Machine Learning is Natural Experiment

Yusuke Narita Kohei Yata*

June 8, 2020

Abstract

Machine learning and other algorithms produce a growing portion of decisions and recommendations. Such algorithmic decisions are natural experiments (conditionally quasi-randomly assigned instruments) since the algorithms make decisions based only on observable input variables. We use this observation to characterize the sources of causal-effect identification for a class of stochastic and deterministic algorithms. Data from almost every algorithm is shown to identify some causal effect. This identification result translates into consistent estimators of causal effects that are easily implemented even with high-dimensional data and complex algorithms.

Preliminary and Not for Circulation

*Narita: Yale University, email: yusuke.narita@yale.edu. Yata: Yale University, email: kohei.yata@yale.edu.

1 Introduction

Today's society increasingly resorts to machine learning ("AI") and other algorithms for decision-making and resource allocation. For example, judges make legal judgements using predictions from supervised machine learning (descriptive regression). Supervised learning is also used by governments to detect potential criminals and terrorists, and financial companies (such as banks and insurance companies) to screen potential customers. Tech companies like Facebook, Microsoft, and Netflix allocate digital content by reinforcement learning and bandit algorithms. Uber and other ride sharing services adjust prices using their surge pricing algorithms to take into account local demand and supply information. Retailers and e-commerce platforms like Amazon engage in algorithmic pricing. Similar algorithms are invading into more and more high-stakes treatment assignment, such as education, health, and military.

All of the above, seemingly diverse examples share a common trait: An algorithm makes decisions based only on observable input variables the data-generating algorithm uses. Conditional on the observable variables, therefore, algorithmic treatment decisions are (quasi-)randomly assigned. This property makes algorithm-based treatment decisions an instrumental variable we can use for measuring the causal effect of the final treatment assignment. The algorithm-based instrument may produce regression-discontinuity-style local variation (e.g. machine judges), stratified randomization (e.g. several bandit and reinforcement leaning algorithms), or mixes of the two.

Based on this observation, this paper first characterizes the whole sources of causal-effect identification (quasi-experimental variation) for a class of algorithms, nesting both stochastic and deterministic ones. This class includes all of the above examples, thus nesting existing insights on quasi-experimental variation from particular algorithms: Surge pricing (Cohen, Hahn, Hall, Levitt and Metcalfe, 2016), bandit (Li, Chu, Langford and Schapire (2010)), reinforcement learning (Precup, 2000), and supervised learning (Cowgill, 2018; Bundorf, Polyakova and Tai-Seale, 2019). Our framework also reveals new sources of identification for algorithms that, at first sight, do not appear to produce natural experiment.

The sources of causal-effect identification turn out to be summarized by a suitable modification of the Propensity Score (Rosenbaum and Rubin, 1983), which we call the Quasi Propensity Score (QPS). The Quasi Propensity Score provides an easy-to-check condition for what causal effects the data from an algorithm allows us to identify.

Based on the identification analysis, we also offer ways of estimating treatment effects using algorithm-based treatment decisions. The treatment effects can be estimated by two-stage least squares (2SLS) or other off-the-shelf IV estimators where we regress the outcome on the treatment with the ML recommendation as an instrument.[1] To make the ML recommendation a conditionally independent instrument, we need to control for appropriate control variables, which we propose to be a linear control for the Quasi Propensity Score. As the main technical result, we prove 2SLS controlling for the Quasi Propensity Score is a consistent estimator of well-defined causal effects (weighted average of conditional Local Average Treatment Effects; Imbens

[1] Recent empirical studies document that algorithmic treatment recommendations have impacts on final treatment assignment by humans (Cowgill, 2018; Bundorf *et al.*, 2019).

and Angrist (1994)). This estimator also induces a multidimensional regression discontinuity design as a special case.

Related Literature. There are heated discussions about whether machine decisions are "better" than human decisions, where "better" is in terms of fairness and efficiency measures (Executive Office of the President, 2016; Hoffman, Kahn and Li, 2017; Horton, 2017; Kleinberg, Lakkaraju, Leskovec, Ludwig and Mullainathan, 2017). In this study, we take a complementary perspective in that our analysis takes a decision algorithm as given, no matter whether it is good or bad, and study the informational value of its produced data.

This paper also relates to the emerging literature on the integration of machine learning, causal inference, and the social sciences. While we are interested in machine learning as a data-*production* tool, the existing literature (except the above mentioned strand) focuses on machine learning as a data-*analysis* tool. For example, a set of predictive studies applies machine learning to make predictions important for social policy questions (Kleinberg *et al.*, 2017; Einav, Finkelstein, Mullainathan and Obermeyer, 2018). Another set of causal and structural work repurposes machine learning to aid with causal inference and structural econometrics (Athey and Imbens, 2017; Belloni, Chernozhukov, Fernández-Val and Hansen, 2017; Bonhomme, Lamadon and Manresa, 2017; Bajari, Nekipelov, Ryan and Yang, 2015). The dual roles of machine learning (data production and data analysis) are complementary to each other.[2]

Methodologically, our framework integrates the classic selection-on-observable scenario with (a multidimensional extension of) the regression discontinuity design, building on the vast literature on the local-average-treatment-effect framework. Important studies on the selection-on-observable scenario include Abadie (2003), Frölich (2007a) and Belloni *et al.* (2017). Hahn, Todd and der Klaauw (2001), Frandsen, Frölich and Melly (2012), Dong (2018) and Arai, Hsu, Kitagawa, Mourifié and Wan (2018), among others, study the local average treatment effect in the regression discontinuity design with a single assignment variable. Our framework also accommodates multidimensional regression discontinuity frameworks. Prior work exploring regression discontinuity designs with multiple running variables and multiple cutoffs includes Papay, Willett and Murnane (2011); Zajonc (2012); Wong, Steiner and Cook (2013); Keele and Titiunik (2015); and Cattaneo, Titiunik, Vazquez-Bare and Keele (2016). This paper also shares some of its spirit with the local random assignment interpretation of regression discontinuity, discussed by Frölich (2007b); Cattaneo, Frandsen and Titiunik (2015); Cattaneo, Titiunik and Vazquez-Bare (2017); Frandsen (2017); Sekhon and Titiunik (2017); Frölich and Huber (2019); and Abdulkadiroğlu, Angrist, Narita and Pathak (2019). These papers study a variety of special cases of this paper's framework.

2 Framework

We are interested in the effect of some binary treatment $D_i \in \{0, 1\}$ on some outcome of interest Y_i. Let $Y_i(1)$ and $Y_i(0)$ represent potential outcomes that would be realized if individual i were

[2]The experimental value of machine learning as algorithmic treatment assignment shares some of its spirit with prior work on mechanism design as another sort of algorithmic treatment assignment (Abdulkadiroğlu, Angrist, Narita and Pathak, 2017; Narita, 2017, 2018).

treated and not treated, respectively. The observed outcome Y_i can therefore be written as $Y_i = D_i Y_i(1) + (1 - D_i)Y_i(0)$.

The treatment assignment D_i may be influenced by a binary treatment recommendation $Z_i \in \{0, 1\}$ made by some machine learning (ML) algorithm. As in Imbens and Angrist (1994), let the binary variable $D_i(z)$ be the potential treatment assignment when $Z_i = z \in \{0, 1\}$. $D_i(z) = 1$ means that individual i is treated when the treatment recommendation were $Z_i = z$. Observed treatment D_i is therefore $D_i = Z_i D_i(1) + (1 - Z_i)D_i(0)$. As is standard in the literature, we impose the exclusion restriction that the treatment recommendation Z_i does not affect the observed outcome other than through the treatment assignment D_i. This allows us to define the potential outcomes indexed against the treatment assignment D_i alone.[3]

We consider algorithms making treatment recommendations based solely on individual i's predetermined, observable covariates $X_i \in \mathbb{R}^p$. Let the function $ML : \mathbb{R}^p \to [0, 1]$ represent the *ML prediction-decision algorithm*, where $ML(X_i) = \Pr(Z_i = 1|X_i)$ is the probability that the treatment recommendation is turned on for individual i with covariates X_i. In typical machine-learning scenarios, ML is the result of applying machine learning on X_i to make some prediction, and then using the prediction to make a final recommendation. The function ML is known to the econometrician. The treatment recommendation Z_i for individual i is then randomly turned on with probability $ML(X_i)$ and independently from everything else.

Assumption 1 (Conditional Independence)**.** $Z_i \perp\!\!\!\perp (Y_i(1), Y_i(0), D_i(1), D_i(0))|X_i$.

Let Y_{zi} defined as $Y_{zi} \equiv D_i(z)Y_i(1) + (1 - D_i(z))Y_i(0)$ for $z \in \{0, 1\}$. Y_{zi} is the outcome that would be realized if the treatment recommendation were $Z_i = z$. Under Assumption 1, $Z_i \perp\!\!\!\perp (Y_{1i}, Y_{0i})|X_i$.

Note that the codomain of ML contains 0 and 1, allowing for deterministic treatment assignments conditional on X_i. Our framework therefore nests the (sharp) regression discontinuity design as a special case.[4] The classic conditional independence scenario with the common support condition ($ML(X_i) \in (0, 1)$ almost surely) is also a special case of our framework. In addition to these simple settings, this framework nests many other situations, such as multidimensional regression discontinuity designs and complex ML algorithms, as illustrated in Section 5.

We put a few assumptions on the covariates X_i and the ML algorithm. To simplify the exposition, the main text assumes that X_i is absolutely continuous with respect to the Lebesgue measure. Appendix B.3 extends the analysis to the case where some covariates in X_i are discrete. Let $\mathcal{X}$ be the support of X_i, $\mathcal{X}_0 = \{x \in \mathcal{X} : ML(x) = 0\}$, $\mathcal{X}_1 = \{x \in \mathcal{X} : ML(x) = 1\}$, and $\mathcal{L}^p$ be the Lebesgue measure on $\mathbb{R}^p$.

[3]Formally, let $Y_i(d, z)$ denote the potential outcome that would be realized if i's treatment assignment and recomendation were d and z, respectively. The exclusion restriction assumes that $Y_i(d, 1) = Y_i(d, 0)$ for $d \in \{0, 1\}$ (Angrist and Pischke, 2008).

[4]Most of the existing studies on regression discontinuity designs define the potential treatment assignment indexed against the running variable like $D_i(x)$, which represents the counterfactual treatment assignment the individual i would have received if her running variable had been set to x. Unlike prior work, we define it indexed against the treatment recommendation z, since X_i is predetermined covariates, and a reasonable thought experiment is what would i's treatment assignment have been if her treatment recommendation had been set differently (holding her covariates fixed).

Assumption 2 (Almost Everywhere Continuity of ML). *(a) ML is continuous almost everywhere with respect to the Lebesgue measure.*

(b) $\mathcal{L}^p(\mathcal{X}_k) = \mathcal{L}^p(\text{int}(\mathcal{X}_k))$ for $k = 0, 1$.

Assumption 2 allows the function ML to be discontinuous on a set of points with the Lebesgue measure zero. For example, ML is allowed to take on a finite number of values as long as it is continuous almost everywhere.

In mechanism designs and other algorithms with capacity constraints, the treatment recommendation for individual i may depend not only on X_i but also on the characteristics of others. These interactive situations can be accommodated by our framework if we consider the following large market setting. Suppose that there is a continuum of individuals $i \in [0, 1]$ and that the recommendation probability for individual i with covariate X_i is determined by a function M as follows:

$$\Pr(Z_i = 1 | X_i; F_{X_{-i}}) = M(X_i; F_{X_{-i}}).$$

Here $F_{X_{-i}} = \Pr(\{j \in [0, 1] \setminus \{i\} : X_j \leq x\})$ is the distribution of X among all individuals $j \in [0, 1] \setminus \{i\}$. The function $M : \mathbb{R}^p \times \mathcal{F} \to [0, 1]$, where $\mathcal{F}$ is a set of distributions on $\mathbb{R}^p$, gives the recommendation probability for each individual in the market. With a continuum of individuals, for any $i \in [0, 1]$, $F_{X_{-i}}$ is the same as the distribution of X in the whole market, denoted by F_X. Therefore, the data generated by the mechanism M are equivalent to the data generated by the algorithm $ML : \mathbb{R}^p \to [0, 1]$ such that $ML(x) \equiv M(x; F_X)$ for all $x \in \mathbb{R}^p$. Our framework is applicable to this interactive setting.

3 Identification

What causal effects can we learn from data (X_i, Z_i, D_i, Y_i) generated by the ML algorithm? A key step toward answering this question is what we call the *Quasi Propensity Score* (QPS). To define it, let:

$$p^{ML}(x; \delta) \equiv \frac{\int_{\mathcal{X} \cap N(x,\delta)} ML(x^*) dx^*}{\int_{\mathcal{X} \cap N(x,\delta)} dx^*},$$

where $N(x, \delta) = \{x^* \in \mathbb{R}^p : \|x - x^*\| \leq \delta\}$ is the $\delta-$ball around $x \in \mathcal{X}$. Here, $\|\cdot\|$ denotes the Euclidean distance on $\mathbb{R}^p$. We assume that ML is a $\mathcal{L}^p$-measurable function so that the integrals exist. We then define QPS as follows:

$$p^{ML}(x) \equiv \lim_{\delta \to 0} p^{ML}(x; \delta).$$

Intuitively, QPS at x is the average probability of a treatment recommendation in a shrinking neighborhood of x.[5] To make common δ for all dimensions reasonable, we normalize X_{ij} to have mean zero and variance one for each $j = 1, ..., p$.[6]

[5]The idea behind QPS shares some of its spirit with the local randomization interpretation of regression discontinuity designs (Frölich, 2007b; Cattaneo *et al.*, 2015, 2017): the treatment assignment is considered as

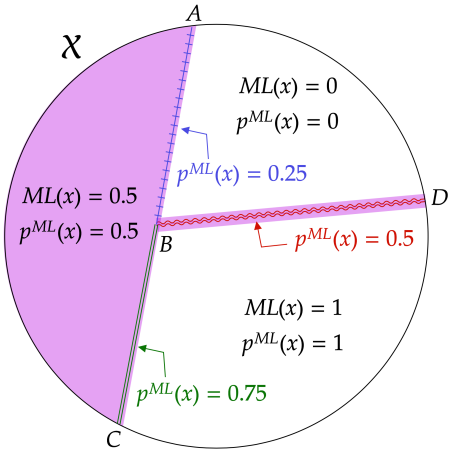

Figure 1: Example of the Quasi Propensity Score

Figure 1 illustrates QPS. In the example, X_i is two dimensional, and the support of X_i is divided into three sets depending on the value of ML. For the interior points of each set, QPS is equal to ML (as formally implied by Part 2 of Corollary 2 below). On the border of any two sets, QPS is the average of the ML values in the two sets. Thus, $p^{ML}(x) = \frac{1}{2}(0 + 0.5) = 0.25$ for any x in the open line segment AB, $p^{ML}(x) = \frac{1}{2}(0.5 + 1) = 0.75$ for any x in the open line segment BC, and $p^{ML}(x) = \frac{1}{2}(0 + 1) = 0.5$ for any x in the open line segment BD.

Our identification analysis uses the following continuity condition.

Assumption 3 (Local Mean Continuity)**.** *For $z \in \{0, 1\}$, the conditional expectation functions $E[Y_{zi}|X_i]$ and $E[D_i(z)|X_i]$ are continuous at any point $x \in \mathcal{X}$ such that $p^{ML}(x) \in (0, 1)$ and $ML(x) \in \{0, 1\}$.*

$ML(x) \in \{0, 1\}$ means that the treatment recommendation Z_i is deterministic. If QPS at the point x is nondegenerate ($p^{ML}(x) \in (0, 1)$), however, there exists a point close to x that has a different value of ML from x's, which creates variation in the treatment recommendation near x. For any such point x, Assumption 3 requires that the points close to x have similar conditional means of the outcome Y_{zi} and treatment assignment $D_i(z)$ for each possible treatment recommendation $z \in \{0, 1\}$.

In the context of the regression discontinuity design with a single running variable, the point x for which $p^{ML}(x) \in (0, 1)$ and $ML(x) \in \{0, 1\}$ is the cutoff point at which the treatment probability discontinuously changes. In this special case, one sufficient condition for continuity of $E[Y_{zi}|X_i]$ is a local independence condition in the spirit of Hahn *et al.* (2001):

---

good as randomly assigned in a neighborhood of the cutoff. However, we do not rely on a local-randomization-type assumption like the one proposed by Cattaneo *et al.* (2015) for identification and estimation.

[6]This normalization is without loss of generality in the following sense. Take a vector X_i^* of any continuous random variables and $ML^* : \mathbb{R}^p \to [0, 1]$. They induce the associated normalized random vector $X_i = A(X_i^* - E[X_i^*])$, where A is a diagonal matrix with diagonal entries $\frac{1}{\mathrm{Var}(X_{i1}^*)^{1/2}}, ..., \frac{1}{\mathrm{Var}(X_{ip}^*)^{1/2}}$. Let $ML(x) = ML^*(A^{-1}x + E[X_i^*])$. Then (X_i^*, ML^*) is equivalent to (X_i, ML) in the sense that $ML(X_i) = ML^*(X_i^*)$ for any individual i.

$(Y_i(1), Y_i(0), D_i(1), D_i(0))$ is independent of X_i near x. This is a strong condition, since it does not allow the distribution of potential outcomes to depend on X_i near x. A weaker sufficient condition, which allows such dependence, is that $E[Y_i(d)|D_i(1) = d_1, D_i(0) = d_0, X_i]$ and $\Pr(D_i(1) = d_1, D_i(0) = d_0|X_i)$ are continuous at x for every $d \in \{0,1\}$ and $(d_1, d_0) \in \{0,1\}^2$ (Dong, 2018). This assumes that the conditional means of the potential outcomes for each of the four types determined based on the potential treatment assignment $D_i(z)$ and the conditional probabilities of those types are continuous at the cutoff.[7] It is easily seen that these two sets of conditions are sufficient for continuity of $E[Y_{zi}|X_i]$ regardless of the dimension of X_i, accommodating multidimensional regression discontinuity designs.

We say that a causal effect is *identified* if it is uniquely determined by the joint distribution of (Y_i, X_i, D_i, Z_i). QPS provides an easy-to-check condition for whether an algorithm allows us to identify causal effects.

Proposition 1 (Identification)**.** *Under Assumptions 1 – 3:*

(a) $E[Y_{1i} - Y_{0i}|X_i = x]$ *and* $E[D_i(1) - D_i(0)|X_i = x]$ *are identified for every* $x \in \mathcal{X}$ *such that* $p^{ML}(x) \in (0,1)$.

(b) Let A *be any open subset of* $\mathcal{X}$ *such that* $p^{ML}(x)$ *exists for all* $x \in A$*. Then either* $E[Y_{1i} - Y_{0i}|X_i \in A]$ *or* $E[D_i(1) - D_i(0)|X_i \in A]$*, or both are identified only if* $p^{ML}(x) \in (0,1)$ *for almost every* $x \in A$ *(with respect to the Lebesgue measure).*[8]

Proposition 1 characterizes a necessary and sufficient condition for identification. Part (a) says that the average effects of the treatment recommendation Z_i on the outcome Y_i and on the treatment assignment D_i for the individuals with $X_i = x$ are both identified if QPS at x is neither 0 nor 1. Non-degeneracy of QPS at x implies that there are both types of individuals who receive $Z_i = 1$ and $Z_i = 0$ among those whose X_i is close to x. Assumption 3 ensures that those individuals are similar in terms of average potential outcomes and treatment assignments. We can therefore identify the average effects conditional on $X_i = x$. In Figure 1, $p^{ML}(x) \in (0,1)$ holds for any x in the shaded region (the union of the minor circular segment made by the chord AC and the line segment BD).

A consequence of Part (a) is that it is possible to identify $\int_{\{x^* \in \mathcal{X}: p^{ML}(x^*) \in (0,1)\}} \omega(x) E[Y_{1i} - Y_{0i}|X_i = x] d\mu(x)$ and $\int_{\{x^* \in \mathcal{X}: p^{ML}(x^*) \in (0,1)\}} \omega(x) E[D_i(1) - D_i(0)|X_i = x] d\mu(x)$ for any known or identified function $\omega : \mathbb{R}^p \to \mathbb{R}$ and any measure μ provided that the integrals exist.

Part (a) nests two well-known identification results as special cases. First, suppose that $ML(x) \in (0,1)$ for every $x \in \mathcal{X}$. This corresponds to the classic conditional independence setting (or stratified randomization setting) with nondegenerate assignment probability, in which conditional average causal effects are identified (see for example Angrist and Pischke (2008)). Second, suppose that $ML(x) \in \{0,1\}$ for all $x \in \mathcal{X}$ but the value of ML discontinuously changes at some point x^* so that $p^{ML}(x^*) \in (0,1)$. This case corresponds to a regression discontinuity

[7] Frandsen *et al.* (2012) and Arai *et al.* (2018) consider similar assumptions. They impose continuity of the distributions of the potential outcomes rather than continuity of the conditional means.

[8] We assume that p^{ML} is a $\mathcal{L}^p$-measurable function so that $\{x \in A : p^{ML}(x) = 0\}$ and $\{x \in A : p^{ML}(x) = 1\}$ are $\mathcal{L}^p$-measurable.

design, in which the average causal effect at a boundary point is identified under continuity of conditional expectation functions of potential outcomes (Hahn *et al.*, 2001; Keele and Titiunik, 2015).

Part (b) provides a necessary condition for identification. It says that if the average effect of the treatment recommendation conditional on X_i being in some open set A is identified, then we must have $p^{ML}(x) \in (0,1)$ for almost every $x \in A$. If, to the contrary, there is a positive mass with $p^{ML}(x) = 1$ (or $p^{ML}(x) = 0$) inside A, then Z_i has no variation in the mass, which makes it impossible to identify the average effect for the mass.

Proposition 1 concerns causal effects of treatment *recommendation*, not of treatment *assignment*. The proposition implies that the conditional average treatment effects and the conditional local average treatment effects (LATEs) are identified under additional assumptions.[9]

Corollary 1 (Perfect and Imperfect Compliance)**.** *Under Assumptions 1 – 3:*

(a) The average treatment effect conditional on $X_i = x$, $E[Y_i(1) - Y_i(0)|X_i = x]$, is identified for every $x \in \mathcal{X}$ such that $p^{ML}(x) \in (0,1)$ and $\Pr(D_i(1) > D_i(0)|X_i = x) = 1$ (perfect compliance).

(b) Let A be any open subset of $\mathcal{X}$ such that $p^{ML}(x)$ exists for all $x \in A$, and $\Pr(D_i(1) > D_i(0)|X_i \in A) = 1$. Then $E[Y_i(1) - Y_i(0)|X_i \in A]$ is identified only if $p^{ML}(x) \in (0,1)$ for almost every $x \in A$.

(c) The local average treatment effect conditional on $X_i = x$, $E[Y_i(1) - Y_i(0)|D_i(1) \neq D_i(0), X_i = x]$, is identified for every $x \in \mathcal{X}$ such that $p^{ML}(x) \in (0,1)$, $\Pr(D_i(1) \geq D_i(0)|X_i = x) = 1$ (monotonicity) and $\Pr(D_i(1) \neq D_i(0)|X_i = x) > 0$ (existence of compliers).

(d) Let A be any open subset of $\mathcal{X}$ such that $p^{ML}(x)$ exists for all $x \in A$, $\Pr(D_i(1) \geq D_i(0)|X_i \in A) = 1$, and $\Pr(D_i(1) \neq D_i(0)|X_i \in A) > 0$. Then $E[Y_i(1) - Y_i(0)|D_i(1) \neq D_i(0), X_i \in A]$ is identified only if $p^{ML}(x) \in (0,1)$ for almost every $x \in A$.

Non-degeneracy of QPS $p^{ML}(x)$ therefore summarizes what causal effects the data from ML identify. Since the key condition ($p^{ML}(x) \in (0,1)$) holds for some points x for almost every algorithm, the data from almost every algorithm identify some causal effect.

3.1 Existence of the Quasi Propensity Score

The above results assume that QPS exists, but is it fair to assume so? In general, QPS may fail to exist; we provide such an example in Appendix Section B.1. Nevertheless, it does for almost every covariate point.

Proposition 2. *Suppose that $\text{int}(\mathcal{X}) \neq \emptyset$. Then $p^{ML}(x)$ exists and is equal to $ML(x)$ for almost every $x \in \text{int}(\mathcal{X})$ (with respect to the Lebesgue measure).*

[9] If ML is a deterministic algorithm and x is a boundary point, the conditions that $\Pr(D_i(1) \geq D_i(0)|X_i = x) = 1$ and that $\Pr(D_i(1) \neq D_i(0)|X_i = x) > 0$ in Part (c) of Corollary 1 imply that the conditional treatment assignment probability $\Pr(D_i = 1|X_i)$ changes discontinuously at x as in regression discontinuity designs. If $\Pr(D_i = 1|X_i)$ does not jump but has a kink at x, it is still possible to identify the conditional LATE under additional smoothness assumptions in the spirit of Dong (2012).

How do we know whether QPS at a particular point exists? If it exists, what is the value of QPS at the point? To answer these questions, for each $x \in \mathcal{X}$ and each $q \in \text{Supp}(ML(X_i))$, define

$$\mathcal{U}_{x,q} \equiv \{u \in N(\mathbf{0}, 1) : \lim_{\delta \to 0} ML(x + \delta u) = q\},$$

where $\mathbf{0} \in \mathbb{R}^p$ is a vector of zeros. $\mathcal{U}_{x,q}$ is the set of vectors in $N(\mathbf{0}, 1)$ such that the value of ML approaches q as we approach x from the direction of the vector. With this notation, we obtain a sufficient condition for the existence of QPS at a point x and its characterization.

Proposition 3. *Take any $x \in \text{int}(\mathcal{X})$. If there exists a countable set $Q \subset \text{Supp}(ML(X_i))$ such that $\mathcal{L}^p(\cup_{q \in Q} \mathcal{U}_{x,q}) = \mathcal{L}^p(N(\mathbf{0}, 1))$ and $\mathcal{U}_{x,q}$ is $\mathcal{L}^p$-measurable for all $q \in Q$, then $p^{ML}(x)$ exists and is given by*

$$p^{ML}(x) = \frac{\sum_{q \in Q} q \mathcal{L}^p(\mathcal{U}_{x,q})}{\mathcal{L}^p(N(\mathbf{0}, 1))}.$$

If almost every point in $N(\mathbf{0}, 1)$ is contained by one of countably many $\mathcal{U}_{x,q}$'s, therefore, QPS exists and is equal to the weighted average of the values of q with the weight proportional to the hypervolume of $\mathcal{U}_{x,q}$. This result implies that QPS exists in practically important cases.

Corollary 2.

1. *If ML is continuous at $x \in \text{int}(\mathcal{X})$, then $p^{ML}(x)$ exists and $p^{ML}(x) = ML(x)$.*
2. (Interior points) *Let $\mathcal{X}_q = \{x \in \mathcal{X} : ML(x) = q\}$ for some $q \in [0, 1]$. Then, for any interior point $x \in \text{int}(\mathcal{X}_q)$, $p^{ML}(x)$ exists and $p^{ML}(x) = q$.*
3. (Smooth boundary points) *Suppose that $\{x \in \mathcal{X} : ML(x) = q_1\} = \{x \in \mathcal{X} : f(x) \geq 0\}$ and $\{x \in \mathcal{X} : ML(x) = q_2\} = \{x \in \mathcal{X} : f(x) < 0\}$ for some $q_1, q_2 \in [0, 1]$, where $f : \mathbb{R}^p \to \mathbb{R}$. Let x be a boundary point in $\text{int}(\mathcal{X})$ such that $f(x) = 0$, f is continuously differentiable in a neighborhood of x, and $\frac{\partial f(x)}{\partial x} \neq 0$. In this case, $p^{ML}(x)$ exists and $p^{ML}(x) = \frac{1}{2}(q_1 + q_2)$.*
4. (Intersection points under CART and random forest) *Let $p = 2$, and suppose that $\{x \in \mathcal{X} : ML(x) = q_1\} = \{(x_1, x_2)' \in \mathcal{X} : x_1 \leq 0 \text{ or } x_2 \leq 0\}$, $\{x \in \mathcal{X} : ML(x) = q_2\} = \{(x_1, x_2)' \in \mathcal{X} : x_1 > 0, x_2 > 0\}$, and $\mathbf{0} = (0, 0)' \in \text{int}(\mathcal{X})$. This is an example in which tree-based algorithms such as CART and random forests are used to create ML. In this case, $p^{ML}(\mathbf{0})$ exists and $p^{ML}(\mathbf{0}) = \frac{3}{4}q_1 + \frac{1}{4}q_2$.*

4 Estimation

The sources of quasi-random assignment characterized in Proposition 1 suggest a way of estimating causal effects of the treatment. In particular, Proposition 1 suggests that conditioning on QPS makes ML-based treatment recommendation quasi randomly assigned. This motivates us to use ML recommendation as an instrument conditional on QPS, which we operationalize as follows.

