Abstract

Econometric Methods for Program Evaluation and Policy Choice

Kohei Yata

2022

This dissertation presents two essays on econometric methods for program evaluation and policy choice.

In Chapter 1, I develop a statistically optimal way of using data to make policy decisions when the performance of counterfactual policies is only partially identified. Specifically, I consider a class of statistical decision problems in which the policy maker must decide between two alternative policies to maximize social welfare (e.g., the population mean of an outcome) based on a finite sample. The central assumption is that the underlying, possibly infinite-dimensional parameter, lies in a known convex set, potentially leading to partial identification of the welfare effect. An example of such restrictions is the smoothness of counterfactual outcome functions. As the main theoretical result, I obtain a finite-sample decision rule (i.e., a function that maps data to a decision) that is optimal under the minimax regret criterion. This rule is easy to compute, yet achieves optimality among all decision rules; no ad hoc restrictions are imposed on the class of decision rules. I then apply my results to the problem of whether to change a policy eligibility cutoff in a regression discontinuity setup. I illustrate my approach in an empirical application to the Burkinabé Response to Improve Girls' Chances to Succeed program, a school construction program in Burkina Faso, where villages were selected to receive schools based on scores computed from their characteristics. Under reasonable restrictions on the smoothness of the counterfactual outcome function, the optimal decision rule implies that it is not cost-effective to expand the program. I empirically compare the performance of the optimal decision rule with alternative decision rules.

In Chapter 2, joint with Yusuke Narita, we show how to use data obtained from algorithmic decision making for impact evaluation. Machine learning and other algorithms produce a growing portion of decisions and recommendations both in policy and in business. This chapter first highlights a valuable aspect of such algorithmic decisions. That

is, algorithmic decisions are natural experiments (conditionally quasi-randomly assigned instruments) since the algorithms make decisions based only on observable input variables. We then 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 were 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.

Econometric Methods for Program Evaluation and Policy Choice

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
in Candidacy for the Degree of
Doctor of Philosophy

By
Kohei Yata

Dissertation Director: Yuichi Kitamura

May 2022

Contents

List of Figures

List of Tables

Acknowledgements

I owe a great debt of gratitude to many people during the process of completing this dissertation. First, I am extremely grateful to my main advisor Prof. Yuichi Kitamura for his encouragement, guidance, and support. Without his help, this research would not have been possible. I am deeply indebted to Prof. Timothy Armstrong, whose insight and knowledge steered me through this research. I would also like to thank Prof. Yusuke Narita for continuously collaborating with me and encouraging me during my doctoral studies. Furthermore, I am grateful to Prof. Donald Andrews and my fellow graduate students, especially Jaewon Lee and Vitor Possebom, for their invaluable comments, which substantially improved the quality of this dissertation. Lastly, I would like to express my appreciation to my parents, Yoko and Yoshiharu Yata, for all the support they gave to me.

Chapter 1

Optimal Decision Rules Under Partial Identification

1.1 Introduction

A fundamental goal of empirical research in economics is to inform policy decisions. Evaluation of counterfactual policies often requires extrapolating from observables to unobservables. Without strong model restrictions such as functional form assumptions or exogeneity of an intervention, the performance of each counterfactual policy may be only partially determined by observed data. In such situations, policy decision making is challenging since we have no clear understanding of which policy is the best.

For example, a regression discontinuity (RD) design only credibly estimates the impact of treatment on the individuals at the eligibility cutoff. Therefore, without restrictive assumptions such as constant treatment effects, whether or not to offer the treatment to those away from the cutoff is ambiguous. Even randomized controlled trials may provide only partial knowledge of the impact of a new intervention, as can happen if participants do not perfectly comply with their assigned treatment or if the experimental sample is an unrepresentative subset of the target population.

This chapter develops an optimal way of using data to make policy decisions when the performance of counterfactual policies is only partially identified. Specifically, I solve a class

of statistical decision problems. The setup is as follows. The policy maker must decide between two alternative policies, policy 1 and policy 0, to maximize social welfare. The difference in welfare between policy 1 and policy 0 is given by $L(\theta) \in \mathbb{R}$. $L(\cdot)$ is a linear function of an unknown, possibly infinite-dimensional parameter θ, where θ belongs to a known parameter space Θ. By construction, it is optimal to choose policy 1 if $L(\theta) \geq 0$ and to choose policy 0 if $L(\theta) < 0$. The policy maker makes a decision after observing a finite sample $(Y_1, ..., Y_n) \in \mathbb{R}^n$ whose expected value is given by $(m_1(\theta), ..., m_n(\theta)) \in \mathbb{R}^n$, where $m_i(\cdot)$'s are linear functions of θ.

A leading example of this setup is a choice between two treatment assignment policies based on data generated by nonparametric regression models, including an RD model. A treatment assignment policy specifies who would receive treatment based on an individual's observable covariates. In this example, the parameter θ is a conditional mean function of a counterfactual outcome given covariates and treatment. $m_i(\theta)$ is the conditional mean counterfactual outcome given individual i's observed covariates and treatment. The welfare difference $L(\theta)$ corresponds to the average treatment effect for the subpopulation that would be affected by the switch from the status quo to a new policy. The parameter space Θ is a class of conditional mean counterfactual outcome functions that satisfy, for example, some smoothness restrictions (e.g., bounds on derivatives or the linearity of a function). The welfare difference $L(\theta)$ may or may not be point identified, depending on which function class the policy maker imposes.[1]

As the main theoretical result, I obtain a finite-sample decision rule (i.e., a function that maps the sample $(Y_1, ..., Y_n)$ to a probability of choosing policy 1) that is optimal under the minimax regret criterion, a standard criterion used in the literature on statistical treatment choice (e.g., Manski, 2004; Stoye, 2009; Kitagawa and Tetenov, 2018). The minimax regret criterion evaluates decision rules based on the maximum regret, that is, the maximum of the expected amount of welfare lost by choosing the worse policy over the parameter space. The decision rule derived in this chapter minimizes the maximum regret over the class of all decision rules. This optimality result holds whether the welfare difference $L(\theta)$ is point or

1. In Section 1.2.1, I will discuss what I mean by identification in this finite-sample setup.

partially identified. To derive the optimality result, I assume that the sample $(Y_1, ..., Y_n)$ is normally distributed with a known variance and that the parameter space Θ is convex and symmetric with respect to the origin, as well as mild regularity conditions.

Importantly, I do not impose any restrictions on the class of decision rules, thus allowing for nonrandomized threshold rules based on a nonlinear function of the sample $(Y_1, ..., Y_n)$ and randomized rules, among others. Solving minimax problems over the class of all decision rules is generally a difficult task. The main tool that I use to solve the minimax regret problem is what is called the *modulus of continuity* (Donoho, 1994). The modulus of continuity at $\epsilon \geq 0$ is the largest possible welfare difference over the parameter space under the constraint that the Euclidean norm of the expected value of the sample $(Y_1, ..., Y_n)$ is at most ϵ, formally defined in Section 1.3. The minimax problem can be simplified into an optimization problem with respect to the modulus of continuity, which is analytically and computationally tractable.

The resulting decision rule is simple and thus easy to compute. It makes a decision based on a linear function of $(Y_1, ..., Y_n)$. The minimax regret rule may be randomized or nonrandomized, depending on the restrictions imposed on the parameter space. Specifically, it is a nonrandomized rule if the length of the identified set of the welfare difference $L(\theta)$ is short relative to the variance of the sample $(Y_1, ..., Y_n)$, including the case where $L(\theta)$ is point identified. Otherwise, it is a randomized rule, assigning a positive probability both to policies 1 and 0.

When the minimax regret rule is nonrandomized, it can be viewed as a rule that plugs a particular linear estimator of the welfare difference $L(\theta)$ into the optimal decision $\mathbf{1}\{L(\theta) \geq 0\}$. I compare this linear estimator with a linear minimax mean squared error (MSE) estimator of $L(\theta)$, which minimizes the maximum of the MSE over the parameter space within the class of all linear estimators. The two estimators are shown to be generally different, which suggests that the plug-in rule based on the linear minimax MSE estimator is not optimal under the minimax regret criterion. More precisely, the linear estimator used by the minimax regret rule places more importance on the bias than on the variance compared to the linear minimax MSE estimator.

This chapter makes new contributions even under point identification in settings with

restricted parameter spaces. When the welfare difference $L(\theta)$ is point identified, the minimax regret rule is insensitive to the choice of the restrictions imposed on the parameter space as long as the restrictions are weak enough. For example, consider linear regression models where $m_i(\theta) = x_i'\theta$, $x_i \in \mathbb{R}^k$ is unit i's fixed regressors, and $\theta \in \Theta \subset \mathbb{R}^k$. The minimax regret rule bases decisions on the sign of $L(\hat{\theta})$, where $\hat{\theta}$ is the best linear unbiased estimator of θ, if the parameter space Θ is sufficiently large (e.g., if $\Theta = \mathbb{R}^k$). When the restrictions on Θ become strong enough, the minimax regret rule starts to use an estimator that optimally trades off the bias and variance.

I then apply my results to the problem of eligibility cutoff choice in an RD setup. In many policy domains, the eligibility for treatment is determined based on an individual's observable characteristics. One crucial policy question is whether we should change the eligibility criterion to achieve better outcomes (Dong and Lewbel, 2015). Specifically, I consider an RD setup and study the problem of whether or not to change the eligibility cutoff from a current value c_0 to a new value c_1. For an illustration of the results, I focus on the case where the new value is smaller than the current one (i.e., $c_1 < c_0$) and the conditional mean counterfactual outcome function belongs to the class of Lipschitz functions with a known Lipschitz constant C. The absolute value of the derivative of any differentiable function in this function class is bounded above by C. Under the Lipschitz constraint, the effect of the cutoff change on the population mean outcome is partially identified.

A closed-form expression for the minimax regret rule can be obtained in this application when the Lipschitz constant C is large enough. In such cases, the minimax regret rule is based on the mean outcome difference between the treated unit closest to the status quo cutoff c_0 and the untreated units between the two cutoffs c_0 and c_1. On the other hand, when C is not sufficiently large, the minimax regret rule may also use outcomes of other units, although it does not generally admit a closed form. I provide a simple procedure to numerically compute it for any choice of C.

Implementation of the minimax regret rule requires choosing the Lipschitz constant C. In principle, it is not possible to choose the Lipschitz constant C that applies to both sides of the status quo cutoff c_0 in a data-driven way since we only observe outcomes either under treatment or under no treatment on each side. It is, however, possible to estimate a

lower bound on C. In practice, I recommend considering a range of plausible choices of C, including the estimated lower bound, to conduct a sensitivity analysis.

Finally, I illustrate my approach in an empirical application to the Burkinabé Response to Improve Girls' Chances to Succeed (BRIGHT) program, a school construction program in Burkina Faso (Kazianga, Levy, Linden and Sloan, 2013). Aiming to improve educational outcomes in rural villages, the program constructed primary schools in 132 villages from 2005 to 2008. To allocate schools, the Ministry of Education first computed a score summarizing village characteristics for each of the nominated 293 villages and then selected the highest-ranking villages to receive a school. This situation fits into an RD setup.

I ask whether we should expand this program or not. The more specific question considered in this analysis is whether or not to construct schools in the top 20% of previously ineligible villages. The analysis uses the enrollment rate as the welfare measure and assumes that the conditional mean counterfactual outcome function belongs to the class of Lipschitz functions with a known Lipschitz constant C. To consider policy costs, I assume that implementing the policy is optimal if it is better in terms of cost-effectiveness than a similar policy, whose cost-effectiveness is available from external studies. Given available estimates of the new policy costs, my approach can be used to consider this decision problem. For a plausible range of the Lipschitz constant C, the minimax regret rule implies that building schools in the top 20% of previously ineligible villages is not cost-effective.

I empirically compare the minimax regret rule with plug-in decision rules that make a decision according to the sign of a policy effect's estimator. The performance of the minimax regret rule is shown to be relatively robust to misspecification of the Lipschitz constant C toward zero, which suggests that the potential loss due to an optimistic choice of C may not be a major concern.

My approach is applicable to many other policy choice problems. One example is the problem of deciding whether to introduce a new policy based on data from a randomized experiment when the experiment has imperfect compliance or when the experimental sample is a selected subset of the target population.

1.1.1 Related Literature

This chapter contributes to the literature on statistical treatment choice, which has been growing in econometrics since the work by Manski (2000, 2004). The literature has intensively studied optimal treatment assignment based on covariates in settings where social welfare under each policy is point identified (Manski, 2004; Dehejia, 2005; Hirano and Porter, 2009; Stoye, 2009, 2012; Bhattacharya and Dupas, 2012; Kitagawa and Tetenov, 2018, 2021; Athey and Wager, 2021; Mbakop and Tabord-Meehan, 2021).[2] In contrast, my approach can be applied to this problem even with partial identification if the choice set consists of two treatment assignment policies. Additionally, while many of the above papers provide finite-sample regret bounds or asymptotic optimality results, I derive a finite-sample optimality result.

This chapter is more closely related to the area of treatment choice under partial identification. In particular, Stoye (2012) and Ishihara and Kitagawa (2021) consider binary treatment choice problems under Gaussian models and derive minimax regret rules. Stoye (2012) provides a special case of my result in a setting where the experiment has imperfect internal or external validity. Ishihara and Kitagawa (2021) consider the problem of deciding whether or not to introduce a new policy to a specific local population based on causal evidence of similar policies implemented in other populations. While they derive a minimax regret rule within the class of plug-in rules based on a linear function of the sample, I derive a minimax regret rule within the class of all decision rules. Other papers studying optimal policy under partial identification include Manski (2007, 2009, 2010, 2011a,b, 2021), Kasy (2016, 2018), Mo, Qi and Liu (2021), Russell (2020), Christensen, Moon and Schorfheide (2020), and Kallus and Zhou (2021) among others.

The Gaussian model used in this chapter has been studied for the problem of optimal estimation and inference in nonparametric regression models. Donoho (1994) uses the modulus of continuity to characterize minimax optimal estimators and confidence intervals on linear functionals of a regression function. The derivation of my result and that of Donoho

2. This problem has also been actively studied in statistics and machine learning. A partial list includes Qian and Murphy (2011), Zhao, Zeng, Rush and Kosorok (2012), Swaminathan and Joachims (2015), and Kallus (2018).

(1994)'s consist of similar steps. However, the proof of each step is nontrivially different since the problem of policy choice and that of estimation and inference are not nested by each other; specifically, the loss function and the action space are different. Recent work on estimation and inference using Donoho (1994)'s framework includes Armstrong and Kolesár (2018), Imbens and Wager (2019), Rambachan and Roth (2020), Armstrong and Kolesár (2021), de Chaisemartin (2021), and Ignatiadis and Wager (2021).

In terms of an application to eligibility cutoff choice, this chapter is also related to the growing literature on extrapolation away from the cutoff in RD designs, including Rokkanen (2015), Angrist and Rokkanen (2015), Dong and Lewbel (2015), Bertanha and Imbens (2020), Bertanha (2020), Bennett (2020), and Cattaneo, Keele, Titiunik and Vazquez-Bare (2020). Unlike these papers, I explicitly consider the decision problem of whether or not to change the cutoff and derive an optimal decision rule. To the best of my knowledge, there are no existing results for optimal policy decisions based on data generated by an RD model.

1.2 Setup, Optimality Criterion, and Motivating Example

In this section, I set up the policy maker's problem of deciding between two alternative policies. This setup allows for the case where the welfare difference between the two policies is only partially identified. I then introduce the minimax regret criterion to evaluate different procedures for using data to make decisions. To illustrate my framework and its applicability, I present the problem of eligibility cutoff choice in an RD setup as an example.

1.2.1 Setup

Data-generating Model. Suppose that the policy maker observes a sample $\boldsymbol{Y} = (Y_1, ..., Y_n)' \in \mathbb{R}^n$ of the form

$$\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{m}(\theta), \boldsymbol{\Sigma}), \tag{1.1}$$

where θ is an unknown parameter that lies in a known subset Θ of a vector space $\mathbb{V}$, $\boldsymbol{m} : \mathbb{V} \to \mathbb{R}^n$ is a known linear function, and $\boldsymbol{\Sigma}$ is a known, positive-definite $n \times n$ matrix.[3] I allow θ to be an infinite-dimensional parameter such as a function.

The linearity of $\boldsymbol{m}$ is not necessarily restrictive. If we specify θ so that it contains each of the expected values of $Y_1, ..., Y_n$ as its element, $\boldsymbol{m}$ is a function that extracts those expected values from θ, which is linear in θ.

This model allows the expected value of $\boldsymbol{Y}$ to depend on other observed variables such as covariates and treatment by treating them as fixed and subsuming them into $\boldsymbol{m}$ and $\boldsymbol{\Sigma}$. For example, a regression model with fixed regressors

$$Y_i = f(x_i) + u_i, \quad u_i \sim \mathcal{N}(0, \sigma^2(x_i)) \text{ independent across } i$$

is a special case where $\boldsymbol{Y} = (Y_1, ..., Y_n)'$, $\theta = f$, Θ is a class of functions, $\boldsymbol{m}(f) = (f(x_1), ..., f(x_n))'$, and $\boldsymbol{\Sigma} = \text{diag}(\sigma^2(x_1), ..., \sigma^2(x_n))$.

The normality of $\boldsymbol{Y}$ and the assumption of known variance are restrictive, but are often imposed to deliver finite-sample optimality results for problems of estimation, inference, and treatment choice. In some cases, it is plausible to assume the normality of $\boldsymbol{Y}$. For example, suppose that unit i represents a group of individuals defined by place, time, and individual characteristics among others and that $Y_1, ..., Y_n$ are group-level mean outcomes. If the number of groups is fixed at n, the distribution of $\boldsymbol{Y}$ approaches a normal distribution as the size of each group grows to infinity by the central limit theorem. The normal model (1.1) can be viewed as an asymptotic approximation if each group is large enough.[4]

I assume that the parameter space Θ is convex and centrosymmetric (i.e., $\theta \in \Theta$ implies $-\theta \in \Theta$) throughout the chapter. Typical parameter spaces considered in empirical analyses

3. Donoho (1994), Low (1995), and Armstrong and Kolesár (2018) investigate optimal estimation and inference of a linear functional of θ in a slightly more general version of this model that allows $\boldsymbol{Y}$ to be infinite dimensional.

4. More generally, suppose that the policy maker observes an n-dimensional vector of statistics of the original data and that it is an asymptotically normal estimator of its population counterpart. For example, the mean outcome difference between the treatment and control groups in a randomized experiment is a statistic that is asymptotically normal for the population mean difference. If we regard the n-dimensional vector of statistics as $\boldsymbol{Y}$, the normal model (1.1) can again be viewed as an asymptotic approximation (Stoye, 2012; Tetenov, 2012; Rambachan and Roth, 2020; Andrews, Kitagawa and McCloskey, 2021; Ishihara and Kitagawa, 2021).

are convex. For example, in the regression model above, classes of functions with bounded derivatives (e.g., the class of Lipschitz functions with a known Lipschitz constant) are convex. The centrosymmetry simplifies the analysis, but rules out some shape restrictions. In the regression model above, Θ fails to be centrosymmetric if we assume the convexity or concavity of the regression function.[5]

Policy Choice Problem. Now, suppose that the policy maker is interested in choosing between two alternative policies, policy 1 and policy 0, to maximize social welfare. The class of binary policy decisions includes, for example, whether to introduce a program to a target population and whether to change a policy from the status quo to a new one. Suppose that the welfare resulting from implementing policy $a \in \{0, 1\}$ under θ is $W_a(\theta)$, where $W_a : \mathbb{V} \to \mathbb{R}$ is a known function specified by the policy maker. The welfare difference between policy 1 and policy 0 is given by

$$L(\theta) \coloneqq W_1(\theta) - W_0(\theta).$$

I assume that $L : \mathbb{V} \to \mathbb{R}$ is a linear function. The optimal policy under θ is policy 1 if $L(\theta) > 0$, policy 0 if $L(\theta) < 0$, and either of the two if $L(\theta) = 0$.

One example of a welfare criterion is a weighted average of an outcome across individuals. For example, suppose that a policy could change the outcome of each individual in the population. Suppose also that we specify $\theta = (f_1(\cdot), f_0(\cdot))$, where $f_a(x)$ represents the counterfactual mean outcome under policy a across individuals whose observed covariates are x. The welfare under policy a can be defined, for example, by the population mean outcome $W_a(\theta) = \int f_a(x) dP_X$, where P_X is the probability measure of covariates in the population and is assumed to be known. In this case, the welfare difference $L(\theta) = \int [f_1(x) - f_0(x)] dP_X$ is linear in $\theta = (f_1(\cdot), f_0(\cdot))$. If we are required to take the policy cost into account, we can

5. On the other hand, in some cases, it is possible to impose the monotonicity of the regression function by normalizing the sample $\boldsymbol{Y}$ so that the new parameter space is centrosymmetric. Suppose, for example, that $\Theta = \{f \in \mathcal{F}_{\text{Lip}}(C) : f(x) \text{ is nondecreasing in } x\}$, where $\mathcal{F}_{\text{Lip}}(C) = \{f : |f(x) - f(\tilde{x})| \leq C|x - \tilde{x}| \text{ for every } x, \tilde{x} \in \mathbb{R}\}$. $\mathcal{F}_{\text{Lip}}(C)$ is centrosymmetric while Θ is not. It is easy to show that $\Theta = \{\tilde{f} + f_0 : \tilde{f} \in \mathcal{F}_{\text{Lip}}(C/2)\}$, where $f_0(x) = \frac{C}{2}x$ for all $x \in \mathbb{R}$. Therefore, the model $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{m}(f), \boldsymbol{\Sigma})$, $f \in \Theta$, is equivalent to the model $\tilde{\boldsymbol{Y}} \sim \mathcal{N}(\boldsymbol{m}(\tilde{f}), \boldsymbol{\Sigma})$, $\tilde{f} \in \mathcal{F}_{\text{Lip}}(C/2)$, where $\tilde{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{m}(f_0) = (Y_1 - f_0(x_1), ..., Y_n - f_0(x_n))'$; the set of distributions of $\boldsymbol{Y}$ over $f \in \Theta$ is identical to the set of distributions of $\tilde{\boldsymbol{Y}} + \boldsymbol{m}(f_0)$ over $\tilde{f} \in \mathcal{F}_{\text{Lip}}(C/2)$.

incorporate it into the welfare by redefining the outcome to be the raw outcome minus the cost. On the other hand, the linearity of L may rule out welfare criteria that depend on the distribution of the counterfactual outcome.[6]

Importantly, this framework allows for cases in which $L(\theta)$ is not point identified in the sense that the identified set of $L(\theta)$ when $\boldsymbol{m}(\theta) = \boldsymbol{\mu}$, namely

$$\{L(\theta) : \boldsymbol{m}(\theta) = \boldsymbol{\mu}, \theta \in \Theta\},$$

is nonsingleton for some or all $\boldsymbol{\mu} \in \mathbb{R}^n$. This is the set of possible values of $L(\theta)$ consistent with the observed value of $\boldsymbol{Y} \in \mathbb{R}^n$ when there is no sampling uncertainty. If the identified set contains both positive and negative values, which policy we should choose is ambiguous even without sampling uncertainty. Whether $L(\theta)$ is point identified or not depends on the parameter space Θ.

This framework nests some existing setups of treatment choice, such as limit experiments under parametric models by Hirano and Porter (2009), Gaussian experiments with limited validity by Stoye (2012), and a setup of policy choice based on multiple studies by Ishihara and Kitagawa (2021).[7] One of the essential departures from these setups is that the parameter θ can be infinite dimensional, accommodating nonparametric regression models, for example.[8]

1.2.2 Optimality Criterion

What is the optimal procedure for using the sample $\boldsymbol{Y}$ to make a policy choice? This chapter considers the minimax regret criterion as an optimality criterion, following existing treatment choice studies (e.g., Manski, 2004, 2007; Stoye, 2009, 2012; Kitagawa and Tetenov,

6. One example of such a criterion is $\int_0^1 w(\tau)F^{-1}(\tau; f_a)d\tau$. Here, $w(\cdot)$ is a known weight function, $F(\cdot; f_a)$ is the distribution of the counterfactual outcome under policy a induced by the normal distribution with the conditional mean function f_a, and $F^{-1}(\tau; f_a)$ is the τ-th quantile of the distribution.

7. Ishihara and Kitagawa (2021) do not impose convexity of the parameter space to derive their results. However, the specific examples of the parameter space that they consider satisfy convexity.

8. Hirano and Porter (2009) consider a model with an infinite Gaussian sequence as a limit experiment under semiparametric models. They do not allow for a partially identified welfare difference—one of the crucial aspects of this chapter's setup.

2018).[9]

I define a few concepts to introduce the minimax regret criterion. A *decision rule* is a measurable function $\delta : \mathbb{R}^n \to [0,1]$, where $\delta(\boldsymbol{y})$ represents the probability of choosing policy 1 when the realization of the sample $\boldsymbol{Y}$ is $\boldsymbol{y}$. The *welfare regret loss* for policy choice $a \in \{0,1\}$ is

$$l(a,\theta) \coloneqq \max_{a' \in \{0,1\}} W_{a'}(\theta) - W_a(\theta) = \begin{cases} L(\theta) \cdot (1-a) & \text{if } L(\theta) \geq 0, \\ -L(\theta) \cdot a & \text{if } L(\theta) < 0. \end{cases}$$

The welfare regret loss $l(a,\theta)$ is the difference between the welfare under the optimal policy and the welfare under policy a under θ. If the policy maker chooses the superior policy, they do not incur any loss; otherwise, they incur a loss of the absolute value of the welfare difference $L(\theta)$. The risk or *regret* of decision rule δ under θ is the expected welfare regret loss

$$R(\delta,\theta) \coloneqq \begin{cases} L(\theta)(1 - \mathbb{E}_\theta[\delta(\boldsymbol{Y})]) & \text{if } L(\theta) \geq 0, \\ -L(\theta)\mathbb{E}_\theta[\delta(\boldsymbol{Y})] & \text{if } L(\theta) < 0, \end{cases}$$

where $\mathbb{E}_\theta$ denotes the expectation taken with respect to $\boldsymbol{Y}$ under θ.

Given a particular choice of Θ, I evaluate decision rules based on the maximum regret over Θ, $\sup_{\theta \in \Theta} R(\delta,\theta)$. My goal is to derive a *minimax regret* decision rule, which achieves

$$\inf_{\delta} \sup_{\theta \in \Theta} R(\delta,\theta),$$

where the infimum is taken over the set of all possible decision rules. I do not impose any restrictions on the class of decision rules.

To sum up, the minimax regret criterion deals with the sampling uncertainty given θ by taking the expectation of the welfare regret loss with respect to the distribution of $\boldsymbol{Y}$. It

9. Alternative criteria include the maximin criterion, which solves $\sup_\delta \inf_{\theta \in \Theta} U(\delta,\theta)$, where $U(\delta,\theta) = W_1(\theta)\mathbb{E}_\theta[\delta(\boldsymbol{Y})] + W_0(\theta)(1 - \mathbb{E}_\theta[\delta(\boldsymbol{Y})])$ is the expected welfare under decision rule δ under θ. It has been pointed out that the maximin criterion is unreasonably pessimistic and can lead to pathological decision rules (Savage, 1951; Manski, 2004). Another approach is the Bayesian one, which solves $\sup_\delta \int U(\delta,\theta) d\pi(\theta)$, where π is a prior on the vector space $\mathbb{V}$ that θ belongs to. In practice, when it is difficult to make a prior, the minimax regret criterion is a reasonable choice.

then deals with the parameter θ by considering the worst-case expected welfare regret loss. It does not distinguish between the case where the welfare difference $L(\theta)$ is point identified and the case where it is not. Nevertheless, as I show in Section 1.3, the minimax regret rule behaves differently in each case.

1.2.3 Motivating Example: Eligibility Cutoff Choice in Regression Discontinuity Designs

In many policy domains, ranging from health to education to social programs, the eligibility for treatment is determined based on an individual's observable characteristics. A critical policy question is whether we should change the eligibility criterion to achieve better welfare (Dong and Lewbel, 2015).[10] My framework can be of use for policy makers interested in utilizing data to make such decisions.

Consider the following RD setup. For each unit $i = 1, .., n$, we observe a fixed running variable $x_i \in \mathbb{R}$, a binary treatment status $d_i \in \{0, 1\}$, and an outcome $Y_i \in \mathbb{R}$. The eligibility for treatment is determined based on whether the running variable exceeds a specific cutoff $c_0 \in \mathbb{R}$, so that $d_i = \mathbf{1}\{x_i \geq c_0\}$. Suppose that the outcome Y_i is of the form

$$Y_i = f(x_i, d_i) + u_i, \quad u_i \sim \mathcal{N}(0, \sigma^2(x_i, d_i)) \text{ independent across } i, \tag{1.2}$$

where $f : \mathbb{R} \times \{0, 1\} \to \mathbb{R}$ is an unknown function and the conditional variance $\sigma^2(x_i, d_i)$ is known for $i = 1, ..., n$.[11] We interpret $f(x, d)$ as the counterfactual mean outcome across individuals with running variable x if their treatment status is set to $d \in \{0, 1\}$.[12] We can

10. For example, there is a heated debate about whether to extend Medicare eligibility in the United States (Song, 2020).

11. I make this assumption to deliver finite-sample optimality results. In practice, one replaces the true conditional variances with their consistent estimators. See Section 1.4.1 for possible estimators.

12. This interpretation is established in a potential outcome model as follows (Armstrong and Kolesár, 2021). Suppose we observe a triple of the outcome, treatment status, and running variable (Y_i, D_i, X_i). The observed outcome is $Y_i = Y_i(1)D_i + Y_i(0)(1 - D_i)$, where $Y_i(1)$ and $Y_i(0)$ are potential outcomes under treatment and no treatment, respectively. Let $f(x, d) = \mathbb{E}[Y_i(d)|X_i = x]$, which is equal to $\mathbb{E}[Y_i|X_i = x, D_i = d]$ if D_i is a deterministic function of X_i. We obtain model (1.2) by conditioning on the realized values $\{(x_i, d_i)\}_{i=1}^n$ and assuming normal conditional errors.

write the model in a vector form:

$$\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{m}(f), \boldsymbol{\Sigma}),$$

where $\boldsymbol{Y} = (Y_1, ..., Y_n)'$, $\boldsymbol{m}(f) = (f(x_1, d_1), ..., f(x_n, d_n))'$, and $\boldsymbol{\Sigma} = \text{diag}(\sigma^2(x_1, d_1), ..., \sigma^2(x_n, d_n))$. Here, f plays the role of the unknown parameter θ.

Now, suppose that we are interested in changing the eligibility cutoff from c_0 to a specific value c_1. For illustration purposes, I assume $c_1 < c_0$. Suppose that the welfare under the cutoff c_a, with $a \in \{0, 1\}$, is an average of the counterfactual mean outcome across different values of the running variable

$$W_a(f) = \int [f(x, 1)\mathbf{1}\{x \geq c_a\} + f(x, 0)\mathbf{1}\{x < c_a\}] d\nu(x)$$

for some known measure ν. One choice of ν is an empirical measure, for which the welfare is the unweighted sample average: $W_a(f) = \frac{1}{n}\sum_{i=1}^{n}[f(x_i, 1)\mathbf{1}\{x_i \geq c_a\} + f(x_i, 0)\mathbf{1}\{x_i < c_a\}]$. The welfare difference between the two cutoffs is

$$L(f) = W_1(f) - W_0(f) = \int \mathbf{1}\{c_1 \leq x < c_0\}[f(x, 1) - f(x, 0)] d\nu(x),$$

which is a linear function of f. $L(f)$ is a weighted sum of the conditional average treatment effect $f(x, 1) - f(x, 0)$ across different values of the running variable between the two cutoffs c_1 and c_0.

To conclude the problem's setup, suppose that $f \in \mathcal{F}$, where $\mathcal{F}$ is a known set of functions and plays the role of the parameter space Θ. For an illustration of the results and empirical application, I focus on the *Lipschitz class* with a known Lipschitz constant $C \geq 0$:

$$\mathcal{F}_{\text{Lip}}(C) = \{f : |f(x, d) - f(\tilde{x}, d)| \leq C|x - \tilde{x}| \text{ for every } x, \tilde{x} \in \mathbb{R} \text{ and } d \in \{0, 1\}\}.$$

The Lipschitz constraint bounds the maximum possible change in $f(x, d)$ in response to a shift in x by one unit. In other words, the absolute value of the derivative of $f(x, d)$ with

respect to x must be at most C if f is differentiable. The Lipschitz class $\mathcal{F}_{\text{Lip}}(C)$ is both convex and centrosymmetric.

Imposing $f \in \mathcal{F}_{\text{Lip}}(C)$ is not strong enough to uniquely determine $L(f)$ from a given value of $\boldsymbol{m}(f) = (f(x_1, d_1), ..., f(x_n, d_n))'$. Nevertheless, it produces an informative identified set of $L(f)$ since it gives finite upper and lower bounds on $f(x, d)$ for every $(x, d) \in \mathbb{R} \times \{0, 1\}$ from the knowledge of $(f(x_1, d_1), ..., f(x_n, d_n))$.[13]

In Section 1.4, I derive a minimax regret rule for this example when the welfare is the sample average outcome and $\mathcal{F} = \mathcal{F}_{\text{Lip}}(C)$.

This example can be easily generalized to a setup where the observed treatment is independent of counterfactual outcomes conditional on multidimensional covariates (i.e., the unconfoundedness assumption holds) and there is no or limited overlap in the covariate distribution between the treatment and control groups. This general setup covers the problem of whether to change an eligibility criterion based on multiple covariates. Limited overlap may also occur, for example, if the policy maker wishes to consider whether to introduce a policy using a composite dataset of treated units from a local randomized experiment and nonexperimental comparison units from national surveys (LaLonde, 1986; Dehejia and Wahba, 1999). See Appendix 1.A.1 for details on the general setup.

1.3 Main Result

In this section, I derive a minimax regret rule among all possible decision rules and then discuss its interpretations and implications. For simplicity, I normalize $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}_n$ for some $\sigma > 0$, where $\boldsymbol{I}_n$ is the identity matrix.[14]

13. Given a value of $(f(x_1, d_1), ..., f(x_n, d_n))$, the upper bound on $f(x, d)$ is $\min_{i:d_i=d}(f(x_i, d) + C|x_i - x|)$. The lower bound on $f(x, d)$ is $\max_{i:d_i=d}(f(x_i, d) - C|x_i - x|)$.

14. This normalization is without loss of information in the following sense. If $\boldsymbol{\Sigma}$ is known, observing $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{m}(\theta), \boldsymbol{\Sigma})$ is equivalent to observing $\tilde{\boldsymbol{Y}} \sim \mathcal{N}(\tilde{\boldsymbol{m}}(\theta), \sigma^2 \boldsymbol{I}_n)$ for any $\sigma > 0$, where $\tilde{\boldsymbol{Y}} = \sigma \boldsymbol{\Sigma}^{-1/2} \boldsymbol{Y}$ and $\tilde{\boldsymbol{m}}(\theta) = \sigma \boldsymbol{\Sigma}^{-1/2} \boldsymbol{m}(\theta)$.

1.3.1 Modulus of Continuity

Solving minimax problems over the class of all decision rules is generally a difficult task. The main tool that I use to solve the minimax regret problem is the *modulus of continuity*, defined as

$$\omega(\epsilon; L, \boldsymbol{m}, \Theta) \coloneqq \sup\{L(\theta) : \|\boldsymbol{m}(\theta)\| \leq \epsilon, \theta \in \Theta\}, \quad \epsilon \geq 0,$$

where $\|\cdot\|$ is the Euclidean norm. The modulus of continuity and its variants have been used in constructing minimax optimal estimators and confidence intervals on linear functionals in Gaussian models (Donoho, 1994; Low, 1995; Cai and Low, 2004; Armstrong and Kolesár, 2018).[15] It has not been used in deriving minimax regret rules for the problem of treatment choice.

By definition, $\omega(\epsilon; L, \boldsymbol{m}, \Theta)$ is nonnegative and nondecreasing in ϵ. Furthermore, $\omega(\epsilon; L, \boldsymbol{m}, \Theta)$ is concave in ϵ if Θ is convex.[16] I say that $\theta_\epsilon \in \Theta$ *attains the modulus of continuity at* ϵ if $L(\theta_\epsilon) = \omega(\epsilon; L, \boldsymbol{m}, \Theta)$ and $\|\boldsymbol{m}(\theta_\epsilon)\| \leq \epsilon$, namely if $\theta_\epsilon \in \arg\max_{\theta \in \Theta} L(\theta)$ *s.t.* $\|\boldsymbol{m}(\theta)\| \leq \epsilon$. Below, I suppress the arguments $L, \boldsymbol{m}$, and Θ if they are clear from the context.

In the context of this chapter, the modulus of continuity at ϵ is the largest possible welfare difference under the constraint that the norm of $\boldsymbol{m}(\theta)$, namely the expected value of $\boldsymbol{Y}$, is less than or equal to ϵ. When $\epsilon = 0$ and hence the expected value of $\boldsymbol{Y}$ must be a vector of zeros, the sample $\boldsymbol{Y}$ is uninformative. When the norm constraint $\|\boldsymbol{m}(\theta)\| \leq \epsilon$ is relaxed, the strength of $\boldsymbol{Y}$ as a signal for $L(\theta)$ may increase, which makes it easier for the policy maker to detect the optimal policy. At the same time, the largest potential welfare loss when choosing the inferior policy may increase since the flexibility of θ increases because of the weaker norm constraint. The modulus of continuity is used to trade off these two criteria and find parameter values that are least favorable for the policy maker.

Here, I briefly formalize the above argument, deferring the statement of the necessary assumptions and the complete proof and discussion to Section 1.3.2 and Section 1.6, respec-

15. Donoho (1994) defines the modulus of continuity as $\tilde{\omega}(\epsilon) = \sup\{|L(\theta) - L(\tilde{\theta})| : \|\boldsymbol{m}(\theta - \tilde{\theta})\| \leq \epsilon, \theta, \tilde{\theta} \in \Theta\}$. If Θ is convex and centrosymmetric, the relationship $\tilde{\omega}(\epsilon) = 2\omega(\epsilon/2)$ holds.

16. See, for example, Donoho (1994, Lemma 3) and Armstrong and Kolesár (2018, Appendix A).

tively.

I introduce some notation. I use $\mathcal{R}(\sigma;\Theta)$ to denote the *minimax risk* $\inf_\delta \sup_{\theta\in\Theta} R(\delta,\theta)$, which may depend on the standard deviation σ and on the choice of the parameter space Θ among others. Given any two parameter values $\tilde{\theta},\bar{\theta}\in\mathbb{V}$, where $\mathbb{V}$ is the vector space that the parameter θ belongs to, I define a *one-dimensional subproblem* as the set of all convex combinations of $\tilde{\theta}$ and $\bar{\theta}$, denoted by $[\tilde{\theta},\bar{\theta}]=\{(1-\lambda)\tilde{\theta}+\lambda\bar{\theta}:\lambda\in[0,1]\}$. Let $\mathcal{R}(\sigma;[\tilde{\theta},\bar{\theta}])$ denote the minimax risk $\inf_\delta \sup_{\theta\in[\tilde{\theta},\bar{\theta}]} R(\delta,\theta)$ for the one-dimensional subproblem $[\tilde{\theta},\bar{\theta}]$. Additionally, let Φ and ϕ denote the cumulative distribution function and the probability density function, respectively, of a standard normal variable. Lastly, let $a^*\in\arg\max_{a\geq 0} a\Phi(-a)$, which is shown to be unique by Lemma 1.B.1 in Appendix 1.B.1.

Below, I first use the modulus of continuity to characterize the hardest one-dimensional subproblem of the form $[-\bar{\theta},\bar{\theta}]$ with $\bar{\theta}\in\Theta$, namely the one that has the largest minimax risk $\mathcal{R}(\sigma;[-\bar{\theta},\bar{\theta}])$ among all one-dimensional subproblems of this form. I then explain that the hardest one-dimensional subproblem is as hard as the original problem with the whole parameter space Θ, suggesting that the hardest one-dimensional subproblem consists of the least favorable parameter values in the original problem.

As shown in Lemmas 1.3 and 1.6 in Section 1.6, the minimax risk for the one-dimensional subproblem $[-\bar{\theta},\bar{\theta}]$ with $L(\bar{\theta})\geq 0$ is given by

$$\mathcal{R}(\sigma;[-\bar{\theta},\bar{\theta}])=\begin{cases} L(\bar{\theta})\Phi\left(-\frac{\|\boldsymbol{m}(\bar{\theta})\|}{\sigma}\right) & \text{if } \|\boldsymbol{m}(\bar{\theta})\|\leq a^*\sigma,\\ a^*\sigma\frac{L(\bar{\theta})}{\|\boldsymbol{m}(\bar{\theta})\|}\Phi\left(-a^*\right) & \text{if } \|\boldsymbol{m}(\bar{\theta})\|> a^*\sigma. \end{cases}$$

For example, if $L(\bar{\theta})\geq 0$ and $\|\boldsymbol{m}(\bar{\theta})\|>0$, the decision rule $\bar{\delta}(\boldsymbol{Y})=\mathbf{1}\{\boldsymbol{m}(\bar{\theta})'\boldsymbol{Y}\geq 0\}$ is shown to be minimax regret for the subproblem $[-\bar{\theta},\bar{\theta}]$. Computing the maximum regret $\sup_{\theta\in[-\bar{\theta},\bar{\theta}]} R(\bar{\delta},\theta)$ yields the above display.

For simplicity, I assume only here that for each $\epsilon\geq 0$, there exists a value of θ that attains the modulus of continuity at ϵ with $\|\boldsymbol{m}(\theta)\|=\epsilon$. The minimax risk for the hardest

one-dimensional subproblem can then be expressed in terms of the modulus of continuity:

$$\begin{aligned}
&\sup_{\bar{\theta}\in\Theta} \mathcal{R}(\sigma;[-\bar{\theta},\bar{\theta}]) = \sup_{\epsilon\geq 0} \sup_{\bar{\theta}\in\Theta:\|\boldsymbol{m}(\bar{\theta})\|=\epsilon, L(\bar{\theta})\geq 0} \mathcal{R}(\sigma;[-\bar{\theta},\bar{\theta}]) \\
&= \sup\left\{ \sup_{\epsilon\in[0,a^*\sigma]} \sup_{\bar{\theta}\in\Theta:\|\boldsymbol{m}(\bar{\theta})\|=\epsilon} L(\bar{\theta})\Phi\left(-\frac{\|\boldsymbol{m}(\bar{\theta})\|}{\sigma}\right), \sup_{\epsilon>a^*\sigma} \sup_{\bar{\theta}\in\Theta:\|\boldsymbol{m}(\bar{\theta})\|=\epsilon} a^*\sigma\frac{L(\bar{\theta})}{\|\boldsymbol{m}(\bar{\theta})\|}\Phi\left(-a^*\right) \right\} \\
&= \sup\left\{ \sup_{\epsilon\in[0,a^*\sigma]} \omega(\epsilon)\Phi(-\epsilon/\sigma), \sup_{\epsilon>a^*\sigma} a^*\sigma\frac{\omega(\epsilon)}{\epsilon}\Phi(-a^*) \right\},
\end{aligned}$$

where the first equality holds since restricting attention to $\bar{\theta}$ with $L(\bar{\theta}) \geq 0$ does not change the supremum by the centrosymmetry of Θ and the last equality follows from the definition of the modulus of continuity. Furthermore, since $\frac{\omega(\epsilon)}{\epsilon}$ is shown to be nonincreasing, $\sup_{\epsilon>a^*\sigma} a^*\sigma\frac{\omega(\epsilon)}{\epsilon}\Phi(-a^*) = \omega(a^*\sigma)\Phi(-a^*)$. The above expression can then be simplified into:

$$\sup_{\bar{\theta}\in\Theta} \mathcal{R}(\sigma;[-\bar{\theta},\bar{\theta}]) = \sup_{\epsilon\in[0,a^*\sigma]} \omega(\epsilon)\Phi(-\epsilon/\sigma).$$

Now, let ϵ^* solve the maximization problem on the right-hand side. ϵ^* balances the potential welfare loss ($\omega(\epsilon)$) and the probability of incurring loss ($\Phi\left(-\epsilon/\sigma\right)$). The corresponding subproblem $[-\theta_{\epsilon^*},\theta_{\epsilon^*}]$ has the largest minimax risk among all one-dimensional subproblems, where θ_{ϵ^*} attains the modulus of continuity at ϵ^*.

It turns out that the subproblem $[-\theta_{\epsilon^*},\theta_{\epsilon^*}]$ is as hard as the original problem with the whole parameter space Θ. That is, the two problems have the same minimax risk:

$$\mathcal{R}(\sigma;[-\theta_{\epsilon^*},\theta_{\epsilon^*}]) = \mathcal{R}(\sigma;\Theta),$$

as shown in Section 1.6. Thus, θ_{ϵ^*} and $-\theta_{\epsilon^*}$ are the least favorable parameter values for the policy maker.

In the next section, I derive a minimax regret rule. The rule protects against the worst case, as I discuss in Sections 1.3.3 and 1.3.4.

1.3.2 Minimax Regret Rules

I now present a minimax regret rule. To derive the result, I impose the following restrictions on L, $\boldsymbol{m}$, and Θ. For expositional purposes, I assume that $a^*\sigma < \sup_{\theta\in\Theta} \|\boldsymbol{m}(\theta)\|$.[17]

Assumption 1.1 (Regularity)**.** *The following holds for some $\bar{\epsilon} > 0$.*

(a) For all $\epsilon \in [0, \bar{\epsilon}]$, there exists $\theta_\epsilon \in \Theta$ that attains the modulus of continuity at ϵ.

(b) There exists $\boldsymbol{w}^* \in \mathbb{R}^n$ such that $\lim_{\epsilon\to 0} \epsilon^{-1} \left(\boldsymbol{w}^* - \frac{\boldsymbol{m}(\theta_\epsilon)}{\|\boldsymbol{m}(\theta_\epsilon)\|}\right) = \mathbf{0}$.

(c) For all $\epsilon \in [0, \bar{\epsilon}]$, there exists $\iota \in \Theta$ such that $L(\iota) \neq 0$ and $\theta_\epsilon + c\iota \in \Theta$ for all c in a neighborhood of zero.

(d) $\omega(\cdot)$ is differentiable at any $\epsilon \in (0, a^*\sigma]$. Furthermore, $\rho(\cdot)$ is differentiable at any $\epsilon \in (\epsilon_1, \epsilon_2)$, where $\rho(\epsilon) = \sup\{L(\theta) : (\boldsymbol{w}^*)'\boldsymbol{m}(\theta) = \epsilon, \theta \in \Theta\}$ for $\epsilon \in \mathbb{R}$, $\epsilon_1 = \inf\{(\boldsymbol{w}^*)'\boldsymbol{m}(\theta) : \theta \in \Theta\}$, and $\epsilon_2 = \sup\{(\boldsymbol{w}^*)'\boldsymbol{m}(\theta) : \theta \in \Theta\}$.

Assumption 1.1(a) says that the modulus of continuity is attained for all sufficiently small $\epsilon \geq 0$, which typically holds if Θ is closed.[18] Assumption 1.1(b) requires that the unit vector $\frac{\boldsymbol{m}(\theta_\epsilon)}{\|\boldsymbol{m}(\theta_\epsilon)\|} \in \mathbb{R}^n$ converge to some constant $\boldsymbol{w}^*$ faster than ϵ as $\epsilon \to 0$. The limit $\boldsymbol{w}^*$ can be viewed as the direction at which the welfare difference $L(\theta)$ increases the most when we move $\boldsymbol{m}(\theta)$ from $\mathbf{0}$. In Section 1.4, I show that Assumption 1.1(b) holds for the example in Section 1.2.3 by calculating a closed-form expression for $\frac{\boldsymbol{m}(\theta_\epsilon)}{\|\boldsymbol{m}(\theta_\epsilon)\|}$ for any sufficiently small $\epsilon > 0$. In principle, it is possible to verify whether Assumption 1.1(b) holds or not by numerically computing the limit of $\frac{\boldsymbol{m}(\theta_\epsilon)}{\|\boldsymbol{m}(\theta_\epsilon)\|}$ as $\epsilon \to 0$ and its convergence rate.

Assumption 1.1(c) and (d) are mild regularity conditions. Assumption 1.1(c) says that θ_ϵ lies in Θ even after receiving a small perturbation in the direction of some ι such that $L(\iota) \neq 0$. Assumption 1.1(d) assumes the differentiability of $\omega(\epsilon)$ and $\rho(\epsilon) = \sup\{L(\theta) : (\boldsymbol{w}^*)'\boldsymbol{m}(\theta) = \epsilon, \theta \in \Theta\}$. I provide sufficient conditions for the differentiability in Appendix 1.A.3. I make Assumption 1.1(c) and (d) to simplify the characterization of a minimax regret decision rule. In Section 1.6, I present the results under relaxed conditions.

17. When $a^*\sigma \geq \sup_{\theta\in\Theta} \|\boldsymbol{m}(\theta)\|$, Theorem 1.1 holds with $a^*\sigma$ replaced with $\sup_{\theta\in\Theta} \|\boldsymbol{m}(\theta)\|$.

18. See Donoho (1994, Lemma 2) for sufficient conditions.

The following theorem derives a minimax regret rule.

Theorem 1.1 (Minimax Regret Rule)**.** *Let Θ be convex and centrosymmetric, and suppose that Assumption 1.1 holds. Let*

$$\epsilon^* \in \arg\max_{\epsilon \in [0, a^*\sigma]} \omega(\epsilon)\Phi(-\epsilon/\sigma),$$

and suppose that there exists $\theta_{\epsilon^*} \in \Theta$ that attains the modulus of continuity at ϵ^*. Then, the following decision rule is minimax regret:

$$\delta^*(\boldsymbol{Y}) = \begin{cases} \mathbf{1}\left\{\boldsymbol{m}(\theta_{\epsilon^*})'\boldsymbol{Y} \geq 0\right\} & \text{if } \sigma > 2\phi(0)\frac{\omega(0)}{\omega'(0)}, \\ \mathbf{1}\left\{(\boldsymbol{w}^*)'\boldsymbol{Y} \geq 0\right\} & \text{if } \sigma = 2\phi(0)\frac{\omega(0)}{\omega'(0)}, \\ \Phi\left(\dfrac{(\boldsymbol{w}^*)'\boldsymbol{Y}}{((2\phi(0)\omega(0)/\omega'(0))^2 - \sigma^2)^{1/2}}\right) & \text{if } \sigma < 2\phi(0)\frac{\omega(0)}{\omega'(0)}, \end{cases}$$

where $\omega'(0)$ is the right derivative of $\omega(\cdot)$ at $\epsilon = 0$.[19] *Here, $\boldsymbol{m}(\theta_{\epsilon^*})$ does not depend on the choice of θ_{ϵ^*} among those that attain the modulus of continuity at ϵ^*. The minimax risk is given by*

$$\mathcal{R}(\sigma; \Theta) = \omega(\epsilon^*)\Phi(-\epsilon^*/\sigma).$$

Proof. See Section 1.6. □

The minimax regret rule takes different forms for the case where $\sigma \geq 2\phi(0)\frac{\omega(0)}{\omega'(0)}$ and for the case where $\sigma < 2\phi(0)\frac{\omega(0)}{\omega'(0)}$. If $\sigma \geq 2\phi(0)\frac{\omega(0)}{\omega'(0)}$, the minimax regret rule is nonrandomized, making a choice according to the sign of a weighted sum of the sample $\boldsymbol{Y}$. If $\sigma < 2\phi(0)\frac{\omega(0)}{\omega'(0)}$, the minimax regret rule is randomized, assigning a positive probability both to policies 1 and 0.

Below, I discuss the interpretations and implications of Theorem 1.1, starting from the condition $\sigma \geq 2\phi(0)\frac{\omega(0)}{\omega'(0)}$.

