Algorithm is Experiment: Machine Learning, Market Design, and Policy Eligibility Rules

Yusuke Narita Kohei Yata*

September 10, 2021

Abstract

Algorithms produce a growing portion of decisions and recommendations both in policy and business. 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 develop a treatment-effect estimator for a class of stochastic and deterministic decision-making algorithms. Our estimator is shown to be consistent and asymptotically normal for well-defined causal effects. A key special case of our estimator is a multidimensional regression discontinuity design. We apply our estimator to evaluate the effect of the Coronavirus Aid, Relief, and Economic Security (CARES) Act, where hundreds of billions of dollars worth of relief funding is allocated to hospitals via an algorithmic rule. Our estimates suggest that the relief funding has little effect on COVID-19-related hospital activity levels. Naive OLS and IV estimates exhibit substantial selection bias.

*Narita: Yale University, email: yusuke.narita@yale.edu. Yata: Yale University, email: kohei.yata@yale.edu. For their suggestions, we are grateful to Josh Angrist, Tim Armstrong, Yingying Dong, Pat Kline, Michal Kolesár, Chris Walters, and seminar participants at the American Economic Association, UC Berkeley, Caltech, Columbia, CEMFI, Counterfactual Machine Learning Workshop, Econometric Society, European Economic Association, Hitotsubashi, JSAI, Stanford, UC Irvine, the University of Tokyo, and Yale. We are especially indebted to Aneesha Parvathaneni, Richard Liu, Richard Gong, and many others for expert research assistance.

1 Introduction

Today's society increasingly resorts to algorithms for decision-making and resource allocation. For example, judges in the US make legal decisions aided by predictions from supervised machine learning algorithms. Supervised learning is also used by governments to detect potential criminals and terrorists, and by banks and insurance companies to screen potential customers. Tech companies like Facebook, Microsoft, and Netflix allocate digital content by reinforcement learning and bandit algorithms. Retailers and e-commerce platforms engage in algorithmic pricing. Similar algorithms are encroaching on high-stakes settings, such as in education, healthcare, and the military.

Other types of algorithms also loom large. School districts, college admissions systems, and labor markets use matching algorithms for position and seat allocations. Objects worth astronomical sums of money change hands every day in algorithmically run auctions. Many public policy domains like Medicaid often use algorithmic rules to decide who are eligible.

All of the above, diverse examples share a common trait: a decision-making algorithm makes decisions based only on its observable input variables. Thus conditional on the observable variables, algorithmic treatment decisions are (quasi-)randomly assigned. That is, they are independent of any potential outcome or unobserved heterogeneity. This property turns algorithm-based treatment decisions into instrumental variables (IVs) that can be used for measuring the causal effect of the final treatment assignment. The algorithm-based instrument may produce stratified randomization, regression-discontinuity-style local variation, or some combination of the two.

This paper shows how to use data obtained from algorithmic decision-making to identify and estimate causal effects. In our framework, the analyst observes a random sample $\{(Y_i, X_i, D_i, Z_i)\}_{i=1}^n$, where Y_i is the outcome of interest, $X_i \in \mathbb{R}^p$ is a vector of pre-treatment covariates used as the algorithm's input variables, D_i is the binary treatment assignment, possibly made by humans, and Z_i is the binary treatment recommendation made by a known algorithm. The algorithm takes X_i as input and computes the probability of the treatment recommendation $A(X_i) = \Pr(Z_i = 1|X_i)$. Z_i is then randomly determined based on the known probability $A(X_i)$ independently of everything else conditional on X_i. The algorithm's recommendation Z_i may influence the final treatment assignment D_i, determined as $D_i = Z_i D_i(1) + (1 - Z_i)D_i(0)$, where $D_i(z)$ is the potential treatment assignment that would be realized if $Z_i = z$. Finally, the observed outcome Y_i is determined as $Y_i = D_i Y_i(1) + (1 - D_i)Y_i(0)$, where $Y_i(1)$ and $Y_i(0)$ are potential outcomes that would be realized if the individual were treated and not treated, respectively. This setup is an IV model where the IV satisfies the conditional independence condition but may not satisfy the overlap (full-support) condition. To our knowledge, there is no standard estimator for this setup.

Within this framework, we first characterize the sources of causal-effect identification for a class of data-generating algorithms. This class includes all of the aforementioned examples, nesting both stochastic and deterministic algorithms. The sources of causal-effect identification turn out to be summarized by a suitable modification of the Propensity Score (Rosenbaum and Rubin, 1983). We call it the *Approximate Propensity Score* (APS). For each covariate value x, the Approximate Propensity Score is the average probability of a treatment recommendation in

a shrinking neighborhood around x, defined as

$$p^A(x) \equiv \lim_{\delta \to 0} \frac{\int_{B(x,\delta)} A(x^*) dx^*}{\int_{B(x,\delta)} dx^*},$$

where $B(x, \delta)$ is a p-dimensional ball with radius δ centered at x. The Approximate Propensity Score provides an easy-to-check condition for what causal effects the data from an algorithm allow us to identify. In particular, we show that the conditional local average treatment effect (LATE; Imbens and Angrist, 1994) at covariate value x is identified if and only if the Approximate Propensity Score is nondegenerate, i.e., $p^A(x) \in (0, 1)$.

The identification analysis suggests a way of estimating treatment effects using the algorithm-produced data. The treatment effects can be estimated by two-stage least squares (2SLS) where we regress the outcome on the treatment with the algorithm's recommendation as an IV. To make the algorithmic recommendation a conditionally independent IV, we propose to control for the Approximate Propensity Score. A more precise definition is as follows.

1. For small bandwidth $\delta > 0$ and a large number of simulation draws S, compute

$$p^s(X_i; \delta) = \frac{1}{S} \sum_{s=1}^{S} A(X^*_{i,s}),$$

   where $X^*_{i,1}, ..., X^*_{i,S}$ are S independent simulation draws from the uniform distribution on $B(X_i, \delta)$.[1] This $p^s(X_i; \delta)$ is a simulation-based approximation to the Approximate Propensity Score $p^A(x)$.

2. Using the observations with $p^s(X_i; \delta) \in (0, 1)$, run the following 2SLS IV regression:

$$D_i = \gamma_0 + \gamma_1 Z_i + \gamma_2 p^s(X_i; \delta) + \nu_i \text{ (First Stage)}$$
$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 p^s(X_i; \delta) + \epsilon_i \text{ (Second Stage)}.$$

   Let $\hat{\beta}^s_1$ be the estimated coefficient on D_i.

As the main theoretical result, we prove the 2SLS estimator $\hat{\beta}^s_1$ is a consistent and asymptotically normal estimator of a well-defined causal effect (weighted average of conditional local average treatment effects). We also show that inference based on the conventional 2SLS heteroskedasticity-robust standard errors is asymptotically valid as long as the bandwidth δ goes to zero at an appropriate rate. There appears to be no existing estimator with these asymptotic properties even for the multidimensional RDD, a special case of our framework where the decision-making algorithm is deterministic and uses multiple input (running) variables for assigning treatment recommendations. Moreover, our result applies to much more general settings with stochastic algorithms, deterministic algorithms, and combinations of the two. We prove the

[1] To make common δ for all dimensions reasonable, we standardize each characteristic X_{ij} $(j = 1, ..., p)$ to have mean zero and variance one, where p is the number of input characteristics. For the bandwidth δ, we suggest that the analyst considers several different values and check if the 2SLS estimates are robust to bandwidth changes, as we often do in regression discontinuity design (RDD) applications.

asymptotic properties by exploiting results from differential geometry and geometric measure theory, which may be of independent interest.

The practical performance of our estimator is demonstrated through simulation and an original application. We first conduct a Monte Carlo simulation mimicking real-world decision-making based on machine learning algorithms. We consider a data-generating process combining stochastic and deterministic algorithms. Treatment recommendations are randomly assigned for a small experimental segment of the population and are determined by a high-dimensional, deterministic machine learning algorithm for the rest of the population. Our estimator is shown to be feasible in this high-dimensional setting and have smaller mean squared errors relative to alternative estimators.

Our empirical application is an analysis of COVID-19 hospital relief funding. The Coronavirus Aid, Relief, and Economic Security (CARES) Act and Paycheck Protection Program designated $175 billion for COVID-19 response efforts and reimbursement to health care entities for expenses or lost revenues (Kakani, Chandra, Mullainathan and Obermeyer, 2020). This policy intended to help hospitals hit hard by the pandemic, as "*financially insecure hospitals may be less capable of investing in COVID-19 response efforts*" (Khullar, Bond and Schpero, 2020). We ask whether this problem is alleviated by the relief funding to hospitals.

We identify the causal effects of the relief funding by exploiting the funding eligibility rule. The government employs an algorithmic rule to decide which hospitals are eligible for funding. This fact allows us to apply our method to estimate the effect of relief funding. Specifically, our 2SLS estimators use eligibility status as an instrumental variable for funding amounts, while controlling for the Approximate Propensity Score induced by the eligibility-determining algorithm.

The resulting estimates suggest that COVID-19 relief funding has little to no effect on outcomes, such as the number of COVID-19 patients hospitalized at each hospital. The estimated causal effects of relief funding are much smaller and less significant than the naive ordinary least squares (OLS) (with and without controls) or 2SLS estimates with no controls. Our finding provides causal evidence for the concern that funding in the CARES Act might not be well targeted to the clinics and hospitals with the greatest needs.[2]

Related Literature

Theoretically, our framework integrates the classic propensity-score (selection-on-observables) scenario with a multidimensional extension of the RDD. We analyze this integrated setup in the IV world with noncompliance. This general setting appears to have no prior established estimator. Armstrong and Kolesár (2020) provide an estimator for a related setting with perfect compliance.[3]

[2]See, for example, Kakani *et al.* (2020) as well as *Forbes*'s article, "Hospital Giant HCA To Return $6 Billion in CARES Act Money," at `https://www.forbes.com/sites/brucejapsen/2020/10/08/hospital-giant-hca-to-return-6-billion-in-cares-act-money`, retrieved September 2021.

[3]Building on their prior work (Armstrong and Kolesár, 2018), Armstrong and Kolesár (2020) consider estimation and inference on average treatment effects under the assumption that the final treatment assignment is independent of potential outcomes conditional on observables. Their estimator is not applicable to the IV world we consider. Their method and our method also achieve different goals; their goal lies in finite-sample optimality and asymptotically valid inference while our goal is to obtain consistency, asymptotic normality, and asymptotically

When we adapt our estimator to the multidimensional RDD case, our estimator has three features. First, it is a consistent and asymptotically normal estimator of a well-interpreted causal effect (average of conditional treatment effects along the RDD boundary) even if treatment effects are heterogeneous. Second, it uses observations near all the boundary points as opposed to using only observations near one specific boundary point, thus avoiding variance explosion even when X_i has many elements. Third, it can be easily implemented even in cases with many covariates and complex algorithms (RDD boundaries). Our method circumvents the difficulty of identifying the decision boundary from a complicated decision-making algorithm. No prior estimator appears to have all of these properties (Papay, Willett and Murnane, 2011; Zajonc, 2012; Keele and Titiunik, 2015; Cattaneo, Titiunik, Vazquez-Bare and Keele, 2016; Imbens and Wager, 2019). Appendix A.1 provides a detailed review of the most closely related papers on the multidimensional RDD.

The Approximate Propensity Score developed in this paper shares its spirit with the local random assignment interpretation of the RDD, discussed by Frölich (2007), Cattaneo, Frandsen and Titiunik (2015), Cattaneo, Titiunik and Vazquez-Bare (2017), Frandsen (2017), Sekhon and Titiunik (2017), Frölich and Huber (2019), Abdulkadiroğlu, Angrist, Narita and Pathak (Forthcoming) and Eckles, Ignatiadis, Wager and Wu (2020). These papers consider settings that fit into this paper's framework.

Our estimator is applicable to a class of data-generating algorithms that includes stochastic and deterministic algorithms used in practice. Our results thus nest existing insights on quasi-experimental variation in particular algorithms, such as surge pricing (Cohen, Hahn, Hall, Levitt and Metcalfe, 2016), bandit (Li, Chu, Langford and Schapire, 2010), reinforcement learning (Precup, 2000), supervised learning (Cowgill, 2018; Bundorf, Polyakova and Tai-Seale, 2019), and market-design algorithms (Abdulkadiroğlu, Angrist, Narita and Pathak, 2017; Abdulkadiroğlu *et al.*, Forthcoming; Abdulkadiroğlu, 2013; Kawai, Nakabayashi, Ortner and Chassang, 2020; Narita, 2020, 2021). Our framework also reveals new sources of identification for algorithms that, at first sight, do not appear to produce a natural experiment.[4]

Our empirical application uses the proposed method to study hospitals receiving CARES Act relief funding. Our empirical finding contributes to emerging work on how health care providers respond to financial shocks (Duggan, 2000; Adelino, Lewellen and Sundaram, 2015; Dranove, Garthwaite and Ody, 2017; Adelino, Lewellen and McCartney, 2021). Our empirical setting

valid inference.

[4]A focal group of decision-making algorithms are machine learning algorithms, as illustrated in our machine-learning simulation. 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, Lakkaraju, Leskovec, Ludwig and Mullainathan, 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, 2019; Mullainathan and Spiess, 2017). We supplement these studies by highlighting the role of machine learning as a data-production tool. This paper also has a conceptual connection to the heated conversation about whether algorithmic decisions are better than human decisions. Here "better" is in terms of fairness and efficiency (Hoffman, Kahn and Li, 2017; Horton, 2017; Kleinberg *et al.*, 2017). In this study, we take a complementary perspective in that we take a decision algorithm as given, no matter whether it is good or bad, and study how to use its produced data for impact evaluation.

is a healthcare crisis, complementing prior work on more normal situations. Our analysis also exploits rule-based locally random assignment of cash flows to hospitals. This feature provides our estimates with additional confidence in their causal interpretation.

2 Framework

Our framework is a mix of the conditional independence, multidimensional RDD, and instrumental variable scenarios. In the setup in the introduction, we are interested in the effect of some binary treatment $D_i \in \{0,1\}$ on some outcome of interest $Y_i \in \mathbb{R}$. As is standard in the literature, we impose the exclusion restriction that the treatment recommendation $Z_i \in \{0,1\}$ 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.[5]

We consider algorithms that make treatment recommendations based solely on individual i's predetermined, observable covariates $X_i = (X_{i1}, ..., X_{ip})' \in \mathbb{R}^p$. Let the function $A : \mathbb{R}^p \to [0,1]$ represent the decision algorithm, where $A(X_i) = \Pr(Z_i = 1|X_i)$ is the probability that the treatment is recommended for individual i with covariates X_i. The central assumption is that the analyst knows function A and is able to simulate it. That is, the analyst is able to compute the recommendation probability $A(x)$ given any input value $x \in \mathbb{R}^p$. The treatment recommendation Z_i for individual i is then randomly determined with probability $A(X_i)$ independently of everything else. Consequently, the following conditional independence property holds.

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

Note that the codomain of A contains 0 and 1, allowing for deterministic treatment assignments conditional on X_i. Our framework therefore nests the RDD as a special case. Another special case of our framework is the classic conditional independence scenario with the common support condition ($A(X_i) \in (0,1)$ almost surely). In addition to these simple settings, this framework nests many other situations, such as multidimensional RDDs and complex machine learning and market-design algorithms, as illustrated in Section 7.

In typical machine-learning scenarios, an algorithm first applies machine learning on X_i to make some prediction and then uses the prediction to output the recommendation probability $A(X_i)$, as in the following example.

Example. Automated disease detection algorithms use machine learning, in particular deep learning, to detect various diseases and to identify patients at risk (Gulshan *et al.*, 2016). Using our framework described above, a detection algorithm predicts whether an individual i has a certain disease ($Z_i = 1$) or not ($Z_i = 0$) based on a digital image $X_i \in \mathbb{R}^p$ of a part of the individual's body, where each $X_{ij} \in \mathbb{R}$ denotes the intensity value of a pixel in the image. The algorithm uses training data to construct a binary classifier $A : \mathbb{R}^p \to \{0,1\}$. The classifier takes an image of individual i as input and makes a binary prediction of whether the individual has

[5]Formally, let $Y_i(d,z)$ denote the potential outcome that would be realized if i's treatment assignment and recommendation were d and z, respectively. The exclusion restriction assumes that $Y_i(d,1) = Y_i(d,0)$ for $d \in \{0,1\}$ (Imbens and Angrist, 1994).

the disease:

$$Z_i \equiv A(X_i).$$

The algorithm's diagnosis Z_i may influence the doctor's treatment decision for the individual, denoted by $D_i \in \{0, 1\}$. We are interested in how the treatment decision D_i affects the individual's outcome Y_i.

Let Y_{zi} be 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 potential outcome when the treatment recommendation is $Z_i = z$. It follows from Property 1 that $Z_i \perp\!\!\!\perp (Y_{1i}, Y_{0i})|X_i$.

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

Assumption 1.

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

(b) (Measure Zero Boundaries of $\mathcal{X}_0$ and $\mathcal{X}_1$) $\mathcal{L}^p(\mathcal{X}_k) = \mathcal{L}^p(\text{int}(\mathcal{X}_k))$ *for* $k = 0, 1$.

Assumption 1 (a) allows the function A to be discontinuous on a set of points with the Lebesgue measure zero. For example, A is allowed to be a discontinuous step function as long as it is continuous almost everywhere. Assumption 1 (b) holds if the Lebesgue measures of the boundaries of $\mathcal{X}_0$ and $\mathcal{X}_1$ are zero.

3 Identification

What causal effects can be learned from data (Y_i, X_i, D_i, Z_i) generated by the algorithm A? A key step toward answering this question is what we call the *Approximate Propensity Score* (APS). To define it, we first define the *fixed-bandwidth Approximate Propensity Score* as follows:

$$p^A(x; \delta) \equiv \frac{\int_{B(x,\delta)} A(x^*)dx^*}{\int_{B(x,\delta)} dx^*},$$

where $B(x, \delta) = \{x^* \in \mathbb{R}^p : \|x - x^*\| < \delta\}$ is the (open) δ-ball around $x \in \mathcal{X}$.[6] Here, $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^p$. To make a common bandwidth δ for all dimensions reasonable, we

[6]Whether we use an open ball or closed ball does not affect $p^A(x; \delta)$. We use a ball for simplicity. When we instead use a rectangle, ellipsoid, or any standard kernel function to define $p^A(x; \delta)$, the limit $\lim_{\delta \to 0} p^A(x; \delta)$ may be different at some points (e.g., at discontinuity points of A), but the same identification results hold under suitable conditions.

normalize X_{ij} to have mean zero and variance one for each $j = 1, ..., p$.[7] We assume that A is a $\mathcal{L}^p$-measurable function so that the integrals exist. We then define APS as follows:

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

APS at x is the average probability of a treatment recommendation in a shrinking ball around x. We call this the *Approximate* Propensity Score, since this score modifies the standard propensity score $A(X_i)$ to incorporate local variation in the score. APS exists for most covariate points and algorithms (see Appendix A.2).

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

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). Our identification analysis uses the following continuity condition.

Assumption 2 (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 $p^A(x) \in (0, 1)$ and $A(x) \in \{0, 1\}$.*

Assumption 2 is a multivariate extension of the local mean continuity condition that is frequently assumed in the RDD.[8] $A(x) \in \{0, 1\}$ means that the treatment recommendation Z_i is deterministic conditional on $X_i = x$. If APS at the point x is nondegenerate ($p^A(x) \in (0, 1)$), however, there exists a point close to x that has a different value of A from x's, which creates variation in the treatment recommendation near x. For any such point x, Assumption 2 requires that the points close to x have similar conditional means of the outcome Y_{zi} and treatment assignment $D_i(z)$.[9] Note that Assumption 2 does not require continuity of the conditional means at x for which $A(x) \in (0, 1)$, since the identification of the conditional means at such points follows from Property 1 without continuity.

Under the above assumptions, APS provides an easy-to-check condition for whether an algorithm allows us to identify causal effects.

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

[8] In the RDD with a single running variable, the point x for which $p^A(x) \in (0, 1)$ and $A(x) \in \{0, 1\}$ is the cutoff point at which the treatment probability discontinuously changes.

[9] In the context of the RDD with a single running variable, one sufficient condition for continuity of $E[Y_{zi}|X_i]$ is a local independence condition in the spirit of Hahn, Todd and van der Klaauw (2001): $(Y_i(1), Y_i(0), D_i(1), D_i(0))$ is independent of 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. These two sets of conditions are sufficient for continuity of $E[Y_{zi}|X_i]$ regardless of the dimension of X_i, accommodating multidimensional RDDs.

Proposition 1 (Identification)**.** *Under Assumptions 1 and 2:*

(a) $E[Y_{1i} - Y_{0i}|X_i = x]$ *and* $E[D_i(1) - D_i(0)|X_i = x]$ *are identified for every* $x \in \text{int}(\mathcal{X})$ *such that* $p^A(x) \in (0, 1)$.[10]

(b) Let S *be any open subset of* $\mathcal{X}$ *such that* $p^A(x)$ *exists for all* $x \in S$*. Then either* $E[Y_{1i} - Y_{0i}|X_i \in S]$ *or* $E[D_i(1) - D_i(0)|X_i \in S]$ *or both are identified only if* $p^A(x) \in (0, 1)$ *for almost every* $x \in S$ *(with respect to the Lebesgue measure).*[11]

Proof. See Appendix C.1. □

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 APS at x is neither 0 nor 1. Non-degeneracy of APS 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 2 ensures that these 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, $p^A(x) \in (0, 1)$ 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).

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 S is identified, then we must have $p^A(x) \in (0, 1)$ for almost every $x \in S$. If, to the contrary, there is a subset of S of nonzero measure for which $p^A(x) = 1$ (or $p^A(x) = 0$), then Z_i has no variation in the subset, which makes it impossible to identify the average effect for the subset.

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.

Corollary 1 (Perfect and Imperfect Compliance)**.** *Under Assumptions 1 and 2:*

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

(b) 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* $x \in \text{int}(\mathcal{X})$ *such that* $p^A(x) \in (0, 1)$*,* $\Pr(D_i(1) \geq D_i(0)|X_i = x) = 1$ *(monotonicity), and* $\Pr(D_i(1) \neq D_i(0)|X_i = x) > 0$ *(existence of compliers).*

Proof. See Appendix C.2. □

[10] The causal effects may not be identified at a boundary point x of $\mathcal{X}$ for which $p^A(x) \in (0, 1)$. For example, if $A(x^*) = 1$ for all $x^* \in B(x, \delta) \cap \mathcal{X}$ and $A(x^*) = 0$ for all $x^* \in B(x, \delta) \setminus \mathcal{X}$ for any sufficiently small $\delta > 0$, $p^A(x) \in (0, 1)$ but the causal effects are not identified at x since $\Pr(Z_i = 0|X_i \in B(x, \delta)) = 0$.

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

Non-degeneracy of APS $p^A(x)$ therefore summarizes what causal effects the data from A identify. Note that the key condition ($p^A(x) \in (0,1)$) holds for some points x for every standard algorithm except trivial algorithms that always recommend a treatment with probability 0 or 1. Therefore, the data from almost every algorithm identify some causal effect.

4 Estimation

The sources of quasi-random assignment characterized in Proposition 1 suggest a way of estimating causal effects of the treatment. In view of Proposition 1, it is possible to nonparametrically estimate conditional average causal effects $E[Y_{1i} - Y_{0i}|X_i = x]$ and $E[D_i(1) - D_i(0)|X_i = x]$ for points x such that $p^A(x) \in (0,1)$. This approach is hard to use in practice, however, when X_i has many elements.

We instead seek an estimator that aggregates conditional effects at different points into a single average causal effect.[12] Proposition 1 suggests that conditioning on APS makes algorithm-based treatment recommendation quasi-randomly assigned. This motivates the use of an algorithm's recommendation as an instrument conditional on APS, which we operationalize as follows.

4.1 Two-Stage Least Squares Meets APS

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 as described in the introduction and Section 2. Consider the following 2SLS regression using the observations with $p^A(X_i;\delta_n) \in (0,1)$:

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

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

where bandwidth δ_n shrinks toward zero as the sample size n increases. Let $I_{i,n} = 1\{p^A(X_i;\delta_n) \in (0,1)\}$, $\mathbf{D}_{i,n} = (1, D_i, p^A(X_i;\delta_n))'$, and $\mathbf{Z}_{i,n} = (1, Z_i, p^A(X_i;\delta_n))'$. The 2SLS estimator $\hat{\beta}$ is then given by

$$\hat{\beta} = (\sum_{i=1}^n \mathbf{Z}_{i,n}\mathbf{D}_{i,n}' I_{i,n})^{-1} \sum_{i=1}^n \mathbf{Z}_{i,n} Y_i I_{i,n}.$$

Let $\hat{\beta}_1$ denote the 2SLS estimator of β_1 in the above regression.[13]

[12] If the analyst is interested in heterogeneity in terms of covariates, it is also possible to split the sample into subgroups based on covariates and apply our method separately to different subgroups.

[13] For the standard RDD with a single running variable $X_i \in \mathbb{R}$ and cutoff c, $p^A(X_i;\delta_n) = \frac{X_i - c}{2\delta_n} + \frac{1}{2}$ if $X_i \in [c - \delta_n, c + \delta_n]$ and $p^A(X_i;\delta_n) \in \{0,1\}$ otherwise. In this special case, the estimator $\hat{\beta}_1$ from the 2SLS regression (1) and (2) is numerically equivalent to a version of the RD local linear estimator (Hahn *et al.*, 2001) which uses a box kernel and places the same slope coefficient of X_i on both sides of the cutoff. It is possible to allow for slope changes at the cutoff by viewing $p^A(X_i;\delta_n)$ as a running variable with cutoff $\frac{1}{2}$ and applying standard RD local linear estimators (i.e., adding interaction terms $D_i(p^A(X_i;\delta_n) - \frac{1}{2})$ and $Z_i(p^A(X_i;\delta_n) - \frac{1}{2})$ to (1) and (2), respectively). However, it is not straightforward to extend this approach to the multidimensional

The above regression uses true fixed-bandwidth APS $p^A(X_i;\delta_n)$, but it may be difficult to analytically compute if A is complex. In such a case, we propose to approximate $p^A(X_i;\delta_n)$ using brute force simulation. We draw a value of x from the uniform distribution on $B(X_i,\delta_n)$ a number of times, compute $A(x)$ for each draw, and take the average of $A(x)$ over the draws.[14] Formally, let $X^*_{i,1},...,X^*_{i,S_n}$ be S_n independent draws from the uniform distribution on $B(X_i,\delta_n)$, and calculate

$$p^s(X_i;\delta_n)=\frac{1}{S_n}\sum_{s=1}^{S_n}A(X^*_{i,s}).$$

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 $p^A(X_i;\delta_n)$ has a $1/\sqrt{S_n}$ rate of convergence.[15] This rate does not depend on the dimension of X_i, so the simulation error can be made negligible even when X_i has many elements.

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

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

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

Let $\hat{\beta}^s_1$ 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 it uses the simulated fixed-bandwidth APS $p^s(X_i;\delta_n)$ in place of $p^A(X_i;\delta_n)$.[16]

4.2 Consistency and Asymptotic Normality

We establish the consistency and asymptotic normality of the 2SLS estimators $\hat{\beta}_1$ and $\hat{\beta}^s_1$. Our consistency and asymptotic normality result uses the following assumptions.

Assumption 3.

(a) (Finite Moment) $E[Y_i^4]<\infty$.

(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* $p^A(x)\in(0,1)$.

RDD, since the value of $p^A(X_i;\delta_n)$ no longer determines whether $Z_i=1$ or $Z_i=0$ unless the RD boundary is linear, which may invalidate the use of $p^A(X_i;\delta_n)$ as a single running variable. We leave to future research how to allow for more flexible 2SLS specifications in the general multi-dimensional setting.

[14] See Appendix A.5 for how to efficiently sample from the uniform distribution on a p-dimensional ball.

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

[16] In many industry and policy applications, the analyst is only able to change the algorithm's recommendation Z_i by redesigning the algorithm. In this case, the effect of recommendation Z_i on outcome Y_i may also be of interest. We can estimate the effect of recommendation by running the following ordinary least squares (OLS) regression using the observations with $p^s(X_i;\delta)\in(0,1)$:

$$Y_i=\alpha_0+\alpha_1 Z_i+\alpha_2 p^s(X_i;\delta)+u_i.$$

The estimated coefficient on Z_i, $\hat{\alpha}^s_1$, is our preferred estimator of the recommendation effect.

(c) (Nonzero Conditional Variance) *If* $\Pr(A(X_i) \in (0,1)) > 0$, *then* $\text{Var}(A(X_i)|A(X_i) \in (0,1)) > 0$.

If $\Pr(A(X_i) \in (0,1)) = 0$, *then the following conditions (d)–(g) hold.*

(d) (Nonzero Variance) $\text{Var}(A(X_i)) > 0$.

For a set $S \subset \mathbb{R}^p$, *let* $\text{cl}(S)$ *denote the closure of* S *and let* ∂S *denote the boundary of* S, *i.e.,* $\partial S = \text{cl}(S) \setminus \text{int}(S)$.

(e) (C^2 Boundary of Ω^*) *There exists a partition* $\{\Omega_1^*, ..., \Omega_M^*\}$ *of* $\Omega^* = \{x \in \mathbb{R}^p : A(x) = 1\}$ *(the set of the covariate points whose* A *value is one) such that*

(i) $\text{dist}(\Omega_m^*, \Omega_{m'}^*) > 0$ *for any* $m, m' \in \{1, ..., M\}$ *such that* $m \neq m'$. *Here* $\text{dist}(S,T) = \inf_{x \in S, y \in T} \|x - y\|$ *is the distance between two sets* S *and* $T \subset \mathbb{R}^p$;

(ii) Ω_m^* *is nonempty, bounded, open, connected and twice continuously differentiable for each* $m \in \{1, ..., M\}$. *Here we say that a bounded open set* $S \subset \mathbb{R}^p$ *is twice continuously differentiable if for every* $x \in S$, *there exists a ball* $B(x, \epsilon)$ *and a one-to-one mapping* ψ *from* $B(x, \epsilon)$ *onto an open set* $D \subset \mathbb{R}^p$ *such that* ψ *and* ψ^{-1} *are twice continuously differentiable,* $\psi(B(x, \epsilon) \cap S) \subset \{(x_1, ..., x_p) \in \mathbb{R}^p : x_p > 0\}$ *and* $\psi(B(x, \epsilon) \cap \partial S) \subset \{(x_1, ..., x_p) \in \mathbb{R}^p : x_p = 0\}$.

Let f_X *denote the probability density function of* X_i *and let* $\mathcal{H}^k$ *denote the* k*-dimensional Hausdorff measure on* $\mathbb{R}^p$.[17]

(f) (Regularity of Deterministic A)

(i) $\mathcal{H}^{p-1}(\partial\Omega^*) < \infty$, *and* $\int_{\partial\Omega^*} f_X(x) d\mathcal{H}^{p-1}(x) > 0$.

(ii) There exists $\delta > 0$ *such that* $A(x) = 0$ *for almost every* $x \in N(\mathcal{X}, \delta) \setminus \Omega^*$, *where* $N(S, \delta) = \{x \in \mathbb{R}^p : \|x - y\| < \delta \text{ for some } y \in S\}$ *for a set* $S \subset \mathbb{R}^p$ *and* $\delta > 0$.

(g) (Conditional Moments and Density near $\partial\Omega^*$) *There exists* $\delta > 0$ *such that*

(i) $E[Y_{1i}|X_i]$, $E[Y_{0i}|X_i]$, $E[D_i(1)|X_i]$, $E[D_i(0)|X_i]$ *and* f_X *are continuously differentiable and have bounded partial derivatives on* $N(\partial\Omega^*, \delta)$;

(ii) $E[Y_{1i}^2|X_i]$, $E[Y_{0i}^2|X_i]$, $E[Y_{1i}D_i(1)|X_i]$ *and* $E[Y_{0i}D_i(0)|X_i]$ *are continuous on* $N(\partial\Omega^*, \delta)$;

(iii) $E[Y_i^4|X_i]$ *is bounded on* $N(\partial\Omega^*, \delta)$.

Assumption 3 is a set of conditions for establishing consistency. Assumption 3 (b) assumes that, conditional on each value of X_i for which APS is nondegenerate, more individuals would change their treatment assignment status from 0 to 1 in response to treatment recommendation

[17]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 $S \in \Sigma$ and $\delta > 0$, let $\mathcal{H}_\delta^k(S) = \inf\{\sum_{j=1}^\infty d(E_j)^k : S \subset \cup_{j=1}^\infty E_j, d(E_j) < \delta, E_j \subset \mathbb{R}^p \text{ for all } j\}$, where $d(E) = \sup\{\|x - y\| : x, y \in E\}$. The k-dimensional Hausdorff measure of S on $\mathbb{R}^p$ is $\mathcal{H}^k(S) = \lim_{\delta \to 0} \mathcal{H}_\delta^k(S)$.

than would change it from 1 to 0.[18] 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 $x \in \mathcal{X}$ with $p^A(x) \in (0,1)$, changing the labels of treatment recommendation makes Assumption 3 (b) hold.

Assumption 3 (c) rules out potential multicollinearity. If the support of $A(X_i)$ contains only one value in $(0,1)$, $p^A(X_i;\delta_n)$ is asymptotically constant and equal to $A(X_i)$ conditional on $p^A(X_i;\delta_n) \in (0,1)$, resulting in the multicollinearity between $p^A(X_i;\delta_n)$ and the constant term. Although dropping the constant term from the 2SLS regression solves this issue, Assumption 3 (c) allows us to only consider the regression with a constant for the purpose of simplifying the presentation. In Appendix C.3, we provide 2SLS estimators that are consistent and asymptotically normal even if we do not know whether Assumption 3 (c) holds.

Assumption 3 (d)–(g) are a set of conditions we require for proving consistency and asymptotic normality of $\hat{\beta}_1$ when A is deterministic and produces only multidimensional RD variation. Assumption 3 (d) says that A produces variation in the treatment recommendation.

Assumption 3 (e) imposes the differentiability of the boundary of $\Omega^* = \{x \in \mathbb{R}^p : A(x) = 1\}$. 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 $\nabla f(x) = (\frac{\partial f(x)}{\partial x_1}, ..., \frac{\partial f(x)}{\partial x_p})' \neq \mathbf{0}$ for all $x \in \mathbb{R}^p$ with $f(x) = 0$. Ω^* takes this form, for example, when the conditional treatment effect $E[Y_i(1) - Y_i(0)|X]$ is predicted by supervised learning based on smooth models such as lasso and ridge regressions, and treatment is recommended 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 Classification And Regression Tree (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 3 (e), the assumptions approximately hold in that Ω^* is arbitrarily well approximated by a set that satisfies the differentiability condition.[19]

Part (i) of Assumption 3 (f) says that the boundary of Ω^* is $(p-1)$ dimensional and that the boundary has nonzero density.[20] Part (ii) puts a weak restriction on the values A takes on outside the support of X_i. It requires that $A(x) = 0$ for almost every $x \notin \Omega^*$ that is outside $\mathcal{X}$ but is in the neighborhood of $\mathcal{X}$. $A(x)$ may take on any value if x is not close to $\mathcal{X}$. These conditions hold in practice. Assumption 3 (g) imposes continuity, continuous differentiability and boundedness on the conditional moments of potential outcomes and the probability density

[18] At the cost of making the presentation more complex, 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 with $p^A(x) \in (0,1)$.

[19] For example, suppose that $p = 2$, $A(x) = 1$ if $x_1 > 0$ and $x_2 > 0$, and $A(x) = 0$ otherwise. In this case, $\Omega^* = \{x \in \mathbb{R}^2 : x_1 > 0, x_2 > 0\}$. Let $\{\Omega_k\}_{k=1}^{\infty}$ be a sequence of subsets of $\mathbb{R}^2$, where $\Omega_k = \{x \in \mathbb{R}^2 : x_2 \geq \frac{1}{kx_1}, x_1 > 0\}$ for each k. Ω_k is twice continuously differentiable for all k, and well approximates Ω^* for a large k in that $d_H(\Omega^*, \Omega_k) \to 0$ as $k \to \infty$, where $d_H(S,T) = \max\{\sup_{x\in S} \inf_{y\in T} \|x-y\|, \sup_{y\in T} \inf_{x\in S} \|x-y\|\}$ is the Hausdorff distance between two sets S and $T \subset \mathbb{R}^p$.

[20] The boundary of Ω^* may fail to be $(p-1)$ dimensional in trivial cases where the Lebesgue measure of Ω^* is zero and hence $A(X_i) = 0$ with probability one. 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, the boundary of Ω^* is the same as Ω^*, and its two-dimensional Hausdorff measure is zero.

near the boundary of Ω^*.

When A is stochastic, asymptotic normality requires additional assumptions. Let

$$C^* = \{x \in \mathbb{R}^p : A \text{ is continuously differentiable at } x\},$$

and let $D^* = \mathbb{R}^p \setminus C^*$ be the set of points at which A is not continuously differentiable.

Assumption 4. *If* $\Pr(A(X_i) \in (0,1)) > 0$*, then the following conditions (a)–(c) hold.*

(a) (Probability of Neighborhood of D^*) $\Pr(X_i \in N(D^*, \delta)) = O(\delta)$.

(b) (Bounded Partial Derivatives of A) *The partial derivatives of A are bounded on C^*.*

(c) (Bounded Conditional Mean) $E[Y_i|X_i]$ *is bounded on $\mathcal{X}$.*

Assumption 4 is required for proving asymptotic normality of $\hat{\beta}_1$ when A is stochastic. To explain the role of Assumption 4 (a), consider a path of covariate points $x_\delta \in N(D^*, \delta) \cap C^*$ indexed by $\delta > 0$. Since A is continuous at x_δ, $p^A(x_\delta) = A(x_\delta)$ (as formally implied by Proposition A.2 in Appendix A.2). However, $p^A(x_\delta; \delta)$ does not necessarily get sufficiently close to $A(x_\delta)$ even as $\delta \to 0$, since x_δ is in the δ-neighborhood of D^* and hence A may discontinuously change within the δ-ball $B(x_\delta, \delta)$. Assumption 4 (a) requires that the probability of X_i being in the δ-neighborhood of D^* shrink to zero at the rate of δ, which makes the points in the neighborhood negligible.

Assumption 4 (a) often holds in practice. If A is continuously differentiable on $\mathcal{X}$, then $D^* \cap \mathcal{X} = \emptyset$, so this condition holds. If, for example, the treatment recommendation is randomly assigned based on a stratified randomized experiment or on the ϵ-Greedy algorithm (see Example A.1 (a) in Appendix A.6), D^* is the boundary at which the recommendation probability changes discontinuously. For any boundary of standard shape, the probability of X_i being in the δ-neighborhood of the boundary vanishes at the rate of δ, and the required condition is satisfied. We provide a sufficient condition for this condition in Appendix A.4. Assumption 4 (b) and (c) are regularity conditions, imposing the boundedness of the partial derivatives of A and of the conditional mean of the outcome.

The following assumption is the key to proving asymptotic normality of the simulation-based estimator $\hat{\beta}_1^s$.

Assumption 5 (The Number of Simulation Draws)**.** $n^{-1/2} S_n \to \infty$*, and* $\Pr(p^A(X_i; \delta_n) \in (0, \gamma \frac{\log n}{S_n}) \cup (1 - \gamma \frac{\log n}{S_n}, 1)) = o(n^{-1/2} \delta_n^{1/2})$ *for some* $\gamma > \frac{1}{2}$.

Assumption 5 says that we need to choose the number of simulation draws S_n so that it grows to infinity faster than $n^{1/2}$, and that the probability that $p^A(X_i; \delta_n)$ lies on the tails $(0, \gamma \frac{\log n}{S_n}) \cup (1 - \gamma \frac{\log n}{S_n}, 1)$ vanishes faster than $n^{-1/2} \delta_n^{1/2}$. This condition makes the bias caused by using $p^s(X_i; \delta_n)$ instead of $p^A(X_i; \delta_n)$ asymptotically negligible. To illustrate how the second part of this assumption restricts the rate at which S_n goes to infinity, consider an example where $\Pr(p^A(X_i; \delta_n) \in (0,1)) = O(\delta_n)$, and $p^A(X_i; \delta_n)$ is approximately uniformly distributed on the tails $(0, \gamma \frac{\log n}{S_n}) \cup (1 - \gamma \frac{\log n}{S_n}, 1)$. In this case, $\Pr(p^A(X_i; \delta_n) \in (0, \gamma \frac{\log n}{S_n}) \cup (1 - \gamma \frac{\log n}{S_n}, 1)) = O(\delta_n \frac{\log n}{S_n})$, and the second part of Assumption 5 requires that S_n grow sufficiently fast so that

$\frac{n^{1/2}\delta_n^{1/2}\log n}{S_n} = o(1)$. One choice of S_n satisfying this is $S_n = \alpha n^{\kappa}\delta_n^{1/2}$ for some $\alpha > 0$ and $\kappa > \frac{1}{2}$, in which case $\frac{n^{1/2}\delta_n^{1/2}\log n}{S_n} = \frac{\log n}{\alpha n^{\kappa-1/2}} = o(1)$.

Under the above conditions, the 2SLS estimators $\hat{\beta}_1$ and $\hat{\beta}_1^s$ are consistent and asymptotically normal estimators of a weighted average treatment effect.

Theorem 1 (Consistency and Asymptotic Normality)**.** *Suppose that Assumptions 1 and 3 hold, 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*

$$\beta_1 \equiv \lim_{\delta \to 0} E[\omega_i(\delta)(Y_i(1) - Y_i(0))],$$

where

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

Suppose, in addition, that Assumptions 4 and 5 hold and that $n\delta_n^2 \to 0$ as $n \to \infty$. Then

$$\hat{\sigma}_n^{-1}(\hat{\beta}_1 - \beta_1) \xrightarrow{d} \mathcal{N}(0,1),$$
$$(\hat{\sigma}_n^s)^{-1}(\hat{\beta}_1^s - \beta_1) \xrightarrow{d} \mathcal{N}(0,1),$$

where we define $\hat{\sigma}_n^{-1}$ and $(\hat{\sigma}_n^s)^{-1}$ as follows. Let

$$\hat{\mathbf{\Sigma}}_n = (\sum_{i=1}^n \mathbf{Z}_{i,n}\mathbf{D}_{i,n}' I_{i,n})^{-1}(\sum_{i=1}^n \hat{\epsilon}_{i,n}^2 \mathbf{Z}_{i,n}\mathbf{Z}_{i,n}' I_{i,n})(\sum_{i=1}^n \mathbf{D}_{i,n}\mathbf{Z}_{i,n}' I_{i,n})^{-1},$$

where

$$\hat{\epsilon}_{i,n} = Y_i - \mathbf{D}_{i,n}'\hat{\beta}.$$

$\hat{\mathbf{\Sigma}}_n$ is the conventional heteroskedasticity-robust estimator for the variance of the 2SLS estimator. $\hat{\sigma}_n^2$ is the second diagonal element of $\hat{\mathbf{\Sigma}}_n$. $(\hat{\sigma}_n^s)^2$ is the analogously-defined estimator for the variance of $\hat{\beta}_1^s$ from the simulation-based regression.

