cemmap

centre for microdata methods and practice

Equality-minded treatment choice

Toru Kitagawa
Aleksey Tetenov

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP71/18

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

An ESRC Research Centre

Equality-Minded Treatment Choice*

Toru Kitagawa[†] and Aleksey Tetenov[‡]

December 8, 2018

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 policy maximizing the sum of individual incomes may not be desirable 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 analog of the rank-dependent welfare over a properly constrained set of policies. 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 when the propensity score is known. 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, John Weymark, Max Tabord-Meehan, and Jeff Rowley for useful comments. We have also benefitted from presenting this work at Academia Sinica, Bristol ESG, RES, Cowles Conference on Econometrics, IAAE, SCW, AMES, Cemmap Advances in Econometrics, CEU, EEA-ESEM, Oxford, and Northwestern Conference of Former Econometrics Students. Financial support from the ESRC through the ESRC Centre for Microdata Methods and Practice (CeMMAP) (grant number RES-589-28-0001) and the European Research Council (Starting grant No. 715940) is gratefully acknowledged. Toru Kitagawa thanks the Visiting Researcher Fellowship from Collegio Carlo Alberto.

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

[‡]University of Bristol, Email: a.tetenov@bristol.ac.uk

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 *additive social welfare* criterion (sometimes called "utilitarian") defined as the mean of the outcomes in the population, even though welfare economics offers a variety of alternative criteria. The additive social welfare criterion offers analytical and computational convenience because 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

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

where Y_i is individual i's outcome, $\text{Rank}(Y_i)$ is the outcome rank of i from the bottom of the outcome distribution, and $\omega(\cdot)$ is a non-negative weight assigned to each rank. The additive SWF is a special case of (1.1) with constant $\omega(\cdot)$. The class of *generalized Gini* SWFs proposed by Mehran (1976) and Weymark (1981) consists of SWFs of the form (1.1) with *non-increasing* $\omega(\cdot)$. It closely relates to *income inequality indices*, including the widely-used Gini index. Blackorby and Donaldson (1978) show that, given a specification of $\omega(\cdot)$, the rank-dependent SWF can be written as a product of the average outcome and one minus the generalized relative index of inequality, e.g., Gini. This implies that these SWFs generate a ranking of outcome distributions that is increasing in the average outcome and decreasing in the chosen index of inequality. 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.[1] Under an additive SWF $\left(\int u(Y_i)di\right)$ (averaging either outcomes Y_i or, more generally, their non-linear transformations $u(Y_i)$), it is optimal to assign for each subgroup the treatment with the highest conditional mean $E(u(Y_d)|X)$. In contrast, a rank-dependent SWF is non-decomposable across subgroups, as the ranking of treatment assignment rules for a given covariate value may change depending on the treatment assignment of other subgroups (see Section B in the online supplement).

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.

[1] In a slight abuse of notation typical in the treatment effects literature, a single subscript on Y will be used to denote either potential outcomes of different treatments (Y_0, Y_1, or Y_d) or to sample realizations Y_i of the random variable $Y \equiv (1 - D)Y_0 + DY_1$, which denotes the outcome of treatment D_i assigned to individual i.

Our aim is to estimate from the sample data a treatment assignment rule $\widehat{G}$ belonging to a constrained (but generally large) *set of feasible policies* $\mathcal{G}$, 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. Kitagawa and Tetenov (2018a) argue that this is not restrictive for many 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 of $W(G)$, the SWF evaluated at the population distribution of Y that would realize if treatment assignment rule G is implemented. The general idea of estimating a policy by maximizing an empirical welfare criterion is in line with the method developed by Kitagawa and Tetenov (2018a) for the additive welfare case, but construction of the sample analog of $W(G)$ and derivation of its properties are substantially different and more challenging. Following the terminology suggested there, we refer to the method proposed in the current paper for rank-dependent SWFs as *equality-minded Empirical Welfare Maximization (EWM)*.

We evaluate the statistical performance of $\widehat{G}$ in terms of *regret* $E_{P^n}\left[\sup_{G\in\mathcal{G}} W(G) - W(\widehat{G})\right]$, which is the average welfare loss relative to the maximum welfare achievable in $\mathcal{G}$ with respect to the sampling distribution P^n of a size n sample. We derive a non-asymptotic and distribution-free upper bound on regret in terms of the sample size n and a measure of complexity of $\mathcal{G}$, and show that it converges to zero at $n^{-1/2}$ rate. We also show that this rate is minimax optimal over a minimally constrained class of population distributions, ensuring that no other data-driven treatment rule can lead to a faster welfare loss convergence rate uniformly over the class of distributions.

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 and show the convergence rate properties of the EWM rule for rank-dependent welfare. Section 4 provides extensions of the model that incorporate cost or capacity constraints and that allow the sampled population to be only a subset of the full population on which social welfare is defined. In Section 5, we apply our method to the experimental data from the National JTPA study. Main proofs are collected in the appendix. An online supplement contains additional proofs, examples, and extensions.

1.1 Related Literature

The analysis of statistical treatment rules 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 (2016, 2018), Kitagawa and Tetenov (2018a), Mbakop and Tabord-Meehan (2018), and Athey and Wager (2018), among others. All the existing works on treatment choice except for Kasy (2016) posit an additive welfare criterion as the objective function of the policy maker. Motivated by policy concerns about economic inequality, the current paper instead analyzes the treatment choice problem for a class of rank-dependent social welfare functions that embody inequality aversion.

The main feature distinguishing the current analysis from the EWM approach for the additive welfare case considered in Kitagawa and Tetenov (2018a) is that the rank-dependent welfare criterion is non-decomposable. Computing the empirical welfare criterion then requires that the whole distribution of outcomes that would be generated by each policy is estimated first, before a nonlinear transformation is applied to this distribution estimate. This problem has not been previously considered in econometrics nor in the machine learning and statistics literatures on empirical risk minimization problems (Vapnik (1998)), where the empirical risk criterion always takes the form of a sample average (with the exception of Wang et al. (2018), who maximize one specific quantile of the outcome distribution). Another novel technical contribution of this paper is that we allow outcomes to be unbounded (which is important for analysis of economic outcomes like earnings) with only a weak restriction on the tail of the potential outcome distribution.

Kasy (2016) analyzes treatment choice for a class of social welfare functions including rank-dependent social welfare. Our approach differs from his in several aspects. First, Kasy

(2016) considers a linear approximation of the rank-dependent welfare function around a status-quo policy in order to discuss (partial) identification of a welfare-improving local policy change. We instead focus on a globally optimal policy without invoking approximations. Second, we assume that the welfare criterion is point-identified by the sampling process, while Kasy (2016) focuses on partial identification of the welfare criterion and construction of the social planner's incomplete preference ordering over policies. Third, we study estimation of an optimal policy and examine optimality of the estimator in terms of welfare regret convergence rate, while Kasy (2016) studies asymptotically valid inference on the welfare rankings.

Aaberge et al. (2017) estimate a rank-dependent social welfare function of two policy alternatives: with and without uniform implementation of the treatment. Firpo and Pinto (2016) estimate the impact of uniform implementation of the treatment on measures of inequality, including the Gini coefficient. In contrast, the focus of the current paper is on estimating the optimal treatment rule from a large class of individualized assignment rules.

We consider social welfare functions that satisfy the axiom of *anonymity*, i.e., social welfare functions that are functionals of the distribution of outcomes after treatment assignment and that are indifferent to reshuffling of the outcomes 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 in the program evaluation literature (Heckman et al. (1997b), Firpo and Ridder (2008), Fan and Park (2010)).

Building social welfare functions satisfying the Pigou-Dalton principle of transfers (that a transfer of income from a higher ranked individual to a lower ranked individual that does not change their ranks is always desirable) 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 is the class of Atkinson-type SWFs (Atkinson (1970)), $W(F) = \int_0^\infty U(y)dF(y)$, where $F(y)$ is the cumulative distribution function (cdf) of the outcome and $U(\cdot)$ is a concave non-decreasing function. Since the Atkinson-type social welfare function is linear in F, the EWM approach of Kitagawa and Tetenov (2018a) 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). 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.[2] 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

We call a SWF *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. The equality-minded SWFs analyzed in this paper are the rank-dependent SWFs with decreasing welfare weights (also called generalized Gini SWFs), introduced by Mehran (1976) and Weymark (1981) and axiomatized by Yaari (1988). An equality-minded rank-dependent SWF admits the following representation:

$$W_\Lambda(F) \equiv \int_0^\infty \Lambda(F(y))dy, \tag{2.1}$$

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 SWF satisfies the Pigou-Dalton Principle of Transfers if and

[2] As noted in Machina (1982), the rank-dependent social welfare function generalizes the Atkinson-type social welfare exactly as 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.

only if $\Lambda(\cdot)$ is convex.

The term *rank-dependent* is due to an equivalent representation of (2.1) as a weighted sum of outcomes. Given that a convex $\Lambda(\cdot)$ is almost everywhere differentiable, we can apply integration by parts to equivalently express $W_\Lambda(F)$ as

$$W_\Lambda(F) = \int_0^1 F^{-1}(\tau)\omega(\tau)d\tau, \tag{2.2}$$

where $F^{-1}(\tau) \equiv \inf\{y : F(y) \geq \tau\}$ is the τ-th quantile of the outcome distribution and $\omega(\tau) \equiv \frac{d[1-\Lambda(\tau)]}{d\tau}$. Thus $W_\Lambda(F)$ is a weighted average of outcomes Y, where $\omega(\tau)$ specifies the *rank-specific* welfare weight assigned to individuals with outcomes at the τ-th quantile. If the SWF is equality-minded then $\Lambda(\cdot)$ is convex, hence $\omega(\cdot)$ is non-increasing and assigns larger welfare weights to individuals with lower outcomes.

Throughout the paper, we consider equality-minded SWFs satisfying the following assumption:

Assumption 2.1 (SWF)**.**
The policy-maker's SWF has representation (2.1), where $\Lambda(\cdot) : [0,1] \to [0,1]$ is a non-increasing, convex function with $\Lambda(0) = 1$, $\Lambda(1) = 0$, and its right derivative at the origin is finite: $|\Lambda'(0)| < \infty$.

An important family of social welfare functions satisfying Assumption 2.1 is the *extended Gini* family considered in Donaldson and Weymark (1980, 1983) and Aaberge et al. (2017):

$$W_k(F) \equiv \int_0^\infty (1 - F(y))^{k-1}dy = \int_0^\infty \Lambda_k(F(y))dy = \int_0^1 F^{-1}(\tau)\omega_k(\tau)d\tau, \tag{2.3}$$

where $\Lambda_k(\tau) \equiv (1-\tau)^{k-1}$ and the welfare weight function is $\omega_k(\tau) \equiv (k-1)(1-\tau)^{k-2}$. Extended Gini social welfare functions are equality-minded for $k > 2$. Setting $k = 2$ yields the additive welfare function $W_2(F) = \int_0^\infty (1 - F(y))dy = E(Y)$, which is not equality-minded.

The standard *Gini social welfare function* (Blackorby and Donaldson, 1978, Weymark, 1981) corresponds to $k = 3$ in the extended Gini family, with $\Lambda_3(\tau) = (1-\tau)^2$ and welfare weights $\omega_3(\tau) = 2(1-\tau)$. It could also be written as

$$W_{Gini}(F) = E(Y)(1 - I_{Gini}(F)), \tag{2.4}$$

where $I_{Gini}(F) = 1 - \frac{\int_0^1 F^{-1}(\tau)\cdot 2(1-\tau)d\tau}{E(Y)}$ is the widely-used Gini inequality index.

Assumption 2.1 implies that the rank-specific weight function $\omega(\cdot)$ defined in (2.2) does not asymptote at the origin, i.e., the welfare weight assigned to the lowest rank is bounded. This restriction rules out SWFs that closely approximate the Rawlsian social welfare.

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 SWF. 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

$$F_\delta(y) \equiv \int_{\mathcal{X}} \left[(1 - \delta(x))F_{Y_0|X=x}(y) + \delta(x)F_{Y_1|X=x}(y)\right] dP_X(x), \tag{2.5}$$

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

$$\delta^* \in \arg\max_{\delta} W_\Lambda(F_\delta). \tag{2.6}$$

For the additive welfare function (the mean of Y), the welfare maximization problem simplifies to

$$\delta^*_{util} \in \arg\max_{\delta} \int_{\mathcal{X}} \left[(1 - \delta(x))E(Y_0|X = x) + \delta(x)E(Y_1|X = x)\right] dP_X(x). \tag{2.7}$$

This social welfare is additive across covariates and depends on the outcome distributions only through their conditional means $E(Y_d|X)$. Then the optimal 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 one maximizing an additive welfare if there is no first-order stochastic dominance relationship between $F_{Y_0|X}$ and $F_{Y_1|X}$ for some covariate values.

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 SWFs. The following theorem mitigates this complication by substantially reducing the set of candidate treatment rules that need to be considered.

Theorem 2.1. *If $W_\Lambda(\cdot)$ 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 $W_\Lambda(F_{\delta_G}) \geq W_\Lambda(F_\delta)$.*

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 $W_\Lambda(F_{\delta_G}) \geq W_\Lambda(F_\delta)$.

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 $W_\Lambda(G)$, suppressing the cumulative distribution function in its argument,

$$\begin{aligned} W_\Lambda(G) \equiv & W_\Lambda(F_G), \\ F_G(y) \equiv & \int_{\mathcal{X}} \left[F_{Y_0|X=x}(y) 1\{x \notin G\} + F_{Y_1|X=x}(y) 1\{x \in G\} \right] dP_X(x). \end{aligned} \tag{2.8}$$

Our goal is to estimate from the sample data a treatment assignment rule that attains the maximum level of social welfare $\sup_{G \in \mathcal{G}} W_\Lambda(G)$ 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 the complexity of $\mathcal{G}$ is constrained by a finite *Vapnik-Cervonenkis (VC) dimension* (defined in the Appendix).

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 analog of $W_\Lambda(G)$ 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 (2018a) for other examples of classes $\mathcal{G}$ that satisfy Assumption 2.2. An example of $\mathcal{G}$ that does not satisfy Assumption 2.2 is the class of all monotone treatment rules $\mathcal{G} \equiv \{G = \{(x_1, x_2) \in \mathbb{R}^2 : x_2 \geq f(x_1)\}, f : \mathbb{R} \to \mathbb{R} \text{ increasing}\}$ considered by Mbakop and Tabord-Meehan (2018) in the additive welfare case.

3 EWM for Equality-Minded Welfare

We proceed to propose our method of estimating the treatment rule in finite samples and analyze its properties.

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 $X_i \in \mathcal{X} \subset \mathbb{R}^{d_x}$ refers to observable pre-treatment covariates of individual i, $Y_{0i}, Y_{1i} \in \mathbb{R}_+$ are the potential outcomes of treatments 0 and 1, and $D_i \in \{0, 1\}$ is a binary indicator of the individual's experimental treatment. The observed experimental outcome is $Y_i = (1 - D_i)Y_{0i} + D_i Y_{1i}$.

The data is a size n random sample from P of observations $Z_i = (Y_i, D_i, X_i)$. 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 the target population will be assigned to treatment 0 or to treatment 1. The following are our maintained assumptions about the class $\mathcal{P}$ of population distributions of (Y_0, Y_1, D, X):

Assumption 3.1.

(UCF) Unconfoundedness: $(Y_0, Y_1) \perp D | X$.

(TC) Tail Condition: There exists $\Upsilon < \infty$ such that for all $P \in \mathcal{P}$

$$\int_0^\infty \sqrt{P(Y_d > y)}dy \leq \Upsilon. \tag{3.1}$$

(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 be associated with their potential outcomes. *Strict Overlap* can also be violated in an observational study if only one of the treatments is assigned in the sampling process for some covariate values. We do not constrain any feature of the joint distribution of (Y_0, Y_1, X) except that the distributions of Y_0 and Y_1 satisfy the tail condition (TC). A sufficient condition for (TC) is that

$$\sup_{P \in \mathcal{P},\ d \in \{0,1\}} E[Y_d^{2+\Delta}] < \infty \tag{3.2}$$

for some $\Delta > 0$. The outcome variable and the covariates can be discrete, continuous, or their combination, and the support of X does not have to be bounded.

