

Semiparametric Learning of Integral Functionals on Submanifolds

Xiaohong Chen^{*} and Wayne Yuan Gao[†]

July 16, 2025

Abstract

This paper studies the semiparametric estimation and inference of integral functionals on submanifolds, which arise naturally in a variety of econometric settings. For linear integral functionals on a regular submanifold, we show that the semiparametric plug-in estimator attains the minimax-optimal convergence rate $n^{-\frac{s}{2s+d-m}}$, where s is the Hölder smoothness order of the underlying nonparametric function, d is the dimension of the first-stage nonparametric estimation, m is the dimension of the submanifold over which the integral is taken. This rate coincides with the standard minimax-optimal rate for a $(d - m)$ -dimensional nonparametric estimation problem, illustrating that integration over the m -dimensional manifold effectively reduces the problem's dimensionality. We then provide a general asymptotic normality theorem for linear/nonlinear submanifold integrals, along with a consistent variance estimator. We provide simulation evidence in support of our theoretical results.

Keywords: semi/non-parametric estimation and inference, integral functional, manifold, contour set, level set

^{*}Chen: Department of Economics and Colwes Foundation for Economic Research, Yale University, 28 Hillhouse Ave, New Haven, CT 06511, USA, xiaohong.chen@yale.edu

[†]Gao: Department of Economics, University of Pennsylvania, 133 S 36th St., Philadelphia, PA 19104, USA, waynegao@upenn.edu.

1 Introduction

The central problem we analyze in this paper is the semiparametric estimation of the following linear integral functional whose domain of integration \mathcal{M} is a submanifold of dimension $m < d$ in \mathbb{R}^d :

$$\Gamma(h_0) := \int_{\mathcal{M}} h_0(x) w_0(x) d\mathcal{H}^m(x), \quad (1)$$

where h_0 is an unknown but nonparametrically estimable function that maps from $\mathcal{X} \subseteq \mathbb{R}^d$ to \mathbb{R} , w_0 is a known weight that maps from \mathcal{X} to \mathbb{R} function, and \mathcal{H}^m denotes the m -dimensional Hausdorff measure on \mathbb{R}^d .¹ Based on our analysis on the linear functional Γ in (1), we also provide results on the estimation and inference on nonlinear integral functionals whose first-order expansion becomes a linear submanifold integral of the form (1).

Submanifold integrals of the form (1) naturally emerge in econometric settings where we take derivatives of a standard integral (such as an expectation) with respect to a changing domain of integration. For example, consider the following treatment assignment problem

$$\max_{\beta \in \mathbb{R}^d: \|\beta\|=1} \int \mathbb{1}\{x'\beta \geq 0\} \text{CATE}(x) p_0(x) dx, \quad (2)$$

where $\text{CATE}(x) := \mathbb{E}[Y_i(1) - Y_i(0) | X_i = x]$ denotes the Conditional Average Treatment Effect function, and p_0 is the marginal density of subjects with characteristics x . Then the first-order condition for the optimal linear treatment assignment parameter β is characterized by the derivative of the integral (2) with respect to β , which affects the integration domain. Mathematically, the derivative of an integral with respect to its domain translates into an integral over the *boundary* of its domain, and the boundary is often a lower-dimensional submanifold of the original domain. For (2), the first-order condition for β is given by

$$\int_{\{x \in \mathbb{R}^d: x'\beta=0\}} \text{CATE}(x) x p_0(x) d\mathcal{H}^{d-1}(x) = \mathbf{0}, \quad (3)$$

which is a submanifold integral of the form (1), with $\mathcal{M} = \{x \in \mathbb{R}^d: x'\beta = 0\}$ being a $(d-1)$ -dimensional hyperplane that defines the boundary of the integral in (3). Here, we

¹The m -dimensional Hausdorff measure \mathcal{H}^m in \mathbb{R}^d is defined as follows. For a set $A \subseteq \mathbb{R}^d$, define $\mathcal{H}^m(A) := \lim_{\delta \rightarrow 0} \mathcal{H}_\delta^m$, where, for any $\delta \in (0, \infty)$,

$$\mathcal{H}_\delta^m(A) := \inf \left\{ \sum_{j=1}^{\infty} \alpha(m) \left(\frac{\text{diam}(C_j)}{2} \right)^m : A \subseteq \cup_{j=1}^{\infty} C_j, \text{diam}(C_j) \leq \delta \right\},$$

with $\alpha(m) = \frac{\pi^{m/2}}{\Gamma(m/2+1)} = \frac{\pi^{m/2}}{\int_0^\infty e^{-x} x^{m/2} dx}$. The m -dimensional Hausdorff measure coincides with the familiar m -dimensional Lebesgue measure in the lower dimensional \mathbb{R}^m , and can be thought as a generalization of a variety of “uniform” measures for lower-dimensional subsets such as line, area, surface, and volume measures. See, for example, [Evans and Gariepy \(2015\)](#) for more details of the Hausdorff measure.

may take $w_0(x) = xp_0(x)$ and $h_0(x) = \text{CATE}(x)$, with $\text{CATE}(x)$ being the unknown function to be nonparametrically estimated.

Differentiation of an integral with respect to its domain also shows up in asymptotic analysis. For an example related to the one above, consider the estimation of the following average welfare, average cost, or average subject characteristics under first-best treatment assignment

$$W(h_0) := \int \mathbb{1}\{h_0(x) \geq 0\} w_0(x) dx = \int_{\{x \in \mathbb{R}^d: h_0(x) \geq 0\}} w_0(x) dx, \quad (4)$$

with $h_0(x) := \text{CATE}(x)$, so that the integral above is taken over the observable types who would weakly benefit from the treatment on average. Then, for example, setting $w_0(x) = xp_0(x)$ would endow $W(h_0)$ with the interpretation as average characteristics under the first-best treatment. If h_0 is nonparametrically estimated, the impact of the estimation error on the plug-in estimation of $W(h_0)$ can be analyzed using the functional (pathwise) derivative of $W(h_0)$ with respect to the CATE function, which again appears in the domain of integration. As a result, the pathwise derivative of $W(h_0)$ in the direction of $h - h_0$ at h_0 again becomes a submanifold integral of the form (1):

$$\int_{\{x \in \mathbb{R}^d: h_0(x)=0\}} \frac{h(x) - h_0(x)}{\|\nabla_x h_0(x)\|} w_0(x) d\mathcal{H}^{d-1}(x),$$

where the level set $\mathcal{M}_0 = \{x \in \mathbb{R}^d: h_0(x) = 0\}$ is again a $(d-1)$ -dimensional submanifold in \mathbb{R}^d (under mild regularity conditions).

Submanifold integrals also show up in a variety of other econometric settings, such as the estimation of marginal treatment effects, the estimation of (generalized) partial means, the estimation of weighted average derivatives, and maximum score estimation for discrete choice models. We provide more details of these motivating examples in Section 2.

Our first set of theoretical results is about the minimax convergence rate for the estimation of $\Gamma(h_0)$. When h_0 is assumed to lie in a Hölder class of smoothness order s , then under some other mild regularity conditions, we show that the (attainable) minimax-optimal convergence rate is given by $n^{-\frac{s}{2s+d-m}}$, which is the usual (minimax-optimal) convergence rate associated with $(d-m)$ -dimensional nonparametric regressions. This rate shows that the integration over the m -dimensional submanifold effectively reduces the dimensionality of the nonparametric estimation problem, and generalizes well-known existing results in the literature on semiparametric estimation. Our main result generalizes classic rates for pointwise evaluation, partial means, and full-dimensional integrals by explicitly incorporating submanifold dimensionality into the convergence rate.

For example, the special case of $m = 0$ translates to the rate of $n^{-\frac{s}{2s+d}}$, which corresponds

to the standard rate for point evaluation functionals of h_0 , which is known to converge at the original d -dimensional nonparametric regression rate. On the other extreme case of $m = d$, our derived rate becomes the parametric rate of $n^{-1/2}$, which is again intuitive since in this case $\Gamma(h_0)$ becomes a standard full-dimensional (Lebesgue) integral functional of h_0 , which is known to enjoy the parametric convergence rate under mild conditions. For the intermediate case, our result also generalizes previous results on partial means (Newey, 1994), which delivers the same rate formula $n^{-\frac{s}{2s+d-m}}$ but focuses on partial integrals over fixed and known coordinates of dimension $d-m$. Our result shows that the minimax-optimal convergence rate for $\Gamma(h_0)$ is given by $n^{-\frac{s}{2s+d-m}}$ as long as the submanifold is of dimension m (provided some mild regularity conditions).

We also provide a general asymptotic normality theorem for sieve-based semiparametric estimators of (potentially nonlinear) integral functionals whose first-order expansion takes the form of (1), along with a consistent asymptotic variance estimator. We also provide lower-level conditions for asymptotic normality in two specific types of nonlinear integral functionals: integral of nonlinear transformations of h_0 on submanifolds, and integrals on upper contour sets of h_0 . For the latter case, it is worth noting that the calculation of the pathwise derivatives is a relatively nonstandard mathematical exercise that utilizes the calculus of moving manifolds in differential geometry.

We conduct Monte Carlo simulations on semiparametric estimation and inference of linear and nonlinear integral functionals on submanifolds, and provide numerical evidence for the good finite-sample performance of the plug-in estimator and the corresponding confidence intervals. In particular, the realized CI coverage rates are close to the nominal 95% level.

Our paper contributes to the literature on the theory of semi/non-parametric estimation and inference, especially on irregular integral functionals. To our best knowledge, our paper is the first in the literature to provide a general result on the plug-in estimation of submanifold integral functionals, in which we show how the dimension of the integral affects the final convergence rate in the derived rate formula $n^{-\frac{s}{2s+d-m}}$ under Hölder smoothness. Our framework and results generalize and complement previous results in Newey (1994) on the estimation of partial means with kernel first stage, and Qiao (2021), who considers kernel estimation of surface integrals on level sets. While we exploit the results in Chen, Liao and Sun (2014); Chen and Christensen (2015), which cover sieve-based semiparametric inference on irregular functionals, our paper provides more explicit rate results by characterizing the growth rate of the norm of the Riesz representer for submanifold integral functionals.

Our paper utilizes the mathematical tools in differential geometry and geometric measure theory to analyze derivatives and integrals on manifolds. In the econometric literature, pre-

vious work by [Kim and Pollard \(1990\)](#) shows that the first-order condition for the maximum score estimator/criterion becomes a submanifold integral with the submanifold given by a $(d - 1)$ -dimensional hyperplane; [Sasaki \(2015\)](#) also noted that differentiation with respect to the domain of integration produces lower-dimensional submanifold integrals when analyzing conditional quantile functions in nonseparable structural models; [Chernozhukov, Fernández-Val and Luo \(2018\)](#) also uses integrals on manifolds on level sets to study sorted partial effects in heterogeneous coefficient models; a very recent paper by [Feng and Hong \(2024\)](#) shows how Hausdorff integrals can be used in the analysis asymptotic properties of value functions of optimal allocation problems as well as two-step estimator of ROC (receiver operating characteristics) curves. While our paper uses many similar differential geometry and geometric measure theory tools as in these aforementioned papers to analyze submanifold integrals, the main objects of interest being analyzed are substantially different across these papers and ours. Notably, the semiparametric estimation problems considered in [Chernozhukov, Fernández-Val and Luo \(2018\)](#) and [Feng and Hong \(2024\)](#) belong mostly to the *regular* (\sqrt{n} -estimable) case, while the semiparametric estimation problem we consider in this paper is mostly *irregular* with slower-than- \sqrt{n} convergence rate. Correspondingly, our paper features a general and explicit rate formula that relates the dimensions of the submanifolds to the convergence rates of the final semiparametric plug-in estimators.

Our paper is also conceptually related to [Khan and Tamer \(2010\)](#) on “thin-set” identification, where the identifying information in an economic model is contained in a “thin set” that has Lebesgue measure (or probability) zero in the population. Our paper provides a complementary, unified, and more refined perspective. First, even though a “thin set”, i.e., a submanifold, may be of Lebesgue measure zero, it may still be of nontrivial internal dimensionality, and integrals wrt Hausdorff measure over the submanifold provides a non-trivial aggregation of information over the submanifold. Second, a point, a curve, a surface, or more generally a submanifold may all be “thin sets” of Lebesgue measure zero, but, as our paper shows, they can possess different dimensionalities and geometric structures that lead to different properties in the estimation and inference on parameters about such “thin sets”.

The rest of the paper is organized as follows. Section 2 provides some motivating examples for submanifold integrals in econometric settings. Section 3 establishes the minimax-optimal convergence rate for the estimation of linear submanifold integrals. Section 4 provides a general asymptotic normality theorem, along a consistency variance estimator, for inference on both linear and nonlinear submanifold integrals, can still be used for more complicated nonlinear integral functionals. Section 5 provides Monte Carlo simulation results.

2 Setup and Motivating Examples

Let $(Y_i, X_i)_{i=1}^n$ be a random sample of data with joint distribution $P_{(X,Y)}$, where Y_i is a scalar-valued outcome variable, and X_i is a vector of observed covariates with support $\mathcal{X} \subseteq \mathbb{R}^d$.

Let $h_0 : \mathcal{X} \rightarrow \mathbb{R}$ be a nonparametric function² that is directly identified from the data and can be estimated using standard nonparametric estimation methods. A leading example of h_0 is the conditional expectation (nonparametric regression) function $h_0(x) = \mathbb{E}[Y_i | X_i = x]$, which will be our focus. That said, h_0 may also take the form of density functions, conditional quantiles and structural regression functions in NPIV models.

We consider submanifolds that take the form of level sets of functions, which arise naturally in a variety of economic problems as we show in the examples below. Specifically, let $g : \mathcal{X} \rightarrow \mathbb{R}^{d-m}$ be a continuously differentiable³ function with $\mathbf{0} \in \text{int}(g(\mathcal{X})) \subseteq \mathbb{R}^{d-m}$ and write

$$\mathcal{M} := \{x \in \mathcal{X} : g(x) = \mathbf{0}\}$$

as the (zero) level set of g .⁴ We maintain the following standard regularity condition that ensures the \mathcal{M} is an m -dimensional submanifold of \mathbb{R}^d (e.g. by Theorem 12.1 of [Loomis and Sternberg, 2014](#)).

Assumption 1 (Regular Level Set). *$\nabla_x g(x)$ has full rank $d-m$ for every $x \in \mathcal{M}$. Depending on the exact problem setup, g may be known or unknown, parametric or nonparametric, and it may be taken to be different from or the same as h_0 , which will be illustrated in the examples below and treated in subsequent sections.*

A central object of interest in this paper is the integral of the nonparametric function h_0 over the submanifold \mathcal{M} :

$$\int_{\mathcal{M}} h_0(x) w_0(x) d\mathcal{H}^m(x), \quad (5)$$

where $w_0 : \mathcal{X} \rightarrow \mathbb{R}$ is a scalar-valued weight function, and $\mathcal{H}^m(x)$ denotes the m -dimensional Hausdorff measure on \mathbb{R}^d . Under Assumption 1, $\{g_0(x) = c\}$ is an m -dimensional submanifold that has Lebesgue measure 0 in \mathbb{R}^d . As a result, the above integral would have been trivially zero if it had been taken with respect to the Lebesgue measure (on \mathbb{R}^d) instead of the (m -dimensional) Hausdorff measure.

We now provide some motivating examples for the study of submanifold integrals in the form of (5) in a variety of economic settings.

²More generally, h_0 may be a vector of nonparametric functions.

³We will impose additional smoothness condition on g when it comes to estimation and inference (Assumption 3 in Section 3), which is not yet needed for the discussion of the examples in this section.

⁴Note that $\mathbf{0}$ may be replaced with any other constant vector $c \in \text{int}(g(\mathcal{X})) \subseteq \mathbb{R}^{d-m}$ without affecting the results in this section.

