values of $X$) and so on. In that case average outcomes would reflect perfect rank correlation between $W$ and $X$. Alternatively, we could be interested in the average outcome if we were to assign $W$ to be negatively perfectly rank correlated with $X$. A third possibility is to assign $W$ so that it is independent of $X$. We are interested in the effect of such policies on the average value of the output. We refer to such effects as Aggregate Redistributional Effects (AREs). The three reallocations mentioned are a special case of a general set of reallocation rules that fix the marginal distributions of $W$ and $X$, but allow for correlation in their joint distribution. For perfect assortative matching the correlation is 1, for negative perfect assortative matching -1, and for random allocation 0. By using a bivariate normal cupola we can trace out the path between these extremes.

We wish to emphasize that there are at least two limitations to our approach. First, we focus on comparing specific assignment rules, rather than searching for the optimal assignment rule within a class. The latter problem is a particularly demanding problem in the current setting with continuously-valued inputs as the optimal assignment for each unit depends both on the characteristics of that unit as well as on the marginal distribution of characteristics in the population. When the inputs are discrete-valued both the problems of inference for a specific rule as well as the problem of finding the optimal rule become considerably more tractable. In that case any rule, corresponding to a joint distribution of the inputs, is characterized by a finite number of parameters. Maximizing estimated average output over all rules evaluated will then generally lead to the optimal rule. Graham, Imbens and Ridder (2006a) and, motivated by an early version of the current paper, Bhattacharya (2008), provide a discussion for the case with discrete covariates.

A second limitation is that of this class of assignment rules leaves the marginal distribution of inputs unchanged. This latter restriction is perfectly appropriate in cases where the inputs are indivisible, as, for example, in the social interactions and educational examples. In other cases one need not be restricted to such assignment rules. A richer class of estimands would allow for assignment rules that maintain some aspects of the marginal distribution of inputs but not others. A particularly interesting class consists of assignment rules that maintain the average (and thus total) level of the input, but allow for its arbitrary distribution across units. This can be interpreted as assignment rules that 'balance the budget'. In such cases one might assign the maximum level of the input to some subpopulation and the minimum level of the input to the remainder of the population. Finally, one may wish to consider arbitrary decision rules where each unit can be assigned any level of the input within a set. In that case interesting questions include both the optimal assignment rule as a function of unit-level characteristics as well as average outcomes of specific assignment rules. In the binary treatment case such problems have been studied by Dehejia (2005), Manski (2004), and Hirano and Porter (2005).

We consider the following four estimands that include four benchmark assignment rules. All leave the marginal distribution of inputs unchanged. This obviously does not exhaust the possibilities within this class. Many other assignment rules are possible, with corresponding estimands. However, the estimands we consider here include focal assignments, indicate of the range of possibilities, and capture many of the methodological issues involved.

# 3 Aggregate Redistributional Effects

Let $h_{W|X,Z}(w|x,z)$ denote a conditional distribution of $W$ given $(X,Z)$. The distribution in the data will be denoted by $f_{W|X,Z}(w|x,z)$. We will allow $h_{W|X,Z}(w|x,z)$ to correspond to a degenerate distribution. In general we are interested in the average outcome that would result from the current distribution of $(X,Z,\varepsilon)$, if the distribution of $W$ given $(X,Z)$ were changed from its current distribution, $f_{W|X,Z}(w|x,z)$ to $h_{W|X,Z}(w|x,z)$. We denote the expected output given such a reallocation by

$$\beta_h^{\text{are}} = \int g(w,x,z) h_{W|X,Z}(w|x,z) f_{X,Z}(x,z) dw dx dz. \tag{3.2}$$

In the next two sections we discuss some specific choices for $h(\cdot)$.

## 3.1 Positive and Negative Assortive Matching Allocations

The first estimand we consider is expected average outcome given perfect assortative matching of $W$ on $X$ conditional on $Z$:

$$\beta^{\text{pam}} = \mathbb{E}[g(F_{W|Z}^{-1}(F_{X|Z}(X|Z)|Z), X, Z)], \tag{3.3}$$

where $F_{X|Z}(X|Z)$ denotes the conditional CDF of $X$ given $Z$ and $F_{W|Z}^{-1}(q|Z)$ is the quantile of order $q \in [0,1]$ associated with the conditional distribution of $W$ given $Z$ (i.e., $F_{W|Z}^{-1}(q|Z)$ is a conditional quantile function). Therefore $F_{W|Z}^{-1}(F_{X|Z}(X|Z)|Z)$ computes a unit's location on the conditional CDF of $X$ given $Z$ and reassigns it the corresponding quantile of the conditional distribution of $W$ given $Z$. Thus among units with the same realization of $Z$, those with the highest value of $X$ are reassigned the highest value of $W$ and so on.

The focus on reallocations within subpopulations defined by $Z$, as opposed to population-wide reallocations, is because the average outcome effects of such reallocations solely reflect complementarity or substitutability between $W$ and $X$.

To see why this is the case consider the alternative estimand

$$\beta^{\text{pam2}} = \mathbb{E}\left[g\left(F_W^{-1}(F_X(X)), X, Z\right)\right]. \tag{3.4}$$

This gives average output associated with population-wide perfect assortative matching of $W$ on $X$. If, for example, $X$ and $Z$ are correlated, then this reallocation, in addition to altering the joint distribution of $W$ and $X$, will alter the joint distribution of $W$ and $Z$. Say $Z$ is also a scalar and is positively correlated with $X$. Population-wide positive assortative matching will induce perfect rank correlated between $W$ and $X$, but it will also increase the degree of correlation between $W$ and $Z$. This complicates interpretation when $g(w, x, z)$ may be non-separable in $w$ and $z$ as well as $w$ and $x$.

An example helps to clarify the issues involved. Let $W$ denote an observable measure of teacher quality, $X$ mean (beginning-of-year) achievement in a classroom, and $Z$ the fraction of the classroom that is female. If begining-of-year achievement varies with gender, (say, with classes with a higher fraction of girls having higher average achievement) then $X$ and $Z$ will be correlated. A reallocation that assigns high quality teachers to high achievement classrooms, will also tend to assign such teachers to classrooms will an above average fraction of females. Average achievement increases observed after implementing such a reallocation may reflect complementarity between teacher quality and begining-of-year student achievement or it may be that the effects of changes in teacher quality vary with gender and that, conditional on gender, their is no complementarity between teacher quality and achievement. By focusing on reallocations of teachers across classrooms with similar gender mixes, but varying baseline achievement, (3.3) provides a more direct avenue to learning about complementarity.[2]

Both (3.3) and (3.4) may be policy relevant, depending on the circumstances, and both are identified under Assumption 2.1. Under the additional assumption that

$$g(w, x, z) = g_1(w, x) + g_2(z),$$

the estimands, although associated with different reallocations, also have the same basic interpretation. Here we focus upon (3.3), although all of our results extend naturally and directly to (3.4).

Our second estimand is the expected average outcome given negative assortative matching:

$$\beta^{\text{nam}} = \mathbb{E}[g(F_{W|Z}^{-1}(1 - F_{X|Z}(X|Z)|Z), X, Z)]. \tag{3.5}$$

If, within subpopulations homogenous in $Z$, $W$ and $X$ are everywhere complements, then the difference $\beta^{\text{pam}} - \beta^{\text{nam}}$ provides a measure of the strength input complementarity. When $g(\cdot)$ is not supermodular interpretation of this difference is not straightforward. In Section **??** below we present a measure of 'local' (relative to the status quo allocation) complementarity between $X$ and $W$.

### 3.2 Correlated Matching Allocations

Average output under the *status quo* allocation is given by

$$\beta^{\text{sq}} = \mathbb{E}[Y] = \mathbb{E}[g(W, X, Z)],$$

[2] We make the connection to complementarity more explicit in Section **??** below.

while average output under the random matching allocation is given by

$$\beta^{\mathrm{rm}} = \int_z \left[ \int_x \int_w g(w,x,z) \mathrm{d}F_{W|Z}(w|z) \mathrm{d}F_{X|Z}(x|z) \right] \mathrm{d}F_Z(z).$$

This last estimand gives average output when $W$ and $X$ are independently assigned within subpopulations.

The perfect positive and negative assortative allocations are focal allocations, being emphasized in theoretical research (e.g., Becker and Murphy 2000). The status quo and random matching allocations are similarly natural benchmarks. However these allocations are just four among the class of feasible allocations. This class is comprised of all joint distributions of inputs consistent with fixed marginal distributions (within subpopulations homogenous in $Z$). As noted in the introduction, if the inputs are continuously distributed this class of joint distributions is very large. For this reason we only consider a subset of these joint distributions. To be specific, we concentrate on a two-parameter subset of the feasible allocations that include as special cases the negative and positive assortative matching allocations, the independent allocation, and the status quo allocation. By changing the two parameters we trace out a 'path' in two directions: further from or closer to the status quo allocation, and further from, or closer to, the perfect sorting allocations. Borrowing a term from the literature on cupolas, we call this class of feasible allocations comprehensive, because it contains all four focal allocations as a special case.

For the purposes of estimation, the correlated matching allocations are redefined using a truncated bivariate normal cupola. The truncation ensures that the denominator in the weights of the correlated matching ARE are bounded from 0, so that we do not require trimming. The bivariate standard normal PDF is

$$\phi(x_1, x_2; \rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}(x_1^2 - 2\rho x_1 x_2 + x_2^2)}, \qquad -\infty < x_1, x_2 < \infty$$

with a corresponding joint CDF denoted by $\Phi(x_1, x_2; \rho)$. Observe that

$$\mathrm{pr}(-c < x_1 \leq c, -c < x_2 \leq c) = \Phi(c,c;\rho) - \Phi(c,-c;\rho) - [\Phi(-c,c;\rho) - \Phi(-c,-c;\rho)],$$

so that the truncated standard bivariate normal PDF is given by

$$\phi_c(x_1, x_2; \rho) = \frac{\phi(x_1, x_2; \rho)}{\Phi(c,c;\rho) - \Phi(c,-c;\rho) - [\Phi(-c,c;\rho) - \Phi(-c,-c;\rho)]}, \qquad -c < x_1, x_2 \leq c.$$

Denote the truncated bivariate CDF by $\Phi_c$.

The truncated normal bivariate CDF gives a comprehensive cupola, because the corresponding joint CDF

$$H_{W,X}(w,x) = \Phi_c\left(\Phi_c^{-1}(F_W(w)), \Phi_c^{-1}(F_X(x)); \rho\right)$$

has marginal CDFs equal to $H_{W,X}(w,\infty) = F_W(w)$ and $H_{W,X}(\infty, x) = F_X(x)$, it reaches the upper and lower Fréchet bounds on the joint CDF for $\rho = 1$ and $\rho = -1$, respectively, and it has independent $W, X$ as a special case for $\rho = 0$.

To obtain an estimate of $\beta^{\mathrm{cm}}(\rho, \tau)$ we note that joint PDF associated with $H_{W,X}(w,x)$ equals

$$h_{W,X}(w,x) = \phi_c\left(\Phi_c^{-1}(F_W(w)), \Phi_c^{-1}(F_X(x)); \rho\right) \frac{f_W(w) f_X(x)}{\phi_c\left(\Phi_c^{-1}(F_W(w))\right) \phi_c\left(\Phi_c^{-1}(F_X(x))\right)},$$

and hence that $\beta^{\mathrm{cm}}(\rho,0)$, redefined in terms of the truncated normal, is given by

$$\beta^{\mathrm{cm}}(\rho,0)=\int_{x,z,w} g(w,x,z)\frac{\phi_c\left(\Phi_c^{-1}(F_W(w)),\Phi_c^{-1}(F_X(x));\rho\right)}{\phi_c\left(\Phi_c^{-1}(F_W(w))\right)\phi_c\left(\Phi_c^{-1}(F_X(x))\right)}f_W(w)f_{X,Z}(x,z)\mathrm{d}w\mathrm{d}x\mathrm{d}z.$$

Average output under the correlated matching allocation is given by

$$\begin{aligned}\beta^{\mathrm{cm}}(\rho,\tau)&=\tau\cdot\mathbb{E}[Y]+(1-\tau)\cdot\beta^{\mathrm{cm}}(\rho,0)\\&=\tau\cdot\mathbb{E}[Y_i]+(1-\tau)\\&\quad\times\int_{x,z,w} g(w,x,z)\frac{\phi_c\left(\Phi_c^{-1}(F_W(w)),\Phi_c^{-1}(F_X(x));\rho\right)}{\phi_c\left(\Phi_c^{-1}(F_W(w))\right)\phi_c\left(\Phi_c^{-1}(F_X(x))\right)}f_W(w)f_{X,Z}(x,z)\mathrm{d}w\mathrm{d}x\mathrm{d}z,\end{aligned}\tag{3.6}$$

for $\tau\in[0,1]$ and $\rho\in(-1,1)$.

The case with $\tau=1$ corresponds to the *status quo*:

$$\beta^{\mathrm{sq}}=\beta^{\mathrm{cm}}(\rho,1)$$

The case with $\tau=\rho=0$ corresponds to random allocation of inputs within sub-populations defined by $Z$:

$$\beta^{\mathrm{rm}}=\beta^{\mathrm{cm}}(0,0)=\int_z\left[\int_x\int_w g(w,x,z)\mathrm{d}F_{W|Z}(w|z)\mathrm{d}F_{X|Z}(x|z)\right]\mathrm{d}F_Z(z).$$

Although the cases with $\tau=0$ and $\rho\to1$ and $-1$ correspond respectively to the perfect positive and negative assortative matching allocations:

$$\beta^{\mathrm{pam}}=\lim_{\rho\to1}\beta^{\mathrm{cm}}(\rho,0),\qquad\text{and}\quad\beta^{\mathrm{nam}}=\lim_{\rho\to-1}\beta^{\mathrm{cm}}(\rho,0).$$

More generally, with $\tau=0$ we allocate the inputs using a normal copula in a way that allows for arbitrary correlation between $W$ and $X$ indexed by the parameter $\rho$. In principle we could use other copulas.

### 3.3 Local Measures of Complementarity

A potential problem with the correlated matching reallocation family of estimands $\beta(\rho,\tau)$ is that the support requirements that allow for precise estimation may be difficult to satisfy in practice. This is particularly relevant for allocations 'distant' from the status quo. For example, if the status quo is characterized by a high degree of correlation between the inputs, evaluating the effect of allocations with a small or even negative correlation between inputs, such as random allocation or negative assortative matching can be difficult because such allocations rely on knowledge of the production at pairs of values $(W,X)$ that are infrequently seen in the data. For this reason a measure of local (close to the status quo) complementarity between $W$ and $X$ would be valuable. To this end we next characterize the expected effect on output associated with a 'small' increase toward either positive or negative assortative matching. The resulting estimand forms the basis of a simple test for local efficiency of the status quo allocation. W derive this local measure by considering matching on a family of transformations of $X_i$ and $W_i$, indexed by a scalar parameter $\lambda$, where for some values of $\lambda$ the matching is on $W_i$ (corresponding to the status quo), and for other values of $\lambda$ the matching is on $X_i$ or $-X_i$, corresponding to positive and negative assortative matching respectively. We then focus on the derivative of the expected outcomes from matching on this family of transformations, evaluated at the value of $\lambda$ that corresponds to the status quo.

For technical reasons, and to be consistent with the subsequent formal statistical analysis in Section 4 of the previously discussed estimands $\beta^{\text{pam}}$ and $\beta^{\text{nam}}$ we assume that the support of $X_i$ is the interval $[x_l, x_u]$, with midpoint $x_m = (x_u + x_l)/2$, and similarly that the support of $W_i$ is the interval $[w_l, w_u]$, with midpoint $w_m = (w_u + w_l)/2$. Without loss of generality we will assume that $x_l = 0$, $x_m = 1/2$, $x_u = 1$, $w_l = 0$, $w_m = 1/2$, and $w_u = 1$. To focus on the key issues we also ignore the presence of additional covariates $Z_i$. First define a smooth function $d(w)$ that goes to zero at the boundary of the support of $W_i$:

$$d(w) = 1_{w>w_m} \cdot (w_u - w) + 1_{w\leq w_m} \cdot (w - w_l).$$

We implement our local reallocation as follows: for $\lambda \in [-1, 1]$, define the random variable $U_\lambda$ as a transformation of $(X, W)$:

$$U_\lambda = \lambda \cdot X \cdot d(W)^{1-|\lambda|} + (\sqrt{1 - \lambda^2}) \cdot W.$$

The average output associated with positive assortative matching on $U_\lambda$ is given by

$$\beta^{\text{lr}}(\lambda) = \mathbb{E}[g(F_W^{-1}(F_{U_\lambda}(U_\lambda)), X)]. \tag{3.7}$$

For $\lambda = 0$ and $\lambda = 1$ we have $U_\lambda = W$ and $U_\lambda = X$ respectively and hence $\beta^{\text{lr}}(0) = \beta^{\text{sq}}$ and $\beta^{\text{lr}}(1) = \beta^{\text{pam}}$. Perfect negative assortative matching is also nested in this framework since

$$\text{pr}\,(-X \leq -x) = \text{pr}\,(X \geq x) = 1 - F_X\,(x)\,,$$

and hence for $\lambda = -1$ we have $\beta^{\text{lr}}(-1) = \beta^{\text{nam}}$. Values of $\lambda$ close to zero induce reallocations of $W$ that are 'local' to the status quo, with $\lambda > 0$ and $\lambda < 0$ generating shifts toward positive and negative assortative matching respectively.

We focus on the effect of such a small reallocation as our local measure of complementarity:

$$\beta^{\text{lc}} = \frac{\partial \beta^{\text{lr}}}{\partial \lambda}(0). \tag{3.8}$$

This local complementarity measure has two alternative representations which are given in the following Theorem. Before stating this we introduce one assumption. This assumption is stronger than needed for this theorem, but its full force will be used later.

**Assumption 3.1** (Distribution of Data)
*(i)* $(Y_1, W_1, X_1), (Y_2, W_2, X_2), \ldots,$ *are independent and identically distributed,*
*(ii) The support of* $W$ *is* $\mathbb{W} = [w_l, w_u]$*, a compact subset of* $\mathbb{R}$*,*
*(iii) the support of* $X$ *is* $\mathbb{X} = [x_l, x_u]$*, a compact subset of* $\mathbb{R}$*,*
*(iv) the joint distribution of* $W$ *and* $X$ *is bounded and bounded away from zero, and* $q$ *times continuously differentiable on* $\mathbb{W} \times \mathbb{X}$*,*
*(v)* $g(w, x)$ *is* $q$ *times continuously differentiable with respect to* $w$ *and* $x$ *on* $\mathbb{W} \times \mathbb{X}$*,*
*(vi)* $\mathbb{E}[|Y_i|^p | X_i = x]$ *is bounded.*

**Theorem 3.1** *Suppose Suppose Assumption 3.1 holds. Then, $\beta^{\text{lc}}$ has two equivalent representations:*

$$\beta^{\text{lc}} = \mathbb{E}\left[\frac{\partial g}{\partial w}(W_i, X_i) \cdot \Big(X_i \cdot d(W_i) - \mathbb{E}\left[X_i \cdot d(W_i) | \, W_i\right]\Big)\right], \tag{3.9}$$

*and,*

$$\beta^{\text{lc}} = \mathbb{E}\left[\delta(W_i, X_i) \cdot \frac{\partial^2 g}{\partial w \partial x}(W_i, X_i)\right], \tag{3.10}$$

*where the weight function $\delta(w, x)$ is non-negative and has the form*

$$\delta(w,x) = d(w) \cdot \frac{F_{X|W}(x|w) \cdot (1 - F_{X|W}(x|w))}{f_{X|W}(x|w)} \cdot \Big(\mathbb{E}\left[X_i | X_i > x, W_i = w\right] - \mathbb{E}\left[X_i | X_i \leq x, W_i = w\right]\Big).$$

The proofs for the Theorems in the body of the text are given in Appendix C.

Representation (3.9), as we demonstrate below, suggests a straightforward method-of-moments approach to estimating $\beta_0^{\text{lc}}$. Representation (3.10) is valuable for interpretation. Equation (3.10) demonstrates that a test of $H_0 : \beta^{\text{lc}} = 0$ is a test of the the null of no complementarity or substitutability between $W$ and $X$. If $\beta^{\text{lc}} > 0$, then in the 'vicinity of the status quo' $W$ and $X$ are complements; if $\beta^{\text{lc}} < 0$ they are substitutes. The precise meaning of the 'vicinity of the status quo' is implicit in the form of the weight function $\delta(w, z)$.

Deviations of $\beta^{\text{lc}}$ from zero imply that the status quo allocation does not maximize average outcomes. For $\beta^{\text{lc}} > 0$ a shift toward positive assortative matching will raise average outcomes, while for $\beta^{\text{lc}} < 0$ a shift toward negative assortative matching will do so. Theorem 3.1 therefore provides the basis of a test of the null hypothesis that the status quo allocation is locally efficient.

# 4 Estimation and inference with continuously-valued inputs

In this section we discuss estimation and inference. We focus on the case without additional exogenous covariates. First, in Section 4.1 we describe the nonparametric estimators for the regression functions. In order to deal with boundary issues we use new kernel estimator. Next, in Section **??** we lay out the key assumptions. Then, in Section 3.1 we present estimators for the first pair of estimands, $\beta^{\text{pam}}$ and $\beta^{\text{nam}}$. In Section 4.3 we discuss estimation and inference for $\beta^{\text{cm}}$ (including $\beta^{\text{rm}}$), and in Section 4.4 we discuss $\beta^{\text{lc}}$. Estimation of and inference for the status quo allocation $\beta^{\text{sq}}$ is straightforward, as this estimand is a simple expectation, estimated by a sample average.

## 4.1 Estimating the Production and Distribution Functions

For the two distributions functions we use the empirical distribution functions:

$$\hat{F}_W(w) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{W_i \leq w}, \qquad \text{and} \quad \hat{F}_X(x) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{X_i \leq x}.$$

For the inverse distribution functions we use the definition:

$$\hat{F}_W^{-1}(q) = \inf_{w \in \mathbb{W}} \mathbf{1}_{\hat{F}_W(w) \geq q}, \qquad \text{and} \quad \hat{F}_X^{-1}(q) = \inf_{x \in \mathbb{X}} \mathbf{1}_{\hat{F}_X(x) \geq q}.$$

The estimands we consider in this paper depend on the regression function $g(w,x)$ (in the case of $\beta^{\rm pam}$, $\beta^{\rm nam}$, $\beta^{\rm cm}$), or its derivative (in the case of $\beta^{\rm lc}$, which also depends on the regression function $m(w)=\mathbb{E}[X_i|W_i=w]$). In order to estimand these objects, we need estimators for the regression functions and their derivatives. Write the regression function as

$$g(w,x)=\mathbb{E}\left[Y|W=w,X=x\right]=\frac{h_2(w,x)}{h_1(w,x)},$$

where

$$h_1(w,x)=f_{WX}(w,x),\qquad \text{and}\quad h_2(w,x)=g(w,x)\cdot f_{WX}(w,x).$$

To simplify the following discussion, we rewrite $h_1(w,x)$ and $h_2(w,x)$ as

$$h_m(w,x)=\mathbb{E}[V_{im}|W_i=w,X_i=x]\cdot f_{WX}(w,x),$$

for $m=1,2$, where $V_{i1}=1$, $V_{i2}=Y_i$.

We focus on estimators for $h_m(w,x)$, and use those to estimate $g(w,x)$ and its derivatives. The standard Nadaraya-Watson estimator for $h_m(w,x)$ is

$$\hat{h}_{{\rm nw},m}(w,x)=\frac{1}{N\cdot b^2}\sum_{i=1}^{N}V_{im}\cdot K\left(\frac{W_i-w}{b},\frac{X_i-x}{b}\right).$$

We denote the resulting nonparametric estimators by $\hat{g}(w,x)$. We estimate the derivative of $g(w,x)$ with respect to $w$ by taking the derivative of the estimator of $g(w,x)$.

