Fair Policy Targeting Note

JCY

2023 年 1 月 23 日

1 What is Fair Policy Targeting

This paper designs fair and efficient targeting rules for applications in welfare and health programs. We construct treatment allocation rules using data from experiments or quasi-experiments, and we develop policies that trade-off efficiency and fairness.

This paper advocates for fair and Pareto optimal treatment rules. We discuss targeting in a setting where decision-makers prefer allocations for which we cannot find any other policy that strictly improves welfare for one of the two sensitive groups without decreasing welfare on the opposite group. Within such a set, she then chooses the fairest allocation.

Our framework has three desirable properties:

(i) it applies to general notions of fairness which may reflect different decision makers' preferences;

(ii) it guarantees Pareto efficiency of the policy-function, with the relative importance of each group solely chosen based on the notion of fairness adopted by the decision-maker;

(iii) it also allows for arbitrary legal or ethical constraints, incorporating as a special case the presence of fairness constraints whenever such constraints are binding due to ethical or legal considerations.

We name our method Fair Policy Targeting.

The decision problem consists of lexicographic preferences of the policymaker of the following form: (i) Pareto dominant allocations are preferred over dominated ones; (ii) Pareto optimal allocations are ranked based on fairness considerations. We identify the Pareto frontier as the set of maximizers over any weighted average of each group' s welfares. Therefore, such an approach embeds as a special case maximizing a weighted combination of welfares of each sensitive group

2 Decision Making and Fairness

For each unit, we denote with $S \in \mathcal{S}$ a sensitive or protected attribute. For expositional convenience, we let $\mathcal{S} = \{0, 1\}$, with $S = 1$ denoting the disadvantaged group, and $X \in \mathcal{X} \subseteq \mathbb{R}^p$ individual characteristics. We define the posttreatment outcome with $Y \in \mathcal{Y} \subseteq \mathbb{R}$ realized only once the sensitive attribute, covariates, and the treatment assignment are realized. We define $Y(d), d \in \{0, 1\}$ the

potential outcomes under treatment d . The observed Y satisfies the Single Unit Treatment Value Assumption (SUTVA) (Rubin, 1990). Let

$$e(x,s) = P(D = 1 \mid X = x, S = s), \quad p_1 = P(S = 1) \tag{1}$$

be the propensity score and the probability of being assigned to the disadvantaged group. Here, treatments are independent of potential outcomes.

Given observables, (Y_i, X_i, D_i, S_i) we seek to design a treatment assignment rule (i.e. policy function) $\pi : \mathcal{X} \times \mathcal{S} \mapsto \mathcal{T} \subseteq [0,1], \pi \in \Pi$ that depends on the individual characteristics and protected attributes.

Π incorporates given and binding legal or economic constraints that restrict the decision space. The welfare generated by a policy π on those individuals with sensitive attribute S=s is defined as

$$W_s(\pi) = \mathbb{E}[(Y(1) - Y(0))\pi(X,S) \mid S = s]. \tag{2}$$

we consider a framework where the policymaker simultaneously maximizes each group's welfare, imposing Pareto efficiency on the estimated policy, under arbitrary legal or economic constraints encoded in . Given the set of efficient policies, the planner then selects the least unfair one. Our approach is designed for social and welfare programs where legal constraints naturally occur and where, given such constraints, the policymaker's preferences align with classical notions of "first do no harm".

2.1 Pareto Principle for Treatment Rules

The set of Pareto optimal choices is defined as Π_o, and it contains all such allocations $\pi \in \Pi$ for which the welfare for one of the two groups cannot be improved without reducing the welfare for the opposite group.

Lemma 2.1 (Pareto Frontier). The set $\Pi_o \subseteq \Pi$ is such that

$$\Pi_o = \left\{ \pi_\alpha : \pi_\alpha \in \arg\sup_{\pi \in \Pi} \alpha W_1(\pi) + (1-\alpha) W_0(\pi), \quad \alpha \in (0,1) \right\}. \tag{3}$$

It will be convenient to define

$$\bar{W}_\alpha = \sup_{\pi \in \Pi} \alpha W_1(\pi) + (1-\alpha) W_0(\pi), \tag{4}$$

