

---

# Regularizing Extrapolation in Causal Inference

---

**David Arbour\***  
Abode Research  
arbour@adobe.com

**Harsh Parikh\***  
Yale University  
harsh.parikh@yale.edu

**Bijan Niknam**  
Johns Hopkins University  
bniknam1@jh.edu

**Elizabeth Stuart**  
Johns Hopkins University  
estuart@jhu.edu

**Kara Rudolph**  
Columbia University  
kr2854@cumc.columbia.edu

**Avi Feller**  
University of California Berkeley  
afeller@berkeley.edu

## Abstract

Many common estimators in machine learning and causal inference are linear smoothers, where the prediction is a weighted average of the training outcomes. Some estimators, such as ordinary least squares and kernel ridge regression, allow for arbitrarily negative weights, which improve feature imbalance but often at the cost of increased dependence on parametric modeling assumptions and higher variance. By contrast, estimators like importance weighting and random forests (sometimes implicitly) restrict weights to be non-negative, reducing dependence on parametric modeling and variance at the cost of worse imbalance. In this paper, we propose a unified framework that directly penalizes the level of extrapolation, replacing the current practice of a hard non-negativity constraint with a soft constraint and corresponding hyperparameter. We derive a worst-case extrapolation error bound and introduce a novel “bias-bias-variance” tradeoff, encompassing biases due to feature imbalance, model misspecification, and estimator variance; this tradeoff is especially pronounced in high dimensions, when positivity is poor. We then develop an optimization procedure that regularizes this bound while minimizing imbalance and outline how to use this approach as a sensitivity analysis for dependence on parametric modeling assumptions. We demonstrate the effectiveness of our approach through synthetic experiments and a real-world application, involving the generalization of randomized controlled trial estimates to a target population of interest.

## 1 Introduction

A core challenge in observational causal inference and domain adaptation is to adjust data distributions so that features are comparable across distinct groups, such as control and treated groups or source and target populations [Imbens and Rubin, 2015, Farahani et al., 2021]. Weighting estimators and linear smoothers, in which the prediction is a weighted average of training outcomes, are widely used for such adjustment; examples include implicit weighting estimators like ordinary least squares (OLS) and random forests and explicit weighting approaches like inverse propensity score weighting [Li et al., 2013] and importance sampling [Thomas and Brunskill, 2017].

---

\*Co-first Authors with Equal Contribution (mentioned in alphabetical order)

An important divide among weighting estimators is whether weights are constrained to be non-negative, such as in traditional IPW, matching [Stuart, 2010], the synthetic control method [Abadie et al., 2010], and stable balancing weights [Zubizarreta, 2015, Ben-Michael et al., 2021a], as well as in the weighting component of popular doubly robust estimators like double machine learning [Chernozhukov et al., 2018]. This constraint limits extrapolation and dependence on parametric modeling assumptions, but typically at the cost of worse feature imbalance between re-weighted groups. This imbalance is especially pronounced in high-dimensional settings, when the curse of dimensionality means that positivity is less likely to hold, leading to further bias [D’Amour et al., 2021]. By contrast, linear smoothers like OLS and kernel ridge regression allow for arbitrarily negative weights [Robins et al., 2007], which can improve feature imbalance but at the cost of greater model dependence and higher estimator variance. Finally, augmented estimators that combine outcome modeling with explicit weighting strategies can therefore be viewed as performing controlled extrapolation, balancing model dependence against feature imbalance. Pure weighting and pure outcome modeling thus represent two extremes: no vs uncontrolled extrapolation.

In this paper, we leverage this geometric perspective to establish a general framework for systematically controlling extrapolation. In particular, we propose a unified approach that directly penalizes the level of extrapolation, replacing the current practice of a hard non-negativity constraint with a soft constraint and corresponding hyperparameter. Unlike prior research on extrapolation in machine learning that emphasizes predictions beyond the observed covariate support, we conceptualize extrapolation through unit weights, a particularly natural framework for handling high-dimensional covariates [Ben-Michael et al., 2021b]. Specifically, our contributions are:

- *Bias-bias-variance tradeoff.* We propose a framework quantifying a “bias-bias-variance” tradeoff, decomposing error into bias from distributional imbalance, bias from outcome model misspecification, and estimator variance. This captures key tradeoffs encountered in common causal inference and distribution shift scenarios.
- *Error bound and constrained optimization.* We derive an error bound based on worst-case Hölder continuity deviations from linearity. We present an optimization approach to minimize this bound, explicitly controlling tradeoffs between biases. We also characterize the asymptotic variance properties of our estimator.
- *Sensitivity analysis framework.* We introduce a sensitivity analysis methodology integrated into our optimization framework, enabling systematic evaluation of distributional imbalance and outcome model misspecification impacts. We illustrate this using synthetic data and a practical application involving the transportation of causal estimates to a novel target population.

## 1.1 Related work

Extrapolation and generalization are core topics in causal inference and machine learning. Recent surveys by Degtiar and Rose [2023] and Johansson et al. [2022] provide comprehensive overviews on generalizability and transportability methods.

**Extrapolation and the synthetic control method.** Extrapolating far from the support of the data is a longstanding concern in statistics and the social sciences especially; see King and Zeng [2006] for a seminal discussion of possible dangers of unchecked extrapolation. Methods that limit extrapolation are common; the synthetic control method [Abadie et al., 2010] is a particularly prominent example. Doudchenko and Imbens [2016] discuss the non-negativity constraint in this context, and explore possible regularization. Most relevant to our approach, Ben-Michael et al. [2021b] developed the augmented synthetic control method, which combines outcome modeling with constrained weights to reduce bias while controlling extrapolation.

**Extrapolation in machine learning.** Within machine learning, there has been substantial recent progress on approaches for addressing extrapolation. Shen and Meinshausen [2024] introduced engression, a framework that views extrapolation through the lens of distributional regression, enabling principled uncertainty quantification outside the training distribution. Kong et al. [2024] developed a causal lens for understanding extrapolation, establishing theoretical connections between causal structure and extrapolation. Netanyahu et al. [2023] proposed a transductive approach for learning to extrapolate, leveraging unlabeled test points to guide the extrapolation process. Dong and Ma [2022] provided foundational analysis toward understanding the extrapolation of nonlinear models

to unseen domains, establishing bounds on extrapolation error. Finally, Pfister and Bühlmann [2024] developed extrapolation-aware nonparametric statistical inference methods, with formal guarantees on validity beyond the support of training data.

Unlike this recent literature, we approach extrapolation from a weighting perspective, which offers particular advantages in high-dimensional settings. Rather than focusing on predictions outside the covariate support, we frame extrapolation in terms of the properties of unit weights, providing a natural parameterization for high-dimensional settings [Ben-Michael et al., 2021b]. This perspective allows us to directly quantify and regularize the degree of extrapolation without relying on complex directional derivatives or high-dimensional density estimation.

**Positivity violations and shifting the target.** Our discussion is closely related to the literature on positivity violations in causal inference. Crump et al. [2006], Li et al. [2018], and Parikh et al. [2025] all proposed to avoid issues due to positivity violations by shifting the estimand to regions with greater overlap. By contrast, our approach directly incorporates the severity of positivity violations into the weight estimation process.

**Weighting representations.** A growing literature highlights the connections between various causal estimators through their weighting representations [Chattopadhyay and Zubizarreta, 2023]. Knaus [2024] provided a unified framework for viewing treatment effect estimators as weighted outcomes. Bruns-Smith et al. [2023] showed that augmented balancing weights can be interpreted as a form of linear regression. Lin and Han [2022] examined regression-adjusted imputation estimators through their weighting properties. Our framework builds on these insights by explicitly parameterizing the degree of extrapolation through weight regularization, providing a continuum of estimators that navigate the bias-variance tradeoff.

## 2 Preliminaries

### 2.1 Setup and notation

To ease exposition, we set up our problem for the causal inference problem of estimating the missing control potential outcome for the Average Treatment Effect on the Treated (ATT). As we note below, however, these results hold for general linear estimands as well as for domain adaptation in ML [Johansson et al., 2022].

For each unit  $i \in [n]$ , we observe the tuple  $(X_i, Y_i, Z_i)$ , with covariates  $X_i \in \mathcal{X}$ , outcome  $Y_i \in \mathbb{R}$ , and binary treatment  $Z_i \in \{0, 1\}$ . Invoking SUTVA, let  $Y_i(0)$  and  $Y_i(1)$  denote the control and treated potential outcomes, respectively, for unit  $i$ . Our estimand of interest is the ATT,  $\mathbb{E}[Y_i(1) - Y_i(0) \mid Z_i = 1]$ . Since we observe  $Y(1)$  for the treated group, the key challenge is to estimate the missing control potential outcome mean,  $\mathbb{E}[Y_i(0) \mid Z_i = 1]$ . Finally, define the density ratio  $dQ/dP(x)$ , where  $Q$  and  $P$  denote the populations of units assigned to treatment and control, respectively. We need the following key assumptions for nonparametric identification:

- A.1. (Exchangeability)  $\mathbb{E}[Y(0) \mid X, Z = 1] = \mathbb{E}[Y(0) \mid X, Z = 0]$
- A.2. (Population overlap)  $dQ/dP(x) < \infty$  for all  $x \in \mathcal{X}$

In our setup, we consider situations when the population overlap assumption A.2. might be violated. In that case, researchers can instead rely on parametric assumptions on  $\mu(x) = \mathbb{E}[Y(0) \mid X_i = x]$ , such as linearity, to identify and estimate the expected outcomes. Finally, following Chattopadhyay and Zubizarreta [2023], we will focus on estimating the mean at a target *covariate profile*,  $\mathbf{x}^* \in \mathcal{X}$ , corresponding to our estimand of interest. For the ATT, this profile is simply the mean of the treated population,  $\mathbf{x}^* = \mathbb{E}[X \mid Z = 1]$ , where  $\mu(\mathbf{x}^*) = \mathbb{E}[Y(0) \mid Z = 1]$ .