4.1 Two-Stage Least Squares Meets QPS

Suppose that we observe a random sample $\{(Y_i, X_i, D_i, Z_i)\}_{i=1}^n$ of size n from the population whose data generating process is described in Section 2. Let I be a dummy random variable which is switched on if there exists a constant $q \in (0,1)$ such that $ML(X_i) \in \{0, q, 1\}$ for all $i \in \{1, ..., n\}$. I is the indicator that $ML(X_i)$ takes on only one nondegenerate value *in the sample*. Consider the following 2SLS regression using the observations with $p^{ML}(X_i; \delta_n) \in (0,1)$:

$$D_i = \gamma_0(1 - I) + \gamma_1 Z_i + \gamma_2 p^{ML}(X_i; \delta_n) + \nu_i \tag{1}$$

$$Y_i = \beta_0(1 - I) + \beta_1 D_i + \beta_2 p^{ML}(X_i; \delta_n) + \epsilon_i, \tag{2}$$

where bandwidth δ_n shrinks toward zero as the sample size n increases. If the support of $ML(X_i)$ (in the population) contains only one value in $(0,1)$, $p^{ML}(X_i; \delta_n)$ is asymptotically constant conditional on $p^{ML}(X_i; \delta_n) \in (0,1)$. To avoid the multicollinearlity between $p^{ML}(X_i; \delta_n)$ and a constant, we do not include the constant term if $I = 1$ and there is a constant $q \in (0,1)$ such that $ML(X_i) \in \{0, q, 1\}$ for all $i \in \{1, ..., n\}$.

The coefficient γ_1 in (1) is the first-stage effect of ML's treatment recommendation on the final treatment assignment, while the coefficient β_1 in (2) is the causal treatment effect of interest. Let $\hat{\beta}_1$ denote the 2SLS estimator of β_1 in the above regression.

The above regression uses true QPS $p^{ML}(X_i; \delta_n)$, but it may be difficult to analytically compute it if ML is complex. In such a case, we suggest to approximate $p^{ML}(X_i; \delta_n)$ with brute force simulation. We draw a value of x from the uniform distribution on $\mathcal{X} \cap N(X_i, \delta_n)$ a number of times, compute $ML(x)$ for each draw, and take the average of $ML(x)$ over the draws.[10]

Specifically, let $X_1^*, ..., X_{S_n}^*$ be S_n independent draws from the uniform distribution on $\mathcal{X} \cap N(X_i, \delta_n)$, and calculate

$$p^s(X_i; \delta_n) = \frac{1}{S_n} \sum_{s=1}^{S_n} ML(X_s^*).$$

We compute $p^s(X_i; \delta_n)$ for each $i = 1, ..., n$ independently across i so that $p^s(X_1; \delta_n), ..., p^s(X_n; \delta_n)$ are independent of each other. For fixed n and X_i, the approximation error relative to true $p^{ML}(X_i; \delta_n)$ has a $1/\sqrt{S_n}$ rate of convergence.[11] This rate does not depend on the dimension of X_i, so the simulation error can be made negligible even when X_i is high dimensional.

Now consider the following simulation version of the 2SLS regression using the observations with $p^s(X_i; \delta_n) \in (0,1)$:

$$D_i = \gamma_0(1 - I) + \gamma_1 Z_i + \gamma_2 p^s(X_i; \delta_n) + \nu_i \tag{3}$$

$$Y_i = \beta_0(1 - I) + \beta_1 D_i + \beta_2 p^s(X_i; \delta_n) + \epsilon_i. \tag{4}$$

Let $\hat{\beta}_1^s$ denote the 2SLS estimator of β_1 in the simulation-based regression. This regression is the same as the 2SLS regression (1) and (2) except that we use the simulated QPS $p^s(X_i; \delta_n)$ in place of $p^{ML}(X_i; \delta_n)$.

[10] See Appendix B.4 for how to sample from the uniform distribution on a p-dimensional ball.

[11] More precisely, we have $|p^s(X_i; \delta_n) - p^{ML}(X_i; \delta_n)| = O_{p^s}(1/\sqrt{S_n})$, where O_{p^s} indicates the stochastic boundedness in terms of the probability distribution of the S_n simulation draws.

4.2 Consistency

We establish the consistency of the 2SLS estimators $\hat{\beta}_1$ and $\hat{\beta}_1^s$. Let $\Omega^* = \{x \in \mathbb{R}^p : ML(x) = 1\}$ be the set of the covariate points whose ML value is one. We denote by f_X the probability density function of X_i in the population and let $\mathcal{H}^k$ denote the k-dimensional Hausdorff measure on $\mathbb{R}^p$.[12] For a set $A \subset \mathbb{R}^p$, $\text{cl}(A)$ denotes the closure of A and ∂A denotes the boundary of A, i.e., $\partial A = \text{cl}(A) \setminus \text{int}(A)$. We say that a bounded open set $A \subset \mathbb{R}^p$ is *twice continuously differentiable* if for every $x \in A$, there exists a ball $B = N(x, \epsilon)$ and a one-to-one mapping ψ from B onto an open set $D \subset \mathbb{R}^p$ such that ψ and ψ^{-1} are twice continuously differentiable, $\psi(B \cap A) \subset \{(x_1, ..., x_p) \in \mathbb{R}^p : x_p > 0\}$ and $\psi(B \cap \partial A) \subset \{(x_1, ..., x_p) \in \mathbb{R}^p : x_p = 0\}$. For two sets $A, B \subset \mathbb{R}^p$, let $\text{dist}(A, B) = \inf_{x \in A, y \in B} \|x - y\|$ be the distance between A and B. Our consistency result uses the following regularity conditions.

Assumption 4. *(a)* (Finite Second Moments) $E[Y_i(1)^2]$ *and* $E[Y_i(0)^2]$ *are finite.*

(b) (Nonzero First Stage) *There exists a constant* $c > 0$ *such that* $E[D_i(1) - D_i(0)|X_i = x] > c$ *for every* $x \in \mathcal{X}$ *such that* $p^{ML}(x) \in (0, 1)$.

If $\Pr(ML(X_i) \in (0, 1)) = 0$*, then the following conditions (c) – (h) hold.*

(c) (Nonzero Variance) $\text{Var}(ML(X_i)) > 0$.

(d) (C^2 Boundary) *There exists a partition* $\{\Omega_1^*, ..., \Omega_M^*\}$ *of* Ω^* *such that*

(i) $\text{dist}(\Omega_m^*, \Omega_{m'}^*) > 0$ *for any* $m, m' \in \{1, ..., M\}$ *such that* $m \neq m'$*;*

(ii) Ω_m^* *is nonempty, bounded, open, connected and twice continuously differentiable for each* $m \in \{1, ..., M\}$.

(e) ($(p-1)$-dimensional Boundary) $\mathcal{H}^{p-1}(\partial\Omega^*) < \infty$, $\mathcal{H}^{p-1}(\partial\Omega^* \cap \partial\mathcal{X}) = 0$, *and* $\int_{\partial\Omega^* \cap \text{int}(\mathcal{X})} f_X(x) d\mathcal{H}^{p-1}(x) > 0$.

(f) (Local Mean Continuity) *For* $z \in \{0, 1\}$, $E[Y_{zi}|X_i]$ *and* $E[D_i(z)|X_i]$ *are continuous at every point* $x \in \partial\Omega^* \cap \text{int}(\mathcal{X})$.

(g) (Continuous and Bounded Density) f_X *is continuous at every point* $x \in \partial\Omega^* \cap \text{int}(\mathcal{X})$. *In addition, there exists* $\delta > 0$ *such that* f_X *is bounded on* $\bar{N}(\partial\Omega^* \cap \mathcal{X}, \delta)$, *where* $\bar{N}(\partial\Omega^* \cap \mathcal{X}, \delta) = \{x \in \mathbb{R}^p : \|x - y\| \leq \delta \text{ for some } y \in \partial\Omega^* \cap \mathcal{X}\}$.

(h) (Bounded Second Moments) *There exists* $\delta > 0$ *such that* $E[Y_i(1)^2|X_i]$ *and* $E[Y_i(0)^2|X_i]$ *are bounded on* $\bar{N}(\partial\Omega^* \cap \mathcal{X}, \delta)$.

Assumption 4 (b) assumes that, conditional on each value of X_i for which QPS is nondegenerate, more individuals would change their treatment assignment status from 0 to 1 in response

[12] The k-dimensional Hausdorff measure on $\mathbb{R}^p$ is defined as follows. Let Σ be the Lebesgue σ-algebra on $\mathbb{R}^p$ (the set of all Lebesgue measurable sets on $\mathbb{R}^p$). For $A \in \Sigma$ and $\delta > 0$, let $\mathcal{H}_\delta^k(A) = \inf\{\sum_{j=1}^\infty d(E_j)^k : A \subset \cup_{j=1}^\infty E_j, d(E_j) < \delta, E_j \subset \mathbb{R}^p \text{ for all } j\}$, where $d(E) = \sup\{\|x - y\| : x, y \in E\}$. The k-dimensional Hausdorff measure on $\mathbb{R}^p$ is $\mathcal{H}^k(A) = \lim_{\delta \to 0} \mathcal{H}_\delta^k(A)$.

to treatment recommendation than would change it from 1 to 0.[13] Under this assumption, the estimated first-stage coefficient on Z_i converges to a positive quantity. Note that, if there exists $c < 0$ such that $E[D_i(1) - D_i(0)|X_i = x] < c$ for every $x \in \mathcal{X}$ such that $p^{ML}(x) \in (0, 1)$, changing the labels of treatment recommendation makes Assumption 4 (b) hold.

Assumption 4 (c) – (h) are a set of conditions we require when ML is deterministic and produces only multidimensional regression-discontinuity variation. Assumption 4 (c) says that ML produces variation in the treatment recommendation.

Assumption 4 (d) imposes the differentiability of the boundary of Ω^*. The conditions are satisfied if, for example, $\Omega^* = \{x \in \mathbb{R}^p : f(x) \geq 0\}$ for some twice continuously differentiable function $f : \mathbb{R}^p \to \mathbb{R}$ such that $\frac{\partial f(x)}{\partial x} \neq 0$ for all $x \in \mathbb{R}^p$ with $f(x) = 0$. Ω^* takes this form when the conditional treatment effect $E[Y_i(1) - Y_i(0)|X]$ is predicted by supervised learning based on linear models such as lasso and ridge without including higher order terms of X_i, and treatment is assigned to individuals who are estimated to experience nonnegative treatment effects.

In general, the differentiability of Ω^* may not hold. For example, if tree-based algorithms such as CART and random forests are used to predict the conditional treatment effect, the predicted conditional treatment effect function is not differentiable at some points. Although the resulting Ω^* does not exactly satisfy Assumption 4 (d), the assumptions approximately hold in that Ω^* is arbitrarily well approximated by a set that satisfies the differentiability condition.

The second part of Assumption 4 (e) requires that the intersection of the boundary of Ω^* and the boundary of the support of X_i be less than $(p-1)$ dimensional. The first and third parts of (e) require that the intersection of the boundary of Ω^* and the interior of the support of X_i be $(p-1)$ dimensional so that the $(p-1)$-dimensional Hausdorff measure of the region is nonzero. These conditions are satisfied in practice.[14] Assumption 4 (f) is a continuity condition for consistent estimation of the causal effects on the boundary.[15]

Under the above regularity conditions, the 2SLS estimators $\hat{\beta}_1$ and $\hat{\beta}_1^s$ consistently estimate a weighted average treatment effect.

Theorem 1. *Suppose that Assumptions 1, 2 and 4 hold, $\delta_n \to 0$, $n\delta_n \to \infty$ and $S_n \to \infty$ as $n \to \infty$. Then the 2SLS estimators $\hat{\beta}_1$ and $\hat{\beta}_1^s$ converge in probability to*

$$\lim_{\delta \to 0} E[\omega_i(\delta)(Y_i(1) - Y_i(0))],$$

where

$$\omega_i(\delta) = \frac{p^{ML}(X_i; \delta)(1 - p^{ML}(X_i; \delta))(D_i(1) - D_i(0))}{E[p^{ML}(X_i; \delta)(1 - p^{ML}(X_i; \delta))(D_i(1) - D_i(0))]}.$$

The limit $\lim_{\delta \to 0} E[\omega_i(\delta)(Y_i(1) - Y_i(0))]$ always exists under the assumptions of Theorem 1. Theorem 1 says that the 2SLS estimators converge to a weighted average of causal effects for the

[13] The assumption can be relaxed so that the sign of $E[D_i(1) - D_i(0)|X_i = x]$ is allowed to vary over x such that $p^{ML}(x) \in (0, 1)$ at the cost of making the presentation more complex.

[14] The boundary of Ω^* fails to be $(p-1)$ dimensional, for example, when the covariate space is three dimensional $(p = 3)$ and Ω^* is a straight line, not a set with nonzero volume nor even a plane. In this example, the boundary is the same as Ω^*, and its two-dimensional Hausdorff measure is zero.

[15] Assumption 3 is not sufficient for Assumption 4 (f), since it is possible, for example, that $ML(x) \in (0, 1)$ for all $x \in \partial\Omega^* \cap \text{int}(\mathcal{X})$ even if $\Pr(ML(X_i) \in (0, 1)) = 0$, and Assumption 3 does not guarantee continuity at the boundary points in this case.

subpopulation whose QPS is nondegenerate ($p^{ML}(X_i;\delta) \in (0,1)$) and who would switch their treatment status in response to the treatment recommendation ($D_i(1) \neq D_i(0)$).

When $\Pr(ML(X_i) \in (0,1)) = 0$, our consistency result requires that δ_n go to zero slower than n^{-1}. The rate condition ensures that we have sufficiently many observations in the δ_n-neighborhood of the boundary of Ω^*. Importantly, the rate condition does not depend on the dimension of X_i, unlike other estimation methods that use a bandwidth such as kernel methods. Intuitively, this is because the dimension of the boundary of Ω^* is the dimension of X_i minus one (that is, $p-1$), and the rate at which the probability of X_i being in the δ-neighborhood of the boundary shrinks to zero is δ^{-1} regardless of the dimension of X_i.

Whether or not $\Pr(ML(X_i) \in (0,1)) = 0$, when we use simulated QPS, the consistency result requires that the number of simulation draws S_n go to infinity as n increases, so that $p^s(X_i;\delta_n)$ approaches $p^{ML}(X_i;\delta_n)$. We do not need any conditions on the rate at which S_n increases regardless of the dimension of X_i.

Importantly, the result in Theorem 1 holds whether ML is stochastic ($\Pr(ML(X_i) \in (0,1)) > 0$) or deterministic ($\Pr(ML(X_i) \in (0,1)) = 0$).[16] As shown in the proof of Theorem 1, if we consider these two underlying cases separately, the probability limit of the 2SLS estimators have more specific expressions. For the presentation of the expressions, we use the fact that the 2SLS estimator $\hat{\beta}_1$ is equivalent to $\hat{\alpha}_1/\hat{\gamma}_1$, where $\hat{\gamma}_1$ is the OLS coefficient on Z_i from the first-stage regression (1), and $\hat{\alpha}_1$ is the OLS coefficient on Z_i from the following reduced-form regression (Khazzoom, 1976):

$$Y_i = \alpha_0(1-I) + \alpha_1 Z_i + \alpha_2 p^{ML}(X_i;\delta_n) + u_i. \tag{5}$$

The 2SLS estimator using simulated QPS $\hat{\beta}_1^s$ is numerically equivalent to $\hat{\alpha}_1^s/\hat{\gamma}_1^s$, where $\hat{\gamma}_1^s$ and $\hat{\alpha}_1^s$ are defined analogously. Now, if $\Pr(ML(X_i) \in (0,1)) > 0$, it is shown that

$$\operatorname{plim} \hat{\alpha}_1 = \operatorname{plim} \hat{\alpha}_1^s = \frac{E[ML(X_i)(1-ML(X_i))(Y_{1i}-Y_{0i})]}{E[ML(X_i)(1-ML(X_i))]} \tag{6}$$

and

$$\operatorname{plim} \hat{\gamma}_1 = \operatorname{plim} \hat{\gamma}_1^s = \frac{E[ML(X_i)(1-ML(X_i))(D_i(1)-D_i(0))]}{E[ML(X_i)(1-ML(X_i))]} \tag{7}$$

so that

$$\operatorname{plim} \hat{\beta}_1 = \operatorname{plim} \hat{\beta}_1^s = \frac{E[ML(X_i)(1-ML(X_i))(Y_{1i}-Y_{0i})]}{E[ML(X_i)(1-ML(X_i))(D_i(1)-D_i(0))]}.$$

Expressions (6) and (7) have causal interpretations as weighted averages of the causal effects of the treatment recommendation Z_i on the outcome Y_i and on the treatment assignment D_i, respectively. The weights are proportional to $ML(X_i)(1-ML(X_i))$, the conditional variance of the treatment recommendation. The probability limit of the 2SLS estimators is the ratio of these two causal effects.

[16] Note that $\omega_i(\delta)$ does not simplify to $\frac{p^{ML}(X_i)(1-p^{ML}(X_i))(D_i(1)-D_i(0))}{E[p^{ML}(X_i)(1-p^{ML}(X_i))(D_i(1)-D_i(0))]}$ as $\delta \to 0$ in general, since if $\Pr(ML(X_i) \in (0,1)) = 0$, we have $p^{ML}(X_i) \in (0,1)$ only for $x \in \partial\Omega^* \cap \mathcal{X}$, and $p^{ML}(X_i) \in \{0,1\}$ almost surely.

To relate this result to existing work, consider the reduced-form and first-stage regressions with the (standard) propensity score $ML(X_i)$ control:

$$\begin{aligned} Y_i &= \alpha_1 Z_i + \alpha_2 ML(X_i) + u_i \\ D_i &= \gamma_1 Z_i + \gamma_2 ML(X_i) + \nu_i. \end{aligned} \tag{8}$$

A classic result shows that under Assumption 1, the OLS coefficients on Z_i from the above regressions converge in probability to the treatment-variance weighted averages of causal effects in (6) and (7) (Angrist and Pischke, 2008, Chapter 3.3).[17] Not surprisingly, for this selection-on-observables case, our result shows that the 2SLS estimator is consistent for the same treatment effect whether we use as a control the propensity score, QPS, or simulated QPS.

Moreover, using QPS as a control allows us to consistently estimate a causal effect even if ML is deterministic and produces multidimensional regression-discontinuity variation only. If $\Pr(ML(X_i) \in (0,1)) = 0$, the proof shows that

$$\operatorname{plim} \hat{\alpha}_1 = \operatorname{plim} \hat{\alpha}_1^s = \frac{\int_{\partial\Omega^* \cap \mathcal{X}} E[Y_{1i} - Y_{0i} | X_i = x] f_X(x) d\mathcal{H}^{p-1}(x)}{\int_{\partial\Omega^* \cap \mathcal{X}} f_X(x) d\mathcal{H}^{p-1}(x)} \tag{10}$$

and

$$\operatorname{plim} \hat{\gamma}_1 = \operatorname{plim} \hat{\gamma}_1^s = \frac{\int_{\partial\Omega^* \cap \mathcal{X}} E[D_i(1) - D_i(0) | X_i = x] f_X(x) d\mathcal{H}^{p-1}(x)}{\int_{\partial\Omega^* \cap \mathcal{X}} f_X(x) d\mathcal{H}^{p-1}(x)} \tag{11}$$

so that

$$\operatorname{plim} \hat{\beta}_1 = \operatorname{plim} \hat{\beta}_1^s = \frac{\int_{\partial\Omega^* \cap \mathcal{X}} E[Y_{1i} - Y_{0i} | X_i = x] f_X(x) d\mathcal{H}^{p-1}(x)}{\int_{\partial\Omega^* \cap \mathcal{X}} E[D_i(1) - D_i(0) | X_i = x] f_X(x) d\mathcal{H}^{p-1}(x)}.$$

Expressions (10) and (11) have causal interpretation as (density-weighted) averages of causal effects of the treatment recommendation Z_i on the outcome Y_i and on the treatment assignment D_i, respectively, for the subpopulation who are on the boundary of the treated region. The probability limit of the 2SLS estimators is the ratio of these two causal effects.

Recall that the 2SLS regression uses the observations with $p^{ML}(X_i; \delta_n) \in (0,1)$ (or $p^s(X_i; \delta_n) \in (0,1)$ when we use simulated QPS) only. By definition, if $p^{ML}(X_i; \delta_n) \in (0,1)$, X_i must be in the δ_n-neighborhood of the boundary between the treated and untreated. Therefore, to derive the probability limits of $\hat{\alpha}_1$ and $\hat{\gamma}_1$, it is necessary to consider where the expectation (of relevant variables) conditional on X_i being in the δ-neighborhood of the boundary converges to as δ shrinks to zero. We develop a novel approach drawing on results from differential geometry and geometric measure theory. In this approach, we write the integral over the δ-neighborhood of

[17] Precisely speaking, Angrist and Pischke (2008) consider regression of Y_i (or D_i) on Z_i controlling a dummy variable for every value taken on by X_i (i.e., the model is saturated in X_i) when X_i is a discrete variable:

$$Y_i = \alpha_1 Z_i + \sum_{x \in \mathcal{X}} \alpha_{2,x} 1\{X_i = x\} + u_i. \tag{9}$$

By the Frisch-Waugh Theorem, the population coefficients on Z_i from (8) and (9) are both given by $\alpha_1 = \frac{E[(Z_i - E[Z_i|X_i])Y_i]}{E[(Z_i - E[Z_i|X_i])^2]}$. Angrist and Pischke (2008) show that this expression is reduced to the treatment-variance weighted average of treatment effects in (6) under the conditional independence assumption. Their derivation follows when X_i is continuous and we control the propensity score as in (8).

the boundary in terms of the iterated integral over the sets of values of equal distance from the boundary (so called levels sets), and take the limit of the iterated integral. Consequently, the probability limits of $\hat{\alpha}_1$ and $\hat{\gamma}_1$ are expressed as integrals over boundary points with respect to the $(p-1)$-dimensional Hausdorff measure as in (10) and (11).[18] The weight on the boundary points is simply proportional to the probability density, since by continuity of the probability density function, the density of the points in any level set is close to that of the boundary points when δ is small.[19]

Finally, note that the weight $\omega_i(\delta)$ given in Theorem 1 is negative if $D_i(1) < D_i(0)$, so $E[\omega_i(\delta)(Y_i(1) - Y_i(0))]$ may not be a causally interpretable convex combination of treatment effects $Y_i(1) - Y_i(0)$. This can happen because the treatment effect of those whose treatment assignment shifts from 1 to 0 when the treatment recommendation Z_i is switched on negatively contributes to $E[\omega_i(\delta)(Y_i(1) - Y_i(0))]$. Additional assumptions prevent this problem. If the treatment effect is constant, for example, the 2SLS estimators are consistent for the treatment effect.

Corollary 3. *Suppose that Assumptions 1, 2, and 4 hold, that the treatment effect is constant, i.e., $Y_i(1) - Y_i(0) = b$ for some constant b, and that $\delta_n \to 0$, $n\delta_n \to \infty$ and $S_n \to \infty$ as $n \to \infty$. Then the 2SLS estimators $\hat{\beta}_1$ and $\hat{\beta}_1^s$ converge in probability to b.*

Another approach is to impose monotonicity (Imbens and Angrist, 1994). If $\Pr(D_i(1) \geq D_i(0)|X_i = x) = 1$, we have

$$E[(D_i(1) - D_i(0))(Y_i(1) - Y_i(0))|X_i = x] = E[D_i(1) - D_i(0)|X_i = x]LATE(x),$$

where $LATE(x) = E[Y_i(1) - Y_i(0)|D_i(1) \neq D_i(0), X_i = x]$ is the LATE conditional on $X_i = x$. The 2SLS estimators are then consistent for a weighted average of conditional LATEs with all weights nonnegative.

Corollary 4. *Suppose that Assumptions 1, 2, and 4 hold, that $\Pr(D_i(1) \geq D_i(0)|X_i = x) = 1$ for any $x \in \mathcal{X}$ such that $p^{ML}(x) \in (0,1)$, and that $\delta_n \to 0$, $n\delta_n \to \infty$ and $S_n \to \infty$ as $n \to \infty$. Then the 2SLS estimators $\hat{\beta}_1$ and $\hat{\beta}_1^s$ converge in probability to*

$$\lim_{\delta \to 0} E[\omega(X_i; \delta)LATE(X_i)],$$

where

$$\omega(x;\delta) = \frac{p^{ML}(x;\delta)(1 - p^{ML}(x;\delta))E[D_i(1) - D_i(0)|X_i = x]}{E[p^{ML}(X_i;\delta)(1 - p^{ML}(X_i;\delta))(D_i(1) - D_i(0))]}.$$

[18] In contrast, prior studies on multidimensional regression discontinuity frequently express treatment effect estimands in terms of expectations conditional on X_i being in the boundary like $E[Y_{1i} - Y_{0i}|X_i \in \partial\Omega^*]$ (Zajonc, 2012; Keele and Titiunik, 2015). However, those conditional expectations are, formally, not well-defined, since $\mathcal{L}^p(\partial\Omega^*) = 0$ and hence $\Pr(X_i \in \partial\Omega^*) = 0$. We therefore prefer our expression in terms of integrals with respect to the Hausdorff measure to any expressions in terms of conditional expectations on the boundary.

[19] Controlling $p^{ML}(X_i; \delta_n)$ linearly in the regression plays a role similar to the role played by the linear term in the local linear regression; it is expected to controll the slopes of the conditional means of the potential outcomes near the boundary. This is different from its role of controlling the heterogeneous propensity scores in the case where $\Pr(ML(X_i) \in (0,1) > 0$.

As before, the result holds whether or not $\Pr(ML(X_i) \in (0,1)) > 0$. If $\Pr(ML(X_i) \in (0,1)) > 0$, we have

$$\operatorname{plim} \hat{\beta}_1 = \operatorname{plim} \hat{\beta}_1^s = E[\omega_1(X_i) LATE(X_i)].$$

where

$$\omega_1(x) = \frac{ML(x)(1 - ML(x))E[D_i(1) - D_i(0)|X_i = x]}{E[ML(X_i)(1 - ML(X_i))(D_i(1) - D_i(0))]}.$$

The probability limit of the 2SLS estimators is a weighted average of conditional LATEs over all values of X_i with nondegenerate $ML(X_i)$. The weights are proportional to the conditional variance of the treatment recommendation, $ML(X_i)(1 - ML(X_i))$, and to the proportion of compliers, $E[D_i(1) - D_i(0)|X_i]$. As in the case without monotonicity, it is also known that the 2SLS regression using the propensity score $ML(X_i)$ as a control instead of QPS $p(X_i; \delta_n)$ consistently estimates the same weighted average of conditional LATEs (Hull, 2018). Our result shows that the 2SLS estimator is consistent for the same weighted average treatment effect whether we use as a control the propensity score, QPS, or simulated QPS.

If $\Pr(ML(X_i) \in (0,1)) = 0$, we have

$$\operatorname{plim} \hat{\beta}_1 = \operatorname{plim} \hat{\beta}_1^s = \int_{\partial\Omega^* \cap \mathcal{X}} \omega_2(x) LATE(x) f_X(x) d\mathcal{H}^{p-1}(x),$$

where

$$\omega_2(x) = \frac{E[D_i(1) - D_i(0)|X_i = x]}{\int_{\partial\Omega^* \cap \mathcal{X}} E[D_i(1) - D_i(0)|X_i = x] f_X(x) d\mathcal{H}^{p-1}(x)}.$$

The probability limit is a weighted average of conditional LATEs over all values of X_i on the boundary of the treated region with the weights simply proportional to the proportion of compliers.

5 Examples

Here we give examples and discuss the applicability of our framework. The examples exploit the manner in which a particular algorithm creates treatment recommendations or assignments using observable variables alone.

Example 1 (Supervised Learning)**.** Millions of times each year, judges make bail-or-release decisions that hinge on a prediction of what a defendant would do if released. Many judges now use proprietary algorithms (like COMPAS criminal risk score) to make such predictions and use the predictions to support bail-or-release decisions. Kleinberg *et al.* (2017) also developed another prediction algorithm.

These algorithms fit into the above framework as a simple special case. Using our notation, assume that a criminal risk algorithm recommends bailing ($Z_i = 1$) and releasing ($Z_i = 0$) to each defendant i. The algorithm uses defendant i's observable characteristics X_i, including criminal history and demographics. The algorithm first translates X_i into a continuous risk score $r(X_i)$, where $r : \mathbb{R}^p \to \mathbb{R}$ is a function estimated by supervised leaning based on past data and assumed

to be fixed. For example, Kleinberg *et al.* (2017) learn $r(X_i)$ using gradient boosted decision trees. The algorithm then uses the risk score to make the final recommendation:

$$Z_i^{SL} \equiv 1\{r(X_i) > c\}, \quad ML^{SL}(X_i) \equiv \Pr(r(X_i) > c|X_i)$$

where $c \in \mathbb{R}$ is a constant threshold.[20] A similar procedure applies to the screening of potential borrowers by banks and insurance companies based on credit scores estimated by supervised learning (Agarwal, Chomsisengphet, Mahoney and Stroebel, 2017).[21]

Suppose that r is continuous and that r is continuously differentiable in a neighborhood of x and $\frac{\partial r(x)}{\partial x} \neq 0$ for any $x \in \text{int}(\mathcal{X})$ with $r(x) = c$. As expected, QPS for this case is given by

$$p^{SL}(x) = \begin{cases} 0 & \text{if } r(x) < c \\ 0.5 & \text{if } r(x) = c \text{ and } x \in \text{int}(\mathcal{X}) \\ 1 & \text{if } r(x) > c. \end{cases}$$

It is therefore possible to identify and estimate causal effects conditional on x's with $r(x) = c$ and $x \in \text{int}(\mathcal{X})$.