Throughout the main text we maintain the assumption that the propensity score $e(X) \equiv P(D = 1|X)$ is known, as is usually the case in experimental data. Section C in the Online Supplement extends the analysis to observational data for which the propensity score is unknown and needs to be estimated.

We estimate the treatment rule by maximizing a sample analog of the population SWF. The *equality-minded EWM treatment rule* $\widehat{G}$ maximizes a sample analog $\widehat{W_\Lambda}(G)$ 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 (2.8) could be estimated by

$$\widehat{F}_G(y) \equiv 1 - \frac{1}{n}\sum_{i=1}^{n}\left[\frac{D_i}{e(X_i)} \cdot 1\{X_i \in G\} + \frac{1-D_i}{1-e(X_i)} \cdot 1\{X_i \notin G\}\right] \cdot 1\{Y_i > y\}. \tag{3.3}$$

Under Assumption 3.1 (UCF), $\widehat{F}_G(y)$ is an unbiased estimator of $F_G(y)$.

The sample analog of welfare (equation (2.1)) is defined as[3]

$$\widehat{W}_{\Lambda}(G) \equiv \int_0^{\infty} \Lambda(\widehat{F}_G(y) \vee 0) dy. \tag{3.4}$$

The equality-minded EWM treatment rule is then

$$\widehat{G} \in \arg\max_{G \in \mathcal{G}} \widehat{W}_{\Lambda}(G). \tag{3.5}$$

We also consider properties of the *normalized equality-minded EWM rule*

$$\widehat{G}^R \in \arg\max_{G \in \mathcal{G}} \widehat{W}_{\Lambda}^R(G), \text{ where } \widehat{W}_{\Lambda}^R(G) \equiv \int_0^{\infty} \Lambda(\widehat{F}_G^R(y)) dy, \tag{3.6}$$

using a normalized CDF sample analog

$$\widehat{F}_G^R(y) \equiv \begin{cases} 1\{y \geq \min\limits_{1 \leq i \leq n} Y_i\} & \text{if } \widehat{F}_G(-\infty) = 1, \\ 1 - \frac{1 - \widehat{F}_G(y)}{1 - \widehat{F}_G(-\infty)} & \text{if } \widehat{F}_G(-\infty) < 1, \end{cases} \tag{3.7}$$

which always yields a proper CDF.[4] The ranking of treatment rules by the normalized criterion $\widehat{W}_{\Lambda}^R(G)$ is invariant to positive affine transformations of outcomes Y, whereas the ranking by $\widehat{W}_{\Lambda}(G)$ is not.

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*

$$\sup_{P \in \mathcal{P}} E_{P^n} \left[\sup_{G \in \mathcal{G}} W_{\Lambda}(G) - W_{\Lambda}(\widehat{G}) \right] \leq C |\Lambda'(0)| \frac{\Upsilon}{\kappa} \sqrt{\frac{v}{n}} \tag{3.8}$$

[3] The maximum ($\vee$) of $\widehat{F}_G(y)$ and 0 is taken because $\widehat{F}_G(y)$ may take values smaller than 0, for which $\Lambda(\cdot)$ is not defined. The summands in (3.3) are non-negative, so $\widehat{F}_G(y) \leq 1$ for all y. $\widehat{F}_G(y) = 1$ for all $y \geq \max\limits_{1 \leq i \leq n} Y_i$. $\widehat{F}_G$ may not be a proper CDF because $\lim\limits_{y \to -\infty} \widehat{F}_G(y) = 1 - \frac{1}{n} \sum_{i=1}^n \left[\frac{D_i}{e(X_i)} \cdot 1\{X_i \in G\} + \frac{1 - D_i}{1 - e(X_i)} \cdot 1\{X_i \notin G\} \right]$ may be either below or above zero in finite samples.

[4] Note that $1 - \widehat{F}_G^R(y) = \frac{1}{n} \sum\limits_{i=1}^n \left[\frac{\frac{D_i}{e(X_i)} \cdot 1\{X_i \in G\} + \frac{1 - D_i}{1 - e(X_i)} \cdot 1\{X_i \notin G\}}{\frac{1}{n} \sum_{i=1}^n \left(\frac{D_i}{e(X_i)} \cdot 1\{X_i \in G\} + \frac{1 - D_i}{1 - e(X_i)} \cdot 1\{X_i \notin G\} \right)} \right] \cdot 1\{Y_i > y\}$. This is similar to the idea of normalizing propensity score weights recommended for the estimation of the average treatment effect (Imbens, 2004).

for all $n > 1$, and

$$\sup_{P \in \mathcal{P}} E_{P^n} \left[\sup_{G \in \mathcal{G}} W_\Lambda(G) - W_\Lambda(\widehat{G}^R) \right] \leq |\Lambda'(0)| \frac{\Upsilon}{\kappa} \left(C_1^R \sqrt{\frac{v}{n}} + 4ne^{-C_2^R \kappa^2 n} \right) \tag{3.9}$$

for all $n > C_3^R \left(\frac{1-\kappa}{\kappa}\right)^2 v$, where $\mathcal{P}$ is the class of all distributions satisfying Assumption 3.1 and $C, C_1^R, C_2^R, C_3^R > 0$ are universal constants.

Proof. The proof of (3.8) is in Appendix A. The proof of (3.9) is in the Online Supplement.

This theorem shows that for a large class of data generating processes characterized by Assumption 3.1, the welfare of the equality-minded EWM rule is guaranteed to converge to the maximal attainable welfare no slower than at $n^{-1/2}$ rate (the second term in bound (3.9) is of a lower order). The uniform convergence rate of $n^{-1/2}$ coincides with that of the EWM rule for the additive welfare shown in Theorem 2.1 of Kitagawa and Tetenov (2018a). 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 additive welfare criterion.

The next theorem 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 $\widehat{G}$ that is a function of the sample, and for any $\tau^* \in (0, 1]$ at which $\Lambda(\cdot)$ is differentiable, it holds*

$$\sup_{P \in \mathcal{P}} E_{P^n} \left[\sup_{G \in \mathcal{G}} W_\Lambda(G) - W_\Lambda(\widehat{G}) \right] \geq \frac{e^{-4}}{2} \Upsilon \, |\Lambda'(\tau^*)| \sqrt{\tau^*} \sqrt{\frac{v-1}{n}} \text{ for all } n \geq 4 \frac{v-1}{\tau^*}, \tag{3.10}$$

where $\mathcal{P}$ is the class of all distributions satisfying Assumption 3.1.

Proof. See the Online Supplement.

Since $\Lambda(\cdot)$ is convex and $|\Lambda'(0)| > 0$, there also exists some $\tau^* > 0$ for which $|\Lambda'(\tau^*)| > 0$. Hence the bound (3.10) is always positive for some $\tau^* > 0$. 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 additive welfare case (Theorems 2.1 and 2.2 in Kitagawa and Tetenov (2018a)), and the minimax optimal rate attained by the equality-minded EWM rule is the same as the optimal rate in the additive 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), maximizing the empirical welfare leads to a policy that, if implemented, is guaranteed to reach the maximum attainable social 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-minded EWM rule. This insight is similar to the EWM rule for the additive welfare case (Remark 2.6 in Kitagawa and Tetenov (2018a)).

4 Extensions

4.1 Social Welfare is Defined on a Population Larger than the Sampled Population

One of the distinguishing features of rank-dependent social welfare is that it is not additive over subpopulations (see Section B in the Online Supplement for an illustration). If the subpopulation for which the policy intervention takes place (e.g., unemployed workers) is only a subset of the whole population on which the rank-dependent SWF is defined (e.g., the population of a country), it is important to explicitly take into account the outcome distribution for the rest of the population in estimating the optimal assignment rule.

Suppose that the social welfare function is defined on a population with distribution J that is a mixture of two subpopulations with distributions F and H:

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

Let F be the outcome distribution on the targeted subpopulation from which the experimental data are sampled and on which the estimated treatment policy is to be implemented. Let H be the outcome distribution for the rest of the population (excluded from the sampling process and unaffected by the chosen treatment assignment rule). 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 the treatment assignment policy applied to subpopulation F, e.g., there are no spillover or general equilibrium effect across F and H.

Implementing treatment assignment rule $\{X \in G\}$ on subpopulation F leads to full population welfare equal to

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

where $F_G(\cdot)$ is the cdf defined in (2.8). The empirical welfare maximization method in this case consists of maximizing a sample analog of $W_\Lambda(J_G)$,

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

where $\widehat{F}_G$ is defined in (3.3).

The uniform convergence proof of Theorem 3.1 can be easily extended to this case, the only change being the proportionality of the bound to η.

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

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

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

4.2 Cost of Treatment

In the preceding sections we did not take into account the cost of treatment even though cost differences among treatments are often an important consideration in practice. In this section we discuss how to take the cost of treatment into account in the estimation of welfare maximizing treatment assignment policies.

Let $0 \leq c(x) < \infty$, $x \in \mathcal{X}$, be the cost of treatment 1 for a subject whose observable characteristics are x. We assume that treatment 0 is cost-free and $c(\cdot)$ is known. For the additive social welfare function, we can easily incorporate treatment costs into the EWM criterion by subtracting the per-capita cost of treatment $C(G) \equiv \int_G c(x)dP_X(x)$ from $\widehat{W}(G)$. The additive social welfare criterion depends only on the average treatment cost, it is invariant to assumptions about *who pays the cost*. For rank-dependent social welfare this invariance does not hold, hence we have to be explicit about who bears the cost in the construction of the social welfare criterion. We illustrate this using two cost allocation scenarios.

In the first scenario, assume that the outcome variable is income and the cost of treatment is *self-financed* by each recipient of the treatment. Specifically, the income of individuals assigned to treatment 1 (individuals with $X \in G$) will be reduced by the full cost of their own treatment $c(X)$. The transformed potential outcomes in this scenario are $\tilde{Y}_{1i} \equiv Y_{1i} + \bar{c} - c(X_i)$ and $\tilde{Y}_{0i} \equiv Y_{0i} + \bar{c}$. We add the constant $\bar{c} \equiv \sup_{x \in \mathcal{X}} c(x) < \infty$ to all outcomes to keep them non-negative in line with Assumption 3.1 (TC). The welfare ranking of policies is unchanged when a constant is added uniformly to all outcomes.

The rank-dependent SWF of policy G with self-financed treatment cost is

$$\begin{aligned} W_\Lambda^{sf}(G) &\equiv \int_0^\infty \Lambda(F_G^{sf}(y))dy, \\ F_G^{sf}(y) &\equiv \int_{\mathcal{X}} \left[F_{\tilde{Y}_0|X=x}(y) 1\{x \notin G\} + F_{\tilde{Y}_1|X=x}(y) 1\{x \in G\} \right] dP_X(x), \end{aligned} \tag{4.3}$$

where $F_{\tilde{Y}_0|X=x}(\cdot)$ and $F_{\tilde{Y}_1|X=x}(\cdot)$ are the cdfs of the transformed potential outcomes. An empirical analog for $W^{sf}(G)$ can be obtained by replacing $\widehat{F}_G(y)$ in (3.4) by

$$\widehat{F}_G(y) \equiv 1 - \frac{1}{n} \sum_{i=1}^n \left[\frac{D_i}{e(X_i)} \cdot 1\{X_i \in G\} + \frac{1 - D_i}{1 - e(X_i)} \cdot 1\{X_i \notin G\} \right] \cdot 1\{\tilde{Y}_i > y\}, \tag{4.4}$$

where $\tilde{Y}_i \equiv Y_i + \bar{c} - D_i \cdot c(X_i)$. Since this modification does not affect the validity of Assumption 3.1, the EWM rule with self-financed treatment cost attains the uniform welfare loss upper bounds of Theorem 3.1 with $\Upsilon + \bar{c}$ in place of Υ.

In the second scenario, suppose that the treatment cost is financed by all of the population members equally via a lump-sum transfer. The average per-capita treatment cost $C(G)$ is subtracted from every individual's income regardless of their covariates and assigned treatment. Using representation (2.2), the rank-dependent SWF with equal lump-sum treatment costs can be expressed as

$$W_\Lambda^{ls}(G) \equiv \int_0^1 [F_G^{-1}(\tau) + \bar{c} - C(G)]\omega(\tau)d\tau = W(G) + \bar{c} - C(G), \tag{4.5}$$

using the fact that $\int_0^1 \omega(\tau)d\tau = \Lambda(0) - \Lambda(1) = 1$ and adding $\bar{c}$ to ensure non-negative outcomes. Per-capita treatment cost of policy G could be estimated using its sample analog $\widehat{C}(G) \equiv \frac{1}{n}\sum_{i=1}^{n} c(X_i) \cdot 1\{X_i \in G\}$ and the EWM rule is obtained by maximizing $\widehat{W}_{\Lambda}^{ls}(G) \equiv \widehat{W}_{\Lambda}(G) + \bar{c} - \widehat{C}(G)$ over $G \in \mathcal{G}$.

In this paper we do not consider the joint optimization of the treatment assignment and cost allocation. However, the comparison of $W_{\Lambda}^{sf}(G)$ and $W_{\Lambda}^{ls}(G)$ shows that the allocation of treatment costs across the population can be used as an additional vehicle of policy intervention to increase a rank-dependent SWF.

4.3 Capacity-Constrained Treatment

Another practical concern ruled out in the preceding sections is a capacity constraint limiting the proportion of population that can be assigned to treatment. Suppose that the proportion of the target population that could receive treatment 1 cannot exceed $K \in (0, 1)$. If P_X is unknown, then policies that seem to satisfy the capacity constraint based on the sample estimates of $P_X(G)$ may not actually satisfy it in the population. The analysis of the welfare loss needs to take into account what happens if the proposed policy is infeasible. For tractability, we continue to restrict attention only to non-randomized treatment rules (the result in Theorem 2.1 need not hold with a capacity constraint).

For the additive welfare case, Kitagawa and Tetenov (2018a) proposed a capacity-constrained EWM procedure assuming that if a proposed treatment rule G violates the capacity constraint ($P_X(G) > K$) then the scarce treatment is randomly rationed to a fraction $\frac{K}{P_X(G)}$ of individuals with $X \in G$ independently of (Y_0, Y_1, X). This random rationing approach can be straightforwardly extended to the EWM for the rank-dependent social welfare.

With the capacity constraint and random rationing, the cdf of outcomes generated by policy G can be written as

$$F_G^K(y) = \int_{\mathcal{X}} \left[\begin{array}{c} F_{Y_0|X=x}(y) + \\ \min\left\{1, \frac{K}{P_X(G)}\right\} \left(F_{Y_1|X=x}(y) - F_{Y_0|X=x}(y)\right) \cdot 1\{x \in G\} \end{array} \right] dP_X(x). \tag{4.6}$$

Hence the social welfare under the capacity constraint and random rationing is $W_{\Lambda}^K(G) \equiv$

$\int_0^\infty \Lambda(F_G^K(y))dy$. Its sample analog can be constructed by replacing $\widehat{F}_G(y)$ in (3.4) with

$$\widehat{F}_G^K(y) \equiv 1 - \frac{1}{n}\sum_{i=1}^{n}\left[\begin{array}{c}\frac{1-D_i}{1-e(X_i)} + \\ \min\left\{1, \frac{K}{P_{X,n}(G)}\right\}\left(\frac{D_i}{e(X_i)} - \frac{1-D_i}{1-e(X_i)}\right)1\{X_i \in G\}\end{array}\right] \cdot 1\{Y_i > y\}, \quad (4.7)$$

where $P_{X,n}(G) \equiv \frac{1}{n}\sum_{i=1}^{n} 1\{X_i \in G\}$ is a sample analog of $P_X(G)$.

Proposition 4.1 establishes a finite sample bound for the capacity-constrained equality-minded EWM rule similar to the bound in Proposition 3.1.

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

$$\sup_{P\in\mathcal{P}} E_{P^n}\left[\sup_{G\in\mathcal{G}} W_\Lambda^K(G) - W_\Lambda^K(\widehat{G}^K)\right] \leq \left(\frac{C_{K1}}{K} + C_{K2}\right)|\Lambda'(0)|\frac{\Upsilon}{\kappa}\sqrt{\frac{v}{n}}, \quad (4.8)$$

where $C_{K1}, C_{K2} > 0$ are universal constants.

