Equality-Minded Treatment Choice*

Toru Kitagawa[†] and Aleksey Tetenov[‡]

9 November 2015

[Preliminary and incomplete, do not distribute]

Abstract

The goal of many randomized experiments and quasi-experimental studies in economics is to inform policies that aim to raise incomes and reduce economic inequality. A utilitarian policy maximizing the sum of individual incomes may not be the best choice if it magnifies economic inequality and post-treatment redistribution of income is infeasible. This paper develops a method to estimate the optimal treatment assignment policy based on observable individual covariates when the policy objective is to maximize an *equality-minded rank-dependent social welfare function*, which puts higher weight on individuals with lower-ranked outcomes. We estimate the optimal policy by maximizing a sample analogue of the rank-dependent welfare over a properly constrained set of policies. Although an analytical characterization of the optimal policy under a rank-dependent social welfare is not available even with the knowledge of potential outcome distributions, we show that the average social welfare attained by our estimated policy converges to the maximal attainable welfare at $n^{-1/2}$ rate uniformly over a large class of data distributions. We also show that this rate is minimax optimal. We provide an application of our method using the data from the National JTPA Study.

Keywords: Program evaluation, Treatment choice, Social welfare, Inequality index, Gini coefficient

*We would like to thank Ignacio Monzón for useful comments. Financial support from the ESRC through the ESRC Centre for Microdata Methods and Practice (CeMMAP) (grant number RES-589-28-0001) is gratefully acknowledged. The first author thanks the Visiting Researcher Fellowship from Collegio Carlo Alberto.

[†]Cemmap/University College London, Department of Economics. Email: t.kitagawa@ucl.ac.uk

[‡]Collegio Carlo Alberto, Email: aleksey.tetenov@carloalberto.org

1 Introduction

In causal inference studies analyzing experimental or quasi-experimental data, treatment response generally varies with individual observable characteristics. Learning about such heterogeneity from the data is essential for designing *individualized treatment rules* that assign treatments on the basis of individual observable characteristics. The optimal individualized treatment rule maximizes a social welfare criterion representing the policy maker's preferences over population distributions of post-treatment outcomes. The literature on statistical treatment choice initiated by Manski (2004) emphasizes this perspective of welfare-based empirical policy design and pursues statistically sound ways to estimate the optimal treatment assignment rule.

Research on statistical treatment rules typically focuses on the *utilitarian social welfare* criterion defined by the mean of the outcomes in the population, even though welfare economics offers a variety of alternative criteria. The utilitarian social welfare criterion offers analytical and computational convenience because it is additive across subgroups of the population and depends only on the mean of the outcome distribution. The optimal treatment rule then depends only on the conditional average treatment effect. Empirical researchers studying causal impacts of social programs have stressed the importance of evaluating distributional impacts, which are overlooked when only mean outcomes are considered (e.g., Bitler et al. (2006)). The distributional impact of a policy is especially important when the policy maker is concerned about *economic inequality* in the population.

We study the problem of treatment assignment that aims to maximize a rank-dependent *social welfare function (SWF)*, which has the form

$$W \equiv \int Y_i \cdot \omega(\mathrm{Rank}(Y_i))di, \tag{1.1}$$

where Y_i is individual i's income and the welfare weights $\omega(\mathrm{Rank}(Y_i))$ are positive and decreasing in the individual's rank in the income distribution. This class of social welfare functions is closely related to *income inequality indices*, including the widely-used Gini index. A rank-dependent SWF (1.1) can be represented as a product of the average income and one minus a relative index of inequality (such as the Gini index), which ranks income distributions that have the same mean. These social welfare functions allow for an explicit ranking of all income distributions that is increasing in the average income and decreasing in the chosen index of inequality. We discuss the theoretical properties of rank-dependent SWFs in greater detail in Section 2. While inequality measures are predominantly applied to net income, our analysis allows Y_i to denote other outcome variables of interest, including functions of income, consumption, wealth, or human capital. We will therefore refer to Y_i simply as "the outcome" in this paper.

The goal is to choose a treatment rule δ that assigns individuals to one of two treatments $d \in \{0, 1\}$ depending on their observable pre-treatment covariates $X \in \mathcal{X}$. This choice is made after experimental data has been collected and analyzed. We do not consider the problem of optimal experimental design in this paper, taking the design as given. We assume that an individual's treatment outcome does not depend on treatments received by others. The policy-maker in our setup can only impact the distribution of outcomes through the choice of a treatment assignment rule and cannot combine it with other redistributive policies.

Finding a policy that maximizes a rank-dependent SWF is a non-trivial problem without a closed-form solution even if the conditional distributions of potential outcomes ($P(Y_0|X)$ and $P(Y_1|X)$) are known. A utilitarian SWF $\left(\int u(Y_i)di\right)$ is additive across subgroups of the population, so it is optimal to assign for each subgroup the treatment with the highest conditional mean outcome in that subgroup. In contrast, a rank-dependent SWF is not separable across subgroups, as the ranking of treatment assignment rules for a given covariate value may change depending on the treatment assigned for other subgroups, as shown in an example in Section 2.1. We show in Theorem 2.1 that an equality-minded rank-dependent SWF is always maximized by a non-randomized treatment rule (assigning the same treatment to all individuals with identical covariates). This result greatly simplifies the space of treatment rules that need to be considered. It also allows us to index treatment rules by their *decision sets* $G \subset \mathcal{X}$, denoting all values of the covariates $\{X \in G\}$ for which treatment 1 is assigned.

Our aim is to estimate from the sample data a treatment assignment rule $\widehat{G} \in \mathcal{G}$ belonging to a constrained (but generally large) *set of feasible policies* $\mathcal{G} \equiv \{G \subset \mathcal{X}\}$, which is a collection of non-randomized treatment rules indexed by their decision sets. Policy makers often face legal, ethical, or political constraints that restrict how individual characteristics can be used to determine treatment assignment. One of the advantages of our framework is that it easily incorporates such exogenous restrictions. Our analytical results also require $\mathcal{G}$ to satisfy a certain complexity restriction (a finite *VC-dimension*) to prevent overfitting. We argue that this is not restrictive for public policy applications and provide rich examples of treatment rule classes that satisfy this complexity restriction.

We propose estimating the treatment rule $\widehat{G}$ by maximizing a sample analog $\widehat{W}(G)$ (see Section 3) of the population social welfare function $W(G)$, and derive consistency and convergence rate properties of this approach. We refer to this method as *Empirical Welfare Maximization (EWM).* Its properties for utilitarian social welfare are considered in Kitagawa and Tetenov (2015). The maximum welfare achievable by treatment rules from the constrained class $\mathcal{G}$ is $W^*_{\mathcal{G}} \equiv \sup_{G \in \mathcal{G}} W(G)$. We assess the statistical performance of $\widehat{G}$ in terms of the average welfare it achieves (with respect to the sampling distribution P^n of a size n sample). Specifically, we focus on the *average loss of social welfare* due to estimation of the treatment rule:

$$W^*_{\mathcal{G}} - E_{P^n}[W(\widehat{G})] \geq 0. \tag{1.2}$$

Under mild regularity conditions, we derive a non-asymptotic and distribution-free upper bound on $W^*_{\mathcal{G}} - E_{P^n}[W(\widehat{G})]$ in terms of the sample size n and a measure of complexity of $\mathcal{G}$, and show that it converges to zero at rate $n^{-1/2}$. We also show that this rate is minimax optimal over a minimally constrained class of distributions, in the sense that no other data-driven treatment rule can lead to a faster welfare loss convergence rate uniformly over the class of data distributions. Even though a rank-dependent SWF depends on the entire conditional distributions of potential outcomes given covariates, the welfare convergence rate is the same as the minimax optimal rate that the EWM rule attains in the utilitarian case (Theorems 2.1 and 2.2 in Kitagawa and Tetenov (2015)).

The remainder of this paper is organized as follows. Section 1.1 provides an overview of related literature. Section 2 discusses the properties of equality-minded rank-dependent social welfare functions and their application to the analysis of treatment choice. In Section 3, we introduce the general analytical framework of the problem and show the convergence rate properties of the EWM rule for rank-dependent welfare. Section 4 provides several extensions of the model. In Section 5, we apply our method to the experimental data from the National JTPA study. We collect all the proofs in the appendix.

1.1 Related Literature

The analysis of statistical treatment rules and empirical policy design was pioneered by Manski (2004), and is a growing area of research in econometrics. Important recent developments can be found in Dehejia (2005), Hirano and Porter (2009), Stoye (2009, 2012), Chamberlain (2011), Bhattacharya and Dupas (2012), Tetenov (2012), Kasy (2014), and Kitagawa and Tetenov (2015), among others. To the best of our knowledge, all of the existing works on treatment choice posit a utilitarian welfare criterion (the sum of the outcomes in the population) as the objective function of the policy maker. Motivated by the policy concerns about economic inequality, the current paper analyzes the treatment choice problem for a class of rank-dependent social welfare functions that embody inequality aversion. The main contributions of the current paper are (i) introducing a rank-dependent social welfare function into the treatment choice problem, (ii) developing an estimation method for the optimal policy, and (iii) characterizing its optimality in terms of the convergence rate of the average welfare loss. Several features distinguish the current analysis from the EWM approach for the utilitarian welfare case considered in Kitagawa and Tetenov (2015): (i) the analytical characterization of the optimal assignment policy is not available even with the knowledge of the distribution of (Y_0, Y_1, X), (ii) the empirical welfare criterion involves a nonlinear transformation of the empirical distribution, and (iii) as illustrated in Section 2.1 below, the rank-dependent welfare criterion is non-decomposable, so optimal assignment for one subpopulation depends on what treatment is assigned to other subpopulations. These features also distinguish this paper from machine learning and statistics literature on empirical risk minimization problems (Vapnik (1998)), where the empirical risk criterion always takes the form of a sample average.

The causal impact on non-utilitarian social welfare has been rarely considered in program evaluation studies. Aaberge et al. (2014) estimate a rank-dependent social welfare function of two policy alternatives: with and without the uniform implementation of the treatment. Firpo and Pinto (2015) estimate the impact of uniform implementation of the treatment on measures of inequality, including the Gini coefficient. Instead, the focus of the current paper is on estimating the optimal treatment rule from a large class of feasible rules.

We consider social welfare functions that satisfy the axiom of *anonymity*, i.e., they are indifferent to reshuffling of incomes between individuals. Thus our objective does not depend on the distribution of individual treatment effects $P(Y_1 - Y_0)$, which has also received attention recently (Heckman et al., 1997b, Firpo and Ridder, 2008, Fan and Park, 2010).

Building social welfare function satisfying the Pigou-Dalton principle of transfers is one of the central themes in the literature of inequality measurement and welfare economics (see Cowell (1995, 2000), Lambert (2001)). Currently, there are two widely-used classes of social welfare functions that meet the Pigou-Dalton principle. The first one is the class of Atkinson-type social welfare functions (Atkinson (1970)), which consists of the social welfare functions in the form of $W(F) = \int_0^{-\infty} U(y)dF(y)$ with $U(\cdot)$ being a concave non-decreasing function. Since the Atkinson-type social welfare function is linear in F, the EWM approach of Kitagawa and Tetenov (2015) can be readily applied by defining the outcome observations as $U(Y)$. The second class, which is this paper's main focus, is the class of rank-dependent social welfare functions introduced by Mehran (1976), Blackorby and Donaldson (1978) and Weymark (1981) and axiomatized by Yaari (1988). As noted in Machina (1982), the rank-dependent social welfare function generalizes the Atkinson-type social welfare exactly as the rank-dependent expected utility theory generalizes the classical von Neumann-Morgenstern expected utility theory (Machina (1982) and Quiggin (1982)) by relaxing the controversial independence axiom. These rich and insightful works in welfare economics have not yet been well linked to econometrics and empirical analysis for policy design. One of the main aims of the current paper is to establish a link between these two important literatures.

2 Treatment Choice with Equality-Minded Social Welfare Functions

In this paper, we propose a method to estimate treatment assignment rules that maximize an equality-minded rank-dependent social welfare function. We call a social welfare function *equality minded* if it satisfies the *Pigou-Dalton Principle of Transfers*: A transfer of income from a higher ranked individual to a lower ranked individual is always desirable when it does not change their ranks in the income distribution. Two general and widely-studied classes of social welfare function have this property. One of these classes was proposed by Atkinson

(1970). An Atkinson-type SWF could be represented as a function

$$W(F) = \int_{\mathbb{R}} U(y)dF(y)$$

of the income distribution with cumulative distribution function (cdf) $F(y)$ for some concave non-decreasing transformation $U(\cdot)$. Since the Atkinson-type social welfare function is linear in F, the EWM approach of Kitagawa and Tetenov (2015) can be readily applied by defining the outcome variable as $U(Y)$ instead of Y.

In this paper we focus on another general class of equality-minded social welfare functions: rank-dependent social welfare functions with decreasing welfare weights (also called generalized Gini social welfare functions), introduced by Mehran (1976) and Weymark (1981) and axiomatized by Yaari (1988). An equality-minded rank-dependent social welfare function admits the following representation:

$$W(F) = \int_{\mathbb{R}} \Lambda(F(y))dy, \tag{2.1}$$

where $\Lambda(\cdot) : [0,1] \to [0,1]$ is a non-increasing, non-negative function with $\Lambda(0) = 1$ and $\Lambda(1) = 0$. A rank-dependent social welfare function meets the Pigou-Dalton Principle of Transfers if and only if $\Lambda(\cdot)$ is convex.

