CATE Estimation with Imbalanced Covariates

Linan (Frank) Zhao[*]

Department of Economics, Stanford University

Advised by Stefan Wager

June 7, 2024

Abstract

Conditional average treatment effect (CATE) estimates have been increasingly used in policy decision-making as they can profile and prioritize individuals who receive the most benefits from a treatment. This paper studies the specific case of an imbalanced covariate in the data set. We posit that standard parametric and non-parametric methods lead to disparate performance for the minority and majority groups, creating bias in CATE estimates. In this paper, we first provide theoretical derivations for reweighting methods in a parametric setting, which will provide deeper intuitions about the problem. Then, we propose a repository of tools that address the issues of imbalanced covariates, including reweighting in causal forests and data augmentation through generative modeling. We demonstrate the effectiveness of these methods through extensive simulation studies. Finally, we apply these novel methods to a real-world data set in the case of job training programs.

Keywords: Causal Inference, Heterogeneous Treatment Effect, Imbalanced Covariate, Program Evaluation

Acknowledgements: I would like to thank Stefan Wager for his invaluable mentorship and guidance, without whom this thesis would not have been possible. I want to thank Marcelo Clerici-Arias for all his work as the honors program director. I am grateful for all the professors who kindly supported me with insightful discussions, including Guido Imbens, Lihua Lei, Jann Spiess, Vasilis Syrgkanis, and Han Hong. I also thank Matt Brown, Lily Liu, friends, and family for their thoughtful suggestions and continued support.

[*] E-mail address: `frankz@outlook.com.au`.

1 Introduction

From job training initiatives to housing choice vouchers to educational programs, many empirical questions in economics require an evaluation of the causal effects of interventions or policies (Imbens and Wooldridge 2009). The key problem studied in causal inference is estimating a treatment's effect on a set of people, usually measured through some outcome-dependent variable of interest. Classic approaches focus on directly comparing the treated units' average outcome and the control units' average outcome. Whilst average treatment effect (ATE) estimates can inform the global effectiveness of the treatment intervention, it does not provide enough granularity for more detailed analysis, especially when heterogeneous treatment effects exist. For instance, suppose half of the participants in a randomized controlled experiment receive a positive effect from the treatment, but the other half receive negative treatment effects; we may observe an ATE of around 0. However, the treatment can have strictly positive impacts by only allocating the treatment to the participants who would benefit. As such, studying individual-level effects is crucial.

Indeed, there has been growing interest in estimating and understanding heterogeneous treatment effects in experimental and observational studies. Recent literature in this area has made profound impacts in the field of program evaluation (Abadie and Cattaneo 2018), where policy intervention decisions are made based on conditional average treatment effect (CATE) estimates. Contrary to the average treatment effect, CATE is concerned with the average treatment effect conditional on an individual's covariates such as age, gender, and race. These CATE estimates can inform "prioritization rules" that allocate interventions to individuals that would benefit the most (Hill 2011; Wager and Athey 2018; Künzel et al. 2019). In theory, precisely determined CATE-based rules could be the gold standard for selecting the prioritizing interventions (Manski 2004; Yadlowsky et al. 2023).

Thus, these CATE estimates can have significant societal repercussions when applied to personalized intervention decisions. This paper explores the specific case where an imbalanced covariate leads to disparate CATE estimation performance. When the CATE estimation is much worse on a minority group than a majority group, the induced prioritization rule will make more mistakes on the minority group members, adversely impacting their welfare.

Motivated by this, we study the problem of CATE estimation with imbalanced covariates in detail and propose methods that mitigate the inferior performance on the minority group. Our main contributions are three-fold:

1. We provide a motivating example and derive theoretical results for reweighting methods with parametric models. This provides intuitions for the problem of imbalanced covariates in CATE estimations.

2. We present a repository of tools for handling imbalanced data in causal inference with non-parametric methods and evaluate their performance extensively through simulation studies.

3. We demonstrate the effectiveness of the proposed methods by applying them to a real-world data set in the context of job training programs.

1.1 Preliminaries and Problem Set-up

In the usual causal inference set-up, we observe a training data set $D = (Y_i, X_i, W_i)_{1 \leq i \leq N}$, where Y_i is the observed outcome, X_i are the covariate vectors, and W_i is the treatment indicator. Considering binary treatment in this paper, $W_i = 1$ means that the unit i is exposed to treatment, and $W_i = 0$ means that the unit i is in the control group. For instance, W_i could be eligibility for a job training program, Y_i could be income in the next year, and X_i could be other covariates like age, race, and gender.

Using notation from the potential outcomes framework (Imbens and Rubin 2015), we denote $Y_i(1)$ to be the outcome of unit i if it were treated and $Y_i(0)$ to be the outcome of unit i if it were controlled. The fundamental problem of causal inference is that we only observe one outcome, namely $Y_i(W_i)$, and not the other (Holland 1986). Throughout this paper, we assume common causal inference assumptions of unconfoundedness, consistency, and positivity (Imbens and Rubin 2015).

In econometrics, we often care about the average treatment effect (ATE), which is denoted as $\tau^p = \mathbb{E}[Y_i(1) - Y_i(0)]$. Although a large corpus of causal inference literature focuses on estimating average treatment effects (Imbens and Rubin 2015; Pearl 2000; Morgan and Winship 2014), the main focus of this paper is conditional average treatment effects (CATE). The CATE is defined as

$$\tau(x) \coloneqq \mathbb{E}[Y(1) - Y(0) \mid X = x], \tag{1}$$

which potentially uncovers the heterogeneous treatment effects based on each unit's attributes X_i.

Estimation of the CATE function can be helpful for policy decision-making. For example, researchers and policymakers may use the CATE function to assign future units to their optimal treatment, e.g., $W_i^{opt} = \mathbf{1}_{\tau(X_i) \geq 0}$ (Athey and Imbens 2015). In addition, the CATE function induces a prioritization rule that allocates the treatment in decreasing order of $\tau(X_i)$. When resources are limited, individuals with the largest treatment effects would be treated first. Thus, by estimating the CATE function, researchers can uncover populations that do benefit and make better-informed decisions.

However, when a covariate is imbalanced, we postulate that the estimation accuracy of the CATE function on the minority group would be significantly lower than on the majority group. This implies that the induced prioritization rule will likely make more mistakes for the minority group. Formally, suppose that X_i^1, the first covariate, is a protected variable. For example, it could be age, race, gender, or nationality. We further assume that $\tau(x)$ is not independent of X^1, which ensures that our variable of interest has a nontrivial effect on the CATE. We first convert X^1 into a dummy variable by grouping related classes for categorical variables or grouping based on some threshold for continuous variables. Then, X^1 is defined to be *imbalanced* if $\mathbb{P}(X^1 = 0) - \mathbb{P}(X^1 = 1) > \delta$, for some δ significantly larger than 0 but below 1 (e.g. $\delta = 0.5$). This allows us to partition the data set into two groups, D_A and D_B, where D_A contains all minority units i whose $X_i^1 = 1$ and D_B contains all majority units j whose $X_j^1 = 0$. Here, we stress that covariate imbalance is inherent in the observed distribution of X^1, and covariate imbalance does not imply different covariate distributions between the treatment and control groups.

Let $\hat{\tau}$ be an estimator for τ. Using standard methods, we hypothesize that the estimation performance (e.g., Mean Squared Error) of $\hat{\tau}$ is likely to be significantly worse on D_A than on D_B. That is, $\frac{1}{|D_A|}\sum_{X_i|X_i^1=1}(\hat{\tau}(X_i)-\tau(X_i))^2 > \frac{1}{|D_B|}\sum_{X_i|X_i^1=0}(\hat{\tau}(X_i)-\tau(X_i))^2$. So, the natural research questions are:

1. **When there exists an imbalanced covariate of interest, do we expect the CATE estimation performance to be much worse on the minority group?**

2. **If so, what are some methods that can balance the CATE estimation performance for the minority group with the majority group?**

3. **Is it possible to balance estimation performance between the two groups without significantly worsening the performance on the majority group?**

There may be many reasons why researchers care about the worse estimation performance on the minority group. For instance, the minority group may have been historically marginalized in related policies; there may be sampling bias, and the minority group is less represented in the data set compared to the population; or the lack of treatment may make the minority group particularly vulnerable. Throughout this paper, we will continuously present simulation studies, methodologies, and their results targeted at each of these three questions. Next, we examine related works and put these research questions into a broader context.

1.2 Literature Review

Parallel to the growing interest in CATE estimations, economists and social scientists have gradually adopted machine learning methods for causal inference. Machine learning methods can be powerful where there may be many attributes of a unit relative to the number of observations and where the structural relationship between the covariates and the treatment effects are unknown (Athey and Imbens 2015). However, traditional machine learning algorithms typically minimize error rates without statistical properties like consistency, normality, and efficiency. In the specific case of identifying heterogeneous treatment effects, novel methods have been developed, studied, and broadly used. These methods utilize the power and generality of machine learning algorithms while allowing some statistical properties to be assessed. Some notable methods include the "CATE-generating transformation of the outcome" by Athey and Imbens (2015), the "generalized random forests" by Athey et al. (2019), the "single tree algorithm," the "two tree" algorithm by Hirano et al. (2003), and the "X-learner" by Künzel et al. (2019). See Athey and Imbens (2019) for a review of machine learning methods used in econometrics. This paper will mostly use these non-parametric machine learning approaches to estimate the CATE function.

In the context of imbalanced data, previous literature has focused mostly on the case of class imbalance in causal inference, whether there are substantially more control units than treatment units or whether the covariate distribution between control and treatment groups is significantly different. Some examples include inverse propensity weighting (Hirano et al. 2003), entropy balancing (Hainmueller 2012), and the X-learner (Künzel et al. 2019). However, few previous works have explored the impact when one protected variable is imbalanced. On the other hand, many techniques have been developed to tackle imbalanced data issues in prediction tasks, like local case-control sampling (Fithian and Hastie 2014), synthetic data augmentation (Chawla et al. 2002), and Yang et al. (2021). However, in most cases, we cannot directly transfer these techniques in prediction tasks to causal inference. Therefore, there exists a gap in methods that deal specifically with covariate imbalance in estimating heterogeneous treatment effects.

Next, we delve deeper into existing methods for handling imbalanced data in different contexts. Most of these methods can be broadly classified into two streams of approaches: reweighting and data augmentation. Finally, we will briefly survey emerging literature on fair estimation of treatment effects and summarize how our setting differs from previous works.

Class imbalance between the treatment and control groups occurs most frequently in observational studies because researchers often have access to large amounts of data that are not exposed to treatment. One common approach is to use inverse propensity weighting (Hirano et al. 2003), where the average treatment effect is estimated using $\hat{\tau}^p = \frac{1}{N}\sum_{i=1}^{N}(\frac{Y_i \cdot W_i}{\hat{e}(X_i)} - \frac{Y_i \cdot (1-W_i)}{1-\hat{e}(X_i)})$ and $\hat{e}(X)$ is the estimated propensity score $\mathbb{E}[W|X]$. However, researchers have found that slight misspecification of the propensity score model can lead to large biases in estimated treatment effects. Thus, Imai and Ratkovic (2014) proposed a "covariate balancing propensity score" that models treatment assignment and optimizes covariate balance between the treatment and control groups. This method mitigates the potential effect of misspecification of the propensity score model and provides better estimates. In a similar vein, Hainmueller (2012) proposed entropy balancing to re-weigh the treatment and control groups that can satisfy a large set of pre-specified balance conditions about known sample moments. Note that these methods focus mostly on average treatment effects.

For heterogeneous treatment effects, Künzel et al. (2019) proposed the X-learner, which estimates the CATE function using a three-step approach. It first estimates the response functions and then imputes the treatment effects before aggregating them using a weighted average. This method still addresses the imbalance between treatment and control units. In fact, there is a surprising lack of literature focusing on covariate imbalance, in the sense that one protected variable has significantly fewer units for one group than the other. Thus, this paper aims to address this gap directly by addressing covariate imbalance.

In addition, notice that the first three methods mentioned to address class imbalance are all re-weighting schemes that aim to balance the smaller and larger groups. In fact, re-weighting methods have been popular in the machine learning literature to address imbalance problems. For instance, Hashemi and Karimi (2018) investigates how common machine learning algorithms perform differentially using weighted data vs. non-weighted data. Further, "case-control sampling," which originated in epidemiology by Mantel and Haenszel (1959), samples uniformly from each class but adjusts the mixture of the classes to ensure balance. Building upon that, Fithian and Hastie (2014) proposed local-case control sampling, which uses a pilot estimate to preferentially select samples whose responses are conditionally rare given their attributes. These two approaches are analogous to re-weighting as locally difficult cases are selected more to fit the model. We should also note that these machine learning methods are mainly for prediction tasks and can only serve as inspirations in the causal inference setting.

In the paper by Lu and Liu (2022), the authors systematically investigated a weighted regression adjustment method for the average treatment effect. They give greater weights to units in the minority group and demonstrate that their re-weighting method is more robust than regression adjustment with treatment-covariate interactions. They discovered that the optimal weight for unit i is $\frac{W_i}{{p_1}^2} + \frac{1-W_i}{{p_0}^2}$, where p_1 and p_0 are the proportions of units assigned to the treatment and control groups, respectively. Again, this work addresses class imbalance rather than covariate imbalance. From the above reviews, we can see that we need a systematic study of how re-weighting can be applied to estimating the CATE function in imbalanced covariate cases.

Orthogonal to re-weighting schemes in machine learning and causal inference, synthetic data augmentation has been popular. In particular, SMOTE, or "Synthetic Minority Over-sampling Technique" (Chawla et al. 2002), has been widely used among machine learning researchers. SMOTE operates in the "covariate space" and over-samples the minority class by creating synthetic data points. These synthetic units are introduced along the line segments that join a minority unit and some of its minority class nearest neighbors. Since this approach is again designed for classification problems, it is not directly applicable to causal inference.

On the other hand, a recent paper on "Using Wasserstein Generative Adversarial Networks for the Design of Monte Carlo Simulations" by Athey et al. (2021) uses a WGAN to create synthetic data that adheres to the original data distribution for model selection in causal inference. This allows researchers to limit the degrees of freedom in Monte Carlo study designs and make more convincing comparisons between causal identification approaches. However, this paper has not explored using the generated data to augment the original data set. In the case of an imbalanced covariate, we can augment the minority group using generated data from the WGAN, similar to the idea of SMOTE.

In fact, with the rapid advancement of generative models, more works (Neal et al. 2020; Qian et al. 2023) are being developed to benchmark causal inference methods and leverage synthetic data for downstream applications.

Lastly, some previous works have examined optimal treatment assignment rules since conditional average treatment effect estimation can be used to make better policy decisions. For example, Hirano and Porter (2009) develops loss functions involving welfare loss and regret loss and derives treatment assignment rules that asymptotically achieve optimal outcomes. Similarly, Kitagawa and Tetenov (2018) assesses the Empirical Welfare Maximization model, which formulates a treatment assignment policy by maximizing the sample-based average social welfare across various potential treatment policies. Closely related to welfare maximization models, recent literature also started to focus on issues of fairness. For example, Kim and Zubizarreta (2023) studies the fair estimations of heterogeneous treatment effects and characterizes the trade-off between fairness and maximum welfare. Liang et al. (2021) develops a framework for the fairness-accuracy frontier and Viviano and Bradic (2024); Lei et al. (2023) study fairness in policy decision making.

Although we do not formalize a notion of fairness or welfare, we still contribute to this line of work as our main motivation is to balance the CATE estimation performance between a minority group and a majority group. In fact, as we will see in the rest of this paper, the proposed methods generally have a parameter of choice (weight on minority units, amount of synthetic data augmented) that produces a trade-off curve between the two groups. Researchers and policy-makers can then analyze this trade-off using their user-defined framework for fairness and welfare to pick the optimal point.

In summary, a surprising lack of literature specifically targets the covariate imbalance issue in CATE estimations. This paper contributes to this novel yet important research direction. Previous literature also inspires us to focus on two main approaches to answer the research questions – reweighting methods and synthetic data augmentation. The proposed methods are general and flexible enough to be applied to many contexts, and this paper contributes to the growing literature on algorithmic fairness.

2 A Synthetic Example

To illustrate the problem setup of imbalanced covariates, we first construct a synthetic example that will be used throughout this paper. We consider a randomized control experiment setting. For each unit i, suppose the covariate vector X_i includes six covariates $(X_i^1, X_i^2, \cdots, X_i^6)$. The first three covariates are categorical and only take on the values of 0 or 1. The last three covariates are continuous, and their ranges are in $\mathbb{R}$. Specifically, we denote X^1 as the protected variable whose probability of being 1 is $P(X^1 = 1) = p$, for some $p \in [0, 1]$. This parameter p controls the level of imbalance as smaller values indicate more imbalance in X^1. Unless otherwise specified, p is equal to 0.1 in our examples. This mimics the scenario in which the first covariate is imbalanced, and the number of data points in the majority group ($X^1 = 0$) is significantly more than in the minority group ($X^1 = 1$). The other two categorical covariates X^2, X^3 are balanced in the sense that $P(X^2 = 1) = P(X^3 = 1) = \frac{1}{2}$. The continuous covariates X^4, X^5, X^6 are all generated from a unit normal distribution $\mathcal{N}(0, 1)$.

Since we assume a randomized control experiment setting, the treatment indicator W is a Bernoulli random variable with $P(W = 1) = \pi = \frac{1}{2}$. Furthermore, we define the CATE function as

$$\tau(X) = X^1 + X^2 + X^4 + (1 - X^1) \cdot X^5. \tag{2}$$

The CATE function non-trivially depends on the protected variable X^1, and it also depends on other categorical and continuous covariates. We include the interaction term at the end to simulate the setting where the CATE functions for the minority and majority groups are slightly different locally but still share many similarities. Indeed, we can see that the CATE function is $X^1 + X^2 + X^4$ for the minority group and $X^2 + X^4 + X^5$ for the majority group. Without the interacting term, the functional form for the CATE function would essentially be identical for both groups, hence making the problem trivial. After specifying the CATE function, we define the outcome variable as

$$Y(X, W) = \max(\sum_{i=1}^{6} X^i, 0) + W \cdot \tau(X) + \mathcal{N}(0, 1). \tag{3}$$

This data generating process creates a data set $D = (Y_i, X_i, W_i)_{1 \leq i \leq N}$, from which we wish to estimate the CATE $\hat{\tau}(X)$.