19. The right derivative exists since $\omega(\cdot)$ is concave. Under Assumption 1.1, it is shown that $\omega'(0) > 0$. See Lemma 1.5 in Section 1.6.2 and its proof for the details.

1.3.3 When and Why Randomize?

The condition $\sigma \geq 2\phi(0)\frac{\omega(0)}{\omega'(0)}$ determines whether the minimax regret rule is randomized or not. This condition is related to the strength of the restrictions imposed on the parameter space Θ. To see this, note first that $\omega(0) = \sup\{L(\theta) : \boldsymbol{m}(\theta) = \mathbf{0}, \theta \in \Theta\}$ by definition. Since L and $\boldsymbol{m}$ are linear and Θ is convex and centrosymmetric, it is shown that the closure of the identified set of $L(\theta)$ when $\boldsymbol{m}(\theta) = \mathbf{0}$ is given by[20]

$$\text{cl}(\{L(\theta) : \boldsymbol{m}(\theta) = \mathbf{0}, \theta \in \Theta\}) = [-\omega(0), \omega(0)].$$

We can thus interpret $\omega(0)$ as half of the length of the identified set of $L(\theta)$ when $\boldsymbol{m}(\theta) = \mathbf{0}$.

If $L(\theta)$ is identified, the length of the identified set is zero, which means that $\omega(0) = 0$. Since $\sigma \geq 2\phi(0)\frac{\omega(0)}{\omega'(0)}$, the minimax regret rule is always nonrandomized for cases where $L(\theta)$ is identified. In the example of Section 1.2.3, $L(\theta)$ is identified if, for example, we specify a polynomial model for f (see Section 1.5.2.1 for details).

On the other hand, if $L(\theta)$ is not identified, the length of the identified set is nonzero, which means that $\omega(0) > 0$. If the identified set is small relative to σ (holding $\omega'(0)$ fixed), the condition $\sigma \geq 2\phi(0)\frac{\omega(0)}{\omega'(0)}$ holds, and the minimax regret rule is nonrandomized. If the identified set is large relative to σ, the minimax regret rule is randomized. In Section 1.4, I show how this condition translates into one regarding the Lipschitz constant C in the example of Section 1.2.3.

For understanding why the policy maker should randomize their decisions when $\omega(0)$ is large relative to σ, it is useful to consider the problem of finding worst-case parameter values for a generic decision rule δ. The worst-case regret is attained at the parameter values that optimally trade off the potential welfare loss and the probability of incurring loss (i.e., $L(\theta)$ and $1 - \mathbb{E}_\theta[\delta(\boldsymbol{Y})]$ when $L(\theta) \geq 0$). Suppose that $\omega(0)$ is large relative to σ and that the policy maker uses a nonrandomized rule. Since σ is small, $\boldsymbol{Y}$ does not vary much across repeated samples, which makes the policy maker's choice based on the nonrandomized rule

20. Since L and $\boldsymbol{m}$ are linear and Θ is centrosymmetric, $-\omega(0) = \inf\{L(\theta) : \boldsymbol{m}(\theta) = \mathbf{0}, \theta \in \Theta\}$. Moreover, for any $\alpha \in (-\omega(0), \omega(0))$, we can find $\theta \in \Theta$ such that $L(\theta) = \alpha$ and $\boldsymbol{m}(\theta) = \mathbf{0}$ by the linearity of L and $\boldsymbol{m}$ and the convexity of Θ.

too predictable. By exploiting it, it is easy to find a value of θ under which the policy maker chooses the inferior policy with a high probability. If $\omega(0)$ is large enough, such choice of θ is not likely associated with a small welfare loss, leading to a large expected welfare loss of the decision rule. The policy maker can avoid this by randomizing their decisions; randomization makes their choice less predictable and protects against the exploitation of predictable choices.[21]

1.3.4 Intuition for Minimax Regret Rule

I now provide intuition for and interpretations of the minimax regret rule separately for the case where the rule is nonrandomized and for the one where the rule is randomized.

Nonrandomized Rule. If $\sigma > 2\phi(0)\frac{\omega(0)}{\omega'(0)}$, we first compute $\epsilon^* \in \arg\max_{\epsilon\in[0,a^*\sigma]} \omega(\epsilon)\Phi(-\epsilon/\sigma)$ to construct the minimax regret rule. This maximization problem corresponds to that of finding the hardest one-dimensional subproblem, as discussed in Section 1.3.1. The corresponding parameter values θ_{ϵ^*} and $-\theta_{\epsilon^*}$ are the least favorable for the policy maker, where θ_{ϵ^*} attains the modulus of continuity at ϵ^*.

The minimax regret rule δ^* protects against the worst case. To see this, note that the optimal policy is policy 1 under θ_{ϵ^*} and policy 0 under $-\theta_{\epsilon^*}$. The decision rule δ^* chooses policy 1 if the signal $\boldsymbol{Y}$ agrees more with θ_{ϵ^*} (i.e., $\boldsymbol{m}(\theta_{\epsilon^*})'\boldsymbol{Y} > 0$) and chooses policy 0 if the signal $\boldsymbol{Y}$ agrees more with $-\theta_{\epsilon^*}$ (i.e., $\boldsymbol{m}(\theta_{\epsilon^*})'\boldsymbol{Y} < 0$). More specifically, $\boldsymbol{m}(\theta_{\epsilon^*})'\boldsymbol{Y}$ is shown to be a sufficient statistic of the sample $\boldsymbol{Y}$ for the parameter θ in the one-dimensional subproblem $[-\theta_{\epsilon^*}, \theta_{\epsilon^*}]$. The sign of $\boldsymbol{m}(\theta_{\epsilon^*})'\boldsymbol{Y}$ provides information about whether the true θ is closer to θ_{ϵ^*} or to $-\theta_{\epsilon^*}$.

Randomized Rule. If $\sigma \leq 2\phi(0)\frac{\omega(0)}{\omega'(0)}$, $\epsilon^* = 0$ as shown in Lemma 1.1, where $\epsilon^* \in \arg\max_{\epsilon\in[0,a^*\sigma]} \omega(\epsilon)\Phi(-\epsilon/\sigma)$. As a result, θ_0 and $-\theta_0$ are the least favorable parameter values for the policy maker, where θ_0 attains the modulus of continuity at $\epsilon = 0$. Unlike in the case where $\epsilon^* > 0$, the minimax regret rule does not base decisions on a weighted

21. The fact that the minimax regret criterion may lead to a randomized rule under partial identification has been documented in the literature on treatment choice. See, for example, Manski (2007, 2009, 2011a,b) and Stoye (2012).

sum $\boldsymbol{m}(\theta_0)'\boldsymbol{Y}$, which is always zero since $\boldsymbol{m}(\theta_0) = \boldsymbol{0}$ by definition. The minimax regret rule instead uses $(\boldsymbol{w}^*)'\boldsymbol{Y}$, where $\boldsymbol{w}^* = \lim_{\epsilon \to 0} \frac{\boldsymbol{m}(\theta_\epsilon)}{\|\boldsymbol{m}(\theta_\epsilon)\|}$.

Simple calculations show that the randomized minimax regret rule is equivalent to

$$\delta^*(\boldsymbol{Y}) = \Pr_{\xi \sim \mathcal{N}(0, (2\phi(0)\omega(0)/\omega'(0))^2 - \sigma^2)} \left((\boldsymbol{w}^*)'\boldsymbol{Y} + \xi \geq 0 \right).$$

This rule is obtained through the following two-step procedure. We first add a noise $\xi \sim \mathcal{N}(0, (2\phi(0)\omega(0)/\omega'(0))^2 - \sigma^2)$ to $(\boldsymbol{w}^*)'\boldsymbol{Y}$. This addition artificially increases the standard deviation of $(\boldsymbol{w}^*)'\boldsymbol{Y}$ from σ to $2\phi(0)\frac{\omega(0)}{\omega'(0)}$, which is the threshold at which we switch from a nonrandomized rule to a randomized rule. We then make a decision according to the sign of $(\boldsymbol{w}^*)'\boldsymbol{Y} + \xi$.

The larger $\omega(0)$ is, the larger the variance of ξ is and the more dependent the choice is on the noise. As a result, given any realization of $\boldsymbol{Y}$, the probabilities of choosing policy 1 and policy 0 approach $1/2$ as $\omega(0)$ increases, which suggests that the decisions become more mixed if we impose weaker restrictions on Θ.

1.3.5 Relation to Existing Results

Theorem 1.1 contains Proposition 7(iii) of Stoye (2012) as a special case. He considers a simple setup with a specific form of partial identification. In the notation of this chapter, we observe a scalar sample $Y \sim \mathcal{N}(m(\theta), \sigma^2)$, $\theta = (\theta_1, \theta_2)' \in \mathbb{R}^2$, $m(\theta) = \theta_1$, $\Theta = \{(\theta_1, \theta_2)' \in [-1, 1]^2 : \theta_2 \in [a\theta_1 - b, a\theta_1 + b]\}$ for some known constants $a \in (0, 1]$ and $b \in (0, 1)$, and $L(\theta) = \theta_2$.[22] Stoye (2012) shows that the following rule is minimax regret:

$$\delta^*(Y) = \begin{cases} \mathbf{1}\{Y \geq 0\} & \text{if } \sigma \geq 2\phi(0)\frac{b}{a}, \\ \Phi\left(\dfrac{Y}{((2\phi(0)b/a)^2 - \sigma^2)^{1/2}} \right) & \text{if } \sigma < 2\phi(0)\frac{b}{a}. \end{cases}$$

22. Strictly speaking, Stoye (2012) also covers the case where $b \geq 1$. This case is not covered by Theorem 1.1 since Assumption 1.1(c) does not hold; $\theta^* = (0, 1)'$ attains the modulus of continuity at $\epsilon = 0$, but there exists no $\theta \in \Theta$ such that $L(\theta) = \theta_2 \neq 0$ and $\theta^* + c\theta \in \Theta$ for any small $c > 0$. Theorem 1.3 in Section 1.6.2 covers this case.

The condition $\sigma \geq 2\phi(0)\frac{b}{a}$ is equivalent to $\sigma \geq 2\phi(0)\frac{\omega(0)}{\omega'(0)}$ since $\omega(\epsilon) = \sup\{\theta_2 : \theta_1 \in [-\epsilon, \epsilon], (\theta_1, \theta_2)' \in \Theta\} = \min\{a\epsilon + b, 1\}$ in this setup. Note that the nonrandomized minimax regret rule $\delta^*(Y) = \mathbf{1}\{Y \geq 0\}$ is insensitive to any of σ, a, and b as long as $\sigma \geq 2\phi(0)\frac{b}{a}$.

Theorem 1.1 confirms that both nonrandomized and randomized rules can be minimax regret even in much more general setups than the above. At the same time, Theorem 1.1 suggests that the nonrandomized minimax regret rule $\mathbf{1}\{\boldsymbol{m}(\theta_{\epsilon^*})'\boldsymbol{Y} \geq 0\}$ may be sensitive to σ and Θ, since ϵ^* and θ_{ϵ^*} depend on them. Therefore, the robustness of the nonrandomized minimax regret rule to the error variance and to the parameter space is not a general property.[23]

In a special case of this chapter's setup, Ishihara and Kitagawa (2021) characterize the decision rule that minimizes the maximum regret within the class of decision rules of the form $\delta(\boldsymbol{Y}) = \mathbf{1}\{\boldsymbol{w}'\boldsymbol{Y} \geq 0\}$, where $\boldsymbol{w} \in \mathbb{R}^n$. Theorem 1.1 shows that this restricted class contains the minimax regret rule when $\sigma \geq 2\phi(0)\frac{\omega(0)}{\omega'(0)}$ and does not when $\sigma < 2\phi(0)\frac{\omega(0)}{\omega'(0)}$.

1.4 Application to Eligibility Cutoff Choice

Theorem 1.1 provides a procedure to compute a minimax regret rule for the example in Section 1.2.3. In this section, I provide the formula for the minimax regret rule and discuss how the rule depends on the specification of the Lipschitz constant C and the new cutoff c_1.

I first normalize $\boldsymbol{Y}$ and $\boldsymbol{m}(\cdot)$ by left multiplying them by $\boldsymbol{\Sigma}^{-1/2}$ so that the variance-covariance matrix of the sample is a diagonal matrix:

$$\tilde{\boldsymbol{Y}} \sim \mathcal{N}(\tilde{\boldsymbol{m}}(f), \boldsymbol{I}_n),$$

where $\tilde{\boldsymbol{Y}} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{Y} = (Y_1/\sigma(x_1, d_1), ..., Y_n/\sigma(x_n, d_n))'$, and $\tilde{\boldsymbol{m}}(f) = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{m}(f) = (f(x_1, d_1)/\sigma(x_1, d_1), ..., f(x_n, d_n)/\sigma(x_n, d_n))'$. For illustration, I focus on the Lipschitz class $\mathcal{F} = \mathcal{F}_{\text{Lip}}(C)$ and suppose that the welfare is the sample average of the expected outcome.

23. It is possible to come up with an example where the nonrandomized rule depends on σ and Θ if $\boldsymbol{Y}$ is two or higher dimensional. This suggests that the robustness property is specific to the problem with a scalar sample, where a reasonable nonrandomized decision rule is only either $\mathbf{1}\{Y \geq 0\}$ or $\mathbf{1}\{Y < 0\}$.

The welfare difference is given by

$$L(f) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{c_1 \leq x_i < c_0\}[f(x_i, 1) - f(x_i, 0)].$$

Below, I first verify that Assumption 1.1 holds and then apply Theorem 1.1 to derive the minimax regret rule. Assumption 1.1(c) straightforwardly holds with $\iota(x, d) = d$ for all $x \in \mathbb{R}$, provided that Assumption 1.1(a) holds. Assumption 1.1(d) is shown to hold in Appendix 1.A.4.

I verify Assumption 1.1(a) and (b) by deriving a closed-form expression for a value of f that attains the modulus of continuity at ϵ when ϵ is sufficiently small. The modulus of continuity $\omega(\epsilon; L, \tilde{\boldsymbol{m}}, \mathcal{F}_{\text{Lip}}(C))$ is computed by solving

$$\sup_{f \in \mathcal{F}_{\text{Lip}}(C)} \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{c_1 \leq x < c_0\}[f(x_i, 1) - f(x_i, 0)] \quad s.t. \quad \sum_{i=1}^{n} \frac{f(x_i, d_i)^2}{\sigma^2(x_i, d_i)} \leq \epsilon^2. \tag{1.3}$$

The unknown parameter f is infinite dimensional, but the objective and the norm constraint $\sum_{i=1}^{n} \frac{f(x_i,d_i)^2}{\sigma^2(x_i,d_i)} \leq \epsilon^2$ depend on f only through its values at $(x_1, 0), ..., (x_n, 0), (x_1, 1), ..., (x_n, 1)$. This optimization problem can be reduced to the following convex optimization problem with $2n$ unknowns and $1 + n(n-1)$ inequality constraints by a slight modification of Theorem 2.2 in Armstrong and Kolesár (2021):

$$\max_{(f(x_i,0), f(x_i,1))_{i=1,...,n} \in \mathbb{R}^{2n}} \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{c_1 \leq x_i < c_0\}[f(x_i, 1) - f(x_i, 0)] \tag{1.4}$$

$$s.t. \quad \sum_{i=1}^{n} \frac{f(x_i, d_i)^2}{\sigma^2(x_i, d_i)} \leq \epsilon^2, \ f(x_i, d) - f(x_j, d) \leq C|x_i - x_j|, \quad d \in \{0, 1\}, i, j \in \{1, ..., n\}.$$

Once we find a solution $(f(x_i, 0), f(x_i, 1))_{i=1,...,n}$ to (1.4), we can always find a function $f \in \mathcal{F}_{\text{Lip}}(C)$ that interpolates the points $(x_i, f(x_i, 0)), (x_i, f(x_i, 1))$, $i = 1, ..., n$ (Beliakov, 2006, Theorem 4), which is a solution to the original problem (1.3).

Now, I show that the problem (1.4) has a closed-form solution for any sufficiently small $\epsilon \geq 0$. The derivation utilizes the specific treatment assignment rule in the RD, namely $d_i = \mathbf{1}\{x_i \geq c_0\}$. Let $\tilde{n} = \sum_{i=1}^{n} \mathbf{1}\{c_1 \leq x_i < c_0\}$ denote the number of units whose treatment status would be changed if the cutoff were changed. Additionally, let $x_{+,\min} =$

$\min\{x_i : x_i \geq c_0\}$ be the value of x of the treated unit closest to the original cutoff c_0, and let $\sigma^2_{+,\min} = \sigma^2(x_{+,\min}, 1)$. To simplify the exposition, I assume that $x_i \neq x_j$ for any $i \neq j$, $i, j = 1, ..., n$, in what follows.[24]

Proposition 1.1 (Solution to Modulus Problem for Cutoff Choice)**.** *Suppose that* $d_i = \mathbf{1}\{x_i \geq c_0\}$ *for all* $i = 1, ..., n$ *and that* $x_i \neq x_j$ *for any* $i \neq j$, $i, j = 1, ..., n$. *Then, there exists* $\bar{\epsilon} > 0$ *such that for any* $\epsilon \in [0, \bar{\epsilon}]$, *one solution to (1.4) is given by*

$$f_\epsilon(x_i, 0) = \begin{cases} 0 & \text{if } x_i < c_1 \text{ or } x_i \geq c_0, \\ -\dfrac{\sigma^2(x_i, 0)\epsilon}{\bar{\sigma}} & \text{if } c_1 \leq x_i < c_0, \end{cases}$$

$$f_\epsilon(x_i, 1) = \begin{cases} 0 & \text{if } x_i > x_{+,\min}, \\ C(x_{+,\min} - x_i) + \dfrac{\tilde{n}\sigma^2_{+,\min}\epsilon}{\bar{\sigma}} & \text{if } x_i \leq x_{+,\min}, \end{cases}$$

and the modulus of continuity is given by

$$\omega(\epsilon; L, \tilde{\boldsymbol{m}}, \mathcal{F}_{\text{Lip}}(C)) = C\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{c_1 \leq x_i < c_0\}[x_{+,\min} - x_i] + \frac{\bar{\sigma}\epsilon}{n},$$

where $\bar{\sigma} = (\tilde{n}^2\sigma^2_{+,\min} + \sum_{i:c_1 \leq x_i < c_0} \sigma^2(x_i, 0))^{1/2}$.

Proof. See Appendix 1.B.2. □

A brief explanation of this result is as follows. Since $d_i = \mathbf{1}\{x_i \geq c_0\}$, the norm constraint of (1.4) does not depend on $f(x_i, 1)$ for i with $x_i < c_0$. The upper bound on $f(x_i, 1)$ for such unit i is $C(x_{+,\min} - x_i) + f(x_{+,\min}, 1)$ under the Lipschitz constraint; $(x_i, f(x_i, 1))$ lies on the straight line with slope $-C$ that goes through $(x_{+,\min}, f(x_{+,\min}, 1))$. Given a value of $f(x_{+,\min}, 1)$, the objective of (1.4) then becomes $C\frac{1}{n}\sum_{i:c_1 \leq x_i < c_0}[x_{+,\min} - x_i] + \frac{\tilde{n}}{n} f(x_{+,\min}, 1) - \frac{1}{n}\sum_{i:c_1 \leq x_i < c_0} f(x_i, 0)$, which is a constant plus a weighted sum of $f(x_{+,\min}, 1)$ and $f(x_i, 0)$ for i with $c_1 \leq x_i < c_0$. By maximizing this under the norm constraint, we obtain the display of f_ϵ in Proposition 1.1, which turns out to satisfy the Lipschitz constraint for any sufficiently small ϵ.

24. It is possible to obtain a closed-form solution without this assumption at the cost of making the presentation more complex.

Assumption 1.1(a) immediately follows from Proposition 1.1. Moreover, for any $\epsilon \in [0, \bar{\epsilon}]$, $\frac{\tilde{\boldsymbol{m}}(f_\epsilon)}{\|\tilde{\boldsymbol{m}}(f_\epsilon)\|} = \frac{\tilde{\boldsymbol{m}}(f_\epsilon)}{\epsilon}$ is constant and equal to $\boldsymbol{w}^* = (w_1^*, ..., w_n^*)'$, where

$$w_i^* = \begin{cases} 0 & \text{if } x_i < c_1 \text{ or } x_i > x_{+,\min}, \\ -\dfrac{\sigma(x_i, 0)}{\bar{\sigma}} & \text{if } c_1 \le x_i < c_0, \\ \dfrac{\tilde{n}\sigma_{+,\min}}{\bar{\sigma}} & \text{if } x_i = x_{+,\min}. \end{cases} \tag{1.5}$$

Therefore, Assumption 1.1(b) holds.

Now, I apply Theorem 1.1 to derive the minimax regret rule. By Proposition 1.1, we obtain closed-form expressions for $\omega(0; L, \tilde{\boldsymbol{m}}, \mathcal{F}_{\text{Lip}}(C))$ and $\omega'(0; L, \tilde{\boldsymbol{m}}, \mathcal{F}_{\text{Lip}}(C))$:

$$\omega(0; L, \tilde{\boldsymbol{m}}, \mathcal{F}_{\text{Lip}}(C)) = \frac{C}{n} \sum_{i=1}^{n} \mathbf{1}\{c_1 \le x_i < c_0\}[x_{+,\min} - x_i], \ \omega'(0; L, \tilde{\boldsymbol{m}}, \mathcal{F}_{\text{Lip}}(C)) = \frac{\bar{\sigma}}{n}.$$

Let

$$\sigma^* := 2\phi(0) \frac{\omega(0; L, \tilde{\boldsymbol{m}}, \mathcal{F}_{\text{Lip}}(C))}{\omega'(0; L, \tilde{\boldsymbol{m}}, \mathcal{F}_{\text{Lip}}(C))} = 2\phi(0) C \sum_{i=1}^{n} \mathbf{1}\{c_1 \le x_i < c_0\}[x_{+,\min} - x_i]/\bar{\sigma}. \tag{1.6}$$

Recall that $\tilde{\boldsymbol{Y}} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{Y} = (Y_1/\sigma(x_1, d_1), ..., Y_n/\sigma(x_n, d_n))'$, $\tilde{\boldsymbol{m}}(f) = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{m}(f) = (f(x_1, d_1)/\sigma(x_1, d_1), ..., f(x_n, d_n)/\sigma(x_n, d_n))'$, and the variance of $\tilde{\boldsymbol{Y}}$ is $\boldsymbol{I}_n$. Let $\epsilon^* \in \arg\max_{\epsilon \in [0, a^*]} \omega(\epsilon; L, \tilde{\boldsymbol{m}}, \mathcal{F}_{\text{Lip}}(C))\Phi(-\epsilon)$ and $(f_{\epsilon^*}(x_i, 0), f_{\epsilon^*}(x_i, 1))_{i=1,...,n}$ solve the problem (1.4) for $\epsilon = \epsilon^*$. By Theorem 1.1, the following rule is minimax regret:

$$\delta^*(\boldsymbol{Y}) = \begin{cases} \mathbf{1}\left\{\sum_{i=1}^n f_{\epsilon^*}(x_i, d_i) Y_i/\sigma^2(x_i, d_i) \ge 0\right\} & \text{if } \ 1 > \sigma^*, \\ \mathbf{1}\left\{\sum_{i=1}^n w_i^* Y_i/\sigma(x_i, d_i) \ge 0\right\} & \text{if } \ 1 = \sigma^*, \\ \Phi\left(\dfrac{\sum_{i=1}^n w_i^* Y_i/\sigma(x_i, d_i)}{((\sigma^*)^2 - 1)^{1/2}}\right) & \text{if } \ 1 < \sigma^*. \end{cases} \tag{1.7}$$

The minimax regret rule makes a decision based on a weighted sum of $Y_1, ..., Y_n$.

To understand how the rule differs across different values of the Lipschitz constant C, suppose first that the magnitude of C is moderate so that σ^* is marginally smaller than 1. In this case, ϵ^* tends to be sufficiently small, which implies that $\frac{f_{\epsilon^*}(x_i, d_i)/\sigma(x_i, d_i)}{\epsilon^*} = w_i^*$ by

Proposition 1.1. The minimax regret rule is given by

$$\begin{aligned}\delta^*(\boldsymbol{Y}) &= \mathbf{1}\left\{\sum_{i=1}^n f_{\epsilon^*}(x_i, d_i) Y_i/\sigma^2(x_i, d_i) \geq 0\right\} \\ &= \mathbf{1}\left\{\sum_{i=1}^n w_i^* Y_i/\sigma(x_i, d_i) \geq 0\right\} = \mathbf{1}\left\{Y_{+,\min} - \frac{1}{\tilde{n}} \sum_{i:c_1 \leq x_i < c_0} Y_i \geq 0\right\},\end{aligned}$$

where $Y_{+,\min} = Y_i$ for i with $x_i = x_{+,\min}$. $Y_{+,\min} - \frac{1}{\tilde{n}} \sum_{i:c_1 \leq x_i < c_0} Y_i$ is the difference between the outcome of the treated unit closest to the status quo cutoff c_0 and the mean outcome across the untreated units between the two cutoffs c_0 and c_1. This difference can be interpreted as an estimator of the effect of the cutoff change. The outcomes of the other units are not used to construct the estimator. The minimax regret rule makes a decision according to its sign.

On the other hand, if the Lipschitz constant C is small enough so that σ^* is substantially smaller than 1, nonzero weights may be assigned to some of the other units, that is, $f_{\epsilon^*}(x_i, d_i)$ may be nonzero for some of the units with $x_i < c_1$ or $x_i > x_{+,\min}$. If the Lipschitz constant C is large enough so that $\sigma^* > 1$, the minimax regret rule is a randomized rule based on $Y_{+,\min} - \frac{1}{\tilde{n}} \sum_{i:c_1 \leq x_i < c_0} Y_i$.

Whether the minimax regret rule is randomized or not depends not only on the Lipschitz constant C but also on the cutoffs c_0 and c_1 and $\bar{\sigma} = (\tilde{n}^2 \sigma^2_{+,\min} + \sum_{i:c_1 \leq x_i < c_0} \sigma^2(x_i, 0))^{1/2}$. To investigate their relationships, suppose that $\sigma^2(x_i, d_i) = \sigma^2$ for all i for some $\sigma^2 > 0$ for simplicity. In this situation, $\bar{\sigma} = (\tilde{n}^2 + \tilde{n})^{1/2}\sigma$, and

$$\sigma^* = \frac{2\phi(0) C \frac{1}{\tilde{n}} \sum_{i=1}^n \mathbf{1}\{c_1 \leq x_i < c_0\}[x_{+,\min} - x_i]}{(1 + 1/\tilde{n})^{1/2}\, \sigma}.$$

σ^* is nonincreasing in c_1 since $\frac{1}{\tilde{n}} \sum_{i=1}^n \mathbf{1}\{c_1 \leq x_i < c_0\}[x_{+,\min} - x_i]$ and $\tilde{n}$ are nonincreasing in c_1.[25] Furthermore, σ^* is decreasing in σ. Therefore, the minimax regret rule is nonrandomized when c_1 is large (i.e., when the cutoff change $c_0 - c_1$ is small) or σ is large. The minimax regret rule is randomized otherwise.

25. Whether σ^* is increasing in c_0 or not depends on the empirical distribution of x_i.

1.4.1 Practical Implementation

Here, I summarize the procedure for computing the minimax regret rule and discuss practical issues. Given the conditional variances $\sigma^2(x_i, d_i)$, $i = 1, ..., n$ and the Lipschitz constant C, the minimax regret rule is computed as follows.

1. Compute σ^* using the closed-form expression (1.6).

2. If $1 > \sigma^*$, find $\epsilon^* \in \arg\max_{\epsilon \in [0, a^*]} \omega(\epsilon)\Phi(-\epsilon)$ and compute f_{ϵ^*} that attains the modulus of continuity at ϵ^*. For each $\epsilon \geq 0$, $\omega(\epsilon)$ is computed by solving the convex optimization problem (1.4).[26] An efficient method for computing ϵ^* is provided in Appendix 1.A.6.

3. If $1 \leq \sigma^*$, compute $\boldsymbol{w}^*$ using the closed-form expression (1.5).

4. Construct the decision rule according to (1.7).

In practice, the conditional variance $\sigma^2(x_i, d_i)$ is unknown. I suggest using a consistent estimator in place of the true $\sigma^2(x_i, d_i)$. The conditional variance can be estimated, for example, by applying a local linear regression to the squared residuals (Fan and Yao, 1998) or by the nearest-neighbor variance estimator (Abadie and Imbens, 2006). In the case where unit i represents a group of individuals and where Y_i is the sample mean outcome within group i, it is natural to use the conventional standard error of the sample mean as $\sigma(x_i, d_i)$.

Implementation of the minimax regret rule requires choosing the Lipschitz constant C. In principle, it is not possible to choose the Lipschitz constant C that applies to both sides of the cutoff c_0 in a data-driven way since we only observe outcomes either under treatment or under no treatment on each side. It is, however, possible to estimate a lower bound on C by using the following observation: if $f \in \mathcal{F}_{\text{Lip}}(C)$ is differentiable, a lower bound on C is given by $\max\left\{\max_{\tilde{x} \geq c_0} \left|\frac{\partial f(\tilde{x},1)}{\partial x}\right|, \max_{\tilde{x} < c_0} \left|\frac{\partial f(\tilde{x},0)}{\partial x}\right|\right\}$ since $\left|\frac{\partial f(\tilde{x},d)}{\partial x}\right| \leq C$ for all $\tilde{x}$ and d. To estimate a lower bound, we could estimate the derivatives $\frac{\partial f(\tilde{x},1)}{\partial x}$ for $\tilde{x} \geq c_0$ and $\frac{\partial f(\tilde{x},0)}{\partial x}$ for $\tilde{x} < c_0$ by a local polynomial regression and then take the maximum of their absolute

26. In the empirical application in Section 1.7, I solve the convex optimization problem using CVXPY, a Python-embedded modeling language for convex optimization problems (Diamond and Boyd, 2016; Agrawal, Verschueren, Diamond and Boyd, 2018).

values over a plausible, relevant interval.[27] In practice, I recommend considering a range of plausible choices of C, including the estimated lower bound, to conduct a sensitivity analysis. I implement this approach for my empirical application in Section 1.7.

1.5 Additional Implications of the Main Result

In this section, I present two implications of Theorem 1.1. First, I discuss the difference between the minimax regret rule and a plug-in decision rule based on a linear minimax mean squared error (MSE) estimator. Second, I investigate the properties of the minimax regret rule when the welfare difference is point identified.

1.5.1 Comparison with a Plug-in Rule Based on a Linear Minimax Mean Squared Error Estimator

When $\sigma > 2\phi(0)\frac{\omega(0)}{\omega'(0)}$, the minimax regret rule is a nonrandomized rule that makes a choice according to the sign of a weighted sum of $\boldsymbol{Y}$. In this section, I compare the nonrandomized minimax regret rule with a plug-in rule based on a linear minimax MSE estimator.[28]

To define the alternative rule, let $\hat{L}_{\mathrm{MSE}}(\boldsymbol{Y}) = \boldsymbol{w}_{\mathrm{MSE}}'\boldsymbol{Y}$ be a *linear minimax MSE estimator*, where

$$\boldsymbol{w}_{\mathrm{MSE}} \in \arg\min_{\boldsymbol{w}\in\mathbb{R}^n} \sup_{\theta\in\Theta} \mathbb{E}_\theta[(\boldsymbol{w}'\boldsymbol{Y} - L(\theta))^2].$$

$\hat{L}_{\mathrm{MSE}}(\boldsymbol{Y})$ is an estimator of $L(\theta)$ that has the smallest worst-case MSE within the class of linear estimators. I define the *plug-in MSE rule* as $\delta_{\mathrm{MSE}}(\boldsymbol{Y}) = \mathbf{1}\{\hat{L}_{\mathrm{MSE}}(\boldsymbol{Y}) \geq 0\}$, which makes a choice according to the sign of the linear minimax MSE estimator of $L(\theta)$.

Donoho (1994) characterizes $\hat{L}_{\mathrm{MSE}}(\boldsymbol{Y})$ using the modulus of continuity. For a simple statement of Donoho (1994)'s result, suppose that $\omega(\cdot)$ is differentiable and that Θ is convex

27. Simply taking the maximum of the estimated derivatives could raise a concern of upward bias. One could use the method for intersection bounds developed by Chernozhukov, Lee and Rosen (2013) to address it.

28. In Appendix 1.A.2, I also compare the minimax regret rule with a hypothesis testing rule that chooses policy 1 if a hypothesis that supports policy 0 is rejected.

and centrosymmetric. Let $\epsilon_{\text{MSE}} > 0$ solve

$$\frac{\epsilon^2}{\epsilon^2 + \sigma^2} = \frac{\omega'(\epsilon)\epsilon}{\omega(\epsilon)}.$$

The linear minimax MSE estimator is then given by $\hat{L}_{\text{MSE}}(\boldsymbol{Y}) = \frac{\omega'(\epsilon_{\text{MSE}})}{\epsilon_{\text{MSE}}} \boldsymbol{m}(\theta_{\epsilon_{\text{MSE}}})'\boldsymbol{Y}$, where $\theta_{\epsilon_{\text{MSE}}}$ attains the modulus of continuity at ϵ_{MSE} with $\|\boldsymbol{m}(\theta_{\epsilon_{\text{MSE}}})\| = \epsilon_{\text{MSE}}$. The plug-in MSE rule is $\delta_{\text{MSE}}(\boldsymbol{Y}) = \mathbf{1}\{\boldsymbol{m}(\theta_{\epsilon_{\text{MSE}}})'\boldsymbol{Y} \geq 0\}$.

Recall that the minimax regret rule is $\delta^*(\boldsymbol{Y}) = \mathbf{1}\{\boldsymbol{m}(\theta_{\epsilon^*})'\boldsymbol{Y} \geq 0\}$ if $\sigma > 2\phi(0)\frac{\omega(0)}{\omega'(0)}$, where ϵ^* solves $\max_{\epsilon \in [0, a^*\sigma]} \omega(\epsilon)\Phi(-\epsilon/\sigma)$.

Proposition 1.2 (Comparison with Plug-in MSE Rule)**.** *Suppose that $\omega(\cdot)$ is differentiable, and let ϵ_{MSE} solve $\frac{\epsilon^2}{\epsilon^2+\sigma^2} = \frac{\omega'(\epsilon)\epsilon}{\omega(\epsilon)}$ and ϵ^* solve $\max_{\epsilon \in [0, a^*\sigma]} \omega(\epsilon)\Phi(-\epsilon/\sigma)$. Then, $\epsilon^* < \epsilon_{\text{MSE}}$.*

Proof. See Appendix 1.B.3. □

I provide an implication of Proposition 1.2 through the following result from Donoho (1994) and Low (1995): the optimal bias-variance frontier in the estimation of $L(\theta)$ can be traced out by a class of linear estimators $\{\hat{L}_\epsilon(\boldsymbol{Y})\}_{\epsilon>0}$ of the form $\hat{L}_\epsilon(\boldsymbol{Y}) = \frac{\omega'(\epsilon)}{\epsilon}\boldsymbol{m}(\theta_\epsilon)'\boldsymbol{Y}$. Here, θ_ϵ attains the modulus of continuity at ϵ with $\|\boldsymbol{m}(\theta_\epsilon)\| = \epsilon$. Specifically, for each $\epsilon > 0$, $\hat{L}_\epsilon(\boldsymbol{Y})$ minimizes the maximum bias among all linear estimators with variance bounded by $\text{Var}(\hat{L}_\epsilon(\boldsymbol{Y})) = (\sigma\omega'(\epsilon))^2$:

$$\frac{\omega'(\epsilon)}{\epsilon}\boldsymbol{m}(\theta_\epsilon) \in \arg\min_{\boldsymbol{w}\in\mathbb{R}^n} \overline{\text{Bias}}_\Theta(\boldsymbol{w}'\boldsymbol{Y}) \;\; s.t. \;\; \text{Var}(\boldsymbol{w}'\boldsymbol{Y}) \leq (\sigma\omega'(\epsilon))^2,$$

where $\overline{\text{Bias}}_\Theta(\boldsymbol{w}'\boldsymbol{Y}) = \sup_{\theta\in\Theta} \mathbb{E}_\theta[\boldsymbol{w}'\boldsymbol{Y} - L(\theta)]$ is the maximum bias of $\boldsymbol{w}'\boldsymbol{Y}$ over Θ. As ϵ increases, the maximum bias $\overline{\text{Bias}}_\Theta(\hat{L}_\epsilon(\boldsymbol{Y}))$ increases and the variance $\text{Var}(\hat{L}_\epsilon(\boldsymbol{Y})) = (\sigma\omega'(\epsilon))^2$ decreases. ϵ_{MSE} minimizes the worst-case MSE $\sup_{\theta\in\Theta} \mathbb{E}_\theta[(\hat{L}_\epsilon(\boldsymbol{Y}) - L(\theta))^2] = \overline{\text{Bias}}_\Theta(\hat{L}_\epsilon(\boldsymbol{Y}))^2 + \text{Var}(\hat{L}_\epsilon(\boldsymbol{Y}))$.

Since $\delta^*(\boldsymbol{Y}) = \mathbf{1}\{\boldsymbol{m}(\theta_{\epsilon^*})'\boldsymbol{Y} \geq 0\} = \mathbf{1}\{\hat{L}_{\epsilon^*}(\boldsymbol{Y}) \geq 0\}$, the minimax regret rule $\delta^*(\boldsymbol{Y})$ can be viewed as a rule that makes a choice according to the sign of the linear estimator $\hat{L}_{\epsilon^*}(\boldsymbol{Y})$. Proposition 1.2 implies that the corresponding linear estimator $\hat{L}_{\epsilon^*}(\boldsymbol{Y})$ places more importance on the bias than on the variance compared with the linear minimax MSE

estimator $\hat{L}_{\text{MSE}}(\boldsymbol{Y})$. In other words, ϵ^* minimizes a particular weighted average of the squared bias and variance $\alpha \cdot \overline{\text{Bias}}_\Theta(\hat{L}_\epsilon(\boldsymbol{Y}))^2 + (1-\alpha) \cdot \text{Var}(\hat{L}_\epsilon(\boldsymbol{Y}))$ for some $\alpha \in [1/2, 1]$.[29] This result suggests that the plug-in MSE rule is not necessarily optimal under the minimax regret criterion.

1.5.2 Minimax Regret Rules Under Point Identification of Welfare Difference

Here, I discuss the properties of the minimax regret rule in special cases where the welfare difference $L(\theta)$ is point identified. The starting point is the following result, a corollary of Theorem 1.1. The minimax regret rule is locally insensitive to Θ in some special cases.

Corollary 1.1 (Robustness)**.** *Let Θ be convex and centrosymmetric, and let $\theta^* \in \mathbb{V}$ solve $\sup_{\theta \in \mathbb{V}: \|\boldsymbol{m}(\theta)\| \leq 1} L(\theta)$, where $\mathbb{V}$ is the vector space that θ belongs to. Suppose that $\{\epsilon\theta^* : 0 \leq \epsilon \leq a^*\sigma\} \subset \Theta$. Then, the decision rule*

$$\delta^*(\boldsymbol{Y}) = \mathbf{1}\left\{\boldsymbol{m}(\theta^*)'\boldsymbol{Y} \geq 0\right\}$$

is minimax regret. The minimax risk is given by $\mathcal{R}(\sigma; \Theta) = a^*\sigma L(\theta^*)\Phi(-a^*)$.

Proof. See Appendix 1.B.4. □

Note that θ^* depends on $\mathbb{V}$, not on Θ. Corollary 1.1 implies that the minimax regret rule and minimax risk are robust to the choice of Θ as long as Θ is large enough to contain $\{\epsilon\theta^* : 0 \leq \epsilon \leq a^*\sigma\}$.

Notably, the above result holds only for cases where the welfare difference $L(\theta)$ is identified when $\boldsymbol{m}(\theta) = \mathbf{0}$. This is because it is shown that $\omega(0) = 0$ under the conditions in Corollary 1.1 (see Appendix 1.B.4), which means that $\{L(\theta) : \boldsymbol{m}(\theta) = \mathbf{0}, \theta \in \Theta\} = \{0\}$.

I illustrate this result through an example with linear regression models. This example nests the one in Section 1.2.3 when $\mathcal{F}$ is a class of polynomial functions. I show that the

29. In Figure 1.10 in Appendix 1.C, I report the weight $\alpha \in [1/2, 1]$ to empirically quantify how much importance the minimax regret rule places on the bias in the empirical illustration in Section 1.7.

minimax regret rule is based on the best linear unbiased estimator of the parameter θ when Θ is sufficiently large.

1.5.2.1 Example: Linear Regression Models

Suppose it is known to the policy maker that the sample $\boldsymbol{Y} = (Y_1, ..., Y_n)'$ is generated by the following linear regression model:

$$\boldsymbol{Y} = \boldsymbol{X}\theta + \boldsymbol{U}, \quad \boldsymbol{U} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{X}$ is a fixed $n \times k$ design matrix that stacks k-dimensional covariate vectors of n units, $\theta \in \Theta \subset \mathbb{V} = \mathbb{R}^k$, and $\boldsymbol{\Sigma}$ is a known variance-covariance matrix. This model is a special case of the general one where $\boldsymbol{m}(\theta) = \boldsymbol{X}\theta$. Suppose also that the welfare difference is given by

$$L(\theta) = \boldsymbol{l}'\theta$$

for some known $\boldsymbol{l} \in \mathbb{R}^k$. This setup covers the example in Section 1.2.3 when $\mathcal{F}$ is a class of polynomial functions.[30]

I normalize $\boldsymbol{Y}$ and $\boldsymbol{m}(\cdot)$ by left multiplying them by $\boldsymbol{\Sigma}^{-1/2}$ so that the variance-covariance matrix of the sample is a diagonal matrix:

$$\tilde{\boldsymbol{Y}} \sim \mathcal{N}(\tilde{\boldsymbol{m}}(\theta), \boldsymbol{I}_n),$$

where $\tilde{\boldsymbol{Y}} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{Y}$ and $\tilde{\boldsymbol{m}}(\theta) = \tilde{\boldsymbol{X}}\theta$ with $\tilde{\boldsymbol{X}} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}$.

In this example, $L(\theta)$ is identified as long as the rank condition holds. To see this, suppose that $\tilde{\boldsymbol{m}}(\theta) = \boldsymbol{\mu}$ for some $\boldsymbol{\mu} \in \mathbb{R}^n$. If $\tilde{\boldsymbol{X}}$ is of rank k so that $\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}}$ is invertible, then $\theta = (\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}'\boldsymbol{\mu}$, and hence $L(\theta) = \boldsymbol{l}'(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}'\boldsymbol{\mu}$.

30. Suppose that $\mathcal{F} = \mathcal{F}_{\text{Pol}}(p) \coloneqq \{f : f(x,d) = (\boldsymbol{x}', d\boldsymbol{x}')\theta \text{ for some } \theta \in \mathbb{R}^{2(p+1)}\}$, where $\boldsymbol{x} = (1, x, x^2, ..., x^p)' \in \mathbb{R}^{p+1}$. $\mathcal{F}_{\text{Pol}}(p)$ is the set of functions such that $f(\cdot, d)$ is a polynomial function of degree at most p for each $d \in \{0, 1\}$. Let $\boldsymbol{X}$ denote the $n \times 2(p+1)$ matrix whose i-th row is $(\boldsymbol{x}_i', d_i\boldsymbol{x}_i')$. The model $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{m}(f), \boldsymbol{\Sigma})$ with $f \in \mathcal{F}_{\text{Pol}}(p)$ then becomes $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{X}\theta, \boldsymbol{\Sigma})$ with $\theta \in \mathbb{R}^{2(p+1)}$. Since $f(x,1) - f(x,0) = (\boldsymbol{x}', 1 \cdot \boldsymbol{x}')\theta - (\boldsymbol{x}', 0 \cdot \boldsymbol{x}')\theta$, the welfare difference is given by $L(\theta) = \boldsymbol{l}'\theta$, where $\boldsymbol{l} = \left(\int \boldsymbol{1}\{c_1 \leq x < c_0\} \left[(\boldsymbol{x}', 1 \cdot \boldsymbol{x}') - (\boldsymbol{x}', 0 \cdot \boldsymbol{x}')\right] d\nu(x)\right)' \in \mathbb{R}^{2(p+1)}$.

To apply Corollary 1.1, consider the following problem:

$$\sup_{\theta \in \mathbb{R}^k} \boldsymbol{l}'\theta \ s.t. \ \left(\theta' \tilde{\boldsymbol{X}}' \tilde{\boldsymbol{X}} \theta\right)^{1/2} \leq 1.$$

Simple calculations show that the solution is $\theta^* = \frac{(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}})^{-1}\boldsymbol{l}}{(\boldsymbol{l}'(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}})^{-1}\boldsymbol{l})^{1/2}}$.

By Corollary 1.1, if $\{\epsilon\theta^* : 0 \leq \epsilon \leq a^*\} \subset \Theta$, the minimax regret rule is given by

$$\delta^*(\boldsymbol{Y}) = \mathbf{1}\left\{(\tilde{\boldsymbol{X}}\theta^*)'\tilde{\boldsymbol{Y}} \geq 0\right\} = \mathbf{1}\left\{\boldsymbol{l}'\hat{\theta}_{\text{WLS}}(\boldsymbol{Y}) \geq 0\right\},$$

where $\hat{\theta}_{\text{WLS}}(\boldsymbol{Y}) = (\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}$ is the weighted least squares (WLS) estimator of θ using $\boldsymbol{\Sigma}^{-1}$ as the weighting matrix. $\hat{\theta}_{\text{WLS}}(\boldsymbol{Y})$ is the best linear unbiased estimator of θ by the Gauss-Markov theorem. $\boldsymbol{l}'\hat{\theta}_{\text{WLS}}(\boldsymbol{Y})$ can be viewed as an estimator of $L(\theta) = \boldsymbol{l}'\theta$.

If Θ does not contain $\{\epsilon\theta^* : 0 \leq \epsilon \leq a^*\}$, the minimax regret rule does not generally admit a closed form. When a closed form is not available, we can directly use Theorem 1.1 to calculate the minimax regret rule. The discussion in Section 1.5.1 suggests that the minimax regret rule may use a linear estimator of $L(\theta) = \boldsymbol{l}'\theta$ that optimally trades off the bias and variance.

1.6 Proof of Theorem 1.1

In this section, I provide the proof of Theorem 1.1. I provide separate arguments for the nonrandomized and randomized rules. For each case, I first state an assumption weaker than the conditions in Theorem 1.1, present a result under the relaxed assumption, and then provide the proof for the more general result.

1.6.1 Nonrandomized Rule

Consider the following assumption.

Assumption 1.2 (Informative Worst Case)**.** *There exists a unique, nonzero solution to the maximization problem* $\max_{\epsilon \in [0, a^*\sigma]} \omega(\epsilon)\Phi(-\epsilon/\sigma)$.

An interpretation of this assumption is as follows. As discussed in Section 1.3.1, the

maximization problem $\max_{\epsilon\in[0,a^*\sigma]}\omega(\epsilon)\Phi(-\epsilon/\sigma)$ corresponds to the problem of finding the hardest one-dimensional subproblem. The hardest one-dimensional subproblem is $[-\theta_{\epsilon^*},\theta_{\epsilon^*}]$, where $\epsilon^*\in\arg\max_{\epsilon\in[0,a^*\sigma]}\omega(\epsilon)\Phi(-\epsilon/\sigma)$ and θ_{ϵ^*} attains the modulus of continuity at ϵ^*. If the constraint $\|\boldsymbol{m}(\theta_{\epsilon^*})\|\leq\epsilon^*$ of the modulus problem holds with equality and Assumption 1.2 holds, $\|\boldsymbol{m}(\theta_{\epsilon^*})\|>0$. Equivalently, $\boldsymbol{m}(\theta_{\epsilon^*})\neq\mathbf{0}$, which means that the signal $\boldsymbol{Y}$ under the worst-case parameter values θ_{ϵ^*} and $-\theta_{\epsilon^*}$ is informative.

The following lemma shows that Assumption 1.2 holds if $\sigma>2\phi(0)\frac{\omega(0)}{\omega'(0)}$ under Assumption 1.1(d).

Lemma 1.1. *Suppose that $\omega(\cdot)$ is differentiable at any $\epsilon\in(0,a^*\sigma]$. Then, there exists a unique solution to the maximization problem $\max_{\epsilon\in[0,a^*\sigma]}\omega(\epsilon)\Phi(-\epsilon/\sigma)$. The solution is nonzero if and only if $\sigma>2\phi(0)\frac{\omega(0)}{\omega'(0)}$.*

Proof. See Appendix 1.B.5. □

I obtain a minimax regret rule under Assumption 1.2. The statement on the nonrandomized rule in Theorem 1.1 immediately follows from the result below.

Theorem 1.2 (Nonrandomized Minimax Regret Rule)**.** *Let Θ be convex and centrosymmetric, and suppose that Assumption 1.2 holds. Let $\epsilon^*\in\arg\max_{\epsilon\in[0,a^*\sigma]}\omega(\epsilon)\Phi(-\epsilon/\sigma)$, and suppose that there exists θ_{ϵ^*} that attains the modulus of continuity at ϵ^*. Then, the decision rule $\delta^*(\boldsymbol{Y})=\mathbf{1}\{\boldsymbol{m}(\theta_{\epsilon^*})'\boldsymbol{Y}\geq 0\}$ is minimax regret. Here, $\boldsymbol{m}(\theta_{\epsilon^*})$ does not depend on the choice of θ_{ϵ^*} among those that attain the modulus of continuity at ϵ^*. The minimax risk is given by $\mathcal{R}(\sigma;\Theta)=\omega(\epsilon^*)\Phi(-\epsilon^*/\sigma)$.*

Now, I provide the proof of Theorem 1.2. The proof consists of four steps. First, I consider the simplest problem with a univariate sample and a bounded scalar parameter. Second, I use the result from the first step to solve one-dimensional subproblems, where the parameter space is restricted to a one-dimensional bounded submodel. Third, I characterize the hardest one-dimensional subproblem, that is, the one-dimensional subproblem that has the largest minimax risk. Lastly, I show that the minimax risk in the original problem is achieved by a minimax regret rule for the hardest one-dimensional subproblem. For the problem of estimation and inference, Donoho (1994) splits his proof into these steps.

However, the proof of each step is nontrivially different since the problem of policy choice and that of estimation and inference are not nested by each other; specifically, the loss function and the action space are different.

Step 1. Minimax Regret Rules for Univariate Problems. I begin with a class of univariate problems. The parameter θ is a scalar and lies in $\Theta = [-\tau, \tau]$ for some $\tau > 0$. We observe a sample $Y \sim \mathcal{N}(\theta, \sigma^2)$. This setup is a special case of the general framework where $\boldsymbol{m}(\theta) = \theta$. The welfare difference is given by $L(\theta) = \theta$.

Lemma 1.2 (Univariate Problems)**.** *Suppose that $\Theta = [-\tau, \tau]$ for some $\tau > 0$, that $\boldsymbol{m}(\theta) = \theta$, and that $L(\theta) = \theta$. Then, the decision rule $\delta^*(Y) = \mathbf{1}\{Y \geq 0\}$ is minimax regret. The minimax risk is given by*

$$\mathcal{R}_{\text{uni}}(\sigma; [-\tau, \tau]) = \begin{cases} \tau\Phi(-\tau/\sigma) & \text{if } \tau \leq a^*\sigma, \\ a^*\sigma\Phi(-a^*) & \text{if } \tau > a^*\sigma. \end{cases}$$

Proof. See Appendix 1.B.6. □

In univariate problems, the minimax regret rule makes a choice according to the sign of the sample Y. The minimax risk does not depend on τ as long as $\tau > a^*\sigma$.

