UNIVERSITY OF CALIFORNIA SAN DIEGO

# Essays on Treatment Effect Heterogeneity and Policy Design

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Economics

by

Yu-Chang Chen

Committee in charge:

Professor Yixiao Sun, Chair
Professor Kaspar Wüthrich, Co-Chair
Professor Gordon Dahl
Professor Young-Han Kim
Professor Craig McIntosh

2022

The Dissertation of Yu-Chang Chen is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2022

DEDICATION

To my parents, San-Pao Chen and Chin-Fen Lee.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor, Yixiao Sun, who has been a role model for my academic career. He has provided me invaluable guidance for my intellectual growth and exemplified the qualities of an eminent researcher.

I would like express my gratitude to Professor Kaspar Wüthrich, who has been nothing but supportive in my intellectual journey. I would also like to thank Gordon Dahl, Craig McIntosh, Young-Han Kim, Graham Elliot, Xinwei Ma, and Ming-Jen Lin for their thoughtful feedback. They have immensely enhanced the quality of the work.

Chapter 1, in full, is submitted for publication. Chen, Yu-Chang; Haitian Xie. "Personalized Subsidy Rules". The dissertation author was the primary investigator and author of this material.

Chapter 2, in full, is a reprint of the material as it appears in Oxford Bulletin of Economics and Statistics. 2009. Chen, Yu-Chang; Xie, Haitian. "Global Representation of the Conditional LATE Model: a Separability Result". The dissertation author was the primary investigator and author of this material.

Chapter 3, in full, is in preparation for submission. Chen, Yu-Chang."Empirical Bayes with Optimal Shrinkage Tree". The dissertation author was the primary investigator and author of this material.

VITA

| | |
|---|---|
| 2014 | Bachelor of Arts, National Taiwan University |
| 2017 | Master of Arts, University of California San Diego |
| 2022 | Doctor of Philosophy, University of California San Diego |

ABSTRACT OF THE DISSERTATION

Essays on Treatment Effect Heterogeneity and Policy Design

by

Yu-Chang Chen

Doctor of Philosophy in Economics

University of California San Diego, 2022

Professor Yixiao Sun, Chair
Professor Kaspar Wüthrich, Co-Chair

My research focuses on the econometrics of policy design in the context of heterogeneous treatment effects. I study both the theoretical and practical aspects of these issues, ranging from equivalent representation of models, identification of optimal policies, and estimation of treatment effects in high-dimensional settings.

Chapter 1 studies the optimal allocation of subsidies based on individual characteristics. We show that the marginal treatment effect is a "sufficient statistics" for the counterfactual welfare and accordingly develop a set of identification results. We also show that subsidy rules weakly dominate treatment rules regarding the attainable social welfare since subsidy rules can

implicitly target individuals based on unobserved heterogeneity.

Chapter 2 establishes that the conditional local average treatment effect (CLATE) models have a selection model representation in which a specific form of separability needs to hold. Based on the representation results, we develop several testable implications of the CLATE model that are sharper than the existing ones in the literature.

Chapter 3 proposes a new empirical Bayes method that utilizes a decision tree to improve the statistical accuracy of the estimates while maintaining interpretability. Instead of the "shrinking toward the grand mean" that conventional empirical Bayes methods typically engage in, the proposed method uses a decision tree to group similar parameters together and shrink each estimator in a group-specific fashion. We also provide a method to select the optimal decision tree that minimizes the estimation errors.

# Chapter 1

# Personalized Subsidy Rules

## 1.1 Introduction

With an aim to encourage self-serving or socially beneficial behaviors, governments worldwide offer various kinds of subsidies to households and individuals. Examples include tuition subsidies for college education, price subsidies for preventive health products, childcare subsidies for certified daycare services, and many others. Due to its relevance and prevalence, a large amount of studies across fields in economics have contributed to the evaluation of subsidy programs. For example, in labor economics, empirical works have exploited exogenous variations In the choice of eduction to estimate the return to tuition subsidy [Ichimura and Taber, 2002, Carneiro et al., 2011]. In development economics, researchers have also conducted experiments to examine the impact of subsidies on the take-up of insecticide-treated bed nets [Cohen and Dupas, 2010].

One important objective of these studies is to inform the policy-maker on the optimal allocation of subsidies. Since both the subsidy's effect on the take-up and its eventual impact on the welfare outcomes could vary across individuals, an efficient subsidy scheme should take the welfare effect, the behavioral response, and the cost of subsidies altogether into account. Intuitively, an ideal allocation of subsidies should increase the take-up among those who are most likely to benefit from it while keeping others away if giving subsidies actually lead to an inefficient outcome.

In this paper, we study subsidy rules which maximizes the welfare of the targeted population by the providing personalized subsidies. A subsidy rule, which we interchangeably refer to as a "policy", is defined as an assignment rule that maps individuals to amounts of subsidies based on their observable characteristics. In line with the treatment choice literature, the notion of welfare we consider is of the class of additive social welfare, which is defined as the average outcome (e.g., employment rate for job-training programs) in the population of interest. Noteworthily, our approach allows flexible specification of the cost functions. The cost of subsidies can be dependent on the take-up, individual characteristics, and the amount of subsidies.

We adopt the marginal treatment effect (MTE) framework [Heckman and Vytlacil, 2005] to study our problem. In our setting, we view subsidies as instrumental variables that only affect the take-up the subsidized behavior (the treatment) but are excluded from the outcome equations. We also follow the tradition in the MTE literature that the instrumental variable (which is the subsidy in our case) is randomly assigned in the population from which our data is sampled. The exclusion restriction and exogeneity of the instrument allow us to causally identify the treatment effect and the individuals's treatment selection as a function of the subsidies. We then seek to identify the subsidy rule that maximizes the counterfactual welfare by allocating subsidies based on heterogeneity.

The MTE framework is suitable for our problem because of its explicit yet flexible modeling of the treatment effects and the treatment take-up. As we will see in our analysis, direct modeling of the treatment selections help improve both the interpretability of the result and the practicality of our procedure. For example, we can incorporate structural assumptions that are backed by economic theory into the model to facilitate the identification and estimation of optimal subsidies. We will demonstrate this by studying the case when there is positive selection in the treatment choice.

We start our analysis by first characterizing the welfare of any subsidy rule using the MTE framework. It is widely known that the marginal treatment effects can serve as building

blocks for other treatment parameters such as average treatment effect (ATE) and policy-relevant treatment effect (PRTE) [Heckman and Vytlacil, 2005] after suitable re-weighting. In line with these results, we show that the welfare of a given subsidy rule can also be expressed by the marginal treatment effects, and we use the expression to derive necessary conditions for a subsidy rule to be optimal. In the optimality condition, it is made clear how the MTE curve, the main parameter of interest in the MTE framework, is connected to the welfare effects of subsidies. Furthermore, we provide simple sufficient conditions for a policy to be optimal when the selection into treatment is monotone.

Given the welfare representation result, the identification of optimal policy becomes straightforward if the marginal treatment effect is known. However, point identification of the MTE curve is not always possible. Therefore, we study conditions that allow one to rank different policies, subject to the fact that the MTE curve may not be identified on its whole support. We also study in-depth the case when the MTE curve is not point-identified but known to be monotone, We show that it is possible to identify the optimal policy even when the MTE curve is not identified on its whole support provided that the selection into the treatment is positive. We also provide primitive conditions from the choice-theoretic perspective to guarantee the monotonicity of the MTE curve.

From the practical perspective, this paper addresses the problem of learning the optimal allocation of subsidies from (quasi-) experimental data. Our characterization result essentially simplifies the optimal policy problem to the identification and estimation of the MTE curve. In this view, we bridge the gap between the policy evaluation problem and the policy design problem. Also, the dual role of subsidies is emphasized: when estimating the effects, subsidies serve as instrumental variables for estimating the treatment effects; and, when used as policy assignments, subsidies can also serve as the subject of assignments. We illustrate this idea with an empirical example on determining the optimal amount for wage subsidies using the experimental data from Groh et al. [2016].

Subsidy-based policies are popular since policies that mandate the treatment are not

always feasible to the policy-maker due to practical issues. This paper establishes results showing that assigning subsidies also achieves higher welfare than directly mandating the treatment, giving another justification of subsidy rules. This is because, since the equilibrium treatment choice is made by individuals, subsidy-based policies implicitly target individuals based on their unobserved (to the policy-maker) heterogeneity. For example, individuals may have private information regarding their returns to higher education, on which their decisions on schooling are based. Consequentially, different subsidies may elicit students with different returns. If the selection into education is positive, smaller amounts of subsidy tend to elicit students with higher returns when all other things being equal. Therefore, a subsidy-based policy, if carefully designed, can leverage individuals' private information on their returns and may achieve higher welfare than compulsory policies which totally neglect this information.

The rest of the paper is organized as follows. Section 1.2 introduces the model setup and the policy design problem. In section 1.3, we present the welfare representation results and optimality conditions for welfare-maximizing policies. A set of results regarding the identifications of optimal rules are stated in section 1.4, including results on ranking policies when MTE is not point identified. In section 1.5, we investigate the welfare properties of subsidy rules, showing that subsidies rules are generally preferred to treatment rules in terms of social welfare. In section 1.6, we provide an empirical example of estimating optimal wage subsidies to illustrate the main ideas of this paper. All proofs are in Appendix.

### 1.1.1 Connection to the literature

This paper aims to contribute to a growing literature on personalized treatment rules, including Manski [2004], Dehejia [2005], Hirano and Porter [2009], Stoye [2009], Bhattacharya and Dupas [2012], Kitagawa and Tetenov [2018], Athey and Wager [2021]. For a recent review of the subject, see Hirano and Porter [2019]. Much of the literature has focused on the assignment of treatments, and the case of assigning subsidies as encouragements to treatment

is less studied.[1] Although it is possible to apply existing methods for treatment assignments to subsidy assignments by adopting an intention-to-treat approach that essentially assigns subsidies based on reduced-form estimates of subsidy effects [Bhattacharya and Dupas, 2012, Kitagawa and Tetenov, 2018], we argue that a selection-model approach that explicitly models treatment choice as a function of the assigned subsidy can have its own advantages. For example, our approach can allow the realized cost of subsidies to be dependent on treatment take-ups, which is not possible for intention-to-treat approaches since a model of treatment take-ups is lacking. Also, by modeling treatment choice explicitly, we can incorporate shape restrictions into the selection equation to aid identification and estimation, such as requiring the subsidy to have positive effects on treatment take-ups [Horowitz and Lee, 2017].

Although the majority of studies on individualized treatment rules consider policy learning in randomized experiments, recent works have looked into cases when endogeneity arises for reasons such as non-compliances or omitted variable bias [Kasy, 2016, Cui and Tchetgen Tchetgen, 2020, Qiu et al., 2020, Athey and Wager, 2021, Byambadalai, 2021, Pu and Zhang, 2021]. The method of instrumental variables is often used for identifications, and this paper is no exception. Similar to Kasy [2016], we study the welfare rankings of policies when the treatment effect is only partially identified, albeit we focus on the assignment of instruments rather than the treatment itself. Pu and Zhang [2021] introduces a new notion of optimality, which they refer to as IV-optimality, for treatment assignment rules that maximize the worst-case welfare among the identification region. They also derive a bound on the loss in the welfare of IV-optimal rules relative to the first-best rule, which assigns treatments whenever the effect is positive. This paper shows that the optimal subsidy rule generally outperforms the first-best policy that assigns treatments.

This paper heavily relies on the MTE framework (Heckman and Vytlacil [1999, 2001, 2005, 2007]) for the theoretical analysis of subsidy rules. The main parameter of interest in

[1] The only exception we can find is Qiu et al. [2020]. They study the optimal assignment of binary instrument variables as "encouragements" to treatment take-ups. This paper can be viewed as an extension in the sense that we allow the instrument to be continuous.

the framework, namely the marginal treatment effects, can be identified by the method of local instrument variables and can later be used for predicting the effects of hypothetical policies. Recent works have proposed new approaches to its identification and estimation [Carneiro and Lee, 2009, Brinch et al., 2017, Mogstad et al., 2018, 2020, Sasaki and Ura, 2021] and to apply MTE framework to various research topics such as unconditional quantile effects [Martínez-Iriarte and Sun, 2020] and external validity [Kowalski, 2018]. Among these, our paper is most closely related to Sasaki and Ura [2020], which also applies MTE to statistical decision rules. However, they focus on the assignment of treatment instead of subsidies. They apply the MTE framework to the method of empirical welfare maximization Kitagawa and Tetenov [2018], in which policies are assumed to lie in a known policy class with finite $VC$-dimension. We do not make such restrictions on candidate policies. In addition, we emphasize more on the identification and welfare properties, while Sasaki and Ura [2020] emphasizes more on the estimation.

## 1.2 Setup

In this section, we introduce our setup. In brief, the policy maker's goal is to find the optimal subsidy assignment rule that assigns amounts of subsidy individually based on a person's covariates We say an assignment rule is optimal if it maximizes the social welfare function, which is defined as the population's average of the outcome variable under the policy. Importantly, the subsidy rule only affects the social welfare through its effect on the behavior response, which can eventually impacts the outcome. Since the majority of our modeling assumptions follow the MTE framework, we first introduce the framework while stating the assumptions we are making.

### 1.2.1 Data generation process: the MTE framework

Let $Y_1$ and $Y_0$ be the potential outcomes under the treatment and control status respectively. The potential outcomes are related to the observable covariates as

$$Y_1 = \mu_1(X, U_1), \text{ and } Y_0 = \mu_0(X, U_0), \tag{1.1}$$

where $X$ is a vector of observed random variables influencing potential outcomes, $\mu_1$ and $\mu_0$ are unknown functions, and $U_1$ and $U_0$ are unobserved random variables. Let D=1 had an individual received the treatment and D=0 had an individual not received the treatment, and $Y = DY_1 + (1-D)Y_0$ as the realized outcome.

In the MTE framework, the treatment take-up is modeled by a latent-index utility model, where the selection into the treatment status depends on the individual characteristics and the instrumental variable. Given $(X, W, Z)$, where $W$ and $Z$ are instrumental variables, the treatment take-up $D$ is determined by

$$D = \mathbf{1}\{g(X, W, Z) \geq U_D\}, \tag{1.2}$$

where $U_D$ is the unobserved heterogeneity in the treatment selection process. We can interpret $U_D$ as resistance to treatment take-ups: holding $(X, W, Z)$ fixed, individuals with lower $U_D$ are more likely to select into the treatment.

In our setup, we distinguish two types of instruments. The first type of instruments, denoted by $Z$, are instruments (or subsidies) that are randomly assigned in the data but could in principle be manipulated by the policy-maker as policy tools. The second type of instruments, denoted as $W$, are instruments that only aid the identification of treatment effects and themselves are not subject to the policy-maker's control.[2]

$Z$ has two roles in our setup. First, $Z$ is an instrumental variable that is exogenously set

[2]It is not necessary to have non-manipulatable instrument $W$ to apply our method.

in the data, facilitating the identification of treatment effects. Second, $Z$ is also a policy tool that the policy-maker can utilize to influence individual's treatment take-ups. An example of $Z$ would be price subsidies to services or goods such as education and preventive health products. And $W$ would be variables such as rainfall, earthquake, or local unemployment rate that are valid instruments but not subject to the control of the policy-maker.

The following assumptions, which are commonly imposed in the MTE literature, will also be used in this paper.

**Assumption 1** (Moment Existence)**.** *The expectations $E[Y_1]$ and $E[Y_0]$ exist, i.e., $E[Y_1] < \infty$ and $E[Y_0] < \infty$.*

**Assumption 2** (Density Existence)**.** *The distribution of $U_D$ is absolute continuous with respect to the Lebesgue measure for every $X = x$.*

**Assumption 3** (Random Assignment)**.** *Conditional on $X$, the random vector $(U_1, U_0, U_D)$ is independent to the random vector $(W, Z)$.*

Following Assumption 2, we can without loss of generality impose the normalization that $U_D \mid X \sim \text{Unif}[0,1]$, as $g(X,W,Z) \geq U_D$ is equivalent to $F_{U_D|X}(g(X,W,Z)) \geq F_{U_D|X}(U_D)$.

In the MTE framework, the main parameter of interest is the "marginal treatment effect", which can be used as building blocks for other conventional treatment effect parameters such as the average treatment effect. Denote $\Delta = Y_1 - Y_0$ as the individual treatment effect. The marginal treatment effect (MTE) is defined as

$$\text{MTE}(x,u) = \mathbb{E}\left[\Delta \mid X = x, U_D = u\right],$$

which can be interpreted as the average treatment effect for individuals at different margins. We can also interpret MTE as the infinitesimal local average treatment effect (LATE) since it is identified by the local instrument variable [Heckman and Vytlacil, 2005]. That is, MTE corresponds to the change in population outcome aroused from an infinitesimal change in the

instrumental variable. Therefore, as we will show later, we can also interpret the MTE as the marginal effect of increasing the subsidy, and MTE plays a fundamental role in our analysis. For example, in the first set of our results, we use MTE to characterize social welfares.

By definition, $\text{MTE}(x,u)$ is the mean treatment effects for individuals with $X = x$ at the selection margin $U_D = u$, where higher $U_D$ implies a lower willingness to select the treatment. Fixing $X = x$, a decreasing MTE curve (along the $u$-dimension) corresponds to the case of "positive-selection", meaning that individuals who benefit more from the treatment are more likely to take the treatment. Positive (or negative selection) selection can often be motivated by economics theory, econometrics specifications, and empirical findings. Typical examples include the choice of education in which individuals with higher returns are more likely to invest in education. While we do not impose the assumption of monotone selection in our main results, we will analyze its implications on the characterization, identification, welfare properties, and estimation of optimal policies in the relevant sections throughout this paper.

We assume that the econometrician observes variables $(Y,X,D,W,Z)$, where $Y = DY_1 + (1-D)Y_0$ is the observed outcome.

**Example 1.1** (Tuition Subsidy)**.** *In the context of the tuition subsidy, one can think of $Y$ as earnings after graduation, $X$ as individual characteristics such as family background, $D$ as levels of education, $W$ as proximity to colleges, and $Z$ as the tuition for attending public college [Kane and Rouse, 1995]. While the government has no direct control over students' place of residence, policy makers may change the tuition subsidy to encourage college enrollment.*

### 1.2.2 Subsidy rules, counterfactual outcome, and welfare

We now describe the policy problem and introduce the definition of subsidy rules and welfare. In our setup, the policy-maker can influence an individual's treatment choice and thus the realized outcome by manipulating the subsidies $Z$. Formally, let $\mathscr{X}$ and $\mathscr{W}$ be supports of $X$

and $W$, a policy $\pi$ is a measurable function

$$\pi : (\mathscr{X}, \mathscr{W}) \to \mathscr{Z}^p \tag{1.3}$$

that maps individuals' characteristics to the action space $Z^p$, [3] where $\mathscr{Z}^p$ is the user-specified set of subsidy assignments under consideration. Unless otherwise stated, $\mathscr{Z}^p$ does not have to be equal to the observed support $\mathscr{Z}$ in the data. An example $\mathscr{Z}^p$ would be $\mathscr{Z}^p = [z_l, z_u] \subset \mathbb{R}$, where the range $[z_l, z_u]$ is specified by the policy maker. Notice that we also allow the subsidy to be negative, which could be thought of as a tax imposed by the policy maker. We denote the set of candidate policies as $\Pi$.

Rather than directly setting a mandatory treatment assignment for each individual, a subsidy rule instead aims at improving the welfare by encouraging individuals to take up the treatment with subsidies. Given a policy $\pi$, we assume that the counterfactual treatment choice is

$$D^\pi = \mathbf{1}\{g(X, W, \pi(X, W)) \geq U_D\} \tag{1.4}$$

and the counterfactual outcome is

$$Y^\pi = D^\pi Y_1 + (1 - D^\pi) Y_0. \tag{1.5}$$

Notice that, compared to the case with no policy intervention (equation 1.2 and 1.1), the only difference is that the variable $Z$ is now replaced by the policy-assigned subsidy $\pi(X, W)$. Equivalently, our definition of counterfactual outcomes implicitly assume the following form of policy-invariances: (1) the structural functions $\mu_0(\cdot), \mu_1(\cdot)$, and $g(\cdot)$; and (2) the distribution of the economic fundamentals $(X, W, U_1, U_0, U_D)$ would remain the same under the policy

[3] Although the instrument variable $W$ does not affect the potential outcomes, we allow the assignment of subsidy to be dependent on $W$ as $W$ has affects on the treatment take-up. Therefore, the optimal subsidy for type $X = x$ may depend on the value of $W$ as well. We will elaborate more on the welfare properties in Section 1.5. Furthermore, we restrict our attention to deterministic policies.

intervention $\pi$.

The cost of subsidies is modeled in the following way. Let $c(x,w,z,d)$ be the cost of assigning subsidy $\pi(x,w) = z$ to $(X = x, W = w)$ individuals who then choose treatment status $D = d$. The counterfactual cost under policy $\pi$ is

$$C^{\pi} = c(X,W,Z,D^{\pi}).$$

Unlike previous works that adopt an intention-to-treat approach [e.g., Kitagawa and Tetenov, 2018], our framework allows the realized cost of subsidies to be dependent on individuals's treatment choice $D$. That is, we allow the cost to be endogenous to the individual's decision. This is possible in our framework since, unlike intention-to-treat approaches, treatment choice is explicitly modeled, and the propensity score is estimated. We assume that the cost function $c(\cdot)$ is known to the policy-maker despite that the realized cost for each individual is ex-ante unknown.