Proof. See Appendix C.3. □

Theorem 1 says that the 2SLS estimators converge to the limit of a weighted average of causal effects for the subpopulation whose fixed-bandwidth APS is nondegenerate ($p^A(X_i;\delta) \in (0,1)$) and who would switch their treatment status in response to the treatment recommendation ($D_i(1) \neq D_i(0)$).[21] The limit $\lim_{\delta\to 0} E[\omega_i(\delta)(Y_i(1) - Y_i(0))]$ always exists under the assumptions of Theorem 1. It also shows that inference based on the conventional 2SLS heteroskedasticity-robust standard errors is asymptotically valid if δ_n goes to zero at an appropriate rate. The convergence rate of $\hat{\beta}_1$ is $O_p(1/\sqrt{n})$ if $\Pr(A(X_i) \in (0,1)) > 0$ and is $O_p(1/\sqrt{n\delta_n})$ if $\Pr(A(X_i) \in (0,1)) = 0$.

Our consistency result requires that δ_n go to zero slower than n^{-1}. The rate condition ensures that, when $\Pr(A(X_i) \in (0,1)) = 0$, we have sufficiently many observations in the δ_n-neighborhood of the boundary of Ω^*. Importantly, the rate condition does not depend on the

[21] In principle, it is possible to estimate other weighted averages and the unweighted average by reweighting different observations appropriately. For example, we can estimate the unweighted average treatment effect by weighting observations by the inverse of fixed-bandwidth APS.

dimension of X_i, unlike other bandwidth-based estimation methods such as kernel methods. This is because we use all the observations in the δ-neighborhood of the boundary, and the number of those observations is of order $n\delta_n$ regardless of the dimension of X_i if the dimension of the boundary is one less than the dimension of X_i, i.e., $(p-1)$.

The asymptotic normality result requires that δ_n go to zero sufficiently quickly so that $n\delta_n^2 \to 0$. When $\Pr(A(X_i) \in (0,1)) > 0$, we need to use a small enough δ_n so that $p^A(X_i;\delta_n)$ converges to $p^A(X_i)$ fast enough and δ_n-neighborhood of D^* is asymptotically small enough. When $\Pr(A(X_i) \in (0,1)) = 0$, the asymptotic normality is based on undersmoothing, which eliminates the asymptotic bias by using the observations sufficiently close to the boundary of Ω^*. In both cases, the bias of our estimator is $O(\delta_n)$. The standard deviation is $O(1/\sqrt{n})$ when $\Pr(A(X_i) \in (0,1)) > 0$ and is $O(1/\sqrt{n\delta_n})$ when $\Pr(A(X_i) \in (0,1)) = 0$. The condition that $n\delta_n^2 \to 0$ ensures that the bias converges to zero faster than the standard deviation in either case.[22]

Whether or not $\Pr(A(X_i) \in (0,1)) = 0$, when we use simulated fixed-bandwidth APS, the consistency result requires that the number of simulation draws S_n go to infinity as n increases. The asymptotic normality result requires a sufficiently fast growth rate of S_n given by Assumption 5 to make the bias caused by using $p^s(X_i;\delta_n)$ negligible.[23]

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 switches from 1 to 0 in response to the treatment recommendation (defiers) 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 2. *Suppose that Assumptions 1 and 3 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). Let $LATE(x) = E[Y_i(1) - Y_i(0) | D_i(1) \neq D_i(0), X_i = x]$ be the local average treatment effect (LATE) conditional on $X_i = x$.

Corollary 3. *Suppose that Assumptions 1 and 3 hold, that $\Pr(D_i(1) \geq D_i(0) | X_i = x) = 1$ for any $x \in \mathcal{X}$ with $p^A(x) \in (0,1)$ (monotonicity), 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*

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

[22] In the special case of the univariate RDD, standard RD local linear estimators are shown to have the same convergence rate under our assumptions (the smoothness of regression functions, in particular).

[23] To sum up, the asymptotic normality result for the simulation-based estimator $\hat{\beta}_1^s$ requires the sequence (δ_n, S_n) to satisfy $n\delta_n \to \infty$, $n\delta_n^2 \to 0$, and Assumption 5. In the preceding example where $\Pr(p^A(X_i;\delta_n) \in (0,1)) = O(\delta_n)$ and $p^A(X_i;\delta_n)$ is approximately uniformly distributed on the tails $(0, \gamma\frac{\log n}{S_n}) \cup (1 - \gamma\frac{\log n}{S_n}, 1)$, one appropriate choice of (δ_n, S_n) that satisfies all conditions is $\delta_n = \alpha_1 n^{-\kappa_1}$ and $S_n = \alpha_2 n^{\kappa_2}$ for some $\alpha_1, \alpha_2 > 0$, $\kappa_1 \in (\frac{1}{2}, 1)$ and $\kappa_2 > \frac{1}{2}$.

where

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

The 2SLS estimators are consistent for the limit of a weighted average of conditional LATEs over all values of X_i with nondegenerate fixed-bandwidth APS $p^A(X_i;\delta_n)$. The weights are proportional to $p^A(X_i;\delta_n)(1-p^A(X_i;\delta_n))$, and to the proportion of compliers, $E[D_i(1)-D_i(0)|X_i]$.

4.3 Intuition and Challenges

The result in Theorem 1 holds whether A is stochastic ($\Pr(A(X_i)\in(0,1))>0$) or deterministic ($\Pr(A(X_i)\in(0,1))=0$). If we consider these two underlying cases separately, the probability limit of the 2SLS estimators has a more specific expression. If $\Pr(A(X_i)\in(0,1))>0$,

$$\operatorname{plim}\hat{\beta}_1 = \operatorname{plim}\hat{\beta}_1^s = \frac{E[A(X_i)(1-A(X_i))(D_i(1)-D_i(0))(Y_i(1)-Y_i(0))]}{E[A(X_i)(1-A(X_i))(D_i(1)-D_i(0))]}. \tag{5}$$

The 2SLS estimators converge to a weighted average of treatment effects for the subpopulation with nondegenerate $A(X_i)$, as shown in the proof of Theorem 1 in Appendix C.3.

To relate this result to existing work, consider the following 2SLS regression with the (standard) propensity score $A(X_i)$ control:

$$D_i = \gamma_0 + \gamma_1 Z_i + \gamma_2 A(X_i) + \nu_i \tag{6}$$

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 A(X_i) + \epsilon_i. \tag{7}$$

Under conditional independence, the 2SLS estimator from this regression converges in probability to the treatment-variance weighted average of treatment effects in (5) (Angrist and Pischke, 2008; Hull, 2018).[24] Not surprisingly, for this selection-on-observables case, our result shows that the 2SLS estimator is consistent for the same treatment effect whether we control for the propensity score, fixed-bandwidth APS, or simulated fixed-bandwidth APS.

Importantly, using fixed-bandwidth APS as a control allows us to consistently estimate a causal effect even if A is deterministic and produces multidimensional regression-discontinuity variation. If $\Pr(A(X_i)\in(0,1))=0$,

$$\operatorname{plim}\hat{\beta}_1 = \operatorname{plim}\hat{\beta}_1^s = \frac{\int_{\partial\Omega^*} E[(D_i(1)-D_i(0))(Y_i(1)-Y_i(0))|X_i=x]f_X(x)d\mathcal{H}^{p-1}(x)}{\int_{\partial\Omega^*} E[D_i(1)-D_i(0)|X_i=x]f_X(x)d\mathcal{H}^{p-1}(x)}. \tag{9}$$

The 2SLS estimators converge to a weighted average of treatment effects for the subpopulation who are on the boundary of the treated region.

[24] Precisely speaking, Angrist and Pischke (2008) consider the OLS 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:

$$Y_i = \alpha_1 Z_i + \sum_{x\in\mathcal{X}} \alpha_{2,x} 1\{X_i = x\} + u_i. \tag{8}$$

By the Frisch-Waugh Theorem, the population coefficient on Z_i from (8) is 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 $\frac{E[A(X_i)(1-A(X_i))(Y_{1i}-Y_{0i})]}{E[A(X_i)(1-A(X_i))}$ under the conditional independence assumption. Their derivation follows even when X_i is continuous and we control the propensity score linearly.

Proving this result requires a technique that may be useful for other problems. Recall that the 2SLS regression uses the observations with $p^A(X_i;\delta_n) \in (0,1)$ (or $p^s(X_i;\delta_n) \in (0,1)$ when we use simulated fixed-bandwidth APS) only. By definition, if $p^A(X_i;\delta) \in (0,1)$, X_i must be in the δ-neighborhood of the boundary of Ω^*. Therefore, to derive the probability limit of $\hat{\beta}_1$, it is necessary to derive the limits of the integrals of relevant variables over the δ-neighborhood (e.g., $\int_{N(\partial\Omega^*,\delta)} E[Y_i|X_i = x]f_X(x)dx$) as δ shrinks to zero. We take an approach drawing on change of variables techniques from differential geometry and geometric measure theory.[25] In this approach, we first use the coarea formula (Lemma B.3 in Appendix B.3) to write the integral of an integrable function g over $N(\partial\Omega^*,\delta)$ in terms of the iterated integral over the levels sets of the signed distance function of Ω^*:

$$\int_{N(\partial\Omega^*,\delta)} g(x)dx = \int_{-\delta}^{\delta} \int_{\{x' \in \mathbb{R}^p : d^s_{\Omega^*}(x') = \lambda\}} g(x) d\mathcal{H}^{p-1}(x) d\lambda, \tag{10}$$

where $d^s_{\Omega^*}$ is the signed distance function of Ω^* (see Appendix B.2 for the definition). The set $\{x' \in \mathbb{R}^p : d^s_{\Omega^*}(x') = \lambda\}$ is a level set of $d^s_{\Omega^*}$, which collects the points in Ω^* when $\lambda > 0$ and the points in $\mathbb{R}^p \setminus \Omega^*$ when $\lambda < 0$ whose distance to the boundary $\partial\Omega^*$ is $|\lambda|$. Figure 2a shows a visual illustration of the level set.

We then use the area formula (Lemma B.4 in Appendix B.3) to write the integral over each level set in terms of the integral over the boundary $\partial\Omega^*$:

$$\int_{\{x' \in \mathbb{R}^p : d^s_{\Omega^*}(x') = \lambda\}} g(x) d\mathcal{H}^{p-1}(x) = \int_{\partial\Omega^*} g(x^* + \lambda\nu_{\Omega^*}(x^*)) J^{\partial\Omega^*}_{p-1}\psi_{\Omega^*}(x^*,\lambda) d\mathcal{H}^{p-1}(x^*), \tag{11}$$

where $\nu_{\Omega^*}(x^*)$ is the inward unit normal vector of $\partial\Omega^*$ at x^* (the unit vector orthogonal to all vectors in the tangent space of $\partial\Omega^*$ at x^* that points toward the inside of Ω^*), and $J^{\partial\Omega^*}_{p-1}\psi_{\Omega^*}(x^*,\lambda)$ is the Jacobian of the transformation $\psi_{\Omega^*}(x^*,\lambda) = x^* + \lambda\nu_{\Omega^*}(x^*)$. Figure 2b illustrates this change of variables formula. Finally, combining (10) and (11) and proceeding with further analysis, we prove in Appendix C.3.3 that when g is continuous,

$$\int_{N(\partial\Omega^*,\delta)} g(x)dx = \delta\left(\int_{\partial\Omega^*} g(x) d\mathcal{H}^{p-1}(x) + o(1)\right).$$

Thus, the integral over the δ-neighborhood of $\partial\Omega^*$ scaled up by δ^{-1} converges to the integral over boundary points with respect to the $(p-1)$-dimensional Hausdorff measure. This result is used to derive the expression of the probability limit of $\hat{\beta}_1$ given by (9).

[25] Our approach using geometric theory shows that $\hat{\beta}_1$ converges to an integral of the conditional treatment effect over boundary points with respect to the Hausdorff measure. In constrast, prior studies on multidimensional RDDs 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). 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 an integral with respect to the Hausdorff measure to any expressions in terms of conditional expectations on the boundary. Arias, Rubio-Ramírez and Waggoner (2018), Bornn, Shephard and Solgi (2019), and Qiao (2021) use similar tools from differential geometry and geometric measure theory, but for different purposes.

5 Machine Learning Simulation

This section assesses the feasibility and performance of our method, by conducting a Monte Carlo experiment motivated by high-dimensional decision making by machine learning. Consider a tech company that applies a machine-learning-based deterministic decision algorithm to a large segment of the population. At the same time, the company conducts a randomized controlled trial (RCT) using the rest of the population. They are interested in estimating treatment effects using data from both segments. Our approach offers a way of exploiting not only the RCT segment but also the deterministic algorithm segment.

We simulate $1,000$ hypothetical samples from the following data-generating process. Each sample $\{(Y_i, X_i, D_i, Z_i)\}_{i=1}^n$ is of size $n = 10,000$. There are 100 covariates ($p = 100$), and $X_i \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$. $Y_i(0)$ is generated as $Y_i(0) = 0.75X_i'\alpha_0 + 0.25\epsilon_{0i}$, where $\alpha_0 \in \mathbb{R}^{100}$, and $\epsilon_{0i} \sim \mathcal{N}(0,1)$. We consider two models for $Y_i(1)$, one in which the treatment effect $Y_i(1) - Y_i(0)$ does not depend on X_i and one in which the treatment effect depends on X_i.

Model A. $Y_i(1) = Y_i(0) + \epsilon_{1i}$, where $\epsilon_{1i} \sim \mathcal{N}(0,1)$.

Model B. $Y_i(1) = Y_i(0) + X_i'\alpha_1$, where $\alpha_1 \in \mathbb{R}^{100}$.

The choice of parameters $\mathbf{\Sigma}$, α_0 and α_1 is explained in Appendix D. $D_i(0)$ and $D_i(1)$ are generated as $D_i(0) = 0$ and $D_i(1) = 1\{Y_i(1) - Y_i(0) > u_i\}$, where $u_i \sim \mathcal{N}(0,1)$.

To generate Z_i, let $q_{0.495}$ and $q_{0.505}$ be the 49.5th and 50.5th (empirical) quantiles of the first covariate X_{i1}. Let $\tau_{pred}(X_i)$ be a real-valued function of X_i, which we regard as a prediction of the effect of recommendation on the outcome for individual i obtained from past data. We construct τ_{pred} by random forests using an independent sample (see Appendix D for the details). Z_i is then generated as

$$Z_i = \begin{cases} Z_i^* \sim \text{Bernoulli}(0.5) & \text{if } X_{i1} \in [q_{0.495}, q_{0.505}] \\ 1 & \text{if } X_{i1} \notin [q_{0.495}, q_{0.505}] \text{ and } \tau_{pred}(X_i) \geq 0 \\ 0 & \text{if } X_{i1} \notin [q_{0.495}, q_{0.505}] \text{ and } \tau_{pred}(X_i) < 0. \end{cases}$$

The first case corresponds to the RCT segment while the latter two cases to the deterministic algorithm segment. The function A is given by

$$A(x) = \begin{cases} 0.5 & \text{if } x_1 \in [q_{0.495}, q_{0.505}] \\ 1 & \text{if } x_1 \notin [q_{0.495}, q_{0.505}] \text{ and } \tau_{pred}(x) \geq 0 \\ 0 & \text{if } x_1 \notin [q_{0.495}, q_{0.505}] \text{ and } \tau_{pred}(x) < 0. \end{cases}$$

Finally, D_i and Y_i are generated as $D_i = Z_iD_i(1) + (1 - Z_i)D_i(0)$ and $Y_i = D_iY_i(1) + (1 - D_i)Y_i(0)$, respectively.

Estimators and Estimands. We use the data $\{(Y_i, X_i, D_i, Z_i)\}_{i=1}^n$ to estimate treatment effect parameters. Our main approach is 2SLS with fixed-bandwidth APS controls in Theorem 1. To compute fixed-bandwidth APS, we use $S = 400$ simulation draws for each observation.

We compare our approach with two naive alternatives. The first alternative is OLS of Y_i on a constant and D_i (i.e., the difference in the sample mean of Y_i between the treated group and

untreated group) using all observations. The second alternative is 2SLS with $A(X_i)$ controls. This method uses the observations with $A(X_i) \in (0,1)$ to run the 2SLS regression of Y_i on a constant, D_i, and $A(X_i)$ using Z_i as an instrument for D_i (see (6) and (7) in Section 4.3) and reports the coefficient on D_i.

We consider four parameters as target estimands: ATE $\equiv E[Y_i(1) - Y_i(0)]$, ATE(RCT) $\equiv E[Y_i(1)-Y_i(0)|X_{i1} \in [q_{0.495}, q_{0.505}]]$, LATE $\equiv E[Y_i(1)-Y_i(0)|D_i(1) \neq D_i(0)]$, and LATE(RCT) $\equiv E[Y_i(1) - Y_i(0)|D_i(1) \neq D_i(0), X_{i1} \in [q_{0.495}, q_{0.505}]]$. In the case where the treatment effect does not depend on X_i (Model A), ATE and LATE are the same as ATE(RCT) and LATE(RCT), respectively. In the case where the treatment effect depends on X_i (Model B), ATE and LATE differ from ATE(RCT) and LATE(RCT), respectively. However, since the RCT segment is a randomly selected subpopulation, the average effect for the RCT segment is close to the unconditional average effect. As a result, ATE is similar to ATE(RCT) and LATE is to LATE(RCT).

For both models, the 2SLS estimator converges in probability to LATE(RCT) (equivalently, the right-hand side of equation (5)) whether we control for fixed-bandwidth APS or $A(X_i)$. However, 2SLS with $A(X_i)$ controls uses only the individuals for the RCT segment while 2SLS with fixed-bandwidth APS controls additionally uses the individuals near the decision boundary of the deterministic algorithm (i.e., the boundary of the region for which $\tau_{pred}(x) \geq 0$). Therefore, 2SLS with fixed-bandwidth APS controls is expected to produce a more precise estimate than 2SLS with $A(X_i)$ controls if the conditional effects for those near the boundary are not far from the target estimand.

Results. Table 1 reports the bias, standard deviation (SD), and root mean squared error (RMSE) of each estimator. Panels A and B present the results for the cases where the conditional effects are homogeneous and heterogeneous, respectively. Note first that OLS with no controls is significantly biased, showing the importance of correcting for omitted variable bias. 2SLS with fixed-bandwidth APS achieves this goal, as demonstrated by its smaller biases across all possible treatment effect models, target parameters, and values of the bandwidth δ. 2SLS with fixed-bandwidth APS controls shows a consistent pattern; as the bandwidth δ grows, the bias increases while the variance declines. For several values of δ, 2SLS with fixed-bandwidth APS controls outperforms 2SLS with $A(X_i)$ controls in terms of the RMSE. This finding implies that exploiting individuals near the multidimensional decision boundary of the deterministic algorithm can lead to better performance than using only the individuals in the RCT segment.

We also evaluate our inference procedure based on Theorem 1. Table 1 reports the coverage probabilities of the 95% confidence intervals for LATE(RCT) constructed from the 2SLS estimates and their heteroskedasticity-robust standard errors. The confidence intervals offer nearly correct coverage when δ is small, which supports the implication of Theorem 1 that the inference procedure is valid when we use a sufficiently small δ. Overall, Table 1 shows that our estimator works well in this high-dimensional setting and performs better than alternative estimators.

6 Empirical Policy Application

6.1 Hospital Relief Funding during the COVID-19 Pandemic

Here we provide our real-world empirical application. As part of the 3-phase Coronavirus Aid, Relief, and Economic Security (CARES) Act, the US government has distributed tens of billions of dollars of relief funding to hospitals since April 2020. This funding intended to help health care providers hit hardest by the COVID-19 outbreak and at a high risk of closing. The bill specified that providers may (but are not required to) use the funds for COVID-19-related expenses, such as construction of temporary structures, leasing of properties, purchasing medical supplies and equipment (including personal protective equipment and testing supplies), increased workforce utilization and training, establishing emergency operation centers, retrofitting facilities and managing the surge in capacity, among others.

We are interested in whether this funding had a causal impact on hospital operation and activities in dealing with COVID-19 patients. We focus on an initial portion of this funding ($10 billion), which was allocated to hospitals that qualified as "safety net hospitals" according to a specific eligibility criterion. This eligibility criterion intends to direct funding towards hospitals that "*disproportionately provide care to the most vulnerable, and operate on thin margins.*" Specifically, an acute care hospital was deemed eligible for funding if the following conditions hold:

- Medicare Disproportionate Patient Percentage (DPP) of 20.2% or greater. DPP is equal to the sum of the percentage of Medicare inpatient days attributable to patients eligible for both Medicare Part A and Supplemental Security Income (SSI), and the percentage of total inpatient days attributable to patients eligible for Medicaid but not Medicare Part A.[26]

- Annual Uncompensated Care (UCC) of at least $\$25,000$ per bed. UCC is a measure of hospital care provided for which no payment was received from the patient or insurer. It is the sum of a hospital's bad debt and the financial assistance it provides.[27]

- Profit Margin (Net income/(Net patient revenue + Total other income)) of 3.0% or less.

Hospitals that do not qualify on any of the three dimensions are funding ineligible. Figure 3 visualizes how the three dimensions determine funding eligibility. As the bottom two-dimensional planes show, eligibility discontinuously changes as hospitals cross the eligibility boundary in the space of the three characteristics. This setting is a three-dimensional RDD, falling under our framework.

The final funding amount is calculated as follows. Each eligible hospital is assigned an individual facility score, which is calculated as the product of DPP and the number of beds in that hospital. This facility score determines the share of funding allocated to the hospital, out of the total $10 billion. The share received by each hospital is determined by the ratio of the

[26] Source: `https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/dsh`

[27] Source: `https://www.aha.org/fact-sheets/2020-01-06-fact-sheet-uncompensated-hospital-care-cost`

hospital's facility score to the sum of facility scores across all eligible hospitals. The amount of funding that can be received is bounded below at $5 million and capped above at $50 million. Figure 4 shows the distribution of funding amounts received by eligible hospitals. A majority of eligible hospitals receive the minimum amount of $5 million. A small mass of hospitals receive amounts close to the maximum of $50 million. We replicate the funding eligibility status as well as the amount of funding received, by using publicly available data from the Healthcare Cost Report Information System (HCRIS) for the 2018 financial year.[28]

To obtain outcome measures of interest, we use the publicly available COVID-19 Reported Patient Impact and Hospital Capacity by Facility dataset. This provides facility-level data on hospital utilization aggregated on a weekly basis, from July 31st 2020 onwards.[29] Summary statistics about hospital outcomes and characteristics are documented in Table 2. Eligible hospitals have higher fractions of inpatient and ICU beds occupied by COVID-19 patients. Eligible hospitals also have a higher disproportionate patient percentage, higher uncompensated care per bed, lower profit margins, more employees and beds, and shorter lengths of inpatient stay. These patterns are consistent with the funding's goal of helping struggling hospitals.

6.2 Covariate Balance Estimates

We first evaluate the balancing property of fixed-bandwidth APS conditioning using fixed-bandwidth-APS-controlled differences in covariate means for hospitals who are and are not deemed eligible for funding. Specifically, we run the following OLS regression of hospital-level characteristics on the eligibility status using observations with $p^s(X_i;\delta_n) \in (0,1)$:

$$W_i = \gamma_0 + \gamma_1 Z_i + \gamma_2 p^s(X_i;\delta_n) + \eta_i,$$

where W_i is one of the predetermined characteristics of the hospital, Z_i is a funding eligibility dummy, X_i is a vector of the three input variables (DPP, UCC, and profit margin) that determine the funding eligibility, and $p^s(X_i;\delta_n)$ is the simulated fixed-bandwidth APS. We compute fixed-bandwidth APS using $S = 10,000$ simulation draws.[30] The estimated coefficient on Z_i is the fixed-bandwidth-APS-controlled difference in the mean of the covariate between eligible and ineligible hospitals. For comparison, we also run the OLS regression of hospital characteristics on the eligibility status with no controls using the whole sample.

Table 3 reports the covariate balance estimates. Column 1 shows that, without controlling for fixed-bandwidth APS, eligible hospitals are significantly different from ineligible hospitals. We find that all the relevant hospital eligibility characteristics are strongly associated with eligibility. Once we control for fixed-bandwidth APS with small enough bandwidth δ, eligible and ineligible hospitals have similar financial and utilization characteristics, as reported in columns 2–6 of

[28] We use the methodology detailed in the CARES ACT website to project funding based on 2018 financial year cost reports. We use the RAND cleaned version of the dataset which can be accessed at `https://www.hospitaldatasets.org/`

[29] Source: `https://healthdata.gov/Hospital/COVID-19-Reported-Patient-Impact-and-Hospital-Capa/anag-cw7u`

[30] Figure A.2 in Appendix E.4 reports fixed-bandwidth APS for several hospitals with varying numbers of simulation draws. We find that $S = 10,000$ is sufficient for well stabilizing fixed-bandwidth APS simulation.

Table 3. These estimates are consistent with our theoretical results, establishing the empirical relevance of fixed-bandwidth APS controls.

6.3 2SLS Estimates

The balancing performance of fixed-bandwidth APS motivates us to estimate causal effects of funding by 2SLS using funding eligibility as an instrument for the amount of funding received. We study the effect of funding on relevant hospital outcomes, such as the number of inpatient beds occupied by adult COVID patients between July 31st 2020 and August 6th 2020. We run the following 2SLS regression on four different hospital-level outcome variables, using hospitals with $p^s(X_i;\delta) \in (0,1)$:

$$D_i = \gamma_0 + \gamma_1 Z_i + \gamma_2 p^s(X_i;\delta) + v_i$$
$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 p^s(X_i;\delta) + \epsilon_i,$$

where Y_i is a hospital-level outcome and D_i is the amount of relief funding received.[31] We also run the OLS and 2SLS regressions with no controls, as well as OLS regression controlling for the three eligibility determinants (disproportionate patient percentage, uncompensated care per bed and profit margin).[32] These alternative regressions are computed using the sample of all hospitals, as benchmark estimators.

The first stage effects of funding eligibility on funding amount (in millions), shown in columns 3–10 of Table 4, suggest that funding eligibility boosts the amount of funding significantly. For example, in column 3 of Table 4, we can see that funding eligibility increases funding by approximately 15 million dollars on average.

OLS estimates of funding effects, reported as the benchmark in column 1 of Table 4, indicate that funding is associated with a higher number of adult inpatient beds and higher number of staffed ICU beds utilized by patients who have lab-confirmed or suspected COVID. The estimates indicate that a million dollar increase in funding is associated with 5.58 more adult inpatient beds occupied by patients with lab-confirmed or suspected COVID. The corresponding increase

[31] This specification uses a continuous treatment, unlike our theoretical framework with a binary treatment. We obtain similar results when the treatment is a binary transformation of the amount of relief funding received (e.g., a dummy indicating whether the amount exceeds a certain value). Results are available upon request.

[32] Precisely speaking, we run the following specification of each alternative estimator for each hospital-level outcome variable Y_i. For the OLS regression without any controls, we estimate:

$$Y_i = \beta_0 + \beta_1 D_i + \epsilon_i.$$

For the 2SLS regression without any controls, we run:

$$D_i = \gamma_0 + \gamma_1 Z_i + v_i$$
$$Y_i = \beta_0 + \beta_1 D_i + \epsilon_i.$$

For the OLS regression controlling for disproportionate patient percentage, uncompensated care per bed and profit margin, we estimate:

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_{i1} + \beta_3 X_{i2} + \beta_4 X_{i3} + \epsilon_i,$$

where X_{i1} is disproportionate patient percentage, X_{i2} is uncompensated care per bed, and X_{i3} is profit margin.

in total adult inpatient beds occupied by those who have lab-confirmed COVID is 4.53 and the increase in staffed ICU beds occupied by those who have lab-confirmed or suspected COVID is 1.67. The estimated increase in staffed ICU beds occupied by lab-confirmed COVID patients is 1.51. These uncontrolled OLS estimates show a similar picture as the descriptive statistics in Table 2. Naive 2SLS estimates with no controls and OLS with covariate controls produce similar significantly positive associations of funding with outcomes.

However, the OLS or uncontrolled 2SLS estimates turn out to be an artifact of selection bias. In contrast with these naive estimates, our preferred 2SLS estimates with fixed-bandwidth APS controls show a different picture (columns 4–10). The gains in the number of inpatient beds and staffed ICU beds occupied by suspected and lab-confirmed COVID patients become much smaller and lose significance across all bandwidth specifications. These results suggest that fixed-bandwidth APS reveals important selection bias in the estimated effects of funding. Once we control for fixed-bandwidth APS to eliminate the bias, funding has little to no effect on the hospital utilization level by COVID-19 patients.[33]

The above analysis looks at the immediate effects of relief funding. However, the effects of relief funding might kick in after a time lag, given that expansion in capacity and staff takes time. To investigate the relevance of this concern, we finally measure the evolving effects of relief funding. We estimate our main 2SLS specification on the 7-day average of each hospital outcome for each week from July 31st, 2020 to April 2nd, 2021. We plot the results in Figure 5. The estimated dynamic effects are similar to the initial null effects in Table 4, even several months after the distribution of relief funding. This dynamic analysis suggests that funding has no substantial effect even in the long run.

We further extend this analysis by estimating the heterogeneous effects of funding for different types of hospitals. Figure 6 plots the resulting estimates by repeating the same dynamic analysis as in Figure 5, but for different groups of hospitals defined by hospital size and ownership type. Overall, hospitals with different characteristics sometimes face different trends of funding effects, but none of the differences is statistically significant at the 5% level. We do not find any strong evidence of heterogeneity in the funding effects at any point in time.

Having said that, there is some suggestive indication of potential heterogeneity. In Panel 6a, for example, the estimated funding effect spiked among the hospitals in the lowest quartile of revenue from December 2020 to February 2021. This trend may suggest that the funding was able to alleviate the financial burden faced by struggling hospitals in this strata and allowed them to take on new patients during the winter surge.

There is also a sizable dip in the funding effect of for-profit hospitals around the same period. This could be due to regional differences in the distribution of hospital ownership. Nonprofits and government-managed hospitals tend to be in rural areas, which both received more funding and experienced a worse surge during the winter. On the other hand, the for-profits that received funding tend to be in urban areas and experienced a less extreme winter wave.

The overall insignificance of the estimates suggests that funding by the CARES Act had

[33]The 2SLS estimates in Table 4 are unlikely to be compromised by differential attrition. Estimates reported in Table A.1 in Appendix E.4 show little difference in outcome availability rates between eligible and ineligible hospitals once we control for fixed-bandwidth APS.

largely no effect on hospital utilization trends during the pandemic. This finding is consistent with policy and media arguments that CARES Act funding was not well targeted toward needy providers. Unlike the previous arguments and descriptive analyses, the analysis here provides causal evidence supporting the concern.

7 Other Examples

Here we give real-world examples of other algorithms and discuss the applicability of our framework.

Example 1 (Bandit Algorithms)**.** 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 using bandit algorithms. Our method is applicable to many popular bandit algorithms. For simplicity, assume a perfect-compliance scenario where the company perfectly controls the treatment assignment ($D_i = Z_i$). 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 $z \in \{0, 1\}$. Let μ_z and σ_z^2 denote the estimated functions. The algorithms 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 sampled potential outcome:

$$Z_i^{TS} \equiv \underset{z\in\{0,1\}}{\arg\max}\, y(z), \quad A^{TS}(X_i) = E[\underset{z\in\{0,1\}}{\arg\max}\, y(z)|X_i],$$

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 has observed (Precup, 2000; Li *et al.*, 2010; Narita, Yasui and Yata, 2019; Saito, Aihara, Matsutani and Narita, 2021). Suppose that the functions μ_0, μ_1, σ_0^2 and σ_1^2 are continuous. The function A and APS have an analytical expression:

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

where Φ is the cumulative distribution function of a standard normal distribution. This APS is nondegenerate, meaning that the data from the algorithm allow for causal-effect identification.

(b) (Upper Confidence Bound, UCB) Unlike the above stochastic algorithm, the UCB algorithm (Li *et al.*, 2010) 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:

$$Z_i^{UCB} \equiv \underset{z=0,1}{\arg\max}\{\mu_z(X_i) + \alpha\sigma_z(X_i)\}, \quad A^{UCB}(x) = \underset{z=0,1}{\arg\max}\{\mu_z(x) + \alpha\sigma_z(x)\},$$

where α is chosen so that $|\mu_z(x) - E[Y_i(z)|X_i = x]| \leq \alpha\sigma_z(x)$ at least with some probability, for example, 0.95, for every x. Suppose that the function $g = \mu_1 - \mu_0 + \alpha(\sigma_1 - \sigma_0)$ is continuous on $\mathcal{X}$ and is continuously differentiable in a neighborhood of x with $\nabla g(x) \neq \mathbf{0}$ for any $x \in \mathcal{X}$ such that $g(x) = 0$. APS for this case is given by

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

This means that the UCB algorithm produces potentially complicated quasi-experimental variation along the boundary in the covariate space where the algorithm's treatment recommendation changes from one to the other. It is possible to identify and estimate causal effects across the boundary.

Example 2 (Unsupervised Learning)**.** Customer segmentation is a core marketing practice that divides a company's customers into groups based on their characteristics and behavior 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 such as K-means clustering or 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:

$$Z_i^{CL} \equiv 1\{X_i \in \cup_{k\in T} S_k\}, \quad A^{CL}(x) = 1\{x \in \cup_{k\in T} S_k\},$$

where $T \subset \{1, .., K\}$ is the set of the indices of the target groups.

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. 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 $C(x) \equiv \arg\min_{k\in\{1,...,K\}} \|x - c_k\|$. If $C(x)$ is a singleton, x belongs to the unique group in $C(x)$. If $C(x)$ contains more than one indices, the group to which x belongs is arbitrarily determined. APS for this case is given by

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

and $p^{CL}(x) \in (0, 1)$ if $|C(x)| \geq 3$ and $x \in \partial(\cup_{k\in T} S_k)$, where $|C(x)|$ is the number of elements in $C(x)$.[34] Thus, it is possible to identify causal effects across the boundary $\partial(\cup_{k\in T} S_k)$.

[34] 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 APS 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 APS is determined by the angles at which the boundaries intersect.

Example 3 (Supervised Learning)**.** Millions of times each year, judges make jail-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 jail-or-release decisions. Using our notation, assume that a criminal risk algorithm recommends jailing ($Z_i = 1$) or releasing ($Z_i = 0$) for 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 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) construct a version of $r(X_i)$ using gradient boosted decision trees. The algorithm then uses the risk score to make the final recommendation:

$$Z_i^{SL} \equiv 1\{r(X_i) > c\}, \quad A^{SL}(x) = 1\{r(x) > c\},$$

where $c \in \mathbb{R}$ is a constant threshold that is set ex ante.[35] 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).

A widely-used approach to identifying and estimating treatment effects in these settings is to use the score $r(X_i)$ as a continuous univariate running variable and apply a univariate RDD method (Cowgill, 2018). However, whether $r(X_i)$ is continuously distributed or not depends on how the function r is constructed. For example, suppose that r is constructed by a tree-based algorithm and is the following simple regression tree with three terminal nodes:

$$r(x) = \begin{cases} r_1 & \text{if } x_1 \leq 0 \\ r_2 & \text{if } x_1 > 0, x_2 \leq 0 \\ r_3 & \text{if } x_1 > 0, x_2 > 0, \end{cases}$$

where $r_1 < r_2 < c < r_3$. In this case, the score $r(X_i)$ is a discrete variable, and hence it may not be suitable to apply a standard univariate RDD method.

Our approach is applicable to this case as long as at least one of the original multi-dimensional covariates X_i are continuously distributed. Since $A^{SL}(x) = 1\{r(x) > c\} = 1\{x_1 > 0, x_2 > 0\}$, APS for this case is given by

$$p^{SL}(x) = \begin{cases} 0 & \text{if } x_1 < 0 \text{ or } x_2 < 0 \\ 0.25 & \text{if } x_1 = x_2 = 0 \\ 0.5 & \text{if } (x_1 = 0, x_2 > 0) \text{ or } (x_1 > 0, x_2 = 0) \\ 1 & \text{if } x_1 > 0, x_2 > 0. \end{cases}$$

It is therefore possible to identify causal effects across the boundary $\{x \in \mathcal{X} : (x_1 = 0, x_2 \geq 0) \text{ or } (x_1 > 0, x_2 = 0)\}$.

[35] The algorithm sometimes discretizes the original 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: $Z_i^{SL} \equiv 1\{d(r(X_i)) > c\}$.

Example 4 (Policy Eligibility Rules)**.** Medicaid and other welfare policies often decide who are eligible based on algorithmic rules, as studied by Currie and Gruber (1996) and Brown, Kowalski and Lurie (2020).[36] 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 $A^{Medicaid}$ 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 3.

Example 5 (Mechanism Design: Matching and Auction)**.** Centralized economic mechanisms such as matching and auction are also suitable examples, as summarized below (Abdulkadiroğlu *et al.*, 2017, Forthcoming; Abdulkadiroğlu, 2013; Kawai *et al.*, 2020; Narita, 2020, 2021):

	Matching (e.g., School Choice)	Auction
i	Student	Bidder
X_i	Preference/Priority/Tie-breaker	Bid
Z_i	Whether student i is assigned treatment school	Whether bidder i wins the good
D_i	Whether student i attends treatment school	Same as Z_i
Y_i	Student i's future test score	Bidder i's future economic performance

In mechanism design 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.[37] 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 $A : \mathbb{R}^p \to [0,1]$ such that $A(x) \equiv M(x; F_X)$ for all $x \in \mathbb{R}^p$. Our framework is applicable to this large-market interactive setting.

The above discussions can be summarized as follows.

[36] These papers estimate the effect of Medicaid eligibility by exploiting variation in the eligibility rule across states and over time (simulated instrumental variable method). In contrast, our method exploits local variation in the eligibility status across different individuals given a fixed eligibility rule.

[37] The approach proposed by Borusyak and Hull (2020) is applicable to finite-sample settings if the treatment recommendation probability, which may depend on all individuals' characteristics, is nondegenerate for multiple individuals.

Corollary 4. *In all the above examples, there exists* $x \in \text{int}(\mathcal{X})$ *such that* $p^A(x) \in (0,1)$. *Therefore, a causal effect is identified under Assumptions 1 and 2.*

8 Conclusion

As algorithmic decisions become the new norm, the world becomes a mountain of natural experiments and instruments. We develop a general method to use these algorithm-produced instruments to identify and estimate causal treatment effects. Our analysis of the CARES Act hospital relief funding uses the proposed method to find that relief funding has little effect on COVID-19-related hospital activities. OLS or uncontrolled 2SLS estimates, by contrast, show considerably larger and more significant effects. The large estimates appear to be an artifact of selection bias; relief funding just went to hospitals with more COVID-19 patients, without helping hospitals accommodate additional patients.

Our analysis clarifies a few implications for policy and management practices around algorithmic decision-making. It is important to record the implementation of algorithms in a replicable, simulatable way, including what input variables X_i are used to make algorithmic recommendation Z_i. Another key lesson is the importance of recording an algorithm's recommendation Z_i even if they are superseded by a human decision D_i. These data retention efforts would go a long way to exploit the full potential of algorithms as natural experiments.

An important topic for future research is estimation details, such as data-driven bandwidth selection. This work needs to extend Imbens and Kalyanaraman (2012) and Calonico, Cattaneo and Titiunik (2014)'s bandwidth selection methods in the univariate RDD to our setting.[38] Inference on treatment effects in our framework relies on conventional large sample reasoning. It seems natural to additionally consider permutation or randomization inference as in Imbens and Rosenbaum (2005). It will also be challenging but interesting to develop finite-sample optimal estimation and inference strategies such as those recently introduced by Armstrong and Kolesár (2018, 2020) and Imbens and Wager (2019). Finally, we look forward to empirical applications of our method in a variety of business, policy, and scientific domains.

[38] For univariate RDDs, Imbens and Kalyanaraman (2012) and Calonico *et al.* (2014) estimate the bandwidth that minimizes the asymptotic mean squared error (AMSE). It is not straightforward to estimate the AMSE-optimal bandwidth in our setting with many running variables and complex IV assignment, since it requires nonparametric estimation of functions on the multidimensional covariate space such as conditional mean functions, their derivatives, the curvature of the RDD boundary, etc.

References

Abdulkadiroğlu, A. (2013). Instrumental Variable Estimation in School Choice. *Private Communication*.

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. (Forthcoming). Breaking Ties: Regression Discontinuity Design Meets Market Design. *Econometrica*.

Adelino, M., Lewellen, K. and McCartney, W. B. (2021). Hospital Financial Health and Clinical Choices: Evidence from the Financial Crisis. *Management Science*.

—, — and Sundaram, A. (2015). Investment Decisions of Nonprofit Firms: Evidence from Hospitals. *Journal of Finance*, **70** (4), 1583–1628.

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.

Arias, J. E., Rubio-Ramírez, J. F. and Waggoner, D. F. (2018). Inference Based on Structural Vector Autoregressions Identified with Sign and Zero Restrictions: Theory and Applications. *Econometrica*, **86** (2), 685–720.

Armstrong, T. B. and Kolesár, M. (2018). Optimal Inference in a Class of Regression Models. *Econometrica*, **86** (2), 655–683.

— and Kolesár, M. (2020). Finite-Sample Optimal Estimation and Inference on Average Treatment Effects Under Unconfoundedness. *Econometrica*.

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

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. (2019). Discretizing Unobserved Heterogeneity. *University of Chicago, Becker Friedman Institute for Economics Working Paper No. 2019-16*, unpublished Manuscript, University of Chicago.

Bornn, L., Shephard, N. and Solgi, R. (2019). Moment conditions and bayesian nonparametrics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **81** (1), 5–43.

Borusyak, K. and Hull, P. (2020). Non-Random Exposure to Exogenous Shocks: Theory and Applications. *NBER Working Paper No. 27845*, unpublished Manuscript, University of Chicago.

Brown, D., Kowalski, A. E. and Lurie, I. Z. (2020). Long-Term Impacts of Childhood Medicaid Expansions on Outcomes in Adulthood. *The Review of Economic Studies*, **87** (2), 729–821.

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*.

Calonico, S., Cattaneo, M. D. and Titiunik, R. (2014). Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs. *Econometrica*, **82** (6), 2295–2326.

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*, unpublished Manuscript, University of Chicago.

Cowgill, B. (2018). The Impact of Algorithms on Judicial Discretion: Evidence from Regression Discontinuities. Unpublished Manuscript, Columbia Business School.

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

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

Dranove, D., Garthwaite, C. and Ody, C. (2017). How Do Nonprofits Respond to Negative Wealth Shocks? The Impact of the 2008 Stock Market Collapse on Hospitals. *RAND Journal of Economics*, **48** (2), 485–525.

Duggan, M. G. (2000). Hospital Ownership and Public Medical Spending. *Quarterly Journal of Economics*, **115** (4), 1343–1373.