We set up our problem as an instance of estimating a linear functional. We begin with the fully general setup before specializing to causal inference. Following Bruns-Smith et al. [2023], for each unit  $i \in [n]$ , we observe the tuple  $(X_i, Y_i, Z_i)$ , with covariates  $X_i \in \mathcal{X}$ , outcome  $Y_i \in \mathbb{R}$ , and treatment  $Z_i \in \mathcal{Z}$ . The target functional is  $\mathbb{E}[h(X_i, Z_i, m)]$  for a function  $h$ , where  $m(x, z) = \mathbb{E}[Y_i \mid X_i = x, Z_i = z]$ . Many common problems in causal inference and domain adaptation [e.g., Johansson et al., 2022] are special cases of this setup, including counterfactual quantities like the average derivative and the expected policy-specific outcome.

To simplify exposition, we focus on the special case of estimating the missing control potential outcome in the Average Treatment Effect on the Treated (ATT) with binary treatment  $Z$ . Assuming SUTVA, let  $Y_i(0)$  and  $Y_i(1)$  denote the control and treated potential outcomes, respectively, for unit  $i$ ; our estimand is then  $\psi = \mathbb{E}[Y_i(0) \mid Z_i = 1]$ . Finally, we define the density ratio  $dQ/dP(x)$ , where  $Q$  and  $P$  denote the populations of units assigned to treatment and control, respectively. Our goal is to estimate  $\psi$  using the observed outcomes of the control units. We need the following key assumptions for nonparametric identification:

A.1. (Exchangeability)  $\mathbb{E}[Y(0) \mid X, Z = 1] = \mathbb{E}[Y(0) \mid X, Z = 0]$

A.2. (Population overlap)  $dQ/dP(x) < \infty$  for all  $x \in \mathcal{X}$

In our setup, we consider situations when the population overlap assumption A.2. might be violated. In that case, researchers must rely on parametric assumptions, such as linearity of the outcome-covariate relationship,  $\mu(x) = \mathbb{E}[Y(0) \mid X_i = x]$ , to extrapolate—and identify and estimate—the expected outcomes. Finally, following Chattopadhyay and Zubizarreta [2023], we will focus on estimating the mean at a target *covariate profile*,  $x^* \in \mathcal{X}$ , corresponding to our estimand of interest. For the ATT, this profile is simply the mean of the treated population,  $x^* = \mathbb{E}[X \mid Z = 1]$ , where  $\mu(x^*) = \mathbb{E}[Y(0) \mid Z = 1]$ .

**Linear in features.** Since we are focused on linear smoothers, we therefore focus on models that are linear in *some* features — but which could be possibly complex functions of the underlying covariates. This is an extremely large model class that ranges from *simple linear models* to the last layer embedding from a *pre-trained large language model*. For our setup, we let  $x$  be the features in the representation implied by the parametric model, rather than simply the raw covariates. We further assume:

**Assumption 2.1.**  $\mu$  is Hölder continuous such that  $|\mu(x) - \mu(x')| \leq a \cdot \|x - x'\|^\alpha$ , with  $a > 0$  and  $\alpha > 0$ .

Parameterizing  $\mu$  in terms of its Hölder constants is useful for characterizing departures from linearity that directly affect the estimation error bound.

**General linear estimands.** We can leverage recent work on the Riesz representer [Chernozhukov et al., 2022a] to immediately generalize our results to any linear functional of the data. Following Bruns-Smith et al. [2023], for each unit  $i \in [n]$ , we observe the tuple  $(X_i, Y_i, Z_i)$ , with covariates  $X_i \in \mathcal{X}$ , outcome  $Y_i \in \mathbb{R}$ , and treatment  $Z_i \in \mathcal{Z}$ . The target functional is then  $\mathbb{E}[h(X_i, Z_i, m)]$  for a function  $h \in L_2$ , where  $m(x, z) = \mathbb{E}[Y_i \mid X_i = x, Z_i = z]$ . Many common problems in causal inference and domain adaptation are special cases of this setup, including counterfactual quantities like the average derivative and the expected policy-specific outcome. Finally, define a feature map  $\phi : \mathcal{X} \times \mathcal{Z} \rightarrow \mathbb{R}^d$ ; the target *feature profile* is then  $\phi^*(x, z) = \mathbb{E}[h(X, Z, \phi)]$ . Our results below apply by replacing the simple covariate profile  $x^*$  with the much more general feature profile  $\phi^*(x, z)$ .

## 2.2 Weighting form of causal inference estimators

Our focus is on weighting estimators or *linear smoothers* [Buja et al., 1989] of the form:

$$\hat{\mu}(x^*) = \sum_{i=1}^n w_{\star \leftarrow i}(x_i) Y_i,$$

with weights  $w_{\star \leftarrow i}(x_i)$ , where  $\star \leftarrow i$  emphasizes that the weights can depend both on the source covariates  $x_i$  and the target covariates  $x^*$  [Lin and Han, 2022]. When there is no ambiguity, we suppress the dependence on the covariates  $x_i$  and the target  $x^*$ .

A broad class of estimators have this form. See Knaus [2024] for a comprehensive discussion of the weighting form for common causal inference estimators. We highlight several special cases here, with a focus on whether the implied weights are constrained to be non-negative.

**Explicit weighting estimators.** The first class of methods estimate the density ratio  $\widehat{dQ/dP}(x)$ , either directly or indirectly.

- *Traditional Inverse Propensity Score Weighting.* In standard IPW [Rosenbaum, 1987], researchers first estimate a propensity score,  $e(\mathbf{x}) = \mathbb{P}[Z_i = 1 \mid \mathbf{X}_i = \mathbf{x}]$  via a binary classifier like logistic regression, and then plug into a known functional form for  $dQ/dP(\mathbf{x})$ . For the ATT,  $\hat{w}(\mathbf{x}) = \hat{e}(\mathbf{x})/(1 - \hat{e}(\mathbf{x}))$ ; since  $\hat{e}(\mathbf{x}) \in (0, 1)$ ,  $\hat{w}_i(\mathbf{x}) > 0$  for all  $i$ .
- *Balancing weights, synthetic control, and matching.* An alternative weighting approach instead directly estimates  $dQ/dP(\mathbf{x})$  via constrained optimization [Ben-Michael et al., 2021a]. For example, consider the minimum variance weights that control imbalance in  $\mathbf{x}$  between  $P$  and  $Q$ :

$$\hat{\mathbf{w}} \in \arg \min_{\mathbf{w} \in \mathcal{W}} \left\| \sum_{i=1}^n w_i \mathbf{X}_i - \mathbf{x}^* \right\|_p^2 + \lambda \|\mathbf{w}\|_2^2, \quad (1)$$

where  $\|\cdot\|_p$  is the  $p$  vector norm and where  $\mathcal{W}$  are possible constraints on the weights. *Stable balancing weights* [Zubizarreta, 2015] and the *Synthetic Control Method* [Abadie et al., 2010] are special cases where  $\mathcal{W}$  is the simplex ( $w_i \geq 0$ ,  $\sum w_i = 1$ ) and the imbalance norm is  $p = \infty$  and  $p = 2$ , respectively. *Matching* is a special case where the weights are also constrained to be discrete.

- *Riesz regression.* A final weighting approach, also known as automatic estimation of the Riesz representer [Chernozhukov et al., 2022b] also finds weights via Problem (1), albeit *without* imposing the constraint that weights are non-negative. For example, minimum distance lasso Riesz regression in Chernozhukov et al. [2022b] solves Equation (1) with  $\mathcal{W} = \mathbb{R}^n$  and  $p = \infty$ .

**Linear smoothers and implicit weighting estimators.** A wide range of popular outcome models are linear smoothers [Buja et al., 1989], which implicitly estimate weights  $w$ , including (kernel ridge) regression,  $k$ -nearest neighbors, random forests, xgboost, and many implementations of neural networks; see Lin and Han [2022], Curth et al. [2024]. We highlight two prominent examples with and without a non-negativity constraint.

- *(Kernel) ridge regression.* For features  $\mathbf{X}$ , the implied ridge regression weights are:

$$w_{\star \leftarrow i} = \mathbf{x}^{\star \top} (\mathbf{X}^\top \mathbf{X} + \lambda \mathbb{I})^{-1} \mathbf{X}_i,$$

where  $\lambda$  is a regularization parameter; ordinary least squares (OLS) as a special case when  $\lambda = 0$ . Kernel ridge regression is instead based on the implied kernel features  $\phi(\mathbf{x})$ ; see Bruns-Smith et al. [2023], Hirshberg et al. [2019]. As Bruns-Smith et al. [2023] discuss, the ridge regression weights are equivalent to solving optimization problem (1) with the imbalance norm set to  $p = 2$  and with  $\mathcal{W} = \mathbb{R}^n$ , which does *not* include a non-negativity constraint.

- *Random forests.* As Athey et al. [2019] discuss in the context of causal inference, (honest) random forests is a locally adaptive linear smoother with *non-negative* weights:

$$\hat{w}_{\star \leftarrow i} = \frac{1}{B} \sum_{b=1}^B \frac{\mathbb{I}\{\mathbf{x}^* \in L_b(\mathbf{x})\}}{|L_b(\mathbf{x})|},$$

where  $L_b$  is the set of units that share a leaf node with the target  $\mathbf{x}^*$  and  $b = 1, \dots, B$  index the trees.