2.2 Decision Problem

Proposition 2.2 (Decision Problem). Under Assumption 2.2, $\pi^\star \in \mathcal{C}(\Pi)$ if and only if

$$\begin{gathered} \pi^\star \in \arg\inf_{\pi \in \Pi} \mathrm{UnFairness}(\pi) \\ \text{subject to } \alpha W_1(\pi) + (1-\alpha) W_0(\pi) \geq \bar{W}_\alpha, \text{ for some } \alpha \in (0,1) \end{gathered} \tag{5}$$

Proposition 2.2 formally characterizes the policymakers decision problem, which consists of minimizing the policy's unfairness criterion, under the condition that the policy is Pareto optimal. The

policy-maker does not maximize a weighted combination of welfares, with some pre-specified and hard-to-justify weights. Instead, each group's importance (i.e., α) is implicitly chosen within the optimization problem to maximize fairness.

2.2.1 Policy with Welfare Maximization

The population equivalent of the EWM problem belongs to the Pareto frontier. Namely, $\arg\max_{\pi\in\Pi}\{p_1 W_1(\pi)+(1-p_1)W_0(\pi)\}\subseteq\Pi_{\mathrm{o}}$. An alternative approach consists in maximizing weighted combinations of the welfare with the weights for each group as given. For instance the allocation (Rambachan et al., 2020)

$$\check{\pi}_\omega \in \arg\max_{\pi\in\Pi}\{\omega W_1(\pi)+(1-\omega)W_0(\pi)\}\subseteq\Pi_{\mathrm{o}}, \tag{6}$$

for some specific weight ω belongs to the Pareto frontier.

2.2.2 Policy with fairness constraints

Define $\Pi(\kappa)=\{\pi\in\Pi:\text{UnFairness}\,(\pi)\leq\kappa\}\subseteq\Pi$, the set of policies with constraint, and

$$\tilde{\pi}\in\arg\max_{\pi\in\Pi(\kappa)} p_1 W_1(\pi)+(1-p_1)W_0(\pi) \tag{7}$$

the policy that maximizes the welfare imposing fairness constraints .

2.2.3 Corollary 1 (Properties).

(1) $\text{UnFairness}\,(\pi^\star)\leq\text{UnFairness}\,(\check{\pi}_\omega)\,,\forall\omega\in(0,1)$.

(2) Suppose that either $\tilde{\pi}\in\Pi_o$ (i.e., it belongs to the Pareto frontier), or fairness constraints are binding to the policy-maker, i.e. $\Pi(\kappa)=\Pi$. Then $\text{UnFairness}\,(\pi^\star)\leq\text{UnFairness}(\widetilde{\pi})$.

(3) Suppose instead that $\widetilde{\pi}\notin\Pi_o$. Then $\text{UnFairness}\,(\pi^\star)\leq\text{UnFairness}\,(\pi_o)$ for all $\pi_o\in\Pi_o$ that Pareto dominate $\tilde{\pi}$.

(4) π_ω and $\tilde{\pi}$ do not Pareto dominate $\pi^\star$.

3 Fair Targeting: Estimation

How to construct an estimator of $\pi^\star$. We define

$$m_{d,s}(x)=\mathbb{E}\left[Y_i(d)\mid X_i=x,S_i=s\right],\quad \Gamma_{d,s,i}=\frac{1\{S_i=s\}}{p_s}\left[\frac{1\{D_i=d\}}{e\,(X_i,S_i)}\left(Y_i-m_{d,s}\,(X_i)\right)+m_{d,s}\,(X_i)\right] \tag{8}$$

the conditional mean of the group s under treatment d , and the doubly robust score (Robins and Rotnitzky, 1995), respectively.

We let $\hat{\Gamma}_{d,s,i}$ the estimated counterpart of $\Gamma_{d,s,i}$. Define

$$\hat{W}_s(\pi)=\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\Gamma}_{1,s,i}-\hat{\Gamma}_{0,s,i}\right)\pi\,(X_i,s) \tag{9}$$

the estimated welfare built upon semi-parametric literature with $\hat{m}_{d,s}(.)$, $\hat{e}(.)$, $\hat{p}_s$, constructed via cross-fitting.