**Example 1.2** (Constant Cost)**.** *Kitagawa and Tetenov [2018] studies the optimal eligibility rule for receiving subsidized training in the Job Training Partnership Act (JTPA) program. In their welfare calculation, they impute the cost of the program as* \$774 *for each eligible individual regardless of the actual take-up. In the notation of this paper, their cost function is effectively* $c(x,w,z,d) = c(z)$*, which is only a function of the policy assignment. However, as reported in Bloom et al. [1997], the program take-up varies substantially across different gender and age groups, implying that the realized cost is in fact heterogeneous.*

**Example 1.3** (Voucher Cost)**.** *Following Example 1.2, a more realistic cost function would be* $c(x,w,z,d) = z \cdot d$*, where z is the amount of subsidy paid by the government. Just like vouchers which are covered only when redeemed, the dependence on the treatment choice d reflects that the subsidy is only paid when the individual actually attends the program. Notice that the heterogeneity of treatment take-up is already embedded in this cost function since both*

*the treatment choice and the propensity score are dependent on the covariates as specified in equation (1.4).*

Following previous works in the treatment assignment literature [see e.g., Manski, 2009, Kitagawa and Tetenov, 2018, Athey and Wager, 2021], we adopt the additive welfare criterion to evaluate performance of policies. Given the cost function $c(\cdot)$, the welfare $S(\pi)$ of a policy $\pi$ is defined as

$$S(\pi) = \mathbb{E}[Y^\pi] - \mathbb{E}[C^\pi]. \tag{1.6}$$

The additive welfare criterion is flexible that it can cover various social preferences by suitably transforming the outcome variable. For example, let $\nu(\cdot)$ be a concave function, we can accommodate inequality-averse preference by replacing $Y$ by $\nu(Y)$ [Atkinson, 1970].[4] Alternatively, we can interpret welfare function $\mathbb{E}[\nu(Y^\pi)] - \mathbb{E}[C^\pi] = \mathbb{E}[\nu(Y^\pi) - C^\pi]$ as the average social function in which individuals have quasi-linear preferences.

Given the set of possible subsidy assignments $\mathscr{Z}^p$, the policy class $\Pi$, and the social welfare function $S(\pi)$, a subsidy rule $\pi^* \in \Pi$ is said to be optimal if it attains the social optimum, namely,

$$S(\pi^*) = \sup_{\pi \in \Pi} S(\pi).$$

For the ease of expositions, we will present our results with cost function $c(x,w,z,d) = z \cdot d$ in the main text. Results for general cost functions can be found in the appendix and the proof therein.

**Remark.** *The optimal policy may not be unique in general. For example, in the trivial case that the treatment effect is always zero* $(Y_1 = Y_0)$ *with zero subsidy cost, any subsidy rule is optimal.*

[4] By maximizing $\mathbb{E}[Y^\pi]$, the optimal policy necessarily maximizes $V(\mathbb{E}[Y^\pi])$ for any $V(\cdot)$ that is an increasing transformation. The invariance property helps dealing with cases when welfare can not be represented by simple averages of individual outcomes. For example, in the insecticide-treated bednet example where the policy goal is to increase the usage of nets ($Y$), we can incorporate externality by choosing $V(\cdot)$ that maps the coverage rate $\mathbb{E}[Y^\pi]$ to the counterfactual infection rate $V(\mathbb{E}[Y^\pi])$. Such $V(\cdot)$ arguably exists if the externality is approximately determined by the average coverage of nets.

*Moreover, if any of the targeting variables is a continuous random variable, we can always construct new optimal rules by modifying existing ones on a measure-zero set without changing the implied welfare.*[5]

## 1.3 Optimality Conditions

### 1.3.1 Necessary conditions for optimality

Our first proposition shows that the welfare of a policy can be expressed by MTE and the propensity score.

**Lemma 1** (Characterization of Welfare)**.** *Under Assumptions 1 and 2, we have*

$$\mathbb{E}[Y^{\pi}] = \mathbb{E}[Y_0] + \mathbb{E}\left[\int_0^{g(X,W,\pi(X,W))} MTE(X,u)du\right], \tag{1.7}$$

*and*

$$\mathbb{E}[C^{\pi}] = \mathbb{E}[\pi(X,W)g(X,W,Z)] \tag{1.8}$$

*for the cost function* $c(x,w,z,d) = z \cdot d$.

Results for general cost functions can be found in the appendix. Lemma 1 states that the welfare of a policy has three parts: the baseline outcome under no treatment, the treatment effects on individuals who are induced into treatment status, and the expected costs of subsidies for the treatment takers. Notice that the baseline outcome $\mathbb{E}[Y_0]$ is irrelevant for welfare comparison between policies since it is unaffected by the policy.

Using Lemma 1, we can find the optimal policy $\pi^*$ by optimizing the welfare function pointwisely for each combination of $(x,w)$. Notice that, as implied by the choice equation (1.4), if

[5]For the same reason, any characterization of optimal policies necessarily only apply almost surely outside a measure-zero set if at least one of the targeting variable is a continuous random variable.

the policy maker assigns subsidy $z^p = \pi(x,w)$ for $(X = x, W = w)$ individuals, the counterfactual take-up rate $u^{\pi}_{x,w}$ among these individuals would be given by

$$\begin{aligned} u^{\pi}_{x,w} &= P(D^{\pi} = 1 \mid X = x, W = w) \\ &= P(g(x,w,\pi(x,w)) \geq U_D \mid X = x, W = w) \\ &= g(x,w,\pi(x,w)). \end{aligned}$$

When the propensity score $g(x,w,z)$ is invertible in the third argument $z$, say, when an increase in subsidy always induce more individuals into the treatment status, there is a one-to-one mapping between the amount of subsidy $\pi(x,w)$ and the counterfactual take-up rate $u^{\pi}(x,w)$. In other words, the connection between the two is governed by the propensity score, and we can apply change of variables to simplify the problem. We first solve the optimal take-up rate problem

$$u^{\pi^*}_{x,w} \in \underset{u_{x,w} \in I_{x,w}}{\operatorname{argmax}} \int_0^{u_{x,w}} \mathrm{MTE}(x,u')du' - c(x,w,g^{-1}_{x,w}(u_{x,w}),1) \cdot u_{x,w}, \tag{1.9}$$

where $I_{x,w} = \{g(x,w,z) : z \in \mathscr{Z}^p\}$ is the image of $g(x,w,\cdot)$ and $g_{x,w}(z) = g(x,w,z)$. $I_{x,w}$ reflects to what degree the policy-maker can influence the treatment take-ups through manipulating the subsidy $Z$, and, $g^{-1}_{x,w}(u)$ is the amount of subsidy needed to induce a take-up rate of $u$. Notice that the integration in equation (1.9) starts from 0 as individuals with low $U_D$ are always induced first.

To avoid technical issues, we assume $I_{x,w}$ is a closed set. We also assume that the propensity score $g_{x,w}(z)$ is strictly increasing in $z$ so that $g_{x,w}(z)$ is invertible. The two assumptions, along with other continuity assumptions stated below, ensure that the optimization defined in equation (1.9) is well-defined and that it has a solution.

**Assumption 4** (Continuity)**.** *The propensity score $g(x,w,z)$ and the cost function $C(x,w,z)$ are continuous in the subsidy z, and the marginal treatment effects $MTE(x,u)$ is continuous in u.*

**Assumption 5** (Invertibility)**.** *The propensity score $g(x,w,z)$ is strictly increasing in $z \; \forall \; x \in \mathscr{X}, w \in \mathscr{W}$ .*

While both variables $X$ and $W$ are used for targeting, they play different roles in the policy problem since $W$ is excluded from the outcome equation. Unlike $X$, which underlies the treatment effect heterogeneity, $W$ only enters the optimization problem through the feasible region $I_{x,w}$ and the propensity score $g(x,w,z)$. That is to say, the only value of $W$ as a targeting variable comes from its effect on treatment take-up, which can not be neglected when the policy is made to provide incentives. On the contrary, if the policy were to assign the treatment $D$ instead, it is needless to target on variable $W$ since the treatment effect does not depend on $W$, a property we would elaborate later.

Once we find optimal take-up rate $u^*_{x,w}$, the second step is to find the optimal subsidy level $\pi^*(x,w)$ that achieves $u^*_{x,w}$ in the sense that

$$g(x,w,\pi^*(x,w)) = u^*_{x,w}. \tag{1.10}$$

By construction, $u^*_{x,w}$ is always achieved by some subsidy level $\pi^*(x,w) \in \mathscr{Z}$ since $u^*_{x,w} \in I_{x,w}$. In fact, $\pi^*(x,w)$ is unique as the propensity score is strictly increasing in the amount of subsidy.

The next proposition combines the above arguments and states the necessary condition for optimal policies.

**Proposition 1.1** (Optimality Condition)**.** *Suppose that Assumptions 1 - 5 hold, cost function $c(x,w,z,d) = z \cdot d$, and the action space $\mathscr{Z}^p$ is a closed interval $[z_l, z_u] \subset \mathbb{R}$. If $\pi^*(\cdot)$ is an optimal policy, then $\pi^*(\cdot)$ either satisfies $\Lambda(x,w,\pi^*(x,w)) = 0$, where*

$$\Lambda(x,w,z) = MTE(x, g(x,w,z)) - z - g(x,w,z) \cdot \left[\frac{d}{dz} g(x,w,z)\right]^{-1}, \tag{1.11}$$

*or $\pi^*(x,w) \in \{z_l, z_u\}$.*

We leave the result for the general cost function $c(\cdot)$ in the proof. The definition of $\Lambda(x,w,z)$ has a familiar marginal revenues minus marginal cost interpretation that arises from a monopolist's profit maximization problem. The the first term is the marginal revenue, and

the last two terms represent the marginal cost of subsidy, which depends on the elasticity of treatment take-up.

As we can see in equation (1.11), MTE can be interpreted as the average marginal benefit of increasing the amount of subsidy for individuals with $(X = x, W = w)$. So, MTE is not only a treatment effect parameter as commonly understood, but MTE per se is also a policy-relevant parameter in the context of personalized subsidy rule.[6]

**Example 1.4.** *Suppose X is constant so we ignore it in this example. Let* $\mathscr{Z} = \mathscr{W} = [0,1]$. *The treatment response is* $g(z,w) = \frac{1}{4}(1+z+w)$. *Let the marginal treatment effect be* $MTE(u) = 4-2u$, *which is decreasing. The image of* $g(\cdot,w)$ *is* $I_w = [\frac{1}{4}(1+w), \frac{1}{4}(2+w)]$. *In this case, the optimal policy is* $\pi^*(w) = 1 - \frac{3}{5}w$.

### 1.3.2 Sufficient conditions when MTE is monotone

By definition, $\text{MTE}(x,u)$ is the mean treatment effects for individuals with $X = x$ at the selection margin $U_D = u$, where higher $U_D$ implies a lower willingness to select the treatment. Fixing $X = x$, a decreasing MTE curve (along the $u$-dimension) corresponds to the case of "positive-selection", meaning that individuals who benefit more from the treatment are more likely to take the treatment.

**Assumption 6** (Positive Selection)**.** *The selection process is said to be positive if MTE(x,u) is weakly decreasing in u.*

**Assumption 7** (Negative Selection)**.** *The selection process is said to be negative if MTE(x,u) is weakly increasing in u.*

Monotonicity of MTE can be motivated by economic theory, implied by econometrics specification, and sometimes supported by empirical evidence [see e.g., Carneiro et al., 2011,

[6] Similar connection between MTE and policy effects has been made in the literature. For example, Carneiro et al. [2010] shows that the marginal policy-relevant treatment effect (MPRTE) is a weighted average of MTE. However, in our case, MTE itself but not the average of it is shown to be the marginal effects of subsidies. The distinction appears since we study personalized subsidy rules, whereas they consider universal changes in the amount of subsidy.

Cornelissen et al., 2018]. We introduce a few examples.

**Example 1.5** (Normal Selection Model)**.** *Suppose* $Y_1 = X'\beta_1 + U_1$, $Y_0 = X'\beta_0 + U_0$, *and* $D = \mathbf{1}\{Z'\theta \geq U_D\}$. *Further assume that* $(U_1, U_0, U_D)$ *is jointly normally distributed and independent to* $(X, Z)$, *and the variance of* $U_D$ *is normalized to one. Then* $MTE(x,u) = x'(\beta_1 - \beta_0) + (\sigma_{1D} - \sigma_{0D})\Phi^{-1}(u)$, *where* $\sigma_{1D} = Cov(U_1, U_D), \sigma_{0D} = Cov(U_0, U_D)$, *and where* $\Phi^{-1}(\cdot)$ *is the inverse of the standard normal cumulative function. For extensions to non-normal selection models, see Heckman et al. [2003].*

**Example 1.6** (Roy Model)**.** *In the Roy [1951] model, the treatment take-up is fully determined by the potential gain in the sense that* $D = \mathbf{1}\{\Delta \geq 0\}$ *where,* $\Delta = Y_1 - Y_0$. *Let* $U_D = F_\Delta(\Delta) \sim Unif[0,1]$ *be the normalized gain. Then* $MTE(u) = \mathbb{E}[Y_1 - Y_0 | U_D = u] = F_\Delta^{-1}(u)$.

**Example 1.7** (Generalized Roy Model with Positive Selection)**.** *Consider a selection model in which the treatment take-up is partially determined by the potential gain in the form* $D = \mathbf{1}\{\phi(X, W, Z, \Delta, V) \geq 0\}$, *where* $\Delta = Y_1 - Y_0$ *is the potential gain and V represents the unobserved heterogeneity. In Appendix A.2, we show that if* $\phi(X, W, Z, \Delta, V)$ *is increasing in* $\Delta$, *then we can find functions* $(\phi_1, \phi_2)$ *and random variables* $U_D \sim Unif[0,1]$ *such that* $D = \mathbf{1}\{\phi_1(X, W, Z) \geq \phi_2(\Delta, U_D)\}$ *and* $MTE(x,u) = \mathbb{E}[\Delta | X = x, U_D = u]$ *is decreasing in u.*

Our next proposition characterizes the optimal policy when the selection is monotone.

**Proposition 1.2** (Optimality under Positive Selection)**.** *Suppose that Assumptions 1 - 6 hold, the action space* $\mathscr{Z}^p = [z_l, z_u]$, *and* $g(x,w,z)$ *is weakly concave in z. Further assume that the cost function* $c(x,w,z,d) = z \cdot d$.*Then the optimal subsidy* $\pi^*(x,w)$ *is given by*

$$\pi^*(x,w) = \begin{cases} z_l & \text{if } \Lambda(x,w,z) < 0, \ \forall\, z \in [z_l, z_u] \\ z^* & \text{if } \Lambda(x,w,z^*) = 0 \text{ for some } z^* \in [z_l, z_u] \\ z_u & \text{if } \Lambda(x,w,z) > 0, \ \forall\, z \in [z_l, z_u] \end{cases}, \tag{1.12}$$

*where* $\Lambda(x,w,z) = MTE(x, g(x,w,z)) - z - g(x,w,z) \cdot \left[\frac{\partial g(x,w,z)}{\partial z}\right]^{-1}$.

Proposition 1.2 is a complete characterization of the optimal policy as one of the three cases in equation (1.12) must hold. The characterization stems from the fact that, when the selection is positive, the subsidy's marginal return decreases since individuals with higher returns are always induced first. Corner solutions arise if the marginal return is always positive or negative. We can also characterize the optimal policy for the case of negative selection, albeit under the assumption of no cost $c(x,w,z,d) = 0$.

**Proposition 1.3** (Optimality under Negative Selection)**.** *Suppose that Assumptions 1 - 5 and 7 hold, the action space $\mathscr{Z}^p = [z_l, z_u]$, and $MTE(x,u)$ is weakly-increasing in u. Furthermore, assume $c(x,w,z,d) = 0$. Then the optimal subsidy $\pi^*(x,w)$ is given by*

$$\pi^*(x,w) = \begin{cases} z_l & \text{if } \int_{g(x,w,z_l)}^{g(x,w,z_u)} MTE(x,u)du \leq 0 \\ z_u & \text{if otherwise} \end{cases}. \tag{1.13}$$

In the case of negative selection, individuals who gain the least from the treatment are always induced first, and the marginal return of subsidy is increasing. Therefore, unless the treatment effect is zero, the optimal subsidy is always a corner solution- it either assigns the highest subsidy under consideration so that individuals with high returns are persuaded to take-up the treatment; or, it assigns the least amount of subsidy so that the potential harm caused by the treatment is minimized for the low-return individuals.

We conclude this section with two remarks. First, as we can see from Example 1.4, even though the instrument $W$ is excluded from the outcome equation, $W$ is still valuable for targeting because it affects the selection into treatment.[7] Second, while the variable $U_D$ is not observed and therefore infeasible for targeting, the policy-maker can partially induce individuals into treatment based on $U_D$ by manipulating $Z$: all else being equal, higher level of subsidy would

[7]Notice that the optimal subsidy is decreasing in $w$ in the example for two reasons. First, individuals with high $w$ are ex-ante more likely to select the treatment, therefore less subsidy is needed. Second, since there is positive selection into the treatment (MTE is decreasing), it is of less interest to induce the high resistance (high $u$) individuals into the treatment status.

induce more individuals with high $U_D$ into treatment. The capability to implicitly targeting with $U_D$ will generally increase the welfare, which is advantage not shared with policies that directly assign treatment status. We elaborate more on the welfare properties of subsidy-based policies in Section 1.5.

## 1.4 Identifying Welfare Rankings

The analysis in Section 1.3 provides conditions for optimal policies in terms of MTE as though the MTE is known to the policy-maker. In this section, we take the identification of MTE into account and address the issue of identifying the optimal policy. We first discuss two scenarios in which the optimal policy can be point-identified and provide results on partial rankings when point identification is not possible.

By identification, we aim at representing the objects of interest, such as welfare ranking and optimal policy, by the joint distribution of $(Y,D,X,W,Z)$. The analysis is conducted under the full knowledge of observable distributions, namely, the joint distribution of $(Y,D,X,W,Z)$. The welfare ranking $\succsim$ is an order on the policy space $\Pi$ that $\pi \succsim \pi'$ if and only if $S(\pi) \geq S(\pi')$. The identified ranking is said to be partial if for some pair of policies $(\pi,\pi')$ it is impossible to determine whether $S(\pi) \geq S(\pi')$ given the observable distribution.

### 1.4.1 Point identification of optimal policies

It follows directly from the welfare representation result that, if the entire MTE curve and propensity score $g$ are identified, then both the welfare ranking and optimal policy are identified.

**Corollary 1.** *If for some $(x,w) \in \mathscr{X} \times \mathscr{W}$, $MTE(x,\cdot)$ is identified on $[0,1]$ and $g(x,w,\cdot)$ is identified on $\mathscr{Z}^p$, then the optimal subsidy $\pi^*(x,w)$ is also identified.*

Heckman and Vytlacil [2005] shows that MTE can be identified using Local Instrument Variables (LIV). For completeness and ease of discussion of later results, we first rephrase the identification result of MTE in our setup. Define a function $m$ that is identified from the data as

follows:

$$m(x,u) = \frac{d}{dp}\Big|_{p=u} \mathbb{E}[Y \mid X = x, g(X,W,Z) = p] \cdot \mathbf{1}_{\text{Supp}(g(x,W,Z))}(u).$$

The function $m$ equals to MTE on $\text{Supp}(X, g(X,W,Z))$ and 0 elsewhere according to the following identification result from Heckman and Vytlacil [2005].

**Lemma 2** (Identification of MTE)**.** *Suppose that Assumptions 1 - 3 hold. Further, assume that* $g(X,W,Z)$ *is a nondegenerate random variable conditional on* $X$ *and that* $0 < P(D=1)|X) < 1$. *We have* $MTE(x,u) = m(x,u) \; \forall \; (x,u) \in Supp(X, g(X,W,Z))$.

For the entire MTE curve $MTE(\cdot)$ to be identified, the support of the propensity score $Supp(g(x,W,Z))$ has to cover unit interval for every $x \in \mathscr{X}$ so that the entire MTE curve can be linked the to identified function $m(\cdot)$. Effectively, this requires that there is large enough variation in the support of the instruments $(W,Z)$ such that the propensity score can be arbitrarily close to zero or one depending on the realized value of the instrument variables. As for the identification of $g(\cdot)$, since the propensity score is simply the observed probability of take-up given the covariates and instruments, we can identify $g(\cdot)$ on the support $\mathscr{X} \times \mathscr{W} \times \mathscr{Z}$ by its empirical counterpart. When $Z^p \not\subset Z$, identification of $g(\cdot)$ on $\mathscr{X} \times \mathscr{W} \times \mathscr{Z}^p$ can be further obtained via imposing parametric restriction on $g()$ such as the probit model.

However, point identification of the MTE curve is not necessary for identifying the optimal policy. We present two scenarios in which point identification of the MTE curve is not needed for identifying the optimal policy. First, if the policies under consideration assign subsidies only from the support of $Z$, that is, when $\mathscr{Z}^p \subset \mathscr{Z}$, then the optimal policy is identified. As we show in the next proposition, it is possible to identify the optimal policy without the identification of MTE. Define $\mathscr{P}^{id} = \{\pi \in \Pi : \pi(x,w) \in \text{Supp}(Z \mid X = x, W = w) \text{ for all } (x,w) \in Supp(X,W)\}$.

**Proposition 1.4.** *Suppose that Assumptions 1-3 hold. Then for any* $\pi \in \mathscr{P}^{id}$*,*

$$\mathbb{E}[Y^{\pi} \mid X, W] = \mathbb{E}[Y \mid X, W, Z = \pi(X, W)], \tag{1.14}$$

*and*

$$\mathbb{E}[C^{\pi} \mid X, W] = \pi(X, W) \cdot \mathbb{E}[D \mid X, W, Z = \pi(X, W)]], \tag{1.15}$$

*where* $\mathbb{E}[Y|X, W, Z = \pi(X, W)]$ *and* $\mathbb{E}[D \mid X, W, Z = \pi(X, W)]$ *are identified as* $\pi(x, w) \in Supp(Z \mid X = x, W = w)$*. As a result, the welfare ranking on* $\mathscr{P}^{id}$ *is identified.*

Proposition 1.4 essentially states the counterfactual welfare can be identified by the empirical welfare as long as the subsidy under consideration is observed in the data.