Step 2. Minimax Regret Rules for One-dimensional Subproblems. Consider the original setup where θ resides in a vector space $\mathbb{V}$. Recall that $[\tilde{\theta}, \bar{\theta}] = \{(1-\lambda)\tilde{\theta} + \lambda\bar{\theta} : \lambda \in [0,1]\}$ for $\tilde{\theta}, \bar{\theta} \in \mathbb{V}$. I use Lemma 1.2 to derive minimax regret rules and the minimax risk for one-dimensional subproblems of the form $[-\bar{\theta}, \bar{\theta}]$ with $L(\bar{\theta}) > 0$ and $\boldsymbol{m}(\bar{\theta}) \neq \mathbf{0}$.

Lemma 1.3 (Informative One-dimensional Subproblems)**.** *Suppose that $\Theta = [-\bar{\theta}, \bar{\theta}]$, where $\bar{\theta} \in \mathbb{V}$, $L(\bar{\theta}) > 0$, and $\boldsymbol{m}(\bar{\theta}) \neq \mathbf{0}$. Then, the decision rule $\delta^*(\boldsymbol{Y}) = \mathbf{1}\{\boldsymbol{m}(\bar{\theta})'\boldsymbol{Y} \geq 0\}$ is minimax regret. The minimax risk is given by*

$$\mathcal{R}(\sigma; [-\bar{\theta}, \bar{\theta}]) = \frac{L(\bar{\theta})}{\|\boldsymbol{m}(\bar{\theta})\|}\mathcal{R}_{\text{uni}}(\sigma; [-\|\boldsymbol{m}(\bar{\theta})\|, \|\boldsymbol{m}(\bar{\theta})\|]).$$

Proof. See Appendix 1.B.7. □

In one-dimensional subproblems, the minimax regret rule chooses policy 1 if the sample $\boldsymbol{Y}$ agrees more with $\bar{\theta}$ (or $\boldsymbol{m}(\bar{\theta})'\boldsymbol{Y} > 0$) and chooses policy 0 if the sample $\boldsymbol{Y}$ agrees more with $-\bar{\theta}$ (or $\boldsymbol{m}(\bar{\theta})'\boldsymbol{Y} < 0$).

Step 3. Hardest One-dimensional Subproblems. Using Lemma 1.3, I characterize the supremum of the minimax risk $\mathcal{R}(\sigma; [-\bar{\theta}, \bar{\theta}])$ over all one-dimensional subproblems of the form $[-\bar{\theta}, \bar{\theta}]$, where $\bar{\theta} \in \Theta$, $L(\bar{\theta}) > 0$, and $\boldsymbol{m}(\bar{\theta}) \neq \mathbf{0}$.

First, let ϵ^* be the unique solution to $\max_{\epsilon \in [0, a^*\sigma]} \omega(\epsilon)\Phi(-\epsilon/\sigma)$, which is positive under Assumption 1.2. Let θ_{ϵ^*} attain the modulus of continuity at ϵ^*. Lemma 1.B.2 in Appendix 1.B.1 shows that the constraint $\|\boldsymbol{m}(\theta_{\epsilon^*})\| \leq \epsilon^*$ of the modulus problem holds with equality. As a result, we obtain

$$\begin{aligned}
\mathcal{R}(\sigma; [-\theta_{\epsilon^*}, \theta_{\epsilon^*}]) &= \frac{L(\theta_{\epsilon^*})}{\|\boldsymbol{m}(\theta_{\epsilon^*})\|} \mathcal{R}_{\text{uni}}(\sigma; [-\|\boldsymbol{m}(\theta_{\epsilon^*})\|, \|\boldsymbol{m}(\theta_{\epsilon^*})\|]) \\
&= \frac{\omega(\epsilon^*)}{\epsilon^*} \mathcal{R}_{\text{uni}}(\sigma; [-\epsilon^*, \epsilon^*]) \\
&= \omega(\epsilon^*)\Phi(-\epsilon^*/\sigma),
\end{aligned}$$

where the first equality follows from Lemma 1.3 and the last follows from Lemma 1.2 and the fact that $\epsilon^* \leq a^*\sigma$.

Now, I use the modulus of continuity $\omega(\epsilon)$ to write

$$\begin{aligned}
\sup_{\bar{\theta} \in \Theta: L(\bar{\theta}) > 0, \boldsymbol{m}(\bar{\theta}) \neq \mathbf{0}} \mathcal{R}(\sigma; [-\bar{\theta}, \bar{\theta}]) &= \sup_{\bar{\theta} \in \Theta: L(\bar{\theta}) > 0, \boldsymbol{m}(\bar{\theta}) \neq \mathbf{0}} \frac{L(\bar{\theta})}{\|\boldsymbol{m}(\bar{\theta})\|} \mathcal{R}_{\text{uni}}(\sigma; [-\|\boldsymbol{m}(\bar{\theta})\|, \|\boldsymbol{m}(\bar{\theta})\|]) \\
&= \sup_{\epsilon > 0} \left\{ \sup_{\bar{\theta} \in \Theta: \|\boldsymbol{m}(\bar{\theta})\| = \epsilon} \frac{L(\bar{\theta})}{\|\boldsymbol{m}(\bar{\theta})\|} \mathcal{R}_{\text{uni}}(\sigma; [-\|\boldsymbol{m}(\bar{\theta})\|, \|\boldsymbol{m}(\bar{\theta})\|]) \right\} \\
&= \sup_{\epsilon > 0} \left\{ \frac{\sup_{\bar{\theta} \in \Theta: \|\boldsymbol{m}(\bar{\theta})\| = \epsilon} L(\bar{\theta})}{\epsilon} \mathcal{R}_{\text{uni}}(\sigma; [-\epsilon, \epsilon]) \right\} \\
&\leq \sup_{\epsilon > 0} \frac{\omega(\epsilon)}{\epsilon} \mathcal{R}_{\text{uni}}(\sigma; [-\epsilon, \epsilon]),
\end{aligned}$$

where the last inequality holds by the definition of $\omega(\epsilon)$. By Lemma 1.2,

$$\frac{\omega(\epsilon)}{\epsilon}\mathcal{R}_{\text{uni}}(\sigma;[-\epsilon,\epsilon]) = \begin{cases} \omega(\epsilon)\Phi(-\epsilon/\sigma) & \text{if } \epsilon \le a^*\sigma, \\ \frac{\omega(\epsilon)}{\epsilon}a^*\sigma\Phi(-a^*) & \text{if } \epsilon > a^*\sigma. \end{cases}$$

Since $\omega(\epsilon)$ is concave, $\frac{\omega(\epsilon)}{\epsilon}$ is nonincreasing, so that $\sup_{\epsilon>a^*\sigma}\frac{\omega(\epsilon)}{\epsilon}a^*\sigma\Phi(-a^*) = \omega(a^*\sigma)\Phi(-a^*)$. Therefore,

$$\sup_{\epsilon>0}\frac{\omega(\epsilon)}{\epsilon}\mathcal{R}_{\text{uni}}(\sigma;[-\epsilon,\epsilon]) = \sup_{0<\epsilon\le a^*\sigma}\omega(\epsilon)\Phi(-\epsilon/\sigma) = \omega(\epsilon^*)\Phi(-\epsilon^*/\sigma).$$

Hence, $\sup_{\bar\theta\in\Theta:L(\bar\theta)>0,\boldsymbol{m}(\bar\theta)\neq\mathbf{0}}\mathcal{R}(\sigma;[-\bar\theta,\bar\theta]) \le \omega(\epsilon^*)\Phi(-\epsilon^*/\sigma)$.

Since $\mathcal{R}(\sigma;[-\theta_{\epsilon^*},\theta_{\epsilon^*}]) = \omega(\epsilon^*)\Phi(-\epsilon^*/\sigma)$, it follows that

$$\sup_{\bar\theta\in\Theta:L(\bar\theta)>0,\boldsymbol{m}(\bar\theta)\neq\mathbf{0}}\mathcal{R}(\sigma;[-\bar\theta,\bar\theta]) = \mathcal{R}(\sigma;[-\theta_{\epsilon^*},\theta_{\epsilon^*}]) = \omega(\epsilon^*)\Phi(-\epsilon^*/\sigma).$$

Therefore, $[-\theta_{\epsilon^*},\theta_{\epsilon^*}]$ is one of the hardest one-dimensional subproblems. Its minimax risk is $\omega(\epsilon^*)\Phi(-\epsilon^*/\sigma)$.

Step 4. Minimax Regret Rules for the Original Problem. By Lemma 1.3, the decision rule $\delta^*(\boldsymbol{Y}) = \mathbf{1}\{\boldsymbol{m}(\theta_{\epsilon^*})'\boldsymbol{Y} \ge 0\}$ is minimax regret for the one-dimensional subproblem $[-\theta_{\epsilon^*},\theta_{\epsilon^*}]$. Since $\boldsymbol{m}(\theta_{\epsilon^*})'\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{m}(\theta_{\epsilon^*})'\boldsymbol{m}(\theta), \sigma^2\|\boldsymbol{m}(\theta_{\epsilon^*})\|^2)$ under θ, the maximum regret of δ^* over Θ is given by

$$\begin{aligned}\max_{\theta\in\Theta}R(\delta^*,\theta) &= \max_{\theta\in\Theta}\left[(L(\theta))^+\Phi\left(-\frac{\boldsymbol{m}(\theta_{\epsilon^*})'\boldsymbol{m}(\theta)}{\sigma\|\boldsymbol{m}(\theta_{\epsilon^*})\|}\right) + (-L(\theta))^+\left(1-\Phi\left(-\frac{\boldsymbol{m}(\theta_{\epsilon^*})'\boldsymbol{m}(\theta)}{\sigma\|\boldsymbol{m}(\theta_{\epsilon^*})\|}\right)\right)\right] \\ &= \max_{\theta\in\Theta:L(\theta)>0}L(\theta)\Phi\left(-\frac{\boldsymbol{m}(\theta_{\epsilon^*})'\boldsymbol{m}(\theta)}{\sigma\|\boldsymbol{m}(\theta_{\epsilon^*})\|}\right),\end{aligned}$$

where $x^+ = \max\{x,0\}$ and the second equality holds by the symmetry of the objective function and the centrosymmetry of Θ.

The following lemma is fundamental to characterizing minimax regret rules for the original problem.

Lemma 1.4 (Worst Case for Nonrandomized Rule)**.** *Under the conditions in Theorem 1.2,*

$$\theta_{\epsilon^*} \in \arg \max_{\theta \in \Theta : L(\theta) > 0} L(\theta) \Phi \left(- \frac{\boldsymbol{m}(\theta_{\epsilon^*})' \boldsymbol{m}(\theta)}{\sigma \| \boldsymbol{m}(\theta_{\epsilon^*}) \|} \right).$$

Proof. See Appendix 1.B.8. □

By Lemma 1.4, the maximum regret of the decision rule δ^* over Θ is attained at θ_{ϵ^*}. Therefore,

$$\max_{\theta \in \Theta} R(\delta^*, \theta) = \max_{\theta \in [-\theta_{\epsilon^*}, \theta_{\epsilon^*}]} R(\delta^*, \theta) = \mathcal{R}(\sigma; [-\theta_{\epsilon^*}, \theta_{\epsilon^*}]),$$

where the last equality holds since δ^* is minimax regret for $[-\theta_{\epsilon^*}, \theta_{\epsilon^*}]$. However, by definition,

$$\max_{\theta \in \Theta} R(\delta^*, \theta) \geq \mathcal{R}(\sigma; \Theta) \geq \mathcal{R}(\sigma; [-\theta_{\epsilon^*}, \theta_{\epsilon^*}]).$$

It follows that $\max_{\theta \in \Theta} R(\delta^*, \theta) = \mathcal{R}(\sigma; \Theta) = \mathcal{R}(\sigma; [-\theta_{\epsilon^*}, \theta_{\epsilon^*}])$, and hence δ^* is minimax regret for Θ. The minimax risk for the original problem is the same as that for the hardest one-dimensional subproblem $[-\theta_{\epsilon^*}, \theta_{\epsilon^*}]$.

Lastly, Lemma 1.B.2 in Appendix 1.B.1 shows that $\boldsymbol{m}(\theta_{\epsilon^*})$ does not depend on the choice of θ_{ϵ^*} among those that attain the modulus of continuity at ϵ^*, which completes the proof of Theorem 1.2.

1.6.2 Randomized Rule

Consider the following assumption.

Assumption 1.3 (Regularity for Randomized Rule)**.** *The following holds for some $\bar{\epsilon} > 0$.*

(a) For all $\epsilon \in [0, \bar{\epsilon}]$, there exists $\theta_\epsilon \in \Theta$ that attains the modulus of continuity at ϵ with $\| \boldsymbol{m}(\theta_\epsilon) \| = \epsilon$.

(b) There exists $\boldsymbol{w}^* \in \mathbb{R}^n$ such that $\lim_{\epsilon \to 0} \epsilon^{-1} \left(\boldsymbol{w}^* - \frac{\boldsymbol{m}(\theta_\epsilon)}{\| \boldsymbol{m}(\theta_\epsilon) \|} \right) = \mathbf{0}$.

(c) There exists $\sigma^* \geq \sigma$ *such that* $0 \in \arg\max_{\epsilon \in \mathbb{R}} \rho(\epsilon)\Phi\left(-\epsilon/\sigma^*\right)$*, where* $\rho(\epsilon) \coloneqq \sup\{L(\theta) : (\boldsymbol{w}^*)'\boldsymbol{m}(\theta) = \epsilon, \theta \in \Theta\}$ *for* $\epsilon \in \mathbb{R}$.[31]

Assumption 1.3(a) is slightly stronger than Assumption 1.1(a) since it requires that the constraint $\|\boldsymbol{m}(\theta_\epsilon)\| \leq \epsilon$ of the modulus problem hold with equality. Assumption 1.3(b) is the same as Assumption 1.1(b).

Given σ^*, the maximization problem $\max_{\epsilon \in \mathbb{R}} \rho(\epsilon)\Phi\left(-\epsilon/\sigma^*\right)$ described in Assumption 1.3(c) corresponds to the problem of finding the worst-case parameter values for a randomized decision rule $\delta(\boldsymbol{Y}) = \Pr_{\xi \sim \mathcal{N}(0,(\sigma^*)^2-\sigma^2)}\left((\boldsymbol{w}^*)'\boldsymbol{Y} + \xi \geq 0\right)$. Later, I will show that, under Assumption 1.3(c), we can find the variance of the artificial noise ξ such that the maximum regret of δ is attained at a value of θ that attains the modulus of continuity at $\epsilon = 0$.

The following lemma shows that these conditions hold if $\sigma \leq 2\phi(0)\frac{\omega(0)}{\omega'(0)}$ under Assumption 1.1.

Lemma 1.5. *Let* Θ *be convex and centrosymmetric, and suppose that Assumption 1.1 holds. Then, Assumption 1.3(a) holds. Moreover,* $0 \in \arg\max_{\epsilon \in \mathbb{R}} \rho(\epsilon)\Phi\left(-\epsilon/\sigma^*\right)$ *with* $\sigma^* = 2\phi(0)\frac{\omega(0)}{\omega'(0)}$.

Proof. See Appendix 1.B.9. □

I obtain a minimax regret rule under Assumption 1.3. The statement on the randomized rule in Theorem 1.1 immediately follows from the following result.

Theorem 1.3 (Randomized Minimax Regret Rule)**.** *Let* Θ *be convex and centrosymmetric, and suppose that Assumption 1.3 holds. Then, the following decision rule is minimax regret:*

$$\delta^*(\boldsymbol{Y}) = \begin{cases} \mathbf{1}\{(\boldsymbol{w}^*)'\boldsymbol{Y} \geq 0\} & \text{if } \sigma^* = \sigma, \\ \Phi\left(\dfrac{(\boldsymbol{w}^*)'\boldsymbol{Y}}{((\sigma^*)^2 - \sigma^2)^{1/2}}\right) & \text{if } \sigma^* > \sigma. \end{cases}$$

The minimax risk is given by $\mathcal{R}(\sigma; \Theta) = \omega(0)/2$.

31. I allow the search space of σ^* to contain ∞, letting $\Phi(x/\infty) = 1/2$ for all $x \in \mathbb{R}$. Assumption 1.3(c) then holds with $\sigma^* = \infty$ in Stoye (2012)'s setup described in Section 1.3.5 when $b \geq 1$.

Note that if $\omega(\cdot)$ is differentiable at any $\epsilon \in (0, a^*\sigma]$ and $\sigma \leq 2\phi(0)\frac{\omega(0)}{\omega'(0)}$, $\epsilon^* = 0$ by Lemma 1.1, where $\epsilon^* \in \arg\max_{\epsilon\in[0,a^*\sigma]} \omega(\epsilon)\Phi(-\epsilon/\sigma)$. The minimax risk $\omega(0)/2$ can then be written as $\omega(\epsilon^*)\Phi(-\epsilon^*/\sigma)$, leading to the expression in Theorem 1.1.

Below, I provide the proof of Theorem 1.3. Note first that $\boldsymbol{w}^*$ is a unit vector by construction. Hence, $(\boldsymbol{w}^*)'\boldsymbol{Y} \sim \mathcal{N}\left((\boldsymbol{w}^*)'\boldsymbol{m}(\theta), \sigma^2\right)$. Simple calculations show that δ^* is equivalent to $\delta^*(\boldsymbol{Y}) = \Pr_{\xi\sim\mathcal{N}(0,(\sigma^*)^2-\sigma^2)}\left((\boldsymbol{w}^*)'\boldsymbol{Y} + \xi \geq 0\right)$. Since $(\boldsymbol{w}^*)'\boldsymbol{Y} + \xi \sim \mathcal{N}\left(0, (\sigma^*)^2\right)$ if $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)$, it follows that $\mathbb{E}_{\boldsymbol{Y}\sim\mathcal{N}(\boldsymbol{0},\sigma^2\boldsymbol{I}_n)}[\delta^*(\boldsymbol{Y})] = \frac{1}{2}$. The following lemma shows that δ^* is minimax regret for the one-dimensional subproblem $[-\theta_0, \theta_0]$, where θ_0 attains the modulus of continuity at 0 and hence $\boldsymbol{m}(\theta_0) = \boldsymbol{0}$.

Lemma 1.6 (Uninformative One-dimensional Subproblems)**.** *Suppose that $\Theta = [-\bar{\theta}, \bar{\theta}]$, where $\bar{\theta} \in \mathbb{V}$, $L(\bar{\theta}) \geq 0$, and $\boldsymbol{m}(\bar{\theta}) = \boldsymbol{0}$. Then, any decision rule δ^* such that $\mathbb{E}_{\boldsymbol{Y}\sim\mathcal{N}(\boldsymbol{0},\sigma^2\boldsymbol{I}_n)}[\delta^*(\boldsymbol{Y})] = \frac{1}{2}$ is minimax regret. The minimax risk is given by $\mathcal{R}(\sigma; [-\bar{\theta}, \bar{\theta}]) = L(\bar{\theta})/2$.*

Proof. See Appendix 1.B.10. □

If $\boldsymbol{m}(\bar{\theta}) = \boldsymbol{0}$, $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)$ under any $\theta \in [-\bar{\theta}, \bar{\theta}]$. Lemma 1.6 shows that choosing each policy with probability one half over the distribution of $\boldsymbol{Y}$ is minimax regret for the subproblem $[-\bar{\theta}, \bar{\theta}]$.

Since $(\boldsymbol{w}^*)'\boldsymbol{Y} + \xi \sim \mathcal{N}\left((\boldsymbol{w}^*)'\boldsymbol{m}(\theta), (\sigma^*)^2\right)$ under θ, the maximum regret of δ^* over Θ is given by

$$\begin{aligned}\max_{\theta\in\Theta} R(\delta^*, \theta) &= \max_{\theta\in\Theta}\left[(L(\theta))^+\Phi\left(-\frac{(\boldsymbol{w}^*)'\boldsymbol{m}(\theta)}{\sigma^*}\right) + (-L(\theta))^+\left(1 - \Phi\left(-\frac{(\boldsymbol{w}^*)'\boldsymbol{m}(\theta)}{\sigma^*}\right)\right)\right] \\ &= \max_{\theta\in\Theta} L(\theta)\Phi\left(-\frac{(\boldsymbol{w}^*)'\boldsymbol{m}(\theta)}{\sigma^*}\right),\end{aligned}$$

where the second equality holds by the symmetry of the objective function and the centrosymmetry of Θ. The following lemma shows that the maximum regret is attained at θ_0.

Lemma 1.7 (Worst Case for Randomized Rule)**.** *Under the conditions in Theorem 1.3,*

$$\theta_0 \in \arg\max_{\theta\in\Theta} L(\theta)\Phi\left(-\frac{(\boldsymbol{w}^*)'\boldsymbol{m}(\theta)}{\sigma^*}\right).$$

Proof. See Appendix 1.B.11. □

From the above results, we obtain

$$\max_{\theta\in\Theta} R(\delta^*,\theta) = \max_{\theta\in[-\theta_0,\theta_0]} R(\delta^*,\theta) = \mathcal{R}(\sigma;[-\theta_0,\theta_0]),$$

where the last equality holds since δ^* is minimax regret for $[-\theta_0,\theta_0]$. However, by definition,

$$\max_{\theta\in\Theta} R(\delta^*,\theta) \geq \mathcal{R}(\sigma;\Theta) \geq \mathcal{R}(\sigma;[-\theta_0,\theta_0]).$$

It follows that $\max_{\theta\in\Theta} R(\delta^*,\theta) = \mathcal{R}(\sigma;\Theta) = \mathcal{R}(\sigma;[-\theta_0,\theta_0])$, and hence δ^* is minimax regret for Θ. The minimax risk is given by $\mathcal{R}(\sigma;[-\theta_0,\theta_0]) = L(\theta_0)/2 = \omega(0)/2$.

1.7 Empirical Policy Application

I now illustrate my approach in an empirical application to the BRIGHT program in Burkina Faso. I consider the hypothetical problem of whether or not to expand the program and empirically compare the performance of the minimax regret rule with alternative decision rules.

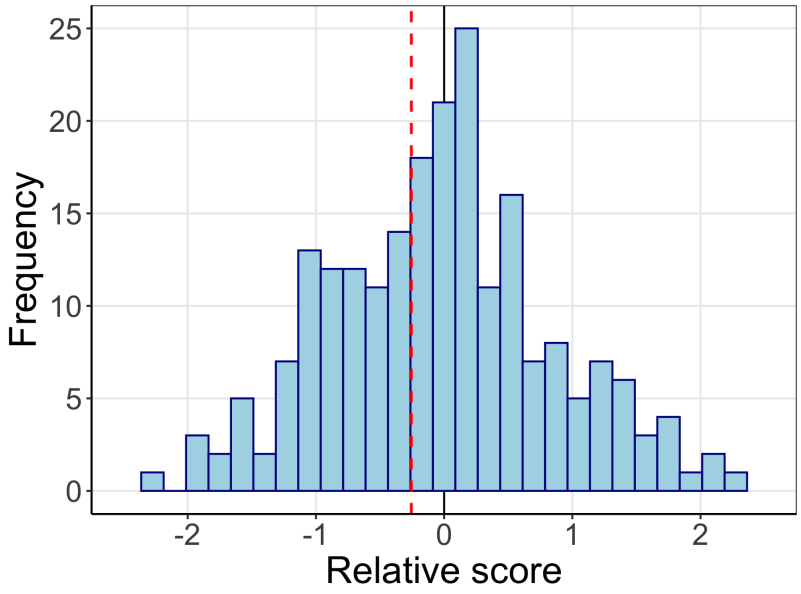
1.7.1 Background and Data

The goal of the BRIGHT program was to improve children's and especially girls' educational outcomes in rural villages by constructing well-resourced village-based schools. The program was funded by the Millennium Challenge Corporation, a U.S. government agency, and implemented by a consortium of non-governmental organizations. The program constructed primary schools with three classrooms for grades 1 to 3 in 132 villages from 47 departments during the period from 2005 to 2008. The Ministry of Education determined the villages where schools would be built through the following process.

1. 293 villages were nominated based on low school enrollment rates.

2. The Ministry administered a survey in each village and assigned each village a score using a set formula. The formula attached large weight to the estimated number of

Figure 1.1: Distribution of Relative Score



Notes: This figure shows the histogram of the relative score of villages on the interval $[-2.5, 2.5]$. The vertical dashed line indicates the new cutoff -0.256, which corresponds to the hypothetical policy of constructing schools in previously ineligible villages whose relative scores are in the top 20%. The villages with zero observed enrollment rates are excluded.

children to be served from the nominated and neighboring villages, giving additional weight to girls.

3. The Ministry ranked villages within each department and selected the top half of the villages to receive a school.

For further details on the BRIGHT program and allocation process, see Levy, Sloan, Linden and Kazianga (2009) and Kazianga *et al.* (2013).

Since the school allocation was determined at department level, the cutoff score for the program eligibility was different across departments. Following Kazianga *et al.* (2013), I define the *relative score* as the score for each village minus the cutoff score for the department that the village belongs to. As a result, a village is eligible for the program when the relative score is larger than zero. Kazianga *et al.* (2013) use the relative score as a running variable and evaluate the causal effect of the program on educational outcomes using a regression discontinuity design. Figure 1.1 reports the distribution of the relative score.

I use the replication data for Kazianga *et al.* (2013)'s results (Kazianga, Levy, Linden

Table 1.1: Child Educational Outcomes and Characteristics

	All	Eligible villages	Ineligible villages
	(1)	(2)	(3)
Panel A. Educational outcomes (child-level means)			
Enrollment	0.366	0.494	0.259
Normalized total test scores	0.000	0.248	−0.209
Highest grade child has achieved	0.876	1.132	0.636
Panel B. Child and household characteristics (child-level means)			
Child's age	8.121	8.174	8.071
Child is female	0.503	0.476	0.525
Head's age	47.653	47.387	47.904
Head years of schooling	0.156	0.198	0.117
Number of members	10.812	10.815	10.808
Number of children	5.971	6.098	5.850
Muslim	0.587	0.576	0.597
Basic roofing	0.516	0.534	0.500
Number of motorbikes	0.299	0.319	0.279
Number of phones	0.185	0.199	0.172
Total number of children	23,282	10,645	12,637
Total number of villages	287	136	151

Notes: This table reports child-level averages of educational outcomes and characteristics by program eligibility in the year 2008, namely 2.5 years after the start of the BRIGHT program. Panel A reports the educational outcomes' means. Panel B reports the means of child and household characteristics. Column (1) shows the means for children in all villages. Columns (2) and (3) show the means for children in villages selected for BRIGHT school and in unselected villages, respectively.

and Sloan, 2019) and consider whether we should expand the program or not. I explain the details of the counterfactual policy in Section 1.7.2. The dataset contains survey results about 30 households from 287 nominated villages, yielding a total sample of 23,282 children between the ages of 5 and 12. The survey was conducted in 2008, namely 2.5 years after the start of the program. Table 1.1 reports summary statistics about child educational outcomes and characteristics. Children in eligible villages are more likely to attend school, achieve higher test scores, and complete a higher grade. Household heads in eligible villages completed slightly more years of schooling. Furthermore, households in eligible villages tend to have more assets such as basic roofing and motorbikes.

I consider school enrollment as the target outcome. Since the score and program eligibility are determined at village level, I use the village-level mean outcome, namely the enrollment rate for each village. This setting fits into the setup in Section 1.2.3, where i represents a village, Y_i is the observed enrollment rate of village i, d_i is the program eligi-

bility, and x_i is the relative score. The original cutoff is $c_0 = 0$, that is, $d_i = \mathbf{1}\{x_i \geq 0\}$. The parameter is a function $f : \mathbb{R} \times \{0, 1\} \to \mathbb{R}$, where $f(x, d)$ represents the counterfactual mean of the enrollment rate conditional on the relative score if the eligibility status were set to $d \in \{0, 1\}$. Since Y_i is a village-level sample mean, it is plausible to assume that Y_i is approximately normally distributed. I use the conventional standard error of the sample mean Y_i as the standard deviation of Y_i.[32]

1.7.2 Hypothetical Policy Choice Problem

I ask whether we should scale up the program and build BRIGHT schools in other villages. Specifically, I consider the following decision problem. The counterfactual policy is to build BRIGHT schools in previously ineligible villages whose relative scores are in the top 20%, which corresponds to lowering the cutoff from 0 to -0.256.[33] I use the average enrollment rate across villages as the welfare criterion, so that the welfare effect of this policy relative to the status quo is

$$L(f) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{-0.256 \leq x_i < 0\}[f(x_i, 1) - f(x_i, 0)].$$

When deciding whether to implement the policy, it is important to consider the benefit relative to the cost. Kazianga *et al.* (2013) provide an estimate of the cost of constructing a BRIGHT school, which is \$4,758 per village.[34] To incorporate the cost into the decision problem, I suppose that the policy maker cares about the cost-effectiveness of this new policy relative to similar programs. Cost-effectiveness is defined as the ratio of the policy cost to the increase in the target outcome, namely the enrollment in the current context. I assume that it is optimal to implement the policy if its cost-effectiveness is smaller than \$83.77,

32. The observed enrollment rate is zero in 21 out of 287 villages. I exclude these villages from the analysis since the standard error of Y_i is zero.

33. Figure 1.11 in Appendix 1.C reports the results when I use 10% and 30% instead of 20%. As predicted by the result in Section 1.4, the minimax regret rule switches from a nonrandomized rule to a randomized rule at a smaller Lipschitz constant C when the fraction of the target villages is larger.

34. I assume that the cost estimate is a known quantity and is constant across villages. It is, however, natural to think of the policy cost as unknown and heterogeneous across villages and to introduce the cost model on top of the outcome model. I leave this for future work.

which is the cost-effectiveness of a school construction program in Indonesia (Duflo, 2001; Kazianga *et al.*, 2013). Specifically, it is optimal to implement the policy if

$$\frac{\$4{,}758}{416 \cdot \frac{1}{\tilde{n}} \sum_{i=1}^{n} \mathbf{1}\{-0.256 \leq x_i < 0\}[f(x_i, 1) - f(x_i, 0)]} \leq \$83.77,$$

where 416 is the number of children per village and $\tilde{n} = \sum_{i=1}^{n} \mathbf{1}\{-0.256 \leq x_i < 0\}$ is the number of villages that would receive a school under the new policy.[35] The denominator represents the increase in the average enrollment across villages that would receive a BRIGHT school under the new policy.

Simple calculations show that the above condition is equivalent to

$$\frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{-0.256 \leq x_i < 0\}[f(x_i, 1) - 0.137 - f(x_i, 0)] \geq 0.$$

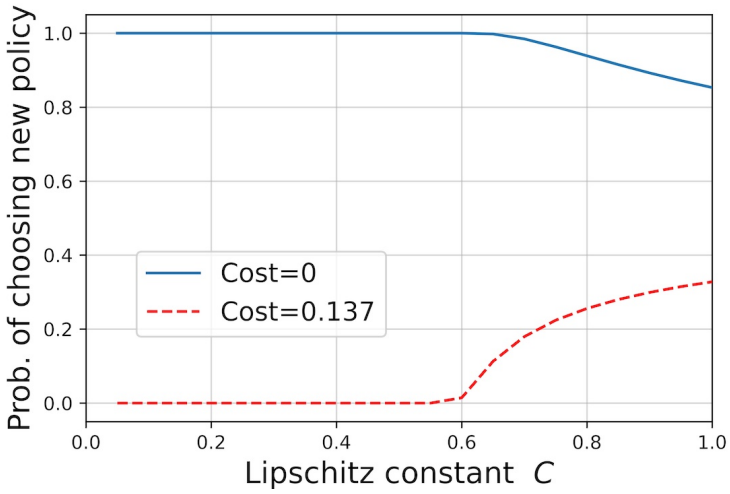
My method can be used to consider this decision problem by setting the outcome to $Y_i - 0.137 d_i$, where 0.137 can be viewed as the policy cost measured in the unit of the enrollment rate. I present the results for this scenario with the cost of 0.137 as well as for the scenario where we ignore the policy cost.

I implement my method assuming that the counterfactual outcome function f belongs to the Lipschitz class $\mathcal{F}_{\text{Lip}}(C)$. Since the relative score x_i is computed based on several village-level characteristics, it is difficult to interpret and specify the Lipschitz constant C using domain-specific knowledge. To obtain a reasonable range of C, I estimate a lower bound on C using the method described in Section 1.4.1, which yields the lower bound estimate of 0.149.[36] I present the results for $C \in \{0.05, 0.1, ..., 0.95, 1\}$ and examine their sensitivity to the choice of C.

35. The cost per village and the cost-effectiveness of a school construction program in Indonesia are found in Tables A18 and A20, respectively, in Online Appendix of Kazianga *et al.* (2013). I compute the number of children per village by dividing the total enrollment by the enrollment rate reported in Table A17 in Online Appendix of Kazianga *et al.* (2013).

36. I estimate $\frac{\partial f(x,0)}{\partial x}$ at $x \in \{-2.5, -2.45, ..., -0.05\}$ and $\frac{\partial f(x,1)}{\partial x}$ at $x \in \{0.05, 0.1, ..., 2.5\}$ by local quadratic regression and take the maximum of their absolute values. For local quadratic regression, I use the MSE-optimal bandwidth selection procedure by Calonico, Cattaneo and Farrell (2018), which can be implemented by R package "nprobust."

Figure 1.2: Optimal Decisions: Probability of Choosing the New Policy



Notes: This figure shows the probability of choosing the new policy computed by the minimax regret rule. The new policy is to construct BRIGHT schools in previously ineligible villages whose relative scores are in the top 20%. The solid line shows the results for the scenario where we ignore the policy cost. The dashed line shows the results for the scenario where the policy cost measured in the unit of the enrollment rate is 0.137. I report the results for the range $[0.05, 0.1, ..., 0.95, 1]$ of the Lipschitz constant C.

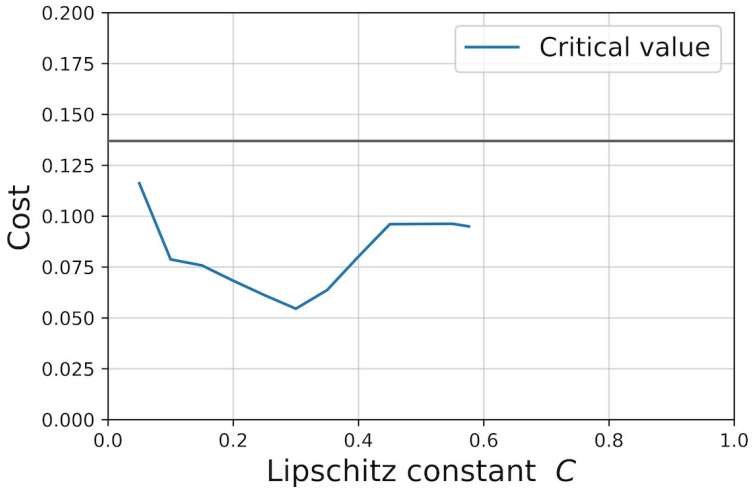
1.7.3 Results

Figure 1.2 plots $\delta^*(\boldsymbol{Y})$, the probability of choosing the new policy computed by the minimax regret rule, against the Lipschitz constant C. When $C < 0.6$, the minimax regret rule is nonrandomized. It chooses the new policy in the no-cost scenario and maintains the status quo in the scenario where the policy cost is 0.137. When $C \geq 0.6$, on the other hand, the minimax regret rule is randomized. The decisions become more mixed as C increases.

Given that the estimate of the lower bound on C is 0.149, the minimax regret rule is nonrandomized when C is less than four times the estimated lower bound. Under this reasonable range of C, the optimal decision is the same in each scenario. If the policy maker wants to be more conservative about the choice of C, they need to randomize their decisions.

The above analysis considers the scenario where the policy cost is fixed at 0.137. To examine the sensitivity of the result to the policy cost, I compute the maximum of the cost values under which the minimax regret rule chooses the new policy with probability one, reported in Figure 1.3. If the policy cost is less than this value, it is optimal to choose the new policy; otherwise, it is optimal to maintain the status quo. The result shows that, when

Figure 1.3: Maximum of Cost Values Under Which Choosing the New Policy is Optimal

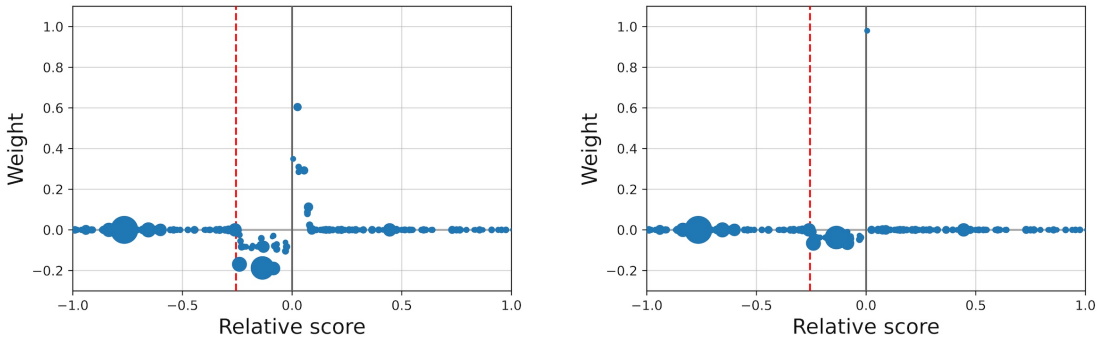


Notes: This figure shows the maximum of the cost values under which the minimax regret rule chooses the new policy with probability one. The horizontal line shows the cost of 0.137, which is my main specification of the policy cost. I only report the results for the Lipschitz constant $C < 0.6$ since the minimax regret rule is randomized for $C \geq 0.6$.

C is above its lower bound 0.149, it is optimal to maintain the status quo as long as the policy cost is higher than 0.10.

If the minimax regret rule is nonrandomized, the rule is of the form $\delta^*(\boldsymbol{Y}) = \mathbf{1}\{\sum_{i=1}^n w_i Y_i \geq 0\}$ for some weights w_i's. Panels (a) and (b) of Figure 1.4 plot the weight w_i attached to each village against the relative score x_i for $C = 0.1$ and $C = 0.5$, respectively. In the plots, the size of circles is proportional to the inverse of the standard error of the enrollment rate Y_i. For both $C = 0.1$ and $C = 0.5$, a few treated units just above the original cutoff (the solid vertical line) receive a positive weight, the untreated units between the original cutoff and the new cutoff (the dashed vertical line) receive a negative weight, and no other units receive any weight. When $C = 0.1$, the weight tends to be larger for units with a smaller standard error. When $C = 0.5$, a positive weight is attached only to the treated unit closest to the original cutoff. Additionally, the weights on the untreated units between the two cutoffs are almost identical. This situation corresponds to the minimax regret rule of the form $\delta^*(\boldsymbol{Y}) = \mathbf{1}\left\{Y_{+,\min} - \frac{1}{\bar{n}} \sum_{i:c_1 \leq x_i < c_0} Y_i \geq 0\right\}$ discussed in Section 1.4.

Figure 1.4: Weight to Each Village Attached by Minimax Regret Rule



(a) $C = 0.1$ (b) $C = 0.5$

Notes: This figure shows the weight w_i attached to each village by the minimax regret rule of the form $\delta^*(\boldsymbol{Y}) = \mathbf{1}\{\sum_{i=1}^n w_i Y_i \geq 0\}$. The weights are normalized so that $\sum_{i=1}^n w_i^2 = 1$. The horizontal axis indicates the relative score of each village. Each circle corresponds to each village. The size of circles is proportional to the inverse of the standard error of the enrollment rate Y_i. The vertical dashed line corresponds to the new cutoff -0.256. Panels (a) and (b) show the results when the Lipschitz constant C is 0.1 and 0.5, respectively.

1.7.3.1 Comparison with Plug-in Rules

I compare the minimax regret rule with plug-in decision rules that make a decision according to the sign of an estimator of the policy effect. I consider three estimators of the policy effect.

1. The linear minimax MSE estimator (Donoho, 1994), described in Section 1.5.1, under the Lipschitz class $\mathcal{F}_{\text{Lip}}(C)$.

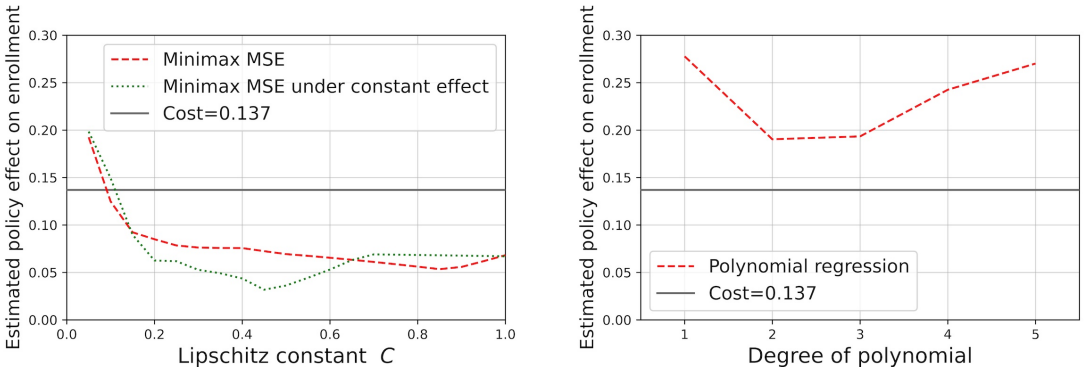
2. The linear minimax MSE estimator under the additional assumption of constant conditional treatment effects. In other words, I construct the estimator assuming that $\mathcal{F} = \{f \in \mathcal{F}_{\text{Lip}}(C) : f(x,1) - f(x,0) = f(\tilde{x},1) - f(\tilde{x},0) \text{ for all } x, \tilde{x}\}$. This estimation corresponds to first nonparametrically estimating the average treatment effect at the original cutoff and then extrapolating the effects on the units between the two cutoffs by the constant effects assumption.

3. The polynomial regression estimator (Kazianga *et al.*, 2013).[37] Given the degree of polynomial p, I first estimate the model $f(x,d) = \alpha_0 + \alpha_1 x + \cdots + \alpha_p x^p + \beta_0 d + \beta_1 d \cdot x + \cdots \beta_p d \cdot x^p$ by the weighted least squares regression using $1/\sigma^2(x_i, d_i)$ as the weight.[38]

37. Kazianga *et al.* (2013) estimate the treatment effect at the cutoff, not the effect on the units away from the cutoff. They apply global polynomial regression RD estimators to child-level data.

38. This is equivalent to the OLS regression of $Y_i/\sigma(x_i,d_i)$ on $(1, x, ..., x^p, d, d \cdot x, ..., d \cdot x^p)'/\sigma(x_i,d_i)$.

Figure 1.5: Estimated Effects of New Policy on Enrollment Rate



(a) Minimax MSE Estimator Under Lipschitz Class

(b) Polynomial Regression Estimator

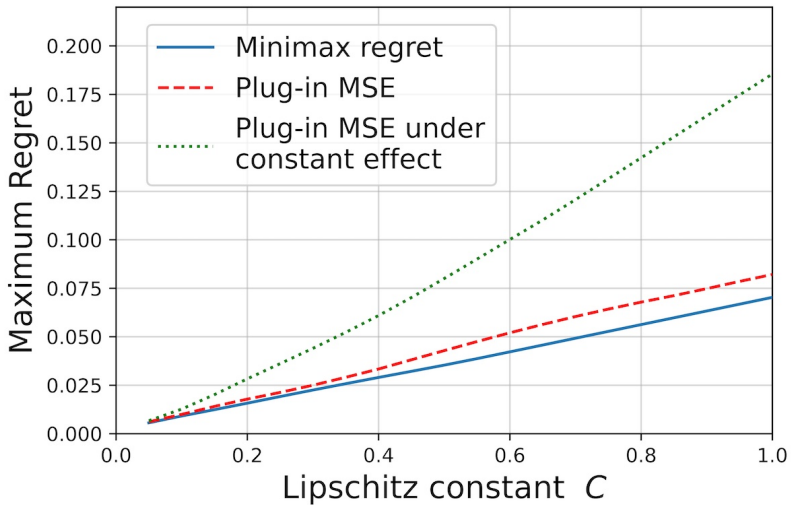
Notes: This figure shows the average effect of the new policy on the enrollment rate across the villages that would receive a school under the new policy. Panel (a) reports the estimates from the linear minimax MSE estimators with and without the assumption of constant conditional treatment effects. I report the results for the range $[0.05, 0.1, ..., 0.95, 1]$ of the Lipschitz constant C. Panel (b) reports the estimates from the polynomial regression estimators of degrees 1 to 5. The horizontal line shows the cost of 0.137, which is my main specification of the policy cost.

I then estimate $L(f)$ by $\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{-0.256 \leq x_i < 0\}[\hat{f}(x_i, 1) - \hat{f}(x_i, 0)]$, where $\hat{f}$ is the estimated polynomial function. This estimator relies on the functional form of f to extrapolate $f(x_i, 1)$ for the untreated units.

Panel (a) of Figure 1.5 reports the estimated policy effects from the linear minimax MSE estimators with and without constant conditional treatment effects. Overall, these two estimators exhibit a similar pattern. While the estimated policy effects are larger than the policy cost when C is close to zero, they are smaller than the policy cost when C is moderate or large. For $C \geq 0.2$, the resulting decisions about whether to choose the new policy are the same as the decision made by the minimax regret rule until C reaches 0.6, where the minimax regret rule starts to randomize. In contrast, the estimated policy effects from the polynomial regression estimators of degrees 1 to 5 exceed the policy cost, as reported in Panel (b) of Figure 1.5. The estimates appear to be close to the simple mean outcome difference between eligible and ineligible villages that can be computed from Table 1.1. The resulting decisions are different from the decision made by the minimax regret rule.[39]

39. The estimators presented here can be written as $\sum_{i=1}^{n} w_i Y_i$ for some weights w_i's. See Figure 1.8 in Appendix 1.C for the plots of these weights. While the linear minimax MSE estimators attach weights to

Figure 1.6: Maximum Regret of Minimax Regret Rule and Plug-in MSE Rules



Notes: This figure shows the maximum regret of the minimax regret rule and the plug-in rules based on the linear minimax MSE estimators with and without the assumption of constant conditional treatment effects. The maximum regret is normalized so that the unit is the same as that of the enrollment rate. I report the results for the range $[0.05, 0.1, ..., 0.95, 1]$ of the Lipschitz constant C.

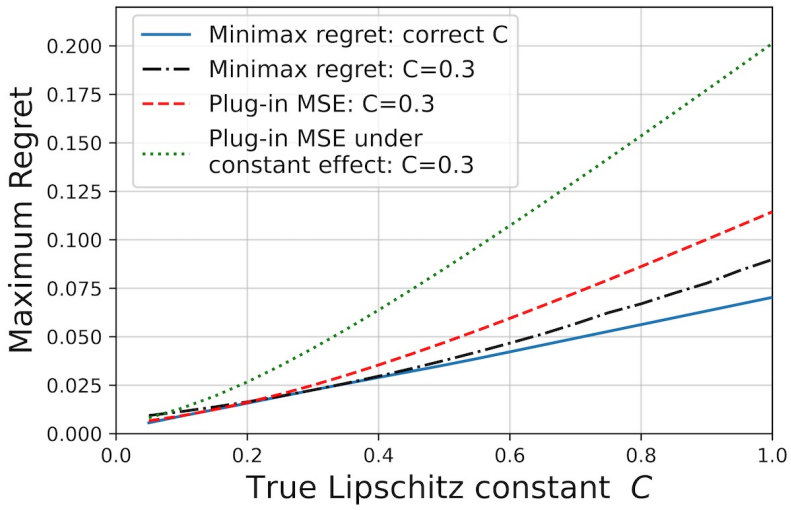
The above estimates and resulting decisions are computed from a particular realization of the sample. To assess the ex ante performance of different decision rules, I compute the maximum regret of these rules when the true function class is $\mathcal{F}_{\text{Lip}}(C)$.[40] Figure 1.6 reports the result for the minimax regret rule and the plug-in rules based on the linear minimax MSE estimators with and without constant conditional treatment effects.[41] The maximum regret of the plug-in MSE rule with constant conditional treatment effects is much larger than that of the other two, especially when the Lipschitz constant C is large. The plug-in MSE rule without constant conditional treatment effects performs worse than the minimax regret rule, as predicted by the theoretical analysis. The ratio of the maximum regret between the two rules is maximized at $C = 0.6$, where the minimax regret rule starts to randomize.

units just above the original cutoff and to units between the two cutoffs, polynomial regression estimators attach weights even to units further away from the cutoffs.

40. I compute the maximum regret of the minimax regret rule using the formula in Theorem 1.1. For the other rules, I adapt the approach by Ishihara and Kitagawa (2021) to numerically calculate the maximum regret in this setup.

41. See Figure 1.9 in Appendix 1.C for the result for the plug-in rules based on polynomial regression estimators. The maximum regret of these rules is significantly larger than that of alternative rules.

Figure 1.7: Maximum Regret Under Misspecification of Lipschitz Constant C



Notes: This figure shows the maximum regret of the minimax regret rule and the plug-in rules based on the linear minimax MSE estimators with and without the assumption of constant conditional treatment effects, which are constructed assuming that the Lipschitz constant C is 0.3. The maximum regret is computed by setting the true C to the value on the horizontal axis. The solid line indicates the maximum regret that can be achieved if C is correctly specified. The maximum regret is normalized so that the unit is the same as that of the enrollment rate. I report the results for the range $[0.05, 0.1, ..., 0.95, 1]$ of the true Lipschitz constant C.

1.7.3.2 Sensitivity to Misspecification of Lipschitz Constant C

So far, I have constructed decision rules assuming that the Lipschitz constant C is known, which is a crucial assumption in my theoretical analysis. To assess the sensitivity of the performance to misspecification of C, I construct decision rules assuming $C = 0.3$ and then compute their maximum regret when the true value of C lies in $\{0.05, 0.1, ..., 0.95, 1\}$.

Figure 1.7 reports the result. The solid line indicates the "oracle" maximum regret, which can be achieved if we correctly specify C. The result shows that the plug-in MSE rule without constant conditional treatment effects performs slightly better than the minimax regret rule when the true C is close to zero. On the other hand, the minimax regret rule outperforms the plug-in MSE rule with nonnegligible differences for any value of the true C greater than 0.2. The result suggests that the minimax regret rule is more robust to misspecification of C toward zero than the plug-in MSE rule.

The potential superiority of the minimax regret rule seems consistent with the theoretical results in the following way. As shown in Section 1.4, when the true value of C is large, the oracle minimax regret rule only uses the treated units just above the original cutoff and

the untreated units between the original and new cutoffs (see Panel (b) of Figure 1.4). If the specified C is smaller than the true value, the resulting minimax regret rule is closer to the oracle rule than the plug-in MSE rule since the minimax regret rule places more importance on the bias than the plug-in MSE rule as discussed in Section 1.5.1. Therefore, it is expected that the minimax regret rule performs better than the plug-in MSE rule under misspecification of C toward zero.

1.8 Conclusion and Future Directions

This chapter develops an optimal procedure for using data to make policy decisions in settings where social welfare under each counterfactual policy is only partially identified. I derive a decision rule that achieves the minimax regret optimality in finite samples and within the class of all decision rules. I apply the result to the problem of eligibility cutoff choice and illustrate it in an empirical application to a school construction program in Burkina Faso.

While my application focuses on eligibility rules based on a scalar variable, it is possible to apply my approach to a choice of treatment assignment policy based on multiple covariates. My method can also be applied to the problem of deciding whether to introduce a new policy using data from a randomized experiment when the experiment has imperfect compliance or when the experimental sample is a selected subset of the target population. I plan to apply my general result to these scenarios and provide an empirical illustration.

Several extensions of my work are possible. First, my result relies on the assumption that the sample is normally distributed with a known variance. It is challenging but natural to consider the asymptotic optimality without the distributional assumption, for example, by extending the limits of experiments framework of Hirano and Porter (2009) to setups with partial identification and restricted parameter spaces. Second, my approach only covers a binary choice problem. It is both theoretically and practically important to extend the analysis to a multiple or continuous policy space. Lastly, while this work focuses on one-shot decision making, it may in practice be possible to make a policy change again after observing the result of a previous policy choice for a certain period of time. It would be interesting to

consider such sequential decision problems.