Example 2 (Reinforcement Learning and Bandit)**.** We are constantly exposed to digital information (movie, music, news, search results, advertisements, and recommendations) through a variety of devices and platforms. Tech companies allocate these pieces of content by reinforcement learning and bandit algorithms (White, 2012; Sutton and Barto, 2018). Our method is also applicable to many popular bandit and reinforcement learning algorithms. For simplicity, we assume that individuals perfectly comply with the treatment assignment ($D_i = Z_i$).

1. (Bandit Algorithms) The algorithms below first use past data and supervised learning to estimate the conditional means and variances of potential outcomes, $E[Y_i(z)|X_i]$ and $\text{Var}(Y_i(z)|X_i)$, for each $z = 0, 1$. Let $\mu_z(X_i)$ and $\sigma_z^2(X_i)$ denote the estimators. The algorithms then use $\mu_z(X_i)$ and $\sigma_z^2(X_i)$ to determine the treatment assignment for individual i.

(a) (Thompson Sampling Using Gaussian Priors) The algorithm first samples potential outcomes from the normal distribution with mean $(\mu_0(X_i), \mu_1(X_i))$ and variance covariance matrix $\text{diag}(\sigma_0^2(X_i), \sigma_1^2(X_i))$. The algorithm then chooses the treatment with the highest

[20] The algorithm sometimes discretizes the continuous risk score $r(X_i)$ into $d(r(X_i))$, where $d : \mathbb{R} \to \mathbb{N}$ (Cowgill, 2018). In this case, the algorithm uses the discretized risk score to make the final recommendation:

$$Z_i^{SL} \equiv 1\{d(r(X_i)) > c\}.$$

[21] A widely-used approach to identifying and estimating treatment effects in these settings is to use $r(X_i)$ as a univariate running variable and apply a univariate regression discontinuity method (Cowgill, 2018; Bundorf *et al.*, 2019). This approach is valid under the assumption that conditional expectation functions of potential outcomes given the score function (e.g., $E[Y_{zi}|r(X_i) = r]$) are continuous. Our approach rather imposes conditions on lower-level conditional expectation functions given the covariates (e.g., $E[Y_{zi}|X_i = x]$) and on ML, and discusses how these assumptions lead to identification and consistent estimation of treatment effects.

sampled potential outcome. As a result, this algorithm chooses the treatment assignment as follows:

$$Z_i^{TS} \equiv \arg\max_{z=0,1} y(z), \quad ML^{TS}(X_i) \equiv E[\arg\max_{z=0,1} y(z)|X_i]$$

where $y(z) \sim \mathcal{N}(\mu_z(X_i), \sigma_z^2(X_i))$ independently across z. These algorithms often induce quasi-experimental variation in treatment assignment, as a strand of the computer science literature observed (Precup, 2000; Li *et al.*, 2010; Narita, Yasui and Yata, 2019). The function ML has an analytical expression:

$$ML^{TS}(x) = 1 - \Phi(\frac{\mu_0(x) - \mu_1(x)}{\sqrt{\sigma_0^2(x) + \sigma_1^2(x)}}),$$

where Φ is the CDF of a standard normal distribution. Suppose that the functions μ_0, μ_1, σ_0^2 and σ_1^2 are continuous on $\text{int}(\mathcal{X})$. QPS for this case is given by

$$p^{TS}(x) = 1 - \Phi(\frac{\mu_0(x) - \mu_1(x)}{\sqrt{\sigma_0^2(x) + \sigma_1^2(x)}})$$

for any $x \in \text{int}(\mathcal{X})$.

(b) (Upper Confidence Bound, UCB) Unlike the above stochastic one, the UCB algorithm is a deterministic algorithm, producing a less obvious example of our framework. This algorithm chooses the treatment with the highest upper confidence bound for the potential outcome:

$$Z_i^{UCB} \equiv \arg\max_{z=0,1}\{\mu_z(X_i) + \alpha\sigma_z(X_i)\},$$

where α is chosen so that $|\mu_z(X_i) - E[Y_i(z)|X_i]| \leq \alpha\sigma_z(X_i)$ at least with some probability, for example, 0.95. Let $ML^{UCB}(X_i) \equiv E[Z_i^{UCB}|X_i]$. Suppose that the function $\mu_1 - \mu_0 + \alpha(\sigma_1 - \sigma_0)$ satisfies the condition imposed on r in Example 1 with $c = 0$. QPS for this case is given by

$$p^{UCB}(x) = \begin{cases} 0 & \text{if } \mu_1(x) + \alpha\sigma_1(x) < \mu_0(x) + \alpha\sigma_0(x) \\ 0.5 & \text{if } \mu_1(x) + \alpha\sigma_1(x) = \mu_0(x) + \alpha\sigma_0(x) \text{ and } x \in \text{int}(\mathcal{X}) \\ 1 & \text{if } \mu_1(x) + \alpha\sigma_1(x) > \mu_0(x) + \alpha\sigma_0(x). \end{cases}$$

This means that the UCB algorithm produces potentially complicated quasi-experimental variation along the boundary in the covariates space where the algorithm's treatment recommendation changes from one to the other.

2. (Reinforcement Learning Algorithms) Reinforcement learning algorithms are used to optimize decisions in dynamic environments, where the current state and action can affect the future states and outcomes. Let $\{(X_{ti}, Z_{ti}, Y_{ti})\}_{t=0}^{\infty}$ denote the trajectory of the states, treatment assignments, and outcomes in periods $t = 0, 1, 2, \cdots$ for individual i. For simplicity, we assume that the trajectory follows a Markov decision process, where the distribution of the state X_{ti} only depends on the last state and treatment assignment $(X_{t-1,i}, Z_{t-1,i})$, the distribution of the outcome Y_{ti} only depends on the current state and treatment assignment (X_{ti}, Z_{ti}), and these distributions

are stationary over periods. Let $Y_{ti}(1)$ and $Y_{ti}(0)$ represent the potential outcomes in period t. Let $Q : \mathcal{X} \times \{0,1\} \to \mathbb{R}$ be the optimal state-action value function, called the *Q-function*: for $(x,z) \in \mathcal{X} \times \{0,1\}$,

$$Q(x,z) \equiv \max_{\pi:\mathcal{X}\to[0,1]} E[\sum_{t=0}^{\infty} \gamma^t (Y_{ti}(1)\pi(X_{ti}) + Y_{ti}(0)(1-\pi(X_{ti}))|X_{0i} = x, Z_{0i} = z],$$

where $\gamma \in [0,1)$ is a discount factor, and π is a policy function that assigns the probability of treatment to each possible state.

(a) (Fitted Q Iteration with ϵ-Greedy) The fitted Q iteration algorithm (Ernst, Geurts and Wehenkel, 2005) is a batch reinforcement learning algorithm that uses a set of L four-tuples $\{(x_{t_l}^l, z_{t_l}^l, y_{t_l}^l, x_{t_l+1}^l) : l = 1, ..., L\}$ collected in the past to yield an approximation of the Q-function. Given $\{(x_{t_l}^l, z_{t_l}^l, y_{t_l}^l, x_{t_l+1}^l) : l = 1, ..., L\}$ and an initial approximation $\hat{Q}$ of Q (e.g., $\hat{Q}(x,z) = 0$ for all (x,z)), the algorithm repeats the following steps until some stopping condition is reached:

i. For each $l = 1, ..., L$, calculate $q^l = y_{t_l}^l + \gamma \max_{z\in\{0,1\}} \hat{Q}(x_{t_l+1}^l, z)$.

ii. Use $\{(x_{t_l}^l, z_{t_l}^l, q^l) : l = 1, ..., L\}$ and a supervised learning method to train a model that predicts q from (x,z). Let the model be a new approximation $\hat{Q}$ of Q.

Possible supervised learning methods used in the second step include tree-based methods (Ernst *et al.*, 2005), neural networks (Riedmiller, 2005, Neural Fitted Q Iteration) and deep neural networks (Lange and Riedmiller, 2010, Deep Fitted Q Iteration).

The algorithm then uses the estimated Q-function to determine the treatment assignment for newly arriving individuals. One standard assignment rule is the ϵ-Greedy algorithm, which chooses the best treatment based on $\hat{Q}(X_{ti}, z)$ with probability $1 - \frac{\epsilon}{2}$ and chooses the other treatment with probability $\frac{\epsilon}{2}$: for each t,

$$Z_{ti}^{\epsilon} \equiv \begin{cases} \arg\max_{z=0,1} \hat{Q}(X_{ti}, z) & \text{with probability } 1 - \frac{\epsilon}{2} \\ 1 - \arg\max_{z=0,1} \hat{Q}(X_{ti}, z) & \text{with probability } \frac{\epsilon}{2}. \end{cases}$$

Let $ML^{\epsilon}(X_{ti}) \equiv E[Z_{ti}^{\epsilon}|X_{ti}]$. Suppose that the function $\hat{Q}(\cdot,1) - \hat{Q}(\cdot,0)$ satisfies the condition imposed on r in Example 1 with $c = 0$. QPS for this case is given by

$$p^{\epsilon}(x) = \begin{cases} \frac{\epsilon}{2} & \text{if } \hat{Q}(x,1) < \hat{Q}(x,0) \\ 0.5 & \text{if } \hat{Q}(x,1) = \hat{Q}(x,0) \text{ and } x \in \text{int}(\mathcal{X}) \\ 1 - \frac{\epsilon}{2} & \text{if } \hat{Q}(x,1) > \hat{Q}(x,0). \end{cases}$$

(b) (Policy Gradient Methods) Policy gradient methods such as REINFORCE (Williams, 1992) approximate the optimal policy function by parametrization and learn the parameter using stochastic gradient ascent. Let $\pi(x;\theta)$ be a parametrization of the policy function that is differentiable with respect to θ. For example, π might be a softmax function with a linear index: $\pi(x;\theta) = \frac{\exp(x'\theta)}{1+\exp(x'\theta)}$. Another example is a neural network whose input is a

representation of the state x, whose output is the treatment assignment probability, and whose weights are represented by the parameter θ.

Suppose that we have collected a set of L trajectories $\{\{(x_t^l, z_t^l, y_t^l)\}_{t=0}^{T_l} : l = 1, ..., L\}$ by running the policy $\pi(x; \theta^0)$ for L individuals. REINFORCE updates the policy parameter in the following way:

$$\theta^1 = \theta^0 + \alpha \sum_{l=1}^{L} \sum_{t=1}^{T_l} (\sum_{t'=t}^{T_l} \gamma^{t'} y_{t'}^l)[z_t^l \frac{\frac{\partial \pi(x_t^l;\theta_0)}{\partial \theta}}{\pi(x_t^l;\theta_0)} - (1 - z_t^l) \frac{\frac{\partial \pi(x_t^l;\theta_0)}{\partial \theta}}{(1 - \pi(x_t^l;\theta_0))}],$$

where α is a step size. Other policy gradient methods such as Actor-Critic Methods substitute a different value for $(\sum_{t'=t}^{T_l} \gamma^{t'} y_{t'}^l)$ in the above updating rule.

The algorithm then uses the updated policy function $\pi(x; \theta^1)$ to determine the treatment assignment for new episodes. For each t,

$$Z_{ti}^{PG} \equiv \begin{cases} 1 & \text{with probability } \pi(X_{ti}; \theta^1) \\ 0 & \text{with probability } 1 - \pi(X_{ti}; \theta^1). \end{cases} \qquad ML^{TG}(X_{ti}) \equiv \pi(X_{ti}; \theta^1).$$

Suppose that the function $\pi(\cdot; \theta^1)$ is continuous on $\text{int}(\mathcal{X})$. QPS for this case is given by

$$p^{TG}(x) = \pi(x; \theta^1)$$

for any $x \in \text{int}(\mathcal{X})$.

Example 3 (Unsupervised Learning)**.** Customer segmentation is a core marketing practice that divides a company's customers into groups based on their characteristics and purchasing behaviors so that the company can effectively target marketing activities at each group. Many businesses today use unsupervised learning algorithms, clustering algorithms in particular, to perform customer segmentation. Using our notation, assume that a company decides whether it targets a campaign at customer i ($Z_i = 1$) or not ($Z_i = 0$). The company first uses a clustering algorithm like K-means clustering and Gaussian mixture model clustering to divide customers into K groups, making a partition $\{S_1, ..., S_K\}$ of the covariate space $\mathbb{R}^p$. The company then conducts the campaign targeted at some of the groups:

$$Z_i^{CL} \equiv 1\{X_i \in \cup_{k \in T} S_k\},$$

where $T \subset \{1, .., K\}$ is the set of the indices of the target groups. Let $ML^{CL}(X_i) \equiv E[Z_i^{CL} | X_i]$.

For example, suppose that the company uses K-means clustering, which creates a partition in which a covariate value x belongs to the group with the nearest centroid. Let $c_1, ..., c_K$ be the centroids of the K groups, and define a set-valued function $C : \mathbb{R}^p \to 2^{\{1,...,K\}}$, where $2^{\{1,...,K\}}$ is the power set of $\{1, ..., K\}$, as

$$C(x) \equiv \underset{k \in \{1,...,K\}}{\arg\min} \ \|x - c_k\|.$$

If $C(x)$ is a singleton, x belongs to the only group in $C(x)$. If $C(x)$ contains more than one indices, which of the groups in $C(x)$ x belongs to is arbitrarily determined. QPS for this case is

given by

$$p^{CL}(x) = \begin{cases} 0 & \text{if } C(x) \cap T = \emptyset \\ 0.5 & \text{if } |C(x)| = 2,\ x \in \partial(\cup_{k \in T} S_k) \text{ and } x \in \text{int}(\mathcal{X}) \\ 1 & \text{if } C(x) \subset T \end{cases}$$

and $p^{CL}(x) \in (0,1)$ if $|C(x)| \geq 3$, $x \in \partial(\cup_{k \in T} S_k)$ and $x \in \text{int}(\mathcal{X})$, where $|C(x)|$ is the number of elements in $C(x)$.[22] Thus, it is possible to identify causal effects conditional on x's on the boundary $\partial(\cup_{k \in T} S_k)$.

Example 4 (Policy Eligibility Rules)**.** Medicaid and other welfare policies often decide eligibility based on algorithmic rules, as studied by Currie and Gruber (1996b,a); Cohodes, Grossman, Kleiner and Lovenheim (2016); Brown, Kowalski and Lurie (2017). Using our notation, the state government determines whether each individual i is eligible ($Z_i = 1$) or not ($Z_i = 0$) for Medicare. The state government's eligibility rule $ML^{Medicaid}$ maps individual characteristics X_i (e.g. income, family composition) into an eligibility decision $Z_i^{Medicare}$. A similar procedure also applies to bankruptcy laws (Mahoney, 2015). These policy eligibility rules produce quasi-experimental variation as in Example 1.

Corollary 5. *In all the above examples, there exists $x \in \mathcal{X}$ such that $p^{ML}(x) \in (0,1)$. Therefore, a causal effect is identified under Assumptions 1 – 3.*

6 Implications

As software (algorithm) eats the world, the world becomes a mountain of natural experiments and instruments. These instruments enable us to estimate causal treatment effects, as we formalize and illustrate in this paper. Our analysis clarifies a few implications for policy and management practices about ML algorithms. It is important to record ML implementation in a replicable, simulatable way, including what input variables X_i are used to make ML recommendation Z_i. Another key is to record ML recommendation Z_i even if they are superseded by human decision D_i. These data retention efforts would go a long way to exploit the full potential of machine learning as natural experiment.

In addition to estimating treatment effects, instruments induced by ML algorithms also inform how to improve the ML algorithms. To see this, suppose some algorithm ML_1 is in use. As we characterize in this paper, this algorithm ML_1 produces instrument IV_1. We can then use instrument IV_1 to make counterfactual predictions about what would happen if we change ML_1 to another algorithm ML_2. We'd then switch to ML_2 if it is predicted to be better than the previous algorithm. This algorithm change in turn would produce another cycle of natural experiment and improvements:

$$ML_1 \to IV_1 \to \textit{Algorithm Improvement}_1 \to ML_2 \to IV_2 \to \textit{Algorithm Improvement}_2 ...$$

[22] If $|C(x)| = 2$ and $x \in \partial(\cup_{k \in T} S_k)$, x is on a linear boundary between one target group and one non-target group, and hence QPS is 0.5. If $|C(x)| \geq 3$ and $x \in \partial(\cup_{k \in T} S_k)$, x is a common endpoint of several group boundaries, and QPS is determined by the angles at which the boundaries intersect.

This cycle of natural experiments and improvements may provide an alternative to well-established A/B testing (randomized experiment), which is often costly and risky. A/B testing is often technically or managerially infeasible, since deploying a new algorithm is time- and money-consuming, and entails a risk of failure. This difficulty with randomized experiment may be alleviated by additionally using machine learning as natural experiment.

References

Abadie, A. (2003). Semiparametric Instrumental Variable Estimation of Treatment Response Models. *Journal of Econometrics*, **113** (2), 231–263.

Abdulkadiroğlu, A., Angrist, J. D., Narita, Y. and Pathak, P. A. (2017). Research Design Meets Market Design: Using Centralized Assignment for Impact Evaluation. *Econometrica*, **85** (5), 1373–1432.

—, —, — and Pathak, P. A. (2019). Breaking Ties: Regression Discontinuity Design Meets Market Design. *Working Paper*.

Agarwal, S., Chomsisengphet, S., Mahoney, N. and Stroebel, J. (2017). Do Banks Pass Through Credit Expansions to Consumers Who Want to Borrow? *Quarterly Journal of Economics*, **133** (1), 129–190.

Angrist, J. D. and Pischke, J.-S. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Arai, Y., Hsu, Y.-C., Kitagawa, T., Mourifié, I. and Wan, Y. (2018). Testing Identifying Assumptions in Fuzzy Regression Discontinuity Designs. *cemmap Working Paper*.

Athey, S. and Imbens, G. W. (2017). The State of Applied Econometrics: Causality and Policy Evaluation. *Journal of Economic Perspectives*, **31** (2), 3–32.

Bajari, P., Nekipelov, D., Ryan, S. P. and Yang, M. (2015). Machine Learning Methods for Demand Estimation. *American Economic Review*, **105** (5), 481–85.

Belloni, A., Chernozhukov, V., Fernández-Val, I. and Hansen, C. (2017). Program Evaluation and Causal Inference with High-Dimensional Data. *Econometrica*, **85** (1), 233–298.

Bonhomme, S., Lamadon, T. and Manresa, E. (2017). Discretizing Unobserved Heterogeneity. *University of Chicago, Becker Friedman Institute for Economics Working Paper No. 2019-16*.

Brown, D., Kowalski, A. E. and Lurie, I. Z. (2017). Long-Term Impacts of Childhood Medicaid Expansions on Outcomes in Adulthood. *NBER Working Paper No. 20835*.

Bundorf, K., Polyakova, M. and Tai-Seale, M. (2019). How Do Humans Interact with Algorithms? Experimental Evidence from Health Insurance. *NBER Working Paper No. 25976*.

Cattaneo, M. D., Frandsen, B. R. and Titiunik, R. (2015). Randomization Inference in the Regression Discontinuity Design: An Application to Party Advantages in the US Senate. *Journal of Causal Inference*, **3** (1), 1–24.

—, Titiunik, R. and Vazquez-Bare, G. (2017). Comparing Inference Approaches for RD Designs: A Reexamination of the Effect of Head Start on Child Mortality. *Journal of Policy Analysis and Management*, **36** (3), 643–681.

—, —, — and Keele, L. (2016). Interpreting Regression Discontinuity Designs with Multiple Cutoffs. *Journal of Politics*, **78** (4), 1229–1248.

Cohen, P., Hahn, R., Hall, J., Levitt, S. and Metcalfe, R. (2016). Using Big Data to Estimate Consumer Surplus: The Case of Uber. *NBER Working Paper No. 22627*.

Cohodes, S. R., Grossman, D. S., Kleiner, S. A. and Lovenheim, M. F. (2016). The Effect of Child Health Insurance Access on Schooling: Evidence from Public Insurance Expansions. *Journal of Human Resources*, **51** (3), 727–759.

Cowgill, B. (2018). The Impact of Algorithms on Judicial Discretion: Evidence from Regression Discontinuities. *Working Paper*.

Crasta, G. and Malusa, A. (2007). The Distance Function from the Boundary in a Minkowski Space. *Transactions of the American Mathematical Society*, **359**, 5725–5759.

Currie, J. and Gruber, J. (1996a). Health Insurance Eligibility, Utilization of Medical Care, and Child Health. *Quarterly Journal of Economics*, **111** (2), 431–466.

— and — (1996b). Saving Babies: The Efficacy and Cost of Recent Changes in the Medicaid Eligibility of Pregnant Women. *Journal of Political Economy*, **104** (6), 1263–1296.

Dong, Y. (2012). Jumpy or Kinky? Regression Discontinuity without the Discontinuity. *Working Paper*.

— (2018). Alternative Assumptions to Identify LATE in Fuzzy Regression Discontinuity Designs. *Oxford Bulletin of Economics and Statistics*, **80** (5), 1020–1027.

Einav, L., Finkelstein, A., Mullainathan, S. and Obermeyer, Z. (2018). Predictive Modeling of U.S. Health Care Spending in Late Life. *Science*, **360** (6396), 1462–1465.

Ernst, D., Geurts, P. and Wehenkel, L. (2005). Tree-Based Batch Mode Reinforcement Learning. *Journal of Machine Learning Research*, **6**, 503–556.

Executive Office of the President (2016). *Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights*. Executive Office of the President.

Frandsen, B. R. (2017). Party Bias in Union Representation Elections: Testing for Manipulation in the Regression Discontinuity Design When the Running Variable is Discrete. In *Regression Discontinuity Designs: Theory and Applications*, Emerald Publishing Limited, pp. 281–315.

—, Frölich, M. and Melly, B. (2012). Quantile Treatment Effects in the Regression Discontinuity Design. *Journal of Econometrics*, **168** (2), 382–395.

Frölich, M. (2007a). Nonparametric IV Estimation of Local Average Treatment Effects with Covariates. *Journal of Econometrics*, **139** (1), 35–75.

Frölich, M. (2007b). Regression Discontinuity Design with Covariates. *IZA Discussion Paper No. 3024*.

Frölich, M. and Huber, M. (2019). Including Covariates in the Regression Discontinuity Design. *Journal of Business and Economic Statistics*, **37(4)**, 736–748.

Hahn, J., Todd, P. and der Klaauw, W. V. (2001). Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design. *Econometrica*, **69** (1), 201–209.

Hoffman, M., Kahn, L. B. and Li, D. (2017). Discretion in Hiring. *Quarterly Journal of Economics*, **133** (2), 765–800.

Horton, J. J. (2017). The Effects of Algorithmic Labor Market Recommendations: Evidence from a Field Experiment. *Journal of Labor Economics*, **35** (2), 345–385.

Hull, P. (2018). Subtracting the Propensity Score in Linear Models. *Working Paper*.

Imbens, G. W. and Angrist, J. D. (1994). Identification and Estimation of Local Average Treatment Effects. *Econometrica*, **62** (2), 467–475.

Keele, L. J. and Titiunik, R. (2015). Geographic Boundaries as Regression Discontinuities. *Political Analysis*, **23** (1), 127–155.

Khazzoom, J. D. (1976). An Indirect Least Squares Estimator for Overidentified Equations. *Econometrica*, **44** (4), 741–750.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J. and Mullainathan, S. (2017). Human Decisions and Machine Predictions. *Quarterly Journal of Economics*, **133** (1), 237–293.

Krantz, S. G. and Parks, H. R. (2008). *Geometric Integration Theory*. Birkhäuser Basel.

Lange, S. and Riedmiller, M. (2010). Deep Auto-Encoder Neural Networks in Reinforcement Learning. *The 2010 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8.

Li, L., Chu, W., Langford, J. and Schapire, R. E. (2010). A Contextual-Bandit Approach to Personalized News Article Recommendation. *Proceedings of the 19th international conference on World Wide Web (WWW)*, pp. 661–670.

Li, S. (2011). Concise Formulas for the Area and Volume of a Hyperspherical Cap. *Asian Journal of Mathematics and Statistics*, **4**, 66–70.

Mahoney, N. (2015). Bankruptcy as Implicit Health Insurance. *American Economic Review*, **105** (2), 710–46.

Narita, Y. (2017). (Non) Randomization: A Theory of Quasi-Experimental Evaluation of School Quality. *Working Paper*.

— (2018). Experiment-as-Market: Incorporating Welfare into Randomized Controlled Trials. *Working Paper*.

—, Yasui, S. and Yata, K. (2019). Efficient Counterfactual Learning from Bandit Feedback. *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, pp. 4634–4641.

Papay, J. P., Willett, J. B. and Murnane, R. J. (2011). Extending the Regression-Discontinuity Approach to Multiple Assignment Variables. *Journal of Econometrics*, **161** (2), 203–207.

Precup, D. (2000). Eligibility Traces for Off-Policy Policy Evaluation. *ICML '00: Proceedings of the Seventeenth International Conference on Machine Learning*, pp. 759–766.

Riedmiller, M. A. (2005). Neural Fitted Q Iteration - First Experiences with a Data Efficient Neural Reinforcement Learning Method. pp. 317–328.

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

Sekhon, J. S. and Titiunik, R. (2017). On Interpreting the Regression Discontinuity Design as a Local Experiment. In *Regression Discontinuity Designs: Theory and Applications*, Emerald Publishing Limited, pp. 1–28.

Stein, E. M. and Shakarchi, R. (2005). *Real analysis: measure theory, integration, and Hilbert spaces*. Princeton lectures in analysis, Princeton, NJ: Princeton Univ. Press.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. MIT Press.

Voelker, A. R., Gosmann, J. and Stewart, T. C. (2017). Efficiently Sampling Vectors and Coordinates from the n-Sphere and n-Ball. *Centre for Theoretical Neuroscience – Technical Report*.

White, J. (2012). *Bandit Algorithms for Website Optimization*. O'Reilly.

Williams, R. J. (1992). Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning. *Machine Learning*, **8**, 229–256.

Wong, V. C., Steiner, P. M. and Cook, T. D. (2013). Analyzing Regression-Discontinuity Designs with Multiple Assignment Variables: A Comparative Study of Four Estimation Methods. *Journal of Educational and Behavioral Statistics*, **38** (2), 107–141.

Zajonc, T. (2012). Regression Discontinuity Design with Multiple Forcing Variables. *Essays on Causal Inference for Public Policy*, pp. 45–81.

A Proofs

A.1 Proof of Proposition 1

Suppose that Assumptions 1 – 3 hold. Here, we only show that

(a) $E[Y_{1i} - Y_{0i}|X_i = x]$ is identified for every $x \in \mathcal{X}$ such that $p^{ML}(x) \in (0,1)$.

(b) Let A be any subset of $\mathcal{X}$. Suppose that A is any open subset of $\mathcal{X}$ such that $p^{ML}(x)$ exists for all $x \in A$. Then $E[Y_{1i} - Y_{0i}|X_i \in A]$ is identified only if $p^{ML}(x) \in (0,1)$ for almost every $x \in A$.

The result for $E[D_i(1) - D_i(0)|X_i \in A]$ is obtained by a similar argument.

Proof of Part (a). Pick an $x \in \mathcal{X}$ such that $p^{ML}(x) \in (0,1)$. In the case where $ML(x) \in (0,1)$, $E[Y_{1i} - Y_{0i}|X_i = x]$ is trivially identified, since by the conditional independence $Z_i \perp\!\!\!\perp (Y_{0i}, Y_{1i})|X_i$, $E[Y_i|X_i = x, Z_i = 1] - E[Y_i|X_i = x, Z_i = 0] = E[Y_{1i} - Y_{0i}|X_i = x]$ (see e.g. Angrist and Pischke (2008)), and the left-hand side is identified since $ML(x) = \Pr(Z_i = 1|X_i = x) \in (0,1)$.

We next consider the case where $ML(x) = 1$. We have that $E[Y_i|X_i = x, Z_i = 1] = E[Y_{1i}|X_i = x]$, so that $E[Y_{1i}|X_i = x]$ is identified. Therefore, it suffices to show that $E[Y_{0i}|X_i = x]$ is identified. Since $p^{ML}(x) = \lim_{\delta \to 0} p^{ML}(x;\delta) < 1$, $p^{ML}(x;\delta) < 1$ for any sufficiently small $\delta > 0$. This implies that there exists $x' \in \mathcal{X} \cap N(x,\delta)$ such that $ML(x') < 1$ for any sufficiently small $\delta > 0$, for otherwise $p^{ML}(x;\delta) = 1$. Hence, we can find a sequence $(x_n)_{n=1}^{\infty}$ such that $\lim_{n\to\infty} x_n = x$ and $ML(x_n) < 1$ for all $n \geq 1$. Since $ML(x_n) < 1$ for all $n \geq 1$, $E[Y_{0i}|X_i = x_n]$ is identified for all $n \geq 1$. By Assumption 3, $E[Y_{0i}|X_i]$ is continuous at x. We therefore obtain

$$\lim_{n\to\infty} E[Y_{0i}|X_i = x_n] = E[Y_{0i}|X_i = x],$$

from which it follows that $E[Y_{1i} - Y_{0i}|X_i = x]$ is identified. In the case where $ML(x) = 0$, identification of $E[Y_{1i} - Y_{0i}|X_i = x]$ is obtained following the procedure above under the assumption that $E[Y_{1i}|X_i = x]$ is continuous. □

Proof of Part (b). Suppose to the contrary that $\mathcal{L}^p(\{x \in A : p^{ML}(x) \in \{0,1\}\}) > 0$. Without loss of generality, assume $\mathcal{L}^p(\{x \in A : p^{ML}(x) = 1\}) > 0$. The proof proceeds in four steps.