A second scenario in which point identification of the MTE curve is not needed is when individuals positively select into the treatment status. Under the assumption of positive selections, individuals with higher returns are always induced first by the subsidies. Therefore, we know an amount of subsidy is optimal if the marginal effect of subsidy is zero.

**Proposition 1.5.** *Suppose the assumptions stated in Proposition 1.2 hold. If there exists* $z^* \in \mathscr{Z}^p$ *such that* $g(x, w, z^*)$ *and* $MTE(x, g(x, w, z^*))$ *are identified and that* $\Lambda(x, w, z^*) = 0$*, then* $\pi^*(x, w) = z^*$*.*

Proposition 1.5 states that, if the selection is positive, we can identify the optimal amount subsidy as long as we know at which point the marginal effect is zero. Therefore, we do not necessarily need the instruments to have large support if it contains the point that has zero marginal effect. Even if the requirement is not met, imposing positive selection still has identification power, as we will show in the next subsection.

### 1.4.2 Partial rankings of policies

While the results in the previous subsections yield point identifications, the requirements can be restrictive. Next, we discuss how to obtain partial rankings of policies by imposing shape restrictions.

Let $\mathscr{M}^o$ and $\mathscr{G}^o$ be sets of functions that represents the functional parameter space of MTE and propensity under possible shape restrictions (e.g. parametric model, monotonicity, boundedness). The true MTE is assumed to be an element of $\mathscr{M}^o$ and the true propensity is assumed to be an element of $\mathscr{G}^o$. Moreover, the true MTE must coincide with the identifiable function $m$ on the identified region. Therefore, the identified set of MTEs $\mathscr{M}$ and propensities $\mathscr{G}$ under shape restrictions are respectively

$$\mathscr{M} = \left\{ \bar{m}(x,u) \in \mathscr{M}^o : \bar{m}(x,u) = m(x,u) \text{ for all } (x,u) \in \text{Supp}(X, g(X,W,Z)) \right\}$$

and

$$\mathscr{G} = \left\{ g \in \mathscr{G}^o : g(x,w,z) = \mathbb{E}[D \mid X = x, W = w, Z = z], (x,w,z) \in \text{Supp}((X,Z,W)) \right\}.$$

Let $\langle \cdot, \cdot \rangle$ be the inner product with respect to the measure underlies the random vector $(X, U_D)$. Define the dual cone and polar cone of $\mathscr{M}$ respectively as

$$\mathscr{M}^* = \{ \ell : \langle l, \bar{m} \rangle \geq 0, \bar{m} \in \mathscr{M} \}, \text{ and } \mathscr{M}^\times = -\mathscr{M}^*. \tag{1.16}$$

For any policy $\pi$ and propensity $g$, we use $F_{g,\pi}(x,u)$ to denote the conditional cumulative distribution function (CDF) of the propensity score $g(X,W,\pi(X,W))$ given $X = x$, that is, $F_{g,\pi}(x,u) = \mathbb{P}(g(X,W,\pi(X,W)) \geq u \mid X = x)$.

**Proposition 1.6** (Geometry Representation of Policy Effects)**.** *Suppose that Assumptions 1-3 hold. Let $(\pi, \pi')$ be a pair of policies. If $\{F_{g,\pi'} - F_{g,\pi} : g \in \mathscr{G}\} \subset \mathscr{M}^*$, then $\mathbb{E}[Y^\pi] \geq \mathbb{E}[Y^{\pi'}]$. If*

$\{F_{g,\pi'} - F_{g,\pi} : g \in \mathscr{G}\} \subset \mathscr{M}^{\times}$, *then* $\mathbb{E}[Y^{\pi'}] \geq \mathbb{E}[Y^{\pi}]$.

This result for incentive-based policy is the analog of Proposition 1 in Kasy [2016].[8] It provides the general identification result of the welfare ranking by considering the geometry in the Hilbert space containing the MTE curves that are data-consistent. For a pair of policies, if the difference between the induced CDFs of the propensity score is orthogonal to the set of plausible MTEs, then the data is uninformative about the welfare ranking between these two policies. The identified welfare rankings constitute an incomplete ordering that admits an expected utility representation [Dubra et al., 2004]. As a consequence, the identified ranking satisfies the independence axiom.

Our next proposition states that, when there is positive selection, we have an easy-to-interpret partial identification result for policy rankings.

**Proposition 1.7** (Direction of Welfare Improvement)**.** *Suppose the assumptions in Proposition 1.2 hold, and let* $\pi^*$ *be an optimal policy. If* $\Lambda(x, g(x,w,z)) \geq 0$ *(resp.* $\leq 0$*) for some* $z \in \mathscr{Z}^p$*, then* $\pi^*(x,w) \geq z$ *(resp.* $\leq z$*).*

The same intuition behind Proposition 1.2 applies here as well- when the selection-into-treatment is positive, the marginal benefit of subsidy is decreasing. Recall that the function $\Lambda(x,w,z)$ only depends on the MTE and the propensity score, and it represents the marginal return of subsidy. Therefore, if the MTE and the propensity score are identified at a specific point $(x,w,z)$, then the optimal subsidy can be bounded from below if the marginal return at that point is positive and can be bounded from above when the marginal return is negative.

## 1.5 Welfare Properties of Subsidy Rules

In economics, instrumental variables are typically used for the identification of treatment effects. However, we argue that the instrument variable also has a fundamental influence on the

[8] Kasy [2016] focuses on the class of direct policies, which directly assigns the (probability of receiving) treatment rather than using incentives to let individuals self-select.

policy design through its function of providing incentives for the treatment take-up. Explicitly, we show that assigning subsidies weakly dominates assigning treatments directly. For ease of presentation, we condition on $X = x$ throughout this subsection, meaning that $g$ and $\pi$ are only functions of the instrumental variables $W$ and $Z$, and MTE is only a function of $u$.

As a benchmark, we introduce a new class of policies in which the policy-maker does not manipulate the subsidy $Z$. These policies, which we refer to as "direct policies',' set the propensity of receiving the treatment based on the observed characteristics $W$ and $Z$ (By contrast, subsidy-based policies can only affect the treatment status indirectly through changing subsidies). Let $T$ denote the treatment status under direct policies. In particular, $T$ is $\sigma(W,Z)$-measurable Bernoulli random variable with parameter $\tau_T(W,Z) = \mathbb{P}(T = 1 \mid W,Z)$, and $T \perp (U_1,U_0,U_D)$ since $(W,Z) \perp (U_1,U_0,U_D)$. Another class of policies of interest, which we refer to as "constant policies", is a collection of direct policies that assign a constant value of propensity for all $(w,z) \in Supp(W,Z)$. That is, the class of constant policies consists of direct policies that has no personalization. Let $Y^T = TY_1 + (1-T)Y_0$ be the counterfactual outcome from a direct policy $T$, and let

$$
\begin{aligned}
S^*_{\text{sub}} &= \sup_{\pi\text{: } \sigma(W)\text{-measurable}} \mathbb{E}[Y^\pi \mathbf{1}\{U_D \in \text{Supp}(g(W,Z))\}] \\
S^*_{\text{dir}} &= \sup_{T\text{: } \sigma(W,Z)\text{-measurable}} \mathbb{E}[Y^T \mathbf{1}\{U_D \in \text{Supp}(g(W,Z))\}] \\
S^*_{\text{con}} &= \sup_{T\text{: } \tau_T(w,z)=\tau_T\in[0,1]} \mathbb{E}[Y^T \mathbf{1}\{U_D \in \text{Supp}(g(W,Z))\}]
\end{aligned}
$$

be the optimal *identified* welfare under different policy settings.[9] The subscript "sub" represents "subsidies", and $S^*_{\text{sub}}$ is the optimal identified welfare when the policy-maker can design subsidy-based policies using subsidy $Z$. The subscript "dir" represents "direct", and $S^*_{\text{dir}}$ is the optimal identified welfare when the policy-maker do not manipulate $Z$ but instead directly assign

[9] In this section, the cost part of the welfare will be ignored as it is unclear how to compare the cost of assigning treatment and the cost of assigning subsidies without specific context.

treatments based on different values of the instrument.[10] The subscript "con" denotes "constant", and $S^*_{\text{con}}$ is the optimal identified welfare when the policy-maker sets the same propensity of treatment for all individuals. By construction, $S^*_{\text{con}} \leq S^*_{\text{dir}}$ since every constant policy is a direct policy.

We restrict the comparison of optimal welfare on the set of individuals whose $U_D$ lies in the region $\text{Supp}(g(W,Z))$, on which the treatment effect can be identified. Individuals outside this region are either "always-takers" or "never-takers" since their treatment take-up can not be affected given the exogenous variation in the instrumental variables. Therefore, the data is inherently uninformative on the treatment effects as well as the counterfactual welfare under different policies for these individuals. Accordingly, we restrict our attention to the identified welfare $\mathbb{E}[Y^\pi \mathbf{1}\{U_D \in \text{Supp}(g(Z,W))\}]$ and $\mathbb{E}[Y^T \mathbf{1}\{U_D \in \text{Supp}(g(Z,W))\}]$. Our next two proposition state that we can rank the optimal welfare under different policy classes.

**Proposition 1.8** (Value of Manipulable Instrument)**.** *Suppose that Assumptions 1-3 hold and* $Supp(Z) \subset \mathscr{Z}^p$*. Then* $S^*_{sub} \geq S^*_{dir}$*. The inequality holds strictly if the supremum in the definition of* $S^*_{sub}$ *is achieved by a unique policy* $\pi^*$ *such that* $g(W,\pi^*(W))$ *lies in the interior of* $Supp(g(W,Z))$ *with positive probability.*

Proposition 1.8 states that the optimal welfare by manipulating the instrument is always preferred to setting treatment directly.[11] The reason why subsidy-based policy weakly dominates the direct policy is that the former affects the treatment status through changes in the treatment selection: smaller subsidy induces only individuals with low $U_D$, while larger subsidy induces both low- and high-$U_D$ into the treatment status. Loosely speaking, the subsidy-based policy implicitly uses $U_D$ as a targeting variable even though $U_D$ is not observable. Since $U_D$ correlates with $(U_1, U_0)$ and thus the treatment effect, targeting on $U_D$ will improve the welfare.

[10] Mathematically, this is means $T$ is dependent on $(W,Z)$ through $\tau_T$.

[11] This is a very interesting case where the policy-maker choose to deliberately stochastic [Cerreia-Vioglio et al., 2019] in terms of the treatment choice.

**Proposition 1.9** (Irrelevance of Instrument in Treatment Rules)**.** *Suppose that Assumptions 1 and 2 hold. Then* $S^*_{dir} = S^*_{con}$.

The result states that, when the policy under consideration is to assign (the propensity of) treatments directly, then the instrument variable is irrelevant as a targeting variable. The instrumental variable $W$ is useful for the identification of treatment effects but, when used for targeting, do not improve the welfare since the instrumental variable is randomly assigned and excluded from the outcome equations. By contrast, as we can see from Example 1.4 in Section 1.3, for subsidy-based policies, the optimal targeting can involve the instrumental variable.

Notice that, although as we see from Proposition 1.8 that subsidies rules implicitly target on unobserved heterogeneity, targeting with subsidies is still not as good as the case when $U_D$ is observed. The subsidy-based policy only targets individuals in a "second-best" sense, as the targeting is restricted to a specific form that individuals with low $U_D$ have to be in the treatment status whenever the high $U_D$ individuals are. However, in the next proposition, we will show that subsidy-based policy can achieve the first-best welfare when MTE is decreasing.

Consider a (infeasible) direct policy that assigns the propensity of treatment based on both the observed characteristics $(W,Z)$ and unobserved heterogeneity $U_D$. The corresponding treatment status, denoted as $\widetilde{T}$, is $\sigma(W,Z,U_D)$-measurable Bernoulli random variable with parameter $\tau_{\widetilde{T}}(W,Z,U_D) = \mathbb{P}(\widetilde{T} = 1 \mid W,Z,U_D)$. In particular, Let $Y^{\widetilde{T}} = \widetilde{T}Y_1 + (1-\widetilde{T})Y_0$ be the counterfactual outcome from the policy. We define optimal welfare $S^*_{\text{inf}}$ with respect to this policy class as

$$S^*_{\text{inf}} = \sup_{\widetilde{T}:\ \sigma(W,Z,U_D)\text{-measurable}} \mathbb{E}[Y^{\widetilde{T}}\mathbf{1}\{U_D \in \text{Supp}(g(W,Z))\}].$$

The next proposition states that subsidy-based policies can achieve the same welfare as direct policies that targets on $U_D$ even though the latter is infeasible.

**Proposition 1.10.** *Suppose that Assumptions 1-3 hold and* $\mathscr{Z}^p = Supp(Z)$*. If MTE*$(u)$ *is de-*

*creasing, then* $S^*_{sub} = S^*_{inf}$.

The intuition behind proposition 1.10 is that the treatment assignment of the infeasible policy can be replicated by a subsidy-based policy when the MTE is decreasing. Specifically, the infeasible first best policy assigns all individuals with $\text{MTE}(u) \geq 0$ to the treatment group. Let $u^*_x$ be a solution to $\text{MTE}(u) = 0$. As individuals with higher MTE are always induced first in the case of positive selection, the subsidy-based policy can achieve the same counterfactual treatment choice if the individuals with $U_D = u^*_x$ are indifferent about the treatment choice under the policy, implying that individuals with $\text{MTE}(u) \geq 0$ are all induced to the treatment group.

## 1.6 Empirical Application

We illustrate our method by applying it to the experimental data from the Jordan New Opportunities for Women (Jordan NOW) pilot study. In the experiment, vouchers for wage subsidies are randomly assigned to female college students who are in their last year of education. These vouchers can be presented to firms while searching for jobs, and, if a student with a voucher is employed, the employer can redeem the voucher for up to six months for an amount equal to the minimum wage. The premise of the program is that wage subsidies can help students land their first jobs in which they can acquire experience and skills helpful for their long-term careers. We refer readers to Groh et al. [2016] for more details about the background and their experiment design.

In their study, Groh et al. [2016] finds that, although wage subsidies substantially increase the employment rate immediately after graduation, wage subsidies have limited effects on the long-term labor market participation. Specifically, wage subsidies increase the employment rate by about 38 percentage points during the subsidized period. However, the effect vanishes rapidly after the subsidy expires. Seventeen months after the subsidies went out, the effect on employment is less than two percentage points and is not statistically significant. Groh et al. [2016] concludes that providing wage subsidies is not an effective measure to promote women's

long-term labor market participation, at least in the context of Jordan.

In our empirical exercise, we investigate whether one can improve the effectiveness of wage subsidies through targeting. The welfare function we consider is the 30-month earnings ($Y$) after the subsidy period minus the cost of wage subsidy.[12] In the model, $Y$ is the realization of one of the potential outcomes $Y_1$ and $Y_0$, depending on whether the student successfully found a job after graduation during the subsidized period ($D = 1$) or not ($D = 0$). In our policy exercise, we consider targeting based on the student's college major. Specifically, we target on whether the student majors in medical assistance ($X$), which includes nursing and pharmacy specializations. We search for the optimal subsidy within the space $\mathscr{Z}^p = [0, 900]$, where $z^p = 0$ refers to no subsidy and $z^p = 900$ is the maximal subsidy an individual could receive in the experiment.

The optimal amount of subsidy critically depends on two factors: (1) how effective can wage subsidy encourage and help students to find their first job and (2) the treatment effect of having a first job after graduation on the long-term labor market outcome. The traditional approach to quantifying these effects starts with imposing joint normality for the error terms $(U_1, U_0, U_D)$ and their independence to $(X, Z)$. Then, we can estimate the outcome and choice equations together with either the method of maximum likelihood or the two-step method proposed by Heckman [1976]. For this illustration, we will proceed accordingly. Although these parametric assumptions could be restrictive, they yield more precise estimates when the sample size is modest. We provide a more flexible method that does not impose normality in the Appendix.

[12] We proxy the 30-month earnings by the monthly earnings reported at the last round, which took place two years after the voucher had expired.

Formally, we estimate the following selection model:

$$
\begin{aligned}
Y_1 &= X'\beta_1 + U_1, \\
Y_0 &= X'\beta_0 + U_0, \\
D &= \mathbf{1}\{X'\beta_D + Z\gamma - \tilde{U}_D \geq 0\}, \\
(X,Z) &\perp (U_1,U_0,\tilde{U_D}), (U_1,U_0,\tilde{U_D}) \sim \mathscr{N}(0,\Sigma),
\end{aligned}
$$

where

$$
\Sigma = \begin{pmatrix} \sigma_1^2 & \rho_{01}\sigma_0\sigma_1 & \rho_1\sigma_1 \\ \rho_{01}\sigma_0\sigma_1 & \sigma_0^2 & \rho_0\sigma_0 \\ \rho_1\sigma_1 & \rho_0\sigma_0 & 1 \end{pmatrix}.
$$

**Table 1.1.** Estimates of the Heckman selection model

| | **Choice equation** | **Outcome equation** $(Y_0)$ | **Outcome equation** $(Y_1)$ |
|---|---|---|---|
| Medical major | 0.30∗ | 1743.90∗ | 2677.21∗ |
| | (0.09) | (351.32) | (420.98) |
| Subsidy | 0.17∗ | | |
| (hundreds of JOD) | (0.00) | | |
| Intercept | −0.94∗ | 607.59∗ | 660.134 |
| | (0.07) | (192.68) | (437.14) |
| $\rho_d$ | | −0.09 | 0.38 |
| $\sigma_d$ | | 2596.77 | 3399.09 |

[1] The sample size is 1347. Standard errors are obtained by bootstraps. The asterisk signifies 1% statistical significance level.

Table 1.1 presents the estimates from the Heckman two-step method. Same as the results in Groh et al. [2016], we find that wage subsidies do significantly increase the chance of finding a job after graduation. In Figure 1.1, we plot the probabilities as functions of subsidies. As we can see, at the maximal amount of subsidy, the employment rate almost triples compared to the case with no subsidy. Also, notice that the employment rate is higher among the students with

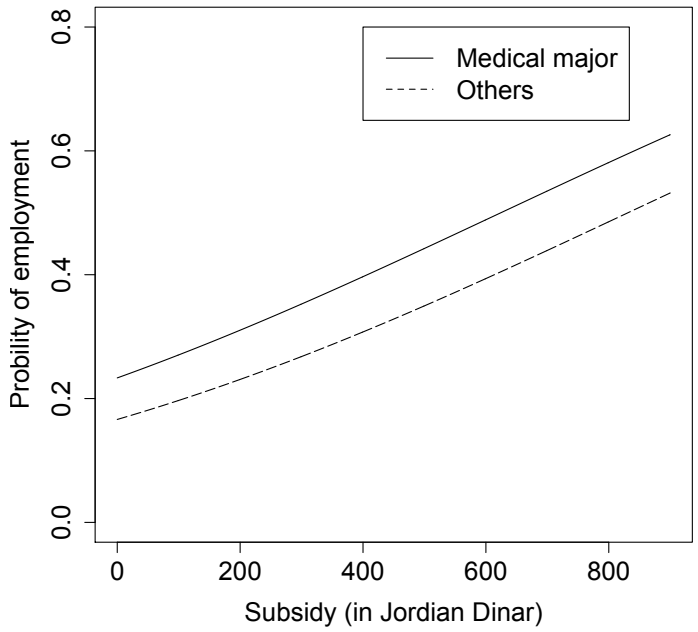



**Figure 1.1.** The fitted probabilities from the choice equation

medical majors at any given level of subsidy. From the policy-maker's perspective, this implies that it is cheaper to help medical students to land their first job.

Under the normal selection model, Heckman et al. [2003] shows that the marginal treatment effect is given by

$$\text{MTE}(x,u) = x'(\beta_1 - \beta_0) - (\rho_1\sigma_1 - \rho_0\sigma_0)\Phi^{-1}(u).$$

We plot the estimated MTE curves in Figure 1.2. There are several things worth noticing. First, the MTE curves are downward-sloping, suggesting that individuals positively select into the treatment. That is, individuals who are more likely to find a job after graduation tend to benefit more from it for their long-term career prospects. Second, the marginal effect can be negative among individuals with high $U_D$, meaning that wage subsidies can potentially do harm for individuals with low willingness to work.[13] From the policy-making perspective, this implies

[13] A possible explanation for the negative effect is the stigmatization toward voucher users [Burtless, 1985].