Example 1 (Maximum Score Estimation of Binary Choice Models). Consider any model that satisfies the following sign alignment restriction

$$h_0(x) \geq 0 \Leftrightarrow x' \beta_0 \geq 0 \quad (6)$$

where h_0 is a nonparametrically identified and estimable function of x and β_0 is a d -dimensional parameter normalized to lie on the unit sphere, i.e., $\beta_0 \in \mathbb{S}^{d-1} := \{\beta \in \mathbb{R}^d : \|\beta\| = 1\}$. For example, in the following binary choice model with a conditional median independence restriction as in [Manski \(1975\)](#),

$$y_i = \mathbb{1} \left\{ X_i' \beta_0 + \epsilon_i \geq 0 \right\}, \quad \text{med}(\epsilon_i | X_i) = 0,$$

the sign alignment restriction (6) is satisfied with

$$h_0(x) := \mathbb{E}[Y_i | X_i = x] - \frac{1}{2}.$$

The population criterion function for maximum score estimator can then be written as

$$W(\beta) := \int h_0(x) \mathbb{1} \left\{ x' \beta \geq 0 \right\} p_0(x) dx, \quad (7)$$

where $p_0(x)$ denotes the density of $X_i = x$. Under appropriate conditions, β_0 can be point identified under scalar normalization,

$$\beta_0 = \arg \max_{\beta \in \mathbb{S}^{d-1}} W(\beta)$$

Even though the indicator $\mathbb{1} \left\{ x' \beta \geq 0 \right\}$ is discontinuous in β , the welfare function $W(\beta)$ remains differentiable in β , and the first-order condition for the optimality of β_0 is given by

$$\mathbf{0} = \nabla_{\beta} W(\beta_0),$$

As shown in [Kim and Pollard \(1990\)](#), the gradient $\nabla_{\beta} W(\beta)$ is “nonstandard” in the sense that it is a derivative of an integral with respect to a parameter that defines the region, or boundary, of the integration. Similarly to the fundamental Theorem of calculus (or the Leibniz rule) for elementary calculus,

$$\frac{d}{dt} \int_0^t f(x) dx = f(t),$$

derivatives of integrals with respect to the integration boundary typically become an “evaluation of the boundary”. Here, the boundary of the region of the integral is given by a $m = (d - 1)$ -dimensional hyperplane, or a “surface”, $\{x \in \mathcal{X} : x' \beta_0 = 0\}$, and the gradient $\nabla_{\beta} W(\beta)$ become a lower-dimensional integral over the boundary hyperplane:

$$\mathbf{0} = \nabla_{\beta} W(\beta_0) := \int_{\mathcal{M}_0} h_0(x) x p_0(x) d\mathcal{H}^{d-1}(x), \quad (8)$$

where the hyperplane $\mathcal{M}_0 = \{x \in \mathcal{X} : x' \beta_0 = 0\}$ has Lebesgue measure 0 in \mathbb{R}^d . Hence the

identification of β_0 is referred to as a type of “thin-set identification” in [Khan and Tamer \(2010\)](#). Consequently, the right-hand side of (8) cannot be represented by a regular Lebesgue integral, but instead by a Hausdorff integral over a $m = (d - 1)$ -dimensional manifold (a hyperplane here) in \mathbb{R}^d .

Example 2 (Optimal Treatment Assignment). It has been recognized, say in [Kitagawa and Tetenov \(2018\)](#), that optimal treatment assignment problem shares some similarity with maximum score estimation. Specifically, consider the problem of optimizing over a parametric family of treatment assignment rules that assigns the treatment status 0/1 according to $\mathbb{1}\{x'\beta \geq 0\}$ for a given observed type x , where β is a d -dimensional choice parameter. Then the welfare function of the assignment rule parameter β is given by

$$W(\beta) := \int \mathbb{1}\{x'\beta \geq 0\} h_0(x) p_0(x) dx, \quad (9)$$

where $h_0(x) := \mathbb{E}[Y_i(1) - Y_i(0) | X_i = x]$ is the conditional average treatment effect (CATE) for type x . Hence (9) is of exactly the same form as the maximum score criterion function (7). As a result, the FOC of (9) for the optimal β_0 (under scale normalization) is again given by the submanifold integral (8). Often times, researchers conduct (costly) experiments on and estimate CATE from a sample of moderate sample size, but the target population on which the treatment in question might be implemented can be of a much larger scale. In such settings p_0 may be known or estimable using a much larger sample size than that used to estimate h_0 , and thus we may focus on the estimation error for h_0 in the optimization of $W(\beta)$.

Example 3 (Aggregate Parameter over Selected Subpopulation). More generally, consider the estimation of the following parameter

$$W(h_0, g_0) := \int_{\{g_0(x) \geq 0\}} h_0(x) p_0(x) dx,$$

where h_0 and g_0 may be both unknown but nonparametrically estimable. For example, if we set $g_0(x)$ to be the CATE for type x and $h_0(x) = x$, then $W(h_0, g_0)$ becomes the average characteristics of the subpopulation with nonnegative CATE. If we again set $g_0(x) = h_0(x) = \text{CATE}(x)$, then $W(h_0, g_0)$ becomes welfare under “first-best” treatment assignment. Alternatively, we may take $h_0(x)$ to any other value/cost function associated type x .

When h_0 and g_0 are real-valued functions and are nonparametrically estimated by \hat{h} and \hat{g} , the pathwise derivative of $W(h_0, g_0)$ with respect to h_0 and g_0 is a key object of interest in the characterization of the asymptotic behaviors of the plug-in estimator $W(\hat{h}, \hat{g})$. Generally, the pathwise derivative of W in the direction of $(h - h_0, g - g_0)$ is given by

$$D_{(h,g)} W(h_0, g_0) [h - h_0, g - g_0]$$

$$\begin{aligned}
&:= \lim_{t \searrow 0} \frac{1}{t} (W(h_0 + t(h - h_0), g_0 + t(g - g_0)) - W(h_0, g_0)) \\
&= \int_{\{g_0(x) \geq 0\}} (h(x) - h_0(x)) p_0(x) dx + \int_{\{g_0(x) = 0\}} (g(x) - g_0(x)) \frac{h_0(x) p_0(x)}{\|\nabla_x g_0(x)\|} d\mathcal{H}^{d-1}(x)
\end{aligned} \tag{10}$$

which consists of two terms: the first is the perturbation of the integrand h_0 in the direction of $h - h_0$ over the true region of integral $\{g_0(x) \geq 0\}$, while the second is the perturbation of (the boundary) region of integration induced by the perturbation of g_0 in the direction of $g - g_0$. While the former is standard in the semiparametric estimation literature, the latter takes the nonstandard form of a Hausdorff integral over the $m = (d - 1)$ -dimensional manifold $\{g_0(x) = 0\}$.⁵ Hence, even if h_0 is known (or parametrically specified and can be estimated at \sqrt{n} -rate), the last term in (10) will still be present and remain as the leading term in the asymptotic behavior of $W(\hat{h}, \hat{g})$, as long as g_0 is nonparametrically specified and needs to be nonparametrically estimated.

Example 4 (Marginal Treatment Effects and Policy Relevant Treatment Effects). [Heckman and Vytlacil \(2005, 2007\)](#) proposes the marginal treatment effect (MTE) as a unifying concept that underlies a wide variety of treatment effects studied in causal inference and program evaluations. It is well-known that MTE can be written as a derivative with respect to propensity scores, or formally,

$$\begin{aligned}
\text{MTE}(x, p) &:= \frac{\partial}{\partial p} \mathbb{E}[Y_i(1) - Y_i(0) | X_i = x, p_0(Z_i) = p], \\
&= \frac{\partial}{\partial p} \int_{\{p_0(z) = p\}} \text{CATE}(x, z) f_0(z) d\mathcal{H}^{d-1}(z)
\end{aligned} \tag{11}$$

where $p_0(z) := \mathbb{E}[D_i = 1 | Z_i = z]$ is the propensity score, $\text{CATE}(x, z)$ is the conditional average treatment effect given both the covariates X_i and the instruments Z_i , and $f_0(z)$ denotes the marginal density of Z_i . It is then clear from (11) that MTE takes the form of the derivative of a submanifold integral, where the submanifold is given by the propensity score level set $p_0(z) = p$. Here, both the CATE function and the propensity score function can be nonparametrically estimated.

As argued in [Carneiro, Heckman and Vytlacil \(2010\)](#), in many scenarios it is useful to consider incremental policy reforms and focus on the analysis of marginal policy changes, whose effects are usually concentrated on individuals “at the margin”. The average marginal treatment effect (AMTE) summarizes the mean benefit of the treatment for the subpopulation who is indifferent between participation in the treatment and nonparticipation. Formally,

⁵In particular, $\|\nabla_x g_0(x)\|$, which measures the “thinness” of the level set, enters into the derivative formula explicitly here. In the previous examples with hyperplane boundaries, we took $g(x) = x'\beta$ with $\beta \in \mathbb{S}^{d-1}$, and thus $\|\nabla_x g(x)\| = \|\beta\| = 1$, which is why this term becomes implicit in formula (8).

this “at-the-margin” population $\{(z, u) \in \mathbb{R}^{d_z+1} : p_0(z) - u = 0\}$ is a submanifold defined by the level set of the propensity score function p_0 , and the average over this subpopulation can again be represented by a submanifold integral in the form of

$$\text{AMTE}(x) := \int_{\{p_0(z)=u\}} \text{MTE}(x, u) f_0(z|x) d\mathcal{H}_{(z,u)}^{d_z}.$$

In addition, MTE is also used to define a wide range of causal parameters, notably the policy-relevant treatment effect (PRTE) proposed in [Heckman and Vytlacil \(2001\)](#). Let F_P denote the CDF of $P_i := p_0(Z_i)$ and suppose that a proposed policy changes this CDF from F_P to G . Then the PRTE is defined by

$$\text{PRTE}(x, G) := \int_0^1 \text{MTE}(x, u) \omega_{\text{PRTE}}(u, G) du, \quad \omega_{\text{PRTE}}(u, G) := \frac{F_P(u) - G(u)}{\mathbb{E}_G(P_i) - \mathbb{E}_{F_P}(P_i)}.$$

Since MTE is defined as a derivative, PRTE takes the form of an average derivative parameter, for which \sqrt{n} estimation is possible under a certain “vanishing-on-boundary” condition [Powell, Stock and Stoker \(1989, PSS thereafter\)](#). [Carneiro, Heckman and Vytlacil \(2010\)](#) pointed out that this type of “vanishing-on-boundary” condition may be violated in plausible scenarios in the analysis of PRTE, which makes it not \sqrt{n} -estimable in general. Our result can be useful in scenarios where “vanishing-on-boundary” conditions are not imposed and \sqrt{n} -consistency cannot be guaranteed. See the example below for a related discussion.

Example 5 (Weighted Average Derivatives). Related to [Example 4](#) above, suppose that we are interested in the following weighted average derivative of a nonparametric function:

$$\text{WAD}(h_0) := \int \nabla h_0(x) w(x) dx$$

For example, PSS focuses on the density-weighted average derivative, which corresponds a special case of the above with $w(x) := p^2(x)$, with $p(x)$ denoting the marginal density of x on its support \mathcal{X} . The standard semiparametric asymptotic analysis of $\text{WAD}(\hat{h})$ above usually exploits the following integration-by-parts formula (a special case of the Divergence Theorem)

$$\int_{\mathcal{X}} \nabla h_0(x) w(x) dx = \int_{\partial\mathcal{X}} \vec{n}(x) \cdot h_0(x) w(x) d\mathcal{H}^{d-1}(x) - \int_{\mathcal{X}} h_0(x) \nabla w(x) dx, \quad (12)$$

where the first term is a Hausdorff integral over the submanifold defined by the support boundary $\partial\mathcal{X}$. The standard semiparametric analysis of average derivative estimation, such as in PSS and [Newey and Stoker \(1993\)](#), exploits a “vanishing-on-boundary” assumption, say, $h_0(x)w(x) = 0$ for all $x \in \partial\mathcal{X}$, so that the first term degenerates to 0, and thus it suffices to only analyze the second Lebesgue integral term $-\int_{\mathcal{X}} h_0(x) \nabla w(x) dx$. For example, PSS assumes that $p(x) = 0$ on $\partial\mathcal{X}$ (Assumption 2), and consequently the density-weighted average derivative becomes equal to $-2 \int h_0(x) \nabla p(x) p(x) dx$. PSS then proceeds

to establish the \sqrt{n} asymptotic normality of the plug-in estimator $\text{WAD}(\hat{h})$. As clear from (12), the \sqrt{n} convergence rate relies crucially on the “vanishing on boundary” assumption, without which the leading asymptotic term would be the first submanifold integral term that converges at slower-than- \sqrt{n} rate as established in our paper.

Example 6 (Generalized Partial Means). Suppose that we are interested in the conditional mean of $\phi(Y_i, X_i)$ given the event $g(X_i) = c \in \mathbb{R}^{d-m}$, where ϕ is some known transformation of Y_i and X_i :

$$\mathbb{E}[\phi(Y_i, X_i) | g(X_i) = c] = \int_{\{g(x)=c\}} h_0(x) w_0(x) d\mathcal{H}^m(x)$$

with $h_0(x) := \mathbb{E}[\phi(Y_i, X_i) | X_i = x]$, $w_0(x) := p_0(x | g(x) = c) = \frac{p_0(x)}{p_{g(X)}(c)}$, $p_0(x)$ being the density of X_i , and $p_{g(X)}$ as defined in Example 6. In particular, when g extracts a subvector of $x \in \mathbb{R}^d$, say, $g(x) = (x_1, \dots, x_{d-m})$ for some $1 \leq m < d$, the above specializes to the partial mean functional of the form $\mathbb{E}[Y_i | X_{i,1} = c_1, \dots, X_{i,d-m} = c_{d-m}]$ as studied in Newey (1994).

Example 7 (Structural Functions in NPIV Regression). The previous examples focus on *exogenous* nonparametric regression functions. Here we present a canonical example of a nonparametric function with endogeneity: the structural functions in the nonparametric instrumental variables (NPIV) model, i.e., the h_0 function below:

$$Y_i = h_0(X_i) + \epsilon_i, \quad \mathbb{E}[\epsilon_i | Z_i] = 0.$$

Previous work by, for example, Ai and Chen (2003, 2007, 2012), Chen and Pouzo (2015), Chen and Christensen (2018), and Chen, Christensen and Kankanala (2025), provides theoretical results on the estimation and inference on h_0 , as well as various (families of) linear and nonlinear functionals of h_0 , including point evaluation, average derivatives/elasticities, and consumer surplus/welfare functionals. In principle, the h_0 function in all the submanifold functionals defined in Examples 1-6 above could be replaced by the structural function in the NPIV model here, and the theory about submanifold integrals developed in the current paper would still be relevant.

Example 8 (Nonparametric Quantile, Density, and Copula). Similarly, the nonparametric function h_0 needs to be restricted as a conditional expectation function, but can be broadly defined as any estimable nonparametric function, for example, nonparametric quantile, density, and copula functions. These general nonparametric estimation problems (without endogeneity) have been studied widely in statistics, econometrics, and beyond in various settings. In the presence of endogeneity issues, Chen and Pouzo (2015), Chen, Pouzo and Powell (2019) and Chen, Liao and Wang (2024) have also provided theoretical results for the nonparametric quantile IV regression. The established theoretical results from these

previous studies can be combined with the theory of submanifold integrals developed here for the study of new functional parameters of interest.

We hope that the above examples illustrate our point that submanifold integrals, which have not been explicitly studied much in econometrics before, actually arise naturally in quite a variety of economic problems.

We should clarify that in the rest of this paper we will not provide detailed solutions to all the problems above. Instead, we will focus on the case where h_0 is the conditional expectation (nonparametric regression) function, and provide a general analysis of the estimation and inference of linear submanifold integrals of form (5) as well as nonlinear submanifold integrals whose first-order linear approximation takes the form of (5).

3 Estimation of Linear Submanifold Integrals

In this section, we focus on the semiparametric estimation of *linear* integrals whose regions of integration are submanifolds/ As we will show in Section 4, our analysis of *linear* integrals also forms the basis for the analysis of *nonlinear* integrals in the more general case.

Formally, we write $\theta_0 := \Gamma(h_0)$, with Γ being the linear integral functional

$$\Gamma(h) := \int_{\mathcal{M}} h(x) w(x) d\mathcal{H}^m(x), \quad (13)$$

Recall that we focus on the leading case of $h_0(x) := \mathbb{E}[Y_i | X_i = x]$. In this section, we treat the level set function g (consequently the manifold \mathcal{M}) and the weight function w both as fixed/known, and show in the next section how to apply the core result in this section to cases where the level set function g and the weight function w may also require estimation.