The term *rank-dependent* is due to an equivalent representation of (2.1) as a weighted sum of incomes. If $\Lambda(\cdot)$ is differentiable, then by integration by parts $W(F)$ can be equivalently expressed as

$$W(F) = \int_0^1 F^{-1}(t)\omega(t)dt, \tag{2.2}$$

where $F^{-1}(t) \equiv \inf\{y : F(y) \geq t\}$ is the income of individuals at quantile t and $\omega(t) = \frac{d[1-\Lambda(t)]}{dt}$ is the welfare weight assigned to that quantile. Thus $W(F)$ is a weighted average of incomes Y, where the $\omega(\cdot)$ specifies the *rank-specific* weight. If the social welfare function is equality-minded then $\Lambda(\cdot)$ is convex, hence $\omega(\cdot)$ is non-decreasing and assigns larger welfare weights to individuals with lower incomes.

Throughout the paper, we will consider equality-minded social welfare functions satisfying the following assumption:

Assumption 2.1 (SWF)**.**
The policy-maker's social welfare function has representation (2.1), where $\Lambda(\cdot) : [0,1] \to [0,1]$

is a non-increasing, convex function with $\Lambda(0) = 1$ and $\Lambda(1) = 0$. We also assume that $\Lambda(\cdot)$ is differentiable and its right derivative at the origin denoted by $\Lambda'(0)$ is bounded, $|\Lambda'(0)| < \infty$.

Assumption 2.1 implies that the rank specific weight function $\omega(\cdot)$ defined in (2.2) does not asymptote at the origin, implying that the rank specific weight assigned to the lowest rank is bounded. For instance, this assumption is satisfied for the extended Gini family of social welfare functions $\{W_k(F) = \int_0^{\infty}(1 - F(y))^{k-1} : 2 < k < \infty\}$ considered in Aaberge et al. (2014). On the other hand, the Rawlsian social welfare that can be approximated by $\lim_{k\to\infty} W_k(F)$ is ruled out from our analysis.

One prominent example from this class is the *Gini social welfare function* (Blackorby and Donaldson (1978), Weymark (1981)),

$$W_{Gini}(F) \equiv E(Y)(1 - I_{Gini}(F)), \tag{2.3}$$

where $I_{Gini}(F) = 1 - \frac{\int_0^1 F^{-1}(t)2(1-t)dt}{E(Y)}$ is the Gini inequality index. This social welfare can be represented in the form of (2.2) with $\omega(t) = 2(1-t)$, implying $\Lambda(t) = (1-t)^2$. The construction of $W_{Gini}(F)$ exhibits the social planner's preference that trades off the mean of outcome Y and the *inequality* of its distribution measured by $I_{Gini}(F)$.

The key axiom of Yaari (1988) that distinguishes rank-dependent social welfare from Atkinson-type social welfare is *Invariance under a Rank-Preserving Lump-Sum Change of Incomes at the Upper Tail,* which means that the preference ordering between two income distributions F and F' is invariant to any fixed lump-sum increase (decrease) in income of all those above (below) the τ-th quantile of F and F' for any $\tau \in (0, 1)$. On the other hand, the key axiom that characterizes the Atkinson-type social welfare is the independence axiom: the preference ordering between F and F' is invariant to any mixing with another common income distribution.

We consider the problem of choosing a policy that assigns individuals to one of two treatments $d \in \{0, 1\}$ in order to maximize the chosen social welfare criterion. A treatment assignment rule $\delta : \mathcal{X} \to [0, 1]$ specifies the proportion of individuals with observable pre-treatment covariates $X \in \mathcal{X} \subset \mathbb{R}^{d_x}$ who will be assigned to treatment 1 by the policy-maker. The policy randomly assigns individuals with covariates X to the two treatments with probabilities $1 - \delta(X)$ and $\delta(X)$). The population distribution of outcomes induced by treatment rule δ has cdf

$$F_\delta(y) \equiv \int_{\mathcal{X}} \left[(1-\delta(x))F_{Y_0|X=x}(y) + \delta(x)F_{Y_1|X=x}(y)\right] dP_X(x), \tag{2.4}$$

where Y_0 and Y_1 denote the potential outcomes of the two treatments with conditional distributions $F_{Y_0|X}$ and $F_{Y_1|X}$ given X and P_X is the marginal distribution of X.

If the population distribution of (Y_0, Y_1, X) were known, the optimal policy maximizing the social welfare function (2.1) would be

$$\delta^* \in \arg\max_\delta W(F_\delta). \tag{2.5}$$

For the utilitarian welfare function $W_{util}(F) = \int_0^1 F^{-1}(t)dt$, the welfare maximization problem simplifies to

$$\delta^*_{util} \in \arg\max_\delta \int_{\mathcal{X}} \left[(1-\delta(x))E(Y_0|X=x) + \delta(x)E(Y_1|X=x)\right] dP_X(x). \tag{2.6}$$

The utilitarian social welfare is additive across covariates and depends on the outcome distributions only through their conditional means $E(Y_d|X)$. The optimal utilitarian policy is

$$\delta^*_{util} = 1\left\{x \in \mathcal{X} : E(Y_1|X=x) > E(Y_0|X=x)\right\}.$$

In contrast, the optimal rule for a rank-dependent welfare function (2.1) depends on the whole conditional distributions of potential outcomes $F_{Y_0|X}$ and $F_{Y_1|X}$, not only on their means. The optimal rule can differ from the utilitarian one if there is no first-order stochastic dominance relationship between $F_{Y_0|X}$ and $F_{Y_1|X}$ for some covariate values. Optimization of a non-utilitarian welfare criterion generally is not separable across covariates, as the desirable treatment assignment for a given covariate value may change depending on the policy chosen for other subpopulations, as shown in Section 2.1.

Even with the knowledge of the distribution of (Y_0, Y_1, X), a simple characterization of the optimal rule does not seem available for rank-dependent social welfare functions. The following theorem mitigates this complication, by substantially reducing the set of candidate treatment rules that need to be considered.

Theorem 2.1. *If $W(\cdot)$ satisfies Assumption 2.1, then for every measurable treatment rule $\delta : \mathcal{X} \to [0,1]$, there exists a non-randomized treatment rule $\delta_G(x) \equiv 1\{x \in G\}$ for some Borel set $G \subset \mathcal{X}$, such that $W(F_{\delta_G}) \geq W(F_\delta)$.*

If all upper level sets of δ belong to a collection $\mathcal{G}$ of Borel subsets of $\mathcal{X}$:

$$\{x : \delta(x) \geq t\} \in \mathcal{G}, \ \forall t \in \mathbb{R},$$

then there exists $\delta_G(x)$, $G \in \mathcal{G}$, such that $W(F_{\delta_G}) \geq W(F_\delta)$.

Proof. See Appendix A. □

This theorem shows that a treatment assignment rule maximizing an equality-minded rank-dependent welfare is non-randomized (assigns all individuals with the same covariates to the same treatment). We can therefore restrict our search for an optimal policy to the set of non-randomized rules that can be succinctly characterized by their *decision sets* $G \subset \mathcal{X}$. Decision set G determines the group of individuals $\{X \in G\}$ to whom treatment 1 is assigned. With abuse of notation, we denote the welfare of a non-randomized treatment rule with decision set G by $W(G)$, suppressing the cumulative distribution function in its argument,

$$\begin{aligned} W(G) \equiv & W(F_G), \\ F_G(y) \equiv & \int_{\mathcal{X}} \left[F_{Y_0|X=x}(y) 1\{x \notin G\} + F_{Y_1|X=x}(y) 1\{x \in G\} \right] dP_X(x). \end{aligned} \tag{2.7}$$

Our goal is to estimate from the sample data a treatment assignment rule that attains the maximum level of social welfare $W^*_{\mathcal{G}} \equiv \sup_{G \in \mathcal{G}} W(G)$ over the *set of feasible policies* $\mathcal{G} \equiv \{G \subset \mathcal{X}\}$, which is a collection of non-randomized treatment rules (subsets of the covariate space $\mathcal{X}$). An important feature of our empirical welfare maximization approach is that $\mathcal{G}$ must have a finite *Vapnik-Cervonenkis (VC) dimension*:

Assumption 2.2 (VC)**.**
The class of decision sets $\mathcal{G}$ has a finite VC-dimension $v < \infty$.

The VC-dimension is a restriction on the complexity of the set of feasible policies. Without it, maximizing a sample analogue of $W(G)$ over G can lead to arbitrarily complicated policies (overfitting) and prevent us from learning the optimal policy on the basis of a finite number of observations. It does not require $\mathcal{G}$ to be finite and allows for very large classes of treatment rules. For example, a class of treatment rules defined by a linear equation in functions of x, $\mathcal{G} \equiv \left\{ G = \{x : \sum_{i=1}^m \beta_j f_j(x) \geq 0\}, \beta \in \mathbb{R}^m \right\}$ has a finite VC-dimension. See Kitagawa and Tetenov (2015) for other examples of $\mathcal{G}$ that satisfy Assumption 2.2.

2.1 Illustrating Example

To illustrate how the optimal rules differ between the rank-dependent welfare case and the utilitarian welfare case, consider the Gini social welfare case, where the social welfare function can be written as $W_{Gini}(F) = \int_0^\infty (1 - F(y))^2 dy$.

Consider first the welfare orderings over the log-normally distributed outcome variable, $Y \sim \log N(\mu, \sigma^2)$, with ignoring the treatment choice problem. It is well known that the mean of Y is given by $E(Y) = \exp(\mu + \sigma^2/2)$. The Gini coefficient for *inequality* of $\log N(\mu, \sigma^2)$ is given by $2\Phi\left(\sigma/\sqrt{2}\right) - 1$ (see, e.g., Cowell (1995)), where $\Phi(\cdot)$ is the cdf of the standard normal. By (2.3), we have

$$W_{Gini}(F) = 2 \exp\left(\mu + \frac{\sigma^2}{2}\right)\left[1 - \Phi\left(\frac{\sigma}{\sqrt{2}}\right)\right]. \quad (2.8)$$

This welfare function is increasing in μ, whereas is not monotonic in σ. For instance, when $\mu = 0$, $W_{Gini}(F_Y)$ is decreasing in σ for $\sigma \leq 0.87$ and increasing for $\sigma > 0.87$. See Figure 1 for a plot of $W_{Gini}(F)$ over $\sigma \in [0, 2]$ with fixing $\mu = 0$. The U-shape of the Gini social welfare indicates that for $\sigma \leq 0.87$, the negative contribution to the social welfare from an increase in the Gini coefficient dominates the positive contribution from the increase of the mean, while for $\sigma > 0.87$, this dominance relationship reverses. In Figure 2, we plot the densities of the log-normal distributions for $\sigma = 0.25$, 0.5, and 1. Since $E(Y)$ is monotonically increasing both in μ and σ, higher σ is more preferable in terms of the utilitarian social welfare. In contrast, as shown in the welfare values plotted in Figure 1, the Gini social welfare yields the complete opposite welfare orderings over the three log-normal distributions in Figure 2.

Consider now the treatment choice problem. Suppose there is only one binary covariate $X \in \{a, b\}$ with $\Pr(X = a) = \Pr(X = b) = 1/2$. Consider the following parameterization of the potential outcome distributions:

$$\begin{aligned} Y_1|(X = a) &\sim \log N(\mu_a, \sigma_a^2), \quad Y_0|(X = a) \sim \log N(0, 0.8^2), \\ Y_1|(X = b) &\sim \log N(\mu_b, \sigma_b^2), \quad Y_0|(X = b) \sim \log N(0, 0.8^2). \end{aligned} \quad (2.9)$$

Since it suffices to consider non-randomized rules to search for an optimal one (Proposition 2.1), we consider ranking the following four policies: $\mathcal{G} = \{\emptyset, \{a\}, \{b\}, \{a, b\}\} \equiv \{G_\emptyset, G_a, G_b, G_{ab}\}$.

Suppose $\sigma_a = \sigma_b = 0.8$ and $\mu_a, \mu_b > 0$. Then, in each subpopulation of $X = a$ and $X = b$, the distribution of Y_1 stochastically dominates the distribution of Y_0. Since the rank-dependent

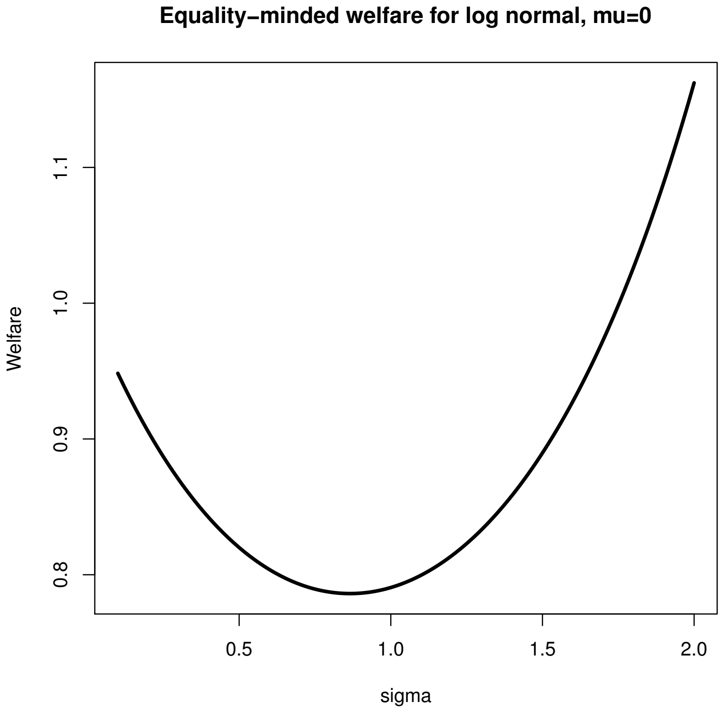

Figure 1: Equality-minded welfare for $\log N(0, \sigma^2)$.