Proof. See the Online Supplement. □

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 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. Our sample consists of 9,223 observations with available data on years of education and pre-program earnings from the sample of adults (22 years and older) used in the original evaluation of the program and in subsequent studies (Bloom et al., 1997, Heckman et al., 1997a, Abadie et al., 2002). The probability of being assigned to the treatment was two thirds in this sample.

For this illustration, total individual earnings in the 30-month period following program assignment serve as the measure of income. We use three social welfare functions from the

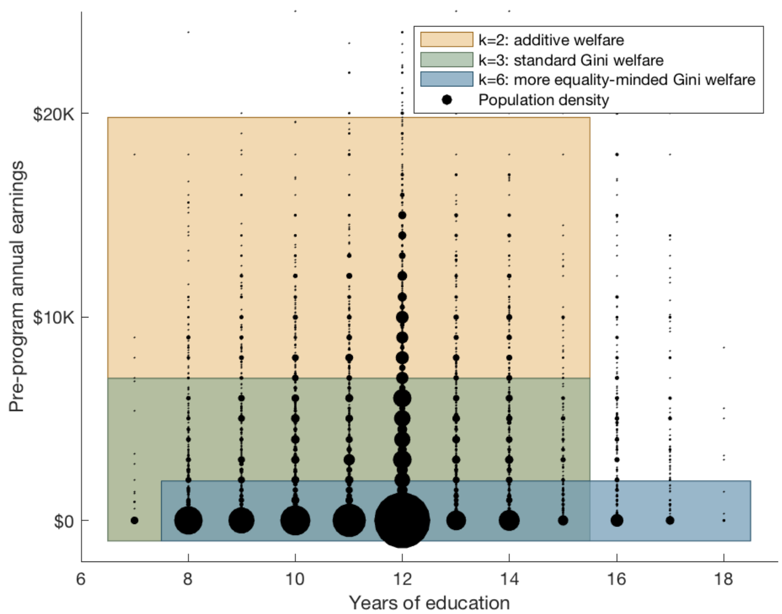

Figure 1: Treatment rules from the quadrant class that maximize welfare functions from the extended Gini family (including the additive welfare)

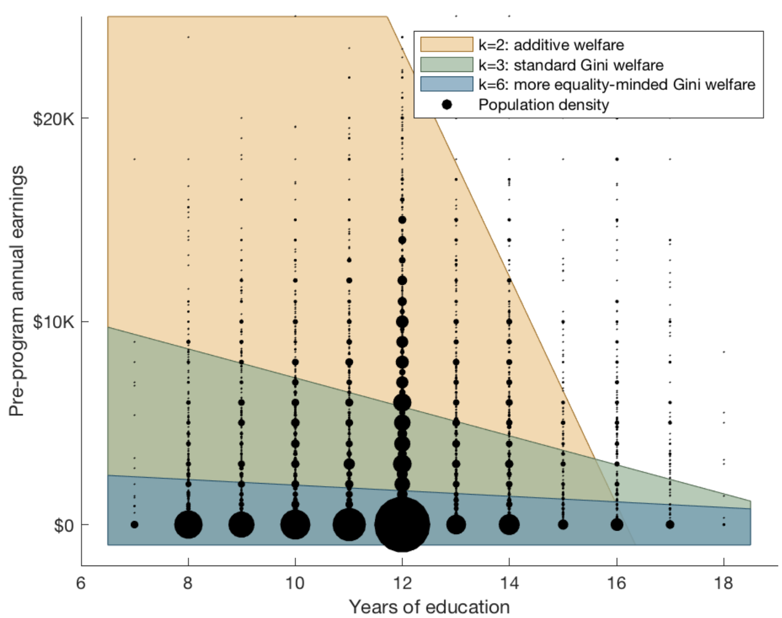

Figure 2: Treatment rules from the linear class that maximize welfare functions from the extended Gini family (including the additive welfare)

Table 1: Estimated representative income under alternative treatment rules that condition on education and pre-program earnings.

	Representative income			Proportion
Treatment rule:	Average (k=2)	Gini (k=3)	Gini (k=6)	to be treated
Treat no one	\$15,311	\$6,769	\$1,561	0
Treat everyone	\$16,487	\$7,423	\$1,786	1
Quadrant class conditioning on years of education and pre-program earnings				
Maximize average income	\$16,646	\$7,490	\$1,807	95%
Maximize standard Gini SWF (k=3)	\$16,462	\$7,522	\$1,828	80%
Maximize extended Gini SWF (k=6)	\$16,153	\$7,388	\$1,835	52%
Linear class conditioning on years of education and pre-program earnings				
Maximize average income	\$16,670	\$7,528	\$1,820	96%
Maximize standard Gini SWF (k=3)	\$16,489	\$7,537	\$1,839	78%
Maximize extended Gini SWF (k=6)	\$16,154	\$7,405	\$1,852	51%

extended Gini family (2.3) with parameters $k \in \{2, 3, 6\}$. $k = 2$ corresponds to the additive social welfare, which is not equality-minded. $k = 3$ corresponds to the standard Gini SWF with welfare weights $\omega_3(\tau) = 2(1-\tau)$ and $k = 6$ corresponds to an extended Gini SWF with welfare weights $\omega_6(\tau) = 5(1-\tau)^4$, which places even greater weight on low-ranked outcomes.

For simplicity, we consider only the distribution of earnings in the population sampled for the experiment in the social welfare function. This embodies concerns about inequality *within* the study population (JTPA-eligible economically disadvantaged adults). In practice, policy makers are more likely to be concerned with inequality in the overall population, which also includes individuals outside of the experiment's sampling frame. Then the social welfare function should be evaluated on the income distribution of the whole population of interest.

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

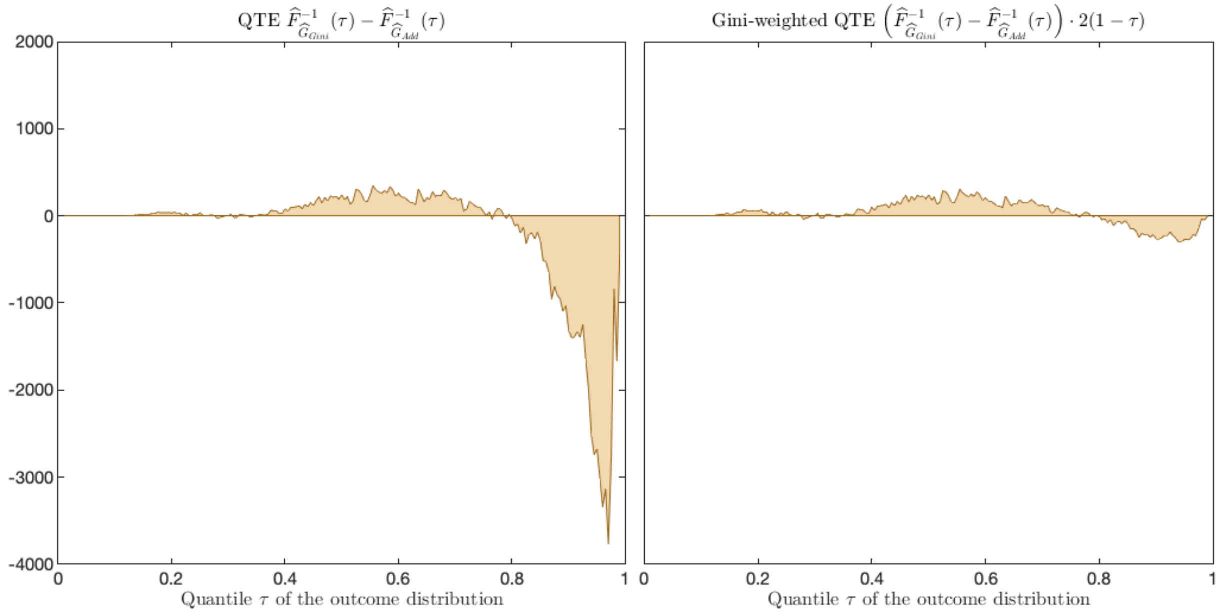

Figure 3: Quantile treatment effects of the Gini welfare maximizing rule compared to the additive welfare maximizing rule. Treatment rules from the quadrant class.

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 (either in the experiment or when the policy is implemented).

Table 1 compares empirical estimates of social welfare measures (representative income) for a few alternative treatment rules. First, we consider simple treatment rules that either assign no one or everyone 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 in Table 1 displays the estimated average income under each treatment rule. The second column shows the standard Gini social welfare, 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 third column shows the representative income under an extended Gini SWF with $k = 6$. The fourth column lists the proportion of the target population assigned to treatment by each policy.

Figure 1 compares the quadrant treatment rules maximizing the average income, the standard Gini SWF, and an extended Gini SWF ($k = 6$). Figure 2 compares the linear treatment rules maximizing the same three criteria. The size of black dots shows the number of individuals with different covariate values. Many individuals would be assigned to treatment by treatment rules maximizing any of the considered welfare functions, but there are also notable differences. Treatment rules maximizing the standard Gini SWF target a smaller proportion of the population, focusing on individuals with lowest pre-program earnings. Treatment rules maximizing the more equality-minded extended Gini SWF assign even fewer individuals to treatment. The estimated treatment rules change discontinuously with the Gini parameter k. For example, the linear treatment rule maximizing the additive SWF ($k = 2$) is also optimal for a range of other Gini parameters ($k = 2.25, 2.5, 2.75$), whereas the linear rule maximizing the extended Gini SWF ($k = 6$) is also optimal for larger parameters ($k = 7, 8, 9$).

Figure 3 explores the trade-off between treatment rules maximizing different social welfare functions. We compute the income distributions generated by $\widehat{G}_{Add}$, the quadrant treatment rule maximizing average income, and by $\widehat{G}_{Gini}$, the quadrant rule maximizing the Gini SWF. The left panel displays the difference between the income distributions generated by these treatment rules at each quantile: $\widehat{F}^{-1}_{\widehat{G}_{Gini}}(\tau) - \widehat{F}^{-1}_{\widehat{G}_{Add}}(\tau)$. The average-maximizing treatment rule $\widehat{G}_{Add}$ generates an income distribution in which top quantiles (0.8 and higher) are substantially higher than in the income distribution generated by the Gini treatment rule. However, the distribution produced by the Gini treatment rule is better at midrange quantiles

(0.4-0.8). The additive welfare criterion equally weights changes of all quantiles, hence it favors $\widehat{G}_{Add}$.

The standard Gini welfare criterion, in contrast, uses decreasing welfare weights $\omega_3(\tau) = 2(1-\tau)$. The right panel of Figure 3 displays the same quantile differences between the two income distributions weighted by $\omega_3(\tau)$. With these equality-minded welfare weights, the gains offered by treatment rule $\widehat{G}_{Add}$ at top quantiles get a lower welfare weight than the gains offered by $\widehat{G}_{Gini}$ in the middle of the income distribution, hence $\widehat{G}_{Gini}$ is preferred under the Gini SWF.

6 Conclusion

This paper develops the first method for individualized treatment choice when the policy maker's objective is to maximize an equality-minded rank-dependent SWF. We showed that the average social welfare obtained by the estimated policy converges at the minimax-optimal $n^{-1/2}$ rate. The key restriction underlying these rate results is the complexity restriction (Assumption 2.2 (VC)) imposed on the set of feasible policies. This complexity restriction still allows for rich classes of individualized treatment rules and offers a flexible and convenient way to incorporate exogenous constraints that policy makers face in realistic settings of policy design. Our analytical results cover a general class of equality-minded rank-dependent SWFs. Computing the equality-minded EWM rule is more challenging than in the additive welfare case and so efficient computation 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 that

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

we can rewrite $F_\delta(y)$ defined in (2.5) as

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

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

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

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

Suppose that $\bar{W}_\Lambda - W_\Lambda(F_{G(u)}) > 0$ 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,

$$0 < \int_0^1 \left(\bar{W}_\Lambda - W_\Lambda(F_{G(u)}) \right) du = \bar{W}_\Lambda - \bar{W}_\Lambda,$$

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

□

The following five lemmas will be used in the proof of Theorem 3.1. The first lemma establishes a quadratic upper bound for the function $t^{-1/2}$ for $t \geq 1$.

Lemma A.1. *Let $t_0 > 1$, define*

$$g(t) \equiv \begin{cases} 0 & \text{for } t = 0, \\ t^{-1/2} & \text{for } t \geq 1, \end{cases} \tag{A.2}$$

$$h(t) \equiv t_0^{-1/2} - \frac{1}{2} t_0^{-3/2}(t - t_0) + t_0^{-2}(t - t_0)^2. \tag{A.3}$$

Then $g(t) \leq h(t)$ for $t = 0$ and for all $t \geq 1$.

Proof of Lemma A.1. For $t = 0$,

$$h(0) = t_0^{-1/2} + \frac{1}{2} t_0^{-3/2} t_0 + t_0^{-2} t_0^2 = \frac{3}{2} t_0^{-1/2} + 1 > 0 = g(0).$$

Now consider the function $(h-g)(t)$ and its derivatives for $t \geq 1$:

$$\begin{aligned}
(h-g)(t) &= t_0^{-1/2} - \frac{1}{2}t_0^{-3/2}(t-t_0) + t_0^{-2}(t-t_0)^2 - t^{-1/2}, \\
(h-g)'(t) &= -\frac{1}{2}t_0^{-3/2} + 2t_0^{-2}(t-t_0) + \frac{1}{2}t^{-3/2}, \\
(h-g)''(t) &= 2t_0^{-2} - \frac{3}{4}t^{-5/2}, \text{ and} \\
(h-g)'''(t) &= \frac{15}{8}t^{-7/2}.
\end{aligned}$$

First, we will show that $(h-g)(t) \geq 0$ for $t \in [1, t_0]$. The function is positive at $t = 1$:

$$\begin{aligned}
(h-g)(1) &= t_0^{-1/2} - \frac{1}{2}t_0^{-3/2}(1-t_0) + t_0^{-2}(1-t_0)^2 - 1 \\
&= t_0^{-1/2} - \frac{1}{2}t_0^{-3/2} + \frac{1}{2}t_0^{-1/2} + t_0^{-2} - 2t_0^{-1} + 1 - 1 \\
&= \frac{1}{2}t_0^{-1/2}\left(3 - 4t_0^{-1/2} - t_0^{-1} + 2t_0^{-3/2}\right) \\
&= \frac{1}{2}t_0^{-1/2}\left(1 - t_0^{-1/2}\right)^2\left(3 + 2t_0^{-1/2}\right) > 0
\end{aligned}$$

because $t_0^{-1/2} > 0$. At $t = t_0$, $(h-g)(t_0) = 0$. We will next show that $(h-g)(t) \geq 0$ between $t = 1$ and $t = t_0$.

The second derivative of $(h-g)$ is positive at $t = t_0$,

$$(h-g)''(t_0) = 2t_0^{-2} - \frac{3}{4}t_0^{-5/2} = t_0^{-2}\left(2 - \frac{3}{4}t_0^{-1/2}\right) > 0 \tag{A.4}$$

because $t_0 > 1$ by assumption, hence $t_0^{-1/2} < 1$. Since the third derivative is positive on $[1, t_0]$, it follows that the second derivative is either positive everywhere on $[1, t_0]$, or it is first negative on some interval $[1, t_2)$ and then positive on $(t_2, t_0]$.

The first derivative of $(h-g)$ equals zero at $t = t_0$:

$$(h-g)'(t_0) = -\frac{1}{2}t_0^{-3/2} + 2t_0^{-2}(t_0 - t_0) + \frac{1}{2}t_0^{-3/2} = 0.$$

If the second derivative is positive everywhere on $[1, t_0]$, then the first derivative must be negative everywhere on $[1, t_0)$. If the second derivative changes sign from negative to positive, then the first derivative must either be negative on $[1, t_0)$ or it could switch sign from positive on some interval $[1, t_1)$ to negative on (t_1, t_0).