3.1 Lower Bound for Minimax Convergence Rate:

We first establish a lower bound for the minimax convergence rate for the estimation of $\Gamma(h_0)$ when h_0 is assumed to belong to the Hölder smoothness class. The established (rate) lower bound holds for any possible estimator of $\Gamma(h_0)$, thus providing a general property of the estimation of the submanifold integral functional Γ .

We first slightly strengthen the regular manifold assumption 1 on the level set function g . Let $\mathcal{J}g(x)$ denotes the Jacobian of $g : \mathbb{R}^d \rightarrow \mathbb{R}^{d-m}$ defined by

$$\mathcal{J}g(x) := \sqrt{\sum_{B(x)} \det(B(x))^2},$$

where B indexes all $(d-m) \times (d-m)$ minors of $\nabla g(x)$.

Assumption 2 (Jacobian Bounded Away from Zero). *There exists a constant $C > 0$ such that $\mathcal{J}g(x) \geq C$ for all $x \in \mathcal{M}$.*

We now introduce the Hölder smooth class for the unknown function h_0 . Formally, let $\mathcal{X} = \mathcal{X}_1 \times \dots \times \mathcal{X}_d$ be the Cartesian product of compact intervals $\mathcal{X}_1, \dots, \mathcal{X}_d$, say, $\mathcal{X} = [0, 1]^d$ for simplicity. A real-valued function h on \mathcal{X} is said to satisfy a Hölder condition with exponent $\gamma \in (0, 1]$ if there is a positive number c such that $|h(x) - h(y)| \leq c \|x - y\|^\gamma$ for all $x, y \in \mathcal{X}$; here $\|x - y\| = \left(\sum_{l=1}^d x_l^2\right)^{1/2}$ is the Euclidean norm of $x = (x_1, \dots, x_d) \in \mathcal{X}$. Given a d -tuple $\alpha = (\alpha_1, \dots, \alpha_d)$ of nonnegative integers, set $[\alpha] = \alpha_1 + \dots + \alpha_d$ and let D^α denote the differential operator defined by

$$D^\alpha = \frac{\partial^{[\alpha]}}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}}.$$

Let $\lfloor s \rfloor$ be a nonnegative integer that is smaller than s , and set $s = \lfloor s \rfloor + \gamma$ for some $\gamma \in (0, 1]$. A real-valued function h on \mathcal{X} is said to be s -smooth if it is $\lfloor s \rfloor$ times continuously differentiable on \mathcal{X} and $D^\alpha h$ satisfies a Hölder condition with exponent γ for all α with $[\alpha] = \lfloor s \rfloor$. Denote the class of all s -smooth real-valued functions on \mathcal{X} by $\Lambda^s(\mathcal{X})$ (called a Hölder class), and the space of all $\lfloor s \rfloor$ -times continuously differentiable real-valued functions on \mathcal{X} by $C^{\lfloor s \rfloor}(\mathcal{X})$. Define a Hölder ball with smoothness $s = \lfloor s \rfloor + \gamma$ as

$$\Lambda_c^s(\mathcal{X}) = \left\{ h \in C^{\lfloor s \rfloor}(\mathcal{X}) : \sup_{[\alpha] \leq \lfloor s \rfloor} \sup_{x \in \mathcal{X}} |D^\alpha h(x)| \leq c, \sup_{[\alpha] = \lfloor s \rfloor} \sup_{x, y \in \mathcal{X}, x \neq y} \frac{|D^\alpha h(x) - D^\alpha h(y)|}{\|x - y\|^\gamma} \leq c \right\}.$$

We will establish our minimax lower bound with h_0 constrained inside $\Lambda_c^s(\mathcal{X})$, but the smoothness of the level function g (and consequently the smoothness of the submanifold \mathcal{M}) also turns out to be relevant, given that the integral is taken over points on the submanifold. Thus we also require g be Hölder smooth as well.

Assumption 3 (Smoothness of Submanifolds). $g \in \Lambda_c^s(\mathcal{X})$.

We also impose the following regularity assumptions.

Assumption 4 (Density on \mathcal{X}). $\mathcal{X} = [0, 1]^d$, with the density of X_i on \mathcal{X} being uniformly bounded away from zero and infinity.

Assumption 5 (Nondegenerate Errors). $\inf_{x \in \mathcal{X}} \mathbb{E}[\epsilon_i^2 | X_i = x] > 0$, where $\epsilon_i := Y_i - h_0(X_i)$.

Theorem 1 (Minimax Convergence Rate: Lower Bound). *Under Assumptions 2-5, the rate $r_n = n^{-\frac{s}{2s+d-m}}$ is the minimax rate lower bound for the estimation of θ_0 , i.e.*

$$\liminf_{n \rightarrow \infty} \inf_{\tilde{\theta}} \sup_{P, w} \mathbb{E}_P \left[n^{\frac{2s}{2s+d-m}} \left(\tilde{\theta} - \theta_0(P, w) \right)^2 \right] \geq c, \quad \text{for some constant } c > 0.$$

where P is any joint probability distribution of (X_i, Y_i) that satisfies $h_0(\cdot) := \mathbb{E}_P[Y_i|X_i = \cdot] \in \Lambda_c^s(\mathcal{X})$ along with Assumptions 4 and 5, w is any uniformly bounded weight function, $\theta_0(P, w) := \int_{\mathcal{M}} h_0(x) w(x) d\mathcal{H}^m(x)$, and $\tilde{\theta}$ is any estimator of $\theta_0 := \theta_0(P, w)$.

Note that our rate formula $r_n = n^{-\frac{s}{2s+d-m}}$ reproduces several well-established results in the literature as special cases: When $m = d$, the lower bound rate becomes $r_n = n^{-\frac{1}{2}}$, reproducing the standard \sqrt{n} rate for regular (full-dimensional) integral functionals.⁶ When $m = 0$, the lower bound becomes $r_n = n^{-\frac{s}{2s+d}}$, reproducing the well-known (Stone, 1982) minimax optimal rate for point evaluation functionals of the d -dimensional nonparametric regression.⁷ When $m = d - 1$, the lower bound becomes $r_n = n^{-\frac{s}{2s+1}}$, which underlies the minimax optimal rate for (smoothed) maximum score estimation in Horowitz (1992).

3.2 Minimax Rate-Optimal Estimation via Sieve First Stage

Next, we show that the lower bound r_n established in Section 3.1 can be attained by the semiparametric plug-in estimator

$$\hat{\theta} := \Gamma(\hat{h}) = \int_{\mathcal{M}} \hat{h}(x) w(x) d\mathcal{H}^m(x), \quad (14)$$

where \hat{h} is a first-stage linear sieve nonparametric estimator. Hence we conclude that the rate r_n is minimax rate optimal, and plug-in estimator with sieve first stage attains this optimal rate.⁸

Following Chen and Christensen (2015), we consider first-stage linear sieve estimator \hat{h} of h_0 given by

$$\hat{h}(x) := \bar{b}^{K_n}(x) \left[\frac{1}{N} \sum_{i=1}^N \bar{b}^{K_n}(X_i) \bar{b}^{K_n}(X_i)' \right]^{-1} \frac{1}{N} \sum_{i=1}^N \bar{b}^{K_n}(X_i) Y_i,$$

where

$$\bar{b}^K(x) := \left(b_1^K(x), \dots, b_{K_n}^K(x) \right)' \equiv \text{vec} \left(\bigotimes_{\ell=1}^d \left(b_1^K(x_\ell), \dots, b_{J_n}^K(x_\ell) \right) \right)$$

is a K -dimensional vector of orthonormalized multivariate sieve basis functions constructed as tensor products of J univariate sieve functions $b_1^J(x_\ell), \dots, b_J^J(x_\ell)$ across each of the d dimensions of \mathcal{X} , with $K = J^d$. The sieve dimension is chosen to increase with n , with $J = J_n \nearrow \infty$ and thus $K = K_n \nearrow \infty$ as $n \rightarrow \infty$, though we will suppress the subscript n in J and K for notational simplicity.

⁶When $m = d$, the Hausdorff measure \mathcal{H}^d coincides with Lebesgue measure in \mathbb{R}^d .

⁷When $m = 0$, the Hausdorff measure \mathcal{H}^0 becomes a point counting measure.

⁸In Appendix D, we also show that this optimal rate is also attained when \hat{h} takes the form of the Nadaraya-Watson kernel regression estimator.

We first impose some mild conditions on the linear sieve \bar{b}^K that are satisfied by a variety of commonly used sieve classes such as splines and wavelets. Define $\zeta_K := \sup_{x \in \mathcal{X}} \|\bar{b}^K(x)\|$, $\lambda_K := \left(\mathbb{E} [\bar{b}^K(X_i) \bar{b}^K(X_i)'] \right)^{-1/2}$, and let \mathcal{B}_K be the closed linear span of the basis functions $\{b_1^K, \dots, b_K^K\}$ and define $P_{K,n}$ be the projection operator

$$P_{K,n}h := \bar{b}^K(x) \left[\frac{1}{N} \sum_{i=1}^N \bar{b}^K(X_i) \bar{b}^K(X_i)' \right]^{-1} \frac{1}{N} \sum_{i=1}^N \bar{b}^K(X_i) h(X_i), \quad \forall h \in L_2(X)$$

with the sup operator norm defined by

$$\|P_{K,n}\|_\infty := \sup_{h: 0 < \|h\|_\infty < \infty} \frac{\|P_{K,n}h\|_\infty}{\|h\|_\infty}.$$

Assumption 6 (Conditions on Linear Sieves). *The linear sieve \bar{b}^K satisfies (i) $\lambda_K = O(1)$, (ii) $\zeta_K = O(\sqrt{K})$, (iii) $\|P_{K,n}\|_\infty = O_p(1)$, and (iv) $\inf_{h \in B_K} \|h - h_0\|_\infty \leq K^{-s/d}$ for any $h_0 \in \Lambda_c^s(\mathcal{X})$.*

Define $\mathcal{V}_K := \mathcal{B}_K - \{h_0\}$. Let $\langle \cdot, \cdot \rangle_2$ denote the $L_2(X)$ inner product and $\|\cdot\|_2$ the $L_2(X)$ norm, and note that $(\mathcal{V}_K, \langle \cdot, \cdot \rangle_2)$ is a Hilbert space. By [Chen and Christensen \(2015\)](#), the linear functional Γ has a closed-form sieve Riesz representer on \mathcal{V}_K given by

$$v_{K_n}^*(\cdot) = \bar{b}^{K_n}(\cdot)' \mathbb{E} [\bar{b}^{K_n}(X_i) \bar{b}^{K_n}(X_i)']^{-1} \Gamma(\bar{b}^{K_n})$$

such that

$$\Gamma(\nu) = \langle v_{K_n}^*, \nu \rangle_2 \equiv \mathbb{E} [v_{K_n}^*(X_i) \nu(X_i)], \quad \forall \nu \in \mathcal{V}_{K_n},$$

Furthermore the rate at which $\|v_{K_n}^*\|_2$ grows is important in determining the variance of the plug-in estimator $\Gamma(\hat{h})$.

Lemma 1. *Under Assumptions 2-6(i),*

$$\|v_{K_n}^*\|_2^2 \asymp \|\Gamma(\bar{b}^{K_n})\|^2 \equiv \sum_{k=1}^{K_n} \Gamma^2(\bar{b}_k^{K_n}) \asymp K_n^{\frac{d-m}{d}} = J_n^{d-m}, \quad \text{as } K_n \rightarrow \infty. \quad (15)$$

Under sieve first stages, Lemma 1 is the core result that demonstrates the rate acceleration provided by the submanifold integral. It shows that the growth rate of the norm of the sieve Riesz representer is asymptotically proportional to $J^{(d-m)}$, where the exponent $(d-m)$ is the codimension of the m -dimensional manifold, but, importantly, not d , the dimension of the ambient space \mathbb{R}^d in which the manifold is embedded. In other words, even though the dimensionality of the first-stage nonparametric estimation of h_0 is d , it is reduced by the integration over the m -dimensional manifold.

Assumption 7. $h_0 \in \Lambda_c^s(\mathcal{X})$.

Assumption 8 (Bounded Variance of Errors). $\sup_{x \in \mathcal{X}} \sigma_\epsilon^2(x) < \infty$, where $\sigma_\epsilon^2(x) := \mathbb{E}[\epsilon_i^2 | X_i = x]$.

Theorem 2 (Convergence Rate under Sieve First Stage). *Under Assumptions 1-8,*

$$\hat{\theta} - \theta_0 \equiv \Gamma(\hat{h} - h_0) = O_p\left(\sqrt{K_n^{\frac{d-m}{d}}/n} + K_n^{-s/d}\right).$$

With optimally set $J_n \equiv K_n^{1/d} = n^{\frac{1}{2s+d-m}}$, we have

$$\hat{\theta} - \theta_0 = O_p\left(n^{-\frac{s}{2s+d-m}}\right),$$

which attains the lower bound rate r_n in Theorem 1 and is thus minimax rate optimal.

Theorems 1 and 2 together demonstrate that the problem of estimating linear integrals over an m -dimensional submanifold is akin to the $d-m$ dimensional nonparametric regression problem in terms of convergence rates. Intuitively, the integration over the m -dimensional submanifold effectively “aggregates out” those m dimensions, leaving only $d-m$ effective dimensions in the nonparametric estimation problem.

In many economic applications as discussed in Section 2, the level set function g is scalar-valued. Correspondingly the level set submanifold is of dimension $m = d-1$ and the co-dimension is $d-m = 1$, where all but one dimensions are “aggregated out” and the consequently the estimation of linear submanifold integrals enjoys the same rate as a 1-dimensional nonparametric regression problem. This illustrates the (potentially) significant “dimension reduction” achieved through integration.

The asymptotic normality of the $\hat{\theta}$ can also be established, but we defer it to Section 4, which provides a general asymptotic normality result that covers both linear and nonlinear submanifold integrals.

4 Inference on (Non)Linear Submanifold Integrals

4.1 Linearization and Asymptotic Normality

In Section 3 we focus on the linear integral functional, but some of the key results there (especially Lemma 1) can also be applied to analyze a variety of nonlinear integral functionals that involve submanifolds. In this subsection, we present high-level conditions and a general theorem for asymptotic normality of plug-in estimator $\Gamma(\hat{h})$ of the potentially nonlinear functional $\Gamma(h_0)$, where \hat{h} is the linear sieve estimator described in Section 3.2. We then apply the results to two specific cases in Sections 4.2 and 4.3.

Specifically, in this section we consider the estimation of a nonlinear integral functional $\Gamma(h_0)$ whose pathwise derivative becomes a linear submanifold integral of form (13).

Assumption 9 (Linearization). *The functional $\Gamma : \mathbb{H} \rightarrow \mathbb{R}$ is such that*

$$D_h \Gamma(h_0)[v] = \int_{\{x: g(x)=\mathbf{0}\}} v(x) \bar{w}(x) \mathcal{H}^m(x), \quad \forall v \in \mathcal{V} := \mathbb{H} - \{h_0\}, \quad (16)$$

for some uniformly bounded function \bar{w} and some level set function $g : \mathcal{X} \rightarrow \mathbb{R}_{d-m}$ that satisfies Assumption 1, where both w and g may depend on the true (and unknown) h_0 .

We emphasize that the level set function g (and thus the submanifold) is *not* necessarily known in Assumption 9. For example, in Section 4.3 below, we consider a case where $g = h_0$.

Next, we impose a condition that controls the remainder term from the pathwise differentiation to be asymptotically negligible, so that the first-order linear term $D_h \Gamma(h_0)[\hat{h} - h_0]$ drives the convergence rate and asymptotic normality of $\Gamma(\hat{h})$.

Assumption 10. *Let $\mathbb{H}_{K_n} := \{h \in \mathcal{B}_{K_n} : \|h - h_0\|_\infty < \epsilon\}$. Suppose that*

$$\sup_{h \in \mathbb{H}_{K_n}} |\Gamma(h) - \Gamma(h_0) - D_h \Gamma(h_0)[h - h_0]| = o_p \left(\sqrt{K_n^{\frac{d-m}{d}}/n} \right). \quad (17)$$

Note that, when Γ is linear, Assumptions 9 and 10 are trivially satisfied and thus the asymptotic normality result in this subsection also covers the linear case. For various non-linear Γ , Assumptions 9 and 10 may be verified given the specific form of Γ with appropriate lower-level conditions: we provide two examples in Sections 4.2 and 4.3 below.