Because the support of $(W_i,X_i)$ is assumed to be bounded, we have to deal with boundary bias of the kernel estimators. Because we also need bias reduction by using higher order kernels we adopt the Nearest Interior Point (NIP) estimator of Imbens and Ridder (2006). This estimator divides, for given bandwidth $b$, the support of $(W,X)$ into an internal part and a boundary part. On the internal part the uniform convergence of the kernel estimators holds, but the estimators must be modified on the boundary part of the support. The NIP estimator coincides with the usual Nadaraya-Watson (NW) kernel estimator on the internal set, but it is equal to a polynomial on the boundary set. The coefficients of this polynomial are those of a Taylor series expansion in a point of the internal set.

To obtain a compact expression for the NIP estimator we adopt the following notation. The vector $z=(w\ x)'$ has $L=2$ components. Let $\mathbb{Z}$ denote the (compact) support of $Z$. Let $\lambda$ denote an $L$ vector of nonnegative integers, with $|\lambda|=\sum_{l=1}^{L}\lambda_l$, and $\lambda!=\prod_{l=1}^{L}\lambda_l!$. For $L$ vectors of nonnegative integers $\lambda$ and $\mu$ let $\mu\leq\lambda$ be equivalent to $\mu_l\leq\lambda_l$ for all $l=1,\ldots,L$, and define

$$\begin{pmatrix}\lambda\\ \mu\end{pmatrix}=\frac{\lambda!}{\mu!(\lambda-\mu)!}=\prod_{l=1}^{L}\frac{\lambda_l!}{\mu_l!(\lambda_l-\mu_l)!}=\prod_{l=1}^{L}\begin{pmatrix}\lambda_l\\ \mu_l\end{pmatrix}.$$

For $L$ vectors $\lambda$ and $z$, let $z^\lambda=\prod_{l=1}^{L}z_l^{\lambda_l}$. As shorthand for partial derivatives of some function $g$ we use $g^{(\lambda)}(z)$:

$$g^{(\lambda)}(z)=\frac{\partial g^{|\lambda|}}{\partial s^\lambda}(z).$$

The definition of the internal set depends on the support of the kernel. Let $K : \mathbb{R}^L \mapsto \mathbb{R}$ denote the kernel function. We will assume that $K(u) = 0$ for $u \notin \mathbb{U}$ with $\mathbb{U}$ compact, and $K(u)$ bounded. For the bandwidth $b$ define the internal set of the support $\mathbb{Z}$ as the subset of $\mathbb{Z}$ such that all $\tilde{z}$ with a distance of up to $b$ times the support of the kernel from $z$ are also in $\mathbb{Z}$

$$\mathbb{Z}_b^I = \left\{ z \in \mathbb{Z} \,\middle|\, \left\{ \tilde{z} \in \mathbb{R}^L \,\middle|\, \frac{z - \tilde{z}}{b} \in \mathbb{U} \right\} \subset \mathbb{Z} \right\} = \{ z \in \mathbb{Z} \mid \{ z - b \cdot u | u \in \mathbb{U} \} \subset \mathbb{Z} \}. \tag{4.11}$$

This is a compact subset of the interior of $\mathbb{Z}$ that contains all points that are sufficiently far away from the boundary that the standard kernel density estimator at those points is not affected by any potential discontinuity of the density at the boundary. If $\mathbb{U} = [-1, 1]^L$ and $\mathbb{Z} = \bigotimes_{l=1}^L [\underline{z}_l, \overline{z}_l]$, we have $\mathbb{Z}_b^I = \bigotimes_{l=1}^L [\underline{z}_l + b, \overline{z}_l - b]$.[3]

Next, we need to develop some notation for Taylor series approximations. Define for a given, $q$ times differentiable function $g : \mathbb{R}^L \mapsto \mathbb{R}$, a point $r \in \mathbb{R}^L$ and an integer $p \leq q$, the $(p-1)$-th order polynomial function $t : \mathbb{Z} \mapsto \mathbb{R}$ based on the Taylor series expansion of order $p - 1$ of $g(\cdot)$ around $r$:

$$t(z; g, r, p) = \sum_{j=0}^{p-1} \sum_{|\lambda|=j} \frac{1}{\lambda!} \cdot g^{(\lambda)}(r) \cdot (z - r)^\lambda. \tag{4.12}$$

Because the function $g$ is $q \geq p$ times continuously differentiable the remainder term in the Taylor series expansion is

$$g(z) - t(z, g, r, p) = \sum_{|\lambda|=p} \frac{1}{\lambda!} g^{(\lambda)}(\overline{r}(s)) \cdot (z - r)^\lambda.$$

with $\overline{r}(z)$ intermediate between $z$ and $r$. Because $\mathbb{Z}$ is compact, and the the $p$-th order continuous, the $p$th order derivative must be bounded, and therefore this remainder term is bounded by $C|z - r|^p$. For the NIP estimator we use this Taylor series expansion around a point that depends on $z$ and the bandwidth. Specifically, we take the expansion around $r_b(z)$, the projection on the internal set

$$r_b(z) = \text{argmin}_{r \in \mathbb{Z}_b^I} \|z - r\|$$

With this preliliminary discussion, the NIP estimator of order $p$ of $h_m(z)$ can be defined:

$$\hat{h}_{m,NIP,p}(z) = \sum_{j=0}^{p-1} \sum_{|\lambda|=j} \frac{1}{\lambda!} \cdot \hat{h}_{m,NW}^{(\mu)}(r_b(z))(z - r_b(z))^\mu \tag{4.13}$$

with $\hat{h}_{m,NW}^{(\lambda)}$ the $\lambda$-th derivative of the kernel estimator $\hat{h}_{m,NW}$. For values of $z \in \mathbb{Z}_b^I$, the NIP estimator is identical to the NW kernel estimator, $\hat{h}_{m,NIP,p}(z) = \hat{h}_{m,NW}(z)$. It is only in the boundary region that a $p$-th order Taylor series expansion is used to address the poor properties of the NS estimator in that region.

Now the NIP estimator for $g(w, x)$ is

$$\hat{g}_{NIP,p}(w, x) = \frac{\hat{h}_{2,NIP,p}(w, x)}{\hat{h}_{1,NIP,p}(w, x)},$$

[3] The set $[-1, 1]^L$ is the set of $L$ vectors with components that are between -1 and 1. The set $\bigotimes_{l=1}^L [\underline{z}_l, \overline{z}_l]$ is the set of $L$ vectors with $l$-th component between $\underline{z}_l$ and $\overline{z}_l$.

and the NIP estimator for the first derivative of $g(w,x)$ with respect to $w$ is

$$\frac{\widehat{\partial g_{NIP,p}}}{\partial w}(w,x) = \frac{\frac{\partial}{\partial w}\hat{h}_{2,NIP,p}(w,x)}{\hat{h}_{1,NIP,p}(w,x)} - \frac{\hat{h}_{2,NIP,p}(w,x)\cdot\frac{\partial}{\partial w}\hat{h}_{1,NIP,p}(w,x)}{\left(\hat{h}_{1,NIP,p}(w,x)\right)^2}.$$

Unlike the NW kernel estimator, the NIP estimator is uniformly consistent. Its properties are discussed in more detail in Imbens and Ridder (2008). A formal statement of the relevant properties for our discussion is given in Lemmas A.20, A.21, and A.22, and Theorems A.1, A.2, and A.3 in Appendix A.

In the remainder of the paper we drop the subscripts from the estimator of the regression function. Unless specifically mentioned, $\hat{g}(w,x)$ will be used to denote $\hat{g}_{NIP,p}(w,x)$.

Next we introduce two more assumptions. Assumption 4.1 describes the properties of the kernel function, and Assumption 4.2 gives the rate on the bandwidth. Before stating the next assumption we need to introduce a class of restrictions on kernel functions. The restrictions govern the rate at which the kernel, which is assumed to have compact support), goes to zero on the boundary of its support. This property allows us to deal with some of the boundary issues. Such properties have previously been used in, for example, Powell, Stock and Stoker (1989).

**Definition 4.1** (Derivative Order of a Kernel) *A Kernel $K:\mathbb{U}\mapsto\mathbb{R}$ is of derivative order $d$ if for all $u$ in the boundary of the set $\mathbb{U}$ and all $|\lambda|\leq d-1$,*

$$\lim_{v\to u}\frac{\partial^\lambda}{\partial u^\lambda}K(v)=0.$$

**Assumption 4.1** (Kernel)
*(i) $K:\mathbb{R}^2\to\mathbb{R}$, with $K(u)=\prod_{l=1}^2\mathcal{K}(u_l)$,*
*(ii) $K(u)=0$ for $u\notin\mathbb{U}$, with $\mathbb{U}=[-1,1]^2$,*
*(iii) $K$ is $r$ times continuously differentiable, with the $r$-th derivative bounded on the interior of $\mathbb{U}$,*
*(iv) $K$ is a kernel of order $s$, so that $\int_{\mathbb{U}}K(u)du=1$ and $\int_{\mathbb{U}}u^\lambda K(u)du=0$ for all $\lambda$ such that $0<|\lambda|<s$, for some $s\geq 1$,*
*(v) $K$ is a kernel of derivative order $d$.*

We refer a kernel satisfying Assumption 4.2 as a derivative kernel of order $(s,d)$.

**Assumption 4.2** (Bandwidth) *The bandwidth $b_N=N^{-\delta}$ for some $\delta>0$.*

### 4.2 Estimation and Inference for $\widehat{\beta}^{\rm pam}$ and $\widehat{\beta}^{\rm nam}$

In this section we introduce the estimators for $\beta^{\rm pam}$ and $\beta^{\rm nam}$ and present results on the large sample properties of the estimators. We estimate $\beta^{\rm pam}$ and $\beta^{\rm nam}$ by substituting nonparametric estimators for the unknown functions $g(w,x)$, $F_W(w)$, and $F_X(x)$:

$$\widehat{\beta}^{\rm pam}=\frac{1}{N}\sum_{i=1}^N\hat{g}\left(\hat{F}_W^{-1}(\hat{F}_X(X_i)),X_i\right),\qquad\text{and}\quad\widehat{\beta}^{\rm nam}=\frac{1}{N}\sum_{i=1}^N\hat{g}\left(\hat{F}_W^{-1}(1-\hat{F}_X(X_i)),X_i\right),$$

The formulation of the large sample properties of these estimators is somewhat complicated. There are four components to its asymptotic approximation. The first component corresponds to the estimation error in $g(w, x)$. This component converges at a rate slower the regular parametric (root-$N$) rate. This is because we estimate a nonparametric regression function with more arguments than we average over in the second stage, so that $\hat{\beta}^{\text{pam}}$ and $\hat{\beta}^{\text{nam}}$ are a partial means in the terminology of Newey (1994). The other three terms converge slightly faster, at the regular root$-N$ rate. There is one term each corresponding to the estimation error in $F_W(w)$ and $F_X(x)$ respectively, and one corresponding to the difference between the average of $g(F_W^{-1}(F_X(X_i, X_i))$ and its expectation. In describing the large sample properties we include all four of these terms, which leaves a remainder that is $o_p(N^{-1/2})$. In principle one could ignore the three terms of order $O_p(N^{-1/2})$, since they will get dominated by the term describing the uncertainty stemming from estimation of $g(w, x)$, but including the additional terms is likely to lead to more accurate confidence intervals. In order to describe the formal properties it is useful to introduce notation for an intermediate quantity. Define:

$$\tilde{\beta}^{\text{pam}} = \frac{1}{N}\sum_{i=1}^{N} g\left(\hat{F}_W^{-1}\left(\hat{F}_X(X_i)\right), X_i\right), \tag{4.14}$$

so that we can write $\hat{\beta}^{\text{pam}} - \beta^{\text{pam}} = (\hat{\beta}^{\text{pam}} - \tilde{\beta}^{\text{pam}}) + (\tilde{\beta}^{\text{pam}} - \beta^{\text{pam}})$. Then $\tilde{\beta}^{\text{pam}} - \beta^{\text{pam}} = O_p(N^{-1/2})$, and the remaining term $\hat{\beta}^{\text{pam}} - \tilde{\beta}^{\text{pam}} = O_p(N^{-1/2} b_N^{-1/2})$.

To describe the large sample properties of $\hat{\beta}^{\text{pam}}$ we need a couple of additional objects. Define

$$g_W(w, x) = \frac{\partial g}{\partial w}(w, x),$$

$$q_{WX}(w, x) = \frac{g_W(F_W^{-1}(F_X(x)), x)}{f_W(F_W^{-1}(F_X(x)))} \cdot \left(\mathbf{1}_{F_W(w) \le F_X(x)} - F_X(x)\right),$$

$$q_W(w) = \mathbb{E}[q_{WX}(w, X)],$$

$$r_{XZ}(x, z) = \frac{g_w(F_W^{-1}(F_X(z)), z)}{f_W(F_W^{-1}(F_X(z)))} \cdot \left(\mathbf{1}_{x \le z} - F_X(z)\right),$$

and

$$r_X(x) = \mathbb{E}[r_{XZ}(x, X)].$$

**Theorem 4.1** (Large Sample Properties of $\hat{\beta}^{\text{pam}}$)
*Suppose Assumptions 2.1, 3.1, 4.1, and 4.2 hold with $q \ge 2s+1$, $r \ge s+3$, $p \ge 4$, $d \ge s-1$, and $1/(2s+1) < \delta < 1/6$. Then*

$$\sqrt{N} \cdot \begin{pmatrix} b_N^{1/2}\left(\hat{\beta}^{\text{pam}} - \tilde{\beta}^{\text{pam}}\right) \\ \tilde{\beta}^{\text{pam}} - \beta^{\text{pam}} \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Omega_{11}^{\text{pam}} & 0 \\ 0 & \Omega_{22}^{\text{pam}} \end{pmatrix}\right),$$

*where*

$$\Omega_{11}^{\text{pam}} = \mathbb{E}\Bigg[\sigma^2\left(F_W^{-1}\left(F_X(X_i)\right), X_i\right) \cdot \int_{u_1}\left(\int_{u_2} K\left(u_1 + \frac{f_X(X_i)}{f_W\left(F_W^{-1}\left(F_X(X_i)\right)\right)} \cdot u_2, u_2\right) du_2\right)^2 du_1$$
$$\cdot f_{W|X}\left(F_W^{-1}\left(F_X(X_i)\right)|X_i\right)\Bigg],$$

*and*

$$\Omega_{22}^{\rm pam} = \mathbb{E}\left[\left(q_W(W_i) + r_X(X_i) + g(F_W^{-1}(F_X(X_i),X_i)) - \beta^{\rm pam}\right)^2\right].$$

Suppose we wish to construct a 95% confidence interval for $\beta^{\rm pam}$. In that case we approximate the variance of $\hat\beta^{\rm pam}-\beta^{\rm pam}$ by $\hat{\mathbb{V}} = \hat\Omega_{11}^{\rm pam}\cdot N^{-1}\cdot b_N^{-1}+\hat\Omega_{22}^{\rm pam}\cdot N^{-1}$, using suitable plug-in estimators $\hat\Omega_{11}^{\rm pam}$ and $\hat\Omega_{22}^{\rm pam}$, and construct the confidence interval as $(\hat\beta^{\rm pam} - 1.96\cdot\sqrt{\hat{\mathbb{V}}}, \hat\beta^{\rm pam} + 1.96\cdot\sqrt{\hat{\mathbb{V}}})$. Although the first term in $\hat{\mathbb{V}}$ will dominate the second term in large samples, in finite samples the second term may still be important.

Similar results hold for $\beta^{\rm nam}$, with some appropriately redefined concepts. Define

$$\tilde\beta^{\rm nam} = \frac{1}{N}\sum_{i=1}^N g\left(\hat F_W^{-1}\left(1-\hat F_X(X_i)\right), X_i\right), \tag{4.15}$$

$$q_{WX}^{\rm nam}(w,x) = \frac{g_W(F_W^{-1}(1-F_X(x)),x)}{f_W(F_W^{-1}(1-F_X(x)))}\cdot\left(\mathbf{1}_{F_W(w)\le F_X(x)} - F_X(x)\right),$$

$$q_W^{\rm nam}(w) = \mathbb{E}\left[q_{WX}^{\rm nam}(w,X)\right],$$

$$r_{XZ}^{\rm nam}(x,z) = \frac{g_w(F_W^{-1}(1-F_X(z)),z)}{f_W(F_W^{-1}(1-F_X(z)))}\cdot\left(\mathbf{1}_{x\le z} - F_X(z)\right),$$

and

$$r_X^{\rm nam}(x) = \mathbb{E}\left[r_{XZ}^{\rm nam}(x,X)\right].$$

**Theorem 4.2** (Large Sample Properties of $\hat\beta^{\rm nam}$)
*Suppose Assumptions 2.1, 3.1, 4.1, and 4.2 hold with $q \ge 2s+1$, $r\ge s+3$, $p\ge 4$, $d\ge s-1$, and $1/(2s+1) < \delta < 1/6$. Then*

$$\sqrt{N}\cdot\begin{pmatrix} b_N^{1/2}\left(\hat\beta^{\rm nam} - \tilde\beta^{\rm nam}\right) \\ \tilde\beta^{\rm nam} - \beta^{\rm nam}\end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\Omega_{11}^{\rm nam} & 0 \\ 0 & \Omega_{22}^{\rm nam}\end{pmatrix}\right),$$

*where*

$$\Omega_{11}^{\rm nam} = \mathbb{E}\Bigg[\sigma^2\left(F_W^{-1}\left(1-F_X(X)\right),X\right)\cdot\int_{u_1}\left(\int_{u_2} K\left(u_1 + \frac{f_X(X)}{f_W\left(F_W^{-1}\left(1-F_X(X)\right)\right)}\cdot u_2, u_2\right)\right)^2 du_1 \cdot f_{W|X}\left(F_W^{-1}\left(1-F_X(X)\right)|X\right)\Bigg],$$

*and*

$$\Omega_{22}^{\rm nam} = \mathbb{E}\left[\left(q_W^{\rm nam}(W) + r_X^{\rm nam}(X) + g(W,X) - \beta^{\rm nam}\right)^2\right].$$

### 4.3 Estimation and Inference for $\beta^{\text{cm}}(\rho, \tau)$

Replacing the integrals with sums over the empirical distribution we get the analog estimator

$$\widehat{\beta}^{\text{cm}}(\rho,0) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\hat{g}(W_i,X_j)\frac{\phi_c\left(\Phi_c^{-1}(\hat{F}_W(W_i)),\Phi_c^{-1}(\hat{F}_X(X_j));\rho\right)}{\phi_c\left(\Phi_c^{-1}(\hat{F}_W(W_i))\right)\phi_c\left(\Phi_c^{-1}(\hat{F}_X(X_j))\right)}.$$

Observe that if $\rho = 0$ (independent matching) the ratio of densities on the right hand side is equal to 1.

For $\tau > 0$, the $\beta^{\text{cm}}(\rho, \tau)$ estimand is a convex combination of average output under the status quo and a correlated matching allocation. The corresponding sample analog is

$$\widehat{\beta}^{\text{cm}}(\rho,\tau) = \tau\cdot\widehat{\beta}^{\text{sq}} + (1-\tau)\cdot\widehat{\beta}^{\text{cm}}(\rho,0).$$

This estimator is linear in the nonparametric regression estimator $\hat{g}$ and nonlinear in the empirical CDFs of $X$ and $W$. This structure simplifies the asymptotic analysis.

A useful and insightful representation of $\beta^{\text{cm}}(\rho, 0)$ is as an average of partial means (c.f., Newey 1994). This representation provides intuition both about the structure of the estimand as well as its large sample properties. Fixing $W$ at $W = w$ but averaging over the joint distribution of $X$ and $Z$ we get the partial mean:

$$\eta\left(w\right) = \mathbb{E}_X\left[g(w,X)\cdot d(w,X)\right], \tag{4.16}$$

where

$$d(w,x) = \frac{\phi_c(\Phi_c^{-1}(F_W(w)),\Phi_c^{-1}(F_X(x));\rho)}{\phi_c(\Phi_c^{-1}(F_W(w)))\phi_c(\Phi_c^{-1}(F_X(x)))}. \tag{4.17}$$

Observe that (4.16) is a weighted averaged of the production function over the joint distribution of $X$ and $Z$ holding the value of the input to be reallocated $W$ fixed at $W = w$. The weight function $d(w, X)$ depends upon the truncated normal cupola. In particular, the weights give greater emphasis to realizations of $g(w, X, Z)$ that are associated with values of $X$ that will be assigned a value of $W$ close to $w$ as part of the correlated matching reallocation. Thus (4.16) equals the average post-reallocation output for those firms being assigned $W = w$. To give a concrete example (4.16) is the post-reallocation expected achievement of those classrooms that will be assigned a teacher of quality $W = w$.

Equation (4.16) also highlights the value of using the truncated normal copula. Doing so ensures that the denominators of the copula 'weights' in (4.16) are bounded from zero. The copula weights thus play the role similar to fixed trimming weights used by Newey (1994).

If we average these partial means over the marginal distribution of $W$ we get $\beta^{\text{cm}}(\rho, 0)$, since

$$\beta^{\text{cm}}(\rho,0) = \mathbb{E}_W\left[\eta\left(W\right)\right],$$

yielding average output under the correlated matching reallocation.