Eckles, D., Ignatiadis, N., Wager, S. and Wu, H. (2020). Noise-induced Randomization in Regression Discontinuity Designs. *arXiv preprint arXiv:2004.09458*.

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.

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. (2007). Regression Discontinuity Design with Covariates. *IZA Discussion Paper No. 3024*, unpublished Manuscript, University of St. Gallen.

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

Gulshan, V. *et al.* (2016). Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *Journal of the American Medical Association*, **316** (22), 2402–2410.

Hahn, J., Todd, P. and van der Klaauw, W. (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. Unpublished Manuscript, MIT.

Imbens, G. and Kalyanaraman, K. (2012). Optimal Bandwidth Choice for the Regression Discontinuity Estimator. *The Review of Economic Studies*, **79** (3), 933–959.

— and Wager, S. (2019). Optimized Regression Discontinuity Designs. *Review of Economics and Statistics*, **101** (2), 264–278.

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

— and Rosenbaum, P. R. (2005). Robust, Accurate Confidence Intervals with a Weak Instrument: Quarter of Birth and Education. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **168** (1), 109–126.

Kakani, P., Chandra, A., Mullainathan, S. and Obermeyer, Z. (2020). Allocation of COVID-19 Relief Funding to Disproportionately Black Counties. *Journal of the American Medical Association (JAMA)*, **324** (10), 1000–1003.

Kawai, K., Nakabayashi, J., Ortner, J. and Chassang, S. (2020). Using bid rotation and incumbency to detect collusion: A regression discontinuity approach.

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

Khullar, D., Bond, A. M. and Schpero, W. L. (2020). COVID-19 and the Financial Health of US Hospitals. *Journal of the American Medical Association (JAMA)*, **323** (21), 2127–2128.

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.

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.

Mullainathan, S. and Spiess, J. (2017). Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives*, **31** (2), 87–106.

Narita, Y. (2020). A Theory of Quasi-Experimental Evaluation of School Quality. *Management Science*.

— (2021). Incorporating ethics and welfare into randomized experiments. *Proceedings of the National Academy of Sciences*, **118** (1).

—, 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. *Proceedings of the Seventeenth International Conference on Machine Learning*, pp. 759–766.

Qiao, W. (2021). Nonparametric Estimation of Surface Integrals on Level Sets. *Bernoulli*, **27** (1), 155–191.

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.

Saito, Y., Aihara, S., Matsutani, M. and Narita, Y. (2021). Open bandit dataset and pipeline: Towards realistic and reproducible off-policy evaluation. Unpublished Manuscript, Tokyo Institute of Technology.

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.

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*.

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

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

Figure 1: Example of the Approximate Propensity Score

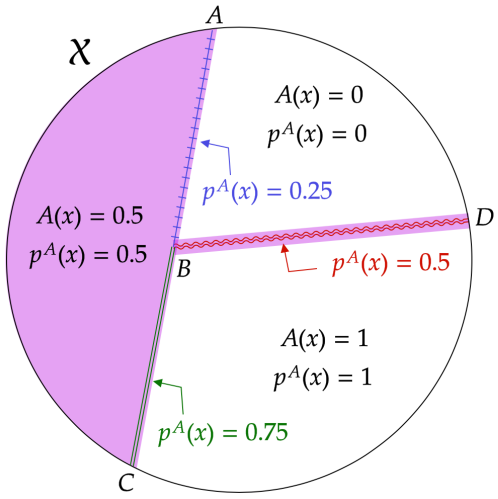
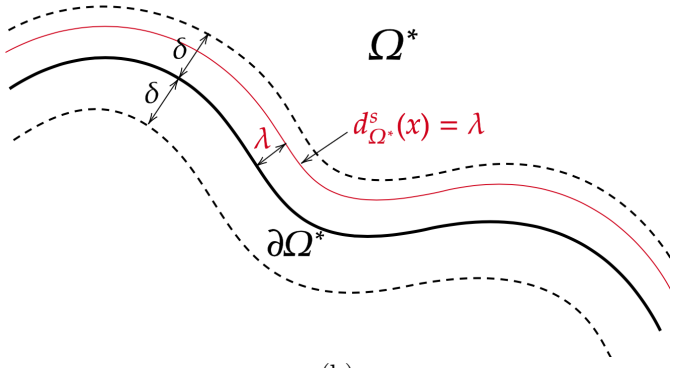


Figure 2: Illustration of the Change of Variables Techniques

(a)



(b)

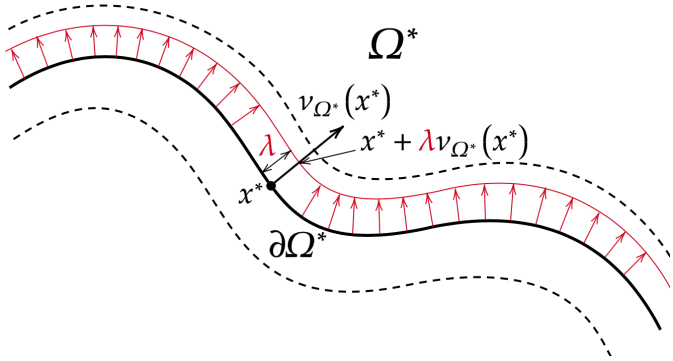
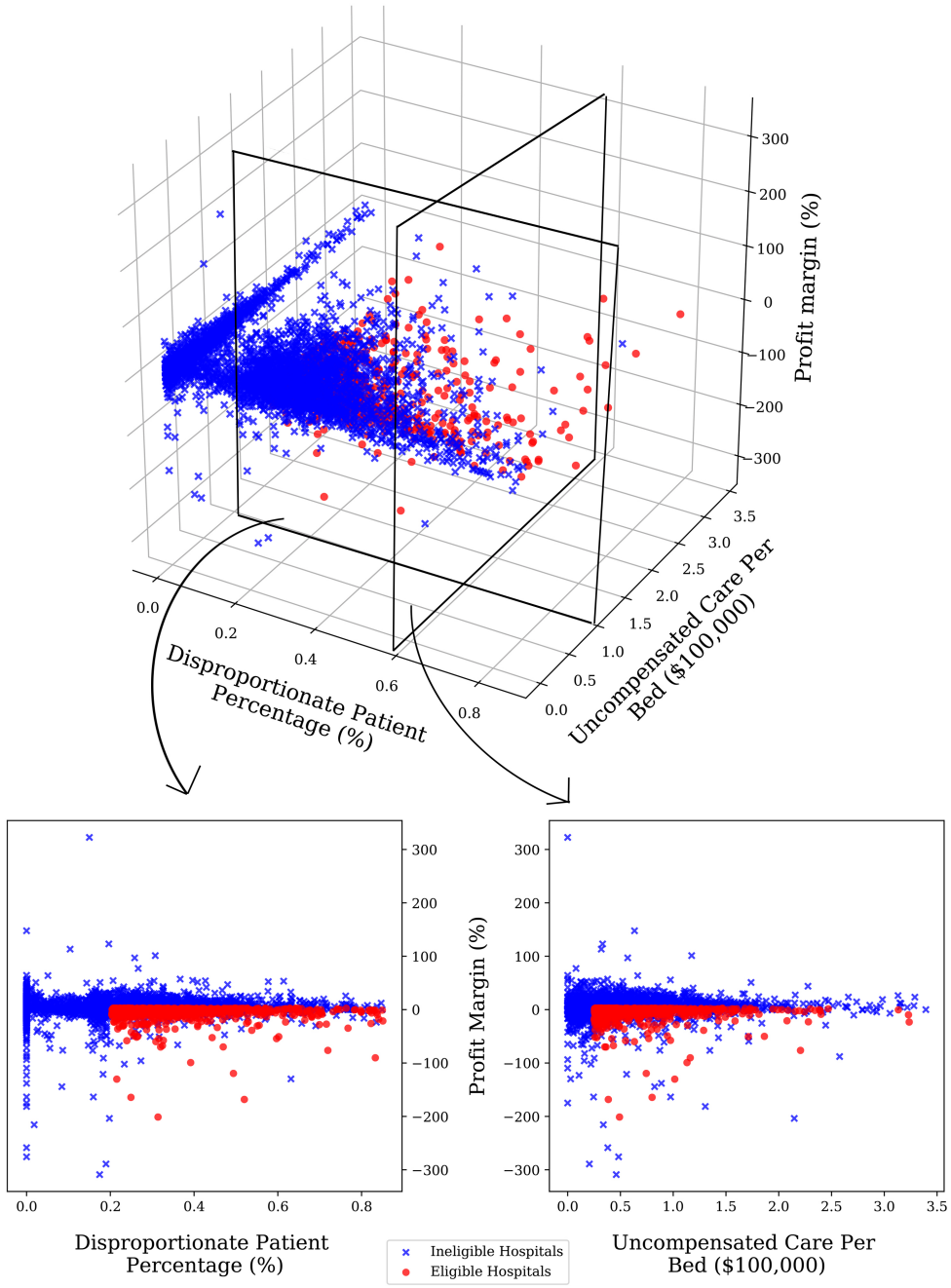


Table 1: Bias, SD, and RMSE of Estimators and Coverage of 95% Confidence Intervals

	Our Method: 2SLS with Approximate Propensity Score Controls						2SLS with $A(X_i)$ Controls	OLS with No Controls
	$\delta = 0.01$	$\delta = 0.05$	$\delta = 0.1$	$\delta = 0.25$	$\delta = 0.5$	$\delta = 1$		
Panel A: Homogeneous Conditional Effects (Model A)								
Estimand: ATE = ATE(RCT) = 0								
Bias	.603	.634	.644	.659	.684	.740	.572	.754
SD	.304	.205	.157	.110	.078	.061	.372	.024
RMSE	.675	.667	.663	.668	.689	.842	.683	.754
Estimand: LATE = LATE(RCT) = 0.564								
Bias	.039	.070	.080	.095	.120	.176	.008	.190
SD	.304	.205	.157	.110	.078	.061	.372	.024
RMSE	.306	.217	.176	.145	.143	.186	.372	.191
Coverage	94.8%	92.8%	92.9%	84.6%	69.6%	18.6%	—	—
Avg N	235	727	1275	2567	3995	5561	100	10000
Panel B: Heterogeneous Conditional Effects (Model B)								
Estimand: ATE = ATE(RCT) = 0								
Bias	.568	.587	.589	.604	.636	.709	.545	1.192
SD	.331	.222	.170	.118	.083	.063	.399	.025
RMSE	.657	.628	.613	.615	.642	.712	.676	1.193
Estimand: LATE = 0.564								
Bias	.004	.023	.025	.040	.072	.145	−.019	.628
SD	.331	.222	.170	.118	.083	.063	.399	.025
RMSE	.331	.223	.172	.125	.110	.158	.399	.629
Estimand: LATE(RCT) = 0.559								
Bias	.009	.028	.030	.045	.077	.150	−.014	.633
SD	.331	.222	.170	.118	.083	.063	.399	.025
RMSE	.331	.224	.173	.127	.114	.163	.399	.634
Coverage	95.9%	94.8%	95.0%	93.2%	87.1%	37.4%	—	—
Avg N	235	723	1274	2567	3993	5561	100	10000

Notes: This table shows the bias, the standard deviation (SD) and the root mean squared error (RMSE) of 2SLS with Approximate Propensity Score controls, 2SLS with $A(X_i)$ controls, and OLS with no controls. These statistics are computed with the estimand set to ATE, ATE(RCT), LATE, or LATE(RCT). The row "Coverage" in each panel shows the probabilities that the 95% confidence intervals of the form $[\hat{\beta}_1^s - 1.96\hat{\sigma}_n^s, \hat{\beta}_1^s + 1.96\hat{\sigma}_n^s]$ contains LATE(RCT), where $\hat{\beta}_1^s$ is the 2SLS estimate with Approximate Propensity Score controls and $\hat{\sigma}_n^s$ is its heteroskedasticity-robust standard error. We use $1,000$ replications of a size $10,000$ simulated sample to compute these statistics. We use several possible values of δ to compute the Approximate Propensity Score. All Approximate Propensity Scores are computed by averaging 400 simulation draws of $A(X_i)$. Panel A reports the results under the model in which the treatment effect does not depend on X_i (Model A). Panel B reports the results under the model in which the treatment effect depends on X_i (Model B). The bottom row "Avg N" in each panel shows the average number of observations used for estimation (i.e., the average number of observations for which the Approximate Propensity Score or $A(X_i)$ is strictly between 0 and 1).

Figure 3: Three-dimensional Regression Discontinuity in Hospital Funding Eligibility



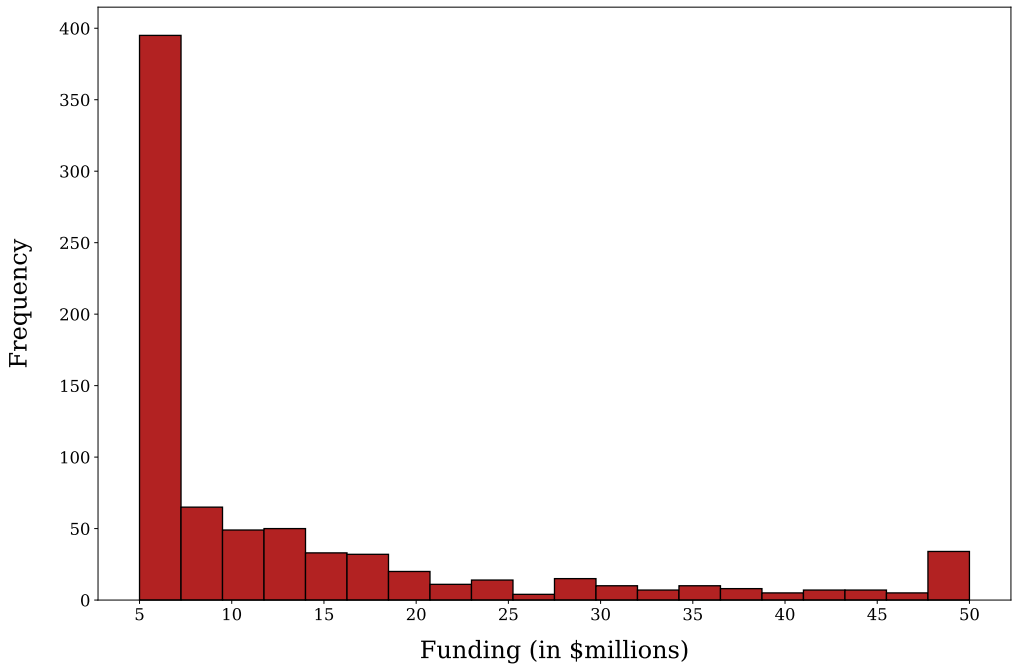
Notes: The top figure visualizes the three dimensions that determine funding eligibility. The bottom figures show the data points plotted along 2 out of 3 dimensions. The bottom left panel plots disproportionate patient percentage against profit margin, while the bottom right panel plots uncompensated care per bed against profit margin. We remove hospitals above the 99th percentile of disproportionate patient percentage and uncompensated care per bed, for visibility purposes.

Table 2: Hospital Characteristics and Outcomes

	All	Ineligible Hospitals	Eligible Hospitals	Hospitals w/ APS ∈ (0,1)
Panel A: Outcome Variable Means				
# Confirmed/Suspected Covid Patients	105.59	98.41	136.61	125.19
# Confirmed Covid Patients	80.1	73.86	107.83	86.78
# Confirmed/Suspected Covid Patients in ICU	31.37	28.92	42.1	36.84
# Confirmed Covid Patients in ICU	26.62	24.41	36.56	31.41
Observations	4,008	3,293	715	429
Panel B: Hospital Characteristics Means				
Beds	143.66	134.6	188.35	206.47
Interns and residents (full-time equivalents) per bed	.06	.05	.11	.09
Adult and pediatric hospital beds	120.26	113.29	154.66	170.49
Ownership: Proprietary (for-profit)	.19	.2	.18	.15
Ownership: Governmental	.22	.22	.23	.16
Ownership: Voluntary (non-profit)	.58	.58	.59	.68
Inpatient length of stay	9.21	10.14	4.66	4.38
Employees on payroll (full-time equivalents)	973.9	897.31	1351.57	1525.06
Disproportionate patient percentage	.21	.18	.38	.36
Uncompensated care per bed	59,850	56,556.03	76,096.31	45,996.48
Profit margin	.02	.04	-.07	-.03
Observations	4,633	3,852	781	485

Notes: This table reports averages of outcome variables and hospital characteristics by safety net eligibility. A safety net hospital is defined as any acute care hospital with disproportionate patient percentage of 20.2% or greater, annual uncompensated care of at least $25,000 per bed and profit margin of 3.0% or less. Panel A reports the outcome variable means. Outcome variable estimates are 7 day sums for the week spanning July 31st 2020 to August 6th 2020. Confirmed or Suspected COVID patients refer to the sum of patients in inpatient beds with lab-confirmed/suspected COVID-19. Confirmed COVID patients refer to the sum of patients in inpatient beds with lab-confirmed COVID-19, including those with both lab-confirmed COVID-19 and influenza. Inpatient bed totals also include observation beds. Similarly, Confirmed/ Suspected COVID patients in ICU refer to the sum of patients in ICU beds with lab-confirmed or suspected COVID-19. Confirmed COVID patients in ICU refers to the sum of patients in ICU beds with lab-confirmed COVID-19, including those with both lab-confirmed COVID-19 and influenza. Panel B reports the means for hospital characteristics for the financial year 2018. Column 1 shows the means for All Hospitals. Columns 2 and 3 show the means for hospitals that are ineligible and eligible to receive safety net funding respectively. Column 4 shows the means for the hospitals with non-degenerate Approximate Propensity Score with bandwidth $\delta = 0.05$. Approximate Propensity Score is computed by averaging 10,000 simulation draws.

Figure 4: Funding Distribution for Eligible Hospitals



Notes: The figure shows the distribution of funding amounts for eligible hospitals. Each eligible hospital is assigned an individual facility score, which is the product of Disproportionate Patient Percentage and number of beds in the hospital. The share of $10 billion received by an eligible hospital is determined by the ratio of the individual facility score of that hospital to the sum of facility scores across all eligible hospitals. The amount of funding that can be received by an eligible hospital is calculated as the product of this ratio and $10 billion, and is bounded below at $5 million and bounded above at $50 million.

Table 3: Covariate Balance Regressions

	Mean (Ineligible) (1)	No Controls (2)	Our Method: OLS with Approximate Propensity Score Controls						
			$\delta = 0.01$ (3)	$\delta = 0.025$ (4)	$\delta = 0.05$ (5)	$\delta = 0.075$ (6)	$\delta = 0.1$ (7)	$\delta = 0.25$ (8)	$\delta = 0.5$ (9)
Panel A: Determinants of Funding Eligibility									
Profit margin	.04	-0.11***	-0.03	-0.01	0.02	0.01	0.02	0.05**	0.06***
		(0.01)	(0.06)	(0.04)	(0.03)	(0.03)	(0.02)	(0.01)	(0.01)
		N=4633	N=90	N=239	N=484	N=670	N=874	N=1726	N=2368
Uncompensated care per bed	56556.02	19,540.28***	4,943.70	10,234.70	-4,182.07	-9,506.62	-10,959.35	-8,009.43	-6,033.20
		(3,827.22)	(12,150.44)	(10,151.35)	(8,666.22)	(7,671.12)	(7,017.85)	(4,538.28)	(3,675.48)
		N=4633	N=90	N=239	N=484	N=670	N=874	N=1726	N=2368
Disproportionate patient percentage	.18	0.21***	-0.09	-0.09	-0.09	-0.08	-0.08	-0.06*	-0.07***
		(0.01)	(0.09)	(0.07)	(0.07)	(0.06)	(0.05)	(0.03)	(0.01)
		N=4633	N=90	N=239	N=484	N=670	N=874	N=1726	N=2368
Panel B: Other Hospital Characteristics									
Full time employees	897.32	454.26***	2,645.18	344.35	62.07	101.29	20.46	206.57	110.85
		(69.23)	(1,635.80)	(1,021.21)	(663.92)	(511.76)	(419.46)	(218.52)	(143.16)
		N=4626	N=90	N=238	N=483	N=669	N=873	N=1723	N=2365
Medicare net revenue (in millions)	20.04	18.36***	34.67	-8.81	-6.39	-1.27	1.55	4.19	-0.42
		(2.39)	(29.56)	(18.59)	(14.25)	(12.03)	(10.83)	(6.64)	(4.68)
		N=4511	N=89	N=238	N=482	N=666	N=870	N=1684	N=2323
Occupancy	.44	0.07***	0.19*	0.07	-0.00	0.01	0.01	0.03	0.04**
		(0.01)	(0.09)	(0.06)	(0.04)	(0.04)	(0.03)	(0.02)	(0.01)
		N=4624	N=90	N=239	N=484	N=670	N=874	N=1726	N=2368
Operating margin	.02	-0.11***	-0.03	0.00	0.03	0.02	0.03	0.06***	0.07***
		(0.01)	(0.06)	(0.05)	(0.03)	(0.03)	(0.03)	(0.02)	(0.01)
		N=4541	N=89	N=238	N=476	N=660	N=863	N=1676	N=2314
Beds	134.6	53.75***	190.00	33.57	4.05	7.37	8.64	16.50	8.30
		(7.05)	(105.37)	(67.12)	(47.58)	(39.07)	(33.42)	(20.05)	(14.46)
		N=4633	N=90	N=239	N=484	N=670	N=874	N=1726	N=2368
Costs per discharge (in thousands)	66.28	-49.95**	3.88	3.29*	1.59	-6.40	-0.62	6.23	6.23
		(17.93)	(2.18)	(1.50)	(1.22)	(8.15)	(2.34)	(4.82)	(4.85)
		N=3539	N=90	N=239	N=484	N=670	N=874	N=1726	N=2368
p-value joint significance		0	.73	.476	.864	.724	.269	0	0

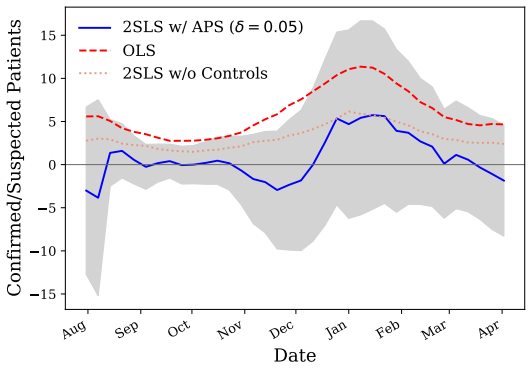
Notes: This table shows the results of the covariate balance regressions at the hospital level. The dependent variables for these regressions are drawn from the Healthcare Cost Report Information System for the financial year 2018. Disproportionate patient percentage, profit margin and uncompensated care per bed are used to determine the hospital's safety net funding eligibility. Other dependent variables shown indicate the financial health and utilization of the hospitals. In column 1, we regress the dependent variables on the safety net eligibility of the hospital with no controls. In columns 2–8, we regress the dependent variables on funding eligibility controlling for the Approximate Propensity Score with different values of bandwidth δ. All Approximate Propensity Scores are computed by averaging 10,000 simulation draws. Column 9 shows the mean of dependent variables for hospitals that are ineligible to receive safety net funding. Robust standard errors are reported in the parenthesis and number of observations are reported separately for each regression. The last row reports the p-value of the joint significance test.

Table 4: Estimated Effects of Funding on Hospital Utilization

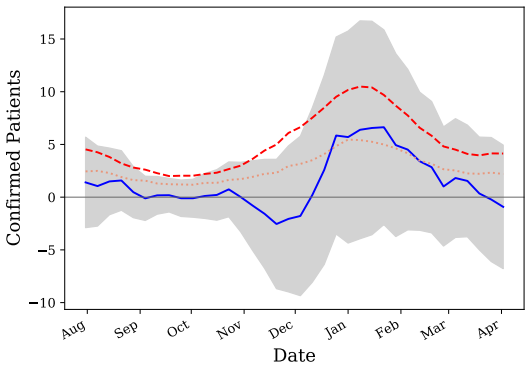
	OLS with No Controls	OLS with Covariate Controls	2SLS with No Controls	Our Method: 2SLS with Approximate Propensity Score Controls						
				$\delta =$ 0.01	$\delta =$ 0.025	$\delta =$ 0.05	$\delta =$ 0.075	$\delta =$ 0.1	$\delta =$ 0.25	$\delta =$ 0.5
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
# Confirmed/Suspected COVID Patients										
First stage			13.78***	15.11*	13.34***	14.28***	14.19***	13.89***	13.96***	13.06***
(in millions)			(0.49)	(5.83)	(3.54)	(2.27)	(1.87)	(1.61)	(1.03)	(0.74)
$1mm of funding	5.58***	3.25***	2.77***	-1.03	-1.86	-3.10	-4.08	-2.91	0.15	-0.31
	(0.68)	(0.89)	(0.58)	(5.64)	(5.40)	(4.99)	(4.57)	(3.58)	(1.59)	(1.21)
Observations	3532	3532	3532	73	195	392	547	719	1389	1947
# Confirmed COVID Patients										
First stage			13.90***	16.55**	14.37***	15.05***	14.81***	14.42***	14.10***	13.19***
(in millions)			(0.50)	(6.11)	(3.66)	(2.33)	(1.91)	(1.64)	(1.04)	(0.75)
$1mm of funding	4.53***	2.50**	2.44***	0.05	-2.14	1.42	0.13	-0.03	-0.09	-0.63
	(0.63)	(0.79)	(0.50)	(4.33)	(3.97)	(2.17)	(1.97)	(1.74)	(1.12)	(0.96)
Observations	3558	3558	3558	70	191	385	539	709	1366	1923
# Confirmed/Suspected COVID Patients in ICU										
First stage			13.88***	14.67*	13.42***	15.75***	15.29***	14.74***	14.31***	13.18***
(in millions)			(0.51)	(5.59)	(3.49)	(2.32)	(1.93)	(1.67)	(1.06)	(0.76)
$1mm of funding	1.67***	0.91**	0.95***	0.93	0.71	0.36	-0.05	0.16	-0.03	-0.32
	(0.21)	(0.28)	(0.18)	(1.47)	(1.27)	(0.74)	(0.70)	(0.60)	(0.40)	(0.36)
Observations	3445	3445	3445	72	186	374	520	678	1314	1846
# Confirmed COVID Patients in ICU										
First stage			13.89***	15.80*	13.79***	15.78***	15.53***	15.08***	14.43***	13.40***
(in millions)			(0.50)	(6.15)	(3.73)	(2.41)	(2.02)	(1.73)	(1.09)	(0.77)
$1mm of funding	1.51***	0.82**	0.88***	0.50	-0.11	0.18	0.04	0.12	-0.13	-0.35
	(0.21)	(0.27)	(0.17)	(1.54)	(1.37)	(0.70)	(0.64)	(0.56)	(0.39)	(0.34)
Observations	3503	3503	3503	67	181	370	514	671	1321	1868

Notes: In this table we regress relevant outcomes at the hospital level on safety net funding. Column 1 presents the results of OLS regression of the outcome variables on safety net funding without any controls. Column 2 presents the results of OLS regression of the outcome variables on safety net funding controlling for disproportionate patient percentage, uncompensated care per bed and profit margin. In columns 3–10, we instrument safety net funding with eligibility to receive this funding and present the results of 2SLS regressions. In columns 3–10, the first stage shows the effect of being deemed eligible on the amount of relief funding received by hospitals, in millions of dollars. Column 3 shows the results of a 2SLS regression with no controls. In columns 4–10, we run this regression controlling for the Approximate Propensity Score with different values of bandwidth δ on the sample with nondegenerate Approximate Propensity Scores. All Approximate Propensity Scores are computed by averaging 10,000 simulation draws. The outcome variables are the 7 day totals for the week spanning July 31st, 2020 to August 6th, 2020. Confirmed or Suspected COVID patients refer to the sum of patients in inpatient beds with lab-confirmed/suspected COVID-19. Confirmed COVID patients refer to the sum of patients in inpatient beds with lab-confirmed COVID-19, including those with both lab-confirmed COVID-19 and influenza. Inpatient bed totals also include observation beds. Similarly, Confirmed/Suspected COVID patients in ICU refer to the sum of patients in ICU beds with lab-confirmed or suspected COVID-19. Confirmed COVID patients in ICU refers to the sum of patients in ICU beds with lab-confirmed COVID-19, including those with both lab-confirmed COVID-19 and influenza. Robust standard errors are reported in parentheses.

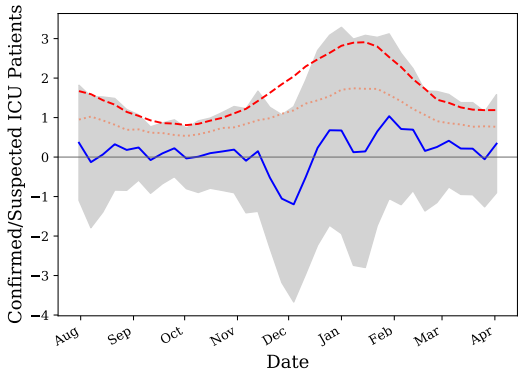
Figure 5: Dynamic Effects of Funding on Weekly Hospital Outcomes



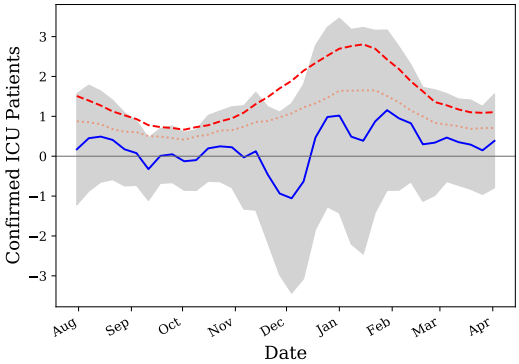
(a) # Confirmed/Suspected COVID Patients



(b) # Confirmed COVID Patients



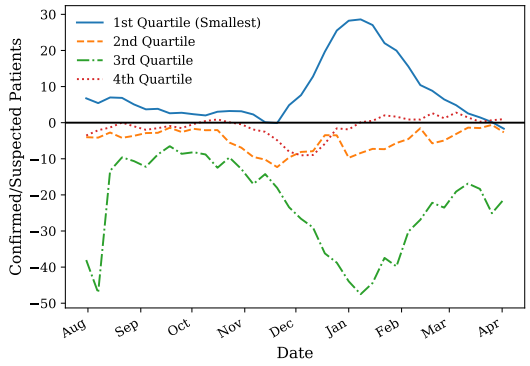
(c) # Confirmed/Suspected COVID Patients in ICU



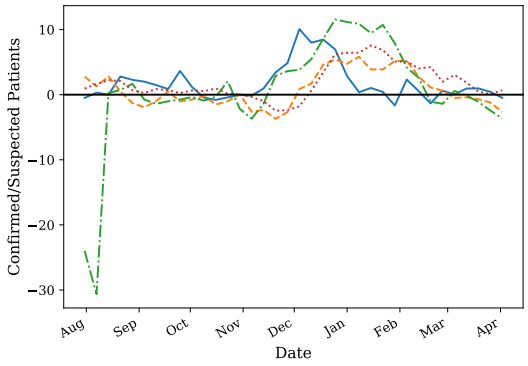
(d) # Confirmed COVID Patients in ICU

Notes: The figure shows the results of estimating our main 2SLS specification about the effect of $1mm of relief funding on weekly hospital outcomes from 07/31/2020 to 04/02/2021. The outcomes record the 7-day sum of the number of hospitalized patients with the specified condition. We compute the Approximate Propensity Score with $S = 10,000$ and $\delta = 0.05$. The estimates from the uncontrolled OLS, uncontrolled 2SLS, and 2SLS with the Approximate Propensity Score controls are plotted on the y-axis. Standard error ribbons are given in grey.

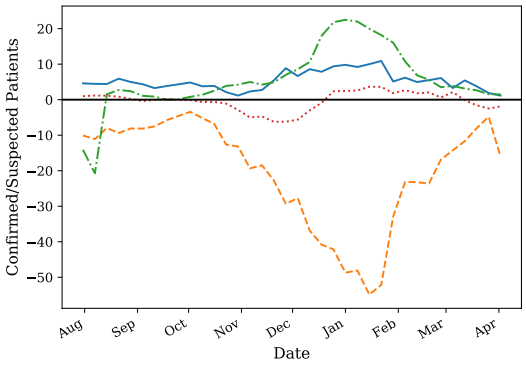
Figure 6: Dynamic Heterogeneous Effects of Hospital Funding by Hospital Characteristics



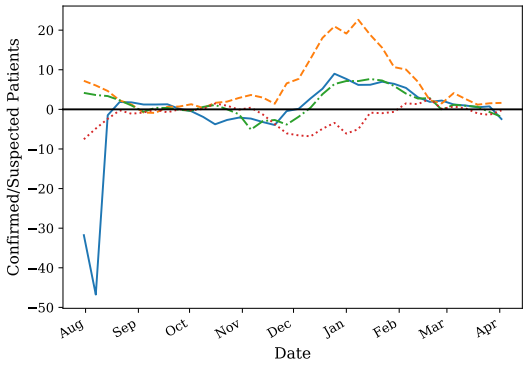
(a) Net Income



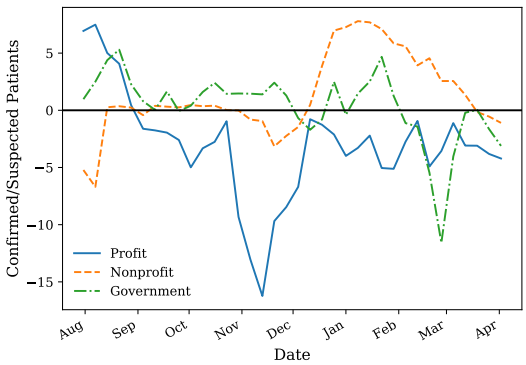
(b) Total Full-Time Employees



(c) Total Beds



(d) Inpatient Length of Stay



(e) Ownership Type

Notes: The figure shows the results of estimating our main 2SLS specification of the effect of $1mm of relief funding on weekly confirmed/suspected Covid-19 patients from 07/31/2020 to 04/12/2021, where the sample is stratified by quartiles of different hospital characteristics, or ownership type. There are no significant estimates at the 5% level. We estimate APS with $S = 10,000$ and $\delta = 0.05$.

A Extensions and Discussions

A.1 Related Literature: Details

In this section, we discuss the related methodological literature on the multidimensional RDD in detail. Imbens and Wager (2019) propose the finite-sample-minimax linear estimator of the form $\sum_{i=1}^{n} \gamma_i Y_i$ and uniform confidence intervals for treatment effects in the multidimensional RDD. One version of their approach constructs a linear estimator by choosing the weight $(\gamma_i)_{i=1}^{n}$ greedily to make the inference as precise as possible. Although their estimator is favorable in terms of precision, it is not obvious what estimand the estimator estimates, without assuming a constant treatment effect. The other version of Imbens and Wager (2019)'s approach and some other existing approaches (Zajonc, 2012; Keele and Titiunik, 2015) consider nonparametric estimation of the conditional average treatment effect $E[Y_i(1) - Y_i(0)|X_i = x]$ for a specified boundary point x. The estimand has a clear interpretation, but "when curvature is nonnegligible, equation (6) can effectively make use of only data near the specified focal point c, thus resulting in relatively long confidence intervals" (Imbens and Wager, 2019, p. 268), where equation (6) defines their estimator.

To obtain more precise estimates while keeping interpretability, several papers studying a two-dimensional RDD, including Zajonc (2012) and Keele and Titiunik (2015), propose to estimate an integral of conditional average treatment effects over the boundary. Their approach first nonparametrically estimates $E[Y_i(1) - Y_i(0)|X_i = x]$ and the density of X_i for a large number of points x in the boundary and then computes the weighted average of the estimated conditional average treatment effects with the weight set to the estimated density.

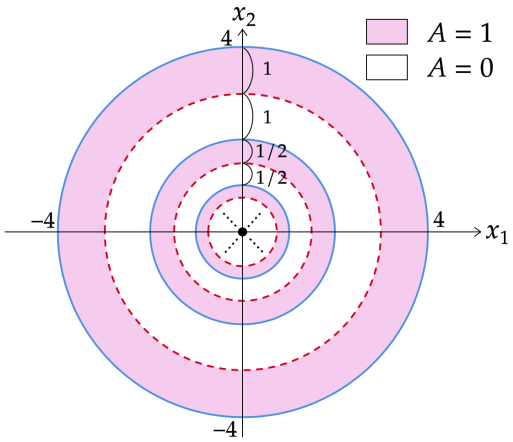
The above approach is difficult to implement, however, when X_i is high dimensional or the decision algorithm is a complex, black box function of X_i, for the following reasons. First, it is computationally demanding to estimate $E[Y_i(1) - Y_i(0)|X_i = x]$ for numerous points in the boundary such that the weighted average well approximates the integral of $E[Y_i(1) - Y_i(0)|X_i = x]$ over the boundary. Second, identifying boundary points from a general decision algorithm itself is hard unless it has a known analytical form. By contrast, we develop an estimator that uses observations near all the boundary points without tracing out the boundary or knowing its analytical form, thus alleviating the limitations of existing estimators.

A.2 Existence of the Approximate Propensity Score

Proposition 1 assumes that APS exists, but is it fair to assume so? In general, APS may fail to exist. Figure A.1 shows such an example. In this example, X_i is two dimensional, and

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

Figure A.1: Example of Algorithm A for which Approximate Propensity Score Does Not Exist



It is shown that

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

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

Nevertheless, APS exists for almost every x, as shown in the following proposition.

Proposition A.1. *$p^A(x)$ exists and is equal to $A(x)$ for almost every $x \in \mathcal{X}$ (with respect to the Lebesgue measure).*

Proof. See Appendix C.6. □

Does APS exist at a specific point x? What is the value of APS at x if it is not equal to $A(x)$? We show that APS exists and is of a particular form for most covariate points and typical algorithms. For each $x \in \mathcal{X}$ and each $q \in \text{Supp}(A(X_i))$, define

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

$\mathcal{U}_{x,q}$ is the set of vectors in $B(\mathbf{0}, 1)$ such that the value of A approaches q as we approach x from the direction of the vector. With this notation, we obtain a sufficient condition for the existence of APS at a point x.

Proposition A.2. *Take any $x \in \mathcal{X}$. If there exists a countable set $Q \subset \text{Supp}(A(X_i))$ such that $\mathcal{L}^p(\cup_{q \in Q} \mathcal{U}_{x,q}) = \mathcal{L}^p(B(\mathbf{0}, 1))$ and $\mathcal{U}_{x,q}$ is $\mathcal{L}^p$-measurable for all $q \in Q$, then $p^A(x)$ exists and is given by*

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

Proof. See Appendix C.4. □

If almost every point in $B(\mathbf{0},1)$ is contained by one of countably many $\mathcal{U}_{x,q}$'s, therefore, APS 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 APS exists in practically important cases.

Corollary A.1.

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

Proof. See Appendix C.5. □

A.3 Discrete Covariates

In this section, we provide the definition of APS and identification and asymptotic normality 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 $\mathcal{X}_{c,0}(x_d) = \{x_c \in \mathcal{X}_c(x_d) : A(x_d, x_c) = 0\}$ and $\mathcal{X}_{c,1}(x_d) = \{x_c \in \mathcal{X}_c(x_d) : A(x_d, x_c) = 1\}$.

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

$$p^A(x;\delta) \equiv \frac{\int_{B(x_c,\delta)} A(x_d, x_c^*) dx_c^*}{\int_{B(x_c,\delta)} dx_c^*},$$
$$p^A(x) \equiv \lim_{\delta \to 0} p^A(x;\delta),$$

where $B(x_c, \delta) = \{x_c^* \in \mathbb{R}^{p_c} : \|x_c - x_c^*\| \leq \delta\}$ is the δ-ball around $x_c \in \mathbb{R}^{p_c}$. In other words, we take the average of the $A(x_d, x_c^*)$ values when x_c^* is uniformly distributed on $B(x_c, \delta)$ holding x_d fixed, and let $\delta \to 0$. Below, we assume that Assumptions 1, 2, 3 and 4 hold conditional on X_{di}.

Assumption A.1 (Almost Everywhere Continuity of A)**.**

(a) For every $x_d \in \mathcal{X}_d$, $A(x_d, \cdot)$ 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 $k = 0, 1$.

A.3.1 Identification

Assumption A.2 (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 $p^A(x_d, x_c) \in (0, 1)$ and $A(x_d, x_c) \in \{0, 1\}$.*

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

Proposition A.3. *Under Assumptions A.1 and A.2:*

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

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

Proof. See Appendix C.7. □

A.3.2 Estimation

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

$$C^*(x_d) = \{x_c \in \mathbb{R}^{p_c} : A(x_d, \cdot) \text{ is continuously differentiable at } x_c\},$$

and let $D^*(x_d) = \mathbb{R}^{p_c} \setminus C^*(x_d)$.

Assumption A.3.

(a) (Finite Moments) $E[Y_i^4] < \infty$.

(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 $p^A(x) \in (0, 1)$.*

(c) (Nonzero Conditional Variance) *If* $\Pr(A(X_i) \in (0, 1)) > 0$, *then* $\text{Var}(A(X_i)|A(X_i) \in (0, 1)) > 0$.

If $\Pr(A(X_i) \in (0, 1)) = 0$, *then the following conditions (d)–(g) hold.*

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

(e) (C^2 Boundary of $\Omega^*(x_d)$) *For each* $x_d \in \mathcal{X}_d^*$, *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* $m \in \{1, ..., M\}$.

(f) (Regularity of Deterministic A)

(i) For each $x_d \in \mathcal{X}_d^*$, $\mathcal{H}^{p_c-1}(\partial\Omega^*(x_d)) < \infty$, *and* $\int_{\partial\Omega^*(x_d)} f_{X_c|X_d}(x_c|x_d) d\mathcal{H}^{p_c-1}(x_c) > 0$.

(ii) There exists $\delta > 0$ *such that* $A(x_d, x_c) = 0$ *for almost every* $x_c \in N(\mathcal{X}_c(x_d), \delta) \backslash \Omega^*(x_d)$.

(g) (Conditional Means and Density near $\partial\Omega^*(x_d)$) *For each* $x_d \in \mathcal{X}_d^*$, *there exists* $\delta > 0$ *such that*

(i) $E[Y_{1i}|X_i = (x_d, \cdot)]$, $E[Y_{0i}|X_i = (x_d, \cdot)]$, $E[D_i(1)|X_i = (x_d, \cdot)]$, $E[D_i(0)|X_i = (x_d, \cdot)]$ *and* $f_{X_c|X_d}(\cdot|x_d)$ *are continuously differentiable and have bounded partial derivatives on* $N(\partial\Omega^*(x_d), \delta)$;

(ii) $E[Y_{1i}^2|X_i = (x_d, \cdot)]$, $E[Y_{0i}^2|X_i = (x_d, \cdot)]$, $E[Y_{1i}D_i(1)|X_i = (x_d, \cdot)]$ *and* $E[Y_{0i}D_i(0)|X_i = (x_d, \cdot)]$ *are continuous on* $N(\partial\Omega^*(x_d), \delta)$;

(iii) $E[Y_i^4|X_i = (x_d, \cdot)]$ *is bounded on* $N(\partial\Omega^*(x_d), \delta)$.