Lastly, we impose a standard Lindeberg condition that helps delivers the asymptotic normality via the Lindeberg CLT.

Assumption 11 (Lindeberg Condition). $\sup_{x \in \mathcal{X}} \mathbb{E}[\epsilon_i^2 \{|\epsilon_i| > c\} | X_i = x] \rightarrow 0$ as $c \rightarrow \infty$.

Theorem 3 (Asymptotic Normality). *Suppose that Assumptions 1-11 hold and the sieve dimension K_n is set to satisfy $K_n \log K_n/n = o(1)$ and $K_n^{-s/d} = o\left(\sqrt{K_n^{(d-m)/d}/n}\right)$. Then:*

$$\frac{\sqrt{n}(\hat{\theta} - \theta_0)}{\|v_{K_n}^*\|_{sd}} \xrightarrow{d} \mathcal{N}(0, 1) \quad \text{with} \quad \|v_{K_n}^*\|_{sd} \asymp \sqrt{K_n^{\frac{d-m}{d}}}.$$

Given the above, there exists a natural estimator of $\|v_{K_n}^*\|_{sd}$ given by

$$\begin{aligned} \widehat{\|v_{K_n}^*\|_{sd}} &:= \frac{1}{n} \sum \hat{v}_{K_n}^*(X_i)^2 (Y_i - \hat{h}(X_i))^2 \\ \hat{v}_{K_n}^*(x) &:= \bar{b}^{K_n}(x) \left(\frac{1}{n} \sum_{i=1}^n \bar{b}^{K_n}(X_i) \bar{b}^{K_n}(X_i) \right)^{-1} D_h \Gamma(\bar{b}^{K_n}) \end{aligned}$$

and its consistency can be easily established following Chen and Christensen (2015) under the following additional assumptions:

Assumption 12. $\|v_{K_n}^*\|^{-1} \|D\Gamma(h)[\tilde{b}^{K_n}] - D\Gamma(h_0)[\tilde{b}^{K_n}]\| = o(1)$ uniformly over \mathbb{H}_{K_n} or $B_{\epsilon, \infty}(h_0)$, where $B_{\epsilon, \infty}(h_0) := \{h \in \mathbb{H} : \|h - h_0\|_\infty < \epsilon\}$.

Assumption 12 corresponds to Assumption 10 in [Chen and Christensen \(2015\)](#) and Assumption 3.1(iii) in [Chen et al. \(2014\)](#): it is trivially satisfied when Γ is linear.

Assumption 13 (Higher Error Moments). $\mathbb{E} [|\epsilon_i|^{2+\delta}] < \infty$ for some $\delta > 0$.

Theorem 4 (Consistent Variance Estimation). *Suppose that Assumptions 1-13 hold and the sieve dimension K_n is set to satisfy $K_n^{(2+\delta)/\delta} \log K_n/n = o(1)$ and $K_n^{-s/d} = o\left(\sqrt{K_n^{(d-m)/d}/n}\right)$. we have*

$$\left| \widehat{\|v_{K_n}^*\|_{sd}} / \|v_{K_n}^*\|_{sd} - 1 \right| = o_p(1) \quad \text{and} \quad \frac{\sqrt{n}(\hat{\theta} - \theta_0)}{\widehat{\|v_{K_n}^*\|_{sd}}} \xrightarrow{d} \mathcal{N}(0, 1).$$

4.2 Integral of Nonlinear Transformation of h_0

In this subsection, we consider the estimation of a specific class of nonlinear Γ and provide lower-level conditions for Assumptions 9 and 10. Specifically, consider Γ defined as

$$\Gamma(h) := \int_{\mathcal{M}} \phi(h(x), x) w(x) d\mathcal{H}^m(x), \quad (18)$$

where $\phi(t, x)$ is a known nonlinear transformation.

Proposition 1. *Suppose that Γ is defined in (18) with $\phi(t, x)$ being L -Lipchitz in t , so that its derivative $\phi_1(t, x) := \frac{\partial}{\partial t} \phi(t, x)$ is well-defined almost everywhere and uniformly bounded by the Lipchitz constant L whenever well-defined. Then (a) Assumption 9 holds with*

$$D_h \Gamma(h_0)[v] = \int_{\mathcal{M}} v(x) \phi_1(h_0(x), x) w(x) d\mathcal{H}^m(x). \quad (19)$$

(b) Furthermore, if:

- i) $\phi_1(t, x)$ is Lipchitz so that the second order derivative $\phi_{11}(t, x)$ is well-defined almost everywhere and uniformly bounded whenever defined,
- ii) the weight w is uniformly bounded,
- iii) the marginal density $p_0(x)$ of X_i is bounded from below on \mathcal{X} , and
- iv) the estimator \hat{h} satisfies the following convergence rate requirement:

$$\|\hat{h} - h_0\|_\infty = o_p\left(n^{-\frac{1}{4}} K_n^{\frac{d-m}{2d}}\right), \quad (20)$$

Then Assumption 10 is satisfied.

We note that our Theorem 3 targets a “ $(d - m)$ -dimensional nonparametric rate” instead of the parametric $1/\sqrt{n}$ rate, so the rate condition (20) here only requires that the linearization remainder term be faster than the square root of the “ $(d - m)$ -dimensional nonparametric rate”, a less stringent requirement than the usual $o_p(n^{-1/4})$ requirement.

4.3 Integral on Upper Contour Set of h_0

In this subsection, we consider the nonlinear functional $\Gamma(h_0)$ with

$$\Gamma(h) := \int_{\{h(x) \geq 0\}} w(x) dx, \quad (21)$$

where $h_0 : \mathbb{R}^d \rightarrow \mathbb{R}$ is an unknown (scalar-valued) function and $w(x)$ is known. Even though here Γ is a full-dimensional (Lebesgue) integral per se, its pathwise derivative w.r.t. h becomes a submanifold integral over the level set $\{x : h_0(x) = 0\}$ as we show below. Hence, in this case the “level set function” $g = h_0$ is taken to be unknown and needs to be nonparametrically estimated. Notice that such “upper contour integrals” show up in several examples in Section 2 that feature subpopulations defined through inequalities, such as the welfare/value under optimal treatment assignment.

Proposition 2. *Suppose that Γ is defined in (21) with*

$$\|\nabla_x h_0(x)\| \geq \underline{\epsilon} > 0, \quad \text{on } \mathcal{M} := \{x \in \mathcal{X} : h_0(x) = 0\}.$$

Then: (a) \mathcal{M} is an $m = (d - 1)$ dimensional submanifold, and Assumption 9 holds with

$$D_h \Gamma(h_0)[v] = \int_{\{h_0(x)=0\}} v(x) \frac{w(x)}{\|\nabla_x h_0(x)\|} d\mathcal{H}^{d-1}(x). \quad (22)$$

(b) Furthermore, if w , ∇w , and the second-order (partial) derivatives of h_0 are all uniformly bounded, and if

$$\|\hat{h} - h_0\|_\infty \|\nabla_x \hat{h} - \nabla_x h_0\|_\infty = o_p\left(\sqrt{\frac{1}{n} K_n^{1/d}}\right), \quad (23)$$

then Assumption 10 is satisfied.

We note that the calculation of the pathwise derivative (22) above is nonstandard, since it involves differentiation w.r.t. the changing boundary of the region of integration:

$$D_h \Gamma(h_0)[v] = \frac{d}{dt} \int_{\{h(x)+tv(x) \geq 0\}} w(x) dx.$$

The derivation of (22) utilizes mathematical techniques in differential geometry and the calculus of moving manifolds, using the “flow” of a vector field. The analysis of the linearization remainder term via the second-order pathwise derivative is even more involved,

since it requires the calculation of derivatives of the form

$$\frac{d}{dt} \int_{\{h(x)+tv(x)=0\}} w(x) d\mathcal{H}^{d-1}(x),$$

which involves moving level sets defined by $\{h(x) + tv(x) = 0\}$. Importantly, the gradient term $\nabla_x v(x)$ shows up in the calculation, so the leading term in the linearization remainder needs to be controlled by $\|\hat{h} - h_0\|_\infty \|\nabla_x \hat{h} - \nabla_x h_0\|_\infty$ instead of $\|\hat{h} - h_0\|_\infty^2$. See Appendix C for more details.

5 Monte Carlo Simulations

5.1 Integral on Known Submanifold

We first report simulations results for the estimation of an integral functional of a nonparametric function over a known submanifold defined by the unit circle: $\theta_0 := \Gamma(h_0)$ with

$$\begin{aligned} \Gamma(h) &:= \int_{\mathbb{S}^1} h(x) d\mathcal{H}^1(x), \\ h_0(x) &:= x_1^2 + 2 \sin(x_1) x_2, \quad x = (x_1, x_2)' \in \mathcal{X} = [-2, 2]^2. \end{aligned}$$

Under the variable transformation $x_1 = \cos(\beta)$ and $x_2 = \sin(\beta)$, we obtain the true value

$$\Gamma(h_0) := \int_0^{2\pi} (\cos(\beta)^2 + 2 \sin(\cos(\beta)) \sin(\beta)) d\beta = \pi.$$

In this exercise, $\Gamma(h_0)$ is a linear functional for which the unknown function h_0 enters only through the integrand, while the region of integration defined by the lower dimensional submanifold (i.e., the unit circle \mathbb{S}^1) is known and given. Hence, this exercise corresponds directly to the theoretical results established in Theorem 1.

We simulate $X_{i1} \sim_{i.i.d.} X_{i2} \sim_{i.i.d.} \text{Uniform}[-2, 2]$ and $Y_i = h_0(X_i) + \epsilon_i$ with $\epsilon_i \sim \mathcal{N}(0, 1)$. We compute a spline nonparametric regression estimator $\hat{h}(x)$ for $h_0(x) = \mathbb{E}[Y_i | X_i = x]$ using $K_n = 36$ tensor-product B-spline terms based as in the NPIV package. Note that the nonparametric estimation of h_0 is carried out in the standard manner, which is not restricted to the unit circle and does not utilize angle reparametrization. Given \hat{h} , the integral unit circle $\int_{\mathbb{S}^1} \hat{h}(x) d\mathcal{H}^1(x)$ is numerically computed using the sample average over $M = 5000$ Sobol sequence points⁹ in the angle space $[0, 2\pi]$. We obtain the plug-in estimator $\hat{\theta}$ and construct the confidence interval $\text{CI} := [\hat{\theta} \pm 1.96\hat{\sigma}_\theta]$ where the standard error is computed

⁹The Sobol sequence sampling, proposed by [Sobol \(1967\)](#), is a well-known quasi-random Monte Carlo sampling method that generates a deterministic sequence of points, whose distribution asymptotically converges to the uniform distribution, but achieves better finite-sample approximation of the population expectation (integral) by the sample mean.

Table 1: Integral on Known Submanifold

n	RMSE	Bias	SD	CI_L	CI_U	U-L	Coverage
500	0.435	-0.0247	0.434	2.266	3.968	1.703	95.4%
1,000	0.305	0.00282	0.306	2.542	3.747	1.204	94.4%
2,000	0.218	0.00391	0.218	2.719	3.582	0.852	94.5%
4,000	0.152	-0.00684	0.152	2.833	3.437	0.604	95.4%
8,000	0.110	9.65×10^{-4}	0.110	2.927	3.354	0.427	94.5%

as

$$\hat{\sigma}_\theta^2 := D_h \Gamma(\hat{h}) [\psi^{(K)}] \hat{\Omega} D_h \Gamma(\hat{h}) [\psi^{(K)}]$$

where the directional derivative $D_h \Gamma(\hat{h})[v] := \int_{\mathbb{S}^1} v(x) d\mathcal{H}^1(x)$ is numerically computed in the same manner as described above, $\psi^{(K)}$ denotes the K spline basis terms, and $\hat{\Omega} := (\Psi^{(K)} \Psi^{(K)'})^{-1} \left(\frac{1}{n} \sum_{i=1}^n u_i^2 \Psi^{(K)} \Psi^{(K)'} \right) (\Psi^{(K)} \Psi^{(K)'})^{-1}$, with the residual $u_i = Y_i - \hat{h}(X_i)$, and $\Psi^{(K)}$ denotes the $n \times K$ matrix of the K spline terms evaluated at the n data points X_1, \dots, X_n ,

Table 1 reports the finite-sample performance of the plug-in estimator and the constructed 95% confidence interval, for five different sample sizes, across $B = 1000$ Monte Carlo replications. We report the square root of the mean squared error (RMSE), the bias (Bias), the standard deviation (SD)¹⁰ of the estimator, the average lower and upper bounds of the confidence interval (CI_L and CI_U), the average length of the confidence interval (U-L), and the realized coverage probability of the CI (Coverage). Overall, the plug-in estimator and the corresponding CI perform very well under all five sample sizes: the RMSE shrinks (almost at \sqrt{n} rate) as the sample size increases, the bias is of negligible order relative to the standard deviation, and the realized coverage probability is close to the nominal 95% level.

5.2 Integral on Estimated Upper Contour Set

In this subsection, we analyze the estimation of an integral functional of a nonparametric function over the submanifold defined by the unit circle: $\theta_0 := \Gamma(h_0)$ with

$$\begin{aligned} \Gamma(h_0) &:= \int \mathbb{1}\{x \in \mathcal{X} : h_0(x) \geq 0\} dx, \quad \mathcal{X} = [-2, 2]^2 \\ h_0(x) &:= (1 - \|x\|^2) (4 + \sin(x_1) x_2 + \cos(x_2)). \end{aligned}$$

Under the above construction, $h_0(x) \geq 0$ if and only if $\|x\| \leq 1$, and thus

$$\theta_0 = \int \mathbb{1}\{\|x\| \leq 1\} dx = \pi.$$

¹⁰The standard deviation is calculated using the standard $\frac{1}{B-1} \sum_{b=1}^B$ formula. For this technical reason, “SD” can be larger than “RMSE”, which is calculated based on the $\frac{1}{B} \sum_{b=1}^B$ formula.

Table 2: Integral on Estimated Upper Contour Set

n	RMSE	Bias	SD	CI_L	CI_U	U-L	Coverage
500	0.0645	8.69×10^{-4}	0.0646	2.987	3.298	0.312	95.9%
1,000	0.0455	-7.68×10^{-4}	0.0456	3.030	3.252	0.222	96.0%
2,000	0.0335	-0.00321	0.0334	3.060	3.217	0.157	94.7%
4,000	0.0233	-0.00258	0.0232	3.083	3.195	0.112	95.8%
8,000	0.0172	-3.44×10^{-4}	0.0172	3.103	3.180	0.0767	95.2%

We compute the spline nonparametric regression estimator $\hat{h}(x)$ for $h_0(x) = \mathbb{E}[Y_i|X_i = x]$ using $K_n = 64$ spline basis terms, and the plug-in estimator $\Gamma(\hat{h}) = \int \mathbb{1}\{\hat{h}(x) \geq 0\} dx$ using the sample average over $M = 5000$ Sobol sequence points, and $\text{CI} := [\hat{\theta} \pm 1.96\hat{\sigma}_\theta]$ using the formula $\hat{\sigma}_\theta^2 := D_h\Gamma(\hat{h})[\psi^{(K)}] \hat{\Omega} D_h\Gamma(\hat{h})[\psi^{(K)}]$. The formula of $\hat{\Omega}$ is the same as in Section 5.1 but the directional derivative $D_h\Gamma(h)[v]$ now takes the following form ($d = 2$ in this example):

$$D_h\Gamma(h)[v] = \int_{\{x \in \mathcal{X}: h(x)=0\}} \frac{v(x)}{\|\nabla_x h(x)\|} d\mathcal{H}^{d-1}(x),$$

which we numerically approximate via

$$\hat{D}_h\Gamma(h)[v] = \frac{1}{2\epsilon} \int_{\{x \in \mathcal{X}: -\epsilon < h(x) < \epsilon\}} v(x) dx$$

based on the mathematical result¹¹ that

$$\lim_{\epsilon \searrow 0} \frac{1}{2\epsilon} \int_{\{x \in \mathcal{X}: -\epsilon < h(x) < \epsilon\}} v(x) dx = \int_{\{x \in \mathcal{X}: h(x)=0\}} \frac{v(x)}{\|\nabla_x h(x)\|} d\mathcal{H}^{d-1}(x).$$

We set $\epsilon = 0.001$ in our simulation, and, given that $\{x \in \mathcal{X}: -\epsilon < h(x) < \epsilon\}$ may occur infrequently for small ϵ , we use the sample average from $M' = 100,000$ Sobol sequence points to approximate $\hat{D}_h\Gamma(h)[v]$.