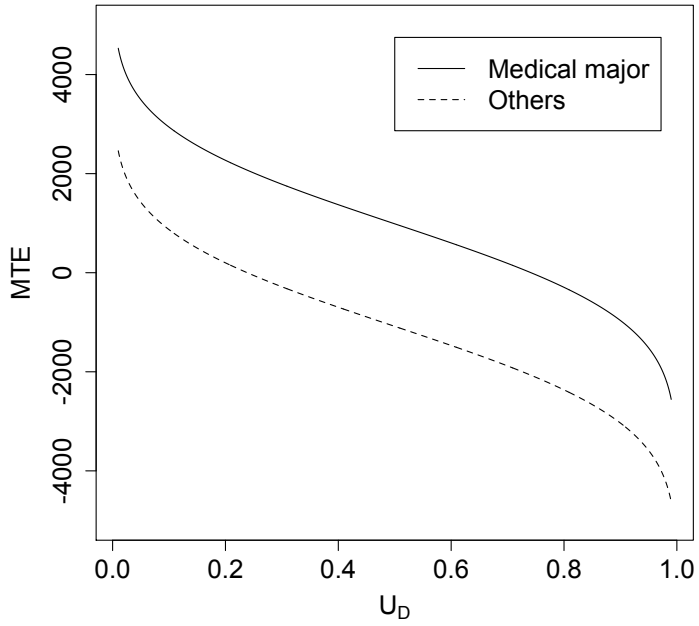
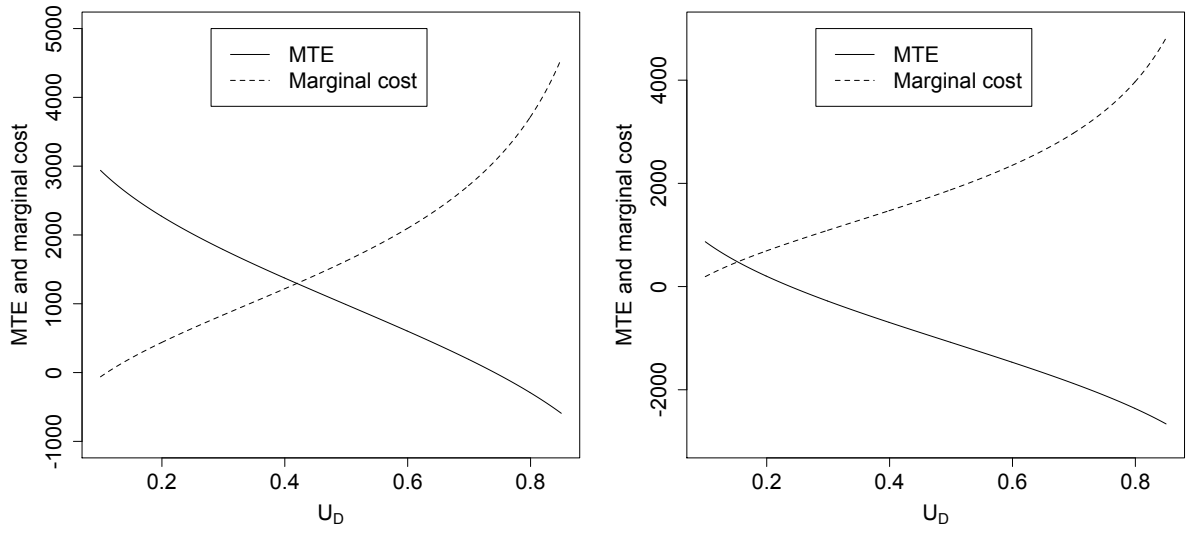



**Figure 1.2.** MTE estimated from a normal selection model

that excessive subsidies could be not only expensive but also harmful to the individuals.

Combining previous results, we plot the MTE curves and marginal cost of subsidies in Figure 1.3. Based on Proposition 1.1, we can look for the optimal take-up rate at the intersection of the MTE and marginal cost curves. The optimal subsidies are substantially different for the two groups of students. As discussed earlier, since students with medical majors tend to benefit more from subsidies, the optimal take-up rate for them is higher than that for students from other majors. In fact, by substituting the optimal take-up rate into the inverse of propensity scores, we can find that the optimal subsidy for medical students is about JOD 375, which is about one-third of the amount provided in the experiment. However, as for students from other majors, the estimated optimal subsidy is negative (JOD $-75$). As negative subsidies are out of consideration in this context, we have a corner solution which implies that the policy-maker should not provide subsidies for this group of students.

Our results suggest the amount of subsidies set in the experiments are much higher than




**(a)** Medical majors

**(b)** Other majors

**Figure 1.3.** MTE curves and marginal costs

their optimal level (in terms of the welfare function we defined). The insignificant long-term effect found in Groh et al. [2016] may stem from the fact that excessive amounts of subsidies may draw individuals with lower returns. By lowering the amount of subsidies, the welfare outcome may actually be improved. Moreover, we can further enhance its efficiency with targeting to exploit the heterogeneity in treatment effects and take-ups.

## 1.7 Conclusion

In this paper, we study the problem of allocating subsidies based on individual characteristics. We adopt the MTE framework to analyze the characterization, identification, and welfare properties of subsidy rules. Our results show that subsidy rules generally outperform policies that mandate the treatment. In our empirical example, we illustrate the main idea of this paper by estimating the optimal wage subsidy using a parametric MTE model. More flexible methods that do not impose parametric assumptions are provided in the appendix, of which their theoretical properties are of interest for future studies.

*Chapter 1, in full, is submitted for publication. Chen, Yu-Chang; Haitian Xie. "Personalized Subsidy Rules". The dissertation author was the primary investigator and author of this material.*

# Chapter 2

# Global Representation of the Conditional LATE model: A Separability Result

## 2.1 Introduction

Self-selection into treatment is a common challenge in causal inference. One approach, pioneered by Heckman [1976], is to impose a model on the selection process. Another approach is to invoke the assumptions of Imbens and Angrist [1994] and use instrumental variables to identify the local average treatment effect (LATE). Vytlacil [2002] shows that the two approaches are equivalent, even though the LATE approach does not provide an explicit model of the selection process. Specifically, Vytlacil [2002] finds that the monotonicity and independence conditions imposed in Imbens and Angrist [1994] together imply a nonparametric binary choice model, in which the instrument and the unobserved heterogeneity are additively separable in the latent index. When conditioning covariates are included, say, for instrument validity, the selection model representation can be established on a given value of covariates. Namely, holding fixed the value of covariates, imposing a (nonparametric) selection model is no stronger than imposing the LATE assumptions.

However, in most empirical settings, a fully nonparametric analysis, conditioning on each value of the covariates, is prohibitively data-demanding. It is typical to pool observations with different characteristics and to incorporate covariates in the selection model (for example, Carneiro et al., 2011 and Cornelissen et al., 2018). These empirical works conduct global analysis

that explicitly models covariates while the theoretical analysis of Vytlacil [2002] is local in the sense that covariates are fixed at a constant level.[1] This paper aims at filling this gap.[2]

Our result extends the representation in Vytlacil [2002] to settings where the covariates are not held fixed. We show that the conditional LATE (CLATE) model [Abadie, 2003] has a threshold-crossing representation in which the instruments are separated from the covariates in the latent index. Loosely speaking, the separability is a result of the monotonicity condition in the CLATE model that requires the instruments to affect the potential treatment status in the same direction for all individuals.[3] In particular, the direction of the monotonicity condition is the same across all values of the covariates, and thus a latent index representation that separates out the covariates is appropriate in this case.

The separability result implies that it is possible to uniformly rank the instruments' values by the propensity score across the covariates' values. Such a ranking has two practical usages. First, it can be used as a single index to construct testable implications for the CLATE model. Second, we can use the ranking to relabel the values of the instruments so that an increase in the ranking never causes individuals to drop the treatment, regardless of their covariates values. Our representation result also implies that the techniques developed in the conditional LATE (CLATE) framework, such as the identification analysis in Abadie [2003], can be applied to selection models that impose separability, and vice versa. As a corollary of the representation theorem. We reformulate and refine the testable implications of Heckman and Vytlacil [2005] for the marginal treatment effect framework.

The remaining of the paper is organized as follows. In section 2.2, we present the

[1]The terminology "local" also appears in the work by Dahl et al. [2020] with a different meaning. They consider the weakening of the LATE assumption based on the outcome distributions rather than the covariates.

[2]In a recent paper, Kline and Walters [2019] compares the selection model and the LATE approach when covariates are present. While they have established that, in the absence of covariates, the selection model and the LATE approach will yield numerically identical estimates, they also point out that their equivalence result does not hold when the covariates are introduced at least in the way covariates are usually modeled in empirical works. This paper addresses the same issue by characterizing the set of selection models that are equivalent to the LATE model when covariates are present.

[3]Heckman and Vytlacil [2005] suggests renaming the monotonicity condition as uniformity because "it is a condition across people than the shape of a function for a particular person."

global latent index representation of the LATE model, which is our main result. Section 2.3 discusses some implications of the representation. Section 2.4 generalizes the result to the case of ordered-discrete choice selection models. The last section concludes.

## 2.2 Representation Results

We first introduce the CLATE model. Let the binary variable $D$ be the receipt of treatment so that $D = 1$ denotes the treatment status, and $D = 0$ denotes the untreated status. The potential outcomes under the treated and untreated status are denoted by $Y_1$ and $Y_0$, respectively. The actual outcome observed by the econometrician is $Y = DY_1 + (1-D)Y_0$. Let $X$ be a random vector containing variables that could potentially affect both the outcome and the treatment choice. The covariates are introduced in the model to make the validity of the instruments plausible. Denote $\mathscr{X}$ as the support of $X$. Let the random vector $Z$ be the collection of variables that affect the treatment choice $D$ but not the potential outcomes. These variables are referred to as instruments or excluded variables. Note that under this specification, $Z$ and $X$ are disjoint sets of variables. Denote $\mathscr{Z}$ as the support of $Z$. For each value $z$ of the instrument, let $D_z$ be the counterfactual treatment status if $Z$ were externally set to $z$. The realized treatment can be represented as $D = D_Z = \sum_{z\in\mathscr{Z}} \mathbf{1}\{Z = z\}D_z$.

To avoid measure-theoretic technicalities, we assume both $\mathscr{Z}$ and $\mathscr{X}$ are countable. We further assume that for any $x \in \mathscr{X}$, $P(D = 1 \mid X = x) \in (0,1)$ and $P(D = 1 \mid Z = z, X = x)$ is not a trivial function of $z$.This means that there exist both treated and untreated individuals, given each value of the covariates value. This assumption is also imposed in Vytlacil [2002]. The assumptions of the CLATE model are listed as follows.

**Assumption 8** (Conditional Independence)**.** $\left(\{D_z : z \in \mathscr{Z}\}, Y_1, Y_0\right) \perp Z \mid X$ .

**Assumption 9** (Monotonicity)**.** *For any* $(z, z') \in \mathscr{Z}^2$*, either*

$$P(D_z \geqslant D_{z'} \mid X = x) = 1, \textit{ for almost all } x$$

*or*

$$P(D_z \leqslant D_{z'} \mid X = x) = 1, \textit{ for almost all } x.$$

Assumption 8 requires the instrument to be "as good as randomly assigned" conditional on the covariates. Assumption 9 is the monotonicity condition that is typically required in the LATE literature. Together, the two assumptions form the CLATE framework. Note that the exclusion restrictions of the instrument on the outcome is already embedded in the notation of the potential outcomes.

We discuss the monotonicity condition in more detail. This condition is global as it requires the direction of monotonicity to be the same across different values of $x$. A weaker and conditional version of monotonicity would be to impose, for any $(z, z') \in \mathscr{Z}^2$, and for each $x$ locally, either

$$P(D_z \geq D_{z'} \mid X = x) = 1, \text{ or } P(D_z \leq D_{z'} \mid X = x) = 1. \tag{2.1}$$

For any $x \in \mathscr{X}$, we can consider the individual with $P(D_z > D_{z'} \mid X = x) = 1$ as the complier and the individual with $P(D_z < D_{z'} \mid X = x) = 1$ as the defier. Then under the local monotonicity condition (2.1), it is possible that for some $x$, there are compliers but no defier; while for other $x$, there are defiers but no complier. This notion of local monotonicity can be found, for example, in Kolesár [2013] and Słoczyński [2020]. However, it is important to have uniformity in the direction of monotonicity in order to obtain the separability result in the global representation.

The main result of this paper is Theorem 1.

**Theorem 1** (Latent Index Representation of CLATE)**.** *The following two representations are equivalent.*

*(i) The CLATE model (Assumptions 8 and 9).*

*(ii) There exist functions m and q, and a random variable U such that* $(Y_1, Y_0, U) \perp Z \mid X$ *and*

*the treatment choice is determined by*

$$D_z = \mathbf{1}\{m(z) \geqslant q(X,U)\} \; w.p.1. \tag{2.2}$$

*Furthermore, if the conditional distribution of* $U \mid X = x$ *is absolute continuous for all* $x \in \mathscr{X}$,[4] *then there exist a function* $q^*$ *and a random variable* $U^* \sim Unif[0,1]$ *such that* $U^* \perp (Z,X)$ *and*

$$D_z = \mathbf{1}\{m(z) \geqslant q^*(X,U^*)\} \; w.p.1. \tag{2.3}$$

This representation result achieves separability between the instrument and covariates in the treatment choice process. The function $m$ ranks the values of the instrument. Moreover, this ranking is invariant to changes in the covariates and is identified up to an increasing transformation. We further explain this ranking in the next section.

The form of Equation (2.2) is to emphasize the separation between the instrument $Z$ and the covariates $X$. Alternatively, we can define $\tilde{U} = q(X,U)$, and write the selection equation as

$$D_z = \mathbf{1}\{m(z) \geqslant \tilde{U}\} \; w.p.1, \tag{2.4}$$

where $(Y_1, Y_0, \tilde{U}) \perp Z \mid X$. Representation (2.4) is used in the proof of Corollary 2. Note that the separation between $Z$ and $X$ in representation (2.2) and (2.3) holds inside the indicator function, and it does not necessarily imply that propensity score is additively separable in $Z$ and $X$. For example, consider the simple treatment selection equation $\mathbf{1}\{Z + X \geq U\}$, where $U \mid (Z,X) \sim N(0,1)$. In this case, the propensity score is equal to $\pi(z,x) = \Phi(z+x)$, with $\Phi$ being the distribution function of the standard normal distribution. This particular propensity is not additively separable between its two arguments.[5]

[4]This assumption is typically imposed in the marginal treatment effect literature (for example, Heckman and Vytlacil, 2005) for the normalization of $U$.

[5]That being said, non-separabilities of $Z$ and $X$ in the propensity score may lead to a contradiction to the monotonicity assumption. For example, this can happen if the marginal effect of increasing $Z$ is positive for some

The intuition of the Theorem is explained along with the following proof, where we make use of the idea presented in Vytlacil [2006a].

*Proof of Theorem 1.* The direction $(ii) \implies (i)$ is obvious. For $(i) \implies (ii)$, consider fixing $X = x$ for any $x \in \mathscr{X}$, then apply the results by Vytlacil [2002]. Formally, let $(\Omega, \mathscr{B}, P)$ be the probability space that underlies the random vector $(Y_1, Y_0, , \{D_z : z \in \mathscr{Z}\}, X, Z)$. Consider the partition $\Omega = \bigcup_{x \in \mathscr{X}} \Omega_x$, where $\Omega_x = \{\omega \in \Omega : X(\omega) = x\}$. The probability $P(\Omega_x)$ is non-zero as $\mathscr{X}$ is assumed to be countable. For each $x \in \mathscr{X}$, we construct the probability space $(\Omega_x, \mathscr{B}_x, P_x)$ where the $\sigma$-algebra

$$\mathscr{B}_x = \{B \cap \Omega_x : B \in \mathscr{B}\}$$

and the probability measure

$$P_x(B) = \frac{P(B \cap \Omega_x)}{P(\Omega_x)} \text{ for } B \in \mathscr{B}_x.$$

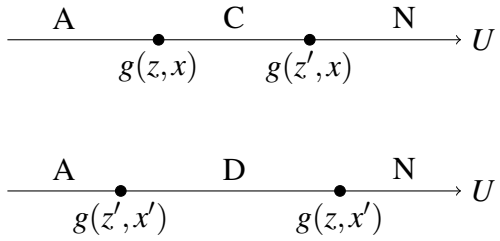
Consider the random variable $D_z^x = D_z\big|_{\Omega_x}$, which is the restriction of $D_z$ to the subdomain $\Omega_x$ Similarly define $Y_1^x, Y_0^x, Z^x$. It is not hard to see that the probability space and the random variables are well-defined and that $P_X(B)$ is the conditional probability of $B$ given $X$.[6] Then Assumption 8 implies that for all $z \in \mathscr{Z}$, $Z^x \perp (Y_1^x, Y_0^x, D_z^x)$ under the probability measure $P_x$. Assumption 9 implies that for all $(z, z') \in \mathscr{Z}^2$, either $D_z^x \geq D_{z'}^x$ or $D_z^x \leq D_{z'}^x$. This means that the variables $(Y_1^x, Y_0^x, \{D_z^x : z \in \mathscr{Z}\}, Z^x)$ satisfy the LATE assumptions defined by Vytlacil [2002] (namely assumptions L-1 and L-2 in that paper). By the result in that paper, an equivalent representation for $D_z^x$ is that there exists a non-trivial function $g$ and a random variable $U^x$, independent of $Z^x$, such that $D_z^x = \mathbf{1}\{g(z,x) \geqslant U^x\}$. Define $U \equiv U^X$. By construction $U \perp Z \mid X$, and $D_z = D_z^X = \mathbf{1}\{g(z,X) \geqslant U\}$.

The global monotonicity condition imposes further restrictions. It forbids the following

$(X,Z) = (x,z)$ but negative for another value $(X,Z) = (x',z)$. For example, if the propensity score is $\pi(z,x) = \Phi(zx)$, then the monotonicity assumption would be violated if $x$ can take both positive and negative values.

[6] We are stating that the followings are true: (1) $(\Omega_x, \mathscr{B}_x, P_x)$ is a probability space, (2) $Y_1^x, Y_0^x, Z^x$ and $\{D_z^x : z \in \mathscr{Z}\}$ are $\mathscr{B}_x$-measurable, and (3) $P_X(B) = \mathbb{E}[\mathbf{1}_B | X]$ a.s. for $B \in \mathscr{B}$.

situation: for some pairs $(z,z') \in \mathscr{Z}^2$ and $(x,x') \in \mathscr{X}^2$, we have $g(z,x) > g(z',x)$ but $g(z,x') < g(z',x')$. This violation is depicted in Figure 2.1, where under $X = x$ we have compliers, but under $X = x'$ we have defiers.




A: always taker, C: complier, D: defier, N: never taker.

**Figure 2.1.** Violation of Assumption 9

In fact, the monotonicity condition implies that $g$ satisfies the following property: for all pairs $(z,z')$, either $g(z,x) > g(z',x)$ for all $x$, or $g(z,x) < g(z',x)$ for all $x$. By Lemma 1 in Vytlacil [2006a], this property implies that there exists a set of strictly increasing functions $\{h_x(\cdot) : x \in \mathscr{X}\}$ and a function $m$ such that $g(z,x) = h_x(m(z))$. Thus, based on the fact that each $h_x$ is strictly increasing hence invertible, we can derive that $D_z = \mathbf{1}\{g(z,X) \geqslant U\} = \mathbf{1}\{h_X(m(z)) \geqslant U\} = \mathbf{1}\{m(z) \geqslant h_X^{-1}(U)\}$ and establish the representation in (2.2).

For representation (2.3), define $U^* = F_{U|X}(U \mid X)$ and $q^*(x,u) = q(x, F_{U|X}^{-1}(u \mid x))$, where $F_{U|X}(\cdot \mid \cdot)$ is the conditional cumulative distribution function of $U$ given X. We have

$$\begin{aligned} D_z &= \mathbf{1}\{m(z) \geqslant q(X,U)\} \\ &= \mathbf{1}\{m(z) \geqslant q(X, F_{U|X}^{-1} \circ F_{U|X}(U))\} \\ &= \mathbf{1}\{m(z) \geqslant q^*(X,U^*)\}. \end{aligned}$$

$U^*$ is independent to $(X,Z)$ and distributed as Unif$[0,1]$ because

$$\begin{aligned} P(U^* \leqslant u \mid X=x, Z=z) &= P(F_{U|X}(U \mid X) \leqslant u \mid X=x, Z=z) \\ &= P(F_{U|X}(U \mid x) \leqslant u \mid X=x) \\ &= P(U \leqslant F_{U|X}^{-1}(u \mid x) \mid X=x) \\ &= F_{U|X}(F_{U|X}^{-1}(u \mid x) \mid x), \\ &= u, \text{ for all } u \in [0,1], \end{aligned}$$

where the second line holds as $U \perp Z \mid X$, and the fourth line holds by the definition of $F_{U|X}$. □

## 2.3 Implications

The separability property between $Z$ and $X$ in the choice equation implies a rank-invariance property of the ranking of the instrument in terms of the propensity score

$$\pi(z,x) \equiv P(D=1 \mid Z=z, X=x).$$

The following corollary also discusses the identification of the function $m$ from the propensity score.

**Corollary 2** (Observable Implications)**.** *Let Assumptions 8 and 9 hold.*

*(i) The propensity score $\pi$ satisfies that for any $z, z' \in \mathscr{Z}$,*

$$\pi(z,x) \geqslant \pi(z',x) \text{ for some } x \implies \pi(z,x) \geqslant \pi(z',x) \text{ for all } x. \tag{2.5}$$

*Further, using representation (2.4), if the CDF of $\tilde{U}$ is strictly increasing conditional on some value $x^*$, then the function $m$ can be ordinally identified as $m(z) = \pi(z,x^*)$. Moreover, if the conditional CDF is strictly increasing for all $x \in \mathscr{X}$, then the above statement also*

*holds when the weak inequalities in equation (2.5) are replaced by strict inequalities (or equalities).*

*(ii) The function m is a sufficient index of the instrument Z in the sense that*

$$P(Y_j \in \mathscr{B} \mid X, Z, D = j) = P(Y_j \in \mathscr{B} \mid X, m(Z), D = j),$$

*for any measurable set $\mathscr{B}$ and $j \in \{0,1\}$. Let $g_1, g_0$ be nonnegative functions, then*

$$\mathbb{E}\left[Dg_1(Y,X) \mid X, m(Z) = \mu\right]$$

*is weakly increasing in $\mu$ $(w.p.1)$ and*

$$\mathbb{E}\left[(1-D)g_0(Y,X) \mid X, m(Z) = \mu\right]$$

*is weakly decreasing in $\mu$ $(w.p.1)$. These implications are testable when the CDF of $\tilde{U}$ is strictly increasing conditional on some value $x^*$, a case in which the function m can be ordinally identified as the propensity score $\pi(z, x^*)$.*

This first implication means that the function *m* provides an observable ordering of the instrument values by their strength of pushing individuals to take up the treatment. That is, in the CLATE model, we can rank the instrument values by their effectiveness of inducing individuals into the treatment status. This ordering remains invariant under different values of $X$ because the monotonicity is assumed to be global (Assumption 9). However, we do not impose the "normalization" that a higher value of the instrument always leads to more treatment take-ups, so the function *m* need not be increasing.