Since $(h-g)(1) > 0$, $(h-g)(t_0) = 0$, and $(h-g)$ is either decreasing on $[1, t_0)$ or increasing on $[1, t_1)$ and then decreasing on (t_1, t_0), it follows that $(h-g)(t) \geq 0$ for all $t \in [1, t_0]$.

Second, consider $t > t_0$. At $t = t_0$, $(h-g)(t_0) = 0$, $(h-g)'(t_0) = 0$, and the second derivative is positive for all $t > t_0$ because it is positive at $t = t_0$ (A.4) and the third derivative is positive for all $t \geq t_0$. It follows that $(h-g)(t) > 0$ for all $t > t_0$. $\square$

The second lemma applies the bound in Lemma A.1 to the expectation of the function $g(\cdot)$ of a binomial variable.

Lemma A.2. *Suppose that random variable $B \sim \text{Binomial}(n,p)$ with $np > 1$ and $g(\cdot)$ is the function defined in (A.2). Then*

$$E[g(B)] < 2(np)^{-1/2}.$$

Proof of Lemma A.2. Let $h(\cdot)$ be the function defined in (A.3) with $t_0 = np = E[B]$. Lemma A.1 shows that $g(t) \leq h(t)$ for all values in the support of B ($t \in \{0,1,2,\dots\}$), therefore

$$\begin{aligned}
E[g(B)] \leq E[h(B)] &= E\left[(np)^{-1/2} - \frac{1}{2}(np)^{-3/2}(B - E[B]) + (np)^{-2}(B - E[B])^2\right] \\
&= (np)^{-1/2} - \frac{1}{2}(np)^{-3/2} \cdot 0 + (np)^{-2}\text{Var}[B] \\
&= (np)^{-1/2} + (np)^{-2}(np)(1-p) \\
&< (np)^{-1/2} + (np)^{-1} < 2(np)^{-1/2}
\end{aligned}$$

because $np > 1$ implies $p > 0$ and $(np)^{-1} < (np)^{-1/2}$. $\square$

Let $\mathbf{x}^l \equiv \{x_1, \dots, x_l\}$ be a finite set with $l \geq 1$ points in $\mathcal{X}$. Given a class of subsets $\mathcal{G}$ in $\mathcal{X}$, define $N(\mathbf{x}^l) = |\{\mathbf{x}^l \cap G : G \in \mathcal{G}\}|$ be the number of different subsets of $\mathbf{x}^l$ picked out by $G \in \mathcal{G}$. The VC-dimension $v \geq 1$ of $\mathcal{G}$ is defined by the largest l such that $\sup_{\mathbf{x}^l} N(\mathbf{x}^l) = 2^l$ holds (Vapnik (1998)). See Vapnik (1998), Dudley (1999, Chapter 4), and van der Vaart and Wellner (1996) for extensive discussions. Note that the VC-dimension is smaller by one compared to the *VC-index* used to measure the complexity of a class of sets in empirical process theory, e.g., van der Vaart and Wellner (1996).

Let $Z_i = (Y_i, D_i, X_i) \in \mathcal{Z}$, where $\mathcal{Z} \equiv \mathbb{R}_+ \times \{0,1\} \times \mathcal{X}$. The *subgraph* of a real-valued function $f : \mathcal{Z} \mapsto \mathbb{R}$ is the set

$$SG(f) \equiv \{(z,t) \in \mathcal{Z} \times \mathbb{R} : 0 \leq t \leq f(z) \text{ or } f(z) \leq t \leq 0\}.$$

The third lemma is reproduced from Kitagawa and Tetenov (2018b) (Lemma A.1). It establishes a link between the VC-dimension of a class of subsets in the covariate space $\mathcal{X}$ and the VC-dimension of a class of subgraphs of functions on $\mathcal{Z}$.

Lemma A.3. *Let $\mathcal{G}$ be a VC-class of subsets of $\mathcal{X}$ with VC-dimension $v < \infty$. Let g and h be two given functions from $\mathcal{Z}$ to $\mathbb{R}$. Then the set of functions from $\mathcal{Z}$ to $\mathbb{R}$*

$$\mathcal{F} = \{f_G(z) = g(z) \cdot 1\{x \in G\} + h(z) \cdot 1\{x \notin G\} : G \in \mathcal{G}\}$$

is a VC-subgraph class of functions with VC-dimension less than or equal to v.

The fourth lemma, reproduced from Kitagawa and Tetenov (2018b) (Lemma A.4), is a maximal inequalitiy that bounds the mean of a supremum of a centered empirical process indexed by a VC-subgraph class of functions.

Lemma A.4. *Let $\mathcal{F}$ be a class of uniformly bounded functions, i.e., there exists $\bar{F} < \infty$ such that $\|f\|_\infty \leq \bar{F}$ for all $f \in \mathcal{F}$. Assume that $\mathcal{F}$ is a VC-subgraph class with VC-dimension $v < \infty$. Then, there is a universal constant C_1 such that*

$$E_{P^n}\left[\sup_{f \in \mathcal{F}} |E_n(f) - E_P(f)|\right] \leq C_1 \bar{F}\sqrt{\frac{v}{n}}$$

holds for all $n \geq 1$.

The last novel lemma allows us to prove Theorem 3.1 for unbounded outcomes.

Lemma A.5. *Let $\mathcal{F}$ be a class of uniformly bounded functions, i.e., there exists $\bar{F} < \infty$ such that $\|f\|_\infty \leq \bar{F}$ for all $f \in \mathcal{F}$. Assume that $\mathcal{F}$ is a VC-subgraph class with VC-dimension $v < \infty$. Let $(Y, Z) \sim P$, where $Y \geq 0$ is a scalar (Y and Z may be dependent). Let $\{(Y_i, Z_i)\}_{i=1}^n \sim P^n$ be an i.i.d. sample from P. Assume that*

$$\int_0^\infty \sqrt{P(Y > y)}dy \leq M. \tag{A.5}$$

Then, there is a universal constant C_T such that

$$\int_0^\infty E_{P^n}\left[\sup_{f \in \mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(Z_i)1\{Y_i > y\} - E_P\left[f(Z)1\{Y > y\}\right]\right|\right] dy \leq C_T \bar{F} M\sqrt{\frac{v}{n}} \tag{A.6}$$

holds for all $n \geq 1$.

Proof of Lemma A.5. We start by deriving upper bounds for each value of y, $y > 0$, on

$$E_{P^n}[\xi_n(y)], \quad \xi_n(y) \equiv \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(Z_i) 1\{Y_i > y\} - E_P[f(Z)1\{Y > y\}] \right|.$$

First, consider values of y for which $nP(Y > y) \leq 1$. Due to the envelope condition, $\left| \frac{1}{n} \sum_{i=1}^{n} f(Z_i) 1\{Y_i > y\} \right| \leq \bar{F} \cdot \frac{1}{n} \sum_{i=1}^{n} 1\{Y_i > y\}$ for any $f \in \mathcal{F}$, hence

$$E_{P^n}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(Z_i) 1\{Y_i > y\} \right| \right] \leq \bar{F} \cdot E_{P^n}\left[\frac{1}{n} \sum_{i=1}^{n} 1\{Y_i > y\} \right] = \bar{F} P(Y > y).$$

Also, $E_P[f(Z)1\{Y > y\}] = P(Y > y) E_P[f(Z) | Y > y] \leq \bar{F} P(Y > y)$. It follows that

$$\begin{aligned} E_{P^n}[\xi_n(y)] &\leq E_{P^n}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(Z_i) 1\{Y_i > y\} \right| \right] + \sup_{f \in \mathcal{F}} E_P[f(Z)1\{Y > y\}] \\ &\leq 2\bar{F} P(Y > y) \leq 2\bar{F} \frac{\sqrt{P(Y > y)}}{\sqrt{n}}, \end{aligned} \tag{A.7}$$

where the last inequality holds because $\sqrt{nP(Y > y)} \leq 1$.

Second, we consider values of y for which $nP(Y > y) > 1$. Define random variables $N_y \equiv \sum_{i=1}^{n} 1\{Y_i > y\}$ for the number of observations in the data with $Y_i > y$, then

$$\frac{1}{n} \sum_{i=1}^{n} f(Z_i) 1\{Y_i > y\} = \begin{cases} 0 & \text{if } N_y = 0, \\ \frac{N_y}{n} \cdot \frac{1}{N_y} \sum_{i=1}^{n} f(Z_i) 1\{Y_i > y\} & \text{if } N_y \geq 1. \end{cases}$$

If $N_y \geq 1$ then

$$\begin{aligned} \xi_n(y) &= \sup_{f \in \mathcal{F}} \left| \frac{N_y}{n} \cdot \frac{1}{N_y} \sum_{i=1}^{n} f(Z_i) 1\{Y_i > y\} - P(Y > y) E_P[f(Z) | Y > y] \right| \qquad \text{(A.8)} \\ &= \sup_{f \in \mathcal{F}} \left| \begin{array}{l} \frac{N_y}{n} \cdot \frac{1}{N_y} \sum_{i=1}^{n} f(Z_i) 1\{Y_i > y\} - P(Y > y) \frac{1}{N_y} \sum_{i=1}^{n} f(Z_i) 1\{Y_i > y\} \\ + P(Y > y) \frac{1}{N_y} \sum_{i=1}^{n} f(Z_i) 1\{Y_i > y\} - P(Y > y) E_P[f(Z) | Y > y] \end{array} \right| \\ &\leq \left| \frac{N_y}{n} - P(Y > y) \right| \sup_{f \in \mathcal{F}} \left| \frac{1}{N_y} \sum_{i=1}^{n} f(Z_i) 1\{Y_i > y\} \right| \\ &\quad + P(Y > y) \sup_{f \in \mathcal{F}} \left| \frac{1}{N_y} \sum_{i=1}^{n} f(Z_i) 1\{Y_i > y\} - E_P[f(Z) | Y > y] \right| \\ &\leq \bar{F} \left| \frac{N_y}{n} - P(Y > y) \right| \text{ (the sum has } N_y \text{ non-zero terms, each bounded by } \bar{F}) \end{aligned}$$

$$+ P(Y > y) \sup_{f \in \mathcal{F}} \left| \frac{1}{N_y} \sum_{i=1}^{n} f(Z_i) 1\{Y_i > y\} - E_P\left[f(Z)|Y > y\right] \right|.$$

Note that $\{Z_i\}_{i:Y_i>y}$ is an i.i.d. sample of size N_y from the conditional distribution $P(Z|Y > y)$. We next apply the bound in Lemma A.4 for each value of $N_y \geq 1$:

$$E_{P^n}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{N_y} \sum_{i=1}^{n} f(Z_i) 1\{Y_i > y\} - E_P\left[f(Z)|Y > y\right] \right| \middle| N_y \right] \leq C_1 \bar{F} \sqrt{\frac{v}{N_y}}. \tag{A.9}$$

Combining inequality (A.8) with bound (A.9), and using the definition $g(t) = t^{-1/2}$ for $t \geq 1$ from (A.2), we obtain a bound on the conditional expectation of $\xi_n(y)$ for $N_y \geq 1$:

$$E_{P^n}\left[\xi_n(y)|N_y\right] \leq \bar{F} \left| \frac{N_y}{n} - P(Y > y) \right| + P(Y > y) C_1 \bar{F} \sqrt{v} g(N_y). \tag{A.10}$$

If $N_y = 0$ then $1\{Y_i > y\} = 0$ for all i, hence

$$\begin{aligned} \xi_n(y) &= \sup_{f \in \mathcal{F}} E_P\left[f(Z) 1\{Y > y\}\right] = P(Y > y) \sup_{f \in \mathcal{F}} E_P\left[f(Z)|Y > y\right] \\ &\leq \bar{F} P(Y > y) = \bar{F} \left| \frac{N_y}{n} - P(Y > y) \right| + P(Y > y) C_1 \bar{F} \sqrt{v} g(N_y), \end{aligned}$$

where the last equality uses the definition $g(0) = 0$ from (A.2). Therefore, the conditional expectation bound (A.10) also holds for $N_y = 0$.

The unconditional expectation of $\xi_n(y)$ is then bounded by

$$E_{P^n}\left[\xi_n(y)\right] \leq \bar{F} E_{P^n} \left| \frac{N_y}{n} - P(Y > y) \right| + P(Y > y) C_1 \bar{F} \sqrt{v} E_{P^n}\left[g(N_y)\right]. \tag{A.11}$$

The random variable N_y has a Binomial$(n, P(Y > y))$ distribution, hence

$$E_{P^n} \left| \frac{N_y}{n} - P(Y > y) \right| \leq \sqrt{\mathrm{Var}\left(\frac{N_y}{n}\right)} = \sqrt{\frac{P(Y > y)(1 - P(Y > y))}{n}} \leq \sqrt{\frac{P(Y > y)}{n}}.$$

Since $nP(Y > y) > 1$, applying Lemma A.2 yields $E_{P^n}\left[g(N_y)\right] \leq \frac{2}{\sqrt{n}\sqrt{P(Y>y)}}$. Combining this inequality with (A.11) and $v \geq 1$ we obtain

$$\begin{aligned} E_{P^n}\left[\xi_n(y)\right] &\leq \bar{F} \sqrt{\frac{P(Y > y)}{n}} + P(Y > y) C_1 \bar{F} \sqrt{v} \frac{2}{\sqrt{n}\sqrt{P(Y > y)}} \\ &\leq 2(1 + C_1) \bar{F} \sqrt{\frac{v}{n}} \sqrt{P(Y > y)} = C_T \bar{F} \sqrt{\frac{v}{n}} \sqrt{P(Y > y)}, \end{aligned} \tag{A.12}$$

where $C_T \equiv 2(1+C_1)$. This bound is higher than the bound (A.7) derived for y such that $nP(Y>y) \leq 1$, hence bound (A.12) holds for all $y \geq 0$.

The last step is to integrate the bound (A.12) over y and apply (A.5):

$$\int_0^\infty E_{P^n}\left[\xi_n(y)\right] dy \leq \int_0^\infty C_T \bar{F} \sqrt{\frac{v}{n}} \sqrt{P(Y>y)} dy \leq C_T \bar{F} M \sqrt{\frac{v}{n}}.$$

□

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

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

where the second line follows since $\widehat{W}_\Lambda(\widehat{G})$ maximizes $\widehat{W}_\Lambda(G)$ over $G \in \mathcal{G}$. It follows that

$$\sup_{G \in \mathcal{G}} W_\Lambda(G) - W_\Lambda(\widehat{G}) \leq 2 \sup_{G \in \mathcal{G}} \left| \widehat{W}_\Lambda(G) - W_\Lambda(G) \right|. \tag{A.13}$$

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

$$\begin{aligned} \sup_{G \in \mathcal{G}} \left| \widehat{W}_\Lambda(G) - W_\Lambda(G) \right| &= \sup_{G \in \mathcal{G}} \left| \int_0^\infty \Lambda(\widehat{F}_G(y) \vee 0) dy - \int_0^\infty \Lambda(F_G(y)) dy \right| \\ &\leq \sup_{G \in \mathcal{G}} \int_0^\infty \left| \Lambda(\widehat{F}_G(y) \vee 0) - \Lambda(F_G(y)) \right| dy \\ &\leq \int_0^\infty \sup_{G \in \mathcal{G}} \left| \Lambda(\widehat{F}_G(y) \vee 0) - \Lambda(F_G(y)) \right| dy \\ &\leq |\Lambda'(0)| \int_0^\infty \sup_{G \in \mathcal{G}} \left| \widehat{F}_G(y) - F_G(y) \right| dy. \end{aligned} \tag{A.14}$$

Combining (A.13) and (A.14), the average welfare loss of $\widehat{G}$ can be bounded by

$$E_{P^n}\left[\sup_{G \in \mathcal{G}} W_\Lambda(G) - W_\Lambda(\widehat{G})\right] \leq 2|\Lambda'(0)| \int_0^\infty E_{P^n}\left[\sup_{G \in \mathcal{G}} \left| \widehat{F}_G(y) - F_G(y) \right|\right] dy. \tag{A.15}$$

By Lemma A.3, the class of functions $\mathcal{W} = \{w_G(\cdot) : G \in \mathcal{G}\}$, where

$$w_G(Z_i) \equiv \left[\frac{D_i}{e(X_i)} \cdot 1\{X_i \in G\} + \frac{1-D_i}{1-e(X_i)} \cdot 1\{X_i \notin G\}\right] \tag{A.16}$$

is a VC-subgraph class with VC-dimension of at most v. Assumption 3.1 (SO) implies that $w_G(Z_i) \in \left[0, \frac{1}{\kappa}\right]$, hence functions in $\mathcal{W}$ are uniformly bounded by $\frac{1}{\kappa}$.