Appendices

1.A Additional Results and Details

1.A.1 Example: Optimal Treatment Assignment Policy Under Unconfoundedness

The basic setup is the same as the one in Section 1.2.3. I generalize it in three ways. First, the covariates x_i are k dimensional, where $k \geq 1$. Second, I remove the assumption that $d_i = \mathbf{1}\{x_i \geq 0\}$ and instead assumes the unconfoundedness (i.e., the observed treatment is independent of potential outcomes conditional on covariates). These first two do not change the notation of the data-generating process:

$$\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{m}(f), \boldsymbol{\Sigma}),$$

where $\boldsymbol{Y} = (Y_1, ..., Y_n)'$, $\boldsymbol{m}(f) = (f(x_1, d_1), ..., f(x_n, d_n))'$, and $\boldsymbol{\Sigma} = \text{diag}(\sigma^2(x_1, d_1), ..., \sigma^2(x_n, d_n))$. Under unconfoundedness, we can interpret $f(x, d)$ as the counterfactual mean outcome for those with covariates x under treatment status d (see footnote 12).

Third, two alternative policies are functions $\pi_a : \mathbb{R}^k \to [0, 1]$, $a \in \{0, 1\}$, where $\pi_a(x)$ is the probability of assigning treatment to individuals whose covariates are $x \in \mathbb{R}^k$. Suppose that the welfare under policy a, $a \in \{0, 1\}$, is an average of the counterfactual mean outcome across different values of covariates

$$W_a(f) = \int [f(x, 1)\pi_a(x) + f(x, 0)(1 - \pi_a(x))] d\nu(x)$$

for some known measure ν. The welfare difference between the two policies is

$$L(f) = W_1(f) - W_0(f) = \int (\pi_1(x) - \pi_0(x))(f(x, 1) - f(x, 0)) d\nu(x).$$

One example of the function class $\mathcal{F}$ is the Lipschitz class with a known Lipschitz constant $C \geq 0$:

$$\mathcal{F}_{\text{Lip}}(C) = \{f : |f(x,d) - f(\tilde{x},d)| \leq C\|x - \tilde{x}\| \text{ for every } x, \tilde{x} \in \mathbb{R}^k \text{ and } d \in \{0,1\}\}.$$

1.A.2 Comparison with Hypothesis Testing Rules

Hypothesis testing can be viewed as an alternative procedure for deciding between two policies. Here, I compare the minimax regret rule with a class of hypothesis testing rules.

To define it, suppose Θ is convex and centrosymmetric, and consider testing

$$H_0 : L(\theta) \leq -b \text{ and } \theta \in \Theta \quad \text{vs.} \quad H_1 : L(\theta) \geq b \text{ and } \theta \in \Theta$$

for some $b > 0$. Let $\theta^{(b)}$ solve $\inf_{\theta \in \Theta : L(\theta) \geq b} \|\boldsymbol{m}(\theta)\|$. For any level $\alpha > 0$, the minimax test, which has the largest minimum power under H_1, is given by the Neyman-Pearson test of $H_0 : \theta = -\theta^{(b)}$ vs. $H_1 : \theta = \theta^{(b)}$ (Armstrong and Kolesár, 2018, Lemma A.2). It rejects H_0 if the test statistic $\boldsymbol{m}(\theta^{(b)})'\boldsymbol{Y}$ is greater than its $1-\alpha$ quantile under $-\theta^{(b)}$. Since $\boldsymbol{m}(\theta^{(b)})'\boldsymbol{Y} \sim \mathcal{N}(-\|\boldsymbol{m}(\theta^{(b)})\|^2, \sigma^2\|\boldsymbol{m}(\theta^{(b)})\|^2)$ under $-\theta^{(b)}$, the critical value is $-\|\boldsymbol{m}(\theta^{(b)})\|^2 + z_{1-\alpha}\sigma\|\boldsymbol{m}(\theta^{(b)})\|$, where $z_{1-\alpha}$ is the $1-\alpha$ quantile of a standard normal variable. The level-α minimax test is then given by

$$\delta_{\alpha,b}(\boldsymbol{Y}) = \mathbf{1}\{\boldsymbol{m}(\theta^{(b)})'\boldsymbol{Y} \geq -\|\boldsymbol{m}(\theta^{(b)})\|^2 + z_{1-\alpha}\sigma\|\boldsymbol{m}(\theta^{(b)})\|\}.$$

I call such tests *hypothesis testing rules*.

Are there any hypothesis testing rules that exactly match the minimax regret rule? Let $\epsilon^* > 0$ solve $\max_{\epsilon \in [0, a^*\sigma]} \omega(\epsilon)\Phi(-\epsilon/\sigma)$, and let θ_{ϵ^*} solve the modulus of continuity at ϵ^* with $\|\boldsymbol{m}(\theta_{\epsilon^*})\| = \epsilon^*$. By the duality of the problem, θ_{ϵ^*} also solves $\inf_{\theta \in \Theta : L(\theta) \geq b^*} \|\boldsymbol{m}(\theta)\|$, where $b^* = \omega(\epsilon^*)$. Let α^* satisfy $-\|\boldsymbol{m}(\theta_{\epsilon^*})\| + z_{1-\alpha^*}\sigma = 0$, i.e., $\alpha^* = \Phi(-\epsilon^*/\sigma)$, so that the critical value is zero. For this choice of α^* and b^*, the hypothesis testing rule is

$$\delta_{\alpha^*,b^*}(\boldsymbol{Y}) = \mathbf{1}\{\boldsymbol{m}(\theta_{\epsilon^*})'\boldsymbol{Y} \geq 0\},$$

which is identical to the minimax regret rule. Since $\epsilon^* \leq a^*\sigma$, we can obtain a lower bound on α^*: $\alpha^* = \Phi(-\epsilon^*/\sigma) \geq \Phi(-a^*) \approx 0.226$. Therefore, the minimax regret rule is less conservative in rejection of the null hypothesis than hypothesis testing rules that use conventional levels such as 0.01 and 0.05. This is consistent with the fact that the minimax regret criterion takes into consideration the potential welfare loss as well as the probability of making a wrong choice.

1.A.3 Sufficient Conditions for Differentiability of $\omega(\cdot)$ and $\rho(\cdot)$

The result below follows from Lemma D.1 in Supplemental Appendix D of Armstrong and Kolesár (2018) in the case where $\mathcal{F} = \mathcal{G}$ in their notation. Note that their definition of the modulus of continuity when $\mathcal{F} = \mathcal{G}$ is the same as Donoho (1994)'s definition, which is different from my definition. See Appendix 1.A.5 for the relationship between their definition and mine.

Lemma 1.A.1. *Let Θ be convex. Let θ_ϵ attain the modulus of continuity at $\epsilon > 0$ with $\|\boldsymbol{m}(\theta_\epsilon)\| = \epsilon$, and suppose that there exists $\iota \in \Theta$ such that $L(\iota) = 1$ and $\theta_\epsilon + c\iota \in \Theta$ for all c in a neighborhood of zero. Then, $\omega(\cdot)$ is differentiable at ϵ with $\omega'(\epsilon) = \frac{\epsilon}{\boldsymbol{m}(\iota)'\boldsymbol{m}(\theta_\epsilon)}$.*

The result below follows from arguments similar to the proof of Lemma D.1 in Armstrong and Kolesár (2018).

Lemma 1.A.2. *Let Θ be convex. Let θ_ϵ satisfy $L(\theta_\epsilon) = \rho(\epsilon)$ and $(\boldsymbol{w}^*)'\boldsymbol{m}(\theta_\epsilon) = \epsilon$, and suppose that there exists $\iota \in \Theta$ such that $L(\iota) = 1$ and $\theta_\epsilon + c\iota \in \Theta$ for all c in a neighborhood of zero. Then, $\rho(\cdot)$ is differentiable at ϵ with $\rho'(\epsilon) = \frac{1}{(\boldsymbol{w}^*)'\boldsymbol{m}(\iota)}$.*

Proof. First, I show that $\rho(\cdot)$ is concave on (ϵ_1, ϵ_2), where $\epsilon_1 = \inf\{(\boldsymbol{w}^*)'\boldsymbol{m}(\theta) : \theta \in \Theta\}$ and $\epsilon_2 = \sup\{(\boldsymbol{w}^*)'\boldsymbol{m}(\theta) : \theta \in \Theta\}$. Pick any $\epsilon, \epsilon' \in (\epsilon_1, \epsilon_2)$, and let $\{\theta_{\epsilon,n}\}_{n=1}^\infty$ and $\{\theta_{\epsilon',n}\}_{n=1}^\infty$ be sequences in Θ such that $(\boldsymbol{w}^*)'\boldsymbol{m}(\theta_{\epsilon,n}) = \epsilon$ and $(\boldsymbol{w}^*)'\boldsymbol{m}(\theta_{\epsilon',n}) = \epsilon'$ for all $n \geq 1$ and that $\lim_{n\to\infty} L(\theta_{\epsilon,n}) = \rho(\epsilon)$ and $\lim_{n\to\infty} L(\theta_{\epsilon',n}) = \rho(\epsilon')$. Then, for each $\lambda \in [0,1]$, $\lambda\theta_{\epsilon,n} + (1-\lambda)\theta_{\epsilon',n} \in \Theta$ by the convexity of Θ, and $(\boldsymbol{w}^*)'\boldsymbol{m}(\lambda\theta_{\epsilon,n} + (1-\lambda)\theta_{\epsilon',n}) = \lambda\epsilon + (1-\lambda)\epsilon'$ so that

$$\rho(\lambda\epsilon + (1-\lambda)\epsilon') \geq L(\lambda\theta_{\epsilon,n} + (1-\lambda)\theta_{\epsilon',n})$$

by the definition of ρ. Taking the limit of the right-hand side as $n \to \infty$ gives

$$\rho(\lambda\epsilon + (1-\lambda)\epsilon') \geq \lambda\rho(\epsilon) + (1-\lambda)\rho(\epsilon').$$

Therefore, $\rho(\cdot)$ is concave.

Since $\rho(\cdot)$ is concave, the superdifferential of $\rho(\cdot)$ at ϵ,

$$\partial\rho(\epsilon) = \{d : \rho(\eta) \leq \rho(\epsilon) + d(\eta - \epsilon) \text{ for all } \eta \in \mathbb{R}\},$$

is nonempty for all $\epsilon \in (\epsilon_1, \epsilon_2)$.

Now, let θ_ϵ satisfy $L(\theta_\epsilon) = \rho(\epsilon)$ and $(\boldsymbol{w}^*)'\boldsymbol{m}(\theta_\epsilon) = \epsilon$ for some ϵ, and suppose that there exists $\iota \in \Theta$ such that $L(\iota) = 1$ and $\theta_\epsilon + c\iota \in \Theta$ for all c in a neighborhood of zero. Then, for any $d \in \partial\rho(\epsilon)$ and for any c in a neighborhood of zero such that $\theta_\epsilon + c\iota \in \Theta$,

$$\rho(\epsilon) + d[(\boldsymbol{w}^*)'\boldsymbol{m}(\theta_\epsilon + c\iota) - \epsilon] \geq \rho((\boldsymbol{w}^*)'\boldsymbol{m}(\theta_\epsilon + c\iota)) \geq L(\theta_\epsilon + c\iota) = L(\theta_\epsilon) + c = \rho(\epsilon) + c,$$

where the first inequality follows since $d \in \partial\rho(\epsilon)$, and the second inequality follows from the definition of ρ. Since $(\boldsymbol{w}^*)'\boldsymbol{m}(\theta_\epsilon + c\iota) = \epsilon + c(\boldsymbol{w}^*)'\boldsymbol{m}(\iota)$, it follows that $cd(\boldsymbol{w}^*)'\boldsymbol{m}(\iota) \geq c$ for all c in a neighborhood of zero. This implies that $d(\boldsymbol{w}^*)'\boldsymbol{m}(\iota) = 1$. The result then follows. □

1.A.4 Differentiability of $\omega(\cdot)$ and $\rho(\cdot)$ for Example in Section 1.4

I apply Lemma 1.A.1 to show the differentiability of $\omega(\cdot)$. Consider the problem (1.4). There exists a solution to this problem for any $\epsilon > 0$, since the objective is continuous, and the set of the vectors of $2n$ unknowns that satisfy the constraints is closed and bounded. The norm constraint must hold with equality, for otherwise we can increase the objective by increasing $f(x_i, 1)$ for all i by a small amount. The differentiability of $\omega(\cdot)$ then follows from Lemma 1.A.1.

I show the differentiability of $\rho(\cdot)$ at any ϵ by deriving its closed-form expression. Observe

that $\rho(\epsilon)$ is obtained by solving

$$\max_{(f(x_i,0),f(x_i,1))_{i=1,\ldots,n}\in\mathbb{R}^{2n}} \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{c_1 \le x_i < c_0\}[f(x_i,1)-f(x_i,0)] \tag{A.1}$$
$$s.t.\ \ \sum_{i=1}^{n} w_i^* \frac{f(x_i,d_i)}{\sigma(x_i,d_i)} = \epsilon,\ \ f(x_i,d)-f(x_j,d) \le C|x_i-x_j|,\ \ \ d\in\{0,1\}, i,j\in\{1,...,n\},$$

where w_i^* is the ith element of $\boldsymbol{w}^*$ given in Section 1.4. By the same logic explained in Section 1.4 for the problem for $\omega(\epsilon)$, it is sufficient to check the Lipschitz constraint among x's in the sample.

The constraint $\sum_{i=1}^{n} w_i^* \frac{f(x_i,d_i)}{\sigma(x_i,d_i)} = \epsilon$ is equivalent to

$$-\sum_{i:c_1\le x_i<c_0} f(x_i,0) + \tilde{n} f(x_{+,\min},1) = \bar{\sigma}\epsilon, \tag{A.2}$$

where $\bar{\sigma} = (\tilde{n}^2\sigma_{+,\min}^2 + \sum_{i:c_1\le x_i<c_0}\sigma^2(x_i,0))^{1/2}$. The objective of (A.1) and the constraint (A.2) depend only on $((f(x_i,0),f(x_i,1))_{i:c_1\le x_i<0}, f(x_{+,\min},1))$, so we can set the other values arbitrarily as long as the Lipschitz constraint holds. Also, given a value of $f(x_{+,\min},1)$, the objective of (A.1) is maximized only when $f(x_i,1) = C(x_{+,\min}-x_i) + f(x_{+,\min},1)$ for all i with $c_1 \le x_i < c_0$ under the Lipschitz constraint. Therefore, the problem (A.1) reduces to

$$\max_{((f(x_i,0))_{i:c_1\le x_i<c_0}, f(x_{+,\min},1))\in\mathbb{R}^{\tilde{n}+1}} \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{c_1\le x_i<c_0\}[C(x_{+,\min}-x_i) + f(x_{+,\min},1) - f(x_i,0)]$$
$$s.t. -\sum_{i:c_1\le x_i<c_0} f(x_i,0) + \tilde{n} f(x_{+,\min},1) = \bar{\sigma}\epsilon.$$

By plugging the constraint to the objective, we obtain that the maximized value is

$$\rho(\epsilon) = C\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{c_1\le x_i<c_0\}[x_{+,\min}-x_i] + \frac{\tilde{\sigma}\epsilon}{n}.$$

Therefore, $\rho(\cdot)$ is differentiable for all $\epsilon\in\mathbb{R}$.

1.A.5 Linear Minimax MSE Estimator and Optimal Bias-Variance Tradeoff

Donoho (1994) defines the modulus of continuity as $\tilde{\omega}(\epsilon) = \sup\{|L(\theta) - L(\tilde{\theta})| : \|\boldsymbol{m}(\theta - \tilde{\theta})\| \leq \epsilon, \theta, \tilde{\theta} \in \Theta\}$. I first discuss some relationships between this definition and the definition in this chapter. If Θ is convex and centrosymmetric, the relationship $\tilde{\omega}(\epsilon) = 2\omega(\epsilon/2)$ holds. Also, if $\tilde{\omega}(\cdot)$ is differentiable, $\tilde{\omega}'(\epsilon) = \omega'(\epsilon/2)$. Let $(-\tilde{\theta}_{\tilde{\epsilon}}, \tilde{\theta}_{\tilde{\epsilon}})$ solve the modulus problem $\sup\{|L(\theta) - L(\tilde{\theta})| : \|\boldsymbol{m}(\theta - \tilde{\theta})\| \leq \tilde{\epsilon}, \theta, \tilde{\theta} \in \Theta\}$, i.e., $\tilde{\omega}(\tilde{\epsilon}) = 2L(\tilde{\theta}_{\tilde{\epsilon}})$ and $2\|\boldsymbol{m}(\tilde{\theta}_{\tilde{\epsilon}})\| \leq \tilde{\epsilon}$. Note that $\tilde{\theta}_{\tilde{\epsilon}}$ solves $\sup\{2L(\theta) : \|2\boldsymbol{m}(\theta)\| \leq \tilde{\epsilon}, \theta \in \Theta\}$, or $\sup\{L(\theta) : \|\boldsymbol{m}(\theta)\| \leq \tilde{\epsilon}/2, \theta \in \Theta\}$, so that $\tilde{\theta}_{\tilde{\epsilon}} = \theta_{\tilde{\epsilon}/2}$, where θ_{ϵ} solves $\sup\{L(\theta) : \|\boldsymbol{m}(\theta)\| \leq \epsilon, \theta \in \Theta\}$ as in the main text.

Linear Minimax MSE Estimators. Let $\tilde{\epsilon}_{\text{MSE}}$ solve

$$\frac{(\epsilon/2)^2}{(\epsilon/2)^2 + \sigma^2} = \frac{\epsilon\tilde{\omega}'(\epsilon)}{\tilde{\omega}(\epsilon)}.$$

The linear minimax MSE estimator of $L(\theta)$ is then given by $\hat{L}_{\text{MSE}}(\boldsymbol{Y}) = \boldsymbol{w}'_{\text{MSE}}\boldsymbol{Y}$ (Donoho, 1994), where

$$\boldsymbol{w}_{\text{MSE}} = \frac{2\tilde{\omega}'(\tilde{\epsilon}_{\text{MSE}})\boldsymbol{m}(\tilde{\theta}_{\tilde{\epsilon}_{\text{MSE}}})}{\tilde{\epsilon}_{\text{MSE}}}.$$

Now, let $\epsilon_{\text{MSE}} = \tilde{\epsilon}_{\text{MSE}}/2$, so that

$$\frac{(\epsilon_{\text{MSE}})^2}{(\epsilon_{\text{MSE}})^2 + \sigma^2} = \frac{2\epsilon_{\text{MSE}}\tilde{\omega}'(2\epsilon_{\text{MSE}})}{\tilde{\omega}(2\epsilon_{\text{MSE}})},$$

which is equivalent to

$$\frac{(\epsilon_{\text{MSE}})^2}{(\epsilon_{\text{MSE}})^2 + \sigma^2} = \frac{\epsilon_{\text{MSE}}\omega'(\epsilon_{\text{MSE}})}{\omega(\epsilon_{\text{MSE}})}.$$

We also have

$$\boldsymbol{w}_{\text{MSE}} = \frac{2\tilde{\omega}'(2\epsilon_{\text{MSE}})\boldsymbol{m}(\tilde{\theta}_{2\epsilon_{\text{MSE}}})}{2\epsilon_{\text{MSE}}} = \frac{\omega'(\epsilon_{\text{MSE}})\boldsymbol{m}(\theta_{\epsilon_{\text{MSE}}})}{\epsilon_{\text{MSE}}}.$$

Optimal Bias-Variance Frontier. The optimal bias-variance frontier in estimation of $L(\theta)$ can be traced out by a class of linear estimators $\{\tilde{L}_{\tilde{\epsilon}}(\boldsymbol{Y})\}_{\tilde{\epsilon}>0}$, where for each $\tilde{\epsilon} > 0$,

$$\tilde{L}_{\tilde{\epsilon}}(\boldsymbol{Y}) = \frac{2\tilde{\omega}'(\tilde{\epsilon})\boldsymbol{m}(\tilde{\theta}_{\tilde{\epsilon}})'}{\tilde{\epsilon}}\boldsymbol{Y}.$$

The maximum bias of $\tilde{L}_{\tilde{\epsilon}}$ is

$$\overline{\text{Bias}}_{\Theta}(\tilde{L}_{\tilde{\epsilon}}(\boldsymbol{Y})) = \frac{1}{2}(\tilde{\omega}(\tilde{\epsilon}) - \tilde{\epsilon}\tilde{\omega}'(\tilde{\epsilon})),$$

and the variance is $(\sigma\tilde{\omega}'(\tilde{\epsilon}))^2$.

For each $\epsilon > 0$, let $\hat{L}_{\epsilon}(\boldsymbol{Y}) = \tilde{L}_{2\epsilon}(\boldsymbol{Y})$. Then,

$$\hat{L}_{\epsilon}(\boldsymbol{Y}) = \frac{2\tilde{\omega}'(2\epsilon)\boldsymbol{m}(\tilde{\theta}_{2\epsilon})'}{2\epsilon}\boldsymbol{Y} = \frac{\omega'(\epsilon)\boldsymbol{m}(\theta_{\epsilon})'}{\epsilon}\boldsymbol{Y}.$$

Therefore, the class of estimators $\{\tilde{L}_{\tilde{\epsilon}}(\boldsymbol{Y})\}_{\tilde{\epsilon}>0}$ is the same as a class of linear estimators $\{\hat{L}_{\epsilon}(\boldsymbol{Y})\}_{\epsilon>0}$. The maximum bias of $\hat{L}_{\epsilon}$ is

$$\overline{\text{Bias}}_{\Theta}(\hat{L}_{\epsilon}(\boldsymbol{Y})) = \frac{1}{2}(\tilde{\omega}(2\epsilon) - 2\epsilon\tilde{\omega}'(2\epsilon)) = \omega(\epsilon) - \epsilon\omega'(\epsilon),$$

and the variance is $(\sigma\omega'(\epsilon))^2$.

Example 1.A.1 (Eligibility Cutoff Choice (Cont.))**.** Consider the setup in Section 1.4. Let $\iota \in \mathcal{F}_{\text{Lip}}(C)$ such that $\iota(x, 0) = 0$ for all x and $\iota(x, 1) = \frac{n}{\bar{n}}$. Then, $L(\iota) = 1$, and $f + c\iota \in \mathcal{F}_{\text{Lip}}(C)$ for all $c \in \mathbb{R}$ and $f \in \mathcal{F}_{\text{Lip}}(C)$. By Lemma 1.A.1 in Appendix 1.A.3, we obtain

$$\omega'(\epsilon; L, \tilde{\boldsymbol{m}}, \mathcal{F}_{\text{Lip}}(C)) = \frac{\epsilon}{\tilde{\boldsymbol{m}}(\iota)'\tilde{\boldsymbol{m}}(f_{\epsilon})} = \frac{\epsilon}{\frac{n}{\bar{n}}\sum_{i=1}^{n} d_i f_{\epsilon}(x_i, d_i)/\sigma^2(x_i, d_i)}, \tag{A.3}$$

where f_{ϵ} solves the modulus problem in this chapter's definition. Then,

$$\hat{L}_{\epsilon}(\boldsymbol{Y}) = \sum_{i=1}^{n} \frac{f_{\epsilon}(x_i, d_i)/\sigma^2(x_i, d_i)}{\frac{n}{\bar{n}}\sum_{j=1}^{n} d_j f_{\epsilon}(x_j, d_j)/\sigma^2(x_j, d_j)} Y_i.$$

The maximum bias of $\hat{L}_\epsilon$ is

$$\begin{aligned}&\overline{\text{Bias}}_{\mathcal{F}_{\text{Lip}}(C)}(\hat{L}_\epsilon(\boldsymbol{Y}))\\&= \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{c_1 \leq x_i < c_0\}[f_\epsilon(x_i,1) - f_\epsilon(x_i,0)] - \frac{\epsilon^2}{\frac{n}{\bar{n}}\sum_{i=1}^{n} d_i f_\epsilon(x_i,d_i)/\sigma^2(x_i,d_i)},\end{aligned}$$

and the variance is $\text{Var}(\hat{L}_\epsilon(\boldsymbol{Y})) = \epsilon^2 / \left(\frac{n}{\bar{n}}\sum_{i=1}^{n} d_i f_\epsilon(x_i,d_i)/\sigma^2(x_i,d_i)\right)^2$.

1.A.6 Computing ϵ^* for Example in Section 1.4

Here, I provide a procedure for computing $\epsilon^* \in \arg\max_{\epsilon\in[0,a^*]} \omega(\epsilon)\Phi(-\epsilon)$ for the example of eligibility cutoff choice under the Lipschitz class.

The procedure is based on the first-order condition. By differentiating $\omega(\epsilon)\Phi(-\epsilon)$, we have

$$\omega'(\epsilon)\Phi(-\epsilon) - \omega(\epsilon)\phi(-\epsilon) = \left[\frac{1-\Phi(\epsilon)}{\phi(\epsilon)} - \frac{\omega(\epsilon)}{\omega'(\epsilon)}\right]\omega'(\epsilon)\phi(\epsilon),$$

where the equality holds since $\Phi(x) = 1-\Phi(-x)$ and $\phi(x) = \phi(-x)$. $\frac{1-\Phi(\epsilon)}{\phi(\epsilon)}$ is the Mills ratio of a standard normal variable, which is strictly decreasing in ϵ. Since $\omega(\epsilon)$ is nondecreasing and concave, $\frac{\omega(\epsilon)}{\omega'(\epsilon)}$ is nondecreasing in ϵ. Therefore, $\frac{1-\Phi(\epsilon)}{\phi(\epsilon)} - \frac{\omega(\epsilon)}{\omega'(\epsilon)}$ is strictly decreasing in ϵ.

I suggest using the following procedure to compute ϵ^*.

1. If $\frac{1-\Phi(a^*)}{\phi(a^*)} - \frac{\omega(a^*)}{\omega'(a^*)} > 0$, $\epsilon^* = a^*$.
2. If not, use the bisection method to find $\epsilon^* \in [0, a^*]$ that solves $\frac{1-\Phi(\epsilon)}{\phi(\epsilon)} - \frac{\omega(\epsilon)}{\omega'(\epsilon)} = 0$.

Note that, for each ϵ, once we solve the convex optimization problem (1.4) to compute $\omega(\epsilon)$ and $(f_\epsilon(x_i,0), f_\epsilon(x_i,1))$, $i = 1,...,n$, we can compute $\omega'(\epsilon)$ using the closed-form expression (A.3) in Appendix 1.A.5.

1.B Proofs

1.B.1 Auxiliary Lemmas

Lemma 1.B.1. *Let $g(t) = h(t)\Phi\left(\frac{b-t}{a}\right)$, where $h(t)$ is nonconstant, nondecreasing, concave, and differentiable on $[\underline{t}, \bar{t}]$, $a > 0$, and $b \in \mathbb{R}$. If $a\frac{1-\Phi\left(\frac{\underline{t}-b}{a}\right)}{\phi\left(\frac{\underline{t}-b}{a}\right)} \leq \frac{h(\underline{t})}{h'(\underline{t})}$, then $g(t)$ is strictly decreasing on $[\underline{t}, \bar{t}]$. If $a\frac{1-\Phi\left(\frac{\bar{t}-b}{a}\right)}{\phi\left(\frac{\bar{t}-b}{a}\right)} \geq \frac{h(\bar{t})}{h'(\bar{t})}$, then $g(t)$ is strictly increasing on $[\underline{t}, \bar{t}]$. If $a\frac{1-\Phi\left(\frac{\underline{t}-b}{a}\right)}{\phi\left(\frac{\underline{t}-b}{a}\right)} > \frac{h(\underline{t})}{h'(\underline{t})}$ and $a\frac{1-\Phi\left(\frac{\bar{t}-b}{a}\right)}{\phi\left(\frac{\bar{t}-b}{a}\right)} < \frac{h(\bar{t})}{h'(\bar{t})}$, then there exists a unique $t^* \in [\underline{t}, \bar{t}]$ such that $g(t)$ is strictly increasing on $[\underline{t}, t^*)$ and strictly decreasing on $(t^*, \bar{t}]$. t^* is the solution to $a\frac{1-\Phi\left(\frac{t-b}{a}\right)}{\phi\left(\frac{t-b}{a}\right)} = \frac{h(t)}{h'(t)}$ if $h'(t)$ is continuous.*

Proof. Note first that $h'(\underline{t}) > 0$; if $h'(\underline{t}) \leq 0$, then $h'(t) = 0$ for all $t \in [\underline{t}, \bar{t}]$ since $h(t)$ is nondecreasing and concave, but this contradicts the assumption that $h(t)$ is nonconstant.

By differentiating $g(t)$, we have for $t \in [\underline{t}, \bar{t}]$,

$$g'(t) = h'(t)\Phi\left(\frac{b-t}{a}\right) - h(t)\phi\left(\frac{b-t}{a}\right)/a = \left[a\frac{1-\Phi\left(\frac{t-b}{a}\right)}{\phi\left(\frac{t-b}{a}\right)} - \frac{h(t)}{h'(t)}\right]h'(t)\phi\left(\frac{t-b}{a}\right)/a,$$

where the second equality holds since $\Phi(x) = 1 - \Phi(-x)$ and $\phi(x) = \phi(-x)$. By the fact that the Mills ratio $\frac{1-\Phi(x)}{\phi(x)}$ of a standard normal variable is strictly decreasing, $a\frac{1-\Phi\left(\frac{t-b}{a}\right)}{\phi\left(\frac{t-b}{a}\right)}$ is strictly decreasing in t. In addition, $a\frac{1-\Phi\left(\frac{t-b}{a}\right)}{\phi\left(\frac{t-b}{a}\right)}$ is continuous. Furthermore, since $h(t)$ is nondecreasing and concave on $[\underline{t}, \bar{t}]$, $\frac{h(t)}{h'(t)}$ is nondecreasing on $[\underline{t}, \bar{t}]$. Therefore, if $a\frac{1-\Phi\left(\frac{\underline{t}-b}{a}\right)}{\phi\left(\frac{\underline{t}-b}{a}\right)} \leq \frac{h(\underline{t})}{h'(\underline{t})}$, then $g'(t) < 0$ for all $t \in (\underline{t}, \bar{t}]$. If $a\frac{1-\Phi\left(\frac{\bar{t}-b}{a}\right)}{\phi\left(\frac{\bar{t}-b}{a}\right)} \geq \frac{h(\bar{t})}{h'(\bar{t})}$, then $g'(t) > 0$ for all $t \in [\underline{t}, \bar{t})$. If $a\frac{1-\Phi\left(\frac{\underline{t}-b}{a}\right)}{\phi\left(\frac{\underline{t}-b}{a}\right)} > \frac{h(\underline{t})}{h'(\underline{t})}$ and $a\frac{1-\Phi\left(\frac{\bar{t}-b}{a}\right)}{\phi\left(\frac{\bar{t}-b}{a}\right)} < \frac{h(\bar{t})}{h'(\bar{t})}$, then $g'(t) > 0$ for $t \in [\underline{t}, t^*)$ and $g'(t) < 0$ for $t \in (t^*, \bar{t}]$, where $t^* = \sup\{t \in [\underline{t}, \bar{t}] : g'(t) \geq 0\}$. t^* is the solution to $a\frac{1-\Phi\left(\frac{t-b}{a}\right)}{\phi\left(\frac{t-b}{a}\right)} = \frac{h(t)}{h'(t)}$ if $h'(t)$ is continuous. The conclusion then follows. □

Lemma 1.B.2. *Suppose that the conditions in Theorem 1.2 hold, and let $\Theta_{\epsilon^*} = \arg\max_{\theta\in\Theta:\|\boldsymbol{m}(\theta)\|\leq\epsilon^*} L(\theta)$. Then, $\|\boldsymbol{m}(\theta)\| = \epsilon^*$ for any $\theta \in \Theta_{\epsilon^*}$, and $\boldsymbol{m}(\theta) = \boldsymbol{m}(\tilde{\theta})$ for any $\theta, \tilde{\theta} \in \Theta_{\epsilon^*}$.*

Proof. First, pick any $\theta \in \Theta_{\epsilon^*}$. Since θ attains the modulus of continuity at ϵ^*, it also attains the modulus at $\|\boldsymbol{m}(\theta)\|$, so that $\omega(\|\boldsymbol{m}(\theta)\|) = \omega(\epsilon^*)$. It follows that $\|\boldsymbol{m}(\theta)\| = \epsilon^*$,

since if $\|\boldsymbol{m}(\theta)\| < \epsilon^*$,

$$\omega(\|\boldsymbol{m}(\theta)\|)\Phi(-\|\boldsymbol{m}(\theta)\|/\sigma) = \omega(\epsilon^*)\Phi(-\|\boldsymbol{m}(\theta)\|/\sigma) > \omega(\epsilon^*)\Phi(-\epsilon^*/\sigma),$$

which contradicts the assumption that ϵ^* maximizes $\omega(\epsilon)\Phi(-\epsilon/\sigma)$ over $[0, a^*\sigma]$.

Now, pick any $\theta, \tilde{\theta} \in \Theta_{\epsilon^*}$. By the above argument, $\|\boldsymbol{m}(\theta)\| = \|\boldsymbol{m}(\tilde{\theta})\| = \epsilon^*$, and hence $\boldsymbol{m}(\theta), \boldsymbol{m}(\tilde{\theta}) \in \{\boldsymbol{\beta} \in \mathbb{R}^n : \|\boldsymbol{\beta}\| \leq \epsilon^*\}$. Suppose that $\boldsymbol{m}(\theta) \neq \boldsymbol{m}(\tilde{\theta})$, and let $\bar{\theta} = \lambda\theta + (1-\lambda)\tilde{\theta}$ for some $\lambda \in (0,1)$. Then $L(\bar{\theta}) = \lambda L(\theta) + (1-\lambda)L(\tilde{\theta}) = \omega(\epsilon^*)$. By the convexity of Θ, $\bar{\theta} \in \Theta$. Furthermore, since $\{\boldsymbol{\beta} \in \mathbb{R}^n : \|\boldsymbol{\beta}\| \leq \epsilon^*\}$ is strictly convex, $\boldsymbol{m}(\bar{\theta}) = \lambda\boldsymbol{m}(\theta) + (1-\lambda)\boldsymbol{m}(\tilde{\theta})$ is an interior point of $\{\boldsymbol{\beta} \in \mathbb{R}^n : \|\boldsymbol{\beta}\| \leq \epsilon^*\}$, which implies that $\|\boldsymbol{m}(\bar{\theta})\| < \epsilon^*$. Thus, $\bar{\theta}$ attains the modulus at ϵ^*, but then it must be the case that $\|\boldsymbol{m}(\bar{\theta})\| = \epsilon^*$. □

Lemma 1.B.3. *Let $\psi(a,b) = a\Phi(-b)$. Then, $\psi(a,b)$ is strictly quasi-concave on $(0,\infty)\times\mathbb{R}$.*

Proof. Take any $a_0, a_1 > 0$ and $b_0, b_1 \in \mathbb{R}$ such that $(a_0, b_0) \neq (a_1, b_1)$. I show that $\psi(a_0 + \lambda(a_1 - a_0), b_0 + \lambda(b_1 - b_0)) > \min\{\psi(a_0, b_0), \psi(a_1, b_1)\}$ for all $\lambda \in (0,1)$.

First, suppose that $a_0 \leq a_1$ and $b_0 \geq b_1$. Since either $a_0 < a_1$ or $b_0 > b_1$ or both must hold, $\psi(a_0 + \lambda(a_1 - a_0), b_0 + \lambda(b_1 - b_0)) = (a_0 + \lambda(a_1 - a_0))\Phi(-b_0 - \lambda(b_1 - b_0))$ is strictly increasing in λ. It then follows that $\psi(a_0 + \lambda(a_1 - a_0), b_0 + \lambda(b_1 - b_0)) > \psi(a_0, b_0)$. Likewise, if $a_0 \geq a_1$ and $b_0 \leq b_1$, then $\psi(a_0 + \lambda(a_1 - a_0), b_0 + \lambda(b_1 - b_0)) > \psi(a_1, b_1)$.

Now suppose that $a_0 < a_1$ and $b_0 < b_1$. Note that the set $\{(a_0, b_0) + \lambda(a_1 - a_0, b_1 - b_0) : \lambda \in (0,1)\}$ is equivalent to

$$\left\{ \left(0, b_0 - a_0\frac{b_1 - b_0}{a_1 - a_0}\right) + t\left(\frac{a_1 - a_0}{b_1 - b_0}, 1\right) : t \in \left(a_0\frac{b_1 - b_0}{a_1 - a_0}, a_1\frac{b_1 - b_0}{a_1 - a_0}\right) \right\}.$$

We have

$$\begin{aligned}\psi\left(\left(0, b_0 - a_0\frac{b_1 - b_0}{a_1 - a_0}\right) + t\left(\frac{a_1 - a_0}{b_1 - b_0}, 1\right)\right) &= t\left(\frac{a_1 - a_0}{b_1 - b_0}\right)\Phi\left(-b_0 + a_0\frac{b_1 - b_0}{a_1 - a_0} - t\right) \\ &= \left(\frac{a_1 - a_0}{b_1 - b_0}\right) g\left(t\right),\end{aligned}$$

where $g\left(t\right) = t\Phi\left(-b_0 + a_0\frac{b_1 - b_0}{a_1 - a_0} - t\right)$. Lemma 1.B.1 implies that the minimum of $g(t)$ over

an interval $[t_0, t_1]$ is attained only at t_0 or t_1 or both. Hence, for all $t \in \left(a_0\frac{b_1-b_0}{a_1-a_0}, a_1\frac{b_1-b_0}{a_1-a_0}\right)$,

$$g(t) > \min\left\{g\left(a_0\frac{b_1-b_0}{a_1-a_0}\right), g\left(a_1\frac{b_1-b_0}{a_1-a_0}\right)\right\}.$$

Thus, for all $t \in \left(a_0\frac{b_1-b_0}{a_1-a_0}, a_1\frac{b_1-b_0}{a_1-a_0}\right)$,

$$\begin{aligned}
&\psi\left(\left(0, b_0 - a_0\frac{b_1-b_0}{a_1-a_0}\right) + t\left(\frac{a_1-a_0}{b_1-b_0}, 1\right)\right) \\
&> \left(\frac{a_1-a_0}{b_1-b_0}\right)\min\left\{g\left(a_0\frac{b_1-b_0}{a_1-a_0}\right), g\left(a_1\frac{b_1-b_0}{a_1-a_0}\right)\right\} \\
&= \min\{\psi(a_0, b_0), \psi(a_1, b_1)\}.
\end{aligned}$$

Therefore, $\psi(a_0 + \lambda(a_1 - a_0), b_0 + \lambda(b_1 - b_0)) > \min\{\psi(a_0, b_0), \psi(a_1, b_1)\}$ for all $\lambda \in (0, 1)$. The same argument holds for the case where $a_0 > a_1$ and $b_0 > b_1$. □

1.B.2 Proof of Proposition 1.1

The problem (1.4) is equivalent to

$$\max_{(f(x_i,0), f(x_i,1))_{i=1,\ldots,n} \in \mathbb{R}^{2n}} \frac{1}{n}\sum_{i:c_1 \le x_i < c_0} [f(x_i, 1) - f(x_i, 0)] \tag{B.1}$$

$$s.t. \quad \sum_{i:x_i<c_1} \frac{f(x_i,0)^2}{\sigma^2(x_i,0)} + \sum_{i:c_1\le x_i<c_0} \frac{f(x_i,0)^2}{\sigma^2(x_i,0)} + \frac{f(x_{+,\min},1)^2}{\sigma^2_{+,\min}} + \sum_{i:x_i>x_{+,\min}} \frac{f(x_i,1)^2}{\sigma^2(x_i,1)} \le \epsilon^2, \tag{B.2}$$

$$f(x_i, d) - f(x_j, d) \le C|x_i - x_j|, \quad d \in \{0, 1\}, i, j \in \{1, ..., n\}. \tag{B.3}$$

First, consider the case where $\epsilon = 0$. Since the left-hand side of (B.2) must be zero, any solution satisfies $f(x_i, 0) = 0$ if $x_i < c_0$ and $f(x_i, 1) = 0$ if $x_i \ge c_0$. The objective (B.1) and constraint (B.2) do not depend on $(f(x_i, 0))_{i:x_i \ge c_0}$, so we can set these values arbitrarily as long as the Lipschitz constraint (B.3) holds. I set all of them to 0, so that $f(x_i, 0) = 0$ for

all i. The above problem then reduces to solving

$$\max_{(f(x_i,1))_{i:x_i<c_0}\in\mathbb{R}^n} \frac{1}{n} \sum_{i:c_1\leq x_i<c_0} f(x_i,1) \tag{B.4}$$

$$s.t.\ \ f(x_i,1) - f(x_j,1) \leq C|x_i - x_j|, \quad i,j \in \{1,...,n\}, \tag{B.5}$$

where $f(x_i,1) = 0$ for any i with $x_i \geq c_0$. The Lipschitz constraint (B.5) implies that $f(x_i,1) \leq C(x_{+,\min} - x_i)$ for any i with $c_1 \leq x_i < c_0$. Therefore, the value of the objective (B.4) is at most $\frac{1}{n}\sum_{i:c_1\leq x_i<c_0} C(x_{+,\min} - x_i)$. This value is attained by setting $f(x_i,1) = C(x_{+,\min} - x_i)$ for all i with $x_i < c_0$, which satisfies the Lipschitz constraint (B.5). In sum, when $\epsilon = 0$, one solution to problem (B.1)–(B.3) is given by

$$f_0(x_i,0) = 0,\ i = 1,...,n, \quad f_0(x_i,1) = \begin{cases} 0 & \text{if } x_i > x_{+,\min}, \\ C(x_{+,\min} - x_i) & \text{if } x_i \leq x_{+,\min}. \end{cases}$$

Now, let $\bar{\epsilon} = (C/\bar{\sigma}) \min_{x,\tilde{x}\in\mathcal{X},x\neq\tilde{x}} |x - \tilde{x}|$, where $\bar{\sigma} = \max_i \sigma(x_i,d_i)$ and $\mathcal{X} = \{x_i : i = 1,...,n\}$ is the set of points of x in the sample. Consider any $\epsilon \in (0,\bar{\epsilon}]$. I claim that problem (B.1)–(B.3) reduces to solving

$$\max_{((f(x_i,0),f(x_i,1))_{i:c_1\leq x_i<c_0},f(x_{+,\min},1))\in\mathbb{R}^{2\tilde{n}+1}} \frac{1}{n} \sum_{i:c_1\leq x_i<c_0} [f(x_i,1) - f(x_i,0)] \tag{B.6}$$

$$s.t. \quad \sum_{i:c_1\leq x_i<c_0} \frac{f(x_i,0)^2}{\sigma^2(x_i,0)} + \frac{f(x_{+,\min},1)^2}{\sigma^2_{+,\min}} \leq \epsilon^2, \tag{B.7}$$

$$f(x_i,1) - f(x_j,1) \leq C|x_i - x_j|, \quad i,j \in \{k : c_1 \leq x_k \leq x_{+,\min}\}. \tag{B.8}$$

To see this, suppose that $((f_\epsilon(x_i,0), f_\epsilon(x_i,1))_{i:c_1\leq x_i<c_0}, f_\epsilon(x_{+,\min},1))$ is a solution to the above problem. First, it must be the case that $f_\epsilon(x_i,1) \geq 0$ for any i with $c_1 \leq x_i < c_0$, since if $f_\epsilon(x_i,1) < 0$ for some i with $c_1 \leq x_i < c_0$, it is possible to strictly increase the objective (B.6) without violating the constraints (B.7) and (B.8) by changing f_ϵ to $\tilde{f}_\epsilon$ such that $\tilde{f}_\epsilon(x_i,1) = \max\{f(x_i,1),0\}$ for any i with $c_1 \leq x_i < c_0$. By similar arguments, it must be the case that $f_\epsilon(x_{+,\min},1) \geq 0$ and $f_\epsilon(x_i,0) \leq 0$ for i with $c_1 \leq x_i < c_0$.

Next, given $((f_\epsilon(x_i,0), f_\epsilon(x_i,1))_{i:c_1\le x_i<c_0}, f_\epsilon(x_{+,\min},1))$, set

$$f_\epsilon(x_i,0) = 0 \quad \text{if } x_i < c_1 \text{ or } x_i \ge c_0, \tag{B.9}$$

$$f_\epsilon(x_i,1) = \begin{cases} 0 & \text{if } x_i > x_{+,\min}, \\ C(x_{-,\min} - x_i) + f_\epsilon(x_{-,\min},1) & \text{if } x_i < c_1, \end{cases} \tag{B.10}$$

where $x_{-,\min} = \min\{x_i : c_1 \le x_i < c_0\}$ is the smallest values of x among those whose treatment status would be changed if the cutoff were changed to c_1 in the sample. I show that $(f_\epsilon(x_i,0), f_\epsilon(x_i,1))_{i=1,\dots,n}$ is a solution to the original problem (B.1)–(B.3). Clearly, it satisfies the constraint (B.2). To see that the Lipschitz constraint (B.3) is satisfied for $d = 0$, it suffices to check that $|f_\epsilon(x_i,0) - f_\epsilon(x_j,0)| \le C|x_i - x_j|$ for any i, j with $x_i, x_j \le x_{+,\min}$, given that the Lipschitz constraint for any i, j with $x_i, x_j \ge x_{+,\min}$ holds by construction. Observe that for any i, j with $x_i, x_j \le x_{+,\min}$ and $x_i \ne x_j$,

$$\begin{aligned} |f_\epsilon(x_i,0) - f_\epsilon(x_j,0)|^2 &= f_\epsilon(x_i,0)^2 + f_\epsilon(x_j,0)^2 - 2f_\epsilon(x_i,0)f_\epsilon(x_j,0) \\ &\le f_\epsilon(x_i,0)^2 + f_\epsilon(x_j,0)^2 \\ &\le \bar\sigma^2\bar\epsilon^2 \\ &= C^2 \min_{x,\tilde x\in\mathcal{X}, x\ne\tilde x} |x-\tilde x|^2 \\ &\le C^2|x_i - x_j|^2, \end{aligned}$$

where the inequality in the second line holds since $f_\epsilon(x_i,0) \le 0$ for all i, the inequality in the third line follows from the constraint (B.7), and the equality in the fourth line from the definition of $\bar\epsilon$. For $d = 1$, it is sufficient to check that $|f_\epsilon(x_i,1) - f_\epsilon(x_{+,\min},1)| \le C(x_i - x_{+,\min})$ for any i with $x_i > x_{+,\min}$ and that $|f_\epsilon(x_i,1) - f_\epsilon(x_{-,\min},1)| \le C(x_{-,\min} - x_i)$ for any i with $x_i < c_1$. The latter immediately follows from the construction of f_ϵ. Regarding the former, for any i with $x_i > x_{+,\min}$,

$$|f_\epsilon(x_i,1) - f_\epsilon(x_{+,\min},1)|^2 = f_\epsilon(x_{+,\min},1)^2 \le \bar\sigma^2\bar\epsilon^2 \le C^2|x_i - x_{+,\min}|^2.$$

Therefore, $(f_\epsilon(x_i,0), f_\epsilon(x_i,1))_{i=1,\dots,n}$ satisfies constraint (B.3). Since the value of the original

problem (B.1)–(B.3) at $(f_\epsilon(x_i,0), f_\epsilon(x_i,1))_{i=1,\ldots,n}$ is equal to the maximized value of the less constrained problem (B.6)–(B.8), $(f_\epsilon(x_i,0), f_\epsilon(x_i,1))_{i=1,\ldots,n}$ is the solution to problem (B.1)–(B.3).

Now I derive a solution to (B.6)–(B.8). Note that, given a value of $f(x_{+,\min},1)$, the objective is maximized only when $f(x_i,1) = C(x_{+,\min}-x_i)+f(x_{+,\min},1)$ under the constraints (B.7) and (B.8). Plugging this into (B.6)–(B.8), one can further simplify the problem to

$$\max_{((f(x_i,0))_{i:c_1\le x_i<c_0}, f(x_{+,\min},1))\in\mathbb{R}^{\tilde{n}+1}} \frac{C}{n}\sum_{i:c_1\le x_i<c_0}(x_{+,\min}-x_i)+\frac{\tilde{n}}{n}f(x_{+,\min},1) - \frac{1}{n}\sum_{i:c_1\le x_i<c_0} f(x_i,0)$$
$$s.t. \quad \sum_{i:c_1\le x_i<c_0}\frac{f(x_i,0)^2}{\sigma^2(x_i,0)}+\frac{f(x_{+,\min},1)^2}{\sigma^2_{+,\min}}\le\epsilon^2.$$

This is a convex optimization problem that maximizes a weighted sum of $\tilde{n}+1$ unknowns under the constraint on the upper bound on a weighted Euclidean norm of the unknowns. Simple calculations show that the solution is given by $f(x_i,0) = -\frac{\sigma^2(x_i,0)\epsilon}{(\tilde{n}^2\sigma^2_{+,\min}+\sum_{i:c_1\le x_i<c_0}\sigma^2(x_i,0))^{1/2}}$ for any i with $c_1\le x_i<c_0$ and $f(x_{+,\min},1) = \frac{\tilde{n}\sigma^2_{+,\min}\epsilon}{(\tilde{n}^2\sigma^2_{+,\min}+\sum_{i:c_1\le x_i<c_0}\sigma^2(x_i,d_i))^{1/2}}$. From (B.9) and (B.10), one solution to the original problem (B.1)–(B.3) is then given by

$$f_\epsilon(x_i,0) = \begin{cases} 0 & \text{if } x_i<c_1 \text{ or } x_i\ge c_0,\\ -\frac{\sigma^2(x_i,0)\epsilon}{(\tilde{n}^2\sigma^2_{+,\min}+\sum_{i:c_1\le x_i<c_0}\sigma^2(x_i,0))^{1/2}} & \text{if } c_1\le x_i<c_0,\end{cases}$$
$$f_\epsilon(x_i,1) = \begin{cases} 0 & \text{if } x_i>x_{+,\min},\\ C(x_{+,\min}-x_i)+\frac{\tilde{n}\sigma^2_{+,\min}\epsilon}{(\tilde{n}^2\sigma^2_{+,\min}+\sum_{i:c_1\le x_i<c_0}\sigma^2(x_i,0))^{1/2}} & \text{if } x_i\le x_{+,\min}.\end{cases}$$

The modulus of continuity is given by

$$\omega(\epsilon) = C\frac{1}{n}\sum_{i=1}^n \mathbf{1}\{c_1\le x_i<c_0\}[x_{+,\min}-x_i]+\frac{1}{n}\left(\tilde{n}^2\sigma^2_{+,\min}+\sum_{i:c_1\le x_i<c_0}\sigma^2(x_i,0)\right)^{1/2}\epsilon.$$

□

1.B.3 Proof of Proposition 1.2

I first show that $\sigma\frac{1-\Phi(\epsilon/\sigma)}{\phi(\epsilon/\sigma)} \geq \frac{\omega(\epsilon)}{\omega'(\epsilon)}$ for all $\epsilon \leq \epsilon^*$. From the arguments in the proof of Lemma 1.1, if $\sigma\frac{1-\Phi(a^*)}{\phi(a^*)} \geq \frac{\omega(a^*\sigma)}{\omega'(a^*\sigma)}$, then $\epsilon^* = a^*\sigma$, and the above statement holds. Suppose that $\sigma\frac{1-\Phi(a^*)}{\phi(a^*)} < \frac{\omega(a^*\sigma)}{\omega'(a^*\sigma)}$. Again from the arguments in the proof of Lemma 1.1, $\sigma\frac{1-\Phi(\epsilon/\sigma)}{\phi(\epsilon/\sigma)} > \frac{\omega(\epsilon)}{\omega'(\epsilon)}$ for $\epsilon < \epsilon^*$ and $\sigma\frac{1-\Phi(\epsilon/\sigma)}{\phi(\epsilon/\sigma)} < \frac{\omega(\epsilon)}{\omega'(\epsilon)}$ for $\epsilon > \epsilon^*$. Since the left-hand side is continuous, and the curve $(\epsilon, \frac{\omega(\epsilon)}{\omega'(\epsilon)})_{\epsilon>0}$ is connected by Lemma 3 of Donoho (1994), $\sigma\frac{1-\Phi(\epsilon^*/\sigma)}{\phi(\epsilon^*/\sigma)} = \frac{\omega(\epsilon^*)}{\omega'(\epsilon^*)}$. This implies that $\sigma\frac{1-\Phi(\epsilon/\sigma)}{\phi(\epsilon/\sigma)} \geq \frac{\omega(\epsilon)}{\omega'(\epsilon)}$ for all $\epsilon \leq \epsilon^*$.