**Augmented and hybrid estimators.** Finally, augmented or hybrid estimators combine initial weights  $w^0$  and outcome model  $\hat{m}$ :

$$\begin{aligned} \hat{\mu}^{dr}(\mathbf{x}^*) &= \sum_{i=1}^N \hat{w}_i^0 Y_i + \left( \hat{m}(\mathbf{x}^*) - \sum_{i=1}^N w_i^0 \hat{m}(\mathbf{x}_i) \right) \\ &= \hat{m}(\mathbf{x}^*) + \sum_{i=1}^N \hat{w}_i^0 (Y_i - \hat{m}(\mathbf{x}_i)). \end{aligned}$$

When  $\hat{m}$  is a linear smoother, then  $\hat{\mu}^{dr}(\mathbf{x})$  also has a weighting representation. Let  $\hat{m}(\mathbf{x}^*) = \sum \hat{\omega}_i(\mathbf{x}) Y_i$  for a weighting function  $\hat{\omega} : \mathbb{R}^d \rightarrow \mathbb{R}^n$ . Following Ben-Michael et al. [2021b]:

$$\hat{\mu}^{dr}(\mathbf{x}^*) = \sum_{i=1}^N \left( \hat{w}_i^0 + \hat{w}_i^{\text{adj}} \right) Y_i \quad \text{where} \quad \hat{w}_i^{\text{adj}} \equiv \hat{\omega}_i(\mathbf{x}_*) - \sum_{j=1}^n \hat{w}_j^0 \hat{\omega}_i(\mathbf{x}_j)$$

For example, when the outcome model is ridge regression, the implied weights for the doubly robust estimator has the following form:

$$\hat{w}_i^{dr} = \hat{w}_i^0 + (\mathbf{x}^* - \mathbf{x}'\hat{\mathbf{w}}^0)'(\mathbf{x}'\mathbf{x} + \lambda\mathbb{I})^{-1}\mathbf{x}_i.$$

Importantly, even if the initial weights  $\mathbf{w}^0$  are constrained to be non-negative, such as in traditional IPW, the implied doubly robust weights  $\mathbf{w}^{dr}$  could be negative. In fact, the combined weights can be negative even if both the initial weights  $\mathbf{w}^0$  and the outcome model-implied weights  $\alpha$  are non-negative.

There are many examples of combined estimators of this form: standard Augmented IPW [Chatopadhyay and Zubizarreta, 2023], bias correction for inexact matching [Lin et al., 2021], augmented synthetic control method [Ben-Michael et al., 2021b], and regression-adjusted imputation estimators more broadly [Lin and Han, 2022]. Finally, both debiased machine learning [Chernozhukov et al., 2018] and *automatic* debiased machine learning [Chernozhukov et al., 2022a]; the former constrains the initial weights to be non-negative, the latter does not.

### 3 Regularizing Worst-Case Extrapolation Bias

Our goal is to bound the estimation error:  $|\mu(\mathbf{x}^*) - \sum_{i=1}^n w_i Y_i|$ . We begin by building intuition for our approach. First, note that, under linearity, a negative weight on training point  $\mathbf{x}_i$  is equivalent to reflecting the training point around the origin:  $-\mu(\mathbf{x}_i) = \mu(-\mathbf{x}_i)$ . We can use this to construct a “reflected” estimator, denoted by  $\hat{\mu}^\ddagger$ , which reflects points with negative weights around the origin:

$$\begin{aligned}\hat{\mu}^\ddagger(\mathbf{x}^*) &= \sum_{i=1}^n w_i \mathbb{1}(w_i \geq 0) \mu(\mathbf{X}_i) + |w_i| \mathbb{1}(w_i < 0) \mu(-\mathbf{X}_i) \\ &= \sum_{i=1}^n |w_i| \mu(\mathbf{X}_i^\ddagger), \quad \mathbf{X}_i^\ddagger = \begin{cases} \mathbf{X}_i, & w_i \geq 0 \\ -\mathbf{X}_i, & w_i < 0 \end{cases},\end{aligned}$$

where  $\hat{\mu}(\mathbf{x}^*) = \hat{\mu}^\ddagger(\mathbf{x}^*)$  if  $\mu$  is an odd-function, and where  $w_i \mathbf{X}_i = |w_i| \mathbf{X}_i^\ddagger$  for all  $i$ .

Second, the difference between  $\hat{\mu}(\mathbf{x}^*)$  and  $\hat{\mu}^\ddagger(\mathbf{x}^*)$  is a measure of nonlinearity. We can decompose  $\mu(-\mathbf{x}_i)$  as  $\delta(\mathbf{x}_i) - \mu(\mathbf{x}_i)$ , where  $\delta(\mathbf{x}_i) = (\mu(-\mathbf{x}_i) + \mu(\mathbf{x}_i))$ : then  $\hat{\mu}(\mathbf{x}^*) = \hat{\mu}^\ddagger(\mathbf{x}^*) + \sum_i |w_i| \mathbb{1}(w_i < 0) \delta(\mathbf{x}_i)$ . If  $\mu$  is an “odd function” (e.g.,  $\mu$  is linear), then  $\delta(\mathbf{X}) = 0$  because  $\mu(-\mathbf{X}) = -\mu(\mathbf{X})$ . Thus,  $\delta(\mathbf{X})$  is a point-specific measure of nonlinearity in the underlying data generating process.

We use this representation to decompose the estimator  $\hat{\mu}(\mathbf{x}^*)$ :

$$\begin{aligned}\hat{\mu}(\mathbf{x}^*) &= \sum_{i=1}^n w_i Y_i = \sum_{i=1}^n w_i (\mu(\mathbf{X}_i) + \epsilon_i) \\ &= \sum_{i=1}^n w_i \mathbb{1}(w_i \geq 0) \mu(\mathbf{X}_i) + |w_i| \mathbb{1}(w_i < 0) (\mu(-\mathbf{X}_i) - \delta(\mathbf{X}_i)) + w_i \epsilon_i \\ &= \underbrace{\sum_{i=1}^n |w_i| \mu(\mathbf{X}_i^\ddagger)}_{\hat{\mu}^\ddagger(\mathbf{x}^*)} + \underbrace{\sum_{i=1}^n |w_i| \mathbb{1}(w_i < 0) \delta(\mathbf{X}_i)}_{\text{nonlinearity}} + \underbrace{\sum_{i=1}^n w_i \epsilon_i}_{\text{noise}}.\end{aligned}$$

Although  $\delta(\mathbf{X})$  is unknown, we can bound its magnitude using the Hölder continuity assumption:  $\|\delta(\mathbf{X})\| \leq 2a\|\mathbf{X}\|^\alpha + 2\|\mu(\mathbf{X})\|$ .

Further, if we assume  $\mu(0) = 0$ .<sup>2</sup> Then  $\|\delta(\mathbf{X})\| \leq 2a\|\mathbf{X}\|^\alpha$ .

---

<sup>2</sup>this can also be achieved via centering

The resulting error bound is therefore  $|\mu_1(\mathbf{x}^*) - \hat{\mu}(\mathbf{x}^*)|$  less than or equal to

$$\begin{aligned} & \underbrace{\left| \sum_{i=1}^n |w_i| \mu(\mathbf{X}_i^\dagger) - \mu(\mathbf{x}^*) \right|}_{\text{error in } \hat{\mu}^\dagger(\mathbf{x}^*)} + \underbrace{\left| \sum_{i=1}^n w_i \epsilon_i \right|}_{\text{noise}} \\ & + 2a \underbrace{\sum_{i=1}^n |w_i| \mathbb{1}(w_i < 0) \|\mathbf{X}_i\|^\alpha}_{\text{error due to nonlinearity}}. \end{aligned} \quad (2)$$

The first term directly depends on the imbalance between the target point  $\mathbf{x}^*$  and the re-weighted (reflected) training points  $|\mathbf{w}|' \mathbf{X}^\dagger$ . The second term captures additional error due to nonlinearity, which corresponds to the  $\delta(\mathbf{X})$  term above. The final term is the noise term.

### 3.1 Characterizing asymmetry-induced bias

Thus far we have presented a conservative nonparametric bound. We now provide a slightly refined characterization by noting that the extent of the bias induced by negative weights is driven by the asymmetry in  $\mu$ . We do so by considering the decomposition of the  $\mu$  into its even and odd components, i.e.,  $\mu(x) = \mu_e(x) + \mu_o(x)$ . By the definition of odd functions, we have  $-\mu_o(x) = \mu_o(-x)$ , so we can bound the risk by bounding the worst-case risk of  $\hat{w}$  using the assumed Hölder constants  $a$  and  $\alpha$  and isolating the effect of the even component. The formal statement is given below in Proposition 3.1 the proof is given in Appendix C.

**Proposition 3.1.** *Let  $\hat{\mu}(x^*) = \sum_{i=1}^n \hat{w}_i Y_i$  be the estimate of  $\mu(x^*)$  with weights estimated via Equation 4. Given  $Y_i = \mu(X_i) + \epsilon_i$  where  $\epsilon_i$  are independent random variables with  $\mathbb{E}[\epsilon_i] = 0$  and finite second moment  $\sigma^2 = \mathbb{E}[\epsilon_i^2]$ , and  $\mu$  is Hölder continuous with constants  $a$  and  $\alpha$ . If  $\epsilon_i$  are sub-Gaussian<sup>3</sup> with parameter  $\sigma$ , then with probability at least  $1 - \delta$ ,*

$$|\mu(x^*) - \hat{\mu}(x^*)| \leq B_{\text{even}}(x^*) + \sigma \|\hat{w}\|_2 \sqrt{2 \log(2/\delta)} \quad (3)$$

where

$$B_{\text{even}}(x^*) = \left| \sum_{i=1}^n \hat{w}_i [\mu_e(X_i) - \mu_e(x^*)] + 2 \sum_{i=1}^n |\hat{w}_i| \mathbf{1}(\hat{w}_i < 0) a \|X_i - x^*\|^\alpha \right|$$

and  $\mu_e(x) = \frac{\mu(x) + \mu(-x)}{2}$  denotes the even part of  $\mu$ .