From the above discussion it is clear that our correlated matching estimator can be viewed as a semiparametric two-step method-of-moments estimator with a moment function of

$$m(Y,W,\beta^{\text{cm}}(\rho,\tau),\eta(W)) = \tau Y + (1-\tau)\,\eta\left(W\right) - \beta^{\text{cm}}(\rho,\tau).$$

Our estimator, $\widehat{\beta}^{\mathrm{cm}}(\rho,\tau)$, is the feasible GMM estimator based upon the above moment function after replacing the partial mean (4.16) with a consistent estimate. While the above representation is less useful for deriving the asymptotic properties of $\widehat{\beta}^{\mathrm{cm}}(\rho,\tau)$ it does provide some insight as to why we are able to achievement parametric rates of convergence.

$$\begin{aligned}
e_W(w,x) &= \frac{\rho\phi_c(\Phi_c^{-1}(F_W(w)),\Phi_c^{-1}(F_X(x));\rho)}{(1-\rho^2)\phi_c(\Phi_c^{-1}(F_W(w)))^2\phi_c(\Phi_c^{-1}(\widehat{F}_X(x)))}\times \\
&\quad \left[\Phi_c^{-1}(F_X(x)) - \rho\Phi_c^{-1}(F_W(w))\right] \\
e_X(w,x) &= \frac{\rho\phi_c(\Phi_c^{-1}(F_W(w)),\Phi_c^{-1}(F_X(x));\rho)}{(1-\rho^2)\phi_c(\Phi_c^{-1}(F_W(w)))\phi_c(\Phi_c^{-1}(\widehat{F}_X(x)))^2}\times \\
&\quad \left[\Phi_c^{-1}(F_W(w)) - \rho\Phi_c^{-1}(F_X(X_k))\right]. \\
\omega(w,x) &= \frac{\phi_c\left(\Phi_c^{-1}(F_W(w)),\Phi_c^{-1}(F_X(x));\rho\right)}{\phi_c\left(\Phi_c^{-1}(F_W(w))\right)\phi_c\left(\Phi_c^{-1}(F_X(x))\right)} \\
\psi_0(y,w,x) &= \left(\mathbb{E}\left[g(W,x)\cdot\omega(W,x)\right] - \beta^{\mathrm{cm}}(\rho,0)\right) + \left(\mathbb{E}\left[g(w,X)\cdot\omega(w,X)\right] - \beta^{\mathrm{cm}}(\rho,0)\right) \\
\psi_g(y,w,x) &= \frac{f_W(w)\cdot f_X(x)}{f_{WX}(w,x)}(y-g(w,x))\omega(w,x) \\
\psi_W(y,w,x) &= \int\int g(s,t)e_W(s,t)\left(1\left(w\le s\right) - F_W(s)\right)f_W(s)\,f_X(t)\,\mathrm{d}s\mathrm{d}t. \\
\psi_X(y,w,x) &= \int\int g(s,t)e_X(s,t)\left(1\left(x\le t\right) - F_X(t)\right)f_W(s)\,f_X(t)\,\mathrm{d}s\mathrm{d}t.
\end{aligned}$$

**Theorem 4.3** *Suppose Assumptions 2.1, 3.1, 4.1, and 4.2 hold with* $q\ge 2s-1$*,* $r\ge s+1$*,* $p\ge 3$*,* $d\ge s-1$*, and* $(1/2s)<\delta<1/4$*, then*

$$\widehat{\beta}^{\mathrm{cm}}(\rho,\tau)\xrightarrow{p}\beta^{\mathrm{cm}}(\rho,\tau)$$

*and*

$$\sqrt{N}(\widehat{\beta}^{\mathrm{cm}}(\rho,\tau)-\beta^{\mathrm{cm}}(\rho,\tau))\xrightarrow{d}\mathcal{N}(0,\Omega^{\mathrm{cm}}),$$

*where*

$$\Omega^{\mathrm{cm}} = \mathbb{E}\left[\left(\tau(Y-\beta^{\mathrm{sq}}) + (1-\tau)\,\psi(Y,W,X)\right)^2\right],$$

*and*

$$\psi(y,w,x) = \psi_0(y,w,x) + \psi_g(y,w,x) + \psi_W(y,w,x) + \psi_X(y,w,x). \tag{4.18}$$

Define

$$\hat{\Omega}^{\mathrm{cm}} = \frac{1}{N}\sum\nolimits_{i=1}^{N}(\tau(Y_i-\beta^{\mathrm{sq}}) + (1-\tau)\,\widehat{\psi}(Y_i,W_i,X_i,Z_i))^2.$$

### 4.4 Estimation and Inference for $\beta^{\text{lc}}$

Estimation of $\beta^{\text{lc}}$ proceeds in two-steps. First we estimate $g(w,x) = \mathbb{E}[Y|W=w, X=x]$ (and its derivative with respect to $w$) and $m(w) = \mathbb{E}[X|W=w]$ using kernel methods as in Section 4.1. In the second step we estimate $\beta^{\text{lc}}$ by method-of-moments using the sample analog of the moment condition

$$\mathbb{E}\left[\frac{\partial}{\partial w} g(W_i, X_i) \cdot d(W_i) \cdot (X_i - m(W_i)) - \beta^{\text{lc}}\right] = 0,$$

where $g(W,X)$ and $m(W)$ are replaced with the first step estimates, i.e.,

$$\hat{\beta}^{\text{lc}} = \frac{1}{N} \sum\nolimits_{i=1}^{N} \frac{\partial}{\partial w} \widehat{g}(W_i, X_i) \cdot d(W_i) \cdot (X_i - \widehat{m}(W_i)). \tag{4.19}$$

The asymptotic properties of $\hat{\beta}^{\text{lc}}$ are summarized by Theorem 4.4.

**Theorem 4.4** *Suppose Assumptions 2.1, 3.1, 4.1, and 4.2 hold with* $q \geq 2s-1$, $r \geq s+1$, $p \geq 4$, $d \geq s-1$, *and* $1/(2s) < \delta < 1/8$. *CHECK CONDITIONS Then*

$$\hat{\beta}^{\text{lc}} \xrightarrow{p} \beta^{\text{lc}},$$

*and*

$$\sqrt{N}(\hat{\beta}^{\text{lc}} - \beta^{\text{lc}}) \xrightarrow{d} \mathcal{N}(0, \Omega^{\text{lc}}),$$

*where*

$$\Omega^{\text{lc}} = \mathbb{E}\left[\left(\frac{\partial}{\partial w} g(W_i, X_i) \cdot d(W_i) \cdot (X_i - m(W_i)) + \delta_g(Y_i, W_i, X_i) + \delta_m(Y_i, W_i, X_i)\right)^2\right],$$

*where*

$$\begin{aligned}\delta_g(y,w,x) = &-\frac{1}{f_{W,X}(w,x)} \frac{\partial f_{W,X}(w,x)}{\partial W} d(w)(y - g(w,x))(x - m(w)) \\ &-\frac{\partial m(W)}{\partial W} d(w)(y - g(w,x)) \\ &-\frac{\partial d(w)}{\partial W}(y - g(w,x))(x - m(w)).\end{aligned}$$

*and*

$$\delta_m(y,w,x) = \mathbb{E}\left[\frac{\partial}{\partial w} g(w, X_i) \middle| W_i = w\right] \cdot d(w) \cdot (x - m(w))$$

The variance component corresponding to $\delta_g(y,w,x)$ and $\delta_m(y,w,x)$ captures the uncertainty from estimating $\frac{\partial}{\partial w} g(w,x)$ and $m(w)$ respectively.

## 5 A Monte Carlo Study

## 6 Conclusions

## References


Angrist, J. D. and A. B. Krueger. (1999). "Empirical Strategies in Labor Economics," *Handbook of Labor Economics 3A*: 1277 - xxxx (O. Ashenfelter and D. Card, Eds). New York: Elsevier Science.

Athey, S., and G. Imbens. (2005). "Identification and Inference in Nonlinear Difference-In-Differences Models," forthcoming *Econometrica*.

Athey, S., and S. Stern. (1998). "An Empirical Framework for Testing Theories About Complementarity in Organizational Design", *NBER Working Paper No. 6600.*

Bhattacharya, D., (2008), "Inferring Optimal Peer Assignment from Experimental Data," Unpublished Manuscript, Dept of Economics, Dartmouth College.

Becker, Gary S. and Kevin M. Murphy. (2000). *Social Economics: Market Behavior in a Social Environment.* Cambridge, MA: Harvard University Press.

Brock, W. and S. Durlauf. (2001). "Interactions-based Models," *Handbook of Econometrics 5*: 3297 - 3380 (J. Heckman & E. Leamer, Eds.). Amsterdam: North-Holland.

Card, D., and A. Krueger. (1992). "Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States, "*Journal of Political Economy* 100 (1): 1 - 40.

Dehejia, R. (2005). "Program evaluation as a decision problem," *Journal of Econometrics* 125 (1-2): 141 - 173.

Flores, C. (2005). "Estimation of Dose-Response Functions and Optimal Doses with a Continuous Treatment," *Mimeo.*

Graham, B., G. Imbens, and G. Ridder,(2006a). "Complementarity and the Optimal Allocation of Inputs," *Mimeo.*

Graham, B., G. Imbens, and G. Ridder,(2006b). "Measuring the Average Outcome and Inequality Effects of Segregation in the Presence of Social Spillovers," *Mimeo.*

Hardle, W. and T.M. Stoker. (1989). "Investigating smooth multiple regression by the method of average derivatives," *Journal of the American Statistical Association* 84 (408): 986 - 995.

Heckman, J., R. Lalonde, and J. Smith. (2000). "The Economics and Econometrics of Active Labor Markets Programs," *Handbook of Labor Economics 3A*: xxxx - xxxx (O. Ashenfelter and D. Card, Eds). New York: Elsevier Science.

Heckman, J., J. Smith, and N. Clements. (1997). "Making The Most Out Of Programme Evaluations and Social Experiments: Accounting For Heterogeneity in Programme Impacts," *Review of Economic Studies* 64 (4): 487 - 535.

Hirano, K. and J. Porter. (2005). "Asymptotics for statistical treatment rules," *Mimeo.*

Imbens, G. (2000). "The Role of the Propensity Score in Estimating Dose-Response Functions," *Biometrika* 87 (3): 706 - 710.

Imbens, G. (2004). "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Survey," *Review of Economics and Statistics* 86 (1): 4 - 30.

Lehman, E. (1998), *Elements of Large-Sample Theory*, Springer-Verlag, New York.

Linton, O., and J. Nielsen. (1995), "A Kernel Method of Estimating Structured Nonparametric Regression Based on Marginal Integration," *Biometrika* 82 (1): 93 - 100.

Manski, C. (1990). "Nonparametric Bounds on Treatment Effects," *American Economic Review* 80 (2): 319 - 323.

Manski, C. (1993). "Identification of Endogenous Social Effects: The Reflection Problem," *Review of Economic Studies* 60 (3): 531 - 542.

Manski, C. (2003). *Partial Identification of Probability Distributions.* New York: Springer-Verlag.

Manski, C. (2004). "Statistical Treatment Rules for Heterogenous Populations," *Econometrica* 72 (4): 1221 - 1246.

Newey, W. (1994). "Kernel Estimation of Partial Means and a General Variance Estimator," *Econometric Theory* 10 (2): 233 - 253.

Powell, J., J. Stock and T. Stoker. (1989). "Semiparametric Estimation of Index Coefficients," *Econometrica* 57 (6): 1403 - 1430.

Rosenbaum, P., and D. Rubin. (1983). "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika* 70 (1): 41 - 55.

Legros, P. and A. Newman. (2004). "Beauty is a beast, frog is a prince: assortative matching with nontransferabilities," Mimeo.

Sacerdote, B. (2001). "Peer effects with random assignment: results for Dartmouth roommates," *Quarterly Journal of Economics* 116 (2): 681 - 704.

## Notation

$b_N = N^{-\delta}$ is bandwidth

$K = \dim(Z)$

$W_\lambda = \lambda \cdot X + (1-\lambda) \cdot W$ for testing ARE's

$\beta$'s are average outcomes under various policies

$\beta^{\text{pam}}$ for positive assortative matching

$\beta^{\text{nam}}$ for negative assortative matching

$\beta^{\text{rm}}$ for random matching

$\beta^{\text{sq}}$ for status quo

$\beta^{\text{cm}}(\rho, \tau)$ for correlated matching

$\beta^{\text{lc}}$ is limit of the local complementarity test statistic

$g(w, x, z) = \mathbb{E}[Y|W = w, X = x, Z = z] = h_2(w, x, z)/h_1(w, x, z)$

$m(w, z) = \mathbb{E}[X|W = w, Z = z]$

$K_b(u) = \frac{1}{b_N^{k+2}} \mathcal{K}(u/\sigma)$ is kernel, where the dimension of $u$ is $k + 2$. Kernel is bounded, with bounded support $\mathbb{U} \subset \mathbb{R}^{k+2}$, and of order $s$.

Support of random variable $Z$ is $\mathcal{Z}$

$V = (W, X, Z')'$ is collection of all random right hand side variables

$N$ observations, $(Y_i, V_i)$ $i = 1, \ldots, N$.

$q$ is number of derivatives of $g(w, x)$ and $f_{WX}(w, x)$

$r$ is number of derivatives of kernel $K(u)$.

$s$ is order of kernel $K(u)$

$d$ is derivative order of kernel $K(u)$

$p$ is number of conditional moments of $Y_i$ that are finite

$t$ is number of derivatives of weight function $\omega(\cdot)$ and $n(\cdot)$ in IR theorems

## Appendix A: Additional Lemmas and Theorems

In this appendix we state a number of additional results that will be used in the proofs of the four Theorems 3.1-4.4. Specifically, Theorem 3.1 uses Lemmas A.1 and A.2. Theorem 4.1 uses Lemmas Theorem 4.2 uses the exact same lemmas. Theorem 4.3 uses Lemmas Theorem 4.4 uses Lemmas and Theorems.

**Definition 6.1** (Sobolev Norm) *The norm that we use for functions $g : \mathbb{Z} \subset \mathbb{R}^L \to \mathbb{R}$ that are at least $j$ times continuously differentiable is the Sobolev norm*

$$|g|_j = \sup_{z\in\mathbb{Z},|\lambda|\leq j} \left| \frac{\partial g^{|\lambda|}}{\partial z^\lambda}(z) \right|.$$

**Lemma A.1** *Let $f : \mathbb{X} \mapsto \mathbb{R}$, with $\mathbb{X} = [x_l, x_u]$ a compact subset of $\mathbb{R}$, be a twice continuously differentiable function, and let $g : \mathbb{R} \mapsto \mathbb{R}$ satisfy a Lipschitz condition, $|g(x+y) - g(x)| \leq c \cdot |y|$. Then*

$$\left| f(g(\lambda)) - \left( f(g(0)) + \frac{\partial}{\partial x} f(g(0)) \cdot (g(\lambda) - g(0)) \right) \right| \leq \frac{1}{2} \cdot \sup_{x\in\mathbb{X}} \left| \frac{\partial^2}{\partial x^2} f(x) \right| \cdot c^2 \cdot \lambda^2.$$

**Lemma A.2** *Let $X$ be a real-valued random variable with support $\mathbb{X} = [x_l, x_u]$, with density $f_X(x) > 0$ for all $x \in \mathbb{X}$, and let $h : \mathbb{X} \mapsto \mathbb{R}$ be a continous function. Suppose that $\mathbb{E}[|h(X) \cdot X]$ is finite. Then*

$$\text{Cov}(h(X), X) = \mathbb{E}\left[ \frac{\partial}{\partial x} h(X) \cdot \gamma(X) \right],$$

*where*

$$\gamma(x) = \frac{F_X(x) \cdot (1 - F_X(x)}{f_X(x)} \cdot \left( \mathbb{E}[X|X > x] - \mathbb{E}[X|X \leq x] \right),$$

*and $F_X(x)$ is the cumulative distribution function of $X$.*

For completeness we state a couple of results from Athey and Imbens (12006, AI from hereon).

**Lemma A.3** (Lemma A.2 in AI) *For a real-valued, continuously distributed, random variable $Y$ with compact support $\mathbb{Y} = [\underline{y}, \overline{y}]$, with the probability density function $f_Y(y)$ continuous, bounded, and bounded away from zero, on $\mathbb{Y}$, for any $\delta < 1/2$:*

$$\sup_{y\in\mathbb{Y}} N^\delta \cdot |\hat{F}_Y(y) - F_Y(y)| \xrightarrow{p} 0.$$

**Lemma A.4** (Lemma A.3 in AI) *For a real-valued, continuously distributed, random variable $Y$ with compact support $\mathbb{Y} = [\underline{y}, \overline{y}]$, with the probability density function $f_Y(y)$ continuous, bounded, and bounded away from zero, on $\mathbb{Y}$, for any $\delta < 1/2$:*

$$\sup_{q\in[0,1]} N^\delta \cdot |\hat{F}_Y^{-1}(q) - F_Y^{-1}(q)| \xrightarrow{p} 0.$$

**Lemma A.5** (Lemma A.5 in AI) *Suppose $\mathbb{Y} = [\underline{y}, \overline{y}]$, and $F_Y(y)$ is twice continuously differentiable on $\mathbb{Y}$, with its first derivative $f_Y(y) = \frac{\partial F_Y}{\partial y}(y)$ bounded away from zero on $\mathbb{Y}$. Then, for $0 < \eta < 3/4$ and $\delta > \max(2\eta - 1, \eta/2)$,*

$$\sup_{y\in\mathbb{Y}, x\leq N^{-\delta}, x+y\in\mathbb{Y}} N^\eta \cdot \left| \hat{F}_Y(y+x) - \hat{F}_Y(y) - (F_Y(y+x) - F_Y(y)) \right| \xrightarrow{p} 0.$$

**Lemma A.6** (Lemma A.6 in AI) *Suppose $\mathbb{Y} = [\underline{y}, \overline{y}]$, and $F_Y(y)$ is twice continuously differentiable on $\mathbb{Y}$, with its first derivative $f_Y(y) = \frac{\partial F_Y}{\partial y}(y)$ bounded away from zero on $\mathbb{Y}$. Then, for all $0 < \eta < 5/7$,*

$$\sup_{q\in[0,1]} N^\eta \cdot \left| \hat{F}_Y^{-1}(q) - F_Y^{-1}(q) + \frac{1}{f_Y(F_Y^{-1}(q))} \left( \hat{F}_Y(F_Y^{-1}(q)) - q \right) \right| \xrightarrow{p} 0.$$

**Lemma A.7** *For real-valued, continuously distributed, random variables $Y$ and $X$, with compact support $\mathbb{Y} = [\underline{y}, \overline{y}]$ and $\mathbb{X} = [\underline{x}, \overline{x}]$, with the probability density functions $f_Y(y)$ and $f_X(x)$ continuous, bounded, and bounded away from zero, on $\mathbb{Y}$ and $\mathbb{X}$, for any $\delta < 1/2$:*

$$\sup_{x\in\mathbb{X}} N^{\delta} \cdot \left| \hat{F}_Y^{-1}\left(\hat{F}_X(x)\right) - F_Y^{-1}\left(F_X(x)\right)\right| \xrightarrow{p} 0.$$

**Lemma A.8** *Suppose $\mathbb{Y} = [\underline{y}, \overline{y}]$, and $F_Y(y)$ is twice continuously differentiable on $\mathbb{Y}$, with its first derivative $f_Y(y) = \frac{\partial F_Y}{\partial y}(y)$ bounded away from zero on $\mathbb{Y}$. Then, for $0 < \eta < 3/4$ and $\delta > \max(2\eta - 1, \eta/2)$,*

$$\sup_{y\in\mathbb{Y}, x\leq N^{-\delta}, x+y\in\mathbb{Y}} N^{\eta} \cdot \left| \hat{F}_Y(y+x) - \hat{F}_Y(y) - f_Y(y)\cdot x\right| \xrightarrow{p} 0.$$

**Lemma A.9** *Suppose Assumptions 3.1-4.2 hold. Moreover, suppose that in these assumptions $q \geq 2s-1$, $r \geq s-1$. Then*

$$\sup_{w\in\mathbb{W}} |\hat{m}(w) - m(w)| = O_p\left(\left(\frac{\ln(N)}{N\cdot b_N}\right)^{1/2} + b_N^s\right).$$

**Lemma A.10** *Suppose Assumptions 3.1-4.2 hold. Moreover, suppose that $q \geq 2s+1$ and $r \geq s+3$, Then (i)*

$$\sup_{w\in\mathbb{W}, x\in\mathbb{X}} |\hat{g}(w,x) - g(w,x)| = O_p\left(\left(\frac{\ln(N)}{N\cdot b_N^2}\right)^{1/2} + b_N^s\right),$$

*(ii)*

$$\sup_{w\in\mathbb{W}, x\in\mathbb{X}} \left|\frac{\partial \hat{g}}{\partial w}(w,x) - \frac{\partial g}{\partial w}(w,x)\right| = O_p\left(\left(\frac{\ln(N)}{N\cdot b_N^4}\right)^{1/2} + b_N^s\right),$$

*and iii),*

$$\sup_{w\in\mathbb{W}, x\in\mathbb{X}} \left|\frac{\partial^2 \hat{g}}{\partial w^2}(w,x) - \frac{\partial^2 g}{\partial w^2}(w,x)\right| = O_p\left(\left(\frac{\ln(N)}{N\cdot b_N^6}\right)^{1/2} + b_N^s\right).$$

The next lemma shows that we can separate out the uncertainty in $\hat{\beta}^{\text{pam}}$ into five components: the uncertainty from estimating $g(\cdot)$, the uncertainty from estimating $\hat{F}_W^{-1}(\cdot)$, the uncertainty from estimating $\hat{F}_X(\cdot)$, and the uncertainty from averaging $g(F_W^{-1}(F_X(X_i)), X_i)$ over the sample, and a remainder term that is $o_p\left(N^{-1/2}\right)$. Define

$$\hat{\beta}_{\text{pam},g} = \frac{1}{N}\sum_{i=1}^{N} \hat{g}\left(F_W^{-1}\left(F_X(X_i)\right), X_i\right),$$

$$\hat{\beta}_{\text{pam},W} = \frac{1}{N}\sum_{i=1}^{N} g\left(\hat{F}_W^{-1}\left(F_X(X_i)\right), X_i\right),$$

$$\hat{\beta}_{\text{pam},X} = \frac{1}{N}\sum_{i=1}^{N} g\left(F_W^{-1}\left(\hat{F}_X(X_i)\right), X_i\right),$$

and

$$\overline{g}_{\text{pam}} = \frac{1}{N}\sum_{i=1}^{N} g\left(F_W^{-1}\left(F_X(X_i)\right), X_i\right).$$

**Lemma A.11** *Suppose Assumptions 3.1, 4.1, and 4.2 hold with $q \geq 2s+1$, $r \geq s+3$, and $0 \leq \delta < 1/6$. Then*

$$\hat{\beta}^{\text{pam}} - \beta^{\text{pam}} =$$
$$\left(\hat{\beta}_{\text{pam},g} - \overline{g}_{\text{pam}}\right) + \left(\hat{\beta}_{\text{pam},W} - \overline{g}_{\text{pam}}\right) + \left(\hat{\beta}_{\text{pam},X} - \overline{g}_{\text{pam}}\right) + \left(\overline{g}_{\text{pam}} - \beta^{\text{pam}}\right) + o_p\left(N^{-1/2}\right). \quad \text{(A.1)}$$

The next two results are special cases of theorems in Imbens and Ridder (2008). The first one refers to the full mean case, and focuses on the case where we take full means of regression functions and their first derivatives. The second result focuses on partial means of regression functions. The results in Imbens and Ridder (2008) allow for more general dependence on higher order derivatives, even in the partial mean case. Here we also restrict the analysis to the case where the regressors are the pair $(W_i, X_i)$.

Let $Z_i = (W_i, X_i)$, with $X_i \in \mathbb{X} \subset \mathbb{R}^{L_X}$, $W_i \in \mathbb{W} \subset \mathbb{R}^{L_W}$, $Z_i \in \mathbb{W} \times \mathbb{X} \subset \mathbb{R}^{L_Z}$, with $L_Z = L_X + L_W$. As before $h(z) = (h_1(z), h_2(z))'$, with $h_1(z) = f_Z(z)$, and $h_2(z) = \mathbb{E}[Y|Z = z] \cdot f_Z(z)$. Let $n : \mathbb{R}^K \mapsto \mathbb{R}$, $t : \mathbb{X} \mapsto \mathbb{W}$, and $\omega : \mathbb{X} \mapsto \mathbb{R}$, and define $\tilde{Y} = (\tilde{Y}_{i1}\ \tilde{Y}_{i2})'$, with $\tilde{Y}_{i1} = 1$ and $\tilde{Y}_{i2} = Y_i$. We are interested in full means (possibly depending on derivatives) of the regression function,

$$\theta_{\text{fm}} = \mathbb{E}\left[\omega(Z) n\left(h^{[\lambda]}(Z)\right)\right],$$

or partial means,

$$\theta_{\text{pm}} = \mathbb{E}\left[\omega(Z_{i1}) n\left(h\left(Z_1, t(Z_1)\right)\right)\right].$$

In the full mean example $h^{[\lambda]}$ denotes the vector with elements including all derivatives $h^{(\mu)}$ for $\mu \leq \lambda$. The estimators we focus on are

$$\hat{\theta}_{\text{fm}} = \frac{1}{N}\sum_{i=1}^{N} \omega(X_i) n(\hat{h}^{[\lambda]}_{NIP,s}(Z_i)), \qquad \text{and} \quad \hat{\theta}_{\text{pm}} = \frac{1}{N}\sum_{i=1}^{N} \omega(X_i) n(\hat{h}_{NIP,s}(t(X_i), X_i)).$$

It will also be useful to define the averages over the true regression functions and their derivatives,

$$\overline{\theta}_{\text{fm}} = \frac{1}{N}\sum_{i=1}^{N} \omega(X_i) n(h^{[\lambda]}(Z_i)), \qquad \text{and} \quad \overline{\theta}_{\text{pm}} = \frac{1}{N}\sum_{i=1}^{N} \omega(X_i) n(h(t(X_i), X_i)).$$

**Assumption A.1** (Distribution)
*(i)* $(Y_1, Z_1), (Y_2, Z_2), \ldots,$ *are independent and identically distributed,*
*(ii) the support of* $Z$ *is* $\mathbb{Z} \subset \mathbb{R}^{L_Z}$, $\mathbb{Z} = \bigotimes_{l=1}^{L_Z} [\underline{z}_l, \overline{z}_l]$, $\underline{z}_l < \overline{z}_l$ *for all* $l = 1, \ldots, L_Z$.
*(iii)* $\sup_{z \in \mathbb{Z}} \mathbb{E}[|Y|^p | Z = z] < \infty$ *for some* $p > 3$.
*(iv)* $g(z) = \mathbb{E}[Y|Z = z]$ *is* $q$ *times continuously differentiable on the interior of* $\mathbb{Z}$ *with the* $q$*-th derivative bounded,*
*(v)* $f_Z(z)$ *is* $q$ *times continuously differentiable on the interior of* $\mathbb{Z}$ *with the* $q$*-th derivative bounded.*

**Assumption A.2** (Kernel)
*(i)* $K : \mathbb{R}^L \rightarrow \mathbb{R}$, *with* $K(u) = \prod_{l=1}^{L} \mathcal{K}(u_l)$,
*(ii)* $K(u) = 0$ *for* $u \notin \mathbb{U}$, *with* $\mathbb{U} = [-1, 1]^L$,
*(iii)* $K$ *is* $r$ *times continuously differentiable, with the* $r$*-th derivative bounded on the interior of* $\mathbb{U}$,
*(iv)* $K$ *is a kernel of order* $s$, *so that* $\int_{\mathbb{U}} K(u)du = 1$ *and* $\int_{\mathbb{U}} u^\lambda K(u)du = 0$ *for all* $\lambda$ *such that* $0 < |\lambda| < s$, *for some* $s \geq 1$,
*(v)* $K$ *is a kernel of derivative order* $d$.