3 Re-weighting Method with Linear Models

To estimate the CATE function in Equation (2) given D, we assume that we can construct and subsequently have access to the "scores" or pseudo-outcomes $\widehat{\Gamma}_i$ that are nearly unbiased but potentially noisy proxies for the CATE:

$$\mathbb{E}\left[\widehat{\Gamma}_i \mid X_i\right] \approx \tau\left(X_i\right) = \mathbb{E}\left[Y_i(1) - Y_i(0) \mid X_i\right]. \tag{4}$$

One choice for constructing these scores is inverse-propensity weighting (IPW) $\widehat{\Gamma}_i = \frac{W_i Y_i}{\pi} - \frac{(1-W_i)Y_i}{1-\pi}$. However, these scores can have large variances, so we can use its doubly-robust counter-part (Robins et al. 1994):

$$\begin{aligned} \widehat{\Gamma}_i &= \hat{m}\left(X_i, 1\right) - \hat{m}\left(X_i, 0\right) + \tfrac{W_i - \hat{e}(X_i)}{\hat{e}(X_i)(1-\hat{e}(X_i))}\left(Y_i - \hat{m}\left(X_i, W_i\right)\right), \\ e(x) &= P\left[W_i = 1 \mid X_i = x\right], \quad m(x, w) = \mathbb{E}\left[Y_i(w) \mid X_i = x\right], \end{aligned} \tag{5}$$

where $e(x)$ is the propensity scores and $m(x, w)$ represents the expected outcome given a unit's covariates and treatment assignment; and $\hat{e}, \hat{m}$ are non-parametric estimates of e, m. In randomized controlled trials, we can directly replace $\hat{e}(x)$ with the probability of treatment assignment π and Equation (5) becomes the Augmented IPW (AIPW) estimator

$$\widehat{\Gamma}_i = \hat{m}\left(X_i, 1\right) - \hat{m}\left(X_i, 0\right) + \frac{W_i - \pi}{\pi\left(1 - \pi\right)}\left(Y_i - \hat{m}\left(X_i, W_i\right)\right). \tag{6}$$

We assume that the non-parametrically estimated nuisance parameter $\hat{m}(X_i, W) \approx \mathbb{E}[Y_i(W)|X_i]$ for $W \in \{0, 1\}$. This represents the expected outcome given covariates and treatment assignment. In a randomized trial, $\mathbb{E}[W|X_i] = \pi$, and so the constructed scores in Equation (6) satisfies Equation (4) whenever cross-fitting is used. Cross-fitting (Chernozhukov et al. 2018) is the process of splitting the data sample into two halves, where one is used for estimating the nuisance parameters, and the other is used for applying the score functions. This process helps to avoid "own observation" bias and helps the constructed scores to adhere to Equation (4) (Yadlowsky et al. 2023).

Once we have access to these scores, we can regress them against the covariates X to obtain an estimate of the CATE function. We start with the linear model $\widehat{\Gamma} \sim X\beta$. For simplicity, let X be a $n \times 7$ matrix, where the last column contains all ones, and let β be a 7-dimensional column vector. We suppose that there are m observations in the minority group. As such, $\mathbb{E}[\frac{m}{n}] = p$. Next, we can fit a linear model and assess its performance.

However, notice that in our synthetic example in Equation (2), the CATE function is not globally linear, and the linear model would be misspecified. The CATE function for the minority group would be $X^1 + X^2 + X^4$, and the CATE function for the majority group would be $X^2 + X^4 + X^5$. Thus, when we fit the coefficients of a linear model, the large sample size of the majority group would bias the coefficient of X^5 to be closer to 1, leading to lower MSE for the majority group and higher MSE for the minority group.

One straightforward way to tackle this imbalance in performance is to use a weighted regression, similar to Lu and Liu (2022). We may assign weights $\alpha(X_i)$ to each observation before running the linear regression. The simplest choice is $\alpha(X_i) = 1$ for all units in the majority group and $\alpha(X_i) = \gamma \in \mathbb{R}^+$ for all units in the minority group. We are interested in studying the estimation performances on both groups as the weight γ on the minority group varies. Note that the weighting function $\alpha(X_i)$ induces a $n \times n$ diagonal matrix A, where the diagonal entries are $\alpha(X_i)$.

Below, we first derive some theoretical results on this weighted linear regression. Then, we run some simulations to confirm the theoretical intuitions numerically.

3.1 Theoretical Derivations

Using least squares, the coefficients of the weighted regression $\widehat{\Gamma} \sim X\beta$ with weights $\alpha(X_i) = \begin{cases} \gamma & X_i^1 = 1 \\ 1 & X_i^1 = 0 \end{cases}$ would be

$$\hat{\beta}_\gamma = \text{argmin}_\beta \mathbb{E}[\alpha(X_i)(\widehat{\Gamma}_i - X_i\beta)^2] = \text{argmin}_\beta (\widehat{\Gamma} - X\beta)^\top A(\widehat{\Gamma} - X\beta). \tag{7}$$

The loss function $(\widehat{\Gamma} - X\beta)^\top A(\widehat{\Gamma} - X\beta)$ can be written as $\widehat{\Gamma}^\top A\widehat{\Gamma} - 2\beta^\top X^\top A\widehat{\Gamma} + \beta^\top X^\top AX\beta$. Then, taking the derivative against β, we obtain $-2X^\top A\widehat{\Gamma} + 2X^\top AX\beta$. Setting the derivative to zero, we see that $\hat{\beta}_\gamma = (X^\top AX)^{-1}X^\top A\widehat{\Gamma}$. Therefore,

$$\begin{aligned}
\hat{\beta}_\gamma &= (X^\top AX)^{-1}X^\top A\widehat{\Gamma} \\
&= (X^\top AX)^{-1}X^\top A\tau(X) + (X^\top AX)^{-1}X^\top A(\widehat{\Gamma} - \tau(X)) \\
&= \left(\frac{1}{n}X^\top AX\right)^{-1}\left(\frac{1}{n}X^\top A\tau(X)\right) + \left(\frac{1}{n}X^\top AX\right)^{-1}\left(\frac{1}{n}X^\top A(\widehat{\Gamma} - \tau(X))\right) \qquad (8)
\end{aligned}$$

Let $G = \frac{1}{n}X^\top AX$, $H = \frac{1}{n}X^\top A\tau(X)$, and $D = \frac{1}{n}X^\top A(\widehat{\Gamma} - \tau(X))$. Note that $G_{i,j} = \frac{1}{n}\sum_{k=1}^{n}\alpha(X_k)X_k^i \cdot X_k^j$. Invoking the central limit theorem, we see that $G - \mathbb{E}[G] \xrightarrow{d} \frac{1}{\sqrt{n}}\mathcal{N}(0, \Sigma)$. As long as $\mathbb{E}[G]$ is invertible, the delta method with the function $f(\cdot)$ being matrix inverse gives

$$G^{-1} - \mathbb{E}[G]^{-1} \xrightarrow{d} \frac{1}{\sqrt{n}}\mathcal{N}(0, f^{'}(\mathbb{E}[G])^\top\Sigma f^{'}(\mathbb{E}[G])). \qquad (9)$$

Thus, we can write $G^{-1} = \mathbb{E}[G]^{-1} + o_p(1)$ because $f^{'}(\mathbb{E}[G])^\top\Sigma f^{'}(\mathbb{E}[G])$ is finite. Similarly, we have that $H = \mathbb{E}[H] + o_p(1)$ by the central-limit theorem. In addition, Robins et al. (1994) shows that the AIPW estimator $\widehat{\Gamma}$ is asymptotically normal under some regularity conditions and when the nuisance parameters are cross-fitted. Thus, $D = \frac{1}{n}X^\top A(\widehat{\Gamma} - \tau(X)) = o_p(1)$. As such,

$$\begin{aligned}
\hat{\beta}_\gamma &= (\mathbb{E}[G]^{-1} + o_p(1))(\mathbb{E}[H] + o_p(1)) + (\mathbb{E}[G]^{-1} + o_p(1))o_p(1)) \\
&= \mathbb{E}[G]^{-1}\mathbb{E}[H] + o_p(1) \qquad (10)
\end{aligned}$$

because $c \cdot o_p(1) = o_p(1)$ for a constant c and $o_p(1) \cdot o_p(1) = o_p(1), o_p(1) + o_p(1) = o_p(1)$.

By computing the expected value of the product of two random variables X^iX^j, we can explicitly compute each entry of $\mathbb{E}[G]_{i,j}$ using the underlying data-generating process.

The full matrix $\mathbb{E}[G]$ is recorded in Appendix B.1. The inverse $\mathbb{E}[G]^{-1}$ is calculated as

$$\begin{pmatrix} \frac{p\gamma+1-p}{p\gamma(1-p)} & 0 & 0 & 0 & 0 & 0 & \frac{1}{p-1} \\ 0 & \frac{4}{p\gamma+1-p} & 0 & 0 & 0 & 0 & -\frac{2}{p\gamma+1-p} \\ 0 & 0 & \frac{4}{p\gamma+1-p} & 0 & 0 & 0 & -\frac{2}{p\gamma+1-p} \\ 0 & 0 & 0 & \frac{1}{p\gamma+1-p} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{p\gamma+1-p} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{p\gamma+1-p} & 0 \\ \frac{1}{p-1} & -\frac{2}{p\gamma+1-p} & -\frac{2}{p\gamma+1-p} & 0 & 0 & 0 & \frac{(\gamma-3)p+3}{(1-p)(p\gamma+1-p)} \end{pmatrix}. \tag{11}$$

Similarly, we can evaluate $\mathbb{E}[H]$, which turns out to be

$$\begin{pmatrix} \frac{3p\gamma}{2} & \frac{2p\gamma+(1-p)}{2} & \frac{3p\gamma+(1-p)}{4} & p\gamma+(1-p) & 1-p & 0 & \frac{3p\gamma+(1-p)}{2} \end{pmatrix}^{\top}. \tag{12}$$

By Equation (10), $\hat{\beta}_\gamma$ converges in probability to $\mathbb{E}[G]^{-1}\mathbb{E}[H]$ as the sample size grows to infinity. Thus, $\beta_\gamma^{\text{pop}} = \mathbb{E}[G]^{-1}\mathbb{E}[H]$, which we can now calculate as

$$\beta_\gamma^{\text{pop}} = \begin{pmatrix} 1 & 1 & 0 & 1 & \frac{1-p}{p\gamma+1-p} & 0 & 0 \end{pmatrix}^{\top}. \tag{13}$$

In fact, we see that as n goes to infinity, we can expect the coefficients for all covariates apart from X^5 to be correct. The weights γ only affect the asymptotic coefficient β_5 on X^5 since that is where the CATE function differs on the minority and majority groups. In particular, we can see that as p decreases, the coefficient β_5 gets closer to 1, indicating the dominance of the majority group. However, as γ increases, this coefficient decreases, which shows that the model is fitting the minority group more. Using the derived β vector in Equation (13), the CATE estimate would be $\hat{\tau}(X) = X^1 + X^2 + X^4 + \frac{1-p}{p\gamma+1-p}X^5$ as the sample size grows to infinity.

Next, we calculate the variances of the linear regression coefficients $\hat{\beta}_\gamma$. The results in White (1980) and Alberto Abadie and Zheng (2014) imply that, under some regularity conditions,

$$\hat{\beta}_\gamma - \beta_\gamma^{\text{pop}} \xrightarrow{d} \frac{1}{\sqrt{n}} \mathcal{N}(0, \mathbb{V}_\gamma^{\text{pop}}), \tag{14}$$

where the asymptotic variance is

$$\mathbb{V}_\gamma^{\text{pop}} = \left(\mathbb{E}\left[(X^\top A X)\right]\right)^{-1} \left(\mathbb{E}\left[X^\top A \varepsilon \varepsilon^\top A^\top X\right]\right) \left(\mathbb{E}\left[X^\top A X\right]\right)^{-1} \tag{15}$$

and $\varepsilon = \widehat{\Gamma} - X\beta_\gamma^{\text{pop}} = (\widehat{\Gamma} - \tau(X)) + (\tau(X) - X\beta_\gamma^{\text{pop}})$. The asymptotic variance can be used to construct confidence intervals on the coefficients and run hypothesis tests. In general, the first part $\widehat{\Gamma} - \tau(X)$ of ε has heteroscedastic variance and so Equation (15) can only be evaluated numerically.

With a sufficiently large sample size, we can use Equation (13) and Equation (15) to calculate the expected mean squared errors on the two groups:

$$\begin{aligned} \mathbb{E}\left[\frac{1}{|D_A|} \sum_{X_i | X_i^1 = 1} (\hat{\tau}(X_i) - \tau(X_i))^2\right] &\approx \left(\frac{1-p}{p\gamma + 1 - p}\right)^2 + \mathrm{V}_A, \\ \mathbb{E}\left[\frac{1}{|D_B|} \sum_{X_i | X_i^1 = 0} (\hat{\tau}(X_i) - \tau(X_i))^2\right] &\approx \left(\frac{p\gamma}{p\gamma + 1 - p}\right)^2 + \mathrm{V}_B, \end{aligned} \tag{16}$$

where $\mathrm{V}_A, \mathrm{V}_B$ are errors caused by the variances of the coefficients. These two values depend on the number of observations n in the training data, the level of imbalance p, the weight γ on the minority units, and the residuals of the constructed scores. On the other hand, the first part of the expected MSE is attributed to the misspecification of our linear model. We refer to the two terms $(\frac{1-p}{p\gamma+1-p})^2, (\frac{p\gamma}{p\gamma+1-p})^2$ as misspecification errors.

When the number of units in the data set n grows, the second components $\mathrm{V}_A, \mathrm{V}_B$ converges to 0. As such, studying the errors attributed to misspecification is more interesting. The theoretical ratio of the misspecification errors on the minority group to the majority group is $r(p, \gamma) = \left(\frac{1-p}{p\gamma}\right)^2$. When the linear regression is not weighted with $\gamma = 1$, a small p value, which induces a large imbalance in the data set, creates a large discrepancy between the expected MSEs on the minority and majority groups. For instance, when $p = 0.1$, the misspecification error on the minority group can be 81 times the misspecification error on the majority group. In addition, Equation (16) provides the optimal weight γ for equalizing the misspecification errors on the two groups. The optimal γ would be $\dfrac{1-p}{p}$, which is the ratio of the majority group size to the minority group size in the data set.

These theoretical results provide intuitions about the linear parametric model with scores constructed using the AIPW estimator. In general, we confirm that when a linear model is misspecified, the estimated CATE function admits significantly higher mean squared errors in the minority group than in the majority group. This performance gap is exaggerated when p decreases and the covariate X^1 is more imbalanced. A weighted linear regression is a valid tool for balancing the CATE estimation performance across the two groups. However, while increasing the weight γ on the minority group decreases the expected MSE on the minority group, it necessitates an increase in the expected MSE on the majority group. Thus, reweighting methods with simple linear models create a trade-off curve between the MSEs on the two groups, visualized in Figure 1. Equation (16) allows decision-makers to deliberate further where the optimal point is on the trade-off curve. For instance, if we want to ensure complete equality, the weight $\gamma = \frac{1-p}{p}$, which balances the misspecification errors on both groups, is a good choice.

When researchers use simple linear models on a real-world data set, the model can be arbitrarily misspecified. When the CATE function may differ across the minority and majority groups, these models can be biased towards the majority group in the sense of a lower MSE. Adding more weight to the minority units can mitigate this problem, but some trade-offs between the two groups are likely to occur. So far, we have made progress towards the first two research questions, but the third question has yet to be explored.

3.2 Correspondence with Simulation Results

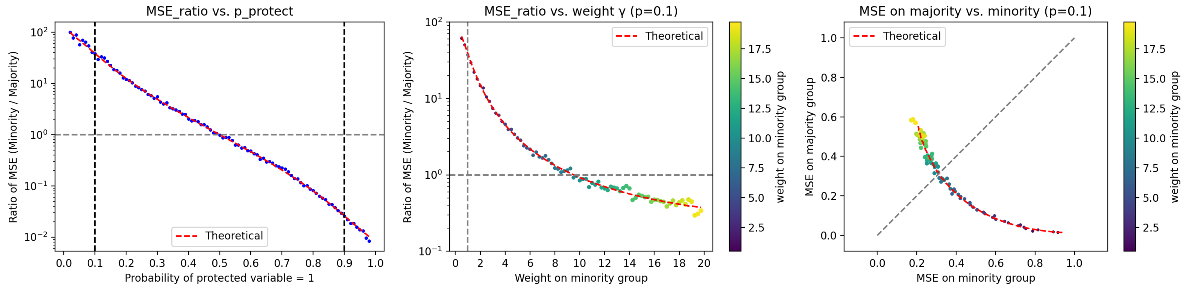
To confirm the theoretical results, we can simulate a data set using the data generating process outlined in Section 2. For different values of p and γ, we generate 4000 data points for the training set and 1000 data points for the testing set. We use Equation (5) to obtain the scores, where the nuisance parameters are cross-fitted using regression forests

(Breiman 2001). Then, the constructed scores are regressed against the covariates using weighted linear regression. Each specification of p and γ are simulated 50 times, and the results are averaged.

We also check that the residuals of the constructed scores have high heteroscedasticity. The residuals obtain a p-value of less than 10^{-49} in the Breusch–Pagan (Breusch and Pagan 1979) test for heteroscedasticity. Similarly, as the probability of the protected covariate p varies, the variance of $\widehat{\Gamma} - \tau(X)$ in each subgroup varies significantly, with higher variance for the minority group. In addition, the variance of this component seems to be much larger than the second component. For a more detailed analysis of this heteroscedastic variance, see Appendix B.2. As such, even though we understand the second part $\tau(X) - X\beta_{\gamma}^{\text{pop}}$ of ε, it remains impossible to fully calculate $\varepsilon\varepsilon^{\top}$ analytically. Thus, to determine the theoretical MSE values, we need to calculate the terms $\mathrm{V}_A, \mathrm{V}_B$ using Equation (15) numerically. The formulas for $\mathrm{V}_A, \mathrm{V}_B$ are included in Appendix B.3.

We can, therefore, plot the empirical simulation results and overlay the theoretical curves to make comparisons, as shown in Figure 1.

Figure 1: The left graph plots the ratio in MSE (MSE on minority / MSE on majority) against the level of imbalance p. The middle graph plots the ratio of MSE on the minority group to the MSE on the majority group against weights γ for $p = 0.1$. The right graph plots the MSE on the majority group against the MSE on the minority group as the weight on the minority group varies. For all three graphs, the dotted red line represents the theoretical curve derived from Equation (16) with numerically calculated Huber-White variance. The other points come from simulations.