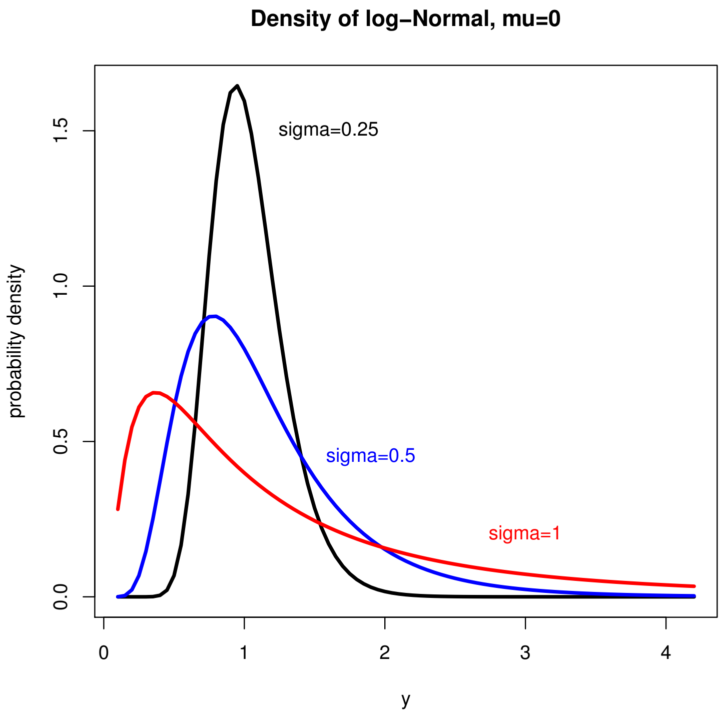

Figure 2: Density of $\log N(0, \sigma^2)$.

social welfare is clearly monotonic in the first-order stochastic dominance relationship, treating both $\{X = a\}$ and $\{X = b\}$ maximizes the Gini social welfare. This optimal rule indeed coincides with that of the utilitarian welfare case. Furthermore, in the current special case, since the stochastic dominance relationship is available between the Y_1- and Y_0-distributions in each subpopulation, the optimal rule can be obtained by focusing on the treatment choice problem separately in each subpopulation even with the Gini social welfare.

These results in case of $\sigma_a = \sigma_b$ changes drastically once we allow $\sigma_a \neq \sigma_b$. Suppose $\mu_a = \mu_b = 0$, while we vary σ_a and σ_b independently over $[0.1, 1.6]$. As the mean of log normal random variable is increasing in σ-parameter, the optimal treatment rule for the utilitarian welfare is obtained by

$$G^*_{util} = \begin{cases} G_\emptyset & \text{if } \sigma_a < 0.8 \text{ and } \sigma_b < 0.8, \\ G_a & \text{if } \sigma_a \geq 0.8 \text{ and } \sigma_b < 0.8, \\ G_b & \text{if } \sigma_a < 0.8 \text{ and } \sigma_b \geq 0.8, \\ G_{ab} & \text{if } \sigma_a \geq 0.8 \text{ and } \sigma_b \geq 0.8. \end{cases}$$

In Figure 3, we plot the utilitarian optimal treatment rule at each grid point of $(\sigma_a, \sigma_b) \in [0.1, 1.6]^2$. Since the utilitarian social welfare is the sum of the welfares over the subpopulations, a treatment preferable for one subpopulation does not depend on what treatment is assigned to the other subpopulation. The region that each rule in $\mathcal{G}$ is optimal creates quadrant partition as shown in Figure 3.

In Figure 4, we plot the optimal policies in terms of the Gini social welfare. The regions that each rule in $\mathcal{G}$ become optimal are strikingly different compared with the utilitarian welfare case (G^*_{util}) shown in Figure 3. In the neighborhood of $(\sigma_a, \sigma_b) = (0.8, 0.8)$, the subpopulations to be treated under the Gini social welfare are the converse of those to be treated under the utilitarian welfare. This is because the Gini social welfare is decreasing in σ in the neighborhood of $\sigma = 0.8$ (Figure 1), while the utilitarian welfare is monotonically increasing in σ. Another notable difference is that in contrast to the quadrant partition observed in the utilitarian welfare case (Figure 3), the partition in the equality-minded welfare case are more complex (Figure 4). One treatment rule becomes optimal in disconnected regions, e.g., G_{ab} is optimal in the south-west and north-east region in the plot. Furthermore, the region that G_a becomes

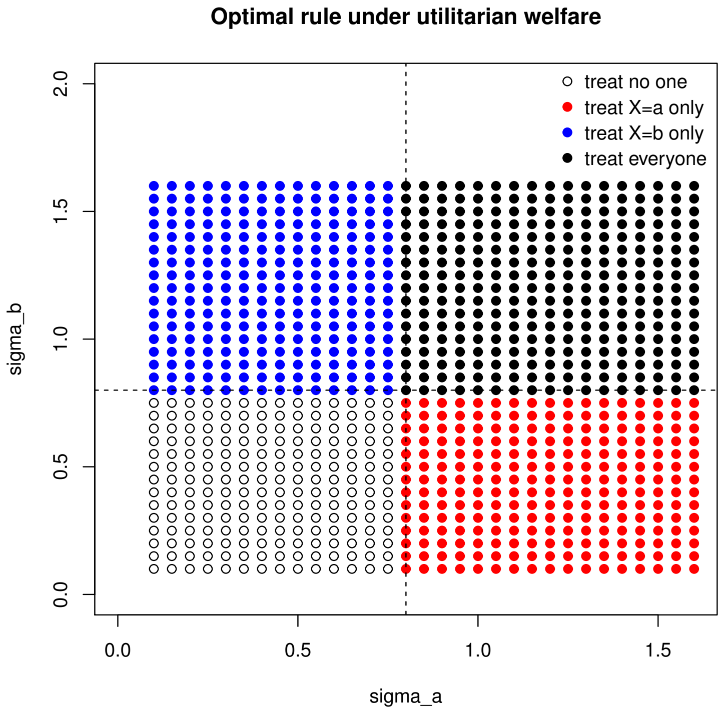

Figure 3: Utilitarian Welfare: Optimal Policies

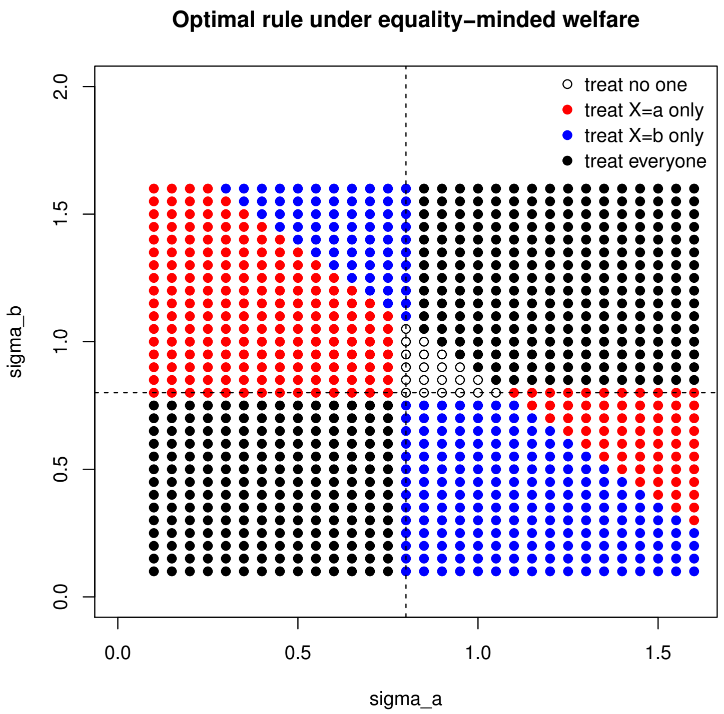

Figure 4: Gini Social Welfare: Optimal Policies

optimal can face the region that G_b become optimal. On the borders of these regions, the policy maker ultimately compares either to treat $X = a$ only or $X = b$ only, rather than whether to additionally treat the other subpopulation.

3 EWM for Equality-Minded Welfare

We proceed to propose our method of estimating the treatment rule in a finite sample setting and analyze its properties. Let the data be a size n random sample of $Z_i = (Y_i, D_i, X_i)$, where X_i refers to the observable pre-treatment covariates of individual i, $D_i \in \{0, 1\}$ is a binary indicator of the individual's experimental treatment, and $Y_i \in \mathbb{R}_+$ is her/his observed post-treatment outcome. The population from which the sample is drawn is characterized by P, a joint distribution of $(Y_{0i}, Y_{1i}, D_i, X_i)$, where Y_{0i} and Y_{1i} are the potential outcomes that would be observed if i's treatment status were $D_i = 0$ and $D_i = 1$, respectively. We assume that the propensity score $e(X) = \Pr(D = 1|X = x)$ is known, Section 4.2 extends the analysis to include propensity score estimation. Based on this data, the policy-maker has to choose a conditional treatment rule $G \in \mathcal{G}$ that determines whether individuals with covariates X in a target population will be assigned to treatment 0 or to treatment 1. The following are our maintained assumptions about the class of population distributions of (Y_0, Y_1, D, X):

Assumption 3.1.
(UCF) Unconfoundedness: $(Y_0, Y_1) \perp D|X$.
(BO) Bounded Outcomes: There exists $M < \infty$ such that the support of Y is contained in $[0, M]$.
(SO) Strict Overlap: There exist $\kappa \in (0, 1/2]$ such that the propensity score satisfies $e(x) \in [\kappa, 1 - \kappa]$ for all $x \in \mathcal{X}$.

These assumptions generally hold if the data come from an experiment with randomized treatment assignment. In observational studies, on the other hand, *Unconfoundedness* rules out situations in which the observed treatment assignments depend on subjects' unobserved characteristics that can drive their potential outcomes. *Strict Overlap* can be also violated in an observational study if only one of the treatments is assigned in the sampling process for some covariate values. It is important to note that we do not constrain any feature of the joint distribution of (Y_0, Y_1, X) except that Y_0 and Y_1 have a bounded support. The outcome variable and the covariates can be discrete, continuous, or their combination, and the support of X does not have to be bounded.

We propose estimating the treatment rule by maximizing a sample analog of the population social welfare criterion and derive consistency and convergence rate properties of this approach. This method is analogous to the Empirical Welfare Maximization (EWM) method proposed by Kitagawa and Tetenov (2015) for the utilitarian social welfare case. The *equality-minded EWM treatment rule* $\widehat{G}$ maximizes a sample analog $\widehat{W}(G)$ of the welfare criterion over the set of feasible rules $G \in \mathcal{G}$. The unknown outcome distribution F_G induced by treatment rule G in (2.7) could be estimated by its sample analog

$$\widehat{F}_G(y) \equiv \frac{1}{n}\sum_{i=1}^{n}\left[\frac{D_i \cdot 1\{Y_i \leq y\}}{e(X_i)} \cdot 1\{X_i \in G\} + \frac{(1-D_i)\cdot 1\{Y_i \leq y\}}{1-e(X_i)} \cdot 1\{X_i \notin G\}\right]. \tag{3.1}$$

The sample analog of welfare, by analogy with equation (2.1) is defined as[1]

$$\widehat{W}(G) \equiv \int_0^M \Lambda(\widehat{F}_G(y) \wedge 1)dy. \tag{3.2}$$

The equality-minded EWM treatment rule is then

$$\widehat{G} \in \arg\max_{G\in\mathcal{G}} \widehat{W}(G). \tag{3.3}$$

The cdf sample analog $\widehat{F}_G(y)$ may not itself be a proper cdf, but it could be rescaled to a proper cdf as follows: [2]

$$\widehat{F}_G^R(y) \equiv \begin{cases} \widehat{F}_G(y)/\widehat{F}_G(M) & \text{if } \widehat{F}_G(M) > 0, \\ 1\{y \geq M\} & \text{if } \widehat{F}_G(M) = 0. \end{cases} \tag{3.4}$$

We also consider properties of the equality-minded EWM rule using rescaled cdfs $\widehat{F}_G^R(y)$

$$\widehat{G}^R \equiv \arg\max_{G\in\mathcal{G}} \widehat{W}^R(G), \text{ where } \widehat{W}^R(G) \equiv \int_0^M \Lambda(\widehat{F}_G^R(y))dy. \tag{3.5}$$

[1]The minimum ($\wedge$) of $\widehat{F}_G(y)$ and 1 is taken because $\widehat{F}_G(y)$ may take values greater than 1, for which $\Lambda(\cdot)$ is not defined.