**Assumption A.3** *The bandwidth* $b_N = N^{-\delta}$ *for some* $\delta > 0$.

**Assumption A.4** (Smoothness of $n$ and $\omega$)
*(i) The function* $n$ *is* $t$ *times continuously differentiable with its* $t$*-th derivative bounded, and*
*(ii) the function* $\omega$ *is* $t$ *times differentiable on* $\mathbb{X}$. *with bounded* $t$*-th derivative.*

**Assumption A.5** (Smoothness of $t$)
*(i) The function* $t : \mathbb{X} \mapsto \mathbb{W}$ *is twice continuously differentiable on* $\mathbb{X}$ *with its first derivative positive, bounded, and bounded away from zero.*

**Theorem A.1** (Generalized Full Mean and Average Derivative, Imbens and Ridder, 2008)
*If Assumptions A.1, A.2, A.3, and A.4 hold with* $q \geq |\lambda| + 2s - 1$, $r \geq |\lambda| + s - 1 + L_Z$, $t \geq |\lambda| + s$, $p \geq 3$, $d \geq \max\{\lambda_1, \ldots, \lambda_L\} + s - 1$, *and*

$$\frac{1}{2s} < \delta < \min\left\{\frac{2 - \frac{4}{p}}{2L + 4\max\{1, |\lambda|\}}, \frac{1}{2L + 4|\lambda|}\right\}$$

*then* $\hat{\theta}_{\rm fm}$ *is asymptotically linear with*

$$\sqrt{N}(\hat{\theta}_{\rm fm}-\theta_{\rm fm}) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(\omega(Z_i)n(h^{[\lambda]}(Z_i)) - \mathbb{E}\left[\omega(Z_i)m\left(h^{[\lambda]}(Z_i)\right)\right]\right)$$

$$+\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(\sum_{\kappa\leq\lambda}(-1)^{|\kappa|}\sum_{m=1}^{2}\left(\alpha_{\kappa m}^{(\kappa)}(X_i)\tilde{Y}_{im} - \mathbb{E}[\alpha_{\kappa m}^{(\kappa)}(X)\tilde{Y}_m]\right)\right) + o_p(1).$$

*with*

$$\alpha_{\kappa 1}(z) = f_X(z)\omega(z)\frac{\partial n}{\partial h_1^{(\kappa)}(z)}(h^{[\lambda]}(z)), \qquad \text{and} \quad \alpha_{\kappa 2}(z) = f_X(z)\omega(z)\frac{\partial n}{\partial h_2^{(\kappa)}(z)}(h^{[\lambda]}(z)).$$

The second theorem from IR gives the asymptotic properties of the GPM estimators

**Theorem A.2** (Generalized Partial Mean, Imbens and Ridder, 2008)
*If Assumptions A.1, A.2, A.3, A.4, and A.5 hold with* $q \geq 2s-1$, $r \geq s-1+L_Z$, $t \geq s$, $p \geq 4$, $d \geq s-1$, *and*

$$\frac{1}{2s} < \delta < \min\left\{\frac{2-\frac{4}{p}}{2L_Z+4}, \frac{1}{2L_Z}\right\},$$

*then* $\hat{\theta}_{\rm pm}$ *is asymptotically linear with*

$$\sqrt{N}(\hat{\theta}_{\rm pm}-\theta_{\rm pm}) = \sqrt{N}\cdot\left(\overline{\theta}_{\rm pm}-\theta_{\rm pm}\right)$$

$$+\frac{1}{{b_N}^{L_W}\sqrt{N}}\cdot\sum_{i=1}^{N}\sum_{m=1}^{2}\left(\alpha_m(X_i)'\tilde{Y}_{im}\int_{\mathbb{U}_2}K\left(\frac{W_i-t(X_i)}{b_N}+\frac{\partial}{\partial x_1'}t(X_i)\cdot u_2, u_2\right)du_2\right.$$

$$\left.-\mathbb{E}\left[\alpha_m(X)'\tilde{Y}_m\int_{\mathbb{U}_2}K\left(\frac{W-t(X)}{b_N}+\frac{\partial}{\partial z_1'}t(X)\cdot u_2, u_2\right)du_2\right]\right) + o_p(1),$$

*with*

$$\alpha_1(x) = f_Z(t(x),x)\omega(x)\frac{\partial n}{\partial h_1}(h(t(x),x)), \qquad \text{and} \quad \alpha_2(x) = f_Z(t(x),x)\omega(x)\frac{\partial n}{\partial h_2}(h(t(x),x)).$$

*Moreover,*

$$\begin{pmatrix} \sqrt{N}\cdot\left(\overline{\theta}_{\rm pm}-\theta_{\rm pm}\right) \\ \sqrt{N}b_N^{L_W/2}(\hat{\theta}_{\rm pm}-\overline{\theta}_{\rm pm}) \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}V_1 & 0\\ 0 & V_2\end{pmatrix}\right),$$

*with*

$$V_1 = \mathbb{E}\left[(\omega(X)n(h(t(X),X)) - \theta_{\rm pm})^2\right],$$

*and*

$$V_2 = \sum_{m=1}^{2}\sum_{m'=1}^{2}\int_{\mathbb{X}_1}\mu_{mm'}(x_1,t(x_1))\alpha_m(x_1)\alpha_{m'}(x_1)\int_{\mathbb{U}_2}\left(\int_{\mathbb{U}_1}K\left(u_1, \frac{\partial t}{\partial x_1}(x_1)u_1+u_2\right)du_1\right)^2 du_2 f_X(x_1,t(x_1))dx_1,$$

*with* $\mu_{mm'}(x) = \mathbb{E}[\tilde{Y}_{im}\tilde{Y}_{im'}|X=x]$ *for* $m, m' = 1, 2$.

**Lemma A.12** *Suppose Assumptions 3.1, 4.1, and 4.2 hold, with* $q \geq 2s-1$, $r \geq s+1$, $p \geq 4$, $d \geq s-1$, *and* $1/(2s) < \delta < 1/8$. *Then*

$$\sqrt{N}\left(\hat{\beta}_g - \overline{g}\right)$$

$$= \frac{1}{\sqrt{N}b_N}\sum_{i=1}^{N}\left(Y_i - g(F_W^{-1}(F_X(X_i)),X_i)\right)\cdot\int_{u_2}K\left(\frac{W_i - F_W^{-1}(F_X(X_i))}{b_N} + \frac{f_X(X_i)}{f_W\left(F_W^{-1}(F_X(X_i))\right)}\cdot u_2, u_2\right)du_2$$

$$+o_p(1).$$

**Lemma A.13** *Suppose Assumptions 3.1, 4.1, and 4.2 hold WITH $q \geq 2$. Then*

$$\hat{\beta}_W - \overline{g} = \frac{1}{N}\sum_{i=1}^{N} q_W(W_i) + o_p\left(N^{-1/2}\right).$$

**Lemma A.14** *Suppose Assumptions 3.1, 4.1, and 4.2 hold, with $q \geq 2$. Then*

$$\hat{\beta}_{\mathrm{pam},X} - \overline{g}_{\mathrm{pam}} = \frac{1}{N}\sum_{i=1}^{N} r_X(X_i) + o_p\left(N^{-1/2}\right).$$

Define

$$\hat{\beta}_{\mathrm{lc},g} = \frac{1}{N}\sum_{i=1}^{N} \frac{\partial \hat{g}}{\partial w}(W_i, X_i) \cdot d(W_i) \cdot (X_i - m(W_i)),$$

$$\hat{\beta}_{\mathrm{lc},m} = \frac{1}{N}\sum_{i=1}^{N} \frac{\partial g}{\partial w}(W_i, X_i) \cdot d(W_i) \cdot (X_i - \hat{m}(W_i)),$$

and

$$\overline{g}_{\mathrm{lc}} = \frac{1}{N}\sum_{i=1}^{N} \frac{\partial g}{\partial w}(W_i, X_i) \cdot d(W_i) \cdot (X_i - m(W_i)).$$

**Lemma A.15** *Suppose Assumption 3.1 hold. Moreoever, suppose that the estimators for $g(w,x)$ and $m(w)$, $\hat{g}(w,x)$ and $\hat{m}(w)$ respectively, satisfy*

$$\sup_{w\in\mathbb{W},x\in\mathbb{X}} \left|\frac{\partial}{\partial w}\hat{g}(w,x) - \frac{\partial}{\partial w} g(w,x)\right| = o_p(N^{-\eta}) \qquad \text{and} \qquad \sup_{w\in\mathbb{W}} |\hat{m}(w) - m(w)| = o_p(N^{-\eta}),$$

*for some $\eta > 1/4$. Then*

$$\hat{\beta}^{\mathrm{lc}} - \beta^{\mathrm{lc}} = \left(\hat{\beta}_{\mathrm{lc},g} - \overline{g}_{\mathrm{lc}}\right) + \left(\hat{\beta}_{\mathrm{lc},m} - \overline{g}_{\mathrm{lc}}\right) + \left(\overline{g}_{\mathrm{lc}} - \beta^{\mathrm{lc}}\right) + o_p\left(N^{-1/2}\right). \tag{A.2}$$

**Lemma A.16**

$$\hat{\beta}_{\mathrm{lc},g} - \overline{g}_{\mathrm{lc}} = \frac{1}{N}\sum_{i=1}^{N} \delta_g(Y_i, W_i, X_i) + o_p\left(N^{-1/2}\right), \tag{A.3}$$

*where*

$$\begin{aligned}\delta_g(Y,W,X) = &-\frac{1}{f_{W,X}(W,X)}\frac{\partial f_{W,X}(W,X)}{\partial W}(Y - g(W,X))\, d(W)\,(X - m(W)) \\ &-\frac{\partial m(W)}{\partial W} d(W)\,(Y - g(W,X)) \\ &+\frac{\partial}{\partial w} d(W)\,(X - m(W))\,(Y - g(W,X)).\end{aligned}$$

**Lemma A.17** *If $\mathbb{E}[\varepsilon_i|X_i] = 0$, $\mathbb{E}[\varepsilon_i^2|X_i]$ finite, $\mathbb{E}[Y_i^2|X_i]$ finite, and $b_N \to 0$, then, for*

$$h(x) = \mathbb{E}[Y_i|X_i = x] \cdot f_X(x), \qquad \text{and} \qquad \hat{h}(x) = \frac{1}{N \cdot b_N}\sum_{i=1}^{N} Y_i \cdot K\left(\frac{X_i - x}{b_N}\right),$$

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N} \varepsilon_i \cdot \left(\hat{h}(X_i) - h(X_i)\right) = o_p(1),$$

**Lemma A.18** *Suppose* $\mathbb{E}[\varepsilon_i|X_i]=0$, $\mathbb{E}[\varepsilon_i^2|X_i]$ *finite,* $\mathbb{E}[Y_i^2|X_i]$ *finite, and* $b_N \to 0$. *With*

$$h_1(x) = f_X(x), \qquad \text{and} \qquad h_2(x) = \mathbb{E}[Y_i|X_i = x]\cdot f_X(x),$$

$$\hat{h}_1(x) = \frac{1}{N\cdot b_N}\sum_{i=1}^N K\left(\frac{X_i - x}{b_N}\right) \qquad \text{and} \qquad \hat{h}(x) = \frac{1}{N\cdot b_N}\sum_{i=1}^N Y_i\cdot K\left(\frac{X_i - x}{b_N}\right),$$

*then, for*

$$g(x) = h_2(x)/h_1(x), \qquad \text{and} \qquad \hat{g}(x) = \hat{h}_2(x)/\hat{h}_1(x),$$

*if* $\sup_{x\in\mathbb{X}}|\hat{h}_m(x) - h_m(x)| = O_p(N^{-\eta})$ *for some* $\eta > 1/4$, *then*

$$\frac{1}{\sqrt{N}}\sum_{i=1}^N \varepsilon_i\cdot(\hat{g}(X_i) - g(X_i)) = o_p(1)$$

**Lemma A.19** *Suppose Assumptions 3.1-4.2 hold, with* $q \geq 2s-1$, $r \geq s$, $p \geq 3$, $d \geq s-1$, *and*

$$\frac{1}{2s} < \delta < \frac{1}{3} - \frac{2}{3p}.$$

*Then*

$$\hat{\beta}_{\mathrm{lc},m} - \overline{g}_{\mathrm{lc}} = \frac{1}{N}\sum_{i=1}^N \mathbb{E}\left[\frac{\partial}{\partial w}g\left(W_i, X_i\right)\middle| W_i\right]\cdot d(W_i)\cdot(X_i - m(W_i)) + o_p\left(N^{-1/2}\right), \tag{A.4}$$

The next lemma gives a bound on the bias of the NIP estimator.

**Lemma A.20** (Bias)
*If for* $m = 1,\ldots,M$ *Assumptions 3.1-4.1 hold, and* $q \geq 2s-1$ *and* $r \geq s-1$, *then*

$$\sup_{z\in\mathbb{Z}}\left|\mathbb{E}\left[\hat{h}_{m,NIP,s}(z)\right] - h_m(z)\right| = O\left(b^s\right).$$

Note that by matching the order of the kernel and the degree of the polynomial in the NIP estimator we obtain the same reduction in the bias on the full support as on the interior set, i.e. the NIP estimator has a bias that is of the same order as that of the NW estimator on the interior set. The variance is bounded in the following lemma.

**Lemma A.21** (Variance)
*If Assumptions 3.1-4.1 hold and* $q \geq s-1, r \geq s-1+L$, *then*

$$\sup_{z\in\mathbb{Z}}\left|\hat{h}_{m,NIP,s}(z) - \mathbb{E}\left[\hat{h}_{m,NIP,s}(z)\right]\right| = O_p\left(\left(\frac{\log N}{N b_N^L}\right)^{1/2}\right).$$

This is the same bound as for the NW estimator on the internal set.
The two lemmas imply a uniform rate for the NIP estimator

**Lemma A.22** (Uniform Convergence)
*If Assumptions 3.1-4.1 hold and* $q \geq 2s-1, r \geq s-1+L$, *then*

$$\sup_{z\in\mathbb{Z}}\left|\hat{h}_{m,NIP,s}(z) - h_m(z)\right| = O_p\left(\left(\frac{\log N}{N\cdot b^L}\right)^{1/2} + b^s\right).$$

Before the next theorem we need some additional definitions. Define

$$\theta = \int_{\mathcal{S}_2}\int_{\mathcal{S}_2} n(h(s_1,s_2)) f_{S_1}(s_1) f_{S_2}(s_2)\mathrm{d}s_1\mathrm{d}s_2 \tag{A.5}$$

with estimator

$$\hat{\theta} = \frac{1}{N^2}\sum_{j=1}^N\sum_{k=1}^N n(\hat{h}_{NIP,s}(S_{1j}, S_{2k})). \tag{A.6}$$

Define also

$$n_1(s) = \mathbb{E}\left[n(h(s,S_i))\right], \qquad \text{and} \qquad n_2(s) = \mathbb{E}\left[n(h(S_i,s))\right].$$

**Theorem A.3** *Suppose that Assumptions 3.1-4.2, hold with $q \geq 2s-1$ and $r \geq s-1+L$, that $n(\cdot)$ is twice continously differentiable, and that*

$$\frac{1}{2s} < \delta < \frac{1}{2L},$$

*then*

$$\sqrt{N}(\hat{\theta}-\theta) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\{(\mathbb{E}[n(h(S_{1i},S_2))]-\theta)+\mathbb{E}[n(h_0(S_1,S_{2i}))]-\theta)\}+ \tag{A.7}$$

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left\{\frac{\partial n}{\partial h}(h_0(S_i))'V_i f_{S_1}(S_{1i})f_{S_2}(S_{2i}) - \mathbb{E}_{VS}\left[\frac{\partial n}{\partial h}(h_0(S))'V f_{S_1}(S_1)f_{S_2}(S_2)\right]\right\}+o_p(1).$$

Before stating some additional lemmas that will be used for proving Theorem 4.3 we need some additional definitions. Define

$$\overline{g}_{\rm cm} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} g(W_i,X_j)\frac{\phi_c\left(\Phi_c^{-1}(F_W(W_i)),\Phi_c^{-1}(F_X(X_j));\rho\right)}{\phi_c\left(\Phi_c^{-1}(F_W(W_i))\right)\phi_c\left(\Phi_c^{-1}(F_X(X_j))\right)}$$

$$\hat{\beta}_g^{\rm cm} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} \hat{g}(W_i,X_j)\frac{\phi_c\left(\Phi_c^{-1}(F_W(W_i)),\Phi_c^{-1}(F_X(X_j));\rho\right)}{\phi_c\left(\Phi_c^{-1}(F_W(W_i))\right)\phi_c\left(\Phi_c^{-1}(F_X(X_j))\right)}$$

$$\hat{\beta}_W^{\rm cm} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} g(W_i,X_j)\frac{\phi_c\left(\Phi_c^{-1}(\hat{F}_W(W_i)),\Phi_c^{-1}(F_X(X_j));\rho\right)}{\phi_c\left(\Phi_c^{-1}(\hat{F}_W(W_i))\right)\phi_c\left(\Phi_c^{-1}(F_X(X_j))\right)}$$

$$\hat{\beta}_X^{\rm cm} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} g(W_i,X_j)\frac{\phi_c\left(\Phi_c^{-1}(F_W(W_i)),\Phi_c^{-1}(\hat{F}_X(X_j));\rho\right)}{\phi_c\left(\Phi_c^{-1}(F_W(W_i))\right)\phi_c\left(\Phi_c^{-1}(\hat{F}_X(X_j))\right)}$$

**Lemma A.23** *Suppose Assumptions 3.1-4.2 hold. Then*

$$\widehat{\beta}^{\rm cm}(\rho,0)-\beta^{\rm cm}(\rho,0) = \left(\hat{\beta}_g^{\rm cm}-\overline{g}^{\rm cm}\right)+\left(\hat{\beta}_W^{\rm cm}-\overline{g}^{\rm cm}\right)+\left(\hat{\beta}_X^{\rm cm}-\overline{g}^{\rm cm}\right)+\left(\overline{g}^{\rm cm}-\beta^{\rm cm}(\rho,0)\right)+o_p\left(N^{-1/2}\right).$$

**Lemma A.24** *Suppose Assumptions 3.1-4.2 hold. Then*

$$\hat{\beta}_g^{\rm cm}-\overline{g}^{\rm cm} = \frac{1}{N}\sum_{i=1}^{N}\psi_g(Y_i,W_i,X_i)+o_p\left(N^{-1/2}\right),$$

*where*

$$\psi_g(y,w,x) = \frac{f_W(w)\cdot f_X(x)}{f_{WX}(w,x)}\cdot(y-g(w,x))\cdot d(w,x). \tag{A.8}$$

**Lemma A.25** *Suppose Assumptions 3.1-4.2 hold. Then*

$$\hat{\beta}_W^{\rm cm}-\overline{g}^{\rm cm} = \frac{1}{N}\sum_{i=1}^{N}\psi_W(Y_i,W_i,X_i)+o_p\left(N^{-1/2}\right),$$

*where*

$$\psi_W(y,w,x) = \mathbb{E}\left[e_W(W_i,X_i)\cdot g(W_i,X_i)\cdot(1_{w\leq W_i}-F_W(w))\right]. \tag{A.9}$$

**Lemma A.26** *Suppose Assumptions 3.1-4.2 hold. Then*

$$\hat{\beta}_X^{\rm cm}-\overline{g}^{\rm cm} = \frac{1}{N}\sum_{i=1}^{N}\psi_X(Y_i,W_i,X_i)+o_p\left(N^{-1/2}\right),$$

*where*

$$\psi_X(y,w,x) = \mathbb{E}\left[e_X(W_i,X_i)\cdot g(W_i,X_i)\cdot(1_{x\leq X_i}-F_W(x))\right]. \tag{A.10}$$

**Lemma A.27** *Suppose Assumptions 3.1-4.2 hold. Then*

$$\overline{g}^{\mathrm{cm}} - \beta^{\mathrm{cm}}(\rho, 0) = \frac{1}{N}\sum_{i=1}^{N} \psi_0(Y_i, W_i, X_i) + o_p\left(N^{-1/2}\right),$$

*where*

$$\psi_0(y, w, x) = \left(\mathbb{E}\left[g(W_i, x) \cdot d(W_i, x)\right] - \beta^{\mathrm{cm}}(\rho, 0)\right) + \left(\mathbb{E}\left[g(w, X_i) \cdot d(Ww, X_i)\right] - \beta^{\mathrm{cm}}(\rho, 0)\right). \tag{A.11}$$

The following theorem is a simplifed version of the V-statistics results in Lehman (1998).