Assumption A.4. *If* $\Pr(A(X_i) \in (0, 1)) > 0$, *then the following conditions (a)–(c) hold.*

(a) (Probability of Neighborhood of $D^*(x_d)$) *For each* $x_d \in \mathcal{X}_d^*$, $\Pr(X_i \in N(D^*(x_d), \delta)) = O(\delta)$.

(b) (Bounded Partial Derivatives of A) *For each* $x_d \in \mathcal{X}_d^*$, *the partial derivatives of* $A(x_d, \cdot)$ *are bounded on* $C^*(x_d)$.

(c) (Bounded Conditional Mean) *For each* $x_d \in \mathcal{X}_d^*$, $E[Y_i|X_i = (x_d, \cdot)]$ *is bounded on* $\mathcal{X}_c(x_d)$.

Theorem A.1. *Suppose that Assumptions A.1 and A.3 hold, 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*

$$\beta_1 \equiv \lim_{\delta \to 0} E[\omega_i(\delta)(Y_i(1) - Y_i(0))],$$

where

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

Suppose, in addition, that Assumptions A.4 and 5 hold and that $n\delta_n^2 \to 0$ *as* $n \to \infty$. *Then*

$$\hat{\sigma}_n^{-1}(\hat{\beta}_1 - \beta_1) \xrightarrow{d} \mathcal{N}(0, 1),$$
$$(\hat{\sigma}_n^s)^{-1}(\hat{\beta}_1^s - \beta_1) \xrightarrow{d} \mathcal{N}(0, 1).$$

Proof. See Appendix C.8. □

As in the case in which all covariates are continuous, the probability limit of the 2SLS estimators has more specific expressions depending on whether $\Pr(A(X_i) \in (0,1)) > 0$ or not. If $\Pr(A(X_i) \in (0,1)) > 0$,

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

If $\Pr(A(X_i) \in (0,1)) = 0$,

$$\begin{aligned}
&\operatorname{plim} \hat{\beta}_1 \\
=\ &\operatorname{plim} \hat{\beta}_1^s \\
=\ &\frac{\sum_{x_d \in \mathcal{X}_d^*} \Pr(X_{di} = x_d) \int_{\partial\Omega^*(x_d)} E[(D_i(1)-D_i(0))(Y_i(1)-Y_i(0))|X_i = x] f_{X_c|X_d}(x_c|x_d) d\mathcal{H}^{p_c-1}(x_c)}{\sum_{x_d \in \mathcal{X}_d^*} \Pr(X_{di} = x_d) \int_{\partial\Omega^*(x_d)} E[D_i(1)-D_i(0)|X_i = x] f_{X_c|X_d}(x_c|x_d) d\mathcal{H}^{p_c-1}(x_c)}.
\end{aligned}$$

A.4 A Sufficient Condition for Assumption 4 (a)

We provide a sufficient condition for Assumption 4 (a).

Assumption A.5.

(a) (Twice Continuous Differentiability of D^*) *There exist* $C_1^*, ..., C_M^* \subset \mathbb{R}^p$ *such that*

- *(i)* $\partial(\tilde{C}^*) = D^*$, *where* $\tilde{C}^* \equiv \cup_{m=1}^M C_m^*$;
- *(ii)* $\operatorname{dist}(C_m^*, C_{m'}^*) > 0$ *for any* $m, m' \in \{1, ..., M\}$ *such that* $m \neq m'$;
- *(iii)* C_m^* *is nonempty, bounded, open, connected and twice continuously differentiable for each* $m \in \{1, ..., M\}$.

(b) (Regularity of D^*) $\mathcal{H}^{p-1}(D^*) < \infty$.

(c) (Bounded Density near D^*) *There exists* $\delta > 0$ *such that* f_X *is bounded on* $N(D^*, \delta)$.

The key condition is the twice continuous differentiability of D^*. This condition holds if, for example, the ϵ-Greedy algorithm described in Example A.1 (a) in Appendix A.6 uses an estimated Q-function that is twice continuously differentiable in x.

Under Assumption A.5 (a), by Lemma B.4 in Appendix B.3 and with change of variables $v = \frac{\lambda}{\delta}$, for any sufficiently small $\delta > 0$,

$$\begin{aligned}
\Pr(X_i \in N(D^*, \delta)) &= \int_{-\delta}^{\delta} \int_{D^*} f_X(u + \lambda \nu_{\tilde{C}^*}(u)) J_{p-1}^{D^*} \psi_{\tilde{C}^*}(u, \lambda) d\mathcal{H}^{p-1}(u) d\lambda \\
&= \delta \int_{-1}^{1} \int_{D^*} f_X(u + \delta v \nu_{\tilde{C}^*}(u)) J_{p-1}^{D^*} \psi_{\tilde{C}^*}(u, \delta v) d\mathcal{H}^{p-1}(u) dv.
\end{aligned}$$

(See Appendix B for the notation.) If f_X is bounded on $N(D^*, \delta)$ and $\mathcal{H}^{p-1}(D^*) < \infty$, the right-hand side is $O(\delta)$.

A.5 Sampling from Uniform Distribution on p-Dimensional Ball

When we calculate fixed-bandwidth APS by simulation, we need to uniformly sample from $B(X_i, \delta)$. We introduce three existing methods to uniformly sample from a p-dimensional unit ball $B(\mathbf{0}, 1)$. By multiplying the sampled vector by δ and adding X_i to it, we can sample from a uniform distribution on $B(X_i, \delta)$.

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 $x = (s_1, ..., s_p)$.

A.6 Additional Examples

Example A.1 (Reinforcement Learning Algorithms)**.** Extending bandit algorithms to dynamically changing environments, reinforcement learning algorithms optimize decisions in dynamic environments, where the state (the set of observables that the agent receives from the environment) and action in the current period 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.[39] 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 $(x, z) \in \mathcal{X} \times \{0, 1\}$,

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

[39] Under a Markov decision process, 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.

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) (ϵ-Greedy) This algorithm first uses past data to yield $\hat{Q}$, an estimate of the Q-function. For example, the fitted Q iteration (Ernst, Geurts and Wehenkel, 2005) is used to estimate Q.[40] The algorithm then 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,

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

$$A^{\epsilon}(x) = \begin{cases} \frac{\epsilon}{2} & \text{if } \hat{Q}(x,1) < \hat{Q}(x,0) \\ 1 - \frac{\epsilon}{2} & \text{if } \hat{Q}(x,1) > \hat{Q}(x,0). \end{cases}$$

Suppose that the function $g(\cdot) = \hat{Q}(\cdot, 1) - \hat{Q}(\cdot, 0)$ is continuous on $\mathcal{X}$ and is continuously differentiable in a neighborhood of x with $\nabla g(x) \neq \mathbf{0}$ for any $x \in \mathcal{X}$ such that $g(x) = 0$. APS for this case is given by

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

(b) (Policy Gradient Methods) Policy gradient methods such as REINFORCE (Williams, 1992) and Actor-Critic 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.[41] 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. Policy gradient methods use the trajectories to update the policy parameter to θ^1 by stochastic gradient ascent. The algorithms then use the updated policy function $\pi(x;\theta^1)$ to determine the treatment assignment for new episodes. For each t,

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

Suppose that the function $\pi(\cdot;\theta^1)$ is continuous. APS for this case is given by

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

[40] Suppose that we have collected a set of L four-tuples $\{(x_{t_l}^l, z_{t_l}^l, y_{t_l}^l, x_{t_l+1}^l)\}_{l=1}^L$ as a result of the agent interacting with the dynamic environment. Given the dataset and an initial approximation $\hat{Q}$ of Q (e.g., $\hat{Q}(x,z) = 0$ for all (x,z)), we repeat the following steps until some stopping condition is reached: 1. For each $l = 1, ..., L$, calculate $q^l = y_{t_l}^l + \gamma \max_{z \in \{0,1\}} \hat{Q}(x_{t_l+1}^l, z)$; 2. 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, neural networks (Neural Fitted Q Iteration) and deep neural networks (Deep Fitted Q Iteration).

[41] 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.

B Notation and Lemmas

B.1 Basic Notations

For a scalar-valued differentiable function $f : S \subset \mathbb{R}^n \to \mathbb{R}$, let $\nabla f : S \to \mathbb{R}^n$ be a gradient of f: for every $x \in S$,

$$\nabla f(x) = \left(\frac{\partial f(x)}{\partial x_1}, \cdots, \frac{\partial f(x)}{\partial x_n} \right)'.$$

Also, when the second-order partial derivatives of f exist, let $D^2 f(x)$ be the Hessian matrix:

$$D^2 f(x) = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f(x)}{\partial x_n^2} \end{bmatrix}$$

for each $x \in S$.

Let $f : S \subset \mathbb{R}^m \to \mathbb{R}^n$ be a function such that its first-order partial derivatives exist. For each $x \in S$, let $Jf(x)$ be the Jacobian matrix of f at x:

$$Jf(x) = \begin{bmatrix} \frac{\partial f_1(x)}{\partial x_1} & \cdots & \frac{\partial f_1(x)}{\partial x_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n(x)}{\partial x_1} & \cdots & \frac{\partial f_n(x)}{\partial x_m} \end{bmatrix}.$$

For a positive integer n, let I_n denote the $n \times n$ identity matrix.

B.2 Differential Geometry

We provide some concepts and facts from differential geometry of twice continuously differentiable sets, following Crasta and Malusa (2007). Let $S \subset \mathbb{R}^p$ be a twice continuously differentiable set. For each $x \in \partial S$, we denote by $\nu_S(x) \in \mathbb{R}^p$ the inward unit normal vector of ∂S at x, that is, the unit vector orthogonal to all vectors in the tangent space of ∂S at x that points toward the inside of A. For a set $S \subset \mathbb{R}^p$, let $d_S^s : \mathbb{R}^p \to \mathbb{R}$ be the signed distance function of S, defined by

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

where $d(x, B) = \inf_{y \in B} \|y - x\|$ for any $x \in \mathbb{R}^p$ for a set $B \subset \mathbb{R}^p$. Note that we can write $N(\partial S, \delta) = \{x \in \mathbb{R}^p : -\delta < d_S^s(x) < \delta\}$ for $\delta > 0$. Lastly, let $\Pi_{\partial S}(x) = \{y \in \partial S : \|y - x\| = d(x, \partial S)\}$ be the set of projections of x on ∂S.

Lemma B.1 (Corollary of Theorem 4.16, Crasta and Malusa (2007))**.** *Let $S \subset \mathbb{R}^p$ be nonempty, bounded, open, connected and twice continuously differentiable. Then the function d_S^s is twice continuously differentiable on $N(\partial S, \mu)$ for some $\mu > 0$. In addition, for every $x_0 \in \partial S$, $\Pi_{\partial S}(x_0 + t\nu_S(x_0)) = \{x_0\}$ for every $t \in (-\mu, \mu)$. Furthermore, for every $x \in N(\partial S, \mu)$, $\Pi_{\partial S}(x)$ is a singleton, $\nabla d_S^s(x) = \nu_S(y)$ and $x = y + d_S^s(x)\nu_S(y)$ for $y \in \Pi_{\partial S}(x)$, and $\|\nabla d_S^s(x)\| = 1$.*

Proof. We apply results from Crasta and Malusa (2007). Let $K = \{x \in \mathbb{R}^p : \|x\| \leq 1\}$. K is nonempty, compact, convex subset of $\mathbb{R}^p$ with the origin as an interior point. The polar body of K, defined as $K_0 = \{y \in \mathbb{R}^p : y \cdot x \leq 1 \text{ for all } x \in K\}$, is K itself. The gauge functions $\rho_K, \rho_{K_0} : \mathbb{R}^p \to [0, \infty]$ of K and K_0 are given by

$$\rho_K(x) \equiv \inf\{t \geq 0 : x \in tK\} = \|x\|,$$
$$\rho_{K_0}(x) \equiv \inf\{t \geq 0 : x \in tK_0\} = \|x\|.$$

Given ρ_{K_0}, the Minkowski distance from a set $S \subset \mathbb{R}^p$ is defined as

$$\delta_S(x) \equiv \inf_{y \in S} \rho_{K_0}(x - y), \quad x \in \mathbb{R}^p.$$

Note that we can write

$$d_S^s(x) = \begin{cases} \delta_{\partial S}(x) & \text{if } x \in \text{cl}(S) \\ -\delta_{\partial S}(x) & \text{if } x \in \mathbb{R}^p \setminus \text{cl}(S). \end{cases}$$

It then follows from Theorem 4.16 of Crasta and Malusa (2007) that d_S^s is twice continuously differentiable on $N(\partial S, \mu)$ for some $\mu > 0$, and for every $x_0 \in \partial S$,

$$\nabla d_S^s(x_0) = \frac{\nu_S(x_0)}{\rho_K(\nu_S(x_0))} = \frac{\nu_S(x_0)}{\|\nu_S(x_0)\|} = \nu_S(x_0),$$

where the last equality follows since $\nu_S(x_0)$ is a unit vector. It then follows that $\|\nabla d_S^s(x_0)\| = \|\nu_S(x_0)\| = 1$ for every $x_0 \in \partial S$. Also, it is obvious that, for every $x_0 \in \partial S$, $\Pi_{\partial S}(x_0) = \{x_0\}$ and $x_0 = x_0 + d_S^s(x_0)\nu_S(x_0)$, since $d_S^s(x_0) = 0$. In addition, as stated in the proof of Theorem 4.16 of Crasta and Malusa (2007), μ is chosen so that (4.7) in Proposition 4.6 of Crasta and Malusa (2007) holds for every $x_0 \in \partial S$ and every $t \in (-\mu, \mu)$. That is, $\Pi_{\partial S}(x_0 + t\nabla\rho_K(\nu_S(x_0))) = \{x_0\}$ for every $x_0 \in \partial S$ and every $t \in (-\mu, \mu)$. Since $\nabla\rho_K(\nu_S(x_0)) = \frac{\nu_S(x_0)}{\|\nu_S(x_0)\|} = \nu_S(x_0)$, $\Pi_{\partial S}(x_0 + t\nu_S(x_0)) = \{x_0\}$ for every $x_0 \in \partial S$ and every $t \in (-\mu, \mu)$.

Furthermore, for every $x \in N(\partial S, \mu) \setminus \partial S$, $\Pi_{\partial S}(x)$ is a singleton as shown in the proof of Theorem 4.16 of Crasta and Malusa (2007). Let $\pi_{\partial S}(x)$ be the unique element in $\Pi_{\partial S}(x)$. By Lemma 4.3 of Crasta and Malusa (2007), for every $x \in N(\partial S, \mu) \setminus \partial S$,

$$\nabla d_S^s(x) = \frac{\nu_S(\pi_{\partial S}(x))}{\rho_K(\nu_S(\pi_{\partial S}(x)))} = \frac{\nu_S(\pi_{\partial S}(x))}{\|\nu_S(\pi_{\partial S}(x))\|} = \nu_S(\pi_{\partial S}(x)),$$

where the last equality follows since $\nu_S(\pi_{\partial S}(x))$ is a unit vector. It then follows that $\|\nabla d_S^s(x)\| = \|\nu_S(\pi_{\partial S}(x))\| = 1$ for every $x \in N(\partial S, \mu) \setminus \partial S$.

Lastly, note that

$$\delta_{\partial S}(x) = \begin{cases} d_S^s(x) & \text{if } x \in N(\partial S, \mu) \cap \text{int}(S) \\ -d_S^s(x) & \text{if } x \in N(\partial S, \mu) \setminus \text{cl}(S), \end{cases}$$

and

$$\nabla\delta_{\partial S}(x) = \begin{cases} \nabla d_S^s(x) & \text{if } x \in N(\partial S, \mu) \cap \text{int}(S) \\ -\nabla d_S^s(x) & \text{if } x \in N(\partial S, \mu) \setminus \text{cl}(S), \end{cases}$$

so $\delta_{\partial S}(x)\nabla\delta_{\partial S}(x) = d_S^s(x)\nabla d_S^s(x) = d_S^s(x)\nu_S(\pi_{\partial S}(x))$ for every $x \in N(\partial S, \mu) \setminus \partial S$. By Proposition 3.3 (i) of Crasta and Malusa (2007), for every $x \in N(\partial S, \mu) \setminus \partial S$,

$$\nabla\rho_K(\nabla\delta_{\partial S}(x)) = \frac{x - \pi_{\partial S}(x)}{\delta_{\partial S}(x)},$$

which implies that

$$\begin{aligned} x &= \pi_{\partial S}(x) + \delta_{\partial S}(x)\nabla\rho_K(\nabla\delta_{\partial S}(x)) \\ &= \pi_{\partial S}(x) + \delta_{\partial S}(x)\frac{\nabla\delta_{\partial S}(x)}{\|\nabla\delta_{\partial S}(x)\|} = \pi_{\partial S}(x) + d_S^s(x)\nu_S(\pi_{\partial S}(x)). \end{aligned}$$

□

We say that a set $S \subset \mathbb{R}^n$ is a *m-dimensional C^1 submanifold of* $\mathbb{R}^n$ if for every point $x \in S$, there exist an open neighborhood $V \subset \mathbb{R}^n$ of x and a one-to-one continuously differentiable function ϕ from an open set $U \subset \mathbb{R}^m$ to $\mathbb{R}^n$ such that the Jacobian matrix $J\phi(u)$ is of rank m for all $u \in U$, and $\phi(U) = V \cap S$.

Lemma B.2. *Let $S \subset \mathbb{R}^p$ be nonempty, bounded, open, connected and twice continuously differentiable. Then ∂S is a $(p-1)$-dimensional C^1 submanifold of $\mathbb{R}^p$,*

Proof. Fix any $x^* \in \partial S$. By Lemma B.1, $\nabla d_S^s(x^*)$ is nonzero. Without loss of generality, let $\frac{\partial d_S^s(x^*)}{\partial x_p} \neq 0$. Let $\psi : \mathbb{R}^p \to \mathbb{R}^p$ be the function such that $\psi(x) = (x_1, ..., x_{p-1}, d_S^s(x))$. ψ is continuously differentiable, and the Jacobian matrix of ψ at x^* is given by

$$J\psi(x^*) = \begin{pmatrix} \frac{\partial\psi_1}{\partial x_1}(x^*) & \cdots & \frac{\partial\psi_1}{\partial x_p}(x^*) \\ \vdots & \ddots & \vdots \\ \frac{\partial\psi_p}{\partial x_1}(x^*) & \cdots & \frac{\partial\psi_p}{\partial x_p}(x^*) \end{pmatrix} = \begin{pmatrix} & & & 0 \\ & I_{p-1} & & \vdots \\ & & & 0 \\ \frac{\partial d_S^s(x^*)}{\partial x_1} & \cdots & \frac{\partial d_S^s(x^*)}{\partial x_{p-1}} & \frac{\partial d_S^s(x^*)}{\partial x_p} \end{pmatrix}.$$

Since $\frac{\partial d_S^s(x^*)}{\partial x_p} \neq 0$, 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 $\frac{\partial d_S^s(x)}{\partial x_p} \neq 0$ 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 $x \in V \cap \partial S$ by the definition of d_S^s. Let $U = \{(x_1, ..., x_{p-1}) \in \mathbb{R}^{p-1} : x \in V \cap \partial S\}$ and $\phi : U \to \mathbb{R}^p$ be a function 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 $\phi(U) = V \cap \partial S$, 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,

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

Lastly, we show that U is open. Pick any $\bar{u} \in U$. Then, there exists $\bar{x}_p \in \mathbb{R}$ such that $(\bar{u}, \bar{x}_p) \in V \cap \partial S$. As $(\bar{u}, \bar{x}_p) \in V \cap \partial S$, $d_S^s((\bar{u}, \bar{x}_p)) = 0$. Since $\frac{\partial d_S^s((\bar{u}, \bar{x}_p))}{\partial x_p} \neq 0$, 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 $d_S^s(u, g(u)) = 0$ 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^s, $d_S^s(x) = 0$ if and only if $x \in \partial S$. Therefore, if $u \in S'$, $(u, g(u))$ must be contained by ∂S, for otherwise $d_S^s(u, g(u)) \neq 0$, which is a contradiction. Thus, $(u, g(u)) \in V \cap \partial S$ 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. □

B.3 Geometric Measure Theory

We provide some concepts and facts from geometric measure theory, following Krantz and Parks (2008). Recall that for a function $f : S \subset \mathbb{R}^m \to \mathbb{R}^n$ and a point $x \in S$ at which f is differentiable, $Jf(x)$ denotes the Jacobian matrix of f at x.

Lemma B.3 (Coarea Formula, Lemma 5.1.4 and Corollary 5.2.6 of Krantz and Parks (2008))**.** *If $f : \mathbb{R}^m \to \mathbb{R}^n$ is a Lipschitz function and $m \geq n$, then*

$$\int_S g(x) J_n f(x) d\mathcal{L}^m(x) = \int_{\mathbb{R}^n} \int_{\{x' \in S : f(x') = y\}} g(x) d\mathcal{H}^{m-n}(x) d\mathcal{L}^n(y)$$

for every Lebesgue measurable subset S of $\mathbb{R}^m$ and every $\mathcal{L}^m$-measurable function $g : S \to \mathbb{R}$, where for each $x \in \mathbb{R}^m$ at which f is differentiable,

$$J_n f(x) = \sqrt{\det((Jf(x))(Jf(x))')}.$$

Let S be an m-dimensional C^1 submanifold of $\mathbb{R}^n$. Let $x \in S$ and let $\phi : U \subset \mathbb{R}^m \to \mathbb{R}^n$ be as in the definition of m-dimensional C^1 submanifold. We denote by $T_S(x)$ the tangent space of S at x, $\{J\phi(u)v : v \in \mathbb{R}^m\}$, where $u = \phi^{-1}(x)$.

Lemma B.4 (Area Formula, Lemma 5.3.5 and Theorem 5.3.7 of Krantz and Parks (2008))**.** *Suppose $m \leq \nu$ and $f : \mathbb{R}^n \to \mathbb{R}^\nu$ is Lipschitz. If S is an m-dimensional C^1 submanifold of $\mathbb{R}^n$, then*

$$\int_S g(x) J_m^S f(x) d\mathcal{H}^m(x) = \int_{\mathbb{R}^\nu} \sum_{x \in S : f(x) = y} g(x) d\mathcal{H}^m(y)$$

for every $\mathcal{H}^m$-measurable function $g : S \to \mathbb{R}$, where for each $x \in \mathbb{R}^n$ at which f is differentiable,

$$J_m^S f(x) = \frac{\mathcal{H}^m(\{Jf(x)y : y \in P\})}{\mathcal{H}^m(P)}$$

for an arbitrary m-dimensional parallelepiped P contained in $T_S(x)$.

Let $S \subset \mathbb{R}^p$. For each $x \in \mathbb{R}^p$ at which d_S^s is differentiable and for each $\lambda \in \mathbb{R}$, let $\psi_S(x, \lambda) = x + \lambda \nabla d_S^s(x)$.

Lemma B.5. *Let $\Omega \subset \mathbb{R}^p$, and suppose that 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 $m \in \{1, ..., M\}$.

Then there exists $\mu > 0$ such that d_Ω^s is twice continuously differentiable on $N(\partial\Omega, \mu)$ and that

$$\int_{N(\partial\Omega,\delta)} g(x)dx = \int_{-\delta}^{\delta} \int_{\partial\Omega} g(u + \lambda\nu_\Omega(u)) J_{p-1}^{\partial\Omega} \psi_\Omega(u, \lambda) d\mathcal{H}^{p-1}(u) d\lambda$$

for every $\delta \in (0, \mu)$ and every function $g : \mathbb{R}^p \to \mathbb{R}$ that is integrable on $N(\partial\Omega, \delta)$, where for each fixed $\lambda \in (-\mu, \mu)$, $J_{p-1}^{\partial\Omega}\psi_\Omega(\cdot, \lambda)$ is calculated by applying the operation $J_{p-1}^{\partial\Omega}$ to the function $\psi_\Omega(\cdot, \lambda)$. Futhermore, $J_{p-1}^{\partial\Omega}\psi_\Omega(x, \cdot)$ is continuously differentiable in λ and $J_{p-1}^{\partial\Omega}\psi_\Omega(x, 0) = 1$ for every $x \in \partial\Omega$, and $J_{p-1}^{\partial\Omega}\psi_\Omega(\cdot, \cdot)$ and $\frac{\partial J_{p-1}^{\partial\Omega}\psi_\Omega(\cdot,\cdot)}{\partial\lambda}$ are bounded on $\partial\Omega \times (-\mu, \mu)$.

Proof. Let $\bar{\mu} = \frac{1}{2} \min_{m,m' \in \{1,...,M\}, m \neq m'} \text{dist}(\Omega_m^*, \Omega_{m'})$ so that $\{N(\partial\Omega_m, \bar{\mu})\}_{m=1}^M$ is a partition of $N(\partial\Omega, \bar{\mu})$. Note that for every $m \in \{1, ..., M\}$, $d_\Omega^s(x) = d_{\Omega_m}^s(x)$ for every $x \in N(\partial\Omega_m, \bar{\mu})$. By Lemma B.1, for every $m \in \{1, ..., M\}$, there exists $\bar{\mu}_m > 0$ such that $d_{\Omega_m}^s$ is twice continuously differentiable on $N(\partial\Omega_m, \bar{\mu}_m)$. Letting $\mu \in (0, \min\{\bar{\mu}, \bar{\mu}_1, ..., \bar{\mu}_M\})$, we have that d_Ω^s is twice continuously differentiable on $N(\partial\Omega, \mu)$. This implies that d_Ω^s is Lipschitz on $N(\partial\Omega, \mu)$. For every $\delta \in (0, \mu)$ and every function $g : \mathbb{R}^p \to \mathbb{R}$ that is integrable on $N(\partial\Omega, \delta)$,

$$\begin{aligned}
\int_{N(\partial\Omega,\delta)} g(x)dx &= \int_{\{x' \in \mathbb{R}^p : d_\Omega^s(x') \in (-\delta,\delta)\}} g(x) \sqrt{\det(\|\nabla d_\Omega^s(x)\|)} dx \\
&= \int_{\{x' \in \mathbb{R}^p : d_\Omega^s(x') \in (-\delta,\delta)\}} g(x) \sqrt{\det(\nabla d_\Omega^s(x)' \nabla d_\Omega^s(x))} dx \\
&= \int_{\{x' \in \mathbb{R}^p : d_\Omega^s(x') \in (-\delta,\delta)\}} g(x) \sqrt{\det((Jd_\Omega^s(x))(Jd_\Omega^s(x))')} dx \\
&= \int_{\mathbb{R}} \int_{\{x' \in \mathbb{R}^p : d_\Omega^s(x') \in (-\delta,\delta), d_\Omega^s(x') = \lambda\}} g(x) d\mathcal{H}^{p-1}(x) d\lambda \\
&= \int_{-\delta}^{\delta} \int_{\{x' \in \mathbb{R}^p : d_\Omega^s(x') = \lambda\}} g(x) d\mathcal{H}^{p-1}(x) d\lambda, \qquad (12)
\end{aligned}$$

where the first equality follows since $\|\nabla d_\Omega^s(x)\| = 1$ for every $x \in N(\partial\Omega, \delta)$ by Lemma B.1, the third equality follows from the definition of the Jacobian matrix, and the fourth equality follows from Lemma B.3.

Let $\Gamma(\lambda) = \{x \in \mathbb{R}^p : d^s_\Omega(x) = \lambda\}$ for each $\lambda \in (-\mu, \mu)$. Since ∇d^s_Ω is differentiable on $N(\partial\Omega, \mu)$, $\psi_\Omega(x, \lambda)$ is defined on $N(\partial\Omega, \mu) \times \mathbb{R}$. We show that $\{\psi_\Omega(x_0, \lambda) : x_0 \in \partial\Omega\} \subset \Gamma(\lambda)$ for every $\lambda \in (-\mu, \mu)$. By Lemma B.1, for every $x_0 \in \partial\Omega$, $\psi_\Omega(x_0, \lambda) = x_0 + \lambda\nu_\Omega(x_0)$ and

$$\Pi_{\partial\Omega}(\psi_\Omega(x_0, \lambda)) = \Pi_{\partial\Omega}(x_0 + \lambda\nu_\Omega(x_0)) = \{x_0\}.$$

Hence,

$$d(\psi_\Omega(x_0, \lambda), \partial\Omega) = \|\psi_\Omega(x_0, \lambda) - x_0\| = \|\lambda\nu_\Omega(x_0)\| = |\lambda|.$$

Since $\nu_\Omega(x_0)$ is an inward normal vector, $\psi_\Omega(x_0, \lambda) \in \mathrm{cl}(\Omega)$ if $0 \le \lambda < \mu$, and $\psi_\Omega(x, \lambda_0) \in \mathbb{R}^p \setminus \mathrm{cl}(\Omega)$ if $-\mu < \lambda < 0$. It follows that

$$\begin{aligned} d^s_\Omega(\psi_\Omega(x_0, \lambda)) &= \begin{cases} |\lambda| & \text{if } 0 \le \lambda < \mu \\ -|\lambda| & \text{if } \mu < \lambda < 0 \end{cases} \\ &= \lambda, \end{aligned}$$

so $\{\psi_\Omega(x_0, \lambda) : x_0 \in \partial\Omega\} \subset \Gamma(\lambda)$. It also holds that $\Gamma(\lambda) \subset \{\psi_\Omega(x_0, \lambda) : x_0 \in \partial\Omega\}$, since by Lemma B.1, for every $x \in \Gamma(\lambda)$,

$$\psi_\Omega(\pi_{\partial\Omega}(x), \lambda) = \pi_{\partial\Omega}(x) + \lambda\nabla d^s_\Omega(\pi_{\partial\Omega}(x)) = \pi_{\partial\Omega}(x) + d^s_\Omega(x)\nu_\Omega(\pi_{\partial\Omega}(x)) = x,$$

where $\pi_{\partial\Omega}(x)$ is the unique element in $\Pi_{\partial\Omega}(x)$. Thus, $\{\psi_\Omega(x_0, \lambda) : x_0 \in \partial\Omega\} = \Gamma(\lambda)$.

Now note that $\{\partial\Omega_m\}_{m=1}^M$ is a partition of $\partial\Omega$, since $\mathrm{dist}(\Omega_m, \Omega_{m'}) > 0$ for any $m, m' \in \{1, ..., M\}$ such that $m \ne m'$. By Lemma B.2, $\partial\Omega_m$ is a $(p-1)$-dimensional C^1 submanifold of $\mathbb{R}^p$ for every $m \in \{1, ..., M\}$, and hence $\partial\Omega$ is a $(p-1)$-dimensional C^1 submanifold of $\mathbb{R}^p$. Furthermore, since ∇d^s_Ω is continuously differentiable on $N(\partial\Omega, \mu)$, $\psi_\Omega(\cdot, \lambda)$ is continuously differentiable on $N(\partial\Omega, \mu)$, which implies that $\psi_\Omega(\cdot, \lambda)$ is Lipschitz on $N(\partial\Omega, \mu)$ for every $\lambda \in \mathbb{R}$. Applying Lemma B.4, we have that for every $\lambda \in (-\mu, \mu)$,

$$\begin{aligned} \int_{\partial\Omega} g(u + \lambda\nu_\Omega(u)) J^{\partial\Omega}_{p-1}\psi_\Omega(u, \lambda) d\mathcal{H}^{p-1}(u) &= \int_{\partial\Omega} g(\psi_\Omega(u, \lambda)) J^{\partial\Omega}_{p-1}\psi_\Omega(u, \lambda) d\mathcal{H}^{p-1}(u) \\ &= \int_{\mathbb{R}^p} \sum_{u \in \partial\Omega : \psi_\Omega(u,\lambda) = x} g(\psi_\Omega(u, \lambda)) d\mathcal{H}^{p-1}(x). \end{aligned} \tag{13}$$

If $x \notin \{\psi_\Omega(u, \lambda) : u \in \partial\Omega\}$, $\{u \in \partial\Omega : \psi_\Omega(u, \lambda) = x\} = \emptyset$. If $x \in \{\psi_\Omega(u, \lambda) : u \in \partial\Omega\}$, there exists $u \in \partial\Omega$ such that $x = \psi_\Omega(u, \lambda)$. Since $\Pi_{\partial\Omega}(x) = \Pi_{\partial\Omega}(u + \lambda\nabla d^s_\Omega(u)) = \Pi_{\partial\Omega}(u + \lambda\nu_\Omega(u)) = \{u\}$ by Lemma B.1, such u is unique, and hence $\{u \in \partial\Omega : \psi_\Omega(u, \lambda) = x\}$ is a singleton. It follow that

$$\begin{aligned} \int_{\mathbb{R}^p} \sum_{u \in \partial\Omega : \psi_\Omega(u,\lambda) = x} g(\psi_\Omega(u, \lambda)) d\mathcal{H}^{p-1}(x) &= \int_{\{\psi_\Omega(u,\lambda) : u \in \partial\Omega\}} g(x) d\mathcal{H}^{p-1}(x) \\ &= \int_{\Gamma(\lambda)} g(x) d\mathcal{H}^{p-1}(x), \end{aligned} \tag{14}$$

where the last equality holds since $\{\psi_\Omega(u,\lambda) : u \in \partial\Omega\} = \Gamma(\lambda)$. Combining (12), (13) and (14), we obtain

$$\int_{N(\partial\Omega,\delta)} g(x)dx = \int_{-\delta}^{\delta}\int_{\partial\Omega} g(u+\lambda\nu_\Omega(u))J_{p-1}^{\partial\Omega}\psi_\Omega(u,\lambda)d\mathcal{H}^{p-1}(u)d\lambda.$$

We next show that $J_{p-1}^{\partial\Omega}\psi_\Omega(x,\cdot)$ is continuously differentiable in λ and $J_{p-1}^{\partial\Omega}\psi_\Omega(x,0) = 1$ for every $x \in \partial\Omega$. Fix an $x \in \partial\Omega$, and let $V_\Omega(x)$ be an arbitrary $p \times (p-1)$ matrix whose columns $v_1(x), ..., v_{p-1}(x) \in \mathbb{R}^p$ form an orthonormal basis of $T_{\partial\Omega}(x)$. Let $P(x) \subset T_{\partial\Omega}(x)$ be a parallelepiped determined by $v_1(x), ..., v_{p-1}(x)$, that is, let $P(x) = \{\sum_{k=1}^{p-1} c_k v_k(x) : 0 \leq c_k \leq 1 \text{ for } k = 1, ..., p-1\}$. Since $v_1(x), ..., v_{p-1}(x)$ are linearly independent, $P(x)$ is a $(p-1)$-dimensional parallelepiped. It follows that for each fixed $\lambda \in \mathbb{R}$,

$$\begin{aligned}
\{J\psi_\Omega(x,\lambda)y : y \in P(x)\} &= \{J\psi_\Omega(x,\lambda)\sum_{k=1}^{p-1} c_k v_k(x) : 0 \leq c_k \leq 1 \text{ for } k = 1, ..., p-1\} \\
&= \{\sum_{k=1}^{p-1} c_k J\psi_\Omega(x,\lambda)v_k(x) : 0 \leq c_k \leq 1 \text{ for } k = 1, ..., p-1\} \\
&= \{\sum_{k=1}^{p-1} c_k w_k(x,\lambda) : 0 \leq c_k \leq 1 \text{ for } k = 1, ..., p-1\},
\end{aligned}$$

where $w_k(x,\lambda) = J\psi_\Omega(x,\lambda)v_k(x)$ for $k = 1, ..., p-1$. Since $J\psi_\Omega(x,\lambda)v_k(x)$ is the k-th column of $J\psi_\Omega(x,\lambda)V_\Omega(x)$, $\{J\psi_\Omega(x,\lambda)y : y \in P(x)\}$ is the parallelepiped determined by the columns of $J\psi_\Omega(x,\lambda)V_\Omega(x)$. By Proposition 5.1.2 of Krantz and Parks (2008), we have that

$$\begin{aligned}
J_{p-1}^{\partial\Omega}\psi_\Omega(x,\lambda) &= \frac{\mathcal{H}^{p-1}(\{\sum_{k=1}^{p-1} c_k w_k(x,\lambda) : 0 \leq c_k \leq 1 \text{ for } k = 1, ..., p-1\})}{\mathcal{H}^{p-1}(P(x))} \\
&= \frac{\sqrt{\det((J\psi_\Omega(x,\lambda)V_\Omega(x))'(J\psi_\Omega(x,\lambda)V_\Omega(x)))}}{\sqrt{\det(V_\Omega(x)'V_\Omega(x))}} \\
&= \frac{\sqrt{\det((V_\Omega(x)+\lambda D^2 d_\Omega^s(x)V_\Omega(x))'(V_\Omega(x)+\lambda D^2 d_\Omega^s(x)V_\Omega(x)))}}{\sqrt{\det(I_{p-1})}} \\
&= \sqrt{\det(V_\Omega(x)'V_\Omega(x) + 2V_\Omega(x)'\lambda D^2 d_\Omega^s(x)V_\Omega(x) + V_\Omega(x)'(\lambda D^2 d_\Omega^s(x))^2 V_\Omega(x))} \\
&= \sqrt{\det(I_{p-1} + \lambda V_\Omega(x)'(2D^2 d_\Omega^s(x) + \lambda(D^2 d_\Omega^s(x))^2)V_\Omega(x)))} \\
&= \sqrt{\det(I_p + \lambda V_\Omega(x)V_\Omega(x)'(2D^2 d_\Omega^s(x) + \lambda(D^2 d_\Omega^s(x))^2))},
\end{aligned}$$

where we use the fact that $V_\Omega(x)'V_\Omega(x) = I_{p-1}$ and the fact that $\det(I_m + AB) = \det(I_n + BA)$ for an $m \times n$ matrix A and an $n \times m$ matrix B (the Weinstein-Aronszajn identity). For every $x \in \partial\Omega$, $J_{p-1}^{\partial\Omega}\psi_\Omega(x,\cdot)$ is continuously differentiable in λ, and $J_{p-1}^{\partial\Omega}\psi_\Omega(x,0) = \sqrt{\det(I_p)} = 1$.

Lastly, we show that $J_{p-1}^{\partial\Omega}\psi_\Omega(\cdot,\cdot)$ and $\frac{\partial J_{p-1}^{\partial\Omega}\psi_\Omega(\cdot,\cdot)}{\partial\lambda}$ are bounded on $\partial\Omega \times (-\mu,\mu)$. Let $f, h : \partial\Omega \times \mathbb{R}^{p\times(p-1)} \to \mathbb{R}^{p\times p}$ be functions such that

$$\begin{aligned}
f(x,A) &= 2AA'D^2 d_\Omega^s(x), \\
h(x,A) &= AA'(D^2 d_\Omega^s(x))^2.
\end{aligned}$$

Also, let $k : \partial\Omega \times \mathbb{R} \times \mathbb{R}^{p\times(p-1)} \to \mathbb{R}$ be a function such that

$$k(x, \lambda, A) = \sqrt{\det(I_p + \lambda f(x, A) + \lambda^2 h(x, A))}.$$

Observe that

$$J^{\partial\Omega}_{p-1}\psi_\Omega(x, \lambda) = k(x, \lambda, V_\Omega(x))$$

and that

$$\begin{aligned}
&\frac{\partial J^{\partial\Omega}_{p-1}\psi_\Omega(x, \lambda)}{\partial\lambda} \\
&= \left.\frac{\partial k(x, \lambda, A)}{\partial\lambda}\right|_{A=V_\Omega(x)} \\
&= \left.\frac{1}{2k(x, \lambda, A)} \sum_{i,j} \frac{\partial \det(I_p + \lambda f(x, A) + \lambda^2 h(x, A))}{\partial b_{ij}} (f_{ij}(x, A) + 2\lambda h_{ij}(x, A))\right|_{A=V_\Omega(x)},
\end{aligned}$$

where $\frac{\partial \det(B)}{\partial b_{ij}}$ denotes the partial derivative of the function $\det : \mathbb{R}^{p\times p} \to \mathbb{R}$ with respect to the (i, j) entry of B.

Note that $k(\cdot, \cdot, \cdot)$ and $\frac{\partial k(\cdot,\cdot,\cdot)}{\partial\lambda}$ are continuous on $\partial\Omega \times \mathbb{R} \times \mathbb{R}^{p\times(p-1)}$ (except at the points for which $k(x, \lambda, A) = 0$), since det is infinitely differentiable, and f and h are continuous on $\partial\Omega \times \mathbb{R}^{p\times(p-1)}$. Let $S = \{(x, \lambda, A) \in \partial\Omega \times [-\mu, \mu] \times \mathbb{R}^{p\times(p-1)} : \|a_j\| = 1 \text{ for } k = 1, ..., p-1\}$, where a_j denotes the jth column of A. Since $k(\cdot, \cdot, \cdot)$ and $\frac{\partial k(\cdot,\cdot,\cdot)}{\partial\lambda}$ are continuous and S is closed and bounded, $\bar{k} = \max_{(x,\lambda,A)\in S} |k(x, \lambda, A)|$ and $\bar{k}' = \max_{(x,\lambda,A)\in S} |\frac{\partial k(x,\lambda,A)}{\partial\lambda}|$ exist. Since $(x, \lambda, V_\Omega(x)) \in S$ for every $(x, \lambda) \in \partial\Omega \times (-\mu, \mu)$, it follows that $|J^{\partial\Omega}_{p-1}\psi_\Omega(x, \lambda)| \leq \bar{k}$ and $|\frac{\partial J^{\partial\Omega}_{p-1}\psi_\Omega(x,\lambda)}{\partial\lambda}| \leq \bar{k}'$ for every $(x, \lambda) \in \partial\Omega \times (-\mu, \mu)$. □

B.4 Other Lemmas

Lemma B.6. *Let $\{V_i\}_{i=1}^\infty$ be i.i.d. random variables such that $E[V_i^2] < \infty$. If Assumption 1 holds, then for $l \geq 0$ and $m = 0, 1$,*

$$E[V_i p^A(X_i; \delta)^l 1\{p^A(X_i; \delta) \in (0, 1)\}^m] \to E[V_i A(X_i)^l 1\{A(X_i) \in (0, 1)\}^m]$$

as $\delta \to 0$. Moreover, if, in addition, $\delta_n \to 0$ as $n \to \infty$, then for $l \geq 0$,

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

as $n \to \infty$.