3.1 (Approximate) Pareto Optimality

We characterize the Pareto frontier using linear inequalities. To construct the Pareto frontier we use the constraint in Equation (5) after discretizing the set of weights α . Namely, in the first step, we discretize the Pareto frontier, and construct a grid of equally spaced values $\alpha_j \in (0,1), j \in \{1,\ldots,N\}$, with $N = \sqrt{n}$. We approximate the Pareto frontier using the set $\left(\hat{W}_0, \hat{W}_1\right.$ are defined in Equation (17)

$$\widehat{\Pi}_o = \left\{ \pi_\alpha \in \Pi : \pi_\alpha \in \arg\sup_{\pi\in\Pi} \left\{ \alpha \hat{W}_0(\pi) + (1-\alpha)\hat{W}_1(\pi) \right\}, \text{ s.t. } \alpha \in \{\alpha_1,\ldots,\alpha_N\} \right\}. \tag{10}$$

Instead of directly estimating $\widehat{\Pi}_{\mathrm{o}}$, we characterize it through linear constraints.

(2) We find the largest empirical welfare achieved on the discretized Pareto Frontier defined as

$$\bar{W}_{j,n} = \sup_{\pi\in\Pi} \left\{ \alpha_j \hat{W}_0(\pi) + (1-\alpha_j)\, \hat{W}_1(\pi) \right\}, \text{ for each } j \in \{1,\ldots,N\}, \tag{11}$$

which can be obtained through standard optimization routines.

(2) We observe that any $\pi \in \widehat{\Pi}_o$, must satisfy $\alpha_j \hat{W}_0(\pi) + (1-\alpha_j)\, \hat{W}_1(\pi) \geq \bar{W}_{j,n}$, for some $j \in \{1,\ldots,N\}$, since$\bar{W}_{j,n}$ defines the largest objective for a given α_j . We impose such constraint up to a small slackness parameters $\lambda/\sqrt{n}$ and construct an approximate Pareto frontier as follows:

$$\widehat{\Pi}_o(\lambda) = \left\{ \pi \in \Pi : \exists j \in \{1,\ldots,N\} \; suchthat \; \alpha_j \hat{W}_{0,n}(\pi) + (1-\alpha_j)\, \hat{W}_{1,n}(\pi) \geq \bar{W}_{j,n} - \frac{\lambda}{\sqrt{n}} \right\}, \tag{12}$$

where $\widehat{\Pi}_o(0) = \widehat{\Pi}_o$, and $\widehat{\Pi}_o \subseteq \widehat{\Pi}_o(\lambda)$ for any $\lambda \geq 0$.

The estimated policy is defined as

$$\hat{\pi}_\lambda \in \arg\min_{\pi\in\hat{\Pi}_{\mathrm{o}}(\lambda)} \widehat{\mathcal{V}}_n(\pi) \tag{13}$$

3.2 Optimization: Mixed Integer Quadratic Program

We provide a mixed-integer quadratic program (MIQP) for optimization. We define $\mathbf{z}_s = (z_{s,1},\cdots,z_{s,n}), z_{s,i} = \pi(X_i,s), \pi \in \Pi$. Here, $z_{s,n}$ defines the treatment assignment under policy π and sensititive attribute s (see the example below); $\mathbf{z}_s$ have simple representation for general classes of policy functions.

(Maximum score). For the maximum score $\pi(X_i,s) = 1\{X_i\beta_x + S\mu \geq 0\}, \beta = (\beta_x,\mu) \in \mathcal{B}$, the indicators $z_{s,n}$ are defined via mixed-integer constraints of the form (Florios and Skouras, 2008) $\frac{X_i^\top\beta+s\mu}{|C_i|} < z_{s,i} \leq \frac{X_i^\top\beta+s\mu}{|C_i|} + 1, C_i \geq \sup_{\beta\in\mathcal{B}} \left|(X_i,S_i)^\top\beta\right|, z_{s,i} \in \{0,1\}$. Such constraint guarantees that $z_{s,i} = 1\left\{X_i^\top\beta_x + s\mu \geq 0\right\}$.