Since $F_G(y) = 1 - E_P[w_G(Z) \cdot 1\{Y > y\}]$ and $\widehat{F}_G(y) = 1 - \frac{1}{n}\sum_{i=1}^{n} w_G(Z_i) \cdot 1\{Y_i > y\}$,

$$\left|\widehat{F}_G(y) - F_G(y)\right| = \left|\frac{1}{n}\sum_{i=1}^{n} w_G(Z_i) \cdot 1\{Y_i > y\} - E_P[w_G(Z) \cdot 1\{Y > y\}]\right|.$$

It follows from Assumption 3.1 (TC) and $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ that

$$\begin{aligned}\int_0^\infty \sqrt{P(Y > y)}dy &= \int_0^\infty \sqrt{P(Y_1 > y, D = 1) + P(Y_0 > y, D = 0)}dy \\ &\leq \int_0^\infty \left[\sqrt{P(Y_1 > y)} + \sqrt{P(Y_0 > y)}\right] dy \leq 2\Upsilon. \end{aligned} \tag{A.17}$$

Applying Lemma A.5 to (A.15) yields

$$E_{P^n}\left[\sup_{G \in \mathcal{G}} W_\Lambda(G) - W_\Lambda(\widehat{G})\right] \leq 4C_T|\Lambda'(0)|\frac{\Upsilon}{\kappa}\sqrt{\frac{v}{n}}.$$

Setting $C = 4C_T$ completes the proof of (3.8).

The proof of (3.9) is found in the online supplement. □

References

Aaberge, R., T. Havnes, and M. Mogstad (2017): "A Partial Identification Approach to Welfare Rankings of Distributions," *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.

Athey, S. and S. Wager (2018): "Efficient Policy Learning," *arXiv preprint*, arXiv 1702.0289.

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

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.

Donaldson, D. and J. A. Weymark (1980): "A Single-Parameter Generalization of the Gini Indices of Inequality," *Journal of Economic Theory*, 22, 67–86.

——— (1983): "Ethically Flexible Gini Indices for Income Distributions in the Continuum," *Journal of Economic Theory*, 29, 353–358.

Dudley, R. (1999): *Uniform Central Limit Theorems*, Cambridge University Press.

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 (2016): "Identification and Estimation of Distributional Impacts of Interventions Using Changes in Inequality Measures," *Journal of Applied Econometrics*, 31, 457–486.

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.

Imbens, G. W. (2004): "Nonparametric estimation of average treatment effects under exogeneity: a review," *Review of Economics and Statistics*, 86, 4–29.

Kasy, M. (2016): "Partial Identification, Distributional Preferences, and the Welfare Ranking of Policies," *Review of Economics and Statistics*, 98, 111–131.

——— (2018): "Optimal taxation and insurance using machine learning – sufficient statistics and beyond," *Journal of Public Economics*, forthcoming.

Kitagawa, T. and A. Tetenov (2018a): "Who Should Be Treated? Empirical Welfare Maximization Methods for Treatment Choice," *Econometrica*, 86, 591–616.

——— (2018b): "Supplement to "Who Should Be Treated? Empirical Welfare Maximization Methods for Treatment Choice"," *Econometrica*.

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

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.

Mbakop, E. and M. Tabord-Meehan (2018): “Model Selection for Treatment Choice: Penalized Welfare Maximization,” *arXiv preprint*, arXiv 1609.03167.

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.

van der Vaart, A. W. and J. A. Wellner (1996): *Weak Convergence and Empirical Processes*, Springer.

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

Wang, L., Y. Zhou, R. Song, and B. Sherwood (2018): “Quantile-Optimal Treatment Regimes,” *Journal of the American Statistical Association*, 113, 1243–1254.

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.

Online Supplement to "Equality-Minded Treatment Choice"

Toru Kitagawa* and Aleksey Tetenov†

December 8, 2018

Abstract

This online supplement contains the materials and proofs omitted from Kitagawa and Tetenov (2018), "Equality-minded Treatment Choice."

B Illustrative Example

In this section, we illustrate the properties of rank-dependent SWFs in comparison with the utilitarian one in a simple setting with the Gini SWF, $W_{Gini}(F) = \int_0^\infty (1 - F(y))^2 dy$. We first compare the welfare ordering on the parametric family of log-normal outcome distributions. Second, we consider a simple treatment choice problem with binary X in order to illustrate how the optimal rules fundamentally differ between the two SWFs.

First, consider the welfare ordering over the family of log-normal distributions of outcomes, $Y \sim \log N(\mu, \sigma^2)$, ignoring the treatment choice problem. The mean of Y is given by $E(Y) = \exp(\mu + \sigma^2/2)$. The Gini inequality coefficient for $\log N(\mu, \sigma^2)$ 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 distribution. By (2.4), we have

$$W_{Gini}(\mu, \sigma) \equiv 2 \exp\left(\mu + \frac{\sigma^2}{2}\right)\left[1 - \Phi\left(\frac{\sigma}{\sqrt{2}}\right)\right]. \tag{B.1}$$

This welfare function is increasing in μ, whereas it is not monotonic in σ. For instance, when $\mu = 0$, $W_{Gini}(\mu, \sigma)$ is decreasing in σ for $\sigma < 0.87$ and increasing for $\sigma > 0.87$. See Figure B.1 for a plot of $W_{Gini}(\mu, \sigma)$ over $\sigma \in [0, 2]$ holding $\mu = 0$ fixed. The U-shape of the Gini social

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

†University of Bristol, Email: a.tetenov@bristol.ac.uk

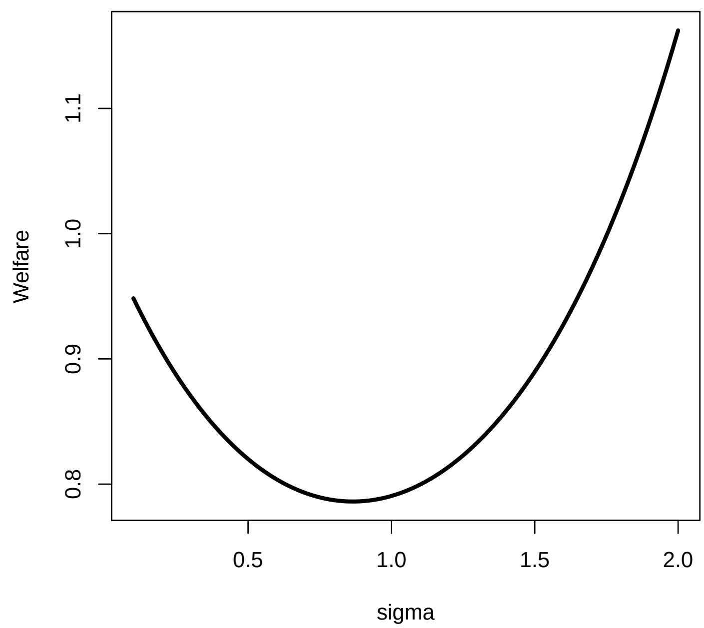

Figure B.1: Equality-minded welfare for $\log N(0, \sigma^2)$.

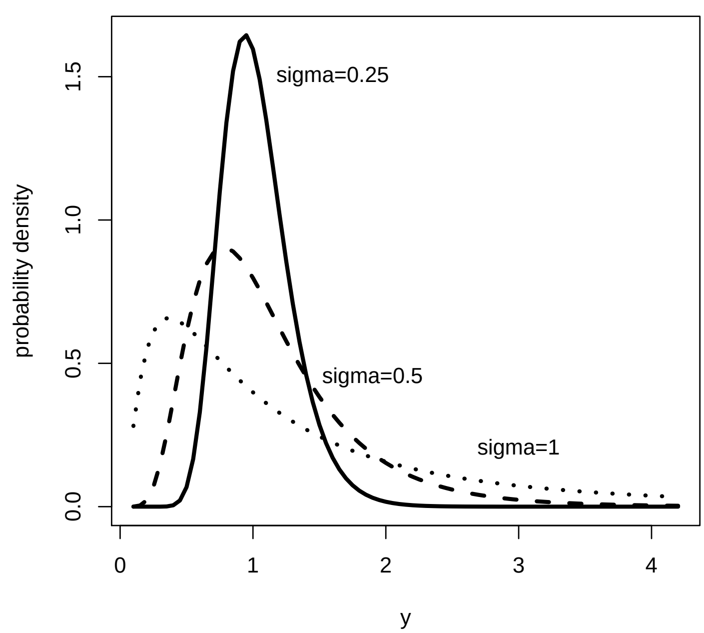

Figure B.2: Density of $\log N(0, \sigma^2)$.

welfare indicates that for $\sigma < 0.87$, the negative contribution to the social welfare from an increase in the Gini coefficient dominates the positive contribution from an increase in the mean, while for $\sigma > 0.87$, this relationship reverses. In Figure B.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 always preferable in terms of the utilitarian social welfare. In contrast, as shown in the welfare values plotted in Figure B.1, the Gini social welfare yields the complete opposite welfare ordering over the three log-normal distributions in Figure B.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:

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

According to Theorem 2.1, it suffices to consider non-randomized rules to search for an optimal one. We therefore 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 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. In general, when stochastic dominance relationships between $Y_1|X$- and $Y_0|X$-distributions are present for all X, the optimal rule for the rank-dependent social welfare agrees with the utilitarian one and can be obtained by solving the treatment choice problem separately in each subpopulation.

These results change drastically once we let $\sigma_a \neq \sigma_b$. Suppose we fix $\mu_a = \mu_b = 0$, while we vary both σ_a and σ_b over $[0.1, 1.6]$. As the mean of a log normal random variable is

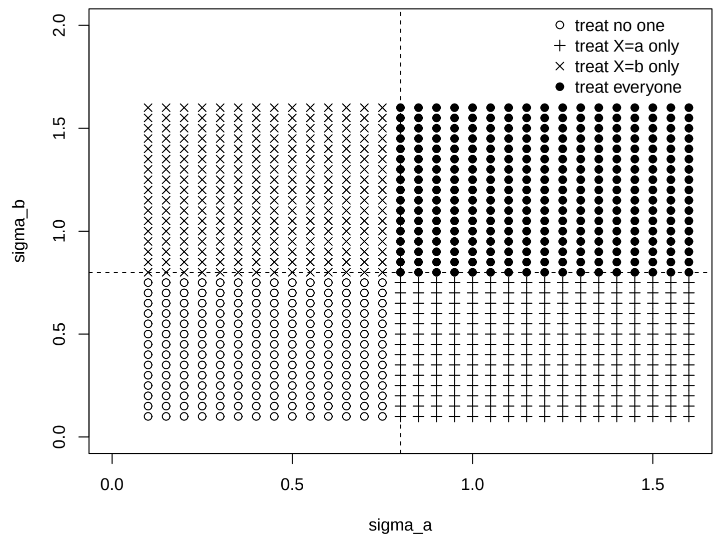

Figure B.3: Optimal policies under the additive welfare. Log-normal potential outcome distributions with $\mu_a = \mu_b = 0$.

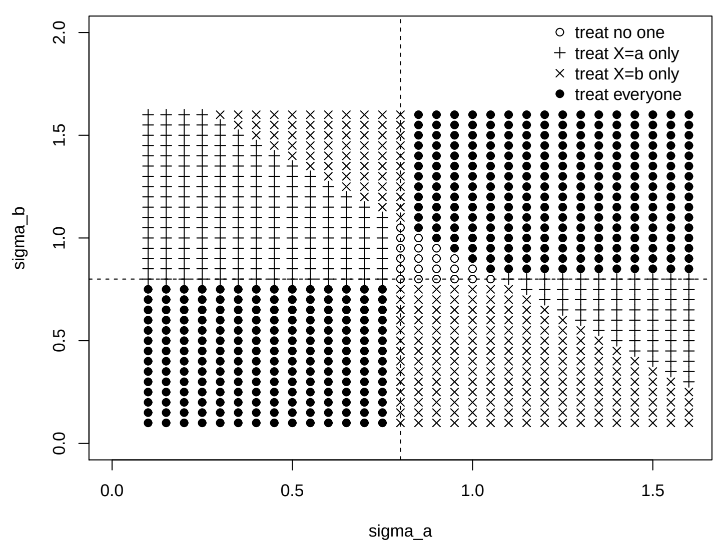

Figure B.4: Optimal policies under the Gini welfare. Log-normal potential outcome distributions with $\mu_a = \mu_b = 0$.

increasing in σ, the optimal treatment rule for the additive welfare is obtained by

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

In Figure B.3, we plot the optimal treatment rule under the additive welfare at each grid point of $(\sigma_a, \sigma_b) \in [0.1, 1.6]^2$. Since the additive social welfare is separable over the subpopulations, a treatment preferable for one subpopulation does not depend on the treatment assigned to the other subpopulation. The regions in which different rules from $\mathcal{G}$ are optimal form a quadrant partition, as shown in Figure B.3.

In Figure B.4, we plot the optimal policies in terms of the Gini social welfare. The regions in which different rules from $\mathcal{G}$ are optimal are strikingly different compared with the additive welfare case (G^*_{Add}) shown in Figure B.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 B.1), while the additive welfare is monotonically increasing in σ. Another notable difference is that in contrast to the quadrant partition observed in the additive welfare case, the partition in the equality-minded welfare case is more complex. Some treatment rules are optimal in disconnected regions, e.g., G_{ab} is optimal in the south-west and the north-east regions of the plot. Furthermore, the region in which G_a is optimal can border the region in which G_b is optimal. On the border between these regions, the policy maker chooses whether to treat $X = a$ only or $X = b$ only, rather than whether to additionally treat the other subpopulation.

C EWM with Estimated Propensity Score

Unknown propensity score is common in observational studies. This section considers the equality-minded EWM approach with estimated propensity scores and investigates the influence of the lack of knowledge on propensity scores on 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

$$\widehat{W}_{\Lambda}^{e}(G) = \int_{0}^{\infty} \Lambda(\widehat{F}_{G}^{e}(y) \vee 0) dy,$$
$$\widehat{F}_{G}^{e}(y) \equiv 1 - \frac{1}{n} \sum_{i=1}^{n} \left[\frac{D_i}{\hat{e}(X_i)} \cdot 1\{X_i \in G\} + \frac{(1 - D_i)}{1 - \hat{e}(X_i)} \cdot 1\{X_i \notin G\} \right] \cdot 1\{Y_i > y\}.$$

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

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

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 C.1. For a class of data generating processes $\mathcal{P}_e$, there exist sequences $\phi_n, \tilde{\phi}_n \to \infty$ as $n \to \infty$ such that

$$\limsup_{n \to \infty} \sup_{P \in \mathcal{P}_e} \phi_n E_{P^n} \left[\frac{1}{n} \sum_{i=1}^{n} \left| \frac{1}{e(X_i)} - \frac{1}{\hat{e}(X_i)} \right| \right] < \infty, \tag{C.1}$$
$$\limsup_{n \to \infty} \sup_{P \in \mathcal{P}_e} \phi_n E_{P^n} \left[\frac{1}{n} \sum_{i=1}^{n} \left| \frac{1}{1 - e(X_i)} - \frac{1}{1 - \hat{e}(X_i)} \right| \right] < \infty,$$
$$\limsup_{n \to \infty} \sup_{P \in \mathcal{P}_e} E_{P^n} \left[\left(\tilde{\phi}_n \max_{1 \le i \le n} \left| \frac{1}{e(X_i)} - \frac{1}{\hat{e}(X_i)} \right| \right)^2 \right] < \infty, \quad \text{and} \tag{C.2}$$
$$\limsup_{n \to \infty} \sup_{P \in \mathcal{P}_e} E_{P^n} \left[\left(\tilde{\phi}_n \max_{1 \le i \le n} \left| \frac{1}{1 - e(X_i)} - \frac{1}{1 - \hat{e}(X_i)} \right| \right)^2 \right] < \infty$$

hold.