[2]$\widehat{F}_G(y)$ could also be rescaled to a proper cdf without specifying the value of M, by replacing M with the largest observed outcome value $\max_{1\leq i\leq n} Y_i$

3.1 Rate Optimality of EWM

The next theorem derives a uniform upper bound of the average welfare loss of the EWM rule.

Theorem 3.1. *Under Assumptions 2.1 and 2.2, the average welfare loss of treatment rules $\widehat{G}$ and $\widehat{G}^R$ satisfies*

$$\sup_{P\in\mathcal{P}} E_{P^n}\left[W_{\mathcal{G}}^* - W(\widehat{G})\right] \leq C|\Lambda'(0)|\frac{M}{\kappa}\sqrt{\frac{v}{n}}, \tag{3.6}$$

$$\sup_{P\in\mathcal{P}} E_{P^n}\left[W_{\mathcal{G}}^* - W(\widehat{G}^R)\right] \leq C^R|\Lambda'(0)|\frac{M}{\kappa}\sqrt{\frac{v}{n}}, \tag{3.7}$$

where $\mathcal{P}$ is the class of all distributions satisfying Assumption 3.1 and $C > 0$, $C^R > 0$ are universal constants.

Proof. See Appendix A.

This theorem shows that for a large class of data generating processes characterized by Assumption 3.1, the welfare loss of the EWM rule is guaranteed to converge to the maximal attainable welfare no slower than at $n^{-1/2}$ rate. The uniform convergence rate of $n^{-1/2}$ coincides with that of the EWM rule for the utilitarian welfare shown in Theorem 2.1 of Kitagawa and Tetenov (2015). This is a nontrivial result, given that the rank-dependent welfare function depends on the whole conditional distributions of the potential outcomes given covariates, rather than only on their conditional means, as is the case for the utilitarian welfare criterion.

The next proposition provides a universal lower bound for the worst-case average welfare loss of any treatment rule.

Theorem 3.2. *Suppose that Assumptions 2.1 and 2.2 hold with $v \geq 2$, then for any non-randomized treatment choice rule $\widehat{G}$ as a function of the sample, and any $t^* \in (0, 1]$, it holds*

$$\sup_{P\in\mathcal{P}} E_{P^n}\left[W_{\mathcal{G}}^* - W(\widehat{G})\right] \geq \frac{e^{-4}}{2}M\,|\Lambda'(t^*)|\,\sqrt{t^*}\sqrt{\frac{v-1}{n}} \text{ for all } n \geq 4(v-1)/t^*. \tag{3.8}$$

Proof. See Appendix A.

By Assumption 2.1, $\Lambda(\cdot)$ is convex and differentiable, hence continuously differentiable. Since $|\Lambda'(0)| > 0$, there also exists some $t^* > 0$ for which $|\Lambda'(t^*)| > 0$, hence the bound (3.8) is always positive for some $t^* > 0$. A comparison of the lower bound of this theorem with the welfare loss upper bound of the EWM rule obtained in Theorem 3.1 shows that the EWM rule is *minimax rate optimal* over the class of data generating processes satisfying Assumption 3.1. We can therefore claim that in the absence of any additional restrictions other than Assumption 3.1, no other data-driven procedure for obtaining a non-randomized rule can outperform the EWM rule in terms of the uniform convergence rate over $\mathcal{P}$. This optimality claim is analogous to that of the EWM rule for the utilitarian welfare case (Theorems 2.1 and 2.2 in Kitagawa and Tetenov (2015)), and the minimax optimal rate attained by the equality-EWM rule is the same as the optimal rate in the utilitarian welfare case. It is remarkable to see that even in the absence of any analytical characterization of the true optimal assignment rule in terms of the population distribution of (Y_0, Y_1, X), only maximizing the empirical welfare leads us to a policy by implementing which the social welfare is guaranteed to reach the attainable maximum welfare at the minimax optimal rate.

It is also worth noting that the VC-dimension of $\mathcal{G}$ appears in the same order both in the upper and lower bound expressions of Theorems 3.1 and 3.2. Since these bounds are non-asymptotic, we can let v increase with the sample size, and we can still conclude the minimax rate optimality of the equality-EWM rule. This insight is similar to the EWM rule for the utilitarian welfare case (Remark 2.2 in Kitagawa and Tetenov (2015)).

4 Extensions

4.1 Social Welfare is Defined on a Population nesting the Target Population

As illustrated in Section 2.1, one of the distinguishing features of the rank-dependent social welfare is that it is *not* additive over the subpopulations. This implies that if the population for which the policy can be implemented (e.g., unemployed workers) is only a subset of the population on which the rank-dependent social welfare function is defined (e.g., population of a country), it is important to explicitly take into account the outcome distribution of the non-targeted subset of the population in estimating the optimal assignment rule.

Suppose that the social welfare function is defined on the population whose outcome distribution is given by a mixture of two subpopulations

$$J = \eta F + (1-\eta)H, \quad \eta \in (0,1), \tag{4.1}$$

where F is the outcome distribution of those to whom a treatment assignment policy applies and H is the outcome distribution of those not affected by the treatment assignment. The mixture weight η represents the size of subpopulation F. For simplicity, we assume that η and H are known to the social planner. We also assume that the outcome distribution H is invariant to a treatment assignment policy applied to the subpopulation F, e.g., no spill-over or general equilibrium effect across F and H. Experimental data of (Y, D, X) with size n are drawn from the subpopulation F.

The social welfare function is defined on J. Hence, implementing treatment assignment rule $\{X \in G\}$ on subpopulation F attains

$$W(J_G) \equiv \int_0^\infty \Lambda(\eta F_G(y) + (1-\eta)H(y))dy,$$

where $F_G(\cdot)$ is the cdf defined in (2.7). The empirical welfare maximization method in the current case is set up as maximizing a sample analogue of $W(J_G)$,

$$\widehat{G} \in \arg\max_{G\in\mathcal{G}} W(\eta \widehat{F}_G(y) + (1-\eta)H(y)),$$

where $\widehat{F}_G(y)$ is as defined in (3.1).

The uniform convergence property of Theorem 3.1 carries over to the current case except for a minor change in the constant term, as shown in the next corollary. Its proof can be obtained similarly to Theorem 3.1.

Corollary 4.1. *Under Assumptions 2.1, 2.2 and 3.1,*

$$\sup_{P\in\mathcal{P}} E_{P^n}\left[\sup_{G\in\mathcal{G}} W(J_G) - W(J_{\widehat{G}})\right] \leq C|\Lambda'(0)|\eta\frac{M}{\kappa}\sqrt{\frac{v}{n}}, \tag{4.2}$$

where $C > 0$ is a universal constant defined in Theorem 3.1.

4.2 EWM with Estimated Propensity Score

Unknown propensity score is common in observational studies. This section considers that the equality-minded EWM approach with estimated propensity scores and investigates the influence of the lack of knowledge on propensity scores to the uniform convergence rate of the welfare loss criterion.

Let $\hat{e}(x)$ be an estimator for the propensity score $\Pr(D = 1|X = x)$. The empirical welfare criterion of assignment policy $\{X \in G\}$ with the estimated propensity scores plugged in is given by

$$W(\widehat{F}^e_G) = \int_0^M \Lambda(\widehat{F}^e_G(y) \wedge 1)dy,$$
$$\widehat{F}^e_G(y) \equiv \frac{1}{n}\sum_{i=1}^n \left[\frac{D_i 1\{Y_i \le y\}}{\hat{e}(X_i)} \cdot 1\{X_i \in G\} + \frac{(1-D_i)1\{Y_i \le y\}}{1-\hat{e}(X_i)} \cdot 1\{X_i \notin G\}\right].$$

The equality-EWM rule with estimated propensity score is defined accordingly as

$$\widehat{G}^e \in \arg\max_{G \in \mathcal{G}} W(\widehat{F}^e_G).$$

To characterize the uniform convergence rate of the welfare loss of $\widehat{G}^e$, we first assume that $\hat{e}(\cdot)$ is uniformly consistent to the true propensity score $e(\cdot)$ in the following sense.

Assumption 4.1. For a class of data generating processes $\mathcal{P}_e$, there exists a sequence $\phi_n \to \infty$ such that

$$\limsup_{n\to\infty} \sup_{P \in \mathcal{P}_e} \phi_n E_{P^n}\left[\frac{1}{n}\sum_{i=1}^n \left|1 - \frac{e(X_i)}{\hat{e}(X_i)}\right|\right] < \infty \quad \text{and} \tag{4.3}$$
$$\limsup_{n\to\infty} \sup_{P \in \mathcal{P}_e} \phi_n E_{P^n}\left[\frac{1}{n}\sum_{i=1}^n \left|1 - \frac{1-e(X_i)}{1-\hat{e}(X_i)}\right|\right] < \infty$$

hold.

When the class of data generating processes $\mathcal{P}_e$ constrains the propensity score to a parametric family and the support of X to being compact, parametric estimator $\hat{e}(X_i)$ satisfies this assumption with $\phi_n = n^{1/2}$. When the propensity scores are estimated nonparametrically instead, ϕ_n is generally slower than $n^{1/2}$. The rate of ϕ_n for nonparametrically estimated propensity scores depends on the smoothness of $e(\cdot)$ and the dimension of X, as we discuss further below.

Theorem 4.1. *Suppose Assumptions 2.1, 2.2 and 3.1 hold. For a class of data generating process $\mathcal{P}_e$, if an estimator for the propensity score satisfies Assumption 4.1, then*

$$\sup_{P \in \mathcal{P}_e \cap \mathcal{P}} E_{P^n}\left[W^*_{\mathcal{G}} - W(F_{\widehat{G}^e})\right] \le O\left(\phi_n^{-1} \vee \sqrt{\frac{v}{n}}\right). \tag{4.4}$$

Proof. See Appendix A.

This theorem extends Theorem 2.5 (e) of Kitagawa and Tetenov (2015) to the rank-dependent social welfare case. The shown uniform convergence rate implies that the parametrically estimated propensity score achieving $\phi_n = n^{1/2}$ does not affect the convergence rate property of the welfare loss. With nonparametrically estimated propensity score, on the other hand, the uniform welfare loss convergence rate can be slower than the one with the known propensity score obtained in Theorem 3.1. For instance, if $\hat{e}(X_i)$ is estimated by local polynomial regression (with proper trimming), then for a suitably defined $\mathcal{P}_e$, we have $\phi_n = n^{\frac{1}{2+d_x/\beta_e}}$, where $\beta_e \geq 1$ is the parameter constraining smoothness of $e(\cdot)$ in terms of the degree of Hölder class of functions and $d_x \geq 1$ is the dimension of X. Since $\frac{1}{2+d_x/\beta_e} < \frac{1}{2}$, the upper bound of the uniform convergence rate shown in Theorem 4.1 implies

$$\sup_{P \in \mathcal{P}_e \cap \mathcal{P}} E_{P^n}\left[W_{\mathcal{G}}^* - W(F_{\widehat{G}^e})\right] \leq O\left(n^{-\frac{1}{2+d_x/\beta_e}}\right), \tag{4.5}$$

as far as the VC-dimension of $\mathcal{G}$ is either constant or does not grow too fast as the sample size. For a formal derivation of (4.5) and the precise construction of the local polynomial estimator for $e(\cdot)$, see Appendix B.

4.3 Capacity Constraint

5 Empirical Illustration

To illustrate equality-minded empirical treatment choice, we apply our method to the experimental data from the National Job Training Partnership Act (JTPA) Study. A detailed description of the study and an assessment of average program effects for five large subgroups of the target population is found in Bloom et al. (1997). The study randomized whether applicants were eligible to receive a mix of training, job-search assistance, and other services provided by the JTPA for a period of 18 months. It collected background information about the applicants prior to random assignment, as well as administrative and survey data on applicants' earnings in the 30-month period following assignment. We use the same sample of 11204 adults (22 years and older) used in the original evaluation of the program and many subsequent studies (Bloom et al., 1997, Heckman et al., 1997a, Abadie et al., 2002). The

Table 1: Gini welfare and average income under selected treatment rules that condition on education and pre-program earnings.