The second implication uses the identified *m* to derive a set of testable implications of the CLATE model. This set of testable implications is a refinement of the testable implications of the marginal treatment effect framework derived in Heckman and Vytlacil [2005] as the role of

Z is fully summarized by the function $m$.[7] The testable implications are also analogous to those presented in Equation (3.3) in Kitagawa [2015] except that, here, $m(Z)$, but not $Z$ itself, enters the conditioning set. That is, the testable implications we derived do not restrict the direction of the effect of $Z$ on the treatment take-up. The distinction appears as we do not explicitly assume that no defier exists as Kitagawa [2015] does. We only assume that defiers and compliers can not both exist. That is, as stated in Assumption 9, we leave the direction of monotonicity unspecified. [8]

*Proof of Corollary 2.* (i) From (2.4), we have $\pi(Z,X) = P(m(Z) \geqslant \tilde{U} \mid Z,X) = F_{\tilde{U}|X}(m(Z))$, where $F_{\tilde{U}|X}$ denotes the conditional CDF of $\tilde{U}$ given $X$. The Condition (2.5) on the propensity score is satisfied since $F_{\tilde{U}|X}$ is non-decreasing. When $F_{\tilde{U}|X=x^*}$ is strictly increasing, the ordinal information contained in $m(\cdot)$ is fully transformed into $\pi(\cdot,x^*)$.

Now suppose that $F_{\tilde{U}|X=x^*}$ is strictly increasing for all $x$. As a result, the function $m$ is ordinally identified by $\pi(z,x)$ for any $x \in \mathscr{X}$. By the definition of being ordinally identified, we have

$$\pi(z,x) > \pi(z',x) \iff m(z) > m(z') \iff \pi(z,x') > \pi(z',x')$$

and

$$\pi(z,x) = \pi(z',x) \iff m(z) = m(z') \iff \pi(z,x') = \pi(z',x')$$

for any $z,z' \in \mathscr{Z}$ and $x,x' \in \mathscr{X}$.

(ii) Consider the case where $j = 1$, the other case can be proved by symmetric arguments. By

[7] Notice that notation "$Z$" in Heckman and Vytlacil [2005] is different from ours as it represents the joint set of the instruments and covariates. By contrast, $Z$ only contains the excluded instruments in our paper. Accordingly, the set of testable implications we derive is also stronger in that $m$ only depends on the excluded instrument, which is a result of the monotonicity condition imposed by the CLATE model.

[8] If the function $m$ were known and increasing, the testable implication in this paper would essentially reduce to Equation (3.3) in Kitagawa [2015], except that $Z$ can possibly be non-binary in our case. Combined with the testing procedure proposed in Section 3.1 in his paper to handle multivalued instruments, we may likewise design a test for our implication.

representation (2.2), we have

$$\begin{aligned}
P(Y_1 \in \mathscr{B} \mid X = x, Z = z, D = 1) &= P(Y_1 \in \mathscr{B} \mid X = x, Z = z, m(z) \geqslant q(x,U)) \\
&= P(Y_1 \in \mathscr{B} \mid X = x, m(z) \geqslant q(x,U)) \\
&= P(Y_1 \in \mathscr{B} \mid X = x, m(Z) = m(z), m(z) \geqslant q(x,U)) \\
&= P(Y_1 \in \mathscr{B} \mid X = x, m(Z) = m(z), m(Z) \geqslant q(X,U)) \\
&= P(Y_1 \in \mathscr{B} \mid X = x, m(Z) = m(z), D = 1),
\end{aligned}$$

where the second and third lines follow from the conditional independence assumption (Assumption 8), and the last line follow from the equivalence result.[9] For the second assertion, let $\mu > \mu'$, then

$$\begin{aligned}
&\mathbb{E}\left[Dg_1(Y,X) \mid X, m(Z) = \mu\right] - \mathbb{E}\left[Dg_1(Y,X) \mid X, m(Z) = \mu'\right] \\
=&\mathbb{E}\left[\mathbf{1}\{h(X,U) \leqslant \mu\} g_1(Y_1,X) \mid X, m(Z) = \mu\right] \\
-&\mathbb{E}\left[\mathbf{1}\{h(X,U) \leqslant \mu'\} g_1(Y_1,X) \mid X, m(Z) = \mu'\right] \\
=&\mathbb{E}\left[\mathbf{1}\{\mu' < h(X,U) \leqslant \mu\} g_1(Y_1,X) \mid X\right] \geqslant 0.
\end{aligned}$$

The case of $\mathbb{E}\left[(1-D)g_0(Y,X) \mid X, m(Z) = \mu\right]$ is similar. When the function $m$ is ordinally identified as $m(z) = \pi(z, x^*)$ for some $x^*$, we can rewrite the implications as (i) the following equality

$$P(Y_j \in \mathscr{B} \mid X, Z, D = j) = P(Y_j \in \mathscr{B} \mid X, \pi(Z, x^*), D = j),$$

hold for any measurable set $\mathscr{B}$ and $j \in \{0,1\}$, (ii) the function

$$\mathbb{E}\left[Dg_1(Y,X) \mid X, \pi(Z,x^*) = p\right]$$

[9] We thank one of the anonymous referee for suggesting this proof to improve the clarity of the original arguments.

is weakly increasing for $p$ in the range of $\pi(Z,x^*)$, and (iii) the function

$$\mathbb{E}\left[(1-D)g_1(Y,X) \mid X, \pi(Z,x^*) = p\right]$$

is weakly decreasing for $p$ in the range of $\pi(Z,x^*)$. These implications are testable as $\pi(Z,x^*)$ is identified by the observed propensity $P(D=1 \mid Z=z, X=x^*)$.

□

## 2.4 Ordered Treatment Levels

This section extends the representation result in Section 2.2 to incorporate multiple ordered levels of treatment. The argument follows from the equivalence results in Vytlacil [2006b]. Let there be $K$ possible levels of treatment. Now the treatment $D$ takes values in an ordered set $\{1,\cdots,K\}$. The counterfactual treatment $D_z$'s are defined accordingly. The corresponding potential outcomes are denoted by $(Y_1,\cdots,Y_K)$.

The CLATE assumptions are modified to incorporate the ordered multiplicity in treatment levels. Although we have a different definition of $D$, the statement of the monotonicity condition does not change.

**Assumption 8'.** $\left(\{D_z : z \in \mathscr{Z}\}, Y_1, \cdots, Y_K\right) \perp Z \mid X$.

**Assumption 9'.** *For any* $(z,z') \in \mathscr{Z}^2$*, either*

$$P(D_z \geqslant D_{z'} \mid X = x) = 1, \textit{ for almost all } x$$

*or*

$$P(D_z \leqslant D_{z'} \mid X = x) = 1, \textit{ for almost all } x.$$

**Corollary 3** (Ordered Treatment Levels)**.** *The ordered CLATE model (Assumptions 8' and 9') is equivalent to the following statements. There exist a function* $m$ *and* $K+1$ *random variables*

$U_0, \cdots, U_K$ *such that for* $k = 1, \cdots, K$,

*(i)* $D_z = k \iff U_{k-1} \leqslant m(z) < U_k$,

*(ii)* $Z \perp (U_1, \cdots, U_{K-1}, Y_1, \cdots, Y_K) \mid X$,

*(iii)* $U_0 = -\infty$, $U_K = \infty$, *and* $U_k \geqslant U_{k-1}$.

This is basically the conditional version of the representation result in Vytlacil [2006b]. The main point is that even though the random thresholds $U_1, \cdots, U_K$ covariates with $X$, the latent index $m(Z)$ does not explicitly depend on $X$. Again, this is because Assumption 9' requires that the direction of monotonicity has to be the same across all values of $X$.

*Proof of Corollary 3.* Proof of direction from the random thresholds model to the CLATE model can be done in the exact same way as in Vytlacil [2006b]. In particular, Assumption 8' is implied by item (ii). For Assumption 9', suppose for some $(z, z') \in \mathscr{Z}^2$ and $x \in \mathscr{X}$, $P(D_z > D_{z'} \mid X) > 0$. This implies that $m(z) > m(z')$, which in turn implies that $D_z \geq D_{z'}$. This in fact means that $P(D_z \geq D_{z'} \mid X = x) = 1$ for all $x \in \mathscr{X}$. Because otherwise, $P(D_z \geq D_{z'} \mid X = x) < 1$ for some $x \in \mathscr{X}$, then by the law of total probability,

$$P(D_z \geq D_{z'}) = \sum_{x \in \mathscr{X}} P(X = x) P(D_z \geq D_{z'} \mid X = x) < 1.$$

For the other direction, define $D_z^k = \mathbf{1}\{D_z > k\}$. Then each $D_z^K$ is a binary treatment whose representation can be analyzed by Theorem 1. So $D_z^k = \mathbf{1}\{m^k(z) \geqslant \tilde{U}^k\}$. For any $z, z' \in \mathscr{Z}$, $m^k(z) \geqslant m^k(z')$ implies that $D_z \geqslant D_{z'}$ by monotonicity. Let $d(z) = \mathbb{E}[D_z]$. Then by Lemma 1 in Vytlacil [2006b], $m^k(z) = g^k(d(z))$ for some non-decreasing $g^k(\cdot)$. The rest of the proof follows from that paper. □

## 2.5 Conclusion

This paper shows that the CLATE model has a latent index representation in which the instrument and the covariates are separable in the treatment choice equation. On the theoretical side, the result more rigorously links the CLATE model to the latent index representation when covariates are present. On the practical side, the result establishes conditions when methods from the two pieces of literature can be used interchangeably. For example, one can employ the nonparametric estimator in Frölich [2007] as robustness checks for the structural estimates in selection models. For future works, one can consider extending this result into the unordered monotonicity model [Heckman and Pinto, 2018].

*Chapter 2, in full, is a reprint of the material as it appears in Oxford Bulletin of Economics and Statistics. 2009. Chen, Yu-Chang; Xie, Haitian. "Global Representation of the Conditional LATE Model: a Separability Result". The dissertation author was the primary investigator and author of this material.*

# Chapter 3

# Empirical Bayes with Optimal Shrinkage Trees

## 3.1 Introduction

Empirical Bayes methods are widely-used for the estimation of treatment effects in applications where the number of treatments is high but the sample size is only modest. Prominent examples include neighborhood effects, hospital effects, and teacher value-added measures (VAM). In these applications, having accurate estimates of the treatment effects are essential as these estimates themselves often serve as important pieces of information in various situations. In teacher VAM, for example, the estimated effects can be inputs to high-stake personnel decisions. In other cases, the researchers would use the estimates to study the distribution of treatment effects as well as their correlation to other variables. Inaccurate estimates of treatment effects would make such analysis prohibitively uninformative.

Similar to other techniques in high-dimensional statistics, the method of empirical Bayes utilizes shrinkage to improve the estimator's performance. In its simplest form, the empirical Bayes estimates can be obtained by shrinking the initial estimators of treatment effects (such as coefficient estimates from a regression) toward a common target. In most cases, the common target is the average of treatment effects, and the extent of shrinkage depends on the signal-to-noise ratio: the noisier the initial treatment effect estimator is, the more it is shrunk toward the target. Although shrinkage introduces bias as the treatment effects can deviate from their

average, it increases the overall precision by downplaying the initial estimators, which are usually unbiased but imprecise due to high-dimensionality. This specific form of shrinkage engaged in empirical Bayes methods is also known as "shrinkage toward the grand mean" in the literature.

This paper extends the classical empirical Bayes method by allowing each treatment effect estimate to be shrunk toward different targets instead of one common target shared by every treatment type. Specifically, our method involves the following steps. First, we use a decision tree to group together treatment effects with similar treatment characteristics. For example, we may group hospitals based on locations or group teachers based on years of experience. We then calculate the average of treatment effects for each group, and these group-specific averages will replace the grand mean and serve as the targets for shrinkage. Lastly, we shrink each treatment effect estimate toward the local average of the group it belongs to to produce the final estimates.

The proposed method improves the classical method since the group-specific averages are often the better shrinkage targets compared the grand mean. The grand mean, which is generic to all treatments, can be distant from each individual treatment effect if the variation among treatment effects is high. Grouping treatments based on treatment characteristics can reduce the unexplained variation, especially when the characteristics are strong predictors of the effects. By introducing the grouping, we can reduce the bias caused by shrinkage, making the resulting empirical Bayes estimates more accurate. On the contrary, shrinkage toward the grand mean totally ignores the information contained in the treatment characteristics.

In this paper, we first extend the classical empirical Bayes to incorporate group-specific shrinkage provided that a decision tree for grouping is already given. We first derive formulas for the group-specific shrinkage factor that decides the optimal degree of shrinkage. The group-specific shrinkage factor can be very different from the one arisen in classical method as conditioning on treatment characteristics can dramatically affect the signal-to-noise ratio. We then provide a consistent model selection method to help researchers decide the optimal decision tree that minimizes the mean-squared errors (MSE) of the corresponding empirical Bayes estimator.The optimal tree is referred to as the "optimal shrinkage tree", hence the title of

this paper "Empirical Bayes with Optimal Shrinkage Tree."

We then conduct a series of simulation experiments to examine the statistical properties of the proposed method. We first verify that the proposed model selection procedure can successfully pick the optimal decision tree with high probability given that number of treatments is large enough. We then compare the tree-based empirical Bayes method to the classical method under various data generation process, in which we quantify the efficiency gains as a function of the treatment characteristics's prediction power on treatment effects. Our simulation result shows that proposed method is most useful when the treatment characteristics can at least partially predict treatment effects.

While there are alternative grouping algorithms, such as nearest neighbor, that we can in principle combine with empirical Bayes, we argue that decision tree is a more suitable choice for its ease of interpretation. As mentioned earlier, empirical Bayes estimates are often used for high-stake decisions, and transparency is often a necessity for practical or moral reasons. It would be hard for the policy-maker to justify and implement the policy if it is too difficult to communicate with the stakeholders. Moreover, as a by-product of the proposed method, researchers can use selected decision tree to investigate the determinants of treatment effects. Without the interpretability of decisions trees, such analysis would be rather uninformative.

The rest of paper is organized as follows. Section 2 presents our model setup and the group-specific empirical Bayes method provided that a decision tree for grouping is given. In section 3, we then propose a data-driven approach for selecting the optimal shrinkage tree and show that the procedure is consistent. Results from a series of simulation experiments are collected in section 4. We present an empirical application using the Tennessee's Student Teacher Achievement Ratio (STAR) dataset in section 4. Section 5 concludes.

### 3.1.1 Literature review

Although the original idea can be traced back as early as to Robbins [1956] and James and Stein [1961], the empirical Bayes methods have recently regained popularity in both econometrics

and statistics [see, e.g., Hansen, 2017, Meager, 2019, Ignatiadis and Wager, 2019, Azevedo et al., 2020, Armstrong et al., 2020, Bonhomme and Weidner, 2021]. The resurgence is partly due to its applications to high-dimensional problems [Efron, 2010] as well as its increasing popularity in empirical studies [see e.g., Chetty et al., 2014, Chetty and Hendren, 2018, Hull, 2018, Angrist et al., 2021]. Our paper provides a novel approach that combines empirical Bayes method with decision trees, seeking to provide a more reliable method for empirical researchers amid the controversies on the usage of empirical Bayes methods, especially in the teacher VAM [Harris, 2009, Rothstein, 2010, Jackson et al., 2014, Guarino et al., 2015]. Our paper also adds to tree-based or more generally partition-based methods in economics [Athey and Imbens, 2016, Cattaneo and Farrell, 2018, Tabord-Meehan, 2018, Cattaneo et al., 2020].

## 3.2 Model Setup

### 3.2.1 Model and motivating examples

This paper considers the following model:

$$
\begin{aligned}
&Y_{ij} = \alpha_i + u_{ij}, \\
&(\alpha_i, Z_i) \overset{i.i.d.}{\sim} F, \\
&\mathrm{E}[u_{ij}|\alpha_i, Z_i] = \mathrm{E}[u_{ij}|\alpha_i] = 0
\end{aligned}
$$

$Y_{i1}$, $Y_{i2}$, ...., $Y_{iJ}$ are the observed outcomes for treatment $i$, and the main parameter of interest are the treatment effects $\{\alpha_i\}_{i=1}^N$. For ease of exposition, we assume that $J$, which is the number of observations for each treatment does not vary across $i$, although it is straightforward to incorporate heterogenous sample size $J_i$ at the cost of extra notation burdens. Each treatment $i$ comes with some treatment characteristics $Z_i$, which is potentially correlated with the treatment effect $\alpha_i$. We assume the error terms $u_{ij}$ are independently and identically distributed across $i$ and $j$.

In below, we list some applications that will motivate our method.

**Example 3.1.** *Teacher Value-Added Measures (VAM): In a typical application of VAM, $Y_{ij}$ is the test score for teacher i's j-th student. The teacher effect $\alpha_i$, i.e., the teacher value-added, is potentially correlated with the teacher characteristics $Z_i$ such as years of teaching experience. In VAM applications, it is a standard practice to apply the empirical Bayes method since J, which is the number of students per teacher, is typically small.*

**Example 3.2.** *Evaluation of Multi-Sites Programs. Nationwide government programs, such as the Head Start Program, typically have multiple experimental sites, and it is common for the researchers to pool samples from all sites in their analysis for the sake of statistical precision. However, since the actual implementations of the program could differ across sites, ignoring such complications may fail to reveal critical determinants of program effectiveness. Our method serves as an alternative to pooling when the heterogeneity across sites is of concern.*

**Example 3.3.** *Meta-Analysis. Similar to Bayesian hierarchical models [for a recent example in development economics, see Meager, 2019] and meta-regression analysis [Thompson and Sharp, 1999, Stanley and Jarrell, 2005], our method can be used to synthesize estimates from different studies. However, unlike previous approaches, our method provides a data-driven approach to uncover the effect heterogeneity across studies.*

### 3.2.2 Group-specific shrinkage when the decision tree is given

In this subsection, we extend the classical empirical Bayes method when a decision tree is specified to capture the correlation between treatment effects and treatment characteristics. We discuss how to select the decision using a data-driven approach in the next section.

We first define some notations. Let $T$ denote a binary tree and $L$ be its depth. A binary tree is induces a partition over $\mathscr{Z}$, the support of treatment characteristics $Z$. Write the partitioning sets (the leaf nodes) as $\{1,2,...,K\}, K = 2^L$.[1] Let $\kappa_T(\cdot) : \mathscr{Z} \to \{1,2,...,K\}$ denote the partition

[1] In this paper, we only consider perfect binary trees.

rule induced by the decision tree $T$ and write $K_i = \kappa_T(Z_i)$ as the leaf node for each treatment. The two conditional variances

$$\sigma_\alpha^2(k) = \text{Var}(\alpha_i | K_i = k)$$

and

$$\sigma_u^2(k) = \text{Var}(u_{ij} | K_i = k)$$

govern the leaf-specific signal-to-noise ratio, where the $\sigma_\alpha^2(\cdot)$ and $\sigma_u^2(\cdot)$ correspond to the strength of signals and the amount of noises respectively.

One obvious candidate to estimate $\alpha_i$ is the sample average $\bar{Y}_i = \frac{1}{J}\sum_{j=1}^{J} Y_{ij}$. However, as shown in the seminal work of Stein [1956], the sample average $\{\bar{Y}_i\}_{i=1}^N$ is actually inadmissible for the estimation of $\{\alpha_i\}_{i=1}^N$ when the error terms are normal, meaning that there is an estimator that dominates the sample average. One such estimator, and the most well-known, is the James-Stein estimator. The surprising finding that the sample average is in fact inadmissible even though it is the maximum likelihood estimator under normality motivates extensive studies on the James-Stein estimator, whose connection to empircal Bayes is later recognized [Efron and Morris, 1973].

The classical empirical Bayes method produce better estimators of $\{\alpha_i\}_{i=1}^N$ by shrinking the sample average toward the "grand mean":

$$\hat{\alpha}_i^{EB} = \bar{Y}_i - \frac{\sigma_u^2/J}{\sigma_\alpha^2 + \sigma_u^2/J} \cdot (\bar{Y}_i - \bar{\bar{Y}}),$$

where $\sigma_\alpha^2$ and $\sigma_u^2$ are the unconditional variances of $\alpha_i$ and $u_{ij}$ and $\bar{\bar{Y}} = \frac{1}{N}\sum_{i=1}^{N} \bar{Y}_i$ is the grand mean. We can simplify the above formula and write

$$\hat{\alpha}_i^{EB} = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_u^2/J} \cdot \bar{Y}_i + \frac{\sigma_u^2/J}{\sigma_\alpha^2 + \sigma_u^2/J} \cdot \bar{\bar{Y}}.$$

Above form resembles the posterior mean in a normal-normal Bayesian model except that the shrinkage is toward the grand mean, which is estimated from the data, instead of the researcher-specified prior mean, hence the name "empirical Bayes".