Now, note that ϵ_{MSE} solves $\frac{\epsilon^2+\sigma^2}{\epsilon} = \frac{\omega(\epsilon)}{\omega'(\epsilon)}$. If $\frac{\epsilon^2+\sigma^2}{\epsilon} > \sigma\frac{1-\Phi(\epsilon/\sigma)}{\phi(\epsilon/\sigma)}$ for all $\epsilon > 0$, then $\frac{\epsilon^2+\sigma^2}{\epsilon} > \frac{\omega(\epsilon)}{\omega'(\epsilon)}$ for all $\epsilon \leq \epsilon^*$, which implies that $\epsilon^* < \epsilon_{\text{MSE}}$. Below, I show that $\frac{\epsilon^2+\sigma^2}{\epsilon} > \sigma\frac{1-\Phi(\epsilon/\sigma)}{\phi(\epsilon/\sigma)}$ for all $\epsilon > 0$. Let $g(\epsilon) = \frac{\epsilon^2+\sigma^2}{\epsilon}$. We have $g'(\epsilon) = 1 - \frac{\sigma^2}{\epsilon^2}$, which is strictly increasing in ϵ. Therefore, $g(\epsilon)$ is minimized at $\epsilon = \sigma$, at which $g'(\epsilon) = 0$. For any $\epsilon > 0$, $g(\epsilon) \geq g(\sigma) = 2\sigma > \sigma\frac{1-\Phi(0)}{\phi(0)} > \sigma\frac{1-\Phi(\epsilon/\sigma)}{\phi(\epsilon/\sigma)}$, where the second inequality holds since $\frac{1-\Phi(0)}{\phi(0)} \approx 1.253$, and the last holds since $\frac{1-\Phi(\epsilon/\sigma)}{\phi(\epsilon/\sigma)}$ is strictly decreasing. □

1.B.4 Proof of Corollary 1.1

I first show that $\omega(\epsilon) = L(\epsilon\theta^*)$ for any $\epsilon \in (0, a^*\sigma]$. Since $\omega(\epsilon) \leq \sup_{\theta\in\mathbb{V}:\|\boldsymbol{m}(\theta)\|\leq\epsilon} L(\theta)$ by definition, and $\epsilon\theta^* \in \Theta$ and $\|\boldsymbol{m}(\epsilon\theta^*)\| \leq \epsilon$ for all $\epsilon \in (0, a^*\sigma]$ by assumption, we have

$$L(\epsilon\theta^*) \leq \omega(\epsilon) \leq \sup_{\theta\in\mathbb{V}:\|\boldsymbol{m}(\theta)\|\leq\epsilon} L(\theta), \quad \epsilon \in (0, a^*\sigma].$$

Therefore, it suffices to show that $\sup_{\theta\in\mathbb{V}:\|\boldsymbol{m}(\theta)\|\leq\epsilon} L(\theta) \leq L(\epsilon\theta^*)$ for all $\epsilon \in (0, a^*\sigma]$. Suppose that $\sup_{\theta\in\mathbb{V}:\|\boldsymbol{m}(\theta)\|\leq\epsilon} L(\theta) > L(\epsilon\theta^*)$ for some $\epsilon \in (0, a^*\sigma]$. Then, there exists $\theta \in \mathbb{V}$ such that $\|\boldsymbol{m}(\theta)\| \leq \epsilon$ and $L(\theta) > L(\epsilon\theta^*)$. It follows that $\|\boldsymbol{m}(\theta/\epsilon)\| \leq 1$ and $L(\theta/\epsilon) > L(\theta^*)$, which contradicts the assumption that θ^* solves $\sup_{\theta\in\mathbb{V}:\|\boldsymbol{m}(\theta)\|\leq 1} L(\theta)$.

By the definition of a^*, we have

$$\arg\max_{0<\epsilon\leq a^*\sigma} \omega(\epsilon)\Phi(-\epsilon/\sigma) = \arg\max_{0<\epsilon\leq a^*\sigma} L(\theta^*)\epsilon\Phi(-\epsilon/\sigma) = \{a^*\sigma\}.$$

It is straightforward to show that Assumptions 1.1 and 1.2 hold. Applying Theorem 1.1 or

Theorem 1.2, it is shown that the minimax regret decision rule is

$$\delta^*(\boldsymbol{Y}) = \mathbf{1}\left\{\boldsymbol{m}(a^*\sigma\theta^*)'\boldsymbol{Y} \geq 0\right\} = \mathbf{1}\left\{\boldsymbol{m}(\theta^*)'\boldsymbol{Y} \geq 0\right\},$$

and the minimax risk is $\mathcal{R}(\sigma;\Theta) = a^*\sigma L(\theta^*)\Phi(-a^*)$. □

1.B.5 Proof of Lemma 1.1

Let $g(\epsilon) = \omega(\epsilon)\Phi(-\epsilon/\sigma)$. As in the proof of Lemma 1.B.1, differentiating g (from the right) at $\epsilon \geq 0$ gives

$$g'(\epsilon) = \left[\sigma\frac{1-\Phi\left(\frac{\epsilon}{\sigma}\right)}{\phi\left(\frac{\epsilon}{\sigma}\right)} - \frac{\omega(\epsilon)}{\omega'(\epsilon)}\right]\omega'(\epsilon)\phi\left(\frac{\epsilon}{\sigma}\right)/\sigma.$$

By the fact that the Mills ratio $\frac{1-\Phi(x)}{\phi(x)}$ of a standard normal variable is strictly decreasing, $\sigma\frac{1-\Phi\left(\frac{\epsilon}{\sigma}\right)}{\phi\left(\frac{\epsilon}{\sigma}\right)}$ is strictly decreasing in ϵ. In addition, $\sigma\frac{1-\Phi\left(\frac{\epsilon}{\sigma}\right)}{\phi\left(\frac{\epsilon}{\sigma}\right)}$ is continuous. Furthermore, since $\omega(\epsilon)$ is nondecreasing and concave, $\frac{\omega(\epsilon)}{\omega'(\epsilon)}$ is nondecreasing.

Suppose that $\sigma > 2\phi(0)\frac{\omega(0)}{\omega'(0)}$. Then, $g'(0) > 0$. This implies that $g(\epsilon) > g(0)$ for any sufficiently small $\epsilon > 0$. If $g'(a^*\sigma) > 0$, $g(\epsilon)$ is strictly increasing on $[0, a^*\sigma]$, so g is uniquely maximized at $a^*\sigma$ over $[0, a^*\sigma]$. If $g'(a^*\sigma) \leq 0$, $g'(\epsilon) > 0$ for $\epsilon \in [0, \epsilon^*)$ and $g'(\epsilon) < 0$ for $\epsilon \in (\epsilon^*, a^*\sigma]$, where $\epsilon^* = \sup\{\epsilon \in [0, a^*\sigma] : g'(\epsilon) \geq 0\}$. Then g is uniquely maximized at ϵ^* over $[0, a^*\sigma]$.

Suppose that $\sigma \leq 2\phi(0)\frac{\omega(0)}{\omega'(0)}$. Then, $g'(0) \leq 0$. Since $\sigma\frac{1-\Phi\left(\frac{\epsilon}{\sigma}\right)}{\phi\left(\frac{\epsilon}{\sigma}\right)} - \frac{\omega(\epsilon)}{\omega'(\epsilon)}$ is strictly decreasing, $g'(\epsilon) < 0$ for any $\epsilon > 0$. By the mean value theorem, for every $\epsilon > 0$, $g(\epsilon) = g(0) + g'(\ddot{\epsilon})\epsilon$ for some $\ddot{\epsilon} \in (0, \epsilon)$, which implies $g(\epsilon) < g(0)$. Therefore, g is uniquely maximized at 0 over $[0, a^*\sigma]$. □

1.B.6 Proof of Lemma 1.2

Here, I prove the following result, which covers Lemma 1.2 as a special case.

Lemma 1.B.4 (Univariate Problems (General))**.** *Suppose that* $\Theta = [-\tau_1, -\tau_0] \cup [\tau_0, \tau_1]$ *for some* τ_0, τ_1 *such that* $\tau_1 \geq \tau_0 \geq 0$ *and* $\tau_1 > 0$*, that* $\boldsymbol{m}(\theta) = \theta$*, and that* $L(\theta) = \theta$*. Then, the*

decision rule $\delta^*(Y) = \mathbf{1}\{Y \geq 0\}$ *is minimax regret. The minimax risk is given by*

$$\mathcal{R}_{\text{uni}}(\sigma; [-\tau_1, -\tau_0] \cup [\tau_0, \tau_1]) = \begin{cases} \tau_1 \Phi(-\tau_1/\sigma) & \text{if } \tau_1 < a^*\sigma, \\ a^*\sigma\Phi(-a^*) & \text{if } a^*\sigma \in [\tau_0, \tau_1], \\ \tau_0\Phi(-\tau_0/\sigma) & \text{if } a^*\sigma < \tau_0. \end{cases}$$

Following Stoye (2009), I use a statistical game to solve the minimax regret problem. Consider the following two-person zero-sum game between the decision maker and nature. The strategy space for the decision maker is $\mathcal{D}$, the set of all decision rules. The strategy space for nature is $\Delta(\Theta)$, the set of probability distributions on $\Theta = [-\tau_1, -\tau_0] \cup [\tau_0, \tau_1]$. If the decision maker chooses $\delta \in \mathcal{D}$ and nature chooses $\pi \in \Delta(\Theta)$, nature's expected payoff (and the decision maker's expected loss) is given by $r(\delta, \pi) = \int R(\delta, \theta) d\pi(\theta)$, the Bayes risk of δ with respect to prior π. By Theorem 17 in Chapter 5 of Berger (1985), if (δ^*, π^*) satisfies

$$\delta^* \in \arg\min_{\delta \in \mathcal{D}} r(\delta, \pi^*), \text{ and } R(\delta^*, \theta) \leq r(\delta^*, \pi^*) \text{ for all } \theta \in \Theta,$$

then δ^* is a minimax regret rule and π^* is a least favorable prior. Below I construct (δ^*, π^*) that satisfies the above conditions.

I first restrict the search space of decision rules to an essentially complete class of decision rules, following Tetenov (2012).[42] Since Y has monotone likelihood ratio and the loss function satisfies $l(1, \theta) - l(0, \theta) \geq 0$ if $\theta < 0$ and $l(1, \theta) - l(0, \theta) \leq 0$ if $\theta > 0$, it follows from Theorem 5 in Chapter 8 of Berger (1985) (which is originally from Karlin and Rubin (1956)) that the class of monotone decision rules

$$\delta(Y) = \begin{cases} 0 & \text{if } Y < t, \\ \lambda & \text{if } Y = t, \\ 1 & \text{if } Y > t, \end{cases}$$

42. A class $\mathcal{C}$ of decision rules is essentially complete if, for any decision rule $\delta \notin \mathcal{C}$, there is a decision rule $\delta' \in \mathcal{C}$ such that $R(\delta, \theta) \geq R(\delta', \theta)$ for all $\theta \in \Theta$.

where $t \in \mathbb{R}$ and $\lambda \in [0,1]$, is essentially complete. Furthermore, since $\mathbb{P}_\theta(Y = t) = 0$, a smaller class of threshold decision rules $\delta(Y) = \mathbf{1}\{Y \geq t\}$, $t \in \mathbb{R}$, is also essentially complete.

Let δ_t denote the threshold rule with threshold t. Since $Y \sim \mathcal{N}(\theta, \sigma^2)$,

$$R(\delta_t, \theta) = \begin{cases} \theta\Phi(\sigma^{-1}(t-\theta)) & \text{if } \theta \geq 0, \\ (-\theta)(1-\Phi(\sigma^{-1}(t-\theta))) & \text{if } \theta < 0. \end{cases}$$

Let $\bar{R}_0(t, \tau_0, \tau_1) = \max_{\theta \in [-\tau_1, -\tau_0]} R(\delta_t, \theta) = \max_{\theta \in [-\tau_1, -\tau_0]} -\theta(1 - \Phi(\sigma^{-1}(t-\theta)))$ and $\bar{R}_1(t, \tau_0, \tau_1) = \max_{\theta \in [\tau_0, \tau_1]} R(\delta_t, \theta) = \max_{\theta \in [\tau_0, \tau_1]} \theta\Phi(\sigma^{-1}(t-\theta))$. By symmetry of $\bar{R}_0(t, \tau_0, \tau_1)$ and $\bar{R}_1(t, \tau_0, \tau_1)$, $\bar{R}_0(0, \tau_0, \tau_1) = \bar{R}_1(0, \tau_0, \tau_1)$.

Now let $\theta_0^* \in \arg\max_{\theta \in [-\tau_1, -\tau_0]} R(\delta_0, \theta)$ and $\theta_1^* \in \arg\max_{\theta \in [\tau_0, \tau_1]} R(\delta_0, \theta)$, where $\delta_0(Y) = \mathbf{1}\{Y \geq 0\}$. By symmetry, $\theta_0^* = -\theta_1^*$. Let $\pi^* \in \Delta(\Theta)$ be such that

$$\pi^*(\theta_1^*) = \frac{-\theta_0^*\phi(\sigma^{-1}(-\theta_0^*))}{-\theta_0^*\phi(\sigma^{-1}(-\theta_0^*)) + \theta_1^*\phi(\sigma^{-1}(-\theta_1^*))} = \frac{1}{2},$$

and $\pi^*(\theta_0^*) = 1 - \pi^*(\theta_1^*) = \frac{1}{2}$. Since $\bar{R}_0(0, \tau_0, \tau_1) = \bar{R}_1(0, \tau_0, \tau_1) = R(\delta_0, \theta_0^*) = R(\delta_0, \theta_1^*)$, $r(\delta_0, \pi^*) = R(\delta_0, \theta_0^*) = R(\delta_0, \theta_1^*) \geq R(\delta_0, \theta)$ for all $\theta \in [-\tau_1, -\tau_0] \cup [\tau_0, \tau_1]$.

Since the class of threshold rules is essentially complete, $\delta_0 \in \arg\min_{\delta \in \mathcal{D}} r(\delta, \pi^*)$ if $0 \in \arg\min_{t \in \mathbb{R}} r(\delta_t, \pi^*)$. Observe

$$\begin{aligned} r(\delta_t, \pi^*) &= R(\delta_t, \theta_0^*)(1 - \pi^*(\theta_1^*)) + R(\delta_t, \theta_1^*)\pi^*(\theta_1^*) \\ &= -\theta_0^*(1 - \Phi(\sigma^{-1}(t - \theta_0^*)))(1 - \pi^*(\theta_1^*)) + \theta_1^*\Phi(\sigma^{-1}(t - \theta_1^*))\pi^*(\theta_1^*), \end{aligned}$$

and

$$\begin{aligned} \frac{\partial r(\delta_t, \pi^*)}{\partial t} &= \sigma^{-1}\theta_0^*\phi(\sigma^{-1}(t-\theta_0^*))(1-\pi^*(\theta_1^*)) + \sigma^{-1}\theta_1^*\phi(\sigma^{-1}(t-\theta_1^*))\pi^*(\theta_1^*) \\ &= \sigma^{-1}\phi(\sigma^{-1}(t-\theta_0^*)) \left[\theta_0^*(1-\pi^*(\theta_1^*)) + \theta_1^*\pi^*(\theta_1^*)\frac{\phi(\sigma^{-1}(t-\theta_1^*))}{\phi(\sigma^{-1}(t-\theta_0^*))}\right]. \end{aligned}$$

Since $\frac{\phi(\sigma^{-1}(t-\theta_1^*))}{\phi(\sigma^{-1}(t-\theta_0^*))}$ is increasing in t by the monotone likelihood ratio property, and $\theta_0^*(1 -$

$\pi^*(\theta_1^*)) + \theta_1^*\pi^*(\theta_1^*)\frac{\phi(\sigma^{-1}(t-\theta_1^*))}{\phi(\sigma^{-1}(t-\theta_0^*))}$ is equal to zero at $t = 0$ by construction of π^*, it follows that

$$\frac{\partial r(\delta_t, \pi^*)}{\partial t} \begin{cases} > 0 & \text{if } t > 0, \\ = 0 & \text{if } t = 0, \\ < 0 & \text{if } t < 0. \end{cases}$$

Therefore, $r(\delta_t, \pi^*)$ is minimized at $t = 0$. Thus, δ_0 is minimax regret.

The minimax risk is given by

$$\begin{aligned} \mathcal{R}_{\text{uni}}(\sigma; [-\tau_1, -\tau_0] \cup [\tau_0, \tau_1]) &= \max_{\theta \in [\tau_0, \tau_1]} \theta\Phi(-\theta/\sigma) \\ &= \max_{a \in [\tau_0/\sigma, \tau_1/\sigma]} \sigma a \Phi(-a). \end{aligned}$$

By Lemma 1.B.1, $a\Phi(-a)$ has a unique maximizer a^* over $[0, \infty)$ and $a\Phi(-a)$ is strictly increasing on $[0, a^*)$ and strictly decreasing on (a^*, ∞). Therefore,

$$\mathcal{R}_{\text{uni}}(\sigma; [-\tau_1, -\tau_0] \cup [\tau_0, \tau_1]) = \begin{cases} \tau_1\Phi(-\tau_1/\sigma) & \text{if } \tau_1 < \sigma a^*, \\ \sigma a^*\Phi(-a^*) & \text{if } \sigma a^* \in [\tau_0, \tau_1], \\ \tau_0\Phi(-\tau_0/\sigma) & \text{if } \sigma a^* < \tau_0. \end{cases}$$

□

1.B.7 Proof of Lemma 1.3

Here, I prove the following result, which covers Lemma 1.3 as a special case.

Lemma 1.B.5 (Informative One-dimensional Subproblems (General))**.** *Suppose that* $\Theta = [-\bar{\theta}, -t\bar{\theta}] \cup [t\bar{\theta}, \bar{\theta}]$*, where* $\bar{\theta} \in \mathbb{V}$*,* $L(\bar{\theta}) > 0$*,* $\boldsymbol{m}(\bar{\theta}) \neq \mathbf{0}$*, and* $t \in [0, 1]$*. Then, the decision rule* $\delta^*(\boldsymbol{Y}) = \mathbf{1}\left\{\boldsymbol{m}(\bar{\theta})'\boldsymbol{Y} \geq 0\right\}$ *is minimax regret. The minimax risk is given by*

$$\mathcal{R}(\sigma; [-\bar{\theta}, -t\bar{\theta}] \cup [t\bar{\theta}, \bar{\theta}]) = \frac{L(\bar{\theta})}{\|\boldsymbol{m}(\bar{\theta})\|}\mathcal{R}_{\text{uni}}(\sigma; [-\|\boldsymbol{m}(\bar{\theta})\|, -t\|\boldsymbol{m}(\bar{\theta})\|] \cup [t\|\boldsymbol{m}(\bar{\theta})\|, \|\boldsymbol{m}(\bar{\theta})\|]).$$

Fix $\bar{\theta} \in \mathbb{V}$, where $L(\bar{\theta}) > 0$ and $\boldsymbol{m}(\bar{\theta}) \neq \mathbf{0}$, and $t \in [0, 1]$. We can write $[-\bar{\theta}, -t\bar{\theta}] \cup [t\bar{\theta}, \bar{\theta}] =$

$\{\lambda\bar{\theta} : \lambda \in [-1, -t] \cup [t, 1]\}$. For $\lambda \in [-1, -t] \cup [t, 1]$, the regret of decision rule δ under $\lambda\bar{\theta}$ equals

$$\begin{aligned} R(\delta, \lambda\bar{\theta}) &= (L(\lambda\bar{\theta}))^+(1 - \mathbb{E}_{\lambda\bar{\theta}}[\delta(\boldsymbol{Y})]) + (-L(\lambda\bar{\theta}))^+\mathbb{E}_{\lambda\bar{\theta}}[\delta(\boldsymbol{Y})] \\ &= L(\bar{\theta})\left(\lambda^+(1 - \mathbb{E}_{\lambda\bar{\theta}}[\delta(\boldsymbol{Y})]) + (-\lambda)^+\mathbb{E}_{\lambda\bar{\theta}}[\delta(\boldsymbol{Y})]\right), \end{aligned}$$

where $x^+ = \max\{x, 0\}$. Minimax regret decision rules thus solve

$$\inf_{\delta} \sup_{\lambda \in [-1, -t] \cup [t, 1]} \left(\lambda^+(1 - \mathbb{E}_{\lambda\bar{\theta}}[\delta(\boldsymbol{Y})]) + (-\lambda)^+\mathbb{E}_{\lambda\bar{\theta}}[\delta(\boldsymbol{Y})]\right).$$

Viewing λ as a parameter, I derive a sufficient statistic of $\boldsymbol{Y}$ for λ. For $\lambda \in [-1, -t] \cup [t, 1]$, $\boldsymbol{Y} \sim \mathcal{N}(\lambda\boldsymbol{m}(\bar{\theta}), \sigma^2\boldsymbol{I}_n)$ under $\lambda\bar{\theta}$. It follows that the probability density of $\boldsymbol{Y}$ is

$$\begin{aligned} p(\boldsymbol{y}) &= \frac{1}{\sqrt{(2\pi)^n}\sigma^n} \exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{y} - \lambda\boldsymbol{m}(\bar{\theta})\|^2\right) \\ &= \frac{1}{\sqrt{(2\pi)^n}\sigma^n} \exp\left(-\frac{1}{2\sigma^2}(\|\boldsymbol{y}\|^2 - 2\lambda\boldsymbol{m}(\bar{\theta})'\boldsymbol{y} + \lambda^2\|\boldsymbol{m}(\bar{\theta})\|^2)\right) \\ &= h(\boldsymbol{y})g(T(\boldsymbol{y}), \lambda), \end{aligned}$$

where $h(\boldsymbol{y}) = \frac{1}{\sqrt{(2\pi)^n}\sigma^n} \exp(-\frac{1}{2\sigma^2}\|\boldsymbol{y}\|^2)$, $g(t, \lambda) = \exp(-\frac{1}{2\sigma^2}(-2\lambda t + \lambda^2)\|\boldsymbol{m}(\bar{\theta})\|^2)$, and $T(\boldsymbol{y}) = \frac{\boldsymbol{m}(\bar{\theta})'\boldsymbol{y}}{\|\boldsymbol{m}(\theta)\|^2}$. By the factorization theorem, $T(\boldsymbol{Y})$ is a sufficient statistic for λ.

It follows from Theorem 1 in Chapter 1 of Berger (1985) that the class of decision rules that only depend on $T(\boldsymbol{Y})$ is essentially complete. Since $T(\boldsymbol{Y}) \sim \mathcal{N}(\lambda, \frac{\sigma^2}{\|\boldsymbol{m}(\bar{\theta})\|^2})$ under $\lambda\bar{\theta}$, minimax regret decision rules that only depend on $T(\boldsymbol{Y})$ solve

$$\inf_{\delta} \sup_{\lambda \in [-1, -t] \cup [t, 1]} \left(\lambda^+(1 - \mathbb{E}_{\lambda}[\delta(T)]) + (-\lambda)^+\mathbb{E}_{\lambda}[\delta(T)]\right),$$

where the expectation is taken with respect to $T \sim \mathcal{N}(\lambda, \frac{\sigma^2}{\|\boldsymbol{m}(\bar{\theta})\|^2})$. This problem is equivalent to the univariate problem where $\Theta = [-1, -t] \cup [t, 1]$, $\boldsymbol{m}(\theta) = \theta$, $L(\theta) = \theta$, and the variance of the observed normal variable is $\frac{\sigma^2}{\|\boldsymbol{m}(\bar{\theta})\|^2}$. Thus, by Lemma 1.B.4, the decision rule

$$\delta^*(\boldsymbol{Y}) = \mathbf{1}\{T(\boldsymbol{Y}) \geq 0\} = \mathbf{1}\left\{\boldsymbol{m}(\bar{\theta})'\boldsymbol{Y} \geq 0\right\}$$

is minimax regret. The minimax risk is given by

$$\begin{aligned}\mathcal{R}(\sigma;[-\bar{\theta},-t\bar{\theta}]\cup[t\bar{\theta},\bar{\theta}]) &= L(\bar{\theta})\mathcal{R}_{\text{uni}}\left(\frac{\sigma}{\|\boldsymbol{m}(\bar{\theta})\|};[-1,-t]\cup[t,1]\right)\\ &= \frac{L(\bar{\theta})}{\|\boldsymbol{m}(\bar{\theta})\|}\mathcal{R}_{\text{uni}}\left(\sigma;[-\|\boldsymbol{m}(\bar{\theta})\|-t\|\boldsymbol{m}(\bar{\theta})\|]\cup[t\|\boldsymbol{m}(\bar{\theta})\|,\|\boldsymbol{m}(\bar{\theta})\|]\right),\end{aligned}$$

where the second equality follows from the fact that $\mathcal{R}_{\text{uni}}(\alpha\sigma;[-\alpha\tau_1,-\alpha\tau_0]\cup[\alpha\tau_0,\alpha\tau_1]) = \alpha\mathcal{R}_{\text{uni}}(\sigma;[-\tau_1,-\tau_0]\cup[\tau_0,\tau_1])$ for all $\alpha>0$. □

1.B.8 Proof of Lemma 1.4

Let ϵ^* be the unique nonzero solution to $\max_{0\le\epsilon\le a^*\sigma}\omega(\epsilon)\Phi(-\epsilon/\sigma)$, and let θ_{ϵ^*} attain the modulus of continuity at ϵ^*. By Lemma 1.B.2, $\|\boldsymbol{m}(\theta_{\epsilon^*})\| = \epsilon^*$. I first introduce some notation. Pick any $\underline{\eta}\in(0,\min\{\omega(\epsilon^*)\Phi(-\epsilon^*/\sigma),\epsilon^*\})$ and any $\bar{\epsilon}>\epsilon^*$, and define

$$\Gamma_{+,\underline{\eta},\bar{\epsilon}} = \left\{(L(\theta),\boldsymbol{m}(\theta)')'\in\mathbb{R}^{n+1}:\theta\in\Theta, L(\theta)\ge\underline{\eta}, \frac{\boldsymbol{m}(\theta_{\epsilon^*})'\boldsymbol{m}(\theta)}{\|\boldsymbol{m}(\theta_{\epsilon^*})\|}\ge\underline{\eta}, \|\boldsymbol{m}(\theta)\|\le\bar{\epsilon}\right\}.$$

Since $\omega(\epsilon^*)$ is finite, $\omega(\bar{\epsilon})$ is also finite by the convexity of $\omega(\epsilon)$. Note that $\Gamma_{+,\underline{\eta},\bar{\epsilon}}$ is bounded, since $\underline{\eta}\le\alpha\le\omega(\bar{\epsilon})$ and $\|\boldsymbol{\beta}\|\le\bar{\epsilon}$ for all $\boldsymbol{\gamma}=(\alpha,\boldsymbol{\beta}')'\in\Gamma_{+,\underline{\eta},\bar{\epsilon}}$. Let

$$\Theta_{+,\underline{\eta},\bar{\epsilon}} = \left\{\theta\in\Theta: L(\theta)\ge\underline{\eta}, \frac{\boldsymbol{m}(\theta_{\epsilon^*})'\boldsymbol{m}(\theta)}{\|\boldsymbol{m}(\theta_{\epsilon^*})\|}\ge\underline{\eta}, \|\boldsymbol{m}(\theta)\|\le\bar{\epsilon}\right\}$$

and

$$\begin{aligned}\Theta_{\underline{\eta},\bar{\epsilon}} &= \{\theta\in\Theta:(\theta\in\Theta_{+,\underline{\eta},\bar{\epsilon}}) \text{ or } (-\theta\in\Theta_{+,\underline{\eta},\bar{\epsilon}})\}\\ &= \left\{\theta\in\Theta:\left[\left(L(\theta)\ge\underline{\eta}, \frac{\boldsymbol{m}(\theta_{\epsilon^*})'\boldsymbol{m}(\theta)}{\|\boldsymbol{m}(\theta_{\epsilon^*})\|}\ge\underline{\eta}\right) \text{ or } \left(L(\theta)\le-\underline{\eta}, \frac{\boldsymbol{m}(\theta_{\epsilon^*})'\boldsymbol{m}(\theta)}{\|\boldsymbol{m}(\theta_{\epsilon^*})\|}\le-\underline{\eta}\right)\right]\right.\\ &\qquad\left.\text{and } \|\boldsymbol{m}(\theta)\|\le\bar{\epsilon}\right\}.\end{aligned}$$

We can then write $\Gamma_{+,\underline{\eta},\bar{\epsilon}} = \{(L(\theta),\boldsymbol{m}(\theta)')'\in\mathbb{R}^{n+1}:\theta\in\Theta_{+,\underline{\eta},\bar{\epsilon}}\}$.

Now, let $\bar{\Gamma}_{+,\underline{\eta},\bar{\epsilon}}$ denote the closure of $\Gamma_{+,\underline{\eta},\bar{\epsilon}}$. Define a set-valued function $\Psi:\bar{\Gamma}_{+,\underline{\eta},\bar{\epsilon}}\to$

$2^{\bar{\Gamma}_{+,\underline{\eta},\bar{\epsilon}}}$ as follows: for $\boldsymbol{\gamma} = (\alpha, \boldsymbol{\beta}')' \in \bar{\Gamma}_{+,\underline{\eta},\bar{\epsilon}}$,

$$\Psi(\boldsymbol{\gamma}) = \arg \max_{\bar{\boldsymbol{\gamma}}=(\bar{\alpha},\bar{\boldsymbol{\beta}}')' \in \bar{\Gamma}_{+,\underline{\eta},\bar{\epsilon}}} \bar{\alpha}\Phi\left(-\frac{\boldsymbol{\beta}'\bar{\boldsymbol{\beta}}}{\sigma\|\boldsymbol{\beta}\|}\right).$$

Note that $\alpha > 0$ and $\boldsymbol{\beta} \neq 0$ for all $(\alpha, \boldsymbol{\beta}')' \in \bar{\Gamma}_{+,\underline{\eta},\bar{\epsilon}}$, since $\tilde{\alpha} \geq \underline{\eta}$ and $\frac{\boldsymbol{m}(\theta_{\epsilon^*})'\tilde{\boldsymbol{\beta}}}{\|\boldsymbol{m}(\theta_{\epsilon^*})\|} \geq \underline{\eta}$ for all $\tilde{\boldsymbol{\gamma}} = (\tilde{\alpha}, \tilde{\boldsymbol{\beta}}')' \in \Gamma_{+,\underline{\eta},\bar{\epsilon}}$, and $(\alpha, \boldsymbol{\beta}')'$ is a point or a limit point of $\Gamma_{+,\underline{\eta},\bar{\epsilon}}$.

The proof consists of six steps.

Step 1. *Ψ has a fixed point, i.e., there exists $\boldsymbol{\gamma} \in \bar{\Gamma}_{+,\underline{\eta},\bar{\epsilon}}$ such that $\boldsymbol{\gamma} \in \Psi(\boldsymbol{\gamma})$.*

Proof. I apply Kakutani's fixed point theorem. First of all, $\bar{\Gamma}_{+,\underline{\eta},\bar{\epsilon}}$ is nonempty, since $(L(\theta_{\epsilon^*}), \boldsymbol{m}(\theta_{\epsilon^*}))')' \in \bar{\Gamma}_{+,\underline{\eta},\bar{\epsilon}}$. Furthermore, $\bar{\Gamma}_{+,\underline{\eta},\bar{\epsilon}}$ is closed and bounded by construction.

I now show that $\bar{\Gamma}_{+,\underline{\eta},\bar{\epsilon}}$ is convex. It suffices to show that $\Gamma_{+,\underline{\eta},\bar{\epsilon}}$ is convex, since the closure of a convex subset of $\mathbb{R}^{n+1}$ is convex. Pick any $\boldsymbol{\gamma}, \tilde{\boldsymbol{\gamma}} \in \Gamma_{+,\underline{\eta},\bar{\epsilon}}$. Let $\theta, \tilde{\theta} \in \Theta_{+,\underline{\eta},\bar{\epsilon}}$ be such that $(L(\theta), \boldsymbol{m}(\theta)')' = \boldsymbol{\gamma}$ and $(L(\tilde{\theta}), \boldsymbol{m}(\tilde{\theta})')' = \tilde{\boldsymbol{\gamma}}$. Fix $\lambda \in [0,1]$. By the linearity of L and $\boldsymbol{m}$, $\lambda\boldsymbol{\gamma} + (1-\lambda)\tilde{\boldsymbol{\gamma}} = (L(\lambda\theta + (1-\lambda)\tilde{\theta}), \boldsymbol{m}(\lambda\theta + (1-\lambda)\tilde{\theta})')'$. Since $\lambda\theta + (1-\lambda)\tilde{\theta} \in \Theta$ by the convexity of Θ, $L(\lambda\theta + (1-\lambda)\tilde{\theta}) = \lambda L(\theta) + (1-\lambda)L(\tilde{\theta}) \geq \underline{\eta}$, $\frac{\boldsymbol{m}(\theta_{\epsilon^*})'\boldsymbol{m}(\lambda\theta+(1-\lambda)\tilde{\theta})}{\|\boldsymbol{m}(\theta_{\epsilon^*})\|} = \lambda\frac{\boldsymbol{m}(\theta_{\epsilon^*})'\boldsymbol{m}(\theta)}{\|\boldsymbol{m}(\theta_{\epsilon^*})\|} + (1-\lambda)\frac{\boldsymbol{m}(\theta_{\epsilon^*})'\boldsymbol{m}(\tilde{\theta})}{\|\boldsymbol{m}(\theta_{\epsilon^*})\|} \geq \underline{\eta}$, and $\|\boldsymbol{m}(\lambda\theta + (1-\lambda)\tilde{\theta})\| \leq \|\boldsymbol{m}(\lambda\theta)\| + \|\boldsymbol{m}((1-\lambda)\tilde{\theta})\| = \lambda\|\boldsymbol{m}(\theta)\| + (1-\lambda)\|\boldsymbol{m}(\tilde{\theta})\| \leq \bar{\epsilon}$, it follows that $\lambda\theta + (1-\lambda)\tilde{\theta} \in \Theta_{+,\underline{\eta},\bar{\epsilon}}$. Therefore, $\lambda\boldsymbol{\gamma} + (1-\lambda)\tilde{\boldsymbol{\gamma}} \in \Gamma_{+,\underline{\eta},\bar{\epsilon}}$.

Next, I show that $\Psi(\boldsymbol{\gamma})$ is nonempty and convex for all $\boldsymbol{\gamma} \in \bar{\Gamma}_{+,\underline{\eta},\bar{\epsilon}}$. Fix $\boldsymbol{\gamma} = (\alpha, \boldsymbol{\beta}')' \in \bar{\Gamma}_{+,\underline{\eta},\bar{\epsilon}}$. Let

$$S_{\boldsymbol{\beta}} = \left\{\left(\bar{\alpha}, \frac{\boldsymbol{\beta}'\bar{\boldsymbol{\beta}}}{\sigma\|\boldsymbol{\beta}\|}\right) \in \mathbb{R}^2 : \bar{\boldsymbol{\gamma}} = (\bar{\alpha}, \bar{\boldsymbol{\beta}}')' \in \bar{\Gamma}_{+,\underline{\eta},\bar{\epsilon}}\right\},$$

which is a subset of $(0,\infty) \times \mathbb{R}$. Using $S_{\boldsymbol{\beta}}$, we can write

$$\Psi(\boldsymbol{\gamma}) = \left\{(\bar{\alpha}, \bar{\boldsymbol{\beta}}')' \in \bar{\Gamma}_{+,\underline{\eta},\bar{\epsilon}} : \left(\bar{\alpha}, \frac{\boldsymbol{\beta}'\bar{\boldsymbol{\beta}}}{\sigma\|\boldsymbol{\beta}\|}\right) \in \arg\max_{(a,b)\in S_{\boldsymbol{\beta}}} a\Phi(-b)\right\}.$$

Since the mapping $\bar{\boldsymbol{\gamma}} \mapsto \left(\bar{\alpha}, \frac{\boldsymbol{\beta}'\bar{\boldsymbol{\beta}}}{\sigma\|\boldsymbol{\beta}\|}\right)$ is continuous and $\bar{\Gamma}_{+,\underline{\eta},\bar{\epsilon}}$ is compact, $S_{\boldsymbol{\beta}}$ is compact. Furthermore, since $\bar{\boldsymbol{\gamma}} \mapsto \left(\bar{\alpha}, \frac{\boldsymbol{\beta}'\bar{\boldsymbol{\beta}}}{\sigma\|\boldsymbol{\beta}\|}\right)$ is linear and $\bar{\Gamma}_{+,\underline{\eta},\bar{\epsilon}}$ is convex, $S_{\boldsymbol{\beta}}$ is convex. It then follows that $\arg\max_{(a,b)\in S_{\boldsymbol{\beta}}} a\Phi(-b)$ is nonempty and singleton, since $a\Phi(-b)$ is continuous and is strictly quasi-concave on $(0,\infty)\times\mathbb{R}$ by Lemma 1.B.3. Let $(a^*_{\boldsymbol{\beta}}, b^*_{\boldsymbol{\beta}}) \in \arg\max_{(a,b)\in S_{\boldsymbol{\beta}}} a\Phi(-b)$.

We can then write

$$\Psi(\boldsymbol{\gamma}) = \left\{ (\bar{\alpha}, \bar{\boldsymbol{\beta}}')' \in \bar{\Gamma}_{+,\underline{\eta},\bar{\epsilon}} : \bar{\alpha} = a^*_{\boldsymbol{\beta}}, \frac{\boldsymbol{\beta}'\bar{\boldsymbol{\beta}}}{\sigma\|\boldsymbol{\beta}\|} = b^*_{\boldsymbol{\beta}} \right\},$$

which is nonempty and convex.

Lastly, I show that Ψ has a closed graph. Let $f(\bar{\boldsymbol{\gamma}}, \boldsymbol{\gamma}) = \bar{\alpha}\Phi\left(-\frac{\boldsymbol{\beta}'\bar{\boldsymbol{\beta}}}{\sigma\|\boldsymbol{\beta}\|}\right)$. Take any sequence $\{(\boldsymbol{\gamma}_n, \boldsymbol{\gamma}^*_n)\}_{n=1}^{\infty}$ such that $\boldsymbol{\gamma}_n, \boldsymbol{\gamma}^*_n \in \bar{\Gamma}_{+,\underline{\eta},\bar{\epsilon}}$ for all n, $(\boldsymbol{\gamma}_n, \boldsymbol{\gamma}^*_n) \to (\boldsymbol{\gamma}, \boldsymbol{\gamma}^*)$, and $\boldsymbol{\gamma}^*_n \in \Psi(\boldsymbol{\gamma}_n)$ for all n. Since $\boldsymbol{\gamma}_n, \boldsymbol{\gamma}^*_n \in \bar{\Gamma}_{+,\underline{\eta},\bar{\epsilon}}$ and $\bar{\Gamma}_{+,\underline{\eta},\bar{\epsilon}}$ is closed, it follows that $\boldsymbol{\gamma}, \boldsymbol{\gamma}^* \in \bar{\Gamma}_{+,\underline{\eta},\bar{\epsilon}}$. This implies that $\boldsymbol{\beta}^* \neq 0$.

I show that $\boldsymbol{\gamma}^* \in \Psi(\boldsymbol{\gamma})$. Suppose this does not hold. Then there exist $\boldsymbol{\gamma}^{**} \in \bar{\Gamma}_{+,\underline{\eta},\bar{\epsilon}}$ and $\epsilon > 0$ such that $f(\boldsymbol{\gamma}^{**}, \boldsymbol{\gamma}) > f(\boldsymbol{\gamma}^*, \boldsymbol{\gamma}) + 3\epsilon$. Also, since f is continuous on $\bar{\Gamma}_{+,\underline{\eta},\bar{\epsilon}} \times \bar{\Gamma}_{+,\underline{\eta},\bar{\epsilon}}$ and $(\boldsymbol{\gamma}_n, \boldsymbol{\gamma}^*_n) \to (\boldsymbol{\gamma}, \boldsymbol{\gamma}^*)$, we have $f(\boldsymbol{\gamma}^{**}, \boldsymbol{\gamma}_n) > f(\boldsymbol{\gamma}^{**}, \boldsymbol{\gamma}) - \epsilon$ and $f(\boldsymbol{\gamma}^*, \boldsymbol{\gamma}) > f(\boldsymbol{\gamma}^*_n, \boldsymbol{\gamma}_n) - \epsilon$ for any sufficiently large n. Combining the preceding inequalities, we obtain for any sufficiently large n,

$$f(\boldsymbol{\gamma}^{**}, \boldsymbol{\gamma}_n) > f(\boldsymbol{\gamma}^*, \boldsymbol{\gamma}) + 2\epsilon > f(\boldsymbol{\gamma}^*_n, \boldsymbol{\gamma}_n) + \epsilon.$$

This contradicts the assumption that $\boldsymbol{\gamma}^*_n \in \Psi(\boldsymbol{\gamma}_n)$ for all n.

Application of Kakutani's fixed point theorem proves the statement. □

Let $\boldsymbol{\gamma}^* = (\alpha^*, (\boldsymbol{\beta}^*)')'$ be a fixed point of Ψ. In Steps 2–5, I prove that $L(\theta_{\epsilon^*}) = \alpha^*$ and $\boldsymbol{m}(\theta_{\epsilon^*}) = \boldsymbol{\beta}^*$.

$\boldsymbol{\gamma}^*$ may not be an element of $\Gamma_{+,\underline{\eta},\bar{\epsilon}}$, but since it is an element of $\bar{\Gamma}_{+,\underline{\eta},\bar{\epsilon}}$, we can take sequences $\{\boldsymbol{\gamma}_n = (\alpha_n, \boldsymbol{\beta}_n')'\}_{n=1}^{\infty}$ and $\{\theta_n\}_{n=1}^{\infty}$ such that $\theta_n \in \Theta_{+,\underline{\eta},\bar{\epsilon}}$ and $\boldsymbol{\gamma}_n = (L(\theta_n), \boldsymbol{m}(\theta_n)')' \in \Gamma_{+,\underline{\eta},\bar{\epsilon}}$ for all $n \geq 1$ and that $\lim_{n\to\infty} \boldsymbol{\gamma}_n = \boldsymbol{\gamma}^*$. Let $\tilde{\delta}(\boldsymbol{Y}) = \mathbf{1}\{(\boldsymbol{\beta}^*)'\boldsymbol{Y} \geq 0\}$. Below, I suppress the argument σ of the minimax risk $\mathcal{R}(\sigma; \cdot)$ for notational brevity.

Step 2. $\sup_{\theta\in\Theta_{\underline{\eta},\bar{\epsilon}}} R(\tilde{\delta}, \theta) = \sup_{\theta\in\Theta_{+,\underline{\eta},\bar{\epsilon}}} \mathcal{R}([-\theta, \theta] \cap \Theta_{\underline{\eta},\bar{\epsilon}}) = \lim_{n\to\infty} \mathcal{R}([-\theta_n, \theta_n] \cap \Theta_{\underline{\eta},\bar{\epsilon}}) = \alpha^*\Phi\left(-\frac{\|\boldsymbol{\beta}^*\|}{\sigma}\right)$.

Proof. Since $\boldsymbol{\gamma}^* \in \arg\max_{\boldsymbol{\gamma}=(\alpha,\boldsymbol{\beta}')'\in\bar{\Gamma}_{+,\underline{\eta},\bar{\epsilon}}} \alpha\Phi\left(-\frac{(\boldsymbol{\beta}^*)'\boldsymbol{\beta}}{\sigma\|\boldsymbol{\beta}^*\|}\right)$,

$$\begin{aligned}
\max_{\boldsymbol{\gamma}=(\alpha,\boldsymbol{\beta}')'\in\bar{\Gamma}_{+,\underline{\eta},\bar{\epsilon}}} \alpha\Phi\left(-\frac{(\boldsymbol{\beta}^*)'\boldsymbol{\beta}}{\sigma\|\boldsymbol{\beta}^*\|}\right) &= \alpha^*\Phi\left(-\frac{(\boldsymbol{\beta}^*)'(\boldsymbol{\beta}^*)}{\sigma\|\boldsymbol{\beta}^*\|}\right) \\
&= \lim_{n\to\infty} \alpha_n\Phi\left(-\frac{(\boldsymbol{\beta}^*)'\boldsymbol{\beta}_n}{\sigma\|\boldsymbol{\beta}^*\|}\right) \\
&= \lim_{n\to\infty} L(\theta_n)\Phi\left(-\frac{(\boldsymbol{\beta}^*)'\boldsymbol{m}(\theta_n)}{\sigma\|\boldsymbol{\beta}^*\|}\right) \\
&= \lim_{n\to\infty} R(\tilde{\delta},\theta_n),
\end{aligned}$$

where the second equality follows by the fact that the mapping $\boldsymbol{\gamma} \mapsto \alpha\Phi\left(-\frac{(\boldsymbol{\beta}^*)'\boldsymbol{\beta}}{\sigma\|\boldsymbol{\beta}^*\|}\right)$ is continuous. Also, by continuity of the mapping $\boldsymbol{\gamma} \mapsto \alpha\Phi\left(-\frac{(\boldsymbol{\beta})'\boldsymbol{\beta}}{\sigma\|\boldsymbol{\beta}\|}\right)$,

$$\begin{aligned}
\alpha^*\Phi\left(-\frac{(\boldsymbol{\beta}^*)'(\boldsymbol{\beta}^*)}{\sigma\|\boldsymbol{\beta}^*\|}\right) &= \lim_{n\to\infty} \alpha_n\Phi\left(-\frac{\boldsymbol{\beta}_n'\boldsymbol{\beta}_n}{\sigma\|\boldsymbol{\beta}_n\|}\right) \\
&= \lim_{n\to\infty} L(\theta_n)\Phi\left(-\frac{\boldsymbol{m}(\theta_n)'\boldsymbol{m}(\theta_n)}{\sigma\|\boldsymbol{m}(\theta_n)\|}\right) \\
&= \lim_{n\to\infty} R(\delta_n,\theta_n),
\end{aligned}$$

where $\delta_n(\boldsymbol{Y}) = \mathbf{1}\{\boldsymbol{m}(\theta_n)'\boldsymbol{Y} \geq 0\}$ for all n.

On the other hand, by definition,

$$\begin{aligned}
\sup_{\boldsymbol{\gamma}=(\alpha,\boldsymbol{\beta}')'\in\bar{\Gamma}_{+,\underline{\eta},\bar{\epsilon}}} \alpha\Phi\left(-\frac{(\boldsymbol{\beta}^*)'\boldsymbol{\beta}}{\sigma\|\boldsymbol{\beta}^*\|}\right) &\geq \sup_{\boldsymbol{\gamma}=(\alpha,\boldsymbol{\beta}')'\in\Gamma_{+,\underline{\eta},\bar{\epsilon}}} \alpha\Phi\left(-\frac{(\boldsymbol{\beta}^*)'\boldsymbol{\beta}}{\sigma\|\boldsymbol{\beta}^*\|}\right) \\
&= \sup_{\theta\in\Theta_{+,\underline{\eta},\bar{\epsilon}}} L(\theta)\Phi\left(-\frac{(\boldsymbol{\beta}^*)'\boldsymbol{m}(\theta)}{\sigma\|\boldsymbol{\beta}^*\|}\right) \\
&= \sup_{\theta\in\Theta_{+,\underline{\eta},\bar{\epsilon}}} R(\tilde{\delta},\theta) \\
&\geq \lim_{n\to\infty} R(\tilde{\delta},\theta_n).
\end{aligned}$$

Therefore,

$$\sup_{\theta\in\Theta_{+,\underline{\eta},\bar{\epsilon}}} R(\tilde{\delta},\theta) = \lim_{n\to\infty} R(\tilde{\delta},\theta_n) = \lim_{n\to\infty} R(\delta_n,\theta_n) = \alpha^*\Phi\left(-\frac{(\boldsymbol{\beta}^*)'(\boldsymbol{\beta}^*)}{\sigma\|\boldsymbol{\beta}^*\|}\right).$$

Note that, $\sup_{\theta\in\Theta_{+,\underline{\eta},\bar{\epsilon}}} R(\tilde{\delta},\theta) = \sup_{\theta\in\Theta_{\underline{\eta},\bar{\epsilon}}} R(\tilde{\delta},\theta)$ by the symmetry of the regret function

and the centrosymmetry of $\Theta_{\underline{\eta},\bar{\epsilon}}$.

We also have

$$\begin{aligned}\lim_{n\to\infty} R(\delta_n,\theta_n) &\leq \lim_{n\to\infty} \sup_{\theta\in[-\theta_n,\theta_n]\cap\Theta_{\underline{\eta},\bar{\epsilon}}} R(\delta_n,\theta)\\ &= \lim_{n\to\infty} \mathcal{R}([-\theta_n,\theta_n]\cap\Theta_{\underline{\eta},\bar{\epsilon}})\\ &\leq \sup_{\theta\in\Theta_{+,\underline{\eta},\bar{\epsilon}}} \mathcal{R}([-\theta,\theta]\cap\Theta_{\underline{\eta},\bar{\epsilon}}),\end{aligned}$$

where the equality in the second line holds since $[-\theta_n,\theta_n]\cap\Theta_{\underline{\eta},\bar{\epsilon}} = [-\theta_n,-t_n\theta_n]\cup[t_n\theta_n,\theta_n]$ with $t_n = \max\{\frac{\underline{\eta}}{L(\theta_n)}, \underline{\eta}\frac{\|\boldsymbol{m}(\theta_{\epsilon^*})\|}{\boldsymbol{m}(\theta_{\epsilon^*})'\boldsymbol{m}(\theta_n)}\}$, and δ_n is minimax regret for $[-\theta_n,-t_n\theta_n]\cup[t_n\theta_n,\theta_n]$ by Lemma 1.B.5. However, by definition,

$$\sup_{\theta\in\Theta_{\underline{\eta},\bar{\epsilon}}} R(\tilde{\delta},\theta) \geq \mathcal{R}(\Theta_{\underline{\eta},\bar{\epsilon}}) \geq \sup_{\theta\in\Theta_{+,\underline{\eta},\bar{\epsilon}}} \mathcal{R}([-\theta,\theta]\cap\Theta_{\underline{\eta},\bar{\epsilon}}).$$

It follows that

$$\sup_{\theta\in\Theta_{\underline{\eta},\bar{\epsilon}}} R(\tilde{\delta},\theta) = \sup_{\theta\in\Theta_{+,\underline{\eta},\bar{\epsilon}}} \mathcal{R}([-\theta,\theta]\cap\Theta_{\underline{\eta},\bar{\epsilon}}) = \lim_{n\to\infty} \mathcal{R}([-\theta_n,\theta_n]\cap\Theta_{\underline{\eta},\bar{\epsilon}}) = \alpha^*\Phi\left(-\frac{(\boldsymbol{\beta}^*)'(\boldsymbol{\beta}^*)}{\sigma\|\boldsymbol{\beta}^*\|}\right).$$

□

Step 3. $\sup_{\theta\in\Theta_{+,\underline{\eta},\bar{\epsilon}}} \mathcal{R}([-\theta,\theta]\cap\Theta_{\underline{\eta},\bar{\epsilon}}) = \omega(\epsilon^*)\Phi(-\epsilon^*/\sigma)$.