Step 1. $\mathcal{L}^p(A \cap \mathcal{X}_1) > 0$.

Proof. By Assumption 2, ML is continuous almost everywhere. Part 1 of Cororally 2 then implies that $p^{ML}(x) = ML(x)$ for almost every $x \in \{x^* \in A : p^{ML}(x^*) = 1\}$. Since $\mathcal{L}^p(\{x \in A : p^{ML}(x) = 1\}) > 0$, $\mathcal{L}^p(\{x \in A : p^{ML}(x) = 1, p^{ML}(x) = ML(x)\}) > 0$, and hence $\mathcal{L}^p(A \cap \mathcal{X}_1) > 0$. □

Step 2. $A \cap \text{int}(\mathcal{X}_1) \neq \emptyset$.

Proof. Suppose that $A \cap \text{int}(\mathcal{X}_1) = \emptyset$. Then, we must have that $A \cap \mathcal{X}_1 \subset \mathcal{X}_1 \setminus \text{int}(\mathcal{X}_1)$. It then follows that $\mathcal{L}^p(A \cap \mathcal{X}_1) \leq \mathcal{L}^p(\mathcal{X}_1 \setminus \text{int}(\mathcal{X}_1)) = \mathcal{L}^p(\mathcal{X}_1) - \mathcal{L}^p(\text{int}(\mathcal{X}_1)) = 0$, where the last equality holds by Assumption 2. But this is a contradiction to the result from Step 1. □

Step 3. $p^{ML}(x) = 1$ *for any* $x \in \text{int}(\mathcal{X}_1)$.

Proof. Pick any $x \in \text{int}(\mathcal{X}_1)$. By the definition of interior, $N(x, \delta) \subset \mathcal{X}_1$ for any sufficiently small $\delta > 0$. Therefore, $p^{ML}(x; \delta) = 1$ for any sufficiently small $\delta > 0$. □

Step 4. $E[Y_{1i} - Y_{0i}|X_i \in A]$ *is not identified.*

Proof. We first introduce some notations. Let $\mathbf{Q}$ be the set of distributions of $(Y_{1i}, Y_{0i}, X_i, Z_i)$ satisfying (i) $Z_i \perp\!\!\!\perp (Y_{1i}, Y_{0i})|X_i$ and (ii) $E[Y_{zi}|X_i]$ is continuous at any point $x \in \mathcal{X}$ such that $p^{ML}(x) \in (0, 1)$ and $ML(x) \in \{0, 1\}$ for $z = 0, 1$. This is the set of possible distributions of $(Y_{1i}, Y_{0i}, X_i, Z_i)$ under our framework and Assumption 3. Let $\mathbf{P}$ be the set of all distributions of (Y_i, X_i, Z_i). Let $T : \mathbf{Q} \to \mathbf{P}$ be a function such that, for $Q \in \mathbf{Q}$, $T(Q)$ is the distribution of $(Z_i Y_{1i} + (1 - Z_i)Y_{0i}, X_i, Z_i)$, where the distribution of $(Y_{1i}, Y_{0i}, X_i, Z_i)$ is Q. Let Q_0 and P_0 denote the true distributions of $(Y_{1i}, Y_{0i}, X_i, Z_i)$ and (Y_i, X_i, Z_i), respectively. Given P_0, the identified set of $E[Y_{1i} - Y_{0i}|X_i \in A]$ is given by $\{E_Q[Y_{1i} - Y_{0i}|X_i \in A] : P_0 = T(Q), Q \in \mathbf{Q}\}$, where $E_Q[\cdot]$ is the expectation operator under distribution Q. We show that this set contains two distinct values. In what follows, $\Pr(\cdot)$ and $E[\cdot]$ without a subscript denote the probability and expectation under the true distributions Q_0 and P_0 as up until now.

Now pick any $x^* \in A \cap \text{int}(\mathcal{X}_1)$. Since A and $\text{int}(\mathcal{X}_1)$ are open, there is some $\delta > 0$ such that $N(x^*, \delta) \subset A \cap \text{int}(\mathcal{X}_1)$. Let $\epsilon = \frac{\delta}{2}$, and consider a function $f : \mathcal{X} \to \mathbb{R}$ such that $f(x) = E[Y_{0i}|X = x]$ for all $x \in \mathcal{X} \setminus N(x^*, \epsilon)$ and $f(x) = E[Y_{0i}|X = x] - 1$ for all $x \in N(x^*, \epsilon)$. Below, we show that f is continuous at any point $x \in \mathcal{X}$ such that $p^{ML}(x) \in (0, 1)$ and $ML(x) \in \{0, 1\}$. Pick any $x \in \mathcal{X}$ such that $p^{ML}(x) \in (0, 1)$ and $ML(x) \in \{0, 1\}$. Since $N(x^*, \delta) \subset \text{int}(\mathcal{X}_1)$ and $\text{int}(\mathcal{X}_1) \subset \{x' \in \mathcal{X} : p^{ML}(x') = 1\}$ by Step 3, $x \notin N(x^*, \delta)$. Hence, $N(x, \epsilon) \subset \mathcal{X} \setminus N(x^*, \epsilon)$. By Assumption 3 and the definition of f, f is continuous at x.

Now take any random vector $(Y_{1i}^*, Y_{0i}^*, X_i^*, Z_i^*)$ that is distributed according to the true distribution Q_0. Let Q be the distribution of $(Y_{1i}^Q, Y_{0i}^Q, X_i^Q, Z_i^Q)$, where $(Y_{1i}^Q, X_i^Q, Z_i^Q) = (Y_{1i}^*, X_i^*, Z_i^*)$, and

$$Y_{0i}^Q = \begin{cases} Y_{0i}^* & \text{if } X_i^* \in \mathcal{X} \setminus N(x^*, \epsilon) \\ Y_{0i}^* - 1 & \text{if } X_i^* \in N(x^*, \epsilon) \end{cases}$$

Note first that $Q \in \mathbf{Q}$, since $E_Q[Y_{1i}^Q|X_i^Q = x] = E[Y_{1i}^*|X_i^* = x]$ and $E_Q[Y_{0i}^Q|X_i^Q = x] = f(x)$, where $E[Y_{1i}^*|X_i^*]$ and f are both continuous at any point $x \in \mathcal{X}$ such that $p^{ML}(x) \in (0, 1)$ and $ML(x) \in \{0, 1\}$. Also, $Z_i^Q = Z_i^* = 1$ if $X_i^* \in N(x^*, \epsilon)$. It then follows that

$$\begin{aligned} Y_i^Q &= Z_i^Q Y_{1i}^Q + (1 - Z_i^Q)Y_{0i}^Q \\ &= \begin{cases} Z_i^* Y_{1i}^* + (1 - Z_i^*)Y_{0i}^* & \text{if } X_i^* \in \mathcal{X} \setminus N(x^*, \epsilon) \\ Z_i^* Y_{1i}^* & \text{if } X_i^* \in N(x^*, \epsilon) \end{cases} \end{aligned}$$

and

$$\begin{aligned} Y_i^* &= Z_i^* Y_{1i}^* + (1 - Z_i^*)Y_{0i}^* \\ &= \begin{cases} Z_i^* Y_{1i}^* + (1 - Z_i^*)Y_{0i}^* & \text{if } X_i^* \in \mathcal{X} \setminus N(x^*, \epsilon) \\ Z_i^* Y_{1i}^* & \text{if } X_i^* \in N(x^*, \epsilon) \end{cases} \end{aligned}$$

Thus, $Y_i^Q = Y_i^*$, and hence $T(Q) = T(Q_0) = P_0$.

Using $E_Q[Y_{1i}^Q|X_i^Q = x] = E[Y_{1i}^*|X_i^* = x]$ and $E_Q[Y_{0i}^Q|X_i^Q = x] = f(x)$, we have

$$\begin{aligned}
&E_Q[Y_{1i}^Q - Y_{0i}^Q|X_i^Q \in A] \\
=&E_Q[E_Q[Y_{1i}^Q|X_i^Q]|X_i^Q \in A] \\
&- E_Q[E_Q[Y_{0i}^Q|X_i^Q]|X_i^Q \in A, X_i^Q \notin N(x^*,\epsilon)]Pr_Q(X_i^Q \notin N(x^*,\epsilon)|X_i^Q \in A) \\
&- E_Q[E_Q[Y_{0i}^Q|X_i^Q]|X_i^Q \in N(x^*,\epsilon)]Pr_Q(X_i^Q \in N(x^*,\epsilon)|X_i^Q \in A) \\
=&E[E[Y_{1i}^*|X_i^*]|X_i^* \in A] - E[f(X_i^*)|X_i^* \in A, X_i^* \notin N(x^*,\epsilon)]\Pr(X_i^* \notin N(x^*,\epsilon)|X_i^* \in A) \\
&- E[f(X_i^*)|X_i^* \in N(x^*,\epsilon)]\Pr(X_i^* \in N(x^*,\epsilon)|X_i^* \in A) \\
=&E[Y_{1i}^*|X_i^* \in A] - E[Y_{0i}^*|X_i^* \in A, X_i^* \notin N(x^*,\epsilon)]\Pr(X_i^* \notin N(x^*,\epsilon)|X_i^* \in A) \\
&- E[Y_{0i}^* - 1|X_i^* \in N(x^*,\epsilon)]\Pr(X_i^* \in N(x^*,\epsilon)|X_i^* \in A) \\
=&E[Y_{1i}^* - Y_{0i}^*|X_i^* \in A] + \Pr(X_i^* \in N(x^*,\epsilon)|X_i^* \in A).
\end{aligned}$$

By the definition of support, $\Pr(X_i^* \in N(x^*,\epsilon)) > 0$. Since $T(Q) = T(Q_0) = P_0$ but $E_Q[Y_{1i}^Q - Y_{0i}^Q|X_i^Q \in A] \neq E[Y_{1i}^* - Y_{0i}^*|X_i^* \in A]$, $E[Y_{1i} - Y_{0i}|X_i \in A]$ is not identified. □ □ □

A.2 Proof of Corollary 1

If $\Pr(D_i(1) - D_i(0) = 1|X_i = x) = 1$, $\Pr(Y_{1i} - Y_{0i} = Y_i(1) - Y_i(0)|X_i = x) = 1$, and hence $E[Y_{1i} - Y_{0i}|X_i = x] = E[Y_i(1) - Y_i(0)|X_i = x]$. Then, Parts (a) and (b) follow from Proposition 1.

If $\Pr(D_i(1) \geq D_i(0)|X_i = x) = 1$, we have

$$\begin{aligned}
E[Y_{1i} - Y_{0i}|X_i = x] &= E[(D_i(1) - D_i(0))(Y_i(1) - Y_i(0))|X_i = x] \\
&= \Pr(D_i(1) \neq D_i(0)|X_i = x)E[Y_i(1) - Y_i(0)|D_i(1) \neq D_i(0), X_i = x].
\end{aligned}$$

If in addition $\Pr(D_i(1) \neq D_i(0)|X_i = x) > 0$, we obtain

$$\begin{aligned}
E[Y_i(1) - Y_i(0)|D_i(1) \neq D_i(0), X_i = x] &= \frac{E[Y_{1i} - Y_{0i}|X_i = x]}{\Pr(D_i(1) \neq D_i(0)|X_i = x)} \\
&= \frac{E[Y_{1i} - Y_{0i}|X_i = x]}{E[D_i(1) - D_i(0)|X_i = x]}.
\end{aligned}$$

Then, Part (c) follows from Proposition 1 (a). Part (d) is established by following the procedure used to show Proposition 1 (b). □

A.3 Proof of Proposition 2

Since ML is a $\mathcal{L}^p$-measurable and bounded function, ML is locally integrable with respect to the Lebesgue measure, i.e., for every ball $B \subset \mathbb{R}^p$, $\int_B ML(x)dx$ exists. An application of the Lebesgue differentiation theorem (see e.g. Theorem 1.4 of Stein and Shakarchi (2005)) to the function ML shows that

$$\lim_{\delta \to 0} \frac{\int_{N(x,\delta)} ML(x^*)dx^*}{\int_{N(x,\delta)} dx^*} = ML(x)$$

for almost every $x \in \mathbb{R}^p$. By the definition of the interior, for any $x \in \text{int}(\mathcal{X})$, $N(x,\delta) \subset \mathcal{X}$ so that

$$p^{ML}(x;\delta) = \frac{\int_{N(x,\delta)} ML(x^*)dx^*}{\int_{N(x,\delta)} dx^*}$$

for any sufficiently small $\delta > 0$. Therefore,

$$p^{ML}(x) \equiv \lim_{\delta \to 0} p^{ML}(x;\delta) = ML(x)$$

for almost every $x \in \text{int}(\mathcal{X})$. □

A.4 Proof of Proposition 3

With change of variables $u = \frac{x^* - x}{\delta}$, we have

$$\begin{aligned}
p^{ML}(x;\delta) &= \frac{\int_{N(x,\delta)} 1\{x^* \in \mathcal{X}\} ML(x^*)dx^*}{\int_{N(x,\delta)} 1\{x^* \in \mathcal{X}\}dx^*} \\
&= \frac{\delta^p \int_{N(\mathbf{0},1)} 1\{x + \delta u \in \mathcal{X}\} ML(x+\delta u)du}{\delta^p \int_{N(\mathbf{0},1)} 1\{x + \delta u \in \mathcal{X}\}du} \\
&= \frac{\int_{\cup_{q\in Q}\mathcal{U}_{x,q}} 1\{x + \delta u \in \mathcal{X}\} ML(x+\delta u)du + \int_{N(\mathbf{0},1)\setminus\cup_{q\in Q}\mathcal{U}_{x,q}} 1\{x + \delta u \in \mathcal{X}\} ML(x+\delta u)du}{\int_{N(\mathbf{0},1)} 1\{x + \delta u \in \mathcal{X}\}du} \\
&= \frac{\sum_{q\in Q} \int_{\mathcal{U}_{x,q}} 1\{x + \delta u \in \mathcal{X}\} ML(x+\delta u)du}{\int_{N(\mathbf{0},1)} 1\{x + \delta u \in \mathcal{X}\}du},
\end{aligned}$$

where the last equality follows from the assumption that $\mathcal{L}^p(\cup_{q\in Q}\mathcal{U}_{x,q}) = \mathcal{L}^p(N(\mathbf{0},1))$. Since $x \in \text{int}(\mathcal{X})$, $\lim_{\delta\to 0} 1\{x + \delta u \in \mathcal{X}\} = 1$ for any $u \in N(\mathbf{0},1)$. By the definition of $\mathcal{U}_{x,q}$, for each $q \in Q$, $\lim_{\delta\to 0} ML(x+\delta u) = q$ for any $u \in \mathcal{U}_{x,q}$. By the Dominated Convergence Theorem,

$$\begin{aligned}
p^{ML}(x) &= \lim_{\delta\to 0} p^{ML}(x;\delta) \\
&= \frac{\sum_{q\in Q} q\mathcal{L}^p(\mathcal{U}_{x,q})}{\mathcal{L}^p(N(\mathbf{0},1))}.
\end{aligned}$$

The numerator exists, since $q \leq 1$ for all $q \in Q$ and $\sum_{q\in Q} \mathcal{L}^p(\mathcal{U}_{x,q}) = \mathcal{L}^p(N(\mathbf{0},1))$. □

A.5 Proof of Corollary 2

1. Suppose that ML is continuous at $x \in \text{int}(\mathcal{X})$, and let $q = ML(x)$. Then, by definition, $\mathcal{U}_{x,q} = N(\mathbf{0},1)$. By Proposition 3, $p^{ML}(x)$ exists, and $p^{ML}(x) = q$. □

2. Pick any $x \in \text{int}(\mathcal{X}_q)$. ML is continuous at x, since there exists $\delta > 0$ such that $N(x,\delta) \subset \mathcal{X}_q$ by the definition of interior. By the previous result, $p^{ML}(x)$ exists, and $p^{ML}(x) = q$. □

3. Let $\mathcal{N}$ be the neighborhood of x on which f is continuously differentiable. Let $\nabla f(x^*) = \frac{\partial f(x^*)}{\partial x}$ for $x^* \in \mathcal{N}$. By the mean value theorem, for any sufficiently small $\delta > 0$,

$$\begin{aligned}
f(x+\delta u) &= f(x) + \nabla f(\tilde{x}_\delta) \cdot \delta u \\
&= \nabla f(\tilde{x}_\delta) \cdot \delta u
\end{aligned}$$

for some $\tilde{x}_\delta$ which is on the line segment connecting x and $x+\delta u$. Since $\tilde{x}_\delta \to x$ as $\delta \to 0$ and ∇f is continuous on $\mathcal{N}$, $\nabla f(\tilde{x}_\delta)\cdot u \to \nabla f(x)\cdot u$ as $\delta \to 0$. Therefore, if $\nabla f(x)\cdot u > 0$, then $f(x+\delta u) = \nabla f(\tilde{x}_\delta)\cdot \delta u > 0$ for any sufficiently small $\delta > 0$, and if $\nabla f(x)\cdot u < 0$, then $f(x+\delta u) = \nabla f(\tilde{x}_\delta)\cdot \delta u < 0$ for any sufficiently small $\delta > 0$. We then have

$$\mathcal{U}_x^+ \equiv \{u \in N(\mathbf{0},1) : \nabla f(x)\cdot u > 0\} \subset \mathcal{U}_{x,q_1}$$
$$\mathcal{U}_x^- \equiv \{u \in N(\mathbf{0},1) : \nabla f(x)\cdot u < 0\} \subset \mathcal{U}_{x,q_2}.$$

Let V be the Lebesgue measure of a half p-dimensional unit ball. Since $V = \mathcal{L}^p(\mathcal{U}_x^+) \leq \mathcal{L}^p(\mathcal{U}_{x,q_1})$, $V = \mathcal{L}^p(\mathcal{U}_x^-) \leq \mathcal{L}^p(\mathcal{U}_{x,q_2})$, and $\mathcal{L}^p(\mathcal{U}_{x,q_1}) + \mathcal{L}^p(\mathcal{U}_{x,q_2}) \leq \mathcal{L}^p(N(\mathbf{0},1)) = 2V$, it follows that $\mathcal{L}^p(\mathcal{U}_{x,q_1}) = \mathcal{L}^p(\mathcal{U}_{x,q_2}) = V$. By Proposition 3, $p^{ML}(x)$ exists, and $p^{ML}(x) = \frac{1}{2}(q_1+q_2)$. □

4. We have that $\mathcal{U}_{\mathbf{0},q_1} = \{(u_1,u_2)' \in N(\mathbf{0},1) : u_1 \leq 0 \text{ or } u_2 \leq 0\}$ and $\mathcal{U}_{\mathbf{0},q_2} = \{(u_1,u_2)' \in N(\mathbf{0},1) : u_1 > 0, u_2 > 0\}$. By Proposition 3, $p^{ML}(x)$ exists, and $p^{ML}(x) = \frac{q_1\mathcal{L}^2(\mathcal{U}_{\mathbf{0},q_1}) + q_2\mathcal{L}^2(\mathcal{U}_{\mathbf{0},q_2})}{\mathcal{L}^2(N(\mathbf{0},1))} = \frac{3}{4}q_1 + \frac{1}{4}q_2$. □

A.6 Proof of Theorem 1

Let $\hat{\alpha}_1$ denote the coefficient on Z_i of the following OLS regression using the observations with $p^{ML}(X_i;\delta_n) \in (0,1)$

$$Y_i = \alpha_0(1-I) + \alpha_1 Z_i + \alpha_2 p^{ML}(X_i;\delta_n) + u_i,$$

and $\hat{\gamma}_1$ denote the coefficient on Z_i of the first stage of the 2SLS regression using the observations with $p^{ML}(X_i;\delta_n) \in (0,1)$

$$D_i = \gamma_0(1-I) + \gamma_1 Z_i + \gamma_2 p^{ML}(X_i;\delta_n) + \nu_i.$$

Let $\hat{\alpha}_1^s$ denote the coefficient on Z_i of the following OLS regression using the observations with $p^s(X_i;\delta_n) \in (0,1)$

$$Y_i = \alpha_0(1-I) + \alpha_1 Z_i + \alpha_2 p^s(X_i;\delta_n) + u_i,$$

and $\hat{\gamma}_1^s$ denote the coefficient on Z_i of the first stage of the 2SLS regression using the observations with $p^s(X_i;\delta_n) \in (0,1)$

$$D_i = \gamma_0(1-I) + \gamma_1 Z_i + \gamma_2 p^s(X_i;\delta_n) + \nu_i.$$

It is known that in the just-identified case, the 2SLS estimator coincides with the indirect least squares estimator, which is obtained by solving for the parameter in terms of the OLS estimator of the reduced form (see Khazzoom (1976) for example). It then holds that $\hat{\beta}_1 = \frac{\hat{\alpha}_1}{\hat{\gamma}_1}$ and $\hat{\beta}_1^s = \frac{\hat{\alpha}_1^s}{\hat{\gamma}_1^s}$. The conclusion in Theorem 1 follows from Lemmas 1 – 6 below and from the fact that $E[Y_{1i} - Y_{0i}|X_i] = E[(D_i(1) - D_i(0))(Y_i(1) - Y_i(0))|X_i]$.

Lemma 1. *Suppose that Assumptions 1, 2 and 4 (a) hold, and that* $\Pr(ML(X_i) \in (0,1)) > 0$. *Then*

$$\hat{\alpha}_1 \xrightarrow{p} \frac{E[ML(X_i)(1-ML(X_i))E[Y_{1i} - Y_{0i}|X_i]|ML(X_i) \in (0,1)]}{E[ML(X_i)(1-ML(X_i))|ML(X_i) \in (0,1)]}$$

and

$$\hat{\gamma}_1 \xrightarrow{p} \frac{E[ML(X_i)(1-ML(X_i))E[D_i(1)-D_i(0)|X_i]|ML(X_i)\in(0,1)]}{E[ML(X_i)(1-ML(X_i))|ML(X_i)\in(0,1)]}$$

as $n \to \infty$.

Lemma 2. *Suppose that Assumptions 1 and 4 (c) – (h) hold, and that* $\Pr(ML(X_i) \in (0,1)) = 0$. *If* $n\delta_n \to \infty$ *and* $\delta_n \to 0$, *then*

$$\hat{\alpha}_1 \xrightarrow{p} \frac{\int_{\partial\Omega^* \cap \mathcal{X}} E[Y_{1i} - Y_{0i}|X_i = x] f_X(x) d\mathcal{H}^{p-1}(x)}{\int_{\partial\Omega^* \cap \mathcal{X}} f_X(x) d\mathcal{H}^{p-1}(x)}$$

and

$$\hat{\gamma}_1 \xrightarrow{p} \frac{\int_{\partial\Omega^* \cap \mathcal{X}} E[D_i(1) - D_i(0)|X_i = x] f_X(x) d\mathcal{H}^{p-1}(x)}{\int_{\partial\Omega^* \cap \mathcal{X}} f_X(x) d\mathcal{H}^{p-1}(x)}.$$

as $n \to \infty$.

Lemma 3. *Suppose that Assumptions 2 and 4 (a) hold, and that* $\Pr(ML(X_i) \in (0,1)) > 0$. *Then* $\hat{\alpha}_1^s = \hat{\alpha}_1 + o_p(1)$ *and* $\hat{\gamma}_1^s = \hat{\gamma}_1 + o_p(1)$ *as* $n \to \infty$.

Lemma 4. *Suppose that Assumption 4 (c) – (h) hold, and that* $\Pr(ML(X_i) \in (0,1)) = 0$. *If* $n\delta_n \to \infty$, $\delta_n \to 0$ *and* $S_n \to \infty$, *then* $\hat{\alpha}_1^s = \hat{\alpha}_1 + o_p(1)$ *and* $\hat{\gamma}_1^s = \hat{\gamma}_1 + o_p(1)$ *as* $n \to \infty$.

Lemma 5. *Suppose that Assumptions 2 and 4 (a) hold, and that* $\Pr(ML(X_i) \in (0,1)) > 0$. *Then*

$$\begin{aligned} &\lim_{\delta \to 0} E[p^{ML}(X_i;\delta)(1-p^{ML}(X_i;\delta))(Y_{1i} - Y_{0i})] \\ =\ & E[ML(X_i)(1-ML(X_i))E[Y_{1i} - Y_{0i}|X_i]|ML(X_i) \in (0,1)] \Pr(ML(X_i) \in (0,1)) \end{aligned}$$

and

$$\begin{aligned} &\lim_{\delta \to 0} E[p^{ML}(X_i;\delta)(1-p^{ML}(X_i;\delta))(D_i(1) - D_i(0))] \\ =\ & E[ML(X_i)(1-ML(X_i))E[D_i(1) - D_i(0)|X_i]|ML(X_i) \in (0,1)] \Pr(ML(X_i) \in (0,1)). \end{aligned}$$

Lemma 6. *Suppose that Assumption 4 (c) – (h) hold, and that* $\Pr(ML(X_i) \in (0,1)) = 0$. *Then*

$$\begin{aligned} &\lim_{\delta \to 0} \delta^{-1} E[p^{ML}(X_i;\delta)(1-p^{ML}(X_i;\delta))(Y_{1i} - Y_{0i})] \\ =\ & C \int_{\partial\Omega^* \cap \mathcal{X}} E[Y_{1i} - Y_{0i}|X_i = x] f_X(x) d\mathcal{H}^{p-1}(x) \end{aligned}$$

and

$$\begin{aligned} &\lim_{\delta \to 0} \delta^{-1} E[p^{ML}(X_i;\delta)(1-p^{ML}(X_i;\delta))(D_i(1) - D_i(0))] \\ =\ & C \int_{\partial\Omega^* \cap \mathcal{X}} E[D_i(1) - D_i(0)|X_i = x] f_X(x) d\mathcal{H}^{p-1}(x) \end{aligned}$$

for some common constant $C > 0$.

A.6.1 Proof of Lemma 1

We only present how we derive the probability limit of $\hat{\alpha}_1$. The probability limit of $\hat{\gamma}_1$ can be derived in the same way. Let $W_i = (1, Z_i, p^{ML}(X_i;\delta_n))'$, $W_i^{nc} = (Z_i, p^{ML}(X_i;\delta_n))'$ and $I_i^p = 1\{p^{ML}(X_i;\delta_n) \in (0,1)\}$. The OLS estimator $\tilde{\alpha} = (\tilde{\alpha}_0, \tilde{\alpha}_1, \tilde{\alpha}_2)'$ of regression of Y_i on 1, Z_i and $p^{ML}(X_i;\delta_n)$ is given by

$$\tilde{\alpha} = (\sum_{i=1}^n W_i W_i' I_i^p)^{-1} \sum_{i=1}^n W_i Y_i I_i^p,$$

and the OLS estimator $\tilde{\alpha}^{nc} = (\tilde{\alpha}_1^{nc}, \tilde{\alpha}_2^{nc})'$ of regression of Y_i on Z_i and $p^{ML}(X_i;\delta_n)$ is given by

$$\tilde{\alpha}^{nc} = (\sum_{i=1}^n W_i^{nc}(W_i^{nc})' I_i^p)^{-1} \sum_{i=1}^n W_i^{nc} Y_i I_i^p.$$

Below we consider two cases: 1. $\text{Var}(ML(X_i)|ML(X_i) \in (0,1)) > 0$ and 2. $\text{Var}(ML(X_i)|ML(X_i) \in (0,1)) = 0$.

Case 1. We first consider the case where $\text{Var}(ML(X_i)|ML(X_i) \in (0,1)) > 0$. The OLS estimator $\hat{\alpha}_1$ is given by

$$\hat{\alpha}_1 = \tilde{\alpha}_1(1 - I) + \tilde{\alpha}_1^{nc} I.$$

We derive the probability limit of $\tilde{\alpha}_1$ and $\tilde{\alpha}_1^{nc}$. Let $\tilde{\mathcal{X}} = \{x \in \mathcal{X} : ML(x) \in (0,1)\}$, and $\mathcal{X}_c = \{x \in \mathcal{X} : ML \text{ is continuous at } x\}$.

Claim 1. *$p^{ML}(x) = ML(x)$ and $\lim_{\delta\to 0} 1\{p^{ML}(x;\delta) \in (0,1)\} = 1\{p^{ML}(x) \in (0,1)\}$ for almost every $x \in \mathcal{X}$.*

Proof. By Assumption 2, $x \in \text{int}(\mathcal{X}_0) \cup \text{int}(\mathcal{X}_1) \cup (\tilde{\mathcal{X}} \cap \mathcal{X}_c)$ for almost every $x \in \mathcal{X}$. Pick any $x \in \text{int}(\mathcal{X}_k)$ for some $k \in \{0,1\}$. By the definition of interior, $N(x,\delta) \subset \mathcal{X}_k$ for any sufficiently small $\delta > 0$, and hence $p^{ML}(x;\delta) = k$ for any sufficiently small $\delta > 0$. Therefore, $p^{ML}(x) = ML(x) = k$ and $\lim_{\delta\to 0} 1\{p^{ML}(x;\delta) \in (0,1)\} = 1\{p^{ML}(x) \in (0,1)\}$.