$\hat{\pi}_\lambda$ satisfies $\hat{\pi}_\lambda \in \arg\min_{\pi\in\hat{\Pi}_{\mathrm{o}}(\lambda)} \widehat{\mathcal{V}}_n(\pi)$ if and only if

$$\hat{\pi}_\lambda \in \arg\min_{\pi} \min_{\mathbf{z}_0,\mathbf{z}_1,\mathbf{u}} \widehat{\mathcal{V}}_n(\pi) \tag{14}$$

$$
\begin{aligned}
\text{subject to} \quad & z_{s,i} = \pi\left(X_i, s\right), \quad 1 \leq i \leq n, (A) \\
& u_j \alpha_j \left\langle \hat{\mathbf{\Gamma}}_{1,0} - \hat{\mathbf{\Gamma}}_{0,0}, \boldsymbol{z}_0 \right\rangle + u_j \left(1 - \alpha_j\right) \left\langle \hat{\mathbf{\Gamma}}_{1,1} - \hat{\mathbf{\Gamma}}_{0,1}, \boldsymbol{z}_1 \right\rangle \geq u_j n \bar{W}_{j,n} - \sqrt{n}\lambda (B) \\
& \langle \mathbf{1}, \mathbf{u} \rangle \geq 1 (C) \\
& \pi \in \Pi (D) \\
& u_j \in \{0, 1\}, \quad 1 \leq j \leq N (E).
\end{aligned} \tag{15}
$$

Constraints (B) and (C) state that the resulting policy is (approximately) Pareto optimal, or, equivalently, it maximizes a weighted combination of groups' welfare for some α_j.

```
## Estimate the Pareto frontier with MILP
## argument: as before
## output: list

estimate_Pareto_frontier <- function(Y, X, D, S, propensity1,
   propensity2, scale_Y = T,
discretization = floor(sqrt(length(Y))),  cost_treatment = 0, params=NA,
max_treated_units = length(Y),
numcores = 10, maxtime = 300, alpha_seq = seq(from = 0, to = 1, length =
    discretization),
m1 = 0, m0 = 0, additional_fairness_constraint = F,
parity_constraint = '>=', probabilistic = F,
threshold_probabilistic = F, parallel = T, tolerance = 10**(-3)){
        library(foreach)
                res <- foreach(i = alpha_seq, .combine = rbind, .export
                   = c('Y', 'D', 'X', 'S', 'propensity1',
                'propensity2', 'params', 'scale_Y',
                'cost_treatment', 'max_treated_units', 'method',
                'maxtime', 'm1', 'm0', 'additional_fairness_constraint',
                'parity_constraint'))%do%{
                        source('./library/helpers.R')
                        library(gurobi)
                        if(probabilistic == F & threshold_probabilistic
                           == F){
                                result <- Est_max_score(Y, X, D, S,
                                   propensity1, propensity2,B=1, params
                                   = params, tolerance_constraint =
                                   tolerance,
                                scale_Y = scale_Y,
                                   additional_fairness_constraint =
                                   additional_fairness_constraint,
```

```
                parity_constraint = parity_constraint,
                   cost_treatment  = cost_treatment,
                   alpha = i,
                max_treated_units = max_treated_units,
                   maxtime = maxtime, m1 = m1, m0 = m0,
                   cores = numcores) } else if (
                   probabilistic == T &
                   threshold_probabilistic == F) {
                result <- Est_max_score_probabilistic(Y,
                    X, D, S, propensity1, propensity2,B
                   =1, params = params,
                   tolerance_constraint =tolerance,
                scale_Y = scale_Y,
                   additional_fairness_constraint =
                   additional_fairness_constraint,
                parity_constraint = parity_constraint,
                cost_treatment  = cost_treatment, alpha
                   = i, max_treated_units =
                   max_treated_units, maxtime = maxtime,
                    m1 = m1, m0 = m0,
                cores = numcores)
            } else {   result <- Est_threshold_probabilistic
               (Y, X, D, S, propensity1, propensity2,B=1,
               params = params, tolerance_constraint =
               tolerance,
                     scale_Y = scale_Y,
                        additional_fairness_constraint =
                        additional_fairness_constraint,
                     parity_constraint = parity_constraint,
                     cost_treatment  = cost_treatment, alpha
                        = i, max_treated_units =
                        max_treated_units, maxtime = maxtime,
                         m1 = m1, m0 = m0, cores = numcores)

            }
            c(result[[2]], result[[1]],
            result[[3]], result[[4]], length(result[[2]]),
            length(result[[3]]))
```

```
                }

        if(threshold_probabilistic == F){
                nn <- res[1, dim(res)[2] - 1]
                nn2 <- res[1, dim(res)[2]]
                res <- res[, -dim(res)[2]]
                aa1 <- res[, 1:nn]
                aa2 <- res[, nn + 1]
                return(list(g_i = res[, 1:nn], objective = res[, nn +
                   1], results = res[, c((nn + 2):(nn + 1 + nn2))],
                policies = res[, c((nn + 2 + nn2):(dim(res)[2]-1))],
                   beta = res[, c((nn + 2 + nn):(nn + 1 + nn2))]))
        } else {
                nn <- res[1, dim(res)[2] - 1]
                nn2 <- res[1, dim(res)[2]]
                res <- res[, -dim(res)[2]]
                aa1 <- res[, 1:nn]
                aa2 <- res[, nn + 1]
                return(list(g_i = res[, 1:nn], objective = res[, nn +
                   1], results = res[, c((nn + 2):(nn + 1 + nn2))],
                policies = res[, c((nn + 2 + nn2):(dim(res)[2]-1))],
                   beta = res[, c((nn + 1 + nn + 1):(2 * nn + 1 + dim(X)
                   [2] + 1 ))],
                probs = res[, c((2 * nn + 2 + dim(X)[2] + 1 ):(2 * nn +
                   2 + dim(X)[2] + 2 ) )]
                ))
        }
}
```