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

Theorem C.1. *Suppose Assumptions 2.1, 2.2 and 3.1 hold. For a class of data generating processes $\mathcal{P}_e$, if an estimator for the propensity score satisfies Assumption C.1, then*

$$\sup_{P\in\mathcal{P}_e\cap\mathcal{P}} E_{P^n}\left[\sup_{G\in\mathcal{G}} W_\Lambda(G) - W_\Lambda(\widehat{G}^e)\right] \leq O\left(\phi_n^{-1} \vee \sqrt{\frac{v}{n}}\right). \tag{C.3}$$

Proof. See Appendix D.

This theorem extends Theorem 2.5 (e) of Kitagawa and Tetenov (2018a) to the cases of rank-dependent social welfare or unbounded outcome or both. 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 $\phi_n = n^{\frac{1}{2+d_x/\beta_e}}$ and $\tilde{\phi}_n = \log n \cdot (\log n/n)^{\frac{1}{2+d_x/\beta_e}}$, where $\beta_e \geq 1$ is the parameter constraining smoothness of $e(\cdot)$ in terms of the degree of the 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 C.1 implies

$$\sup_{P\in\mathcal{P}_e\cap\mathcal{P}} E_{P^n}\left[\sup_{G\in\mathcal{G}} W_\Lambda(G) - W_\Lambda(\widehat{G}^e)\right] \leq O\left(n^{-\frac{1}{2+d_x/\beta_e}}\right), \tag{C.4}$$

as long as the VC-dimension of $\mathcal{G}$ is either constant or does not grow too fast as the sample size increases. For a formal derivation of (C.4) and the precise construction of the local polynomial estimator for $e(\cdot)$, see Appendix D.

D Additional proofs

Proof of Theorem 3.1 (ii). Similarly to inequalities (A.14) and (A.15) shown in Appendix A, the average welfare regret for the normalized cdf case can be bounded by

$$E_{P^n}\left[\sup_{G\in\mathcal{G}} W_\Lambda(G) - W_\Lambda(\widehat{G}^R)\right] \leq 2|\Lambda'(0)| \int_0^\infty E_{P^n}\left[\sup_{G\in\mathcal{G}}\left|\widehat{F}_G^R(y) - F_G(y)\right|\right] dy. \tag{D.1}$$

We hence focus on bounding $\int_0^\infty E_{P^n}\left[\sup_{G\in\mathcal{G}}\left|\widehat{F}_G^R(y) - F_G(y)\right|\right] dy$.

Let $w_G(Z_i)$ be as defined in (A.16), and let

$$A_{n,G} \equiv \{\widehat{F}_G(-\infty) < 1\} = \{n^{-1}\sum_{i=1}^{n} w_G(Z_i) > 0\}$$

denote the event that the normalizing term in $\widehat{F}^R_G(y)$ at policy G is nonzero, and $A^c_{n,G} \equiv \{\widehat{F}_G(-\infty) = 1\} = \{n^{-1}\sum_{i=1}^{n} w_G(Z_i) = 0\}$ be the complement of $A_{n,G}$. Using the indicator functions for $A_{n,G}$ and $A^c_{n,G}$, $\widehat{F}^R_G(y)$ can be written as

$$\widehat{F}^R_G(y) = \left[1 - \frac{1}{n}\sum_{i=1}^{n} w^R_{G,i} 1\{Y_i > y\}\right] \cdot 1\{A_{n,G}\} + [1 - 1\{y < \min_{1\le i\le n} Y_i\}] \cdot 1\{A^c_{n,G}\}, \quad \text{(D.2)}$$

where

$$w^R_{G,i} = \frac{w_G(Z_i)}{n^{-1}\sum_{i=1}^{n} f_G(Z_i) + 1}, \qquad f_G(Z_i) = w_G(Z_i) - 1. \quad \text{(D.3)}$$

By the triangle inequality,

$$\begin{aligned} &\left|\widehat{F}^R_G(y) - F_G(y)\right| \le \left|\widehat{F}^R_G(y) - \widehat{F}_G(y)\right| + \left|\widehat{F}_G(y) - F_G(y)\right| \\ &\le \left|\frac{1}{n}\sum_{i=1}^{n}\left[w^R_{G,i} - w_G(Z_i)\right] 1\{Y_i > y\}\right| \cdot 1\{A_{n,G}\} + 1\{y < \min_{1\le i\le n} Y_i\} \cdot 1\{A^c_{n,G}\} + \left|\widehat{F}_G(y) - F_G(y)\right|. \end{aligned} \quad \text{(D.4)}$$

Let

$$S^-_n \equiv \inf_{G\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} f_G(Z_i),$$
$$S_n \equiv \sup_{G\in\mathcal{G}} \left|\frac{1}{n}\sum_{i=1}^{n} f_G(Z_i)\right|.$$

For $\delta \in (0,1)$ fixed, define

$$\Omega_{n,\delta} \equiv \{S_n - E_{P^n}(S_n) \le \delta/2\} = \{-S_n \ge -E_{P^n}(S_n) - \delta/2\}.$$

By Lemma A.3, $\{f_G : G \in \mathcal{G}\}$ is a VC-subgraph class of functions with VC-dimension at most v with $E_P(f_G) = 0$ and an envelope $\|f_G\|_\infty \le \frac{1-\kappa}{\kappa}$. Hence, by Lemma A.4,

$$E_{P^n}(S_n) \le C_1 \frac{1-\kappa}{\kappa}\sqrt{\frac{v}{n}}$$

holds, where C_1 is the universal constant defined in Lemma A.4. Accordingly, for all $n > n(\delta, v) \equiv \left(\frac{C_1(1-\kappa)}{\kappa(1-\delta)}\right)^2 v$, $-E_{P_n}(S_n) - \delta/2 > -1 + \delta/2$ holds. Since $S_n^- \geq -S_n$ holds, $\Omega_{n,\delta}$ being true and $n > n(\delta, v)$ imply $S_n^- > -1 + \delta/2$. Hence, on $\Omega_{n,\delta}$ and for $n > n(\delta, v)$, we have $0 \leq w_{G,i}^R \leq (2/\delta) w_G(Z_i)$ and

$$\left|w_{G,i}^R - w_G(Z_i)\right| = w_{G,i}^R \left|\frac{1}{n}\sum_{i=1}^n f_G(Z_i)\right| \leq \frac{2}{\delta} \cdot w_G(Z_i) S_n. \tag{D.5}$$

On $\Omega_{n,\delta}^c$ and for G such that $A_{n,G}$ is true, we have $0 \leq w_{G,i}^R \leq n$ and

$$\left|w_{G,i}^R - w_G(Z_i)\right| \leq n\frac{1-\kappa}{\kappa}. \tag{D.6}$$

Combining (D.5) and (D.6), (D.4) can be rewritten as

$$\begin{aligned}
&\left|\widehat{F}_G^R(y) - F_G(y)\right| \\
\leq &\frac{2}{\delta} \cdot S_n \cdot \frac{1}{n}\sum_{i=1}^n w_G(Z_i) 1\{Y_i > y\} \cdot 1\{\Omega_{n,\delta} \cap A_{n,G}\} + n\frac{1-\kappa}{\kappa} \cdot \frac{1}{n}\sum_{i=1}^n 1\{Y_i > y\} \cdot 1\{\Omega_{n,\delta}^c \cap A_{n,G}\} \\
&+ 1\{y < \min_{1\leq i\leq n} Y_i\} \cdot 1\{\Omega_{n,\delta} \cap A_{n,G}^c\} + 1\{y < \min_{1\leq i\leq n} Y_i\} \cdot 1\{\Omega_{n,\delta}^c \cap A_{n,G}^c\} + \left|\widehat{F}_G(y) - F_G(y)\right|
\end{aligned} \tag{D.7}$$

Note that $\{S_n^- > -1\}$ is equivalent to $\left\{\inf_{G\in\mathcal{G}} n^{-1}\sum_{i=1}^n w_G(Z_i) > 0\right\}$, implying that $A_{n,G}$ is true for all $G \in \mathcal{G}$. Hence, for $n > n(\delta, v)$, $\Omega_{n,\delta} \cap A_{n,G} = \Omega_{n,\delta}$, and $\Omega_{n,\delta} \cap A_{n,G}^c = \emptyset$ hold for all $G \in \mathcal{G}$. By also noting $w_G(Z_i) \leq \frac{D_i}{e(X_i)} + \frac{1-D_i}{1-e(X_i)}$, (D.7) can be further bounded by

$$\begin{aligned}
\leq &\frac{2}{\delta} \cdot S_n \cdot \frac{1}{n}\sum_{i=1}^n \left[\frac{D_i}{e(X_i)} + \frac{1-D_i}{1-e(X_i)}\right] 1\{Y_i > y\} \cdot 1\{\Omega_{n,\delta}\} \\
&+ n\frac{1-\kappa}{\kappa} \cdot \frac{1}{n}\sum_{i=1}^n 1\{Y_i > y\} \cdot 1\{\Omega_{n,\delta}^c \cap A_{n,G}\} \\
&+ 1\{y < \min_{1\leq i\leq n} Y_i\} \cdot 1\{\Omega_{n,\delta}^c \cap A_{n,G}^c\} + \left|\widehat{F}_G(y) - F_G(y)\right| \\
\leq &\frac{2}{\delta} \cdot S_n \left[P(Y_1 > y) + P(Y_0 > y)\right] \cdot 1\{\Omega_{n,\delta}\} \\
&+ \frac{2}{\delta} \cdot \left(1 - \frac{\delta}{2}\right) \cdot \frac{1}{n}\sum_{i=1}^n \left[\left(\frac{D_i}{e(X_i)} + \frac{1-D_i}{1-e(X_i)}\right) 1\{Y_i > y\} - P(Y_1 > y) - P(Y_0 > y)\right] \cdot 1\{\Omega_{n,\delta}\} \\
&+ n\frac{1-\kappa}{\kappa} \cdot \frac{1}{n}\sum_{i=1}^n 1\{Y_i > y\} \cdot 1\{\Omega_{n,\delta}^c\} + \left|\widehat{F}_G(y) - F_G(y)\right|,
\end{aligned}$$

$$
\begin{aligned}
\leq & \frac{2}{\delta} \cdot S_n \left[P(Y_1 > y) + P(Y_0 > y) \right] \\
& + \frac{2}{\delta} \cdot \left(1 - \frac{\delta}{2}\right) \cdot \frac{1}{n} \sum_{i=1}^{n} \left[\left(\frac{D_i}{e(X_i)} + \frac{1 - D_i}{1 - e(X_i)} \right) 1\{Y_i > y\} - P(Y_1 > y) - P(Y_0 > y) \right] \\
& + n \frac{1-\kappa}{\kappa} \cdot \frac{1}{n} \sum_{i=1}^{n} 1\{Y_i > y\} \cdot 1\{\Omega_{n,\delta}^c\} + \left| \widehat{F}_G(y) - F_G(y) \right|,
\end{aligned} \tag{D.8}
$$

where the second inequality follows from the fact that $S_n \leq \left(1 - \frac{\delta}{2}\right)$ holds on $\Omega_{n,\delta}$ and for $n > n(\delta, v)$, and $n\frac{1-\kappa}{\kappa} \geq 1$ and $n^{-1} \sum_{i=1}^{n} 1\{Y_i > y\} \geq 1\{y < \min_{1 \leq i \leq n} Y_i\}$ hold for all y. Since the second term in (D.8) has mean zero, $E_{P^n}\left[\sup_{G \in \mathcal{G}} \left| \widehat{F}_G^R(y) - F_G(y) \right|\right]$ can be bounded by

$$
\begin{aligned}
& E_{P^n}\left[\sup_{G \in \mathcal{G}} \left| \widehat{F}_G^R(y) - F_G(y) \right|\right] \\
\leq & \underbrace{\frac{2}{\delta} \cdot E_{P^n}[S_n] \cdot (P(Y_1 > y) + P(Y_0 > y))}_{\text{(i)}} + \underbrace{n\frac{1-\kappa}{\kappa} \cdot E_{P^n}\left[\frac{1}{n}\sum_{i=1}^{n} 1\{Y_i > y\} \cdot 1\{\Omega_{n,\delta}^c\}\right]}_{\text{(ii)}} \\
& + \underbrace{E_{P^n}\left[\sup_{G \in \mathcal{G}} \left| \widehat{F}_G(y) - F_G(y) \right|\right]}_{\text{(iii)}}.
\end{aligned} \tag{D.9}
$$

By Assumption 3.1 (TC) and Lemma A.4, the integral of term (i) in (D.9) can be bounded as

$$
\int_0^\infty \text{(i)} dy \leq \frac{4C_1}{\delta} \cdot \frac{1-\kappa}{\kappa} \sqrt{\frac{v}{n}} \Upsilon, \tag{D.10}
$$

where we use $E_{P^n}[S_n] \leq C_1 \frac{1-\kappa}{\kappa}\sqrt{\frac{v}{n}}$ and $\int_0^\infty P(Y_d > y) dy \leq \int_0^\infty \sqrt{P(Y_d > y)} dy \leq \Upsilon$.

Consider term (ii); by the Cauchy-Schwarz inequality,

$$
\begin{aligned}
\text{(ii)} & \leq n\frac{1-\kappa}{\kappa} \sqrt{E_{P^n}\left[\left(\frac{1}{n}\sum_{i=1}^{n} 1\{Y_i > y\}\right)^2\right]} \sqrt{P^n(\Omega_{n,\delta}^c)} \\
& \leq n\frac{1-\kappa}{\kappa} \sqrt{P(Y > y)} \sqrt{P^n(\Omega_{n,\delta}^c)}
\end{aligned}
$$

Bernstein's inequality (see, e.g., Theorem 12.2 in Boucheron et al. (2013)) implies that

$$
P^n(\Omega_{n,\delta}^c) \leq 2P^n\left(-S_n^- - E_{P^n}(-S_n^-) \geq \frac{\delta}{2}\right) \leq 2\exp\left\{-\frac{(\delta/2)^2 n}{2[2(\Sigma_f^2 + \sigma_f^2) + \bar{f}\delta/2]}\right\},
$$

where $\Sigma_f^2 \equiv E_{P^n}\left[\sup_{G\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^n f_G^2(Z_i)\right] \leq \left(\frac{1-\kappa}{\kappa}\right)^2$, $\sigma_f^2 \equiv \sup_{G\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^n E_P(f_G^2(Z_i)) \leq \left(\frac{1-\kappa}{\kappa}\right)^2$, and $\bar{f} \equiv \sup_{G\in\mathcal{G}} \|f_G\|_\infty \leq \frac{1-\kappa}{\kappa}$. Hence,

$$\begin{aligned}
\sqrt{P^n(\Omega_{n,\delta})} &\leq \sqrt{2}\exp\left\{-\frac{\delta^2 n}{16[2(\Sigma_f^2+\sigma_f^2)+\bar{f}\delta/2]}\right\} \\
&\leq \sqrt{2}\exp\left\{-\frac{\delta^2\kappa^2 n}{16[4(1-\kappa)^2+(1-\kappa)\kappa\delta/2]}\right\} \\
&\leq \sqrt{2}\exp\left\{-c_1(\delta)\kappa^2 n\right\}
\end{aligned}$$

holds, where $c_1(\delta) = \delta^2/(64+8\delta) > 0$. The integral of term (ii) can be therefore bounded by

$$\int_0^\infty (\text{ii})dy \leq \frac{2\sqrt{2}(1-\kappa)\Upsilon}{\kappa} \cdot n\exp\left\{-c_1(\delta)\kappa^2 n\right\}. \tag{D.11}$$

As shown in the proof of Theorem 3.1 (i), Lemma A.5 applies to term (iii) to yield

$$\int_0^\infty (\text{iii})dy \leq \frac{2C_T\cdot\Upsilon}{\kappa}\sqrt{\frac{v}{n}} \tag{D.12}$$

Combining (D.1), (D.9), (D.10), (D.11), and (D.12), and setting $\delta = 1/2$, we conclude

$$E_{P^n}\left[\sup_{G\in\mathcal{G}} W_\Lambda(G) - W_\Lambda(\widehat{G}^R)\right] \leq \frac{\Lambda'(0)\Upsilon}{\kappa}\left[C_1^R\sqrt{\frac{v}{n}} + 4\sqrt{2}n\exp\{-C_2^R\kappa^2 n\}\right]$$

for all $n > n(1/2, v) = C_3^R\left(\frac{1-\kappa}{\kappa}\right)^2 v$, where $C_1^R = 16C_1 + 4C_T$, $C_2^R = c_1(1/2) = 1/272$, and $C_3^R = 4C_1^2$. □

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 $n^{-1/2}$. Specifically, we restrict distributions of potential outcomes to those whose supports are restricted to $[0,\Upsilon]$. Any such distribution satisfies Assumption 3.1 (TC), since $\int_0^\infty \sqrt{P(Y_d > y)}dy = \int_0^\Upsilon \sqrt{P(Y_d > y)}dy \leq \Upsilon$.