Treatment rule:	Representative income for Gini welfare function	Average income	Share of population assigned to treatment
Simple treatment rules			
Treat nobody	\$6,610	\$14,969	0
Treat everyone	\$7,310	\$16,238	1
Quadrant class conditioning on years of education and pre-program earnings			
Maximize average income	\$7,406	\$16,583	0.8266
Maximize Gini welfare	\$7,440	\$15,953	0.7266
Linear class conditioning on years of education and pre-program earnings			
Maximize average income	\$7,411	\$16,626	0.8197
Maximize Gini welfare	\$7,478	\$16,063	0.7760

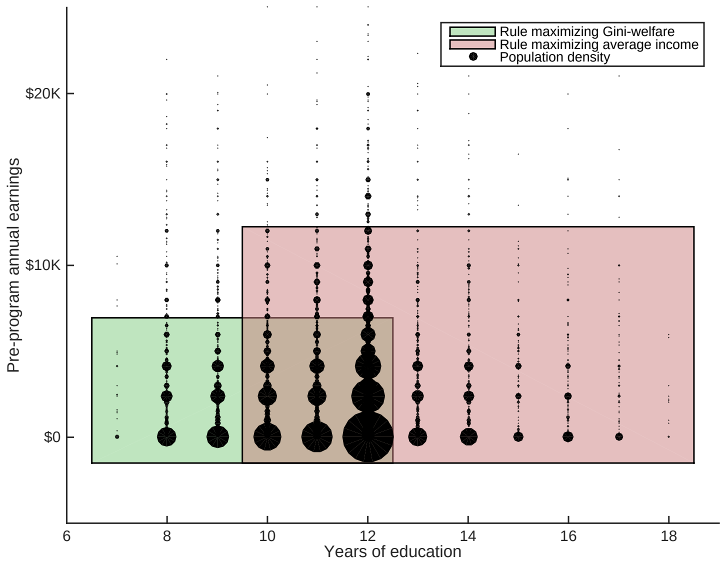

Figure 5: Treatment rules that maximize Gini welfare function and average income within the quadrant class conditioning on years of education and pre-program earnings

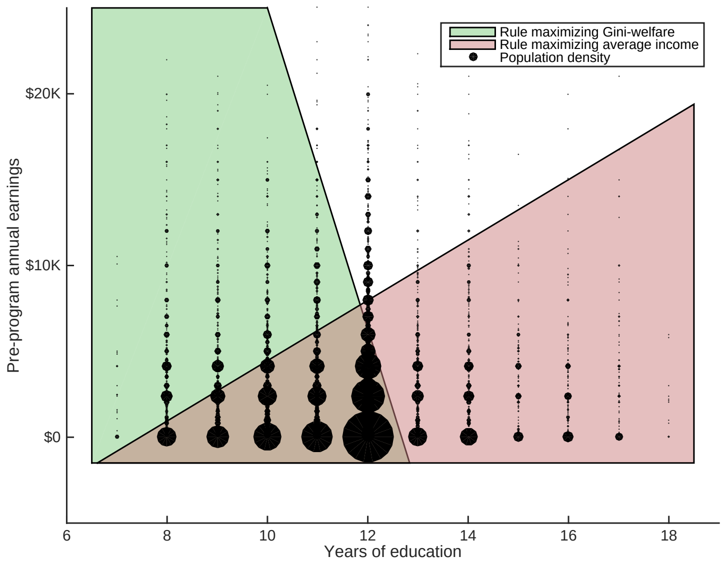

Figure 6: Treatment rules that maximize Gini welfare function and average income within the linear class conditioning on years of education and pre-program earnings

probability of being assigned to treatment was one third in this sample.

For this illustration, total individual earnings in the 30-month period following program assignment serve as the measure of income. We evaluate the Gini social welfare function as an example of an equality-minded criterion. We use average income as an example of a utilitarian criterion. For simplicity, we consider only the distribution of earnings in the population sampled for the experiment in the social welfare function. In practice, the targeted population is only part of the total population. The social welfare function should be evaluated on the full income distribution that combines both targeted and non-targeted individuals.

Pre-treatment variables on which we consider conditioning treatment assignment are the individual's years of education and earnings in the year prior to assignment. We do not use race, sex, or age to condition treatment assignment. Though treatment effects may vary with these characteristics, using them to condition treatment assignment is often socially unacceptable and illegal. Education and earnings are verifiable characteristics, which is also important for conditioning treatment assignment. The performance of treatment rules that condition on unverifiable characteristics is hard to evaluate if individuals change their self-reported characteristics to obtain their desired treatment assignment.

Table 1 compares empirical estimates of social welfare measures for a few alternative treatment rules. First, we consider simple treatment rules that either assign nobody or everybody to treatment. Second, we consider empirically optimal rules from the class of *quadrant treatment rules:*

$$\mathcal{G}_Q \equiv \left\{ \begin{array}{c} \{x : s_1(\text{education} - t_1) > 0 \ \&\ s_2(\text{prior earnings} - t_2) > 0\}\,, \\ s_1, s_2 \in \{-1, 0, 1\}, t_1, t_2 \in \mathbb{R} \end{array} \right\}. \tag{5.1}$$

This class of treatment eligibility rules is easily implementable and is often used in practice. To be assigned to treatment according to such rules, an individual's education and pre-program earnings both have to be above (or below) some specific thresholds. Third, we consider empirically optimal rules from the class of *linear treatment rules:*

$$\mathcal{G}_{LES} \equiv \left\{ \{x : \beta_0 + \beta_1 \cdot \text{education} + \beta_2 \cdot \text{prior earnings} > 0\}\,, \beta_0, \beta_1, \beta_2 \in \mathbb{R} \right\}. \tag{5.2}$$

The first column evaluates the Gini social welfare function, expressed in terms of the *representative income* of the policy. The income distribution generated by the policy is valued as much as an equal income distribution with the representative income. The second column lists the average income. The third column lists the proportion of the target population assigned to treatment by each policy.

Figure 5 compares the quadrant treatment rules maximizing the sample analogs of the Gini SWF and of the average income. Figure 6 compares the linear treatment rules maximizing these criteria. The size of black dots shows the number of individuals with different covariate values. Many individuals would be assigned to treatment by both rules, but there are also notable differences. Both quadrant and linear treatment rules maximizing the Gini SWF are focused on assigning low income, low education individuals to treatment. Treatment rules maximizing the average income instead favor low income, but more educated individuals. In both cases, treatment rules maximizing the Gini SWF would assign the program treatment to a lower proportion of the target population.

6 Conclusion

This paper develops an empirical welfare maximization method for treatment choice when the policy maker's objective is to maximize a rank-dependent social welfare function. We showed that we can learn the optimal feasible assignment policy under a rank-dependent social welfare at the same uniform convergence rate as we can for the optimal utilitarian assignment rule. The key restriction underlying these rate results is the complexity restriction (Assumption 2.2) imposed on the set of feasible policies. This complexity restriction should be seen as an attractive feature rather than a disadvantage of the method, as it offers a flexible and convenient way to incorporate the exogenous constraints that the policy maker faces in realistic settings of policy design. Our analytical results cover a general class of equality-minded rank-dependent social welfare functions. Computing the EWM rule is more challenging than in the utilitarian welfare case. Efficient computation for the equality-minded EWM rule remains an open question.

A Appendix: Lemmas and Proofs

Proof of Theorem 2.1. Denote an upper level set of $\delta(x)$ at level $u \in [0,1]$ by $G(u) \equiv \{x \in \mathcal{X} : \delta(x) \geq u\}$. By noting

$$\delta(x) = \int_0^1 1\{x \in G(u)\} du,$$

we can rewrite $F_\delta(y)$ defined in (2.4) as

$$\begin{aligned} F_\delta(y) &= \int_{\mathcal{X}} \left[\int_0^1 1\{x \notin G(u)\} du \cdot F_{Y_0|X=x}(y) + \int_0^1 1\{x \in G(u)\} du \cdot F_{Y_1|X=x}(y) \right] dP_X(x) \\ &= \int_0^1 \left[\int_{\mathcal{X}} \big(1\{x \notin G(u)\} \cdot F_{Y_0|X=x}(y) + 1\{x \in G(u)\} \cdot F_{Y_1|X=x}(y) \big) dP_X(x) \right] du \\ &= \int_0^1 F_{G(u)}(y) du, \end{aligned}$$

where $F_{G(u)}(y)$ is the distribution of outcomes induced by treatment rule $\delta_{G(u)} \equiv 1\{x \in G(u)\}$. By convexity of $\Lambda(\cdot)$, we obtain

$$\Lambda(F_\delta(y)) \leq \int_0^1 \Lambda(F_{G(u)}(y)) du,$$

and this leads to

$$W(F_\delta) \leq \int_0^1 W(F_{G(u)}) du \equiv \bar{W}. \tag{A.1}$$

Suppose that $\bar{W} - W(F_{G(u)}) > 0$ for all $u \in [0,1]$. Then the integral of this function over the set $u \in [0,1]$ of positive measure must also be strictly positive,

$$0 < \int_0^1 \big(\bar{W} - W(F_{G(u)}) \big) du = \bar{W} - \bar{W},$$

which is a contradiction. Therefore, there exists $u^* \in [0,1]$ for which $W(F_{G(u^*)}) \geq \bar{W}$, hence $W(F_{G(u^*)}) \geq W(F_\delta)$. If all upper level sets $G(u)$ of δ belong to $\mathcal{G}$, then also $G(u^*) \in \mathcal{G}$.

□

Proof of Theorem 3.1. Take an arbitrary set $G^* \in \mathcal{G}$, then

$$\begin{aligned} W(G^*) - W(\widehat{G}) &= W(G^*) - \widehat{W}(\widehat{G}) + \widehat{W}(\widehat{G}) - W(\widehat{G}) \\ &\leq W(G^*) - \widehat{W}(G^*) + \widehat{W}(\widehat{G}) - W(\widehat{G}) \\ &\leq 2 \sup_{G \in \mathcal{G}} \left| \widehat{W}(G) - W(G) \right|, \end{aligned}$$

where the second line follows since $\widehat{W}(\widehat{G})$ maximizes $\widehat{W}(G)$ over $G \in \mathcal{G}$. Since $W_{\mathcal{G}}^* = \max_{G\in\mathcal{G}} W(G)$, it follows that

$$W_{\mathcal{G}}^* - W(\widehat{G}) \leq 2 \sup_{G\in\mathcal{G}} \left|\widehat{W}(G) - W(G)\right|. \tag{A.2}$$

Since $\Lambda(\cdot)$ is convex and non-increasing,

$$\begin{aligned}\sup_{G\in\mathcal{G}} \left|\widehat{W}(G) - W(G)\right| &\leq \int_0^M \sup_{G\in\mathcal{G}} \left|\Lambda(\widehat{F}_G(y)) - \Lambda(F_G(y))\right| dy \\ &\leq |\Lambda'(0)| \int_0^M \sup_{G\in\mathcal{G}} \left|\widehat{F}_G(y) - F_G(y)\right| dy. \end{aligned}\tag{A.3}$$

Combining (A.2) and (A.3), the average welfare loss of $\widehat{G}$ can be bounded by

$$E_{P^n}\left[W_{\mathcal{G}}^* - W(\widehat{G})\right] \leq 2|\Lambda'(0)| \int_0^M E_{P^n}\left[\sup_{G\in\mathcal{G}} \left|\widehat{F}_G(y) - F_G(y)\right|\right] dy. \tag{A.4}$$

The same arguments apply to $\widehat{G}^R$, whose average loss can by bounded by

$$E_{P^n}\left[W_{\mathcal{G}}^* - W(\widehat{G}^R)\right] \leq 2|\Lambda'(0)| \int_0^M E_{P^n}\left[\sup_{G\in\mathcal{G}} \left|\widehat{F}_G^R(y) - F_G(y)\right|\right] dy. \tag{A.5}$$

It follows from the definition (3.4) that for any $y \in [0, M]$,

$$\left|\widehat{F}_G(y) - \widehat{F}_G^R(y)\right| = \begin{cases} \frac{\widehat{F}_G(y)}{\widehat{F}_G(M)} \left|\widehat{F}_G(M) - 1\right| \leq \left|\widehat{F}_G(M) - F_G(M)\right| & \text{if } \widehat{F}_G(M) > 0, \\ 0 & \text{if } \widehat{F}_G(M) = 0, \end{cases}$$

since $\widehat{F}_G(y) \leq \widehat{F}_G(M)$ and $F_G(M) = 1$. Then it follows that

$$\begin{aligned}\left|\widehat{F}_G^R(y) - F_G(y)\right| &\leq \left|\widehat{F}_G(y) - F_G(y)\right| + \left|\widehat{F}_G(y) - \widehat{F}_G^R(y)\right| \\ &\leq \left|\widehat{F}_G(y) - F_G(y)\right| + \left|\widehat{F}_G(M) - F_G(M)\right|. \end{aligned}\tag{A.6}$$

By combining (A.5), and (A.6), the average welfare loss of the rescaled EWM treatment rule $\widehat{G}^R$ can be bounded by

$$\begin{aligned} E_{P^n}\left[W_{\mathcal{G}}^* - W(\widehat{G}^R)\right] \leq\ & 2|\Lambda'(0)| \int_0^M E_{P^n}\left[\sup_{G\in\mathcal{G}} \left|\widehat{F}_G(y) - F_G(y)\right|\right] dy \\ & + 2|\Lambda'(0)| \cdot M \cdot E_{P^n}\left[\sup_{G\in\mathcal{G}} \left|\widehat{F}_G(M) - F_G(M)\right|\right]. \end{aligned}\tag{A.7}$$

Denote the observations by $Z_i = (Y_i, D_i, X_i)$, $i = 1, \ldots, n$ and the sample average with respect to them by $E_n(\cdot)$. To apply a maximal inequality for the centered empirical process to $E_{P^n}\left[\sup_{G\in\mathcal{G}}\left|\widehat{F}_G(y) - F_G(y)\right|\right]$, note that with y fixed, it can be written as

$$E_{P^n}\left[\sup_{f\in\mathcal{F}}|E_n(f) - E_P(f)|\right],$$

where

$$\mathcal{F} \equiv \left\{ f(Z_i) = \frac{D_i \cdot 1\{Y_i \leq y\}}{e(X_i)} \cdot 1\{X_i \in G\} + \frac{(1-D_i)\cdot 1\{Y_i \leq y\}}{1-e(X_i)} \cdot 1\{X_i \notin G\},\ G \in \mathcal{G} \right\}.$$

$\mathcal{F}$ is a VC-subgraph class of functions with envelope κ^{-1} and VC-dimension less than or equal to v by Lemma A.1 of Kitagawa and Tetenov (2015). Accordingly, by applying the maximal inequality for the centered empirical processes over the VC-subgraph class of functions in the form given in Lemma A.4 of Kitagawa and Tetenov (2015), we obtain

$$E_{P^n}\left[\sup_{G\in\mathcal{G}}\left|\widehat{F}_G(y) - F_G(y)\right|\right] \leq C_1\kappa^{-1}\sqrt{\frac{v}{n}}, \tag{A.8}$$

where C_1 is a universal constant.