The simulation results are identical to the theoretical curves. This again confirms that as the probability of the protected variable $X^1 = 1$ increases from small values up to 0.5 (complete balance), the discrepancy between the majority and minority groups diminishes. As the weight on the minority group increases, the MSE on minority units decreases while the MSE on majority units increases. In fact, the CATE estimation performance is equalized right around $\gamma = 9$, which confirms our theoretical intuition that $\gamma = \frac{1-p}{p} = 9$ balances the misspecification errors. Most importantly, the right graph illustrates the trade-off curve between the MSEs of the two groups, enabling policymakers to make informed decisions within specific settings by choosing an optimal point.

4 Non-parametric Methods

In this section, we turn to the third research question and explore whether it is possible to increase CATE estimation accuracy on the minority group without harming the majority group significantly. To this end, we propose two non-parametric methods – reweighting with causal forests and synthetic data augmentation. We evaluate each method's performance using simulation studies. The simulations are set up according to Section 2, where we generate 4000 data points for the training set and 1000 data points for the testing set. Again, every specification is repeated 50 times, and the results are averaged.

4.1 Reweighting with Causal Forests

Similar to reweighting in linear models, we can apply the same analysis to causal forests (Athey et al. 2019). Causal forest extends the random forest (Breiman 2001) algorithm and recursively partitions the covariate space to maximize treatment heterogeneity. First, denote $e(x) = \mathbb{P}[W_i|X_i = x]$ for the propensity score and $m(x) = \mathbb{E}[Y/_i|X_i = x]$ for the expected outcome. Essentially, the causal forest tries to learn the heterogeneous treatment effects through the R-Learner (Athey and Wager 2019):

$$\hat{\tau}(\cdot) = \operatorname{argmin}_{\tau} \left\{ \sum_{i=1}^{n} \left(\left(Y_i - \hat{m}^{(-i)}\left(X_i\right) \right) - \tau\left(X_i\right) \left(W_i - \hat{e}^{(-i)}\left(X_i\right) \right) \right)^2 + \Lambda_n(\tau(\cdot)) \right\}, \quad (17)$$

where the propensity scores $\hat{e}(\cdot)$ and marginal outcomes $\hat{m}(\cdot)$ are cross-fitted with separate regression forests, and $\Lambda_n(\tau(\cdot))$ is a regularizer for the complexity of learned $\hat{\tau}$. In a randomized controlled experiment, $\hat{e}(\cdot)$ can be replaced with the treatment probability

π. The $^{(-i)}$ superscript denotes out-of-bag predictions, where for example Y_i is not used to train $\hat{m}^{(-i)}(X_i)$.

Given a subsample S of the data, a causal tree aims to split the sample into L and R leaves based on some covariate. It assumes homogeneous treatment effects in the two children and computes $\hat{\tau}_L$ and $\hat{\tau}_R$ through residual-on-residual regression:

$$\hat{\tau}_L \leftarrow \operatorname{lm}\left(\left(Y_i - \hat{m}^{(-i)}(X_i)\right) \sim \left(W_i - \hat{e}^{(-i)}(X_i)\right) : X_i \in L\right), \tag{18}$$

and likewise for $\hat{\tau}_R$.

The algorithm greedily finds a split that maximizes the weighted difference $n_L n_R(\hat{\tau}_L - \hat{\tau}_R)^2$. Intuitively, this procedure allows causal trees to find leaves where the treatment effect is constant but different from others. This is because when the treatment effect is roughly constant over some neighborhood $N(X)$, we can solve a partially linear model over the neighborhood using Equation (18). After growing the trees and aggregating them into a forest, given a new test point x, the weights induced by the causal forest are

$$\omega_i(x) = \frac{1}{B}\sum_{b=1}^{B} \frac{\mathbf{1}\left(\{X_i \in L_b(x), i \in \mathcal{S}_b\}\right)}{\left|\{i : X_i \in L_b(x), i \in \mathcal{S}_b\}\right|}, \tag{19}$$

where B is the number of trees, $L_b(x)$ is the leaf that the new test point x falls in tree b, and $\mathcal{S}_b$ is the subsample used to construct tree b. Intuitively, these weights measure the similarity between the new test and training data points.

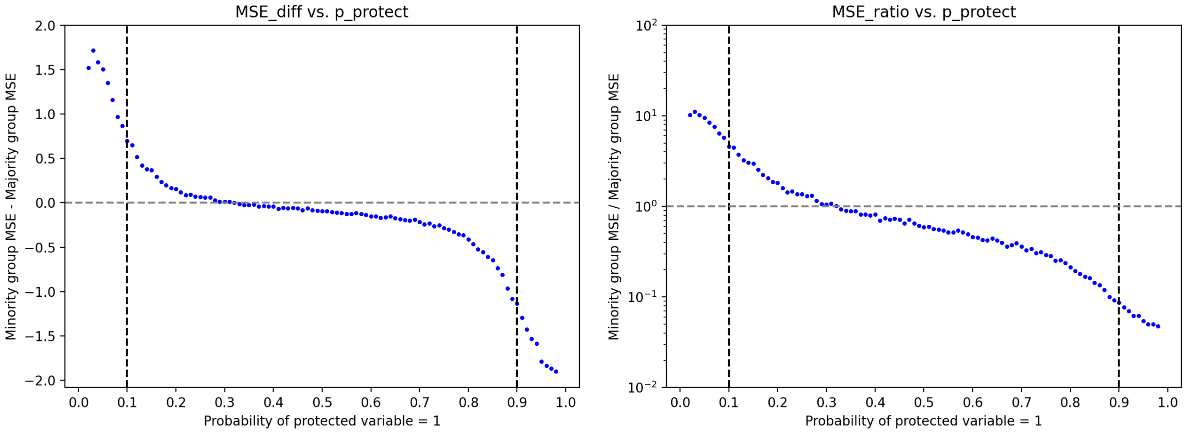
Then, causal forest acts as an adaptive kernel method and calculates the final $\hat{\tau}$ as:

$$\hat{\tau} = \frac{\sum_{i=1}^{n} \omega_i(x)\left(Y_i - \hat{m}^{(-i)}(X_i)\right)\left(W_i - \hat{e}^{(-i)}(X_i)\right)}{\sum_{i=1}^{n} \omega_i(x)\left(W_i - \hat{e}^{(-i)}(X_i)\right)^2}. \tag{20}$$

In addition, a causal forest uses an honesty condition to avoid over-fitting and confounding effects. Half of the sample would be used to build the trees, and the other half of the sample would be used for inference. In this paper, we use the *EconML* package to run causal forests for simulation studies and *grf* for application to a real data set.

Then, we can present graphs similar to Figure 1. We first plot the CATE estimation performance disparity against p in Figure 2, which controls the amount of minority data points compared to the majority group.

Figure 2: Performance disparity against different levels of imbalance using the causal forest model. The left graph plots the difference in MSE (MSE on minority - MSE on majority) against p, and the right graph plots the ratio in MSE (MSE on minority / MSE on majority) against p.

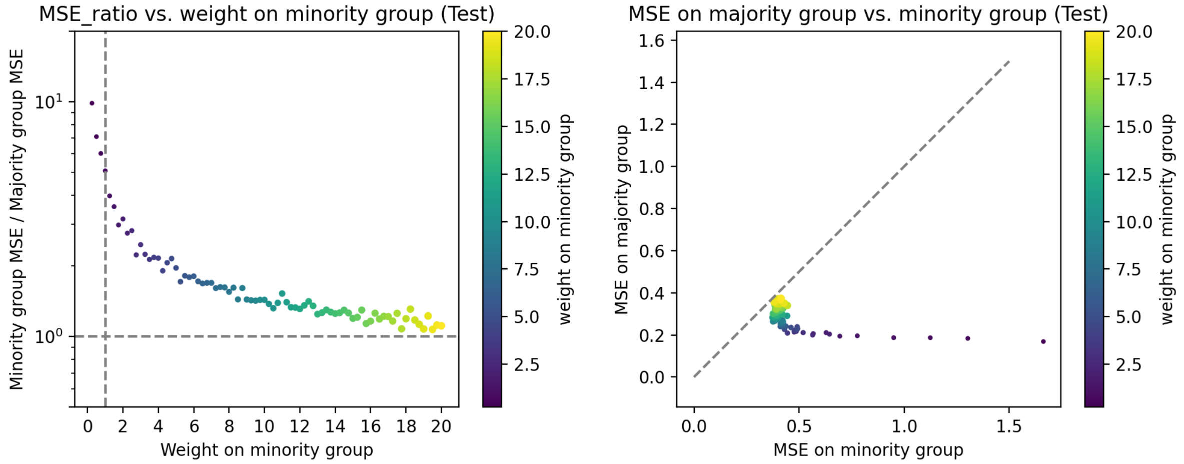


Similar to our results before, as the number of minority data points increases, the performance gap between the two groups first closes and then widens after the equilibrium point of balance. Interestingly, in the causal forest case, the point of balance seems to be around $p = 0.3$, where there are slightly more majority data points than minority ones. This could be because the estimation is slightly more difficult locally for the majority group as their CATE function involves the additional term X^5.

Next, we apply weights γ to the minority data points when running the causal forest algorithm, as shown in Figure 3. When sample weights are added to the algorithm, the objective in Equation (17) becomes a weighted sum, and Equation (18) becomes a weighted linear regression. During inference, Equation (20) adds the additional sample weights $\alpha(x)$ to the weighted regression, i.e. $\omega_i(x)$ becomes $\alpha(x)\omega_i(x)$. Intuitively, this prioritizes fit on the units that are weighted more.

Similar to the linear model result, as weight increases, MSE on the minority group decreases, and MSE on the majority group increases slightly. However, contrary to the linear model, the increasing MSE on the majority group is much less prevalent. In fact, the MSE on the majority group never becomes larger than the MSE on the minority group.

Figure 3: Performance disparity against different minority group weights γ using the causal forest approach. The left graph plots the ratio in MSE (MSE on minority / MSE on majority) against different weights γ. The right graph plots the MSE on the majority group against the MSE on the minority group for different weights γ. Both plots depict performances on the test set.



Intuitively, this could be because when the weights on the minority data points become sufficiently large, the causal trees always split on the covariate X^1. After this split, the CATE function is locally linear for both groups; hence, the estimation would be relatively straightforward. Nonetheless, the fact that increasing minority weights γ creates large positive improvements for the minority group at the cost of minute loss for the majority group is desirable. This might indicate that a close-to-Pareto improvement is possible for the whole population. However, we should still be careful about the exact value of γ to use. It seems that when $\gamma \geq 10$, the MSE on the minority group stops improving, and it only worsens the performance on the majority group. But for a weight $\gamma < 10$ in this case, reweighting with causal forest achieves the goal of balancing performance among the two groups while not hurting the majority group significantly.

4.2 Data Augmentation with Generative Models

In classification tasks, synthetic data augmentation methods like SMOTE (Chawla et al. 2002) have been popularly used in class imbalance settings. Similarly, we can envision an approach that generates synthetic data to augment the minority group in the covariate imbalance setting. The advantage could be that re-weighting methods are discrete, but adding synthetic data brings additional information and leads to smoother improvements.

There has not been much work done directly on augmenting the data set with synthetic data points in CATE estimations. However, Athey et al. (2021) proposes a viable way to generate synthetic data for causal inference tasks. The authors propose Wasserstein Generative Adversarial Networks (WGAN) to systematically generate artificial data that closely mimics existing data sets to improve the credibility of Monte Carlo studies. Although the paper utilizes synthetic data for model comparisons, we can leverage the generated data for augmentation purposes. Indeed, it remains unexplored how adding synthetic data units will influence CATE estimation performance.

The generative adversarial framework was first introduced in Goodfellow et al. (2014). It is a two-player, non-cooperative game between a generator g and a discriminator d. The generator is often parameterized as a neural network with parameters θ_g. It takes some Gaussian noise $Z \sim N(0, I)$ as input and tries to generate data points similar to the input data distribution from $X_1, \cdots, X_N$. The discriminator is also often parameterized as a neural network with parameters θ_d. The discriminator's goal is to discern whether a data point is "real" or "fake," meaning discerning whether the data is from the input data set or the generator. It outputs the probability of any data point being "real." The classic GAN training min-max objective is:

$$\min_{\theta_g} \max_{\theta_d} L\left(\theta_d, \theta_g\right), \tag{21}$$

where the objective function is

$$L\left(\theta_d, \theta_g\right) = \frac{1}{N_R} \sum_{i=1}^{N_R} \ln d\left(X_i; \theta_d\right) + \frac{1}{N_F} \sum_{i=1}^{N_F} \ln\left[1 - d\left(g\left(Z_i; \theta_g\right); \theta_d\right)\right]. \tag{22}$$

Here, $d(\cdot)$ represents the output of the discriminator, $g(\cdot)$ represents the generator's output, and N_R, N_F are the number of real and fake units, respectively.

Then, we can use gradient descent methods to update the generator and discriminator iteratively. When the training converges, the generated sample distribution should be near-identical to the training data distribution, and the discriminator will not be able to discern them apart.

However, GAN training is notoriously hard and unstable. Difficulties can arise when the discriminator becomes too proficient early on in detecting generated observations and fails to provide useful gradient information for the generator to improve. To remedy that, Arjovsky et al. (2017) proposes an alternative using the Earth-Mover or Wasserstein distance. The optimization problem of the WGAN becomes

$$\min_{\theta_g} \max_{\theta_c} \left\{ \frac{1}{N_R} \sum_{i=1}^{N_R} f\left(X_i; \theta_c\right) - \frac{1}{N_F} \sum_{i=1}^{N_F} f\left(g\left(Z_i; \theta_g\right); \theta_c\right) \right\}, \tag{23}$$

where the function f is a critic parameterized by θ_c, and its optimized value implies an upper bound on how much any Lipschitz-continuous moment can differ between the real and fake distributions. For WGAN to work, we require the critic to be 1-Lipschitz. To enforce that, it is common to add a penalty term in the following form

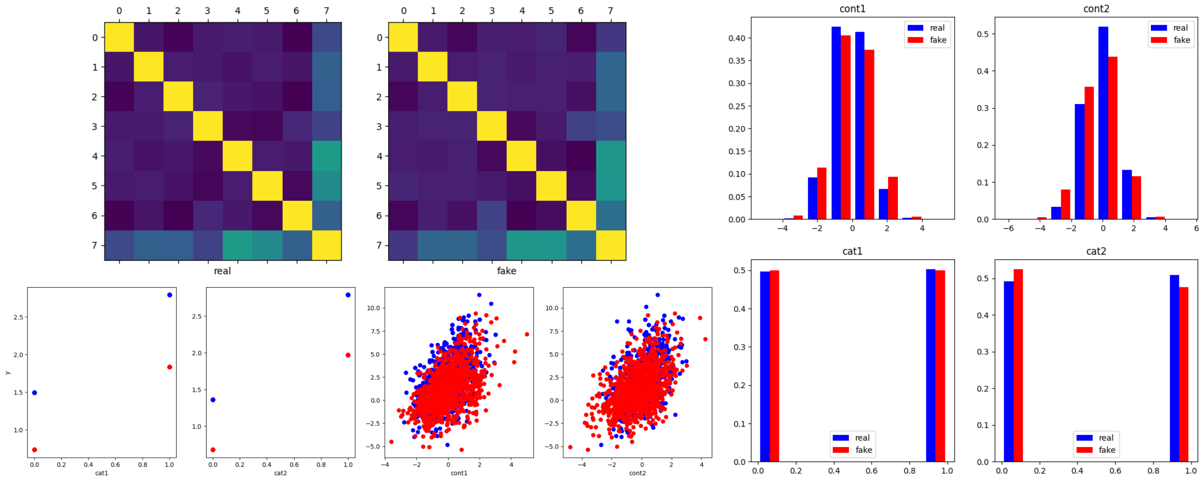
$$\lambda \left\{ \frac{1}{m} \sum_{i=1}^{m} \left[\max\left(0, \left\| \nabla_{\hat{x}} f\left(\hat{X}_i; \theta_c\right) \right\|_2 - 1\right) \right]^2 \right\}. \tag{24}$$

Further, WGAN can be conditioned on some input V by simply concatenating V to the input noise. For the exact WGAN algorithm, please refer to page 11 in Athey et al. (2021). The first step is to generate X conditional on W, where W contains the same fraction of treated units as in the input data. The second step is to train another generator that generates $Y(W)$ given the treatment indicator W and the covariates X. This results in two generators with parameters $\theta_{g,X|W}$ and $\theta_{g,Y(W)|X,W}$. Modeling the data generation process in this way also allows us to have access to both potential outcomes $Y(0)$ and $Y(1)$.

After training the generators, we can synthesize realistic data points that are similar to the original data set. Although Athey et al. (2021) does not explore whether the synthetic data could be used as augmentation, it serves as promising evidence that we can, in principle, create realistic synthetic data sets. Thus, exploring what happens when we use synthetic data augmentation to address covariate imbalance would be intriguing.

To explore the idea of synthetic data augmentation, we again use the simulation setup in Section 2. In particular, after we simulate a training data set of 4000 units, we use the approach outlined above to train a generator (regarding $\{\theta_{g,X|W}, \theta_{g,Y(W)|X,W}\}$ as a single entity) through adversarial loss. Importantly, to aid the generator in learning the minority group distribution, we over-sample the minority group to balance with the majority group before training the generator. The graphs in Figure 4 show the generation results.