Proof. By Lemma 1.B.5,

$$\begin{aligned}\mathcal{R}([-\theta_{\epsilon^*},\theta_{\epsilon^*}]) &= \frac{L(\theta_{\epsilon^*})}{\|\boldsymbol{m}(\theta_{\epsilon^*})\|}\mathcal{R}_{\text{uni}}\left(\sigma;[-\|\boldsymbol{m}(\theta_{\epsilon^*})\|,\|\boldsymbol{m}(\theta_{\epsilon^*})\|]\right)\\ &= \frac{\omega(\epsilon^*)}{\epsilon^*}\mathcal{R}_{\text{uni}}\left(\sigma;[-\epsilon^*,\epsilon^*]\right).\end{aligned}$$

On the other hand, with $t = \max\{\frac{\underline{\eta}}{L(\theta_{\epsilon^*})}, \frac{\underline{\eta}}{\epsilon^*}\}$,

$$\begin{aligned}\mathcal{R}([-\theta_{\epsilon^*},\theta_{\epsilon^*}]\cap\Theta_{\underline{\eta},\bar{\epsilon}}) &= \mathcal{R}([-\theta_{\epsilon^*},-t\theta_{\epsilon^*}]\cup[t\theta_{\epsilon^*},\theta_{\epsilon^*}])\\ &= \frac{L(\theta_{\epsilon^*})}{\|\boldsymbol{m}(\theta_{\epsilon^*})\|}\mathcal{R}_{\text{uni}}\left(\sigma;[-\|\boldsymbol{m}(\theta_{\epsilon^*})\|,-t\|\boldsymbol{m}(\theta_{\epsilon^*})\|]\cup[t\|\boldsymbol{m}(\theta_{\epsilon^*})\|,\|\boldsymbol{m}(\theta_{\epsilon^*})\|]\right)\\ &= \frac{\omega(\epsilon^*)}{\epsilon^*}\mathcal{R}_{\text{uni}}\left(\sigma;[-\epsilon^*,-t\epsilon^*]\cup[t\epsilon^*,\epsilon^*]\right).\end{aligned}$$

Since $\epsilon^* \leq a^*\sigma$, $\mathcal{R}_{\text{uni}}(\sigma; [-\epsilon^*, \epsilon^*]) = \mathcal{R}_{\text{uni}}(\sigma; [-\epsilon^*, -t\epsilon^*] \cup [t\epsilon^*, \epsilon^*])$ by Lemma 1.B.4. Therefore, $\mathcal{R}([-\theta_{\epsilon^*}, \theta_{\epsilon^*}]) = \mathcal{R}([-\theta_{\epsilon^*}, \theta_{\epsilon^*}] \cap \Theta_{\underline{\eta},\bar{\epsilon}})$. Note that this equals $\sup_{\theta\in\Theta: L(\theta)>0, \boldsymbol{m}(\theta)\neq\mathbf{0}} \mathcal{R}([-\theta, \theta]) = \omega(\epsilon^*)\Phi(-\epsilon^*/\sigma)$ as discussed in Step 3 in Section 1.6.1.

Now, since $\theta_{\epsilon^*} \in \Theta_{+,\underline{\eta},\bar{\epsilon}}$,

$$\sup_{\theta\in\Theta_{+,\underline{\eta},\bar{\epsilon}}} \mathcal{R}([-\theta, \theta] \cap \Theta_{\underline{\eta},\bar{\epsilon}}) \geq \mathcal{R}([-\theta_{\epsilon^*}, \theta_{\epsilon^*}] \cap \Theta_{\underline{\eta},\bar{\epsilon}}).$$

However, by definition and the above result,

$$\begin{aligned}\sup_{\theta\in\Theta_{+,\underline{\eta},\bar{\epsilon}}} \mathcal{R}([-\theta, \theta] \cap \Theta_{\underline{\eta},\bar{\epsilon}}) &\leq \sup_{\theta\in\Theta_{+,\underline{\eta},\bar{\epsilon}}} \mathcal{R}([-\theta, \theta]) \\ &\leq \sup_{\theta\in\Theta: L(\theta)>0, \boldsymbol{m}(\theta)\neq\mathbf{0}} \mathcal{R}([-\theta, \theta]) = \mathcal{R}([-\theta_{\epsilon^*}, \theta_{\epsilon^*}] \cap \Theta_{\underline{\eta},\bar{\epsilon}}).\end{aligned}$$

Therefore, I obtain

$$\sup_{\theta\in\Theta_{+,\underline{\eta},\bar{\epsilon}}} \mathcal{R}([-\theta, \theta] \cap \Theta_{\underline{\eta},\bar{\epsilon}}) = \sup_{\theta\in\Theta_{+,\underline{\eta},\bar{\epsilon}}} \mathcal{R}([-\theta, \theta]) = \omega(\epsilon^*)\Phi(-\epsilon^*/\sigma).$$

□

Step 4. *$\alpha^* = \omega(\epsilon^*)$ and $\|\boldsymbol{\beta}^*\| = \epsilon^*$.*

Proof. First, with $t_n = \max\{\frac{\underline{\eta}}{L(\theta_n)}, \underline{\eta}\frac{\|\boldsymbol{m}(\theta_{\epsilon^*})\|}{\boldsymbol{m}(\theta_{\epsilon^*})'\boldsymbol{m}(\theta_n)}\}$ and $t^* = \lim_{n\to\infty} t_n = \max\{\frac{\underline{\eta}}{\alpha^*}, \underline{\eta}\frac{\|\boldsymbol{m}(\theta_{\epsilon^*})\|}{\boldsymbol{m}(\theta_{\epsilon^*})'\boldsymbol{\beta}^*}\}$,

$$\begin{aligned}&\lim_{n\to\infty} \mathcal{R}([-\theta_n, \theta_n] \cap \Theta_{\underline{\eta},\bar{\epsilon}}) \\ =& \lim_{n\to\infty} \mathcal{R}([-\theta_n, -t_n\theta_n] \cup [t_n\theta_n, \theta_n]) \\ =& \lim_{n\to\infty} \frac{L(\theta_n)}{\|\boldsymbol{m}(\theta_n)\|} \mathcal{R}_{\text{uni}}(\sigma; [-\|\boldsymbol{m}(\theta_n)\|, -t_n\|\boldsymbol{m}(\theta_n)\|] \cup [t_n\|\boldsymbol{m}(\theta_n)\|, \|\boldsymbol{m}(\theta_n)\|]) \\ =& \frac{\alpha^*}{\|\boldsymbol{\beta}^*\|} \mathcal{R}_{\text{uni}}(\sigma; [-\|\boldsymbol{\beta}^*\|, -t^*\|\boldsymbol{\beta}^*\|] \cup [t^*\|\boldsymbol{\beta}^*\|, \|\boldsymbol{\beta}^*\|]),\end{aligned}$$

where the last equality follows from the fact that the mapping $(\tau, t) \mapsto \mathcal{R}_{\text{uni}}(\sigma; [-\tau, -t\tau] \cup [t\tau, \tau])$ is continuous. By Steps 2–3, $\omega(\epsilon^*)\Phi(-\epsilon^*/\sigma) = \lim_{n\to\infty} \mathcal{R}([-\theta_n, \theta_n] \cap \Theta_{\underline{\eta},\bar{\epsilon}})$. There-

fore,

$$\begin{aligned}\omega(\epsilon^*)\Phi(-\epsilon^*/\sigma) &= \frac{\alpha^*}{\|\boldsymbol{\beta}^*\|}\mathcal{R}_{\text{uni}}\left(\sigma; [-\|\boldsymbol{\beta}^*\|, -t^*\|\boldsymbol{\beta}^*\|] \cup [t^*\|\boldsymbol{\beta}^*\|, \|\boldsymbol{\beta}^*\|]\right) \\ &\leq \frac{\alpha^*}{\|\boldsymbol{\beta}^*\|}\mathcal{R}_{\text{uni}}\left(\sigma; [-\|\boldsymbol{\beta}^*\|, \|\boldsymbol{\beta}^*\|]\right).\end{aligned}$$

Note that $\alpha_n \leq \omega(\|\boldsymbol{\beta}_n\|)$ for all $n \geq 1$ by the definition of $\omega(\cdot)$. Since $\omega(\cdot)$ is continuous by the concavity, taking the limit of both sides yields $\alpha^* \leq \omega(\|\boldsymbol{\beta}^*\|)$. It follows that

$$\omega(\epsilon^*)\Phi(-\epsilon^*/\sigma) \leq \frac{\omega(\|\boldsymbol{\beta}^*\|)}{\|\boldsymbol{\beta}^*\|}\mathcal{R}_{\text{uni}}\left(\sigma; [-\|\boldsymbol{\beta}^*\|, \|\boldsymbol{\beta}^*\|]\right).$$

On the other hand, as discussed in Step 3 in Section 1.6.1,

$$\omega(\epsilon^*)\Phi(-\epsilon^*/\sigma) = \max_{\epsilon>0}\frac{\omega(\epsilon)}{\epsilon}\mathcal{R}_{\text{uni}}\left(\sigma; [-\epsilon, \epsilon]\right) \geq \frac{\omega(\|\boldsymbol{\beta}^*\|)}{\|\boldsymbol{\beta}^*\|}\mathcal{R}_{\text{uni}}\left(\sigma; [-\|\boldsymbol{\beta}^*\|, \|\boldsymbol{\beta}^*\|]\right).$$

It follows that

$$\begin{aligned}\max_{\epsilon>0}\frac{\omega(\epsilon)}{\epsilon}\mathcal{R}_{\text{uni}}\left(\sigma; [-\epsilon, \epsilon]\right) &= \frac{\alpha^*}{\|\boldsymbol{\beta}^*\|}\mathcal{R}_{\text{uni}}\left(\sigma; [-\|\boldsymbol{\beta}^*\|, -t^*\|\boldsymbol{\beta}^*\|] \cup [t^*\|\boldsymbol{\beta}^*\|, \|\boldsymbol{\beta}^*\|]\right) \\ &= \frac{\alpha^*}{\|\boldsymbol{\beta}^*\|}\mathcal{R}_{\text{uni}}\left(\sigma; [-\|\boldsymbol{\beta}^*\|, \|\boldsymbol{\beta}^*\|]\right) \\ &= \frac{\omega(\|\boldsymbol{\beta}^*\|)}{\|\boldsymbol{\beta}^*\|}\mathcal{R}_{\text{uni}}\left(\sigma; [-\|\boldsymbol{\beta}^*\|, \|\boldsymbol{\beta}^*\|]\right).\end{aligned}$$

Therefore, $\alpha^* = \omega(\|\boldsymbol{\beta}^*\|)$ and $\|\boldsymbol{\beta}^*\| \in \arg\max_{\epsilon>0}\frac{\omega(\epsilon)}{\epsilon}\mathcal{R}_{\text{uni}}\left(\sigma; [-\epsilon, \epsilon]\right)$. Furthermore,

$$\mathcal{R}_{\text{uni}}\left(\sigma; [-\|\boldsymbol{\beta}^*\|, -t^*\|\boldsymbol{\beta}^*\|] \cup [t^*\|\boldsymbol{\beta}^*\|, \|\boldsymbol{\beta}^*\|]\right) = \mathcal{R}_{\text{uni}}\left(\sigma; [-\|\boldsymbol{\beta}^*\|, \|\boldsymbol{\beta}^*\|]\right). \tag{B.11}$$

If it is shown that $\|\boldsymbol{\beta}^*\| \leq a^*\sigma$, then

$$\|\boldsymbol{\beta}^*\| \in \arg\max_{0<\epsilon\leq a^*\sigma}\frac{\omega(\epsilon)}{\epsilon}\mathcal{R}_{\text{uni}}\left(\sigma; [-\epsilon, \epsilon]\right) = \arg\max_{0<\epsilon\leq a^*\sigma}\omega(\epsilon)\Phi(-\epsilon/\sigma) = \{\epsilon^*\}.$$

Suppose to the contrary that $\|\boldsymbol{\beta}^*\| > a^*\sigma$. By inspection of the form of $\mathcal{R}_{\text{uni}}$ given in Lemma 1.B.4, it is necessary that $t^*\|\boldsymbol{\beta}^*\| \leq a^*\sigma$ for Eq. (B.11) to hold. Therefore, $t^* \leq \frac{a^*\sigma}{\|\boldsymbol{\beta}^*\|} < 1$ since $\|\boldsymbol{\beta}^*\| > a^*\sigma$. It follows that $t^* = \max\{\frac{\eta}{\alpha^*}, \underline{\eta}\frac{\|\boldsymbol{m}(\theta_{\epsilon^*})\|}{\boldsymbol{m}(\theta_{\epsilon^*})'\boldsymbol{\beta}^*}\} < 1$, so $\underline{\eta} < \frac{\boldsymbol{m}(\theta_{\epsilon^*})'\boldsymbol{\beta}^*}{\|\boldsymbol{m}(\theta_{\epsilon^*})\|}$. Note also

that

$$\alpha^* > \alpha^* \Phi\left(-\frac{\|\boldsymbol{\beta}^*\|}{\sigma}\right) = \omega(\epsilon^*)\Phi(-\epsilon^*/\sigma) > \underline{\eta},$$

where the equality follows from Steps 2–3, and the last inequality by the choice of $\underline{\eta}$. We can pick $\underline{t} \in (0,1)$ sufficiently close to 1 so that for all $t \in [\underline{t}, 1]$, $\underline{\eta} < \frac{\boldsymbol{m}(\theta_{\epsilon^*})' t\boldsymbol{\beta}^*}{\|\boldsymbol{m}(\theta_{\epsilon^*})\|} = \lim_{n\to\infty} \frac{\boldsymbol{m}(\theta_{\epsilon^*})'\boldsymbol{m}(t\theta_n)}{\|\boldsymbol{m}(\theta_{\epsilon^*})\|}$ and $\underline{\eta} < t\alpha^* = \lim_{n\to\infty} L(t\theta_n)$. It follows that $t\boldsymbol{\gamma}^* = (t\alpha^*, t(\boldsymbol{\beta}^*)')' = \lim_{n\to\infty}(L(t\theta_n), \boldsymbol{m}(t\theta_n)')' \in \bar{\Gamma}_{+,\underline{\eta},\bar{\epsilon}}$ for all $t \in [\underline{t}, 1]$. Since $\boldsymbol{\gamma}^* \in \arg\max_{\boldsymbol{\gamma}=(\alpha,\boldsymbol{\beta}')'\in\bar{\Gamma}_{+,\underline{\eta},\bar{\epsilon}}} \alpha\Phi\left(-\frac{(\boldsymbol{\beta}^*)'\boldsymbol{\beta}}{\sigma\|\boldsymbol{\beta}^*\|}\right)$, we have

$$\boldsymbol{\gamma}^* \in \arg\max_{\boldsymbol{\gamma}\in[\underline{t}\boldsymbol{\gamma}^*, \boldsymbol{\gamma}^*]} \alpha\Phi\left(-\frac{(\boldsymbol{\beta}^*)'\boldsymbol{\beta}}{\sigma\|\boldsymbol{\beta}^*\|}\right).$$

This implies that

$$1 \in \arg\max_{t\in[\underline{t},1]} t\alpha^*\Phi\left(-\frac{(\boldsymbol{\beta}^*)' t\boldsymbol{\beta}^*}{\sigma\|\boldsymbol{\beta}^*\|}\right) = \arg\max_{t\in[\underline{t},1]} t\Phi\left(-\frac{t\|\boldsymbol{\beta}^*\|}{\sigma}\right).$$

By Lemma 1.B.1, $t \mapsto t\Phi\left(-\frac{t\|\boldsymbol{\beta}^*\|}{\sigma}\right)$ is strictly increasing on $[0, \frac{a^*\sigma}{\|\boldsymbol{\beta}^*\|})$ and strictly decreasing on $(\frac{a^*\sigma}{\|\boldsymbol{\beta}^*\|}, \infty)$. Since $\frac{a^*\sigma}{\|\boldsymbol{\beta}^*\|} < 1$ by the hypothesis, $1 \notin \arg\max_{t\in[\underline{t},1]} t\Phi\left(-\frac{t\|\boldsymbol{\beta}^*\|}{\sigma}\right)$, which is a contradiction. □

Step 5. $\boldsymbol{m}(\theta_{\epsilon^*}) = \boldsymbol{\beta}^*$.

Proof. Suppose that $\boldsymbol{m}(\theta_{\epsilon^*}) \neq \boldsymbol{\beta}^*$. Pick any $\lambda \in (0,1)$, and consider the sequence $\{\theta_{\lambda,n}\}_{n=1}^{\infty}$, where $\theta_{\lambda,n} = \lambda\theta_{\epsilon^*} + (1-\lambda)\theta_n$. Since $\Theta_{+,\underline{\eta},\bar{\epsilon}}$ is convex, $\theta_{\lambda,n} \in \Theta_{+,\underline{\eta},\bar{\epsilon}}$ for all n. We have

$$\lim_{n\to\infty} L(\theta_{\lambda,n}) = \lambda L(\theta_{\epsilon^*}) + (1-\lambda)\lim_{n\to\infty} L(\theta_n) = \lambda\omega(\epsilon^*) + (1-\lambda)\alpha^* = \omega(\epsilon^*),$$

and

$$\lim_{n\to\infty} \|\boldsymbol{m}(\theta_{\lambda,n})\| = \|\lambda\boldsymbol{m}(\theta_{\epsilon^*}) + (1-\lambda)\lim_{n\to\infty} \boldsymbol{m}(\theta_n)\| = \|\lambda\boldsymbol{m}(\theta_{\epsilon^*}) + (1-\lambda)\boldsymbol{\beta}^*\| < \epsilon^*,$$

where the last inequality holds since $\{\boldsymbol{\beta} \in \mathbb{R}^n : \|\boldsymbol{\beta}\| \leq \epsilon^*\}$ is strictly convex. These imply that

$$\omega(\tilde{\epsilon}) = \sup\{L(\theta) : \theta \in \Theta, \|\boldsymbol{m}(\theta)\| \leq \tilde{\epsilon}\} = \omega(\epsilon^*)$$

for any $\tilde{\epsilon} \in (\lim_{n\to\infty} \|\boldsymbol{m}(\theta_{\lambda,n})\|, \epsilon^*)$. It follows that

$$\omega(\tilde{\epsilon})\Phi(-\tilde{\epsilon}/\sigma) = \omega(\epsilon^*)\Phi(-\tilde{\epsilon}/\sigma) > \omega(\epsilon^*)\Phi(-\epsilon^*/\sigma),$$

which contradicts the fact that ϵ^* maximizes $\omega(\epsilon)\Phi(-\epsilon/\sigma)$ over $[0, a^*\sigma]$. □

Step 6. $\theta_{\epsilon^*} \in \arg\max_{\theta\in\Theta: L(\theta)>0} L(\theta)\Phi\left(-\frac{\boldsymbol{m}(\theta_{\epsilon^*})'\boldsymbol{m}(\theta)}{\sigma\|\boldsymbol{m}(\theta_{\epsilon^*})\|}\right)$.

Proof. Since $\alpha^* = L(\theta_{\epsilon^*})$, $\boldsymbol{\beta}^* = \boldsymbol{m}(\theta_{\epsilon^*})$, and $\boldsymbol{\gamma}^* \in \arg\max_{\boldsymbol{\gamma}=(\alpha,\boldsymbol{\beta}')'\in\bar{\Gamma}_{+,\underline{\eta},\bar{\epsilon}}} \alpha\Phi\left(-\frac{(\boldsymbol{\beta}^*)'\boldsymbol{\beta}}{\sigma\|\boldsymbol{\beta}^*\|}\right)$,

$$\boldsymbol{\gamma}^* = (L(\theta_{\epsilon^*}), \boldsymbol{m}(\theta_{\epsilon^*})')' \in \arg\max_{\boldsymbol{\gamma}=(\alpha,\boldsymbol{\beta}')'\in\bar{\Gamma}_{+,\underline{\eta},\bar{\epsilon}}} \alpha\Phi\left(-\frac{\boldsymbol{m}(\theta_{\epsilon^*})'\boldsymbol{\beta}}{\sigma\|\boldsymbol{m}(\theta_{\epsilon^*})\|}\right), \tag{B.12}$$

which implies

$$\theta_{\epsilon^*} \in \arg\max_{\theta\in\Theta_{+,\underline{\eta},\bar{\epsilon}}} L(\theta)\Phi\left(-\frac{\boldsymbol{m}(\theta_{\epsilon^*})'\boldsymbol{m}(\theta)}{\sigma\|\boldsymbol{m}(\theta_{\epsilon^*})\|}\right).$$

Pick any $\theta \in \Theta$ such that $L(\theta) > 0$ and $\theta \notin \Theta_{+,\underline{\eta},\bar{\epsilon}}$. Let $\boldsymbol{\gamma} = (\alpha, \boldsymbol{\beta}')' = (L(\theta), \boldsymbol{m}(\theta)')'$. Since $\underline{\eta} < \omega(\epsilon^*) = \alpha^*$ and $\underline{\eta} < \epsilon^* < \bar{\epsilon}$ by the choice of $\underline{\eta}$ and $\bar{\epsilon}$, we can pick $t \in (0,1)$ sufficiently close to 1 so that

$$L((1-t)\theta + t\theta_{\epsilon^*}) = (1-t)\alpha + t\alpha^* > \underline{\eta},$$

$$\frac{\boldsymbol{m}(\theta_{\epsilon^*})'\boldsymbol{m}((1-t)\theta + t\theta_{\epsilon^*})}{\|\boldsymbol{m}(\theta_{\epsilon^*})\|} = \frac{\boldsymbol{m}(\theta_{\epsilon^*})'((1-t)\boldsymbol{\beta} + t\boldsymbol{\beta}^*)}{\|\boldsymbol{m}(\theta_{\epsilon^*})\|} = (1-t)\frac{\boldsymbol{m}(\theta_{\epsilon^*})'\boldsymbol{\beta}}{\|\boldsymbol{m}(\theta_{\epsilon^*})\|} + t\epsilon^* > \underline{\eta},$$

and

$$\|\boldsymbol{m}((1-t)\theta + t\theta_{\epsilon^*})\| = \|(1-t)\boldsymbol{\beta} + t\boldsymbol{\beta}^*\| \le (1-t)\|\boldsymbol{\beta}\| + t\epsilon^* < \bar{\epsilon}.$$

It follows that $(1-t)\theta + t\theta_{\epsilon^*} \in \Theta_{+,\underline{\eta},\bar{\epsilon}}$ and $(1-t)\boldsymbol{\gamma} + t\boldsymbol{\gamma}^* \in \Gamma_{+,\underline{\eta},\bar{\epsilon}}$. By Eq. (B.12), this implies that

$$\alpha^*\Phi\left(-\frac{\boldsymbol{m}(\theta_{\epsilon^*})'\boldsymbol{\beta}^*}{\sigma\|\boldsymbol{m}(\theta_{\epsilon^*})\|}\right) \ge [(1-t)\alpha + t\alpha^*]\Phi\left(-(1-t)\frac{\boldsymbol{m}(\theta_{\epsilon^*})'\boldsymbol{\beta}}{\sigma\|\boldsymbol{m}(\theta_{\epsilon^*})\|} - t\frac{\boldsymbol{m}(\theta_{\epsilon^*})'\boldsymbol{\beta}^*}{\sigma\|\boldsymbol{m}(\theta_{\epsilon^*})\|}\right).$$

Since the function $(a,b) \mapsto a\Phi(-b)$ is strictly quasi-concave on $(0,\infty)\times\mathbb{R}$ by Lemma 1.B.3,

$$\begin{aligned}&[(1-t)\alpha + t\alpha^*]\Phi\left(-(1-t)\frac{\boldsymbol{m}(\theta_{\epsilon^*})'\boldsymbol{\beta}}{\sigma\|\boldsymbol{m}(\theta_{\epsilon^*})\|} - t\frac{\boldsymbol{m}(\theta_{\epsilon^*})'\boldsymbol{\beta}^*}{\sigma\|\boldsymbol{m}(\theta_{\epsilon^*})\|}\right)\\ &> \min\left\{\alpha\Phi\left(-\frac{\boldsymbol{m}(\theta_{\epsilon^*})'\boldsymbol{\beta}}{\sigma\|\boldsymbol{m}(\theta_{\epsilon^*})\|}\right), \alpha^*\Phi\left(-\frac{\boldsymbol{m}(\theta_{\epsilon^*})'\boldsymbol{\beta}^*}{\sigma\|\boldsymbol{m}(\theta_{\epsilon^*})\|}\right)\right\}.\end{aligned}$$

Therefore,

$$\alpha^* \Phi\left(-\frac{\boldsymbol{m}(\theta_{\epsilon^*})'\boldsymbol{\beta^*}}{\sigma\|\boldsymbol{m}(\theta_{\epsilon^*})\|}\right) > \alpha\Phi\left(-\frac{\boldsymbol{m}(\theta_{\epsilon^*})'\boldsymbol{\beta}}{\sigma\|\boldsymbol{m}(\theta_{\epsilon^*})\|}\right),$$

which implies that

$$L(\theta_{\epsilon^*})\Phi\left(-\frac{\boldsymbol{m}(\theta_{\epsilon^*})'\boldsymbol{m}(\theta_{\epsilon^*})}{\sigma\|\boldsymbol{m}(\theta_{\epsilon^*})\|}\right) > L(\theta)\Phi\left(-\frac{\boldsymbol{m}(\theta_{\epsilon^*})'\boldsymbol{m}(\theta)}{\sigma\|\boldsymbol{m}(\theta_{\epsilon^*})\|}\right).$$

The conclusion then follows.

□

1.B.9 Proof of Lemma 1.5

Let θ_ϵ attain the modulus of continuity at $\epsilon \in [0, \bar{\epsilon}]$. Also, let $\iota \in \Theta$ be a parameter value that satisfies Assumption 1.1(c). Without loss of generality, I normalize ι so that $L(\iota) = 1$.

First, Assumption 1.3(a) holds under Assumption 1.1(a) and (c), since if $\|\boldsymbol{m}(\theta_\epsilon)\| < \epsilon$, then there exists $c > 0$ such that $\theta_\epsilon + c\iota \in \Theta$, $L(\theta_\epsilon + c\iota) > L(\theta_\epsilon)$, and $\|\boldsymbol{m}(\theta_\epsilon + c\iota)\| < \epsilon$, which contradicts the definition of θ_ϵ.

Under Assumption 1.1(b), θ_0 satisfies $L(\theta_0) = \omega(0) = \rho(0)$ as shown in the proof of Lemma 1.7. Additionally, $(\boldsymbol{w}^*)'\boldsymbol{m}(\theta_0) = 0$ by construction. Applying Lemma 1.A.2, we have that $\rho'(0) = \frac{1}{(\boldsymbol{w}^*)'\boldsymbol{m}(\iota)}$.

By Lemma 1.A.1, for any sufficiently small $\epsilon > 0$, $\omega'(\epsilon) = \frac{\epsilon}{\boldsymbol{m}(\iota)'\boldsymbol{m}(\theta_\epsilon)}$. Since $\omega(\cdot)$ is differentiable and concave, it is continuously differentiable. Therefore,

$$\omega'(0) = \lim_{\epsilon \to 0} \omega'(\epsilon) = \lim_{\epsilon \to 0} \frac{1}{\boldsymbol{m}(\iota)'\boldsymbol{m}(\theta_\epsilon/\epsilon)} = \frac{1}{\boldsymbol{m}(\iota)'\boldsymbol{w}^*} = \rho'(0),$$

where the second last equality holds by Assumption 1.1(b) and the fact that $\|\boldsymbol{m}(\theta_\epsilon)\| = \epsilon$. Since $\omega'(0)$ is nonnegative by the definition of the modulus and $\boldsymbol{m}(\iota)'\boldsymbol{w}^*$ is finite, $\omega'(0)$ and $\rho'(0)$ must be positive.

Let $\sigma^* = 2\phi(0)\frac{\rho(0)}{\rho'(0)} = 2\phi(0)\frac{\omega(0)}{\omega'(0)}$ and $g(\epsilon) = \rho(\epsilon)\Phi\left(-\frac{\epsilon}{\sigma^*}\right)$. By differentiating $g(\cdot)$, we have

$$g'(\epsilon) = \rho'(\epsilon)\Phi\left(-\frac{\epsilon}{\sigma^*}\right) - \rho(\epsilon)\phi\left(-\frac{\epsilon}{\sigma^*}\right)/\sigma^*.$$

$g'(0) = 0$ by the choice of σ^*. Since $\rho(\cdot)$ is concave (as shown in the proof of Lemma 1.A.2) and differentiable, $\rho'(\cdot)$ is continuous. Let $\epsilon_1 = \inf\{(\boldsymbol{w}^*)'\boldsymbol{m}(\theta) : \theta \in \Theta\}$, $\epsilon_1' = \sup\{\epsilon : \rho'(\epsilon) \geq 0\}$, and $\epsilon_2' = \sup\{\epsilon : \rho(\epsilon) \geq 0\}$. Since $\rho(\cdot)$ is concave and hence $\rho'(\cdot)$ is nonincreasing, $\rho(\cdot)$ is nondecreasing on $(\epsilon_1, \epsilon_1']$. Also, since $\rho'(0) > 0$, we have $\epsilon_1' > 0$, and $\rho(\cdot)$ is nonconstant on $(\epsilon_1, \epsilon_1']$. By Lemma 1.B.1, $g(\cdot)$ is maximized at 0 over $(\epsilon_1, \epsilon_1']$. For $\epsilon \in (\epsilon_1', \epsilon_2']$, $\rho'(\epsilon) \leq 0$ and $\rho(\epsilon) \geq 0$, so that $g'(\epsilon) \leq 0$. Therefore, $g(\cdot)$ is maximized at 0 over $(\epsilon_1, \epsilon_2']$. Finally, $g(\epsilon) \leq 0$ for all $\epsilon > \epsilon_2'$. Since $g(0) = \rho(0)/2 = \omega(0)/2 \geq 0$, $g(\cdot)$ is maximized at 0 globally. □

1.B.10 Proof of Lemma 1.6

Since $\boldsymbol{Y} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_n)$ under θ for all $\theta \in [-\bar{\theta}, \bar{\theta}]$, $\mathbb{E}_\theta[\delta(\boldsymbol{Y})] = \mathbb{E}_{\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)}[\delta(\boldsymbol{Y})]$ is constant over $\theta \in [-\bar{\theta}, \bar{\theta}]$. Therefore, the maximum regret of decision rule δ is

$$\sup_{\theta \in [-\bar{\theta}, \bar{\theta}]} R(\delta, \theta) = \begin{cases} L(\bar{\theta})(1 - \mathbb{E}_{\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)}[\delta(\boldsymbol{Y})]) & \text{if } \mathbb{E}_{\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)}[\delta(\boldsymbol{Y})] < 1/2, \\ L(\bar{\theta})/2 & \text{if } \mathbb{E}_{\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)}[\delta(\boldsymbol{Y})] = 1/2, \\ (-L(-\bar{\theta}))\mathbb{E}_{\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)}[\delta(\boldsymbol{Y})] & \text{if } \mathbb{E}_{\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)}[\delta(\boldsymbol{Y})] > 1/2. \end{cases}$$

Thus, any decision rule δ^* such that $\mathbb{E}_{\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)}[\delta^*(\boldsymbol{Y})] = \frac{1}{2}$ is minimax regret. The minimax risk is given by

$$\mathcal{R}(\sigma; [-\bar{\theta}, \bar{\theta}]) = \frac{L(\bar{\theta})}{2}.$$

□

1.B.11 Proof of Lemma 1.7

The maximum regret of δ^* over Θ is

$$\begin{aligned} \sup_{\theta \in \Theta} R(\delta^*, \theta) &= \sup_{\theta \in \Theta} L(\theta) \Phi\left(-\frac{(\boldsymbol{w}^*)'\boldsymbol{m}(\theta)}{\sigma^*}\right) \\ &= \sup_{\epsilon \in \mathbb{R}} \sup_{\theta \in \Theta : (\boldsymbol{w}^*)'\boldsymbol{m}(\theta) = \epsilon} L(\theta) \Phi\left(-\frac{\epsilon}{\sigma^*}\right) \\ &= \sup_{\epsilon \in \mathbb{R}} \rho(\epsilon) \Phi\left(-\frac{\epsilon}{\sigma^*}\right) \\ &= \frac{1}{2}\rho(0), \end{aligned}$$

where the last equality holds by Assumption 1.3(c).

Since $(\boldsymbol{w}^*)'\boldsymbol{m}(\theta) = 0$ for any $\theta \in \Theta$ such that $\boldsymbol{m}(\theta) = \mathbf{0}$, it follows that $\rho(0) \geq \omega(0)$ by the definition of $\omega(\cdot)$ and $\rho(\cdot)$. If $\rho(0) = \omega(0)$, then $\sup_{\theta\in\Theta} R(\delta^*, \theta)$ is attained at θ_0, and $\sup_{\theta\in\Theta} R(\delta^*, \theta) = \frac{1}{2}\omega(0)$. Below, I show that $\rho(0) = \omega(0)$. Suppose to the contrary that $\rho(0) > \omega(0)$. Then there exists $\theta \in \Theta$ such that $(\boldsymbol{w}^*)'\boldsymbol{m}(\theta) = 0$, $\boldsymbol{m}(\theta) \neq \mathbf{0}$, and $L(\theta) > \omega(0)$. For $\epsilon \in (0, \bar{\epsilon}]$, by the Cauchy-Schwarz inequality,

$$\epsilon^{-1}\left|\frac{\boldsymbol{m}(\theta_\epsilon)'}{\epsilon}\boldsymbol{m}(\theta)\right| = \epsilon^{-1}\left|\left(\frac{\boldsymbol{m}(\theta_\epsilon)}{\|\boldsymbol{m}(\theta_\epsilon)\|} - \boldsymbol{w}^*\right)'\boldsymbol{m}(\theta)\right| \leq \epsilon^{-1}\left\|\frac{\boldsymbol{m}(\theta_\epsilon)}{\|\boldsymbol{m}(\theta_\epsilon)\|} - \boldsymbol{w}^*\right\|\|\boldsymbol{m}(\theta)\|,$$

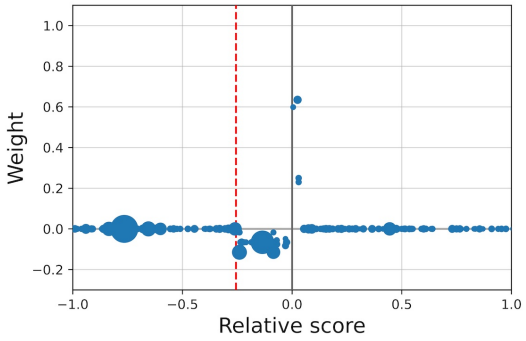
where θ_ϵ attains the modulus of continuity at ϵ, and the first equality follows from Assumption 1.3(a) and the fact that $(\boldsymbol{w}^*)'\boldsymbol{m}(\theta) = 0$. The right-hand side converges to zero as $\epsilon \to 0$ by Assumption 1.3(b). Also, $\omega(\cdot)$ is concave and hence is continuous. Therefore, $\epsilon^{-1}\frac{\boldsymbol{m}(\theta_\epsilon)'}{\epsilon}\boldsymbol{m}(\theta) \leq 1/2$ and $L(\theta) > \omega(\epsilon)$ for any sufficiently small $\epsilon > 0$. Pick such an $\epsilon > 0$, and let $\theta_\lambda = \lambda\theta_\epsilon + (1-\lambda)\theta$ for $\lambda \in \mathbb{R}$. By simple algebra,

$$\begin{aligned}
\|\boldsymbol{m}(\theta_\lambda)\|^2 &= \lambda^2\|\boldsymbol{m}(\theta_\epsilon)\|^2 + 2\lambda(1-\lambda)\boldsymbol{m}(\theta_\epsilon)'\boldsymbol{m}(\theta) + (1-\lambda)^2\|\boldsymbol{m}(\theta)\|^2 \\
&\leq \lambda^2\epsilon^2 + \lambda(1-\lambda)\epsilon^2 + (1-\lambda)^2\|\boldsymbol{m}(\theta)\|^2 \\
&= \|\boldsymbol{m}(\theta)\|^2\lambda^2 - (2\|\boldsymbol{m}(\theta)\|^2 - \epsilon^2)\lambda + \|\boldsymbol{m}(\theta)\|^2.
\end{aligned}$$

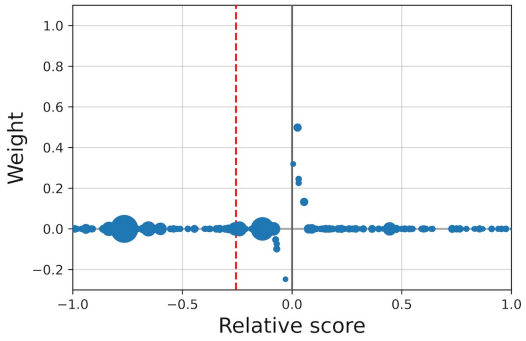
Observe that the right-hand side is quadratic in λ, minimized at $\lambda = \frac{2\|\boldsymbol{m}(\theta)\|^2 - \epsilon^2}{2\|\boldsymbol{m}(\theta)\|^2} < 1$, and equal to ϵ^2 when $\lambda = 1$. This implies that $\|\boldsymbol{m}(\theta_\lambda)\|^2 < \epsilon^2$ for any λ close to one. However, $L(\theta_\lambda) = \lambda L(\theta_\epsilon) + (1-\lambda)L(\theta) > \omega(\epsilon)$ for all $\lambda \in (0,1)$, which contradicts the assumption that θ_ϵ attains the modulus of continuity at ϵ. □

1.C Empirical Policy Application: Additional Figures

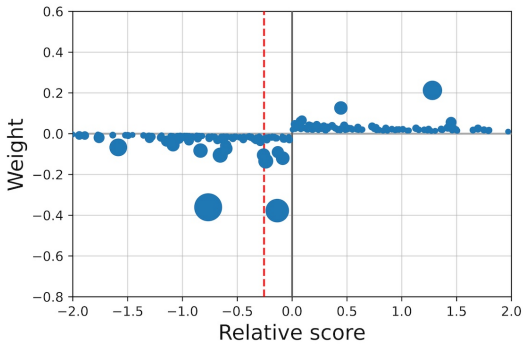
Figure 1.8: Weight to Each Village Attached by Plug-in Rules



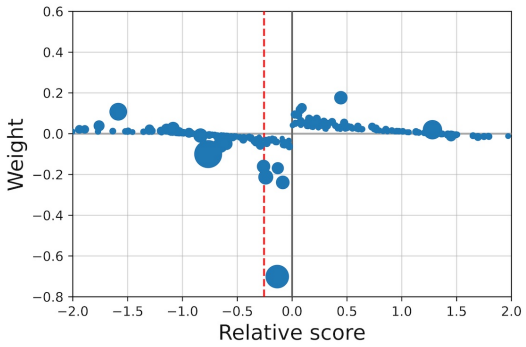
(a) Plug-in MSE Rule, $C = 0.5$



(b) Plug-in MSE Rule Under Constant Effect, $C = 0.5$



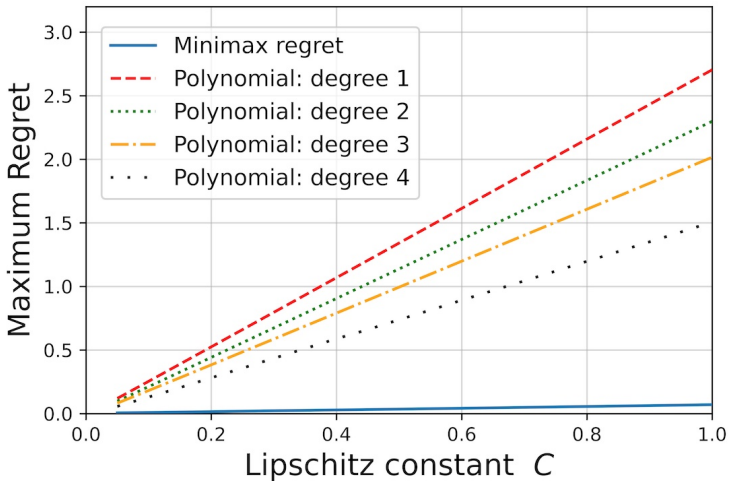
(c) Polynomial of Degree 2



(d) Polynomial of Degree 4

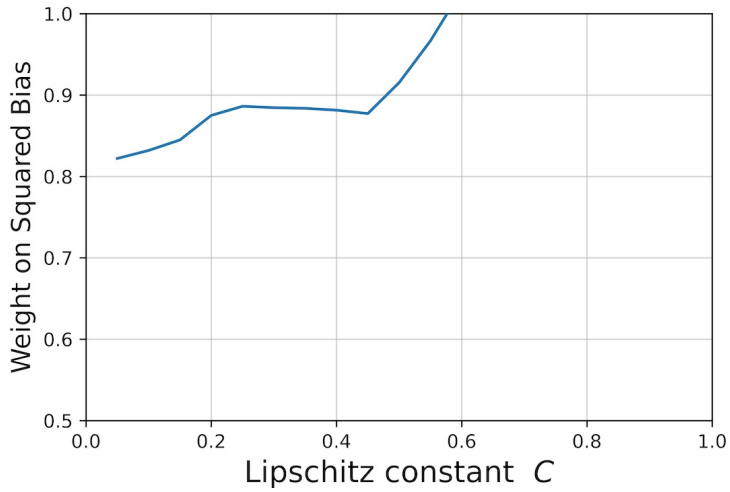
Notes: This figure shows the weight w_i attached to each village by the plug-in decision rules of the form $\delta(\boldsymbol{Y}) = \mathbf{1}\{\sum_{i=1}^n w_i Y_i \geq 0\}$. The weights are normalized so that $\sum_{i=1}^n w_i^2 = 1$. The horizontal axis indicates the relative score of each village. Each circle corresponds to each village. The size of circles is proportional to the inverse of the standard error of the enrollment rate Y_i. The vertical dashed line corresponds to the new cutoff -0.256. Panels (a) and (b) show the results for the plug-in rules based on the linear minimax MSE estimators with or without the assumption of constant conditional treatment effects when the Lipschitz constant C is 0.5. Panels (c) and (d) show the results for the plug-in rules based on the polynomial regression estimators of degrees 2 and 4, respectively.

Figure 1.9: Maximum Regret of Minimax Regret Rule and Plug-in Rules Based on Polynomial Regression Estimators



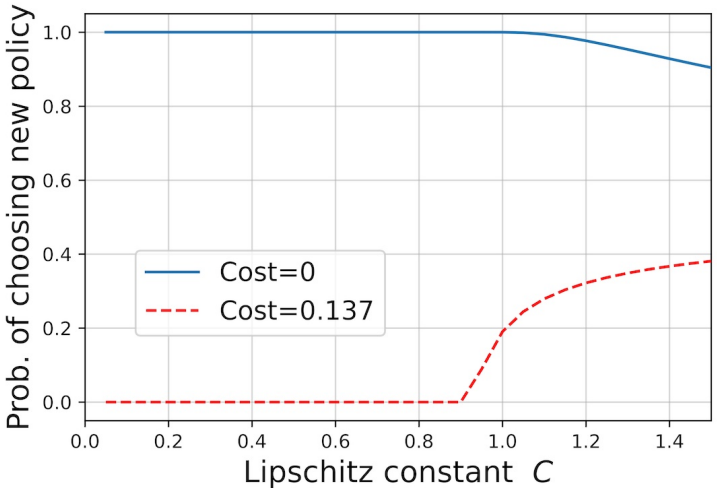
Notes: This figure shows the maximum regret of the minimax regret rule and the plug-in rules based on the polynomial regression estimators of degrees 1 to 4. The maximum regret is computed by setting the true function class of the counterfactual outcome function to the Lipschitz class. The maximum regret is normalized so that the unit is the same as that of the enrollment rate. I report the results for the range $[0.05, 0.1, ..., 0.95, 1]$ of the Lipschitz constant C.

Figure 1.10: Weight on Bias Placed by Minimax Regret Rule

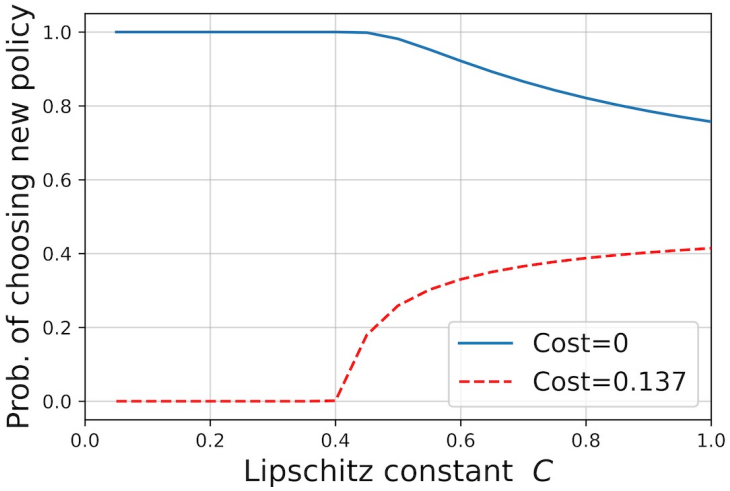


Notes: This figure shows the weight $\alpha \in [1/2, 1]$ on the squared worst-case bias placed by the minimax regret rule of the form $\delta^*(\boldsymbol{Y}) = \mathbf{1}\{\tilde{\boldsymbol{w}}'\boldsymbol{Y} \geq 0\}$, where $\tilde{\boldsymbol{w}} \in \arg\min_{\boldsymbol{w}\in\mathbb{R}^n} \left\{\alpha \cdot \left(\sup_{f\in\mathcal{F}_{\mathrm{Lip}}(C)} \mathbb{E}_f[\boldsymbol{w}'\boldsymbol{Y} - L(f)]\right)^2 + (1-\alpha)\cdot \mathrm{Var}(\boldsymbol{w}'\boldsymbol{Y})\right\}$. I only report the results for the Lipschitz constant $C < 0.6$ since the minimax regret rule is randomized for $C \geq 0.6$.

Figure 1.11: Optimal Decisions for Alternative New Policies



(a) New Policy of Constructing Schools in 10% of Villages



(b) New Policy of Constructing Schools in 30% of Villages

Notes: This figure shows the probability of choosing the new policy computed by the minimax regret rule. The new policy is to construct BRIGHT schools in previously ineligible villages whose relative scores are in the top 10% (Panel (a)) or in the top 30% (Panel (b)). The solid line shows the results for the scenario where we ignore the policy cost. The dashed line shows the results for the scenario where the policy cost measured in the unit of the enrollment rate is 0.137. I report the results for the range $[0.05, 0.1, ..., 1.45, 1.5]$ of the Lipschitz constant C in Panel (a) and for the range $[0.05, 0.1, ..., 0.95, 1]$ in Panel (b).

Chapter 2

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

Joint with Yusuke Narita

2.1 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 is 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 assigned independently 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 IV may produce stratified randomization, regression-discontinuity-style local variation, or some combination of the two.

This chapter shows how to use data obtained from algorithmic decision making to identify and estimate causal effects. In our framework, the analyst observes a random iid 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. This setup nests the classic propensity-score and regression-discontinuity-design (RDD) setups.

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 of our estimator is as follows.[1]

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 $B(X_i, \delta)$.[2] This $p^s(X_i; \delta)$ is a simulation-based approximation to the Approximate Propensity Score $p^A(x)$.

1. Code implementing this procedure in Python, R, and Stata is available at https://github.com/rfgong/IVaps

2. 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. To eliminate selection bias, we suggest selecting a bandwidth value small enough to balance a large set of pre-treatment covariates between groups with and without treatment recommendations. We also suggest that the analyst considers several different values and check if the 2SLS estimates are robust to bandwidth changes, as we often do in RDD applications.

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 $\hat{\beta}_1^s$ be the estimated coefficient on D_i.

As the main theoretical result, we prove the 2SLS estimator $\hat{\beta}_1^s$ 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 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 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 runs an algorithmic rule on hospital characteristics 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 funding eligibility status as an IV for funding amounts, while controlling for the Approximate Propensity Score induced by the eligibility-determining algorithm. The funding eligibility IV boosts the funding amount by about $14 million on average.

The resulting 2SLS estimates with Approximate-Propensity-Score controls 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 controlling for hospital characteristics) or 2SLS estimates with no controls. The OLS estimates, for example, imply that a $1 million increase in funding allows hospitals to accommodate 5.58 more COVID-19 patients. The uncontrolled 2SLS estimates produce similar, slightly smaller effects (3.25 more patients per $1 million of funding). In contrast, the 2SLS estimates with Approximate-Propensity-Score controls show no or even negative effects (from 1.03 to 4.08 *less* patients for every $1 million of funding).

The null effect of funding also turns out to persist several months after the distribution of funding. We also find no clear heterogeneity in the null funding effect across different subgroups of hospitals. 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.[3]

3. 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.

Related Literature

Theoretically, our framework integrates the classic propensity-score (selection-on-observables) scenario with a multidimensional extension of the fuzzy 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 (2021) provide an estimator for a related setting with perfect compliance.[4]

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 supervised learning (Cowgill, 2018; Bundorf, Polyakova and Tai-Seale, 2019), surge pricing (Cohen, Hahn, Hall, Levitt and Metcalfe, 2016), bandit (Li, Chu, Langford and Schapire, 2010), reinforcement learning (Precup, 2000), and market-design algorithms (Abdulkadiroğlu, Angrist, Narita and Pathak, 2017, 2022; Abdulkadiroğlu, 2013; Kawai, Nakabayashi, Ortner and Chassang, 2022; 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.[5]

When we specialize 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 bound-

4. Building on their prior work (Armstrong and Kolesár, 2018), Armstrong and Kolesár (2021) 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 valid inference.

5. A focal group of decision-making algorithms are machine learning algorithms, as illustrated in our machine-learning simulation in Section 2.5. 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 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; Mullainathan and Spiess, 2017). We supplement these studies by highlighting the role of machine learning as a data-production tool. This chapter also has a conceptual connection to the heated conversation about whether algorithmic decisions are better than human decisions (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.

ary 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). 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).

A popular approach to the two-dimensional RDD is to use the shortest (Euclidean) distance from a point to the boundary as a univariate running variable and apply a univariate RDD method (Black, 1999; Wong, Steiner and Cook, 2013). However, there appears to be no established general approaches to computing the shortest distance for arbitrary covariate spaces and arbitrary decision boundaries. Our method circumvents the difficulty; computing the Approximate Propensity Score is feasible for any problem, since it only requires simulating the decision-making algorithm.[6]

The Approximate Propensity Score developed in this chapter shares its spirit with the local random assignment interpretation of the RDD, discussed by Cattaneo, Frandsen and Titiunik (2015), Cattaneo, Titiunik and Vazquez-Bare (2017), Frandsen (2017), Sekhon and Titiunik (2017), Frölich and Huber (2019), Abdulkadiroğlu *et al.* (2022) and Eckles, Ignatiadis, Wager and Wu (2020). These papers consider settings that fit into this chapter's framework.

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 is a healthcare crisis, so our work complements prior work on more normal situations. Our analysis also exploits rule-based locally random assignment of cash flows

6. Another common approach to the two-dimensional RDD is to first estimate the conditional average treatment effect $E[Y_i(1) - Y_i(0)|X_i = x]$ for a large number of boundary points x (either by the univariate local polynomial regression using the distance to the point x as a univariate covariate or by the bivariate local polynomial regression). It then computes a weighted average of the estimated conditional average treatment effects over the boundary (Zajonc, 2012; Keele and Titiunik, 2015). However, identifying boundary points from a general decision algorithm itself is hard unless it has a known analytical form. In addition, even if we can trace out the boundary, it is not straightforward to select a grid of points along the boundary.

to hospitals. This feature provides our estimates with additional confidence in their causal interpretation.

2.2 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.[7] $Y_i(1)$ and $Y_i(0)$ denote potential outcomes when the individual is treated and not treated, respectively.