To simplify the proof, we normalize the range of outcomes to $Y \in [0,1]$. We rescale the ourcome to $Y \in [0,\Upsilon]$ in the final step of the proof by multiplying Υ to the regret lower bound, as the rank-dependent SWF is equivariant to a multiplicative positive constant to Y.

The construction of $\mathcal{P}^*$ proceeds as follows. We restrict the range of outcomes to binary $Y \in \{0, 1\}$. 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 be supported only on $(x_1, \ldots, x_v)$. Let $\tau^* \in (0, 1]$ stated in the current theorem be given. We put mass $p \equiv \frac{\tau^*}{v-1}$ at x_i for all $i < v$, and mass $1 - \tau^*$ at x_v. The constructed marginal distribution of X is common in $\mathcal{P}^*$. Let the distribution of the 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 $b_j = 1$,

$$Y_0|(X = x_j) \sim Ber\left(\frac{1-\gamma}{2}\right), \quad Y_1|(X = x_j) \sim Ber\left(\frac{1+\gamma}{2}\right), \tag{D.13}$$

and, if $b_j = 0$,

$$Y_0|(X = x_j) \sim Ber\left(\frac{1+\gamma}{2}\right), \quad Y_1|(X = x_j) \sim Ber\left(\frac{1-\gamma}{2}\right), \tag{D.14}$$

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 $W_\Lambda = \Lambda(1 - \mu)$, 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

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

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 the treatment assignment rule $\widehat{G}$ is implemented for a given realization of the sample. Outcomes are binary for all $P \in \mathcal{P}^*$, 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:

$$\begin{aligned}
\sup_{P \in \mathcal{P}} E_{P^n}\left[\sup_{G \in \mathcal{G}} W_\Lambda(G) - W_\Lambda(\widehat{G})\right] &\geq \sup_{P_{\mathbf{b}} \in \mathcal{P}^*} E_{P^n_{\mathbf{b}}}\left[W_\Lambda(G^*_{\mathbf{b}}) - W_\Lambda(\widehat{G})\right] \\
&\geq \int_{\mathbf{b}} E_{P^n_{\mathbf{b}}}\left[W_\Lambda(G^*_{\mathbf{b}}) - W_\Lambda(\widehat{G})\right] d\pi(\mathbf{b}) \\
&= \int_{\mathbf{b}} E_{P^n_{\mathbf{b}}}\left[\Lambda(1-\mu^*) - \Lambda(1-\mu_{\widehat{G}})\right] d\pi(\mathbf{b}) \\
&\geq \int_{\mathbf{b}} E_{P^n_{\mathbf{b}}}\left[|\Lambda'(1-\mu_{\widehat{G}})|(\mu^* - \mu_{\widehat{G}})\right] d\pi(\mathbf{b}) \\
&\geq |\Lambda'(\tau^*)| \int_{\mathbf{b}} E_{P^n_{\mathbf{b}}}\left[\mu^* - \mu_{\widehat{G}}\right] d\pi(\mathbf{b}), \qquad \text{(D.15)}
\end{aligned}$$

where the fourth line follows since $\Lambda(\cdot)$ is convex and non-increasing. The fifth line follows from the observation that for all $P \in \mathcal{P}^*$, $\mu_G \geq 1 - \tau^*$ for any treatment rule G, therefore $|\Lambda'(1-\mu_{\widehat{G}})| \geq |\Lambda'(\tau^*)|$.

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 (2018a) considers bounding a similar quantity, though the current construction of $\mathcal{P}^*$ is different from the construction in that paper. Therefore, in what follows, we reproduce the proof of Theorem 2.2 in Kitagawa and Tetenov (2018a) with some necessary modifications.

Consider

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

$$\geq \inf_{\widehat{G}} \gamma \int_{\mathbf{b}} \int_{Z_1,\ldots,Z_n} P_X\left(\{b(X) \neq \hat{b}(X)\}\right) dP^n\left(Z_1,\ldots,Z_n|\mathbf{b}\right) d\pi(\mathbf{b})$$

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

$$\begin{aligned} & \gamma \int_{Z_1,\ldots,Z_n} E_X\left[\min\left\{\pi\left(b(X)=1|Z_1,\ldots,Z_n\right), 1-\pi\left(b(X)=1|Z_1,\ldots,Z_n\right)\right\}\right] d\tilde{P}^n \\ = & \quad \gamma \int_{Z_1,\ldots,Z_n} \sum_{j=1}^{v-1} p\left[\min\left\{\pi\left(b_j=1|Z_1,\ldots,Z_n\right), 1-\pi(b_j=1|Z_1,\ldots,Z_n)\right\}\right] d\tilde{P}^n, \end{aligned} \tag{D.16}$$

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,

$$\begin{aligned} & \min\left\{\pi\left(b_j=1|Z_1,\ldots,Z_n\right), 1-\pi(b_j=1|Z_1,\ldots,Z_n)\right\} \\ = & \frac{\min\left\{\left(\frac{1+\gamma}{2}\right)^{k_j^+}\left(\frac{1-\gamma}{2}\right)^{k_j^-}, \left(\frac{1+\gamma}{2}\right)^{k_j^-}\left(\frac{1-\gamma}{2}\right)^{k_j^+}\right\}}{\left(\frac{1+\gamma}{2}\right)^{k_j^+}\left(\frac{1-\gamma}{2}\right)^{k_j^-}+\left(\frac{1+\gamma}{2}\right)^{k_j^-}\left(\frac{1-\gamma}{2}\right)^{k_j^+}} \\ = & \frac{\min\left\{1, \left(\frac{1+\gamma}{1-\gamma}\right)^{k_j^+-k_j^-}\right\}}{1+\left(\frac{1+\gamma}{1-\gamma}\right)^{k_j^+-k_j^-}} \end{aligned}$$

$$= \frac{1}{1+a^{|k_j^+-k_j^-|}}, \text{ where } a = \frac{1+\gamma}{1-\gamma} > 1. \tag{D.17}$$

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

$$\tilde{Y}_i = \begin{cases} 1 & \text{if } Y_i D_i + (1-Y_i)(1-D_i) = 1, \\ -1 & \text{otherwise.} \end{cases} \tag{D.18}$$

Since $k_j^+ - k_j^- = \sum_{i:X_i=x_j} \tilde{Y}_i$, plugging (D.17) into (D.16) 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 inequality follows by $a > 1$, and the third inequality 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,

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

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

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

Hence, the Bayes risk (D.16) is bounded from below by

$$\begin{aligned} & \frac{\gamma}{2} p(v-1) a^{-\sqrt{np}} \\ \geq\ & \frac{\gamma}{2} p(v-1) e^{-(a-1)\sqrt{np}} \quad (\because 1 + x \leq e^x \ \forall x) \\ =\ & \frac{p\gamma}{2}(v-1) e^{-\frac{2\gamma}{1-\gamma}\sqrt{np}}, \end{aligned} \tag{D.19}$$

therefore

$$\int_{\mathbf{b}} E_{P^n_{\mathbf{b}}} \left[\mu^* - \mu_{\widehat{G}}\right] d\pi(\mathbf{b}) \geq \frac{p\gamma}{2}(v-1)e^{-\frac{2\gamma}{1-\gamma}\sqrt{np}}. \tag{D.20}$$

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 $\gamma = \sqrt{\frac{v-1}{n\tau^*}}$. Then for all $n \geq 4(v-1)/\tau^*$, $\gamma \leq 1/2$ and since $p = \frac{\tau^*}{v-1}$,

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

Then

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

Inserting this bound into (D.20) and multiplying by Υ provides a lower bound for (D.15). This completes the proof. □

Proof of Proposition 4.1. Similarly to inequality (A.15) shown in Appendix A, the average welfare regret of the capacity-constrained estimated policy satisfies

$$E_{P^n}\left[\sup_{G\in\mathcal{G}} W^K_{\Lambda}(G) - W^K_{\Lambda}(\widehat{G}^K)\right] \leq 2|\Lambda'(0)| \int_0^\infty E_{P^n}\left[\sup_{G\in\mathcal{G}}\left|\widehat{F}^K_G(y) - F^K_G(y)\right|\right] dy. \tag{D.21}$$

We hence focus on bounding $E_{P^n}\left[\sup_{G\in\mathcal{G}}\left|\widehat{F}^K_G(y) - F^K_G(y)\right|\right]$.

Expressing $\widehat{F}^K_G(y)$ and $F^K_G(y)$ as

$$\widehat{F}^K_G(y) = 1 - \frac{1}{n}\sum_{i=1}^n \widehat{w}^K_{G,i} \cdot 1\{Y_i > y\},$$
$$F^K_G(y) = 1 - E_P[w^K_G(Z) \cdot 1\{Y > y\}],$$

where

$$w^K_G(Z) = \frac{1-D}{1-e(X)} + \min\left\{1, \frac{K}{P_X(G)}\right\}\tilde{w}_G(Z),$$
$$\widehat{w}^K_{G,i} = \frac{1-D_i}{1-e(X_i)} + \min\left\{1, \frac{K}{P_{X,n}(G)}\right\}\tilde{w}_G(Z_i),$$

where $\tilde{w}_G(Z_i) = \left[\frac{D_i}{e(X_i)} - \frac{1-D_i}{1-e(X_i)}\right] \cdot 1\{X_i \in G\}$. Note that $\|\tilde{w}_G\|_\infty \leq \kappa^{-1}$. Define

$$\tilde{F}_G^K(y) = 1 - \frac{1}{n}\sum_{i=1}^n w_G^K(Z_i) \cdot 1\{Y_i > y\}.$$

We consider

$$\sup_{G\in\mathcal{G}} \left|\widehat{F}_G^K(y) - F_G^K(y)\right| \leq \underbrace{\sup_{G\in\mathcal{G}} \left|\widehat{F}_G^K(y) - \tilde{F}_G^K(y)\right|}_{\text{(iv)}} + \underbrace{\sup_{G\in\mathcal{G}} \left|\tilde{F}_G^K(y) - F_G^K(y)\right|}_{\text{(v)}}, \tag{D.22}$$

and derive bounds for $\int_0^\infty E_{P^n}[\text{(iv)}]dy$ and $\int_0^\infty E_{P^n}[\text{(v)}]dy$.

For term (iv), we have

$$\begin{aligned}
\left|\widehat{F}_G^K(y) - \tilde{F}_G^K(y)\right| &\leq \frac{1}{n}\sum_{i=1}^n \left|\widehat{w}_{G,i}^K - w_G^K(Z_i)\right| \cdot 1\{Y_i > y\} \\
&\leq \kappa^{-1}\left|\frac{K}{\max\{K, P_{X,n}(G)\}} - \frac{K}{\max\{K, P_X(G)\}}\right| \cdot \frac{1}{n}\sum_{i=1}^n 1\{Y_i > y\} \\
&\leq \frac{1}{\kappa K}\left|P_{X,n}(G) - P_X(G)\right| \cdot \frac{1}{n}\sum_{i=1}^n 1\{Y_i > y\}. \\
&= \frac{1}{\kappa K}\left|P_{X,n}(G) - P_X(G)\right| \cdot P(Y > y) \\
&\quad + \frac{1}{\kappa K}\left|P_{X,n}(G) - P_X(G)\right| \cdot \frac{1}{n}\sum_{i=1}^n \left[1\{Y_i > y\} - P(Y > y)\right].
\end{aligned} \tag{D.23}$$

Note that by Lemma A.4, $E_{P^n}\left[\sup_{G\in\mathcal{G}} |P_{X,n}(G) - P_X(G)|\right] \leq C_1\sqrt{v/n}$. By the Cauchy-Schwarz inequality,

$$\begin{aligned}
&E_{P^n}\left[\sup_{G\in\mathcal{G}} |P_{X,n}(G) - P_X(G)| \cdot \frac{1}{n}\sum_{i=1}^n \left[1\{Y_i > y\} - P(Y > y)\right]\right] \\
\leq&\sqrt{E_{P^n}\left[\left(\sup_{G\in\mathcal{G}} |P_{X,n}(G) - P_X(G)|\right)^2\right]} \cdot \sqrt{\frac{P(Y > y)(1 - P(Y > y))}{n}} \\
\leq&\sqrt{\frac{P(Y > y)}{n}},
\end{aligned} \tag{D.24}$$

where the second inequality follows from

$$\sup_{G\in\mathcal{G}} |P_{X,n}(G) - P_X(G)|^2 \leq \frac{1}{n}\sum_{i=1}^n (1\{X_i \in G\} - P_X(G))^2 \leq 1.$$

Hence, by noting $1 \leq \sqrt{v}$,

$$\int_0^\infty E_{P^n}[(\text{iv})] \leq \frac{2\,(C_1+1)\,\Upsilon}{\kappa K} \cdot \sqrt{\frac{v}{n}} \tag{D.25}$$

Next, consider term (v). Let $\tilde{F}_\emptyset^K(y) = 1 - n^{-1}\sum_{i=1}^n \frac{1-D_i}{1-e(X_i)} \cdot 1\{Y_i > y\}$. We decompose term (v) as follows:

$$\begin{aligned}
\left|\tilde{F}_G^K(y) - F_G^K(y)\right| \leq& \left|(\tilde{F}_G^K(y) - \tilde{F}_\emptyset^K(y)) - (F_G^K(y) - F_\emptyset^K(y))\right| + |\tilde{F}_\emptyset^K(y) - F_\emptyset^K(y)| \\
=& \min\left\{1, \frac{K}{P_X(G)}\right\}\left|\frac{1}{n}\sum_{i=1}^n \tilde{w}_G(Z_i)\cdot 1\{Y_i > y\} - E_P[\tilde{w}_G(Z)\cdot 1\{Y > y\}]\right| \\
&+ |\tilde{F}_\emptyset^K(y) - F_\emptyset^K(y)| \\
\leq& \left|\frac{1}{n}\sum_{i=1}^n \tilde{w}_G(Z_i)\cdot 1\{Y_i > y\} - E_P[\tilde{w}_G(Z)\cdot 1\{Y > y\}]\right| + |\tilde{F}_\emptyset^K(y) - F_\emptyset^K(y)|.
\end{aligned} \tag{D.26}$$

Hence,

$$\begin{aligned}
&\int_0^\infty E_{P^n}\left[\sup_{G\in\mathcal{G}}\left|\tilde{F}_G^K(y) - F_G^K(y)\right|\right] \\
\leq& \int_0^\infty E_{P^n}\left[\sup_{G\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^n \tilde{w}_G(Z_i)\cdot 1\{Y_i > y\} - E_P[\tilde{w}_G(Z)\cdot 1\{Y > y\}]\right|\right] dy \\
&+ \int_0^\infty E_{P^n}\left[|\tilde{F}_\emptyset^K(y) - F_\emptyset^K(y)|\right]\, dy \\
\leq& 2C_T\frac{\Upsilon}{\kappa}\sqrt{\frac{v}{n}} + \int_0^\infty E_{P^n}\left[|\tilde{F}_\emptyset^K(y) - F_\emptyset^K(y)|\right]\, dy,
\end{aligned} \tag{D.27}$$

where the second inequality follows from Lemma A.5 with $M = 2\Upsilon$ and $\bar{F} = \kappa^{-1}$, where C_T is the universal constant defined there.