Table 2 reports the finite-sample performance of the estimator and the CI based on $B = 1000$ Monte Carlo replications. The results in Table 2 display a similar pattern as those in Table 1: Again, the RMSE shrinks (almost at \sqrt{n} rate) as the sample size increases, the bias is of smaller order relative to the standard deviation, and the realized CI coverage is close to the nominal 95% level. One noticeable difference between the two tables, however, lies in that the RMSEs in Table 2 are substantially smaller than those in Table 1. Heuristically, this might have been due to the fact that the upper contour set integral being estimated in Table 2 is itself a full-dimensional integral: even though its asymptotic behavior is theoretically driven by the lower-dimensional submanifold, in finite sample the full-dimensional nature of the integral may have made it overall easier to estimate.

¹¹See Theorem 3.13.(iii) of Evans and Gariepy (2015).

6 Conclusion

This paper studies the semiparametric plug-in estimation of integral functionals on submanifolds, and establishes under a standard set of regularity conditions that the semiparametric plug-in estimator is asymptotically normal with the minimax-optimal convergence rate $n^{-\frac{s}{2s+d-m}}$, which corresponds to the usual (minimax-optimal) rate of a $(d-m)$ -dimensional nonparametric estimation problem. This shows that the “Lebesgue measure zero” or “probability zero” subsets of a population can still provide meaningful dimensional reduction through integration over such subpopulations with respect to the Hausdorff measure.

As discussed in Section 2, our formulation and analysis of submanifold integrals is relevant in a variety of econometric settings, immediately suggesting several directions for future research. It would be interesting to investigate specific economic problems of interest, such as optimal treatment assignment and marginal/policy-relevant treatment effects, based on the analysis of submanifold integrals in combination of specific economic ingredients and institutional details pertinent to the problem in question.

References

- AI, C. and CHEN, X. (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, **71** (6), 1795–1843.
- and — (2007). Estimation of possibly misspecified semiparametric conditional moment restriction models with different conditioning variables. *Journal of Econometrics*, **141** (1), 5–43.
- and — (2012). The semiparametric efficiency bound for models of sequential moment restrictions containing unknown functions. *Journal of Econometrics*, **170** (2), 442–457.
- CARNEIRO, P., HECKMAN, J. J. and VYTLACIL, E. (2010). Evaluating marginal policy changes and the average effect of treatment for individuals at the margin. *Econometrica*, **78** (1), 377–394.
- CHEN, X., CHRISTENSEN, T. and KANKANALA, S. (2025). Adaptive estimation and uniform confidence bands for nonparametric structural functions and elasticities. *Review of Economic Studies*, **92** (1), 162–196.
- and CHRISTENSEN, T. M. (2015). Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions. *Journal of Econometrics*, **188** (2), 447–465.

- and — (2018). Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric iv regression. *Quantitative Economics*, **9** (1), 39–84.
- , LIAO, Y. and WANG, W. (2024). Inference on time series nonparametric conditional moment restrictions using nonlinear sieves. *Journal of Econometrics*, p. 105920.
- , LIAO, Z. and SUN, Y. (2014). Sieve inference on possibly misspecified semi-nonparametric time series models. *Journal of Econometrics*, **178**, 639–658.
- and POUZO, D. (2015). Sieve wald and qlr inferences on semi/nonparametric conditional moment models. *Econometrica*, **83** (3), 1013–1079.
- , — and POWELL, J. L. (2019). Penalized sieve gel for weighted average derivatives of nonparametric quantile iv regressions. *Journal of Econometrics*, **213** (1), 30–53.
- CHERNOZHUKOV, V., FERNÁNDEZ-VAL, I. and LUO, Y. (2018). The sorted effects method: Discovering heterogeneous effects beyond their averages. *Econometrica*, **86** (6), 1911–1938.
- EVANS, L. C. and GARIEPY, R. F. (2015). *Measure Theory and Fine Properties of Functions*. CRC Press.
- FEDERER, H. (1996). Geometric measure theory. *Classics in Mathematics*.
- FENG, K. and HONG, H. (2024). Statistical inference of optimal allocations i: Regularities and their implications. *arXiv preprint arXiv:2403.18248*.
- HECKMAN, J. J. and VYTLACIL, E. (2001). Policy-relevant treatment effects. *American Economic Review*, **91** (2), 107–111.
- and — (2005). Structural equations, treatment effects, and econometric policy evaluation 1. *Econometrica*, **73** (3), 669–738.
- and VYTLACIL, E. J. (2007). Econometric evaluation of social programs, part ii: Using the marginal treatment effect to organize alternative econometric estimators to evaluate social programs, and to forecast their effects in new environments. *Handbook of econometrics*, **6**, 4875–5143.
- HOROWITZ, J. L. (1992). A smoothed maximum score estimator for the binary response model. *Econometrica: journal of the Econometric Society*, pp. 505–531.
- KHAN, S. and TAMER, E. (2010). Irregular identification, support conditions, and inverse weight estimation. *Econometrica*, **78** (6), 2021–2042.

- KIM, J. and POLLARD, D. (1990). Cube root asymptotics. *The Annals of Statistics*, pp. 191–219.
- KITAGAWA, T. and TETENOV, A. (2018). Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, **86** (2), 591–616.
- LANG, S. (2002). *Introduction to differentiable manifolds*, vol. 32. Springer.
- LOOMIS, L. H. and STERNBERG, S. Z. (2014). *Advanced Calculus (Revised Edition)*. World Scientific Publishing Company.
- MANSKI, C. F. (1975). Maximum score estimation of the stochastic utility model of choice. *Journal of econometrics*, **3** (3), 205–228.
- MUNKRES, J. R. (1991). *Analysis On Manifolds*. Westview Press.
- NEWBY, W. K. (1994). Kernel estimation of partial means and a general variance estimator. *Econometric Theory*, **10** (2), 1–21.
- and STOKER, T. M. (1993). Efficiency of weighted average derivative estimators and index models. *Econometrica: Journal of the Econometric Society*, pp. 1199–1223.
- POWELL, J. L., STOCK, J. H. and STOKER, T. M. (1989). Semiparametric estimation of index coefficients. *Econometrica: Journal of the Econometric Society*, pp. 1403–1430.
- QIAO, W. (2021). Nonparametric estimation of surface integrals on level sets. *Bernoulli*, **27** (1), 155–191.
- SASAKI, Y. (2015). What do quantile regressions identify for general structural functions? *Econometric Theory*, **31** (5), 1102–1116.
- SOBOL, I. M. (1967). The distribution of points in a cube and the approximate evaluation of integrals. *USSR Computational mathematics and mathematical physics*, **7**, 86–112.
- STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The annals of statistics*, pp. 1040–1053.

Mathematical Appendix

This Appendix consists of four sections. Section A provides an instrumental transformation of submanifold integrals wrt Hausdorff measures into sums of lower-dimensional Lebesgue integrals using basic differential geometry techniques. Appendix B provides proofs of the results in Section 3. Appendix C provides proofs of the results in Section 4. The last section establishes additional results using a kernel nonparametric regression in the first stage.

A Decomposition of Hausdorff Integrals on Submanifolds via Partition of Unity

In this section, we explain how we can decompose a Hausdorff integral on a regular submanifold into a sum of lower-dimensional Lebesgue integrals, which can be analyzed more straightforwardly based on the existing theory on the semiparametric estimation of Lebesgue integral functionals.

This is done using a combination of standard mathematical tools from differential manifold theory and geometric measure theory: see, for example, [Munkres \(1991\)](#) and [Lang \(2002\)](#) for textbook treatments of differentiable manifolds, as well as [Federer \(1996\)](#) and [Evans and Gariepy \(2015\)](#) textbook treatments of geometric measure theory. Specifically, the key idea is as follows:

- (i) Construct a finite open cover of the manifold \mathcal{M} in question.
- (ii) On each piece of the open cover in (i), apply the implicit function theorem to obtain a parametrization (i.e., a coordinate chart) of the m -dimensional manifold by an open subset of an m -dimensional Euclidean space.
- (iii) On each piece of the open cover in (i), use the “change of variable” formula in geometric measure theory to convert Hausdorff integral on the manifold into an m -dimensional Lebesgue integral via the parametrization in (ii).
- (iv) Use the “partition of unity” method to combine the results in (iii) across different pieces of the open cover in (i).

We now proceed through (i)-(iv) formally with more details.

By the regularity of \mathcal{M} (Assumption 1), for every $\bar{x} \in \mathcal{M}$, $\nabla_x g(\bar{x})$ has rank $d-m$. Hence, we can decompose \bar{x} , potentially with a permutation of coordinate indexes, as $(\bar{x}_{(m)}, \bar{x}_{-(m)})$,

where $\bar{x}_{(m)}$ is an m -dimensional subvector and $\bar{x}_{-(m)}$ is the remaining $(d - m)$ -dimensional subvector, and correspondingly decompose $\nabla_x g(\bar{x})$ into

$$\nabla_x g(\bar{x}) = \begin{bmatrix} \underbrace{\nabla_{x_{(m)}} g(\bar{x})}_{m \times (d-m)}, \underbrace{\nabla_{x_{-(m)}} g(\bar{x})}_{(d-m) \times (d-m)} \end{bmatrix}$$

so that $\text{rank}(\nabla_{x_{-(m)}} g(\bar{x})) = d - m$. Then, by the Implicit Function Theorem (see, e.g. Theorem 9.2 in [Munkres, 1991](#)), there exists an open neighborhood $\mathcal{U}_{\bar{x}}$ around $\bar{x}_{(m)}$ and a unique r -times continuously differentiable function $\psi_{\bar{x}} : \mathcal{U}_{\bar{x}} \rightarrow \mathbb{R}^{-(m)}$ such that

$$x_{-(m)} = \psi_{\bar{x}}(x_{(m)}), \quad \forall x_{(m)} \in \mathcal{U}_{\bar{x}}.$$

Then, define $\varphi_{\bar{x}} : \mathcal{U}_{\bar{x}} \rightarrow \mathcal{M} \subseteq \mathbb{R}^d$ by

$$\varphi_{\bar{x}}(x_{(m)}) := (x_{(m)}, \psi_{\bar{x}}(x_{(m)})).$$

Then $\varphi_{\bar{x}}$ is a diffeomorphism (r -times differentiable bijection) between $\mathcal{U}_{\bar{x}}$ and $\varphi_{\bar{x}}(\mathcal{U}_{\bar{x}}) \subseteq \mathcal{M}$: $(\mathcal{U}_{\bar{x}}, \varphi_{\bar{x}})$ is thus a local coordinate chart of \mathcal{M} at \bar{x} , which parameterizes each point x in $\varphi_{\bar{x}}(\mathcal{U}_{\bar{x}})$ on the manifold with a vector $x_{(m)} \in \mathcal{U}_{\bar{x}}$, where $\mathcal{U}_{\bar{x}}$ is an open set in an m -dimensional Euclidean space \mathbb{R}^m .

Clearly $\{\varphi_{\bar{x}}(\mathcal{U}_{\bar{x}})\}_{\bar{x} \in \mathcal{M}}$ is an open cover for \mathcal{M} . Since \mathcal{M} is compact, there exists a finite sub-cover $\{\varphi_j(\mathcal{U}_j) : j = 1, \dots, \bar{j}\}$, where, for each j , $(\mathcal{U}_j, \varphi_j) = (\mathcal{U}_{\bar{x}^{(j)}}, \varphi_{\bar{x}^{(j)}})$ for some point $\bar{x}^{(j)} \in \mathcal{M}$. From now on, we also write ψ_j as the function $\psi_{\bar{x}^{(j)}}$ associated with $(\mathcal{U}_{\bar{x}^{(j)}}, \varphi_{\bar{x}^{(j)}})$, and we write $x_{(j,m)}$ for a generic point in $\mathcal{U}_{\bar{x}^{(j)}}$. The “ j ” in the subscript “ (j, m) ” emphasizes that, while $x_{(j,m)}$ is always an m -dimensional vector that corresponds to m different coordinates in \mathbb{R}^d , the exact sequence of the m coordinates may differ across different j ’s.

In general, the open cover $\{\varphi_j(\mathcal{U}_j) : j = 1, \dots, \bar{j}\}$ may have nonempty intersections. The standard mathematical tool to avoid “double counting” under potentially overlapping covers is the so-called “partition of unity”, i.e., a collection of smooth real-valued functions $\{\rho_j : j = 1, \dots, \bar{j}\}$ on \mathbb{R}^d such that (i) each ρ_j is nonnegative (ii) $\sum_{j=1}^{\bar{j}} \rho_j(x) = 1$ for all $x \in \mathcal{M}$, and (iii) $\rho_j(x) = 0$ for all $x \notin \varphi_j(\mathcal{U}_j)$. Such a “partition of unity” (ρ_j) is guaranteed to exist, say, by Lemma 25.2 in [Munkres \(1991\)](#).

Given $(\mathcal{U}_j, \varphi_j, \rho_j)_{j=1}^{\bar{j}}$, we may then decompose an integral on the manifold \mathcal{M} into a sum of integrals on $\varphi_j(\mathcal{U}_j)$: for any map $\omega : \mathcal{X} \mapsto \mathbb{R}$,

$$\int_{\mathcal{M}} \omega(x) d\mathcal{H}^m(x) = \sum_{j=1}^{\bar{j}} \int_{\varphi_j(\mathcal{U}_j)} \rho_j(x) \omega(x) d\mathcal{H}^m(x) \quad (24)$$

Then, since φ_j is a bijection between \mathcal{U}_j and $\varphi_j(\mathcal{U}_j)$ by construction, each $x \in \varphi_j(\mathcal{U}_j)$ can

be expressed as $x = \varphi_j(x_{(j,m)})$ for $x_{(j,m)} \in \mathbb{R}^m$.

By the “change-of-variable” formula in geometric measure theory, e.g., Theorem 3.9 in Evans and Gariepy (2015), each of the \bar{j} integrals on $\varphi_j(\mathcal{U}_j)$ can be evaluated as Lebesgue integrals on $\mathcal{U}_j \subset \mathbb{R}^m$ through the parametrization φ_j , i.e.,

$$\int_{\varphi_j(\mathcal{U}_j)} \rho_j(x) w(x) d\mathcal{H}^m(x) = \int_{\mathcal{U}_j} \rho_j(\varphi_j(x_{(j,m)})) w(\varphi_j(x_{(j,m)})) \mathcal{J}\varphi_j(x_{(j,m)}) dx_{(j,m)}, \quad (25)$$

where $\mathcal{J}\varphi_j := \sqrt{\det(\varphi_j' \varphi_j)}$ denotes the Jacobian of φ_j . Combining (24) and (25) we obtain:

$$\int_{\mathcal{M}} \omega(x) d\mathcal{H}^m(x) = \sum_{j=1}^{\bar{j}} \int_{\mathcal{U}_j} \rho_j(\varphi_j(x_{(j,m)})) \omega(\varphi_j(x_{(j,m)})) \mathcal{J}\varphi_j(x_{(j,m)}) dx_{(j,m)}. \quad (26)$$

The decomposition in (26) converts a Hausdorff integral on a m -dimensional manifold \mathcal{M} into a finite sum of m -dimensional Lebesgue integrals on \mathbb{R}^m , the latter of which are easier to analyze using existing results on semi/nonparametric estimation. Below we show how we use (26) to establish the convergence rate and asymptotic normality of the semiparametric estimator when the first stage nonparametric function h_0 could be learned/estimated using any machine learning/AI algorithms. The key is to use a sieve Riesz representation theory to establish the asymptotic local influence function representation for $\theta = \Gamma(h)$.

B Proofs of Theoretical Results in Section 3

To establish the lower bound for the convergence rate in Theorem 1, we use Le Cam’s two-point comparison approach based on KL divergence.