Now pick any $x \in \tilde{\mathcal{X}} \cap \mathcal{X}_c$. By Part 1 of Corollary 2, $p^{ML}(x) = ML(x)$. Since $ML(x) \in (0,1)$ and ML is continuous at x, $N(x,\delta) \subset \tilde{\mathcal{X}}$ for any sufficiently small $\delta > 0$, and hence $p^{ML}(x;\delta) \in (0,1)$ for any sufficiently small $\delta > 0$. Therefore, $\lim_{\delta\to 0} 1\{p^{ML}(x;\delta) \in (0,1)\} = 1 = 1\{p^{ML}(x) \in (0,1)\}$. □

Claim 2. *Let $\{V_i\}_{i=1}^\infty$ be i.i.d. random variables such that $E[V_i^2] < \infty$. Then,*

$$\frac{1}{n}\sum_{i=1}^n V_i p^{ML}(X_i;\delta_n)^l I_i^p \xrightarrow{p} E[V_i ML(X_i)^l | ML(X_i) \in (0,1)] \Pr(ML(X_i) \in (0,1))$$

for $l = 0, 1, 2$.

Proof. We show that $E[\frac{1}{n}\sum_{i=1}^n V_i p^{ML}(X_i;\delta_n)^l I_i^p] \to E[V_i ML(X_i)^l | ML(X_i)\in(0,1)]\Pr(ML(X_i)\in(0,1))$ and that $\mathrm{Var}(\frac{1}{n}\sum_{i=1}^n V_i p^{ML}(X_i;\delta_n)^l I_i^p)\to 0$ for $l=0,1,2$. For the first part, we have

$$\begin{aligned}
E[\frac{1}{n}\sum_{i=1}^n V_i p^{ML}(X_i;\delta_n)^l I_i^p] &= E[V_i p^{ML}(X_i;\delta_n)^l I_i^p] \\
&= \int_{\mathcal{X}} E[V_i|X_i=x] p^{ML}(x;\delta_n)^l 1\{p^{ML}(x;\delta_n)\in(0,1)\} f_X(x)dx \\
&\to \int_{\mathcal{X}} E[V_i|X_i=x] ML(x)^l 1\{ML(x)\in(0,1)\} f_X(x)dx \\
&= E[V_i ML(X_i)^l 1\{ML(X_i)\in(0,1)\}] \\
&= E[V_i ML(X_i)^l | ML(X_i)\in(0,1)]\Pr(ML(X_i)\in(0,1)),
\end{aligned}$$

where the convergence follows by Claim 1 and by the Dominated Convergence Theorem.

As for variance,

$$\begin{aligned}
\mathrm{Var}[\frac{1}{n}\sum_{i=1}^n V_i p^{ML}(X_i;\delta_n)^l I_i^p] &\le \frac{1}{n} E[V_i^2 p^{ML}(X_i;\delta_n)^{2l} I_i^{p2}] \\
&\le \frac{1}{n} E[V_i^2] \\
&\to 0.
\end{aligned}$$

□

Let I_i^{ML} denote the event $\{ML(X_i)\in(0,1)\}$. By Claim 2, we obtain

$$\tilde{\alpha} \xrightarrow{p} S_W^{-1} S_Y,$$

where

$$S_W = \begin{bmatrix} 1 & E[ML(X_i)|I_i^{ML}] & E[ML(X_i)|I_i^{ML}] \\ E[ML(X_i)|I_i^{ML}] & E[ML(X_i)|I_i^{ML}] & E[ML(X_i)^2|I_i^{ML}] \\ E[ML(X_i)|I_i^{ML}] & E[ML(X_i)^2|I_i^{ML}] & E[ML(X_i)^2|I_i^{ML}] \end{bmatrix},$$

and $S_Y = (E[Y_i|I_i^{ML}] \;\; E[Z_iY_i|I_i^{ML}] \;\; E[ML(X_i)Y_i|I_i^{ML}])'$. Also,

$$\tilde{\alpha}^{nc} \xrightarrow{p} (S_W^{nc})^{-1} S_Y^{nc},$$

where

$$S_W^{nc} = \begin{bmatrix} E[ML(X_i)|I_i^{ML}] & E[ML(X_i)^2|I_i^{ML}] \\ E[ML(X_i)^2|I_i^{ML}] & E[ML(X_i)^2|I_i^{ML}] \end{bmatrix},$$

and $S_Y^{nc} = (E[Z_iY_i|I_i^{ML}] \;\; E[ML(X_i)Y_i|I_i^{ML}])'$.

The determinant of S_W is $\mathrm{Var}(ML(X_i)|I_i^{ML})E[ML(X_i)(1-ML(X_i))|I_i^{ML}]>0$. Also, the determinant of S_W^{nc} is $E[ML(X_i)^2|I_i^{ML}]E[ML(X_i)(1-ML(X_i))|I_i^{ML}]>0$. After a few lines of

algebra, we have that both $\tilde{\alpha}_1$ and $\tilde{\alpha}_1^{nc}$ converge in probability to

$$
\begin{aligned}
&\frac{E[Z_iY_i|I_i^{ML}] - E[ML(X_i)Y_i|I_i^{ML}]}{E[ML(X_i)(1-ML(X_i))|I_i^{ML}]} \\
=&\frac{E[Z_iY_{1i}|I_i^{ML}] - E[ML(X_i)(Z_iY_{1i}+(1-Z_i)Y_{0i})|I_i^{ML}]}{E[ML(X_i)(1-ML(X_i))|I_i^{ML}]} \\
=&\frac{E[ML(X_i)E[Y_{1i}|X_i]|I_i^{ML}] - E[ML(X_i)(ML(X_i)E[Y_{1i}|X_i]+(1-ML(X_i))E[Y_{0i}|X_i])|I_i^{ML}]}{E[ML(X_i)(1-ML(X_i))|I_i^{ML}]} \\
=&\frac{E[ML(X_i)(1-ML(X_i))E[Y_{1i}-Y_{0i}|X_i]|I_i^{ML}]}{E[ML(X_i)(1-ML(X_i))|I_i^{ML}]},
\end{aligned}
$$

where the second last equality follows by the fact that (Y_{1i}, Y_{0i}) is independent of Z_i conditional on X_i. It then follows that

$$
\begin{aligned}
\hat{\alpha}_1 &= \tilde{\alpha}_1(1-I) + \tilde{\alpha}_1^{nc} I \\
&= \tilde{\alpha}_1 + (\tilde{\alpha}_1^{nc} - \tilde{\alpha}_1) I \\
&\xrightarrow{p} \frac{E[ML(X_i)(1-ML(X_i))E[Y_{1i}-Y_{0i}|X_i]|I_i^{ML}]}{E[ML(X_i)(1-ML(X_i))|I_i^{ML}]}.
\end{aligned}
$$

Case 2. We next consider the case where $\Pr(ML(X_i) \in (0,1)) > 0$ and $\text{Var}(ML(X_i)|ML(X_i) \in (0,1)) = 0$. Since $\text{Var}(ML(X_i)|ML(X_i) \in (0,1)) = 0$, there exists $q \in (0,1)$ such that $\Pr(ML(X_i) = q | ML(X_i) \in (0,1)) = 1$. It follows that

$$
\begin{aligned}
\Pr(I=0) &= \Pr(ML(X_i) \in \{0,1\} \text{ for all } i = 1,...,n) \\
&\quad + \Pr(ML(X_i) = q' \text{ and } ML(X_j) = q'' \text{ for some } q', q'' \in (0,1) \text{ with } q' \neq q'' \\
&\qquad\qquad \text{for some } i, j \in \{1,...,n\}) \\
&= \Pr(ML(X_i) \in \{0,1\} \text{ for all } i = 1,...,n) \\
&= (1 - \Pr(ML(X_i) \in (0,1)))^n,
\end{aligned}
$$

which converges to zero as $n \to \infty$, since $\Pr(ML(X_i) \in (0,1)) > 0$. Therefore, $\hat{\alpha}_1 = \tilde{\alpha}_1^{nc}$ with probability approaching one, and $\hat{\alpha}_1 \xrightarrow{p} \frac{E[ML(X_i)(1-ML(X_i))E[Y_{1i}-Y_{0i}|X_i]|I_i^{ML}]}{E[ML(X_i)(1-ML(X_i))|I_i^{ML}]}$. □

A.6.2 Proof of Lemma 2

We only present how we derive the probability limit of $\hat{\alpha}_1$. The probability limit of $\hat{\gamma}_1$ can be derived in the same way. Since $\Pr(ML(X_i) \in (0,1)) = 0$, $I = 0$ with probability one. Let $W_i = (1, Z_i, p^{ML}(X_i;\delta_n))'$ and $I_i^p = 1\{p^{ML}(X_i;\delta_n) \in (0,1)\}$. The OLS estimator $\hat{\alpha} = (\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2)'$ of regression of Y_i on 1, Z_i and $p^{ML}(X_i;\delta_n)$ is given by

$$
\hat{\alpha} = (\sum_{i=1}^n W_i W_i' I_i^p)^{-1} \sum_{i=1}^n W_i Y_i I_i^p.
$$

We first introduce some notations. Note that by Assumption 4 (d), Ω^* is twice continuouslly differentiable. For each $x \in \partial\Omega^*$, we denote by $T(x) \subset \mathbb{R}^p$ the tangent space of $\partial\Omega^*$ at x and by $\nu(x) \in \mathbb{R}$ the inward unit normal vector of $\partial\Omega^*$ at x, that is, the unit vector orthogonal to all vectors in $T(x)$ that points toward the inside of Ω^* (see e.g. the summary in Crasta and Malusa (2007)). For a set $A \subset \mathbb{R}^p$, let $d_A^s : \mathbb{R}^p \to \mathbb{R}$ be the signed distance function of A, defined by

$$d_A^s(x) = \begin{cases} d(x, \partial A) & \text{if } x \in \mathrm{cl}(A) \\ -d(x, \partial A) & \text{if } x \in \mathbb{R}^p \setminus \mathrm{cl}(A), \end{cases}$$

where $d(x, B) = \inf_{y \in B} \|y - x\|$ for any $x \in \mathbb{R}^p$ for a set $B \subset \mathbb{R}^p$, and $\mathrm{cl}(A)$ denotes the closure of A. In addition, let $N(B, \delta) = \{x \in \mathbb{R}^p : d(x, B) < \delta\}$ for $B \subset \mathbb{R}^p$ and $\delta > 0$.

Claim 3. *There exists $\bar{\mu} > 0$ such that $d^s_{\Omega^*}$ is twice continuously differentiable on $N(\partial\Omega^*, \bar{\mu})$. Moreover, $\|\nabla d^s_{\Omega^*}(x)\| = 1$ for all $x \in N(\partial\Omega^*, \bar{\mu})$, and $\nabla d^s_{\Omega^*}(x) = \nu(x)$ for all $x \in \partial\Omega^*$, where $\nabla d^s_{\Omega^*}(x) = (\frac{\partial d^s_{\Omega^*}(x)}{\partial x_1}, ..., \frac{\partial d^s_{\Omega^*}(x)}{\partial x_p})'$.*

Proof. Let $\mu = \frac{1}{2} \min_{m,m' \in \{1,...,M\}, m \neq m'} \mathrm{dist}(\Omega^*_m, \Omega^*_{m'})$ so that $\{N(\partial\Omega^*_m, \mu)\}_{m=1}^M$ is a partition of $N(\partial\Omega^*, \mu)$. Note that for every $m \in \{1, ..., M\}$, $d^s_{\Omega^*}(x) = d^s_{\Omega^*_m}(x)$ for any $x \in N(\partial\Omega^*_m, \mu)$. Let $C = \{y \in \mathbb{R}^p : \|y\| \leq 1\}$. It holds that $\|x\| = x \cdot \frac{x}{\|x\|} = \sup_{y \in C}(x \cdot y)$ for any $x \in \mathbb{R}^p$. By Theorem 4.16 of Crasta and Malusa (2007), for every $m \in \{1, ..., M\}$, there exists $\bar{\mu}_m > 0$ such that $d^s_{\Omega^*_m}$ is twice continuously differentiable on $N(\partial\Omega^*_m, \mu_m)$. Letting $\bar{\mu} = \min\{\mu, \mu_1, ..., \mu_M\}$, we have that $d^s_{\Omega^*}$ is twice continuously differentiable on $N(\partial\Omega^*, \bar{\mu})$.

Now note that, for every $x \in N(\partial\Omega^*, \bar{\mu})$, there exists a unique $\pi(x) \in \partial\Omega^*$ such that $d(x, \partial\Omega^*) = \|\pi(x) - x\|$ as stated in the proof of Theorem 4.16 of Crasta and Malusa (2007). By Lemma 4.3 and the second part of Theorem 4.16 of Crasta and Malusa (2007), for every $x \in N(\partial\Omega^*, \bar{\mu})$

$$\begin{aligned} \nabla d^s_{\Omega^*}(x) &= \frac{\nu(\pi(x))}{\inf\{t \geq 0 : \nu(\pi(x)) = ty \text{ for some } y \in C\}} \\ &= \nu(\pi(x)). \end{aligned}$$

Therefore, $\|\nabla d^s_{\Omega^*}(x)\| = 1$ for all $x \in N(\partial\Omega^*, \bar{\mu})$, and $\nabla d^s_{\Omega^*}(x) = \nu(x)$ for all $x \in \partial\Omega^*$. □

Claim 4. *$\partial\Omega^*$ is $(p-1)$-dimensional continuously differentiable submanifold of $\mathbb{R}^p$, that is, for every point $x \in \partial\Omega^*$, there exist an open neighborhood $V \subset \mathbb{R}^p$ of x and a one-to-one continuously differentiable function ϕ from an open set $U \subset \mathbb{R}^{p-1}$ to $\mathbb{R}^p$ such that the Jacobian matrix $J_\phi(u)$ is of rank $p-1$ for all $u \in U$, and $\phi(U) = V \cap \partial\Omega^*$.*

Proof. Let $\bar{\mu}$ be the one given by Claim 3. Fix any $x^* \in \partial\Omega^*$. By Claim 3, $\nabla d^s_{\Omega^*}(x^*)$ is nonzero. Without loss of generality, let $\frac{\partial d^s_{\Omega^*}(x^*)}{\partial x_p} \neq 0$. Let $\psi : \mathbb{R}^p \to \mathbb{R}^p$ be the function such that $\psi(x) = (x_1, ..., x_{p-1}, d^s_{\Omega^*}(x))$. The Jacobian matrix of ψ at x^* is given by

$$J_\psi(x^*) = \begin{pmatrix} \frac{\partial \psi_1}{\partial x_1}(x^*) & \cdots & \frac{\partial \psi_1}{\partial x_p}(x^*) \\ \vdots & \ddots & \vdots \\ \frac{\partial \psi_p}{\partial x_1}(x^*) & \cdots & \frac{\partial \psi_p}{\partial x_p}(x^*) \end{pmatrix} = \begin{pmatrix} & & & 0 \\ & I_{p-1} & & \vdots \\ & & & 0 \\ \frac{\partial d^s_{\Omega^*}(x^*)}{\partial x_1} & \cdots & \frac{\partial d^s_{\Omega^*}(x^*)}{\partial x_{p-1}} & \frac{\partial d^s_{\Omega^*}(x^*)}{\partial x_p} \end{pmatrix},$$

where I_{p-1} is the $(p-1)\times(p-1)$ identify matrix. Since $\frac{\partial d^s_{\Omega^*}(x^*)}{\partial x_p} \neq 0$, the Jacobian matrix is invertible. By the Inverse Function Theorem, there exist an open set V containing x^* and an open set W containing $\psi(x^*)$ such that $\psi : V \to W$ has an inverse function $\psi^{-1} : W \to V$ that is continuously differentiable. We make V small enough so that $\frac{\partial d^s_{\Omega^*}(x)}{\partial x_p} \neq 0$ for every $x \in V$. The Jacobian matrix of ψ^{-1} is given by $J_{\psi^{-1}}(y) = J_\psi(\psi^{-1}(y))^{-1}$ for all $y \in W$.

Now note that $\psi(x) = (x_1, ..., x_{p-1}, 0)$ for all $x \in V \cap \partial\Omega^*$ by the definition of $d^s_{\Omega^*}$. Let $U = \{(x_1, ..., x_{p-1}) \in \mathbb{R}^{p-1} : x \in V \cap \partial\Omega^*\}$ and $\phi : U \to \mathbb{R}^p$ such that $\phi(u) = \psi^{-1}((u, 0))$ for all $u \in U$. Below we verify that ϕ is one-to-one and continously differentiable, that $J_\phi(u)$ is of rank $p-1$ for all $u \in U$, that $\phi(U) = V \cap \partial\Omega^*$, and that U is open.

First, ϕ is one-to-one, since ψ^{-1} is one-to-one, and $(u, 0) \neq (u', 0)$ if $u \neq u'$. Second, ϕ is continuously differentiable, since ψ^{-1} is so. The Jacobian matrix of ϕ at $u \in U$ is by definition

$$J_\phi(u) = \begin{pmatrix} \frac{\partial \psi_1^{-1}}{\partial y_1}((u,0)) & \cdots & \frac{\partial \psi_1^{-1}}{\partial y_{p-1}}((u,0)) \\ \vdots & \ddots & \vdots \\ \frac{\partial \psi_p^{-1}}{\partial y_1}((u,0)) & \cdots & \frac{\partial \psi_p^{-1}}{\partial y_{p-1}}((u,0)) \end{pmatrix}.$$

Note that this is the left $p \times (p-1)$ submatrix of $J_{\psi^{-1}}((u, 0))$. Since $J_{\psi^{-1}}((u, 0))$ has full rank, $J_\phi(u)$ is of rank $p-1$. Moreover,

$$\begin{aligned} \phi(U) &= \{\psi^{-1}((u, 0)) : u \in U\} \\ &= \{\psi^{-1}((x_1, ..., x_{p-1}, 0)) : x \in V \cap \partial\Omega^*\} \\ &= \{\psi^{-1}(\psi(x)) : x \in V \cap \partial\Omega^*\} \\ &= V \cap \partial\Omega^*. \end{aligned}$$

Lastly, we show that U is open. Pick any $\bar{u} \in U$. Then, there exists $\bar{x}_p \in \mathbb{R}$ such that $(\bar{u}, \bar{x}_p) \in V \cap \partial\Omega^*$. As $(\bar{u}, \bar{x}_p) \in V \cap \partial\Omega^*$, $d^s_{\Omega^*}((\bar{u}, \bar{x}_p)) = 0$. Since $\frac{\partial d^s_{\Omega^*}((\bar{u}, \bar{x}_p))}{\partial x_p} \neq 0$, it follows by the Implicit Function Theorem that there exist an open set $S \subset \mathbb{R}^{p-1}$ containing $\bar{u}$ and a continuously differentiable function $g : S \to \mathbb{R}$ such that $g(\bar{u}) = \bar{x}_p$ and $d^s_{\Omega^*}(u, g(u)) = 0$ for all $u \in S$. Since g is continuous, $(\bar{u}, g(\bar{u})) \in V$ and V is open, there exists an open set $S' \subset S$ containing $\bar{u}$ such that $(u, g(u)) \in V$ for all $u \in S'$. By the definition of $d^s_{\Omega^*}$, $d^s_{\Omega^*}(x) = 0$ if and only if $x \in \partial\Omega^*$. Therefore, if $u \in S'$, $(u, g(u))$ must be contained by $\partial\Omega^*$, for otherwise $d^s_{\Omega^*}(u, g(u)) \neq 0$, which is a contradiction. Thus, $(u, g(u)) \in V \cap \partial\Omega^*$ and hence $u \in U$ for all $u \in S'$. This implies that S' is an open subset of U containing $\bar{u}$, which proves that U is open. □

Let $\pi : N(\partial\Omega^*, \bar{\mu}) \to \partial\Omega^*$ be a function such that for every $x \in N(\partial\Omega^*, \bar{\mu})$, $y = \pi(x)$ is the unique point in $\partial\Omega^*$ such that $d(x, \partial\Omega^*) = \|y - x\|$ as defined in the proof of Claim 3. Let $\Gamma(\lambda) = \{x^* \in \mathbb{R}^p : d^s_{\Omega^*}(x^*) = \lambda\}$ for $\lambda \in \mathbb{R}$. We denote by $D^2 d^s_{\Omega^*}(x)$ the $p \times p$ matrix whose (i, j) element is $\frac{\partial^2 d^s_{\Omega^*}(x)}{\partial x_i \partial x_j}$.

Claim 5. *For any $\delta \in (0, \bar{\mu})$ and any function $g : \mathbb{R}^p \to \mathbb{R}$ that is integrable on $N(\partial\Omega^*, \delta)$,*

$$\int_{N(\partial\Omega^*, \delta)} g(x)dx = \int_{-\delta}^{\delta} \int_{\partial\Omega^*} g(u + \lambda\nu(u)) \det(I_{p-1} + \lambda V_u' D^2 d^s_{\Omega^*}(u) V_u) d\mathcal{H}^{p-1}(u) d\lambda,$$

where for every $u \in \partial\Omega^*$, V_u is an $p \times (p-1)$ matrix whose columns are an orthonormal basis of the tangent space $T(u)$, and $\det(I_{p-1} + \lambda V_u' D^2 d^s_{\Omega^*}(u) V_u)$ does not depend on the choice of the orthonormal basis.

Proof. For any $\delta \in (0, \bar{\mu})$,

$$\begin{aligned}
\int_{N(\partial\Omega^*,\delta)} g(x)dx &= \int_{\{x^* \in \mathbb{R}^p : d^s_{\Omega^*}(x^*) \in (-\delta,\delta)\}} g(x) \|\nabla d^s_{\Omega^*}(x)\| dx \\
&= \int_{\{x^* \in \mathbb{R}^p : d^s_{\Omega^*}(x^*) \in (-\delta,\delta)\}} g(x) (\det(\nabla d^s_{\Omega^*}(x)' \nabla d^s_{\Omega^*}(x)))^{1/2} dx \\
&= \int_{\mathbb{R}} \int_{\{x^* \in \mathbb{R}^p : d^s_{\Omega^*}(x^*) \in (-\delta,\delta), d^s_{\Omega^*}(x^*) = \lambda\}} g(x) d\mathcal{H}^{p-1}(x) d\lambda \\
&= \int_{-\delta}^{\delta} \int_{\Gamma(\lambda)} g(x) d\mathcal{H}^{p-1}(x) d\lambda,
\end{aligned}$$

where the first equality follows from Claim 3, and the third equality follows from the coarea formula (Corollary 5.2.6 of Krantz and Parks (2008)).

We next apply Theorem 5.3.7 of Krantz and Parks (2008). For a fixed $\lambda \in (-\delta, \delta)$, consider a function $f_\lambda : N(\partial\Omega^*, \delta) \to \mathbb{R}^p$ such that $f_\lambda(x) = x + \lambda \nabla d^s_{\Omega^*}(x)$. By construction, $d^s_{\Omega^*}(f_\lambda(x)) = \lambda$ for all $x \in \partial\Omega^*$ (see Remark 4.5 and Proposition 4.6 of Crasta and Malusa (2007)), and hence $f_\lambda(\partial\Omega^*) \subset \Gamma(\lambda)$. It also holds that $\Gamma(\lambda) \subset f_\lambda(\partial\Omega^*)$, since $x = f_\lambda(\pi(x))$ for all $x \in \Gamma(\lambda)$. Therefore, $f_\lambda(\partial\Omega^*) = \Gamma(\lambda)$.

Since $\nabla d^s_{\Omega^*}$ is differentiable by Claim 3, f_λ is differentiable. For every $x \in N(\partial\Omega^*, \delta)$, let $D_x f_\lambda : \mathbb{R}^p \to \mathbb{R}^p$ denote the differential of f_λ at x, i.e., $D_x f_\lambda$ is a linear map such that

$$\lim_{v \to \mathbf{0}} \frac{\|f_\lambda(x+v) - f_\lambda(x) - D_x f_\lambda(v)\|}{\|v\|} = 0.$$

Let $D_x^{T_x} f_\lambda : T_x \to \mathbb{R}^p$ denote the restriction of $D_x f_\lambda$ to T_x. Choose an arbitrary orthonormal basis of T_x, $\{v_1, ..., v_{p-1}\} \in \mathbb{R}^{p \times (p-1)}$, and let $\{e_1, ..., e_p\} \in \mathbb{R}^{p \times p}$ be the standard basis of $\mathbb{R}^p$. With these two bases, the linear map $D_x^{T_x} f_\lambda$ is represented by the $p \times (p-1)$ matrix $(I_p + \lambda \nabla^2 d^s_{\Omega^*}(x))V$, where V is the $p \times (p-1)$ matrix having $v_1, ..., v_{p-1}$ as its columns. By Definition 5.3.3 of Krantz and Parks (2008), the $(p-1)$-dimensional Jacobian of f_λ relative to $\partial\Omega^*$ at x is given by

$$\begin{aligned}
J^{\partial\Omega^*}_{p-1} f_\lambda(x) = \sup\Big\{ &\frac{\mathcal{H}^{p-1}(D_x^{T_x} f_\lambda(P))}{\mathcal{H}^{p-1}(P)} : \\
&P \text{ is a } (p-1)\text{-dimensional parallelepiped contained in } T_x \Big\}.
\end{aligned}$$

Let $P \subset T_x$ be a parallelepiped determined by $v_1, ..., v_{p-1}$. Then $D_x^{T_x} f_\lambda(P)$ is the parallelepiped determined by the columns of $(I_p + \lambda \nabla^2 d^s_{\Omega^*}(x))V$. By Proposition 5.1.2 and Lemma 5.3.5 of Krantz and Parks (2008), we have that

$$\begin{aligned}
J^{\partial\Omega^*}_{p-1} f_\lambda(x) &= \frac{(\det(V'(I_p + \lambda \nabla^2 d^s_{\Omega^*}(x))^2 V))^{1/2}}{(\det(V'V))^{1/2}} \\
&= (\det(V'(I_p + \lambda \nabla^2 d^s_{\Omega^*}(x))^2 V))^{1/2} \\
&= (\det((I_{p-1} + \lambda V' \nabla^2 d^s_{\Omega^*}(x) V)^2))^{1/2} \\
&= \det(I_{p-1} + \lambda V' \nabla^2 d^s_{\Omega^*}(x) V),
\end{aligned}$$

where we use the fact that $V'V = I_{p-1}$. Moreover, $J^{\partial\Omega^*}_{p-1} f_\lambda(x)$ does not depend on the choice of the orthonormal basis $\{v_1, ..., v_{p-1}\}$ by Lemma 5.3.5 of Krantz and Parks (2008).

By Claim 4, $\partial\Omega^*$ is $(p-1)$-dimensional continuously differentiable submanifold of $\mathbb{R}^p$. Applying Theorem 5.3.7 of Krantz and Parks (2008), we have that

$$\begin{aligned}
&\int_{\partial\Omega^*} g(f_\lambda(u)) J^{\partial\Omega^*}_{p-1} f_\lambda(u) d\mathcal{H}^{p-1}(u) \\
&= \int_{\mathbb{R}^p} \int_{\{u\in\partial\Omega^*: f_\lambda(u)=x\}} g(f_\lambda(u)) d\mathcal{H}^0(u) d\mathcal{H}^{p-1}(x) \\
&= \int_{\Gamma(\lambda)} \int_{\{u\in\partial\Omega^*: f_\lambda(u)=x\}} g(f_\lambda(u)) d\mathcal{H}^0(u) d\mathcal{H}^{p-1}(x) \\
&= \int_{\Gamma(\lambda)} g(x) d\mathcal{H}^{p-1}(x),
\end{aligned}$$

where the second equality holds since $f_\lambda(\partial\Omega^*) = \Gamma(\lambda)$, the last equality follows from the fact that $\mathcal{H}^0$ is equivalent to the counting measure and that f_λ is bijective on $\Gamma(\lambda)$ so that $|\{u \in \partial\Omega^* : f_\lambda(u) = x\}| = 1$ for all $x \in \Gamma(\lambda)$. □

Claim 6. *For any $\delta > 0$, $\{x \in \mathcal{X} : p^{ML}(x;\delta) \in (0,1)\} \subset N(\partial\Omega^*, \delta)$.*

Proof. Pick any $\delta > 0$ and any $x \in \mathcal{X}\backslash N(\partial\Omega^*, \delta)$. Since $x \in \mathcal{X}\backslash N(\partial\Omega^*, \delta)$, $N(x,\delta)\cap\partial\Omega^* = \emptyset$, and hence either $N(x,\delta) \subset \text{int}(\Omega^*)$ or $N(x,\delta) \subset \text{int}(\mathbb{R}^p \setminus \Omega^*)$. If $N(x,\delta) \subset \text{int}(\Omega^*)$, $p^{ML}(x;\delta) = 1$.