Same as the normal-normal Bayesian model, the amount of shrinkage depends on the signal-to-noise ratio:

$$\frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_u^2/J}.$$

When $\sigma_u^2$ is large, that is, when the measurements of $\alpha_i$ is relatively noisy, $\bar{Y}_i$ is shrunk more to reduce the impact of noise. On the contrary, if $\sigma_\alpha^2$ is large, the signal contained in the measurements $Y_{ij}$ is more predominant and therefore more weight is put on the individual estimate $\bar{Y}_i$.

A second way to understand the shrinkage is through the perspective of forecast combination. That is, we can view $\bar{Y}_i$ and $\bar{\bar{Y}}$ as two "forecasters" for the treatment effect $\alpha_i$ and would like to combine the two forecasters to improve precision. It is not hard to see that for the optimal linear combination, the weight for each forecaster is inversely proportional to their respective variances. For $\bar{Y}_i$, the variance comes from the noise $\{u_{ij}\}_{j=1}^J$, which results in an expected forecast error of $\frac{\sigma_u^2}{J}$. As for the grand mean $\bar{\bar{Y}}$, its forecast error stems from the fact that individual treatment effects can deviate from their mean, and therefore it's expected forecast error is $\sigma_\alpha^2$. Combining the two forecasters inversely proportional to their errors, we can obtain the empirical Bayes estimator.

The forecast combination perspective sheds light on why group-specific shrinkage can further improve the statistical precision of empirical Bayes estimator. Recall that $K_i = \kappa_T(Z_i)$ is the leaf node observation $i$ belongs to in the decision tree $T$. As a forecaster for $\alpha_i$, the expected forecast error of the group-specific average is the conditional variance $\sigma_\alpha^2(K_i)$, and the conditional variance is generally less than the unconditional variance as long as the treatment characteristics are not entirely uncorrelated with the treatment effects. As a result, the combined forecaster with the group-specific average has higher precision, and the gain in accuracy is

especially high if the prediction power of treatment characteristics is strong since the conditional variance $\sigma_\alpha(k)$ will be small.

We now formally introduce the tree-based empirical Bayes estimator given that a decision tree $T$ is given. Intuitively, the tree-based empirical Bayes estimator is the empirical Bayes estimator except that the shrinkage target and shrinkage factor are now group-specific. The group-specific average is given by

$$\hat{\mu}_k = \frac{1}{N_k} \sum_{i:K_i=k} \bar{Y}_i \mathbf{1}\{K_i = k\},\ k = 1, 2, ...., K,$$

where $N_k = \sum_{i=1}^{N} \mathbf{1}\{K_i = k\}$ is the number of observation in leaf node $k$. We then estimate the group-specific signal-to-noise ratio from the within- and across-treatment variations of $Y_{ij}$. Specifically, for each leaf node $k \in \{1, 2, ...., K\}$, we use the observed within-treatment variation

$$\hat{\sigma}_u^2(k) = \frac{1}{N_k} \sum_{i:K_i=k} \sum_{j=1}^{J} \frac{(Y_{ij} - \bar{Y}_i)^2}{J-1},\ k = 1, 2, ...., K,$$

to estimate the variance of noise and observed across-treatment variation

$$\hat{\sigma}_\alpha^2(k) = \frac{1}{N_k} \sum_{i:K_i=k} \frac{(\bar{Y}_i - \bar{\bar{Y}}_k)^2}{N_k - 1} - \frac{\hat{\sigma}_u^2(k)}{J},\ k = 1, 2, ...., K,$$

to estimate the variance of treatment effects. Notice that it is necessary to subtract the within-treatment variation from the across-treatment variation since we do not directly observe $\alpha_i$.

Given the above quantities, the tree-based empirical Bayes estimator corresponds to the decision tree $T$ is

$$\hat{\alpha}_i^{EB,T} = \frac{\hat{\sigma}_\alpha^2(K_i)}{\hat{\sigma}_\alpha^2(K_i) + \hat{\sigma}_u^2(K_i)/J} \cdot \bar{Y}_i + \frac{\hat{\sigma}_u^2(K_i)/J}{\hat{\sigma}_\alpha^2(K_i) + \hat{\sigma}_u^2(K_i)/J} \cdot \hat{\mu}_{K_i}.$$

## 3.3 Optimal Shrinkage Trees

We now discuss how to select the decision tree that generates the empirical Bayes estimators with minimum estimation errors possible. In this section, we provide a data-driven method to select a tree and show that it can consistently select the optimal shrinkage tree when the number of treatments go to infinity.

For any tree algorithm, or more generally for machine learning methods, the first step is to specify an appropriate loss function. The criteria we consider is quadratic loss function, and we seek to find the decision tree $T$ and shrinkage factors $\{\lambda_i\}_{i=1}^N$ that minimizes the risk function

$$R(T) = \mathbb{E}[\sum_{i=1}^{N}(\hat{\alpha}_i^{EB,T} - \alpha_i)^2].$$

Notice that the function $R$ as it is can not be directly used for optimization since $\alpha_i$ is unobserved. Our first proposition simplifies and characterizes the risk function in a way that we can construct the empirical risk function.

**Proposition 3.1.** *Under above assumptions,*

$$\frac{1}{N}R(T) = \sum_{k=1}^{K} P(K_i = k)\left[\lambda_k^2 \frac{\sigma_u^2(k)}{J} + (1-\lambda_k)^2\sigma_\alpha^2(k)\right] + o(1),$$

*where* $\lambda_k = \frac{\sigma_\alpha^2(K_i)}{\sigma_\alpha^2(K_i)+\sigma_u^2(K_i)/J}$.

*Proof.* We derive the expression by simplifying the in the summands $(\hat{\alpha}_i^{EB,T} - \alpha_i)^2$, which can

decomposed in the following part. We first analyze the estimation error from the raw estimates.

$$
\begin{aligned}
E[\sum_i \lambda_{K_i}^2(\bar{Y}_i-\alpha_i)^2] &= E[\sum_i\sum_k 1_k(K_i)\lambda_k^2(\bar{Y}_i-\alpha_i)^2] \\
&= \sum_i\sum_k \lambda_k^2 EE[1_k(K_i)(\bar{Y}_i-\alpha_i)^2|\{K_i\}] \\
&= \sum_i\sum_k \lambda_k^2 E[1_k(K_i)E[\bar{\varepsilon}_i^2|\{K_i\}]] \\
&= \sum_i\sum_k \lambda_k^2 E1_k(K_i)\frac{\sigma_u^2(K_i)}{J} \\
&= \sum_k\sum_i \lambda_k^2 P(K_i=k)\frac{\sigma_u^2(k)}{J} \\
&= \sum_k \lambda_k^2 NP(K_i=k)\frac{\sigma_u^2(k)}{J}
\end{aligned}
$$

The second term is error from estimating group mean:

$$
\begin{aligned}
&E[\sum_i(1-\lambda_{K_i})^2(\bar{\bar{Y}}_{K_i}-\mu_{K_i})^2] \\
=&E[\sum_i\sum_k(1-\lambda_k)^2 1_k(K_i)(\bar{\bar{Y}}_k-\mu_k)^2] \\
=&\sum_k\sum_i(1-\lambda_k)^2E[1_k(K_i)E[(\bar{\bar{Y}}_k-\mu_k)^2|\{K_i\}]] \\
=&\sum_k\sum_i(1-\lambda_k)^2E[1_k(K_i)E[\bar{\alpha}_k+\bar{\bar{u}}_k-\mu_k)^2|\{K_i\}]] \\
=&\sum_k\sum_i(1-\lambda_k)^2E[1_k(K_i)E[(\bar{\alpha}_k-\mu_k)^2+\bar{\bar{u}}_k^2+2(\bar{\alpha}_k-\mu_k)\bar{\bar{u}}_k]|\{K_i\}]] \\
=&\sum_k\sum_i(1-\lambda_k)^2E[1_k(K_i)[\frac{\sigma_\alpha^2(k)}{N_k}+\frac{\sigma_u^2(k)}{N_kT}]] \\
=&\sum_k(1-\lambda_k)^2E[N_k[\frac{\sigma_\alpha^2(k)}{N_k}+\frac{\sigma_u^2(k)}{N_kT}]] \\
=&\sum_k(1-\lambda_k)^2[\sigma_\alpha^2(k)+\frac{\sigma_u^2(k)}{T}]
\end{aligned}
$$

The third term is the unexplained portion of the individual effects.

$$
\begin{aligned}
E[\sum_i (1-\lambda_{K_i})^2(\mu_{K_i}-\alpha_i)^2] &= E[\sum_i \sum_k 1_k(K_i)(1-\lambda_k)^2(\mu_k-\alpha_i)^2] \\
&= E[\sum_k \sum_i 1_k(K_i)(1-\lambda_k)^2(\mu_k-\alpha_i)^2] \\
&= \sum_k (1-\lambda_k)^2 \sum_i EE[1_k(K_i)(\mu_k-\alpha_i)^2|K_i] \\
&= \sum_k (1-\lambda_k)^2 \sum_i E[1_k(K_i)\sigma_\alpha^2(K_i)] \\
&= \sum_k (1-\lambda_k)^2 NE[1_k(K_i)\sigma_\alpha^2(K_i)] \\
&= \sum_k (1-\lambda_k)^2 NP(K_i=k)\sigma_\alpha^2(k) \\
&= \sum_k NP(K_i=k)(1-\lambda_k)^2\sigma_\alpha^2(k)
\end{aligned}
$$

Combine above quantities we have the desired results. □

Using the result from Proposition 1, we can construct the empirical risk function as follows:

$$
\hat{R}(T) = \hat{P}(K_i=k)\left[\hat{\lambda}_k^2 \frac{\hat{\sigma}_u^2(k)}{J} + (1-\hat{\lambda}_k)^2\hat{\sigma}_\alpha^2(k)\right],
$$

where

$$
\begin{aligned}
\hat{\sigma}_u^2(k) &= \frac{1}{N_k}\sum_{i:K_i=k}\sum_{j=1}^{J}\frac{(Y_{ij}-\bar{Y}_i)^2}{J-1}, \\
\hat{\sigma}_\alpha^2(k) &= \frac{1}{N_k}\sum_{i:K_i=k}\frac{(\bar{Y}_i-\bar{\bar{Y}}_k)^2}{N_k-1} - \frac{\hat{\sigma}_u^2(k)}{J},
\end{aligned}
$$

and

$$\hat{P}(K_i = k) = \frac{N_k}{N}.$$

Let $\mathscr{T}_L$ denote the set of perfect binary trees with depth at most $L$. Without loss of generality for practical purposes, we can assume that $\mathscr{T}_L$ is countable. Our next proposition shows that we can consistently estimate the optimal shrinkage tree by empirical risk minimization (ERM).

**Proposition 3.2.** *Let* $\hat{T} = \text{argmin}_{T \in \mathscr{T}} \hat{R}(T)$. *Under above assumptions, we have*

$$\frac{R(\hat{T})}{\inf_{T \in \mathscr{T}_L} R(T)} \xrightarrow{a.s.} 1 \text{ as } N \to \infty.$$

*Proof.* First rewrite $R(T)$ as

$$R(T) = E[r_T(Z)],$$

where

$$r_T(z) = \lambda_{k(z)}^2 \frac{\sigma_u^2(k(z))}{J} + (1 - \lambda_{k(z)})^2 \sigma_\alpha^2(k(z)),$$
$$\sigma_u^2(k) = Var(u_{it} | K_i = k), \sigma_\alpha^2(k) = Var(\alpha_i | K_i = k)$$

Write $\hat{R}_n(T)$ as

$$\hat{R}_N(T) = \frac{1}{N} \sum_{n=1}^{N} \hat{r}_T(Z_i),$$

where

$$\hat{r}_T(z) = \lambda_{k(z)}^2 \frac{\hat{\sigma}_u^2(k(z))}{T} + (1 - \lambda_{k(z)})^2 \hat{\sigma}_\alpha^2(k(z))$$

Now, let $\tilde{T}$ be any tree in $\mathscr{T}_L$

$$\begin{aligned} R(\hat{T}) - R(\tilde{T}) &= R(\hat{T}) - \hat{R}_N(\hat{T}) + \hat{R}_N(\hat{T}) - R(\tilde{T}) \\ &\leq R(\hat{T}) - \hat{R}_N(\hat{T}) + \hat{R}_N(\tilde{T}) - R(\tilde{T}) \\ &\leq 2 \sup_{T \in \mathscr{T}_L} |\hat{R}_N(T) - R(T)| \end{aligned}$$

This implies that

$$|R(\hat{T}) - \inf_{T \in \mathscr{T}} R(T)| \leq 2 \sup_{T \in \mathscr{T}_L} |\hat{R}_N(T) - R(T)|$$

So it suffices to show

$$\sup_{T \in \mathscr{T}_L} |\hat{R}_N(T) - R(T)| \xrightarrow{a.s.} 0 \text{ as } N \to 0$$

We also introduce the following intermediate quantity for the sake of the proof:

$$R_N(T) = \frac{1}{N} \sum_{i=1}^{N} r_T(Z_i)$$

To prove ULLN, we utilize the intermediate quantity $R_N(T)$ (an infeasible $R$ estimator with known conditional variance) and apply the triangular inequality:

$$\begin{aligned} \sup_{T \in \mathscr{T}_L} |\hat{R}_N(T) - R(T)| \leq & \sup_{T \in \mathscr{T}_L} |\hat{R}_N(T) - R_N(T)| \\ & + \sup_{T \in \mathscr{T}_L} |R_N(T) - R(T)| \end{aligned}$$

Note that the second term converges to zero because the function class $\{r_T(\cdot) : T \in \mathscr{T}\}$

is GC because it has finite VC dimension. Now we focus on the first term. By definition

$$\begin{aligned}\hat{R}_N(T) &= \frac{1}{N}\sum_{i=1}^{N}\hat{r}_T(Z_i) \\ &= \frac{1}{N}\sum_{i=1}^{N}[\lambda^2_{k(Z_i)}\frac{\hat{\sigma}^2_u(k(Z_i))}{T} + (1-\lambda_{k(Z_i)})^2\hat{\sigma}^2_\alpha(k(Z_i))] \\ &= \sum_{k=1}^{K}\left[\left(\frac{1}{N}\sum_{i=1}^{N}1\{k(Z_i)=k\}\right)\left(\lambda_k^2\frac{\hat{\sigma}^2_u(k)}{J} + (1-\lambda_k)^2\hat{\sigma}^2_\alpha(k)\right)\right].\end{aligned}$$

Lastly, by triangular inequality,

$$\begin{aligned}&|\hat{R}_N(T) - R_N(T)| \\ =&\left|\sum_{k=1}^{K}\left[\left(\frac{1}{N}\sum_{i=1}^{N}1\{k(Z_i)=k\}\right)\left(\lambda_k^2\frac{\sigma^2_u(k)-\hat{\sigma}^2_u(k)}{T} + (1-\lambda_k)^2(\hat{\sigma}^2_\alpha(k)) - \sigma^2_\alpha(k))\right)\right]\right| \\ \leq&\sum_{k=1}^{K}\left[\left(\frac{1}{N}\sum_{i=1}^{N}1\{k(Z_i)=k\}\right)\left|\lambda_k^2\frac{\sigma^2_u(k)-\hat{\sigma}^2_u(k)}{J} + (1-\lambda_k)^2(\hat{\sigma}^2_\alpha(k)) - \sigma^2_\alpha(k))\right|\right] \\ \leq&\sum_{k=1}^{K}\left|\lambda_k^2\frac{\sigma^2_u(k)-\hat{\sigma}^2_u(k)}{J} + (1-\lambda_k)^2(\hat{\sigma}^2_\alpha(k)) - \sigma^2_\alpha(k))\right|.\end{aligned}$$

Then we apply Gilvenko-Cantelli theorem again to establish uniform convergence. □

**Remark.** *We briefly discuss why conventional tree algorithm may be inappropriate for our appications. Specifically, if we apply the classification and regression tree (CART), the algorithm will search for a tree $T$ and corresponding the leaf averages $\hat{\mu}_k$ to minimize*

$$\min_{T\in\mathscr{T}_L}\mathbb{E}[\sum_{j=1}^{J}(\hat{\alpha}_j - \hat{\mu}_k)^2]$$

*However, minimizing the above criterion does not generate a tree that exactly fits our purpose. The primary reason is that above criterion function ignores the fact that we're estimating the*

*treatment effects* $\{\alpha_j\}_{j=1}^J$ *not only based on treatment characteristics* $\{Z_j\}_{j=1}^J$ *but also the outcome data* $\{Y_{ij}\}_{i=1}^N$.

**Remark.** *Although this paper mainly concerns with the estimation problem, inference on the treatment effects can be carried out by adopting the honest approach proposed by Athey and Imbens [2016]. The basic idea is to split the samples into two parts: one for determining the tree structure and one for estimating the tree. In the example of teacher value-added estimations, we might use teacher data from additional sources to estimate the tree structure.* [2]

## 3.4 Simulation

In this section, we investigate the performance of propose estimators by simulation. We first examine the model selection property. The data generation process (DGP) is the following:

$$
\begin{aligned}
Y_{ij} &= \alpha_i + u_{ij}, \\
\alpha_i &= Z_{1i}\beta + v_i, \\
u_{ij} &\overset{i.i.d}{\sim} N(0,\sigma_u^2), v_i \overset{i.i.d.}{\sim} N(0,\sigma_v^2) \\
Z_{1i} &\overset{i.i.d.}{\sim} Ber(0.5)
\end{aligned}
$$

We also generate irrelevant binary variables $\tilde{Z}_{1i},...,\tilde{Z}_{Pi}$ that does not have prediction power over treatment effects.

In our first exercise, we investigate the model selection property of the tree algorithm, that is, how frequently the algorithm will correctly choose the only the relevant variable $Z_1$. In the simulation, we try out various parameter value. First, we alter the $R^2$ of regressing $\alpha_i$ on $Z_{1i}$, which is the prediction power of $Z_{1i}$ on $\alpha_i$. Intuitively, we should expect the proposed method to

[2] To fully justify the use of teacher data from other sources, we might want to assume that the effect model is stable across different settings. Such an assumption may be too strong to be true in most situations. However, even if the assumption is violated, its consequence is not fatal because only the efficiency but not the consistency is compromised.

work well if $R^2$ is high. We also experiment with different $P$ which is the number of irrelevant binary variables.

**Table 3.1.** Probability of selecting the correct variable

| $\frac{P}{N}$ | $R^2$ 0.05 | $R^2$ 0.1 | $R^2$ 0.25 |
|---|---|---|---|
| 0.05 | 0.836 | 0.975 | 0.968 |
| 0.1 | 0.716 | 0.931 | 0.998 |
| 1 | 0.275 | 0.721 | 0.983 |
| 2.5 | 0.170 | 0.588 | 0.981 |

We then compare statistical accuracy of classical empirical Bayes to that of tree-based empirical Bayes estimator. We use the unshrunk raw averages as the benchmark. It is important to note that the relative performance of the methods critically depends on the $R^2$. To make the comparison more realistic and meaningful, we experiment with a few different values of $R^2$ reported in the empirical study of teacher VAM [Rivkin et al., 2005, Rockoff et al., 2011, Dobbie, 2011].

Table 3.1 and 3.2 present the simulation result. In line with our intuition, the model selection is most accurate when $R^2$ is high. Interestingly, as long as the prediction power of $Z_{1i}$ is high, the proposed method can always pick the correct tree regardless of the number of irrelevant variables. Also, as we can in Table 3.2, the efficiency gain from the proposed method is especially high when the treatment characteristics is more correlated with the treatment effects.

**Table 3.2.** Comparison of mean-squared errors

| $R^2$ | $\bar{Y}_i$ | Classical EB | Tree EB | Reference |
|---|---|---|---|---|
| 0.03 | 1 | 0.441 | 0.439 | Rivkin et al. (2005) |
| 0.1 | 1 | 0.469 | 0.442 | Rockoff et al. (2011) |
| 0.5 | 1 | 0.462 | 0.229 | Dobbie (2011) |

## 3.5 Empirical example: the STAR project

Tennessee's Student Teacher Achievement Ratio (STAR) project was a randomized experiment designed to study the effect of class size on students' academic outcome. Among the 79 participating schools, teachers and students are randomly assigned to classes of different sizes. We use its public data set to illustrate the usage of the proposed method on teacher value-added measures.

We restrict our sample to students in first grade in small classes and focus on their outcomes for the math test score. The sample contains 332 teachers and 3979 students, and the student-teacher ratio is about 12.

### 3.5.1 Empirical strategy

We first obtain the raw estimates of teacher effect based on the following OLS specification:

$$Y_i = X_i'\beta_1 + S_i\beta_2 + \sum_{j=1}^{J} \alpha_j D_i^j + u_i$$

$Y$ is the math test score at grade 1. $X$ contains student demographics such as gender, race and free/reduced lunch status, and past test score. $S$ indicates whether a school is at urban area. Finally, $D$ is a vector of teacher assignment dummies.

We then use the OLS estimates $\{\hat{\alpha}\}_{j=1}^{J}$ and teacher characteristics data to select a tree model. Due to data limitation, teacher experience and level of education are the only available teacher characteristics that can be used. Nevertheless, the data-driven method will be more helpful when the number of available variables is higher.

Figure 1 plots the tree generated by the algorithm. It first categorizes teachers based on education level then based on years of experience. Note that for master's degree holders, the cutoff for experience is smaller than that for bachelor's degree holders. One possible explanation is that experience after the fourth-year does not matter so much for master's degree holders. Figure 2 plots the raw teacher value-added for the two education levels. We can see that teachers

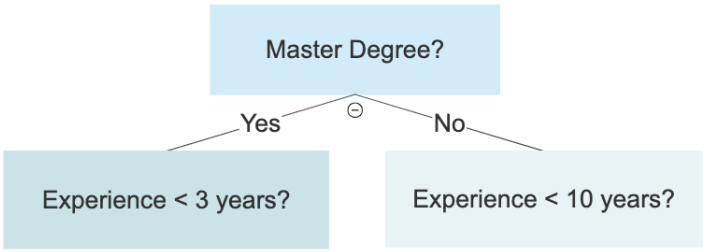



**Figure 3.1.** Estimated Tree

with master's degree are indeed on average more effective.