Lemma 2 (Le Cam’s Minimax Rate Bounds based on KL divergence). *Suppose that there exist P_0, P_1 such that $KL(P_0, P_1) \leq \frac{\log 2}{n}$. Then*

$$R_n := \inf_{\hat{\theta}} \sup_P \mathbb{E}_P [d(\hat{\theta}, \theta(P))] \geq \frac{1}{16} d(\theta(P_0), \theta(P_1)).$$

Proof of Theorem 1. (i) We first prove the result for the nonparametric regression case $h_0(x) = \mathbb{E}[Y_i | X_i = x]$. Given any $h_0 \in \Lambda_c^s(\mathcal{X})$, let b_n be a small positive number and define

$$h_1(x) := h_0(x) + b_n^s K_{d-m} \left(\frac{g(x)}{b_n} \right)$$

where $K_{d-m}(x) = \prod_{\ell=1}^{d-m} K(x_\ell)$ and

$$K(t) := a \exp \left(-\frac{1}{1-t^2} \right) \mathbb{1}\{|t| \leq 1\} \quad (27)$$

for some sufficiently small $a > 0$ such that h_1 stays in Hölder class of smoothness order s . To see this, notice that K and K_{d-m} are infinitely differentiable with uniformly bounded

derivatives, and for any $k \leq s$ we have:

$$\frac{\partial^k}{\partial x_j^k} h_1(x) = \frac{\partial^k}{\partial x_j^k} h_0(x) + b_n^{s-k} \left(K_{d-m}^{(1)}(\cdot) \frac{\partial^k}{\partial x_j^k} g(x) + \dots + K_{d-m}^{(k)}(\cdot) \frac{\partial}{\partial x_j} g(x) \right).$$

Since $h_0, g \in \Lambda_c^s(\mathcal{X})$, all derivatives of K are uniformly bounded, and b_n is small, $\frac{\partial^k}{\partial x_j^k} h_1(x)$ is uniformly bounded as well. The Hölder condition for the fractional exponent $s - [s]$ can also be similarly verified.

Note that, to derive a lower bound for the minimax rate for a class of estimation problems, it suffices to establish the lower bound under one admissible example of the problem. Specifically, we consider the example setting where X_i be uniformly distributed on $[0, 1]^d$, and $w(x) \equiv 1$. In addition, let P_0 be the joint distribution of $(X_i, Y_{i0})_{i=1}^n$ with

$$Y_{i0} = h_0(X_i) + \epsilon_i, \quad \epsilon_i \sim_{i.i.d.} \mathcal{N}(0, 1).$$

and let P_1 be the joint distribution of $(X_i, Y_{i1})_{i=1}^n$ with

$$Y_{i1} = h_1(X_i) + \epsilon_i, \quad \epsilon_i \sim_{i.i.d.} \mathcal{N}(0, 1).$$

Then, the KL divergence between P_0^n and P_1^n is given by:

$$\begin{aligned} KL(P_0^n, P_1^n) &= nKL(P_0, P_1) \\ &= n \int_{[0,1]^d} \int p_0(x, y) \log \left(\frac{p_0(x, y)}{p_1(x, y)} \right) dy dx \\ &= n \int_{[0,1]^d} \int p_0(y|x) \log \left(\frac{p_0(y|x)}{p_1(y|x)} \right) dy dx \\ &= n \int_{[0,1]^d} \int \phi(y) \log \left(\frac{\phi(y - h_0(x))}{\phi(y - h_1(x))} \right) dy dx \\ &= n \int_{[0,1]^d} \int KL(\mathcal{N}(h_0(x), 1), \mathcal{N}(h_1(x), 1)) dy dx \\ &= n \frac{1}{2} \int_{[0,1]^d} (h_1(x) - h_0(x))^2 dx \\ &= n \frac{1}{2} \int_{[0,1]^d} b_n^{2s} K_{d-m}^2 \left(\frac{g(x)}{b_n} \right) dx \\ &= n \frac{1}{2} b_n^{2s} \int_{[0,1]^d} K_{d-m}^2 \left(\frac{g(x)}{b_n} \right) dx \\ &\leq n \frac{a^2}{2} b_n^{2s} \int_{[0,1]^d} \prod_{\ell=1}^{d-m} \mathbb{1} \left\{ \frac{|g_\ell(x)|}{b_n} \leq 1 \right\} dx \\ &\leq C n b_n^{2s} \mathbb{P}(|g(X_i)| \leq b_n) \\ &\leq C' n b_n^{2s+d-m} \end{aligned} \tag{28}$$

provided that

$$\mathbb{P}(\mathbb{P}(|g_\ell(X_i)| \leq b_n)) \leq Mb_n^{d-m}. \quad (29)$$

We now show that (29) holds. Writing $W_i := g(X_i)$, note that the density of W_i

$$p_W(w) := \int_{\{g(x)=w\}} p_X(x) \frac{1}{\mathcal{J}g(x)} d\mathcal{H}^m(x) \leq M,$$

is uniformly bounded since $p_X(x) \equiv 1$ and $\mathcal{J}g(x)$ is uniformly bounded away from zero as below by Assumption 2. Consequently,

$$\begin{aligned} \mathbb{P}(\|g(X_i)\| \leq b_n) &= \int_{\mathbb{1}_{\{|w| \leq b_n\}}} p_W(w) dw \\ &\leq M \int_{\mathbb{1}_{\{|w| \leq b_n\}}} dw \\ &\leq Mb_n^{d-m} \end{aligned}$$

verifying condition (29).

Hence, we can set $b_n = \left(\frac{\log 2}{C'_n}\right)^{\frac{1}{2s+d-m}} \asymp n^{-\frac{1}{2s+d-m}}$ so that

$$KL(P_0, P_1) \leq C' b_n^{2s+d-m} = \frac{\log 2}{n}$$

as required in Lemma 2.

In the meanwhile, notice that

$$|\Gamma(h_1) - \Gamma(h_0)| = \int_{\{g_0(x)=\mathbf{0}\}} [h_1(x) - h_0(x)] w(x) d\mathcal{H}^m(x) = C'' b_n^s K(0)$$

Hence, define $\theta(P) := \Gamma(\mathbb{E}_P[Y|X = \cdot])$ and $d(\theta, \theta') = |\theta - \theta'|$, we deduce from Lemma 2 that the minimax rate

$$\begin{aligned} R_n &\geq \frac{1}{16} d(\theta(P_0), \theta(P_1)) \\ &= \frac{1}{16} |\Gamma(h_1) - \Gamma(h_0)| \\ &= \frac{1}{16} \int_{\{g_0(x)=\mathbf{0}\}} |h_1(x) - h_0(x)| d\mathcal{H}^m(x) \\ &= C'' b_n^s |K(0)| \\ &\asymp b_n^s \asymp n^{-\frac{s}{2s+d-m}}. \end{aligned}$$

□

Proof of Lemma 1. Write $K_n = K = J^d$. Note that

$$\|\Gamma(\bar{b}^K)\|^2 = \sum_{k=1}^K \Gamma^2(\bar{b}_k^K) = \sum_{k=1}^{K_n} \left[\int_{\mathcal{M}} \bar{b}_k^{K_n}(x) w(x) d\mathcal{H}^m(x) \right]^2.$$

Recall that $(\bar{b}_k^{K_n})_{j=1}^{K_n}$ is constructed as tensor products of univariate basis functions

$$\bar{b}_k^{K_n}(x) = \prod_{\ell=1}^d b_{k_\ell}(x_\ell)$$

for some $1 \leq k_1, \dots, k_d \leq J$: in other words, each index k is bijectively identified by the vector (k_1, \dots, k_d) . Hence,

$$\sum_{k=1}^K \Gamma^2(\bar{b}_k^K) = \sum_{k_1, \dots, k_d=1}^J \Gamma^2\left(\prod_{\ell=1}^d b_{k_\ell}(x_\ell)\right).$$

For each $\Gamma(\bar{b}_k^K)$, we apply the decomposition (26) and obtain

$$\Gamma(\bar{b}_k^K) = \int_{\mathcal{M}} \bar{b}_k^K(x) w(x) p(x) d\mathcal{H}^m(x) = \sum_{j=1}^{\bar{j}} T_{kj},$$

where each j corresponds to a piece from the local charts $(\mathcal{U}_j, \varphi_j)$ for \mathcal{M} along with the partition of unit function ρ_j and

$$T_{kj} := \int_{\mathcal{U}_j} \rho_j(\varphi_j(x_{(j,m)})) \bar{b}_k^K(\varphi_j(x_{(j,m)})) w(\varphi_j(x_{(j,m)})) \mathcal{J}\varphi_j(x_{(j,m)}) dx_{(j,m)}.$$

From now on, to simplify notation, we will suppress the subscript j whenever there is no ambiguity. Furthermore, define

$$\bar{w}(x_{(m)}) := \rho(\varphi(x_{(m)})) w(\varphi(x_{(m)})) \mathcal{J}\varphi(x_{(m)}),$$

so that we may write T_{kj} more succinctly as

$$T_{kj} = \int_{\mathcal{U}} \bar{b}_k^K(\varphi(x_{(m)})) \bar{w}(x_{(m)}) dx_{(m)}.$$

Again, since $(\bar{b}_k^{K_n})$ is constructed as tensor products of univariate basis functions, we can decompose

$$\bar{b}_k^K(\varphi(x_{(m)})) = \bar{b}_{k,(m)}^K(x_{(m)}) \cdot \bar{b}_{k,-(m)}^K(\psi(x_{(m)}))$$

where $\bar{b}_{k,(m)}^K(x_{(m)}) = \prod_{\ell \in (m)} b_{k_\ell}(x_\ell)$ and $\bar{b}_{k,-(m)}^K(x_{(m)}) = \prod_{\ell \notin (m)} b_{k_\ell}([\psi(x_{(m)})]_\ell)$ correspond to the coordinates in $x_{(m)}$ and $x_{-(m)}$, respectively. Then, we have

$$\begin{aligned} T_{kj} &= \int_{\mathcal{U}} \bar{b}_{k,(m)}^K(x_{(m)}) \bar{b}_{k,-(m)}^K(\psi(x_{(m)})) \bar{w}(\varphi(x_{(m)})) dx_{(m)} \\ &= \langle \bar{b}_{(m)}^K(\cdot), \bar{b}_{-(m)}^K(\psi(\cdot)) \bar{w}_j(\varphi(\cdot)) \rangle_{L_2(\mathcal{U})} \end{aligned}$$

where $\langle f_1, f_2 \rangle_{L_2(\mathcal{U})} := \int_{\mathcal{U}} f_1(x_{(m)}) f_2(x_{(m)}) dx_{(m)}$.

Since $\{\bar{b}_{(m)}^K(\cdot)\}$ is a sequence of orthonormal basis functions on $\mathcal{X} \subseteq \mathbb{R}^d$, its restriction

to \mathcal{U} is a frame,¹² which

$$\begin{aligned}
\sum_{l:l \in (m)} \sum_{k_\ell=1}^J T_{kj}^2 &= \sum_{k_\ell:l \in (m)} \langle \bar{b}_{(m)}^K(\cdot), \bar{b}_{(m)}^K(\psi(\cdot)) \bar{w}_j(\varphi(\cdot)) \rangle_{L_2}^2 \\
&\asymp M \cdot \left\| \bar{b}_{(m)}^K(\psi(\cdot)) \bar{w}_j(\varphi(\cdot)) \right\|_{L_2}^2 \text{ by the frame condition} \\
&= M \cdot \int \left[\bar{b}_{(m)}^K(\psi(x_{(m)})) \right]^2 \bar{w}_j^2(\varphi(x_{(m)})) dx_{(m)} \\
&\asymp M.
\end{aligned}$$

Hence,

$$\sum_{k=1}^K T_{kj}^2 = \sum_{l': l' \notin (m)} \sum_{k_\ell=1}^J \left[\sum_{l:l \in (m)} \sum_{k_\ell=1}^J T_{4(k_1, \dots, k_d)j}^2 \right] \asymp \sum_{l': l' \notin (m)} \sum_{k_\ell=1}^J M = J^{(d-m)} M$$

Now, since $\Gamma(\bar{b}_k^K) = \sum_{j=1}^{\bar{j}} T_{kj}$, and $\Gamma^2(\bar{b}_k^K) = \left(\sum_{j=1}^{\bar{j}} T_{kj} \right)^2 \asymp \sum_{j=1}^{\bar{j}} T_{kj}^2$, we have

$$\left\| \Gamma(\bar{b}_k^K) \right\|^2 = \sum_{k=1}^K \Gamma^2(\bar{b}_k^K) \asymp \bar{j} \sum_{j=1}^{\bar{j}} J^{(d-m)} M \asymp M J^{(d-m)}.$$

□

Proof of Theorem 2. Note that

$$\begin{aligned}
|\hat{\theta} - \theta_0| &= \left| \Gamma(\hat{h} - h_0) \right| \\
&\leq \left| \Gamma(\hat{h} - \tilde{h}) \right| + \left| \Gamma(\tilde{h} - h_0) \right| \\
&\leq \left| \Gamma(\hat{h} - \tilde{h}) \right| + \left\| \tilde{h} - h_0 \right\|_\infty
\end{aligned}$$

By Theorem 3.1 of [Chen and Christensen \(2015\)](#), under the i.i.d. setting, under Assumptions 6-8, we have

$$\Gamma(\hat{h} - \tilde{h}) = O_p \left(\frac{1}{\sqrt{n}} \|v_{K_n}^*\|_{sd} \right) = O_p \left(\frac{1}{\sqrt{n}} \|v_{K_n}^*\|_2 \right).$$

Since $\|v_{K_n}^*\|_2^2 \asymp K_n^{\frac{d-m}{d}}$ by Lemma 1, we have

$$\Gamma(\hat{h} - \tilde{h}) = O_p \left(\sqrt{K_n^{\frac{d-m}{d}} / n} \right)$$

In addition, by Theorem 2.1 of [Chen and Christensen \(2015\)](#), we have

$$\left\| \tilde{h} - h_0 \right\|_\infty = O_p \left(K_n^{-s/d} \right).$$

¹²A (finite or countable) sequence of functions $(b_k(\cdot))$ is said to be a frame on a Hilbert space $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ of functions if it satisfies the *frame condition*: there exist constants $\underline{M}, \overline{M} > 0$ s.t. $\underline{M} \|f\|^2 \leq \sum_k \langle b_k(\cdot), f(\cdot) \rangle^2 \leq \overline{M} \|f\|^2$ for every $f \in \mathcal{H}$.