The worst-case bound provided earlier is recovered if  $\mu$  is an even function, i.e., contains no odd component.

One of the fundamental limitations of Proposition 3.1 is that it requires access to the separate even and odd functions constituting  $\mu$  which are latent in practice. In our proposed estimator and empirical application we address this by preferring the conservative form which assumes the worst case form. For completeness we provide the following proposition that provides an empirical analog of Proposition 3.1. Intuitively, when the observations of  $X$  are symmetric, i.e., for every  $X_i - X_i$  is also in the dataset, then we can recover both the even and odd functions. In practice, because this symmetry is unlikely to hold we approximate it with a one-nearest neighbor and incorporate the induced uncertainty into the bound using the Hölder constants. The formal statement is below, the proof is deferred to Appendix C.

**Proposition 3.2** (Approximate Bounds). *Let  $\hat{\mu}(x^*) = \sum_{i=1}^n \hat{w}_i Y_i$  be the estimate of  $\mu(x^*)$  with weights estimated via Equation (5). Given  $Y_i = \mu(X_i) + \epsilon_i$  where  $\epsilon_i$  are independent sub-Gaussian random variables with parameter  $\sigma$ , and  $\mu$  is Hölder continuous with constants  $a$  and  $\alpha$ . Let  $I_{\text{paired}} = \{i : -X_i \in \{X_1, \dots, X_n\}\}$ ,  $I_{\text{nn}} = \{1, \dots, n\} \setminus I_{\text{paired}}$ , and for each  $i \in I_{\text{nn}}$ , define  $j^*(i) = \arg \min_{j \neq i} \|X_j - (-X_i)\|$ . Then with probability at least  $1 - \delta$ ,  $|\mu(x^*) - \hat{\mu}(x^*)|$  is less*

<sup>3</sup>We assume mean zero sub-Gaussian noise, analogous results can be obtained with this assumption replaced by bounded noise.

than or equal to

$$\begin{aligned}
& \left| \sum_{i \in I_{\text{paired}}} \hat{w}_i \frac{Y_i + Y_{-i}}{2} + \sum_{i \in I_{nn}} \hat{w}_i \frac{Y_i + Y_{j^*(i)}}{2} - \mu_e(x^*) \sum_{i=1}^n \hat{w}_i \right| \\
& + \sum_{i \in I_{nn}} |\hat{w}_i| a \|X_{j^*(i)} - (-X_i)\|^\alpha \\
& + 2 \sum_{i=1}^n |\hat{w}_i| \mathbf{1}(\hat{w}_i < 0) a \|X_i - x^*\|^\alpha + \sigma \|\hat{w}\|_2 \sqrt{2 \log(6/\delta)} \\
& + \frac{\sigma}{\sqrt{2}} (\|\hat{w}_{I_{\text{paired}}}\|_2 + \|\hat{w}_{I_{nn}}\|_2) \sqrt{2 \log(6/\delta)}
\end{aligned}$$

where  $\mu_e(x^*)$  is bounded using the closest observation  $j^* = \arg \min_{j=1, \dots, n} \|X_j - x^*\|$ ,  $|\mu_e(x^*)|$  less than or equal to

$$\begin{aligned}
& \left| \frac{Y_{j^*} + Y_{j^*(j^*)}}{2} \right| + \sigma \sqrt{2 \log(6/\delta)} \\
& + a \|X_{j^*(j^*)} - (-X_{j^*})\|^\alpha + a \|X_{j^*} - x^*\|^\alpha
\end{aligned}$$

Proposition 3.2 can be used as a companion to Chattopadhyay and Zubizarreta [2023], who propose computing the effective sample size of units with negative weight,  $\frac{\sum_i \mathbf{1}[w_i < 0] |w_i|}{\sum_i |w_i|}$ , to indicate the extent to which the estimate relies on extrapolation and parametric assumptions – we refer to this as *negative influence*.

### 3.2 Learning Estimator

We now propose an estimator to learn weights  $\mathbf{w}$  that directly control the error bound in Equation (2). To do so, we modify the standard balancing weights optimization problem in Equation (1) by using the Lagrangian form of the non-negative restriction, rather than the hard constraint. Thus, the combined estimator minimizes the error bound by controlling three terms: covariate imbalance, dispersion of the weights, and level of extrapolation:

$$\hat{\mathbf{w}} \in \arg \min_{\mathbf{w}} \left( \underbrace{\left\| \sum_i w_i \mathbf{X}_i - \mathbf{x}^* \right\|_p^2}_{(a)} + \lambda \underbrace{\|\mathbf{w}\|_2^2}_{(b)} + \gamma \underbrace{\|\mathbf{1}(w_i < 0) |w_i| (\|\mathbf{X}_i\|_2^\alpha)\|_p}_{(c)} \right),$$

where

- Term (a): Enforces balance between the target point  $\mathbf{x}^*$  and the re-weighted training points  $\{\mathbf{X}_1, \dots, \mathbf{X}_n\}$ , recalling that  $w_i \mathbf{X}_i = |w_i| \mathbf{X}_i^\dagger$  for all  $i$ . We will focus on the case where  $p = 2$ , but this setup immediately generalizes to  $p = \infty$ .
- Term (b): Regularizes the dispersion of the weights  $\mathbf{w}$  to control the variance of the overall estimator
- Term (c): Controls extrapolation, particularly through penalization of negative weights.

The standard balancing weights problem in Equation (1) only focuses on a single bias-variance tradeoff: trading off covariate imbalance in the first term — which directly introduces bias — and the norm of the weights in the second term — which directly controls the estimator variance. By contrast, the new optimization problem (4) has a more elaborate bias-bias-variance trade-off. Allowing for negative weights introduces an additional trade-off between the first two terms and term (c): when the target unit lies outside the convex hull of the training points, controlling imbalance often requires some  $w_i$  values to be negative, which also increases the norm  $\|\mathbf{w}\|_2$ . Allowing negative weights also necessitates reliance on parametric assumptions for extrapolation.

Term (c) mitigates the risk of biased estimation by regulating the contribution of negative weights. For  $\gamma = 0$ , Equation (4) recovers a standard, unconstrained balancing weights problem as in Equation (1). At the other extreme  $\gamma \rightarrow \infty$  is equivalent to a hard non-negativity constraint. Increasing  $\gamma$



constrains extrapolation, reducing bias due to possible violations of parametric assumptions and limiting  $\|w\|_2$ , but worsening bias due to insufficient balance in term (a).

**Hyperparameter choice.** We argue that our proposed framework is best viewed as a form of sensitivity analysis, and encourage researchers to examine the full set of estimates spanned by the extrapolation hyperparameter  $\gamma$ . To supplement this, we propose a heuristic for choosing a default  $\gamma^*$ , starting with  $\gamma = 0$  (i.e., no penalty) and then gradually increasing  $\gamma$  until all weights are non-negative i.e.,  $w \geq 0$ . Alternatively, one can choose hyperparameter, in the spirit of Lepski’s method, by looking at when the change in the point estimate is sufficiently small, for a researcher-defined cutoff.

Finally, we provide the following proposition which specializes Proposition 3.2 for our proposed regularized estimator.

**Proposition 3.3** (Regularized Bound). *Consider the regularized estimator,*

$$\hat{w} \in \arg \min_w \left\| \sum_{i=1}^n w_i X_i - x^* \right\|_2^2 + \lambda \|w\|_2^2 + \gamma \sum_{i=1}^n \mathbf{1}(w_i < 0) |w_i| \|X_i\|^\alpha$$

and  $\hat{\mu}(x^*) = \sum_{i=1}^n \hat{w}_i Y_i$  where  $Y_i = \mu(X_i) + \epsilon_i$ . Under the assumptions of Proposition 3.2 and  $\|X_i\| \leq R$ , with probability at least  $1 - \delta$ :

$$|\mu(x^*) - \hat{\mu}(x^*)| \leq B_{\text{even}}(x^*) + \sigma \left( \frac{\|x^*\|_2}{\sqrt{\lambda}} + \frac{\gamma R^\alpha \sqrt{n}}{\lambda} \right) \sqrt{2 \log(2/\delta)}$$

where  $B_{\text{even}}(x^*)$  is defined as in Proposition 3.2.

The proof is provided in the appendix, proceeding by providing a more explicit form for the weight norm in proposition 3.2. This formulation explicitly shows the trade-offs introduced by the nonnegative regularizer.

## 4 Generalizing Medication for Opioid Use Disorder Trial Evidence

The Starting Treatment With Agonist Replacement Therapies (START) trial, initiated in 2006, was a multi-center study comparing buprenorphine versus methadone in treating opioid use disorder [Saxon et al., 2013, Hser et al., 2014]. The trial enrolled 1,271 participants, who were randomized in a 2:1 ratio to receive either buprenorphine or methadone. Methadone was found to have higher rates of patient retention in treatment compared to buprenorphine [Hser et al., 2014]. Our analysis focuses on the outcome of relapse to regular opioid use within 24 weeks of medication assignment, defined as non-study opioid use for four consecutive weeks or daily use for seven consecutive days.