**Theorem A.4** (V-STATISTICS) *Suppose $Z_1, \ldots, Z_N$ are independent and identically distributed random vectors with dimension $k$, with support $\mathbb{Z} \subset \mathbb{R}^k$. Let $\psi : \mathbb{Z}^k \times \mathbb{Z} \mapsto \mathbb{R}$ be a function. Define*

$$\theta = \mathbb{E}\left[\psi(Z_1, Z_2)\right],$$

$$\sigma^2 = \mathrm{Cov}(\psi(Z_1, Z_2), \psi(Z_1, Z_3)) + \mathrm{Cov}(\psi(Z_2, Z_1), \psi(Z_1, Z_3)) + \mathrm{Cov}(\psi(Z_1, Z_2), \psi(Z_3, Z_1)) + \mathrm{Cov}(\psi(Z_2, Z_2), \psi(Z_3, Z_1)).$$

*and*

$$V = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} \psi(Z_i, Z_j).$$

*Then, if $0 < \sigma^2 < \infty$,*

$$\sqrt{N} \cdot (V - \theta) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right).$$

## Appendix B: Proofs of Additional Lemmas and Theorems

**Proof of Lemma A.1:** Because $f(\cdot)$ is twice continuously differentiable on $\mathbb{X}$, a compact subset of $\mathbb{R}$, it follows that for all $a, b \in \mathbb{X}$, by a Taylor series expansion,

$$f(b) = f(a) + \frac{\partial f}{\partial x}(a) \cdot (b-a) + \frac{1}{2} \cdot \frac{\partial^2 f}{\partial x^2}(c) \cdot (b-a)^2,$$

for some $c \in \mathbb{X}$. Hence

$$\left| f(g(\lambda)) - \left( f(g(0)) + \frac{\partial f}{\partial x}(g(0)) \cdot (g(\lambda) - g(0)) \right) \right| \leq \frac{1}{2} \cdot \sup_{x \in \mathbb{X}} \left| \frac{\partial^2 f}{\partial x^2}(x) \right| \cdot (g(\lambda) - g(0))^2.$$

By the Lipschitz condition on $g(\lambda)$, this is bounded by

$$\frac{1}{2} \cdot \sup_{x \in \mathbb{X}} \left| \frac{\partial^2 f}{\partial x^2}(x) \right| \cdot c^2 \cdot \lambda^2.$$

□

**Proof of Lemma A.2:** Let $\mu = \mathbb{E}[X]$, nd write $h(x) = h(x_l) + \int_{x_l}^{x} \frac{\partial}{\partial x} h(z) dz$. Then:

$$\text{Cov}(h(X), X) = \mathbb{E}\left[ h(X) \cdot (X - \mu) \right] = \mathbb{E}\left[ h(x_l) + \int_{x_l}^{X} \frac{\partial}{\partial x} h(z) dz \cdot (X - \mu) \right]$$

$$= \mathbb{E}\left[ \int_{x_l}^{x_u} 1_{X \geq z} \cdot \frac{\partial}{\partial x} h(z) dz \cdot (X - \mu) \right]$$

$$= \int_{x_l}^{x_u} \frac{\partial}{\partial x} h(z) \cdot \mathbb{E}\left[ 1_{X \geq z} \cdot (X - \mu) \right] dz$$

$$= \int_{x_l}^{x_u} \frac{\partial}{\partial x} h(z) \cdot \mathbb{E}\left[ X - \mu | X > z \right] \cdot \text{pr}(X > z) dz$$

$$= \int_{x_l}^{x_u} \frac{\partial}{\partial x} h(z) \cdot F_X(z) \cdot (1 - F_X(z)) \cdot \left( \mathbb{E}[X | X > z] - \mathbb{E}[X | X \leq z] \right) dz$$

$$= \int_{x_l}^{x_u} \frac{\partial}{\partial x} h(z) \cdot \frac{F_X(z) \cdot (1 - F_X(z))}{f_X(z)} \cdot \left( \mathbb{E}[X | X > z] - \mathbb{E}[X | X \leq z] \right) f_X(z) dz$$

$$= \mathbb{E}\left[ \frac{\partial}{\partial x} h(X) \cdot \gamma(X) \right].$$

□

**Proof of Lemma A.7:** By the triangle inequality

$$\sup_{x \in \mathbb{X}} N^{\delta} \cdot \left| \hat{F}_Y^{-1}\left( \hat{F}_X(x) \right) - F_Y^{-1}\left( F_X(x) \right) \right|$$

$$\leq \sup_{x \in \mathbb{X}} N^{\delta} \cdot \left| \hat{F}_Y^{-1}\left( \hat{F}_X(x) \right) - F_Y^{-1}\left( \hat{F}_X(x) \right) \right|$$

$$+ \sup_{x \in \mathbb{X}} N^{\delta} \cdot \left| F_Y^{-1}\left( \hat{F}_X(x) \right) - F_Y^{-1}\left( \hat{F}_X(x) \right) \right|$$

$$\leq \sup_{q \in [0,1]} N^{\delta} \cdot \left| \hat{F}_Y^{-1}(q) - F_Y^{-1}(q) \right|$$

$$+ \sup_{x \in \mathbb{X}, y \in \mathbb{Y}} N^{\delta} \cdot \frac{1}{f_Y(y)} \left| \hat{F}_X(x) - F_X(x) \right|.$$

The first term is $o_p(1)$ by Lemma A.4, and the second by the fact that $F_Y(y)$ is continuous differentiable with its derivative bounded away from zero, in combination with Lemma A.3. □

**Proof of Lemma A.8:** By the triangle inequality

$$\sup_{y \in \mathbb{Y}, x \leq N^{-\delta}, x+y \in \mathbb{Y}} N^{\eta} \cdot \left| \hat{F}_Y(y+x) - \hat{F}_Y(y) - f_Y(y) \cdot x \right|$$

$$\leq \sup_{y \in \mathbb{Y}, x \leq N^{-\delta}, x+y \in \mathbb{Y}} N^{\eta} \cdot \left| \hat{F}_Y(y+x) - \left( F_Y(y+x) - F_Y(y) \right) \right|$$

$$+ \sup_{y\in\mathbb{Y},x\leq N^{-\delta},x+y\in\mathbb{Y}} N^{\eta}\cdot |F_Y(y+x)-F_Y(y)-f_Y(y)\cdot x|\,.$$

The first term on the right-hand side converges to zero in probability by Lemma A.5. To show that the second term converges to zero note that

$$\sup_{y\in\mathbb{Y},x\leq N^{-\delta},x+y\in\mathbb{Y}} N^{\eta}\cdot |F_Y(y+x)-F_Y(y)-f_Y(y)\cdot x|$$
$$\leq \sup_{y\in\mathbb{Y},x\leq N^{-\delta},x+y\in\mathbb{Y},\lambda\in[0,1]} N^{\eta}\cdot |f_Y(y+\lambda x)\cdot x-f_Y(y)\cdot x|$$
$$\leq \sup_{y\in\mathbb{Y},x\leq N^{-\delta},x+y\in\mathbb{Y}} N^{\eta}\cdot |F_Y(y+x)-F_Y(y)-f_Y(y)\cdot x|$$
$$\leq \sup_{y\in\mathbb{Y},z\in\mathbb{Y},x\leq N^{-\delta},x+y\in\mathbb{Y}} N^{\eta}\cdot \left|\frac{\partial f_Y}{\partial y}(z)\cdot \lambda x^2\right|$$
$$\leq \sup_{y\in\mathbb{Y},x\leq N^{-\delta}} N^{\eta}x^2\frac{\partial f_Y}{\partial y}(y)\to 0,$$

because $\frac{\partial f_Y}{\partial y}(y)$ is bounded, $x < N^{-\delta}$, and $\delta > \eta/2$. □

**Proof of Lemma A.9:** This follows directly from Theorem 7.1 in IR □

**Proof of Lemma A.10:** This follows directly from Theorem 7.1 in IR. □

**Proof of Lemma A.11:**

First note that by the assumptions in the Lemma the conditions for Lemma A.10 are satisfied. Moreoever, by the assumption that $0<\delta<1/6$, it follows that$O_p(b^N)=o_p(N^{-\eta})$ for $\eta<\delta\cdot s$, and $O_p(\ln(N)N^{-1}b_N^{-2}) = O_p\ln(N)N^{-1+2\delta}) = o_p(1)$, $O_p(\ln(N)N^{-1}b_N^{-4}) = O_p\ln(N)N^{-1+4\delta}) = o_p(N^{-\eta})$ for $\eta<1-4\delta$, and $O_p(\ln(N)N^{-1}b_N^{-6}) = O_p\ln(N)N^{-1+6\delta}) = o_p(1)$. Hence the results from Lemma A.10 imply

$$\sup_{w\in\mathbb{W},x\in\mathbb{X}} |\hat{g}(w,x)-g(w,x)| = O_p\left(\left(\frac{\ln(N)}{N\cdot b_N^2}\right)^{1/2}+b_N^s\right) = o_p(1), \tag{B.1}$$

$$\sup_{w\in\mathbb{W},x\in\mathbb{X}} \left|\frac{\partial\hat{g}}{\partial w}(w,x)-\frac{\partial g}{\partial w}(w,x)\right| = O_p\left(\left(\frac{\ln(N)}{N\cdot b_N^4}\right)^{1/2}+b_N^s\right) = o_p\left(N^{-\eta}\right), \tag{B.2}$$

for $\eta<\min(1-4\delta,\delta\cdot s)$, and

$$\sup_{w\in\mathbb{W},x\in\mathbb{X}} \left|\frac{\partial^2\hat{g}}{\partial w^2}(w,x)-\frac{\partial^2 g}{\partial w^2}(w,x)\right| = O_p\left(\left(\frac{\ln(N)}{N\cdot b_N^6}\right)^{1/2}+b_N^s\right) = o_p(1). \tag{B.3}$$

Now,

$$\hat{\beta}^{\text{pam}}-\beta^{\text{pam}} = \frac{1}{N}\sum_{i=1}^N \hat{g}\left(\hat{F}_W^{-1}\left(\hat{F}_X(X_i)\right),X_i\right) - \mathbb{E}\left[g\left(F_W^{-1}\left(F_X(X)\right),X\right)\right]$$

$$= \frac{1}{N}\sum_{i=1}^N \hat{g}\left(\hat{F}_W^{-1}\left(\hat{F}_X(X_i)\right),X_i\right) - \frac{1}{N}\sum_{i=1}^N g\left(\hat{F}_W^{-1}\left(\hat{F}_X(X_i)\right),X_i\right) \tag{B.4}$$

$$-\left(\frac{1}{N}\sum_{i=1}^N \hat{g}\left(F_W^{-1}\left(F_X(X_i)\right),X_i\right) - \frac{1}{N}\sum_{i=1}^N g\left(F_W^{-1}\left(F_X(X_i)\right),X_i\right)\right) \tag{B.5}$$

$$+\frac{1}{N}\sum_{i=1}^N \hat{g}\left(F_W^{-1}\left(F_X(X_i)\right),X_i\right) - \frac{1}{N}\sum_{i=1}^N g\left(F_W^{-1}\left(F_X(X_i)\right),X_i\right) \tag{B.6}$$

$$+\frac{1}{N}\sum_{i=1}^N g\left(\hat{F}_W^{-1}\left(\hat{F}_X(X_i)\right),X_i\right) - \frac{1}{N}\sum_{i=1}^N g\left(F_W^{-1}\left(\hat{F}_X(X_i)\right),X_i\right) \tag{B.7}$$

$$-\left(\frac{1}{N}\sum_{i=1}^N g\left(\hat{F}_W^{-1}\left(F_X(X_i)\right),X_i\right) - \frac{1}{N}\sum_{i=1}^N g\left(F_W^{-1}\left(F_X(X_i)\right),X_i\right)\right) \tag{B.8}$$

$$+\frac{1}{N}\sum_{i=1}^N g\left(\hat{F}_W^{-1}\left(F_X(X_i)\right),X_i\right) - \frac{1}{N}\sum_{i=1}^N g\left(F_W^{-1}\left(F_X(X_i)\right),X_i\right) \tag{B.9}$$

$$+\frac{1}{N}\sum_{i=1}^{N} g\left(F_W^{-1}\left(\hat{F}_X(X_i)\right), X_i\right) - \frac{1}{N}\sum_{i=1}^{N} g\left(F_W^{-1}\left(F_X(X_i)\right), X_i\right) \tag{B.10}$$

$$+\frac{1}{N}\sum_{i=1}^{N} g\left(F_W^{-1}\left(F_X(X_i)\right), X_i\right) - \mathbb{E}\left[g\left(F_W^{-1}\left(F_X(X)\right), X\right)\right]. \tag{B.11}$$

Since (B.6) is equal to $\hat{\beta}_{pam,g} - \overline{g}_{pam}$, (B.9) equals $\hat{\beta}_{pam,W} - \overline{g}_{pam}$, (B.10) equals $\hat{\beta}_{pam,X} - \overline{g}_{pam}$, and (B.11) equals $\overline{g}_{pam} - \beta^{\text{pam}}$, we only need to show that the sum of (B.4), (B.5), and that of (B.7), (B.8) are $o_p(N^{-1/2})$. First consider the sum of (B.4) and (B.5) that is equal to

$$\frac{1}{N}\sum_{i=1}^{N} \hat{g}\left(\hat{F}_W^{-1}\left(\hat{F}_X(X_i)\right), X_i\right) - \frac{1}{N}\sum_{i=1}^{N} \hat{g}\left(F_W^{-1}\left(F_X(X_i)\right), X_i\right)$$
$$-\left(\frac{1}{N}\sum_{i=1}^{N} g\left(\hat{F}_W^{-1}\left(\hat{F}_X(X_i)\right), X_i\right) - \frac{1}{N}\sum_{i=1}^{N} g\left(F_W^{-1}\left(F_X(X_i)\right), X_i\right)\right).$$

By a second order Taylor series expansion of $\hat{g}$ and $g$ in $F_W^{-1}\left(F_X(X_i)\right)$ this is, for some $\tilde{W}_i$ and $\bar{W}_i$, equal to

$$\frac{1}{N}\sum_{i=1}^{N} \frac{\partial \hat{g}}{\partial w}\left(F_W^{-1}\left(F_X(X_i)\right), X_i\right)\left(\hat{F}_W^{-1}\left(\hat{F}_X(X_i)\right) - F_W^{-1}\left(F_X(X_i)\right)\right)$$

$$+\frac{1}{2N}\sum_{i=1}^{N} \frac{\partial^2 \hat{g}}{\partial w^2}\left(\tilde{W}_i, X_i\right)\left(\hat{F}_W^{-1}\left(\hat{F}_X(X_i)\right) - F_W^{-1}\left(F_X(X_i)\right)\right)^2 \tag{B.12}$$

$$-\frac{1}{N}\sum_{i=1}^{N} \frac{\partial g}{\partial w}\left(F_W^{-1}\left(F_X(X_i)\right), X_i\right)\left(\hat{F}_W^{-1}\left(\hat{F}_X(X_i)\right) - F_W^{-1}\left(F_X(X_i)\right)\right)$$

$$-\frac{1}{2N}\sum_{i=1}^{N} \frac{\partial^2 g}{\partial w^2}\left(\bar{W}_i, X_i\right)\cdot\left(\hat{F}_W^{-1}\left(\hat{F}_X(X_i)\right) - F_W^{-1}\left(F_X(X_i)\right)\right)^2. \tag{B.13}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\partial \hat{g}}{\partial w}\left(F_W^{-1}\left(F_X(X_i)\right), X_i\right) - \frac{\partial g}{\partial w}\left(F_W^{-1}\left(F_X(X_i)\right), X_i\right)\right)$$
$$\times\left(\hat{F}_W^{-1}\left(\hat{F}_X(X_i)\right) - F_W^{-1}\left(F_X(X_i)\right)\right) + o_p\left(N^{-1/2}\right).$$

$$\leq \sup_{x\in\mathbb{X}}\left|\frac{\partial \hat{g}}{\partial w}\left(F_W^{-1}\left(F_X(x)\right), x\right) - \frac{\partial g}{\partial w}\left(F_W^{-1}\left(F_X(x)\right), x\right)\right| \tag{B.14}$$

$$\times\sup_{x\in\mathbb{X}}\left|\left(\hat{F}_W^{-1}\left(\hat{F}_X(x)\right) - F_W^{-1}\left(F_X(x)\right)\right)\right| + o_p\left(N^{-1/2}\right). \tag{B.15}$$

We used the fact that (B.13) is $o_p(N^{-1/2})$ because $\partial^2 g(w,x)/\partial w^2$ is bounded and because $\sup_{x\in\mathbb{X}}(\hat{F}_W^{-1}(\hat{F}_X(x)) - F_W^{-1}(F_X(x)))^2$ is $o_p(N^{-1/2})$ by Lemma A.7. Also (B.12) is $o_p(N^{-1/2})$ by the same argument because the bandwidth choice implies $\sup_{w\in\mathbb{W},x\in\mathbb{X}}|\partial^2\hat{g}(w,x)/\partial w^2 - \partial^2 g(w,x)/\partial w^2| = o_p(1)$ by B.3, so that

$$\sup_{w\in\mathbb{W},x\in\mathbb{X}}\left|\frac{\partial^2\hat{g}(w,x)}{\partial w^2}\right| \leq \sup_{w\in\mathbb{W},x\in\mathbb{X}}\left|\frac{\partial^2 g(w,x)}{\partial w^2}\right| + \sup_{w\in\mathbb{W},x\in\mathbb{X}}\left|\frac{\partial^2\hat{g}(w,x)}{\partial w^2} - \frac{\partial^2 g(w,x)}{\partial w^2}\right| = \sup_{w\in\mathbb{W},x\in\mathbb{X}}\left|\frac{\partial^2 g(w,x)}{\partial w^2}\right| + o_p(1)$$

Finally by Lemma A.6

$$\sup_{x\in\mathbb{X}}\left|\hat{F}_W^{-1}\left(\hat{F}_X(x)\right) - F_W^{-1}\left(F_X(x)\right)\right| = o_p\left(N^{-1/2+\eta}\right)$$

for all $\eta > 0$. By the assumption of the lemma

$$\sup_{x\in\mathbb{X}}\left|\frac{\partial \hat{g}}{\partial w}\left(F_W^{-1}\left(F_X(x)\right), x\right) - \frac{\partial g}{\partial w}\left(F_W^{-1}\left(F_X(x)\right), x\right)\right| = o_p(N^{-\eta})$$

for some $\eta > 0$. We conclude that the sum of (B.4) and (B.5) is $o_p\left(N^{-1/2}\right)$.

Next, consider the sum of (B.7) and (B.8) that is bounded by

$$\sup_{x\in\mathbb{X}}\left|\left[g\left(\hat{F}_W^{-1}\left(\hat{F}_X(x)\right),x\right)-g\left(F_W^{-1}\left(\hat{F}_X(x)\right),x\right)\right]-\left[g\left(\hat{F}_W^{-1}\left(F_X(x)\right),x\right)-g\left(F_W^{-1}\left(F_X(x)\right),x\right)\right]\right|$$

By a second order Taylor series expansion with intermediate values $\tilde{W}(x)$ and $\bar{W}(x)$ and the triangle inequality this is bounded by

$$\sup_{x\in\mathbb{X}}\left|\frac{\partial g}{\partial w}\left(F_W^{-1}\left(\hat{F}_X(x)\right),x\right)\left[\hat{F}_W^{-1}\left(\hat{F}_X(x)\right)-F_W^{-1}\left(\hat{F}_X(x)\right)\right]\right.$$

$$\left.-\frac{\partial g}{\partial w}\left(F_W^{-1}\left(F_X(x)\right),x\right)\left[\hat{F}_W^{-1}\left(F_X(x)\right)-F_W^{-1}\left(F_X(x)\right)\right]\right|+$$

$$\sup_{x\in\mathbb{X}}\frac{1}{2}\left|\frac{\partial^2 g}{\partial w^2}\left(\tilde{W}(x),x\right)\left[\hat{F}_W^{-1}\left(\hat{F}_X(x)\right)-F_W^{-1}\left(\hat{F}_X(x)\right)\right]^2\right|+\frac{1}{2}\sup_{x\in\mathbb{X}}\left|\frac{\partial^2 g}{\partial w^2}\left(\bar{W}(x),x\right)\left[\hat{F}_W^{-1}\left(F_X(x)\right)-F_W^{-1}\left(F_X(x)\right)\right]^2\right|$$

where because the second derivative of $g(w,x)$ is bounded on $\mathbb{W}\times\mathbb{X}$, by Lemma A.4 the expression on the last line is $o_p(N^{-1/2})$. The first term is bounded by

$$\sup_{x\in\mathbb{X}}\left|\left[\frac{\partial g}{\partial w}\left(F_W^{-1}\left(\hat{F}_X(x)\right),x\right)-\frac{\partial g}{\partial w}\left(F_W^{-1}\left(F_X(x)\right),x\right)\right]\left[\hat{F}_W^{-1}\left(\hat{F}_X(x)\right)-F_W^{-1}\left(\hat{F}_X(x)\right)\right]\right|$$

$$+\sup_{x\in\mathbb{X}}\left|\frac{\partial g}{\partial w}\left(F_W^{-1}\left(F_X(x)\right),x\right)\left[\hat{F}_W^{-1}\left(\hat{F}_X(x)\right)-F_W^{-1}\left(\hat{F}_X(x)\right)-\hat{F}_W^{-1}\left(F_X(x)\right)+F_W^{-1}\left(F_X(x)\right)\right]\right|$$

By a first order Taylor series expansion of $\frac{\partial g}{\partial w}\left(F_W^{-1}\left(\hat{F}_X(x)\right),x\right)$ in $F_X(x)$ we have, because the second derivative of $g(w,x)$ is bounded and the density of $W$ is bounded from 0 on its support, that by Lemmas A.4 and A.3 (Lemmas A.3 and A.4), the expression on the first line is $o_p(N^{-1/2})$. The bound on the expression in the second line is proportional to

$$\sup_{x\in\mathbb{X}}\left|\hat{F}_W^{-1}\left(\hat{F}_X(x)\right)-F_W^{-1}\left(\hat{F}_X(x)\right)-\hat{F}_W^{-1}\left(F_X(x)\right)+F_W^{-1}\left(F_X(x)\right)\right|$$

This expression is bounded by

$$\sup_{x\in\mathbb{X}}\left|\frac{1}{f_W\left(F_W^{-1}\left(\hat{F}_X(x)\right)\right)}\left[\hat{F}_W\left(F_W^{-1}\left(\hat{F}_X(x)\right)\right)-\hat{F}_X(x)\right]-\frac{1}{f_W\left(F_W^{-1}\left(F_X(x)\right)\right)}\left[\hat{F}_W\left(F_W^{-1}\left(F_X(x)\right)\right)-F_X(x)\right]\right|$$

$$+\sup_{x\in\mathbb{X}}\left|\hat{F}_W^{-1}\left(\hat{F}_X(x)\right)-F_W^{-1}\left(\hat{F}_X(x)\right)-\frac{1}{f_W\left(F_W^{-1}\left(\hat{F}_X(x)\right)\right)}\left[\hat{F}_W\left(F_W^{-1}\left(\hat{F}_X(x)\right)\right)-\hat{F}_X(x)\right]\right|$$

$$+\sup_{x\in\mathbb{X}}\left|\hat{F}_W^{-1}\left(F_X(x)\right)-F_W^{-1}\left(F_X(x)\right)-\frac{1}{f_W\left(F_W^{-1}\left(F_X(x)\right)\right)}\left[\hat{F}_W\left(F_W^{-1}\left(F_X(x)\right)\right)-F_X(x)\right]\right|$$

By Lemma A.6 the expressions in the last two lines are $o_p(N^{-1/2})$. The expression in the first line is bounded by

$$\sup_{x\in\mathbb{X}}\left|\left[\frac{1}{f_W\left(F_W^{-1}\left(\hat{F}_X(x)\right)\right)}-\frac{1}{f_W\left(F_W^{-1}\left(F_X(x)\right)\right)}\right]\left[\hat{F}_W\left(F_W^{-1}\left(\hat{F}_X(x)\right)\right)-\hat{F}_X(x)\right]\right|$$

$$+\sup_{x\in\mathbb{X}}\left|\frac{1}{f_W\left(F_W^{-1}\left(F_X(x)\right)\right)}\left[\hat{F}_W\left(F_W^{-1}\left(\hat{F}_X(x)\right)\right)-\hat{F}_X(x)-\hat{F}_W\left(F_W^{-1}\left(F_X(x)\right)\right)+F_X(x)\right]\right|$$