Plugging inequality (A.8) into (A.4) yields

$$E_{P^n}\left[W^*_{\mathcal{G}} - W(\widehat{G})\right] \leq 2C_1|\Lambda'(0)|\frac{M}{\kappa}\sqrt{\frac{v}{n}}.$$

Since this upper bound does not depend on $P \in \mathcal{P}$, this completes the proof of (3.6). Similarly, plugging (A.8) into (A.7) completes the proof of (3.7)

$$E_{P^n}\left[W^*_{\mathcal{G}} - W(\widehat{G}^R)\right] \leq 4C_1|\Lambda'(0)|\frac{M}{\kappa}\sqrt{\frac{v}{n}}.$$

□

Proof of Theorem 3.2. We consider a suitable subclass $\mathcal{P}^* \subset \mathcal{P}$, for which the worst case welfare loss can be bounded from below by a distribution-free term that converges at rate $n^{-1/2}$. To simplify the proof, we restrict the range of outcomes to $Y \in [0,1]$, when $Y \in [0, M]$, the lower bound is proportional to M.

The construction of $\mathcal{P}^*$ proceeds as follows. By the definition of VC-dimension, there exists a set of v points in $\mathcal{X}$, denoted $x_1, \ldots, x_v \in \mathcal{X}$ that are shattered by $\mathcal{G}$. We constrain the marginal distribution of X to being supported only on $(x_1, \ldots, x_v)$. We put mass $p \equiv \frac{t^*}{v-1}$

at x_i, $i < v$, and mass $1 - t^*$ at x_v. Thus-constructed marginal distribution of X is common in $\mathcal{P}^*$. Let the distribution of treatment indicator D be independent of (Y_0, Y_1, X), and let D follow the Bernoulli distribution with $\Pr(D = 1) = 1/2$. Let $\mathbf{b} = (b_1, \ldots, b_{v-1}) \in \{0, 1\}^{v-1}$ be a bit vector used to index a member of $\mathcal{P}^*$, i.e., $\mathcal{P}^* = \{P_{\mathbf{b}} : \mathbf{b} \in \{0, 1\}^{v-1}\}$ consists of a finite number of DGPs. For each $j = 1, \ldots, (v-1)$, and depending on $\mathbf{b}$, construct the following conditional distributions of potential outcomes given $X = x_j$; if $b_j = 1$,

$$Y_0|(X = x_j) \sim Ber\left(\frac{1-\gamma}{2}\right), \quad Y_1|(X = x_j) \sim Ber\left(\frac{1+\gamma}{2}\right), \tag{A.9}$$

and, if $b_j = 0$,

$$Y_0|(X = x_j) \sim Ber\left(\frac{1+\gamma}{2}\right), \quad Y_1|(X = x_j) \sim Ber\left(\frac{1-\gamma}{2}\right), \tag{A.10}$$

where $Ber(m)$ denotes the Bernoulli distribution with mean m and $\gamma \in (0, 1)$ is chosen properly in a later step of the proof. For $j = v$, we set the distribution of potential outcomes to be degenerate at the maximum value of Y, $P(Y_0 = Y_1 = 1|X = x_v) = 1$. Clearly, $P_{\mathbf{b}} \in \mathcal{P}$ for every $\mathbf{b} \in \{0, 1\}^{v-1}$. We accordingly define $\mathcal{P}^* = \left\{P_{\mathbf{b}} : \mathbf{b} \in \{0, 1\}^{v-1}\right\} \subset \mathcal{P}$.

Note that when the outcome distribution is Bernoulli with mean μ, the equality-minded welfare function equals $W(F) = \Lambda(1 - \mu)$, which is a non-decreasing function of μ. Hence, given knowledge of $P_{\mathbf{b}}$, an optimal treatment assignment rule for the equality-minded welfare coincides with that for the utilitarian welfare case,

$$G^*_{\mathbf{b}} = \{x_j : j < v, b_j = 1\},$$

which is feasible, since $G^*_{\mathbf{b}} \in \mathcal{G}$ by the construction of the support points of X. The maximized social welfare is accordingly obtained as

$$W(G^*_{\mathbf{b}}) = \Lambda(1 - \mu^*),$$
$$\mu^* \equiv p(v-1)\left(\frac{1+\gamma}{2}\right) + (1 - t^*) = t^*\left(\frac{1+\gamma}{2}\right) + (1 - t^*),$$

which does not depend on $\mathbf{b}$.

Let $\widehat{G}$ be an arbitrary treatment choice rule as a function of observations $Z_i \equiv (Y_i, D_i, X_i)$, $i = 1, \ldots, n$, and $\hat{\mathbf{b}} \in \{0, 1\}^{(v-1)}$ be a binary vector whose j-th element is $\hat{b}_j = 1\{x_j \in \widehat{G}\}$. Let $\mu_{\widehat{G}}$ be the mean of outcome Y when treatment assignment rule $\widehat{G}$ is implemented for a given realization of the sample. For $P \in \mathcal{P}^*$ outcomes are binary, hence

$$\mu_{\widehat{G}} \equiv \int_{\widehat{G}} \Pr(Y_1 = 1|X = x)dP_X(x) + \int_{\widehat{G}^c} \Pr(Y_0 = 1|X = x)dP_X(x).$$

Consider $\pi(\mathbf{b})$, a prior distribution for $\mathbf{b}$, such that $b_1, \ldots, b_{v-1}$ are iid and $b_1 \sim Ber(1/2)$. The welfare loss satisfies the following inequalities:

$$\begin{aligned}
\sup_{P \in \mathcal{P}} E_{P^n}\left[W^*_{\mathcal{G}} - W(\widehat{G})\right] &\geq \sup_{P_{\mathbf{b}} \in \mathcal{P}^*} E_{P^n_{\mathbf{b}}}\left[W(G^*_{\mathbf{b}}) - W(\widehat{G})\right] \\
&\geq \int_{\mathbf{b}} E_{P^n_{\mathbf{b}}}\left[W(G^*_{\mathbf{b}}) - W(\widehat{G})\right] d\pi(\mathbf{b}) \\
&= \int_{\mathbf{b}} E_{P^n_{\mathbf{b}}}\left[\Lambda(1-\mu^*) - \Lambda(1-\mu_{\widehat{G}})\right] d\pi(\mathbf{b}) \\
&\geq \int_{\mathbf{b}} E_{P^n_{\mathbf{b}}}\left[|\Lambda'(1-\mu_{\widehat{G}})|(\mu^* - \mu_{\widehat{G}})\right] d\pi(\mathbf{b}) \\
&\geq |\Lambda'(t^*)| \int_{\mathbf{b}} E_{P^n_{\mathbf{b}}}\left[\mu^* - \mu_{\widehat{G}}\right] d\pi(\mathbf{b}),
\end{aligned} \tag{A.11}$$

where the fourth line follows since $\Lambda(\cdot)$ is convex and non-increasing. The fifth line follows from the observation that for all $P \in \mathcal{P}^*$, $\mu_G \geq 1 - t^*$ for any treatment rule G, therefore $|\Lambda'(1-\mu_{\widehat{G}})| \geq |\Lambda'(t^*)|$.

Consider now bounding $\int_{\mathbf{b}} E_{P^n_{\mathbf{b}}}\left[\mu^* - \mu_{\widehat{G}}\right] d\pi(\mathbf{b})$ from below. Building on the lower bound calculation for the classification risk of the empirical risk minimizing classifier in Lugosi (2002), the proof of Theorem 2.2 in Kitagawa and Tetenov (2015) considers bounding a similar quantity, though the current construction of $\mathcal{P}^*$ is different from theirs. Therefore, in what follows, we reproduce the proof of Theorem 2.2 in Kitagawa and Tetenov (2015) with some necessary modifications.

Consider

$$\begin{aligned}
\int_{\mathbf{b}} E_{P^n_{\mathbf{b}}}\left[\mu^* - \mu_{\widehat{G}}\right] d\pi(\mathbf{b}) &\geq \gamma \int_{\mathbf{b}} E_{P^n_{\mathbf{b}}}\left[P_X(G^*_{\mathbf{b}} \triangle \widehat{G})\right] d\pi(\mathbf{b}) \\
&= \gamma \int_{\mathbf{b}} \int_{Z_1,\ldots,Z_n} P_X\left(\{b(X) \neq \hat{b}(X)\}\right) dP^n\left(Z_1, \ldots, Z_n|\mathbf{b}\right) d\pi(\mathbf{b}) \\
&\geq \inf_{\widehat{G}} \gamma \int_{\mathbf{b}} \int_{Z_1,\ldots,Z_n} P_X\left(\{b(X) \neq \hat{b}(X)\}\right) dP^n\left(Z_1, \ldots, Z_n|\mathbf{b}\right) d\pi(\mathbf{b})
\end{aligned}$$

where each $b(X)$ and $\hat{b}(X)$ is an element of $\mathbf{b}$ and $\hat{\mathbf{b}}$ such that $b(x_j) = b_j$, $\hat{b}(x_j) = \hat{b}_j$, and $b(x_v) = \hat{b}(x_v) = 0$. Note that the last expression can be seen as the minimized Bayes risk with the loss function corresponding to the classification error for predicting binary unknown random variable $b(X)$. Hence, the minimizer of the Bayes risk is attained by the Bayes classifier,

$$\widehat{G}^* = \left\{x_j : \pi(b_j = 1|Z_1, \ldots, Z_n) \geq \frac{1}{2}, j < v\right\},$$

where $\pi(b_j|Z_1, \ldots, Z_n)$ is the posterior of b_j. The minimized Bayes risk is given by

$$\begin{aligned} & \gamma \int_{Z_1,\ldots,Z_n} E_X \left[\min\left\{\pi\left(b(X)=1|Z_1,\ldots,Z_n\right), 1-\pi\left(b(X)=1|Z_1,\ldots,Z_n\right)\right\}\right] d\tilde{P}^n \\ = \; & \gamma \int_{Z_1,\ldots,Z_n} \sum_{j=1}^{v-1} p\left[\min\left\{\pi\left(b_j=1|Z_1,\ldots,Z_n\right), 1-\pi(b_j=1|Z_1,\ldots,Z_n)\right\}\right] d\tilde{P}^n, \qquad \text{(A.12)} \end{aligned}$$

where $\tilde{P}^n$ is the marginal likelihood of $\{(Y_i, D_i, X_i) : i = 1, \ldots, n\}$ corresponding to prior $\pi(\mathbf{b})$. For each $j = 1, \ldots, (v-1)$ let

$$\begin{aligned} k_j^+ &= \#\left\{i : X_i = x_j, Y_i D_i = 1 \text{ or } (1-Y_i)(1-D_i) = 1\right\}, \\ k_j^- &= \#\left\{i : X_i = x_j, (1-Y_i) D_i = 1 \text{ or } Y_i(1-D_i) = 1\right\}. \end{aligned}$$

The posterior for $b_j = 1$ can be written as

$$\pi(b_j = 1|Z_1, \ldots, Z_n) = \begin{cases} \frac{1}{2} & \text{if } \#\{i : X_i = x_j\} = 0, \\ \frac{\left(\frac{1+\gamma}{2}\right)^{k_j^+}\left(\frac{1-\gamma}{2}\right)^{k_j^-}}{\left(\frac{1+\gamma}{2}\right)^{k_j^+}\left(\frac{1-\gamma}{2}\right)^{k_j^-} + \left(\frac{1+\gamma}{2}\right)^{k_j^-}\left(\frac{1-\gamma}{2}\right)^{k_j^+}} & \text{otherwise.} \end{cases}$$

Hence,

$$\begin{aligned} & \min\left\{\pi\left(b_j=1|Z_1,\ldots,Z_n\right), 1-\pi(b_j=1|Z_1,\ldots,Z_n)\right\} \\ = \; & \frac{\min\left\{\left(\frac{1+\gamma}{2}\right)^{k_j^+}\left(\frac{1-\gamma}{2}\right)^{k_j^-}, \left(\frac{1+\gamma}{2}\right)^{k_j^-}\left(\frac{1-\gamma}{2}\right)^{k_j^+}\right\}}{\left(\frac{1+\gamma}{2}\right)^{k_j^+}\left(\frac{1-\gamma}{2}\right)^{k_j^-} + \left(\frac{1+\gamma}{2}\right)^{k_j^-}\left(\frac{1-\gamma}{2}\right)^{k_j^+}} \\ = \; & \frac{\min\left\{1, \left(\frac{1+\gamma}{1-\gamma}\right)^{k_j^+ - k_j^-}\right\}}{1 + \left(\frac{1+\gamma}{1-\gamma}\right)^{k_j^+ - k_j^-}} \\ = \; & \frac{1}{1 + a^{|k_j^+ - k_j^-|}}, \text{ where } a = \frac{1+\gamma}{1-\gamma} > 1. \qquad \text{(A.13)} \end{aligned}$$