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 2.1 (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

7. 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).

settings, this framework nests many other situations, such as multidimensional RDDs and complex machine learning and market-design algorithms, as illustrated in Sections 2.5-2.7.

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 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 health 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 2.1 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 2.A.2 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 2.1.

(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 2.1 (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 2.1 (b) holds if the Lebesgue measures of the boundaries of $\mathcal{X}_0$ and $\mathcal{X}_1$ are zero. We assume Assumption 2.1 (b) only to rule out perverse cases such as the case where $A(x) = 1$ if $x \in \mathbb{R}$ is a irrational number and $A(x) \neq 1$ otherwise.

2.3 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:

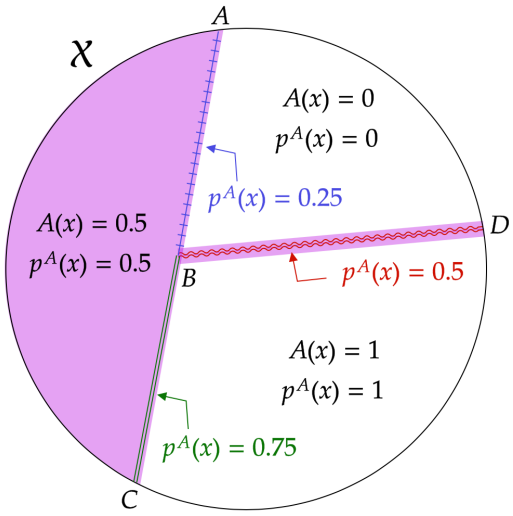
$$p^A(x; \delta) \equiv \frac{\int_{B(x,\delta)} A(x^*) dx^*}{\int_{B(x,\delta)} dx^*},$$

where $B(x, \delta) = \{x^* \in \mathbb{R}^p : \|x - x^*\| < \delta\}$ is the (open) δ-ball around $x \in \mathcal{X}$.[8] Here, $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^p$. To make a common bandwidth δ for all dimensions reasonable, we normalize X_{ij} to have mean zero and variance one for each $j = 1, ..., p$.[9] We

8. 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.

9. 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.

Figure 2.1: Example of the Approximate Propensity Score



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 2.A.1).

Figure 2.1 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 2.2 (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 2.2 is a multivariate extension of the local mean continuity condition frequently assumed in the RDD; 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. $A(x) \in \{0,1\}$ 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 2.2 requires that the points close to x have similar conditional means of the outcome Y_{zi} and treatment assignment $D_i(z)$.[10] Note that Assumption 2.2 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 2.1 without continuity.

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

Proposition 2.1 (Identification)**.** *Under Assumptions 2.1 and 2.2:*

(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* $p^A(x) \in (0,1)$.[11]

(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* $p^A(x) \in$

10. 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.

11. 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 $\Pr(Z_i = 0|X_i \in B(x,\delta)) = 0$.

$(0,1)$ *for almost every* $x \in S$ *(with respect to the Lebesgue measure).*[12]

Proof. See Appendix 2.C.1. □

Proposition 2.1 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 2.2 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 2.1, $p^A(x) \in (0,1)$ 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 2.1 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 2.1 (Perfect and Imperfect Compliance)**.** *Under Assumptions 2.1 and 2.2:*

(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* $p^A(x) \in (0,1)$, $\Pr(D_i(1) \geq$

12. 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.

$D_i(0)|X_i = x) = 1$ *(monotonicity), and* $\Pr(D_i(1) \neq D_i(0)|X_i = x) > 0$ *(existence of compliers).*

Proof. See Appendix 2.C.2. □

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 every nondegenerate algorithm identify some causal effect, as formalized in the following proposition.

Proposition 2.2. *For simplicity, suppose that* $\mathcal{X} = \mathbb{R}^p$ *and that the conditional expectation functions* $E[Y_{zi}|X_i]$ *and* $E[D_i(z)|X_i]$ *are continuous for* $z \in \{0,1\}$.[13] *If* $\text{Var}(A(X_i)) > 0$, *there exists* $x \in \mathcal{X}$ *such that* $E[Y_{1i} - Y_{0i}|X_i = x]$ *and* $E[D_i(1) - D_i(0)|X_i = x]$ *are identified.*

Proof. See Appendix 2.C.3. □

2.4 Estimation

The sources of quasi-random assignment characterized in Proposition 2.1 suggest a way of estimating causal effects of the treatment. In view of Proposition 2.1, 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. Proposition 2.1 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.

13. We assume $\mathcal{X} = \mathbb{R}^p$ to rule out cases such as the case where $\mathcal{X} = \mathcal{X}_0 \cup \mathcal{X}_1$ and the closures of $\mathcal{X}_0$ and $\mathcal{X}_1$ are nonempty and disjoint. In this case, $A(X_i)$ has variation, but no causal effect may be identified.

2.4.1 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 2.2. Consider the following 2SLS regression using the observations with $p^A(X_i; \delta_n) \in (0, 1)$:

$$D_i = \gamma_0 + \gamma_1 Z_i + \gamma_2 p^A(X_i; \delta_n) + \nu_i \tag{2.1}$$

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 p^A(X_i; \delta_n) + \epsilon_i, \tag{2.2}$$

where bandwidth δ_n shrinks toward zero as the sample size n increases. We drop the constant term if $A(X_i)$ takes only one nondegenerate value in the sample. 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.[14]

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. Formally, let $X_{i,1}^*, ..., X_{i,S_n}^*$ be S_n independent draws from the uniform

14. For the RDD special case, the 2SLS specification can be interpreted and be made more flexible in the following way. 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 (2.1) and (2.2) is numerically equivalent to a version of the regression discontinuity (RD) local linear estimator (Hahn *et al.*, 2001) that 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 (2.1) and (2.2), respectively). For the multidimensional RDD with a linear boundary, $p^A(X_i; \delta_n) \geq 1/2$ if and only if $Z_i = 1$, and hence we may use $p^A(X_i; \delta_n)$ as a single running variable with cutoff $\frac{1}{2}$. However, if the boundary is nonlinear, Z_i is not a deterministic function of $p^A(X_i; \delta_n)$. In this case, it is not straightforward to use $p^A(X_i; \delta_n)$ as a single running variable, since no appropriate cutoff value exists. We leave to future research how to allow for more flexible 2SLS specifications in the general multidimensional setting.

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 $p^s(X_i; \delta_n) \in (0, 1)$:

$$D_i = \gamma_0 + \gamma_1 Z_i + \gamma_2 p^s(X_i; \delta_n) + \nu_i \tag{2.3}$$

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 p^s(X_i; \delta_n) + \epsilon_i. \tag{2.4}$$

Let $\hat{\beta}_1^s$ denote the 2SLS estimator of β_1 in the simulation-based regression. This regression is the same as the 2SLS regression (2.1) and (2.2) 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]

2.4.2 Consistency and Asymptotic Normality

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

Assumption 2.3.

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

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 $p^s(X_i; \delta) \in (0, 1)$:

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

The estimated coefficient on Z_i, $\hat{\alpha}_1^s$, is our preferred estimator of the recommendation effect.

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]

(b) (Nonzero First Stage) $\int_{\mathcal{X}} p^A(x)(1-p^A(x))E[D_i(1)-D_i(0)|X_i=x]f_X(x)d\mu(x) \neq 0$, *where μ is the Lebesgue measure $\mathcal{L}^p$ when $\Pr(A(X_i) \in (0,1)) > 0$ and is the $(p-1)$-dimensional Hausdorff measure $\mathcal{H}^{p-1}$ when $\Pr(A(X_i) \in (0,1)) = 0$.*

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

(c) (Nonzero Variance) $\mathrm{Var}(A(X_i)) > 0$.

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

(d) (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) $\mathrm{dist}(\Omega_m^*, \Omega_{m'}^*) > 0$ for any $m, m' \in \{1, ..., M\}$ such that $m \neq m'$. Here $\mathrm{dist}(S,T) = \inf_{x \in S, y \in T} \|x-y\|$ 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 $\psi(B(x,\epsilon) \cap \partial S) \subset \{(x_1, ..., x_p) \in \mathbb{R}^p : x_p = 0\}$.

(e) (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 $\delta > 0$.

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 $\mathcal{H}^k(S) = \lim_{\delta \to 0} \mathcal{H}_\delta^k(S)$.

(f) (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 2.3 is a set of conditions for establishing consistency. Assumption 2.3 (b) assumes that the weighted average effect of the algorithm's recommendation on the treatment assignment is nonzero.[18] Under this assumption, the estimated first-stage coefficient on Z_i converges to a nonzero quantity.

Assumptions 2.3 (c)–(f) 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 regression-discontinuity variation. Assumption 2.3 (c) says that A produces variation in the treatment recommendation.

Assumption 2.3 (d) 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

18. The average is taken over the covariate values for which $p^A(x)$ is nondegenerate (i.e., $p^A(x)(1-p^A(x)) \in (0,1)$). There is a positive mass of such covariates values when A is stochastic ($\Pr(A(X_i) \in (0,1)) > 0$). When A is deterministic ($\Pr(A(X_i) \in (0,1)) = 0$), APS is nondegenerate only for boundary points at which the treatment recommendation changes from one to the other. Typically, the Lebesgue measure of the boundary is zero, and hence we compute the integral with respect to the $(p-1)$-dimensional Hausdorff measure instead.

2.3 (d), the assumptions approximately hold in that Ω^* is arbitrarily well approximated by a set that satisfies the differentiability condition.[19]

Part (i) of Assumption 2.3 (e) 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 2.3 (f) imposes continuity, continuous differentiability and boundedness on the conditional moments of potential outcomes and the probability density near the boundary of Ω^*. Note that Part (i) of Assumption 2.3 (f) implies Assumption 2.2.

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 2.4. *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 2.4 is required for proving asymptotic normality of $\hat{\beta}_1$ when A is stochastic. To explain the role of Assumption 2.4 (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 2.A.2 in Appendix 2.A.1). However, $p^A(x_\delta; \delta)$ does not

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.

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 2.4 (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 2.4 (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, 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 2.A.3. Assumption 2.4 (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 2.5 (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 2.5 imposes the condition on the growth rate of the number of simulation draws S_n. This assumption ensures that the bias caused by using $p^s(X_i; \delta_n)$ instead of $p^A(X_i; \delta_n)$ is asymptotically negligible. To understand this condition, note that $p^s(X_i; \delta_n)$ enters the 2SLS first-order condition, $\sum_{i=1}^n (1, Z_i, p^s(X_i; \delta_n))'(Y_i - \beta_0 - \beta_1 D_i - \beta_2 p^s(X_i; \delta_n)) 1\{p^s(X_i; \delta_n) \in (0, 1)\} = 0$, in two ways. First, $p^s(X_i; \delta_n)$ enters the condition in a nonlinear but smooth way through the $p^s(X_i; \delta_n)^2$ term. The asymptotic bias due to simulation errors is $O(\sqrt{n}/S_n)$. The bias diminishes under the first part of Assumption 2.5. Second, $p^s(X_i; \delta_n)$ also enters the first-order condition in a nonsmooth way, since we only use observations for which $p^s(X_i; \delta_n) \in (0, 1)$. If $p^A(X_i; \delta_n)$ is nondegenerate but close to zero or one, $p^s(X_i; \delta_n)$ may be degenerate (i.e., $A(X_{i,s}^*) = 0$ for all s or $A(X_{i,s}^*) = 1$ for all s) with a large probability. The second part of Assumption 2.5 ensures that the fraction of such observations goes to zero sufficiently fast, which eliminates the asymptotic bias caused

by not using observations with $p^A(X_i;\delta_n) \in (0,1)$.[21]

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 2.5 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 (δ_n, S_n) that satisfies both parts of Assumption 2.5 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 $\kappa_2 > \frac{1}{2}$.

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 2.1 (Consistency and Asymptotic Normality)**.** *Suppose that Assumptions 2.1 and 2.3 hold and $\delta_n \to 0$, $n\delta_n \to \infty$ 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 2.4 and 2.5 hold and $n\delta_n^2 \to 0$ 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

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

21. We also use Assumption 2.4 (c) to eliminate the asymptotic bias. By Assumption 2.4 (c), the conditional mean $E[Y_i | p^A(X_i;\delta_n) \in (0,\gamma\frac{\log n}{S_n}) \cup (1-\gamma\frac{\log n}{S_n}, 1)]$ is bounded by the same constant for all n, which implies that $E[Y_i 1\{p^A(X_i;\delta_n) \in (0,\gamma\frac{\log n}{S_n}) \cup (1-\gamma\frac{\log n}{S_n},1)\}] = o(n^{-1/2}\delta_n^{1/2})$ under Assumption 2.5.

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 2.C.4. □

Theorem 2.1 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 ($D_i(1) \neq D_i(0)$).[22] The limit $\lim_{\delta\to 0} E[\omega_i(\delta)(Y_i(1) - Y_i(0))]$ always exists under the assumptions of Theorem 2.1. Theorem 2.1 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 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_n-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

22. 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. Under monotonicity ($\Pr(D_i(1) \geq D_i(0)|X_i) = 1$), we could also apply Abadie (2003)âĂŹs Kappa weighting method using fixed-bandwidth APS instead of the standard propensity score to estimate other weighted averages of treatment effects for compliers.

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.[23]

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 2.5 to make the bias caused by using $p^s(X_i;\delta_n)$ negligible.

Finally, note that the weight $\omega_i(\delta)$ given in Theorem 2.1 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 (i.e., 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 2.2. *Suppose that Assumptions 2.1 and 2.3 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 2.3. *Suppose that Assumptions 2.1 and 2.3 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*

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

23. 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).

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, $E[D_i(1)-D_i(0)|X_i]$.

2.4.3 Intuition and Challenges

Theorem 2.1 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, as shown in the proof of Theorem 2.1 in Appendix 2.C.4. If $\Pr(A(X_i) \in (0,1)) > 0$,

$$\operatorname{plim} \hat{\beta}_1 = \operatorname{plim} \hat{\beta}_1^s = \frac{E[A(X_i)(1-A(X_i))(D_i(1)-D_i(0))(Y_i(1)-Y_i(0))]}{E[A(X_i)(1-A(X_i))(D_i(1)-D_i(0))]}. \tag{2.5}$$

The 2SLS estimators converge to a weighted average of treatment effects for the subpopulation with nondegenerate $A(X_i)$.

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

$$D_i = \gamma_0 + \gamma_1 Z_i + \gamma_2 A(X_i) + \nu_i \tag{2.6}$$

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 A(X_i) + \epsilon_i. \tag{2.7}$$

Under conditional independence, the 2SLS estimator from this regression converges in probability to the treatment-variance weighted average of treatment effects in (2.5) (Angrist and Pischke, 2008; Hull, 2018). 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 $\Pr(A(X_i) \in (0,1)) = 0$,

$$\operatorname{plim} \hat{\beta}_1 = \operatorname{plim} \hat{\beta}_1^s = \frac{\int_{\partial\Omega^*} E[(D_i(1) - D_i(0))(Y_i(1) - Y_i(0))|X_i = x] f_X(x) d\mathcal{H}^{p-1}(x)}{\int_{\partial\Omega^*} E[D_i(1) - D_i(0)|X_i = x] f_X(x) d\mathcal{H}^{p-1}(x)}. \quad (2.8)$$

The 2SLS estimators converge to a weighted average of treatment effects for the subpopulation who are on the boundary of the treated region.[24]

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 2.B.3 in Appendix 2.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^*:

$$\int_{N(\partial\Omega^*,\delta)} g(x) dx = \int_{-\delta}^{\delta} \int_{\{x' \in \mathbb{R}^p : d^s_{\Omega^*}(x') = \lambda\}} g(x) d\mathcal{H}^{p-1}(x) d\lambda, \quad (2.9)$$

where $d^s_{\Omega^*}$ is the signed distance function of Ω^* (see Appendix 2.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 2.2a shows a visual illustration of the level set.

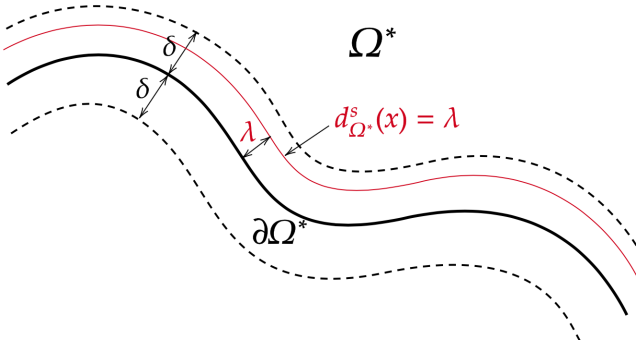
We then use the area formula (Lemma 2.B.4 in Appendix 2.B.3) to write the integral

24. The numerator and denominator are invariant to the scaling and shifting of covariates (i.e., multiplying by a constant vector element wise and adding a constant vector), since they are density-weighted averages.

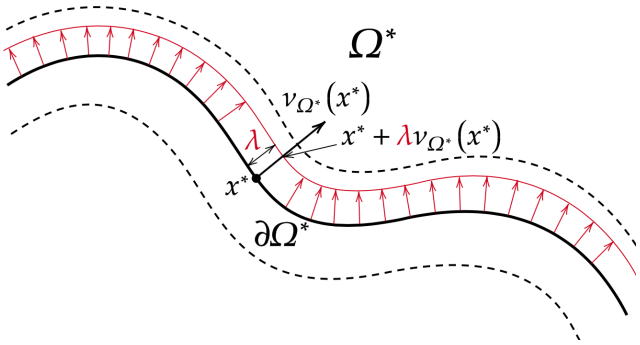
25. Our approach using geometric theory shows that $\hat{\beta}_1$ converges to an integral of the conditional treatment effect over boundary points with respect to the Hausdorff measure. In contrast, prior studies on multidimensional RDDs express treatment effect estimands in terms of expectations conditional 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.

Figure 2.2: Illustration of the Change of Variables Techniques

(a)



(b)



over each level set in terms of the integral over the boundary $\partial\Omega^*$:

$$\int_{\{x' \in \mathbb{R}^p : d^s_{\Omega^*}(x') = \lambda\}} g(x) d\mathcal{H}^{p-1}(x) = \int_{\partial\Omega^*} g(x^* + \lambda\nu_{\Omega^*}(x^*)) J^{\partial\Omega^*}_{p-1}\psi_{\Omega^*}(x^*, \lambda) d\mathcal{H}^{p-1}(x^*), \quad (2.10)$$

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^*). $J^{\partial\Omega^*}_{p-1}\psi_{\Omega^*}(x^*, \lambda)$ is the Jacobian of the transformation $\psi_{\Omega^*}(x^*, \lambda) = x^* + \lambda\nu_{\Omega^*}(x^*)$. Figure 2.2b illustrates this change of variables formula. Finally, combining (2.9) and (2.10) and

proceeding with further analysis, we prove in Appendix 2.C.4.3 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 (2.8).

2.5 Decision Making by Machine Learning

This section assesses the feasibility and performance of our method, by conducting a Monte Carlo experiment motivated by decision making by machine learning with high-dimensional data. 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 2.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 2.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 $D_i = Z_i D_i(1) + (1 - Z_i) D_i(0)$ and $Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$, respectively.

Estimands and Estimators. 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), the conditional effects are heterogeneous. However, since the RCT segment consists of those in the middle of the distribution of X_{1i}, 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 similar to LATE(RCT).

We use the data $\{(Y_i, X_i, D_i, Z_i)\}_{i=1}^n$ to estimate the treatment effect parameters. Our main approach is 2SLS with fixed-bandwidth APS controls in Theorem 2.1. 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 (2.6) and (2.7) in Section 2.4.3) and reports the coefficient on D_i.

For both models, the 2SLS estimator converges in probability to LATE(RCT) (equivalently, the right-hand side of equation (2.5)) 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 2.1 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 controls achieves this goal, as demonstrated by its smaller biases across possible treatment effect models, target parameters, and values of the bandwidth δ that are sufficiently small.

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 2.1. Table 2.1 reports the coverage probabilities of the 95% confidence intervals for LATE(RCT) constructed from the estimates and their heteroskedasticity-robust standard errors. The confidence intervals

Table 2.1: Bias, RMSE, and SD of Estimators and Coverage of 95% Confidence Intervals

	OLS with No Controls	2SLS with $A(X_i)$ Controls	Our Method: OLS with Approximate Propensity Score 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 = 0								
Bias	0.663	0.558	0.564	0.615	0.652	0.716	0.810	0.965
RMSE	0.663	0.661	0.596	0.625	0.658	0.719	0.813	0.967
Estimand: ATE(RCT) = −0.001								
Bias	0.663	0.558	0.564	0.616	0.653	0.716	0.811	0.965
RMSE	0.664	0.661	0.596	0.625	0.659	0.720	0.813	0.967
Estimand: LATE = 0.564								
Bias	0.098	−0.007	−0.001	0.051	0.088	0.152	0.246	0.400
RMSE	0.101	0.354	0.194	0.122	0.126	0.167	0.254	0.405
Estimand: LATE(RCT) = 0.566								
Bias	0.096	−0.009	−0.003	0.048	0.086	0.149	0.244	0.398
RMSE	0.099	0.354	0.194	0.121	0.124	0.165	0.252	0.403
SD	0.021	0.354	0.194	0.111	0.090	0.070	0.062	0.060
Coverage	0.4%	95.2%	94.4%	92.7%	84.0%	46.0%	3.1%	0.0%
Avg N	10000.0	100.0	397.0	1175.0	1722.0	2613.0	3349.0	3994.0
Panel B: Heterogeneous Conditional Effects (Model B)								
Estimand: ATE = 0								
Bias	1.010	0.541	0.470	0.491	0.521	0.589	0.696	0.883
RMSE	1.010	0.667	0.519	0.507	0.532	0.594	0.700	0.885
Estimand: ATE(RCT) = −0.004								
Bias	1.014	0.546	0.474	0.496	0.526	0.593	0.701	0.887
RMSE	1.014	0.670	0.523	0.512	0.536	0.599	0.704	0.890
Estimand: LATE = 0.564								
Bias	0.446	−0.023	−0.094	−0.073	−0.042	0.025	0.132	0.319
RMSE	0.446	0.390	0.240	0.146	0.112	0.084	0.150	0.325
Estimand: LATE(RCT) = 0.559								
Bias	0.450	−0.018	−0.090	−0.068	−0.038	0.029	0.137	0.323
RMSE	0.451	0.390	0.238	0.144	0.110	0.085	0.154	0.330
SD	0.018	0.389	0.221	0.127	0.104	0.080	0.071	0.065
Coverage	0.0%	94.6%	92.4%	91.7%	94.1%	93.5%	51.8%	0.3%
Avg N	10000.0	100.0	397.0	1175.0	1722.0	2613.0	3349.0	3994.0

Notes: This table shows the bias, root mean squared error (RMSE), and standard deviation (SD) of OLS with no controls, 2SLS with $A(X_i)$ controls, and 2SLS with Approximate Propensity Score controls. These statistics are computed with the estimand set to ATE, ATE(RCT), LATE, or LATE(RCT). The row 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 estimate 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 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).

for 2SLS offer nearly correct coverage when δ is small, which supports the implication of Theorem 2.1 that the inference procedure is valid when we use a sufficiently small δ. Overall,

Table 2.1 shows that our estimator works well in this high-dimensional setting and performs better than alternative estimators.

2.6 Empirical Policy Application

2.6.1 Hospital Relief Funding during the COVID-19 Pandemic

Here we provide our real-world empirical application. The COVID-19 pandemic has afflicted millions of people across the country and has imposed historic challenges for the hospitals and health systems. The pandemic led to revenue losses coupled with skyrocketing expenses, pushing many already overburdened hospitals further to their financial brink.

To deal with this crisis, 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 ask whether this funding had a causal impact on hospital operation and activities in dealing with COVID-19 patients. Answering this question would help the government respond to future healthcare crises in more effective ways. We focus on an initial portion of this funding ($10 billion). This portion 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 (1) the percentage of Medicare inpatient days attributable to patients eligible for both Medicare Part A and Supplemental Security Income (SSI),

and (2) the percentage of total inpatient days attributable to patients eligible for Medicaid but not Medicare Part A.

- 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.

- 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 2.3 visualizes how the three dimensions determine funding eligibility. From the original space of the three eligibility determinants, we extract two-dimensional planes to better visualize the structure of quasi-experimental variation. As the bottom two-dimensional planes show, eligibility discontinuously changes as hospitals cross the eligibility boundary in the characteristic space. This setting is a three-dimensional fuzzy RDD, falling under our framework.

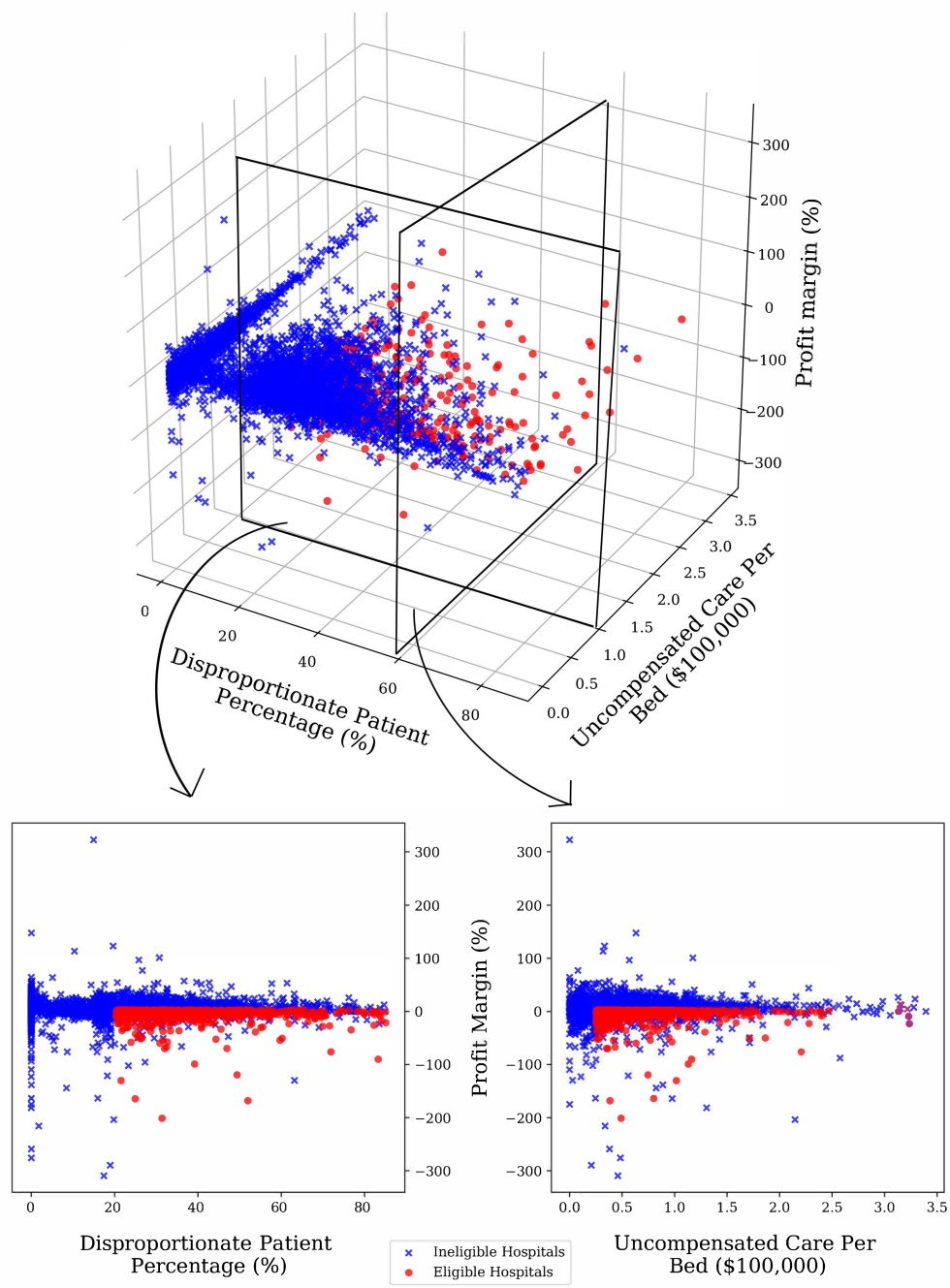
Our treatment is the funding amount, which is calculated as follows. Each eligible hospital is assigned an individual facility score, which is calculated as the product of DPP and the number of beds in that hospital. This facility score determines the share of funding allocated to the hospital, out of the total $10 billion. The share received by each hospital is determined by the ratio of the hospital's facility score to the sum of facility scores across all eligible hospitals. The amount of funding that can be received by a hospital is bounded below at $5 million and capped above at $50 million.[26]

Figure 2.4 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.[27]

26. We avoid using the founding amount as Z_i since the founding amount for a hospital is continuous and determined not only by that hospital's characteristics but also by characteristics of other hospitals.

27. 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/

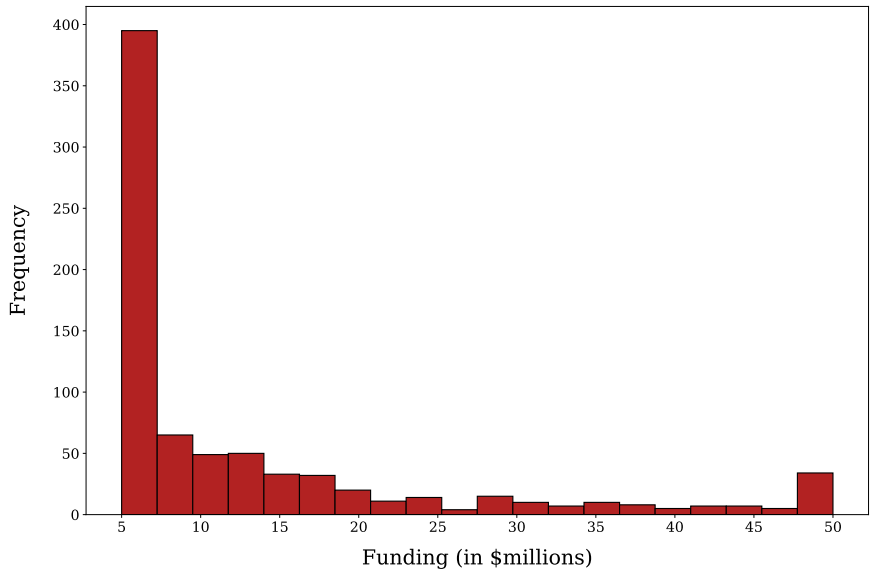
Figure 2.3: Three-dimensional Regression Discontinuity in Hospital Funding Eligibility



Notes: The top figure visualizes the three hospital characteristics 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.

Our primary outcomes are a few different versions of the number of COVID patients hospitalized at each hospital. To obtain these outcomes, we use the publicly available COVID-19

Figure 2.4: 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.

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. Summary statistics about hospital outcomes and characteristics are documented in Table 2.2. Eligible hospitals have larger numbers 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.

2.6.2 Covariate Balance Estimates

We first validate our method by evaluating the balancing property of fixed-bandwidth APS conditioning. To do so, we calculate 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

Table 2.2: 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	80.10	73.86	107.83	86.78
# Confirmed/Suspected COVID Patients in ICU	31.37	28.92	42.10	36.84
# Confirmed COVID Patients in ICU	26.62	24.41	36.56	31.41
N	4,008	3,293	715	429
Panel B: Hospital Characteristics Means				
Beds	143.66	134.60	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	.20	.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)	973.90	897.31	1351.57	1525.06
Disproportionate patient percentage	.21	.18	.38	.36
Uncompensated care per bed ($)	59,850.00	56,556.03	76,096.31	45,996.48
Profit margin	.02	.04	−.07	−.03
N	4,633	3,852	781	485

Notes: This table reports averages of outcome variables and hospital characteristics by funding eligibility. 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. Confirmed COVID patients refer to the sum of patients in inpatient beds with lab-confirmed COVID, including those with both lab-confirmed COVID 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. Confirmed COVID patients in ICU refers to the sum of patients in ICU beds with lab-confirmed COVID, including those with both lab-confirmed COVID 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 funding respectively. Column 4 shows the means for the hospitals with nondegenerate Approximate Propensity Score with bandwidth $\delta = 0.05$. Approximate Propensity Score is computed by averaging 10,000 simulation draws.

observations with $p^s(X_i;\delta) \in (0,1)$:

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

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 $p^s(X_i;\delta)$ is the simulated fixed-bandwidth APS. We compute fixed-bandwidth APS using $S = 10,000$ simulation draws for different bandwidth values.[28] The estimated coefficient on Z_i is the fixed-bandwidth-APS-controlled difference in

28. Figure 2.7 in Appendix 2.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.

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 2.3 reports the covariate balance estimates. Column 2 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 3–7 of Table 2.3. These estimates are consistent with our theoretical results, establishing the empirical ability of fixed-bandwidth APS controls to eliminate selection bias.

2.6.3 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.[29] 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

29. 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.

Table 2.3: Covariate Balance Regressions

	Mean (Ineligible Hospitals)	No Controls	Our Method: OLS 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)
Panel A: Determinants of Funding Eligibility									
Profit margin	.04	−0.11***	−0.03	−0.00	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=89	N=239	N=485	N=671	N=879	N=1726	N=2368
Uncompensated care per bed ($)	56556	19,540***	4,905	10,761	−4,229	−9,600	−11,001	−8,005*	−6,121*
		(3,827)	(12,161)	(10,356)	(8,611)	(7,651)	(6,976)	(4,498)	(3,660)
		N=4633	N=89	N=239	N=485	N=671	N=879	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)	(0.02)	(0.02)
		N=4633	N=89	N=239	N=485	N=671	N=879	N=1726	N=2368
Panel B: Other Hospital Characteristics									
Full time employees	897.32	454.26***	2,670.64	387.75	37.55	90.96	64.41	214.79	115.94
		(69.23)	(1,652.55)	(997.64)	(668.32)	(515.72)	(414.25)	(218.77)	(143.35)
		N=4626	N=89	N=238	N=484	N=670	N=878	N=1723	N=2365
Medicare net revenue (in millions $)	20.04	18.36***	35.34	−7.92	−6.16	−1.47	2.41	4.70	−0.47
		(2.39)	(29.74)	(18.36)	(14.34)	(12.09)	(10.81)	(6.63)	(4.68)
		N=4511	N=88	N=238	N=483	N=667	N=875	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=89	N=239	N=485	N=671	N=879	N=1726	N=2368
Operating margin	.02	−0.11***	−0.03	0.00	0.02	0.02	0.03	0.06***	0.07***
		(0.01)	(0.06)	(0.04)	(0.03)	(0.03)	(0.03)	(0.02)	(0.01)
		N=4541	N=88	N=238	N=477	N=661	N=868	N=1676	N=2314
Beds	134.6	53.75***	198.67*	35.86	2.93	7.02	11.93	17.47	8.92
		(7.05)	(105.74)	(66.42)	(47.75)	(39.24)	(33.06)	(20.05)	(14.46)
		N=4633	N=89	N=239	N=485	N=671	N=879	N=1726	N=2368
Costs per discharge (in thousands $)	66.28	−49.95***	3.83*	3.37**	1.65	−6.42	−0.88	6.06	6.76
		(17.93)	(2.18)	(1.49)	(1.23)	(8.12)	(2.58)	(4.63)	(5.09)
		N=3539	N=89	N=239	N=485	N=671	N=879	N=1726	N=2368
p-value joint significance		0	.74	.457	.87	.745	.286	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 funding eligibility. Other dependent variables shown indicate the financial health and utilization of the hospitals. In column 2, we regress the dependent variables on the eligibility of the hospital with no controls. In columns 3–9, 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 1 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 the number of observations is reported separately for each regression. The last row reports the p-value of the joint significance test. */**/*** indicate $p < 0.10/0.05/0.01$.

care per bed and profit margin).[30] These alternative regressions are computed using the sample of all hospitals, as benchmark estimators.

30. 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:

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

The first-stage effects of funding eligibility on funding amount (in millions of dollars) suggest that funding eligibility boosts the amount of funding significantly (columns 3–10 of Table 2.4). For example, in column 3 of Table 2.4, we 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 2.4, 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. For example, 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 in staffed ICU beds occupied by those who have lab-confirmed or suspected COVID is 1.67. These uncontrolled OLS estimates show a similar picture as the descriptive statistics in Table 2.2. 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 or lab-confirmed COVID patients become much smaller and lose significance across all bandwidth specifications. In fact, even the sign of the estimated funding effects is reversed for several combinations of the outcome and bandwidth. Once we control for fixed-bandwidth APS to eliminate the bias, therefore, funding has little to no effect on the hospital utilization level by COVID-19 patients. These results suggest that fixed-bandwidth APS reveals important selection bias

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.

Table 2.4: Estimated Effects of Funding on Hospital Utilization

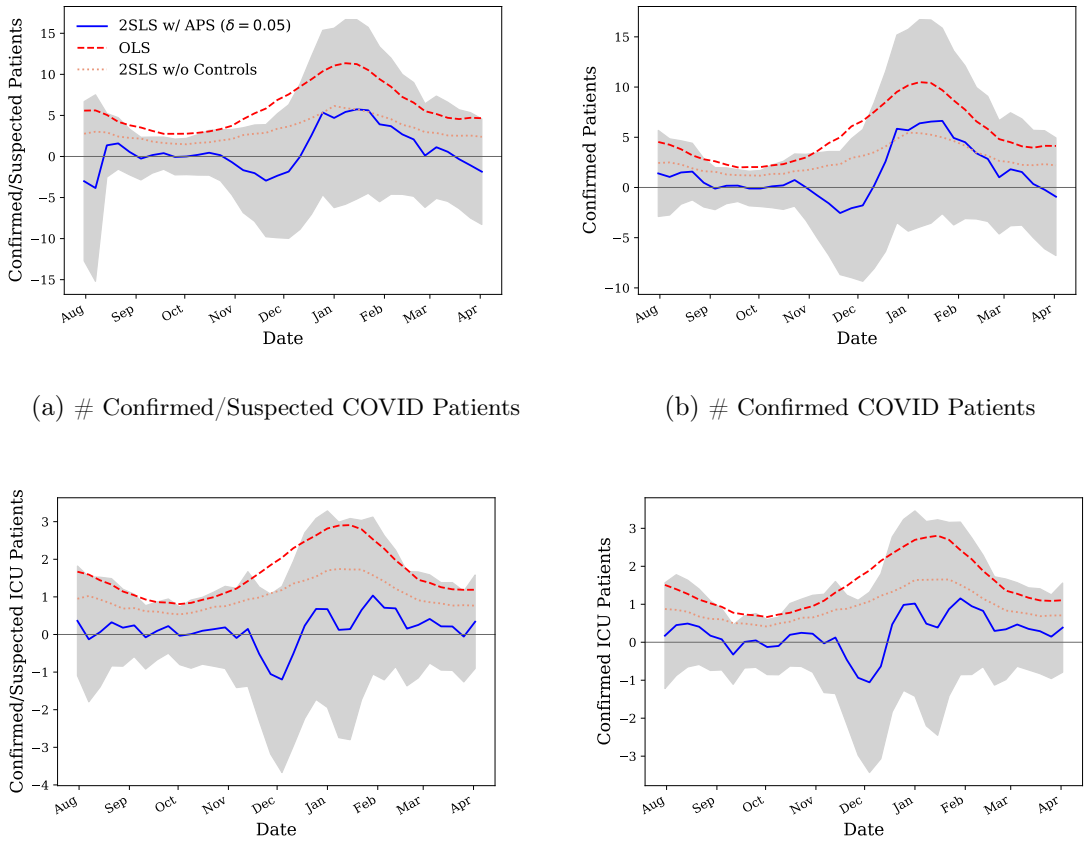
	OLS with No Controls	OLS with Covariate Controls	2SLS with No Controls	Our Method: 2SLS with Approximate Propensity Score Controls						
				$\delta =$ 0.01	$\delta =$ 0.025	$\delta =$ 0.05	$\delta =$ 0.075	$\delta =$ 0.1	$\delta =$ 0.25	$\delta =$ 0.5
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
# Confirmed/Suspected 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)
N	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)
N	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)
N	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)
N	3503	3503	3503	67	181	370	514	671	1321	1868

Notes: In this table we regress relevant outcomes at the hospital level on the amount of funding. Column 1 presents the results of OLS regression of the outcome variables on funding without any controls. Column 2 presents the results of OLS regression of the outcome variables on funding controlling for disproportionate patient percentage, uncompensated care per bed and profit margin. In columns 3–10, we instrument the amount of 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. */**/*** indicate $p < 0.10/0.05/0.01$.

in the naively estimated effects of funding.[31]

31. The 2SLS estimates in Table 2.4 are unlikely to be compromised by differential attrition. Estimates reported in Table 2.5 in Appendix 2.E.4 show little difference in outcome availability rates between eligible and ineligible hospitals once we control for fixed-bandwidth APS.

Figure 2.5: 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. Grey areas are 95% confidence intervals.

2.6.4 Persistence and Heterogeneity

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 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 estimated dynamic effects in Figure 2.5. The estimated dynamic effects are similar to the initial null effects in Table 2.4, 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 2.6 plots the resulting estimates by repeating the same dynamic analysis as in Figure 2.5, 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 Figure 2.6a, 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.

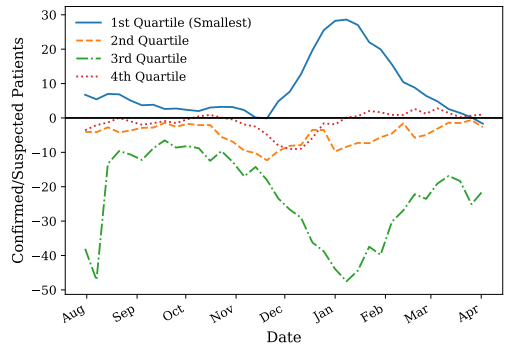
There is also a sizable dip in the funding effect of for-profit hospitals around the same period. This could be due to regional differences in the distribution of hospital ownership. Nonprofits and government-managed hospitals tend to be in rural areas, which both received more funding and experienced 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 largely no effect on hospital utilization trends during the pandemic. The null effect is widely observed for subgroups of hospitals at different points in time. 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.

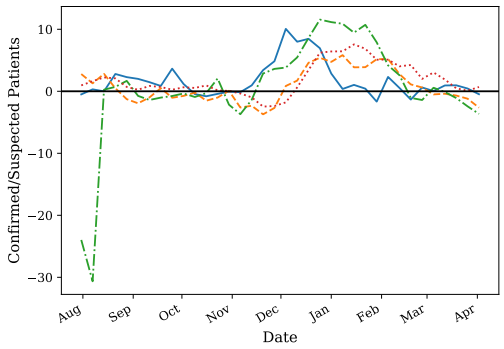
2.7 Other Examples

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

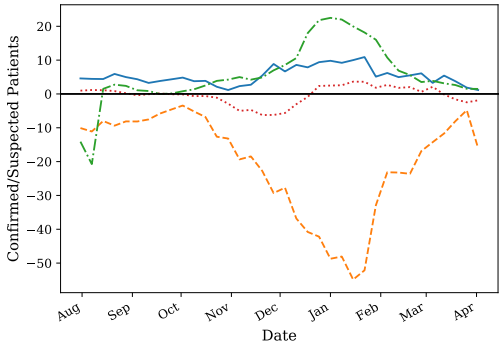
Figure 2.6: Dynamic Heterogeneous Effects of Hospital Funding by Hospital Characteristics



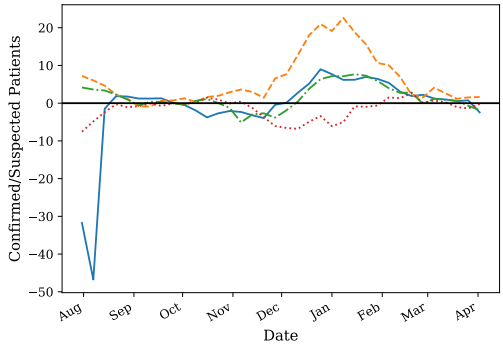
(a) Net Income



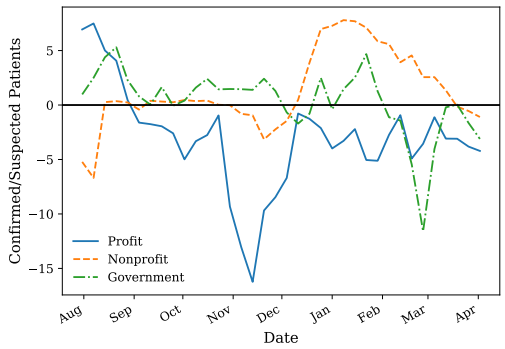
(b) Total Full-Time Employees



(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. No effect estimates are significantly different from each other at the 5% level. We compute APS with $S = 10,000$ and $\delta = 0.05$.

Example 2.1 (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 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 $\mathrm{Var}(Y_i(z)|X_i)$, 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 $\mathrm{diag}(\sigma_0^2(X_i), \sigma_1^2(X_i))$. 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 $\mu_0(\cdot)$, $\mu_1(\cdot)$, $\sigma_0^2(\cdot)$ and $\sigma_1^2(\cdot)$ 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. Furthermore, if the functions $\mu_0(\cdot)$, $\mu_1(\cdot)$, $\sigma_0^2(\cdot)$ and $\sigma_1^2(\cdot)$ are continuously differentiable, this algorithm satisfies Assumption 2.4 (a) and (b), which are required for asymptotic normality when $\Pr(A(X_i) \in (0, 1)) > 0$.

(b) (Upper Confidence Bound, UCB) Unlike the above stochastic algorithm, the UCB

algorithm is a deterministic algorithm, producing a less obvious example of our framework. This algorithm chooses the treatment with the highest upper confidence bound for the potential outcome:

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

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. If, in addition, g is twice continuously differentiable along the boundary, this algorithm satisfies Assumption 2.3 (d), which is required for consistency and asymptotic normality when $\Pr(A(X_i) \in (0, 1)) = 0$. It is possible to identify and estimate causal effects across the boundary.

Example 2.2 (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 $C(x)$.[32] Thus, it is possible to identify causal effects across the boundary $\partial(\cup_{k\in T} S_k)$. Assumption 2.3 (d) approximately holds in that the target group $\cup_{k\in T} S_k$ is arbitrarily well approximated by a set that satisfies the differentiability condition.

Example 2.3 (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 learning based on past data and

32. 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.

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. 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. It may not be suitable to apply a standard univariate RDD method.

In contrast, our approach is applicable as long as X_{1i} and X_{2i} 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 $\{x \in \mathcal{X} : (x_1 = 0, x_2 > 0) \text{ or } (x_1 > 0, x_2 = 0)\}$. Assumption 2.3 (d) approximately holds in that the set

$\{x \in \mathbb{R}^2 : x_1 > 0, x_2 > 0\}$ is arbitrarily well approximated by a set that satisfies the differentiability condition.

Example 2.4 (Policy Eligibility Rules)**.** Similar to our empirical application about hospital funding, Medicaid and other welfare policies often decide who are eligible based on algorithmic rules (Currie and Gruber, 1996; Brown, Kowalski and Lurie, 2020).[33] 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 2.3.

Example 2.5 (Mechanism Design: Matching and Auction)**.** Centralized economic mechanisms such as matching and auction are also suitable examples, as summarized below (Abdulkadiroğlu *et al.*, 2017, 2022; Abdulkadiroğlu, 2013; Kawai *et al.*, 2022; 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 con-

33. 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.

sider the following large market setting.[34] 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.

Corollary 2.4. *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 2.1 and 2.2.*

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.

34. 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.

Our analysis provides a few implications for policy and management practices of decision-making algorithms. 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 and inference 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.[35] 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, 2021) and Imbens and Wager (2019). Finally, we look forward to empirical applications of our method in a variety of business, policy, and scientific domains.

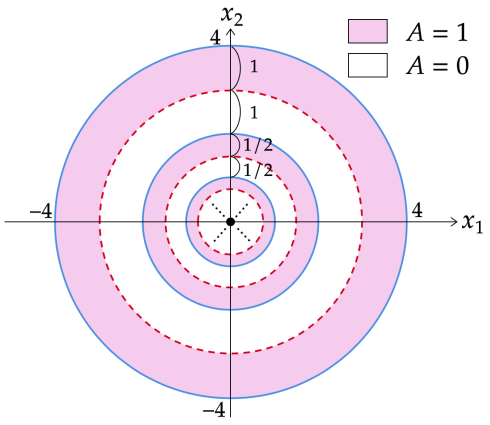
Appendices

2.A Extensions and Discussions

2.A.1 Existence of the Approximate Propensity Score

Proposition 2.1 assumes that APS exists, but is it fair to assume so? In general, APS may fail to exist. The following figure shows such an example.

35. 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.



In this example, X_i is two dimensional, and

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

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 2.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 2.C.5. □

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 2.A.2. *Take any $x \in \mathcal{X}$. If there exists a countable set $Q \subset \mathrm{Supp}(A(X_i))$ 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 2.C.6. □

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 2.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 $x \in \mathrm{int}(\mathcal{X}_q)$, $p^A(x)$ 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 2.C.7. □

2.A.2 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 2.1, 2.2, 2.3 and 2.4 hold conditional on X_{di}.

Assumption 2.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 $k = 0, 1$.

2.A.2.1 Identification

Assumption 2.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 2.A.3. *Under Assumptions 2.A.1 and 2.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 2.C.8. □

2.A.2.2 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 2.A.3.

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

(b) (Nonzero First Stage) $\int_{\mathcal{X}} p^A(x)(1 - p^A(x))E[D_i(1) - D_i(0)|X_i = x]f_X(x)d\mu(x) \neq 0$, *where* μ *is the Lebesgue measure* $\mathcal{L}^p$ *when* $\Pr(A(X_i) \in (0, 1)) > 0$ *and is the* $(p-1)$*-dimensional Hausdorff measure* $\mathcal{H}^{p-1}$ *when* $\Pr(A(X_i) \in (0, 1)) = 0$.

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

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

(d) (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* $m \in \{1, ..., M\}$.

(e) (Regularity of Deterministic A) *For each* $x_d \in \mathcal{X}_d^*$, *the following holds.*

(i) $\mathcal{H}^{p_c-1}(\partial\Omega^*(x_d)) < \infty$, *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* $x_c \in N(\mathcal{X}_c(x_d), \delta) \setminus \Omega^*(x_d)$.

(f) (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 2.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 2.A.1. *Suppose that Assumptions 2.A.1 and 2.A.3 hold and $\delta_n \to 0$, $n\delta_n \to \infty$ 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 2.A.4 and 2.5 hold and $n\delta_n^2 \to 0$ 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 2.C.9. □

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

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

2.A.3 A Sufficient Condition for Assumption 2.4 (a)

We provide a sufficient condition for Assumption 2.4 (a).

Assumption 2.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) $\text{dist}(C_m^*, C_{m'}^*) > 0$ *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^*. Under Assumption 2.A.5 (a), by Lemma 2.B.4 in Appendix 2.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 2.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 $O(\delta)$.

2.B Notation and Lemmas

2.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.

2.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 S. 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 2.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

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

and

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

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 an *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 2.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 2.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. □

2.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 2.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 2.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 2.B.5. *Let $\Omega \subset \mathbb{R}^p$, and suppose that there exists a partition $\{\Omega_1, ..., \Omega_M\}$ of Ω such that*

(i) $\operatorname{dist}(\Omega_m, \Omega_{m'}) > 0$ *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 2.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 $N(\partial\Omega, \delta)$,

$$\begin{aligned}
\int_{N(\partial\Omega,\delta)} g(x)dx &= \int_{\{x' \in \mathbb{R}^p : d_\Omega^s(x') \in (-\delta,\delta)\}} g(x)\sqrt{\det(\|\nabla d_\Omega^s(x)\|)}dx \\
&= \int_{\{x' \in \mathbb{R}^p : d_\Omega^s(x') \in (-\delta,\delta)\}} g(x)\sqrt{\det(\nabla d_\Omega^s(x)'\nabla d_\Omega^s(x))}dx \\
&= \int_{\{x' \in \mathbb{R}^p : d_\Omega^s(x') \in (-\delta,\delta)\}} g(x)\sqrt{\det((Jd_\Omega^s(x))(Jd_\Omega^s(x))')}dx \\
&= \int_{\mathbb{R}} \int_{\{x' \in \mathbb{R}^p : d_\Omega^s(x') \in (-\delta,\delta), d_\Omega^s(x') = \lambda\}} g(x)d\mathcal{H}^{p-1}(x)d\lambda \\
&= \int_{-\delta}^{\delta} \int_{\{x' \in \mathbb{R}^p : d_\Omega^s(x') = \lambda\}} g(x)d\mathcal{H}^{p-1}(x)d\lambda, \qquad (2.11)
\end{aligned}$$

where the first equality follows since $\|\nabla d_\Omega^s(x)\| = 1$ for every $x \in N(\partial\Omega, \delta)$ by Lemma 2.B.1, the third equality follows from the definition of the Jacobian matrix, and the fourth equality follows from Lemma 2.B.3.