## 3.6 Conclusion

This paper proposes a tree-based empirical Bayes method. We extend the classical empirical Bayes estimator to allow for a tree model for effect heterogeneity. A data-driven tree algorithm for model selection is provided to optimize the performance of tree-based shrinkage estimator.

In future works, we plan to further study the theoretical properties of the proposed methods. Moreover, it would be interesting to see if the tree-based method can be used for purposes other than estimations such as testing for effect heterogeneity.

*Chapter 3, in full, is in preparation for submission. Chen, Yu-Chang."Empirical Bayes with Optimal Shrinkage Tree". The dissertation author was the primary investigator and author of this material.*

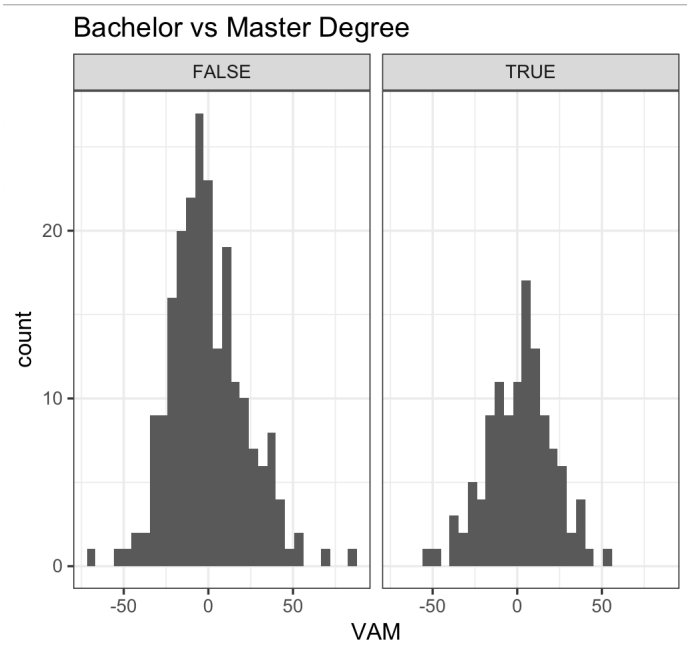



**Figure 3.2.** Distribution of Teacher VAM by Level of Education

Left: teacher VAM for bachelor-degree holders.

# Bibliography


Alberto Abadie. Semiparametric instrumental variable estimation of treatment response models. *Journal of econometrics*, 113(2):231–263, 2003.

Joshua Angrist, Peter Hull, Parag A Pathak, and Christopher Walters. Credible school value-added with undersubscribed school lotteries. *The Review of Economics and Statistics*, pages 1–46, 2021.

Timothy B. Armstrong, Michal Kolesár, and Mikkel Plagborg-Møller. Robust empirical bayes confidence intervals, 2020. URL https://arxiv.org/abs/2004.03448.

Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

Susan Athey and Stefan Wager. Policy learning with observational data. *Econometrica*, 89(1): 133–161, 2021.

Anthony B Atkinson. On the measurement of inequality. *Journal of economic theory*, 2(3): 244–263, 1970.

Eduardo M Azevedo, Alex Deng, José Luis Montiel Olea, Justin Rao, and E Glen Weyl. A/b testing with fat tails. *Journal of Political Economy*, 128(12):4614–000, 2020.

Debopam Bhattacharya and Pascaline Dupas. Inferring welfare maximizing treatment assignment under budget constraints. *Journal of Econometrics*, 167(1):168–196, 2012.

Howard S Bloom, Larry L Orr, Stephen H Bell, George Cave, Fred Doolittle, Winston Lin, and Johannes M Bos. The benefits and costs of jtpa title ii-a programs: Key findings from the national job training partnership act study. *Journal of human resources*, pages 549–576, 1997.

Stéphane Bonhomme and Martin Weidner. Posterior average effects. *Journal of Business & Economic Statistics*, 0(0):1–14, 2021. doi: 10.1080/07350015.2021.1984928. URL https://doi.org/10.1080/07350015.2021.1984928.

Christian N Brinch, Magne Mogstad, and Matthew Wiswall. Beyond late with a discrete

instrument. *Journal of Political Economy*, 125(4):985–1039, 2017.

Gary Burtless. Are targeted wage subsidies harmful? evidence from a wage voucher experiment. *ILR Review*, 39(1):105–114, 1985.

Undral Byambadalai. Identification and inference for welfare gains without unconfoundedness. 2021.

Pedro Carneiro and Sokbae Lee. Estimating distributions of potential outcomes using local instrumental variables with an application to changes in college enrollment and wage inequality. *Journal of Econometrics*, 149(2):191–208, 2009.

Pedro Carneiro, James J Heckman, and Edward Vytlacil. Evaluating marginal policy changes and the average effect of treatment for individuals at the margin. *Econometrica*, 78(1):377–394, 2010.

Pedro Carneiro, James J Heckman, and Edward J Vytlacil. Estimating marginal returns to education. *American Economic Review*, 101(6):2754–81, 2011.

Matias D Cattaneo and Max H Farrell. Large Sample Properties of Partitioning-Based Series Estimators. 2018.

Matias D Cattaneo, Max H Farrell, and Yingjie Feng. Large sample properties of partitioning-based series estimators. *The Annals of Statistics*, 48(3):1718–1741, 2020.

Simone Cerreia-Vioglio, David Dillenberger, Pietro Ortoleva, and Gil Riella. Deliberately stochastic. *American Economic Review*, 109(7):2425–45, 2019.

Raj Chetty and Nathaniel Hendren. The impacts of neighborhoods on intergenerational mobility i: Childhood exposure effects. *The Quarterly Journal of Economics*, 133(3):1107–1162, 2018.

Raj Chetty, John N Friedman, and Jonah E Rockoff. Measuring the impacts of teachers ii: Teacher value-added and student outcomes in adulthood. *American economic review*, 104(9): 2633–79, 2014.

Jessica Cohen and Pascaline Dupas. Free distribution or cost-sharing? evidence from a randomized malaria prevention experiment. *Quarterly journal of Economics*, 125(1):1, 2010.

Thomas Cornelissen, Christian Dustmann, Anna Raute, and Uta Schönberg. Who benefits from universal child care? estimating marginal returns to early child care attendance. *Journal of Political Economy*, 126(6):2356–2409, 2018.

Yifan Cui and Eric Tchetgen Tchetgen. A semiparametric instrumental variable approach to optimal treatment regimes under endogeneity. *Journal of the American Statistical Association*,

pages 1–12, 2020.

Christian M Dahl, Martin Huber, and Giovanni Mellace. It's never too late: A new look at local average treatment effects with or without defiers. Technical report, Working paper, 2020.

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

Will Dobbie. Teacher characteristics and student achievement: Evidence from teach for america. *Unpublished manuscript, Harvard University*, 2011.

Juan Dubra, Fabio Maccheroni, and Efe A Ok. Expected utility theory without the completeness axiom. *Journal of Economic Theory*, 115(1):118–133, 2004.

Bradley Efron. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press, 2010.

Bradley Efron and Carl Morris. Stein's estimation rule and its competitors—an empirical bayes approach. *Journal of the American Statistical Association*, 68(341):117–130, 1973.

Markus Frölich. Nonparametric IV estimation of local average treatment effects with covariates. *Journal of Econometrics*, 139(1):35–75, 2007.

Promit Ghosal and Bodhisattva Sen. On univariate convex regression. *Sankhya A*, 79(2):215–253, 2017.

Matthew Groh, Nandini Krishnan, David McKenzie, and Tara Vishwanath. Do wage subsidies provide a stepping-stone to employment for recent college graduates? evidence from a randomized experiment in jordan. *Review of Economics and Statistics*, 98(3):488–502, 2016.

Cassandra M Guarino, Michelle Maxfield, Mark D Reckase, Paul N Thompson, and Jeffrey M Wooldridge. An evaluation of empirical bayes's estimation of value-added teacher performance measures. *Journal of Educational and Behavioral Statistics*, 40(2):190–222, 2015.

Bruce E Hansen. Stein-like 2sls estimator. *Econometric Reviews*, 36(6-9):840–852, 2017.

Douglas N Harris. Would accountability based on teacher value added be smart policy? an examination of the statistical properties and policy alternatives. *Education finance and policy*, 4(4):319–350, 2009.

James Heckman, Justin L Tobias, and Edward Vytlacil. Simple estimators for treatment parameters in a latent-variable framework. *Review of Economics and Statistics*, 85(3):748–755, 2003.

James J Heckman. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In *Annals of economic and social measurement, volume 5, number 4*, pages 475–492. NBER, 1976.

James J Heckman and Rodrigo Pinto. Unordered monotonicity. *Econometrica*, 86(1):1–35, 2018.

James J Heckman and Edward Vytlacil. Structural equations, treatment effects, and econometric policy evaluation 1. *Econometrica*, 73(3):669–738, 2005.

James J Heckman and Edward J Vytlacil. Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proceedings of the national Academy of Sciences*, 96(8):4730–4734, 1999.

James J. Heckman and Edward J. Vytlacil. *Local instrumental variables*, page 1–46. International Symposia in Economic Theory and Econometrics. Cambridge University Press, 2001. doi: 10.1017/CBO9781139175203.003.

James J Heckman and Edward J Vytlacil. Econometric evaluation of social programs, part i: Causal models, structural models and econometric policy evaluation. *Handbook of econometrics*, 6:4779–4874, 2007.

Keisuke Hirano and Jack R Porter. Asymptotics for statistical treatment rules. *Econometrica*, 77 (5):1683–1701, 2009.

Keisuke Hirano and Jack R Porter. Statistical decision rules in econometrics. *Handbook of Econometrics*, 7, 2019.

Joel L Horowitz and Sokbae Lee. Nonparametric estimation and inference under shape restrictions. *Journal of Econometrics*, 201(1):108–126, 2017.

Peter Hull. Estimating hospital quality with quasi-experimental data. *Available at SSRN 3118358*, 2018.

Hidehiko Ichimura and Christopher Taber. Semiparametric reduced-form estimation of tuition subsidies. *American Economic Review*, 92(2):286–292, 2002.

Nikolaos Ignatiadis and Stefan Wager. Covariate-powered empirical bayes estimation. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/48f7d3043bc03e6c48a6f0ebc0f258a8-Paper.pdf.

Guido W. Imbens and Joshua D. Angrist. Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475, 1994. ISSN 00129682, 14680262. URL http://www.

jstor.org/stable/2951620.

C Kirabo Jackson, Jonah E Rockoff, and Douglas O Staiger. Teacher effects and teacher-related policies. *Annu. Rev. Econ.*, 6(1):801–825, 2014.

W James and Charles Stein. Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, volume 4, pages 361–380. University of California Press, 1961.

Thomas J Kane and Cecilia Elena Rouse. Labor-market returns to two-and four-year college. *The American Economic Review*, 85(3):600–614, 1995.

Maximilian Kasy. Partial identification, distributional preferences, and the welfare ranking of policies. *Review of Economics and Statistics*, 98(1):111–131, 2016.

Toru Kitagawa. A test for instrument validity. *Econometrica*, 83(5):2043–2063, 2015.

Toru Kitagawa and Aleksey Tetenov. Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616, 2018.

Patrick Kline and Christopher R Walters. On heckits, late, and numerical equivalence. *Econometrica*, 87(2):677–696, 2019.

Michal Kolesár. Estimation in an instrumental variables model with treatment effect heterogeneity. *Unpublished Working Paper*, 2013.

Amanda E Kowalski. How to examine external validity within an experiment. Technical report, National Bureau of Economic Research, 2018.

Charles F Manski. Statistical treatment rules for heterogeneous populations. *Econometrica*, 72 (4):1221–1246, 2004.

Charles F Manski. *Identification for prediction and decision*. Harvard University Press, 2009.

Julian Martínez-Iriarte and Yixiao Sun. Identification and estimation of unconditional policy effects of an endogenous binary treatment. *arXiv preprint arXiv:2010.15864*, 2020.

Rachael Meager. Understanding the average impact of microcredit expansions: A bayesian hierarchical analysis of seven randomized experiments. *American Economic Journal: Applied Economics*, 11(1):57–91, 2019.

Magne Mogstad, Andres Santos, and Alexander Torgovitsky. Using instrumental variables for inference about policy relevant treatment parameters. *Econometrica*, 86(5):1589–1619, 2018.

Magne Mogstad, Alexander Torgovitsky, and Christopher R Walters. Policy evaluation with multiple instrumental variables. Technical report, National Bureau of Economic Research, 2020.

Hongming Pu and Bo Zhang. Estimating optimal treatment rules with an instrumental variable: A partial identification learning approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(2):318–345, 2021.

Hongxiang Qiu, Marco Carone, Ekaterina Sadikova, Maria Petukhova, Ronald C Kessler, and Alex Luedtke. Optimal individualized decision rules using instrumental variable methods. *Journal of the American Statistical Association*, pages 1–18, 2020.

Steven G Rivkin, Eric A Hanushek, and John F Kain. Teachers, schools, and academic achievement. *Econometrica*, 73(2):417–458, 2005.

Herbert Robbins. An empirical bayes approach to statistics. Technical report, COLUMBIA UNIVERSITY New York City United States, 1956.

Peter M Robinson. Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, pages 931–954, 1988.

Jonah E Rockoff, Brian A Jacob, Thomas J Kane, and Douglas O Staiger. Can you recognize an effective teacher when you recruit one? *Education finance and Policy*, 6(1):43–74, 2011.

Jesse Rothstein. Teacher quality in educational production: Tracking, decay, and student achievement. *The Quarterly Journal of Economics*, 125(1):175–214, 2010.

Andrew Donald Roy. Some thoughts on the distribution of earnings. *Oxford economic papers*, 3 (2):135–146, 1951.

Yuya Sasaki and Takuya Ura. Welfare analysis via marginal treatment effects. *arXiv preprint arXiv:2012.07624*, 2020.

Yuya Sasaki and Takuya Ura. Estimation and inference for policy relevant treatment effects. *Journal of Econometrics*, 2021.

Emilio Seijo and Bodhisattva Sen. Nonparametric least squares estimation of a multivariate convex regression function. *The Annals of Statistics*, 39(3):1633–1657, 2011.

Tymon Słoczyński. When should we (not) interpret linear iv estimands as late? *arXiv preprint arXiv:2011.06695*, 2020.

Tom D Stanley and Stephen B Jarrell. Meta-regression analysis: a quantitative method of literature surveys. *Journal of economic surveys*, 19(3):299–308, 2005.

Charles Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley symposium on mathematical statistics and probability*, volume 1, pages 197–206, 1956.

Jörg Stoye. Minimax regret treatment choice with finite samples. *Journal of Econometrics*, 151 (1):70–81, 2009.

Max Tabord-Meehan. Stratification trees for adaptive randomization in randomized controlled trials. *arXiv preprint arXiv:1806.05127*, 2018.

Simon G Thompson and Stephen J Sharp. Explaining heterogeneity in meta-analysis: a comparison of methods. *Statistics in medicine*, 18(20):2693–2708, 1999.

Edward Vytlacil. Independence, monotonicity, and latent index models: An equivalence result. *Econometrica*, 70(1):331–341, 2002.

Edward Vytlacil. A note on additive separability and latent index models of binary choice: Representation results. *Oxford Bulletin of Economics and Statistics*, 68(4):515–518, 2006a.

Edward Vytlacil. Ordered discrete-choice selection models and local average treatment effect assumptions: Equivalence, nonequivalence, and representation results. *The Review of Economics and Statistics*, 88(3):578–581, 2006b.

Xiang Zhou and Yu Xie. Marginal treatment effects from a propensity score perspective. *Journal of Political Economy*, 127(6):3070–3084, 2019.

# Appendix A

# Appendix for chapter 1

## A.1 Semi-parametric and non-parametric methods of policy learning

### A.1.1 Semi-parametric method

Although the local IV approach to estimate MTE is fully-nonparametric in principle, it comes with the cost of slower convergence. In applications with smaller sample size, the nonparametric approach might be inappropriate. In this section of Appendix, we describe a semi-parametric approach which is less data-demanding than the nonparametric method but more flexible than parametric methods.

First, to facilitate the estimation of MTE, it is common in the MTE literature [Carneiro and Lee, 2009, Carneiro et al., 2011, Brinch et al., 2017, Zhou and Xie, 2019] to specify the potential outcomes in Equation (1.1) as a linear representation

$$Y_1 = X\beta_1 + U_1, \text{ and } Y_0 = X\beta_0 + U_0,$$

where $\beta_1$ and $\beta_0$ are unknown parameters. Another useful simplification that is often used is to assume that $(X, W, Z)$ are jointly independent of $(U_1, U_0, U_D)$. Under above specification, we

have

$$\begin{aligned}\mathbb{E}\left[Y \mid X = x, g(X,W,Z) = p\right] &= x\left(p\beta_1 + (1-p)\beta_0\right) + \mathbb{E}\left[(U_1 - U_0)\mathbf{1}\{U_D \leq p\}\right] \\ &= x\beta_0 + xp\left(\beta_1 - \beta_0\right) + \lambda(p),\end{aligned}$$

where $\lambda(p) = \int_0^p \mathbb{E}\left[U_1 - U_0 \mid U_D = u\right] du$ has an unknown form. Taking partial derivative with respect to $p$, we get a partial linear specification of the MTE

$$\text{MTE}(x,u) = x(\beta_1 - \beta_0) + \lambda'(u). \tag{A.1}$$

We consider the semiparametric estimators $\hat{\beta}_1$ and $\hat{\beta}_0$ defined in Carneiro and Lee [2009], which is based on the semiparametric estimation procedure developed by Robinson [1988]. Under mild regularity conditions, they are $\sqrt{n}$-consistent.

The semi-parametric approach relies on the parametrization of the selection correction term

$$\lambda(p) = \lambda(p \mid \theta)$$

and of the propensity score

$$g(X,W,Z) = g(X,W,Z \mid \gamma),$$

where $\theta$ and $\gamma$ are the parameters. Common choices of $\lambda(p)$ include finite-order polynomials [Brinch et al., 2017] and the inverse of normal cumulative distribution function [Carneiro et al., 2011]. As for the propensity score $g(X,W,Z \mid \gamma)$, popular choices include the logit and probit models.

The above parametrization implies that

$$\mathbb{E}\left[Y \mid X = x, g(X,W,Z) = p\right] = x\beta_0 + xp\,(\beta_1 - \beta_0) + \lambda\,(p \mid \theta), \tag{A.2}$$

which can be estimated in the following way.

Step 1: estimate $\gamma$ in accordance the specified binary choice model. For example, run probit regression if the probit model is assumed

Step 2: plug in the estimates $\hat{\gamma}$ to obtain the generated regressor $\hat{P} = g(X,W,Z \mid \hat{\gamma})$

Step 3: estimate $(\beta_1, \beta_0, \theta)$ by solving the non-linear least-square problem

$$\min_{\beta_0,\beta_1,\theta} \sum_{i=1}^{N} [Y_i - X\beta_0 - X\hat{P}\beta_1 - \lambda\,(\hat{P} \mid \theta)]^2$$

If desired, additional constraints can be incorporated in the regression to guarantee that the MTE curve $\text{MTE}(x,u) = x(\beta_1 - \beta_0) + \lambda'(u)$ satisfies certain shape restriction such as monotonicity.

Step 4: for each $X = x$ and given the estimates $(\hat{\beta}_0, \hat{\beta}_1, \hat{\gamma})$, we can estimate the optimal participation rate $\hat{u}^*_{x,w}$ by solving the equation

$$\text{MTE}(x,u) = x(\beta_1 - \beta_0) + \lambda'(u) = 0. \tag{A.3}$$

The optimal incentive assignment $z^*_{x,w}$ is then estimated by inverting the propensity score

$$\hat{z}^*_{x,w} = g^{-1}_{x,w}(u^*), \tag{A.4}$$

where $g_{x,w}(z) = g(x,w,z \mid \hat{\gamma})$.

Notice that Step 1-3 adds up to a two-stage regression, which can be formulated as a generalized method of moments estimator that is asymptotically normal under regularity conditions. Since $\hat{z}^*_{x,w}$ is a function of the estimates $(\hat{\beta}_0, \hat{\beta}_1, \hat{\gamma})$, we can derive its asymptotic properties based on the asymptotic distribution of $(\hat{\beta}_0, \hat{\beta}_1, \hat{\gamma})$ and the delta method.

### A.1.2 Policy learning under monotonicity

In this subsection, we briefly describe how monotonicity of MTE can aid the estimation for optimal policy and outline an estimation procedure. Specifically, the approach has the advantage of being fully automated as no tuning parameter is required, even though we left the MTE to be non-parametrically unspecified. The approach is simple in particularly when the propensity score is concave.[1] Noteworthily, the method also serves as a new approach for estimating the MTE.

When MTE is monotone, say, decreasing, the conditional mean $\mathbb{E}[Y|P]$ is concave in $P$. If the propensity score $P$ were known, concavity suggests that we can employ a concave regression of $Y$ on $P$. Since concave regression is uniformly consistent under regularity condition [Seijo and Sen, 2011], the optimal participation rate can also be consistently estimated from maximizing the fitted regression. Formally, let $Q_n$ nonparametric least square estimator that solves

$$\min_{Q_n(\cdot) \text{ concave}} \sum_{i=1}^{n} (Y_i - Q_n(P_i)).$$

Since $Q_n$ is only uniquely defined on the observed values of $P_i$, we linearly interpolate between points of $P_i$. Also, let $\partial Q_n(p)$ denote the subdifferentail of $Q_n$. Seijo and Sen [2011] shows that, under mild conditions, $Q_n$ is uniformly consistent of $\mathbb{E}[Y|P]$. Furthermore, if $\mathbb{E}[Y|P]$ is differentiable, then the estimated subdifferential is also uniformly consistent to the derivative. Since differentiability is typically assumed for the identification of MTE, we can expect to construct an uniformly consistent estimator of the MTE curve from the convex regression.