The expression in the first line is bounded by

$$\sup_{x\in\mathbb{X}}\left|\frac{1}{f_W\left(F_W^{-1}\left(\hat{F}_X(x)\right)\right)}-\frac{1}{f_W\left(F_W^{-1}\left(F_X(x)\right)\right)}\right|\times\sup_{x\in\mathbb{X}}\left|\hat{F}_W\left(F_W^{-1}\left(\hat{F}_X(x)\right)\right)-\hat{F}_X(x)\right|$$

By a first order Taylor series expansion of $\frac{1}{f_W\left(F_W^{-1}\left(\hat{F}_X(x)\right)\right)}$ in $F_X(x)$, the fact that $f_W(w)$ is bounded from 0 and its derivative bounded on $\mathbb{W}$, and Lemma A.3 the first factor is $o_p(N^{-\delta})$ for all $\delta > 0$ and by Lemma A.3 the same is true for the second factor, so that the product is $o_p(N^{-1/2})$. Because $f_W(w)$ is bounded from 0 on $\mathbb{W}$, the expression on the second line has a bound that is proportional to

$$\sup_{x\in\mathbb{X}}\left|\hat{F}_W\left(F_W^{-1}\left(\hat{F}_X(x)\right)\right) - \hat{F}_X(x) - \hat{F}_W\left(F_W^{-1}\left(F_X(x)\right)\right) + F_X(x)\right|$$

We rewrite this as

$$\sup_{x\in\mathbb{X}}\left|\hat{F}_W\left(F_W^{-1}\left(\hat{F}_X(x)\right)\right) - \hat{F}_W\left(F_W^{-1}\left(F_X(x)\right)\right) - \left(F_W\left(F_W^{-1}\left(\hat{F}_X(x)\right) - F_W\left(F_W^{-1}\left(F_X(x)\right)\right)\right)\right)\right| \leq$$

$$\sup_{x\in\mathbb{X}}\left|\hat{F}_W\left(F_W^{-1}\left(\hat{F}_X(x)\right)\right) - \hat{F}_W\left(F_W^{-1}\left(F_X(x)\right)\right) - \left(F_W\left(F_W^{-1}\left(\hat{F}_X(x)\right) - F_W\left(F_W^{-1}\left(F_X(x)\right)\right)\right)\right)\right| \times$$

$$1_{\sup_{x\in\mathbb{X}}\left|F_W^{-1}(\hat{F}_X(x))-F_W^{-1}(F_X(x))\right|\leq N^{-\delta}} + 4\cdot 1_{\sup_{x\in\mathbb{X}}\left|F_W^{-1}(\hat{F}_X(x))-F_W^{-1}(F_X(x))\right|>N^{-\delta}}$$

By Lemma A.7 the final term is $o_p(1)$ if $\delta = 1/3$. By

$$F_W^{-1}\left(\hat{F}_X(x)\right) = F_W^{-1}\left(F_X(x)\right) + \left[F_W^{-1}\left(\hat{F}_X(x)\right) - F_W^{-1}\left(F_X(x)\right)\right]$$

and defining $\overline{w} = F_W^{-1}\left(F_X(x)\right)$ and $\tilde{w} = F_W^{-1}\left(\hat{F}_X(x)\right) - F_W^{-1}\left(F_X(x)\right)$ we have that the first term on the right hand side is bounded by

$$\sup_{\overline{w}\in\mathbb{W},|\tilde{w}|\leq N^{-\delta},\tilde{w}+\overline{w}\in\mathbb{W}}\left|\hat{F}_W\left(\overline{w}+\tilde{w}\right) - \hat{F}_W\left(\overline{w}\right) - \left(F_W\left(\overline{w}+\tilde{w}\right) - F_W\left(\overline{w}\right)\right)\right| = o_p(N^{-2/3})$$

by Lemma A.5 with $\delta = 1/3, \eta = 2/3$, so that we finally conclude that the sum of (B.7) and (B.8) is $o_p(N^{-1/2})$. □

**Proof of Lemma A.12:** The proof involves checking the conditions for Theorem A.2 from IR (given in Appendix A in the current paper), and simplifying the conclusions from that Theorem to the case at hand.
Define

$$h_1(w,x) = f_{WX}(w,x), \qquad \text{and} \quad h_2(w,x) = f_{WX}(w,x)\cdot g(w,x),$$

$$n(h) = \frac{h_2}{h_1},$$

so that

$$\frac{\partial n}{\partial h_1}(h) = -\frac{h_2}{(h_1)^2} = -\frac{g(\left(F_W^{-1}\left(F_X(x)\right),x\right)}{f_{WX}(\left(F_W^{-1}\left(F_X(x)\right),x\right)},$$

$$\frac{\partial n}{\partial h_2}(h) = \frac{1}{h_1} = \frac{1}{f_{WX}(\left(F_W^{-1}\left(F_X(x)\right),x\right)},$$

$$t(x) = F_W^{-1}\left(F_X(x)\right), \qquad \frac{\partial}{\partial x}t(x) = -\frac{f_X(x)}{f_W(F_W^{-1}(F_X(x)))},$$

$$\alpha_1(x) = -g(\left(F_W^{-1}\left(F_X(x)\right),x\right), \qquad \alpha_2(x) = 1.$$

With $\tilde{Y}_i = (\tilde{Y}_{i1}\ \tilde{Y}_{i2})' = (1\ Y_i)'$, we have

$$\alpha(x)'\tilde{y} = y - g(\left(F_W^{-1}\left(F_X(x)\right),x\right).$$

Applying the results in Theorem A.2, we have

$$\int_{\mathbb{U}_2} K\left(\frac{W_i - t(X_i)}{b_N} + \frac{\partial t}{\partial x}(X_i)\cdot u_2, u_2\right) du_2 = \int_u K\left(u, \frac{W_i - F_W^{-1}(F_X(X_i))}{b_N} + \frac{f_X(X_i)}{f_W(F_W^{-1}(F_X(X_i)))}\cdot u\right) du,$$

Substituting this into the result from Theorem A.2 we get

$$\sqrt{N}\left(\hat{\theta}_{\text{pam},g} - \overline{g}_{\text{pam}}\right)$$

$$= \frac{1}{\sqrt{N} b_N} \sum_{i=1}^{N} \left( \left(Y_i - g(\left(F_W^{-1}\left(F_X(X_i)\right), X_i\right)\right) \cdot \int_u K\left( \frac{W_i - F_W^{-1}\left(F_X(X_i)\right)}{b_N} + \frac{f_X(X_i)}{f_W\left(F_W^{-1}\left(F_X(X_i)\right)\right)} \cdot u, u \right) du \right.$$
$$\left. -\mathbb{E}\left[ \left(Y - g(\left(F_W^{-1}\left(F_X(X)\right), X\right)\right) \cdot \int_u K\left( \frac{W - F_W^{-1}\left(F_X(X)\right)}{b_N} + \frac{f_X(X)}{f_W\left(F_W^{-1}\left(F_X(X)\right)\right)} \cdot u, u \right) du \right] \right)$$
$$+ o_p(1).$$

□

**Proof of Lemma A.13:** We prove the result in three parts. First, we show

$$\frac{1}{N}\sum_{i=1}^{N} g\left(\hat{F}_W^{-1}\left(F_X\left(X_i\right)\right), X_i\right) - \frac{1}{N}\sum_{i=1}^{N} g\left(F_W^{-1}\left(F_X\left(X_i\right)\right), X_i\right)$$
$$= \frac{1}{N}\sum_{i=1}^{N} g_w\left(F_W^{-1}\left(F_X\left(X_i\right)\right), X_i\right) \cdot \left(\hat{F}_W^{-1}\left(F_X\left(X_i\right)\right) - F_W^{-1}\left(F_X\left(X_i\right)\right)\right) + o_p\left(N^{-1/2}\right) \tag{B.16}$$

Second, we will prove that

$$\frac{1}{N}\sum_{i=1}^{N} g_w\left(F_W^{-1}\left(F_X\left(X_i\right)\right), X_i\right) \cdot \left(\hat{F}_W^{-1}\left(F_X\left(X_i\right)\right) - F_W^{-1}\left(F_X\left(X_i\right)\right)\right)$$
$$= \frac{1}{N}\sum_{i=1}^{N} \frac{g_w\left(F_W^{-1}\left(F_X\left(X_i\right)\right), X_i\right)}{f_W\left(F_W^{-1}\left(F_X\left(X_i\right)\right)\right)} \cdot \left(\hat{F}_W\left(F_W^{-1}\left(F_X\left(X_i\right)\right)\right) - F_X\left(X_i\right)\right) + o_p\left(N^{-1/2}\right). \tag{B.17}$$

Third, we will show that

$$\frac{1}{N}\sum_{i=1}^{N} \frac{g_w\left(F_W^{-1}\left(F_X\left(X_i\right)\right), X_i\right)}{f_W\left(F_W^{-1}\left(F_X\left(X_i\right)\right)\right)} \cdot \left(\hat{F}_W\left(F_W^{-1}\left(F_X\left(X_i\right)\right)\right) - F_X\left(X_i\right)\right)$$
$$= \frac{1}{N}\sum_{i=1}^{N} q_W(W_i) + o_p\left(N^{-1/2}\right). \tag{B.18}$$

Together these three claims, (B.16)-(B.18), imply the result in the Lemma.
First we prove (B.16).

$$\left| \frac{1}{N}\sum_{i=1}^{N} g\left(\hat{F}_W^{-1}\left(F_X\left(X_i\right)\right), X_i\right) - \frac{1}{N}\sum_{i=1}^{N} g\left(F_W^{-1}\left(F_X\left(X_i\right)\right), X_i\right) \right.$$
$$\left. - \frac{1}{N}\sum_{i=1}^{N} g_w\left(F_W^{-1}\left(F_X\left(X_i\right)\right), X_i\right) \cdot \left(\hat{F}_W^{-1}\left(F_X\left(X_i\right)\right) - F_W^{-1}\left(F_X\left(X_i\right)\right)\right) \right|$$
$$\le \sup_x \left| g\left(\hat{F}_W^{-1}\left(F_X\left(x\right)\right), x\right) - g\left(F_W^{-1}\left(F_X\left(x\right)\right), x\right) \right.$$
$$\left. - g_w\left(F_W^{-1}\left(F_X\left(x\right)\right), x\right) \cdot \left(\hat{F}_W^{-1}\left(F_X\left(x\right)\right) - F_W^{-1}\left(F_X\left(x\right)\right)\right) \right|$$
$$\le \sup_{w,x} \left| \frac{\partial^2}{\partial w^2} g(w,x) \right| \cdot \sup_{q\in[0,1]} \left| \hat{F}_W^{-1}\left(F_X\left(x\right)\right) - F_W^{-1}\left(F_X\left(x\right)\right) \right|^2.$$

By Lemma A.3 it follows that for all $\delta < 1/2$, $\sup_{q\in[0,1]} N^{\delta} \cdot \left| \hat{F}_W^{-1}\left(F_X\left(x\right)\right) - F_W^{-1}\left(F_X\left(x\right)\right) \right| = o_p(1)$. In combination with the fact that $\frac{\partial^2}{\partial w^2} g(w,x)$ is bounded this implies that

$$\sup_{w,x} \left| \frac{\partial^2}{\partial w^2} g(w,x) \right| \cdot \sup_{q\in[0,1]} \left| \hat{F}_W^{-1}\left(F_X\left(x\right)\right) - F_W^{-1}\left(F_X\left(x\right)\right) \right|^2 = o_p\left(N^{-1/2}\right).$$

This finishes the proof of (B.16).
Next, we prove (B.17).

$$\left| \frac{1}{N}\sum_{i=1}^{N} g_w\left(F_W^{-1}\left(F_X\left(X_i\right)\right), X_i\right) \cdot \left(\hat{F}_W^{-1}\left(F_X\left(X_i\right)\right) - F_W^{-1}\left(F_X\left(X_i\right)\right)\right) \right.$$

$$+\frac{1}{N}\sum_{i=1}^{N}\frac{g_w\left(F_W^{-1}\left(F_X\left(X_i\right)\right),X_i\right)}{f_W\left(F_W^{-1}\left(F_X\left(X_i\right)\right)\right)}\cdot\left(\hat{F}_W\left(F_W^{-1}\left(F_X\left(X_i\right)\right)\right)-F_X\left(X_i\right)\right)\Bigg|$$

$$\leq \sup_{w,x,q}\left|g_w\left(w,x\right)\cdot\left(\hat{F}_W^{-1}\left(q\right)-F_W^{-1}\left(q\right)\right)+\frac{g_w\left(w,x\right)}{f_W\left(F_W^{-1}(q)\right)}\cdot\left(\hat{F}_W\left(F_W^{-1}\left(q\right)\right)-q\right)\right|$$

$$\leq \sup_{w,x}\left|g_w\left(w,x\right)\right|\cdot\sup_{q\in[0,1]}\left|\left(\hat{F}_W^{-1}\left(q\right)-F_W^{-1}\left(q\right)\right)+\frac{1}{f_W\left(F_W^{-1}(q)\right)}\cdot\left(\hat{F}_W\left(F_W^{-1}\left(q\right)\right)-q\right)\right|$$

By Lemma A.6, for all $0<\eta<5/7$,

$$\sup_{q\in[0,1]}N^{\eta}\cdot\left|\hat{F}_W^{-1}(q)-F_W^{-1}(q)+\frac{1}{f_W(F_W^{-1}(q))}\left(\hat{F}_W(F_W^{-1}(q))-q\right)\right|\xrightarrow{p}0.$$

So that (B.17) holds.
Finally, let us prove (B.18).

$$\frac{1}{N}\sum_{i=1}^{N}\frac{g_w\left(F_W^{-1}\left(F_X\left(X_i\right)\right),X_i\right)}{f_W\left(F_W^{-1}\left(F_X\left(X_i\right)\right)\right)}\cdot\left(\hat{F}_W\left(F_W^{-1}\left(F_X\left(X_i\right)\right)\right)-F_X\left(X_i\right)\right)$$

$$=\frac{1}{N}\sum_{i=1}^{N}\frac{g_w\left(F_W^{-1}\left(F_X\left(X_i\right)\right),X_i\right)}{f_W\left(F_W^{-1}\left(F_X\left(X_i\right)\right)\right)}\cdot\left(\frac{1}{N}\sum_{j=1}^{N}1_{W_j\leq F_W^{-1}\left(F_X\left(X_i\right)\right)}-F_X\left(X_i\right)\right)$$

$$=\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{g_w\left(F_W^{-1}\left(F_X\left(X_i\right)\right),X_i\right)}{f_W\left(F_W^{-1}\left(F_X\left(X_i\right)\right)\right)}\cdot\left(1_{F_W\left(W_j\right)\leq F_X\left(X_i\right)}-F_X\left(X_i\right)\right).$$

This is a two-sample V-statistic. The projection is obtained if we fix $X_i=x$ and take the expectation over $W_j$ which, because $F(W_j)$ has a standard uniform distribution, gives 0. Second if we fix $W_j=w$ and take the expectation over $X_i$ which gives $q_W(w)$ defined above. Thus,

$$\frac{1}{N}\sum_{i=1}^{N}\frac{g_w\left(F_W^{-1}\left(F_X\left(X_i\right)\right),X_i\right)}{f_W\left(F_W^{-1}\left(F_X\left(X_i\right)\right)\right)}\cdot\left(\hat{F}_W\left(F_W^{-1}\left(F_X\left(X_i\right)\right)\right)-F_X\left(X_i\right)\right)=\frac{1}{N}\sum_{i=1}^{N}q_W(W_i)+o_p\left(N^{-1/2}\right),$$

which is the claim in (B.18). □
**Proof of Lemma A.14:** We prove this result in two steps. First we prove

$$\Bigg|\frac{1}{N}\sum_{i=1}^{N}g\left(F_w^{-1}\left(\hat{F}_X\left(X_i\right)\right),X_i\right)-\frac{1}{N}\sum_{i=1}^{N}g\left(F_w^{-1}\left(F_X\left(X_i\right)\right),X_i\right)$$
$$-\frac{g_w\left(F_W^{-1}\left(F_X\left(X_i\right)\right),X_i\right)}{f_w\left(F_W^{-1}\left(F_X\left(X_i\right)\right)\right)}\cdot\left(\hat{F}_X(X_i)-F_X(X_i)\right)\Bigg|=o_p\left(N^{-1/2}\right). \quad \text{(B.19)}$$

Second, we prove

$$\frac{1}{N}\sum_{i=1}^{N}\frac{g_w\left(F_W^{-1}\left(F_X\left(X_i\right)\right),X_i\right)}{f_w\left(F_W^{-1}\left(F_X\left(X_i\right)\right)\right)}\cdot\left(\hat{F}_X(X_i)-F_X(X_i)\right)=\frac{1}{N}\sum_{i=1}^{N}r_X\left(X_i\right)+o_p\left(N^{-1/2}\right). \quad \text{(B.20)}$$

Together these two results imply the claim in Lemma A.14.
First we prove (B.19). By a second order Taylor series expansion, using the fact that $g(w,x)$ is at least twice continuously differentiable,

$$\Bigg|\frac{1}{N}\sum_{i=1}^{N}g\left(F_w^{-1}\left(\hat{F}_X\left(X_i\right)\right),X_i\right)-\frac{1}{N}\sum_{i=1}^{N}g\left(F_w^{-1}\left(F_X\left(X_i\right)\right),X_i\right)$$
$$-\frac{g_w\left(F_W^{-1}\left(F_X\left(X_i\right)\right),X_i\right)}{f_w\left(F_W^{-1}\left(F_X\left(X_i\right)\right)\right)}\cdot\left(\hat{F}_X(X_i)-F_X(X_i)\right)\Bigg|$$

$$\leq\sup_{x}\left|g\left(F_w^{-1}\left(\hat{F}_X\left(x\right)\right),x\right)-g\left(F_w^{-1}\left(F_X\left(x\right)\right),x\right)-\frac{g_w\left(F_W^{-1}\left(F_X\left(x\right)\right),x\right)}{f_w\left(F_W^{-1}\left(F_X\left(x\right)\right)\right)}\cdot\left(\hat{F}_X(x)-F_X(x)\right)\right|$$

$$\leq \frac{1}{2}\sup_{w,x}\left|\frac{g_{ww}(w,x)}{f_W(w)} - \frac{g_w(w,x)\cdot f_W'(w)}{(f_W(w))^2}\right|\sup_x\left|\hat{F}_X(x) - F_X(x)\right|^2 = o_p\left(N^{-1/2}\right),$$

by Lemma A.3 and where $g_{ww}$ is the second derivative of $g$ with respect to $w$ and $f_W'$ the derivative of the density of $W$. This finishes the proof of (B.19).
Second we prove (B.20).

$$\frac{1}{N}\sum_{i=1}^N \frac{g_w\left(F_W^{-1}\left(F_X\left(X_i\right)\right),X_i\right)}{f_w\left(F_W^{-1}\left(F_X\left(X_i\right)\right)\right)}\cdot\left(\hat{F}_X(X_i) - F_X(X_i)\right)$$

$$= \frac{1}{N^2}\sum_{i=1}^N\sum_{j=1}^N \frac{g_w\left(F_W^{-1}\left(F_X\left(X_i\right)\right),X_i\right)}{f_w\left(F_W^{-1}\left(F_X\left(X_i\right)\right)\right)}\cdot\left(\mathbf{1}_{X_j\leq X_i} - F_X(X_i)\right)$$

This is a one-sample V-statistic. To obtain the projection we first fix $X_i = x$ and take the expectation over $X_j$. This gives 0 for all $x$. Second, we fix $X_j = x$ and take the expectation over $X_i$. This gives $r_X(x)$ defined above. This finishes the proof of (B.20), and thus completes the proof of Lemma A.14. □

**Proof of Lemma A.15:**
Adding and subtracting terms we have

$$\hat{\beta}^{\rm lc} - \beta^{\rm lc} = \frac{1}{N}\sum_{i=1}^N \frac{\partial\hat{g}}{\partial w}(W_i,X_i)\cdot d(W_i)\cdot(X_i - \hat{m}(W_i)) - \frac{1}{N}\sum_{i=1}^N \frac{\partial g}{\partial w}(W_i,X_i)\cdot d(W_i)\cdot(X_i - m(W_i))$$

$$= \frac{1}{N}\sum_{i=1}^N \frac{\partial\hat{g}}{\partial w}(W_i,X_i)\cdot d(W_i)\cdot(X_i - \hat{m}(W_i)) - \frac{1}{N}\sum_{i=1}^N \frac{\partial g}{\partial w}(W_i,X_i)\cdot d(W_i)\cdot(X_i - \hat{m}(W_i)) \quad (B.21)$$

$$-\left(\frac{1}{N}\sum_{i=1}^N \frac{\partial\hat{g}}{\partial w}(W_i,X_i)\cdot d(W_i)\cdot(X_i - m(W_i)) - \frac{1}{N}\sum_{i=1}^N \frac{\partial g}{\partial w}(W_i,X_i)\cdot d(W_i)\cdot(X_i - m(W_i))\right) \quad (B.22)$$

$$+\frac{1}{N}\sum_{i=1}^N \frac{\partial\hat{g}}{\partial w}(W_i,X_i)\cdot d(W_i)\cdot(X_i - m(W_i)) - \frac{1}{N}\sum_{i=1}^N \frac{\partial g}{\partial w}(W_i,X_i)\cdot d(W_i)\cdot(X_i - m(W_i)) \quad (B.23)$$

$$+\frac{1}{N}\sum_{i=1}^N \frac{\partial g}{\partial w}(W_i,X_i)\cdot d(W_i)\cdot(X_i - \hat{m}(W_i)) - \frac{1}{N}\sum_{i=1}^N \frac{\partial g}{\partial w}(W_i,X_i)\cdot d(W_i)\cdot(X_i - m(W_i)) \quad (B.24)$$

$$+\frac{1}{N}\sum_{i=1}^N \frac{\partial g}{\partial w}(W_i,X_i)\cdot d(W_i)\cdot(X_i - m(W_i)) - \beta^{\rm lc}. \quad (B.25)$$

Because (B.23) is equal to $\beta_{{\rm lc},g} - \overline{g}_{\rm lc}$, (B.24) is equal to $\beta_{{\rm lc},m} - \overline{g}_{\rm lc}$, and (B.25) is equal to $\overline{g}_{\rm lc} - \beta^{\rm lc}$, it follows that it is sufficient for the proof of Lemma A.15 to show that the sum of (B.21) and (B.22) is $o_p(N^{-1/2})$. We can write the sum of (B.21) and (B.22) as

$$\frac{1}{N}\sum_{i=1}^N \frac{\partial\hat{g}}{\partial w}(W_i,X_i)\cdot d(W_i)\cdot(X_i - \hat{m}(W_i)) - \frac{1}{N}\sum_{i=1}^N \frac{\partial g}{\partial w}(W_i,X_i)\cdot d(W_i)\cdot(X_i - \hat{m}(W_i))$$

$$-\left(\frac{1}{N}\sum_{i=1}^N \frac{\partial\hat{g}}{\partial w}(W_i,X_i)\cdot d(W_i)\cdot(X_i - m(W_i)) - \frac{1}{N}\sum_{i=1}^N \frac{\partial g}{\partial w}(W_i,X_i)\cdot d(W_i)\cdot(X_i - m(W_i))\right)$$

$$= \frac{1}{N}\sum_{i=1}^N d(W_i)\cdot\left(\frac{\partial\hat{g}}{\partial w}(W_i,X_i) - \frac{\partial g}{\partial w}(W_i,X_i)\right)\cdot(m(W_i) - \hat{m}(W_i))$$

$$\leq \sup_{w\in\mathbb{W}}|d(w)|\cdot\sup_{w\in\mathbb{W},x\in\mathbb{X}}\left|\frac{\partial\hat{g}}{\partial w}(w,x) - \frac{\partial g}{\partial w}(x,x)\right|\cdot\sup_{w\in\mathbb{W}}|\hat{m}(w) - m(w)| = C\cdot o_p\left(N^{-\eta}\right)\cdot o_p\left(N^{-\eta}\right),$$

for some $\eta > 1/4$, and so this expression is $o_P(N^{-1/2})$. □

**Proof of Lemma A.16:**
The proof consists of checking the conditions for Theorem A.1, and specializing the result in Theorem A.1 to the case in the Lemma.