Hence,

$$\hat{\theta} - \theta_0 = O_p \left(\sqrt{K_n^{\frac{d-m}{d}}/n} + K_n^{-s/d} \right).$$

The rate is minimized by setting K_n such that $\sqrt{K_n^{\frac{d-m}{d}}/n} \asymp K_n^{-s/d}$, or

$$J_n = K_n^{1/d} = n^{\frac{1}{2s+d-m}},$$

we have

$$\hat{\theta} - \theta_0 = O_p \left(K_n^{-s/d} \right) = O_p \left(n^{-\frac{s}{2s+d-m}} \right).$$

□

C Proofs of Theoretical Results in Section 4

Proof of Theorem 3. The asymptotic normality result follows from Theorem 3.1 of [Chen and Christensen \(2015\)](#) [CC15 for short] with its stated assumptions and additional conditions verified (for i.i.d. setting). Specifically, Assumption 1(i) and 2(i) in CC15 are satisfied under the iid setting with $\mathcal{X} = [0, 1]^d$, Assumption 2(ii)(iv)(v) in CC15 is satisfied with Assumptions 8, 5, and 11. Assumption 4(iii) is verified with Assumption 6(i). Assumption 5 in CC15 is satisfied with Assumption 6 and the choice of K_n s.t. $K_n \log K_n/n = o(1)$. Assumption 9(i)(ii) are verified with Assumptions 9 and 10. Assumption 9(iii) is verified with the “undersmoothing” choice of K_n s.t.

$$\left| D\Gamma(h_0)[\tilde{h} - h_0] \right| \leq \left\| \tilde{h} - h_0 \right\|_{\infty} = O_p(K_n^{s/d}) = o_p \left(\sqrt{\frac{K_n^{\frac{d-m}{d}}}{n}} \right),$$

where the first inequality follows from Assumption 9 and the uniform boundedness of \bar{w} . Given Lemma 1, the rate for $\left\| v_{K_n}^* \right\|_{sd}$ follows from and Assumptions 5 and 11, which imply that $\left\| v_{K_n}^* \right\|_{sd} \asymp \left\| v_{K_n}^* \right\|$. □

Proof of Theorem 4. We apply Theorem 3.2 in CC15. In addition to the assumptions already verified above in the proof of Theorem 3, note that: Assumptions 2(iii) and 10 in CC15 are satisfied given Assumptions 13 and 12, and the additional condition $\left(\zeta_{K,n} \lambda_{K,n} \sqrt{\log(n)/n} \right) = o(1)$ is satisfied with Assumption 6(i)(ii) and the stated condition on the rate of K_n . □

Proof of Proposition 1. (19) follows from

$$\frac{d}{dt} \int_{\mathcal{M}} \phi(h_0(x) + tv(x), x) w(x) d\mathcal{H}^m(x) = \int_{\mathcal{M}} \phi_1(h_0(x) + tv(x), x) v(x) w(x) d\mathcal{H}^m(x).$$

Since $\phi(t, x)$ is almost everywhere twice differentiable in t with its second derivative ϕ_{11} bounded, we have

$$|\phi(h(x), x) - \phi(h_0(x), x) - \phi_1(h_0(x), x)(h(x) - h_0(x))| \leq M |h(x) - h_0(x)|^2$$

and thus

$$|\Gamma(h) - \Gamma(h_0) - D_h \Gamma(h_0)[h - h_0]| \leq M \int_{\mathcal{M}} |h(x) - h_0(x)|^2 w(x) d\mathcal{H}^m(x) \leq M \|h - h_0\|_{\infty}^2$$

Hence, (17) is satisfied if

$$\|\hat{h} - h_0\|_{\infty}^2 = o_p \left(\sqrt{\frac{1}{n} K_n^{\frac{d-m}{d}}} \right).$$

□

Proof of Proposition 2(a). Defining

$$\mathcal{V}_t := \{x : h_0(x) + tv(x) \geq 0\},$$

whose boundary is given by

$$\partial \mathcal{V}_t = \mathcal{M}_t := \{x : h_0(x) + tv(x) = 0\}$$

with

$$\mathcal{V}_0 = \{x : h_0(x) \geq 0\}, \quad \partial \mathcal{V}_0 = \mathcal{M}_0 = \{x : h_0(x) = 0\}.$$

By the generalized Stokes Theorem,

$$\begin{aligned} \frac{d}{dt} \int_{\mathcal{V}_t} w(x) dx \Big|_{t=0} &= \int_{\partial \mathcal{V}_0} w(x) (\dot{\mathbf{X}}(x, 0) \cdot \mathbf{n}(x, 0)) d\mathcal{H}^{d-1}(x) \\ &= \int w(x) (\dot{\mathbf{X}}(x, 0) \cdot \mathbf{n}(x, 0)) d\mathcal{H}^{d-1}(x) \end{aligned}$$

where

$$\dot{\mathbf{X}}(x, 0) := \frac{\partial}{\partial t} \mathbf{X}(x, t) \Big|_{t=0}, \quad \mathbf{n}(x, 0) := -\frac{\nabla h_0(x)}{\|\nabla h_0(x)\|}$$

with $\mathbf{n}(x, t)$ denoting the outward-pointing unit normal at $x \in \mathcal{M}_t$, and $\mathbf{X}(x, t)$ being a diffeomorphism from \mathcal{M}_0 to \mathcal{M}_t , which by definition satisfies $\mathbf{X}(x, 0) \equiv x$ and

$$h_0(\mathbf{X}(x, t)) + tv(\mathbf{X}(x, t)) = 0, \tag{30}$$

for each $x \in \mathcal{M}_0$ and $t \in [0, \epsilon]$ for some $\epsilon > 0$.

Taking derivatives of (30) with respect to t yields

$$\nabla_x h_0(\mathbf{X}(x, t)) \cdot \dot{\mathbf{X}}(x, t) + t \nabla_x v(\mathbf{X}(x, t)) \cdot \dot{\mathbf{X}}(x, t) + v(\mathbf{X}(x, t)) = 0$$

and thus, evaluating the above at $t = 0$, we have

$$\nabla_x h_0(\mathbf{X}(x, 0)) \cdot \dot{\mathbf{X}}(x, 0) + v(\mathbf{X}(x, 0)) = 0$$

or equivalently,

$$\nabla_x h_0(x) \cdot \dot{\mathbf{X}}(x, 0) = -v(x), \quad (31)$$

Hence,

$$\dot{\mathbf{X}}(x, 0) \cdot \mathbf{n}(x, 0) = -\dot{\mathbf{X}}(x, 0) \cdot \frac{\nabla h_0(x)}{\|\nabla h_0(x)\|} = \frac{v(x)}{\|\nabla_x h_0(x)\|}$$

and thus

$$\left. \frac{d}{dt} \int_{\mathcal{V}_t} w(x) dx \right|_{t=0} = \int_{\mathcal{M}_0} w(x) \frac{v(x)}{\|\nabla_x h_0(x)\|} d\mathcal{H}^{d-1}(x).$$

□

Lemma on the Pathwise Derivative of Level Set Integrals

Lemma 3. *Given h, v , define $h_t(x) := h(x) + tv(x)$ and*

$$I(t) := \int_{\{h_t(x)=0\}} w(x) d\mathcal{H}^{d-1}(x) \quad (32)$$

Then $I(t)$ is continuously differentiable in t with

$$I'(0) = - \int_{\mathcal{M}_0} \left[\frac{\nabla_x h_0(x)}{\|\nabla_x h_0(x)\|} \cdot \nabla_x \left(w(x) \frac{v(x)}{\|\nabla_x h_0(x)\|} \right) + w(x) \frac{v(x)}{\|\nabla_x h_0(x)\|} \operatorname{div} \left(\frac{\nabla_x h_0(x)}{\|\nabla_x h_0(x)\|} \right) \right] d\mathcal{H}^{d-1}(x) \quad (33)$$

$$= - \int_{\mathcal{M}_0} \left[\frac{v(x)}{\|\nabla_x h_0(x)\|} \frac{\nabla_x h_0(x)}{\|\nabla_x h_0(x)\|} \cdot \nabla_x w(x) + w(x) \frac{\nabla_x h_0(x)}{\|\nabla_x h_0(x)\|} \cdot \frac{\nabla_x v(x)}{\|\nabla_x h_0(x)\|} + w(x) \frac{v(x)}{\|\nabla_x h_0(x)\|} \operatorname{div} \left(\frac{\nabla_x h_0(x)}{\|\nabla_x h_0(x)\|} \right) \right] d\mathcal{H}^{d-1}(x) \quad (34)$$

$$= - \int_{\mathcal{M}_0} \operatorname{div} \left(w(x) \frac{v(x)}{\|\nabla_x h_0(x)\|} \frac{\nabla_x h_0(x)}{\|\nabla_x h_0(x)\|} \right) d\mathcal{H}^{d-1}(x) \quad (35)$$

and

$$\|I'(t)\| \leq M(\|\nabla w\|_\infty \|v\|_\infty + \|w\|_\infty \|\nabla_x v\|_\infty + \|w\|_\infty \|v\|_\infty). \quad (36)$$

In (33) we may write

$$H(x) := \operatorname{div} \left(\frac{\nabla_x h_0(x)}{\|\nabla_x h_0(x)\|} \right),$$

for the *mean curvature* of \mathcal{M}_0 at x (also known as the trace of the Weingarten map, a.k.a., shape operator). Hence the formula (33) reflects the effect of the curvature of the level set on functional derivative, which is as expected.

In addition, the integrand in (33) features three key elements:

- $w(x)$, which is the original function being integrated over at x .

- $\frac{h(x)-h_0(x)}{\|\nabla_x h_0(x)\|}$ reflects the normalized intensity of change in the direction $h(x) - h_0(x)$.
- $\frac{\nabla_x h_0(x)}{\|\nabla_x h_0(x)\|}$ is the unit normal at $x \in \mathcal{M}_0$.

Given that “divergence” has the intuitive interpretation of “outgoingness”, the formula also seems intuitive: (33) captures the “outward flow” of w from \mathcal{M}_0 as the level set function h_0 is perturbed in the direction of $h - h_0$.

Proof of Lemma 3. We work with a diffeomorphism $\mathbf{X}(\cdot, t)$ that maps $\mathcal{M}_0 = \{x : h_0(x) = 0\}$ to $\mathcal{M}_t = \{x : h_0(x) + t(h(x) - h_0(x)) = 0\}$. Then

$$I(t) := \int_{\mathcal{M}_t} w(x) d\mathcal{H}^{d-1}(x) = \int_{\mathcal{M}_0} w(\mathbf{X}(x, t)) J_{\mathbf{X}}(x, t) d\mathcal{H}^{d-1}(x)$$

where $J_{\mathbf{X}}(x, t)$ denotes the Jacobian of $\mathbf{X}(\cdot, t)$. Then,

$$\begin{aligned} I'(t) &= \int_{\mathcal{M}_0} \frac{d}{dt} [w(\mathbf{X}(x, t)) J_{\mathbf{X}}(x, t)] d\mathcal{H}^{d-1}(x) \\ &= \int_{\mathcal{M}_0} \left[\nabla_x w(\mathbf{X}(x, t)) \cdot \dot{\mathbf{X}}(x, t) J_{\mathbf{X}}(x, t) + w(\mathbf{X}(x, t)) \dot{J}_{\mathbf{X}}(x, t) \right] d\mathcal{H}^{d-1}(x) \\ &= \int_{\mathcal{M}_0} \left[\nabla_x w(\mathbf{X}(x, t)) \cdot \dot{\mathbf{X}}(x, t) + w(\mathbf{X}(x, t)) \operatorname{div}(\dot{\mathbf{X}}(x, t)) \right] J_{\mathbf{X}}(x, t) d\mathcal{H}^{d-1}(x) \\ &= \int_{\mathcal{M}_0} \operatorname{div} [w(\mathbf{X}(x, t)) \dot{\mathbf{X}}(x, t)] J_{\mathbf{X}}(x, t) d\mathcal{H}^{d-1}(x) \end{aligned} \quad (37)$$

where the last line follows from the property that

$$\dot{J}_{\mathbf{X}}(x, t) = J_{\mathbf{X}}(x, t) \operatorname{div}(\dot{\mathbf{X}}(x, t))$$

with $\operatorname{div}(\mathbf{x}) := \sum_j \frac{\partial x_j}{\partial x_j}$ denoting the divergence operator.

In addition, it is without loss of generality to take \mathbf{X} so that $\dot{\mathbf{X}}(x, t)$ is along the direction (or the opposite direction) of the unit normal $\mathbf{n}(x, t)$. In other words,

$$\dot{\mathbf{X}}(x, t) = -\bar{u}(x, t) \frac{\nabla_x h_t(x)}{\|\nabla_x h_t(x)\|}$$

for some scalar-valued function $\bar{u}(x, t)$. Plugging the above into (31), we have

$$-v(x) = -\nabla_x h_t(x) \cdot \bar{u}(x, 0) \frac{\nabla_x h_t(x)}{\|\nabla_x h_t(x)\|} = -\bar{u}(x, t) \frac{\|\nabla_x h_t(x)\|^2}{\|\nabla_x h_t(x)\|}$$

and thus

$$\bar{u}(x, t) = \frac{v(x)}{\|\nabla_x h_t(x)\|}.$$

and

$$\dot{\mathbf{X}}(x, t) = -\frac{v(x)}{\|\nabla_x h_t(x)\|} \frac{\nabla_x h_t(x)}{\|\nabla_x h_t(x)\|}$$

Plugging this into (37) and evaluating at $t = 0$, we have

$$\begin{aligned}
I'(0) &= - \int_{\mathcal{M}_0} \left[\nabla_x w(x) \cdot \frac{v(x)}{\|\nabla_x h_0(x)\|} \frac{\nabla_x h_0(x)}{\|\nabla_x h_0(x)\|} + w(x) \operatorname{div} \left(\frac{v(x)}{\|\nabla_x h_0(x)\|} \frac{\nabla_x h_0(x)}{\|\nabla_x h_0(x)\|} \right) \right] d\mathcal{H}^{d-1}(x) \\
&= - \int_{\mathcal{M}_0} \left[\begin{aligned} &\nabla_x w(x) \cdot \frac{v(x)}{\|\nabla_x h_0(x)\|} \frac{\nabla_x h_0(x)}{\|\nabla_x h_0(x)\|} \\ &+ w(x) \cdot \frac{v(x)}{\|\nabla_x h_0(x)\|} \operatorname{div} \left(\frac{\nabla_x h_0(x)}{\|\nabla_x h_0(x)\|} \right) \\ &+ w(x) \frac{\nabla_x h_0(x)}{\|\nabla_x h_0(x)\|} \cdot \nabla_x \left(\frac{v(x)}{\|\nabla_x h_0(x)\|} \right) \end{aligned} \right] d\mathcal{H}^{d-1}(x) \\
&= - \int_{\mathcal{M}_0} \left[\frac{\nabla_x h_0(x)}{\|\nabla_x h_0(x)\|} \cdot \nabla_x \left(w(x) \frac{v(x)}{\|\nabla_x h_0(x)\|} \right) + w(x) \frac{v(x)}{\|\nabla_x h_0(x)\|} \operatorname{div} \left(\frac{\nabla_x h_0(x)}{\|\nabla_x h_0(x)\|} \right) \right] d\mathcal{H}^{d-1}(x) \\
&= - \int_{\mathcal{M}_0} \operatorname{div} \left(w(x) \frac{v(x)}{\|\nabla_x h_0(x)\|} \frac{\nabla_x h_0(x)}{\|\nabla_x h_0(x)\|} \right) d\mathcal{H}^{d-1}(x)
\end{aligned}$$

where equalities above use the product rule for vector calculus and the product rule for divergence:

$$\operatorname{div}(w(x) \nabla_x h(x)) = w(x) \operatorname{div}(\nabla_x h(x)) + \nabla_x h(x) \cdot \nabla_x w(x).$$

For $t > 0$,

$$\begin{aligned}
|I'(t)| &\leq \left| \int_{\mathcal{M}_0} \nabla_x w(\mathbf{X}(x, t)) \cdot \dot{\mathbf{X}}(x, t) J_{\mathbf{X}}(x, t) d\mathcal{H}^{d-1}(x) \right| \\
&\quad + \left| \int_{\mathcal{M}_0} w(\mathbf{X}(x, t)) \operatorname{div}(\dot{\mathbf{X}}(x, t)) J_{\mathbf{X}}(x, t) d\mathcal{H}^{d-1}(x) \right| \\
&\leq \|\nabla_x w\|_{\infty} \|\dot{\mathbf{X}}\|_{\infty} + \|w\|_{\infty} \|\operatorname{div}(\dot{\mathbf{X}})\|_{\infty}
\end{aligned}$$

with

$$\|\dot{\mathbf{X}}\|_{\infty} = \left\| \frac{v(x)}{\|\nabla_x h_t(x)\|} \frac{\nabla_x h_t(x)}{\|\nabla_x h_t(x)\|} \right\|_{\infty} \leq M \|v\|_{\infty}$$

and

$$\begin{aligned}
\|\operatorname{div}(\dot{\mathbf{X}})\|_{\infty} &= \left\| \operatorname{div} \left(\frac{v(\mathbf{X})}{\|\nabla_x h_t(\mathbf{X})\|} \frac{\nabla_x h_t(\mathbf{X})}{\|\nabla_x h_t(\mathbf{X})\|} \right) \right\| \\
&= \left\| v(\mathbf{X}) \operatorname{div} \left(\frac{\nabla_x h_t(\mathbf{X})}{\|\nabla_x h_t(\mathbf{X})\|^2} \right) + \frac{\nabla_x h_t(\mathbf{X})}{\|\nabla_x h_t(\mathbf{X})\|^2} \cdot \nabla_x v(\mathbf{X}) \right\| \\
&\leq M (\|v\|_{\infty} + \|\nabla_x v\|_{\infty})
\end{aligned}$$

Hence,

$$|I'(t)| \leq M (\|\nabla_x w\|_{\infty} \|v\|_{\infty} + \|w\|_{\infty} \|\nabla_x v\|_{\infty} + \|w\|_{\infty} \|v\|_{\infty})$$

□

Proof of Proposition 2(b). Note that

$$D_h \Gamma(h_0 + tv_2)[v_1] = \int_{\{h(x) + tv_2(x) = 0\}} v_1(x) \frac{w(x)}{\|\nabla_x h_0(x)\|} d\mathcal{H}^{d-1}(x)$$

is a level set integral of the form (32). Applying Lemma 3, we obtain

$$\begin{aligned} & |\Gamma(h) - \Gamma(h_0) - D_h \Gamma(h_0)[h - h_0]| \\ & \leq M \|\hat{h} - h_0\|_\infty (\|\nabla_x \hat{h} - \nabla_x h_0\|_\infty + \|\hat{h} - h_0\|_\infty) \\ & \leq \|\hat{h} - h_0\|_\infty \|\nabla_x \hat{h} - \nabla_x h_0\|_\infty, \end{aligned}$$

which is asymptotically negligible under condition (23). \square

D Asymptotics with Nadaraya-Watson First Stage

In this section, we show that the key rate-acceleration result in the previous subsection, as one may expect, also holds for kernel-based nonparametric methods. To illustrate this point, we consider the case where h_0 is given by a conditional expectation and \hat{h} is given by the Nadaraya-Watson kernel estimator and show that our key result continues to hold.