Proof. Note that $E[\frac{1}{n}\sum_{i=1}^n V_i p^A(X_i; \delta_n)^l I_{i,n}] = E[V_i p^A(X_i; \delta_n)^l 1\{p^A(X_i; \delta_n) \in (0, 1)\}]$. We show that

$$E[V_i p^A(X_i; \delta)^l 1\{p^A(X_i; \delta) \in (0, 1)\}^m] \to E[V_i A(X_i)^l 1\{A(X_i) \in (0, 1)\}^m]$$

for $l \geq 0$ and $m = 0, 1$ as $\delta \to 0$, and that

$$\text{Var}(\frac{1}{n}\sum_{i=1}^{n} V_i p^A(X_i;\delta_n)^l I_{i,n}) \to 0$$

for $l \geq 0$ as $n \to \infty$. For the first part, we have

$$E[V_i p^A(X_i;\delta)^l 1\{p^A(X_i;\delta) \in (0,1)\}^m] = \int_{\mathcal{X}} E[V_i|X_i = x] p^A(x;\delta)^l 1\{p^A(x;\delta) \in (0,1)\}^m f_X(x)dx.$$

Suppose A is continuous at x and $A(x) \in (0,1)$. Then $\lim_{\delta\to 0} p^A(x;\delta) = A(x)$ by Part 1 of Corollary A.1, and hence $p^A(x;\delta) \in (0,1)$ for sufficiently small $\delta > 0$. It follows that $1\{p^A(x;\delta) \in (0,1)\} \to 1 = 1\{A(x) \in (0,1)\}$ as $\delta \to 0$. Suppose $x \in \text{int}(\mathcal{X}_0) \cup \text{int}(\mathcal{X}_1)$. Then $B(x,\delta) \subset \mathcal{X}_0$ or $B(x,\delta) \subset \mathcal{X}_1$ for sufficiently small $\delta > 0$ by the fact that $\text{int}(\mathcal{X}_0)$ and $\text{int}(\mathcal{X}_1)$ are open, and hence $1\{p^A(x;\delta) \in (0,1)\} \to 0 = 1\{A(x) \in (0,1)\}$ as $\delta \to 0$. Therefore, $\lim_{\delta\to 0} p^A(x;\delta) = A(x)$ and $\lim_{\delta\to 0} 1\{p^A(x;\delta) \in (0,1)\} = 1\{A(x) \in (0,1)\}$ for almost every $x \in \mathcal{X}$, since A is continuous at x for almost every $x \in \mathcal{X}$ by Assumption 1 (a), and either $A(x) \in (0,1)$ or $x \in \text{int}(\mathcal{X}_0) \cup \text{int}(\mathcal{X}_1)$ for almost every $x \in \mathcal{X}$ by Assumption 1 (b). By the Dominated Convergence Theorem,

$$\begin{aligned} E[V_i p^A(X_i;\delta)^l 1\{p^A(X_i;\delta) \in (0,1)\}^m] &\to \int_{\mathcal{X}} E[V_i|X_i = x] A(x)^l 1\{A(x) \in (0,1)\}^m f_X(x)dx \\ &= E[V_i A(X_i)^l 1\{A(X_i) \in (0,1)\}^m] \end{aligned}$$

as $\delta \to 0$. As for variance,

$$\begin{aligned} \text{Var}(\frac{1}{n}\sum_{i=1}^{n} V_i p^A(X_i;\delta_n)^l I_{i,n}) &\leq \frac{1}{n} E[V_i^2 p^A(X_i;\delta_n)^{2l} (I_{i,n})^2] \\ &\leq \frac{1}{n} E[V_i^2] \\ &\to 0 \end{aligned}$$

as $n \to \infty$. □

Lemma B.7. *Let $\{(\delta_n, S_n)\}_{n=1}^{\infty}$ be any sequence of positive numbers and positive integers. Fix $x \in \mathcal{X}$, and let $X_1^*, ..., X_{S_n}^*$ be S_n independent draws from the uniform distribution on $B(x,\delta_n)$ so that*

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

Then,

$$\begin{aligned} E[p^s(x;\delta_n) - p^A(x;\delta_n)] &= 0, \\ E[(p^s(x;\delta_n) - p^A(x;\delta_n))^2] &\leq \frac{1}{S_n}, \\ |E[p^s(x;\delta_n)^2 - p^A(x;\delta_n)^2]| &\leq \frac{1}{S_n}, \\ E[(p^s(x;\delta_n)^2 - p^A(x;\delta_n)^2)^2] &\leq \frac{4}{S_n}, \\ \Pr(p^s(x;\delta_n) \in \{0,1\}) &\leq (1 - p^A(x;\delta_n))^{S_n} + p^A(x;\delta_n)^{S_n}. \end{aligned}$$

Moreover, for any $\epsilon > 0$,

$$E[|p^s(x;\delta_n) - p^A(x;\delta_n)|] \leq \frac{1}{S_n\epsilon^2} + \epsilon,$$

and if $S_n \to \infty$, *then*

$$E[|p^s(x;\delta_n) - p^A(x;\delta_n)|] \to 0$$

as $n \to \infty$.

Proof. By construction, $E[A(X_s^*)] = p^A(x;\delta_n)$, so

$$\begin{aligned} E[p^s(x;\delta_n) - p^A(x;\delta_n)] &= E[\frac{1}{S_n}\sum_{s=1}^{S_n} A(X_s^*)] - p^A(x;\delta_n) \\ &= E[A(X_s^*)] - p^A(x;\delta_n) \\ &= 0. \end{aligned}$$

We have

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

$$\begin{aligned} |E[p^s(x;\delta_n)^2 - p^A(x;\delta_n)^2]| &= |\mathrm{Var}(p^s(x;\delta_n)) + (E[p^s(x;\delta_n)])^2 - p^A(x;\delta_n)^2| \\ &\leq \frac{1}{S_n} + |(p^A(x;\delta_n))^2 - p^A(x;\delta_n)^2| \\ &= \frac{1}{S_n}, \end{aligned}$$

and

$$\begin{aligned} & E[(p^s(x;\delta_n)^2 - p^A(x;\delta_n)^2)^2] \\ &= E[(p^s(x;\delta_n) + p^A(x;\delta_n))^2(p^s(x;\delta_n) - p^A(x;\delta_n))^2] \\ &\leq 4E[(p^s(x;\delta_n) - p^A(x;\delta_n))^2] \\ &\leq \frac{4}{S_n}. \end{aligned}$$

Now note that we have the following bounds on $\Pr(A(X_s^*) = 0)$ and $\Pr(A(X_s^*) = 1)$:

$$\begin{aligned} 0 \leq \Pr(A(X_s^*) = 0) \leq 1 - p^A(x;\delta_n), \\ 0 \leq \Pr(A(X_s^*) = 1) \leq p^A(x;\delta_n). \end{aligned}$$

It follows that

$$\begin{aligned}0 &\leq \Pr(p^s(x;\delta_n) \in \{0,1\}) \\ &= \Pr(A(X_s^*) = 0)^{S_n} + \Pr(A(X_s^*) = 1)^{S_n} \\ &\leq (1 - p^A(x;\delta_n))^{S_n} + p^A(x;\delta_n)^{S_n}.\end{aligned}$$

Lastly, for any $\epsilon > 0$,

$$\begin{aligned}&E[|p^s(x;\delta_n) - p^A(x;\delta_n)|] \\ &= E[|p^s(x;\delta_n) - p^A(x;\delta_n)||\,|p^s(x;\delta_n) - p^A(x;\delta_n)| \geq \epsilon]\Pr(|p^s(x;\delta_n) - p^A(x;\delta_n)| \geq \epsilon) \\ &\quad + E[|p^s(x;\delta_n) - p^A(x;\delta_n)||\,|p^s(x;\delta_n) - p^A(x;\delta_n)| < \epsilon]\Pr(|p^s(x;\delta_n) - p^A(x;\delta_n)| < \epsilon) \\ &< 1 \cdot \frac{\mathrm{Var}(p^s(x;\delta_n))}{\epsilon^2} + \epsilon \cdot 1 \\ &\leq \frac{1}{S_n\epsilon^2} + \epsilon,\end{aligned}$$

where we use Chebyshev's inequality for the first inequality. We can make $E[|p^s(x;\delta_n) - p^A(x;\delta_n)|]$ arbitrarily close to zero by taking sufficiently small $\epsilon > 0$ and sufficiently large S_n, which implies that $E[|p^s(x;\delta_n) - p^A(x;\delta_n)|] = o(1)$ if $S_n \to \infty$.

□

Lemma B.8. *Let $I_{i,n}^s = 1\{p^s(X_i;\delta_n) \in (0,1)\}$, and let $\{V_i\}_{i=1}^\infty$ be i.i.d. random variables such that $E[V_i^2] < \infty$. If Assumption 1 holds, $S_n \to \infty$, and $\delta_n \to 0$, then*

$$\frac{1}{n}\sum_{i=1}^n V_i p^s(X_i;\delta_n)^l I_{i,n}^s - \frac{1}{n}\sum_{i=1}^n V_i p^A(X_i;\delta_n)^l I_{i,n} = o_p(1)$$

for $l = 0,1,2,3,4$. If, in addition, Assumption 5 holds, and $E[V_i|X_i]$ is bounded, then

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n V_i p^s(X_i;\delta_n)^l I_{i,n}^s - \frac{1}{\sqrt{n}}\sum_{i=1}^n V_i p^A(X_i;\delta_n)^l I_{i,n} = o_p(1)$$

for $l = 0,1,2$.

Proof. We have

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n V_i p^s(X_i;\delta_n)^l I_{i,n}^s - \frac{1}{n}\sum_{i=1}^n V_i p^A(X_i;\delta_n)^l I_{i,n} \\ =&\frac{1}{n}\sum_{i=1}^n V_i p^s(X_i;\delta_n)^l (I_{i,n}^s - I_{i,n}) + \frac{1}{n}\sum_{i=1}^n V_i (p^s(X_i;\delta_n)^l - p^A(X_i;\delta_n)^l) I_{i,n}.\end{aligned}$$

We first consider $\frac{1}{n}\sum_{i=1}^n V_i(p^s(X_i;\delta_n)^l - p^A(X_i;\delta_n)^l)I_{i,n}$. By Lemma B.7, for $l = 0, 1, 2$,

$$\begin{aligned}
&|E[\frac{1}{n}\sum_{i=1}^n V_i(p^s(X_i;\delta_n)^l - p^A(X_i;\delta_n)^l)I_{i,n}]| \\
&= |E[V_i(p^s(X_i;\delta_n)^l - p^A(X_i;\delta_n)^l)I_{i,n}]| \\
&= E[|E[V_i|X_i]||E[p^s(X_i;\delta_n)^l - p^A(X_i;\delta_n)^l|X_i]|I_{i,n}] \\
&\leq \frac{1}{S_n}E[|E[V_i|X_i]|I_{i,n}] \\
&= O(S_n^{-1}).
\end{aligned}$$

Also, by Lemma B.7,

$$\begin{aligned}
&|E[\frac{1}{n}\sum_{i=1}^n V_i(p^s(X_i;\delta_n)^3 - p^A(X_i;\delta_n)^3)I_{i,n}]| \\
&= |E[V_i(p^s(X_i;\delta_n) - p^A(X_i;\delta_n))(p^s(X_i;\delta_n)^2 + p^s(X_i;\delta_n)p^A(X_i;\delta_n) + p^A(X_i;\delta_n)^2)I_{i,n}]| \\
&\leq E[|E[V_i|X_i]||E[(p^s(X_i;\delta_n) - p^A(X_i;\delta_n))(p^s(X_i;\delta_n)^2 + p^s(X_i;\delta_n)p^A(X_i;\delta_n) + p^A(X_i;\delta_n)^2)|X_i]|I_{i,n}] \\
&\leq 3E[|E[V_i|X_i]|E[|p^s(X_i;\delta_n) - p^A(X_i;\delta_n)||X_i]I_{i,n}] \\
&= o(1),
\end{aligned}$$

and

$$\begin{aligned}
&|E[\frac{1}{n}\sum_{i=1}^n V_i(p^s(X_i;\delta_n)^4 - p^A(X_i;\delta_n)^4)I_{i,n}]| \\
&= |E[V_i(p^s(X_i;\delta_n)^2 + p^A(X_i;\delta_n)^2)(p^s(X_i;\delta_n) + p^A(X_i;\delta_n))(p^s(X_i;\delta_n) - p^A(X_i;\delta_n))I_{i,n}]| \\
&\leq E[|E[V_i|X_i]||E[(p^s(X_i;\delta_n)^2 + p^A(X_i;\delta_n)^2)(p^s(X_i;\delta_n) + p^A(X_i;\delta_n))(p^s(X_i;\delta_n) - p^A(X_i;\delta_n))|X_i]|I_{i,n}] \\
&\leq 4E[|E[V_i|X_i]|E[|p^s(X_i;\delta_n) - p^A(X_i;\delta_n)||X_i]I_{i,n}] \\
&= o(1).
\end{aligned}$$

As for variance, for $l = 0, 1, 2$,

$$\begin{aligned}
\mathrm{Var}(\frac{1}{n}\sum_{i=1}^n V_i(p^s(X_i;\delta_n)^l - p^A(X_i;\delta_n)^l)I_{i,n}) &\leq \frac{1}{n}E[V_i^2(p^s(X_i;\delta_n)^l - p^A(X_i;\delta_n)^l)^2 I_{i,n}] \\
&\leq \frac{1}{n}E[E[V_i^2|X_i]E[(p^s(X_i;\delta_n)^l - p^A(X_i;\delta_n)^l)^2|X_i]I_{i,n}] \\
&\leq \frac{4}{nS_n}E[E[V_i^2|X_i]I_{i,n}] \\
&= O((nS_n)^{-1}),
\end{aligned}$$

and for $l = 3, 4$,

$$\begin{aligned}
\mathrm{Var}(\frac{1}{n}\sum_{i=1}^n V_i(p^s(X_i;\delta_n)^l - p^A(X_i;\delta_n)^l)I_{i,n}) &\leq \frac{1}{n}E[V_i^2(p^s(X_i;\delta_n)^l - p^A(X_i;\delta_n)^l)^2 I_{i,n}] \\
&\leq \frac{1}{n}E[V_i^2 I_{i,n}] \\
&= o(1).
\end{aligned}$$

Therefore, $\frac{1}{n}\sum_{i=1}^n V_i(p^s(X_i;\delta_n)^l - p^A(X_i;\delta_n)^l)I_{i,n} = o_p(1)$ if $S_n \to \infty$ for $l = 0,1,2,3,4$, and $\frac{1}{\sqrt{n}}\sum_{i=1}^n V_i(p^s(X_i;\delta_n)^l - p^A(X_i;\delta_n)^l)I_{i,n} = o_p(1)$ if $n^{-1/2}S_n \to \infty$ for $l = 0,1,2$.

We next show that $\frac{1}{n}\sum_{i=1}^n V_i p^s(X_i;\delta_n)^l(I^s_{i,n} - I_{i,n}) = o_p(1)$ if $S_n \to \infty$ and $\delta_n \to 0$ for $l \geq 0$. We have

$$\begin{aligned}
|E[\frac{1}{n}\sum_{i=1}^n V_i p^s(X_i;\delta_n)^l(I^s_{i,n} - I_{i,n})]| &= |E[V_i p^s(X_i;\delta_n)^l(I^s_{i,n} - I_{i,n})]| \\
&\leq E[|E[V_i|X_i]||E[p^s(X_i;\delta_n)^l(I^s_{i,n} - I_{i,n})|X_i]|] \\
&\leq E[|E[V_i|X_i]|E[|I^s_{i,n} - I_{i,n}||X_i]].
\end{aligned}$$

Note that by construction, $1\{p^s(X_i;\delta_n) \in (0,1)\} \leq 1\{p^A(X_i;\delta_n) \in (0,1)\}$ with probability one conditional on $X_i = x$, so that

$$E[|I^s_{i,n} - I_{i,n}||X_i = x] = -E[I^s_{i,n} - I_{i,n}|X_i = x].$$

Suppose A is continuous at x and $A(x) \in (0,1)$. Then $\lim_{\delta\to 0} p^A(x;\delta) = A(x) \in (0,1)$ by Part 1 of Corollary A.1, and hence $p^A(x;\delta_n) \in [\epsilon, 1-\epsilon]$ for sufficiently small $\delta_n > 0$ for some constant $\epsilon \in (0,1/2)$. It follows that

$$\begin{aligned}
E[I^s_{i,n}|X_i = x] &= 1 - \Pr(p^s(x;\delta_n) \in \{0,1\}) \\
&\geq 1 - (1 - p^A(x;\delta_n))^{S_n} - p^A(x;\delta_n)^{S_n} \\
&\geq 1 - 2(1-\epsilon)^{S_n} \\
&\to 1
\end{aligned}$$

as $S_n \to \infty$, where the first inequality follows from Lemma B.7. This implies that $E[I^s_{i,n} - I_{i,n}|X_i = x] \to 0$ as $n \to \infty$. Suppose $x \in \text{int}(\mathcal{X}_0) \cup \text{int}(\mathcal{X}_1)$. Then $B(x,\delta_n) \subset \mathcal{X}_0$ or $B(x,\delta_n) \subset \mathcal{X}_1$ for sufficiently small $\delta_n > 0$ by the fact that $\text{int}(\mathcal{X}_0)$ and $\text{int}(\mathcal{X}_1)$ are open, and hence $p^A(x;\delta_n) \in \{0,1\}$ and $p^s(x;\delta_n) \in \{0,1\}$ for sufficiently small $\delta_n > 0$, so that $E[I^s_{i,n} - I_{i,n}|X_i = x] \to 0$ as $n \to \infty$. Therefore, $E[I^s_{i,n} - I_{i,n}|X_i = x] \to 0$ for almost every $x \in \mathcal{X}$, since A is continuous at x for almost every $x \in \mathcal{X}$ by Assumption 1 (a), and either $A(x) \in (0,1)$ or $x \in \text{int}(\mathcal{X}_0) \cup \text{int}(\mathcal{X}_1)$ for almost every $x \in \mathcal{X}$ by Assumption 1 (b). By the Dominated Convergence Theorem,

$$-E[|E[V_i|X_i]|E[I^s_{i,n} - I_{i,n}|X_i]] \to 0$$

as $n \to \infty$.

As for variance,

$$\begin{aligned}
\text{Var}(\frac{1}{n}\sum_{i=1}^n V_i p^s(X_i;\delta_n)^l(I^s_{i,n} - I_{i,n})) &\leq \frac{1}{n}E[V_i^2 p^s(X_i;\delta_n)^{2l}(I^s_{i,n} - I_{i,n})^2] \\
&\leq \frac{1}{n}E[V_i^2] \\
&\to 0.
\end{aligned}$$

Lastly, we show that, for $l \geq 0$, $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} V_i p^s(X_i;\delta_n)^l (I^s_{i,n} - I_{i,n}) = o_p(1)$ if Assumption 5 holds, and $E[V_i|X_i]$ is bounded. Let $\eta_n = \gamma \frac{\log n}{S_n}$, where γ is the one satisfying Assumption 5. We have

$$
\begin{aligned}
&|E[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} V_i p^s(X_i;\delta_n)^l (I^s_{i,n} - I_{i,n})]| \\
&\leq \sqrt{n} E[|E[V_i|X_i]| E[|I^s_{i,n} - I_{i,n}||X_i]] \\
&= -\sqrt{n} E[|E[V_i|X_i]| E[I^s_{i,n} - 1|X_i] I_{i,n}] \\
&\leq \sqrt{n} E[|E[V_i|X_i]| ((1 - p^A(X_i;\delta_n))^{S_n} + p^A(X_i;\delta_n)^{S_n}) I_{i,n}] \\
&= \sqrt{n} E[|E[V_i|X_i]| ((1 - p^A(X_i;\delta_n))^{S_n} + p^A(X_i;\delta_n)^{S_n}) 1\{p^A(X_i;\delta_n) \in (0,\eta_n) \cup (1-\eta_n, 1)\}] \\
&\quad + \sqrt{n} E[|E[V_i|X_i]| ((1 - p^A(X_i;\delta_n))^{S_n} + p^A(X_i;\delta_n)^{S_n}) 1\{p^A(X_i;\delta_n) \in [\eta_n, 1-\eta_n]\}] \\
&\leq (\sup_{x \in \mathcal{X}} |E[V_i|X_i = x]|)(\sqrt{n} \Pr(p^A(X_i;\delta_n) \in (0,\eta_n) \cup (1-\eta_n, 1)) + 2\sqrt{n}(1-\eta_n)^{S_n}),
\end{aligned}
$$

where the second equality follows from the fact that $I^s_{i,n} \leq I_{i,n}$ with strict inequality only if $I_{i,n} = 1$. By Assumption 5, $\sqrt{n} \Pr(p^A(X_i;\delta_n) \in (0,\eta_n) \cup (1-\eta_n, 1)) = o(1)$. As for $\sqrt{n}(1-\eta_n)^{S_n}$, first observe that $\eta_n = \gamma \frac{\log n}{S_n} = \gamma \frac{\log n}{n^{1/2}} \frac{1}{n^{-1/2} S_n} \to 0$, since $n^{-1/2} S_n \to \infty$ and $\frac{\log n}{n^{1/2}} \to 0$. Using the fact that $e^t \geq 1 + t$ for every $t \in \mathbb{R}$, we have

$$
\begin{aligned}
\sqrt{n}(1-\eta_n)^{S_n} &\leq \sqrt{n}(e^{-\eta_n})^{S_n} \\
&= \sqrt{n} e^{-\eta_n S_n} \\
&= \sqrt{n} e^{-\gamma \log n} \\
&= \sqrt{n} n^{-\gamma} \\
&= n^{1/2-\gamma} \\
&\to 0,
\end{aligned}
$$

since $\gamma > 1/2$. As for variance,

$$
\begin{aligned}
\mathrm{Var}(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} V_i p^s(X_i;\delta_n)^l (I^s_{i,n} - I_{i,n})) &\leq E[V_i^2 p^s(X_i;\delta_n)^{2l} (I^s_{i,n} - I_{i,n})^2] \\
&\leq E[V_i^2 |I^s_{i,n} - I_{i,n}|] \\
&= E[E[V_i^2|X_i] E[|I^s_{i,n} - I_{i,n}||X_i]] \\
&= o(1).
\end{aligned}
$$

□

C Proofs

C.1 Proof of Proposition 1

Suppose that Assumptions 1 and 2 hold. Here, we only show that

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

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

The results for $E[D_i(1) - D_i(0)|X_i = x]$ and $E[D_i(1) - D_i(0)|X_i \in S]$ are obtained by a similar argument.

Proof of Part (a). Pick an $x \in \text{int}(\mathcal{X})$ such that $p^A(x) \in (0,1)$. If $A(x) \in (0,1)$, $E[Y_{1i} - Y_{0i}|X_i = x]$ and $E[D_i(1) - D_i(0)|X_i = x]$ are trivially identified by Property 1:

$$E[Y_i|X_i = x, Z_i = 1] - E[Y_i|X_i = x, Z_i = 0] = E[Y_{1i} - Y_{0i}|X_i = x].$$

We next consider the case where $A(x) \in \{0,1\}$. Since $x \in \text{int}(\mathcal{X})$, $B(x,\delta) \subset \mathcal{X}$ for any sufficiently small $\delta > 0$. Moreover, since $p^A(x) = \lim_{\delta\to 0} p^A(x;\delta) \in (0,1)$, $p^A(x;\delta) \in (0,1)$ for any sufficiently small $\delta > 0$. This implies that we can find points $x_{0,\delta}, x_{1,\delta} \in B(x,\delta)(\subset \mathcal{X})$ such that $A(x_{0,\delta}) < 1$ and $A(x_{1,\delta}) > 0$ for any sufficiently small $\delta > 0$, for otherwise $p^A(x;\delta) \in \{0,1\}$. Noting that $x_{0,\delta} \to x$ and $x_{1,\delta} \to x$ as $\delta \to 0$,

$$\begin{aligned}\lim_{\delta\to 0}(E[Y_i|X_i = x_{1,\delta}, Z_i = 1] - E[Y_i|X_i = x_{0,\delta}, Z_i = 0]) &= \lim_{\delta\to 0}(E[Y_{i1}|X_i = x_{1,\delta}] - E[Y_{i0}|X_i = x_{0,\delta}]) \\ &= E[Y_{1i} - Y_{0i}|X_i = x],\end{aligned}$$

where the first equality follows from Property 1, and the second from Assumption 2. □

Proof of Part (b).

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

Step C.1.1. $\mathcal{L}^p(S \cap \mathcal{X}_1) > 0$.

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

Step C.1.2. $S \cap \text{int}(\mathcal{X}_1) \neq \emptyset$.

Proof. Suppose that $S \cap \text{int}(\mathcal{X}_1) = \emptyset$. Then, we must have that $S \cap \mathcal{X}_1 \subset \mathcal{X}_1 \setminus \text{int}(\mathcal{X}_1)$. It then follows that $\mathcal{L}^p(S \cap \mathcal{X}_1) \leq \mathcal{L}^p(\mathcal{X}_1 \setminus \text{int}(\mathcal{X}_1)) = \mathcal{L}^p(\mathcal{X}_1) - \mathcal{L}^p(\text{int}(\mathcal{X}_1)) = 0$, where the last equality holds by Assumption 1. But this is a contradiction to the result from Step C.1.1. □

Step C.1.3. $p^A(x) = 1$ *for any* $x \in \text{int}(\mathcal{X}_1)$.

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

Step C.1.4. $E[Y_{1i} - Y_{0i}|X_i \in S]$ *is not identified.*

Proof. We first introduce some notation. Let $\mathbf{Q}$ be the set of all distributions of $(Y_{1i}, Y_{0i}, X_i, Z_i)$ satisfying Property 1 and Assumptions 1 and 2. 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 $E[Y_{1i} - Y_{0i}|X_i \in S]$ is given by $\{E_Q[Y_{1i} - Y_{0i}|X_i \in S] : P_0 = T(Q), Q \in \mathbf{Q}\}$, 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 $x^* \in S \cap \text{int}(\mathcal{X}_1)$. Since A and $\text{int}(\mathcal{X}_1)$ are open, there is some $\delta > 0$ such that $B(x^*, \delta) \subset S \cap \text{int}(\mathcal{X}_1)$. 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 $x \in \mathcal{X} \setminus B(x^*, \epsilon)$ and $f(x) = E[Y_{0i}|X = x] - 1$ for all $x \in B(x^*, \epsilon)$. Below, we show that f is continuous at any point $x \in \mathcal{X}$ such that $p^A(x) \in (0, 1)$ and $A(x) \in \{0, 1\}$. Pick any $x \in \mathcal{X}$ such that $p^A(x) \in (0, 1)$ and $A(x) \in \{0, 1\}$. Since $B(x^*, \delta) \subset \text{int}(\mathcal{X}_1)$ and $\text{int}(\mathcal{X}_1) \subset \{x' \in \mathcal{X} : p^A(x') = 1\}$ by Step C.1.3, $x \notin B(x^*, \delta)$. Hence, $B(x, \epsilon) \subset \mathcal{X} \setminus B(x^*, \epsilon)$. By Assumption 2 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

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

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 $p^A(x) \in (0, 1)$ and $A(x) \in \{0, 1\}$. Also, $Z_i^Q = Z_i^* = 1$ if $X_i^* \in B(x^*, \epsilon)$. It then follows that

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

and

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

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

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

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

□

□

C.2 Proof of Corollary 1

If $\mathrm{Pr}(D_i(1) - D_i(0) = 1|X_i = x) = 1$, $\mathrm{Pr}(Y_{1i} - Y_{0i} = Y_i(1) - Y_i(0)|X_i = x) = 1$, and hence $E[Y_{1i} - Y_{0i}|X_i = x] = E[Y_i(1) - Y_i(0)|X_i = x]$. Then, Part (a) follows from Proposition 1 (a). If $\mathrm{Pr}(D_i(1) \geq D_i(0)|X_i = x) = 1$, we have

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

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

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

Then, Part (b) follows from Proposition 1 (a). □

C.3 Proof of Theorem 1

We prove consistency and asymptotic normality of the following estimators without imposing Assumption 3 (c). These estimators are aymptotically equivalent to the estimators defined in Section 4.1 if Assumption 3 (c) holds.

First, consider the following 2SLS regression using the observations with $p^A(X_i;\delta_n) \in (0,1)$:

$$D_i = \gamma_0(1 - \mathbf{I}_n) + \gamma_1 Z_i + \gamma_2 p^A(X_i;\delta_n) + \nu_i \tag{15}$$

$$Y_i = \beta_0(1 - \mathbf{I}_n) + \beta_1 D_i + \beta_2 p^A(X_i;\delta_n) + \epsilon_i. \tag{16}$$

Here $\mathbf{I}_n$ is a dummy random variable which equals one if there exists a constant $q \in (0,1)$ such that $A(X_i) \in \{0, q, 1\}$ for all $i \in \{1, ..., n\}$. $\mathbf{I}_n$ is the indicator that $A(X_i)$ takes on only one nondegenerate value *in the sample*. If the support of $A(X_i)$ (in the population) contains only one value in $(0,1)$, $p^A(X_i;\delta_n)$ is asymptotically constant conditional on $p^A(X_i;\delta_n) \in (0,1)$. To avoid the multicollinearity between asymptotically constant $p^A(X_i;\delta_n)$ and a constant, we do not include the constant term if $\mathbf{I}_n = 1$. Let $I_{i,n} = 1\{p^A(X_i;\delta_n) \in (0,1)\}$, $\mathbf{D}_{i,n} = (1, D_i, p^A(X_i;\delta_n))'$, $\mathbf{Z}_{i,n} = (1, Z_i, p^A(X_i;\delta_n))'$, $\mathbf{D}^{nc}_{i,n} = (D_i, p^A(X_i;\delta_n))'$, and $\mathbf{Z}^{nc}_{i,n} = (Z_i, p^A(X_i;\delta_n))'$. The 2SLS estimator $\hat{\beta}$ from this regression is then given by

$$\hat{\beta} = \begin{cases} (\sum_{i=1}^n \mathbf{Z}_{i,n}\mathbf{D}'_{i,n}I_{i,n})^{-1}\sum_{i=1}^n \mathbf{Z}_{i,n}Y_iI_{i,n} & \text{if } \mathbf{I}_n{=}0 \\ (\sum_{i=1}^n \mathbf{Z}^{nc}_{i,n}(\mathbf{D}^{nc}_{i,n})'I_{i,n})^{-1}\sum_{i=1}^n \mathbf{Z}^{nc}_{i,n}Y_iI_{i,n} & \text{if } \mathbf{I}_n{=}1. \end{cases}$$

Let $\hat{\beta}_1$ denote the 2SLS estimator of β_1 in the above regression.

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

$$D_i = \gamma_0(1 - \mathbf{I}_n) + \gamma_1 Z_i + \gamma_2 p^s(X_i;\delta_n) + \nu_i \tag{17}$$

$$Y_i = \beta_0(1 - \mathbf{I}_n) + \beta_1 D_i + \beta_2 p^s(X_i;\delta_n) + \epsilon_i. \tag{18}$$

Let $\hat{\beta}^s_1$ denote the 2SLS estimator of β_1 in the simulation-based regression.

Below, we prove the following result.

Theorem C.1. *Suppose that Assumptions 1 and 3 hold except Assumption 3 (c), 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 $\hat{\beta}^s_1$ converge in probability to*

$$\beta_1 \equiv \lim_{\delta \to 0} E[\omega_i(\delta)(Y_i(1) - Y_i(0))],$$

where

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

Suppose, in addition, that Assumptions 4 and 5 hold and that $n\delta_n^2 \to 0$ as $n \to \infty$. Then

$$\hat{\sigma}_n^{-1}(\hat{\beta}_1 - \beta_1) \xrightarrow{d} \mathcal{N}(0,1),$$

$$(\hat{\sigma}^s_n)^{-1}(\hat{\beta}^s_1 - \beta_1) \xrightarrow{d} \mathcal{N}(0,1).$$

where we define $\hat{\sigma}_n^{-1}$ and $(\hat{\sigma}^s_n)^{-1}$ as follows: let

$$\hat{\mathbf{\Sigma}}_n = \begin{cases} (\sum_{i=1}^n \mathbf{Z}_{i,n}\mathbf{D}'_{i,n}I_{i,n})^{-1}(\sum_{i=1}^n \hat{\epsilon}^2_{i,n}\mathbf{Z}_{i,n}\mathbf{Z}'_{i,n}I_{i,n})(\sum_{i=1}^n \mathbf{D}_{i,n}\mathbf{Z}'_{i,n}I_{i,n})^{-1} & \text{if } \mathbf{I}_n = 0 \\ (\sum_{i=1}^n \mathbf{Z}^{nc}_{i,n}(\mathbf{D}^{nc}_{i,n})'I_{i,n})^{-1}(\sum_{i=1}^n \hat{\epsilon}^2_{i,n}\mathbf{Z}^{nc}_{i,n}(\mathbf{Z}^{nc}_{i,n})'I_{i,n})(\sum_{i=1}^n \mathbf{D}^{nc}_{i,n}(\mathbf{Z}^{nc}_{i,n})'I_{i,n})^{-1} & \text{if } \mathbf{I}_n = 1, \end{cases}$$

where

$$\hat{\epsilon}_{i,n} = \begin{cases} Y_i - \mathbf{D}'_{i,n}\hat{\beta} & \text{if } \mathbf{I}_n = 0 \\ Y_i - (\mathbf{D}^{nc}_{i,n})'\hat{\beta} & \text{if } \mathbf{I}_n = 1. \end{cases}$$

Let $\hat{\sigma}^2_n$ denote the estimator for the variance of $\hat{\beta}_1$. That is, $\hat{\sigma}^2_n$ is the second diagonal element of $\hat{\mathbf{\Sigma}}_n$ when $\mathbf{I}_n = 0$ and is the first diagonal element of $\hat{\mathbf{\Sigma}}_n$ when $\mathbf{I}_n = 1$. $(\hat{\sigma}^s_n)^2$ is the analogously-defined estimator for the variance of $\hat{\beta}^s_1$ from the simulation-based regression.

Throughout the proof, we omit the subscript n from $I_{i,n}$, $\mathbf{D}_{i,n}$, $\mathbf{Z}_{i,n}$, $\hat{\epsilon}_{i,n}$, $\hat{\mathbf{\Sigma}}_n$, $\hat{\sigma}_n$, etc. for notational brevity. We provide proofs separately for the two cases, the case in which $\Pr(A(X_i) \in (0,1)) > 0$ and the case in which $\Pr(A(X_i) \in (0,1)) = 0$. For each case, we first prove consistency and asymptotic normality of $\hat{\beta}_1$, and then prove consistency and asymptotic normality of $\hat{\beta}_1^s$.

C.3.1 Consistency and Asymptotic Normality of $\hat{\beta}_1$ When $\Pr(A(X_i) \in (0,1)) > 0$

By Lemma B.6,

$$\lim_{\delta \to 0} E[p^A(X_i;\delta)(1-p^A(X_i;\delta))(D_i(1)-D_i(0))] = E[A(X_i)(1-A(X_i))(D_i(1)-D_i(0))].$$

When $\Pr(A(X_i) \in (0,1)) > 0$, $E[A(X_i)(1-A(X_i))(D_i(1)-D_i(0))] = E[p^A(X_i)(1-p^A(X_i))(D_i(1)-D_i(0))]$, since $p^A(x) = A(x)$ for almost every $x \in \mathcal{X}$ by Proposition A.1. Under Assumption 3 (b), $E[p^A(X_i)(1-p^A(X_i))(D_i(1)-D_i(0))] > 0$. Again by Lemma B.6,

$$\lim_{\delta \to 0} E[\omega_i(\delta)(Y_i(1)-Y_i(0))] = \frac{E[A(X_i)(1-A(X_i))(D_i(1)-D_i(0))(Y_i(1)-Y_i(0)]}{E[A(X_i)(1-A(X_i))(D_i(1)-D_i(0))]}.$$

Let $\beta_1 = \frac{E[A(X_i)(1-A(X_i))(D_i(1)-D_i(0))(Y_i(1)-Y_i(0)]}{E[A(X_i)(1-A(X_i))(D_i(1)-D_i(0))]}$. Let

$$\hat{\beta}^c = (\sum_{i=1}^n \mathbf{Z}_i \mathbf{D}_i' I_i)^{-1} \sum_{i=1}^n \mathbf{Z}_i Y_i I_i$$
$$\hat{\beta}^{nc} = (\sum_{i=1}^n \mathbf{Z}_i^{nc} (\mathbf{D}_i^{nc})' I_i)^{-1} \sum_{i=1}^n \mathbf{Z}_i^{nc} Y_i I_i,$$

and let $\hat{\beta}_1^c = (0,1,0)\hat{\beta}^c$ and $\hat{\beta}_1^{nc} = (1,0)\hat{\beta}^{nc}$. $\hat{\beta}_1$ is given by

$$\hat{\beta}_1 = \hat{\beta}_1^c(1-\mathbf{I}_n) + \hat{\beta}_1^{nc}\mathbf{I}_n.$$

Also, let $\tilde{\mathbf{D}}_i = (1, D_i, A(X_i))'$, $\tilde{\mathbf{Z}}_i = (1, Z_i, A(X_i))'$, $\tilde{\mathbf{D}}_i^{nc} = (D_i, A(X_i))'$, $\tilde{\mathbf{Z}}_i^{nc} = (Z_i, A(X_i))'$, and $I_i^A = 1\{A(X_i) \in (0,1)\}$.

We claim that $\Pr(\mathbf{I}_n = 1) \to 0$ when $\mathrm{Var}(A(X_i)|I_i^A = 1) > 0$, and that $\Pr(\mathbf{I}_n = 1) \to 1$ when $\mathrm{Var}(A(X_i)|I_i^A = 1) = 0$. To show the first claim, observe that $\mathbf{I}_n = 1$ if and only if $\hat{V}_n = 0$, where

$$\hat{V}_n = \frac{\sum_{i=1}^n (A(X_i) - \frac{\sum_{i=1}^n A(X_i) I_i^A}{\sum_{i=1}^n I_i^A})^2 I_i^A}{\sum_{i=1}^n I_i^A}$$

is the sample variance of $A(X_i)$ conditional on $I_i^A = 1$. When $\mathrm{Var}(A(X_i)|I_i^A = 1) > 0$,

$$\begin{aligned}\Pr(\mathbf{I}_n = 1) &= \Pr(\hat{V}_n = 0) \\ &\leq \Pr(|\hat{V}_n - \mathrm{Var}(A(X_i)|I_i^A = 1)| \geq \mathrm{Var}(A(X_i)|I_i^A = 1)) \\ &\to 0,\end{aligned}$$

where the convergence follows since $\hat{V}_n \xrightarrow{p} \mathrm{Var}(A(X_i)|I_i^A = 1) > 0$.

To show the second claim, note that, when $\text{Var}(A(X_i)|I_i^A = 1) = 0$, there exists $q \in (0,1)$ such that $\Pr(A(X_i) = q|I_i^A = 1) = 1$. It follows that

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

which converges to zero as $n \to \infty$, since $\Pr(A(X_i) \in (0,1)) > 0$.

The above claims imply that $\hat{\beta}_1 = \hat{\beta}_1^c$ with probability approaching one when $\text{Var}(A(X_i)|I_i^A = 1) > 0$, and that $\hat{\beta}_1 = \hat{\beta}_1^{nc}$ with probability approaching one when $\text{Var}(A(X_i)|I_i^A = 1) = 0$. Therefore, to prove consistency and asymptotic normality of $\hat{\beta}_1$, it suffices to show those of $\hat{\beta}_1^c$ when $\text{Var}(A(X_i)|I_i^A = 1) > 0$ and those of $\hat{\beta}_1^{nc}$ when $\text{Var}(A(X_i)|I_i^A = 1) = 0$.

Below we first show that, if Assumptions 1 and 3 hold and $\delta_n \to 0$ as $n \to \infty$, then $\hat{\beta}_1 \xrightarrow{p} \beta_1$. We then show that, if, in addition, Assumption 4 holds and $n\delta_n^2 \to 0$ as $n \to \infty$, then $\hat{\sigma}^{-1}(\hat{\beta}_1 - \beta_1) \xrightarrow{d} \mathcal{N}(0,1)$.

Proof of Consistency. To prove consistency of $\hat{\beta}_1$, we first show that $\hat{\beta}_1^c \xrightarrow{p} \beta_1$ when $\text{Var}(A(X_i)|I_i^A = 1) > 0$. We then show that $\hat{\beta}_1^{nc} \xrightarrow{p} \beta_1$ whether or not $\text{Var}(A(X_i)|I_i^A = 1) > 0$. By Lemma B.6,

$$\hat{\beta}^c = (\sum_{i=1}^n \mathbf{Z}_i \mathbf{D}_i' I_i)^{-1} \sum_{i=1}^n \mathbf{Z}_i Y_i I_i \xrightarrow{p} (E[\tilde{\mathbf{Z}}_i \tilde{\mathbf{D}}_i' I_i^A])^{-1} E[\tilde{\mathbf{Z}}_i Y_i I_i^A]$$

provided that $E[\tilde{\mathbf{Z}}_i \tilde{\mathbf{D}}_i' I_i^A]$ is invertible. After a few lines of algebra, we have

$$\begin{aligned}&\det(E[\tilde{\mathbf{Z}}_i \tilde{\mathbf{D}}_i' I_i^A]) \\ =& \Pr(I_i^A = 1)^2 \text{Var}(A(X_i)|I_i^A = 1) E[D_i(Z_i - A(X_i))I_i^A] \\ =& \Pr(I_i^A = 1)^2 \text{Var}(A(X_i)|I_i^A = 1) E[(Z_i D_i(1) + (1 - Z_i)D_i(0))(Z_i - A(X_i))I_i^A] \\ =& \Pr(I_i^A = 1)^2 \text{Var}(A(X_i)|I_i^A = 1) E[((Z_i - Z_i A(X_i))D_i(1) - (1 - Z_i)A(X_i)D_i(0))I_i^A] \\ =& \Pr(I_i^A = 1)^2 \text{Var}(A(X_i)|I_i^A = 1) E[((A(X_i) - A(X_i)^2)D_i(1) - (1 - A(X_i))A(X_i)D_i(0))I_i^A] \\ =& \Pr(I_i^A = 1)^2 \text{Var}(A(X_i)|I_i^A = 1) E[A(X_i)(1 - A(X_i))(D_i(1) - D_i(0))I_i^A] \\ =& \Pr(I_i^A = 1)^2 \text{Var}(A(X_i)|I_i^A = 1) E[A(X_i)(1 - A(X_i))(D_i(1) - D_i(0))],\end{aligned}$$

where the fourth equality follows from Property 1. Therefore, $E[\tilde{\mathbf{Z}}_i \tilde{\mathbf{D}}_i' I_i^A]$ is invertible when $\text{Var}(A(X_i)|I_i^A = 1) > 0$. Another few lines of algebra gives

$$(E[\tilde{\mathbf{Z}}_i \tilde{\mathbf{D}}_i' I_i^A])^{-1} = \frac{1}{E[A(X_i)(1 - A(X_i))(D_i(1) - D_i(0))]} \begin{bmatrix} * & * & * \\ 0 & 1 & -1 \\ * & * & * \end{bmatrix}$$

when $\text{Var}(A(X_i)|I_i^A = 1) > 0$. Therefore, when $\text{Var}(A(X_i)|I_i^A = 1) > 0$,

$$\begin{aligned}
\hat{\beta}_1^c &\xrightarrow{p} \frac{E[Z_i Y_i I_i^A] - E[A(X_i) Y_i I_i^A]}{E[A(X_i)(1 - A(X_i))(D_i(1) - D_i(0))]} \\
&= \frac{E[Z_i Y_{1i} I_i^A] - E[A(X_i)(Z_i Y_{1i} + (1 - Z_i) Y_{0i}) I_i^A]}{E[A(X_i)(1 - A(X_i))(D_i(1) - D_i(0))]} \\
&= \frac{E[A(X_i) Y_{1i} I_i^A] - E[A(X_i)(A(X_i) Y_{1i} + (1 - A(X_i)) Y_{0i}) I_i^A]}{E[A(X_i)(1 - A(X_i))(D_i(1) - D_i(0))]} \\
&= \frac{E[A(X_i)(1 - A(X_i))(Y_{1i} - Y_{0i}) I_i^A]}{E[A(X_i)(1 - A(X_i))(D_i(1) - D_i(0))]} \\
&= \frac{E[A(X_i)(1 - A(X_i))((D_i(1) - D_i(0))(Y_i(1) - Y_i(0))]}{E[A(X_i)(1 - A(X_i))(D_i(1) - D_i(0))]} \\
&= \beta_1,
\end{aligned}$$

where the third line follows from Property 1, and the second last follows from the definitions of Y_{1i} and Y_{0i}.