To bound the second term in (D.27),

$$\begin{aligned}
\int_0^\infty E_{P^n}\left[|\tilde{F}_\emptyset^K(y) - F_\emptyset^K(y)|\right] dy \leq& \int_0^\infty \sqrt{Var\left(\frac{1}{n}\sum_{i=1}^n \frac{1-D_i}{1-e(X_i)}\cdot 1\{Y_{0i} > y\}\right)}dy \\
=& \int_0^\infty \sqrt{\frac{1}{n}\left\{E_P\left[\left(\frac{1}{1-e(X)}\right)P(Y_0 > y|X)\right] - P(Y_0 > y)^2\right\}} \\
\leq& \int_0^\infty \sqrt{\frac{1}{n}E_P\left[\left(\frac{1}{1-e(X)}\right)P(Y_0 > y|X)\right]}
\end{aligned}$$

$$\leq \frac{1}{\kappa\sqrt{n}} \int_0^\infty \sqrt{P(Y_0 > y)}dy \leq \frac{\Upsilon}{\kappa}\sqrt{\frac{v}{n}}. \tag{D.28}$$

Combining (D.21), (D.22), (D.25), (D.27), and (D.28), and noting $1 \leq \sqrt{v}$, we conclude

$$E_{P^n}\left[\sup_{G\in\mathcal{G}} W_\Lambda^K(G) - W_\Lambda^K(\widehat{G}^K)\right] \leq \left(\frac{C_{K1}}{K} + C_{K2}\right)|\Lambda'(0)|\frac{\Upsilon}{\kappa}\sqrt{\frac{v}{n}},$$

where $C_{K1} = 4(C_1 + 1)$ and $C_{K2} = 2(2C_T + 1)$. □

Proof of Theorem C.1. For any $G \in \mathcal{G}$, it holds

$$\begin{aligned} W_\Lambda(G) - W_\Lambda(\widehat{G}^e) &\leq \widehat{W}_\Lambda(G) - \widehat{W}_\Lambda^e(G) - \widehat{W}_\Lambda(\widehat{G}^e) + \widehat{W}_\Lambda^e(\widehat{G}^e) \\ &\quad + W_\Lambda(G) - W_\Lambda(\widehat{G}^e) - \widehat{W}_\Lambda(G) + \widehat{W}_\Lambda(\widehat{G}^e) \\ &\leq 2\sup_{G\in\mathcal{G}} |\widehat{W}_\Lambda(G) - \widehat{W}_\Lambda^e(G)| + 2\sup_{G\in\mathcal{G}} \left|\widehat{W}_\Lambda(G) - W_\Lambda(G)\right|, \end{aligned} \tag{D.29}$$

where the first inequality uses $\widehat{W}_\Lambda^e(\widehat{G}^e) - \widehat{W}_\Lambda^e(G) \geq 0$. The mean of the second term in the right-hand side of (D.29) is $O(n^{-1/2})$ as shown in Theorem 3.1 (i).

For the first term in the right-hand side of (D.29), following the inequalities shown in (A.14), we have

$$|\widehat{W}_\Lambda(G) - \widehat{W}_\Lambda^e(G)| \leq |\Lambda'(0)| \int_0^\infty |\widehat{F}_G(y) - \widehat{F}_G^e(y)|dy. \tag{D.30}$$

For every y, the upper bound of $|\widehat{F}_G(y) - \widehat{F}_G^e(y)|$ uniform in G can be obtained as

$$\begin{aligned} &|\widehat{F}_G(y) - \widehat{F}_G^e(y)| \\ \leq &\frac{1}{n}\sum_{i=1}^n \left|\frac{1}{e(X_i)} - \frac{1}{\hat{e}(X_i)}\right| D_i 1\{Y_i > y\}1\{X_i \in G\} \\ &+ \frac{1}{n}\sum_{i=1}^n \left|\frac{1}{1-e(X_i)} - \frac{1}{1-\hat{e}(X_i)}\right| (1-D_i)1\{Y_i > y\}1\{X_i \notin G\} \\ \leq &\frac{1}{n}\sum_{i=1}^n \left|\frac{1}{e(X_i)} - \frac{1}{\hat{e}(X_i)}\right| \cdot 1\{Y_{1i} > y\} + \frac{1}{n}\sum_{i=1}^n \left|\frac{1}{1-e(X_i)} - \frac{1}{1-\hat{e}(X_i)}\right| \cdot 1\{Y_{0i} > y\} \\ = &\underbrace{\frac{1}{n}\sum_{i=1}^n \left|\frac{1}{e(X_i)} - \frac{1}{\hat{e}(X_i)}\right| \cdot P(Y_{1i} > y)}_{\text{(vi)}} + \underbrace{\frac{1}{n}\sum_{i=1}^n \left|\frac{1}{e(X_i)} - \frac{1}{\hat{e}(X_i)}\right| \cdot [1\{Y_{1i} > y\} - P(Y_{1i} > y)]}_{\text{(vii)}} \end{aligned}$$

$$+ \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left|\frac{1}{1-e(X_i)} - \frac{1}{1-\hat{e}(X_i)}\right| \cdot P(Y_{0i} > y)}_{\text{(viii)}}$$

$$+ \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left|\frac{1}{1-e(X_i)} - \frac{1}{1-\hat{e}(X_i)}\right| \cdot [1\{Y_{0i} > y\} - P(Y_{0i} > y)]}_{\text{(ix)}}. \tag{D.31}$$

We derive the convergence rates of the integrated means of terms (vi) - (ix) in (D.31), separately; by Assumption C.1,

$$\int_0^\infty E_{P^n}[(\text{vi})]dy \le E_{P^n}\left[\frac{1}{n}\sum_{i=1}^{n}\left|\frac{1}{e(X_i)} - \frac{1}{\hat{e}(X_i)}\right|\right] \cdot \Upsilon = O(\phi_n^{-1}).$$

$$\begin{aligned}\int_0^\infty E_{P^n}[(\text{vii})]dy &\le \int_0^\infty E_{P^n}\left[\max_{1\le i\le n}\left|\frac{1}{e(X_i)} - \frac{1}{\hat{e}(X_i)}\right| \cdot \frac{1}{n}\sum_{i=1}^{n}[1\{Y_{1i} > y\} - P(Y_{1i} > y)]\right] dy \\ &\le \sqrt{E_{P^n}\left[\left(\max_{1\le i\le n}\left|\frac{1}{e(X_i)} - \frac{1}{\hat{e}(X_i)}\right|\right)^2\right]} \int_0^\infty \sqrt{\frac{P(Y_{1i} > y)(1 - P(Y_{1i} > y))}{n}} dy \\ &\le O(\tilde{\phi}_n^{-1}) \cdot \frac{\Upsilon}{\sqrt{n}} = O\left(\tilde{\phi}_n^{-1}/\sqrt{n}\right).\end{aligned}$$

Similarly, we obtain $\int_0^\infty E_{P^n}[(\text{viii})]dy \le O(\phi_n^{-1})$ and $\int_0^\infty E_{P^n}[(\text{ix})]dy \le O\left(\tilde{\phi}_n^{-1}/\sqrt{n}\right)$.

These convergence rates for terms (vi) - (ix) and (D.30) imply that

$$E_{P^n}\left[\sup_{G\in\mathcal{G}}\left|\widehat{W}_\Lambda(G) - \widehat{W}_\Lambda^e(G)\right|\right] = O\left(\phi_n^{-1} + \frac{\tilde{\phi}_n^{-1}}{\sqrt{n}}\right)$$

Hence, by (D.29) and noting that $\tilde{\phi}_n^{-1} n^{-1/2}$ converges faster than $n^{-1/2}$, we conclude

$$E_{P^n}\left[\sup_{G\in\mathcal{G}} W_\Lambda(G) - W_\Lambda(\widehat{G}^e)\right] \le O\left((\phi_n^{-1} + \tilde{\phi}_n^{-1} n^{-1/2}) \vee n^{-1/2}\right) = O\left(\phi_n^{-1} \vee n^{-1/2}\right).$$

□

E 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 C.1 with an explicit characterization of ϕ_n and $\tilde{\phi}_n$.

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 the leave-one-out local polynomial fit for $e(X_i)$ by

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

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 C.1, we impose the following restrictions, which are identical to Assumption E.2 in Kitagawa and Tetenov (2018b).

Assumption E.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 $L_e < \infty$. [1]

[1]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 $|f_x(x') - f(x)| \leq L \|x - x'\|^{\beta}$.

(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 $\mathbb{R}^{d_x}$ and $B(x,r)$ be the open ball centered at $x \in \mathbb{R}^{d_x}$ with radius r. There exist constants $\underline{c}$ and r_0 such that

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

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 $K(\cdot)$ has support $[-1,1]^{d_x}$, $\int_{\mathbb{R}^{d_x}} K(u)du = 1$, and $\sup_u K(u) \leq K_{\max} < \infty$.

Assumption E.1 (PX) is borrowed from Audibert and Tsybakov (2007), and it provides regularity conditions on the marginal distribution of X. Inequality condition (E.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 on 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 (C.4). These claims are borrowed from Theorem 3.2 in Audibert and Tsybakov (2007) and Lemma E.4 in Kitagawa and Tetenov (2018b).

Lemma E.1. *Let $\mathcal{P}_e$ consist of the data generating processes satisfying Assumption E.1 (Smooth-e) and (PX). Let $\tilde{e}(X_i)$ be the leave-one-out estimator for the propensity score defined in (E.1) whose kernel function satisfies E.1 (Ker).*

(i) There exist positive constants c_2, c_3, and c_4 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 $c_4 h^{\beta_e} < \delta$, and any $n \geq 2$,

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

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)

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

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

$$\sup_{P\in\mathcal{P}_e}\int_{\mathcal{X}} E_{P^{n-1}}\left[\left|\tilde{e}(x)-e\left(x\right)\right|\right]dP_X(x) \le O\left(n^{-\frac{1}{2+d_x/\beta_e}}\right). \tag{E.3}$$

(iii)

$$\sup_{P\in\mathcal{P}_e} E_{P^n}\left[\left(\max_{1\le i\le n}\left|\tilde{e}(X_i)-e\left(X_i\right)\right|\right)^2\right] \le O\left(\frac{h^{2\beta_e}}{t_n^2}\right)+O\left(\frac{\log n}{nh^{d_x}t_n^2}\right)$$

holds. In particular, when the bandwidth is chosen as in claim (ii) of the current proposition, the resulting uniform convergence rate is

$$\sup_{P\in\mathcal{P}_e} E_{P^n}\left[\left(\max_{1\le i\le n}\left|\tilde{e}(X_i)-e\left(X_i\right)\right|\right)^2\right] \le O\left(t_n^{-2}\log n\cdot n^{-\frac{2}{2+d_x/\beta_e}}\right). \tag{E.4}$$

Making use of Lemma E.1, the next proposition shows a propensity score estimator constructed by suitably trimming $\tilde{e}(X_i)$ satisfies Assumption C.1 with an explicit characterization of the growing sequences ϕ_n and $\tilde{\phi}_n$.

Proposition E.1. *Let* $\mathcal{P}_e$ *consist of data generating processes that satisfy Assumption E.1 (Smooth-e) and (PX). Let* $\tilde{e}(X_i)$ *be the leave-one-out local polynomial estimator with degree* $l=(\beta_e-1)$*, trimming sequence for the least eigenvalue* $t_n=(\log n)^{-1}$*, bandwidth sequence* $h\propto n^{-\frac{1}{2\beta_e+d_x}}$*, and whose kernel satisfies Assumption E.1 (Ker). Let*

$$\hat{e}(X_i)\equiv\min\left\{1-\epsilon_n,\max\{\epsilon_n,\tilde{e}(X_i)\}\right\}\in[\epsilon_n,1-\epsilon_n] \tag{E.5}$$

with a sequence of trimming constants ϵ_n *that satisfies* $\epsilon_n=O(n^{-a})$ *for some* $a>0$*. Then,* $\hat{e}(X_i)$ *satisfies Assumption C.1 with* $\phi_n=n^{\frac{1}{2+d_x/\beta_e}}$ *and* $\tilde{\phi}_n=(\log n)^{-3/2}\cdot n^{\frac{1}{2+d_x/\beta_e}}$.

Proof of Proposition E.1. Assume that n is large enough so that $\varepsilon_n\le\kappa/2$ 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

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

Hence,

$$E_{P^n}\left[\frac{1}{n}\sum_{i=1}^{n}\left|\frac{1}{e(X_i)}-\frac{1}{\hat{e}(X_i)}\right|\right]=E_{P^n}\left[\left|\frac{1}{e(X_n)}-\frac{1}{\hat{e}(X_n)}\right|\right]$$

$$\leq \frac{2}{\kappa^2} E_{P^n} |\tilde{e}(X_n) - e(X_n)| + (\kappa\epsilon_n)^{-1} P^n \left(\tilde{e}(X_n) \notin \left[\frac{\kappa}{2}, 1 - \frac{\kappa}{2}\right]\right) \tag{E.6}$$

By Lemma E.1 (ii),

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

Furthermore, by Lemma E.1 (i),

$$\begin{aligned} P^n\left(\tilde{e}\left(X_n\right) \notin \left[\frac{\kappa}{2}, 1 - \frac{\kappa}{2}\right]\right) &= \int_{\mathcal{X}} P^{n-1}\left(\tilde{e}\left(x\right) \notin \left[\frac{\kappa}{2}, 1 - \frac{\kappa}{2}\right]\right) dP_X\left(x\right) \\ &\leq \int_{\mathcal{X}} P^{n-1}\left(|\tilde{e}\left(x\right) - e(x)| > \frac{\kappa}{2}\right) dP_X\left(x\right) \\ &\leq c_2 \exp\left(-\frac{c_3\kappa^2}{4} n h^{d_x}\right) \end{aligned} \tag{E.7}$$

holds for all n satisfying $c_4 h^{\beta_e} < \kappa/2$, where c_2, c_3, and c_4 are the constants defined in Lemma E.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 (E.6), we conclude $\sup\limits_{P \in \mathcal{P}_e} E_{P^n}\left[\frac{1}{n}\sum_{i=1}^n |\hat{e}(X_i) - e(X_i)|\right] \leq O\left(n^{-\frac{1}{2+d_x/\beta_e}}\right)$, i.e., $\phi_n = n^{\frac{1}{2+d_x/\beta_e}}$.

For the bounds for the mean of the squared maximum, we have

$$\begin{aligned} &E_{P^n}\left[\left(\max_{1\leq i\leq n}\left|\frac{1}{e(X_i)} - \frac{1}{\hat{e}(X_i)}\right|\right)^2\right] \\ \leq &\frac{4}{\kappa^4} E_{P^n}\left[\left(\max_{1\leq i\leq n} |\tilde{e}(X_n) - e(X_n)|\right)^2\right] + (\kappa\epsilon_n)^{-2} P^n\left(\tilde{e}(X_n) \notin \left[\frac{\kappa}{2}, 1 - \frac{\kappa}{2}\right]\right) \end{aligned}$$

By Lemma E.1 (iii) and (E.7), $E_{P^n}\left[\left(\max_{1\leq i\leq n}\left|\frac{1}{e(X_i)} - \frac{1}{\hat{e}(X_i)}\right|\right)^2\right] \leq O\left((\log n)^3 \cdot n^{-\frac{2}{2+d_x/\beta_e}}\right)$, i.e., $\tilde{\phi}_n = (\log n)^{-3/2} \cdot n^{\frac{1}{2+d_x/\beta_e}}$.

The other convergence rate bounds in Assumption C.1 can be shown similarly. □

Combining Proposition E.1 with Theorem C.1 proves the claim made in equation (C.4).

References

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

BOUCHERON, S., G. LUGOSI, AND P. MASSART (2013): *Concentration Inequalities, A Nonasymptotic Theory of Independence*, Oxford University Press.

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

KITAGAWA, T. AND A. TETENOV (2018a): "Who Should Be Treated? Empirical Welfare Maximization Methods for Treatment Choice," *Econometrica*, 86, 591–616.

——— (2018b): "Online Appendix to "Who Should Be Treated? Empirical Welfare Maximization Methods for Treatment Choice"," URL: https://www.econometricsociety.org/sites/default/files/13288_Data_and_Programs.zip.

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