Since $Q_n$ is uniformly consistent, we can consistently estimate the optimal participation rate by finding $p_n = arg\max_p Q_n(p)$.[2] Inference-wise, [Ghosal and Sen, 2017] derived the asymptotic distribution of $p_n$ given homoscedasticity and smoothness of $\mathbb{E}[Y|P]$.[3] However,

[1] A third approach is to assume the monotone treatment response ($Y_1 \geq Y_0$) and apply the monotone regression of $Y$ on $Z$.

[2] Alternatively, we can obtain $p_n$ by solving for the root of the estimated MTE curve.

[3] In our context, their smoothness assumption requires $MTE$ is not flat at its root and is continuously differentiable.

it is not trivial to construct confidence interval based the asymptotic theory since the limiting distribution is not pivotal and depends the second order derivate of $\mathbb{E}[Y|P]$.

So far, we have been treating the propensity score as known, which is certainly not case in most applications. Extensions of the existing theory in convex regression is needed to apply the above approach. However, if both $E[Y|P]$ and $p(Z) = \mathbb{E}[D|Z]$ are concave, then $\mathbb{E}[Y|p(Z)]$ is also concave in $Z$. Therefore, we can estimate $\mathbb{E}[Y|Z]$ via concave regression, and solve for the optimal policy by maximizing the regression function $\mathbb{E}[Y|Z]$. The aforementioned theory of convex regression is readily applicable in this case. Notice that this method is similar to an intent-to-treat approach with [Kitagawa and Tetenov, 2018], but the shape restriction that arose from the MTE framework allows us to deal with non-discrete assignments. That is to say, the MTE framework provides a convenient yet economics-sensible way to impose structures over the instruments $Z$.

## A.2 Primitive Conditions for Monotone MTE

**Proposition A.1** (Primitive Conditions for Monotone MTE)**.** *Let the treatment be determined by*

$$D = \mathbf{1}\{\phi(X, W, Z, \Delta, V) \geq 0\},$$

*where* $(Z, W) \perp (\Delta, V)$ *and* $\Delta = Y_1 - Y_0$*. If the function* $\phi$ *satisfies the following two conditions:*

*(i)* $\phi$ *is increasing (resp. decreasing) in* $\Delta$*, and*

*(ii) for any* $(x, w, z), (x', w', z')$*,* $\phi(x, w, z, \delta_0, v_0) > \phi(x', w', z', \delta_0, v_0)$ *for some* $\delta_0, v_0 \implies \phi(x, w, z, \delta, v) > \phi(x', w', z', \delta, v)$ *for all* $\delta, v$*.*

*Then we can construct a random variable* $U_D$ *such that* $(W, Z) \perp U_D \mid X$ *and* $U_D \mid X \sim Unif[0, 1]$*, and a function* $g$ *such that the treatment selection is represented by*

$$D = \mathbf{1}\{g(X, W, Z) \geq U_D\},$$

*and* $MTE(x,u) = \mathbb{E}\left[\Delta \mid X = x, U_D = u\right]$ *is decreasing (resp. increasing) in u.*

The function $\phi(X,W,Z,\Delta,V)$ represents the utility achieved. The individual is going to select into treatment iff the utility is positive. Condition (i) essentially means the utility is monotonic in the individual treatment effect. In the return to schooling study, $\Delta$ represents the change in earnings after receiving certain level of education, in which case $\phi$ is increasing in $\Delta$. Condition (ii) means that the rank of the instrument based on individual's utility is invariant to the values of individual treatment effect and the unobserved heterogeneity $V$.

As a side note, under a rank-invariance condition in Vytlacil [2006a], we can always find a transformation, such that the treatment choice $g$ is increasing in the transformed value of the instrument. This would be helpful when the monotonicity of $g$ facilitates identification.

## A.3 Proofs

*Proof of Lemma 1.* Since $Y = (1-D)Y_0 + DY_1$, we have

$$\mathbb{E}[Y^{\pi}] = \mathbb{E}[D^{\pi}(Y_1 - Y_0)] + \mathbb{E}[Y_0].$$

By the law of iterated expectations,

$$\begin{aligned}
\mathbb{E}\left[D^{\pi}(Y_1 - Y_0)\right] &= \mathbb{E}\left[\mathbb{E}\left[\mathbf{1}\{g(X,W,\pi(X,W)) \geq U_D\}(Y_1 - Y_0) \mid U_D, X, W\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbf{1}\{g(X,W,\pi(X,W)) \geq U_D\} \mid U_D, X, W\right]\mathbb{E}\left[(Y_1 - Y_0) \mid U_D, X\right]\right] \\
&= \mathbb{E}\left[\mathbf{1}\{g(X,W,\pi(X,W)) \geq U_D\}\mathrm{MTE}(X,U_D)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbf{1}\{g(X,W,\pi(X,W)) \geq U_D\}\mathrm{MTE}(X,U_D) \mid X, W\right]\right] \\
&= \mathbb{E}\left[\int_0^{g(X,W,\pi(X,W))} \mathrm{MTE}(X,u)du\right].
\end{aligned}$$

For the cost function, we have

$$\begin{aligned}
\mathbb{E}\left[C^{\pi}\right] &= \mathbb{E}\left[c(X,W,\pi(X,W),D^{\pi})\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[c(X,W,\pi(X,W),D^{\pi}) \mid X,W\right]\right], \\
&= \mathbb{E}\left[c(X,W,\pi(X,W),1)\cdot g(X,W,\pi(X,W))\right] \\
&\quad + \mathbb{E}\left[c(X,W,\pi(X,W),0)\cdot(1-g(X,W,\pi(X,W)))\right]
\end{aligned}$$

where the last equality follows from that $\mathbb{E}\left[D \mid X,W,Z\right] = g(X,W,Z)$ and that $D$ is binary. □

*Proof of Proposition 1.1.* Notice that since $g_{x,w}$ is an one-to-one mapping from $\mathscr{Z}$ to the $[0,1]$, maximizing over $z$ can be equivalently solved by maximizing over $u$ with change of variables. Formally, following the expressions derived in the proof of Lemma 1, we have

$$\begin{aligned}
\mathbb{E}[Y^{\pi} \mid X=x,W=w] &= \mathbb{E}[Y_0 \mid X=x,W=w] + \int_0^{g(x,w,\pi(x,w))} \mathrm{MTE}(x,u')du' \\
&= \mathbb{E}[Y_0 \mid X=x,W=w] + \int_0^{u_{x,w}} \mathrm{MTE}(x,u')du'
\end{aligned}$$

where $u_{x,w} = g_{x,w}(\pi(x,w)) = g(x,w,\pi(x,w))$. Similarly, we can write

$$\mathbb{E}\left[C^{\pi} \mid X=x,W=w\right] = c(x,w,g_{x,w}^{-1}(u_{x,w}),1)\cdot u_{x,w} + c(x,w,g_{x,w}^{-1}(u_{x,w}),0)\cdot(1-u_{x,w}).$$

Since $g_{x,w}$ is assumed to be $1-1$, for each $X=x,W=w$, we can find the welfare-maximizing subsidy $\pi(x,w) \in [z_l,z_u]$ by maximizing the welfare over $u_{x,w} \in I_{x,w} = \{g(x,w,z) : z \in [z_l,z_u]\}$.

The first-order condition for the optimization problem then can be found by differentiating

$\mathbb{E}[Y^{\pi} - C^{\pi}]$ with respect to $\pi^*(x,w) \in [z_l, z_u]$ and substitute $u_{x,w} = g_{x,w}(\pi(x,w))$ as below:

$$\begin{aligned}
&\text{MTE}(x, u^*_{x,w}) \cdot \frac{d}{dz} g_{x,w}(z)\Big|_{z=g^{-1}_{x,w}(u^*_{x,w})} \\
&= u^*_{x,w} \cdot \frac{d}{dz} c(x,w,z,1)\Big|_{z=g^{-1}_{x,w}(u^*_{x,w})} + c(x,w,g^{-1}_{x,w}(u_{x,w}),1) \cdot \frac{d}{dz} g_{x,w}(z)\Big|_{z=g^{-1}_{x,w}(u^*_{x,w})} \\
&+(1-u^*_{x,w}) \cdot \frac{d}{dz} c(x,w,z,0)\Big|_{z=g^{-1}_{x,w}(u^*_{x,w})} - c(x,w,g^{-1}_{x,w}(u^*_{x,w}),0) \cdot \frac{d}{dz} g_{x,w}(z)\Big|_{z=g^{-1}_{x,w}(u^*_{x,w})}.
\end{aligned}$$

Proposition 1.1 results from calculating the corresponding derivatives for $c(x,w,z,d) = z \cdot d$. □

*Proof of Proposition 1.2.* Following the proof of Proposition 1.1, the first-order and of the optimal take-up rate problem which we define in equation (1.9) is

$$\text{MTE}(x,u) - g^{-1}_{x,w}(u) - u \cdot \frac{d}{du} g^{-1}_{x,w}(u).$$

Direct calculation yields that the second order derivate is

$$\frac{\partial}{\partial u} MTE(x,u) - 2 \cdot [\frac{d}{du} g_{x,w}(u)]^{-1}\Big|_{u=g^{-1}(u)} - u \cdot \frac{d^2}{du^2} g^{-1}_{x,w}(u).$$

By applying the formula for derivatives of inverse functions, it is not hard to see that the second order derivative is guaranteed to be non-positive since $\frac{\partial}{\partial u} MTE(x,u) \leq 0$ (Assumption 6), $\frac{d}{dz} g_{x,w}(z) \geq 0$ (Assumption 5), and $\frac{d^2}{dz^2} g_{x,w}(z) \leq 0$. The results follows from that the objective function of the optimal take-up rate problem is concave. □

*Proof of Propostion 1.3.* From Lemma 1, we know that the objective function of the optimal take-up rate problem (1.9) is given by

$$f(u_{x,w}) = \int_0^{u_{x,w}} \text{MTE}(x,u')du'$$

and that its first-order derivative is MTE. Therefore, by Assumption 7, the objective function is

convex since $MTE$ is increasing, and its solution is either the left endpoint $\underline{u}_{x,w} = g(x,w,z_l)$ or the right endpoint $\bar{u}_{x,w} = g(x,w,z_u)$ of the feasible region $I_{x,w}$. To complete the proof, notice that $f(\bar{u}_{x,w}) - f(\underline{u}_{x,w}) = \int_{\underline{u}_{x,w}}^{\bar{u}_{x,w}} \text{MTE}(x,u')du' = \int_{g(x,w,z_l)}^{g(x,w,z_u)} \text{MTE}(x,u')du'$. □

*Proof of Proposition 1.4.* Since $Y^\pi = D^\pi Y_1 + (1-D^\pi)Y_0$, we have

$$\begin{aligned}
\mathbb{E}[D^\pi Y_1 \mid X,W] &= \mathbb{E}[D^\pi Y \mid X,W] \\
&= \mathbb{E}[\mathbf{1}\{g(X,W,\pi(X,W)) \geq U_D\}Y \mid X,W] \\
&= \mathbb{E}[\mathbf{1}\{g(X,W,Z) \geq U_D\}Y \mid X,W,Z=\pi(X,W)] \\
&= \mathbb{E}[DY \mid X,W,Z=\pi(X,W)]
\end{aligned}$$

Similarly, we can show that $\mathbb{E}[(1-D^\pi)Y_1 \mid X,W] = \mathbb{E}[(1-D)Y \mid X,W,Z=\pi(X,W)]$. Combining the two steps, we have

$$\begin{aligned}
\mathbb{E}[Y^\pi \mid X,W] &= \mathbb{E}[D^\pi Y_1 + (1-D^\pi)Y_0 \mid X,W] \\
&= \mathbb{E}[DY + (1-D)Y \mid X,W,Z=\pi(X,W)] \\
&= \mathbb{E}[Y \mid X,W,Z=\pi(X,W)]
\end{aligned}$$

The second part of the proposition follows from Lemma 1 and that

$$\mathbb{E}[D^\pi \mid X,W] = \mathbb{E}[D \mid X,W,Z=\pi(X,W)].$$

$$\begin{aligned}
\mathbb{E}[C(X,W,\pi(X,W),D) \mid X,W] = {} & C(X,W,\pi(X,W),0)(1-\mathbb{E}[D \mid X,W,Z=\pi(X,W)]) \\
& + C(X,W,\pi(X,W),1)\mathbb{E}[D \mid X,W,Z=\pi(X,W)],
\end{aligned}$$

□

*Proof of Proposition 1.5.* The result follows directly from Proposition 1.2. □

*Proof of Proposition 1.6.* In the proof of Lemma 1, we have shown that

$$\mathbb{E}\left[D^{\pi}(Y_1 - Y_0)\right] = \mathbb{E}\left[\mathbf{1}\{g(X,W,\pi(X,W)) \geq U_D\}\text{MTE}(X,U_D)\right].$$

By the law of iterated expectations and $U_D \perp (X,W)$, we have

$$\begin{aligned}
\mathbb{E}\left[D^{\pi}(Y_1 - Y_0)\right] &= \mathbb{E}\left[\mathbb{E}\left[\mathbf{1}\{g(X,W,\pi(X,W)) \geq U_D\} \mid X, U_D\right]\right] \\
&= \mathbb{E}\left[\text{MTE}(X,U_D)\mathbb{E}\left[\mathbf{1}\{g(X,W,\pi(X,W)) \geq U_D\} \mid X, U_D\right]\right] \\
&= \mathbb{E}\left[\text{MTE}(X,U_D)(1 - F_{g,1}(X,U_D))\right] \\
&= \langle 1 - F_{g,\pi}, \text{MTE}\rangle.
\end{aligned}$$

So the difference in welfare between the two policies is

$$S(\pi) - S(\pi') = \langle F_{g,\pi'} - F_{g,\pi}, \text{MTE}\rangle.$$

The result follows from the definitions of $\mathscr{M}^*$ and $\mathscr{M}^\times$ in Equation (1.16). □

*Proof of Proposition 1.7.* Fix $X = x$. Since the first-order derivative of the objective

$$f'(u) = \frac{d}{du}\int_0^u \text{MTE}(x,u')du' = MTE(x,u),$$

$MTE(x,u_0) > 0$ implies that the objective is increasing at $u_0$. Combined with the fact that $f(\cdot)$ is concave when $MTE(x,\cdot)$ is decreasing, we know $\text{argmax}_{u\in[0,1]} \int_0^u \text{MTE}(x,u')du' \geq u_0$. The proof for the case of $MTE(x,u_0) < 0$ is similar. □

*Proof of Proposition 1.8 and 1.9.* Following the proof of Lemma 1, we have

$$\begin{aligned} S^*_{\text{sub}} &= \sup_{\pi}\{\mathbb{E}\left[\mathbf{1}\{g(W,\pi(W)) \geq U_D, U_D \in \text{Supp}(g(W,Z))\}\text{MTE}(U_D)\right] + \mathbb{E}[Y_0]\} \\ &= \sup_{\pi}\left\{\mathbb{E}\left[\int_{B_\pi(W)} \text{MTE}(u)du\right] + \mathbb{E}[Y_0]\right\}, \end{aligned}$$

where $B_\pi(w) = [0, g(w,\pi(w))] \cap \text{Supp}(g(w,Z))$. On the other hand, by the independence between $W$ and $(U_1, U_0, U_D)$, we have

$$\begin{aligned} S^*_{\text{dir}} &= \sup_{T}\left\{\mathbb{E}\left[T\Delta\mathbf{1}\{U \in \text{Supp}(g(Z,W))\}\right] + \mathbb{E}[Y_0]\right\} \\ &= \sup_{T}\left\{\mathbb{E}[T]\mathbb{E}\left[\int_{B(W)} \text{MTE}(u)du\right] + \mathbb{E}[Y_0]\right\} \\ &= S^*_{\text{con}}, \end{aligned}$$

where $B(w) = \text{Supp}(g(w,Z))$. Since we can choose $\pi$ such that $B \subset B_\pi$, we have $S^*_{\text{sub}} \geq S^*_{\text{dir}}$.

If the supremum in the definition of $S^*_{\text{sub}}$ is achieved by a unique policy $\pi^*$ such that $g(W,\pi^*(W))$ lies in the interior of $\text{Supp}(g(W,Z))$ with positive probability, then $S^*_{\text{sub}} > S^*_{\text{dir}}$. □

*Proof of Proposition 1.10.* To reduce the burden of notation, let $S = \mathbf{1}\{U_D \in \text{Supp}(g(W,Z))\}$. The welfare of an infeasible policy $\widetilde{T}$

$$\begin{aligned} \mathbb{E}\left[Y^{\widetilde{T}}S\right] &= \mathbb{E}\left[(Y_1 - Y_0)\widetilde{T}S\right] + \mathbb{E}[Y_0 S] \\ &= \mathbb{E}\left[\mathbb{E}\left[(Y_1 - Y_0)\widetilde{T}S \mid X, W, U_D\right]\right] + \mathbb{E}[Y_0 S] \\ &= \mathbb{E}\left[\widetilde{T}S\mathbb{E}\left[(Y_1 - Y_0) \mid X, W, U_D\right]\right] + \mathbb{E}[Y_0 S] \\ &= \mathbb{E}\left[\widetilde{T}S \cdot \text{MTE}(X, U_D)\right] + \mathbb{E}[Y_0 S], \end{aligned}$$

where the third equality holds since $\widetilde{T}$ is $\sigma(X,W,U_D)$-measurable and the fourth equality holds

as $W \perp (U_1, U_0, U_D) \mid X$. We can further rewrite the first term, which is the only part relevant for policy comparison, as

$$\begin{aligned}
\mathbb{E}\left[\widetilde{T}S \cdot \text{MTE}(X, U_D)\right] &= \mathbb{E}\left[\widetilde{T}S \cdot \text{MTE}(X, U_D)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\widetilde{T}S \cdot \text{MTE}(X, U_D) \mid X, W, U_D\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\widetilde{T}S \cdot \text{MTE}(X, U_D) \mid X, W, U_D\right]\right]
\end{aligned}$$

From here, it is clear to see that the optimal infeasible policy assigns treatment status whenever $\text{MTE}(x,u) \geq 0$. Therefore, $\widetilde{T}^* = \mathbf{1}\{\text{MTE}(x,u) \geq 0\}$. Because MTE is decreasing, fix $X = x$, we have $\{u \mid \text{MTE}(x,u) \geq 0\} = [0, u_x^*]$ for some $u_x^*$. The subsidy rule that sets subsidy equal to $g_{x,w}^{-1}(u_x^*)$ can induce the same treatment assignments and hence achieving the same welfare. □

*Proof of Proposition A.1.* Consider $\phi$ to be increasing in $\Delta$. Based on Vytlacil [2006a], we have $\phi(x,w,z,\delta,v) = \tilde{\phi}_2(\tilde{\phi}_1(x,w,z),\delta,v)$ where the functions $\tilde{\phi}_1$ and $\tilde{\phi}_2$ are constructed as follows. Pick any $\delta_0, v_0$, then define $\tilde{\phi}_1(x,w,z) = \phi(x,w,z,\delta_0,v_0)$. Define a correspondence $\tilde{\phi}_2$ as

$$\tilde{\phi}_2(\cdot,\delta,v) = \{\phi(x,w,z,\delta,v) : \tilde{\phi}_1(x,w,z) = \cdot\}.$$

It's straightforward to show that $\tilde{\phi}_2$ is a single-valued function, strictly increasing in the first argument, and increasing in the second argument. Define $\tilde{\phi}_3(\delta,v)$ by $\tilde{\phi}_2(\tilde{\phi}_3(\delta,v),\delta,v) = 0$. $\tilde{\phi}_3$ is well-defined since $\tilde{\phi}_2$ is strictly increasing in its first argument. Then we have

$$\begin{aligned}
D &= \mathbf{1}\{\phi(X,W,Z,\Delta,V) \geq 0\} \\
&= \mathbf{1}\{\tilde{\phi}_2(\tilde{\phi}_1(X,W,Z),\Delta,V) \geq 0\} \\
&= \mathbf{1}\{\tilde{\phi}_1(X,W,Z) \geq \tilde{\phi}_3(\Delta,V)\}.
\end{aligned}$$

Note that $\tilde{\phi}_3$ is decreasing in $\delta$, because for any $v$, $\tilde{\phi}_2$ increases with $\delta$, so by definition $\tilde{\phi}_3$ has to decrease when $\delta$ increases. Define $F_{3|X}$ as the conditional CDF of $\tilde{\phi}_3(\Delta, V) \mid X$. Then let $g(X,W,Z) = F_{3|X}(\tilde{\phi}_1(X,W,Z))$, and $U_D = F_{3|x}(\tilde{\phi}_3(\Delta,V))$. So it holds that $U_D \perp (W,Z) \mid X$, $U_D \mid X \sim \text{Unif}[0,1]$, and $D = \mathbf{1}\{g(X,W,Z) \geq U_D\}$. The MTE

$$\begin{aligned} \text{MTE}(x,u) &= \mathbb{E}\left[\Delta \mid X = x, U_D = u\right] \\ &= \mathbb{E}\left[\Delta \mid X = x, F_{3|x}(\tilde{\phi}_3(\Delta,V)) = u\right] \end{aligned}$$

is increasing in $u$ since both the function $F_{3|X}(\cdot)$ and $\tilde{\phi}_3(\cdot, v)$ are increasing.

□