Coarsen an observation of (Y_i, D_i) into $\tilde{Y}_i$ defined as

$$\tilde{Y}_i = \begin{cases} 1 & \text{if } \; Y_i D_i + (1-Y_i)(1-D_i) = 1, \\ -1 & \text{otherwise.} \end{cases} \qquad \text{(A.14)}$$

Since $k_j^+ - k_j^- = \sum_{i:X_i=x_j} \tilde{Y}_i$, plugging (A.13) into (A.12) yields

$$\gamma \sum_{j=1}^{v-1} p E_{\tilde{P}^n}\left[\frac{1}{1 + a^{\left|\sum_{i:X_i=x_j} \tilde{Y}_i\right|}}\right] \geq \frac{\gamma}{2} \sum_{j=1}^{v-1} p E_{\tilde{P}^n}\left[\frac{1}{a^{\left|\sum_{i:X_i=x_j} \tilde{Y}_i\right|}}\right] \geq \frac{\gamma}{2} p \sum_{i=1}^{v-1} a^{-E_{\tilde{P}^n}\left|\sum_{i:X_i=x_j} \tilde{Y}_i\right|},$$

where $E_{\tilde{P}^n}(\cdot)$ is the expectation with respect to the marginal likelihood of $\{(Y_i, D_i, X_i), i = 1, \ldots, n\}$. The second line follows by $a > 1$, and the third line follows by Jensen's inequality. Given our prior specification for $\mathbf{b}$, the marginal distribution of Y_i is $\Pr(\tilde{Y}_i = 1) = \Pr(\tilde{Y}_i = -1) = 1/2$. Hence,

$$E_{\tilde{P}^n} \left| \sum_{i:X_i=x_j} \tilde{Y}_i \right| = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} E \left| 2B(k, \frac{1}{2}) - k \right|$$

holds, where $B(k, \frac{1}{2})$ is a random variable following the binomial distribution with parameters k and $\frac{1}{2}$. By noting

$$\begin{aligned} E \left| B(k, \frac{1}{2}) - \frac{k}{2} \right| &\leq \sqrt{E \left(B(k, \frac{1}{2}) - \frac{k}{2} \right)^2} \quad (\because \text{Cauchy-Schwartz inequality}) \\ &= \sqrt{\frac{k}{4}}, \end{aligned}$$

we obtain

$$\begin{aligned} E_{\tilde{P}^n} \left| \sum_{i:X_i=x_j} \tilde{Y}_i \right| &\leq \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} \sqrt{k} \\ &= E \sqrt{B(n,p)} \\ &\leq \sqrt{np}. \quad (\because \text{ Jensen's inequality}). \end{aligned}$$

Hence, the Bayes risk (A.12) is bounded from below by

$$\begin{aligned} & \frac{\gamma}{2} p(v-1) a^{-\sqrt{np}} \\ \geq \; & \frac{\gamma}{2} p(v-1) e^{-(a-1)\sqrt{np}} \quad (\because 1 + x \leq e^x \; \forall x) \\ = \; & \frac{p\gamma}{2}(v-1) e^{-\frac{2\gamma}{1-\gamma}\sqrt{np}}, \end{aligned} \tag{A.15}$$

therefore

$$\int_{\mathbf{b}} E_{P^n_{\mathbf{b}}} \left[\mu^* - \mu_{\widehat{G}} \right] d\pi(\mathbf{b}) \geq \frac{p\gamma}{2}(v-1) e^{-\frac{2\gamma}{1-\gamma}\sqrt{np}}. \tag{A.16}$$

This lower bound of the Bayes risk has the slowest convergence rate when γ is set to be proportional to $n^{-1/2}$. Specifically, let $\gamma = \sqrt{\frac{v-1}{nt^*}}$. Then for all $n \geq 4(v-1)/t^*$, $\gamma \leq 1/2$ and since $p = \frac{t^*}{v-1}$,

$$-\frac{2\gamma}{1-\gamma}\sqrt{np} = -\frac{2}{1-\gamma}\sqrt{\frac{v-1}{nt^*}}\sqrt{\frac{nt^*}{v-1}} = -\frac{2}{1-\gamma} \geq -4.$$

Then

$$\frac{p\gamma}{2}(v-1)e^{-\frac{2\gamma}{1-\gamma}\sqrt{np}} \geq \frac{p\gamma}{2}(v-1)e^{-4} = \frac{t^*}{2}\sqrt{\frac{v-1}{nt^*}}e^{-4} = \frac{e^{-4}}{2}\sqrt{t^*}\sqrt{\frac{v-1}{n}}$$

Inserting this bound into A.11 and multiplying by M completes the proof. □

Proof of Theorem 4.1. For any $G \in \mathcal{G}$, it holds

$$\begin{aligned} W(F_G) - W(F_{\widehat{G}^e}) &\leq W(\widehat{F}_G) - W(\widehat{F}^e_G) - W(\widehat{F}_{\widehat{G}^e}) + W(\widehat{F}^e_{\widehat{G}^e}) \\ &\quad + W(F_G) - W(F_{\widehat{G}^e}) - W(\widehat{F}_G) + W(\widehat{F}_{\widehat{G}^e}) \qquad \text{(A.17)} \\ &\leq 2\sup_{G\in\mathcal{G}} |W(\widehat{F}_G) - W(\widehat{F}^e_G)| + 2\sup_{G\in\mathcal{G}} \left| W(\widehat{F}_G) - W(F_G) \right|, \end{aligned}$$

where the first inequality uses $W(\widehat{F}^e_{\widehat{G}^e}) - W(\widehat{F}^e_G) \geq 0$. By convexity and monotonicity of $\Lambda(\cdot)$, we have

$$|W(\widehat{F}_G) - W(\widehat{F}^e_G)| \leq |\Lambda'(0)| \int_0^M |\widehat{F}_G(y) - \widehat{F}^e_G(y)| dy. \qquad \text{(A.18)}$$

For every y, $|\widehat{F}_G(y) - \widehat{F}^e_G(y)|$ can be bounded by

$$\begin{aligned} |\widehat{F}_G(y) - \widehat{F}^e_G(y)| \leq & \frac{1}{n}\sum_{i=1}^n \left| \frac{1}{e(X_i)} - \frac{1}{\hat{e}(X_i)} \right| D_i 1\{Y_i \leq y\} 1\{X_i \in G\} \\ & + \frac{1}{n}\sum_{i=1}^n \left| \frac{1}{1-e(X_i)} - \frac{1}{1-\hat{e}(X_i)} \right| (1-D_i) 1\{Y_i \leq y\} 1\{X_i \notin G\} \\ = & \frac{1}{n}\sum_{i=1}^n \left| \frac{1}{e(X_i)} - \frac{1}{\hat{e}(X_i)} \right| D_i + \frac{1}{n}\sum_{i=1}^n \left| \frac{1}{1-e(X_i)} - \frac{1}{1-\hat{e}(X_i)} \right| (1-D_i) \end{aligned} \qquad \text{(A.19)}$$

Combining (A.18) and (A.19) gives

$$E_{P^n}\left[\sup_{G\in\mathcal{G}} |W(\widehat{F}_G) - W(\widehat{F}^e_G)|\right] \leq E_{P^n}\left[\frac{1}{n}\sum_{i=1}^n \left| 1 - \frac{e(X_i)}{\hat{e}(X_i)} \right|\right] + E_{P^n}\left[\frac{1}{n}\sum_{i=1}^n \left| 1 - \frac{1-e(X_i)}{1-\hat{e}(X_i)} \right|\right]. \qquad \text{(A.20)}$$

Hence, from (A.17) and (A.20), we obtain

$$\begin{aligned} &\sup_{P\in\mathcal{P}_e\cap\mathcal{P}} E_{P^n}\left[W^*_{\mathcal{G}} - W(F_{\widehat{G}^e})\right] \\ \leq & 2\sup_{P\in\mathcal{P}_e} E_{P^n}\left[\frac{1}{n}\sum_{i=1}^n \left| 1 - \frac{e(X_i)}{\hat{e}(X_i)} \right|\right] + \sup_{P\in\mathcal{P}_e} E_{P^n}\left[\frac{1}{n}\sum_{i=1}^n \left| 1 - \frac{1-e(X_i)}{1-\hat{e}(X_i)} \right|\right] \\ & + 2\sup_{P\in\mathcal{P}} E_{P^n}\left[\sup_{G\in\mathcal{G}} \left| W(\widehat{F}_G) - W(F_G) \right|\right] \end{aligned}$$

Assumption 4.1 and $\sup_{P\in\mathcal{P}} E_{P^n}\left[\sup_{G\in\mathcal{G}}\left|W(\widehat{F}_G)-W(F_G)\right|\right]=O(\sqrt{\frac{v}{n}})$ shown in the proof of Theorem 3.1 lead to the conclusion. □

B Equality-minded EWM with Nonparametrically Estimated Propensity Score

In this appendix, we consider the equality-minded EWM approach with unknown propensity score estimated nonparametrically by local polynomial regressions. We provide regularity conditions under which the nonparametric estimator of the propensity score satisfies Assumption 4.1 with certain ϕ_n.

We consider the leave-one-out local polynomial estimator for $e(\cdot)$, i.e., $\hat{e}(X_i)$ is constructed by fitting the local polynomials excluding the i-th observation. For any multi-index $s=(s_1,\ldots,s_{d_x})\in\mathbb{N}^{d_x}$ and any $(x_1,\ldots,x_{d_x})\in\mathbb{R}^{d_x}$, we define $|s|\equiv\sum_{i=1}^{d_x}s_i$, $s!\equiv s_1!\cdots s_{d_x}!$, $x^s\equiv x_1^{s_1}\cdots x_{d_x}^{s_{d_x}}$, and $\|x\|\equiv\left(x_1^2+\cdots+x_{d_x}^2\right)$. Let $K(\cdot):\mathbb{R}^{d_x}\to\mathbb{R}$ be a kernel function and $h>0$ be a bandwidth, whose dependence on the sample size is implicit in the notation. At each X_i, $i=1,\ldots,n$, we define the leave-one-out local polynomial coefficient estimators with degree $l\geq 0$ as

$$\hat{\theta}(X_i)=\arg\min_{\theta}\sum_{j\neq i}\left[D_j-\theta^T U\left(\frac{X_j-X_i}{h}\right)\right]^2 K\left(\frac{X_j-X_i}{h}\right),$$

where $U\left(\frac{X_j-X_i}{h}\right)$ is the vector with elements indexed by the multi-index s, i.e., $U\left(\frac{X_j-X_i}{h}\right)\equiv\left(\left(\frac{X_j-X_i}{h}\right)^s\right)_{|s|\leq l}$. With a slight abuse of notation, we define $U(0)=(1,0,\ldots,0)^T$. Let $\lambda_n(X_i)$ be the smallest eigenvalue of $B(X_i)\equiv\left(nh^{d_x}\right)^{-1}\sum_{j\neq i}U\left(\frac{X_j-X_i}{h}\right)U^T\left(\frac{X_j-X_i}{h}\right)K\left(\frac{X_i-X_j}{h}\right)$. Accordingly, we construct leave-one-out local polynomial fit for $e(X_i)$ by

$$\tilde{e}(X_i)=U^T(0)\hat{\theta}(X_i)\cdot 1\left\{\lambda_n(X_i)\geq t_n\right\} \tag{B.1}$$

where t_n is a positive sequence that slowly converges to zero, such as $t_n\propto(\log n)^{-1}$. This trimming constant regularizes the regressor matrix of the local polynomial regression and simplifies the proof of the uniform consistency of the local polynomial estimator.

To characterize $\mathcal{P}_e$ in Assumption 4.1, we impose the following restrictions, which are identical to Assumption 2.4 in Kitagawa and Tetenov (2015).

Assumption B.1. *(Smooth-e) Smoothness of the propensity score:* The propensity score $e(\cdot)$ belongs to a Hölder class of functions with degree $\beta_e \geq 1$ and constant $L_e < \infty$. [3]
(PX) Support and Density Restrictions on P_X: Let $\mathcal{X} \subset \mathbb{R}^{d_x}$ be the support of P_X. Let $Leb(\cdot)$ be the Lebesgue measure on $\mathbb{R}^{d_x}$. There exist constants $\underline{c}$ and r_0 such that

$$Leb\left(\mathcal{X} \cap B(x,r)\right) \geq \underline{c} Leb(B(x,r)) \quad \forall 0 < r \leq r_0, \forall x \in \mathcal{X}, \tag{B.2}$$

and P_X has the density function $\frac{dP_X}{dx}(\cdot)$ with respect to the Lebesgue measure of $\mathbb{R}^{d_x}$ that is bounded from above and bounded away from zero, $0 < \underline{p}_X \leq \frac{dP_X}{dx}(x) \leq \bar{p}_X < \infty$ for all $x \in \mathcal{X}$.
(Ker) Bounded Kernel with Compact Support: The kernel function $K(\cdot)$ have support $[-1,1]^{d_x}$, $\int_{\mathbb{R}^{d_x}} K(u)du = 1$, and $\sup_u K\left(u\right) \leq K_{\max} < \infty$.