We next consider $\hat{\beta}_1^{nc}$. By Lemma B.6,

$$\hat{\beta}^{nc} = (\sum_{i=1}^n \mathbf{Z}_i^{nc} (\mathbf{D}_i^{nc})' I_i)^{-1} \sum_{i=1}^n \mathbf{Z}_i^{nc} Y_i I_i \xrightarrow{p} (E[\tilde{\mathbf{Z}}_i^{nc} (\tilde{\mathbf{D}}_i^{nc})' I_i^A])^{-1} E[\tilde{\mathbf{Z}}_i^{nc} Y_i I_i^A]$$

provided that $E[\tilde{\mathbf{Z}}_i^{nc} (\tilde{\mathbf{D}}_i^{nc})' I_i^A]$ is invertible. After a few lines of algebra, we have

$$\begin{aligned}
\det(E[\tilde{\mathbf{Z}}_i^{nc} (\tilde{\mathbf{D}}_i^{nc})' I_i^A]) &= E[A(X_i)^2 I_i^A] E[D_i(Z_i - A(X_i)) I_i^A] \\
&= E[A(X_i)^2 I_i^A] E[A(X_i)(1 - A(X_i))(D_i(1) - D_i(0))] \\
&> 0.
\end{aligned}$$

Another few lines of algebra gives

$$(E[\tilde{\mathbf{Z}}_i^{nc} (\tilde{\mathbf{D}}_i^{nc})' I_i^A])^{-1} = \frac{1}{E[A(X_i)(1 - A(X_i))(D_i(1) - D_i(0))]} \begin{bmatrix} 1 & -1 \\ * & * \end{bmatrix}.$$

Therefore,

$$\hat{\beta}_1^{nc} \xrightarrow{p} \frac{E[Z_i Y_i I_i^A] - E[A(X_i) Y_i I_i^A]}{E[A(X_i)(1 - A(X_i))(D_i(1) - D_i(0))]} = \beta_1.$$

□

Proof of Asymptotic Normality. Let $(\hat{\sigma}^c)^2$ be the second diagonal element of

$$\hat{\mathbf{\Sigma}}^c = (\sum_{i=1}^n \mathbf{Z}_i \mathbf{D}_i' I_i)^{-1} (\sum_{i=1}^n \hat{\epsilon}_i^2 \mathbf{Z}_i \mathbf{Z}_i' I_i) (\sum_{i=1}^n \mathbf{D}_i \mathbf{Z}_i' I_i)^{-1}$$

and $(\hat{\sigma}^{nc})^2$ be the first diagonal element of

$$\hat{\mathbf{\Sigma}}^{nc} = (\sum_{i=1}^n \mathbf{Z}_{i,n}^{nc} (\mathbf{D}_{i,n}^{nc})' I_i)^{-1} (\sum_{i=1}^n \hat{\epsilon}_{i,n}^2 \mathbf{Z}_{i,n}^{nc} (\mathbf{Z}_{i,n}^{nc})' I_i) (\sum_{i=1}^n \mathbf{D}_{i,n}^{nc} (\mathbf{Z}_{i,n}^{nc})' I_i)^{-1}.$$

We only show that $(\hat{\sigma}^c)^{-1}(\hat{\beta}_1^c - \beta_1) \xrightarrow{d} \mathcal{N}(0,1)$ when $\text{Var}(A(X_i)|I_i^A = 1) > 0$. We can show that $(\hat{\sigma}^{nc})^{-1}(\hat{\beta}_1^{nc} - \beta_1) \xrightarrow{d} \mathcal{N}(0,1)$ by an analogous argument. The proof proceeds in six steps.

Step C.3.1.1. *Let $\tilde{\beta}_n = (E[\tilde{\mathbf{Z}}_i\tilde{\mathbf{D}}_i' I_i])^{-1}E[\tilde{\mathbf{Z}}_i Y_i I_i]$, and let $\tilde{\beta}_{1,n}$ denote the second element of $\tilde{\beta}_n$. Then $\tilde{\beta}_{1,n} = \beta_1$ for any choice of $\delta_n > 0$.*

Proof. Note first that, for every $\delta > 0$, $p^A(x;\delta) \in (0,1)$ for almost every $x \in \{x' \in \mathcal{X} : A(x') \in (0,1)\}$, since by almost everywhere continuity of A, for almost every $x \in \{x' \in \mathcal{X} : A(x') \in (0,1)\}$, there exists an open ball $B \subset B(x,\delta)$ such that $A(x') \in (0,1)$ for every $x' \in B$. After a few lines of algebra, we have

$$\begin{aligned}
\det(E[\tilde{\mathbf{Z}}_i\tilde{\mathbf{D}}_i' I_i]) &= \Pr(I_i = 1)^2 \text{Var}(A(X_i)|I_i = 1) E[D_i(Z_i - A(X_i))I_i] \\
&= \Pr(I_i = 1)^2 \text{Var}(A(X_i)|I_i = 1) E[A(X_i)(1 - A(X_i))(D_i(1) - D_i(0))I_i] \\
&= \Pr(I_i = 1)^2 \text{Var}(A(X_i)|I_i = 1) E[A(X_i)(1 - A(X_i))(D_i(1) - D_i(0))],
\end{aligned}$$

where the last equality holds since $p^A(x;\delta) \in (0,1)$ for almost every $x \in \{x' \in \mathcal{X} : A(x') \in (0,1)\}$. By the law of total conditional variance,

$$\begin{aligned}
&\text{Var}(A(X_i)|I_i = 1) \\
&= E[\text{Var}(A(X_i)|I_i = 1, I_i^A)|I_i = 1] + \text{Var}(E[A(X_i)|I_i = 1, I_i^A]|I_i = 1) \\
&\geq \sum_{t \in \{0,1\}} \text{Var}(A(X_i)|I_i = 1, I_i^A = t) \Pr(I_i^A = t|I_i = 1) \\
&\geq \text{Var}(A(X_i)|I_i = 1, I_i^A = 1) \Pr(I_i^A = 1|I_i = 1) \\
&= \text{Var}(A(X_i)|I_i^A = 1) \Pr(I_i^A = 1|I_i = 1) \\
&> 0.
\end{aligned}$$

Therefore, $E[\tilde{\mathbf{Z}}_i\tilde{\mathbf{D}}_i' I_i]$ is invertible. Another few lines of algebra gives

$$(E[\tilde{\mathbf{Z}}_i\tilde{\mathbf{D}}_i' I_i])^{-1} = \frac{1}{E[A(X_i)(1 - A(X_i))(D_i(1) - D_i(0))]} \begin{bmatrix} * & * & * \\ 0 & 1 & -1 \\ * & * & * \end{bmatrix}.$$

It follows that

$$\begin{aligned}
\tilde{\beta}_{1,n} &= \frac{E[Z_i Y_i I_i] - E[A(X_i) Y_i I_i]}{E[A(X_i)(1 - A(X_i))(D_i(1) - D_i(0))]} \\
&= \frac{E[A(X_i)(1 - A(X_i))(D_i(1) - D_i(0))(Y_i(1) - Y_i(0))I_i]}{E[A(X_i)(1 - A(X_i))(D_i(1) - D_i(0))]} \\
&= \frac{E[A(X_i)(1 - A(X_i))(D_i(1) - D_i(0))(Y_i(1) - Y_i(0))]}{E[A(X_i)(1 - A(X_i))(D_i(1) - D_i(0))]} \\
&= \beta_1.
\end{aligned}$$

□

We can write

$$\sqrt{n}(\hat{\beta}^c - \tilde{\beta}_n) = \underbrace{(\frac{1}{n}\sum_{i=1}^n \mathbf{Z}_i\mathbf{D}_i' I_i)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{Z}_i Y_i I_i - (\frac{1}{n}\sum_{i=1}^n \tilde{\mathbf{Z}}_i\tilde{\mathbf{D}}_i' I_i)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde{\mathbf{Z}}_i Y_i I_i}_{=(A)}$$
$$+ \underbrace{(\frac{1}{n}\sum_{i=1}^n \tilde{\mathbf{Z}}_i\tilde{\mathbf{D}}_i' I_i)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde{\mathbf{Z}}_i Y_i I_i - (E[\tilde{\mathbf{Z}}_i\tilde{\mathbf{D}}_i' I_i])^{-1}\sqrt{n}E[\tilde{\mathbf{Z}}_i Y_i I_i]}_{=(B)}.$$

We first consider (B). Let $\tilde{\epsilon}_{i,n} = Y_i - \tilde{\mathbf{D}}_i'\tilde{\beta}_n$ so that

$$E[\tilde{\mathbf{Z}}_i\tilde{\epsilon}_{i,n}I_i] = E[\tilde{\mathbf{Z}}_i(Y_i - \tilde{\mathbf{D}}_i'\tilde{\beta}_n)I_i] = E[\tilde{\mathbf{Z}}_i Y_i I_i] - E[\tilde{\mathbf{Z}}_i\tilde{\mathbf{D}}_i' I_i]\tilde{\beta}_n = 0.$$

Then

$$\begin{aligned}(B) &= (\frac{1}{n}\sum_{i=1}^n \tilde{\mathbf{Z}}_i\tilde{\mathbf{D}}_i' I_i)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde{\mathbf{Z}}_i(\tilde{\mathbf{D}}_i'\tilde{\beta}_n + \tilde{\epsilon}_{i,n})I_i - (E[\tilde{\mathbf{Z}}_i\tilde{\mathbf{D}}_i' I_i])^{-1}\sqrt{n}E[\tilde{\mathbf{Z}}_i(\tilde{\mathbf{D}}_i'\tilde{\beta}_n + \tilde{\epsilon}_{i,n})I_i] \\ &= \sqrt{n}(\tilde{\beta}_n - \tilde{\beta}_n) + (\frac{1}{n}\sum_{i=1}^n \tilde{\mathbf{Z}}_i\tilde{\mathbf{D}}_i' I_i)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde{\mathbf{Z}}_i\tilde{\epsilon}_{i,n}I_i - (E[\tilde{\mathbf{Z}}_i\tilde{\mathbf{D}}_i' I_i])^{-1}\sqrt{n}E[\tilde{\mathbf{Z}}_i\tilde{\epsilon}_{i,n}I_i] \\ &= (\frac{1}{n}\sum_{i=1}^n \tilde{\mathbf{Z}}_i\tilde{\mathbf{D}}_i' I_i)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde{\mathbf{Z}}_i\tilde{\epsilon}_{i,n}I_i.\end{aligned}$$

Step C.3.1.2.

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde{\mathbf{Z}}_i\tilde{\epsilon}_{i,n}I_i \xrightarrow{d} \mathcal{N}(0, E[\tilde{\epsilon}_i^2\tilde{\mathbf{Z}}_i\tilde{\mathbf{Z}}_i' I_i^A]).$$

Proof. We use the triangular-array Lyapunov CLT and the Cramér-Wold device. Pick a nonzero $\lambda \in \mathbb{R}^p$, and let $V_{i,n} = \frac{1}{\sqrt{n}}\lambda'\tilde{\mathbf{Z}}_i\tilde{\epsilon}_{i,n}I_i$. First, we have

$$\sum_{i=1}^n E[V_{i,n}^2] = \lambda' E[\tilde{\epsilon}_{i,n}^2\tilde{\mathbf{Z}}_i\tilde{\mathbf{Z}}_i' I_i]\lambda.$$

By Lemma B.6,

$$\tilde{\beta}_n \to (E[\tilde{\mathbf{Z}}_i\tilde{\mathbf{D}}_i' I_i^A])^{-1}E[\tilde{\mathbf{Z}}_i Y_i I_i^A]$$

as $n \to \infty$. Let $\beta = (E[\tilde{\mathbf{Z}}_i\tilde{\mathbf{D}}_i' I_i^A])^{-1}E[\tilde{\mathbf{Z}}_i Y_i I_i^A]$ and $\tilde{\epsilon}_i = Y_i - \tilde{\mathbf{D}}_i'\beta$. We have

$$\begin{aligned}E[\tilde{\epsilon}_{i,n}^2\tilde{\mathbf{Z}}_i\tilde{\mathbf{Z}}_i' I_i] &= E[(Y_i - \tilde{\mathbf{D}}_i'\tilde{\beta}_n)^2\tilde{\mathbf{Z}}_i\tilde{\mathbf{Z}}_i' I_i] \\ &= E[(\tilde{\epsilon}_i - \tilde{\mathbf{D}}_i'(\tilde{\beta}_n - \beta))^2\tilde{\mathbf{Z}}_i\tilde{\mathbf{Z}}_i' I_i] \\ &= E[\tilde{\epsilon}_i^2\tilde{\mathbf{Z}}_i\tilde{\mathbf{Z}}_i' I_i] - 2E[\tilde{\epsilon}_i((\tilde{\beta}_{0,n} - \beta_0) + D_i(\tilde{\beta}_{1,n} - \beta_1) + A(X_i)(\tilde{\beta}_{2,n} - \beta_2))\tilde{\mathbf{Z}}_i\tilde{\mathbf{Z}}_i' I_i] \\ &\quad + E[((\tilde{\beta}_{0,n} - \beta_0) + D_i(\tilde{\beta}_{1,n} - \beta_1) + A(X_i)(\tilde{\beta}_{2,n} - \beta_2))^2\tilde{\mathbf{Z}}_i\tilde{\mathbf{Z}}_i' I_i] \\ &\to E[\tilde{\epsilon}_i^2\tilde{\mathbf{Z}}_i\tilde{\mathbf{Z}}_i' I_i^A]\end{aligned}$$

as $n \to \infty$, where the convergence follows from Lemma B.6 and from the fact that $\tilde{\beta}_n \to \beta$. Therefore,

$$\sum_{i=1}^{n} E[V_{i,n}^2] \to \lambda' E[\tilde{\epsilon}_i^2 \tilde{\mathbf{Z}}_i \tilde{\mathbf{Z}}_i' I_i^A]\lambda.$$

We next verify the Lyapunov condition: for some $t > 0$,

$$\sum_{i=1}^{n} E[|V_{i,n}|^{2+t}] \to 0.$$

We have

$$\sum_{i=1}^{n} E[|V_{i,n}|^4] = \frac{1}{n} E[|\lambda' \tilde{\mathbf{Z}}_i \tilde{\epsilon}_{i,n} I_i|^4].$$

We use the c_r-inequality: $E[|X+Y|^r] \leq 2^{r-1} E[|X|^r + |Y|^r]$ for $r \geq 1$. Repeating using the c_r-inequality gives

$$\begin{aligned} E[|\lambda' \tilde{\mathbf{Z}}_i \tilde{\epsilon}_{i,n} I_i|^4] &= E[|\lambda' \tilde{\mathbf{Z}}_i (Y_i - \tilde{\beta}_{0,n} - \tilde{\beta}_{1,n} D_i - \tilde{\beta}_{2,n} A(X_i))|^4 I_i] \\ &\leq 2^{3c} E[(|\lambda' \tilde{\mathbf{Z}}_i|^4)(|Y_i|^4 + |\tilde{\beta}_{0,n}|^4 + |\tilde{\beta}_{1,n}|^4 D_i + |\tilde{\beta}_{2,n}|^4 A(X_i)^4) I_i] \\ &\leq 2^{3c} (\lambda_1 + \lambda_2 + \lambda_3)^4 (E[Y_i^4] + \tilde{\beta}_{0,n}^4 + \tilde{\beta}_{1,n}^4 + \tilde{\beta}_{2,n}^4) \end{aligned}$$

for some finite constant c, and the right-hand side converges to

$$2^{3c} (\lambda_1 + \lambda_2 + \lambda_3)^4 (E[Y_i^4] + \tilde{\beta}_0^4 + \tilde{\beta}_1^4 + \tilde{\beta}_2^4),$$

which is finite under Assumption 3 (a). Therefore,

$$\sum_{i=1}^{n} E[|V_{i,n}|^4] \to 0,$$

and the conclusion follows from the Lyapunov CLT and the Cramér-Wold device. □

We next consider (A). We can write

$$\begin{aligned} (A) = &(\frac{1}{n} \sum_{i=1}^{n} \mathbf{Z}_i \mathbf{D}_i' I_i)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\mathbf{Z}_i Y_i I_i - \tilde{\mathbf{Z}}_i Y_i I_i) \\ &- (\frac{1}{n} \sum_{i=1}^{n} \mathbf{Z}_i \mathbf{D}_i' I_i)^{-1} [\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\mathbf{Z}_i \mathbf{D}_i' I_i - \tilde{\mathbf{Z}}_i \tilde{\mathbf{D}}_i' I_i)] (\frac{1}{n} \sum_{i=1}^{n} \tilde{\mathbf{Z}}_i \tilde{\mathbf{D}}_i' I_i)^{-1} \frac{1}{n} \sum_{i=1}^{n} \tilde{\mathbf{Z}}_i Y_i I_i. \end{aligned}$$

Step C.3.1.3. *Let $\{V_i\}_{i=1}^{\infty}$ be i.i.d. random variables such that $E[|V_i|] < \infty$ and that $E[V_i|X_i]$ is bounded on $N(D^*, \delta') \cap \mathcal{X}$ for some $\delta' > 0$. Then,*

$$E[V_i p^A(X_i;\delta)^l (p^A(X_i;\delta) - A(X_i)) 1\{p^A(X_i;\delta) \in (0,1)\}] = O(\delta)$$

for $l = 0, 1$.

Proof. For every $x \notin N(D^*, \delta)$, $B(x, \delta) \cap D^* = \emptyset$, so A is continuously differentiable on $B(x, \delta)$. By the mean value theorem, for every $x \notin N(D^*, \delta)$ and $a \in B(\mathbf{0}, \delta)$,

$$A(x + a) = A(x) + \nabla A(y(x, a))' a$$

for some point $y(x, a)$ on the line segment connecting x and $x + a$. For every $x \notin N(D^*, \delta)$,

$$\begin{aligned} p^A(x; \delta) &= \frac{\int_{B(\mathbf{0},1)} A(x + \delta u) du}{\int_{B(\mathbf{0},1)} du} \\ &= \frac{\int_{B(\mathbf{0},1)} (A(x) + \delta \nabla A(y(x, \delta u))' u) du}{\int_{B(\mathbf{0},1)} du} \\ &= A(x) + \delta \frac{\int_{B(\mathbf{0},1)} \nabla A(y(x, \delta u))' u du}{\int_{B(\mathbf{0},1)} du}. \end{aligned}$$

Now, we can write

$$\begin{aligned} & E[V_i p^A(X_i; \delta)^l (p^A(X_i; \delta) - A(X_i)) 1\{p^A(X_i; \delta) \in (0, 1)\}] \\ = \; & E[V_i p^A(X_i; \delta)^l (p^A(X_i; \delta) - A(X_i)) 1\{p^A(X_i; \delta) \in (0, 1)\} 1\{X_i \notin N(D^*, \delta)\}] \\ & + E[V_i p^A(X_i; \delta)^l (p^A(X_i; \delta) - A(X_i)) 1\{p^A(X_i; \delta) \in (0, 1)\} 1\{X_i \in N(D^*, \delta)\}]. \end{aligned}$$

For the first term,

$$\begin{aligned} & |E[V_i p^A(X_i; \delta)^l (p^A(X_i; \delta) - A(X_i)) 1\{p^A(X_i; \delta) \in (0, 1)\} 1\{X_i \notin N(D^*, \delta)\}]| \\ = \; & \delta |E[V_i p^A(X_i; \delta)^l \frac{\int_{B(\mathbf{0},1)} \nabla A(y(X_i, \delta u))' u du}{\int_{B(\mathbf{0},1)} du} 1\{p^A(X_i; \delta) \in (0, 1)\} 1\{X_i \notin N(D^*, \delta)\}]| \\ \leq \; & \delta E[|V_i| p^A(X_i; \delta)^l \frac{\int_{B(\mathbf{0},1)} \sum_{k=1}^p |\frac{\partial A(y(X_i, \delta u))}{\partial x_k}| |u_k| du}{\int_{B(\mathbf{0},1)} du} 1\{p^A(X_i; \delta) \in (0, 1)\} 1\{X_i \notin N(D^*, \delta)\}] \\ \leq \; & \delta E[|V_i|] \sum_{k=1}^p \sup_{x \in C^*} \left| \frac{\partial A(x)}{\partial x_k} \right| \frac{\int_{B(\mathbf{0},1)} |u_k| du}{\int_{B(\mathbf{0},1)} du} \\ = \; & O(\delta), \end{aligned}$$

where we use the assumption that the partial derivatives of A is bounded on C^*. For the second term, for sufficiently small $\delta > 0$,

$$\begin{aligned} & |E[V_i p^A(X_i; \delta)^l (p^A(X_i; \delta) - A(X_i)) 1\{p^A(X_i; \delta) \in (0, 1)\} 1\{X_i \in N(D^*, \delta)\}]| \\ \leq \; & E[|E[V_i | X_i]| 1\{X_i \in N(D^*, \delta)\}] \\ \leq \; & C E[1\{X_i \in N(D^*, \delta)\}] \\ = \; & C \Pr(X_i \in N(D^*, \delta)) \\ = \; & O(\delta), \end{aligned}$$

where C is some constant, the second inequality follows from the assumption that $E[V_i | X_i]$ is bounded on $N(D^*, \delta') \cap \mathcal{X}$ for some $\delta' > 0$, and the last equality follows from Assumption 4 (a). □

Step C.3.1.4. $\frac{1}{\sqrt{n}}\sum_{i=1}^n(\mathbf{Z}_iY_iI_i - \tilde{\mathbf{Z}}_iY_iI_i) = o_p(1)$ *and* $\frac{1}{\sqrt{n}}\sum_{i=1}^n(\mathbf{Z}_i\mathbf{D}_i'I_i - \tilde{\mathbf{Z}}_i\tilde{\mathbf{D}}_i'I_i) = o_p(1)$.

Proof. We only show that $\frac{1}{\sqrt{n}}\sum_{i=1}^n(p^A(X_i;\delta_n)^2 - A(X_i)^2)I_i = o_p(1)$. The proofs for the other elements are similar. As for bias,

$$\begin{aligned}
&E[\frac{1}{\sqrt{n}}\sum_{i=1}^n(p^A(X_i;\delta_n)^2 - A(X_i)^2)I_i] \\
=&\sqrt{n}E[(p^A(X_i;\delta_n)^2 - A(X_i)^2)I_i] \\
=&\sqrt{n}E[(p^A(X_i;\delta_n) + A(X_i))(p^A(X_i;\delta_n) - A(X_i))I_i] \\
=&\sqrt{n}O(\delta_n) \\
=&0,
\end{aligned}$$

where the third equality follows from Step C.3.1.3 and the last from the assumption that $n\delta_n^2 \to 0$. As for variance, by Lemma B.6,

$$\begin{aligned}
&\mathrm{Var}(\frac{1}{\sqrt{n}}\sum_{i=1}^n(p^A(X_i;\delta_n)^2 - A(X_i)^2)I_i) \\
\leq& E[(p^A(X_i;\delta_n)^2 - A(X_i)^2)^2I_i] \\
=& E[(p^A(X_i;\delta_n)^4 - 2p^A(X_i;\delta_n)^2A(X_i)^2 + A(X_i)^4)I_i] \\
\to& E[(A(X_i)^4 - 2A(X_i)^2A(X_i)^2 + A(X_i)^4)I_i^A] \\
=&0.
\end{aligned}$$

□

Step C.3.1.5. $n\hat{\mathbf{\Sigma}}^c \xrightarrow{p} (E[\tilde{\mathbf{Z}}_i\tilde{\mathbf{D}}_i'I_i^A])^{-1}E[\tilde{\epsilon}_i^2\tilde{\mathbf{Z}}_i\tilde{\mathbf{Z}}_i'I_i^A](E[\tilde{\mathbf{D}}_i\tilde{\mathbf{Z}}_i'I_i^A])^{-1}$.

Proof. Let $\epsilon_i = Y_i - \mathbf{D}_i'\beta$. We have

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n\hat{\epsilon}_i^2\mathbf{Z}_i\mathbf{Z}_i'I_i =&\frac{1}{n}\sum_{i=1}^n(Y_i - \mathbf{D}_i'\hat{\beta}^c)^2\mathbf{Z}_i\mathbf{Z}_i'I_i \\
=&\frac{1}{n}\sum_{i=1}^n(\epsilon_i - \mathbf{D}_i'(\hat{\beta}^c - \beta))^2\mathbf{Z}_i\mathbf{Z}_i'I_i \\
=&\frac{1}{n}\sum_{i=1}^n\epsilon_i^2\mathbf{Z}_i\mathbf{Z}_i'I_i \\
&-\frac{2}{n}\sum_{i=1}^n(Y_i - \mathbf{D}_i'\beta)((\hat{\beta}_0^c - \beta_0) + D_i(\hat{\beta}_1^c - \beta_1) + p^A(X_i;\delta_n)(\hat{\beta}_2^c - \beta_2))\mathbf{Z}_i\mathbf{Z}_i'I_i \\
&+\frac{1}{n}\sum_{i=1}^n((\hat{\beta}_0^c - \beta_0) + D_i(\hat{\beta}_1^c - \beta_1) + p^A(X_i;\delta_n)(\hat{\beta}_2^c - \beta_2))^2\mathbf{Z}_i\mathbf{Z}_i'I_i \\
=&\frac{1}{n}\sum_{i=1}^n\epsilon_i^2\mathbf{Z}_i\mathbf{Z}_i'I_i + o_p(1)O_p(1),
\end{aligned}$$

where the last equality follows from the result that $\hat{\beta}^c - \beta = o_p(1)$ and from Lemma B.6. Again by Lemma B.6,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2\mathbf{Z}_i\mathbf{Z}_i'I_i &= \frac{1}{n}\sum_{i=1}^{n}(Y_i^2 - 2Y_i\mathbf{D}_i'\beta + \beta'\mathbf{D}_i\mathbf{D}_i'\beta)\mathbf{Z}_i\mathbf{Z}_i'I_i \\ &\xrightarrow{p} E[(Y_i^2 - 2Y_i\tilde{\mathbf{D}}_i'\beta + \beta'\tilde{\mathbf{D}}_i\tilde{\mathbf{D}}_i'\beta)\tilde{\mathbf{Z}}_i\tilde{\mathbf{Z}}_i'I_i^A] \\ &= E[\tilde{\epsilon}_i^2\tilde{\mathbf{Z}}_i\tilde{\mathbf{Z}}_i'I_i^A],\end{aligned}$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{Z}_i\mathbf{D}_i'I_i \xrightarrow{p} E[\tilde{\mathbf{Z}}_i\tilde{\mathbf{D}}_i'I_i^A].$$

The conclusion then follows. □

Step C.3.1.6. $(\hat{\sigma}^c)^{-1}(\hat{\beta}_1^c - \beta_1) \xrightarrow{d} \mathcal{N}(0,1)$.

Proof. By combining the results from Steps C.3.1.2–C.3.1.4 and by Lemma B.6,

$$\begin{aligned}&(A) \xrightarrow{p} 0, \\ &(B) \xrightarrow{d} \mathcal{N}(0, (E[\tilde{\mathbf{Z}}_i\tilde{\mathbf{D}}_i'I_i^A])^{-1}E[\tilde{\epsilon}_i^2\tilde{\mathbf{Z}}_i\tilde{\mathbf{Z}}_i'I_i^A](E[\tilde{\mathbf{D}}_i\tilde{\mathbf{Z}}_i'I_i^A])^{-1}),\end{aligned}$$

and therefore,

$$\sqrt{n}(\hat{\beta}^c - \tilde{\beta}_n) \xrightarrow{d} \mathcal{N}(0, (E[\tilde{\mathbf{Z}}_i\tilde{\mathbf{D}}_i'I_i^A])^{-1}E[\tilde{\epsilon}_i^2\tilde{\mathbf{Z}}_i\tilde{\mathbf{Z}}_i'I_i^A](E[\tilde{\mathbf{D}}_i\tilde{\mathbf{Z}}_i'I_i^A])^{-1}).$$

The conclusion then follows from Steps C.3.1.1 and C.3.1.5. □

□

□

C.3.2 Consistency and Asymptotic Normality of $\hat{\beta}_1^s$ When $\Pr(A(X_i) \in (0,1)) > 0$

Let $I_i^s = 1\{p^s(X_i;\delta_n) \in (0,1)\}$, $\mathbf{D}_i^s = (1, D_i, p^s(X_i;\delta_n))'$ and $\mathbf{Z}_i^s = (1, Z_i, p^s(X_i;\delta_n))'$. Let

$$\hat{\beta}^{c,s} = (\sum_{i=1}^{n}\mathbf{Z}_i^s(\mathbf{D}_i^s)'I_i^s)^{-1}\sum_{i=1}^{n}\mathbf{Z}_i^sY_iI_i^s$$

and

$$\hat{\mathbf{\Sigma}}^{c,s} = (\sum_{i=1}^{n}\mathbf{Z}_i^s(\mathbf{D}_i^s)'I_i^s)^{-1}(\sum_{i=1}^{n}(\hat{\epsilon}_i^s)^2\mathbf{Z}_i^s(\mathbf{Z}_i^s)'I_i^s)(\sum_{i=1}^{n}\mathbf{D}_i^s(\mathbf{Z}_i^s)'I_i^s)^{-1},$$

where $\hat{\epsilon}_i^s = Y_i - (\mathbf{D}_i^s)'\hat{\beta}^{c,s}$. Here, we only show that $\hat{\beta}_1^{c,s} \xrightarrow{p} \beta_1$ if $S_n \to \infty$ and that $(\hat{\sigma}^s)^{-1}(\hat{\beta}_1^{c,s} - \beta_1) \xrightarrow{d} \mathcal{N}(0,1)$ if Assumption 5 holds when $\mathrm{Var}(A(X_i)|I_i^A = 1) > 0$. For that, it suffices to show that

$$\hat{\beta}^{c,s} - \hat{\beta}^c = o_p(1)$$

if $S_n \to \infty$ and that

$$\sqrt{n}(\hat{\beta}^{c,s} - \hat{\beta}^c) = o_p(1),$$
$$n\hat{\mathbf{\Sigma}}^{c,s} \xrightarrow{p} (E[\tilde{\mathbf{Z}}_i\tilde{\mathbf{D}}_i' I_i^A])^{-1} E[\tilde{\epsilon}_i^2 \tilde{\mathbf{Z}}_i\tilde{\mathbf{Z}}_i' I_i^A](E[\tilde{\mathbf{D}}_i\tilde{\mathbf{Z}}_i' I_i^A])^{-1}$$

if Assumption 5 holds. We have

$$\begin{aligned}
\hat{\beta}^{c,s} - \hat{\beta}^c &= (\frac{1}{n}\sum_{i=1}^n \mathbf{Z}_i^s(\mathbf{D}_i^s)' I_i^s)^{-1} \frac{1}{n}\sum_{i=1}^n \mathbf{Z}_i^s Y_i I_i^s - (\frac{1}{n}\sum_{i=1}^n \mathbf{Z}_i\mathbf{D}_i' I_i)^{-1}\frac{1}{n}\sum_{i=1}^n \mathbf{Z}_i Y_i I_i \\
&= (\frac{1}{n}\sum_{i=1}^n \mathbf{Z}_i^s(\mathbf{D}_i^s)' I_i^s)^{-1}(\frac{1}{n}\sum_{i=1}^n \mathbf{Z}_i^s Y_i I_i^s - \frac{1}{n}\sum_{i=1}^n \mathbf{Z}_i Y_i I_i) \\
&\quad - (\frac{1}{n}\sum_{i=1}^n \mathbf{Z}_i^s(\mathbf{D}_i^s)' I_i^s)^{-1}(\frac{1}{n}\sum_{i=1}^n \mathbf{Z}_i^s(\mathbf{D}_i^s)' I_i^s - \frac{1}{n}\sum_{i=1}^n \mathbf{Z}_i\mathbf{D}_i' I_i)(\frac{1}{n}\sum_{i=1}^n \mathbf{Z}_i\mathbf{D}_i' I_i)^{-1}\frac{1}{n}\sum_{i=1}^n \mathbf{Z}_i Y_i I_i.
\end{aligned}$$

By Lemma B.8, $\hat{\beta}^{c,s} - \hat{\beta}^c = o_p(1)$ if $S_n \to \infty$, and $\sqrt{n}(\hat{\beta}^{c,s} - \hat{\beta}^c) = o_p(1)$ under the boundedness imposed by Assumption 4 (c) if Assumption 5 holds.

By proceeding as in Step C.3.1.5 in Section C.3.1, we have

$$\frac{1}{n}\sum_{i=1}^n (\hat{\epsilon}_i^s)^2 \mathbf{Z}_i^s(\mathbf{Z}_i^s)' I_i^s = \frac{1}{n}\sum_{i=1}^n (\epsilon_i^s)^2 \mathbf{Z}_i^s(\mathbf{Z}_i^s)' I_i^s + o_p(1),$$

where $\epsilon_i^s = Y_i - (\mathbf{D}_i^s)'\beta$. Then, by Lemma B.8,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n (\hat{\epsilon}_i^s)^2 \mathbf{Z}_i^s(\mathbf{Z}_i^s)' I_i^s - \frac{1}{n}\sum_{i=1}^n \epsilon_i^2 \mathbf{Z}_i\mathbf{Z}_i' I_i \\
&= \frac{1}{n}\sum_{i=1}^n (Y_i^2 - 2Y_i(\mathbf{D}_i^s)'\beta + \beta'\mathbf{D}_i^s(\mathbf{D}_i^s)'\beta)\mathbf{Z}_i^s(\mathbf{Z}_i^s)' I_i^s - \frac{1}{n}\sum_{i=1}^n (Y_i^2 - 2Y_i\mathbf{D}_i'\beta + \beta'\mathbf{D}_i\mathbf{D}_i'\beta)\mathbf{Z}_i\mathbf{Z}_i' I_i + o_p(1) \\
&= o_p(1)
\end{aligned}$$

so that

$$\frac{1}{n}\sum_{i=1}^n (\hat{\epsilon}_i^s)^2 \mathbf{Z}_i^s(\mathbf{Z}_i^s)' I_i^s \xrightarrow{p} E[\tilde{\epsilon}_i^2 \tilde{\mathbf{Z}}_i\tilde{\mathbf{Z}}_i' I_i^A].$$

Also, $\frac{1}{n}\sum_{i=1}^n \mathbf{Z}_i^s(\mathbf{D}_i^s)' I_i^s \xrightarrow{p} E[\tilde{\mathbf{Z}}_i\tilde{\mathbf{D}}_i' I_i^A]$ by using Lemma B.8. The conclusion then follows. □

C.3.3 Consistency and Asymptotic Normality of $\hat{\beta}_1$ When $\Pr(A(X_i) \in (0,1)) = 0$

Since $\Pr(A(X_i) \in (0,1)) = 0$, $\mathbf{I}_n = 0$ with probability one. Hence,

$$\hat{\beta} = (\sum_{i=1}^n \mathbf{Z}_i\mathbf{D}_i' I_i)^{-1}\sum_{i=1}^n \mathbf{Z}_i Y_i I_i$$

with probability one. We use the notation and results provided in Appendix B. By Lemma B.5, under Assumption 3 (e), there exists $\mu > 0$ such that $d_{\Omega^*}^s$ is twice continuously differentiable on $N(\partial\Omega^*, \mu)$ and that

$$\int_{N(\partial\Omega^*,\delta)} g(x)dx = \int_{-\delta}^{\delta}\int_{\partial\Omega^*} g(u + \lambda\nu_{\Omega^*}(u)) J_{p-1}^{\partial\Omega^*}\psi_{\Omega^*}(u,\lambda) d\mathcal{H}^{p-1}(u)d\lambda$$

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

Below we show that $\hat{\beta}_1 \xrightarrow{p} \beta_1$ if $n\delta_n \to \infty$ and $\delta_n \to 0$ and that $\hat{\sigma}^{-1}(\hat{\beta}_1 - \beta_1) \xrightarrow{d} \mathcal{N}(0,1)$ if $n\delta_n^3 \to 0$ in addition. The proof proceeds in eight steps.

Step C.3.3.1. *There exist $\bar{\delta} > 0$ and a bounded function $r : \partial\Omega^* \cap N(\mathcal{X}, \bar{\delta}) \times (-1,1) \times (0, \bar{\delta}) \to \mathbb{R}$ such that*

$$p^A(u + \delta v \nu_{\Omega^*}(u); \delta) = k(v) + \delta r(u, v, \delta)$$

for every $(u, v, \delta) \in \partial\Omega^* \cap N(\mathcal{X}, \bar{\delta}) \times (-1,1) \times (0, \bar{\delta})$, where

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

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

Proof. By Assumption 3 (f) (ii), there exists $\bar{\delta} \in (0, \frac{\mu}{2})$ such that $A(x) = 0$ for almost every $x \in N(\mathcal{X}, 3\bar{\delta}) \setminus \Omega^*$. By Taylor's theorem, for every $u \in \partial\Omega^* \cap N(\mathcal{X}, \bar{\delta})$ and $a \in B(\mathbf{0}, 2\bar{\delta})$,

$$d^s_{\Omega^*}(u + a) = d^s_{\Omega^*}(u) + \nabla d^s_{\Omega^*}(u)' a + a' R(u, a) a,$$

where

$$R(u, a) = \int_0^1 (1 - t) D^2 d^s_{\Omega^*}(u + ta) dt.$$

Since $D^2 d^s_{\Omega^*}$ is continuous and $\mathrm{cl}(N(\partial\Omega^*, 2\bar{\delta}))$ is bounded and closed, $D^2 d^s_{\Omega^*}$ is bounded on $\mathrm{cl}(N(\partial\Omega^*, 2\bar{\delta}))$. Therefore, $R(\cdot, \cdot)$ is bounded on $\partial\Omega^* \cap N(\mathcal{X}, \bar{\delta}) \times B(\mathbf{0}, 2\bar{\delta})$. It also follows that

$$d^s_{\Omega^*}(u + a) = \nu_{\Omega^*}(u)' a + a' R(u, a) a,$$

since $d^s_{\Omega^*}(u) = 0$ and $\nabla d^s_{\Omega^*}(u) = \nu_{\Omega^*}(u)$ for every $u \in \partial\Omega^* \cap N(\mathcal{X}, 2\bar{\delta})$ by Lemma B.1. For $(u, v, \delta) \in \partial\Omega^* \cap N(\mathcal{X}, \bar{\delta}) \times (-1,1) \times (0, \bar{\delta})$,

$$\begin{aligned}
& p^A(u + \delta v \nu_{\Omega^*}(u); \delta) \\
&= \frac{\int_{B(\mathbf{0},1)} A(u + \delta v \nu_{\Omega^*}(u) + \delta w) dw}{\int_{B(\mathbf{0},1)} dw} \\
&= \frac{\int_{B(\mathbf{0},1)} 1\{u + \delta v \nu_{\Omega^*}(u) + \delta w \in \Omega^*\} dw}{\mathrm{Vol}_p} \\
&= \frac{\int_{B(\mathbf{0},1)} 1\{d^s_{\Omega^*}(u + \delta(v \nu_{\Omega^*}(u) + w)) \geq 0)\} dw}{\mathrm{Vol}_p} \\
&= \frac{\int_{B(\mathbf{0},1)} 1\{\delta \nu_{\Omega^*}(u)'(v \nu_{\Omega^*}(u) + w) + \delta^2 (v \nu_{\Omega^*}(u) + w)' R(u, \delta(v \nu_{\Omega^*}(u) + w))(v \nu_{\Omega^*}(u) + w) \geq 0\} dw}{\mathrm{Vol}_p},
\end{aligned}$$

where Vol_p denotes the volume of the p-dimensional unit ball, and the second equality follows since $u + \delta v \nu_{\Omega^*}(u) + \delta w \in N(\mathcal{X}, 3\bar{\delta})$ and hence $A(u + \delta v \nu_{\Omega^*}(u) + \delta w) = 0$ for almost every

$w \in B(\mathbf{0},1)$ such that $u + \delta v \nu_{\Omega^*}(u) + \delta w \notin \Omega^*$. Observe that

$$
\begin{aligned}
&1\{\delta \nu_{\Omega^*}(u)'(v\nu_{\Omega^*}(u)+w) + \delta^2 (v\nu_{\Omega^*}(u)+w)' R(u, \delta(v\nu_{\Omega^*}(u)+w))(v\nu_{\Omega^*}(u)+w) \geq 0\} \\
&= 1\{v + \nu_{\Omega^*}(u)\cdot w + \delta(v\nu_{\Omega^*}(u)+w)' R(u, \delta(v\nu_{\Omega^*}(u)+w))(v\nu_{\Omega^*}(u)+w) \geq 0\} \\
&= 1\{v + \nu_{\Omega^*}(u)\cdot w \geq 0\} \\
&\quad - \underbrace{1\{v + \nu_{\Omega^*}(u)\cdot w \geq 0, v + \nu_{\Omega^*}(u)\cdot w + \delta(v\nu_{\Omega^*}(u)+w)' R(u, \delta(v\nu_{\Omega^*}(u)+w))(v\nu_{\Omega^*}(u)+w) < 0\}}_{=a(u,v,w,\delta)} \\
&\quad + \underbrace{1\{v + \nu_{\Omega^*}(u)\cdot w < 0, v + \nu_{\Omega^*}(u)\cdot w + \delta(v\nu_{\Omega^*}(u)+w)' R(u, \delta(v\nu_{\Omega^*}(u)+w))(v\nu_{\Omega^*}(u)+w) \geq 0\}}_{=b(u,v,w,\delta)}.
\end{aligned}
$$

Note that the set $\{w \in B(\mathbf{0},1) : v + \nu(u)\cdot w \geq 0\}$ 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

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

Therefore, for every $(u,v,\delta) \in \partial\Omega^* \cap N(\mathcal{X}, \bar{\delta}) \times (-1,1) \times (0,\bar{\delta})$,

$$
p^A(u + \delta v \nu_{\Omega^*}(u); \delta) = k(v) + \frac{\int_{B(\mathbf{0},1)} (-a(u,v,w,\delta) + b(u,v,w,\delta)) dw}{\mathrm{Vol}_p}.
$$

Now let $r(u,v,\delta) = \delta^{-1}(p^A(u + \delta v \nu_{\Omega^*}(u); \delta) - k(v))$. Since $R(\cdot,\cdot)$ is bounded on $\partial\Omega^* \cap N(\mathcal{X}, \bar{\delta}) \times B(\mathbf{0}, 2\bar{\delta})$ and $\|\nu_{\Omega^*}(u)\| = 1$, there exists $\bar{r} > 0$ such that

$$
|(v\nu_{\Omega^*}(u)+w)' R(u, \delta(v\nu_{\Omega^*}(u)+w))(v\nu_{\Omega^*}(u)+w)| \leq \bar{r}
$$

for every $(u,v,w,\delta) \in \partial\Omega^* \cap N(\mathcal{X}, \bar{\delta}) \times (-1,1) \times B(\mathbf{0},1) \times (0,\bar{\delta})$. Therefore,

$$
0 \leq a(u,v,w,\delta) \leq 1\{0 \leq v + \nu_{\Omega^*}(u)\cdot w < \delta\bar{r}\}
$$

and

$$
0 \leq b(u,v,w,\delta) \leq 1\{-\delta\bar{r} \leq v + \nu_{\Omega^*}(u)\cdot w < 0\}.
$$

It then follows that

$$
\begin{aligned}
-\frac{\int_{B(\mathbf{0},1)} 1\{0 \leq v + \nu_{\Omega^*}(u)\cdot w < \delta\bar{r}\} dw}{\mathrm{Vol}_p} &\leq \frac{\int_{B(\mathbf{0},1)} (-a(u,v,w,\delta) + b(u,v,w,\delta)) dw}{\mathrm{Vol}_p} \\
&\leq \frac{\int_{B(\mathbf{0},1)} 1\{-\delta\bar{r} \leq v + \nu_{\Omega^*}(u)\cdot w < 0\} dw}{\mathrm{Vol}_p}.
\end{aligned}
$$

The set $\{w \in B(\mathbf{0},1) : 0 \leq v + \nu_{\Omega^*}(u)\cdot w < \delta\bar{r}\}$ is a region of the p-dimensional unit ball cut off by the two planes $\{w \in \mathbb{R}^p : v + \nu_{\Omega^*}(u)\cdot w = 0\}$ and $\{w \in \mathbb{R}^p : v + \nu_{\Omega^*}(u)\cdot w = \delta\bar{r}\}$. Its

Lebesgue measure is at most the volume of the $(p-1)$-dimensional unit ball times the distance between the two planes, so

$$-\delta \text{Vol}_{p-1}\bar{r} \leq -\int_{B(\mathbf{0},1)} 1\{0 \leq v + \nu_{\Omega^*}(u) \cdot w < \delta\bar{r}\}dw.$$

Likewise,

$$\int_{B(\mathbf{0},1)} 1\{-\delta\bar{r} \leq v + \nu_{\Omega^*}(u) \cdot w < 0\}dw \leq \delta \text{Vol}_{p-1}\bar{r}.$$

Therefore,

$$-\frac{\delta \text{Vol}_{p-1}\bar{r}}{\text{Vol}_p} \leq \frac{\int_{B(\mathbf{0},1)}(-a(u,v,w,\delta) + b(u,v,w,\delta))dw}{\text{Vol}_p} \leq \frac{\delta \text{Vol}_{p-1}\bar{r}}{\text{Vol}_p}.$$

It follows that

$$\begin{aligned} r(u,v,\delta) &= \delta^{-1}\frac{\int_{B(\mathbf{0},1)}(-a(u,v,w,\delta) + b(u,v,w,\delta))dw}{\text{Vol}_p} \\ &\in [-\frac{\text{Vol}_{p-1}\bar{r}}{\text{Vol}_p}, \frac{\text{Vol}_{p-1}\bar{r}}{\text{Vol}_p}], \end{aligned}$$

and hence r is bounded on $\partial\Omega^* \cap N(\mathcal{X}, \bar{\delta}) \times (-1,1) \times (0, \bar{\delta})$. □

Step C.3.3.2. *For every $(u,v,\delta) \in \partial\Omega^* \cap N(\mathcal{X}, \bar{\delta}) \times (-1,1) \times (0,\bar{\delta})$, $p^A(u + \delta v \nu_{\Omega^*}(u); \delta) \in (0,1)$.*

Proof. Fix $(u,v,\delta) \in \partial\Omega^* \cap N(\mathcal{X}, \bar{\delta}) \times (-1,1) \times (0,\bar{\delta})$. Suppose $v = 0$. By Step C.3.3.1, $p^A(u) = \lim_{\delta' \to 0} p^A(u;\delta') = k(0) = \frac{1}{2}$. This implies that there exists $\delta' \in (0,\delta)$ such that $p^A(u;\delta') \in (0,1)$. It then follows that $0 < \mathcal{L}^p(B(u,\delta') \cap \Omega^*) \leq \mathcal{L}^p(B(x,\delta) \cap \Omega^*)$ and that $0 < \mathcal{L}^p(B(x,\delta') \setminus \Omega^*) \leq \mathcal{L}^p(B(x,\delta) \setminus \Omega^*)$. Therefore, $p^A(u;\delta) = \frac{\mathcal{L}^p(B(u,\delta) \cap \Omega^*)}{\mathcal{L}^p(B(u,\delta))} \in (0,1)$.