Parikh et al. [2025] identified that Hispanic women with a pre-treatment history of amphetamine and benzodiazepine use were underrepresented in the START trial relative to the target population, highlighting a practical violation of the positivity assumption. In this study, we estimate the target average treatment effects (TATE) for this underrepresented subgroup using our proposed framework alongside standard linear regression, gradient boosting regression (GBR), inverse probability weighting (IPW) and double machine learning estimators.

The target sample is drawn from the 2015–2017 Treatment Episode Dataset - Admissions (TEDS-A), which includes data on individuals entering publicly funded substance use treatment programs across 48 states (excluding Oregon and Georgia) and the District of Columbia. Our analysis focuses on Hispanic women with a pre-treatment history of amphetamine and benzodiazepine use.

We code methadone as  $T = 1$  and buprenorphine as  $T = 0$ , with  $Y = 1$  representing relapse. Pretreatment covariates include age, race, biological sex, and substance use history (amphetamine, benzodiazepines, cannabis, and intravenous drug use) measured at the initiation of medication for opioid use disorder (MOUD) treatment.

We apply our proposed framework to address these issues, which regularizes extrapolation to mitigate reliance on extreme weights. By varying  $\gamma$  from 0.01 to 10, we examine how treatment effect estimates shift with increasing regularization of negative weights. Without regularization, the point estimates converge to those from linear regression. However, as regularization intensifies, the estimates smoothly shift towards zero and occasionally change the sign from negative to positive for smaller values of  $\lambda$ . This sensitivity underscores the influence of assumptions on the point

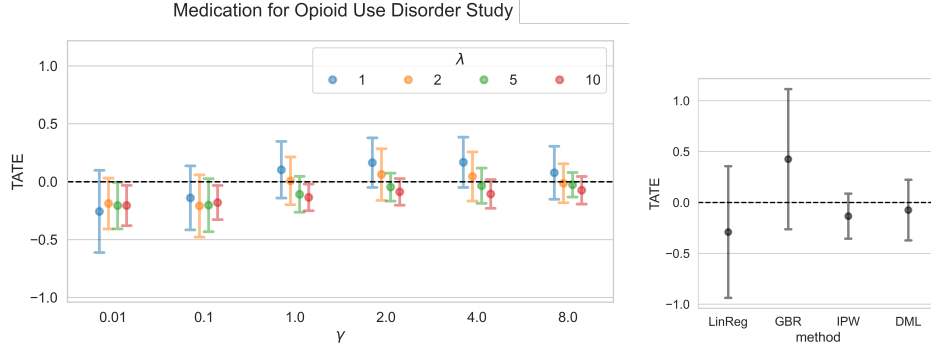


Figure 1: Target Average Treatment Effects for the Target Sample for Hispanic Females who have a history of Amphetamine and Benzodiazepine use in TEDS-A population. Each hue corresponds to a value of  $\lambda$  and the x-axis corresponds to different values of  $\gamma$  (on log scale).

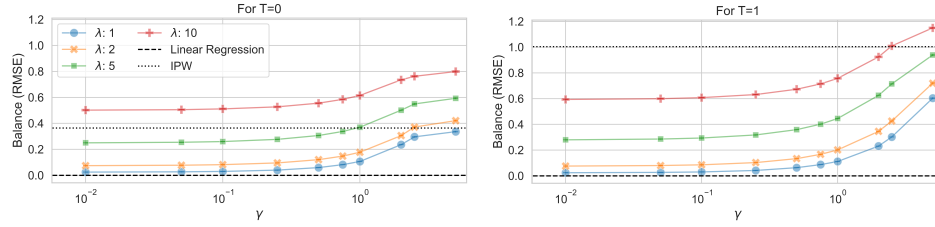


Figure 2: Balance between the trial and the target samples measured as the root mean squared error (RMSE) for different values of  $\gamma$  and  $\lambda$ .

estimates. While increasing  $\gamma$  reduces negative influence (Figure 3), it worsens covariate balance, as reflected in higher RMSE values (Figure 2). Thus, our framework highlights a trade-off between minimizing reliance on parametric assumptions and achieving optimal covariate balance. Our analysis highlights that these estimates are sensitivity to modeling assumptions. Thus, applied researchers should interpret treatment effect estimates among this under-represented subgroups with caution. As Parikh et al. [2025] emphasized, collecting more representative trial data is critical to credibly estimate treatment effects for this underrepresented subgroup.

## 5 Conclusion

This work proposes a framework for regularizing extrapolation in causal inference by replacing hard non-negativity constraints with soft penalties on negative weights. Our approach reveals a fundamental “bias-bias-variance” tradeoff between distributional imbalance, model misspecification, and estimator variance. We provide theoretical error bounds that decompose extrapolation bias through a novel reflection perspective, and demonstrate empirically on both synthetic data and a

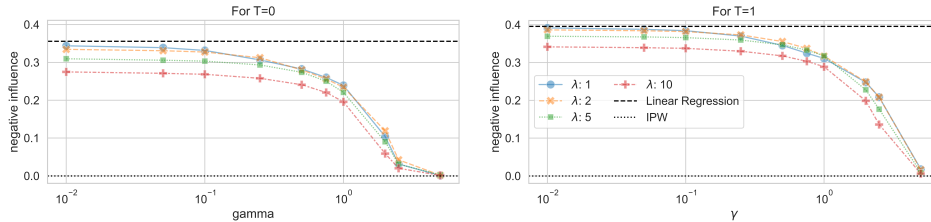


Figure 3: Negative influence, defined as the contribution of negative weights in estimation, for different values of  $\gamma$  and  $\lambda$ .

real-world medication trial that controlled extrapolation can outperform both fully constrained and unconstrained approaches, particularly in high-dimensional settings with poor positivity.

**Limitations and Future Work.** Our approach focuses primarily on weighting-type estimators and relies on Hölder continuity assumptions for expected outcomes and conditional ignorability, which may not hold in all practical settings. Future research should extend the bias-bias-variance tradeoff analysis to broader estimator classes, explore weaker continuity assumptions, and most importantly, operationalize sensitivity analysis for unmeasured confounding. While existing proposals for sensitivity analysis with balancing weights [Soriano et al., 2023] cannot be applied directly to our framework, adapting such methods represents a critical next step for practical implementation.

The framework presented here represents an important step toward more nuanced approaches to positivity violations in causal inference, moving beyond binary perspectives on extrapolation toward a continuous spectrum of regularization strategies that can be tailored to specific research contexts and risk tolerances.

## References

- A. Abadie, A. Diamond, and J. Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of california’s tobacco control program. *Journal of the American statistical Association*, 105(490):493–505, 2010.
- S. Athey, J. Tibshirani, and S. Wager. Generalized random forests. 2019.
- E. Ben-Michael, A. Feller, D. A. Hirshberg, and J. R. Zubizarreta. The balancing act in causal inference. *arXiv preprint arXiv:2110.14831*, 2021a.
- E. Ben-Michael, A. Feller, and J. Rothstein. The augmented synthetic control method. *Journal of the American Statistical Association*, 116(536):1789–1803, 2021b.
- D. Bruns-Smith, O. Dukes, A. Feller, and E. L. Ogburn. Augmented balancing weights as linear regression. *arXiv preprint arXiv:2304.14545*, 2023.
- A. Buja, T. Hastie, and R. Tibshirani. Linear smoothers and additive models. *The Annals of Statistics*, pages 453–510, 1989.
- A. Chattopadhyay and J. R. Zubizarreta. On the implied weights of linear regression for causal inference. *Biometrika*, 110(3):615–629, 2023.
- V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.
- V. Chernozhukov, W. K. Newey, and R. Singh. Automatic debiased machine learning of causal and structural effects. *Econometrica*, 90(3):967–1027, 2022a.
- V. Chernozhukov, W. K. Newey, and R. Singh. Debiased machine learning of global and local parameters using regularized riesz representers. *The Econometrics Journal*, 25(3):576–601, 2022b.
- R. K. Crump, V. J. Hotz, G. Imbens, and O. Mitnik. Moving the goalposts: Addressing limited overlap in the estimation of average treatment effects by changing the estimand, 2006.
- A. Curth, A. Jeffares, and M. van der Schaar. Why do random forests work? understanding tree ensembles as self-regularizing adaptive smoothers. *arXiv preprint arXiv:2402.01502*, 2024.
- I. Degtiar and S. Rose. A review of generalizability and transportability. *Annual Review of Statistics and Its Application*, 10(1):501–524, 2023.
- K. Dong and T. Ma. First steps toward understanding the extrapolation of nonlinear models to unseen domains. *arXiv preprint arXiv:2211.11719*, 2022.
- N. Doudchenko and G. W. Imbens. Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. Technical report, National Bureau of Economic Research, 2016.