Suppose that $N(x,\delta) \subset \text{int}(\mathbb{R}^p \setminus \Omega^*)$. Since $\Pr(ML(X_i) \in (0,1)) = 0$, $ML(x') \in \{0,1\}$ for almost every $x' \in \mathcal{X} \cap N(x,\delta)$. Then, $ML(x') = 0$ for almost every $x' \in \mathcal{X} \cap N(x,\delta)$, since $N(x,\delta) \subset \text{int}(\mathbb{R}^p \setminus \Omega^*)$. Therefore, $p^{ML}(x;\delta) = 0$. □

Claim 7. *For any $u \in \partial\Omega^*$ and $v \in \mathbb{R}^p$,*

$$1\{u + \delta v \in \Omega^*\} \to \begin{cases} 1 & \text{if } \nu(u)\cdot v > 0 \\ 0 & \text{if } \nu(u)\cdot v < 0 \end{cases}$$

as $\delta \to 0$ from the right.

Proof. Pick $u \in \partial\Omega^*$ and $v \in \mathbb{R}^p$. By the mean value theorem, for any sufficiently small $\delta > 0$,

$$\begin{aligned}
d^s_{\Omega^*}(u + \delta v) &= d^s_{\Omega^*}(u) + \nabla d^s_{\Omega^*}(\tilde{u}_\delta)\cdot\delta v \\
&= \nabla d^s_{\Omega^*}(\tilde{u}_\delta)\cdot\delta v.
\end{aligned}$$

for some $\tilde{u}_\delta$ which is on the line segment connecting u and $u + \delta v$. Since $\tilde{u}_\delta \to u$ as $\delta \to 0$ and $\nabla d^s_{\Omega^*}$ is continuous by Claim 3, $\nabla d^s_{\Omega^*}(\tilde{u}_\delta)\cdot v \to \nabla d^s_{\Omega^*}(u)\cdot v = \nu(u)\cdot v$ as $\delta \to 0$. Therefore, if $\nu(u)\cdot v > 0$, then $\nabla d^s_{\Omega^*}(\tilde{u}_\delta)\cdot\delta v > 0$ for any sufficiently small $\delta > 0$, and if $\nu(u)\cdot v < 0$, then $\nabla d^s_{\Omega^*}(\tilde{u}_\delta)\cdot\delta v < 0$ for any sufficiently small $\delta > 0$. The conclusion follows from the fact that $x \in \Omega^*$ if and only if $d^s_{\Omega^*}(x) > 0$.

□

Claim 8. *For any* $(u, v) \in (\partial\Omega^* \cap \text{int}(\mathcal{X})) \times (-1, 1)$, $p^{ML}(u + \delta v\nu(u); \delta) \to k(v)$ *as* $\delta \to 0$, *where*

$$k(v) = \begin{cases} 1 - \frac{1}{2} I_{(2(1-v)-(1-v)^2)}(\frac{p+1}{2}, \frac{1}{2}) & \text{for } v \in [0, 1) \\ \frac{1}{2} I_{(2(1+v)-(1+v)^2)}(\frac{p+1}{2}, \frac{1}{2}) & \text{for } v \in (-1, 0) \end{cases},$$

where $I_x(\alpha, \beta)$ *is the regularized incomplete beta function, or the cumulative distribution function of the beta distribution with shape parameters* α *and* β.

Proof. Since $\Pr(ML(X_i) \in (0, 1)) = 0$, $ML(x') \in \{0, 1\}$ for almost every $x' \in \mathcal{X} \cap N(x, \delta)$ for any $x \in \mathcal{X}$. Then,

$$p^{ML}(u + \delta v\nu(u); \delta) = \frac{\int_{\mathcal{X} \cap \Omega^*} 1\{x \in N(u + \delta v\nu(u), \delta)\} dx}{\int_{\mathcal{X}} 1\{x \in N(u + \delta v\nu(u), \delta)\} dx}.$$

Note that $x \in N(u + \delta v\nu(u), \delta)$ if and only if $\|\frac{x-(u+\delta v\nu(u))}{\delta}\| < 1$. Since $u \in \text{int}(\mathcal{X})$, $1\{u + \delta y \in \mathcal{X}\} \to 1$ as $\delta \to 0$ for any $y \in \mathbb{R}^p$. With change of variables $w = \frac{x-(u+\delta v\nu(u))}{\delta}$,

$$\begin{aligned} & \delta^{-p} \int_{\mathcal{X}} 1\{x \in N(u + v\delta\nu(u), \delta)\} dx \\ = & \int 1\{u + \delta(v\nu(u) + w) \in \mathcal{X}\} 1\{\|w\| < 1\} dw \\ \to & \int 1\{\|w\| < 1\} dw \\ = & \text{Vol}_p, \end{aligned}$$

where the convergence follows by the Dominated Convergence Theorem, and Vol_p is the volume of the p-dimensional unit ball.

For the numerator, we have

$$\begin{aligned} & \delta^{-p} \int_{\mathcal{X} \cap \Omega^*} 1\{x \in N(u + v\delta\nu(u), \delta)\} dx \\ = & \int 1\{u + \delta(v\nu(u) + w) \in \mathcal{X} \cap \Omega^*\} 1\{\|w\| < 1\} dw. \end{aligned}$$

By Claim 7,

$$1\{u + \delta(v\nu(u) + w) \in \Omega^*\} \to \begin{cases} 1 & \text{if } \nu(u) \cdot (v\nu(u) + w) > 0 \\ 0 & \text{if } \nu(u) \cdot (v\nu(u) + w) < 0 \end{cases}$$

as $\delta \to 0$. Note that the Lebesgue measure of $\{w \in \mathbb{R}^p : \nu(u) \cdot (v\nu(u) + w) = 0\}$ is zero, since the set is written as $\{w \in \mathbb{R}^p : v + \nu(u) \cdot w = 0\}$ and it is a $(p-1)$-dimensional plane. It follows that

$$\begin{aligned} & \delta^{-p} \int_{\mathcal{X} \cap \Omega^*} 1\{x \in N(u + v\delta\nu(u), \delta)\} dx \\ \to & \int 1\{\|w\| < 1, \nu(u) \cdot (v\nu(u) + w) > 0\} dw \\ = & \int 1\{\|w\| < 1, v + \nu(u) \cdot w > 0\} dw. \end{aligned}$$

Note that the set $\{w \in \mathbb{R}^p : \|w\| < 1, v + \nu(u) \cdot w > 0\}$ is a region of the p-dimensional unit ball cut off by the plane $\{w \in \mathbb{R}^p : v + \nu(u) \cdot w = 0\}$. The distance from the center of the unit ball to the plane is $|v|$. Using the formula for the volume of a hyperspherical cap (see e.g. Li (2011)), we have

$$\int 1\{\|w\| < 1, v + \nu(u) \cdot w < 0\} dw = \begin{cases} \mathrm{Vol}_p - \frac{1}{2}\mathrm{Vol}_p I_{(2(1-v)-(1-v)^2)}(\frac{p+1}{2}, \frac{1}{2}) & \text{for } v \in [0, 1) \\ \frac{1}{2}\mathrm{Vol}_p I_{(2(1+v)-(1+v)^2)}(\frac{p+1}{2}, \frac{1}{2}) & \text{for } v \in (-1, 0) \end{cases},$$

which completes the proof. □

By Claim 8, for every point $x \in \partial\Omega^* \cap \mathrm{int}(\mathcal{X})$, $p^{ML}(x) = \lim_{\delta \to 0} p^{ML}(x; \delta) = k(0) = \frac{1}{2}$.

Claim 9. *The following holds.*

(a) For $l = 0, 1, 2$,

$$\frac{1}{n\delta_n} \sum_{i=1}^n p^{ML}(X_i; \delta_n)^l I_i^p \xrightarrow{p} \int_{-1}^1 k(v)^l dv \int_{\partial\Omega^* \cap \mathrm{int}(\mathcal{X})} f_X(x) d\mathcal{H}^{p-1}(x).$$

(b) For $l = 0, 1$,

$$\frac{1}{n\delta_n} \sum_{i=1}^n Z_i p^{ML}(X_i; \delta_n)^l I_i^p \xrightarrow{p} \int_0^1 k(v)^l dv \int_{\partial\Omega^* \cap \mathrm{int}(\mathcal{X})} f_X(x) d\mathcal{H}^{p-1}(x).$$

(c) For $l = 0, 1$,

$$\frac{1}{n\delta_n} \sum_{i=1}^n Y_i p^{ML}(X_i; \delta_n)^l I_i^p \xrightarrow{p} \int_0^1 k(v)^l dv \int_{\partial\Omega^* \cap \mathrm{int}(\mathcal{X})} E[Y_{1i} | X_i = x] f_X(x) d\mathcal{H}^{p-1}(x) + \int_{-1}^0 k(v)^l dv \int_{\partial\Omega^* \cap \mathrm{int}(\mathcal{X})} E[Y_{0i} | X_i = x] f_X(x) d\mathcal{H}^{p-1}(x).$$

(d)

$$\frac{1}{n\delta_n} \sum_{i=1}^n Z_i Y_i I_i^p \xrightarrow{p} \int_{\partial\Omega^* \cap \mathrm{int}(\mathcal{X})} E[Y_{1i} | X_i = x] f_X(x) d\mathcal{H}^{p-1}(x).$$

Proof. For Part (a),

$$\begin{aligned}
&E[\frac{1}{n\delta_n} \sum_{i=1}^n p^{ML}(X_i; \delta_n)^l I_i^p] \\
=&\delta_n^{-1} E[p^{ML}(X_i; \delta_n)^l I_i^p] \\
=&\delta_n^{-1} \int_{\mathcal{X}} p^{ML}(x; \delta_n)^l 1\{p^{ML}(x; \delta_n) \in (0, 1)\} f_X(x) dx \\
=&\delta_n^{-1} \int_{N(\partial\Omega^*, \delta_n)} p^{ML}(x; \delta_n)^l 1\{p^{ML}(x; \delta_n) \in (0, 1)\} 1\{x \in \mathcal{X}\} f_X(x) dx \\
=&\delta_n^{-1} \int_{-\delta_n}^{\delta_n} \int_{\partial\Omega^*} p^{ML}(u + \lambda\nu(u); \delta_n)^l 1\{p^{ML}(u + \lambda\nu(u); \delta_n) \in (0, 1)\} 1\{u + \lambda\nu(u) \in \mathcal{X}\} \\
&\qquad \times f_X(u + \lambda\nu(u)) \det(I_{p-1} + \lambda V_u' D^2 d_{\Omega^*}^s(u) V_u) d\mathcal{H}^{p-1}(u) d\lambda,
\end{aligned}$$

where the third equality follows from Claim 6, and the last equality follows from Claim 5. With change of variables $v = \frac{\lambda}{\delta_n}$, we have

$$
\begin{aligned}
&E[\frac{1}{n\delta_n}\sum_{i=1}^{n} p^{ML}(X_i;\delta_n)^l I_i^p] \\
&= \int_{-1}^{1}\int_{\partial\Omega^*} p^{ML}(u+\delta_n v\nu(u);\delta_n)^l 1\{p^{ML}(u+\delta_n v\nu(u);\delta_n)\in(0,1)\}1\{u+\delta_n v\nu(u)\in\mathcal{X}\} \\
&\qquad\qquad \times f_X(u+\delta_n v\nu(u))\det(I_{p-1}+\delta_n v V_u' D^2 d^s_{\Omega^*}(u)V_u)d\mathcal{H}^{p-1}(u)dv.
\end{aligned}
$$

By Assumption 4 (g), there exist $\delta^* > 0$ such that $f_X(x) \le C_1$ for all $x \in \bar{N}(\partial\Omega^* \cap \mathcal{X}, \delta^*)$. Let $\bar{\delta} = \min\{\bar{\mu}, \delta^*\}$. Consider $\det(I_{p-1} + \lambda V' D^2 d^s_{\Omega^*}(u)V)$ as a function of $(\lambda, v_1', ..., v_{p-1}', u')' \in \mathbb{R}^{1+p(p-1)+p}$. Since this function is continuous at every point in the compact set $C = \{(\lambda, v_1', ..., v_{p-1}', u')' \in \mathbb{R}^{1+p(p-1)+p} : \lambda \in [-\bar{\delta}, \bar{\delta}], \|v_i\| = 1 \text{ for } i = 1, ..., p-1, u \in \partial\Omega^*\}$,

$$
\begin{aligned}
&\sup_{u\in\partial\Omega^*\cap\bar{N}(\mathcal{X},\bar{\delta}),v\in[-1,1],\delta\in(0,\bar{\delta}]} |\det(I_{p-1}+\delta v V_u' D^2 d^s_{\Omega^*}(u)V_u)| \\
&\le \sup_{(\lambda,v_1',...,v_{p-1}',u')'\in C} |\det(I_{p-1}+\lambda V' D^2 d^s_{\Omega^*}(u)V)| \\
&= C_2
\end{aligned}
$$

for some constant $C_2 \ge 0$. In addition, for any $\delta \in (0, \bar{\delta})$, if $u \in \partial\Omega^*$ and $u + \delta v\nu(u) \in \mathcal{X}$ for some $v \in (-1, 1)$, then $u \in \partial\Omega^* \cap \bar{N}(\mathcal{X}, \bar{\delta})$, since if there exists such (u, v), then $\|\delta v\nu(u)\| = |\delta v| < \bar{\delta}$. Therefore, for any sufficiently small $\delta_n > 0$,

$$
\begin{aligned}
&E[\frac{1}{n\delta_n}\sum_{i=1}^{n} p^{ML}(X_i;\delta_n)^l I_i^p] \\
&= \int_{-1}^{1}\int_{\partial\Omega^*\cap\bar{N}(\mathcal{X},\bar{\delta})} p^{ML}(u+\delta_n v\nu(u);\delta_n)^l 1\{p^{ML}(u+\delta_n v\nu(u);\delta_n)\in(0,1)\}1\{u+\delta_n v\nu(u)\in\mathcal{X}\} \\
&\qquad\qquad \times f_X(u+\delta_n v\nu(u))\det(I_{p-1}+\delta_n v V_u' D^2 d^s_{\Omega^*}(u)V_u)d\mathcal{H}^{p-1}(u)dv \\
&= \int_{-1}^{1}\int_{\partial\Omega^*\cap\mathrm{int}(\mathcal{X})} p^{ML}(u+\delta_n v\nu(u);\delta_n)^l 1\{p^{ML}(u+\delta_n v\nu(u);\delta_n)\in(0,1)\}1\{u+\delta_n v\nu(u)\in\mathcal{X}\} \\
&\qquad\qquad \times f_X(u+\delta_n v\nu(u))\det(I_{p-1}+\delta_n v V_u' D^2 d^s_{\Omega^*}(u)V_u)d\mathcal{H}^{p-1}(u)dv \\
&\quad + \int_{-1}^{1}\int_{\partial\Omega^*\cap\partial\mathcal{X}} p^{ML}(u+\delta_n v\nu(u);\delta_n)^l 1\{p^{ML}(u+\delta_n v\nu(u);\delta_n)\in(0,1)\}1\{u+\delta_n v\nu(u)\in\mathcal{X}\} \\
&\qquad\qquad \times f_X(u+\delta_n v\nu(u))\det(I_{p-1}+\delta_n v V_u' D^2 d^s_{\Omega^*}(u)V_u)d\mathcal{H}^{p-1}(u)dv \\
&\quad + \int_{-1}^{1}\int_{\partial\Omega^*\cap(\bar{N}(\mathcal{X},\bar{\delta})\setminus\mathcal{X})} p^{ML}(u+\delta_n v\nu(u);\delta_n)^l 1\{p^{ML}(u+\delta_n v\nu(u);\delta_n)\in(0,1)\}1\{u+\delta_n v\nu(u)\in\mathcal{X}\} \\
&\qquad\qquad \times f_X(u+\delta_n v\nu(u))\det(I_{p-1}+\delta_n v V_u' D^2 d^s_{\Omega^*}(u)V_u)d\mathcal{H}^{p-1}(u)dv. \qquad (12)
\end{aligned}
$$

The second integral of (12) is zero for any sufficiently small $\delta_n > 0$, since

$$\begin{aligned}
&\int_{-1}^{1}\int_{\partial\Omega^*\cap\partial\mathcal{X}} |p^{ML}(u+\delta_n v\nu(u);\delta_n)^l 1\{p^{ML}(u+\delta_n v\nu(u);\delta_n)\in(0,1)\}1\{u+\delta_n v\nu(u)\in\mathcal{X}\} \\
&\qquad\qquad \times f_X(u+\delta_n v\nu(u))\det(I_{p-1}+\delta_n v V_u' D^2 d^s_{\Omega^*}(u)V_u)|d\mathcal{H}^{p-1}(u)dv \\
\leq\ & \int_{-1}^{1}\int_{\partial\Omega^*\cap\partial\mathcal{X}} C_1C_2 d\mathcal{H}^{p-1}(u)dv \\
=\ & 2C_1C_2\int_{\partial\Omega^*\cap\partial\mathcal{X}} d\mathcal{H}^{p-1}(u) \\
=\ & 0,
\end{aligned}$$

where the last equality follows from Assumption 4 (e).

By the Dominated Convergence Theorem, the third integral of (12) converges to zero as $\delta_n \to 0$, since for any $(u,v)\in(\mathbb{R}^p\setminus\mathcal{X})\times(-1,1)$, $u+\delta_n v\nu(u)\in\mathbb{R}^p\setminus\mathcal{X}$ for any sufficiently small $\delta_n>0$ by the openness of $\mathbb{R}^p\setminus\mathcal{X}$.

Now consider the limit of the first integral of (12). By Claim 8, $p^{ML}(u+\delta_n v\nu(u);\delta_n)\to k(v)$ for any $(u,v)\in(\partial\Omega^*\cap\mathrm{int}(\mathcal{X}))\times(-1,1)$. Since $k(v)\in(0,1)$ for all $v\in(-1,1)$, $1\{p^{ML}(u+\delta_n v\nu(u);\delta_n)\in(0,1)\}\to 1$ for any $(u,v)\in(\partial\Omega^*\cap\mathrm{int}(\mathcal{X}))\times(-1,1)$. By definition of the interior, $1\{u+\delta_n v\nu(u)\in\mathcal{X}\}\to 1$ for any $(u,v)\in(\partial\Omega^*\cap\mathrm{int}(\mathcal{X}))\times(-1,1)$. Lastly, $f_X(u+\delta_n v\nu(u))\to f_X(u)$ and $\det(I_{p-1}+\delta_n v V_u' D^2 d^s_{\Omega^*}(u)V_u)\to\det(I_{p-1})=1$ for any $(u,v)\in(\partial\Omega^*\cap\mathrm{int}(\mathcal{X}))\times(-1,1)$ by continuity. Therefore, we have

$$\begin{aligned}
E[\frac{1}{n\delta_n}\sum_{i=1}^n p^{ML}(X_i;\delta_n)^l I_i^p] &\to \int_{-1}^{1}\int_{\partial\Omega^*\cap\mathrm{int}(\mathcal{X})} k(v)^l f_X(u)d\mathcal{H}^{p-1}(u)dv \\
&= \int_{-1}^{1} k(v)^l dv \int_{\partial\Omega^*\cap\mathrm{int}(\mathcal{X})} f_X(u)d\mathcal{H}^{p-1}(u).
\end{aligned}$$

As for variance,

$$\begin{aligned}
\mathrm{Var}[\frac{1}{n\delta_n}\sum_{i=1}^n p^{ML}(X_i;\delta_n)^l I_i^p] &\leq \frac{1}{n\delta_n^2}E[p^{ML}(X_i;\delta_n)^{2l} I_i^{p2}] \\
&= \frac{1}{n\delta_n}\left(\int_{-1}^{1} k(v)^{2l}dv \int_{\partial\Omega^*\cap\mathrm{int}(\mathcal{X})} f_X(u)d\mathcal{H}^{p-1} + o(1)\right) \\
&\to 0.
\end{aligned}$$

For Part (b),

$$
\begin{aligned}
&E[\frac{1}{n\delta_n}\sum_{i=1}^{n} Z_i p^{ML}(X_i;\delta_n)^l I_i^p] \\
=&\delta_n^{-1} E[Z_i p^{ML}(X_i;\delta_n)^l I_i^p] \\
=&\delta_n^{-1} E[1\{X_i \in \Omega^*\} p^{ML}(X_i;\delta_n)^l I_i^p] \\
=&\delta_n^{-1} \int_{\Omega^* \cap \mathcal{X}} ML(x) p^{ML}(x;\delta_n)^l 1\{p^{ML}(x;\delta_n) \in (0,1)\} f_X(x) dx \\
=&\delta_n^{-1} \int_{-\delta_n}^{\delta_n} \int_{\partial\Omega^*} p^{ML}(u+\lambda\nu(u);\delta_n)^l 1\{p^{ML}(u+\lambda\nu(u);\delta_n) \in (0,1)\} 1\{u+\lambda\nu(u) \in \Omega^* \cap \mathcal{X}\} \\
&\qquad \times f_X(u+\lambda\nu(u)) \det(I_{p-1} + \lambda V_u' D^2 d^s_{\Omega^*}(u) V_u) d\mathcal{H}^{p-1}(u) d\lambda,
\end{aligned}
$$

where the second equality holds, since $\Pr(ML(X_i) \in (0,1)) = 0$ so that $Z_i = 1\{X_i \in \Omega^*\}$ with probability one. Since $u + \lambda\nu(u) \in \Omega^*$ for all $(u,\lambda) \in \partial\Omega^* \times (0,\delta_n)$, and $u + \lambda\nu(u) \notin \Omega^*$ for all $(u,\lambda) \in \partial\Omega^* \times (-\delta_n, 0)$ for any sufficiently small $\delta_n > 0$,

$$
\begin{aligned}
&E[\frac{1}{n\delta_n}\sum_{i=1}^{n} Z_i p^{ML}(X_i;\delta_n)^l I_i^p] \\
=&\delta_n^{-1} \int_{0}^{\delta_n} \int_{\partial\Omega^*} p^{ML}(u+\lambda\nu(u);\delta_n)^l 1\{p^{ML}(u+\lambda\nu(u);\delta_n) \in (0,1)\} 1\{u+\lambda\nu(u) \in \mathcal{X}\} \\
&\qquad \times f_X(u+\lambda\nu(u)) \det(I_{p-1} + \lambda V_u' D^2 d^s_{\Omega^*}(u) V_u) d\mathcal{H}^{p-1}(u) d\lambda.
\end{aligned}
$$

Part (b) is then established in the same way as Part (a).

For Part (c),

$$
\begin{aligned}
&E[\frac{1}{n\delta_n}\sum_{i=1}^{n} Y_i p^{ML}(X_i;\delta_n)^l I_i^p] \\
=&\delta_n^{-1} E[Y_i p^{ML}(X_i;\delta_n)^l I_i^p] \\
=&\delta_n^{-1} E[1\{X_i \in \Omega^*\} Y_i p^{ML}(X_i;\delta_n)^l I_i^p] + \delta_n^{-1} E[1\{X_i \in \mathbb{R}^p \setminus \Omega^*\} Y_i p^{ML}(X_i;\delta_n)^l I_i^p] \\
=&\delta_n^{-1} E[1\{X_i \in \Omega^*\} Y_{1i} p^{ML}(X_i;\delta_n)^l I_i^p] + \delta_n^{-1} E[1\{X_i \in \mathbb{R}^p \setminus \Omega^*\} Y_{0i} p^{ML}(X_i;\delta_n)^l I_i^p],
\end{aligned}
$$

where the last equality holds, since $\Pr(ML(X_i) \in (0,1)) = 0$ so that $1\{X_i \in \Omega^*\} Y_i = 1\{X_i \in$

$\Omega^*\}Y_{1i}$ and $1\{X_i \in \mathbb{R}^p \setminus \Omega^*\}Y_i = 1\{X_i \in \mathbb{R}^p \setminus \Omega^*\}Y_{0i}$ with probability one. For the first term,

$$
\begin{aligned}
&\delta_n^{-1}E[1\{X_i \in \Omega^*\}Y_{1i}p^{ML}(X_i;\delta_n)^l I_i^p] \\
=&\delta_n^{-1}E[1\{X_i \in \Omega^*\}E[Y_{1i}|X_i]p^{ML}(X_i;\delta_n)^l I_i^p] \\
=&\delta_n^{-1}\int_{\Omega^*\cap\mathcal{X}} E[Y_{1i}|X_i = x]p^{ML}(x;\delta_n)^l 1\{p^{ML}(x;\delta_n) \in (0,1)\}f_X(x)dx \\
=&\delta_n^{-1}\int_{-\delta_n}^{\delta_n}\int_{\partial\Omega^*} E[Y_{1i}|X_i = u+\lambda\nu(u)]p^{ML}(u+\lambda\nu(u);\delta_n)^l 1\{p^{ML}(u+\lambda\nu(u);\delta_n) \in (0,1)\} \\
&\qquad \times 1\{u+\lambda\nu(u) \in \Omega^*\cap\mathcal{X}\}f_X(u+\lambda\nu(u))\det(I_{p-1}+\lambda V_u' D^2 d^s_{\Omega^*}(u)V_u)d\mathcal{H}^{p-1}(u)d\lambda \\
=&\delta_n^{-1}\int_{0}^{\delta_n}\int_{\partial\Omega^*} E[Y_{1i}|X_i = u+\lambda\nu(u)]p^{ML}(u+\lambda\nu(u);\delta_n)^l 1\{p^{ML}(u+\lambda\nu(u);\delta_n) \in (0,1)\} \\
&\qquad \times 1\{u+\lambda\nu(u) \in \mathcal{X}\}f_X(u+\lambda\nu(u))\det(I_{p-1}+\lambda V_u' D^2 d^s_{\Omega^*}(u)V_u)d\mathcal{H}^{p-1}(u)d\lambda,
\end{aligned}
$$

where the last equality holds, since $u+\lambda\nu(u) \in \Omega^*$ for all $(u,\lambda) \in \partial\Omega^*\times(0,\delta_n)$, and $u+\lambda\nu(u) \notin \Omega^*$ for all $(u,\lambda) \in \partial\Omega^* \times (-\delta_n, 0)$ for any sufficiently small $\delta_n > 0$. Similariy, for the second term, we have

$$
\begin{aligned}
&\delta_n^{-1}E[1\{X_i \in \mathbb{R}^p \setminus \Omega^*\}Y_{0i}p^{ML}(X_i;\delta_n)^l I_i^p] \\
=&\delta_n^{-1}\int_{-\delta_n}^{0}\int_{\partial\Omega^*} E[Y_{0i}|X_i = u+\lambda\nu(u)]p^{ML}(u+\lambda\nu(u);\delta_n)^l 1\{p^{ML}(u+\lambda\nu(u);\delta_n) \in (0,1)\} \\
&\qquad \times 1\{u+\lambda\nu(u) \in \mathcal{X}\}f_X(u+\lambda\nu(u))\det(I_{p-1}+\lambda V_u' D^2 d^s_{\Omega^*}(u)V_u)d\mathcal{H}^{p-1}(u)d\lambda.
\end{aligned}
$$

Part (c) is then established similarly under the continuity of $E[Y_{1i}|X_i]$ and $E[Y_{0i}|X_i]$ and the boundedness of $E[Y_{1i}^2|X_i]$ and $E[Y_{0i}^2|X_i]$ imposed by Assumption 4 (f) (h).