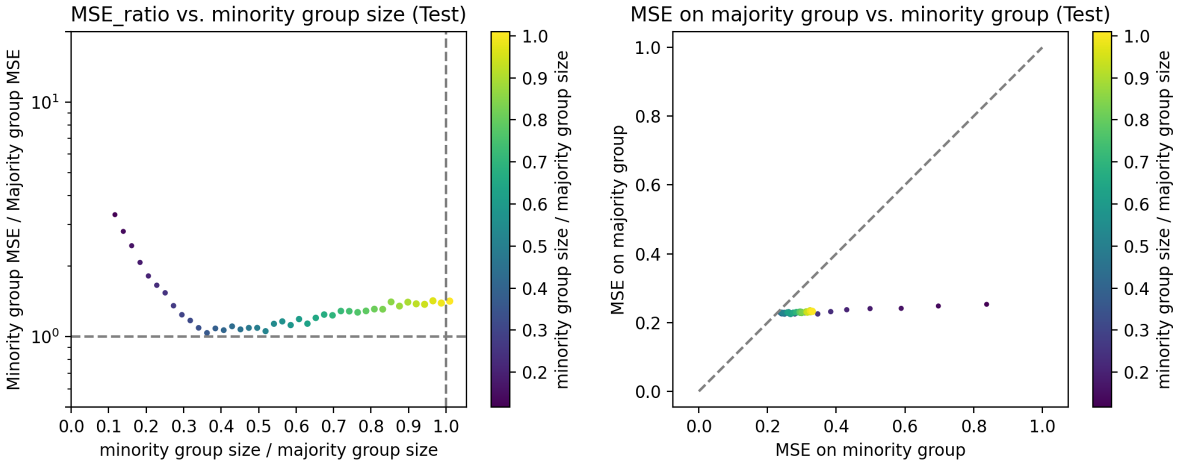
Figure 4: The top-left graph shows the correlation plot among covariates in both the real and fake data sets. The bottom-left plot shows the scatter plot of selected covariates with the outcome variable, where the fake data is overlaid on top of the real data. The plot on the right shows the histogram of selected covariates. The red bars and points indicate fake samples, whereas the blue bars and points indicate real samples.



The generated fake data stays reasonably truthful to the real data distribution. The generator is able to learn the normal and categorical distributions effectively. However, the average of the generated outcome variable Y for the categorical covariates seems to be systematically higher. Yet this shortcoming of the generator does not seem to impede its ability in downstream applications when we use the generated data for augmentation. Since the second generator $\theta_{g,Y(W)|X,W}$ is able to generate both potential outcomes, the generated data comes with an estimation of individual treatment effects. This generated treatment effect and the ground-truth treatment effect share a correlation coefficient of 0.594, which is decent. We still should not use the generator to directly predict individual units' treatment effects. This is because the mean squared error of the generated and ground-truth treatment effects is 1.753, significantly worse than the causal forest or linear model's performance.

Nonetheless, we can take the generated data for which $X^1 = 1$ and augment them to the training data. This would strictly increase the total number of data points for the minority class without affecting the majority class. As more synthetic data is added, we can again plot the performance disparity among the two groups against the amount of data added. Note that this synthetic data augmentation method can be plugged into any CATE estimation method, but we choose to use the causal forest approach. The simulation result is shown in Figure 5. Paradoxically, although the generated treatment effects have higher mean squared errors than causal forests, the synthetic data can still provide useful signals to the causal forest model that improve the performance on the minority group.

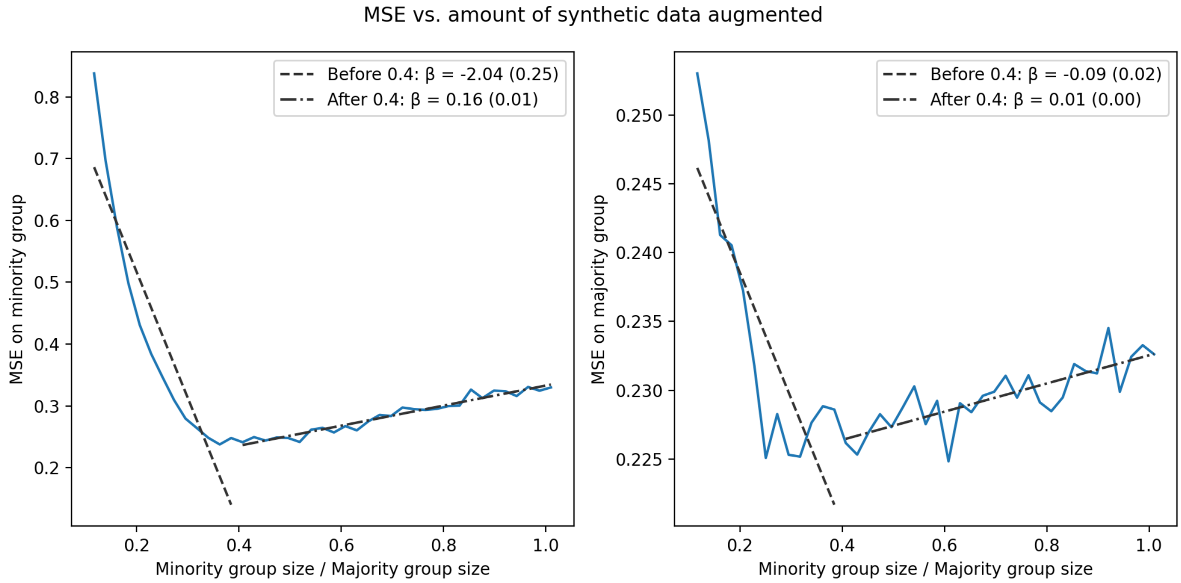
Figure 5: Performance disparity against different amounts of synthetic data augmented using the causal forest approach. The left graph plots the ratio in MSE (MSE on minority / MSE on majority) against the different amounts of synthetic data added. The right graph plots the MSE on the majority group against the MSE on the minority group against the different amounts of synthetic data added. Both plots are performance on the test set.



As more data is added to the minority group, the estimation performance on the minority group first decreases and then increases while the estimation performance on the majority group stays mostly constant. When the minority group size is augmented to be around 0.4 times the majority group size, the estimation accuracy between the two groups is balanced. This demonstrates that the right amount of synthetic data augmentation can effectively balance CATE estimation performance without hurting the performance on the majority group. However, adding too much synthetic data may create too much noise and begin to hurt the performance on both groups. Comparing the specific MSE values to the result on the weighted linear model and weighted causal forest model, this approach achieves the lowest MSE on the minority group while maintaining a relatively low MSE value on the majority group.

We can also look at the impact of synthetic data augmentation more closely. In Figure 6, when synthetic data is added until the minority group size equals 0.4 times the majority group size, the MSE on both groups decreases. In particular, the decrease for the minority group is large, and the decrease for the majority group is more minute. This shows that when the correct amount of synthetic data is added to the minority units, both groups can benefit from Pareto improvements. In addition, even when too much synthetic data is added, the increasing MSEs for both groups are relatively small compared to the initial gain. This shows that synthetic data augmentation methods are reasonably robust regarding the amount of data added.

Figure 6: The left graph plots the MSE on the minority group against the amount of synthetic data augmented. The right graph plots the MSE on the majority group against the amount of synthetic data augmented. Both plots are for the test set. In each plot, two lines of best fit are presented for data points before and after the group sizes ratio reaches 0.4.



Overall, the simulation results show that reweighting with causal forest and synthetic data augmentation through WGAN are two effective methods to address the third research question. Non-parametric machine-learning approaches can indeed enable us to increase performance on the minority group with little or no deterioration on the majority group. The current results for the specific setup in Section 2 imply that the synthetic data generation pipeline has a slight edge. It allows for the possibility of a Pareto improvement and achieves a better overall performance.

5 Application to a Real-World Data Set

As mentioned in Section 1, the broad motivation for fair CATE estimation comes from its application in program evaluation and policy intervention decisions. Thus, we demonstrate the effectiveness of the proposed methods by applying them to a real-world data set. These methods can be applied to arbitrary situations where there is some evidence for treatment heterogeneity and where a variable of interest is imbalanced.

5.1 JTPA Job Training RCT

We choose to use the data from a randomized controlled trial from the National Job Training Partnership Act (JTPA) Study. In 1986, the US Department of Labor commissioned the study to evaluate the benefits and costs of specific employment and training programs for economically disadvantaged adults and out-of-school youths (Bloom et al. 1997). The experiment was conducted for about 20,000 individuals across 16 local JTPA programs. One-third of the participants were randomly assigned to the control group, where they were not allowed to enroll in JTPA programs for 18 months. The other two-thirds were allowed to enroll. Subsequent data like earnings and days until re-employment were collected through unemployment insurance agencies and follow-up surveys.

Since JTPA aims to enhance the skill sets of disadvantaged workers by providing job training and assistance with job searches (Devine and Heckman 1996), its role in the US Labor Market is paramount. As such, it is vitally important to identify people who may benefit the most from such training programs and why they may be ineffective for others. Thus, CATE estimations are beneficial in this setting as they can inform the public employment services about which individuals to prioritize for scarce job training opportunities. Previous studies like (Bloom et al. 1997) and (Sverdrup and Wager 2024) have mostly concluded that JTPA programs are, on average, useful. They have also identified evidence for treatment heterogeneity, where specific subgroups seem to benefit more than others. For example, treatment group members gained modestly more earnings throughout the follow-up period compared to their control group counterparts if they were adults, but the positive effect is almost negligible for youths (Bloom et al. 1997). On the other hand, being assigned to the treatment group positively decreased the amount of time it took for study participants to find a job, and the treatment effect shows clear signs of heterogeneity (Sverdrup and Wager 2024).

For our analysis, we mostly follow Bloom et al. (1997) and consider a subset of the data processed from the raw source, which is included in the link in Appendix A. The outcome variable is the total 30-month earnings during the follow-up period. We include six relevant covariates to estimate the effect of eligibility for JTPA programs on income. These covariates include age, gender, race (White, Black, Hispanic, Native, Asian), high school degree, marital status, and dependent information. Out of these covariates, we found that age is heavily imbalanced, with half of the participants being below 27 years old and only 14.3% at least 40 years old. Detailed descriptive statistics of the age covariate are included in Appendix C.1. Therefore, we pick individuals who are at least 40 years old to be the minority group and everyone who is less than 40 years old to be the majority group. This creates an imbalanced variable, for which the probability of being in the minority group is roughly $p = 14.3\%$.

There are many reasons why policymakers may pay special attention to this vulnerable group of older people. For instance, it is significantly more difficult for older unemployed workers to find new jobs and reintegrate into the labor market (Axelrad et al. 2018). Unemployed older job seekers are more likely to be long-term unemployed and face sharper wage declines (Van Horn et al. 2011), and there are biased stereotypes about older workers

(Axelrad et al. 2013). When CATE-based prioritization rules make more mistakes on the older population (the minority group), it adversely impacts older job seekers and worsens their situation. Therefore, identifying methods that can prevent inferior results for this minority group is paramount.

5.2 RATE

To quantitatively evaluate treatment prioritization rules, we introduce the metric termed Rank-Weighted Average Treatment Effects (RATE) (Yadlowsky et al. 2023). The RATE metrics evaluate the extent to which individuals who are highly ranked by the prioritization rule are more responsive to treatment than random individuals. In fact, in real data sets, researchers do not have access to the ground-truth data-generating process or the CATE function. We would not be able to calculate the exact MSEs on the two subgroups to compare estimation performance. The RATE metric can serve as a proxy that reflects the accuracy of CATE estimations. The RATE metric is more suitable for our setting because it directly evaluates the quality of CATE-induced prioritization rules.

Following Yadlowsky et al. (2023), a prioritization rule is defined using a priority scoring function $S : X \to \mathbb{R}$, where the units are prioritized for treatment in decreasing order of $S(X_i)$. Further denote $F_S(\cdot)$ to be the cumulative distribution function of $S(X_i)$. In our setting, this priority scoring function is $\tau(X_i)$, where individuals with the highest treatment effects are prioritized. RATE metrics are calculated from targeting operator characteristic (TOC). For any prioritization rule with priority score $S(\cdot)$ and any threshold $u \in (0, 1]$ where $F_S^{-1}(u)$ exists, the TOC is defined as:

$$\mathrm{TOC}(u; S) = \mathbb{E}\left[Y_i(1) - Y_i(0) \mid F_S(S(X_i)) \geq 1 - u\right] - \mathbb{E}\left[Y_i(1) - Y_i(0)\right]. \tag{25}$$

Essentially, TOC measures the better-than-average effects of the prioritization rule when treatment is assigned to the top $u\%$ individuals. The next section shows an example of the TOC curve in Figure 7. If the prioritization rule is effective, the top few percentiles should observe much higher treatment effects than average, given treatment heterogeneity.

The RATE metric of a priority score $S(\cdot)$, which is induced by the CATE estimations in this case, is defined as:

$$\theta_\alpha(S) = \int_0^1 \alpha(u)\mathrm{TOC}(u; S)du, \tag{26}$$

where $\alpha : (0, 1] \to \mathbb{R}$ is any weighting function. When $\alpha(u)$ is uniformly 1, the RATE metric is called AUTOC, and when $\alpha(u) = u$, the RATE metric is called Qini.

Empirically, when we do not have access to oracle scores $\mathbb{E}[Y_i(1) - Y_i(0)|X_i]$, we construct scores $\widehat{\Gamma}$ that approximately satisfy Equation (4). One choice is the AIPW estimator in Equation (6) with cross-fitting discussed before. Yadlowsky et al. (2023) shows that the RATE metric can be empirically estimated using

$$\hat{\eta}_{w_\alpha}(S) = \frac{1}{n}\sum_{j=1}^{n} w_\alpha(\frac{j}{n})\widehat{\Gamma}_{i(j)}, \tag{27}$$

where $w_\alpha(t) = \int_t^1 \frac{\alpha(u)}{u}du - \int_0^1 \alpha(u)du$ and $\widehat{\Gamma}_{i(j)}$ is the j^{th}-ranked estimated score.

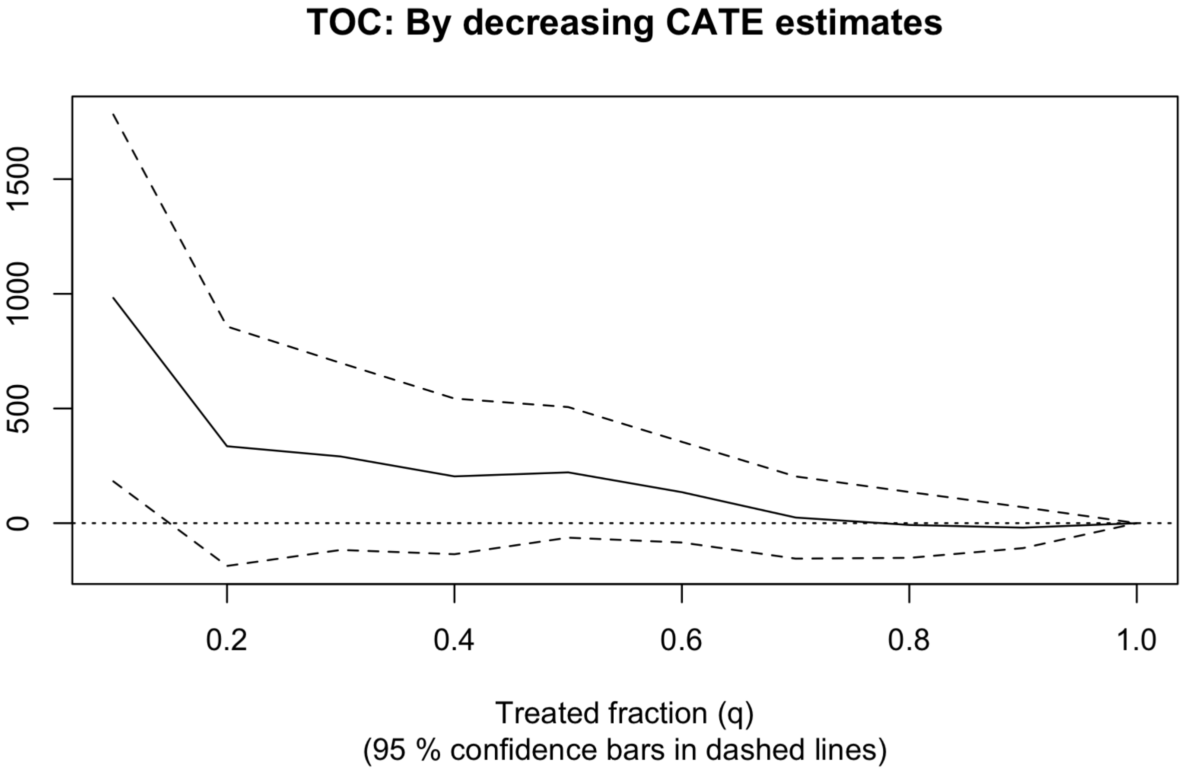
5.3 Results on a Semi-Synthetic JTPA Data Set

The real JTPA data set is somewhat noisy and admits to having a relatively small treatment effect. Score estimation tends to create large approximation errors. As such, we first present results on a semi-synthetic data set that adheres to the distribution of the original JTPA data.

After filtering out missing entries, the JTPA study contains 13198 rows. Inspired by the works of Athey et al. (2021), we feed the entire data set into the Wasserstein Generative Adversarial Networks pipeline outlined in Section 4.2 and train two generators. Then, we can regard the trained generators as the underlying data-generating process. Similar to the result in the simulation studies, the generated data closely resembles the real JTPA dataset, with visualizations shown in Appendix C.2.

We use WGAN to generate large samples $(50, 000)$ for the training and testing sets. We can fit causal forests on the training set and evaluate its CATE-induced prioritization rule on the test set through RATE metrics. First, we show the targeting operator characteristic curve on this semi-synthetic data set as a starting point.

Figure 7: The targeting operator characteristic against different treated fractions. This shows the better-than-average effect when allocating treatment to the top $q\%$ of the population.