Formally, let $h_0(x) = \mathbb{E}[Y_i | X_i = x]$ and

$$\hat{h}(x) = \frac{\frac{1}{nb_n^d} \sum_{i=1}^n K_d\left(\frac{X_i - x}{b_n}\right) Y_i}{\frac{1}{nb_n^d} \sum_{i=1}^n K_d\left(\frac{X_i - x}{b_n}\right)}$$

where $b_n \searrow 0$ is a bandwidth parameter and K_d is a multivariate kernel of smoothness order s .

We estimate $\theta_0 = \int_{\mathcal{M}} h_0(x) w(x) p_0(x) d\mathcal{H}^m(x)$ by $\hat{\theta} := \int_{\mathcal{M}} \hat{h}(x) w(x) \hat{p}(x) d\mathcal{H}^m(x)$.

Assumption 14 (Kernel Smoothness). K_d is a d -dimensional product kernel

$$K_d(x) = \prod_{j=1}^d K(x_j),$$

where K is a univariate kernel of smoothness order s , i.e., (i) $K(u) = K(-u)$, (ii) $\int K(u) du = 1$, (iii) $|K(u)| \leq M < \infty$, (iv) $\int u^j K(u) du = 0$ for $j = 1, \dots, s-1$, and (v) $\kappa_s := \int x_j^s K(x) dx \in (0, \infty)$.

Theorem 5 (Kernel Nadaraya-Watson First Stage). Let $h_0(\cdot) = \mathbb{E}[Y_i | X_i = \cdot] \in \Lambda^s(\mathcal{X})$. Under Assumptions 1 and 14, we have

$$\text{Bias}(\hat{\theta}) = O(b_n^s), \quad \text{Var}(\hat{\theta}) = O\left(\frac{1}{nb_n^{d-m}}\right)$$

and thus

$$\|\hat{\theta} - \theta_0\| = O_p\left(n^{-\frac{s}{2s+d-m}}\right).$$

Proof of Theorem 5. Write $a(x) := h_0(x) p(x)$, we have

$$\begin{aligned}
\hat{\theta} - \theta_0 &= \int_{\mathcal{M}} [\hat{h}(x) - h_0(x)] w(x) p(x) d\mathcal{H}^m(x) \\
&= \int_{\mathcal{M}} \left[\frac{\hat{a}(x)}{\hat{p}(x)} - \frac{a(x)}{p(x)} \right] w(x) p(x) d\mathcal{H}^m(x) \\
&= \int_{\mathcal{M}} \left[\frac{\hat{a}(x) - a(x)}{p(x)} - \frac{a(x)}{p^2(x)} (\hat{p}(x) - p(x)) \right] w(x) p(x) d\mathcal{H}^m(x) + R_1 \\
&= \int_{\mathcal{M}} [\hat{a}(x) - a(x) - h_0(x) (\hat{p}(x) - p(x))] w(x) d\mathcal{H}^m(x) + R_1 \\
&= \int_{\mathcal{M}} [\hat{a}(x) - h_0(x) \hat{p}(x)] w(x) d\mathcal{H}^m(x) + R_1 \\
&= \int_{\mathcal{M}} \frac{1}{nb_n^d} \sum_{i=1}^n K\left(\frac{x - X_i}{b_n}\right) (Y_i - h_0(x)) w(x) d\mathcal{H}^m(x) + R_1 \\
&= \underbrace{\frac{1}{nb_n^d} \sum_{i=1}^n \int_{\mathcal{M}} K\left(\frac{x - X_i}{b_n}\right) (Y_i - h_0(x)) w(x) d\mathcal{H}^m(x)}_{T_1} + R_1, \tag{38}
\end{aligned}$$

where the remainder term R_1 is asymptotically negligible.

By (26), we may write

$$T_1 = \sum_{j=1}^{\bar{j}} \frac{1}{n} \sum_{i=1}^n T_{1ij}$$

with

$$T_{1ij} := \frac{1}{b_n^d} \int_{\varphi_j(\mathcal{U}_j)} K\left(\frac{\varphi_j(x_{(j,m)}) - X_i}{b_n}\right) (Y_i - h_0(\varphi_j(x_{(j,m)}))) \bar{w}_j(Y_i, \varphi_j(x_{(j,m)})) dx_{(j,m)}$$

where

$$\bar{w}_j(Y_i, x) := \rho_j(x) w(x) \mathcal{J}\varphi_j(x)$$

Subsequently, we consider each T_{1ij} separately and suppress the subscript j for simpler notation.

We carry out the kernel change of variables from $x_{(m)}$ to u by setting

$$u := \frac{x_{(m)} - X_{i,(m)}}{b_n}, \quad \Leftrightarrow \quad x_{(m)} = X_{i,(m)} + b_n u.$$

We write $\mathcal{V}_u := \frac{\mathcal{V} - X_{i,(m)}}{b_n}$.

$$\begin{aligned}
T_{1ij} &= \frac{1}{b_n^d} \int_{\mathcal{V}_u} K\left(u, \frac{\psi(X_{i,(m)} + b_n u) - X_{i,-(m)}}{b_n}\right) (Y_i - h_0(\varphi(X_{i,(m)} + b_n u))) \bar{w}(Y_i, \varphi(X_{i,(m)} + b_n u)) d(X_{i,(m)} + b_n u) \\
&= \frac{b_n^m}{b_n^d} \int_{\mathcal{V}_u} K_{(m)}(u) K_{-(m)}\left(\frac{\psi(X_{i,(m)} + b_n u) - X_{i,-(m)}}{b_n}\right) (Y_i - h_0(\varphi(X_{i,(m)} + b_n u))) \bar{w}(\varphi(X_{i,(m)} + b_n u)) d(X_{i,(m)} + b_n u)
\end{aligned}$$

$$\begin{aligned}
&= \frac{1}{b_n^{d-m}} \cdot \left[\int_{\mathcal{V}_u} K_{(m)}(u) du \right] \cdot K_{-(m)} \left(\frac{\psi(X_{i,(m)}) - X_{i,-(m)}}{b_n} \right) (Y_i - h_0(\varphi(X_{i,(m)}))) \bar{w}(\varphi(X_{i,(m)})) + o(b_n^{m-d}) \\
&= \underbrace{\frac{1}{b_n^{d-m}} K_{-(m)} \left(\frac{\psi(X_{i,(m)}) - X_{i,-(m)}}{b_n} \right) (Y_i - h_0(\varphi^{-1}(X_{i,(m)}))) \bar{w}(\varphi^{-1}(X_{i,(m)}))}_{T_{2,ij}} + o(b_n^{m-d})
\end{aligned}$$

where $\int_{\mathcal{V}_u} K_{(m)}(u) du = 1 + o(1)$, since $b_n = o(1)$ and V is open, which implies that $[-M, M]^m \subseteq \mathcal{V}_u$ eventually for any $M < \infty$ as $n \rightarrow \infty$.

Recall that $K_{-(m)}$ is a $(d-m)$ -dimensional kernel. Hence,

$$\frac{1}{n} \sum_i T_{2,ij} = \frac{1}{nb_n^{d-m}} \sum_i K_{-(m)} \left(\frac{\psi(X_{i,(m)}) - X_{i,-(m)}}{b_n} \right) (Y_i - h_0(\varphi^{-1}(X_{i,(m)}))) \bar{w}(\varphi^{-1}(X_{i,(m)}))$$

can be heuristically viewed as an $(d-m)$ -dimensional kernel estimator, and thus shares the asymptotic properties (bias, viance and converage rate) of the $(d-m)$ -dimension kernel estimator. We formalize this and complete the proof via Lemmas 4 and 5 below, which show that the bias and variance rates indeed coincide with the corresponding rates in $(d-m)$ -dimensional kernel regressions. \square

Lemma 4 (Bias). $\mathbb{E}[T_{2,ij}] = O(b_n^s)$.

Proof. Clearly,

$$\begin{aligned}
\mathbb{E}[T_{2ij}] &= \frac{1}{b_n^{d-m}} \int \left[K_{-(m)} \left(\frac{\psi(z_{(m)}) - z_{-(m)}}{b_n} \right) (Y_i - h_0(\varphi(z_{(m)}))) \bar{w}(\varphi(z_{(m)})) \right] p(z) dz \\
&= \frac{1}{b_n^{d-m}} \int \left[K_{-(m)} \left(\frac{\psi(z_{(m)}) - z_{-(m)}}{b_n} \right) (h_0(z) - h_0(\varphi(z_{(m)}))) \bar{w}(\varphi(z_{(m)})) \right] p(z) dz \\
&= \frac{1}{b_n^{d-m}} \int \int \left[K_{-(m)} \left(\frac{\psi(z_{(m)}) - z_{-(m)}}{b_n} \right) (h_0(z_{(m)}, z_{-(m)}) - h_0(\varphi(z_{(m)}))) \bar{w}(\varphi(z_{(m)})) \right] p(z_{-(m)}) dz_{-(m)}
\end{aligned}$$

Define the change of variable from $z_{-(m)}$ to ζ as

$$\zeta := \frac{z_{-(m)} - \psi_j(x_{(m)})}{b_n}, \quad z_{-(m)} = \psi_j(x_{(m)}) + b_n \zeta.$$

We have

$$\begin{aligned}
\mathbb{E}[T_{2ij}] &= \frac{1}{b_n^{d-m}} \int \int \left[K_{-(m)}(\zeta) (h_0(z_{(m)}, \psi_j(x_{(m)}) + b_n \zeta) - h_0(\varphi(z_{(m)}))) \right] \\
&\quad p(\psi_j(x_{(m)}) + b_n \zeta_{-(m)} | z_{(m)}) b_n^{d-m} d\zeta p(z_{(m)}) \bar{w}(\varphi(z_{(m)})) dz_{(m)} \\
&= \int \left[\int K_{-(m)}(\zeta) (h_0(z_{(m)}, \psi_j(x_{(m)}) + b_n \zeta) - h_0(\varphi(z_{(m)}))) (\psi_j(x_{(m)}) + b_n \zeta_{-(m)} | z_{(m)}) d\zeta \right] \\
&\quad p(z_{(m)}) \bar{w}(\varphi(z_{(m)})) dz_{(m)}
\end{aligned}$$

$$\begin{aligned}
&= \int \left[\int K_{-(m)}(\zeta) \left(B(z_{(m)}) b_n^s \zeta^s + o(b_n^s) \right) d\zeta \right] p(z_{(m)}) \bar{w}(\varphi(z_{(m)})) dz_{(m)} \\
&= b_n^s \cdot \int K_{-(m)}(\zeta) \zeta^s d\zeta \cdot \int B(z_{(m)}) p(z_{(m)}) \bar{w}(\varphi(z_{(m)})) dz_{(m)} + o(b_n^s) \\
&= O(b_n^s)
\end{aligned}$$

□

Lemma 5 (Variance). $\text{Var}[T_{2,ij}] = O\left(\frac{1}{nb_n^{d-m}}\right)$.

Proof. Then,

$$\begin{aligned}
\mathbb{E}(T_{2ij}^2) &= \frac{1}{b_n^{2(d-m)}} \int K_{-(m)}^2 \left(\frac{\psi_j(z_{(m)}) - z_{-(m)}}{b_n} \right) \left(Y_i - h_0(\varphi^{-1}(z_{(m)})) \right)^2 \bar{w}^2(\varphi^{-1}(z_{(m)})) p(z_{(m)}, z_{-(m)}) dz_{(m)} \\
&= \frac{1}{b_n^{2(d-m)}} \int K_{-(m)}^2 \left(\frac{\psi_j(z_{(m)}) - z_{-(m)}}{b_n} \right) \left[\sigma_0^2(z) + \left(h_0(z) - h_0(\varphi^{-1}(z_{(m)})) \right)^2 \right] \bar{w}^2(\varphi^{-1}(z_{(m)})) p(z_{(m)}, z_{-(m)}) dz_{(m)} \\
&= \frac{1}{b_n^{2(d-m)}} \int \underbrace{\int K_{-(m)}^2 \left(\frac{\psi_j(z_{(m)}) - z_{-(m)}}{b_n} \right) \left[\sigma_0^2(z) + \left(h_0(z) - h_0(\varphi^{-1}(z_{(m)})) \right)^2 \right] p(z_{-(m)} | z_{(m)}) dz_{-(m)}}_{T_{3j}} dz_{(m)}
\end{aligned}$$

For T_{3j} , we can carry out the change of variable from $z_{-(m)}$ to v as below:

$$v := \frac{z_{-(m)} - \psi_j(z_{(m)})}{b_n} \Leftrightarrow z_{-(m)} = \psi_j(z_{(m)}) + b_n v$$

and thus

$$\begin{aligned}
T_{3j} &= \int K_{-(m)}^2(v) \left[\sigma_0^2(z_{(m)}, \psi_j(z_{(m)}) + b_n v) + \left(h_0(z_{(m)}, \psi_j(z_{(m)}) + b_n v) - h_0(\varphi^{-1}(z_{(m)})) \right)^2 \right] p(\psi_j(z_{(m)}) | z_{(m)}) dz_{(m)} \\
&= b_n^{d-m} \int K_{-(m)}^2(v) \left[\sigma_0^2(z_{(m)}, \psi_j(z_{(m)})) + \left(h_0(z_{(m)}, \psi_j(z_{(m)})) - h_0(\varphi^{-1}(z_{(m)})) \right)^2 + O(b_n) \right] [p(\psi_j(z_{(m)}) | z_{(m)}) + o(b_n^{d-m})] dz_{(m)} \\
&= b_n^{d-m} \int K_{-(m)}^2(v) dv \cdot \sigma_0^2(z_{(m)}, \psi_j(z_{(m)})) p(\psi_j(z_{(m)}) | z_{(m)}) + o(b_n^{d-m}) \\
&= b_n^{d-m} \cdot C \cdot \sigma_0^2(z_{(m)}, \psi_j(z_{(m)})) p(\psi_j(z_{(m)}) | z_{(m)}) + o(b_n^{d-m})
\end{aligned}$$

Hence,

$$\begin{aligned}
\frac{1}{b_n^{2(d-m)}} \mathbb{E}(T_{2ij}^2) &= \frac{1}{b_n^{2(d-m)}} \int \left[b_n^{d-m} \cdot C \cdot \sigma_0^2(z_{(m)}, \psi_j(z_{(m)})) p(\psi_j(z_{(m)}) | z_{(m)}) + o(b_n^{d-m}) \right] \bar{w}^2(\varphi^{-1}(z_{(m)})) p(z_{(m)}) dz_{(m)} \\
&= \frac{1}{b_n^{d-m}} C \cdot \int \sigma_0^2(z_{(m)}, \psi_j(z_{(m)})) p(\psi_j(z_{(m)}) | z_{(m)}) \bar{w}^2(\varphi^{-1}(z_{(m)})) p(z_{(m)}) dz_{(m)} + o\left(\frac{1}{b_n^{d-m}}\right) \\
&= O\left(\frac{1}{b_n^{d-m}}\right)
\end{aligned}$$

Hence,

$$\text{Var}(T_1) \leq \frac{C\bar{j}}{n^2} \sum_{i=1}^n O\left(\frac{1}{b_n^{d-m}}\right) + R_3 = O\left(\frac{1}{nb_n^{d-m}}\right).$$

□