For Part (d),

$$
\begin{aligned}
&E[\frac{1}{n\delta_n}\sum_{i=1}^n Z_i Y_i I_i^p] \\
=&\delta_n^{-1}E[1\{X_i \in \Omega^*\}E[Y_{1i}|X_i]I_i^p] \\
=&\delta_n^{-1}\int_{\Omega^*\cap\mathcal{X}} E[Y_{1i}|X_i = x]1\{p^{ML}(x;\delta_n) \in (0,1)\}f_X(x)dx \\
=&\delta_n^{-1}\int_{-\delta_n}^{\delta_n}\int_{\partial\Omega^*} E[Y_{1i}|X_i = u+\lambda\nu(u)]1\{p^{ML}(u+\lambda\nu(u);\delta_n) \in (0,1)\}1\{u+\lambda\nu(u) \in \Omega^*\cap\mathcal{X}\} \\
&\qquad \times f_X(u+\lambda\nu(u))\det(I_{p-1}+\lambda V_u' D^2 d^s_{\Omega^*}(u)V_u)d\mathcal{H}^{p-1}(u)d\lambda \\
=&\delta_n^{-1}\int_{0}^{\delta_n}\int_{\partial\Omega^*} E[Y_{1i}|X_i = u+\lambda\nu(u)]1\{p^{ML}(u+\lambda\nu(u);\delta_n) \in (0,1)\}1\{u+\lambda\nu(u) \in \mathcal{X}\} \\
&\qquad \times f_X(u+\lambda\nu(u))\det(I_{p-1}+\lambda V_u' D^2 d^s_{\Omega^*}(u)V_u)d\mathcal{H}^{p-1}(u)d\lambda.
\end{aligned}
$$

Part (d) is then established similarly. □

By Claim 9, we obtain

$$\hat{\alpha} \xrightarrow{p} T_W^{-1}T_Y,$$

where

$$T_W = \int_{\partial\Omega^* \cap \text{int}(\mathcal{X})} f_X(x) d\mathcal{H}^{p-1}(x) \begin{bmatrix} 2 & 1 & \int_{-1}^1 k(v)dv \\ 1 & 1 & \int_0^1 k(v)dv \\ \int_{-1}^1 k(v)dv & \int_0^1 k(v)dv & \int_{-1}^1 k(v)^2 dv \end{bmatrix},$$

and

$$T_Y = \begin{bmatrix} \int_{\partial\Omega^* \cap \text{int}(\mathcal{X})} E[Y_{1i} + Y_{0i} | X_i = x] f_X(x) d\mathcal{H}^{p-1}(x) \\ \int_{\partial\Omega^* \cap \text{int}(\mathcal{X})} E[Y_{1i} | X_i = x] f_X(x) d\mathcal{H}^{p-1}(x) \\ \int_{\partial\Omega^* \cap \text{int}(\mathcal{X})} (\int_0^1 k(v)dv E[Y_{1i} | X_i = x] + \int_{-1}^0 k(v)dv E[Y_{0i} | X_i = x]) f_X(x) d\mathcal{H}^{p-1}(x) \end{bmatrix}.$$

After a few lines of algebra, we have that the determinant of T_W is

$$(\int_{\partial\Omega^* \cap \text{int}(\mathcal{X})} f_X(x) d\mathcal{H}^{p-1}(x))^{-1} (\int_{-1}^0 (k(v) - \int_{-1}^0 k(s)ds)^2 dv + \int_0^1 (k(v) - \int_0^1 k(s)ds)^2 dv),$$

which is nonzero under Assumption 4 (e). In addition, we obtain

$$\hat{\alpha}_1 \xrightarrow{p} \frac{\int_{\partial\Omega^* \cap \text{int}(\mathcal{X})} E[Y_{1i} - Y_{0i} | X_i = x] f_X(x) d\mathcal{H}^{p-1}(x)}{\int_{\partial\Omega^* \cap \text{int}(\mathcal{X})} f_X(x) d\mathcal{H}^{p-1}(x)}.$$

Since $\mathcal{H}^{p-1}(\partial\Omega^* \cap \partial\mathcal{X}) = 0$ by Assumption 4 (e), it follows that

$$\hat{\alpha}_1 \xrightarrow{p} \frac{\int_{\partial\Omega^* \cap \mathcal{X}} E[Y_{1i} - Y_{0i} | X_i = x] f_X(x) d\mathcal{H}^{p-1}(x)}{\int_{\partial\Omega^* \cap \mathcal{X}} f_X(x) d\mathcal{H}^{p-1}(x)}.$$

□

A.6.3 Proof of Lemma 3

We only show that $\hat{\alpha}_1^s = \hat{\alpha}_1 + o_p(1)$. Let $I_i^{p,s} = 1\{p^s(X_i, \delta_n) \in (0,1)\}$. As described in the proof of Lemma 1, $\hat{\alpha}$ takes the form of $\hat{S}_W^{-1}\hat{S}_Y$ for some matrix $\hat{S}_W$ and vector $\hat{S}_Y$, where each component of the matrix and vector is written as $\frac{1}{n}\sum_{i=1}^n V_i p^{ML}(X_i;\delta_n)^l I_i^p$ for some i.i.d random variables $\{V_i\}_{i=1}^\infty$ with finite variance and some $l = 0, 1, 2$. $\hat{\alpha}^s$ is obtained by replacing $p^{ML}(X_i;\delta_n)$ and I_i^p with $p^s(X_i;\delta_n)$ and $I_i^{p,s}$, respectively. It is therefore sufficient to show that $\frac{1}{n}\sum_{i=1}^n V_i p^s(X_i;\delta_n)^l I_i^{p.s} - \frac{1}{n}\sum_{i=1}^n V_i p^{ML}(X_i;\delta_n)^l I_i^p = o_p(1)$ for any $l = 0, 1, 2$ for any i.i.d. random variables $\{V_i\}_{i=1}^\infty$ with finite variance such that $\{V_i\}_{i=1}^n$ is independent of simulation draws conditional on $\{X_i\}_{i=1}^n$ for all n.

Let $\tilde{\mathcal{X}} = \{x \in \mathcal{X} : ML(x) \in (0,1)\}$, and $\mathcal{X}_c = \{x \in \mathcal{X} : ML \text{ is continuous at } x\}$. We first prove the following claim.

Claim 10. *$E[p^s(X_i,\delta_n)^l | X_i = x] \to p^{ML}(x)^l$ and $E[p^s(X_i,\delta_n)^l (I_i^{p,s} - I_i^p) | X_i = x] \to 0$ as $n \to \infty$ for any $l = 0, 1, 2$ for almost every $x \in \mathcal{X}$.*

Proof. By Assumption 2, $x \in \text{int}(\mathcal{X}_0) \cup \text{int}(\mathcal{X}_1) \cup (\tilde{\mathcal{X}} \cap \mathcal{X}_c)$ for almost every $x \in \mathcal{X}$. Pick any $x \in \text{int}(\mathcal{X}_k)$ for some $k \in \{0,1\}$, and let $X_1^*, ..., X_{S_n}^*$ be the S_n independent draws from the uniform distribution on $\mathcal{X} \cap N(x, \delta_n)$. Since $x \in \text{int}(\mathcal{X}_k)$, there exists $n_x \in \mathbb{N}$ such that for

all $n \geq n_x$, $N(x, \delta_n) \subset \mathcal{X}_k$ so that $p^{ML}(x) = ML(x)$ and that $ML(X_s^*) = ML(x)$ for all $s = 1, ..., S_n$. Then, for all $n \geq n_x$,

$$E[p^s(X_i, \delta_n)^l | X_i = x] = E[(\frac{1}{S_n} \sum_{s=1}^{S_n} ML(X_s^*))^l] = p^{ML}(x)^l$$

and

$$\begin{aligned}
&E[p^s(X_i, \delta_n)^l (I_i^{p,s} - I_i^p) | X_i = x] \\
=&E[(\frac{1}{S_n} \sum_{s=1}^{S_n} ML(X_s^*))^l (1\{\frac{1}{S_n} \sum_{s=1}^{S_n} ML(X_s^*) \in (0,1)\} - 1\{p^{ML}(x) \in (0,1)\})] \\
=&ML(x)^l (0 - 0) \\
=&0.
\end{aligned}$$

Now pick any $x \in \tilde{\mathcal{X}} \cap \mathcal{X}_c$, and let $X_1^*, ..., X_{S_n}^*$ be the S_n independent draws from the uniform distribution on $\mathcal{X} \cap N(x, \delta_n)$. First, we have

$$E[p^s(X_i, \delta_n)^0 | X_i = x] = 1 = p^{ML}(x)^0$$

for all $n \geq 1$, and that

$$\begin{aligned}
E[p^s(X_i, \delta_n) | X_i = x] &= p^{ML}(x; \delta_n) \\
&\to p^{ML}(x).
\end{aligned}$$

For $l = 2$,

$$\begin{aligned}
E[p^s(X_i, \delta_n)^2 | X_i = x] &= E[(\frac{1}{S_n} \sum_{s=1}^{S_n} ML(X_s^*))^2] \\
&= \text{Var}(\frac{1}{S_n} \sum_{s=1}^{S_n} ML(X_s^*)) + (E[\frac{1}{S_n} \sum_{s=1}^{S_n} ML(X_s^*)])^2 \\
&= \frac{1}{S_n} \text{Var}(ML(X_1^*)) + p^{ML}(x; \delta_n)^2 \\
&= \frac{1}{S_n} \frac{\int_{\mathcal{X} \cap N(x, \delta_n)} (ML(x^*) - p^{ML}(x; \delta_n))^2 dx^*}{\int_{\mathcal{X} \cap N(x, \delta_n)} dx^*} + p^{ML}(x; \delta_n)^2
\end{aligned}$$

With change of variables $u = \frac{x^* - x}{\delta}$, we have

$$\begin{aligned}
\frac{\int_{\mathcal{X} \cap N(x, \delta)} (ML(x^*) - p^{ML}(x; \delta))^2 dx^*}{\int_{\mathcal{X} \cap N(x, \delta)} dx^*} &= \frac{\int_{N(x, \delta)} 1\{x^* \in \mathcal{X}\} (ML(x^*) - p^{ML}(x; \delta))^2 dx^*}{\int_{N(x, \delta)} 1\{x^* \in \mathcal{X}\} dx^*} \\
&= \frac{\delta^p \int_{N(\mathbf{0}, 1)} 1\{x + \delta u \in \mathcal{X}\} (ML(x + \delta u) - p^{ML}(x; \delta))^2 du}{\delta^p \int_{N(\mathbf{0}, 1)} 1\{x + \delta u \in \mathcal{X}\} du}
\end{aligned}$$

Since $ML(x) \in (0,1)$ and ML is continuous at x, $N(x,\delta) \subset \tilde{\mathcal{X}}$ for any sufficiently small $\delta > 0$ so that $\lim_{\delta \to 0} 1\{x + \delta u \in \mathcal{X}\} = 1$ for any $u \in N(\mathbf{0},1)$. Also, by Part 1 of Corollary 2, $p^{ML}(x) = ML(x)$. By the Dominated Convergence Theorem,

$$\begin{aligned}\frac{\int_{N(\mathbf{0},1)} 1\{x+\delta u \in \mathcal{X}\}(ML(x+\delta u) - p^{ML}(x;\delta))^2 du}{\int_{N(\mathbf{0},1)} 1\{x+\delta u \in \mathcal{X}\} du} &\to \frac{\int_{N(\mathbf{0},1)} (ML(x) - p^{ML}(x))^2 du}{\int_{N(\mathbf{0},1)} du} \\ &= 0.\end{aligned}$$

Thus,

$$\lim_{n\to\infty} E[p^s(X_i,\delta_n)^2 | X_i = x] = p^{ML}(x)^2.$$

Since $N(x,\delta) \subset \tilde{\mathcal{X}}$ for any sufficiently small $\delta > 0$, there exists $n_x \in \mathbb{N}$ such that $ML(X_s^*) \in (0,1)$ for all $s = 1, ..., S_n$. Then, for all $n \geq n_x$,

$$\begin{aligned}&E[p^s(X_i,\delta_n)^l (I_i^{p,s} - I_i^p) | X_i = x] \\ =&E[(\frac{1}{S_n}\sum_{s=1}^{S_n} ML(X_s^*))^l (1\{\frac{1}{S_n}\sum_{s=1}^{S_n} ML(X_s^*) \in (0,1)\} - 1\{p^{ML}(x) \in (0,1)\})] \\ =&E[(\frac{1}{S_n}\sum_{s=1}^{S_n} ML(X_s^*))^l (1-1)] \\ =&0.\end{aligned}$$

□

Below, we proceed in the way similar to the one used in the proof of Claim 2. First, we have

$$\begin{aligned}&E[\frac{1}{n}\sum_{i=1}^n V_i p^s(X_i;\delta_n)^l I_i^{p.s} - \frac{1}{n}\sum_{i=1}^n V_i p^{ML}(X_i;\delta_n)^l I_i^p] \\ =&E[V_i\{p^s(X_i;\delta_n)^l (I_i^{p.s} - I_i^p) + (p^s(X_i;\delta_n)^l - p^{ML}(X_i;\delta_n)^l) I_i^p\}] \\ =&\int_{\mathcal{X}} E[V_i|X_i = x]\{E[p^s(X_i,\delta_n)^l (I_i^{p,s} - I_i^p)|X_i = x] \\ &\qquad - (E[p^s(X_i,\delta_n)^l | X_i = x] - p^{ML}(x;\delta_n)^l) 1\{p^{ML}(x;\delta_n) \in (0,1)\} f_X(x) dx \\ \to&0\end{aligned}$$

as $n \to \infty$, where the convergence follows by Claims 1 and 10 and by the Dominated Convergence Theorem.

The variance of $\frac{1}{n}\sum_{i=1}^n V_i p^s(X_i;\delta_n)^l I_i^{p.s} - \frac{1}{n}\sum_{i=1}^n V_i p^{ML}(X_i;\delta_n)^l I_i^p$ also converges to zero, since $p^s(X_i;\delta_n) \in [0,1]$, $p^{ML}(X_i;\delta_n) \in [0,1]$, $I_i^{p,s} \in [0,1]$, $I_i^p \in [0,1]$ and $E[V_i^2] < \infty$. □

A.6.4 Proof of Lemma 4

We only show that $\hat{\alpha}_1^s = \hat{\alpha}_1 + o_p(1)$. Let $I_i^{p,s} = 1\{p^s(X_i,\delta_n) \in (0,1)\}$. It is sufficient to show that the following holds.

(a) For $l = 0, 1, 2$, $\frac{1}{n\delta_n}\sum_{i=1}^n p^s(X_i;\delta_n)^l I_i^{p,s} - \frac{1}{n\delta_n}\sum_{i=1}^n p^{ML}(X_i;\delta_n)^l I_i^p = o_p(1)$.

(b) For $l = 0, 1$, $\frac{1}{n\delta_n}\sum_{i=1}^n Z_i p^s(X_i;\delta_n)^l I_i^{p,s} - \frac{1}{n\delta_n}\sum_{i=1}^n Z_i p^{ML}(X_i;\delta_n)^l I_i^p = o_p(1)$.

(c) For $l = 0, 1$, $\frac{1}{n\delta_n}\sum_{i=1}^n Y_i p^s(X_i;\delta_n)^l I_i^{p,s} - \frac{1}{n\delta_n}\sum_{i=1}^n Y_i p^{ML}(X_i;\delta_n)^l I_i^p = o_p(1)$.

(d) $\frac{1}{n\delta_n}\sum_{i=1}^n Z_i Y_i I_i^{p,s} - \frac{1}{n\delta_n}\sum_{i=1}^n Z_i Y_i I_i^p = o_p(1)$.

We only prove Part (a), since the proof for the other parts is analogous. We use the following claim.

Claim 11. *For any $x \in \mathcal{X}$,*

(a) $E[p^s(X_i;\delta_n)|X_i = x] = p^{ML}(x;\delta_n)$;

(b) $E[p^s(X_i;\delta_n)^2|X_i = x] - p^{ML}(x;\delta_n)^2 = \frac{1}{S_n}p^{ML}(x;\delta_n)(1 - p^{ML}(x;\delta_n))$;

(c) $E[I_i^{p,s} - I_i^p|X_i = x] = -1\{p^{ML}(x;\delta_n) \in (0,1)\}((1 - p^{ML}(x;\delta_n))^{S_n} + p^{ML}(x;\delta_n)^{S_n})$;

(d) for $l = 1, 2$, $E[p^s(X_i;\delta_n)^l(I_i^{p,s} - I_i^p)|X_i = x] = -1\{p^{ML}(x;\delta_n) \in (0,1)\}p^{ML}(x;\delta_n)^{S_n}$.

Proof. Pick any $x \in \mathcal{X}$, and let $X_1^*, ..., X_{S_n}^*$ be the S_n independent draws from the uniform distribution on $\mathcal{X} \cap N(x,\delta_n)$. By the definition of $p^{ML}(x;\delta_n)$, $E[ML(X_s^*)] = p^{ML}(x;\delta_n)$. Moreover, $ML(X_s^*)$ follows Bernoulli distribution with mean $p^{ML}(x;\delta_n)$, since $\Pr(ML(X_i) \in \{0,1\}) = 1$. Then,

$$E[p^s(X_i;\delta_n)|X_i = x] = E[\frac{1}{S_n}\sum_{s=1}^{S_n} ML(X_s^*)] = p^{ML}(x;\delta_n),$$

and

$$\begin{aligned}
E[p^s(X_i;\delta_n)^2|X_i = x] - p^{ML}(x;\delta_n)^2 &= E[(\frac{1}{S_n}\sum_{s=1}^{S_n} ML(X_s^*))^2] - E[\frac{1}{S_n}\sum_{s=1}^{S_n} ML(X_s^*)]^2 \\
&= \mathrm{Var}(\frac{1}{S_n}\sum_{s=1}^{S_n} ML(X_s^*)) \\
&= \frac{1}{S_n}\mathrm{Var}(ML(X_s^*)) \\
&= \frac{1}{S_n}p^{ML}(x;\delta_n)(1 - p^{ML}(x;\delta_n)).
\end{aligned}$$

For Parts (c) and (d), suppose first that $p^{ML}(x;\delta_n) \in \{0,1\}$. Then, $I_i^{p,s} = 0$ with probability one (conditional on $X_i = x$), so $E[I_i^{p,s} - I_i^p|X_i = x] = 0$ and $E[p^s(X_i;\delta_n)^l(I_i^{p,s} - I_i^p)|X_i = x] = 0$ for $l = 1, 2$. Next, suppose that $p^{ML}(x;\delta_n) \in (0,1)$. Then,

$$\begin{aligned}
E[I_i^{p,s} - I_i^p|X_i = x] &= E[1\{\frac{1}{S_n}\sum_{s=1}^{S_n} ML(X_s^*) \in (0,1)\} - 1] \\
&= (0-1)(\Pr(X_1^* = \cdots = X_{S_n}^* = 0) + \Pr(X_1^* = \cdots = X_{S_n}^* = 1)) \\
&= -((1 - p^{ML}(x;\delta_n))^{S_n} + p^{ML}(x;\delta_n)^{S_n}),
\end{aligned}$$

and for $l = 1, 2$,

$$\begin{aligned} E[p^s(X_i;\delta_n)^l(I_i^{p,s} - I_i^p)|X_i = x] &= E[(\frac{1}{S_n}\sum_{s=1}^{S_n} ML(X_s^*))^l(1\{\frac{1}{S_n}\sum_{s=1}^{S_n} ML(X_s^*) \in (0,1)\} - 1)] \\ &= 1(0-1)\Pr(X_1^* = \cdots = X_{S_n}^* = 1) \\ &= -p^{ML}(x;\delta_n)^{S_n}. \end{aligned}$$

□

By proceeding as in the proof of Claim 9, we have for any sufficiently small $\delta_n > 0$,

$$\begin{aligned} &E[\frac{1}{n\delta_n}\sum_{i=1}^n p^s(X_i;\delta_n)^l I_i^{p,s} - \frac{1}{n\delta_n}\sum_{i=1}^n p^{ML}(X_i;\delta_n)^l I_i^p] \\ =&\int_{-1}^1\int_{\partial\Omega^*\cap\mathrm{int}(\mathcal{X})} E[p^s(X_i;\delta_n)^l I_i^{p,s}|X_i = u+\delta_n v\nu(u)]1\{u+\delta_n v\nu(u)\in\mathcal{X}\} \\ &\qquad\qquad \times f_X(u+\delta_n v\nu(u))\det(I_{p-1}+\delta_n v V_u' D^2 d^s_{\Omega^*}(u)V_u)d\mathcal{H}^{p-1}(u)dv \\ &-\int_{-1}^1\int_{\partial\Omega^*\cap\mathrm{int}(\mathcal{X})} p^{ML}(u+\delta_n v\nu(u);\delta_n)^l 1\{p^{ML}(u+\delta_n v\nu(u);\delta_n)\in(0,1)\}1\{u+\delta_n v\nu(u)\in\mathcal{X}\} \\ &\qquad\qquad \times f_X(u+\delta_n v\nu(u))\det(I_{p-1}+\delta_n v V_u' D^2 d^s_{\Omega^*}(u)V_u)d\mathcal{H}^{p-1}(u)dv \\ =&\int_{-1}^1\int_{\partial\Omega^*\cap\mathrm{int}(\mathcal{X})} \big[E[p^s(X_i;\delta_n)^l(I_i^{p,s}-I_i^p)|X_i = u+\delta_n v\nu(u)] \\ &\quad + (E[p^s(X_i;\delta_n)^l|X_i = u+\delta_n v\nu(u)] - p^{ML}(u+\delta_n v\nu(u);\delta_n)^l)1\{p^{ML}(u+\delta_n v\nu(u);\delta_n)\in(0,1)\}\big] \\ &\quad \times 1\{u+\delta_n v\nu(u)\in\mathcal{X}\}f_X(u+\delta_n v\nu(u))\det(I_{p-1}+\delta_n v V_u' D^2 d^s_{\Omega^*}(u)V_u)d\mathcal{H}^{p-1}(u)dv. \end{aligned}$$

Recall that for any $(u,v)\in(\partial\Omega^*\cap\mathrm{int}(\mathcal{X}))\times(-1,1)$, $p^{ML}(u+\delta v\nu(u);\delta)\to k(v)\in(0,1)$ as $\delta\to 0$ by Claim 8. When $l=0$,

$$\begin{aligned} &E[\frac{1}{n\delta_n}\sum_{i=1}^n p^s(X_i;\delta_n)^l I_i^{p,s} - \frac{1}{n\delta_n}\sum_{i=1}^n p^{ML}(X_i;\delta_n)^l I_i^p] \\ =&\int_{-1}^1\int_{\partial\Omega^*\cap\mathrm{int}(\mathcal{X})} \big[-1\{p^{ML}(u+\delta_n v\nu(u);\delta_n)\in(0,1)\}((1-p^{ML}(u+\delta_n v\nu(u);\delta_n))^{S_n} + p^{ML}(u+\delta_n v\nu(u);\delta_n)^{S_n})\big] \\ &\qquad \times 1\{u+\delta_n v\nu(u)\in\mathcal{X}\}f_X(u+\delta_n v\nu(u))\det(I_{p-1}+\delta_n v V_u' D^2 d^s_{\Omega^*}(u)V_u)d\mathcal{H}^{p-1}(u)dv \\ \to&-\int_{-1}^1\int_{\partial\Omega^*\cap\mathrm{int}(\mathcal{X})}((1-k(v))^{\lim_{n\to\infty}S_n} + k(v)^{\lim_{n\to\infty}S_n})f_X(u)d\mathcal{H}^{p-1}(u)dv \\ =&0, \end{aligned}$$

where the second equality holds by Claim 11 (c), and the last equality holds since $k(v)\in(0,1)$

and $S_n \to \infty$. When $l = 1$,

$$
\begin{aligned}
&E[\frac{1}{n\delta_n}\sum_{i=1}^{n} p^s(X_i;\delta_n)^l I_i^{p,s} - \frac{1}{n\delta_n}\sum_{i=1}^{n} p^{ML}(X_i;\delta_n)^l I_i^p] \\
=&\int_{-1}^{1}\int_{\partial\Omega^*\cap \mathrm{int}(\mathcal{X})} \big[-1\{p^{ML}(u+\delta_n v\nu(u);\delta_n)\in(0,1)\}p^{ML}(u+\delta_n v\nu(u);\delta_n)^{S_n}\big] \\
&\qquad \times 1\{u+\delta_n v\nu(u)\in\mathcal{X}\} f_X(u+\delta_n v\nu(u))\det(I_{p-1}+\delta_n v V_u' D^2 d^s_{\Omega^*}(u) V_u) d\mathcal{H}^{p-1}(u)dv \\
\to& -\int_{-1}^{1}\int_{\partial\Omega^*\cap \mathrm{int}(\mathcal{X})} (k(v)^{\lim_{n\to\infty} S_n}) f_X(u) d\mathcal{H}^{p-1}(u)dv \\
=&0,
\end{aligned}
$$

where the second equality holds by Claim 11 (a) and (d). When $l = 2$,

$$
\begin{aligned}
&E[\frac{1}{n\delta_n}\sum_{i=1}^{n} p^s(X_i;\delta_n)^l I_i^{p,s} - \frac{1}{n\delta_n}\sum_{i=1}^{n} p^{ML}(X_i;\delta_n)^l I_i^p] \\
=&\int_{-1}^{1}\int_{\partial\Omega^*\cap \mathrm{int}(\mathcal{X})} \big[-1\{p^{ML}(u+\delta_n v\nu(u);\delta_n)\in(0,1)\}p^{ML}(u+\delta_n v\nu(u);\delta_n)^{S_n} \\
&\qquad + \frac{1}{S_n} p^{ML}(u+\delta_n v\nu(u);\delta_n)(1-p^{ML}(u+\delta_n v\nu(u);\delta_n))\big] \\
&\qquad \times 1\{u+\delta_n v\nu(u)\in\mathcal{X}\} f_X(u+\delta_n v\nu(u))\det(I_{p-1}+\delta_n v V_u' D^2 d^s_{\Omega^*}(u) V_u) d\mathcal{H}^{p-1}(u)dv \\
\to&\int_{-1}^{1}\int_{\partial\Omega^*\cap \mathrm{int}(\mathcal{X})} (-k(v)^{\lim_{n\to\infty} S_n} + \frac{1}{\lim_{n\to\infty} S_n} k(v)(1-k(v))) f_X(u) d\mathcal{H}^{p-1}(u)dv \\
=&0,
\end{aligned}
$$

where the second equality holds by Claim 11 (b) and (d).

As for variance,

$$
\begin{aligned}
&\mathrm{Var}(\frac{1}{n\delta_n}\sum_{i=1}^{n} p^s(X_i;\delta_n)^l I_i^{p,s} - \frac{1}{n\delta_n}\sum_{i=1}^{n} p^{ML}(X_i;\delta_n)^l I_i^p) \\
\le&\frac{1}{n\delta_n^2} E[(p^s(X_i;\delta_n)^l I_i^{p,s} - p^{ML}(X_i;\delta_n)^l I_i^p)^2] \\
=&\frac{1}{n\delta_n}\int_{-1}^{1}\int_{\partial\Omega^*\cap \mathrm{int}(\mathcal{X})} E[(p^s(X_i;\delta_n)^l I_i^{p,s} - p^{ML}(X_i;\delta_n)^l I_i^p)^2 | X_i = u+\delta_n v\nu(u)] 1\{u+\delta_n v\nu(u)\in\mathcal{X}\} \\
&\qquad\qquad \times f_X(u+\delta_n v\nu(u))\det(I_{p-1}+\delta_n v V_u' D^2 d^s_{\Omega^*}(u) V_u) d\mathcal{H}^{p-1}(u)dv \\
\le&\frac{1}{n\delta_n}\int_{-1}^{1}\int_{\partial\Omega^*\cap \mathrm{int}(\mathcal{X})} f_X(u+\delta_n v\nu(u))\det(I_{p-1}+\delta_n v V_u' D^2 d^s_{\Omega^*}(u) V_u) d\mathcal{H}^{p-1}(u)dv \\
\to&0.
\end{aligned}
$$

□

A.6.5 Proof of Lemma 5

Note that $p^{ML}(X_i;\delta)(1-p^{ML}(X_i;\delta)) = p^{ML}(X_i;\delta)(1-p^{ML}(X_i;\delta))1\{p^{ML}(X_i;\delta)\in(0,1)\}$. The conclusion then follows by the argument used in the proof of Claim 2. □

A.6.6 Proof of Lemma 6

Note that $p^{ML}(X_i;\delta)(1-p^{ML}(X_i;\delta)) = p^{ML}(X_i;\delta)(1-p^{ML}(X_i;\delta))1\{p^{ML}(X_i;\delta) \in (0,1)\}$. Using the argument used in the proof of Claim 9, we can show that

$$\begin{aligned}&\lim_{\delta\to 0}\delta^{-1}E[p^{ML}(X_i;\delta)(1-p^{ML}(X_i;\delta))1\{p^{ML}(X_i;\delta)\in(0,1)\}(Y_{1i}-Y_{0i})]\\ =\ &\int_{-1}^{1}k(v)(1-k(v))dv\int_{\partial\Omega^*\cap\mathcal{X}}E[Y_{1i}-Y_{0i}|X_i=x]f_X(x)d\mathcal{H}^{p-1}(x)\end{aligned}$$

and that

$$\begin{aligned}&\lim_{\delta\to 0}\delta^{-1}E[p^{ML}(X_i;\delta)(1-p^{ML}(X_i;\delta))1\{p^{ML}(X_i;\delta)\in(0,1)\}(D_i(1)-D_i(0))]\\ =\ &\int_{-1}^{1}k(v)(1-k(v))dv\int_{\partial\Omega^*\cap\mathcal{X}}E[D_i(1)-D_i(0)|X_i=x]f_X(x)d\mathcal{H}^{p-1}(x),\end{aligned}$$

where $k(v)$ is defined in Claim 8. □

A.7 Proof of Corollary 4

First, by Lemmas 1 – 4, the 2SLS estimators $\hat{\beta}_1$ and $\hat{\beta}_1^s$ converge in probability to

$$\frac{E[ML(X_i)(1-ML(X_i))(D_i(1)-D_i(0))(Y_i(1)-Y_i(0))|ML(X_i)\in(0,1)]}{E[ML(X_i)(1-ML(X_i))(D_i(1)-D_i(0))|ML(X_i)\in(0,1)]}$$

if $\Pr(ML(X_i)\in(0,1))>0$, and to

$$\frac{\int_{\partial\Omega^*\cap\mathcal{X}}E[(D_i(1)-D_i(0))(Y_i(1)-Y_i(0))|X_i=x]f_X(x)d\mathcal{H}^{p-1}(x)}{\int_{\partial\Omega^*\cap\mathcal{X}}E[D_i(1)-D_i(0)|X_i=x]f_X(x)d\mathcal{H}^{p-1}(x)}$$

if $\Pr(ML(X_i)\in(0,1))=0$.