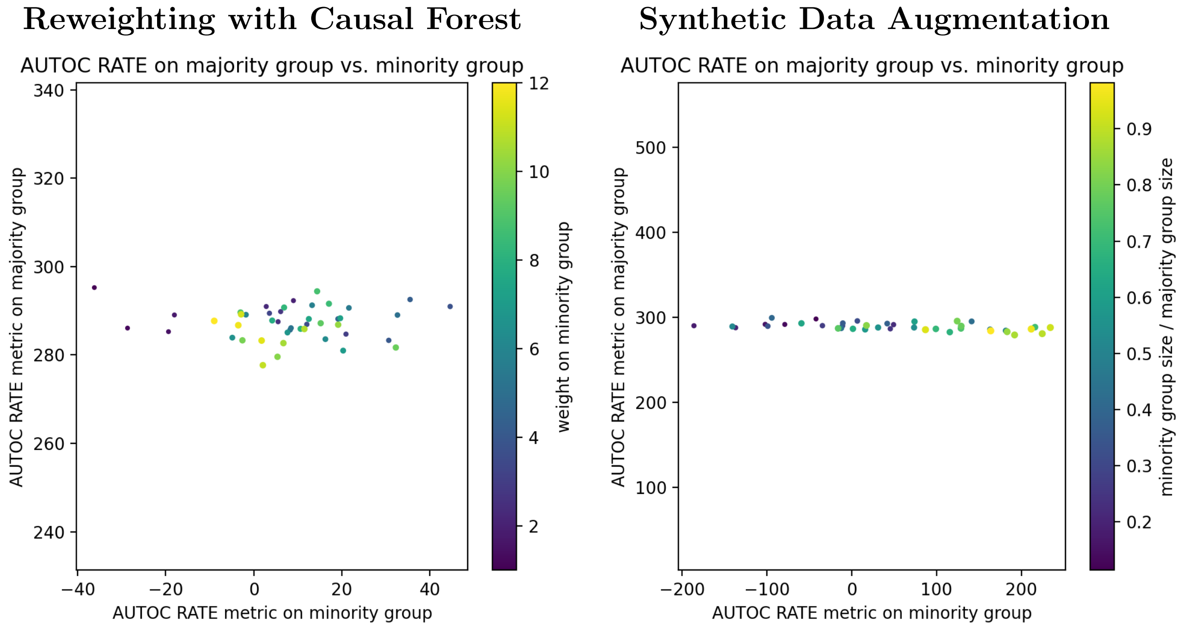


On the semi-synthetic data set, the proportion of units at least 40 years old is $p = 10.0\%$. The average treatment effect is 824.9 dollars with a standard error of 132.0. The 95% confidence interval for the area under the TOC curve (AUTOC) is 312.7 ± 296.1, which does not include 0. These estimates show that the synthetic data has a positive average treatment effect and has treatment heterogeneity. In addition, the TOC curve shows that only a small subset of individuals have a significantly non-zero heterogeneous treatment effect. Following the advice of Yadlowsky et al. (2023), we choose to use AUTOC as the RATE metric. The benefit of the semi-synthetic data set is that we can use oracle scores $Y_i(1) - Y_i(0)$ in RATE metric calculation.

For applications, we focus our analysis on non-parametric methods and omit the linear model. Generally, there is no reason to assume linearity and non-parametric models perform better for complex real-world data sets. Below, we present results on the causal forest reweighting and synthetic data augmentation methods in Figure 8.

As the weight on the minority group increases, the AUTOC metric on the minority

Figure 8: The left graph shows the AUTOC metric on the majority group against the AUTOC metric on the minority group for different weights γ on the minority units. The right graph shows the AUTOC comparisons between the two groups as more synthetic data are added to the minority group. Both methods use the causal forest model, and evaluations are done using the semi-synthetic JTPA test set.


Reweighting with Causal Forest

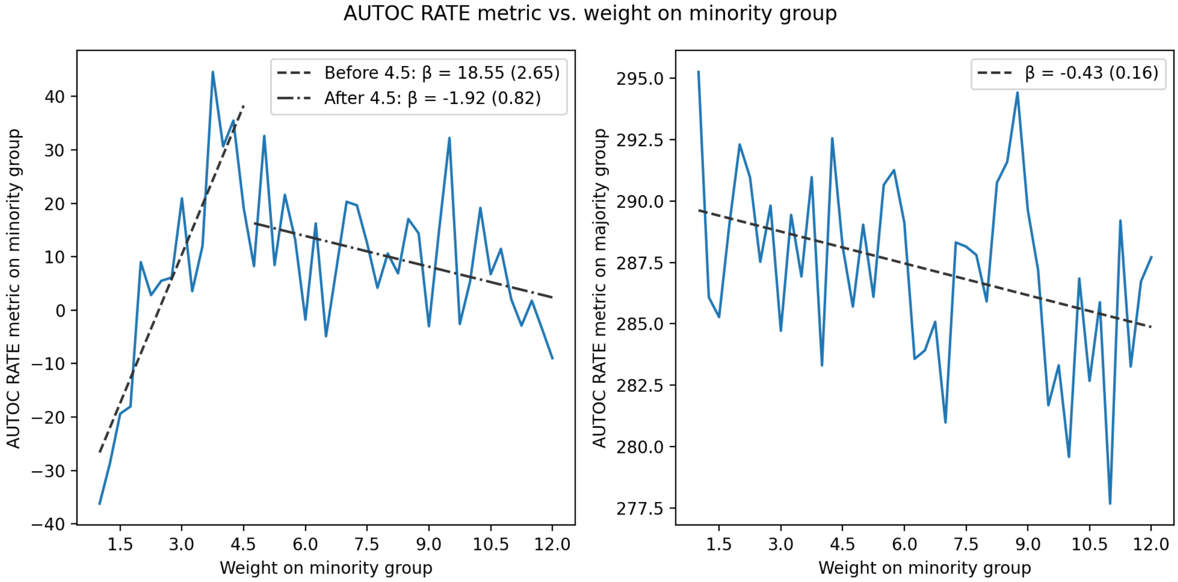
Synthetic Data Augmentation

group first increases and then decreases. The metric on the majority group stays relatively constant. This shows that increasing the weight can benefit the minority group slightly without affecting the majority group. On the other hand, when we re-train WGAN on the semi-synthetic training set and add more synthetic data to the minority group, the AUTOC metric is monotonically improving. In addition, the impact of synthetic data augmented to the minority group has small effects on the majority group. Judging from the immense improvement on the minority group, synthetic data augmentation via WGAN appears to be a superior method.

We first zoom in to the AUTOC plot against minority group weights and study the effects of reweighting more carefully in Figure 9. When the weight is below 4.5, the AUTOC metric on the minority group significantly increases. However, it slowly deteriorates past that point. On the other hand, as more weight is put on the minority group, the AUTOC metric slowly decreases in the majority group, but the decline is almost negligible. This illustrates that a small weight on the minority group can significantly boost the RATE metric on the minority group without hurting the majority group materially.

Still, researchers should be cautious not to add too much weight, as the benefits reverse beyond a specific point.

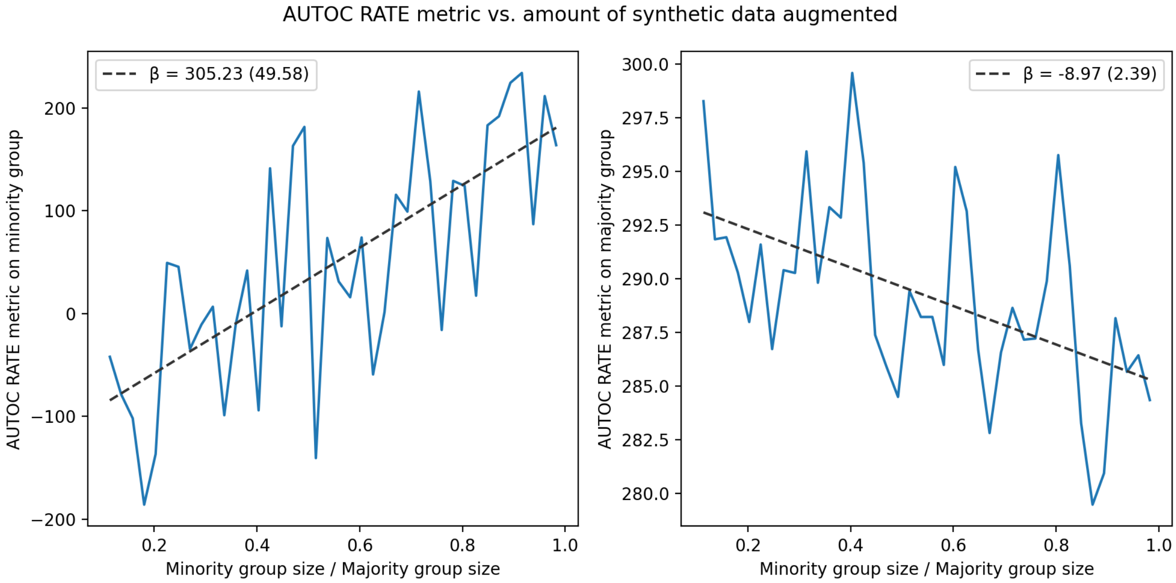
Figure 9: These two plots show the AUTOC metric on the minority and majority groups as more weight is added to the minority units in the causal forest model. For the left graph (minority group), two lines of best fit are shown before and after a chosen turning point. One line of best fit is shown for the right graph (majority group).



Next, we zoom in to the AUTOC plot for synthetic data augmentation in Figure 10. As more synthetic data is added to the minority group, the AUTOC metric monotonically increases from a negative value to almost 200. The AUTOC metric for the majority group slowly declines, but the rate of decline is dwarfed by the rate of improvement for the minority group. As the minority and majority group sizes equalize, the performance measured by AUTOC This incredible gain from synthetic augmentation is perhaps less surprising because the training data is also generated from a WGAN. Thus, it might be easier for WGAN to re-learn the data distribution from the semi-synthetic training samples.

Both of the proposed methods can significantly improve the AUTOC RATE metric on the minority group without notably hurting the majority group. This is beneficial for older people because they can receive higher better-than-average effects through the better CATE-induced prioritization rules. In particular, one should be more careful with the weight being put on minority units, as higher weights can potentially start to worsen

Figure 10: These two plots show the AUTOC metric on the minority and majority groups as more synthetic data is augmented to the minority group in the semi-synthetic JTPA data. In each graph, a line of best fit is shown.



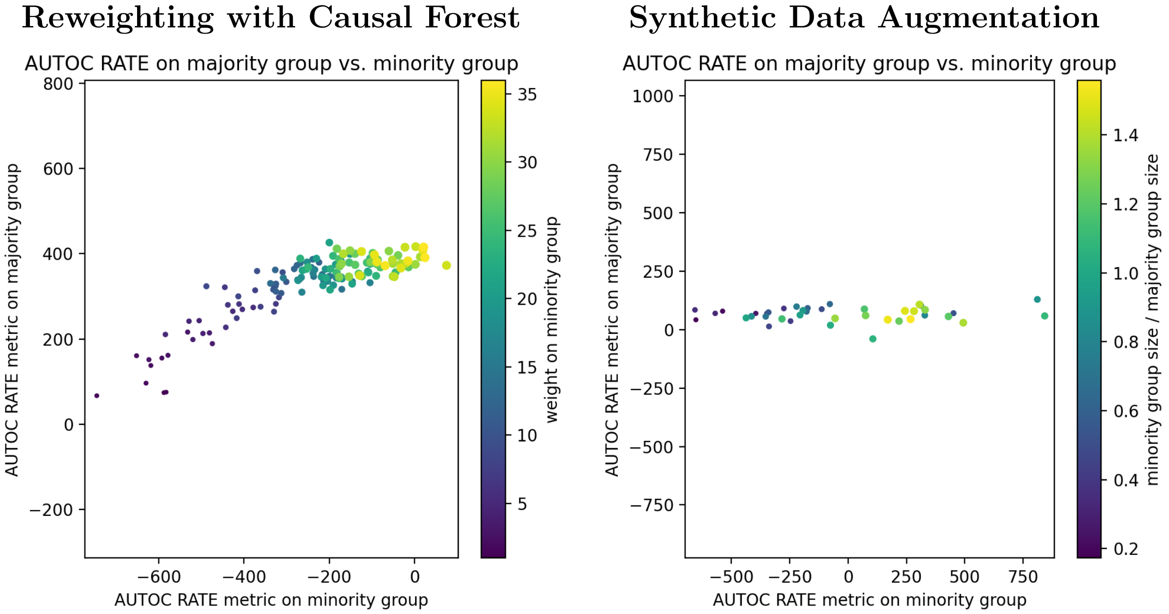
the performance on the minority group.

Overall, we have demonstrated the effectiveness of both methods on a semi-synthetic data set grounded in realism. This again contributes to answering the third research question, where it is possible to balance performance between the two groups without significantly worsening the performance on the majority group. Crucially, the weight added to the minority group and the amount of synthetic data added to the minority units can be seen as hyper-parameters. Researchers can visualize performance as these hyperparameters vary to determine a sensible range and make policy decisions based on the trade-off.

5.4 Real JTPA Data

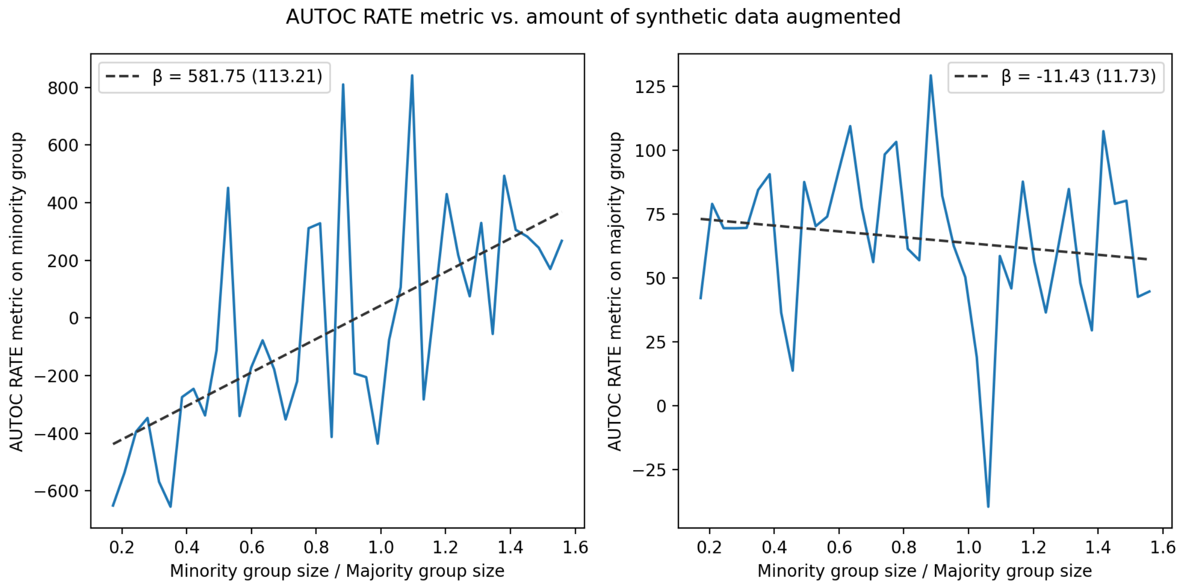
We repeat the same procedure and evaluate the two proposed methods on the real JTPA data set. To estimate RATE metrics on the real data, we must construct scores for the test set according to Equation (6) with cross-fitted nuisance parameters. Note that the constructed scores are held constant irrespective of the methods used. These scores can have large approximation errors compared to ground-truth values. Thus, we should interpret the results in this section with appropriate caution. Figure 11 demonstrates the result on the real JTPA data set.

Figure 11: The left graph shows the AUTOC metric on the majority group against the AUTOC metric on the minority group for different weights γ on the minority units. The right graph shows the AUTOC comparisons between the two groups as more synthetic data are added to the minority group. Both methods use the causal forest model, and evaluations are done on the real JTPA test set.



Surprisingly, the re-weighting method improves performance in both groups, with large improvements for both the minority and majority groups. The AUTOC metric increases from large negative numbers to 0 for the minority group and increases from around 100 to around 400 for the majority group. As more weight is added, the performance on the two groups seems to stabilize. This finding is unintuitive, and more robust tests may be required. One explanation could be that the signal-to-noise ratio in the minority group is somehow higher, and learning in the minority group has positive spillover effects on the majority group. On the other hand, synthetic data augmentation keeps the majority group's AUTOC mostly constant while significantly improving the AUTOC on the minority group. Zooming in on the result of this method in Figure 12, the improvement on the minority group is substantial, and the decline of AUTOC on the majority group is slight (the linear regression coefficient is not significantly different from 0). In fact, synthetic data augmentation can boost the negative RATE on the minority group to a reasonably large positive number.

Figure 12: These two plots show the AUTOC metric on the minority and majority groups as more synthetic data is augmented to the minority group in the real JTPA data. In each graph, a line of best fit is shown.



The result on the real JTPA data set again confirms the effectiveness of the two proposed methods. This demonstrates that researchers who are concerned about an imbalanced covariate should attempt to correct the potentially biased performance on the minority group through these methods.

6 Additional Results and Discussions

For the majority of simulation studies conducted so far in this paper, we rely on the motivating example in Section 2. In this section, we explore other simulation settings and test the limits of the two proposed methods. Since the key focus is on the behavior of the weighted causal forest and synthetic data augmentation methods, we delay other secondary results to the appendix.

6.1 Signal to Noise Ratio

A natural objection to the proposed methods is that we can fit two separate models on the two subgroups, and this might achieve better results than trying to balance one model on both groups. In fact, this can be true when the sample size is large or when the signal-to-noise ratio is high. In those scenarios, learning from the minority group itself may be sufficient. However, in cases where the signal-to-noise ratio is low, which is common in most real-world data sets, this approach of fitting two separate models may not work well.

We test this hypothesis using the simulation setup in Section 2 but we modify Equation (3) to be:

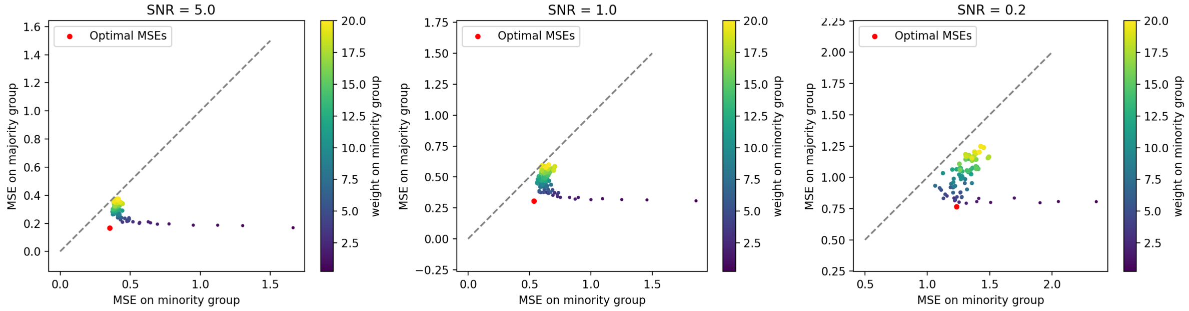
$$Y(X, W) = \max\left(\sum_{i=1}^{6} X^i, 0\right) + W \cdot \tau(X) + \mathcal{N}(0, \sigma^2), \tag{28}$$

where σ^2 controls the noise level in the generated data set. Following Lu and Liu (2022), the signal-to-noise ratio (SNR) is defined as the ratio of the finite-population variance of $\max(\sum_{i=1}^{6} X^i, 0) + W \cdot \tau(X)$ to that of $\mathcal{N}(0, \sigma^2)$. Empirically, the variance of $\max(\sum_{i=1}^{6} X^i, 0) + W \cdot \tau(X)$ is around 4.95 and is similar between the minority and majority groups. Thus, we use $\sigma^2 = 1.0, 5.0, 25.0$ to simulate signal-to-noise ratios that are roughly $5.0, 1.0$, and 0.2.