- A. D’Amour, P. Ding, A. Feller, L. Lei, and J. Sekhon. Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 221(2):644–654, 2021.
- A. Farahani, S. Voghoei, K. Rasheed, and H. R. Arabnia. A brief review of domain adaptation. *Advances in data science and information engineering: proceedings from ICDATA 2020 and IKE 2020*, pages 877–894, 2021.
- D. A. Hirshberg, A. Maleki, and J. R. Zubizarreta. Minimax linear estimation of the retargeted mean. *arXiv preprint arXiv:1901.10296*, 2019.
- Y.-I. Hser, A. J. Saxon, D. Huang, A. Hasson, C. Thomas, M. Hillhouse, P. Jacobs, C. Teruya, P. McLaughlin, K. Wiest, et al. Treatment retention among patients randomized to buprenorphine/naloxone compared to methadone in a multi-site trial. *Addiction*, 109(1):79–87, 2014.
- G. W. Imbens and D. B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.
- F. D. Johansson, U. Shalit, N. Kallus, and D. Sontag. Generalization bounds and representation learning for estimation of potential outcomes and causal effects. *Journal of Machine Learning Research*, 23(166):1–50, 2022.
- S. Kapoor and P. M. Vaidya. Fast algorithms for convex quadratic programming and multicommodity flows. In *Proceedings of the eighteenth annual ACM symposium on Theory of computing*, pages 147–159, 1986.
- G. King and L. Zeng. The dangers of extreme counterfactuals. *Political analysis*, 14(2):131–159, 2006.
- M. C. Knaus. Treatment effect estimators as weighted outcomes. *arXiv preprint arXiv:2411.11559*, 2024.
- L. Kong, G. Chen, P. Stojanov, H. Li, E. Xing, and K. Zhang. Towards understanding extrapolation: a causal lens. *Advances in Neural Information Processing Systems*, 37:123534–123562, 2024.
- F. Li, A. M. Zaslavsky, and M. B. Landrum. Propensity score weighting with multilevel data. *Statistics in Medicine*, 32(19):3373–3387, 2013.
- F. Li, K. L. Morgan, and A. M. Zaslavsky. Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521):390–400, 2018.
- Z. Lin and F. Han. On regression-adjusted imputation estimators of the average treatment effect. *arXiv preprint arXiv:2212.05424*, 2022.
- Z. Lin, P. Ding, and F. Han. Estimation based on nearest neighbor matching: from density ratio to average treatment effect. *arXiv preprint arXiv:2112.13506*, 2021.
- A. Netanyahu, A. Gupta, M. Simchowitz, K. Zhang, and P. Agrawal. Learning to extrapolate: A transductive approach. *arXiv preprint arXiv:2304.14329*, 2023.
- H. Parikh, R. Ross, E. Stuart, and K. Rudolph. Who are we missing?: A principled approach to characterizing the underrepresented population. *Journal of the American Statistical Association*, 0(ja):1–32, 2025. doi:10.1080/01621459.2025.2495319.
- N. Pfister and P. Bühlmann. Extrapolation-aware nonparametric statistical inference. *arXiv preprint arXiv:2402.09758*, 2024.
- J. Robins, M. Sued, Q. Lei-Gomez, and A. Rotnitzky. Comment: Performance of double-robust estimators when” inverse probability” weights are highly variable. *Statistical Science*, 22(4): 544–559, 2007.
- P. R. Rosenbaum. Model-based direct adjustment. *Journal of the American statistical Association*, 82(398):387–394, 1987.

- A. J. Saxon, W. Ling, M. Hillhouse, C. Thomas, A. Hasson, A. Ang, G. Doraimani, G. Tasissa, Y. Lokhnygina, J. Leimberger, et al. Buprenorphine/naloxone and methadone effects on laboratory indices of liver health: a randomized trial. *Drug and alcohol dependence*, 128(1-2):71–76, 2013.
- X. Shen and N. Meinshausen. Engression: extrapolation through the lens of distributional regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkae108, 2024.
- D. Soriano, E. Ben-Michael, P. J. Bickel, A. Feller, and S. D. Pimentel. Interpretable sensitivity analysis for balancing weights. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 186(4):707–721, 2023.
- E. A. Stuart. Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25(1):1–21, 2010.
- P. S. Thomas and E. Brunskill. Importance sampling with unequal support. In *AAAI*, pages 2646–2652, 2017.
- J. R. Zubizarreta. Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511):910–922, 2015.

## A Implementational Details

We implement the methods and case studies in this paper using Python 3.10. We implemented our weight estimation framework using PyTorch (version 2.7.0) for efficient automatic differentiation and optimization. The optimization is performed using Adam optimizer with default learning rate of 0.01 for default of 5,000 epochs. By default the weights are normalized to sum to one after each optimization step – however, user can choose otherwise. The implementation includes comprehensive visualization tools, including Love plots for covariate balance assessment and 2D scatter plots with convex hull visualization for geometric interpretation. All random operations are seeded for reproducibility, and the code supports both single and multiple outcome variables. For the MOUD case study in Section 4, we scale the pre-treatment data to ensure that the maximum value for each covariate is 1 and the minimum is 0; it is important to note that most covariates in this instance are discrete binary. *Computational details:* Our objective mimics a classic quadratic program (QP). QPs are well studied in the literature, and their run time complexity is known to be  $O(n^{3.67} * \log n)$  where  $n$  is the number of training units [Kapoor and Vaidya, 1986]. Our new simulation study also demonstrates that our method is scalable to high-dimensional data.

## B Connecting with Existing Estimators

Our proposed framework establishes meaningful connections with existing weighting and doubly robust estimators. When  $\gamma = 0$  and  $p = 2$ , our estimator reduces to single ridge regression with hyperparameter  $\lambda$ . In this context, the parameter  $\gamma$  can be interpreted as controlling the degree of extrapolation in the doubly robust (DR) estimator; see Ben-Michael et al. [2021b]. This relationship can be further understood through the lens of ridge-augmented balancing weights, which constitute a doubly robust estimator under certain regularity conditions [Bruns-Smith et al., 2023]. Building on the work of Knaus [2024], which demonstrated that many common causal inference estimators—including doubly robust methods like AIPW—can be expressed as weighting estimators, we can extend our framework to accommodate and regularize many existing approaches. Specifically, we can adapt our framework to work with implied weights from baseline estimators by modifying the original objective function as follows:

$$\mathbf{w}^* \in \arg \min_{\mathbf{w}} \|\mathbf{w} - \mathbf{w}'\| + \gamma \|\mathbb{1}(w_i < 0)|w_i|(\|\mathbf{X}_i\|_2^\alpha)\|$$

where  $\mathbf{w}'$  represents the implied weights from a baseline estimator. This formulation allows us to regularize extrapolation in existing estimators such as AIPW and IPW, providing a principled approach to control their behavior in regions where overlap is limited.

## C Proofs

### C.1 Proof of Proposition 3.1

*Proof.* First taking the bound of the estimation error

$$\begin{aligned} |\mu(x^*) - \hat{\mu}(x^*)| &= \left| \mu(x^*) - \sum_{i=1}^n \hat{w}_i Y_i \right| \\ &= \left| \mu(x^*) - \sum_{i=1}^n \hat{w}_i (\mu(X_i) + \epsilon_i) \right| \\ &\leq \left| \mu(x^*) - \sum_{i=1}^n \hat{w}_i \mu(X_i) \right| + \left| \sum_{i=1}^n \hat{w}_i \epsilon_i \right| \end{aligned}$$

Substituting in the even-odd decomposition gives

$$\begin{aligned} \left| \mu(x^*) - \sum_{i=1}^n \hat{w}_i \mu(X_i) \right| &= \left| [\mu_e(x^*) + \mu_o(x^*)] - \sum_{i=1}^n \hat{w}_i [\mu_e(X_i) + \mu_o(X_i)] \right| \\ &\leq \left| \mu_e(x^*) - \sum_{i=1}^n \hat{w}_i \mu_e(X_i) \right| + \left| \mu_o(x^*) - \sum_{i=1}^n \hat{w}_i \mu_o(X_i) \right| \end{aligned}$$

Then decomposing based on the sign of the weights gives

$$\begin{aligned}\sum_{i=1}^n \hat{w}_i \mu_o(X_i) &= \sum_{i=1}^n \hat{w}_i [\mathbf{1}(\hat{w}_i \geq 0) + \mathbf{1}(\hat{w}_i < 0)] \mu_o(X_i) \\ &= \sum_{i=1}^n \hat{w}_i \mathbf{1}(\hat{w}_i \geq 0) \mu_o(X_i) + \sum_{i=1}^n \hat{w}_i \mathbf{1}(\hat{w}_i < 0) \mu_o(X_i)\end{aligned}$$

By definition  $\mu_o$  is odd, which gives  $\mu_o(-X_i) = -\mu_o(X_i)$ . The additional worst-case bias from negative weights would be:

$$\begin{aligned}&\sum_{i=1}^n |\hat{w}_i| \mathbf{1}(\hat{w}_i < 0) [\mu_o(-X_i) - \mu_o(X_i)] \\ &= \sum_{i=1}^n |\hat{w}_i| \mathbf{1}(\hat{w}_i < 0) [-\mu_o(X_i) - \mu_o(X_i)] = -2 \sum_{i=1}^n |\hat{w}_i| \mathbf{1}(\hat{w}_i < 0) \mu_o(X_i)\end{aligned}$$

By applying the definition of the odd function the bias cancels out with the original bias term giving

$$\left| \mu_o(x^*) - \sum_{i=1}^n \hat{w}_i \mu_o(X_i) \right| = \left| \sum_{i=1}^n \hat{w}_i [\mu_o(X_i) - \mu_o(x^*)] \right|$$

For the even component we have

$$\begin{aligned}\left| \mu_e(x^*) - \sum_{i=1}^n \hat{w}_i \mu_e(X_i) \right| &\leq \left| \sum_{i=1}^n \hat{w}_i [\mu_e(X_i) - \mu_e(x^*)] \right| \\ &\quad + 2 \sum_{i=1}^n |\hat{w}_i| \mathbf{1}(\hat{w}_i < 0) |\mu_e(-X_i) - \mu_e(x^*)|\end{aligned}$$

Applying Hölder continuity of  $\mu_e$  and using the worst-case bias terms gives

$$\left| \mu_e(x^*) - \sum_{i=1}^n \hat{w}_i \mu_e(X_i) \right| \leq \left| \sum_{i=1}^n \hat{w}_i [\mu_e(X_i) - \mu_e(x^*)] \right| + 2 \sum_{i=1}^n |\hat{w}_i| \mathbf{1}(\hat{w}_i < 0) a \|X_i - x^*\|^\alpha$$