We apply Theorem A.1 with $z = (w\ x)'$, $Z_i = (W_i\ X_i)'$, $\omega(z) = d(w) \cdot (x - m(w))$, $L = 2$, and $\lambda = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$. Then $\{\kappa : \kappa \le \lambda\} = \{\kappa_0, \kappa_1\} = \left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right\}$, and

$$h^{[\lambda]}(w,x) = \begin{pmatrix} h_1^{(\kappa_0)}(w,x) \\ h_2^{(\kappa_0)}(w,x) \\ h_1^{(\kappa_1)}(w,x) \\ h_2^{(\kappa_1)}(w,x) \end{pmatrix},$$

with

$$\begin{aligned} h_1^{(\kappa_0)}(w,x) &= f_{WX}(w,x) \\ h_2^{(\kappa_0)}(w,x) &= f_{WX}(w,x) \cdot g(w,x) \\ h_1^{(\kappa_1)}(w,x) &= \frac{\partial}{\partial w} f_{WX}(w,x) \\ h_2^{(\kappa_1)}(w,x) &= g(w,x) \cdot \frac{\partial}{\partial w} f_{WX}(w,x) + f_{WX}(w,,x) \cdot \frac{\partial}{\partial w} g(w,x). \end{aligned}$$

The functional of interest is

$$n\left(h^{[\lambda]}\right) = \frac{\partial}{\partial w} g(\cdot) = \frac{h_2^{(\kappa_1)}}{h_1^{(\kappa_0)}} - \frac{h_2^{(\kappa_0)} \cdot h_1^{(\kappa_1)}}{\left(h_1^{(\kappa_0)}\right)^2}$$

The derivatives of this functional are

$$\begin{aligned} \frac{\partial}{\partial h_1^{(\kappa_0)}} n\left(h^{[\lambda]}\right) &= -\frac{h_2^{(\kappa_1)}}{\left(h_1^{(\kappa_0)}\right)^2} + 2\frac{h_2^{(\kappa_0)} \cdot h_1^{(\kappa_1)}}{\left(h_1^{(\kappa_0)}\right)^3} \\ &= -\frac{f_{WX}(w,x) \cdot \frac{\partial}{\partial w} g(w,x) + g(w,x) \cdot \frac{\partial}{\partial w} f_{WX}(w,x)}{\left(f_{WX}(w,x)\right)^2} + 2\frac{g(w,x) \cdot f_{WX}(w,x) \cdot \frac{\partial}{\partial w} f_{WX}(w,x)}{\left(f_{WX}(w,x)\right)^3} \\ &= -\frac{\frac{\partial}{\partial w} g(w,x)}{f_{WX}(w,x)} + \frac{g(w,x) \cdot \frac{\partial}{\partial w} f_{WX}(w,x)}{\left(f_{WX}(w,x)\right)^2} \end{aligned}$$

$$\frac{\partial}{\partial h_2^{(\kappa_0)}} n\left(h^{[\lambda]}\right) = -\frac{h_1^{(\kappa_1)}}{\left(h_1^{(\kappa_0)}\right)^2} = -\frac{\frac{\partial}{\partial w} f_{WX}(w,x)}{\left(f_{WX}(w,x)\right)^2}$$

$$\frac{\partial}{\partial h_1^{(\kappa_1)}} n\left(h^{[\lambda]}\right) = -\frac{h_2^{(\kappa_0)}}{\left(h_1^{(\kappa_0)}\right)^2} = -\frac{g(w,x) \cdot f_{WX}(w,x)}{\left(f_{WX}(w,x)\right)^2} = -\frac{g(w,x)}{f_{WX}(w,x)}$$

$$\frac{\partial}{\partial h_2^{(\kappa_1)}} n\left(h^{[\lambda]}\right) = \frac{1}{h_1^{(\kappa_0)}} = \frac{1}{f_{WX}(w,x)}.$$

$$\begin{aligned} \alpha_{\kappa_0,1} &= d(w) \cdot (x - m(w)) \cdot f_W(w,x) \cdot \left( -\frac{\frac{\partial}{\partial w} g(w,x)}{f_{WX}(w,x)} + \frac{g(w,x) \cdot \frac{\partial}{\partial w} f_{WX}(w,x)}{\left(f_{WX}(w,x)\right)^2} \right) \\ &= d(w) \cdot (x - m(w)) \cdot \left( -\frac{\partial}{\partial w} g(w,x) + \frac{g(w,x) \cdot \frac{\partial}{\partial w} f_{WX}(w,x)}{f_{WX}(w,x)} \right) \end{aligned}$$

$$\alpha_{\kappa_0,2} = d(w) \cdot (x - m(w)) \cdot f_{WX}(w,x) \cdot \left( -\frac{\frac{\partial}{\partial w} f_{WX}(w,x)}{\left(f_{WX}(w,x)\right)^2} \right) = -d(w) \cdot (x - m(w)) \cdot \frac{\frac{\partial}{\partial w} f_{WX}(w,x)}{f_{WX}(w,x)}$$

$$\alpha_{\kappa_1,1} = d(w)\cdot(x-m(w))\cdot f_{WX}(w,x)\cdot\left(-\frac{g(w,x)}{f_{WX}(w,x)}\right) = -d(w)\cdot(x-m(w))\cdot g(w,x)$$

$$\alpha_{\kappa_1,2} = d(w)\cdot(x-m(w))\cdot f_{WX}(w,x)\cdot\frac{1}{f_{WX}(w,x)} = d(w)\cdot(x-m(w))$$

$$(-1)^{|\kappa_0|}\alpha^{(\kappa_0)}_{\kappa_0,1} = \alpha_{\kappa_0,1} = d(w)\cdot(x-m(w))\cdot\left(-\frac{\partial}{\partial w}g(w,x)+\frac{g(w,x)\cdot\frac{\partial}{\partial w}f_{WX}(w,x)}{f_{WX}(w,x)}\right)$$

$$(-1)^{|\kappa_0|}\alpha^{(\kappa_0)}_{\kappa_0,2} = \alpha_{\kappa_0,2} = -d(w)\cdot(x-m(w))\cdot\frac{\frac{\partial}{\partial w}f_{WX}(w,x)}{f_{WX}(w,x)}$$

$$\begin{aligned}(-1)^{|\kappa_1|}\alpha^{(\kappa_1)}_{\kappa_1,1} &= \frac{\partial}{\partial w}\Big(d(w)\cdot(x-m(w))\cdot g(w,x)\Big)\\ &= d(w)\cdot(x-m(w))\cdot\frac{\partial}{\partial w}g(w,x)+g(w,x)\cdot(x-m(w))\cdot\frac{\partial}{\partial w}d(w)-g(w,x)\cdot d(w)\cdot\frac{\partial}{\partial w}m(w)\end{aligned}$$

$$(-1)^{|\kappa_1|}\alpha^{(\kappa_1)}_{\kappa_1,2} = -\frac{\partial}{\partial w}\left(d(w)\cdot(x-m(w))\right) = -(x-m(w))\cdot\frac{\partial}{\partial w}d(w)+d(w)\cdot\frac{\partial}{\partial w}m(w)$$

Then

$$\begin{aligned}&\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(\sum_{\kappa\le\lambda}(-1)^{|\kappa|}\sum_{m=1}^{2}\alpha^{(\kappa)}_{\kappa m}(X_i)\tilde{Y}_{im}\right)\\ &\quad= (-1)^{|\kappa_0|}\alpha^{(\kappa_0)}_{\kappa_0,1}+Y_i\cdot(-1)^{|\kappa_0|}\alpha^{(\kappa_0)}_{\kappa_0,2}+(-1)^{|\kappa_1|}\alpha^{(\kappa_1)}_{\kappa_1,1}+Y_i\cdot(-1)^{|\kappa_1|}\alpha^{(\kappa_1)}_{\kappa_1,2}\\ &\quad= d(w)\cdot(x-m(w))\cdot\left(-\frac{\partial}{\partial w}g(w,x)+\frac{g(w,x)\cdot\frac{\partial}{\partial w}f_{WX}(w,x)}{f_{WX}(w,x)}\right)\\ &\qquad -Y_i\cdot d(w)\cdot(x-m(w))\cdot\frac{\frac{\partial}{\partial w}f_{WX}(w,x)}{f_{WX}(w,x)}\\ &\qquad +d(w)\cdot(x-m(w))\cdot\frac{\partial}{\partial w}g(w,x)+g(w,x)\cdot(x-m(w))\cdot\frac{\partial}{\partial w}d(w)-g(w,x)\cdot d(w)\cdot\frac{\partial}{\partial w}m(w)\\ &\qquad +Y_i\cdot\left(-(x-m(w))\cdot\frac{\partial}{\partial w}d(w)+d(w)\cdot\frac{\partial}{\partial w}m(w)\right)\\ &\quad= -\left(y-g(w,x)\right)\cdot\left(\frac{\frac{\partial}{\partial w}f_{WX}(w,x)}{f_{WX}(w,x)}\cdot d(w)\cdot(x-m(w))+(x-m(w))\cdot\frac{\partial}{\partial w}d(w)-d(w)\cdot\frac{\partial}{\partial w}m(w)\right).\end{aligned}$$

Since this expression has expectation zero, it follows that

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(\sum_{\kappa\le\lambda}(-1)^{|\kappa|}\sum_{m=1}^{2}\mathbb{E}\left[\alpha^{(\kappa)}_{\kappa m}(X_i)\tilde{Y}_{im}\right]\right) = 0,$$

and therefore

$$\sqrt{N}(\hat{\beta}^{\rm lc}-\beta^{\rm lc}) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(\sum_{\kappa\le\lambda}(-1)^{|\kappa|}\sum_{m=1}^{2}\alpha^{(\kappa)}_{\kappa m}(X_i)\tilde{Y}_{im}\right) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\delta(Y_i,W_i,X_i).$$

where

$$\delta(y,w,x) = -\left(y-g(w,x)\right)\cdot\left(\frac{\frac{\partial}{\partial w}f_{WX}(w,x)}{f_{WX}(w,x)}\cdot d(w)\cdot(x-m(w))+(x-m(w))\cdot\frac{\partial}{\partial w}d(w)-d(w)\cdot\frac{\partial}{\partial w}m(w)\right).$$

□.

**Proof of Lemma A.17:**

$$\mathbb{E}\left[\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\varepsilon_i\cdot\left(\hat{h}(X_i)-h(X_i)\right)\right)^2\right]$$

$$=\mathbb{E}\left[\left(\frac{1}{N^{3/2}}\sum_{i=1}^{N}\sum_{j=1}^{N}\varepsilon_i\cdot\left(\frac{Y_j}{b_N}\cdot K\left(\frac{X_i-X_j}{b_N}\right)-h(X_i)\right)\right)^2\right]$$

$$=\frac{1}{N^3}\cdot\mathbb{E}\left[\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N}\sum_{l=1}^{N}\varepsilon_i\cdot\left(\frac{Y_j}{b_N}\cdot K\left(\frac{X_i-X_j}{b_N}\right)-h(X_i)\right)\cdot\varepsilon_k\cdot\left(\frac{Y_l}{b_N}\cdot K\left(\frac{X_k-X_l}{b_N}\right)-h(X_k)\right)\right].$$

The terms with $i$, $j$, $k$, and $l$ all distinct have expectation zero. Ignoring terms of the type of which there are only $N^2$, which are of even lower order, the leading terms are of the type $i=k$. There are $N^3$ of those terms, of the form

$$\mathbb{E}\left[\varepsilon_i^2\cdot\left(\frac{Y_j}{b_N}\cdot K\left(\frac{X_i-X_j}{b_N}\right)-h(X_i)\right)\cdot\left(\frac{Y_l}{b_N}\cdot K\left(\frac{X_i-X_l}{b_N}\right)-h(X_i)\right)\right]$$

$$=\mathbb{E}\left[\varepsilon_i^2\cdot\frac{Y_j}{b_N}\cdot K\left(\frac{X_i-X_j}{b_N}\right)\cdot\frac{Y_l}{b_N}\cdot K\left(\frac{X_i-X_l}{b_N}\right)\right]$$

$$-\mathbb{E}\left[\varepsilon_i^2\frac{Y_j}{b_N}\cdot K\left(\frac{X_i-X_j}{b_N}\right)\cdot h(X_i\right]$$

$$-\mathbb{E}\left[\varepsilon_i^2 h(X_i)\cdot\frac{Y_l}{b_N}\cdot K\left(\frac{X_i-X_l}{b_N}\right)\right]$$

$$+\mathbb{E}\left[\varepsilon_i^2 h(X_i)\cdot h(X_i)\right].$$

Define $m_1(x)=\mathbb{E}[\varepsilon_i^2|X_i=x]$, and $m_2(x)=\mathbb{E}[Y_i|X_i=x]$, (so that $h(x)=m_2(x)\cdot f_X(x)$) so that these four expectations can be written as

$$\mathbb{E}\left[m_1(X_i)\cdot\frac{m_2(X_j)}{b_N}\cdot K\left(\frac{X_i-X_j}{b_N}\right)\cdot\frac{m_2(X_l)}{b_N}\cdot K\left(\frac{X_i-X_l}{b_N}\right)\right]$$

$$-\mathbb{E}\left[m_1(X_i)\frac{m_2(X_j}{b_N}\cdot K\left(\frac{X_i-X_j}{b_N}\right)\cdot h(X_i\right]$$

$$-\mathbb{E}\left[m_1(X_i)h(X_i)\cdot\frac{m_2(X_l)}{b_N}\cdot K\left(\frac{X_i-X_l}{b_N}\right)\right]$$

$$+\mathbb{E}\left[m_1(X_i)h(X_i)\cdot h(X_i)\right].$$

Let us look at the first term:

$$\mathbb{E}\left[m_1(X_i)\cdot\frac{m_2(X_j)}{b_N}\cdot K\left(\frac{X_i-X_j}{b_N}\right)\cdot\frac{m_2(X_l)}{b_N}\cdot K\left(\frac{X_i-X_l}{b_N}\right)\right]$$

$$=\int_{x_1}\int_{x_2}\int_{x_3}m_1(x_1)\cdot\frac{m_2(x_2)}{b_N}\cdot K\left(\frac{x_1-x_2}{b_N}\right)\cdot\frac{m_2(x_3)}{b_N}\cdot K\left(\frac{x_1-x_3}{b_N}\right)f_X(x_3)f_X(x_2)f_X(x_1)dx_3dx_2dx_1.$$

Change variables from $x_2$ to $u=(x_2-x_1)/b_N$ and from $x_3$ to $v=(x_3-x_1)/b_n$, both with Jacobian $b_N$, to get

$$\int_{x_1}\int_u\int_v m_1(x_1)\cdot m_2(x_1+b_N\cdot u)\cdot K(u)\cdot m_2(x_1+b_N\cdot u)\cdot K(v)\,f_X(x_1+b_N\cdot v)f_X(x_1+b_N\cdot u)f_X(x_1)dvdudx_1$$

$$=\int_x m_1(x)\cdot(m_2(x))^2\cdot(f_X(x))^3+O_p(b_N)$$

$$=\int_x m_1(x)\cdot(h(x))^2\cdot f_X(x)+O_p(b_N).$$

$$=\mathbb{E}\left[m_1(X_i)\cdot(h(X_i))^2\right]+O_p(b_N).$$

By the same argument the second and the third terms are equal to

$$-\mathbb{E}\left[m_1(X_i)\cdot(h(X_i))^2\right]+O_p(b_N),$$

so that the sum is $O_p(b_N) = o_p(1)$. □

**Proof of Lemma A.18:**

Define $h_1(x) = F_X(x)$, and $h_2(x) = g(x) \cdot f_X(x)$, so we can write

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \varepsilon_i \cdot (\hat{g}(X_i) - g(X_i)) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \varepsilon_i \cdot \left( \frac{\hat{h}_2(X_i)}{\hat{h}_1(X_i)} - \frac{h_2(X_i)}{h_1(X_i)} \right).$$

First we linearize to show that this is equal to

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \varepsilon_i \cdot \frac{\hat{h}_2(X_i) - h_2(X_i))}{h_1(X_i)} - \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \varepsilon_i \cdot \frac{h_2(X_i) \cdot (\hat{h}_1(X_i) - h_1(X_i))}{h_1^2(X_i)} + o_p(1).$$

Then we apply Lemma A.17, first with $\tilde{\varepsilon}_i = \varepsilon_i / h_1(X_i)$ and then with $\tilde{\varepsilon}_i = \varepsilon_i h_2(X_i)/h_1^2(X_i)$, to show that the first two terms at $o_p(1)$.

For the linearization,

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \varepsilon_i \cdot \left( \frac{\hat{h}_2(X_i)}{\hat{h}_1(X_i)} - \frac{h_2(X_i)}{h_1(X_i)} \right) - \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \varepsilon_i \cdot \frac{\hat{h}_2(X_i) - h_2(X_i))}{h_1(X_i)} + \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \varepsilon_i \cdot \frac{h_2(X_i) \cdot (\hat{h}_1(X_i) - h_1(X_i))}{h_1^2(X_i)}$$

$$= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \varepsilon_i^2 \cdot \left( -h_1(X_i) \left( \hat{h}_2(X_i) - h_2(X_i) \right) \cdot \left( \hat{h}_2(X_i) - h_2(X_i) \right) + h_2(X_i) \cdot \left( \hat{h}_1(X_i) - h_1(X_i) \right)^2 \right)$$

$$\leq N^{1/2} \sup_{x \in \mathbb{X}} \mathbb{E}[\varepsilon_i^2 | X_i = x] \cdot \sup_{x \in \mathbb{X}} \left| \hat{h}_2(x) - h_2(x) \right| \cdot \sup_{x \in \mathbb{X}} \left| \hat{h}_1(x) - h_1(x) \right|$$

$$+ N^{1/2} \sup_{x \in \mathbb{X}} \mathbb{E}[\varepsilon_i^2 | X_i = x] \cdot \sup_{x \in \mathbb{X}} \left| \hat{h}_1(x) - h_1(x) \right|^2 = o_p(1).$$

□

**Proof of Lemma A.19:** Define

$$r(w) = \mathbb{E}\left[ \frac{\partial}{\partial w} g(W_i, X_i) \cdot d(W_i) \right| W_i = w \Bigg]$$

Then

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{\partial}{\partial w} g(W_i, X_i) \cdot d(W_i) \cdot \widehat{m}(W_i) - \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{\partial}{\partial w} g(W_i, X_i) \cdot d(W_i) \cdot m(W_i)$$

$$= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{\partial}{\partial w} g(W_i, X_i) \cdot d(W_i) \cdot (\widehat{m}(W_i) - m(W_i))$$

$$= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left( \frac{\partial}{\partial w} g(W_i, X_i) \cdot d(W_i) - r(W_i) \right) \cdot (\widehat{m}(W_i) - m(W_i)) + \frac{1}{\sqrt{N}} \sum_{i=1}^{N} r(W_i) \cdot (\widehat{m}(W_i) - m(W_i)).$$

Define

$$\varepsilon_i = \frac{\partial}{\partial w} g(W_i, X_i) \cdot d(W_i) - r(W_i),$$

so that $\mathbb{E}[\varepsilon_i | W_i = w] = 0$, and we can apply Lemma A.18 to show that the first term is $o_p(1)$.

For the second term we apply Theorem A.1, which leads to

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} r(W_i) \cdot (\widehat{m}(W_i) - m(W_i)) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} r(W_i) \cdot (X_i - m(W_i)) + o_p(1)$$

$$= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \mathbb{E}\left[ \frac{\partial}{\partial w} g(W_i, X_i) \cdot d(W_i) \right| W_i \Bigg] \cdot (X_i - m(W_i)) + o_p(1).$$

□

**Proof of Theorem A.4:** Define $\phi(z_1, z_2) = (\psi(z_1, z_2) + \psi(z_2, z_1))/2$. Then $V = \sum_{i=1}^{N} \sum_{j=1}^{N} \phi(Z_i, Z_j)/N^2$ is a V-statistic with a symmetric kernel. In the notation of Lehman (1998),

$$\sigma_1^2 = \text{Cov}\left( \phi(Z_i, Z_j), \phi(Z_i, Z_k) \right),$$

for $i, j, k$ distinct, which simplifies to $\sigma_1^2 = \sigma^2/4$. Therefore, by Theorems 6.1.2 (with $a = 2$) and 6.2.1 in Lehman (1998), the result follows. □.