For the linear model, as SNR decreases, the performance of fitting two separate models quickly deteriorates and becomes much worse than fitting one combined model (even without reweighting). This result is shown in Appendix D.1. Below, we show the results of causal forest with reweighting for varying signal-to-noise ratios in Figure 13.

For large signal-to-noise ratios, the separately-fitted models' performance is strictly better than any point on the trade-off curve of the jointly trained model. However, as the SNR decreases, the performance of the separately-fitted model moves closer to the trade-off curve. In the case of a 0.2 signal-to-noise ratio, the right weight on the minority group with a single model yields a lower MSE on the minority group than the separately-fitted models while maintaining similar performance on the majority group. The separately-fitted models are still performing reasonably well, which may be due to the robustness of the causal forest. However, when the training sample size is reduced from 4000 to 1000 with 0.2 SNR, the separately-fitted models produce worse results than a single forest (shown in Appendix D.1).

Figure 13: The three graphs show the performance on the majority and minority groups using the causal forest model with different minority weights under different SNRs. The red dots plot the MSEs obtained from training two separate causal forest models on the two groups.

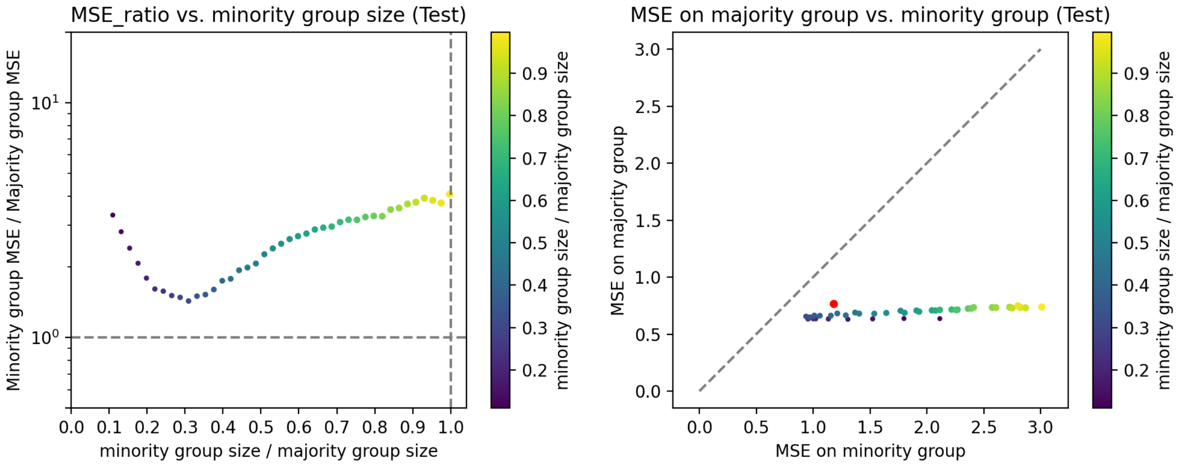


In addition, Figure 13 shows that when the signal-to-noise ratio is low, we need to be more careful with the optimal weight being added to the minority group. A higher-than-necessary weight can cause performance to deteriorate for both groups.

Next, we present the result using synthetic data augmentation when SNR = 0.2 in Figure 14. The generation quality of WGAN is discussed in Appendix D.1. As more synthetic data is added to the minority group, we observe a similar pattern of the MSE dropping initially on the minority group and then increasing. The MSE on the majority group also increases slightly, but at a much slower rate. This illustrates that synthetic data augmentation can still be effective when the signal-to-noise ratio is low, as it almost balances performance. Importantly, this method can achieve performance that is strictly better than fitting two separate models on the two groups.

Comparing this result to the causal forest with reweighting, the lowest MSE achieved on the minority group is much lower. The synthetic data augmentation approach might also be more desirable in the sense that even when an inappropriately large amount of synthetic data is added, the performance on the majority group would not be affected too much. Overall, we have shown that separately training two models on the two subgroups is an inferior approach when the signal-to-noise ratio is low, or the sample size is small, which are likely conditions for real-world data sets. When the signal-to-noise ratio is low, we need to pay more attention to the exact weight on the minority group or the amount of

Figure 14: Performance disparity between the two groups against the amount of synthetic data added when the signal-to-noise ratio is 0.2. The red dot plots the MSEs obtained from training two separate causal forests.



synthetic data added, as higher-than-necessary values lead to deteriorating performance for both groups.

In addition, training two separate models can cause issues downstream when CATE estimations are used for prioritization rules. This is because we cannot easily aggregate the outputs of two models to design priorities for the whole population. For instance, imagine that one of the models is slightly biased to produce higher treatment effects. This would mislead the decision maker to allocate more resources to that group than optimal.

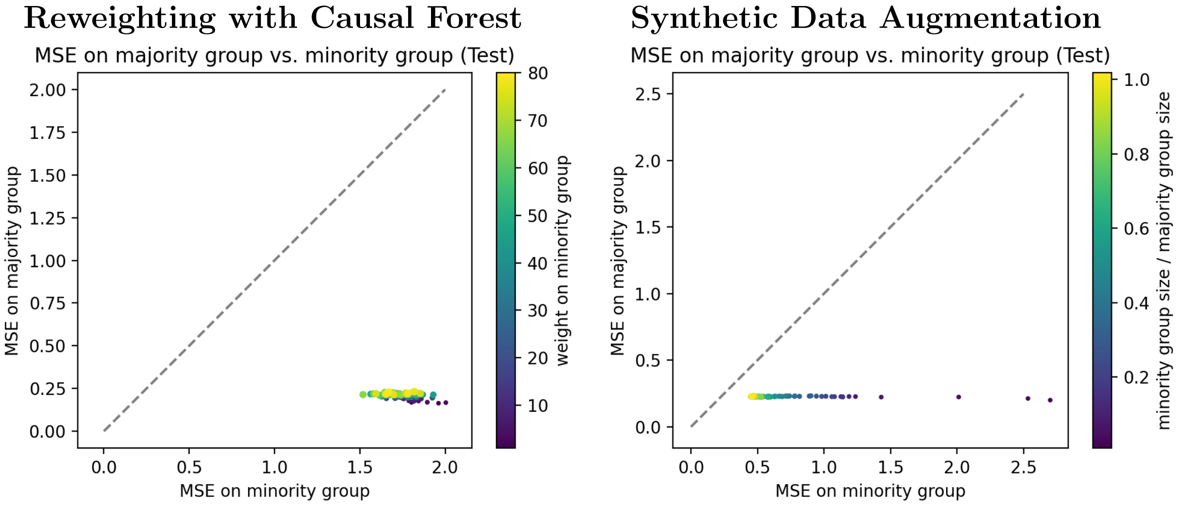
6.2 Extremely Imbalanced Covariates

The examples in the rest of the paper mostly focus on a level of imbalance $p = 0.1$, where the probability of being a minority unit is around 10%. This section considers a more extremely imbalanced case with $p = 0.01$. In the simulation setup where 4000 data points are generated for the training set, only around 40 units would be in the minority group. This presents an extreme case where the amount of information available on the minority group is exceptionally scarce. Using the same setup as Section 2 with $p = 0.01$, we test the robustness of our methods in this setting.

Figure 15 presents the performance of the two proposed methods in this setting.

When small weights are added to the minority group, the MSE on the minority group decreases slightly. However, as more weight is added, the MSE on the minority group hovers between 1.5 and 2.0 and fails to improve below 1.5. In this case, although the performance on the majority group is not affected significantly, the improvement on the minority group is also small.

Figure 15: Results when data is extremely imbalanced ($p = 0.01$). The left graph shows the MSE on the majority group against the MSE on the minority group for different weights γ on the minority units. The right graph shows the MSE comparisons between the two groups as more synthetic data are added to the minority group.



On the other hand, when more synthetic samples are added to the minority group, the performance on the minority group increases significantly while the performance on the majority group stays mostly constant. As more synthetic minority units are added, the minority group MSE drops to around 0.4 and stabilizes there. Surprisingly, even with roughly 40 observations, the WGAN can still accurately learn the underlying distribution and contribute useful information through data augmentation. For visualizations of the WGAN generation quality, please refer to Appendix D.2. In extreme cases where reweighting methods do not work well, Figure 15 is testimony to the robustness of synthetic data augmentation methods. Indeed, with the recent rapidly growing progress in generative modeling, generating synthetic data for augmentation can become an increasingly powerful tool.

6.3 Independent CATE on the Two Groups

In the motivating example in Equation (2), the minority and majority groups share some commonalities in their CATE function. This serves as intuition for the possibility that focusing more on the minority group may not deteriorate the performance on the majority group too much. Essentially, learning from the minority group can potentially also facilitate learning about the majority group.

Then, what about a case where the CATE functions are completely different for the two groups? Would it still be possible to improve on the minority group without harming the majority group? To answer this question, we modify the CATE function below and run the same simulation studies with $p = 0.1$.

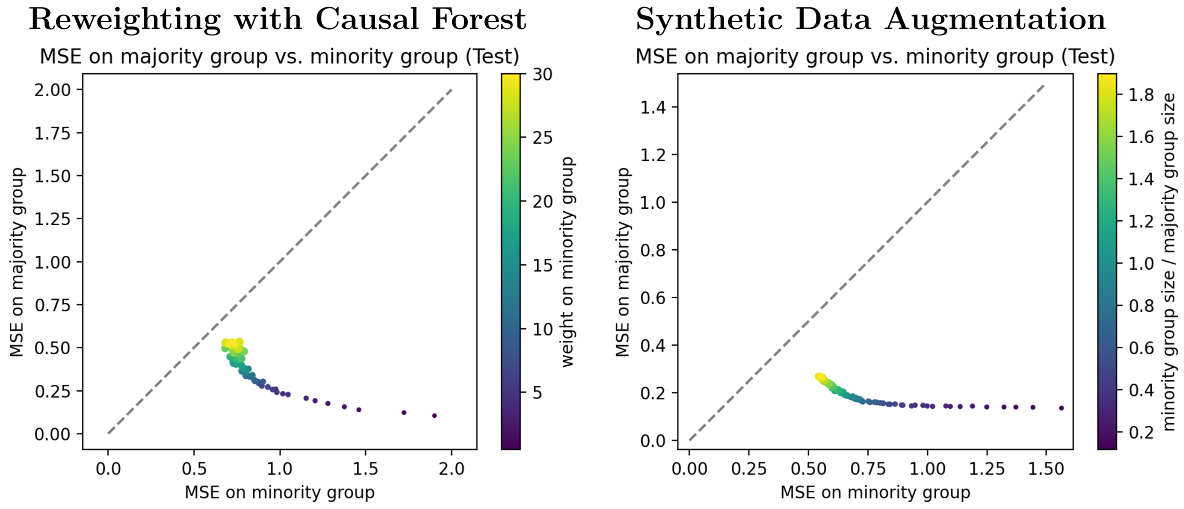
$$\tau(X) = \begin{cases} X^2 + X^4, & X^1 = 1 \\ X^3 + X^5, & X^1 = 0 \end{cases} \tag{29}$$

Unsurprisingly, the result of the weighted linear model using AIPW scores and the performance disparity for different levels of imbalance is almost identical to the result in Figure 1. This is because the CATE is still locally linear for the two groups. As such, we omit these results. The results for the two non-parametric methods are shown in Figure 16.

Comparing with Figure 3 and Figure 5, we do observe more trade-offs when the CATE is independent across the two groups. This is more apparent in the reweighting with causal forest case, where increasing minority weights monotonically increases MSE on the majority group. In the synthetic data augmentation case, when the added data is less than the number of majority units, the MSE decreases significantly for the minority group without affecting the majority group. However, when too much synthetic data is added, the performance on the majority units begins to deteriorate. Again, note that we can obtain better results with the data augmentation method.

Although both methods reduce MSE on the minority group drastically, it is more difficult to avoid a trade-off. In the context of research question 3, reweighting with

Figure 16: Results when the CATE is completely independent between the two groups. The left graph shows the MSE on the majority group against the MSE on the minority group for different weights γ on the minority units. The right graph shows the MSE comparisons between the two groups as more synthetic data are added to the minority group.



Reweighting with Causal Forest

Synthetic Data Augmentation

causal forest no longer works well when the CATE between two groups is independent. The synthetic data augmentation method still produces satisfactory results.

6.4 Complex Nonlinear CATE Function

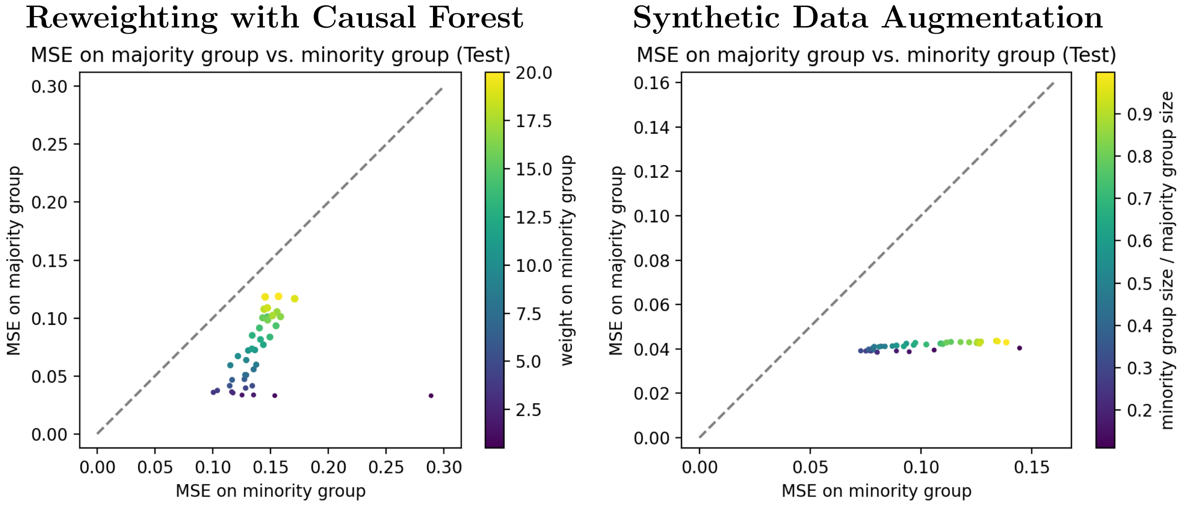
So far, we have worked mostly with locally linear CATE functions, apart from the application to the JTPA study. In this section, we explicitly consider a complicated nonlinear CATE inspired from Künzel et al. (2019). The underlying data generation process is still the same as Section 2 except the CATE function is defined using:

$$\mu_1(X_i) = \begin{cases} \frac{1}{2}\zeta(X_i^2)\zeta(X_i^4), & X_i^1 = 1 \\ \frac{1}{2}\zeta(X_i^2)\zeta(X_i^5), & X_i^1 = 0 \end{cases} \qquad \mu_0(X_i) = -\frac{1}{2}\zeta(X_i^2)\zeta(X_i^4), \tag{30}$$

where $\zeta(x) = \frac{2}{1+\exp(-12(x-\frac{1}{2}))}$, $Y_i(1) = \mu_1(X_i) + \mathcal{N}(0,1)$, and $Y_i(0) = \mu_0(X_i) + \mathcal{N}(0,1)$. The CATE function is thus equal to $\mu_1(X_i) - \mu_0(X_i)$, which is a complicated nonlinear function.

As expected, linear regression with AIPW scores creates larger-than-necessary errors in this case, as shown in Appendix D.4. Below, Figure 17 shows the result of the two proposed methods. Both methods are quite sensitive to the amount of weight or the amount of synthetic data added. Small values improve the MSE on the minority group without hurting the majority group's performance. However, larger values begin to deteriorate the performance on the minority group. Again, even when too much data is added, the MSE on the majority group does not increase significantly. On the contrary, reweighting with causal forests can significantly worsen the performance on the majority group.

Figure 17: Results when the CATE is a complicated nonlinear function. The left graph shows the MSE on the majority group against the MSE on the minority group for different weights γ on the minority units. The right graph shows the MSE comparisons between the two groups as more synthetic data are added to the minority group.



In complicated settings, both methods require researchers to pick the best hyperparameter – weight or amount of synthetic data augmented. With the correct choice, we can usually improve minority group performance by a notable amount without hurting the majority group.

7 Conclusion

In this paper, we tackle the under-explored problem of CATE estimation in the presence of an imbalanced covariate. When the CATE is heterogeneous across the minority and majority groups, we posit that standard methods produce CATE estimations that are substantially less accurate on the minority units. Since CATE induces prioritization rules for policy assignment, this performance disparity can be inequitable for people in the minority group. Through theoretical derivations in a weighted linear regression model, we gain deeper insights into the effect of misspecification on estimation performance disparity. By adding more weight on the minority units, weighted least squares with AIPW scores can improve performance on the minority group at the cost of deteriorating accuracy on the majority group.

In an effort to improve upon this baseline, we turn to non-parametric methods like the causal forest model. Inspired by previous literature, we adapt two popular approaches to this setting – reweighting and synthetic data augmentation. Through extensive simulation studies, we demonstrate the effectiveness and versatility of these two methods. Under all the circumstances explored in this paper, setting the right weight and adding the right amount of synthetic data on minority units always produce gains for the minority group with zero or negligible effect on the majority group. In many situations, the improvement for minority units is substantial.

However, in difficult situations where the signal-to-noise ratio is small, or the CATE function is complicated, one needs to be more careful about the chosen weight or the amount of synthetic data added. After specific thresholds, higher weight or more data added can cause performance to deteriorate for both groups. To this end, synthetic data augmentation seems to be more robust. Even when it may hurt minority group performance beyond a specific point, it rarely significantly worsens the performance on the majority group. In fact, in some edge cases like extreme imbalance and independent CATE functions where the reweighting method struggles, synthetic data augmentation via WGAN still performs favorably. Compared to reweighting with causal forest, synthetic data augmentation almost always produces superior results in the sense that it can obtain better performance on the minority group while maintaining comparable performance on the majority group. As such, researchers may prefer the latter approach in most contexts. The downsides to training a WGAN for generating synthetic data are the loss of statistical properties, extra time, and computational resource requirements.