Suppose $v \neq 0$ and let $\epsilon \in (0, \delta(1-|v|))$. Note that $B(u,\epsilon) \subset B(u + \delta v \nu_{\Omega^*}(u), \delta)$, since for any $x \in B(u,\epsilon)$, $\|u + \delta v \nu_{\Omega^*}(u) - x\| \leq \|\delta v \nu_{\Omega^*}(u)\| + \|u - x\| \leq \delta|v| + \epsilon < \delta$. Since $p^A(u) = \frac{1}{2}$, there exists $\epsilon' \in (0,\epsilon)$ such that $p^A(u;\epsilon') \in (0,1)$. It then follows that $0 < \mathcal{L}^p(B(u,\epsilon') \cap \Omega^*) \leq \mathcal{L}^p(B(u,\epsilon) \cap \Omega^*) \leq \mathcal{L}^p(B(u + \delta v \nu_{\Omega^*}(u), \delta) \cap \Omega^*)$ and that $0 < \mathcal{L}^p(B(x,\epsilon') \setminus \Omega^*) \leq \mathcal{L}^p(B(x,\epsilon) \setminus \Omega^*) \leq \mathcal{L}^p(B(u + \delta v \nu_{\Omega^*}(u), \delta) \setminus \Omega^*)$. Therefore, $p^A(u + \delta v \nu_{\Omega^*}(u); \delta) = \frac{\mathcal{L}^p(B(u + \delta v \nu_{\Omega^*}(u), \delta) \cap \Omega^*)}{\mathcal{L}^p(B(u + \delta v \nu_{\Omega^*}(u), \delta))} \in (0,1)$. □

Step C.3.3.3. *Let $g : \mathbb{R}^p \to \mathbb{R}$ be a function that is bounded on $N(\partial\Omega^*, \delta') \cap N(\mathcal{X}, \delta')$ for some $\delta' > 0$. Then, for $l \geq 0$, there exist $\tilde{\delta} > 0$ and constant $C > 0$ such that*

$$|\delta^{-1} E[p^A(X_i;\delta)^l g(X_i) 1\{p^A(X_i;\delta) \in (0,1)\}]| \leq C$$

for every $\delta \in (0,\tilde{\delta})$. If g is continuous on $N(\partial\Omega^*, \delta') \cap N(\mathcal{X}, \delta')$ for some $\delta' > 0$, then

$$\begin{aligned} \delta^{-1} E[p^A(X_i;\delta)^l g(X_i) 1\{p^A(X_i;\delta) \in (0,1)\}] &= \int_{-1}^{1} k(v)^l dv \int_{\partial\Omega^*} g(x) f_X(x) d\mathcal{H}^{p-1}(x) + o(1) \\ \delta^{-1} E[Z_i p^A(X_i;\delta)^l g(X_i) 1\{p^A(X_i;\delta) \in (0,1)\}] &= \int_{0}^{1} k(v)^l dv \int_{\partial\Omega^*} g(x) f_X(x) d\mathcal{H}^{p-1}(x) + o(1) \end{aligned}$$

for $l \geq 0$. Furthermore, if g is continuously differentiable and ∇g is bounded on $N(\partial\Omega^*, \delta') \cap N(\mathcal{X}, \delta')$ for some $\delta' > 0$, then

$$\delta^{-1}E[p^A(X_i;\delta)^l g(X_i) 1\{p^A(X_i;\delta) \in (0,1)\}] = \int_{-1}^{1} k(v)^l dv \int_{\partial\Omega^*} g(x) f_X(x) d\mathcal{H}^{p-1}(x) + O(\delta)$$
$$\delta^{-1}E[Z_i p^A(X_i;\delta)^l g(X_i) 1\{p^A(X_i;\delta) \in (0,1)\}] = \int_{0}^{1} k(v)^l dv \int_{\partial\Omega^*} g(x) f_X(x) d\mathcal{H}^{p-1}(x) + O(\delta)$$

for $l \geq 0$.

Proof. Let $\bar{\delta}$ be given in Step C.3.3.1. Under Assumption 3 (g), there exists $\tilde{\delta} \in (0, \bar{\delta})$ such that f_X is bounded, is continuously differentiable, and has bounded partial derivatives on $N(\partial\Omega^*, 2\tilde{\delta}) \cap N(\mathcal{X}, 2\tilde{\delta})$. Let $\tilde{\delta} \in (0, \bar{\delta})$ be such that both g and f_X are bounded on $N(\partial\Omega^*, 2\tilde{\delta}) \cap N(\mathcal{X}, 2\tilde{\delta})$. We first show that $p^A(x;\delta) \in \{0,1\}$ for every $x \in \mathcal{X} \setminus N(\partial\Omega^*, \delta)$ for every $\delta \in (0, \tilde{\delta})$. Pick $x \in \mathcal{X} \setminus N(\partial\Omega^*, \delta)$ and $\delta \in (0, \tilde{\delta})$. Since $B(x,\delta) \cap \partial\Omega^* = \emptyset$, either $B(x,\delta) \subset \text{int}(\Omega^*)$ or $B(x,\delta) \subset \text{int}(\mathbb{R}^p \setminus \Omega^*)$. If $B(x,\delta) \subset \text{int}(\Omega^*)$, $p^A(x;\delta) = 1$. If $B(x,\delta) \subset \text{int}(\mathbb{R}^p \setminus \Omega^*)$, $p^A(x;\delta) = 0$, since $A(x') = 0$ for almost every $x' \in B(x,\delta) \subset N(\mathcal{X}, 3\bar{\delta}) \setminus \Omega^*$ by the choice of $\bar{\delta}$. Therefore, $\{x \in \mathcal{X} : p^A(x;\delta) \in (0,1)\} \subset N(\partial\Omega^*, \delta)$ for every $\delta \in (0, \tilde{\delta})$. By this and Lemma B.5, for $\delta \in (0, \tilde{\delta})$,

$$\begin{aligned}
&\delta^{-1}E[p^A(X_i;\delta)^l g(X_i) 1\{p^A(X_i;\delta) \in (0,1)\}] \\
&= \delta^{-1} \int p^A(x;\delta)^l g(x) 1\{p^A(x;\delta) \in (0,1)\} f_X(x) dx \\
&= \delta^{-1} \int_{N(\partial\Omega^*,\delta)} p^A(x;\delta)^l g(x) 1\{p^A(x;\delta) \in (0,1)\} f_X(x) dx \\
&= \delta^{-1} \int_{-\delta}^{\delta} \int_{\partial\Omega^*} p^A(u + \lambda\nu_{\Omega^*}(u);\delta)^l g(u + \lambda\nu_{\Omega^*}(u)) 1\{p^A(u + \lambda\nu_{\Omega^*}(u);\delta) \in (0,1)\} \\
&\qquad \times f_X(u + \lambda\nu_{\Omega^*}(u)) J_{p-1}^{\partial\Omega^*}\psi_{\Omega^*}(u,\lambda) d\mathcal{H}^{p-1}(u) d\lambda.
\end{aligned}$$

With change of variables $v = \frac{\lambda}{\delta}$, we have

$$\begin{aligned}
&\delta^{-1}E[p^A(X_i;\delta)^l g(X_i) 1\{p^A(X_i;\delta) \in (0,1)\}] \\
&= \int_{-1}^{1} \int_{\partial\Omega^*} p^A(u + \delta v\nu_{\Omega^*}(u);\delta)^l 1\{p^A(u + \delta v\nu_{\Omega^*}(u);\delta) \in (0,1)\} \\
&\qquad \times g(u + \delta v\nu_{\Omega^*}(u)) f_X(u + \delta v\nu_{\Omega^*}(u)) J_{p-1}^{\partial\Omega^*}\psi_{\Omega^*}(u,\delta v) d\mathcal{H}^{p-1}(u) dv.
\end{aligned}$$

For every $(u,v,\delta) \in \partial\Omega^* \setminus N(\mathcal{X}, \tilde{\delta}) \times (-1,1) \times (0, \tilde{\delta})$, $u + \delta v\nu_{\Omega^*}(u) \notin \mathcal{X}$, so

$$\begin{aligned}
&\delta^{-1}E[p^A(X_i;\delta)^l g(X_i) 1\{p^A(X_i;\delta) \in (0,1)\}] \\
&= \int_{-1}^{1} \int_{\partial\Omega^* \cap N(\mathcal{X},\tilde{\delta})} p^A(u + \delta v\nu_{\Omega^*}(u);\delta)^l 1\{p^A(u + \delta v\nu_{\Omega^*}(u);\delta) \in (0,1)\} \\
&\qquad \times g(u + \delta v\nu_{\Omega^*}(u)) f_X(u + \delta v\nu_{\Omega^*}(u)) J_{p-1}^{\partial\Omega^*}\psi_{\Omega^*}(u,\delta v) d\mathcal{H}^{p-1}(u) dv \\
&= \int_{-1}^{1} \int_{\partial\Omega^* \cap N(\mathcal{X},\tilde{\delta})} (k(v) + \delta r(u,v,\delta))^l g(u + \delta v\nu_{\Omega^*}(u)) f_X(u + \delta v\nu_{\Omega^*}(u)) J_{p-1}^{\partial\Omega^*}\psi_{\Omega^*}(u,\delta v) d\mathcal{H}^{p-1}(u) dv,
\end{aligned}$$

where the second equality follows from Steps C.3.3.1 and C.3.3.2. By Lemma B.5, $J_{p-1}^{\partial\Omega^*}\psi_{\Omega^*}(\cdot,\cdot)$ is bounded on $\partial\Omega^* \times (-\tilde{\delta}, \tilde{\delta})$. Since r, g and f_X are also bounded, for some constant $C > 0$,

$$|\delta^{-1}E[p^A(X_i;\delta)^l g(X_i) 1\{p^A(X_i;\delta) \in (0,1)\}]| \leq C \int_{-1}^{1} \int_{\partial\Omega^* \cap N(\mathcal{X},\tilde{\delta})} d\mathcal{H}^{p-1}(u)dv,$$

which is finite by Assumption 3 (f) (i). Moreover, if g and f_X are continuous on $N(\partial\Omega^*, 2\tilde{\delta}) \cap N(\mathcal{X}, 2\tilde{\delta})$, by the Dominated Convergence Theorem,

$$\delta^{-1}E[p^A(X_i;\delta)^l g(X_i) 1\{p^A(X_i;\delta) \in (0,1)\}] \to \int_{-1}^{1} k(v)^l dv \int_{\partial\Omega^*} g(u) f_X(u) d\mathcal{H}^{p-1}(u),$$

where we use the fact from Lemma B.5 that $J_{p-1}^{\partial\Omega^*}\psi_{\Omega^*}(u,\lambda)$ is continuous in λ and $J_{p-1}^{\partial\Omega^*}\psi_{\Omega^*}(u,0) = 1$.

Note that $A(x) = 1$ for every $x \in \Omega^*$ and $A(x) = 0$ for almost every $x \in N(\mathcal{X}, 2\tilde{\delta}) \setminus \Omega^*$. Also, for every $(u,v,\delta) \in \partial\Omega^* \cap N(\mathcal{X},\tilde{\delta}) \times (-1,1) \times (0,\tilde{\delta})$, $u + \delta v \nu_{\Omega^*}(u) \in \Omega^*$ if $v \in (0,1)$ and $u + \delta v \nu_{\Omega^*}(u) \in N(\mathcal{X}, 2\tilde{\delta}) \setminus \Omega^*$ if $v \in (-1,0]$. Therefore,

$$\begin{aligned}
&\delta^{-1}E[Z_i p^A(X_i;\delta)^l g(X_i) 1\{p^A(X_i;\delta) \in (0,1)\}] \\
&= \delta^{-1}E[A(X_i) p^A(X_i;\delta)^l g(X_i) 1\{p^A(X_i;\delta) \in (0,1)\}] \\
&= \int_{-1}^{1} \int_{\partial\Omega^* \cap N(\mathcal{X},\tilde{\delta})} A(u + \delta v \nu_{\Omega^*}(u))(k(v) + \delta r(u,v,\delta))^l g(u + \delta v \nu_{\Omega^*}(u)) \\
&\qquad\qquad \times f_X(u + \delta v \nu_{\Omega^*}(u)) J_{p-1}^{\partial\Omega^*}\psi_{\Omega^*}(u,\delta v) d\mathcal{H}^{p-1}(u)dv \\
&= \int_{0}^{1} \int_{\partial\Omega^* \cap N(\mathcal{X},\tilde{\delta})} (k(v) + \delta r(u,v,\delta))^l g(u + \delta v \nu_{\Omega^*}(u)) f_X(u + \delta v \nu_{\Omega^*}(u)) J_{p-1}^{\partial\Omega^*}\psi_{\Omega^*}(u,\delta v) d\mathcal{H}^{p-1}(u)dv \\
&\to \int_{0}^{1} k(v)^l dv \int_{\partial\Omega^*} g(u) f_X(u) d\mathcal{H}^{p-1}(u).
\end{aligned}$$

Now suppose that g and f_X are continuously differentiable on $N(\partial\Omega^*, 2\tilde{\delta}) \cap N(\mathcal{X}, 2\tilde{\delta})$ and that ∇g and ∇f are bounded on $N(\partial\Omega^*, 2\tilde{\delta}) \cap N(\mathcal{X}, 2\tilde{\delta})$. Using the mean-value theorem, we obtain that, for any $(u,v,\delta) \in \partial\Omega^* \cap N(\mathcal{X},\tilde{\delta}) \times (-1,1) \times (0,\tilde{\delta})$,

$$\begin{aligned}
g(u + \delta v \nu_{\Omega^*}(u)) &= g(u) + \nabla g(y_g(u, \delta v \nu_{\Omega^*}(u)))' \delta v \nu_{\Omega^*}(u), \\
f_X(u + \delta v \nu_{\Omega^*}(u)) &= f_X(u) + \nabla f_X(y_f(u, \delta v \nu_{\Omega^*}(u)))' \delta v \nu_{\Omega^*}(u)
\end{aligned}$$

for some $y_g(u, \delta v \nu_{\Omega^*}(u))$ and $y_f(u, \delta v \nu_{\Omega^*}(u))$ that are on the line segment connecting u and $u + \delta v \nu_{\Omega^*}(u)$. In addition,

$$\begin{aligned}
J_{p-1}^{\partial\Omega^*}\psi_{\Omega^*}(u,\delta v) &= J_{p-1}^{\partial\Omega^*}\psi_{\Omega^*}(u,0) + \frac{\partial J_{p-1}^{\partial\Omega^*}\psi_{\Omega^*}(u, y_J(u,\delta v))}{\partial\lambda}\delta v \\
&= 1 + \frac{\partial J_{p-1}^{\partial\Omega^*}\psi_{\Omega^*}(u, y_J(u,\delta v))}{\partial\lambda}\delta v
\end{aligned}$$

for some $y_J(u,\delta v)$ that is on the line segment connecting 0 and δv. By Lemma B.5, $\frac{\partial J_{p-1}^{\partial\Omega^*}\psi_{\Omega^*}(\cdot,\cdot)}{\partial\lambda}$ is bounded on $\partial\Omega^*\times(-\tilde{\delta},\tilde{\delta})$. We then have

$$\begin{aligned}
&\delta^{-1}E[p^A(X_i;\delta)^l g(X_i)1\{p^A(X_i;\delta)\in(0,1)\}]\\
=&\int_{-1}^{1}\int_{\partial\Omega^*\cap N(\mathcal{X},\tilde{\delta})}(k(v)+\delta r(u,v,\delta))^l(g(u)+\nabla g(y_g(u,\delta v\nu_{\Omega^*}(u)))'\delta v\nu_{\Omega^*}(u))\\
&\qquad\times(f_X(u)+\nabla f_X(y_f(u,\delta v\nu_{\Omega^*}(u)))'\delta v\nu_{\Omega^*}(u))(1+\tfrac{\partial J_{p-1}^{\partial\Omega^*}\psi_{\Omega^*}(u,y_J(u,\delta v))}{\partial\lambda}\delta v)d\mathcal{H}^{p-1}(u)dv\\
=&\int_{-1}^{1}\int_{\partial\Omega^*\cap N(\mathcal{X},\tilde{\delta})}(k(v)^l g(u)f_X(u)+\delta h(u,v,\delta))d\mathcal{H}^{p-1}(u)dv\\
=&\int_{-1}^{1}k(v)^l dv\int_{\partial\Omega^*}g(u)f_X(u)d\mathcal{H}^{p-1}(u)+\delta\int_{-1}^{1}\int_{\partial\Omega^*\cap N(\mathcal{X},\tilde{\delta})}h(u,v,\delta)d\mathcal{H}^{p-1}(u)dv
\end{aligned}$$

for some function h bounded on $\partial\Omega^*\cap N(\mathcal{X},\tilde{\delta})\times(-1,1)\times(0,\tilde{\delta})$. It then follows that

$$\delta^{-1}E[p^A(X_i;\delta)^l g(X_i)1\{p^A(X_i;\delta)\in(0,1)\}]=\int_{-1}^{1}k(v)^l dv\int_{\partial\Omega^*}g(u)f_X(u)d\mathcal{H}^{p-1}(u)+O(\delta).$$

Also,

$$\begin{aligned}
&\delta^{-1}E[Z_i p^A(X_i;\delta)^l g(X_i)1\{p^A(X_i;\delta)\in(0,1)\}]\\
=&\int_{0}^{1}\int_{\partial\Omega^*\cap N(\mathcal{X},\tilde{\delta})}(k(v)+\delta r(u,v,\delta))^l g(u+\delta v\nu_{\Omega^*}(u))f_X(u+\delta v\nu_{\Omega^*}(u))J_{p-1}^{\partial\Omega^*}\psi_{\Omega^*}(u,\delta v)d\mathcal{H}^{p-1}(u)dv\\
=&\int_{0}^{1}k(v)^l dv\int_{\partial\Omega^*}g(u)f_X(u)d\mathcal{H}^{p-1}(u)+O(\delta).
\end{aligned}$$

□

Step C.3.3.4. *Let* $S_{\mathbf{D}}=\lim_{\delta\to0}\delta^{-1}E[\mathbf{Z}_i\mathbf{D}_i'1\{p^A(X_i;\delta)\in(0,1)\}]$ *and* $S_Y=\lim_{\delta\to0}\delta^{-1}E[\mathbf{Z}_iY_i1\{p^A(X_i;\delta)\in(0,1)\}]$. *Then the second element of* $S_{\mathbf{D}}^{-1}S_Y$ *is* β_1.

Proof. Note that $D_i=Z_iD_i(1)+(1-Z_i)D_i(0)$ and $Y_i=Z_iY_{1i}+(1-Z_i)Y_{0i}$. By Step C.3.3.3,

$$\begin{aligned}
&S_{\mathbf{D}}\\
=&\begin{bmatrix}
2\bar{f}_X & \int_{\partial\Omega^*}E[D_i(1)+D_i(0)|X_i=x]f_X(x)d\mathcal{H}^{p-1}(x) & \int_{-1}^{1}k(v)dv\bar{f}_X\\
\bar{f}_X & \int_{\partial\Omega^*}E[D_i(1)|X_i=x]f_X(x)d\mathcal{H}^{p-1}(x) & \int_{0}^{1}k(v)dv\bar{f}_X\\
\int_{-1}^{1}k(v)dv\bar{f}_X & \int_{\partial\Omega^*}(\int_{0}^{1}k(v)dvE[D_i(1)|X_i=x]+\int_{-1}^{0}k(v)dvE[D_i(0)|X_i=x])f_X(x)d\mathcal{H}^{p-1}(x) & \int_{-1}^{1}k(v)^2dv\bar{f}_X
\end{bmatrix},
\end{aligned}$$

where $\bar{f}_X=\int_{\partial\Omega^*}f_X(x)d\mathcal{H}^{p-1}(x)$, and

$$S_Y=\begin{bmatrix}
\int_{\partial\Omega^*}E[Y_{1i}+Y_{0i}|X_i=x]f_X(x)d\mathcal{H}^{p-1}(x)\\
\int_{\partial\Omega^*}E[Y_{1i}|X_i=x]f_X(x)d\mathcal{H}^{p-1}(x)\\
\int_{\partial\Omega^*}(\int_{0}^{1}k(v)dvE[Y_{1i}|X_i=x]+\int_{-1}^{0}k(v)dvE[Y_{0i}|X_i=x])f_X(x)d\mathcal{H}^{p-1}(x)
\end{bmatrix}.$$

After a few lines of algebra, we have

$$\begin{aligned}
\det(S_{\mathbf{D}})=&\bar{f}_X^{-2}\int_{\partial\Omega^*}E[D_i(1)-D_i(0)|X_i=x]f_X(x)d\mathcal{H}^{p-1}(x)\\
&\times(\int_{-1}^{0}(k(v)-\int_{-1}^{0}k(s)ds)^2dv+\int_{0}^{1}(k(v)-\int_{0}^{1}k(s)ds)^2dv),
\end{aligned}$$

which is nonzero under Assumption 3 (b) and (f) (i). After another few lines of algebra, we obtain that the second element of $S_{\mathbf{D}}^{-1}S_Y$ is

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

On the other hand, by Step C.3.3.3,

$$\begin{aligned}
\beta_1 &= \lim_{\delta\to 0} E[\omega_i(\delta)(Y_i(1)-Y_i(0))] \\
&= \lim_{\delta\to 0} \frac{\delta^{-1}E[p^A(X_i;\delta)(1-p^A(X_i;\delta))(D_i(1)-D_i(0))(Y_i(1)-Y_i(0))1\{p^A(X_i;\delta)\in(0,1)\}]}{\delta^{-1}E[p^A(X_i;\delta)(1-p^A(X_i;\delta))(D_i(1)-D_i(0))1\{p^A(X_i;\delta)\in(0,1)\}]} \\
&= \frac{\int_{-1}^1 k(v)(1-k(v))dv\int_{\partial\Omega^*} E[(D_i(1)-D_i(0))(Y_i(1)-Y_i(0))|X_i=x]f_X(x)d\mathcal{H}^{p-1}(x)}{\int_{-1}^1 k(v)(1-k(v))dv\int_{\partial\Omega^*} E[D_i(1)-D_i(0)|X_i=x]f_X(x)d\mathcal{H}^{p-1}(x)} \\
&= \frac{\int_{\partial\Omega^*} E[(D_i(1)-D_i(0))(Y_i(1)-Y_i(0))|X_i=x]f_X(x)d\mathcal{H}^{p-1}(x)}{\int_{\partial\Omega^*} E[D_i(1)-D_i(0)|X_i=x]f_X(x)d\mathcal{H}^{p-1}(x)}.
\end{aligned}$$

□

Step C.3.3.5. *If $n\delta_n\to\infty$ as $n\to\infty$, then $\hat{\beta}_1 \xrightarrow{p} \beta_1$.*

Proof. It suffices to verify that the variance of each element of $\frac{1}{n\delta_n}\sum_{i=1}^n \mathbf{Z}_i\mathbf{D}_i'I_i$ and $\frac{1}{n\delta_n}\sum_{i=1}^n \mathbf{Z}_iYI_i$ is $o(1)$. Here, we only verify that $\mathrm{Var}(\frac{1}{n\delta_n}\sum_{i=1}^n p^A(X_i;\delta_n)Y_iI_i)=o(1)$. Note that

$$E[Y_i^2|X_i] = E[Z_iY_{1i}^2+(1-Z_i)Y_{0i}^2|X_i] \le E[Y_{1i}^2+Y_{0i}^2|X_i].$$

Under Assumption 3 (g), there exists $\delta'>0$ such that $E[Y_{1i}^2+Y_{0i}^2|X_i]$ is continuous on $N(\partial\Omega^*,\delta')$. Since $\mathrm{cl}(N(\partial\Omega^*,\frac{1}{2}\delta'))$ is closed and bounded, $E[Y_{1i}^2+Y_{0i}^2|X_i]$ is bounded on $\mathrm{cl}(N(\partial\Omega^*,\frac{1}{2}\delta'))$. We have

$$\begin{aligned}
\mathrm{Var}(\frac{1}{n\delta_n}\sum_{i=1}^n p^A(X_i;\delta_n)Y_iI_i) &\le \frac{1}{n\delta_n}\delta_n^{-1}E[p^A(X_i;\delta_n)^2Y_i^2I_i] \\
&= \frac{1}{n\delta_n}\delta_n^{-1}E[p^A(X_i;\delta_n)^2E[Y_i^2|X_i]I_i] \\
&\le \frac{1}{n\delta_n}C
\end{aligned}$$

for some $C>0$, where the last inequality follows from Step C.3.3.3. The conclusion follows since $n\delta_n\to\infty$. □

Now let $\beta=(\beta_0,\beta_1,\beta_2)'=S_{\mathbf{D}}^{-1}S_Y$ and let $\epsilon_i=Y_i-\mathbf{D}_i'\beta$. We can write

$$\begin{aligned}
\sqrt{n\delta_n}(\hat{\beta}-\beta) &= (\frac{1}{n\delta_n}\sum_{i=1}^n \mathbf{Z}_i\mathbf{D}_i'I_i)^{-1}\frac{1}{\sqrt{n\delta_n}}\sum_{i=1}^n \mathbf{Z}_i\epsilon_iI_i \\
&= (\frac{1}{n\delta_n}\sum_{i=1}^n \mathbf{Z}_i\mathbf{D}_i'I_i)^{-1}\frac{1}{\sqrt{n\delta_n}}\sum_{i=1}^n \{(\mathbf{Z}_i\epsilon_iI_i - E[\mathbf{Z}_i\epsilon_iI_i]) + E[\mathbf{Z}_i\epsilon_iI_i]\}.
\end{aligned}$$

Step C.3.3.6.

$$\frac{1}{\sqrt{n\delta_n}}\sum_{i=1}^{n}(\mathbf{Z}_i\epsilon_i I_i - E[\mathbf{Z}_i\epsilon_i I_i]) \xrightarrow{d} \mathcal{N}(0,\mathbf{V}),$$

where $\mathbf{V} = \lim_{n\to\infty}\delta_n^{-1}E[\epsilon_i^2\mathbf{Z}_i\mathbf{Z}_i I_i]$.

Proof. We use the triangular-array Lyapunov CLT and the Cramér-Wold device. Pick a nonzero $\lambda \in \mathbb{R}^p$, and let $V_{i,n} = \frac{1}{\sqrt{n\delta_n}}\lambda'(\mathbf{Z}_i\epsilon_i I_i - E[\mathbf{Z}_i\epsilon_i I_i])$. First,

$$\sum_{i=1}^{n}E[V_{i,n}^2] = \delta_n^{-1}\lambda'(E[\epsilon_i^2\mathbf{Z}_i\mathbf{Z}_i' I_i] - E[\mathbf{Z}_i\epsilon_i I_i]E[\mathbf{Z}_i'\epsilon_i I_i])\lambda.$$

By Step C.3.3.3,

$$E[\mathbf{Z}_i\epsilon_i I_i] = E[\mathbf{Z}_i(Y_i - \mathbf{D}_i'\beta)I_i] = O(\delta_n),$$

so

$$\delta_n^{-1}E[\mathbf{Z}_i\epsilon_i I_i]E[\mathbf{Z}_i'\epsilon_i I_i] = o(1).$$

We have

$$\begin{aligned}
E[\epsilon_i^2\mathbf{Z}_i\mathbf{Z}_i' I_i] &= E[(Y_i - \beta_0 - \beta_1 D_i - \beta_2 p^A(X_i;\delta_n))^2\mathbf{Z}_i\mathbf{Z}_i' I_i] \\
&= E[Z_i(Y_{1i} - \beta_0 - \beta_1 D_i(1) - \beta_2 p^A(X_i;\delta_n))^2\mathbf{Z}_i\mathbf{Z}_i' I_i] \\
&\quad + E[(1-Z_i)(Y_{0i} - \beta_0 - \beta_1 D_i(0) - \beta_2 p^A(X_i;\delta_n))^2\mathbf{Z}_i\mathbf{Z}_i' I_i].
\end{aligned}$$

Since $E[Y_{1i}|X_i]$, $E[Y_{0i}|X_i]$, $E[D_i(1)|X_i]$, $E[D_i(0)|X_i]$, $E[Y_{1i}^2|X_i]$, $E[Y_{0i}^2|X_i]$, $E[Y_{1i}D_i(1)|X_i]$ and $E[Y_{0i}D_i(0)|X_i]$ are continuous on $N(\partial\Omega^*,\delta')$ for some $\delta' > 0$ under Assumption 3 (g), $\lim_{n\to\infty}\delta_n^{-1}E[\epsilon_i^2\mathbf{Z}_i\mathbf{Z}_i' I_i]$ exists and finite. Therefore,

$$\sum_{i=1}^{n}E[V_{i,n}^2] \to \lambda'\mathbf{V}\lambda < 0.$$

We next verify the Lyapunov condition: for some $t > 0$,

$$\sum_{i=1}^{n}E[|V_{i,n}|^{2+t}] \to 0.$$

We have

$$\begin{aligned}
\sum_{i=1}^{n}E[|V_{i,n}|^4] &= \frac{1}{n\delta_n}\delta_n^{-1}E[|\lambda'(\mathbf{Z}_i\epsilon_i I_i - E[\mathbf{Z}_i\epsilon_i I_i])|^4] \\
&\le \frac{1}{n\delta_n}2^{3c}\delta_n^{-1}\{E[|\lambda'\mathbf{Z}_i\epsilon_i I_i|^4] + |\lambda'E[\mathbf{Z}_i\epsilon_i I_i]|^4\}
\end{aligned}$$

by the c_r-inequality. Repeating using the c_r-inequality gives

$$\begin{aligned}
\delta_n^{-1}E[|\lambda'\mathbf{Z}_i\epsilon_i I_i|^4] &= \delta_n^{-1}E[|\lambda'\mathbf{Z}_i(Y_i - \beta_0 - \beta_1 D_i - \beta_2 p^A(X_i;\delta_n))|^4 I_i] \\
&\le 2^{3c}\delta_n^{-1}E[(|\lambda'\mathbf{Z}_i|^4)(|Y_i|^4 + |\beta_0|^4 + |\beta_1|^4 D_i + |\beta_2|^4 p^A(X_i;\delta_n)^4)I_i] \\
&\le 2^{3c}(\lambda_1+\lambda_2+\lambda_3)^4\delta_n^{-1}E[(Y_i^4 + \beta_0^4 + \beta_1^4 + \beta_2^4)I_i] \\
&= 2^{3c}O(1)
\end{aligned}$$

for some finite constant c, where the last equality holds by Step C.3.3.3 under Assumption 3 (g). Moreover,

$$\begin{aligned}\delta_n^{-1}|\lambda' E[\mathbf{Z}_i\epsilon_i I_i]|^4 &= \delta_n^3|\lambda'\delta_n^{-1}E[\mathbf{Z}_i\epsilon_i I_i]|^4 \\ &= \delta_n^3 O(1) \\ &= o(1).\end{aligned}$$

Therefore, when $n\delta_n \to \infty$,

$$\sum_{i=1}^n E[|V_{i,n}|^4] \to 0,$$

and the conclusion follows from the Lyapunov CLT and the Cramér-Wold device. □

Step C.3.3.7. $n\delta_n\hat{\mathbf{\Sigma}} \xrightarrow{p} S_{\mathbf{D}}^{-1}\mathbf{V}(S_{\mathbf{D}}')^{-1}$.

Proof. We have

$$\begin{aligned}\frac{1}{n\delta_n}\sum_{i=1}^n \hat{\epsilon}_i^2\mathbf{Z}_i\mathbf{Z}_i'I_i =& \frac{1}{n\delta_n}\sum_{i=1}^n (Y_i - \mathbf{D}_i'\hat{\beta})^2\mathbf{Z}_i\mathbf{Z}_i'I_i \\ =& \frac{1}{n\delta_n}\sum_{i=1}^n (\epsilon_i - \mathbf{D}_i'(\hat{\beta}-\beta))^2\mathbf{Z}_i\mathbf{Z}_i'I_i \\ =& \frac{1}{n\delta_n}\sum_{i=1}^n \epsilon_i^2\mathbf{Z}_i\mathbf{Z}_i'I_i \\ & - \frac{2}{n\delta_n}\sum_{i=1}^n (Y_i - \mathbf{D}_i'\beta)((\hat{\beta}_0-\beta_0) + D_i(\hat{\beta}_1-\beta_1) + p^A(X_i;\delta_n)(\hat{\beta}_2-\beta_2))\mathbf{Z}_i\mathbf{Z}_i'I_i \\ & + \frac{1}{n\delta_n}\sum_{i=1}^n ((\hat{\beta}_0-\beta_0) + D_i(\hat{\beta}_1-\beta_1) + p^A(X_i;\delta_n)(\hat{\beta}_2-\beta_2))^2\mathbf{Z}_i\mathbf{Z}_i'I_i \\ =& \frac{1}{n\delta_n}\sum_{i=1}^n \epsilon_i^2\mathbf{Z}_i\mathbf{Z}_i'I_i + o_p(1)O_p(1),\end{aligned}$$

where the last equality follows from the result that $\hat{\beta} - \beta = o_p(1)$ and from application of Step C.3.3.3 as in Steps C.3.3.5 and C.3.3.6. To show $\frac{1}{n\delta_n}\sum_{i=1}^n \epsilon_i^2\mathbf{Z}_i\mathbf{Z}_i'I_i \xrightarrow{p} \mathbf{V}$, it suffices to verify that the variance of each element of $\frac{1}{n\delta_n}\sum_{i=1}^n \epsilon_i^2\mathbf{Z}_i\mathbf{Z}_i'I_i$ is $o(1)$. We only verify that

$\mathrm{Var}(\frac{1}{n\delta_n}\sum_{i=1}^n \epsilon_i^2 p^A(X_i;\delta_n)^2 I_i) = o(1)$. Using the c_r-inequality, we have that for some constant c,

$$\begin{aligned}
\mathrm{Var}(\frac{1}{n\delta_n}\sum_{i=1}^n \epsilon_i^2 p^A(X_i;\delta_n)^2 I_i) &\leq \frac{1}{n\delta_n}\delta_n^{-1}E[\epsilon_i^4 I_i] \\
&= \frac{1}{n\delta_n}\delta_n^{-1}E[(Y_i - \beta_0 - \beta_1 D_i - \beta_2 p^A(X_i))^4 I_i] \\
&\leq \frac{1}{n\delta_n}2^{3c}\delta_n^{-1}E[(Y_i^4 + \beta_0^4 + \beta_1^4 D_i + \beta_2^4 p^A(X_i)^4) I_i] \\
&\leq \frac{1}{n\delta_n}2^{3c}\delta_n^{-1}E[(Y_i^4 + \beta_0^4 + \beta_1^4 + \beta_2^4) I_i] \\
&= \frac{1}{n\delta_n}2^{3c}O(1) \\
&= o(1),
\end{aligned}$$

where the second last equality holds by Step C.3.3.3 under Assumption 3 (g). Therefore,

$$\frac{1}{n\delta_n}\sum_{i=1}^n \hat{\epsilon}_i^2 \mathbf{Z}_i\mathbf{Z}_i' I_i \xrightarrow{p} \mathbf{V}.$$

It follows that

$$n\delta_n\hat{\mathbf{\Sigma}} = (\frac{1}{n\delta_n}\sum_{i=1}^n \mathbf{Z}_i\mathbf{D}_i' I_i)^{-1}(\frac{1}{n\delta_n}\sum_{i=1}^n \hat{\epsilon}_i^2\mathbf{Z}_i\mathbf{Z}_i' I_i)(\frac{1}{n\delta_n}\sum_{i=1}^n \mathbf{D}_i\mathbf{Z}_i' I_i)^{-1} \xrightarrow{p} S_{\mathbf{D}}^{-1}\mathbf{V}(S_{\mathbf{D}}')^{-1}.$$

□

Step C.3.3.8. $\hat{\sigma}^{-1}(\hat{\beta}_1 - \beta_1) \xrightarrow{d} \mathcal{N}(0,1)$.