Let $\Gamma(\lambda) = \{x \in \mathbb{R}^p : d_\Omega^s(x) = \lambda\}$ for each $\lambda \in (-\mu, \mu)$. Since ∇d_Ω^s 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 2.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 $\psi_\Omega(x_0,\lambda) \in \mathbb{R}^p \setminus \mathrm{cl}(\Omega)$ 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 2.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 2.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 2.B.4, we have that for every $\lambda \in (-\mu,\mu)$,

$$\begin{aligned} \int_{\partial\Omega} g(u+\lambda\nu_\Omega(u)) J^{\partial\Omega}_{p-1}\psi_\Omega(u,\lambda)\, d\mathcal{H}^{p-1}(u) &= \int_{\partial\Omega} g(\psi_\Omega(u,\lambda)) J^{\partial\Omega}_{p-1}\psi_\Omega(u,\lambda)\, d\mathcal{H}^{p-1}(u) \\ &= \int_{\mathbb{R}^p} \sum_{u\in\partial\Omega:\psi_\Omega(u,\lambda)=x} g(\psi_\Omega(u,\lambda))\, d\mathcal{H}^{p-1}(x). \end{aligned} \tag{2.12}$$

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 2.B.1, such u is unique, and hence $\{u \in \partial\Omega : \psi_\Omega(u,\lambda) = x\}$ is a singleton. It follow that

$$\begin{aligned}\int_{\mathbb{R}^p} \sum_{u\in\partial\Omega:\psi_\Omega(u,\lambda)=x} g(\psi_\Omega(u,\lambda))d\mathcal{H}^{p-1}(x) &= \int_{\{\psi_\Omega(u,\lambda):u\in\partial\Omega\}} g(x)d\mathcal{H}^{p-1}(x) \\ &= \int_{\Gamma(\lambda)} g(x)d\mathcal{H}^{p-1}(x), \qquad (2.13)\end{aligned}$$

where the last equality holds since $\{\psi_\Omega(u,\lambda) : u \in \partial\Omega\} = \Gamma(\lambda)$. Combining (2.11), (2.12) and (2.13), 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 $P(x) = \{\sum_{k=1}^{p-1} c_k v_k(x) : 0 \le c_k \le 1 \text{ for } k = 1, ..., p-1\}$. 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 $\lambda \in \mathbb{R}$,

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

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

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

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 $x \in \partial\Omega$, $J^{\partial\Omega}_{p-1}\psi_\Omega(x,\cdot)$ is continuously differentiable in λ, and $J^{\partial\Omega}_{p-1}\psi_\Omega(x,0) = \sqrt{\det(I_p)} = 1$.

Lastly, we show that $J^{\partial\Omega}_{p-1}\psi_\Omega(\cdot,\cdot)$ and $\frac{\partial J^{\partial\Omega}_{p-1}\psi_\Omega(\cdot,\cdot)}{\partial\lambda}$ are bounded on $\partial\Omega \times (-\mu^*, \mu^*)$ for some $\mu^* \in (0,\mu)$. Let $f : \mathbb{R} \times \mathbb{R}^{p\times(p-1)} \times \mathbb{R}^{p\times p} \to \mathbb{R}$ be a function such that

$$f(\lambda, V, D) = \det(I_p + 2\lambda VV'D + \lambda^2 VV'D^2).$$

Note that $J^{\partial\Omega}_{p-1}\psi_\Omega(x,\lambda) = \sqrt{f(\lambda, V_\Omega(x), D^2 d^s_\Omega(x))}$.

Let $S = \{(V, D^2 d^s_\Omega(x)) \in \mathbb{R}^{p\times(p-1)} \times \mathbb{R}^{p\times p} : \|v_k\| = 1 \text{ for } k = 1,...,p-1, x \in \partial\Omega\}$, where v_k denotes the kth column of V. Since $D^2 d^s_\Omega(\cdot)$ is continuous on $\partial\Omega$, and $\partial\Omega$ is closed and bounded, S is closed and bounded. Observe that

$$\frac{\partial f(\lambda,V,D)}{\partial\lambda} = \sum_{i,j} \frac{\partial\det(I_p + 2\lambda VV'D + \lambda^2 VV'D^2)}{\partial b_{ij}}(2(VV'D)_{ij} + 2\lambda(VV'D^2)_{ij}),$$

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, which is continuous. Since the right-hand side is continuous in (λ, V, D), there exists $\bar{M} > 0$ such that $|\frac{\partial f(\lambda,V,D)}{\partial\lambda}| \le \bar{M}$ for all $(\lambda, V, D) \in [-\mu,\mu] \times S$.

By the mean value theorem, for every $(\lambda, V, D) \in [-\mu, \mu] \times S$,

$$\begin{aligned} f(\lambda, V, D) &= f(0, V, D) + \frac{\partial f(\tilde{\lambda}, V, D)}{\partial \lambda}\lambda \\ &\in [1 - \bar{M}|\lambda|, 1 + \bar{M}|\lambda|], \end{aligned}$$

where $\tilde{\lambda}$ lies on the line segment connecting 0 and λ and the second line holds since $f(0, V, D) = 1$ by construction. Pick $\mu^* \in (0, \bar{\mu}]$ such that $1 - \bar{M}\mu^* > 0$. Since $\{(V_\Omega(x), D^2 d^s_\Omega(x)) : x \in \partial\Omega\} \subset S$, it follows that $J^{\partial\Omega}_{p-1}\psi_\Omega(x, \lambda) = \sqrt{f(\lambda, V_\Omega(x), D^2 d^s_\Omega(x))}$ is bounded on $\partial\Omega \times (-\mu^*, \mu^*)$. Moreover, for every $(x, \lambda) \in \partial\Omega \times (-\mu^*, \mu^*)$,

$$\begin{aligned} \frac{\partial J^{\partial\Omega}_{p-1}\psi_\Omega(x, \lambda)}{\partial \lambda} &= \frac{1}{2\sqrt{f(\lambda, V_\Omega(x), D^2 d^s_\Omega(x))}} \frac{\partial f(\lambda, V_\Omega(x), D^2 d^s_\Omega(x))}{\partial \lambda} \\ &\in \left(-\frac{\bar{M}}{2\sqrt{1 - \bar{M}\mu^*}}, \frac{\bar{M}}{2\sqrt{1 - \bar{M}\mu^*}} \right). \end{aligned}$$

Thus, $\frac{\partial J^{\partial\Omega}_{p-1}\psi_\Omega(x,\lambda)}{\partial \lambda}$ is bounded on $\partial\Omega \times (-\mu^*, \mu^*)$.

□

2.B.4 Other Lemmas

Lemma 2.B.6. *Let $\{V_i\}_{i=1}^\infty$ be i.i.d. random variables such that $E[V_i^2] < \infty$. If Assumption 2.1 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 2.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 2.1 (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 2.1 (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 2.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 2.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 2.1 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 2.5 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

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

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 2.B.7, for $l = 0, 1, 2$,

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

Also, by Lemma 2.B.7,

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

and

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

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

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

and for $l = 3, 4$,

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

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 $\frac{1}{n}\sum_{i=1}^{n} V_i p^s(X_i;\delta_n)^l(I_{i,n}^s - I_{i,n}) = o_p(1)$ 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 2.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 2.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 2.1 (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 2.1 (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,

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

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 2.5 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 2.5. 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 2.5, $\sqrt{n}\Pr(p^A(X_i;\delta_n) \in (0,\eta_n)\cup(1-\eta_n,1)) = o(1)$. 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,

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

□

2.C Proofs

2.C.1 Proof of Proposition 2.1

Suppose that Assumptions 2.1 and 2.2 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 $A(x) \in (0,1)$, $E[Y_{1i} - Y_{0i}|X_i = x]$ is trivially identified by Property 2.1:

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

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 2.1, and the second from Assumption 2.2. □

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

Proof. By Assumption 2.1, A is continuous almost everywhere. Part 1 of Cororally 2.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 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 2.1. But this is a contradiction to the result from Step 1. □

Step 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 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 2.1 and Assumptions 2.1 and 2.2. 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 S 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 3, $x \notin B(x^*, \delta)$. Hence, $B(x, \epsilon) \subset \mathcal{X} \setminus B(x^*, \epsilon)$. By Assumption 2.2 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

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

By the definition of support, $\Pr(X_i^* \in B(x^*, \epsilon)) > 0$. 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. □

□

□

2.C.2 Proof of Corollary 2.1

If $\Pr(D_i(1) - D_i(0) = 1|X_i = x) = 1$, $\Pr(Y_{1i} - Y_{0i} = Y_i(1) - Y_i(0)|X_i = x) = 1$, and hence $E[Y_{1i} - Y_{0i}|X_i = x] = E[Y_i(1) - Y_i(0)|X_i = x]$. Then, Part (a) follows from Proposition 2.1 (a). If $\Pr(D_i(1) \geq D_i(0)|X_i = x) = 1$, we have

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

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

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

Then, Part (b) follows from Proposition 2.1 (a). □

2.C.3 Proof of Proposition 2.2

Here, we only show that there exists $x \in \mathcal{X}$ such that $E[Y_{1i} - Y_{0i}|X_i = x]$ under stated conditions. The result for $E[D_i(1) - D_i(0)|X_i = x]$ is obtained analogously.

First, consider the case where $\Pr(A(X_i) \in (0, 1)) > 0$. In this case, there exists $x \in \mathcal{X}$ such that $A(x) \in (0, 1)$. By Property 1, $E[Y_{1i} - Y_{0i}|X_i = x]$ is trivially identified.

Second, consider the case where $\Pr(A(X_i) \in (0, 1)) = 0$. In this case, $\Pr(A(X_i) = 1) \in (0, 1)$, since otherwise $\mathrm{Var}(A(X_i)) = 0$. Therefore, $\mathcal{X}_1$ is nonempty and $\mathcal{X}_1 \neq \mathbb{R}^p$. Now, note that the boundary of $\mathcal{X}_1$, denoted by $\partial\mathcal{X}_1$, is nonempty by the following two facts: (1) The boundary of a set $S \subset \mathbb{R}^p$ is empty if and only if S is a closed and open set in $\mathbb{R}^p$; (2) If a set S is a closed and open set in $\mathbb{R}^p$, then either S is empty or $S = \mathbb{R}^p$. Since $\partial\mathcal{X}_1$ is nonempty, there exists $x \in \partial\mathcal{X}_1$. By the definition of the boundary, for any $\delta > 0$, $B(x, \delta) \cap \mathcal{X}_1$ and $B(x, \delta) \cap (\mathcal{X} \setminus \mathcal{X}_1)$ are nonempty. This implies that we can find points $x_{0,\delta}, x_{1,\delta} \in B(x, \delta)$ such that $A(x_{0,\delta}) < 1$ and $A(x_{1,\delta}) = 1$ for any sufficiently small $\delta > 0$. Noting that $x_{0,\delta} \to x$

and $x_{1,\delta} \to x$ as $\delta \to 0$,

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

where the first equality follows from Property 1, and the second from the continuity of $E[Y_{zi}|X_i]$ for $z \in \{0, 1\}$. □

2.C.4 Proof of Theorem 2.1

We prove consistency and asymptotic normality of the following estimators. First, consider the following 2SLS regression using the observations with $p^A(X_i; \delta_n) \in (0, 1)$:

$$D_i = \gamma_0(1 - \mathbf{I}_n) + \gamma_1 Z_i + \gamma_2 p^A(X_i; \delta_n) + \nu_i \tag{2.14}$$

$$Y_i = \beta_0(1 - \mathbf{I}_n) + \beta_1 D_i + \beta_2 p^A(X_i; \delta_n) + \epsilon_i. \tag{2.15}$$

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

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

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 $p^s(X_i;\delta_n) \in (0,1)$:

$$D_i = \gamma_0(1-\mathbf{I}_n) + \gamma_1 Z_i + \gamma_2 p^s(X_i;\delta_n) + \nu_i \tag{2.16}$$

$$Y_i = \beta_0(1-\mathbf{I}_n) + \beta_1 D_i + \beta_2 p^s(X_i;\delta_n) + \epsilon_i. \tag{2.17}$$

Let $\hat{\beta}_1^s$ denote the 2SLS estimator of β_1 in the simulation-based regression.

Below, we prove the following result.

Theorem 2.C.1. *Suppose that Assumptions 2.1 and 2.3 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 2.4 and 2.5 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

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

where

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

Let $\hat{\sigma}_n^2$ denote the estimator for the variance of $\hat{\beta}_1$. That is, $\hat{\sigma}_n^2$ 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* $\mathbf{I}_n = 1$. $(\hat{\sigma}_n^s)^2$ *is the analogously-defined estimator for the variance of* $\hat{\beta}_1^s$ *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 $\hat{\beta}_1^s$.

2.C.4.1 Consistency and Asymptotic Normality of $\hat{\beta}_1$ When $\Pr(A(X_i) \in (0,1)) > 0$

By Lemma 2.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 2.A.1. Note that $E[p^A(X_i)(1-p^A(X_i))(D_i(1)-D_i(0))] = \int_{\mathcal{X}} p^A(x)(1-p^A(x))E[D_i(1)-D_i(0)|X_i = x]f_X(x)dx$. Hence, under Assumption 2.3 (b), $E[p^A(X_i)(1 - p^A(X_i))(D_i(1) - D_i(0))] > 0$. Again by Lemma 2.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 $\text{Var}(A(X_i)|I_i^A = 1) > 0$, and that $\Pr(\mathbf{I}_n = 1) \to 1$ when $\text{Var}(A(X_i)|I_i^A = 1) = 0$. 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 $\text{Var}(A(X_i)|I_i^A = 1) > 0$,

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

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

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 2.1 and 2.3 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 2.4 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 2.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 2.1. 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 2.1, and the second last follows from the definitions of Y_{1i} and Y_{0i}.

We next consider $\hat{\beta}_1^{nc}$. By Lemma 2.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 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,

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

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

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

□

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 2. *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* $\tilde{\epsilon}_i = Y_i - \tilde{\mathbf{D}}_i'\beta$. *Then*

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

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 2.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 $n \to \infty$. 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 2.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 2.3 (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 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

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

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 2.4 (a). □

Step 4. $\frac{1}{\sqrt{n}} \sum_{i=1}^n (\mathbf{Z}_i Y_i I_i - \tilde{\mathbf{Z}}_i Y_i I_i) = o_p(1)$ *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 3 and the last from the assumption that $n\delta_n^2 \to 0$. As for variance, by Lemma 2.B.6,

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

□

Step 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 2.B.6. Again by Lemma 2.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 6. $(\hat{\sigma}^c)^{-1}(\hat{\beta}_1^c - \beta_1) \xrightarrow{d} \mathcal{N}(0,1)$.

Proof. By combining the results from Steps 2–4 and by Lemma 2.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 1 and 5. □

□

□

2.C.4.2 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

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

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 2.5 holds when $\text{Var}(A(X_i)|I_i^A = 1) > 0$. 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 2.5 holds. We have

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

By Lemma 2.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 2.4 (c) if Assumption 2.5 holds.

By proceeding as in Step 5 in Section 2.C.4.1, 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 2.B.8,

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

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 2.B.8. The conclusion then follows.

□

2.C.4.3 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 2.B. By Lemma 2.B.5, under Assumption 2.3 (d), there exists $\mu > 0$ such that $d^s_{\Omega^*}$ is twice continuously differentiable on $N(\partial\Omega^*, \mu)$ and that

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

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 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 2.3 (e) (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 2.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

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

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

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

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

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

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 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 $(u,v,\delta) \in \partial\Omega^* \cap N(\mathcal{X},\bar{\delta}) \times (-1,1) \times (0,\bar{\delta})$. Let $\epsilon \in (0, \delta(1-|v|))$. Note that $B(u,\epsilon) \subset B(u + \delta v \nu_{\Omega^*}(u), \delta)$, since for any $x \in B(u,\epsilon)$, $\|u + \delta v \nu_{\Omega^*}(u) - x\| \leq \|\delta v \nu_{\Omega^*}(u)\| + \|u - x\| \leq \delta|v| + \epsilon < \delta$. By Step 1, $p^A(u) = \lim_{\delta' \to 0} p^A(u;\delta') = k(0) = \frac{1}{2}$. This implies that 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 $0 < \mathcal{L}^p(B(u,\epsilon') \setminus \Omega^*) \leq \mathcal{L}^p(B(u,\epsilon) \setminus \Omega^*) \leq \mathcal{L}^p(B(u + \delta v \nu_{\Omega^*}(u), \delta) \setminus \Omega^*)$. 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 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*

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

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 1. Under Assumption 2.3 (f), 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 2.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

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

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

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

where the second equality follows from Steps 1 and 2. By Lemma 2.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 2.3 (e) (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 2.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,

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

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 2.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,

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

□

Step 4. *Let* $S_{\mathbf{D}} = \lim_{\delta\to 0} \delta^{-1}E[\mathbf{Z}_i \mathbf{D}_i' 1\{p^A(X_i;\delta) \in (0,1)\}]$ *and* $S_Y = \lim_{\delta\to 0} \delta^{-1}E[\mathbf{Z}_i Y_i 1\{p^A(X_i;\delta) \in (0,1)\}]$. *Then the second element of* $S_{\mathbf{D}}^{-1} S_Y$ *is* β_1.

Proof. Note that $D_i = Z_i D_i(1) + (1 - Z_i) D_i(0)$ and $Y_i = Z_i Y_{1i} + (1 - Z_i) Y_{0i}$. By Step 3,

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

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

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

After a few lines of algebra, we have

$$\begin{aligned} \det(S_{\mathbf{D}}) =& \bar{f}_X^2 \int_{\partial\Omega^*} E[D_i(1) - D_i(0) | X_i = x] f_X(x) d\mathcal{H}^{p-1}(x) \\ & \times (\int_{-1}^0 (k(v) - \int_{-1}^0 k(s) ds)^2 dv + \int_0^1 (k(v) - \int_0^1 k(s) ds)^2 dv). \end{aligned}$$

We verify that $\det(S_{\mathbf{D}})$ is nonzero. Since $\bar{f}_X > 0$ by Assumption 2.3 (e) (i), it suffices to show that $\int_{\partial\Omega^*} E[D_i(1) - D_i(0) | X_i = x] f_X(x) d\mathcal{H}^{p-1}(x) > 0$. To do so, we first show that $p^A(x) \in \{0, 1\}$ for every $x \in \mathcal{X} \setminus \partial\Omega^*$. Pick $x \in \mathcal{X} \setminus \partial\Omega^*$. By definition, either $x \in \mathcal{X} \cap \text{int}(\Omega^*)$ or $x \in \mathcal{X} \cap (\mathbb{R}^p \setminus \text{cl}(\Omega^*))$. If $x \in \mathcal{X} \cap \text{int}(\Omega^*)$, then $B(x, \delta) \subset \text{int}(\Omega^*)$ for any sufficiently small $\delta > 0$ so that $p^A(x) = 1$. If $x \in \mathcal{X} \cap (\mathbb{R}^p \setminus \text{cl}(\Omega^*))$, then $B(x, \delta) \subset N(\mathcal{X}, \delta') \cap (\mathbb{R}^p \setminus \text{cl}(\Omega^*))$ for any sufficiently small $\delta > 0$, where $\delta' > 0$ satisfies Assumption 2.3 (e) (ii). Since $A(x') = 0$ for almost every $x' \in N(\mathcal{X}, \delta') \setminus \Omega^*$, $p^A(x) = 0$. Note also that $p^A(x) = \lim_{\delta \to 0} p^A(x; \delta) = k(0) = \frac{1}{2}$ for every $x \in \partial\Omega^* \cap \mathcal{X}$ by Step 1. It then follows that

$$\begin{aligned} & \int_{\partial\Omega^*} E[D_i(1) - D_i(0) | X_i = x] f_X(x) d\mathcal{H}^{p-1}(x) \\ =& 4 \int_{\partial\Omega^* \cap \mathcal{X}} p^A(x)(1 - p^A(x)) E[D_i(1) - D_i(0) | X_i = x] f_X(x) d\mathcal{H}^{p-1}(x) \\ =& 4 \int_{\mathcal{X}} p^A(x)(1 - p^A(x)) E[D_i(1) - D_i(0) | X_i = x] f_X(x) d\mathcal{H}^{p-1}(x), \end{aligned}$$

which is nonzero under Assumption 2.3 (b).

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 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 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 $\frac{1}{n\delta_n}\sum_{i=1}^n \mathbf{Z}_i Y I_i$ is $o(1)$. Here, we only verify that $\mathrm{Var}(\frac{1}{n\delta_n}\sum_{i=1}^n p^A(X_i;\delta_n) Y_i I_i) = o(1)$. Note that

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

Under Assumption 2.3 (f), 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

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

for some $C > 0$, where the last inequality follows from Step 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

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

Step 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 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 2.3 (f), $\lim_{n\to\infty} \delta_n^{-1} E[\epsilon_i^2 \mathbf{Z}_i \mathbf{Z}_i' I_i]$ 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

$$\begin{aligned}\sum_{i=1}^{n} E[|V_{i,n}|^4] &= \frac{1}{n\delta_n}\delta_n^{-1}E[|\lambda'(\mathbf{Z}_i\epsilon_i I_i - E[\mathbf{Z}_i\epsilon_i I_i])|^4] \\ &\leq \frac{1}{n\delta_n}2^{3c}\delta_n^{-1}\{E[|\lambda'\mathbf{Z}_i\epsilon_i I_i|^4] + |\lambda' E[\mathbf{Z}_i\epsilon_i I_i]|^4\}\end{aligned}$$

by the c_r-inequality. Repeating using the c_r-inequality gives

$$\begin{aligned}\delta_n^{-1}E[|\lambda'\mathbf{Z}_i\epsilon_i I_i|^4] &= \delta_n^{-1}E[|\lambda'\mathbf{Z}_i(Y_i - \beta_0 - \beta_1 D_i - \beta_2 p^A(X_i;\delta_n))|^4 I_i] \\ &\leq 2^{3c}\delta_n^{-1}E[(|\lambda'\mathbf{Z}_i|^4)(|Y_i|^4 + |\beta_0|^4 + |\beta_1|^4 D_i + |\beta_2|^4 p^A(X_i;\delta_n)^4)I_i] \\ &\leq 2^{3c}(|\lambda_1| + |\lambda_2| + |\lambda_3|)^4\delta_n^{-1}E[(Y_i^4 + \beta_0^4 + \beta_1^4 + \beta_2^4)I_i] \\ &= 2^{3c}O(1)\end{aligned}$$

for some finite constant c, where the last equality holds by Step 3 under Assumption 2.3 (f). 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 7. $n\delta_n\hat{\mathbf{\Sigma}} \xrightarrow{p} S_{\mathbf{D}}^{-1}\mathbf{V}(S_{\mathbf{D}}')^{-1}$.

Proof. We have

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

where the last equality follows from the result that $\hat{\beta}-\beta = o_p(1)$ and from application of Step 3 as in Steps 5 and 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 3 under Assumption 2.3 (f). 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 8. $\hat{\sigma}^{-1}(\hat{\beta}_1 - \beta_1) \xrightarrow{d} \mathcal{N}(0,1)$.

Proof. Let $\beta_n = S_{\mathbf{D}}^{-1}\delta_n^{-1}E[\mathbf{Z}_iY_iI_i]$. We then have

$$\begin{aligned}
\frac{1}{\sqrt{n\delta_n}}\sum_{i=1}^{n}E[\mathbf{Z}_i\epsilon_iI_i] &= \sqrt{n\delta_n}\delta_n^{-1}E[\mathbf{Z}_i(Y_i - \mathbf{D}'\beta)I_i] \\
&= \sqrt{n\delta_n}\delta_n^{-1}E[\mathbf{Z}_i(Y_i - \mathbf{D}_i'\beta_n + \mathbf{D}_i'(\beta_n - \beta))I_i] \\
&= \sqrt{n\delta_n}\delta_n^{-1}\{E[\mathbf{Z}_iY_iI_i] - E[\mathbf{Z}_i\mathbf{D}_i'I_i]\beta_n + E[\mathbf{Z}_i\mathbf{D}_i'I_i](\beta_n - \beta)\} \\
&= \sqrt{n\delta_n}\{(S_{\mathbf{D}} - \delta_n^{-1}E[\mathbf{Z}_i\mathbf{D}_i'I_i])S_{\mathbf{D}}^{-1}\delta_n^{-1}E[\mathbf{Z}_iY_iI_i] \\
&\qquad + \delta_n^{-1}E[\mathbf{Z}_i\mathbf{D}_i'I_i]S_{\mathbf{D}}^{-1}(\delta_n^{-1}E[\mathbf{Z}_iY_iI_i] - S_Y)\} \\
&= \sqrt{n\delta_n}(O(\delta_n)O(1) + O(1)O(\delta_n)) \\
&= O(\sqrt{n\delta_n}\delta_n),
\end{aligned}$$

where we use Step 3 for the second last equality. Thus, when $n\delta_n^3 \to 0$,

$$\begin{aligned}
\sqrt{n\delta_n}(\hat{\beta} - \beta) &= (\frac{1}{n\delta_n}\sum_{i=1}^{n}\mathbf{Z}_i\mathbf{D}_i'I_i)^{-1}\frac{1}{\sqrt{n\delta_n}}\sum_{i=1}^{n}\{(\mathbf{Z}_i\epsilon_iI_i - E[\mathbf{Z}_i\epsilon_iI_i]) + E[\mathbf{Z}_i\epsilon_iI_i]\} \\
&\xrightarrow{d} \mathcal{N}(0, S_{\mathbf{D}}^{-1}\mathbf{V}(S_{\mathbf{D}}')^{-1}).
\end{aligned}$$

The conclusion then follows from Step 7. □

□

2.C.4.4 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 2.5 holds.

Step 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 2.5 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 3 in Section 2.C.4.3, we have

$$\begin{aligned}
&|E[\frac{1}{n\delta_n}\sum_{i=1}^{n} V_i(p^s(X_i;\delta_n)^l - p^A(X_i;\delta_n)^l)I_i]| \\
=& \delta_n^{-1}|E[E[V_i|X_i]E[p^s(X_i;\delta_n)^l - p^A(X_i;\delta_n)^l|X_i]I_i]| \\
\leq& \delta_n^{-1}E[|E[V_i|X_i]||E[p^s(X_i;\delta_n)^l - p^A(X_i;\delta_n)^l|X_i]|I_i] \\
=& \int_{-1}^{1}\int_{\partial\Omega^*\cap N(\mathcal{X},\tilde{\delta})} |E[V_i|X_i = u + \delta_n v\nu_{\Omega^*}(u)]| \\
&\qquad \times |E[p^s(u+\delta_n v\nu_{\Omega^*}(u);\delta_n)^l - p^A(u+\delta_n v\nu_{\Omega^*}(u);\delta_n)^l]| \\
&\qquad \times f_X(u+\delta_n v\nu_{\Omega^*}(u)) J_{p-1}^{\partial\Omega^*}\psi_{\Omega^*}(u,\delta_n v) d\mathcal{H}^{p-1}(u)dv,
\end{aligned}$$

where the choice of $\tilde{\delta}$ is as in the proof of Step 3 in Section 2.C.4.3. By Lemma 2.B.7, for $l = 0, 1, 2,$

$$\begin{aligned}
&|E[\frac{1}{n\delta_n}\sum_{i=1}^{n} V_i(p^s(X_i;\delta_n)^l - p^A(X_i;\delta_n)^l)I_i]| \\
\leq& \frac{1}{S_n}\int_{-1}^{1}\int_{\partial\Omega^*\cap N(\mathcal{X},\tilde{\delta})} |E[V_i|X_i = u + \delta_n v\nu_{\Omega^*}(u)]| \\
&\qquad \times f_X(u+\delta_n v\nu_{\Omega^*}(u)) J_{p-1}^{\partial\Omega^*}\psi_{\Omega^*}(u,\delta_n v) d\mathcal{H}^{p-1}(u)dv \\
=& O(S_n^{-1}).
\end{aligned}$$

Also, by Lemma 2.B.7,

$$
\begin{aligned}
&|E[\frac{1}{n\delta_n}\sum_{i=1}^{n} V_i(p^s(X_i;\delta_n)^3 - p^A(X_i;\delta_n)^3)I_i]| \\
&= |\delta_n^{-1}E[V_i(p^s(X_i;\delta_n) - p^A(X_i;\delta_n)) \\
&\qquad\qquad \times (p^s(X_i;\delta_n)^2 + p^s(X_i;\delta_n)p^A(X_i;\delta_n) + p^A(X_i;\delta_n)^2)I_i]| \\
&\leq \delta_n^{-1}E[|E[V_i|X_i]||E[(p^s(X_i;\delta_n) - p^A(X_i;\delta_n)) \\
&\qquad\qquad \times (p^s(X_i;\delta_n)^2 + p^s(X_i;\delta_n)p^A(X_i;\delta_n) + p^A(X_i;\delta_n)^2)|X_i]|I_i] \\
&\leq 3\delta_n^{-1}E[|E[V_i|X_i]|E[|p^s(X_i;\delta_n) - p^A(X_i;\delta_n)||X_i]I_i] \\
&= \int_{-1}^{1}\int_{\partial\Omega^*\cap N(\mathcal{X},\tilde{\delta})} |E[V_i|X_i = u + \delta_n v\nu_{\Omega^*}(u)]| \\
&\qquad\qquad \times E[|p^s(u + \delta_n v\nu_{\Omega^*}(u);\delta_n) - p^A(u + \delta_n v\nu_{\Omega^*}(u);\delta_n)|] \\
&\qquad\qquad \times f_X(u + \delta_n v\nu_{\Omega^*}(u))J_{p-1}^{\partial\Omega^*}\psi_{\Omega^*}(u,\delta_n v)d\mathcal{H}^{p-1}(u)dv \\
&\leq (\frac{1}{S_n\epsilon^2} + \epsilon)O(1)
\end{aligned}
$$

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,

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

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

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

and for $l = 3, 4$,

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

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 $l \geq 0$. We have

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

Since $I_i^s - I_i \leq 0$ 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

$$
\begin{aligned}
&|E[\frac{1}{n\delta_n}\sum_{i=1}^{n} V_i p^s(X_i;\delta_n)^l(I_i^s - I_i)]| \\
&\leq \delta_n^{-1} E[|E[V_i|X_i]|\Pr(p^s(X_i;\delta_n)\in\{0,1\}|X_i)I_i] \\
&\leq \delta_n^{-1} E[|E[V_i|X_i]|((1-p^A(X_i;\delta_n))^{S_n} + p^A(X_i;\delta_n)^{S_n})I_i] \\
&\leq \int_{-1}^{1}\int_{\partial\Omega^*\cap N(\mathcal{X},\tilde{\delta})} |E[V_i|X_i = u+\delta_n v\nu_{\Omega^*}(u)]|\{(1-p^A(u+\delta_n v\nu_{\Omega^*}(u);\delta_n))^{S_n} \\
&\qquad\qquad + p^A(u+\delta_n v\nu_{\Omega^*}(u);\delta_n)^{S_n}\} f_X(u+\delta_n v\nu_{\Omega^*}(u)) J_{p-1}^{\partial\Omega^*}\psi_{\Omega^*}(u,\delta_n v) d\mathcal{H}^{p-1}(u)dv,
\end{aligned}
$$

where the second inequality follows from Lemma 2.B.7. Note that for every $(u,v)\in\partial\Omega^*\cap N(\mathcal{X},\tilde{\delta})\times(-1,1)$, $\lim_{\delta\to 0} p^A(u+\delta v\nu_{\Omega^*}(u);\delta) = k(v)\in(0,1)$ by Step 1 in Section 2.C.4.3. 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

2.5 holds. Let $\eta_n = \gamma \frac{\log n}{S_n}$, where γ is the one satisfying Assumption 2.5. We have

$$
\begin{aligned}
&|E[\frac{1}{\sqrt{n\delta_n}} \sum_{i=1}^{n} V_i p^s(X_i;\delta_n)^l (I_i^s - I_i)]| \\
&\leq \sqrt{n\delta_n^{-1}} E[|E[V_i|X_i]|((1-p^A(X_i;\delta_n))^{S_n} + p^A(X_i;\delta_n)^{S_n}) I_i] \\
&= \sqrt{n\delta_n^{-1}} E[|E[V_i|X_i]|((1-p^A(X_i;\delta_n))^{S_n} + p^A(X_i;\delta_n)^{S_n}) \\
&\qquad\qquad \times 1\{p^A(X_i;\delta_n) \in (0,\eta_n) \cup (1-\eta_n, 1)\}] \\
&\quad + \sqrt{n\delta_n^{-1}} E[|E[V_i|X_i]|((1-p^A(X_i;\delta_n))^{S_n} + p^A(X_i;\delta_n)^{S_n}) 1\{p^A(X_i;\delta_n) \in (\eta_n, 1-\eta_n)\}] \\
&\leq (\sup_{x \in N(\partial\Omega^*, 2\tilde{\delta}) \cap N(\mathcal{X}, 2\tilde{\delta})} |E[V_i|X_i = x]|)(\sqrt{n\delta_n^{-1}} \Pr(p^A(X_i;\delta_n) \in (0,\eta_n) \cup (1-\eta_n, 1)) \\
&\qquad\qquad + 2\sqrt{n\delta_n}(1-\eta_n)^{S_n} \delta_n^{-1} E[1\{p^A(X_i;\delta_n) \in (\eta_n, 1-\eta_n)\}]).
\end{aligned}
$$

By Assumption 2.5, $\sqrt{n\delta_n^{-1}} \Pr(p^A(X_i;\delta_n) \in (0,\eta_n) \cup (1-\eta_n, 1)) = o(1)$. 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

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

By Step 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 2.5 holds.

By proceeding as in Step 7 in Section 2.C.4.3, 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 1,

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

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 1. The conclusion then follows. □

□

2.C.5 Proof of Proposition 2.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 $x \in \mathbb{R}^p$. □

2.C.6 Proof of Proposition 2.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 $q \leq 1$ for all $q \in Q$ and $\sum_{q \in Q} \mathcal{L}^p(\mathcal{U}_{x,q}) = \mathcal{L}^p(B(\mathbf{0},1))$. □

2.C.7 Proof of Corollary 2.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 2.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 2.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 2.A.2, $p^A(x)$ exists, and $p^A(x) = \frac{q_1\mathcal{L}^2(\mathcal{U}_{\mathbf{0},q_1}) + q_2\mathcal{L}^2(\mathcal{U}_{\mathbf{0},q_2})}{\mathcal{L}^2(B(\mathbf{0},1))} = \frac{3}{4}q_1 + \frac{1}{4}q_2$. □

2.C.8 Proof of Proposition 2.A.3

We can prove Part (a) using the same argument in the proof of Proposition 2.1 (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 1. $\mathcal{L}^{p_c}(\mathcal{X}_c^S(x_d) \cap \mathcal{X}_{c,1}(x_d)) > 0$.

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

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

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

Following the argument in the proof of Proposition 2.1 (b), we can prove Steps 1–3. Once Step 4 is established, we prove Step 5 by following the proof of Step 4 in Section 2.C.1 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 4.

Proof of Step 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 $B(x_c^*, \delta) \subset \mathcal{X}_c^S(x_d) \cap \text{int}(\mathcal{X}_{c,1}(x_d))$. □

□

2.C.9 Proof of Theorem 2.A.1

The proof is analogous to the proof of Theorem 2.1. 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}. □

2.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 $\text{Unif}(-0.5, 0.5)$ 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 $\text{Unif}(-100, 100)$ independently across j, and draw $\tilde{\alpha}_{1j}$, $j = 1, ..., 100$, from $\text{Unif}(-150, 200)$ 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 $\tilde{Z}_i \sim \text{Bernoulli}(0.5)$. 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 80.5% on the test data.

2.E Empirical Policy Application: Details

2.E.1 Hospital Cost Data

We use publicly available Healthcare Cost Report Information System (HCRIS) data, to project funding eligibility and 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. We use the methodology detailed in the CARES Act website to project funding based on 2018 financial year cost reports.

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.[36] In the data, this variable is missing for 1560 hospitals. We impute the disproportionate patient percentage to 0 when it is missing.

36. For the precise definition, see https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/dsh.

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 hospitalâĂŹs bad debt and the financial assistance it provides.[37] 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 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.

2.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

37. The precise definition can be found at https://www.aha.org/fact-sheets/2020-01-06-fact-sheet-uncompensated-hospital-care-cost.

Protect by state/territorial health departments on behalf of health care facilities.[38]

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.[39]

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.[40]

38. Source: https://healthdata.gov/Hospital/COVID-19-Reported-Patient-Impact-and-Hospital-Capa/anag-cw7u.

39. 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.

40. 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

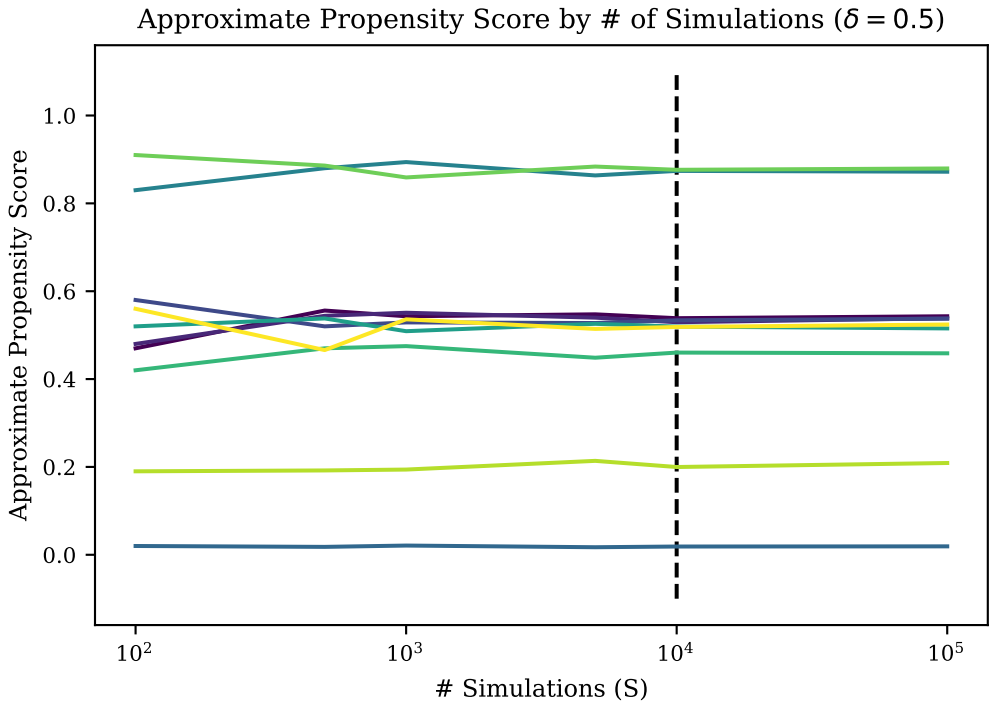
2.E.3 Computing Fixed-Bandwidth Approximate Propensity Score

As the three determinants of funding eligibility are continuous variables, we can think of this setting as a multidimensional regression discontinuity design and a suitable setting to apply our method. In this setting, X_i are disproportionate patient percentage, uncompensated care per bed and profit margin. Funding eligibility (Z_i) is determined algorithmically using these three dimensions. D_i is the amount of funding received by hospital i, which depends on 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.

2.E.4 Additional Empirical Results

impute the suppressed values (coded as -999,999) with 0s. Results are available upon request.

Figure 2.7: 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 2.5: 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. */**/*** indicate $p < 0.10/0.05/0.01$.

Bibliography

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

— and Imbens, G. W. (2006). Large Sample Properties of Matching Estimators for Average Treatment Effects. *Econometrica*, **74** (1), 235–267.

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. (2022). Breaking Ties: Regression Discontinuity Design Meets Market Design. *Econometrica*, **90** (1), 117–151.

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.

Agrawal, A., Verschueren, R., Diamond, S. and Boyd, S. (2018). A Rewriting System for Convex Optimization Problems. *Journal of Control and Decision*, **5** (1), 42–60.

Andrews, I., Kitagawa, T. and McCloskey, A. (2021). Inference on Winners. *Working Paper*.

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

— and Rokkanen, M. (2015). Wanna Get Away? Regression Discontinuity Estimation of Exam School Effects Away From the Cutoff. *Journal of the American Statistical Association*, **110** (512), 1331–1344.

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

— and Kolesár, M. (2021). Finite-Sample Optimal Estimation and Inference on Average Treatment Effects Under Unconfoundedness. *Econometrica*, **89** (3), 1141–1177.

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

— and Wager, S. (2021). Policy Learning With Observational Data. *Econometrica*, **89** (1), 133–161.

Beliakov, G. (2006). Interpolation of Lipschitz Functions. *Journal of Computational and Applied Mathematics*, **196** (1), 20–44.

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.

Bennett, M. (2020). How Far is Too Far? Estimation of an Interval for Generalization of a Regression Discontinuity Design Away from the Cutoff. *Working Paper*.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis; 2nd ed.* Springer Series in Statistics, New York: Springer.

Bertanha, M. (2020). Regression Discontinuity Design with Many Thresholds. *Journal of Econometrics*, **218** (1), 216–241.

— and Imbens, G. W. (2020). External Validity in Fuzzy Regression Discontinuity Designs. *Journal of Business & Economic Statistics*, **38** (3), 593–612.

Bhattacharya, D. and Dupas, P. (2012). Inferring Welfare Maximizing Treatment Assignment Under Budget Constraints. *Journal of Econometrics*, **167** (1), 168–196.

Black, S. E. (1999). Do Better Schools Matter? Parental Valuation of Elementary Education. *Quarterly Journal of Economics*, **114** (2), 577–599.

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

Brown, D., Kowalski, A. E. and Lurie, I. Z. (2020). Long-Term Impacts of Childhood Medicaid Expansions on Outcomes in Adulthood. *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*.

Cai, T. T. and Low, M. G. (2004). An Adaptation Theory for Nonparametric Confidence Intervals. *The Annals of Statistics*, **32** (5), 1805–1840.

Calonico, S., Cattaneo, M. D. and Farrell, M. H. (2018). On the Effect of Bias Estimation on Coverage Accuracy in Nonparametric Inference. *Journal of the American Statistical Association*, **113** (522), 767–779.

—, — 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.

—, Keele, L., Titiunik, R. and Vazquez-Bare, G. (2020). Extrapolating Treatment

Effects in Multi-Cutoff Regression Discontinuity Designs. *Journal of the American Statistical Association*, **0** (0), 1–12.

—, 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.

Chernozhukov, V., Lee, S. and Rosen, A. M. (2013). Intersection Bounds: Estimation and Inference. *Econometrica*, **81** (2), 667–737.

Christensen, T., Moon, H. R. and Schorfheide, F. (2020). Robust Forecasting. *arXiv:2011.03153*.

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*.

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

de Chaisemartin, C. (2021). The Minimax Estimator of the Average Treatment Effect, among Linear Combinations of Estimators of Bounded Conditional Average Treatment Effects. *arXiv:2105.08766*.

Dehejia, R. H. (2005). Program Evaluation as a Decision Problem. *Journal of Econometrics*, **125** (1), 141–173.

— and Wahba, S. (1999). Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs. *Journal of the American Statistical Association*, **94** (448), 1053–1062.

Diamond, S. and Boyd, S. (2016). CVXPY: A Python-embedded Modeling Language for Convex Optimization. *Journal of Machine Learning Research*, **17** (83), 1–5.

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

— and Lewbel, A. (2015). Identifying The Effect of Changing The Policy Threshold in Regression Discontinuity Models. *The Review of Economics and Statistics*, **97** (5), 1081–1092.

Donoho, D. L. (1994). Statistical Estimation and Optimal Recovery. *The Annals of Statistics*, **22** (1), 238–270.

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.

Duflo, E. (2001). Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment. *American Economic Review*, **91** (4), 795–813.

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.

Fan, J. and Yao, Q. (1998). Efficient Estimation of Conditional Variance Functions in Stochastic Regression. *Biometrika*, **85** (3), 645–660.

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

Frölich, M. and 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.

Hirano, K. and Porter, J. R. (2009). Asymptotics for Statistical Treatment Rules. *Econometrica*, **77** (5), 1683–1701.

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

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

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

Ignatiadis, N. and Wager, S. (2021). Confidence Intervals for Nonparametric Empirical Bayes Analysis. *arXiv:1902.02774*.

Imbens, G. and Kalyanaraman, K. (2012). Optimal Bandwidth Choice for the Regression Discontinuity Estimator. *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.

Ishihara, T. and Kitagawa, T. (2021). Evidence Aggregation for Treatment Choice. *arXiv:2108.06473*.

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.

Kallus, N. (2018). Balanced Policy Evaluation and Learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 8909–8920.

— and Zhou, A. (2021). Minimax-Optimal Policy Learning Under Unobserved Confounding. *Management Science*, **67** (5), 2870–2890.

Karlin, S. and Rubin, H. (1956). The Theory of Decision Procedures for Distributions with Monotone Likelihood Ratio. *The Annals of Mathematical Statistics*, **27** (2), 272–299.

Kasy, M. (2016). Partial Identification, Distributional Preferences, and the Welfare Ranking of Policies. *The Review of Economics and Statistics*, **98** (1), 111–131.

— (2018). Optimal Taxation and Insurance Using Machine Learning — Sufficient Statistics and Beyond. *Journal of Public Economics*, **167**, 205–219.

Kawai, K., Nakabayashi, J., Ortner, J. and Chassang, S. (2022). Robust Screens for Non-Competitive Bidding in Procurement Auctions. *Econometrica*, **90** (1), 315–346.

Kazianga, H., Levy, D., Linden, L. L. and Sloan, M. (2013). The Effects of "Girl-Friendly" Schools: Evidence from the BRIGHT School Construction Program in Burkina Faso. *American Economic Journal: Applied Economics*, **5** (3), 41–62.

—, —, — and — (2019). Replication Data for: The Effects of "Girl-Friendly" Schools: Evidence from the BRIGHT School Construction Program in Burkina Faso, TN: American Economic Association [publisher], 2013. Ann Arbor, MI: Inter-university Consortium

for Political and Social Research [distributor], 2019-10-12. https://doi.org/10.3886/E113862V1.

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.

Kitagawa, T. and Tetenov, A. (2018). Who Should Be Treated? Empirical Welfare Maximization Methods for Treatment Choice. *Econometrica*, **86** (2), 591–616.

— and — (2021). Equality-Minded Treatment Choice. *Journal of Business & Economic Statistics*, **39** (2), 561–574.

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.

LaLonde, R. J. (1986). Evaluating the Econometric Evaluations of Training Programs with Experimental Data. *American Economic Review*, **76** (4), 604–620.

Levy, D., Sloan, M., Linden, L. L. and Kazianga, H. (2009). Impact Evaluation of Burkina Faso's BRIGHT Program: Final Report. Mathematica Policy Research, Washington, DC, USA.

Li, L., Chu, W., Langford, J. and Schapire, R. E. (2010). A Contextual-Bandit Approach to Personalized News Article Recommendation. *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.

Low, M. G. (1995). Bias-Variance Tradeoffs in Functional Estimation Problems. *The Annals of Statistics*, **23** (3), 824–835.

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

Manski, C. F. (2000). Identification Problems and Decisions Under Ambiguity: Empirical Analysis of Treatment Response and Normative Analysis of Treatment Choice. *Journal of Econometrics*, **95** (2), 415–442.

— (2004). Statistical Treatment Rules for Heterogeneous Populations. *Econometrica*, **72** (4), 1221–1246.

— (2007). Minimax-regret Treatment Choice with Missing Outcome Data. *Journal of Econometrics*, **139** (1), 105–115.

— (2009). The 2009 Lawrence R. Klein Lecture: Diversified Treatment Under Ambiguity. *International Economic Review*, **50** (4), 1013–1041.

— (2010). Vaccination with Partial Knowledge of External Effectiveness. *Proceedings of the National Academy of Sciences*, **107** (9), 3953–3960.

— (2011a). Choosing Treatment Policies Under Ambiguity. *Annual Review of Economics*, **3** (1), 25–49.

— (2011b). Policy Choice with Partial Knowledge of Policy Effectiveness. *Journal of Experimental Criminology*, **7**, 111–125.

— (2021). Econometrics For Decision Making: Building Foundations Sketched By Haavelmo And Wald. *Econometrica*, **89** (6), 2827–2853.

Mbakop, E. and Tabord-Meehan, M. (2021). Model Selection for Treatment Choice: Penalized Welfare Maximization. *Econometrica*, **89** (2), 825–848.

Mo, W., Qi, Z. and Liu, Y. (2021). Learning Optimal Distributionally Robust Individualized Treatment Rules. *Journal of the American Statistical Association*, **116** (534), 659–674.

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. *Association of the Advancement in Artificial Intelligence (AAAI)*, 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. *International Conference on Machine Learning (ICML)*, pp. 759–766.

Qian, M. and Murphy, S. A. (2011). Performance Guarantees for Individualized Treatment Rules. *The Annals of Statistics*, **39** (2), 1180–1210.

Rambachan, A. and Roth, J. (2020). An Honest Approach to Parallel Trends. *Working Paper*.

Rokkanen, M. (2015). Exam Schools, Ability, and the Effects of Affirmative Action: Latent Factor Extrapolation in the Regression Discontinuity Design. Discussion Paper 1415-03, Columbia University, Department of Economics, New York, NY.

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.

Russell, T. M. (2020). Policy Transforms and Learning Optimal Policies. *arXiv:2012.11046*.

Saito, Y., Aihara, S., Matsutani, M. and Narita, Y. (2021). Open Bandit Dataset and Pipeline: Towards Realistic and Reproducible Off-Policy Evaluation. *Neural Information Processing Systems (NeurIPS)*.

Savage, L. J. (1951). The Theory of Statistical Decision. *Journal of the American Statistical Association*, **46** (253), 55–67.

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.

Song, Z. (2020). Potential Implications of Lowering the Medicare Eligibility Age to 60. *Journal of the American Medical Association (JAMA)*, **323** (24), 2472–2473.

Stein, E. M. and Shakarchi, R. (2005). *Real Analysis: Measure Theory, Integration, and Hilbert Spaces*. Princeton Lectures in Analysis, Princeton, NJ: Princeton Univ. Press.

Stoye, J. (2009). Minimax Regret Treatment Choice with Finite Samples. *Journal of Econometrics*, **151** (1), 70–81.

— (2012). Minimax Regret Treatment Choice with Covariates or with Limited Validity of Experiments. *Journal of Econometrics*, **166** (1), 138–156.

Swaminathan, A. and Joachims, T. (2015). Counterfactual Risk Minimization: Learning from Logged Bandit Feedback. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning*, vol. 37, pp. 814–823.

Tetenov, A. (2012). Statistical Treatment Choice Based on Asymmetric Minimax Regret Criteria. *Journal of Econometrics*, **166** (1), 157–165.

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

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

Zhao, Y., Zeng, D., Rush, A. J. and Kosorok, M. R. (2012). Estimating Individualized Treatment Rules Using Outcome Weighted Learning. *Journal of the American Statistical Association*, **107** (499), 1106–1118.