Finally, we apply these two methods to a real-world data set in the context of job training programs and evaluate their performance through RATE metrics. We provide two realistic ways for assessing CATE estimation performance on real data – creating a semi-synthetic data set or constructing scores. Our results show that both methods can significantly improve the minority group performance without hurting the majority group substantially. Crucially, our methods equip policymakers with a hyperparameter (weight or amount of added data) that controls the potential "trade-off" between the two groups. This empowers decision-makers to weigh fine-grained considerations based on specific fairness and welfare constraints to determine the optimal point.

In conclusion, the proposed methods effectively tackle the issues of CATE estimation with an imbalanced covariate, striking fairness and balance between the two groups.

7.1 Future Research

Even though this paper proposes two effective methods for handling imbalanced covariates, more remains to be researched. First of all, more real-data evaluations should be conducted to test the robustness of both methods. In particular, we should study if the same performance patterns are observed across a diverse range of data sets. To this end, the proposed methods extend naturally to observational studies without confounding, and we should test with observational settings. As an extension to this problem setup, we may also consider the case where multiple covariates of interest are imbalanced.

So far, the proposed method's weighting scheme is uniform, with the same weights for minority units and the same weights for majority units. An improved version could involve a more dynamic weighting scheme. For instance, rarer occurrences in the minority group could be weighted even more than other units. In addition, we can even merge both approaches and use weighting with synthetic data augmentation. For instance, we may be motivated to downweight the synthetic samples.

With the rapid progress in generative models, we can experiment with other generation methods like Neal et al. (2020) and Qian et al. (2023) to compare with the WGAN implementation. Recently, diffusion models (Ho et al. 2020) have demonstrated incredible generation capabilities, and it would be interesting to see how they perform. Lastly, we can also investigate a subsample and ensemble approach. We can train causal trees on balanced (in terms of minority and majority groups) subsets of the training data and aggregate them into a causal forest.

In summary, future research should be conducted to study the generality of the proposed methods, and experiments should be conducted in other novel directions.

References

Abadie, Alberto and Matias D. Cattaneo. 2018. "Econometric Methods for Program Evaluation," *Annual Review of Economics*, *10* (1): 465–503.

Abadie, Guido W. Imbens Alberto and Fanyin Zheng. 2014. "Inference for Misspecified Models With Fixed Regressors," *Journal of the American Statistical Association*, *109* (508): 1601–1614.

Arjovsky, Martin, Soumith Chintala, and Léon Bottou. 2017. "Wasserstein Generative Adversarial Networks," in Doina Precup and Yee Whye Teh, eds., *Proceedings of the 34th International Conference on Machine Learning*, Vol. 70 of *Proceedings of Machine Learning Research* PMLR 06–11 Aug pp. 214–223.

Athey, Susan and Guido W Imbens. 2015. "Machine learning methods for estimating heterogeneous causal effects," *stat*, *1050* (5): 1–26.

_ **and Guido W. Imbens.** 2019. "Machine Learning Methods That Economists Should Know About," *Annual Review of Economics*, *11* (1): 685–725.

_ **and Stefan Wager.** 2019. "Estimating treatment effects with causal forests: An application," *Observational studies*, *5* (2): 37–51.

_ **, Guido W Imbens, Jonas Metzger, and Evan Munro.** 2021. "Using wasserstein generative adversarial networks for the design of monte carlo simulations," *Journal of Econometrics:* p. 105076.

_ **, Julie Tibshirani, and Stefan Wager.** 2019. "Generalized Random Forests," *The Annals of Statistics*, *47* (2): 1148–1178.

Axelrad, Hila, Israel Luski, and Miki Malul. 2013. "Difficulties of integrating older workers into the labor market: Exploring the Israeli labor market," *International Journal of Social Economics*, 10 *40.*

_ , Miki Malul, and Israel Luski. 2018. "Unemployment among younger and older individuals: does conventional data about unemployment tell us the whole story?," *Journal for Labour Market Research*, *52* (1): 3.

Bloom, Howard S., Larry L. Orr, Stephen H. Bell, George Cave, Fred Doolittle, Winston Lin, and Johannes M. Bos. 1997. "The Benefits and Costs of JTPA Title II-A Programs: Key Findings from the National Job Training Partnership Act Study," *The Journal of Human Resources*, *32* (3): 549–576.

Breiman, Leo. 2001. "Random forests," *Machine learning*, *45:* 5–32.

Breusch, T. S. and A. R. Pagan. 1979. "A Simple Test for Heteroscedasticity and Random Coefficient Variation," *Econometrica*, *47* (5): 1287–1294.

Chawla, Nitesh V, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. "SMOTE: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, *16:* 321–357.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. "Double/debiased machine learning for treatment and structural parameters: Double/debiased machine learning," *The Econometrics Journal*, *21* (1).

Devine, Theresa J. and James J. Heckman. 1996. "The Economics of Eligibility Rules for a Social Program: A Study of the Job Training Partnership Act (JTPA)–A Summary Report," *The Canadian Journal of Economics / Revue canadienne d'Economique*, *29:* S99–S104.

Fithian, William and Trevor Hastie. 2014. "Local case-control sampling: Efficient subsampling in imbalanced data sets," *The Annals of Statistics*, oct *42* (5).

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. "Generative adversarial nets," *Advances in neural information processing systems*, *27.*

Hainmueller, Jens. 2012. “Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies,” *Political Analysis*, *20* (1): 25–46.

Hashemi, Mahdi and Hassan Karimi. 2018. “Weighted machine learning,” *Statistics, Optimization and Information Computing*, *6* (4): 497–525.

Hill, Jennifer L. 2011. “Bayesian Nonparametric Modeling for Causal Inference,” *Journal of Computational and Graphical Statistics*, *20* (1): 217–240.

Hirano, Keisuke and Jack R. Porter. 2009. “Asymptotics for Statistical Treatment Rules,” *Econometrica*, *77* (5): 1683–1701.

_ , Guido W. Imbens, and Geert Ridder. 2003. “Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score,” *Econometrica*, *71* (4): 1161–1189.

Ho, Jonathan, Ajay Jain, and Pieter Abbeel. 2020. “Denoising diffusion probabilistic models,” *Advances in neural information processing systems*, *33:* 6840–6851.

Holland, Paul W. 1986. “Statistics and Causal Inference,” *Journal of the American Statistical Association*, *81* (396): 945–960.

Horn, Carl E Van, Nicole Corre, and Maria Heidkamp. 2011. “Older workers, the great recession, and the impact of long-term unemployment,” *Public Policy & Aging Report*, *21* (1): 29–33.

Imai, Kosuke and Marc Ratkovic. 2014. “Covariate balancing propensity score,” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *76* (1): 243–263.

Imbens, Guido W. and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, Cambridge University Press.

_ and Jeffrey M. Wooldridge. 2009. “Recent Developments in the Econometrics of Program Evaluation,” *Journal of Economic Literature*, *47* (1): 5–86.

Kim, Kwangho and José R Zubizarreta. 2023. "Fair and robust estimation of heterogeneous treatment effects for policy learning," in "International Conference on Machine Learning" PMLR pp. 16997–17014.

Kitagawa, Toru and Aleksey Tetenov. 2018. "Who Should Be Treated? Empirical Welfare Maximization Methods for Treatment Choice," *Econometrica*, *86* (2): 591–616.

Künzel, Sören R., Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. 2019. "Metalearners for estimating heterogeneous treatment effects using machine learning," *Proceedings of the National Academy of Sciences*, feb *116* (10): 4156–4165.

Lei, Lihua, Roshni Sahoo, and Stefan Wager. 2023. "Policy learning under biased sample selection," *arXiv preprint arXiv:2304.11735.*

Liang, Annie, Jay Lu, and Xiaosheng Mu. 2021. "Algorithm design: A fairness-accuracy frontier," *arXiv preprint arXiv:2112.09975.*

Lu, Xin and Hanzhong Liu. 2022. "Tyranny-of-the-minority regression adjustment in randomized experiments," *arXiv preprint arXiv:2210.00261.*

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

Mantel, Nathan and William Haenszel. 1959. "Statistical Aspects of the Analysis of Data From Retrospective Studies of Disease," *JNCI: Journal of the National Cancer Institute*, 04 *22* (4): 719–748.

Morgan, Stephen L. and Christopher Winship. 2014. *Counterfactuals and Causal Inference: Methods and Principles for Social Research* Analytical Methods for Social Research, 2 ed., Cambridge University Press.

Neal, Brady, Chin-Wei Huang, and Sunand Raghupathi. 2020. "Realcause: Realistic causal inference benchmarking," *arXiv preprint arXiv:2011.15007.*

Pearl, Judea. 2000. *Causality: Models, Reasoning, and Inference*, Cambridge University Press.

Qian, Zhaozhi, Bogdan-Constantin Cebere, and Mihaela van der Schaar. 2023. "Synthcity: facilitating innovative use cases of synthetic data in different data modalities," *arXiv preprint arXiv:2301.07573.*

Robins, James M., Andrea Rotnitzky, and Lue Ping Zhao. 1994. "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed," *Journal of the American Statistical Association*, *89* (427): 846–866.

Sverdrup, Erik and Stefan Wager. 2024. "Treatment heterogeneity with right-censored outcomes using grf," *arXiv preprint arXiv:2312.02482.*

Viviano, Davide and Jelena Bradic. 2024. "Fair policy targeting," *Journal of the American Statistical Association*, *119* (545): 730–743.

Wager, Stefan and Susan Athey. 2018. "Estimation and inference of heterogeneous treatment effects using random forests," *Journal of the American Statistical Association*, *113* (523): 1228–1242.

White, Halbert. 1980. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, *48* (4): 817–838.

Yadlowsky, Steve, Scott Fleming, Nigam Shah, Emma Brunskill, and Stefan Wager. 2023. "Evaluating Treatment Prioritization Rules via Rank-Weighted Average Treatment Effects," *arXiv preprint arXiv:2111.07966.*

Yang, Yuzhe, Kaiwen Zha, Yingcong Chen, Hao Wang, and Dina Katabi. 2021. "Delving into deep imbalanced regression," in "International conference on machine learning" PMLR pp. 11842–11851.

Appendix

A Replication Code

To augment the usefulness of our findings and methodology, we have made all the replication code publicly accessible. `github.com/Frankz24/CATE-Imbalanced-Covariate/` contains all codebooks for running the simulations, processing the data, training the WGANs, and applying the proposed methods to a real data set.

B Supplementary Material for Theoretical Derivations

B.1 Intermediate Calculation Step

The matrix $\mathbb{E}[G] = \mathbb{E}[\frac{1}{n}X^\top AX]$ is

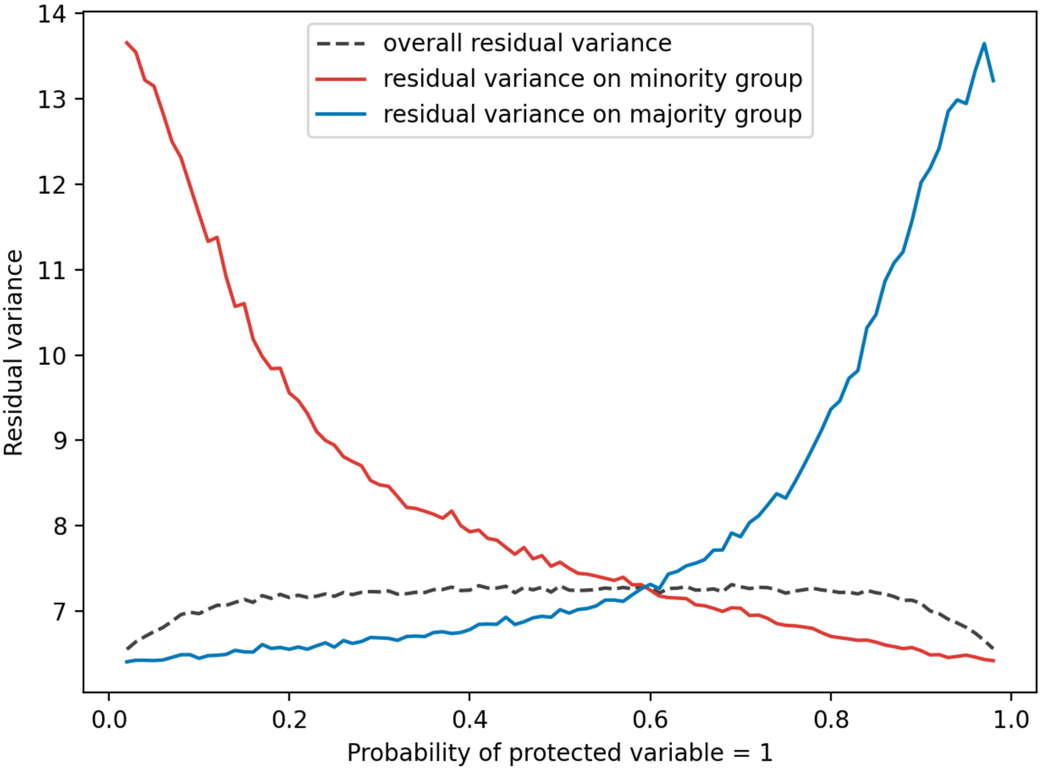
$$\begin{pmatrix}
p\gamma & \frac{p\gamma}{2} & \frac{p\gamma}{2} & 0 & 0 & 0 & p\gamma \\
\frac{p\gamma}{2} & \frac{p\gamma+(1-p)}{2} & \frac{p\gamma+(1-p)}{4} & 0 & 0 & 0 & \frac{p\gamma+(1-p)}{2} \\
\frac{p\gamma}{2} & \frac{p\gamma+(1-p)}{4} & \frac{p\gamma+(1-p)}{2} & 0 & 0 & 0 & \frac{p\gamma+(1-p)}{2} \\
0 & 0 & 0 & p\gamma+(1-p) & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & p\gamma+(1-p) & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & p\gamma+(1-p) & 0 \\
p\gamma & \frac{p\gamma+(1-p)}{2} & \frac{p\gamma+(1-p)}{2} & 0 & 0 & 0 & p\gamma+(1-p)
\end{pmatrix}.$$

B.2 Residual Variance of Constructed Score

To test the heteroscedasticity of $\widehat{\Gamma} - \tau(X)$, we run a simulation study using the data generating process outlined in Section 2. We use 4000 independently generated samples. We construct the scores $\widehat{\Gamma}$ using the AIPW estimator Equation (6), where the nuisance parameters are cross-fitted using regression forests (Breiman 2001).

After constructing the scores, the residuals $\widehat{\Gamma} - \tau(X)$ are recorded. We repeat the simulation for different values of p, the probability of being in the minority group. For each value of p, 50 simulation runs are conducted, and the calculated variances are averaged over all 50 simulations. The overall residual variance, the residual variance on the minority group, and the residual variance on the majority group for different values of p are shown in Figure A1. As we can see, when the data set is imbalanced, the variance of $\widehat{\Gamma} - \tau(X)$ is significantly larger on the minority group than on the majority group.

A1: The red line shows the residual variance of the constructed scores on the minority group against different values of p. The blue line shows the residual variance on the majority group, and the black line shows the overall residual variance.



Further, notice that the scale of the overall variance of $\widehat{\Gamma} - \tau(X)$ is around 7. The variance of $(\tau(X) - X\beta_{\gamma}^{\text{pop}})$, on the other hand, is strictly between 0 and 1. To see why, note that $\text{Var}(\tau(X) - X\beta_{\gamma}^{\text{pop}}) = \text{Var}((\frac{1-p}{p\gamma+1-p})X^5) = (\frac{1-p}{p\gamma+1-p})^2\text{Var}(X^5) = (\frac{1-p}{p\gamma+1-p})^2$ for the minority group, which is strictly below 1. Similarly, $\text{Var}(\tau(X) - X\beta_{\gamma}^{\text{pop}}) = \frac{p\gamma}{p\gamma+1-p})^2$, is also between 0 and 1. This illustrates that the component $\widehat{\Gamma} - \tau(X)$ is dominating in the residuals ε in Equation (15).

B.3 Calculation for $\mathbf{V}_A, \mathbf{V}_B$

To calculate $\mathrm{V}_A, \mathrm{V}_B$ in the expected MSE, we first calculate the Huber-White variance estimator in Equation (15). Given the variance-covariance matrix of the coefficients,

$$\begin{aligned}
\mathrm{V}_A &= \mathrm{Var}(\beta_0) + 2\mathrm{Cov}(\beta_0, \beta_1) + \mathrm{Cov}(\beta_0, \beta_2) + \mathrm{Cov}(\beta_0, \beta_3) + \mathrm{Var}(\beta_1) + \mathrm{Cov}(\beta_1, \beta_2) \\
&\quad + \mathrm{Cov}(\beta_1, \beta_2) + \frac{1}{2}\mathrm{Var}(\beta_2) + \frac{1}{4}\mathrm{Cov}(\beta_2, \beta_3) + \frac{1}{2}\mathrm{Var}(\beta_3) + \mathrm{Var}(\beta_4) + \mathrm{Var}(\beta_5) + \mathrm{Var}(\beta_6), \\
\mathrm{V}_B &= \mathrm{Var}(\beta_0) + \mathrm{Cov}(\beta_0, \beta_2) + \mathrm{Cov}(\beta_0, \beta_3) + \frac{1}{2}\mathrm{Var}(\beta_2) + \frac{1}{4}\mathrm{Cov}(\beta_2, \beta_3) + \frac{1}{2}\mathrm{Var}(\beta_3) \\
&\quad + \mathrm{Var}(\beta_4) + \mathrm{Var}(\beta_5) + \mathrm{Var}(\beta_6),
\end{aligned}$$

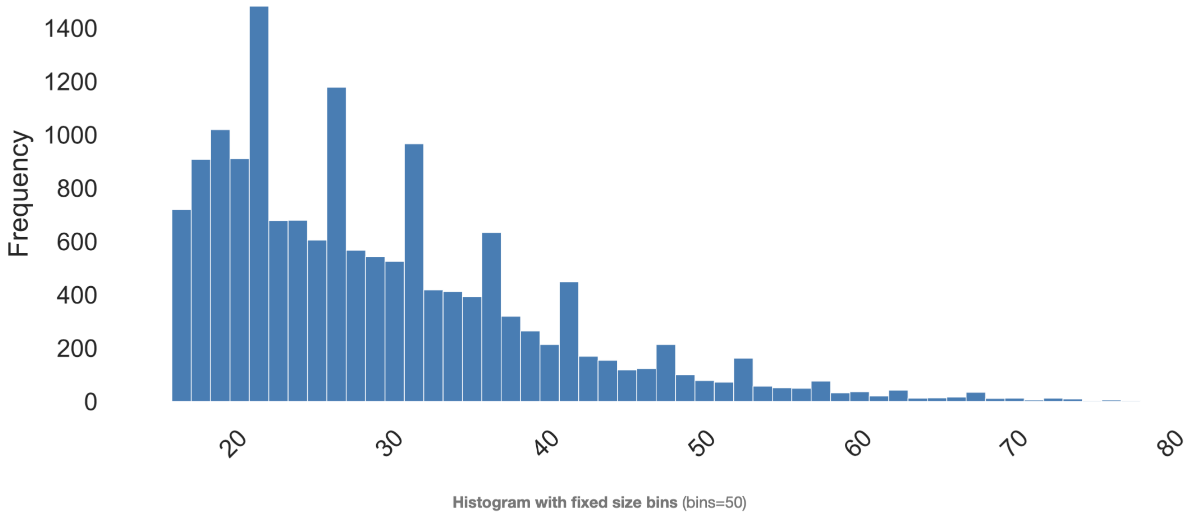
where β_0 is constant term and β_1 to β_6 are coefficients of X^1 to X^6 respectively.