Proof. Let $\beta_n = S_{\mathbf{D}}^{-1}\delta_n^{-1}E[\mathbf{Z}_i Y_i I_i]$. We then have

$$\begin{aligned}
\frac{1}{\sqrt{n\delta_n}}\sum_{i=1}^n E[\mathbf{Z}_i\epsilon_i I_i] &= \sqrt{n\delta_n}\delta_n^{-1}E[\mathbf{Z}_i(Y_i - \mathbf{D}'\beta)I_i] \\
&= \sqrt{n\delta_n}\delta_n^{-1}E[\mathbf{Z}_i(Y_i - \mathbf{D}_i'\beta_n + \mathbf{D}_i'(\beta_n - \beta))I_i] \\
&= \sqrt{n\delta_n}\delta_n^{-1}\{E[\mathbf{Z}_i Y_i I_i] - E[\mathbf{Z}_i\mathbf{D}_i' I_i]\beta_n + E[\mathbf{Z}_i\mathbf{D}_i' I_i](\beta_n - \beta)\} \\
&= \sqrt{n\delta_n}\{(S_{\mathbf{D}} - \delta_n^{-1}E[\mathbf{Z}_i\mathbf{D}_i' I_i])S_{\mathbf{D}}^{-1}\delta_n^{-1}E[\mathbf{Z}_i Y_i I_i] \\
&\qquad + \delta_n^{-1}E[\mathbf{Z}_i\mathbf{D}_i' I_i]S_{\mathbf{D}}^{-1}(\delta_n^{-1}E[\mathbf{Z}_i Y_i I_i] - S_Y)\} \\
&= \sqrt{n\delta_n}(O(\delta_n)O(1) + O(1)O(\delta_n)) \\
&= O(\sqrt{n\delta_n}\delta_n),
\end{aligned}$$

where we use Step C.3.3.3 for the second last equality. Thus, when $n\delta_n^3 \to 0$,

$$\begin{aligned}
\sqrt{n\delta_n}(\hat{\beta} - \beta) &= (\frac{1}{n\delta_n}\sum_{i=1}^n \mathbf{Z}_i\mathbf{D}_i' I_i)^{-1}\frac{1}{\sqrt{n\delta_n}}\sum_{i=1}^n\{(\mathbf{Z}_i\epsilon_i I_i - E[\mathbf{Z}_i\epsilon_i I_i]) + E[\mathbf{Z}_i\epsilon_i I_i]\} \\
&\xrightarrow{d} \mathcal{N}(0, S_{\mathbf{D}}^{-1}\mathbf{V}(S_{\mathbf{D}}')^{-1}).
\end{aligned}$$

The conclusion then follows from Step C.3.3.7. □

□

C.3.4 Consistency and Asymptotic Normality of $\hat{\beta}_1^s$ When $\Pr(A(X_i) \in (0,1)) = 0$

Let $I_i^s = 1\{p^s(X_i;\delta_n) \in (0,1)\}$, $\mathbf{D}_i^s = (1, D_i, p^s(X_i;\delta_n))'$ and $\mathbf{Z}_i^s = (1, Z_i, p^s(X_i;\delta_n))'$. $\hat{\beta}^s$ and $\hat{\mathbf{\Sigma}}^s$ are given by

$$\hat{\beta}^s = (\sum_{i=1}^n \mathbf{Z}_i^s (\mathbf{D}_i^s)' I_i^s)^{-1} \sum_{i=1}^n \mathbf{Z}_i^s Y_i I_i^s.$$

and

$$\hat{\mathbf{\Sigma}}^s = (\sum_{i=1}^n \mathbf{Z}_i^s (\mathbf{D}_i^s)' I_i^s)^{-1} (\sum_{i=1}^n (\hat{\epsilon}_i^s)^2 \mathbf{Z}_i^s (\mathbf{Z}_i^s)' I_i^s) (\sum_{i=1}^n \mathbf{D}_i^s (\mathbf{Z}_i^s)' I_i^s)^{-1},$$

where $\hat{\epsilon}_i^s = Y_i - (\mathbf{D}_i^s)'\hat{\beta}^s$. It is sufficient to show that

$$\hat{\beta}^s - \hat{\beta} = o_p(1),$$

if $S_n \to \infty$ and that

$$\sqrt{n\delta_n}(\hat{\beta}^s - \hat{\beta}) = o_p(1),$$
$$n\delta_n \hat{\mathbf{\Sigma}}^s \xrightarrow{p} S_{\mathbf{D}}^{-1} \mathbf{V} (S_{\mathbf{D}}')^{-1}$$

if Assumption 5 holds.

Step C.3.4.1. *Let $\{V_i\}_{i=1}^\infty$ be i.i.d. random variables. If $E[V_i|X_i]$ and $E[V_i^2|X_i]$ are bounded on $N(\partial\Omega^*, \delta') \cap N(\mathcal{X}, \delta')$ for some $\delta' > 0$, and $S_n \to \infty$, then*

$$\frac{1}{n\delta_n} \sum_{i=1}^n V_i p^s(X_i;\delta_n)^l I_i^s - \frac{1}{n\delta_n} \sum_{i=1}^n V_i p^A(X_i;\delta_n)^l I_i = o_p(1)$$

for $l = 0, 1, 2, 3, 4$. If, in addition, Assumption 5 holds, then

$$\frac{1}{\sqrt{n\delta_n}} \sum_{i=1}^n V_i p^s(X_i;\delta_n)^l I_i^s - \frac{1}{\sqrt{n\delta_n}} \sum_{i=1}^n V_i p^A(X_i;\delta_n)^l I_i = o_p(1)$$

for $l = 0, 1, 2$.

Proof. We have

$$\begin{aligned}
&\frac{1}{n\delta_n} \sum_{i=1}^n V_i p^s(X_i;\delta_n)^l I_i^s - \frac{1}{n\delta_n} \sum_{i=1}^n V_i p^A(X_i;\delta_n)^l I_i \\
=&\frac{1}{n\delta_n} \sum_{i=1}^n V_i p^s(X_i;\delta_n)^l (I_i^s - I_i) + \frac{1}{n\delta_n} \sum_{i=1}^n V_i (p^s(X_i;\delta_n)^l - p^A(X_i;\delta_n)^l) I_i.
\end{aligned}$$

We first consider $\frac{1}{n\delta_n}\sum_{i=1}^n V_i(p^s(X_i;\delta_n)^l - p^A(X_i;\delta_n)^l)I_i$. By using the argument in the proof of Step C.3.3.3 in Section C.3.3, we have

$$
\begin{aligned}
&|E[\frac{1}{n\delta_n}\sum_{i=1}^n V_i(p^s(X_i;\delta_n)^l - p^A(X_i;\delta_n)^l)I_i]| \\
&= \delta_n^{-1}|E[E[V_i|X_i]E[p^s(X_i;\delta_n)^l - p^A(X_i;\delta_n)^l|X_i]I_i]| \\
&\le \delta_n^{-1}E[|E[V_i|X_i]||E[p^s(X_i;\delta_n)^l - p^A(X_i;\delta_n)^l|X_i]|I_i] \\
&= \int_{-1}^1\int_{\partial\Omega^*\cap N(\mathcal{X},\tilde{\delta})} |E[V_i|X_i = u+\delta_n v\nu_{\Omega^*}(u)]||E[p^s(u+\delta_n v\nu_{\Omega^*}(u);\delta_n)^l - p^A(u+\delta_n v\nu_{\Omega^*}(u);\delta_n)^l]| \\
&\qquad\qquad \times f_X(u+\delta_n v\nu_{\Omega^*}(u))J_{p-1}^{\partial\Omega^*}\psi_{\Omega^*}(u,\delta_n v)d\mathcal{H}^{p-1}(u)dv,
\end{aligned}
$$

where the choice of $\tilde{\delta}$ is as in the proof of Step C.3.3.3. By Lemma B.7, for $l = 0, 1, 2$,

$$
\begin{aligned}
&|E[\frac{1}{n\delta_n}\sum_{i=1}^n V_i(p^s(X_i;\delta_n)^l - p^A(X_i;\delta_n)^l)I_i]| \\
&\le \frac{1}{S_n}\int_{-1}^1\int_{\partial\Omega^*\cap N(\mathcal{X},\tilde{\delta})} |E[V_i|X_i = u+\delta_n v\nu_{\Omega^*}(u)]|f_X(u+\delta_n v\nu_{\Omega^*}(u))J_{p-1}^{\partial\Omega^*}\psi_{\Omega^*}(u,\delta_n v)d\mathcal{H}^{p-1}(u)dv \\
&= O(S_n^{-1}).
\end{aligned}
$$

Also, by Lemma B.7,

$$
\begin{aligned}
&|E[\frac{1}{n\delta_n}\sum_{i=1}^n V_i(p^s(X_i;\delta_n)^3 - p^A(X_i;\delta_n)^3)I_i]| \\
&= |\delta_n^{-1}E[V_i(p^s(X_i;\delta_n) - p^A(X_i;\delta_n))(p^s(X_i;\delta_n)^2 + p^s(X_i;\delta_n)p^A(X_i;\delta_n) + p^A(X_i;\delta_n)^2)I_i]| \\
&\le \delta_n^{-1}E[|E[V_i|X_i]||E[(p^s(X_i;\delta_n) - p^A(X_i;\delta_n))(p^s(X_i;\delta_n)^2 + p^s(X_i;\delta_n)p^A(X_i;\delta_n) + p^A(X_i;\delta_n)^2)|X_i]|I_i] \\
&\le 3\delta_n^{-1}E[|E[V_i|X_i]|E[|p^s(X_i;\delta_n) - p^A(X_i;\delta_n)||X_i]I_i] \\
&= \int_{-1}^1\int_{\partial\Omega^*\cap N(\mathcal{X},\tilde{\delta})} |E[V_i|X_i = u+\delta_n v\nu_{\Omega^*}(u)]|E[|p^s(u+\delta_n v\nu_{\Omega^*}(u);\delta_n) - p^A(u+\delta_n v\nu_{\Omega^*}(u);\delta_n)|] \\
&\qquad\qquad \times f_X(u+\delta_n v\nu_{\Omega^*}(u))J_{p-1}^{\partial\Omega^*}\psi_{\Omega^*}(u,\delta_n v)d\mathcal{H}^{p-1}(u)dv \\
&\le (\frac{1}{S_n\epsilon^2} + \epsilon)O(1)
\end{aligned}
$$

for every $\epsilon > 0$. We can make the right-hand side arbitrarily close to zero by taking sufficiently small $\epsilon > 0$ and sufficiently large S_n, which implies that $|E[\frac{1}{n\delta_n}\sum_{i=1}^n V_i(p^s(X_i;\delta_n)^3 - p^A(X_i;\delta_n)^3)I_i]| = o(1)$ if $S_n \to \infty$. Likewise,

$$
\begin{aligned}
&|E[\frac{1}{n\delta_n}\sum_{i=1}^n V_i(p^s(X_i;\delta_n)^4 - p^A(X_i;\delta_n)^4)I_i]| \\
&= |\delta_n^{-1}E[V_i(p^s(X_i;\delta_n)^2 + p^A(X_i;\delta_n)^2)(p^s(X_i;\delta_n) + p^A(X_i;\delta_n))(p^s(X_i;\delta_n) - p^A(X_i;\delta_n))I_i]| \\
&\le \delta_n^{-1}E[|E[V_i|X_i]||E[(p^s(X_i;\delta_n)^2 + p^A(X_i;\delta_n)^2)(p^s(X_i;\delta_n) + p^A(X_i;\delta_n))(p^s(X_i;\delta_n) - p^A(X_i;\delta_n))|X_i]|I_i] \\
&\le 8\delta_n^{-1}E[|E[V_i|X_i]|E[|p^s(X_i;\delta_n) - p^A(X_i;\delta_n)||X_i]I_i] \\
&= o(1).
\end{aligned}
$$

As for variance, for $l = 0, 1, 2$,

$$\begin{aligned}\mathrm{Var}(\frac{1}{n\delta_n}\sum_{i=1}^n V_i(p^s(X_i;\delta_n)^l - p^A(X_i;\delta_n)^l)I_i) &\le \frac{1}{n\delta_n}\delta_n^{-1}E[V_i^2(p^s(X_i;\delta_n)^l - p^A(X_i;\delta_n)^l)^2 I_i] \\ &\le \frac{1}{n\delta_n}\delta_n^{-1}E[E[V_i^2|X_i]E[(p^s(X_i;\delta_n)^l - p^A(X_i;\delta_n)^l)^2|X_i]I_i] \\ &\le \frac{4}{n\delta_n S_n}\delta_n^{-1}E[E[V_i^2|X_i]I_i] \\ &= O((n\delta_n S_n)^{-1}),\end{aligned}$$

and for $l = 3, 4$,

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

Therefore, $\frac{1}{n\delta_n}\sum_{i=1}^n V_i(p^s(X_i;\delta_n)^l - p^A(X_i;\delta_n)^l)I_i = o_p(1)$ if $S_n \to \infty$ for $l = 0, 1, 2, 3, 4$, and $\frac{1}{\sqrt{n\delta_n}}\sum_{i=1}^n V_i(p^s(X_i;\delta_n)^l - p^A(X_i;\delta_n)^l)I_i = o_p(1)$ if $n^{-1/2}S_n \to \infty$ for $l = 0, 1, 2$.

We next show that $\frac{1}{n\delta_n}\sum_{i=1}^n V_i p^s(X_i;\delta_n)^l(I_i^s - I_i) = o_p(1)$ if $S_n \to \infty$ for $l \ge 0$. We have

$$\begin{aligned}|E[\frac{1}{n\delta_n}\sum_{i=1}^n V_i p^s(X_i;\delta_n)^l(I_i^s - I_i)]| &= \delta_n^{-1}|E[V_i p^s(X_i;\delta_n)^l(I_i^s - I_i)]| \\ &\le \delta_n^{-1}E[|E[V_i|X_i]||E[p^s(X_i;\delta_n)^l(I_i^s - I_i)|X_i]|] \\ &= \delta_n^{-1}E[|E[V_i|X_i]|E[|I_i^s - I_i||X_i]].\end{aligned}$$

Since $I_i^s - I_i \le 0$ with strict inequality only if $I_i = 1$,

$$E[|I_i^s - I_i||X_i] = -E[I_i^s - I_i|X_i]I_i = (1 - E[I_i^s|X_i])I_i = \Pr(p^s(X_i;\delta_n) \in \{0,1\}|X_i)I_i.$$

We then have

$$\begin{aligned}&|E[\frac{1}{n\delta_n}\sum_{i=1}^n V_i p^s(X_i;\delta_n)^l(I_i^s - I_i)]| \\ &\le \delta_n^{-1}E[|E[V_i|X_i]|\Pr(p^s(X_i;\delta_n) \in \{0,1\}|X_i)I_i] \\ &\le \delta_n^{-1}E[|E[V_i|X_i]|((1 - p^A(X_i;\delta_n))^{S_n} + p^A(X_i;\delta_n)^{S_n})I_i] \\ &\le \int_{-1}^1 \int_{\partial\Omega^* \cap N(\mathcal{X},\tilde{\delta})} |E[V_i|X_i = u + \delta_n v\nu_{\Omega^*}(u)]|\{(1 - p^A(u + \delta_n v\nu_{\Omega^*}(u);\delta_n))^{S_n} \\ &\qquad + p^A(u + \delta_n v\nu_{\Omega^*}(u);\delta_n)^{S_n}\} f_X(u + \delta_n v\nu_{\Omega^*}(u)) J_{p-1}^{\partial\Omega^*}\psi_{\Omega^*}(u, \delta_n v)d\mathcal{H}^{p-1}(u)dv,\end{aligned}$$

where the second inequality follows from Lemma B.7. Note that for every $(u, v) \in \partial\Omega^* \cap N(\mathcal{X}, \tilde{\delta}) \times (-1, 1)$, $\lim_{\delta\to 0} p^A(u + \delta_n v\nu_{\Omega^*}(u);\delta_n) = k(v) \in (0, 1)$ by Step C.3.3.1 in Section C.3.3. Since $E[V_i|X_i]$, f_X and $J_{p-1}^{\partial\Omega^*}\psi_{\Omega^*}$ are bounded, by the Bounded Convergence Theorem,

$$|E[\frac{1}{n\delta_n}\sum_{i=1}^n V_i p^s(X_i;\delta_n)^l(I_i^s - I_i)]| = o(1)$$

if $S_n \to \infty$.

As for variance,

$$\begin{aligned}
\mathrm{Var}(\frac{1}{n\delta_n}\sum_{i=1}^{n} V_i p^s(X_i;\delta_n)^l(I_i^s - I_i)) &\leq \frac{1}{n\delta_n}\delta_n^{-1}E[V_i^2 p^s(X_i;\delta_n)^{2l}(I_i^s - I_i)^2] \\
&\leq \frac{1}{n\delta_n}\delta_n^{-1}E[V_i^2|I_i^s - I_i|] \\
&= \frac{1}{n\delta_n}\delta_n^{-1}E[E[V_i^2|X_i]E[|I_i^s - I_i||X_i]] \\
&= o(1).
\end{aligned}$$

Lastly, we show that, for $l \geq 0$, $\frac{1}{\sqrt{n\delta_n}}\sum_{i=1}^{n} V_i p^s(X_i;\delta_n)^l(I_i^s - I_i) = o_p(1)$ if Assumption 5 holds. Let $\eta_n = \gamma\frac{\log n}{S_n}$, where γ is the one satisfying Assumption 5. We have

$$\begin{aligned}
&|E[\frac{1}{\sqrt{n\delta_n}}\sum_{i=1}^{n} V_i p^s(X_i;\delta_n)^l(I_i^s - I_i)]| \\
&\leq \sqrt{n\delta_n^{-1}}E[|E[V_i|X_i]|((1-p^A(X_i;\delta_n))^{S_n} + p^A(X_i;\delta_n)^{S_n})I_i] \\
&= \sqrt{n\delta_n^{-1}}E[|E[V_i|X_i]|((1-p^A(X_i;\delta_n))^{S_n} + p^A(X_i;\delta_n)^{S_n}))1\{p^A(X_i;\delta_n) \in (0,\eta_n)\cup(1-\eta_n,1)\}] \\
&\quad + \sqrt{n\delta_n^{-1}}E[|E[V_i|X_i]|((1-p^A(X_i;\delta_n))^{S_n} + p^A(X_i;\delta_n)^{S_n}))1\{p^A(X_i;\delta_n) \in (\eta_n,1-\eta_n)\}] \\
&\leq (\sup_{x\in N(\partial\Omega^*,2\tilde{\delta})\cap N(\mathcal{X},2\tilde{\delta})} |E[V_i|X_i = x]|)(\sqrt{n\delta_n^{-1}}\Pr(p^A(X_i;\delta_n) \in (0,\eta_n)\cup(1-\eta_n,1)) \\
&\qquad + 2\sqrt{n\delta_n}(1-\eta_n)^{S_n}\delta_n^{-1}E[1\{p^A(X_i;\delta_n) \in (\eta_n,1-\eta_n)\}]).
\end{aligned}$$

By Assumption 5, $\sqrt{n\delta_n^{-1}}\Pr(p^A(X_i;\delta_n) \in (0,\eta_n)\cup(1-\eta_n,1)) = o(1)$. For the second term,

$$\begin{aligned}
2\sqrt{n\delta_n}(1-\eta_n)^{S_n}\delta_n^{-1}E[1\{p^A(X_i;\delta_n) \in (\eta_n,1-\eta_n)\}] &\leq 2\sqrt{n\delta_n}(1-\eta_n)^{S_n}\delta_n^{-1}E[I_i] \\
&= 2\sqrt{n\delta_n}(1-\eta_n)^{S_n}O(1).
\end{aligned}$$

Observe that $\eta_n = \gamma\frac{\log n}{S_n} = \gamma\frac{\log n}{n^{1/2}}\frac{1}{n^{-1/2}S_n} \to 0$, since $n^{-1/2}S_n \to \infty$ and $\frac{\log n}{n^{1/2}} \to 0$. Using the fact that $e^t \geq 1+t$ for every $t \in \mathbb{R}$, we have

$$\begin{aligned}
\sqrt{n\delta_n}(1-\eta_n)^{S_n} &\leq \sqrt{n\delta_n}(e^{-\eta_n})^{S_n} \\
&= \sqrt{n\delta_n}e^{-\eta_n S_n} \\
&= \sqrt{n\delta_n}e^{-\gamma\log n} \\
&= \sqrt{n\delta_n}n^{-\gamma} \\
&= n^{1/2-\gamma}\delta_n^{1/2} \\
&\to 0,
\end{aligned}$$

since $\gamma > 1/2$. As for variance,

$$\begin{aligned}\mathrm{Var}(\frac{1}{\sqrt{n\delta_n}}\sum_{i=1}^n V_i p^s(X_i;\delta_n)^l(I_i^s - I_i)) &\leq \delta_n^{-1}E[V_i^2 p^s(X_i;\delta_n)^{2l}(I_i^s - I_i)^2] \\ &\leq \delta_n^{-1}E[E[V_i^2|X_i]E[|I_i^s - I_i||X_i]I_i] \\ &= o(1).\end{aligned}$$

□

We have

$$\begin{aligned}&\hat{\beta}^s - \hat{\beta} \\ &= (\frac{1}{n\delta_n}\sum_{i=1}^n \mathbf{Z}_i^s(\mathbf{D}_i^s)'I_i^s)^{-1}\frac{1}{n\delta_n}\sum_{i=1}^n \mathbf{Z}_i^s Y_i I_i^s - (\frac{1}{n\delta_n}\sum_{i=1}^n \mathbf{Z}_i\mathbf{D}_i'I_i)^{-1}\frac{1}{n\delta_n}\sum_{i=1}^n \mathbf{Z}_i Y_i I_i \\ &= (\frac{1}{n\delta_n}\sum_{i=1}^n \mathbf{Z}_i^s(\mathbf{D}_i^s)'I_i^s)^{-1}(\frac{1}{n\delta_n}\sum_{i=1}^n \mathbf{Z}_i^s Y_i I_i^s - \frac{1}{n\delta_n}\sum_{i=1}^n \mathbf{Z}_i Y_i I_i) \\ &\quad - (\frac{1}{n\delta_n}\sum_{i=1}^n \mathbf{Z}_i^s(\mathbf{D}_i^s)'I_i^s)^{-1}(\frac{1}{n\delta_n}\sum_{i=1}^n \mathbf{Z}_i^s(\mathbf{D}_i^s)'I_i^s - \frac{1}{n\delta_n}\sum_{i=1}^n \mathbf{Z}_i\mathbf{D}_i'I_i)(\frac{1}{n\delta_n}\sum_{i=1}^n \mathbf{Z}_i\mathbf{D}_i'I_i)^{-1}\frac{1}{n\delta_n}\sum_{i=1}^n \mathbf{Z}_i Y_i I_i.\end{aligned}$$

By Step C.3.4.1, $\hat{\beta}^s - \hat{\beta} = o_p(1)$ if $S_n \to \infty$, and $\sqrt{n\delta_n}(\hat{\beta}^s - \hat{\beta}) = o_p(1)$ if Assumption 5 holds.

By proceeding as in Step C.3.3.7 in Section C.3.3, we have

$$\frac{1}{n\delta_n}\sum_{i=1}^n(\hat{\epsilon}_i^s)^2\mathbf{Z}_i^s(\mathbf{Z}_i^s)'I_i^s = \frac{1}{n\delta_n}\sum_{i=1}^n(\epsilon_i^s)^2\mathbf{Z}_i^s(\mathbf{Z}_i^s)'I_i^s + o_p(1),$$

where $\epsilon_i^s = Y_i - (\mathbf{D}_i^s)'\beta$. Then, by Step C.3.4.1,

$$\begin{aligned}&\frac{1}{n\delta_n}\sum_{i=1}^n(\hat{\epsilon}_i^s)^2\mathbf{Z}_i^s(\mathbf{Z}_i^s)'I_i^s - \frac{1}{n\delta_n}\sum_{i=1}^n \epsilon_i^2\mathbf{Z}_i\mathbf{Z}_i'I_i \\ &= \frac{1}{n\delta_n}\sum_{i=1}^n(Y_i^2 - 2Y_i(\mathbf{D}_i^s)'\beta + \beta'\mathbf{D}_i^s(\mathbf{D}_i^s)'\beta)\mathbf{Z}_i^s(\mathbf{Z}_i^s)'I_i^s - \frac{1}{n\delta_n}\sum_{i=1}^n(Y_i^2 - 2Y_i\mathbf{D}_i'\beta + \beta'\mathbf{D}_i\mathbf{D}_i'\beta)\mathbf{Z}_i\mathbf{Z}_i'I_i + o_p(1) \\ &= o_p(1)\end{aligned}$$

so that

$$\frac{1}{n\delta_n}\sum_{i=1}^n(\hat{\epsilon}_i^s)^2\mathbf{Z}_i^s(\mathbf{Z}_i^s)'I_i^s \xrightarrow{p} \mathbf{V}.$$

Also, $\frac{1}{n\delta_n}\sum_{i=1}^n \mathbf{Z}_i^s(\mathbf{D}_i^s)'I_i^s \xrightarrow{p} S_{\mathbf{D}}$ by using Step C.3.4.1. The conclusion then follows. □

□

C.4 Proof of Proposition A.2

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

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

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

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

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

C.5 Proof of Corollary A.1

1. Suppose that A is continuous at $x \in \mathcal{X}$, and let $q = A(x)$. Then, by definition, $\mathcal{U}_{x,q} = B(\mathbf{0},1)$. By Proposition A.2, $p^A(x)$ exists, and $p^A(x) = q$. □

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

3. Let $\mathcal{N}$ be the neighborhood of x on which f is continuously differentiable. 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

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

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 $\mathcal{L}^p(\mathcal{U}_{x,q_1}) + \mathcal{L}^p(\mathcal{U}_{x,q_2}) \leq \mathcal{L}^p(B(\mathbf{0},1)) = 2V$, it follows that $\mathcal{L}^p(\mathcal{U}_{x,q_1}) = \mathcal{L}^p(\mathcal{U}_{x,q_2}) = V$. By Proposition A.2, $p^A(x)$ exists, and $p^A(x) = \frac{1}{2}(q_1 + q_2)$. □

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

C.6 Proof of Proposition A.1

Since A is a $\mathcal{L}^p$-measurable and bounded function, A is locally integrable with respect to the Lebesgue measure, i.e., for every ball $B \subset \mathbb{R}^p$, $\int_B A(x)dx$ exists. An application of the Lebesgue differentiation theorem (see e.g. Theorem 1.4 in Chapter 3 of Stein and Shakarchi (2005)) to the function A shows that

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

for almost every $x \in \mathbb{R}^p$. □

C.7 Proof of Proposition A.3

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 $x_d \in \mathcal{X}_d^S$ such that $\mathcal{L}^{p_c}(\{x_c \in \mathcal{X}_c^S(x_d) : p^A(x_d,x_c) \in \{0,1\}\}) > 0$. Without loss of generality, assume $\mathcal{L}^{p_c}(\{x_c \in \mathcal{X}_c^S(x_d) : p^A(x_d,x_c) = 1\}) > 0$. The proof proceeds in five steps.

Step C.7.1. $\mathcal{L}^{p_c}(\mathcal{X}_c^S(x_d) \cap \mathcal{X}_{c,1}(x_d)) > 0$.

Step C.7.2. $\mathcal{X}_c^S(x_d) \cap \text{int}(\mathcal{X}_{c,1}(x_d)) \neq \emptyset$.

Step C.7.3. $p^A(x_d,x_c) = 1$ *for any* $x_c \in \text{int}(\mathcal{X}_{c,1}(x_d))$.

Step C.7.4. *For every* $x_c^* \in \mathcal{X}_c^S(x_d) \cap \text{int}(\mathcal{X}_{c,1}(x_d))$*, there exists* $\delta > 0$ *such that* $B(x_c^*,\delta) \subset \mathcal{X}_c^S(x_d) \cap \text{int}(\mathcal{X}_{c,1}(x_d))$.

Step C.7.5. $E[Y_{1i} - Y_{0i}|X_i \in S]$ *is not identified.*

Following the argument in the proof of Proposition 1 (b), we can prove Steps C.7.1–C.7.3. Once Step C.7.4 is established, we prove Step C.7.5 by following the proof of Step C.1.4 in Proposition 1 (b) with $B(x_c^*,\delta)$ and $B(x_c^*,\epsilon)$ in place of $B(x^*,\delta)$ and $B(x^*,\epsilon)$, respectively, using the fact that $\Pr(X_{ci} \in B(x_c^*,\epsilon)|X_{di} = x_d) > 0$ by the definition of support. Here, we provide the proof of Step C.7.4.

Proof of Step C.7.4. Pick an $x_c^* \in \mathcal{X}_c^S(x_d) \cap \text{int}(\mathcal{X}_{c,1})$. Then, $x^* = (x_d, x_c^*) \in S$. Since S is open relative to $\mathcal{X}$, there exists an open set $U \in \mathbb{R}^p$ such that $S = U \cap \mathcal{X}$. This implies that for any sufficiently small $\delta > 0$, $B(x^*,\delta) \cap \mathcal{X} \subset U \cap \mathcal{X} = S$. It then follows that $\{x_c \in \mathbb{R}^{p_c} :$

$(x_d, x_c) \in B(x^*, \delta) \cap \mathcal{X}\} \subset \{x_c \in \mathbb{R}^{p_c} : (x_d, x_c) \in S\}$, equivalently, $B(x_c^*, \delta) \cap \mathcal{X}_c(x_d) \subset \mathcal{X}_c^S(x_d)$. By choosing a sufficiently small $\delta > 0$ so that $B(x_c^*, \delta) \subset \text{int}(\mathcal{X}_{c,1}(x_d)) \subset \mathcal{X}_c(x_d)$, we have $B(x_c^*, \delta) \subset \mathcal{X}_c^S(x_d) \cap \text{int}(\mathcal{X}_{c,1}(x_d))$. □

□

C.8 Proof of Theorem A.1

The proof is analogous to the proof of Theorem 1. The only difference is that, when we prove the convergence of expectations, we show the convergence of the expectations conditional on X_{di}, and then take the expectations over X_{di}. □

D Machine Learning Simulation: Details

Parameter Choice. For the variance-covariance matrix $\mathbf{\Sigma}$ of X_i, we first create a 100×100 symmetric matrix $\mathbf{V}$ such that the diagonal elements are one, $\mathbf{V}_{ij}$ is nonzero and equal to $\mathbf{V}_{ji}$ for $(i, j) \in \{2, 3, 4, 5, 6\} \times \{35, 66, 78\}$, and everything else is zero. We draw values from Unif$(-0.5, 0.5)$ independently for the nonzero off-diagonal elements of $\mathbf{V}$. We then create matrix $\mathbf{\Sigma} = \mathbf{V} \times \mathbf{V}$, which is a positive semidefinite matrix.

For α_0 and α_1, we first draw $\tilde{\alpha}_{0j}$, $j = 51, ..., 100$, from Unif$(-100, 100)$ independently across j, and draw $\tilde{\alpha}_{1j}$, $j = 1, ..., 100$, from Unif$(-150, 200)$ independently across j. We then set $\tilde{\alpha}_{0j} = \tilde{\alpha}_{1j}$ for $j = 1, ..., 50$, and calculate α_0 and α_1 by normalizing $\tilde{\alpha}_0$ and $\tilde{\alpha}_1$ so that $\text{Var}(X_i'\alpha_0) = \text{Var}(X_i'\alpha_1) = 1$.

Training of Prediction Model. We construct τ_{pred} using an independent sample $\{(\tilde{Y}_i, \tilde{X}_i, \tilde{D}_i, \tilde{Z}_i)\}_{i=1}^{\tilde{n}}$ of size $\tilde{n} = 2,000$. The distribution of $(\tilde{Y}_i, \tilde{X}_i, \tilde{D}_i, \tilde{Z}_i)$ is the same as that of (Y_i, X_i, D_i, Z_i) except (1) that $\tilde{Y}_i(1)$ is generated as $\tilde{Y}_i(1) = \tilde{Y}_i(0) + 0.5\tilde{X}_i'\alpha_1 + 0.5\epsilon_{1i}$, where $\epsilon_{1i} \sim \mathcal{N}(0, 1)$ and (2) that $\tilde{Z}_i \sim$ Bernoulli(0.5). This can be viewed as data from a past randomized experiment conducted to construct the algorithm.

We then use random forests separately for the subsamples with $\tilde{Z}_i = 1$ and $\tilde{Z}_i = 0$ to predict $\tilde{Y}_i$ from $\tilde{X}_i$. Let $\mu_z(x)$ be the trained prediction model. Set $\tau_{pred}(x) = \mu_1(x) - \mu_0(x)$. We generate the sample $\{(\tilde{Y}_i, \tilde{X}_i, \tilde{D}_i, \tilde{Z}_i)\}_{i=1}^{\tilde{n}}$ and construct τ_{pred} only once, and we use it for all of the 1,000 simulation samples. The distribution of the sample $\{(Y_i, X_i, D_i, Z_i)\}_{i=1}^{n}$ is thus held fixed for all simulations.

When training μ_z, we first randomly split the sample $\{(\tilde{Y}_i, \tilde{X}_i, \tilde{D}_i, \tilde{Z}_i)\}_{i=1}^{\tilde{n}}$ into train (80%) and test datasets (20%). We use random forests on the training sample to obtain the prediction model μ_z and validate its performance on the test sample. The trained algorithm has an accuracy of 97% on the test data.

E Empirical Policy Application: Details

E.1 Hospital Cost Data

We use publicly available Healthcare Cost Report Information System (HCRIS) data,[42] to project[43] funding eligibility and funding amounts for all hospitals in the dataset. This data set contains information on various hospital characteristics including utilization, number of employees, medicare cost data and financial statement data.

The data is available from financial year 1996 to 2019. As the coverage is higher for 2018 (compared to 2019), we utilize the data corresponding to the 2018 financial year. Hospitals are uniquely identified in a financial year by their CMS (Center for Medicaid and Medicare Services) Certification Number. We have data for 4,705 providers for the 2018 financial year. We focus on 4,648 acute care and critical access hospitals that are either located in one of the 50 states or Washington DC.

Disproportionate patient percentage. Disproportionate patient percentage is equal to the percentage of Medicare inpatient days attributable to patients eligible for both Medicare Part A and Supplemental Security Income (SSI) summed with the percentage of total inpatient days attributable to patients eligible for Medicaid but not Medicare Part A.[44] In the data, this variable is missing for 1560 hospitals. We impute the disproportionate patient percentage to 0 when it is missing.

Uncompensated care per bed. Cost of uncompensated care refers to the care provided by the hospital for which no compensation was received from the patient or the insurer. It is the sum of a hospital's bad debt and the financial assistance it provides.[45] The cost of uncompensated care is missing for 86 hospitals, which we impute to 0. We divide the cost of uncompensated care by the number of beds in the hospital to obtain the cost per bed. The data on bed count is missing for 15 hospitals, which we drop from the analysis, leaving us with 4,633 hospitals in 2,473 counties.

Profit Margin. Hospital profit margins are indicative of the financial health of the hospitals. We calculate profit margins as the ratio of net income to total revenue where total revenue is the sum of net patient revenue and total other income. After the calculation, profit margins are missing for 92 hospitals, which we impute to 0.

Funding. We calculate the projected funding using the formula on the CARES ACT website. Hospitals that do not qualify on any of the three dimensions are not given any funding. Each eligible hospital is assigned an individual facility score, which is calculated as the product of disproportionate patient percentage and number of beds in that hospital. We calculate cumulative facility score as the sum of all individual facility scores in the dataset. Each hospital receives

[42] We use the RAND cleaned version of this dataset, which can be accessed `https://www.hospitaldatasets.org/`

[43] We use the methodology detailed in the CARES ACT website to project funding based on 2018 financial year cost reports.

[44] For the precise definition, see `https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/dsh`.

[45] The precise definition can be found at `https://www.aha.org/fact-sheets/2020-01-06-fact-sheet-uncompensated-hospital-care-cost`.

a share of $10 billion, where the share is determined by the ratio of individual facility score of that hospital to the cumulative facility score. The amount of funding received by hospitals is bounded below at $5 million and capped above at $50 million.

E.2 Hospital Utilization Data

We use the publicly available COVID-19 Reported Patient Impact and Hospital Capacity by Facility dataset for our outcome variables. This provides facility level data on hospital utilization aggregated on a weekly basis, from July 31st onwards. These reports are derived from two main sources – (1) HHS TeleTracking and (2) reporting provided directly to HHS Protect by state/territorial health departments on behalf of health care facilities.[46]

The hospitals are uniquely identified for a given collection week (which goes from Friday to Thursday) by their CMS Certification number. All hospitals that are registered with CMS by June 1st 2020 are included in the population. We merge the hospital cost report data with the utilization data using the CMS certification number. According to the terms and conditions of the CARES Health Care Act, the recipients may use the relief funds only to "prevent, prepare for, and respond to coronavirus" and for "health care related expenses or lost revenues that are attributable to coronavirus". Therefore, for our analysis we focus on 4 outcomes that were directly affected by COVID-19, for the week spanning July 31st to August 6th 2020. The outcome measures are described below.[47]

1. Total reports of patients currently hospitalized in an adult inpatient bed who have laboratory-confirmed or suspected COVID-19, including those in observation beds reported during the 7-day period.

2. Total reports of patients currently hospitalized in an adult inpatient bed who have laboratory-confirmed COVID-19 or influenza, including those in observation beds. Including patients who have both laboratory-confirmed COVID-19 and laboratory confirmed influenza during the 7-day period.

3. Total reports of patients currently hospitalized in a designated adult ICU bed who have suspected or laboratory-confirmed COVID-19.

4. Total reports of patients currently hospitalized in a designated adult ICU bed who have laboratory-confirmed COVID-19 or influenza, including patients who have both laboratory-confirmed COVID-19 and laboratory-confirmed influenza.[48]

[46] Source: `https://healthdata.gov/Hospital/COVID-19-Reported-Patient-Impact-and-Hospital-Capa/anag-cw7u`.

[47] We conduct sanity checks and impute observations to missing if they fail our checks. For example, we impute the value # Confirmed/ Suspected COVID Patients and # Confirmed COVID Patients to missing when the latter is greater than the former. # Confirmed/ Suspected COVID Patients should be greater than or equal to # Confirmed COVID Patients as the former includes the latter. Similarly, we impute # Confirmed/ Suspected COVID Patients in ICU and # Confirmed COVID Patients in ICU to be missing when the latter is greater than the former.

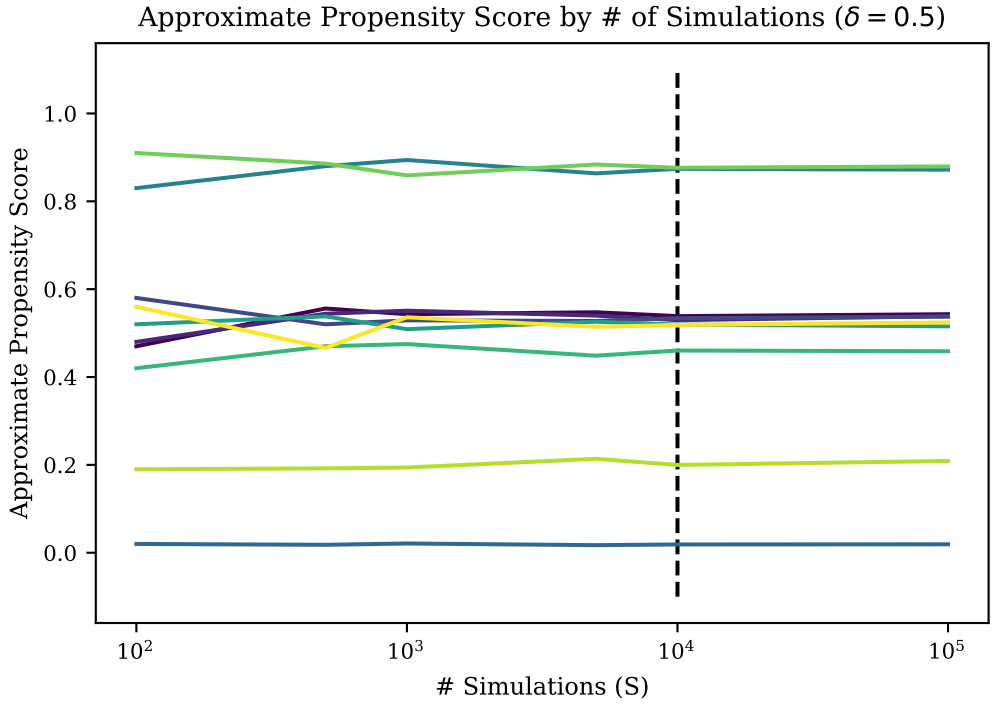
[48] In the dataset, when the values of the 7 day sum are reported to be less than 4, they are replaced with -999,999. We recode these values to be missing. The results in Table 4 remain almost the same even if we impute the suppressed values (coded as -999,999) with 0s. Results are available upon request.

E.3 Computing Fixed-Bandwidth Approximate Propensity Score

As the three determinants of funding eligibility are continuous variables, we can think of this setting as a multi-dimensional regression discontinuity design and a suitable setting to apply our method. In this setting, the X_i are disproportionate patient percentage, uncompensated care per bed and profit margin. Funding eligibility (Z_i) is determined algorithmically using these 3 dimensions. D_i is the amount of funding received by hospital i, which depends on both funding eligibility status Z_i, number of beds in the hospital, and disproportionate patient percentage. Before calculating fixed-bandwidth APS, we normalize each characteristic of X_i to have mean 0 and variance 1. For each hospital and every $\delta \in \{0.01, 0.025, 0.05, 0.075, 0.1, 0.25, 0.5\}$, we draw 10,000 times from a δ-ball around the normalized covariate space and calculate fixed-bandwidth APS by averaging funding eligibility Z_i over these draws.

E.4 Additional Empirical Results

Figure A.2: Fixed-bandwidth APS Estimation with Varying Simulations S



Notes: The above figure plots the fixed-bandwidth APS estimates for 10 randomly selected hospitals along the eligibility margin for varying numbers of simulations S. Each line represents a different hospital. The dotted line at 10^4 indicates the number of simulations we use for our main analysis.

Table A.1: Differential Attrition

			Our Method with Approximate Propensity Score Controls						
	Ineligible Hospitals	No Controls	$\delta =$ 0.01	$\delta =$ 0.025	$\delta =$ 0.05	$\delta =$ 0.075	$\delta =$ 0.1	$\delta =$ 0.25	$\delta =$ 0.5
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
#Confirmed/Suspected Covid Patients	.745	38.19***	-15.51	-24.80	-44.34	-57.95	-40.34	2.05	-4.08
		(8.55)	(85.67)	(70.81)	(70.09)	(63.06)	(48.58)	(22.20)	(15.67)
		N=3532	N=73	N=195	N=392	N=547	N=719	N=1389	N=1947
#Confirmed Covid Patients	.754	33.97***	0.85	-30.81	21.32	1.96	-0.39	-1.28	-8.25
		(7.44)	(73.28)	(55.22)	(33.46)	(29.41)	(25.14)	(15.75)	(12.56)
		N=3558	N=70	N=191	N=385	N=539	N=709	N=1366	N=1923
#Confirmed/Suspected Covid Patients in ICU	.728	13.18***	13.68	9.54	5.71	-0.83	2.34	-0.46	-4.21
		(2.74)	(23.41)	(17.74)	(11.91)	(10.68)	(9.01)	(5.78)	(4.64)
		N=3445	N=72	N=186	N=374	N=520	N=678	N=1314	N=1846
#Confirmed Covid Patients in ICU	.744	12.16***	7.97	-1.54	2.79	0.65	1.87	-1.94	-4.66
		(2.58)	(25.63)	(18.89)	(11.25)	(9.97)	(8.52)	(5.57)	(4.43)
		N=3503	N=67	N=181	N=370	N=514	N=671	N=1321	N=1868

Notes: This table reports differential safety net eligibility effects on the availability of outcome data at the hospital level. Column 1 presents the average of the availability indicators of the outcome variables for the ineligible hospitals. In column 2, we regress the availability indicator on dummy for safety net eligibility without any controls. In columns 3-9, we run this regression controlling for the Approximate Propensity Score with different values of bandwidth δ on the sample with nondegenerate Approximate Propensity Score. All Approximate Propensity Scores are computed by averaging 10,000 simulation draws. The outcome variables are the 7 day totals for the week spanning July 31st, 2020 to August 6th, 2020. Confirmed or Suspected COVID patients refer to the sum of patients in inpatient beds with lab-confirmed/suspected COVID-19. Confirmed COVID patients refer to the sum of patients in inpatient beds with lab-confirmed COVID-19, including those with both lab-confirmed COVID-19 and influenza. Inpatient bed totals also include observation beds. Similarly, Confirmed/Suspected COVID patients in ICU refer to the sum of patients in ICU beds with lab-confirmed or suspected COVID-19. Confirmed COVID patients in ICU refers to the sum of patients in ICU beds with lab-confirmed COVID-19, including those with both lab-confirmed COVID-19 and influenza. Robust standard errors are reported in parenthesis.