Putting the even and odd component portions together gives

$$|\mu(x^*) - \hat{\mu}(x^*)| \leq B_{\text{even}}(x^*) + \left| \sum_{i=1}^n \hat{w}_i \epsilon_i \right|$$

Now turning to the noise term, we have by assumption that the sum  $\sum_{i=1}^n \hat{w}_i \epsilon_i$  is sub-Gaussian with parameter  $\sigma \|\hat{w}\|_2$ . Using standard sub-Gaussian concentration,

$$P\left(\left|\sum_{i=1}^n \hat{w}_i \epsilon_i\right| > t \mid \hat{w}\right) \leq 2 \exp\left(-\frac{t^2}{2\sigma^2 \|\hat{w}\|_2^2}\right)$$

Solving for  $t$ , setting the right hand side to  $\delta$

$$\begin{aligned}2 \exp\left(-\frac{t^2}{2\sigma^2 \|\hat{w}\|_2^2}\right) &= \delta \\ \exp\left(-\frac{t^2}{2\sigma^2 \|\hat{w}\|_2^2}\right) &= \frac{\delta}{2} \\ -\frac{t^2}{2\sigma^2 \|\hat{w}\|_2^2} &= \log\left(\frac{\delta}{2}\right) \\ t^2 &= -2\sigma^2 \|\hat{w}\|_2^2 \log\left(\frac{\delta}{2}\right) = 2\sigma^2 \|\hat{w}\|_2^2 \log\left(\frac{2}{\delta}\right)\end{aligned}$$

We then have,  $t = \sigma \|\hat{w}\|_2 \sqrt{2 \log(2/\delta)}$ , and in turn that with probability at least  $1 - \delta$  we have

$$\left| \sum_{i=1}^n \hat{w}_i \epsilon_i \right| \leq \sigma \|\hat{w}\|_2 \sqrt{2 \log(2/\delta)}$$

We can then obtain our desired statement by taking the expectation over  $\hat{w}$  and combining it with the bias bound.  $\square$

## C.2 Proof of Proposition 3.2

*Proof.* Our approach will be to modify Proposition 3.1 which requires access to the even and odd components, with empirical estimates of the even components. Recall the previous statement was

$$|\mu(x^*) - \hat{\mu}(x^*)| \leq B_{\text{even}}(x^*) + \sigma \|\hat{w}\|_2 \sqrt{2 \log(2/\delta)}$$

where

$$B_{\text{even}}(x^*) = \left| \sum_{i=1}^n \hat{w}_i [\mu_e(X_i) - \mu_e(x^*)] \right| + 2 \sum_{i=1}^n |\hat{w}_i| \mathbf{1}(\hat{w}_i < 0) a \|X_i - x^*\|^\alpha.$$

We first rewrite the bias as

$$\left| \sum_{i=1}^n \hat{w}_i \mu_e(X_i) - \mu_e(x^*) \sum_{i=1}^n \hat{w}_i \right|$$

and replace  $\mu_e(X_i)$  with observable approximations. For  $i \in I_{\text{paired}}$ ,

$$\mu_e(X_i) = \frac{Y_i + Y_{-i}}{2} - \frac{\epsilon_i + \epsilon_{-i}}{2}$$

. For  $i \in I_{\text{nn}}$ ,

$$\mu_e(X_i) = \frac{Y_i + Y_{j^*(i)}}{2} - \frac{\epsilon_i + \epsilon_{j^*(i)}}{2} + \frac{\mu(X_{j^*(i)}) - \mu(-X_i)}{2}$$

where  $Y_{j^*(i)}$  is the approximation of  $Y_{-i}$  using NN, and  $\left| \frac{\mu(X_{j^*(i)}) - \mu(-X_i)}{2} \right| \leq a \|X_{j^*(i)} - (-X_i)\|^\alpha$  by Hölder continuity. Substituting those terms in gives

$$\begin{aligned} \left| \sum_{i=1}^n \hat{w}_i \mu_e(X_i) - \mu_e(x^*) \sum_{i=1}^n \hat{w}_i \right| &\leq \left| \sum_{i \in I_{\text{paired}}} \hat{w}_i \frac{Y_i + Y_{-i}}{2} + \sum_{i \in I_{\text{nn}}} \hat{w}_i \frac{Y_i + Y_{j^*(i)}}{2} - \mu_e(x^*) \sum_{i=1}^n \hat{w}_i \right| \\ &+ \left| \sum_{i \in I_{\text{paired}}} \hat{w}_i \frac{\epsilon_i + \epsilon_{-i}}{2} \right| + \left| \sum_{i \in I_{\text{nn}}} \hat{w}_i \frac{\epsilon_i + \epsilon_{j^*(i)}}{2} \right| + \sum_{i \in I_{\text{nn}}} |\hat{w}_i| a \|X_{j^*(i)} - (-X_i)\|^\alpha. \end{aligned}$$

To address the fact that  $\mu_e(x^*)$  is unobservable, we bound it using the closest observation  $j^* = \arg \min_j \|X_j - x^*\|$ . By Hölder continuity:

$$|\mu_e(x^*)| \leq |\mu_e(X_{j^*})| + a \|X_{j^*} - x^*\|^\alpha.$$

For  $j^* \in I_{\text{paired}}$ ,

$$|\mu_e(X_{j^*})| \leq \left| \frac{Y_{j^*} + Y_{-j^*}}{2} \right| + \left| \frac{\epsilon_{j^*} + \epsilon_{-j^*}}{2} \right|.$$

A similar analysis holds for  $j^* \in I_{\text{nn}}$ .

Finally for the noise terms, we apply concentration with  $\delta/4$  allocation:

$$\begin{aligned} \left| \sum_{i=1}^n \hat{w}_i \epsilon_i \right| &\leq \sigma \|\hat{w}\|_2 \sqrt{2 \log(6/\delta)}, \\ \left| \sum_{i \in I_{\text{paired/nn}}} \hat{w}_i \frac{\epsilon_i + \epsilon_j}{2} \right| &\leq \frac{\sigma}{\sqrt{2}} \|\hat{w}_{I_{\text{paired/nn}}}\|_2 \sqrt{2 \log(6/\delta)} \end{aligned}$$



since  $\frac{\epsilon_i + \epsilon_j}{2}$  is sub-Gaussian with parameter  $\frac{\sigma}{\sqrt{2}}$ ,

$$\left| \frac{\epsilon_{j^*} + \epsilon_{-j^*}}{2} \right| \leq \sigma \sqrt{2 \log(6/\delta)}.$$

By union bound, all hold with probability  $1 - \delta$ . The final result follows by combining each constituent term.  $\square$

### C.3 Proof of proposition 3.3

*Proof.* By Proposition 3.2, with probability at least  $1 - \delta$ :

$$|\mu(x^*) - \hat{\mu}(x^*)| \leq B_{\text{even}}(x^*) + \sigma \|\hat{w}\|_2 \sqrt{2 \log(2/\delta)} \quad (4)$$

We bound  $\|\hat{w}\|_2$  using the first-order optimality conditions. For each  $i$ :

$$2 \sum_{j=1}^n \hat{w}_j X_j^T X_i - 2(x^*)^T X_i + 2\lambda \hat{w}_i + \gamma \|X_i\|^\alpha \xi_i = 0 \quad (5)$$

where  $\xi_i \in \partial(|w_i| \mathbf{1}(w_i < 0))|_{w_i = \hat{w}_i}$ . Multiplying by  $\hat{w}_i$  and summing over all  $i$ :

$$2 \left\| \sum_{j=1}^n \hat{w}_j X_j \right\|_2^2 - 2(x^*)^T \sum_{i=1}^n \hat{w}_i X_i + 2\lambda \|\hat{w}\|_2^2 - \gamma \sum_{i=1}^n |\hat{w}_i| \mathbf{1}(\hat{w}_i < 0) \|X_i\|^\alpha = 0 \quad (6)$$

Rearranging and applying Young's inequality  $(x^*)^T \sum_i \hat{w}_i X_i \leq \frac{1}{2} \|x^*\|_2^2 + \frac{1}{2} \|\sum_i \hat{w}_i X_i\|_2^2$ :

$$\lambda \|\hat{w}\|_2^2 \leq \frac{1}{2} \|x^*\|_2^2 + \frac{\gamma R^\alpha}{2} \sum_{i=1}^n |\hat{w}_i| \mathbf{1}(\hat{w}_i < 0) \quad (7)$$

By Cauchy-Schwarz,  $\sum_{i=1}^n |\hat{w}_i| \mathbf{1}(\hat{w}_i < 0) \leq \sqrt{n} \|\hat{w}\|_2$ . This gives:

$$\|\hat{w}\|_2^2 - \frac{\gamma R^\alpha \sqrt{n}}{2\lambda} \|\hat{w}\|_2 - \frac{\|x^*\|_2^2}{2\lambda} \leq 0 \quad (8)$$

Solving this quadratic inequality:

$$\|\hat{w}\|_2 \leq \frac{\gamma R^\alpha \sqrt{n}}{4\lambda} + \sqrt{\frac{\gamma^2 R^{2\alpha} n}{16\lambda^2} + \frac{\|x^*\|_2^2}{2\lambda}} \leq \frac{\|x^*\|_2}{\sqrt{\lambda}} + \frac{\gamma R^\alpha \sqrt{n}}{\lambda} \quad (9)$$

where we used  $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$  and simplified. Substituting into Proposition 3.2 completes the proof.  $\square$

## D Synthetic Data Study

### D.1 Challenge Study

In this section, we study the performance of our estimator via two data generative process (DGP), one using a linear DGP and the second one using a non-linear DGP. In this simulation study, we specifically consider a challenging case when the target point  $x^*$  is outside the convex hull of the training points  $\{X_1, \dots, X_n\}$ . For linear DGP, the outcome  $Y$  is a linear function of  $X$  while for nonlinear DGP, the outcome is a quadratic function of  $X$ . The two DGPs are as follows:

$$\text{Linear DGP: } \mu(X) = \beta^T X$$

$$\text{Nonlinear DGP: } \mu(X) = 2X_0^2 + X_1 + X_0 X_1 + \epsilon$$

To further make the setup more challenging, we consider a relatively small sample with 10 training units and a target unit as shown in Figure 4. This scenario is particularly interesting because of the limited sample size compared to the problem's dimensionality:  $n/p = 5$ .