## Appendix C: Proofs of Theorems in Text

**Proof of Theorem 3.1** Define

$$V_{\lambda,i} = \lambda \cdot X_i \cdot d(W_i) + W_i,$$

$$h(\lambda, a) = \mathrm{pr}(V_\lambda \le a) = F_{V_\lambda}(a), \qquad \text{and} \quad k(w, x, \lambda) = h(\lambda, \lambda \cdot x \cdot d(w, x) + w).$$

First we focus on

$$\beta^{\mathrm{lr,v}}(\lambda) = \mathbb{E}\left[g(F_W^{-1}(F_{V_\lambda}(V_{\lambda,i})), X)\right] = \mathbb{E}\left[g(F_W^{-1}(k(W_i, X_i, \lambda)), X_i)\right].$$

We then prove four results. First, we show that for small $\lambda$, $\beta^{\mathrm{lr,v}}(\lambda)$ and $\beta^{\mathrm{lr}}(\lambda)$ are close, or

$$\beta^{\mathrm{lr,v}}(\lambda) = \beta^{\mathrm{lr}}(\lambda) + o(\lambda). \tag{C.1}$$

Second, we show that

$$\beta^{\mathrm{lr,v}}(\lambda) = \mathbb{E}[g(W, X)]$$
$$+\mathbb{E}\left[\frac{\partial g}{\partial w}(W_i, X_i)\frac{1}{f_W(W_i)}\left(k(W_i, X_i, \lambda) - k(W_i, X_i, 0)\right)\right] + o(\lambda). \tag{C.2}$$

Next we show that $\beta^{\mathrm{lr,v}}(\lambda)$ has the two representations in Theorem 3.1. In particular, the third part of the proof shows that $\beta^{\mathrm{lc,v}} = \frac{\partial \beta^{\mathrm{lr,v}}}{\partial \lambda}(0)$ satisfies

$$\beta^{\mathrm{lc,v}} = \mathbb{E}\left[\frac{\partial g}{\partial w}(W_i, X_i) \cdot \left(X_i \cdot d(W_i, X_i) - \mathbb{E}\left[X_i \cdot d(W_i, X_i) \middle| W_i\right]\right)\right]. \tag{C.3}$$

Fourth, we show that $\beta^{\mathrm{lc,v}}$ satisfies

$$\beta^{\mathrm{lc,v}} = \mathbb{E}\left[\delta(W_i, X_i) \cdot \frac{\partial^2 g}{\partial w \partial x}(W_i, X_i)\right]. \tag{C.4}$$

We start with the proof of (C.1). Define

$$u(w, x, \lambda) = \lambda \cdot x \cdot d(w, x)^{1-|\lambda|} + \sqrt{1-\lambda^2} \cdot w, \qquad \text{and} \quad u(w, x, \lambda) = \lambda \cdot x \cdot d(w, x) + w.$$

Then

$$\sup_{w,x} |u(w, x, \lambda) - v(w, x, \lambda)| = O\left(\lambda^2\right).$$

Define also

$$h_U(\lambda, a) = \mathrm{pr}(U_\lambda \le a), \qquad \text{and} \quad k_U(w, x, \lambda) = h_U(\lambda, u(w, x, \lambda)).$$

Then

$$\sup_{a} |h_U(\lambda, a) - h(\lambda, a)| = O\left(\lambda^2\right),$$

and

$$\sup_{w,x} |k_U(w, x, \lambda) - k(w, x, \lambda)| = O\left(\lambda^2\right).$$

Combined with the smoothness assumptions, this implies that

$$\beta^{\mathrm{lr,v}}(\lambda) - \beta^{\mathrm{lr}}(\lambda) = \mathbb{E}\left[g(F_W^{-1}(k(W_i, X_i, \lambda)), X_i)\right] - \mathbb{E}\left[g(F_W^{-1}(k_U(W_i, X_i, \lambda)), X_i)\right] = O(\lambda^2).$$

This finishes the proof of (C.1).
Next, we prove (C.2). Let $c_1$ and $c_2$ satisfy

$$\sup_{x,w,\gamma,\lambda} |k(w, x, \lambda + \gamma) - k(w, x, \lambda)| \le c_1 \cdot \gamma,$$

and

$$\sup_{w,x}\left|\frac{\partial^2}{\partial w^2}g(w,x)\right|\le c_2,$$

respectively. Then, applying Lemma A.1 with $f(a)=g(F_W^{-1}(a),x)$ and $h(\lambda)=k(w,x,\lambda)$, we obtain

$$\left|g(F_W^{-1}(k(w,x,\lambda),x)-\left(g(F_W^{-1}(k(w,x,0),x)+\frac{\frac{\partial}{\partial w}g(F_W^{-1}(k(w,x,0),x)}{f_W(F_W^{-1}(k(w,x,0)))}\left(k(w,x,\lambda)-k(w,x,0)\right)\right)\right|$$
$$\le c_2c_1^2\lambda^2=o(\lambda).$$

Since the bound does not depend on $x$ and $w$, we can average over $W$ and $X$ and it follows that

$$\left|\mathbb{E}\left[g(F_W^{-1}(k(W,X,\lambda),X)\right]-\mathbb{E}\left[g(W,X)\right]-\mathbb{E}\left[\frac{\frac{\partial}{\partial w}g(W,X)}{f_W(W)}\left(k(W,X,\lambda)-W\right)\right]\right|,$$

where we also use the fact that $k(w,x,0)=F_W(w)$. This finishes the proof of (C.2).
Now we prove (C.3). By definition,

$$\begin{aligned}
h(\lambda,a)&=\text{pr}\left(V_{\lambda,i}<a\right)=\text{pr}\left(V_{\lambda,i}<a,W_i<w_m\right)+\text{pr}\left(V_{\lambda,i}<a,W_i\ge w_m\right)\\
&=\text{pr}\left(\lambda\cdot X_i\cdot d(W_i,X_i)+W_i\le a,W_i<w_m\right)\\
&\quad+\text{pr}\left(\lambda\cdot X_i\cdot d(W_i,X_i)+W_i\le a,W_i\ge w_m\right).\\
&=\text{pr}\left(\lambda\cdot X_i\cdot(W_i-w_l)+W_i\le a,W_i<w_m\right)\\
&\quad+\text{pr}\left(\lambda\cdot X_i\cdot(w_u-W_i)+W_i\le a,W_i\ge w_m\right)\\
&=\text{pr}\left(W_i\le\min\left(w_m,\frac{a+\lambda\cdot X_i\cdot w_l}{1+\lambda\cdot X_i}\right)\right)\\
&\quad+\text{pr}\left(w_m\le W_i\le\frac{a-\lambda\cdot X_i\cdot w_u}{1-\lambda\cdot X_i}\right).
\end{aligned}$$

For $\lambda$ sufficiently close to zero, we can write this as

$$\begin{aligned}
h(\lambda,a)&=1_{a>w_m}\cdot\text{pr}(W_i\le w_m)+1_{a\le w_m}\cdot\text{pr}\left(W_i\le\frac{a+\lambda\cdot X_i\cdot w_l}{1+\lambda\cdot X_i}\right)\\
&\quad+1_{a>w_m}\cdot\text{pr}\left(w_m<W_i\le\frac{a-\lambda\cdot X_i\cdot w_u}{1-\lambda\cdot X_i}\right)\\
&=1_{a\le w_m}\cdot\text{pr}\left(W_i\le\frac{a+\lambda\cdot X_i\cdot w_l}{1+\lambda\cdot X_i}\right)\\
&\quad+1_{a>w_m}\cdot\text{pr}\left(W_i\le\frac{a-\lambda\cdot X_i\cdot w_u}{1-\lambda\cdot X_i}\right)\\
&=1_{a\le w_m}\cdot\mathbb{E}\left[\text{pr}\left(W_i\le\frac{a+\lambda\cdot X_i\cdot w_l}{1+\lambda\cdot X_i}\,\middle|\,X_i\right)\right]\\
&\quad+1_{a>w_m}\cdot\mathbb{E}\left[\text{pr}\left(W_i\le\frac{a-\lambda\cdot X_i\cdot w_u}{1-\lambda\cdot X_i}\,\middle|\,X_i\right)\right]\\
&=1_{a\le w_m}\cdot\int F_{W|X}\left(\frac{a+\lambda\cdot z\cdot w_l}{1+\lambda\cdot z}\,\middle|\,z\right)f_X(z)dz\\
&\quad+1_{a>w_m}\cdot\int F_{W|X}\left(\frac{a-\lambda\cdot z\cdot w_u}{1-\lambda\cdot z}\,\middle|\,z\right)f_X(z)dz
\end{aligned}$$

Substituting $a=\lambda\cdot x\cdot d(w,x)+w$, we get

$$\begin{aligned}
k(w,x,\lambda)&=1_{\lambda\cdot x\cdot d(w,x)+w\le w_m}\cdot\int F_{W|X}\left(\frac{\lambda\cdot x\cdot d(w,x)+w+\lambda\cdot z\cdot w_l}{1+\lambda\cdot z}\,\middle|\,z\right)f_X(z)dz\\
&\quad+1_{\lambda\cdot x\cdot d(w,x)+w>w_m}\cdot\int F_{W|X}\left(\frac{\lambda\cdot x\cdot d(w,x)+w-\lambda\cdot z\cdot w_u}{1-\lambda\cdot z}\,\middle|\,z\right)f_X(z)dz
\end{aligned}$$

$$= 1_{\lambda\cdot x\cdot(w-w_l)+w\le w_m} 1_{w\le w_m}\cdot\int F_{W|X}\left(\frac{\lambda\cdot x\cdot(w-w_m)+w+\lambda\cdot z\cdot w_l}{1+\lambda\cdot z}\middle| z\right) f_X(z)dz$$
$$+1_{\lambda\cdot x\cdot(w_u-w)+w\le w_m} 1_{w>w_m}\cdot\int F_{W|X}\left(\frac{\lambda\cdot x\cdot(w_u-w)+w+\lambda\cdot z\cdot w_l}{1+\lambda\cdot z}\middle| z\right) f_X(z)dz$$
$$+1_{\lambda\cdot x\cdot(w-w_l)+w>w_m}\cdot 1_{w\le w_m}\int F_{W|X}\left(\frac{\lambda\cdot x\cdot(w-w_m)+w-\lambda\cdot z\cdot w_u}{1-\lambda\cdot z}\middle| z\right) f_X(z)dz$$
$$+1_{\lambda\cdot x\cdot(w_u-w)+w>w_m}\cdot 1_{w>w_m}\int F_{W|X}\left(\frac{\lambda\cdot x\cdot(w_u-w)+w-\lambda\cdot z\cdot w_u}{1-\lambda\cdot z}\middle| z\right) f_X(z)dz$$
$$= 1_{w\le w_m(1+\lambda x w_l/w_m)/(1+\lambda x)}\cdot\int F_{W|X}\left(\frac{\lambda\cdot x\cdot(w-w_m)+w+\lambda\cdot z\cdot w_l}{1+\lambda\cdot z}\middle| z\right) f_X(z)dz$$
$$+0\cdot\int F_{W|X}\left(\frac{\lambda\cdot x\cdot(w_u-w)+w+\lambda\cdot z\cdot w_l}{1+\lambda\cdot z}\middle| z\right) f_X(z)dz$$
$$+1_{w_m(1+\lambda x w_l/w_m)/(1+\lambda x)\le w\le w_m}\cdot\int F_{W|X}\left(\frac{\lambda\cdot x\cdot(w-w_m)+w-\lambda\cdot z\cdot w_u}{1-\lambda\cdot z}\middle| z\right) f_X(z)dz$$
$$+1_{w\ge w_m}\cdot\int F_{W|X}\left(\frac{\lambda\cdot x\cdot(w_u-w)+w-\lambda\cdot z\cdot w_u}{1-\lambda\cdot z}\middle| z\right) f_X(z)dz.$$

The last equality uses the following four facts: $(i)$, $\lambda\cdot x\cdot(w-w_l)+w\le w_m$ implies $w\le w_m(1+\lambda x w_l/w_m)/(1+\lambda x)\le w_m$, $(ii)$ $\lambda\cdot x\cdot(w_u-w)+w\le w_m$ implies $w\le w_m(1-\lambda x w_u/w_m)/(1-\lambda x)<w_m$, $(iii)$, $\lambda\cdot x\cdot(w-w_l)+w>w_m$ implies $w\ge w_m(1+\lambda x w_l/w_m)/(1+\lambda x)$, and $(iv)$ $\lambda\cdot x\cdot(w_u-w)+w>w_m$ implies $w\ge w_m(1-\lambda x w_u/w_m)/(1-\lambda x)$. Now we will look at

$$\mathbb{E}\left[\frac{\partial g}{\partial w}(W_i,X_i)\frac{1}{f_W(W_i)}k(W_i,X_i,\lambda)\right]$$
$$=\int_{x_l}^{x_u}\int_{w_l}^{w_u}\frac{\partial g}{\partial w}(w,x)\frac{1}{f_W(w)}k(w,x,\lambda)f_{W,X}(w,x)dwdx.$$

Substituting the three terms of $k(w,x,\lambda)$ in here we get

$$\mathbb{E}\left[\frac{\partial g}{\partial w}(W_i,X_i)\frac{1}{f_W(W_i)}k(W_i,X_i,\lambda)\right]$$

$$=\int_{x_l}^{x_u}\int_{w_l}^{w_m\frac{1+\lambda x w_l/w_m}{1+\lambda x}}\frac{\frac{\partial g}{\partial w}(w,x)}{f_W(w)}\int F_{W|X}\left(\frac{\lambda\cdot x\cdot(w-w_l)+w+\lambda\cdot z\cdot w_l}{1+\lambda\cdot z}\middle| z\right) f_X(z)dz f_{W,X}(w,x)dwdx \tag{C.5}$$

$$+\int_{x_l}^{x_u}\int_{w_m\frac{(1+\lambda x w_l/w_m}{1+\lambda x}}^{w_m}\frac{\frac{\partial g}{\partial w}(w,x)}{f_W(w)}\int F_{W|X}\left(\frac{\lambda\cdot x\cdot(w-w_l)+w-\lambda\cdot z\cdot w_u}{1-\lambda\cdot z}\middle| z\right) f_X(z)dz f_{W,X}(w,x)dwdx \tag{C.6}$$

$$+\int_{x_l}^{x_u}\int_{w_m}^{w_u}\frac{\frac{\partial g}{\partial w}(w,x)}{f_W(w)}\int F_{W|X}\left(\frac{\lambda\cdot x\cdot(w_u-w)+w-\lambda\cdot z\cdot w_u}{1-\lambda\cdot z}\middle| z\right) f_X(z)dz f_{W,X}(w,x)dwdx \tag{C.7}$$

Next, we take the derivative with respect to $\lambda$ for each of these three terms, and evaluate that derivative at $\lambda=0$. For the first term, (C.5) this derivative consists of two terms, one corresponding to the derivative with respect to the $\lambda$ in the bounds of the integral, and one corresponding to the derivative with respect to $\lambda$ in the integrand. For the second term we only have the term corresponding to the derivative with respect to the $\lambda$ in the bounds of the integral since the other term vanishes when we evaluate it at $\lambda=0$. The third term, (C.6) only has $\lambda$ in the integrand. So,

$$\frac{\partial}{\partial\lambda}\mathbb{E}\left[\frac{\partial g}{\partial w}(W_i,X_i)\frac{1}{f_W(W_i)}k(W_i,X_i,\lambda)\right]\Bigg|_{\lambda=0}$$
$$=(w_l-w_m)\cdot\mathbb{E}\left[\frac{\partial}{\partial w}g(w_m,X_i)\middle| W_i=w_m\right]$$
$$+\int_{x_l}^{x_u}\int_{w_l}^{w_m}\frac{\partial}{\partial w}g(w,x)\frac{1}{f_W(w)}\int_{x_l}^{x_u}f_{W|X}(w|z)\left(x\cdot(w-w_l)+z\cdot w_l-z\cdot w\right)f_X(z)dz f_{W,X}(w,x)dwdx$$

$$-(w_l - w_m) \cdot \mathbb{E}\left[ \frac{\partial}{\partial w} g(w_m, X_i) \middle| W_i = w_m \right]$$

$$+ \int_{x_l}^{x_u} \int_{w_l}^{w_m} \frac{\partial}{\partial w} g(w,x) \frac{1}{f_W(w)} \int_{x_l}^{x_u} f_{W|X}(w|z) \left( x \cdot (w_u - w) + z \cdot w - z \cdot w_u \right) f_X(z) dz f_{W,X}(w,x) dw dx$$

$$= \int_{x_l}^{x_u} \int_{w_l}^{w_m} \frac{\partial}{\partial w} g(w,x) \int_{x_l}^{x_u} f_{X|W}(z|w) \left( x \cdot d(w,x) - z \cdot d(w,z) \right) dz f_{W,X}(w,x) dw dx$$

$$+ \int_{x_l}^{x_u} \int_{w_l}^{w_m} \frac{\partial}{\partial w} g(w,x) \int_{x_l}^{x_u} f_{X|W}(z|w) \left( x \cdot d(w,x) \right) - z \cdot d(w,z) ) dz f_{W,X}(w,x) dw dx$$

$$= \int_{x_l}^{x_u} \int_{w_l}^{w_u} \frac{\partial}{\partial w} g(w,x) \int_{x_l}^{x_u} f_{X|W}(z|w) \left( x \cdot d(w,x) - z \cdot d(w,z) \right) dz f_{W,X}(w,x) dw dx$$

$$= \int_{x_l}^{x_u} \int_{w_l}^{w_u} \frac{\partial}{\partial w} g(w,x) \left( (X_i \cdot d(w,X_i) - \mathbb{E}\left[ X_i \cdot d(w,X_i) \right) \middle| W_i = w \right] \right) f_{W,X}(w,x) dw dx$$

$$= \mathbb{E}\left[ \frac{\partial g}{\partial w} (W_i, X_i) \cdot \left( X_i \cdot d(W_i, X_i) - \mathbb{E}\left[ X_i \cdot d(W_i, X_i) \middle| W_i \right] \right) \right] = \beta^{\mathrm{lc,v}}.$$

This finishes the proof of (C.3).
Finally, we show (C.4), by showing the equality of

$$\beta^{\mathrm{lc,v}} = \mathbb{E}\left[ \frac{\partial g}{\partial w} (W_i, X_i) \cdot \left( X_i \cdot d(W_i) - \mathbb{E}\left[ X_i \cdot d(W_i) \middle| W_i \right] \right) \right], \tag{C.8}$$

and

$$\mathbb{E}\left[ \delta(W_i, X_i) \cdot \frac{\partial^2 g}{\partial w \partial x} (W_i, X_i) \right]. \tag{C.9}$$

Define

$$b(w) = \mathbb{E}\left[ \frac{\partial g}{\partial w} (w, X_i) \cdot \left( X_i \cdot d(w) - \mathbb{E}\left[ X_i \cdot d(w) \middle| W_i = w \right] \right) \middle| W_i = w \right]$$

$$= \mathbb{E}\left[ \frac{\partial g}{\partial w} (w, X_i) \cdot d(w) \cdot \left( X_i - \mathbb{E}\left[ X_i \middle| W_i = w \right] \right) \middle| W_i = w \right],$$

so that $\beta^{\mathrm{lc,v}} = \mathbb{E}[b(W)]$. Apply Lemma A.2, with $h(x) = \frac{\partial g}{\partial w}(w,x) \cdot d(w)$, to get

$$b(w) = \mathbb{E}\left[ \frac{\partial^2}{\partial w \partial x} g(w, X) \cdot \delta(w, X) \right]$$

with

$$\delta(w,x) = d(w) \cdot \frac{F_{X|W}(x|w) \cdot (1 - F_{X|W}(x|w))}{f_{X|W}(x|w)} \cdot \left( \mathbb{E}\left[ X | X > x, W = w \right] - \mathbb{E}\left[ X | X \le x, W = w \right] \right).$$

Thus

$$\beta^{\mathrm{lc,v}} = \mathbb{E}[b(W)] = \mathbb{E}\left[ \frac{\partial^2}{\partial w \partial x} g(W, X) \cdot \delta(W, X) \right].$$

□

**Proof of Theorem 4.1:** We apply Lemmas A.11-A.14. The assumptions in the theorem imply that the conditions for those lemmas are satisfied. □

**Proof of Theorem 4.2:** The proof is essentially the same as that for Theorem 4.1 and is omitted. □

**Proof of Theorem 4.3:** We apply Lemma's A.23-A.27 to get an asymptotic linear representation for $\hat{\beta}^{\mathrm{cm}}(\rho, \tau)$. The assumptions in the Theorem imply that the conditions for the applications of these five lemmas are satisfied. Therefore, by Lemma A.23, we have

$$\hat{\beta}^{\mathrm{cm}}(\rho, 0) = \beta^{\mathrm{cm}}(\rho, 0) + \left( \hat{\beta}^{\mathrm{cm}}_g - \overline{g}^{\mathrm{cm}} \right) + \left( \hat{\beta}^{\mathrm{cm}}_W - \overline{g}^{\mathrm{cm}} \right) + \left( \hat{\beta}^{\mathrm{cm}}_X - \overline{g}^{\mathrm{cm}} \right) + \left( \overline{g}^{\mathrm{cm}} - \beta^{\mathrm{cm}}(\rho, 0) \right) + o_p\left( N^{-1/2} \right).$$

By Lemmas A.24-A.27, this is equal to

$$\beta^{\mathrm{cm}}(\rho, 0) + \frac{1}{N} \sum_{i=1}^{N} \left( \psi_g(Y_i, W_i, X_i) + \psi_W(Y_i, W_i, X_i) + \psi_X(Y_i, W_i, X_i) + \psi_0(Y_i, W_i, X_i) \right) + o_p\left( N^{-1/2} \right)$$

$$= \beta^{\text{cm}}(\rho,0) + \frac{1}{N}\sum_{i=1}^{N}\psi(Y_i,W_i,X_i) + o_p\left(N^{-1/2}\right),$$

with $\psi_g(y,w,x)$ given in (A.8), $\psi_W(y,w,x)$ given in (A.9), $\psi_X(y,w,x)$ given in (A.10), $\psi_0(y,w,x)$ given in (A.11), and $\psi(y,w,x)$ given in (4.18). Then we have an asymptotic linear representation for $\hat{\beta}^{\text{cm}}(\rho,\tau)$:

$$\begin{aligned}
\hat{\beta}^{\text{cm}}(\rho,\tau) &= \tau\cdot\overline{Y} + (1-\tau)\cdot\hat{\beta}^{\text{cm}}(\rho,0)\\
&= \beta^{\text{cm}}(\rho,\tau) + \tau\cdot(\overline{Y} - \beta^{\text{cm}}(\rho,1)) + (1-\tau)\cdot\left(\hat{\beta}^{\text{cm}}(\rho,0) - \beta^{\text{cm}}(\rho,0)\right)\\
&= \beta^{\text{cm}}(\rho,\tau) + \tau\cdot(\overline{Y} - \beta^{\text{cm}}(\rho,1)) + (1-\tau)\cdot\frac{1}{N}\sum_{i=1}^{N}\psi(Y_i,W_i,X_i).
\end{aligned}$$

Since by a law of large numbers $\overline{Y}\to\beta^{\text{cm}}(\rho,1)$, and $\sum_i \psi(Y_i,W_i,X_i)/N \to \mathbb{E}[\psi(Y_i,W_i,X_i)] = 0$, it follows that $\hat{\beta}^{\text{cm}}(\rho,\tau)\to\beta^{\text{cm}}(\rho,\tau)$.
By a central limit theorem the second part of the Theorem follows. □
**Proof of Theorem 4.4:** By Lemma A.15 we can write

$$\begin{aligned}
&\sqrt{N}\left(\hat{\beta}^{\text{lc}} - \beta^{\text{lc}}\right)\\
&= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{\partial g}{\partial w}(W_i,X_i)\cdot d(W_i)\cdot(X-m(W_i)) - \sqrt{N}\cdot\beta^{\text{lc}}\\
&\quad + \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{\partial \hat{g}}{\partial w}(W_i,X_i)\cdot d(W_i)\cdot(X-m(W_i)) - \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{\partial g}{\partial w}(W_i,X_i)\cdot d(W_i)\cdot(X-m(W_i))\\
&\quad + \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{\partial g}{\partial w}(W_i,X_i)\cdot d(W_i)\cdot(X-\hat{m}(W_i)) - \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{\partial g}{\partial w}(W_i,X_i)\cdot d(W_i)\cdot(X-m(W_i)).
\end{aligned}$$

By Lemma A.16

$$\begin{aligned}
&\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{\partial}{\partial w}\widehat{g}\left(W_i,X_i\right)\cdot d(W_i)\cdot\left(X_i - m\left(W_i\right)\right) - \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{\partial}{\partial w}g\left(W_i,X_i\right)\cdot d(W_i)\cdot\left(X_i - m\left(W_i\right)\right)\\
&-\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\delta_g(Y_i,W_i,X_i) = o_p(1),
\end{aligned}$$

where

$$\begin{aligned}
\delta_g\left(Y,W,X\right) = &-\frac{1}{f_{W,X}\left(W,X\right)}\frac{\partial f_{W,X}\left(W,X\right)}{\partial W}\left(Y - g\left(W,X\right)\right)d(W)\left(X - m\left(W\right)\right)\\
&-\frac{\partial m\left(W\right)}{\partial W}d(W)\left(Y - g\left(W,X\right)\right)\\
&+\frac{\partial}{\partial w}d(W)\left(X - m(W)\right)\left(Y - g(W,X)\right).
\end{aligned}$$

By Lemma A.19

$$\begin{aligned}
&\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{\partial}{\partial w}g\left(W_i,X_i\right)\cdot d(W_i)\cdot\widehat{m}\left(W_i\right) - \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{\partial}{\partial w}g\left(W_i,X_i\right)\cdot d(W_i)\cdot m\left(W_i\right)\\
&\qquad -\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\delta_m(Y_i,W_i,X_i) = o_p(1),
\end{aligned}$$

where

$$\delta_m(y,w,x) = \mathbb{E}\left[\left.\frac{\partial g(w,X_i)}{\partial W}\right| W_i = w\right]\cdot d(w)\cdot(x - m(w)).$$

Combining these results implies that

$$\sqrt{N}\left(\hat{\beta}^{\text{lc}} - \beta^{\text{lc}}\right) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\psi(Y_i,W_i,X_i),$$

with

$$\psi(y, w, x) = \frac{\partial g(w, x)}{\partial w} \cdot d(w) \cdot (x - m(w)) - \beta^{\mathrm{lc}} + \delta_g(y, w, x) + \delta_m(y, w, x).$$

Using a law of large numbers then implies the first result in the theorem, and using a central limit theorem implies the second result in the Theorem. □