Note that if $\Pr(ML(X_i)\in(0,1))>0$, $\Pr(D_i(1)\geq D_i(0)|X_i=x)=1$ for almost every $x\in\mathcal{X}$ such that $ML(x)\in(0,1)$, since $p^{ML}(x)=ML(x)$ for almost every $x\in\mathcal{X}$ by Claim 1 and $\Pr(D_i(1)\geq D_i(0)|X_i=x)=1$ for any $x\in\mathcal{X}$ such that $p^{ML}(x)\in(0,1)$ by assumption. Therefore,

$$\begin{aligned}&\frac{E[ML(X_i)(1-ML(X_i))(D_i(1)-D_i(0))(Y_i(1)-Y_i(0))|ML(X_i)\in(0,1)]}{E[ML(X_i)(1-ML(X_i))(D_i(1)-D_i(0))|ML(X_i)\in(0,1)]}\\ =\ &\frac{E[ML(X_i)(1-ML(X_i))E[D_i(1)-D_i(0)|X_i]LATE(X_i)|ML(X_i)\in(0,1)]}{E[ML(X_i)(1-ML(X_i))(D_i(1)-D_i(0))|ML(X_i)\in(0,1)]}.\end{aligned}$$

If $\Pr(ML(X_i)\in(0,1))=0$, $\Pr(D_i(1)\geq D_i(0)|X_i=x)=1$ for any $x\in\partial\Omega^*\cap\mathrm{int}(\mathcal{X})$, since $p^{ML}(x)=\frac{1}{2}$ for any $x\in\partial\Omega^*\cap\mathrm{int}(\mathcal{X})$ by Claim 8 and $\Pr(D_i(1)\geq D_i(0)|X_i=x)=1$ for any $x\in\mathcal{X}$ such that $p^{ML}(x)\in(0,1)$ by assumption. Therefore,

$$\begin{aligned}&\frac{\int_{\partial\Omega^*\cap\mathcal{X}}E[(D_i(1)-D_i(0))(Y_i(1)-Y_i(0))|X_i=x]f_X(x)d\mathcal{H}^{p-1}(x)}{\int_{\partial\Omega^*\cap\mathcal{X}}E[D_i(1)-D_i(0)|X_i=x]f_X(x)d\mathcal{H}^{p-1}(x)}\\ =\ &\frac{\int_{\partial\Omega^*\cap\mathcal{X}}E[D_i(1)-D_i(0)|X_i=x]LATE(x)f_X(x)d\mathcal{H}^{p-1}(x)}{\int_{\partial\Omega^*\cap\mathcal{X}}E[D_i(1)-D_i(0)|X_i=x]f_X(x)d\mathcal{H}^{p-1}(x)}.\end{aligned}$$

As in the proofs of Lemmas 5 and 6, we can show that

$$\lim_{\delta \to 0} E[\omega(X_i;\delta) LATE(X_i)] = \frac{E[ML(X_i)(1-ML(X_i))E[D_i(1)-D_i(0)|X_i] LATE(X_i) | ML(X_i) \in (0,1)]}{E[ML(X_i)(1-ML(X_i))(D_i(1)-D_i(0))|ML(X_i) \in (0,1)]}$$

if $\Pr(ML(X_i) \in (0,1)) > 0$, and that

$$\lim_{\delta \to 0} E[\omega(X_i;\delta) LATE(X_i)] = \frac{\int_{\partial\Omega^* \cap \mathcal{X}} E[D_i(1)-D_i(0)|X_i = x] LATE(x) f_X(x) d\mathcal{H}^{p-1}(x)}{\int_{\partial\Omega^* \cap \mathcal{X}} E[D_i(1)-D_i(0)|X_i = x] f_X(x) d\mathcal{H}^{p-1}(x)}$$

if $\Pr(ML(X_i) \in (0,1)) = 0$. □

A.8 Proof of Proposition 4

We can prove Part (a) using the same argument in the proof of Proposition 1 (a). For Part (b), suppose to the contrary that there exists $x_d \in \mathcal{X}_d^A$ such that $\mathcal{L}^{p_c}(\{x_c \in \mathcal{X}_c^A(x_d) : p^{ML}(x_d, x_c) \in \{0,1\}\}) > 0$. Without loss of generality, assume $\mathcal{L}^{p_c}(\{x_c \in \mathcal{X}_c^A(x_d) : p^{ML}(x_d, x_c) = 1\}) > 0$. The proof proceeds in five steps.

Step 1. $\mathcal{L}^{p_c}(\mathcal{X}_c^A(x_d) \cap \mathcal{X}_{c,1}) > 0$.

Step 2. $\mathcal{X}_c^A(x_d) \cap \text{int}(\mathcal{X}_{c,1}) \neq \emptyset$.

Step 3. $p^{ML}(x_d, x_c) = 1$ *for any* $x_c \in \text{int}(\mathcal{X}_{c,1})$.

Step 4. *For every* $x_c^* \in \mathcal{X}_c^A(x_d) \cap \text{int}(\mathcal{X}_{c,1})$*, there exists* $\delta > 0$ *such that* $N(x_c^*, \delta) \cap \mathcal{X}_c(x_d) \subset \mathcal{X}_c^A(x_d) \cap \text{int}(\mathcal{X}_{c,1})$.

Step 5. $E[Y_{1i} - Y_{0i} | X_i \in A]$ *is not identified.*

Following the argument in the proof of Proposition 1 (b), we can prove Steps 1-3. Once Step 4 is established, we prove Step 5 by following the proof of Step 4 in Proposition 1 (b) with $N(x_c^*, \epsilon) \cap \mathcal{X}_c(x_d)$ in place of $N(x^*, \epsilon)$, using the fact that $\Pr(X_{ci} \in N(x_c^*, \epsilon) \cap \mathcal{X}_c(x_d) | X_{di} = x_d) > 0$ by the definition of support. Here, we provide the proof of Step 4.

Proof of Step 4. Pick an $x_c^* \in \mathcal{X}_c^A(x_d) \cap \text{int}(\mathcal{X}_{c,1})$. Then, $x^* = (x_d, x_c^*) \in A$. Since A is open relative to $\mathcal{X}$, there exists an open set $U \in \mathbb{R}^p$ such that $A = U \cap \mathcal{X}$. This implies that for any sufficiently small $\delta > 0$, $N(x^*, \delta) \cap \mathcal{X} \subset U \cap \mathcal{X} = A$. It then follows that $\{x_c \in \mathbb{R}^{p_c} : (x_d, x_c) \in N(x^*, \delta) \cap \mathcal{X}\} \subset \{x_c \in \mathbb{R}^{p_c} : (x_d, x_c) \in A\}$, equivalently, $N(x_c^*, \delta) \cap \mathcal{X}_c(x_d) \subset \mathcal{X}_c^A(x_d)$. By choosing a sufficiently small $\delta > 0$ so that $N(x_c^*, \delta) \subset \text{int}(\mathcal{X}_{c,1})$, we have $N(x_c^*, \delta) \cap \mathcal{X}_c(x_d) \subset \mathcal{X}_c^A(x_d) \cap \text{int}(\mathcal{X}_{c,1})$. □

A.9 Proof of Theorem 2

We only provide a brief sketch of the proof, since it is mostly based on the proof of Theorem 1. Here, we consider the probability limit of $\tilde{\alpha}$, which is given by

$$\tilde{\alpha} = (\sum_{i=1}^{n} W_i W_i' I_i^p)^{-1} \sum_{i=1}^{n} W_i Y_i I_i^p,$$

where $W_i = (1, Z_i, p^{ML}(X_i; \delta_n))'$ and $I_i^p = 1\{p^{ML}(X_i; \delta_n) \in (0,1)\}$.

Case 1. We first consider the case where $\Pr(ML(X_i) \in (0,1)) > 0$. Note that

$$E[\frac{1}{n}\sum_{i=1}^{n} W_i W_i' I_i^p] = \sum_{x_d \in \mathcal{X}_d} \Pr(X_{di} = x_d) E[W_i W_i' I_i^p | X_{di} = x_d]$$

and

$$E[\frac{1}{n}\sum_{i=1}^{n} W_i Y_i' I_i^p] = \sum_{x_d \in \mathcal{X}_d} \Pr(X_{di} = x_d) E[W_i Y_i' I_i^p | X_{di} = x_d].$$

Let I_i^{ML} denote the event $\{ML(X_i) \in (0,1)\}$. Applying Claim 2 to each of the conditional expectations given X_{di}, it follows that

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} W_i W_i' I_i^p \\
&\xrightarrow{p} \sum_{x_d \in \mathcal{X}_d} \Pr(X_{di} = x_d) E\left[\begin{pmatrix} 1 & ML(X_i) & ML(X_i) \\ ML(X_i) & ML(X_i) & ML(X_i)^2 \\ ML(X_i) & ML(X_i)^2 & ML(X_i)^2 \end{pmatrix} \middle| I_i^{ML}, X_{di} = x_d \right] \Pr(I_i^{ML} | X_{di} = x_d) \\
&= E\left[\begin{pmatrix} 1 & ML(X_i) & ML(X_i) \\ ML(X_i) & ML(X_i) & ML(X_i)^2 \\ ML(X_i) & ML(X_i)^2 & ML(X_i)^2 \end{pmatrix} \middle| I_i^{ML} \right] \Pr(I_i^{ML}),
\end{aligned}$$

where we use the fact that $\Pr(I_i^{ML} | X_{di} = x_d) = \Pr(X_{di} = x_d | I_i^{ML}) \Pr(I_i^{ML}) / \Pr(X_{di} = x_d)$ for the equality. Also,

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} W_i W_i' I_i^p &\xrightarrow{p} \sum_{x_d \in \mathcal{X}_d} \Pr(X_{di} = x_d) E[(Y_i, Z_i Y_i, ML(X_i) Y_i)' | I_i^{ML}, X_{di} = x_d] \Pr(I_i^{ML} | X_{di} = x_d) \\
&= E[(Y_i, Z_i Y_i, ML(X_i) Y_i)' | I_i^{ML}] \Pr(I_i^{ML}).
\end{aligned}$$

Therefore, after a few lines of algebra, we obtain

$$\tilde{\alpha}_1 \xrightarrow{p} \frac{E[ML(X_i)(1 - ML(X_i))(E[Y_{1i} - Y_{0i} | X_i] | ML(X_i) \in (0,1)]}{E[ML(X_i)(1 - ML(X_i)) | ML(X_i) \in (0,1)]}$$

as shown in the proof of Lemma 1.

Case 2. We next consider the case where $\Pr(ML(X_i) \in (0,1)) = 0$. Note that $I_i^p = 0$ almost surely if $X_{di} \notin \mathcal{X}_d^*$. Then,

$$E[\frac{1}{n\delta_n}\sum_{i=1}^n W_i W_i' I_i^p] = \delta_n^{-1} \sum_{x_d \in \mathcal{X}_d^*} \Pr(X_{di} = x_d) E[W_i W_i' I_i^p | X_{di} = x_d]$$

and

$$E[\frac{1}{n\delta_n}\sum_{i=1}^n W_i Y_i' I_i^p] = \delta_n^{-1} \sum_{x_d \in \mathcal{X}_d^*} \Pr(X_{di} = x_d) E[W_i Y_i' I_i^p | X_{di} = x_d].$$

Applying Claim 9 to each of the conditional expectations given X_{di}, it follows that

$$\begin{aligned}
&\frac{1}{n\delta_n}\sum_{i=1}^n W_i W_i' I_i^p \\
&\xrightarrow{p} \sum_{x_d \in \mathcal{X}_d^*} \Pr(X_{di} = x_d) \int_{\partial\Omega^*(x_d) \cap \mathrm{int}(\mathcal{X}_c(x_d))} f_{X_c|X_d}(x_c|x_d) d\mathcal{H}^{p_c-1}(x_c) \begin{bmatrix} 2 & 1 & \int_{-1}^1 k(v)dv \\ 1 & 1 & \int_0^1 k(v)dv \\ \int_{-1}^1 k(v)dv & \int_0^1 k(v)dv & \int_{-1}^1 k(v)^2 dv \end{bmatrix}
\end{aligned}$$

and

$$\begin{aligned}
&\frac{1}{n\delta_n}\sum_{i=1}^n W_i Y_i' I_i^p \\
&\xrightarrow{p} \sum_{x_d \in \mathcal{X}_d^*} \Pr(X_{di} = x_d) \\
&\quad \times \begin{bmatrix} \int_{\partial\Omega^*(x_d) \cap \mathrm{int}(\mathcal{X}_c(x_d))} E[Y_{1i} + Y_{0i} | X_i = x] f_{X_c|X_d}(x_c|x_d) d\mathcal{H}^{p_c-1}(x_c) \\ \int_{\partial\Omega^*(x_d) \cap \mathrm{int}(\mathcal{X}_c(x_d))} E[Y_{1i} | X_i = x] f_{X_c|X_d}(x_c|x_d) d\mathcal{H}^{p_c-1}(x_c) \\ \int_{\partial\Omega^*(x_d) \cap \mathrm{int}(\mathcal{X}_c(x_d))} (\int_0^1 k(v)dv E[Y_{1i}|X_i = x] + \int_{-1}^0 k(v)dv E[Y_{0i}|X_i = x]) f_{X_c|X_d}(x_c|x_d) d\mathcal{H}^{p_c-1}(x_c) \end{bmatrix}.
\end{aligned}$$

After a few lines of algebra, we obtain

$$\tilde{\alpha}_1 \xrightarrow{p} \frac{\sum_{x_d \in \mathcal{X}_d^*} \Pr(X_{di} = x_d) \int_{\partial\Omega^*(x_d) \cap \mathrm{int}(\mathcal{X}_c(x_d))} E[Y_{1i} - Y_{0i} | X_i = x] f_{X_c|X_d}(x_c|x_d) d\mathcal{H}^{p_c-1}(x_c)}{\sum_{x_d \in \mathcal{X}_d^*} \Pr(X_{di} = x_d) \int_{\partial\Omega^*(x_d) \cap \mathrm{int}(\mathcal{X}_c(x_d))} f_{X_c|X_d}(x_c|x_d) d\mathcal{H}^{p_c-1}(x_c)}.$$

as shown in the proof of Lemma 2.

Finally, we can show that $\hat{\beta}_1$ and $\hat{\beta}_1^s$ converge in probability to $\lim_{\delta \to 0} E[\omega_i(\delta)(Y_i(1) - Y_i(0))]$ by the argument used in the proofs of Lemmas 5 and 6. □

B Extensions and Discussions

B.1 The Quasi Propensity Score May Not Exist

Figure 2 shows an example where QPS does not exist at $\mathbf{0}$. In this example, X_i is two dimensional, and

$$ML(x) = \begin{cases} 1 & \text{if } 3(\frac{1}{2})^{k-1} < \|x\| \le 4(\frac{1}{2})^{k-1} \text{ for some } k = 1, 2, \cdots \\ 0 & \text{if } 2(\frac{1}{2})^{k-1} < \|x\| \le 3(\frac{1}{2})^{k-1} \text{ for some } k = 1, 2, \cdots. \end{cases}$$

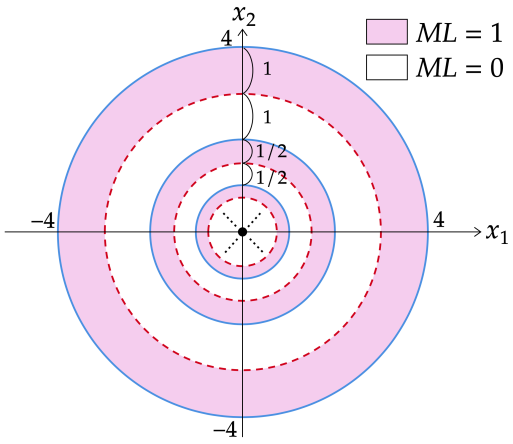

Figure 2: An example of the ML algorithm for which the Quasi Propensity Score fails to exist

It is shown that

$$p^{ML}(\mathbf{0};\delta) = \begin{cases} \frac{7}{12} & \text{if } \delta = 4(\frac{1}{2})^{k-1} \text{ for some } k = 1, 2, \cdots \\ \frac{7}{27} & \text{if } \delta = 3(\frac{1}{2})^{k-1} \text{ for some } k = 1, 2, \cdots. \end{cases}$$

Therefore, $\lim_{\delta \to 0} p^{ML}(\mathbf{0};\delta)$ does not exist.

B.2 Additional Examples

Example 5 (Movie "Eagle Eye")**.** In addition to the above algorithms already in use, our framework includes many others, including fictitious ones that may be realized in the future. For example, *Eagle Eye* is a 2008 American scientific fiction film that grossed $178 million worldwide. This movie highlights a fictitious algorithm (AI) called Eagle Eye, which is originally developed by the US government for surveillance purposes. In the film, based on its surveillance data, Eagle Eye plots the assassination of the Presidential Cabinet on the ground that Eagle Eye observes the President engaged in unconstitutional activities.

This SF scenario is another example of the above framework, though the movie does not provide a functional form for Eagle Eye. Eagle Eye determines whether to assassinate country i's President ($Z_i = 1$) or not ($Z_i = 0$). Eagle Eye uses country i's surveillance data X_i, especially information about the President's behavior, to determine its final assassination decision:

$$Z_i^{EE} \equiv 1\{\text{Eagle Eye classifies country } i\text{'s surveillance data } X_i \\ \text{as containing unconstitutional acts by its President}\}.$$

Example 6 (Surge Pricing)**.** Uber and other ride sharing services use surge pricing algorithms to adjust prices based on demand and supply. Using our notation, assume that Uber assigns one of two surge price levels, high ($Z_i = 1$) and low ($Z_i = 0$), to each session i, a company-defined measure that captures a particular rider trying to order a particular ride. Uber's surge pricing

algorithm maps local demand and supply information X_i about each session into a surge price level as follows:

$$Z_i^{SP} \equiv 1\{s(X_i) > c\}, ML^{SP}(X_i) \equiv \Pr(s(X_i) > c|X_i),$$

where $s : \mathbb{R}^p \to \mathbb{R}$ is a surge generator function and $c \in \mathbb{R}$ is a constant threshold. The surge pricing algorithms often produce regression-discontinuity-style variation, as Cohen *et al.* (2016) point out and exploit. Suppose that the function s satisfies the condition imposed on r in Example 1. The quasi propensity score for this case is given by

$$p^{SP}(x) = \begin{cases} 0 & \text{if } s(x) < c \\ 0.5 & \text{if } s(x) = c \text{ and } x \in \text{int}(\mathcal{X}) \\ 1 & \text{if } s(x) > c. \end{cases}$$

B.3 Discrete Covariates

In this section, we provide the definition of QPS and identification and consistency results when X_i includes discrete covariates. Suppose that $X_i = (X_{di}, X_{ci})$, where $X_{di} \in \mathbb{R}^{p_d}$ is a vector of discrete covariates, and $X_{ci} \in \mathbb{R}^{p_c}$ is a vector of continuous covariates. Let $\mathcal{X}_d$ denote the support of X_{di} and be assumed to be finite. We also assume that X_{ci} is continuously distributed conditional on X_{di}, and let $\mathcal{X}_c(x_d)$ denote the support of X_{ci} conditional on $X_{di} = x_d$ for each $x_d \in \mathcal{X}_d$. Let $\mathcal{X}_{c,0}(x_d) = \{x_c \in \mathcal{X}_c(x_d) : ML(x_d, x_c) = 0\}$ and $\mathcal{X}_{c,1}(x_d) = \{x_c \in \mathcal{X}_c(x_d) : ML(x_d, x_c) = 1\}$.

Define QPS as follows: for each $x = (x_d, x_c) \in \mathcal{X}$,

$$p^{ML}(x; \delta) \equiv \frac{\int_{\mathcal{X}_c(x_d) \cap N(x_c, \delta)} ML(x_d, x_c^*) dx_c^*}{\int_{\mathcal{X}_c(x_d) \cap N(x_c, \delta)} dx_c^*},$$
$$p^{ML}(x) \equiv \lim_{\delta \to 0} p^{ML}(x; \delta),$$

where $N(x_c, \delta) = \{x_c^* \in \mathbb{R}^{p_c} : \|x_c - x_c^*\| \leq \delta\}$ is the $\delta-$ball around $x_c \in \mathbb{R}^{p_c}$. Below, we assume that Assumptions 2, 3 and 4 hold conditional on X_{di}.

Assumption 2′ (Almost Everywhere Continuity of ML). *(a) For every $x_d \in \mathcal{X}_d$, $ML(x_d, \cdot)$ is continuous almost everywhere with respect to the Lebesgue measure $\mathcal{L}^{p_c}$.*

(b) For every $x_d \in \mathcal{X}_d$, $\mathcal{L}^{p_c}(\mathcal{X}_{c,k}(x_d)) = \mathcal{L}^{p_c}(\text{int}(\mathcal{X}_{c,k}(x_d)))$ for $k = 0, 1$.

B.3.1 Identification

Assumption 3′ (Local Mean Continuity). *For every $x_d \in \mathcal{X}_d$ and $z \in \{0, 1\}$, the conditional expectation functions $E[Y_{zi}|X_i = (x_d, x_c)]$ and $E[D_i(z)|X_i = (x_d, x_c)]$ are continuous in x_c at any point $x_c \in \mathcal{X}_c(x_d)$ such that $p^{ML}(x_d, x_c) \in (0, 1)$ and $ML(x_d, x_c) \in \{0, 1\}$.*

We say that a set $A \subset \mathbb{R}^p$ *is open relative to* $\mathcal{X}$ if there exists an open set $U \subset \mathbb{R}^p$ such that $A = U \cap \mathcal{X}$. For a set $A \subset \mathbb{R}^p$, let $\mathcal{X}_d^A = \{x_d \in \mathcal{X}_d : (x_d, x_c) \in A \text{ for some } x_c \in \mathbb{R}^{p_c}\}$ and $\mathcal{X}_c^A(x_d) = \{x_c \in \mathcal{X}_c : (x_d, x_c) \in A\}$ for each $x_d \in \mathcal{X}_d^A$.

Proposition 4. *Under Assumptions 1, 2′ and 3′:*

(a) $E[Y_{1i} - Y_{0i}|X_i = x]$ *and* $E[D_i(1) - D_i(0)|X_i = x]$ *are identified for every* $x \in \mathcal{X}$ *such that* $p^{ML}(x) \in (0,1)$.

(b) Let A *be any subset of* $\mathcal{X}$ *open relative to* $\mathcal{X}$ *such that* $p^{ML}(x)$ *exists for all* $x \in A$. *Then either* $E[Y_{1i} - Y_{0i}|X_i \in A]$ *or* $E[D_i(1) - D_i(0)|X_i \in A]$, *or both are identified only if* $p^{ML}(x) \in (0,1)$ *for almost every* $x_c \in \mathcal{X}_c^A(x_d)$ *for every* $x_d \in \mathcal{X}_d^A$.

B.3.2 Estimation

For each $x_d \in \mathcal{X}_d$, let $\Omega^*(x_d) = \{x_c \in \mathbb{R}^{p_c} : ML(x_d, x_c) = 1\}$. Also, let $\mathcal{X}_d^* = \{x_d \in \mathcal{X}_d : \Pr(ML(x_d, X_{ci}) = 1|X_{di} = x_d) \in (0,1)\}$, and let $f_{X_c|X_d}$ denote the probability density function of X_{ci} conditional on X_{di}.

Assumption 4′. *(a)* (Finite Second Moments) $E[Y_i(1)^2]$ *and* $E[Y_i(0)^2]$ *are finite.*

(b) (Nonzero First Stage) *There exists a constant* $c > 0$ *such that* $E[D_i(1) - D_i(0)|X_i = x] > c$ *for every* $x \in \mathcal{X}$ *such that* $p^{ML}(x) \in (0,1)$.

If $\Pr(ML(X_i) \in (0,1)) = 0$, *then the following conditions (c) – (h) hold.*

(c) (Nonzero Variance) $\mathcal{X}_d^* \neq \emptyset$.

(d) (C^2 Boundary) *For every* $x_d \in \mathcal{X}_d^*$, *there exists a partition* $\{\Omega_1^*(x_d), ..., \Omega_M^*(x_d)\}$ *of* $\Omega^*(x_d)$ *such that*

(i) $\text{dist}(\Omega_m^*(x_d), \Omega_{m'}^*(x_d)) > 0$ *for any* $m, m' \in \{1, ..., M\}$ *such that* $m \neq m'$;

(ii) $\Omega_m^*(x_d)$ *is nonempty, bounded, open, connected and twice continuously differentiable for each* $m \in \{1, ..., M\}$.

(e) ((p_c-1)-dimensional Boundary) *For every* $x_d \in \mathcal{X}_d^*$, $\mathcal{H}^{p_c-1}(\partial\Omega^*(x_d)) < \infty$, $\mathcal{H}^{p_c-1}(\partial\Omega^*(x_d) \cap \partial\mathcal{X}_c(x_d)) = 0$, *and* $\int_{\partial\Omega^*(x_d)\cap\text{int}(\mathcal{X}_c(x_d))} f_{X_c|X_d}(x_c|x_d) d\mathcal{H}^{p_c-1}(x) > 0$.

(f) (Local Mean Continuity) *For every* $x_d \in \mathcal{X}_d^*$ *and* $z \in \{0,1\}$, $E[Y_{zi}|X_i = (x_d, x_c)]$ *and* $E[D_i(z)|X_i = (x_d, x_c)]$ *are continuous at every point* $x_c \in \partial\Omega^*(x_d) \cap \text{int}(\mathcal{X}_c(x_d))$.

(g) (Continuous and Bounded Density) *For every* $x_d \in \mathcal{X}_d^*$, $f_{X_c|X_d}(\cdot|x_d)$ *is continuous at every point* $x_c \in \partial\Omega^*(x_d) \cap \text{int}(\mathcal{X}_c(x_d))$. *In addition, there exists* $\delta > 0$ *such that* $f_{X_c|X_d}(\cdot|x_d)$ *is bounded on* $\bar{N}(\partial\Omega^*(x_d) \cap \mathcal{X}_c(x_d), \delta)$.

(h) (Bounded Second Moments) *For every* $x_d \in \mathcal{X}_d^*$, *there exists* $\delta > 0$ *such that* $E[Y_i(1)^2|X_i = (x_d, x_c)]$ *and* $E[Y_i(0)^2|X_i = (x_d, x_c)]$ *are bounded on* $\bar{N}(\partial\Omega^*(x_d) \cap \mathcal{X}_c(x_d), \delta)$.

Theorem 2. *Suppose that Assumptions 1, 2′ and 4′ hold,* $\delta_n \to 0$, $n\delta_n \to \infty$ *and* $S_n \to \infty$ *as* $n \to \infty$. *Then the 2SLS estimators* $\hat{\beta}_1$ *and* $\hat{\beta}_1^s$ *converge in probability to*

$$\lim_{\delta \to 0} E[\omega_i(\delta)(Y_i(1) - Y_i(0))],$$

where

$$\omega_i(\delta) = \frac{p^{ML}(X_i;\delta)(1-p^{ML}(X_i;\delta))(D_i(1)-D_i(0))}{E[p^{ML}(X_i;\delta)(1-p^{ML}(X_i;\delta))(D_i(1)-D_i(0))]}.$$

More specifically, if $\Pr(ML(X_i) \in (0,1)) > 0$*, then the 2SLS estimators* $\hat{\beta}_1$ *and* $\hat{\beta}_1^s$ *converge in probability to*

$$\frac{E[ML(X_i)(1-ML(X_i))(D_i(1)-D_i(0))(Y_i(1)-Y_i(0))]}{E[ML(X_i)(1-ML(X_i))(D_i(1)-D_i(0))]}.$$

If $\Pr(ML(X_i) \in (0,1)) = 0$*, then* $\hat{\beta}_1$ *and* $\hat{\beta}_1^s$ *converge in probability to*

$$\frac{\sum_{x_d \in \mathcal{X}_d^*} \Pr(X_{di} = x_d) \int_{\partial\Omega^*(x_d)\cap\mathcal{X}_c(x_d)} E[(D_i(1)-D_i(0))(Y_i(1)-Y_i(0))|X_i = x] f_{X_c|X_d}(x_c|x_d) d\mathcal{H}^{p_c-1}(x_c)}{\sum_{x_d \in \mathcal{X}_d^*} \Pr(X_{di} = x_d) \int_{\partial\Omega^*(x_d)\cap\mathcal{X}_c(x_d)} E[D_i(1)-D_i(0)|X_i = x] f_{X_c|X_d}(x_c|x_d) d\mathcal{H}^{p_c-1}(x_c)}.$$

B.4 Sampling from Uniform Distribution on p-Dimensional Ball

When we calculate QPS by simulation, we need to uniformly sample from $N(X_i;\delta)$. We introduce three existing methods to uniformly sample from a p-dimensional unit ball $N(\mathbf{0},1)$.

Method 1.
1. Sample $x_1, ..., x_p$ independently from the uniform distribution on $[-1,1]$.
2. Accept the vector $x = (x_1, ..., x_p)$ if $\sum_{k=1}^p x_k^2 \leq 1$ and reject it otherwise.

Method 1 is a practical choice when p is small (e.g. $p = 2, 3$), but is inefficient for higher dimensions, since the acceptance rate decreases to zero quickly as p increases. The conventional method used for higher dimensions is the following.

Method 2.
1. Sample $x_1^*, ..., x_p^*$ independently from the standard normal distribution, and compute the vector $s = (x_1^*, ..., x_p^*)/\sqrt{\sum_{k=1}^p (x_k^*)^2}$.
2. Sample u from the uniform distribution on $[0,1]$.
3. Return the vector $x = u^{1/p}s$.

There is yet another method efficient for higher dimensions, which is recently proposed by Voelker, Gosmann and Stewart (2017).

Method 3.
1. Sample $x_1^*, ..., x_{p+2}^*$ independently from the standard normal distribution, and compute the vector $s = (x_1^*, ..., x_{p+2}^*)/\sqrt{\sum_{k=1}^{p+2} (x_k^*)^2}$.
2. Return the vector $x = (s_1, ..., s_p)$.