C Application to the JTPA Study

C.1 The "Age" Covariate

The "age" covariate in the JTPA data set is imbalanced, with the 25th percentile being 21 years old, the median being 27 years old, and the 75th percentile being 35 years old. Only 14.3% of the sample is at least 40 years old. The full histogram of the "age" covariate is shown in Figure A2.

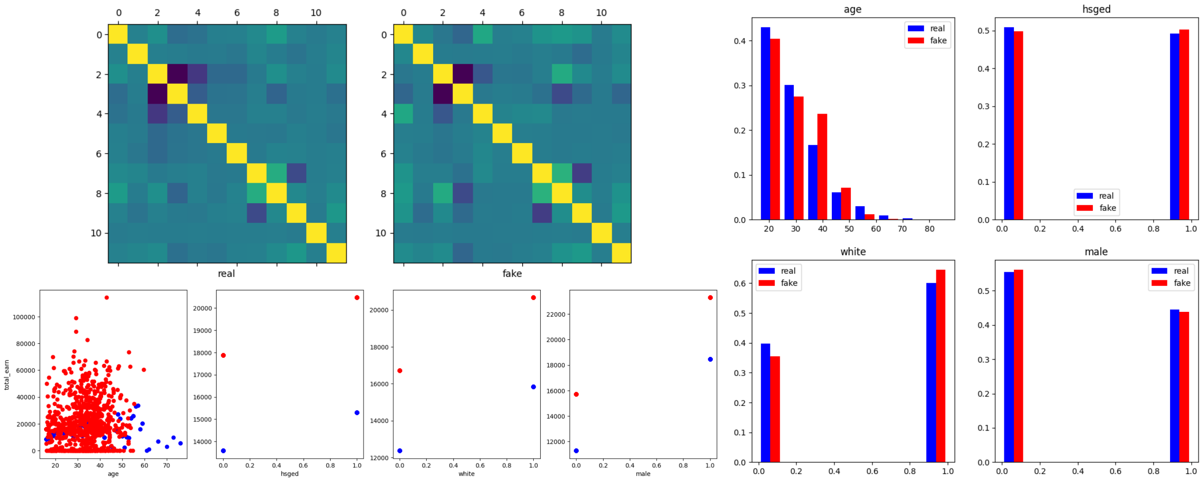
A2: This histogram shows the frequency of different age groups in the JTPA data set.



C.2 Semi-Synthetic JTPA Data Set

Below, we compare the real JTPA data set and the generated semi-synthetic data set. The generated data captures the complex dependencies among covariates and their underlying distributions well. Again, we observe that the fake data consistently outputs a higher outcome variable. However, since this effect is visible for both control and treatment units, it would not affect CATE estimation too much.

A3: The top-left graph shows the correlation plot among covariates in both the real and fake data sets. The bottom-left plot shows the scatter plot of selected covariates with the outcome variable, where the fake data is overlaid on top of the real data. The plot on the right shows the histogram of selected covariates. The red bars and points indicate fake samples, whereas the blue bars and points indicate real samples.



D Additional Simulation Results

D.1 Signal to Noise Ratio

The following Figure A4 shows the performance of the weighted linear model using the constructed AIPW scores when the signal-to-noise ratio varies in the data set. When the SNR is large, fitting two models is, in fact, desirable. This is not surprising as the CATE function is linear for the two groups separately. However, as the signal-to-noise ratio decreases, the MSEs obtained from fitting two models become worse than fitting one model. In fact, when SNR is 0.2, the single weighted linear regression can achieve much lower MSE on the minority group while maintaining similar performance on the majority group.

A4: The three graphs show the performance on the majority and minority groups using the weighted linear model with different minority weights under different SNRs. The red dots plot the MSEs obtained from training two separate linear models on the two groups.

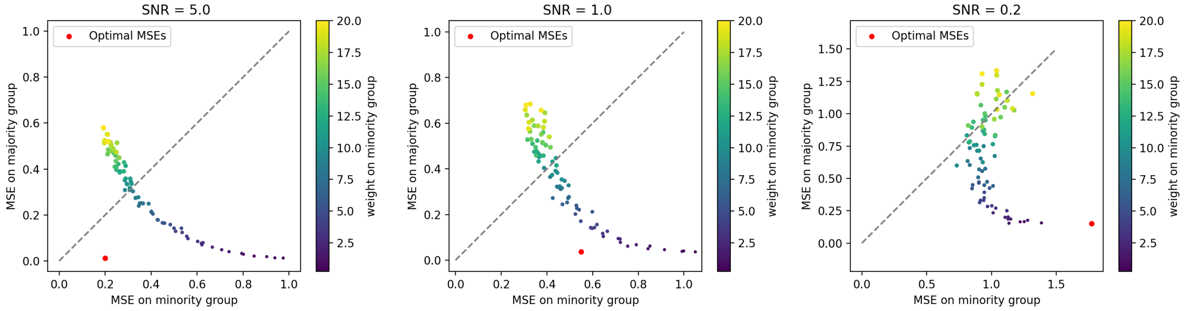


Figure A5 shows the performance of the causal forest with different minority weights for 0.2 SNR and a smaller training set of size 1000. In this case, a single causal forest with the appropriate minority weight can achieve much better results than fitting two separate models. Still, one needs to be careful with the final chosen weight, as a large weight can worsen both groups.

A5: Performance disparity against different minority weights γ using the causal forest approach when the SNR is 0.2 and the training sample size is 1000. The red dot plots the MSEs obtained from training two separate causal forests.

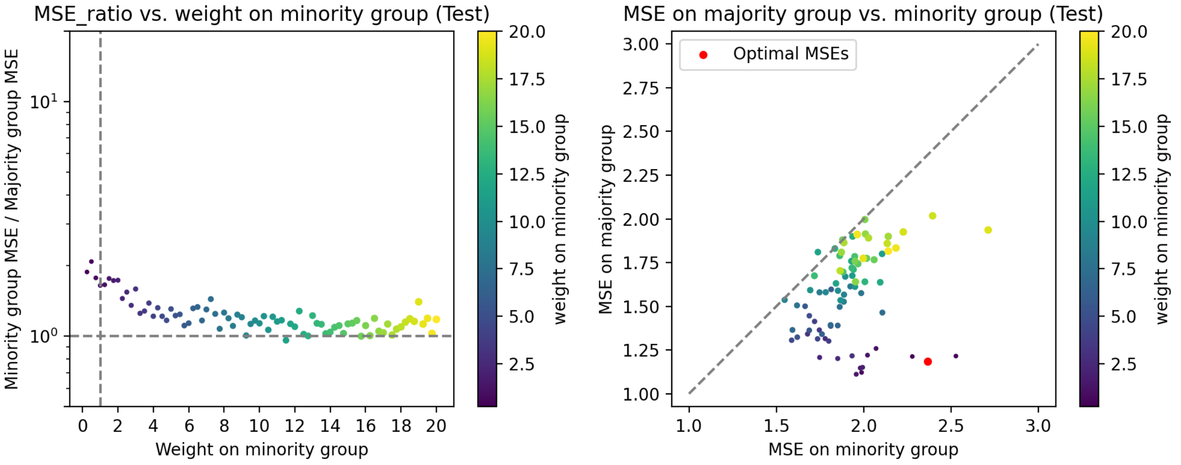
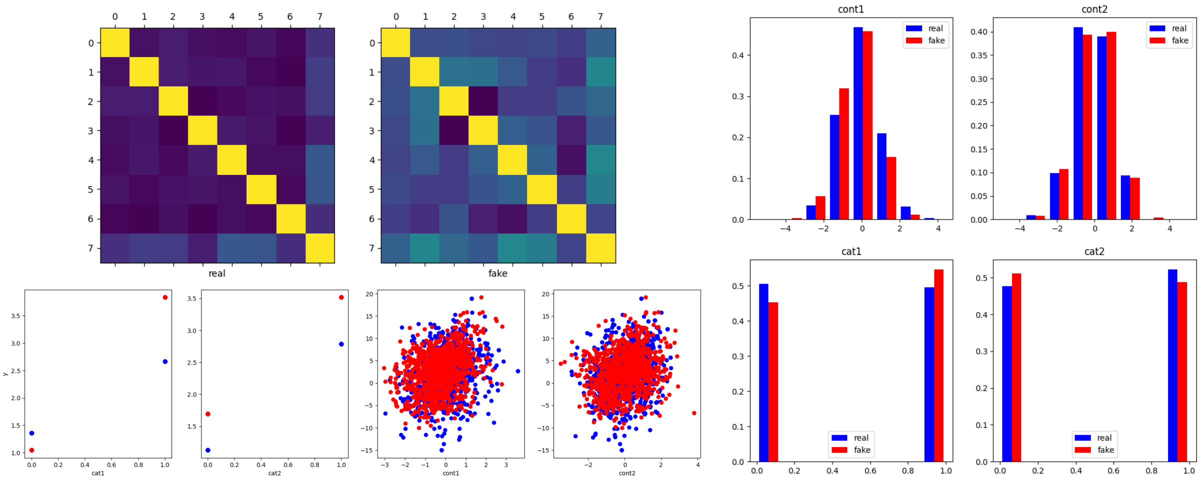


Figure A6 shows the generation quality of the WGAN being trained on low SNR (0.2) input data. The distributions of the covariates and the generated outcome variables are still well-learned. However, due to the noisy input, the generator hallucinates correlations among input covariates that are absent in the data generation process. This can be potentially harmful and may explain the observation that a smaller amount of synthetic data should be added.

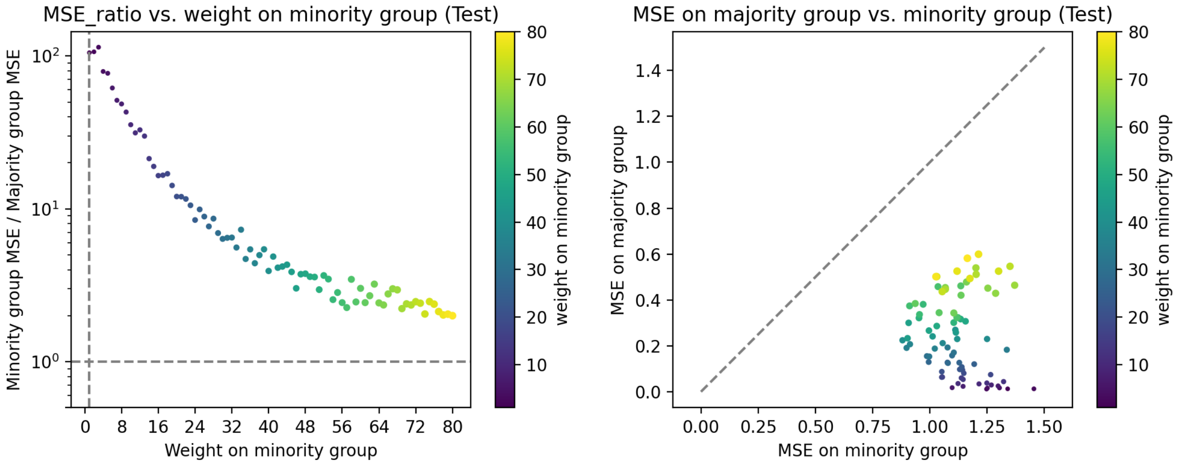
A6: The WGAN generation quality when the signal-to-noise ratio is low (0.2). The graph follows the same structure as Figure A3.



D.2 Extremely Imbalanced Covariates

Here, we present additional results on the weighted linear model when the first covariate is extremely imbalanced ($p = 0.4$). This is the case where only about 40 units are present in the minority group. Contrary to the smooth trade-off curve in Figure 1, Figure A7 demonstrates the potential insufficiency of the weighted linear model. When small weights are put on the minority units, the model does provide slightly better MSEs for the minority group. However, larger weights cause the performance on both groups to deteriorate significantly. This behavior may also correlate with the fact that score construction would be much harder on the small minority group.

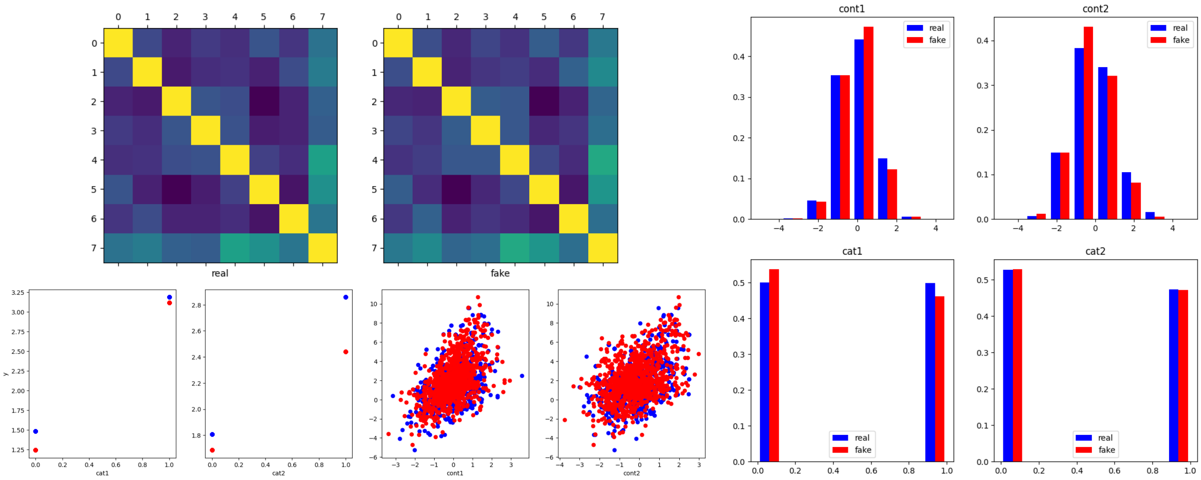
A7: Result for weighted linear regression when the covariate is extremely imbalanced ($p = 0.01$). The plots show the performance disparity between the two groups as the weight on minority units varies.



Combined with the results in Figure 15, reweighting methods tend to fail when the amount of information for the minority group is substantially scarce.

Figure A8 shows the generation quality of the WGAN being trained on the extremely imbalanced input data. Note that we oversampled the minority group before training the generators. The results show that the generators can learn the underlying distribution faithfully, even with highly limited data.

A8: The WGAN generation quality when the covariate is extremely imbalanced ($p = 0.01$). The graph follows the same structure as Figure A3.



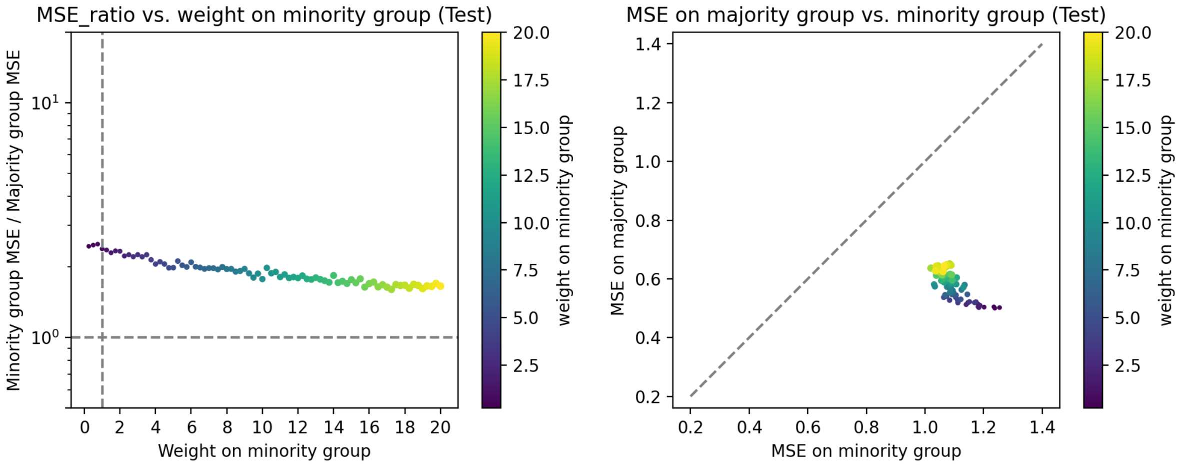
D.3 Independent CATE on the Two Groups

There are no additional results for this section. As mentioned, the performance for the weighted linear model closely resembles Figure 1. And since the CATE function is still relatively simple, the WGAN generation quality is also decent, similar to Figure 4 and hence omitted.

D.4 Complex Nonlinear CATE Function

Figure A9 shows the performance of the weighted linear regression with AIPW scores in the setting where the CATE is complicated and nonlinear. The MSEs produced with linear regression are much larger than causal forest approaches (Figure 17). This demonstrates the generality and power of non-parametric methods.

A9: Result for weighted linear regression when the CATE function is nonlinear and complicated. The plots show the performance disparity between the two groups as the weight on minority units varies.


MSE_ratio vs. weight on minority group (Test)
Minority group MSE / Majority group MSE
10^1
10^0
Weight on minority group
weight on minority group
MSE on majority group vs. minority group (Test)
MSE on majority group
MSE on minority group
weight on minority group