Assumption B.1 (PX) is borrowed from Audibert and Tsybakov (2007), and it provides regularity conditions on the marginal distribution of X. Inequality condition (B.2) constrains the shape of the support of X, and it essentially rules out the case where $\mathcal{X}$ has "sharp" spikes, i.e., $\mathcal{X} \cap B(x,r)$ has an empty interior or $Leb\left(\mathcal{X} \cap B(x,r)\right)$ converges to zero as $r \to 0$ faster than the rate of r^2 for some x in the boundary of $\mathcal{X}$.

The next lemma collects several properties of the local polynomial estimators that are useful to prove the bound shown in (4.5). These claims are borrowed from Theorem 3.2 in Audibert and Tsybakov (2007) and Lemma B.4 in Kitagawa and Tetenov (2015).

Lemma B.1. *Let $\mathcal{P}_e$ consist of the data generating processes satisfying Assumption B.1 (Smooth-e) and (PX). Let $\tilde{e}(X_i)$ be the leave-one-out estimator for the propensity score defined in (B.1) whose kernel function satisfies B.1 (Ker).*
(i) There exist positive constants c_1, c_2, and c_3 that depend only on β_e, d_x, L_e, $\underline{c}$, r_0, $\underline{p}_X$, and $\bar{p}_X$, such that, for any $0 < h < r_0/\underline{c}$, any $c_1 h^\beta < \delta$, and any $n \geq 2$,

$$P^{n-1}\left(\left|\tilde{e}(x) - e\left(x\right)\right| > \delta\right) \leq c_2 \exp\left(-c_3 n h^{d_x} \delta^2\right),$$

[3] Let D^s denote the differential operator $D^s \equiv \frac{\partial^{s_1+\cdots+s_{d_x}}}{\partial x_1^{s_1}\cdots x_{d_x}^{s_{d_x}}}$. Let $\beta \geq 1$ be an integer. For any $x \in \mathbb{R}^{d_x}$ and any $(\beta-1)$ times continuously differentiable function $f: \mathbb{R}^{d_x} \to \mathbb{R}$, we denote the Taylor expansion polynomial of degree $(\beta-1)$ at point x by $f_x(x') \equiv \sum_{|s|\leq\beta-1} \frac{(x'-x)^s}{s!} D^s f(x)$. Let $L > 0$. The Hölder class of functions in $\mathbb{R}^{d_x}$ with degree β and constant $0 < L < \infty$ is defined as the set of function $f: \mathbb{R}^{d_x} \to \mathbb{R}$ that are $(\beta-1)$ times continuously differentiable and satisfy, for any x and $x' \in \mathbb{R}^{d_x}$, the inequality $|f_x(x') - f(x)| \leq L\left\|x - x'\right\|^\beta$.

holds for almost all x with respect to P_X, where $P^{n-1}(\cdot)$ is the distribution of $\left\{(Y_i, D_i, X_i)_{i=1}^{n-1}\right\}$.

(ii)

$$\sup_{P \in \mathcal{P}_e} \int_{\mathcal{X}} E_{P^{n-1}} \left[|\tilde{e}(x) - e(x)| \right] dP_X(x) \leq O(h^{\beta_e}) + O\left(\frac{1}{\sqrt{nh^{d_x}}} \right) \tag{B.3}$$

holds. Hence, an optimal choice of bandwidth that leads to the fastest convergence rate of the uniform upper bound is $h \propto n^{-\frac{1}{2\beta_e + d_x}}$ and the resulting uniform convergence rate is

$$\sup_{P \in \mathcal{P}_e} \int_{\mathcal{X}} E_{P^{n-1}} \left[|\tilde{e}(x) - e(x)| \right] dP_X(x) \leq O\left(n^{-\frac{1}{2 + d_x/\beta_e}} \right).$$

Making use of Lemma B.1, the next proposition shows a propensity score estimator constructed by suitably trimming $\tilde{e}(X_i)$ satisfies Assumption 4.1.

Proposition B.1. *Let $\mathcal{P}_e$ consist of data generating processes that satisfy Assumption B.1 (Smooth-e) and (PX). Let $\tilde{e}(X_i)$ be the leave-one-out local polynomial estimator with degree $l = (\beta_e - 1)$ whose kernel satisfies Assumption B.1 (Ker). Let*

$$\hat{e}(X_i) \equiv \min\left\{1 - \epsilon_n, \max\{\epsilon_n, \tilde{e}(X_i)\}\right\} \in [\epsilon_n, 1 - \epsilon_n] \tag{B.4}$$

with a sequence of trimming constants ϵ_n satisfies $\epsilon_n = O(n^{-a})$ for some $a > 0$. Then, $\hat{e}(X_i)$ satisfies Assumption 4.1 with $\phi_n = n^{\frac{1}{2 + d_x/\beta_e}}$.

Proof of Proposition B.1. Assume that n is large enough so that $\varepsilon_n \leq \kappa/2$ holds. Since $\hat{e}(X_i) = \tilde{e}(X_i)$ whenever $\tilde{e}(X_i) \in \left[\frac{\kappa}{2}, 1 - \frac{\kappa}{2}\right] \subset [\epsilon_n, 1 - \epsilon_n]$, the following bounds are valid

$$\left| 1 - \frac{e(X_i)}{\hat{e}(X_i)} \right| \leq \begin{cases} \frac{2}{\kappa} |\tilde{e}(X_i) - e(X_i)| & \text{if } \tilde{e}(X_i) \in \left[\frac{\kappa}{2}, 1 - \frac{\kappa}{2}\right] \\ \epsilon_n^{-1} & \text{if } \tilde{e}(X_i) \notin \left[\frac{\kappa}{2}, 1 - \frac{\kappa}{2}\right]. \end{cases}$$

Hence,

$$\begin{aligned} E_{P^n}\left[\frac{1}{n} \sum_{i=1}^{n} \left| 1 - \frac{e(X_i)}{\hat{e}(X_i)} \right| \right] &= E_{P^n}\left[\left| 1 - \frac{e(X_n)}{\hat{e}(X_n)} \right| \right] \\ &\leq \frac{2}{\kappa} E_{P^n} |\tilde{e}(X_n) - e(X_n)| + \epsilon_n^{-1} P^n \left(\tilde{e}(X_n) \notin \left[\frac{\kappa}{2}, 1 - \frac{\kappa}{2} \right] \right) \end{aligned} \tag{B.5}$$

By Lemma B.1 (ii),

$$\sup_{P \in \mathcal{P}_e} E_{P^n} |\tilde{e}(X_n) - e(X_n)| = \sup_{P \in \mathcal{P}_e} \int_{\mathcal{X}} E_{P^{n-1}} \left[\left| 1 - \frac{e(x)}{\hat{e}(x)} \right| \right] dP_X(x) \leq O\left(n^{-\frac{1}{2 + d_x/\beta_e}} \right)$$

Furthermore, by Lemma B.1 (i),

$$
\begin{aligned}
P^n\left(\tilde{e}\left(X_n\right) \notin \left[\frac{\kappa}{2}, 1-\frac{\kappa}{2}\right]\right) &= \int_{\mathcal{X}} P^{n-1}\left(\tilde{e}\left(x\right) \notin \left[\frac{\kappa}{2}, 1-\frac{\kappa}{2}\right]\right) dP_X\left(x\right) \\
&\leq \int_{\mathcal{X}} P^{n-1}\left(\left|\tilde{e}\left(x\right) - e(x)\right| \geq \frac{\kappa}{2}\right) dP_X\left(x\right) \\
&\leq c_2 \exp\left(-\frac{c_3\kappa^2}{4} n h^{d_x}\right)
\end{aligned}
$$

holds for all n satisfying $c_1 h^{\beta} < \kappa/2$, where the c_1, c_2, and c_3 are the constants defined in Lemma B.1 (i). Since ε_n is assumed to converge at a polynomial rate, $\epsilon_n^{-1} P^n\left(\hat{e}\left(X_n\right) \notin \left[\frac{\kappa}{2}, 1-\frac{\kappa}{2}\right]\right)$ converges faster than $O(n^{-\frac{1}{2+d_x/\beta_e}})$. Thus, from (B.5), we conclude $\sup_{P \in \mathcal{P}_e} E_{P^n}\left[\frac{1}{n}\sum_{i=1}^{n}\left|1-\frac{e(X_i)}{\hat{e}(X_i)}\right|\right] \leq O\left(n^{-\frac{1}{2+d_x/\beta_e}}\right)$. The other bound of Assumption 4.1 can be shown similarly. □

Combining Proposition B.1 with Theorem 4.1 proves equation (4.5) in the main text.

References

Aaberge, R., T. Havnes, and M. Mogstad (2014): "A Theory for Ranking Distribution Functions," *unpublished manuscript*.

Abadie, A., J. Angrist, and G. Imbens (2002): "Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings," *Econometrica*, 70, 91–117.

Atkinson, A. B. (1970): "On the measurement of inequality," *Journal of Economic Theory*, 2, 244 – 263.

Audibert, J.-Y. and A. B. Tsybakov (2007): "Fast Learning Rates for Plug-in Classifiers," *The Annals of Statistics*, 35, 608–633.

Bhattacharya, D. and P. Dupas (2012): "Inferring Welfare Maximizing Treatment Assignment under Budget Constraints," *Journal of Econometrics*, 167, 168 – 196.

Bitler, M. P., J. B. Gelbach, and H. W. Hoynes (2006): "What Mean Impacts Miss: Distributional Effects of Welfare Reform Experiments," *American Economic Review*, 96, 988–1012.

Blackorby, C. and D. Donaldson (1978): "Measures of Relative Equality and Their Meaning in Terms of Social Welfare," *Journal of Economic Theory*, 18, 59–80.

Bloom, H. S., L. L. Orr, S. H. Bell, G. Cave, F. Doolittle, W. Lin, and J. M. Bos (1997): "The Benefits and Costs of JTPA Title II-A Programs: Key Findings from the National Job Training Partnership Act Study," *Journal of Human Resources*, 32, 549–576.

Chamberlain, G. (2011): "Bayesian Aspects of Treatment Choice," in *The Oxford Handbook of Bayesian Econometrics*, ed. by J. Geweke, G. Koop, and H. van Dijk, Oxford University Press, 11–39.

Cowell, F. (1995): *Measuring Inequality*, Hemel Hempstead: Harvester Wheatsheaf, 2 ed.

——— (2000): "Measurement of Inequality," in *Handbook of Income Distribution*, ed. by A. Atkinson and F. Bourguignon, Elsevier, vol. 1.

Dehejia (2005): "Program Evaluation as a Decision Problem," *Journal of Econometrics*, 125, 141–173.

Fan, Y. and S. S. Park (2010): "Sharp Bounds on the Distribution of Treatment Effects and Their Statistical Inference," *Econometric Theory*, 26, 931–951.

Firpo, S. and C. Pinto (2015): "Identification and Estimation of Distributional Impacts of Interventions Using Changes in Inequality Measures," *Journal of Applied Econometrics*, forthcoming.

Firpo, S. and G. Ridder (2008): "Bounds on functionals of the distribution of treatment effects," *unpublished manuscript.*

Heckman, J. J., H. Ichimura, and P. E. Todd (1997a): "Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme," *The Review of Economic Studies*, 64, 605–654.

Heckman, J. J., J. Smith, and N. Clements (1997b): "Making The Most Out Of Programme Evaluations and Social Experiments: Accounting For Heterogeneity in Programme Impacts," *The Review of Economic Studies*, 64, 487–535.

Hirano, K. and J. R. Porter (2009): "Asymptotics for Statistical Treatment Rules," *Econometrica*, 77, 1683–1701.

Kasy, M. (2014): "Using Data to Inform Policy," *unpublished manuscript.*

Kitagawa, T. and A. Tetenov (2015): "Who should be Treated? Empirical Welfare Maximization Methods for Treatment Choice," *unpublished manuscript.*

Lambert, P. J. (2001): *The Distribution and Redistribution of Income*, Manchester University Press, 3 ed.

Lugosi, G. (2002): "Pattern Classification and Learning Theory," in *Principles of Nonparametric Learning*, ed. by L. Györfi, Vienna: Springer, 1–56.

Machina, M. (1982): ""Expected Utility" Analysis without the Independence Axiom," *Econometrica*, 50, 277–323.

Manski, C. F. (2004): "Statistical Treatment Rules for Heterogeneous Populations," *Econometrica*, 72, 1221–1246.

Mehran, F. (1976): "Linear Measures of Income Inequality," *Econometrica*, 44, 805–09.

Quiggin, J. (1982): "A Theory of Anticipated Utility," *Journal of Economic Behavior and Organization*, 3, 323–343.

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

——— (2012): "Minimax regret treatment choice with covariates or with limited validity of experiments," *Journal of Econometrics*, 166, 138 – 156.

Tetenov, A. (2012): "Statistical treatment choice based on asymmetric minimax regret criteria," *Journal of Econometrics*, 166, 157–165.

Vapnik, V. N. (1998): *Statistical Learning Theory*, John Wiley & Sons.

Weymark, J. A. (1981): "Generalized gini inequality indices," *Mathematical Social Sciences*, 1, 409 – 430.

Yaari, M. E. (1988): "A controversial proposal concerning inequality measurement," *Journal of Economic Theory*, 44, 381 – 397.