For the linear DGP, our parametric assumption holds. We observe that increasing the regularization on extrapolation ( $\gamma$ ) decreases the negative influence while increasing the balance RMSE, as shown in

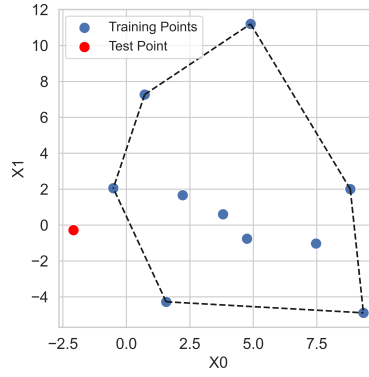


Figure 4: Convex Hull of source and target units in the simulation Study

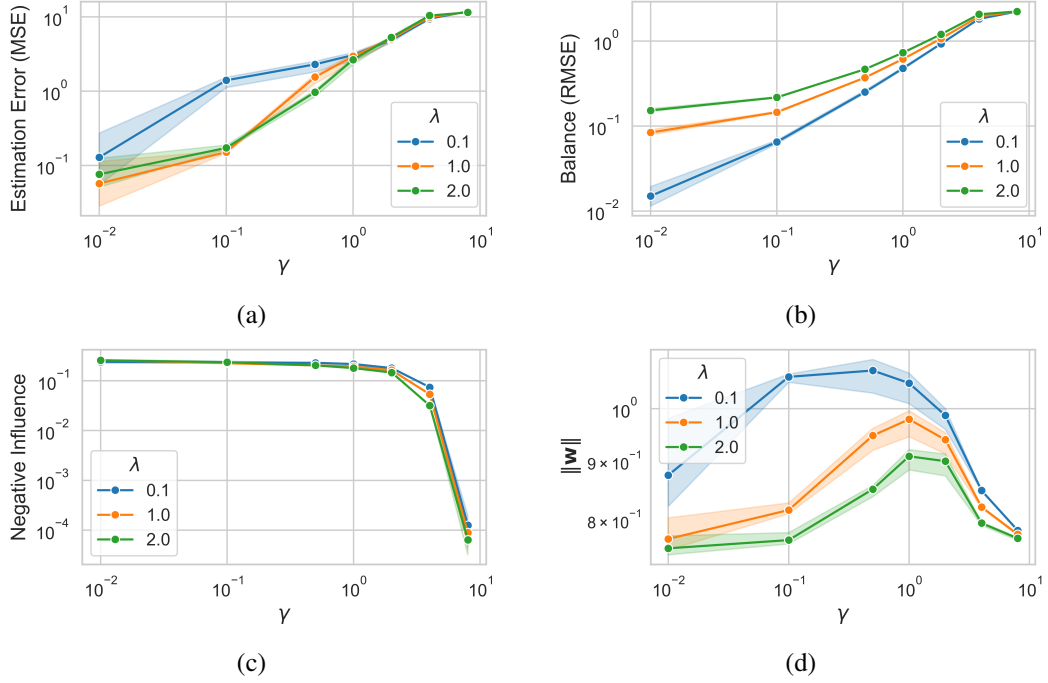


Figure 5: Results on synthetic data generated using linear DGP. (a) Estimation error measured as mean squared error, (b) balance between the weighted source and target populations, (c) extent of extrapolation measured as negative influence – contribution on units with negative weights, (d) L2 norm  $w$  capturing asymptotic variance.

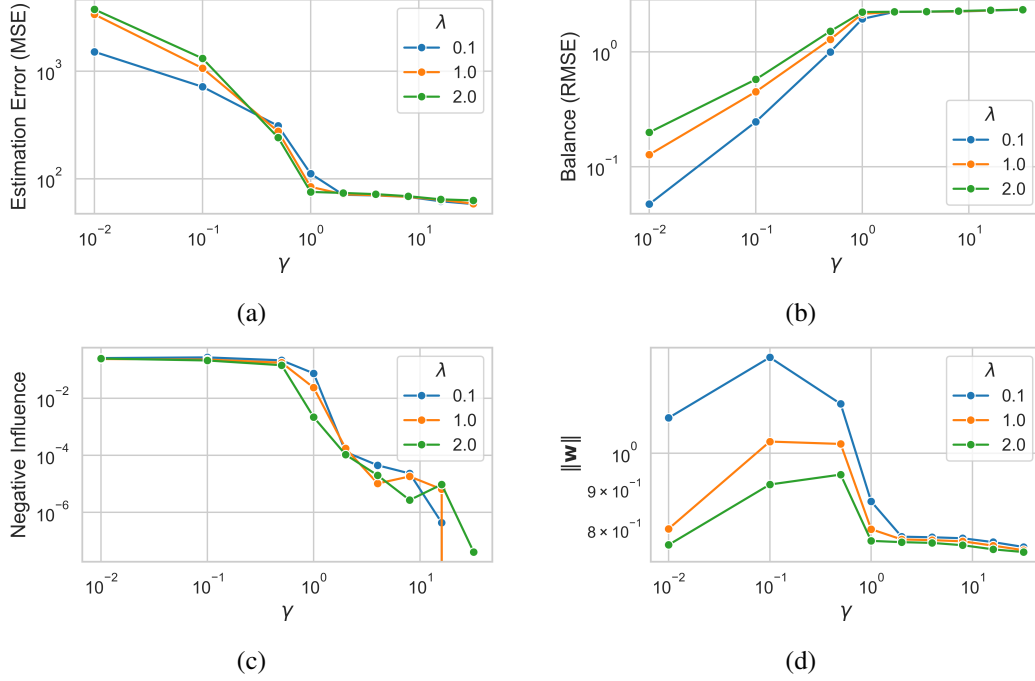


Figure 6: Results on synthetic data generated using non-linear DGP. (a) Estimation error measured as mean squared error, (b) balance between the weighted source and target populations, (c) extent of extrapolation measured as negative influence – contribution on units with negative weights, (d) L2 norm  $w$  capturing asymptotic variance.

Figures 5(b) and (c) – consistent with theoretical expectations. As underlying DGP is linear, relying on parametric assumptions for extrapolation yields optimal estimates with the smallest estimation error corresponding to least level of regularization on extrapolation (see Figure 5(a)).

Unlike linear DGP, the outcome function in the nonlinear DGP is not an odd function and thus the parametric assumption is violated. The outcome function has a quadratic term, an interaction term, and a linear term. Intuitively, we expect that a small amount of extrapolation might be beneficial due to the linear component however large amount of extrapolation may result in a high error rate due to violation of parametric assumption. The estimation error rate shown in Figure 6(a) is in congruency with the above-discussed expectation – thus highlighting tradeoff between two different kinds of biases .

## D.2 High-dimensional Large Scale Study

We now expand our analysis to a large scale high-dimensional data. In particular, we consider a scenario where  $n/p$  is large. We generate synthetic data via Friedman DGP which is commonly used across machine learning and causal inference literature. We generate  $n_{control} = 1000$  control units and  $n_{treat} = 1000$  with  $p = 5000$  covariates. The DGP is Friedman DGP is defined as follows:

$$Y(0) = 10 * \sin(\pi X_0 X_1) + 20(X_2 - 0.5)^2 + 10X_3 + 5X_4 + \epsilon,$$

and  $Y(1) = Y(0)$ . In this example, we are interested in estimating average treatment effect on the treated – the truth is 0. Here, the control units are used as the training set estimate counterfactual for the treated units which act as estimation set here –  $n/p = 1000/5000 = 0.2$ . We compare our approach with IPW, AIPW and Random Forest (RF). One main difference between our framework and AIPW or RF is that AIPW and RF both uses outcome information to learn the estimator while our framework (similar to IPW) is outcome data agnostic and only uses covariate information to ensure balance while also regularizing extrapolation. Figure 7 shows the mean squared error in estimating ATT and compares it with IPW, AIPW and RF. Our results shows that sweeping over the range of  $\gamma$  (i.e. regularizing extrapolation) smoothly ‘interpolates’ between AIPW and IPW estimators (former allows extrapolation using linear regression outcome model while other only relies on interpolation.)

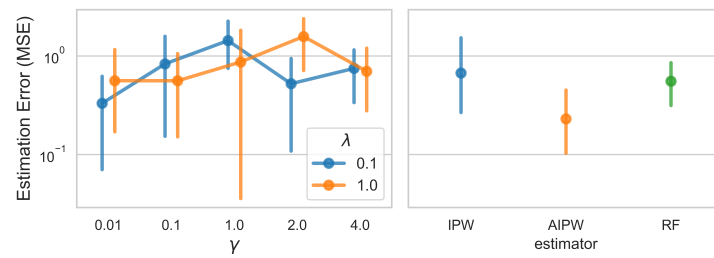


Figure 7: Results on synthetic data generated using High-dimensional Friedman's DGP.