4 Counterfactual UnFairness

We discuss a novel notion of UnFairness which connects the literature on causal fairness (Kilbertus et al., 2017) and the economic literature on envy-freeness.

Define Y(d, s), X(s) the potential outcome and covariates as functions of the sensitive attribute s . The following causal model is considered.

Assumption 5.1. Let (A) $Y(d, s) \perp\!\!\!\perp (D, S) \mid X(s)$, (B)$X(s) \perp\!\!\!\perp S$. Condition (A) and (B) state that the sensitive attribute is independent of potential outcomes and covariates, while it allows for the dependence of observed covariates and outcomes with the sensitive attribute. Indexing

potential outcomes and covariates captures this dependence by the sensitive attribute.

Let the conditional welfare, for the policy function being assigned to the opposite attribute, i.e., the effect of $\pi(x, s_1)$, on the group s_2 , conditional on covariates, be

$$V_{\pi(x,s_1)}(x, s_2) = \mathbb{E}\left[\pi(x, s_1) Y_i(1, s_2) + (1 - \pi(x, s_1)) Y_i(0, s_2) \mid X_i(s_2) = x\right] \tag{16}$$

We say that the agent with attribute s_2 envies the agent with attribute s_1, if her welfare (on the right-hand side of Equation (17)) exceeds the welfare she would have received had her covariate and policy been assigned the opposite attribute (left-hand side of Equation (17)), namely

$$\mathbb{E}_{X(s_1)}\left[V_{\pi(X(s_1),s_1)}(X(s_1), s_2)\right] > \mathbb{E}_{X(s_2)}\left[V_{\pi(X(s_2),s_2)}(X(s_2), s_2)\right] \tag{17}$$

We then measure the unfairness towards an individual with attribute s_2 as

$$\mathcal{A}(s_1, s_2; \pi) = \mathbb{E}_{X(s_1)}\left[V_{\pi(X(s_1),s_1)}(X(s_1), s_2)\right] - \mathbb{E}_{X(s_2)}\left[V_{\pi(X(s_2),s_2)}(X(s_2), s_2)\right]. \tag{18}$$

Whenever we aim not to discriminate in either direction, we take the sum of the effects $\mathcal{A}(s_1, s_2; \pi)$ and $\mathcal{A}(s_2, s_1; \pi)$ it connects to previous notions of counterfactual fairness (Kilbertus et al., 2017).

(Prediction disparity) : Prediction disparity and its empirical counterpart take the following form

$$C(\pi) = \mathbb{E}[\pi(X, S) \mid S = 0] - \mathbb{E}[\pi(X, S) \mid S = 1], \quad \hat{C}(\pi) = \frac{\sum_{i=1}^{n} \pi(X_i)(1 - S_i)}{n(1 - \hat{p}_1)} - \frac{\sum_{i=1}^{n} \pi(X_i) S_i}{n\hat{p}_1}, \tag{19}$$

Prediction disparity captures disparity in the treatment probability between groups. The second notion of UnFairness measures welfare disparities between the two groups.