Regularizing Extrapolation in Causal Inference

David Arbour[*]
Adobe Research

Harsh Parikh[*]
Yale University

Bijan Niknam
Johns Hopkins University

Elizabeth Stuart
Johns Hopkins University

Kara Rudolph
Columbia University

Avi Feller
University of California, Berkeley

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.

[*]co-first authors; alphabetical order

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

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 arms 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).

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 be viewed as performing controlled extrapolation, balancing model dependence against feature imbalance. Pure weighting and pure outcome modeling thus represent the two extremes of no versus 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 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 characterize the finite-sample variance through our error bounds.
- *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 focus on 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, $\mathbb{E}[Y_i(1) - Y_i(0) \mid Z_i = 1]$; we discuss generalizations below. 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 $\mu(\boldsymbol{x}) = \mathbb{E}[Y(0) \mid X_i = \boldsymbol{x}]$, 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*, $\boldsymbol{x}^\star \in \mathcal{X}$, corresponding to our estimand of interest. For the ATT, this profile is simply the mean of the treated population, $\boldsymbol{x}^\star = \mathbb{E}[X \mid Z = 1]$, where $\mu(\boldsymbol{x}^\star) = \mathbb{E}[Y(0) \mid Z = 1]$.

Linear in features. Since we are focused on linear smoothers, we consider models that are linear in *some* features, but which could be 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 $\boldsymbol{x}$ be the features in the representation implied by the parametric model, rather than simply the raw covariates. We further assume:

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

Parameterizing μ 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 $\phi : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}^d$; 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 $\boldsymbol{x}^\star$ with the much more general feature profile $\phi^\star(x, z)$.

2.2 Weighting form of causal inference estimators

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

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

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

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, $e(\boldsymbol{x}) = \mathbb{P}[Z_i = 1 \mid \boldsymbol{X}_i = \boldsymbol{x}]$ via a binary classifier like logistic regression, and then plug into a known functional form for $dQ/dP(x)$. For the ATT, $\hat{w}(\boldsymbol{x}) = \hat{e}(\boldsymbol{x})/(1 - \hat{e}(\boldsymbol{x}))$; since $\hat{e}(\boldsymbol{x}) \in (0, 1)$, $\hat{w}_i(\boldsymbol{x}) > 0$ for all i.

- *Balancing weights, synthetic control, and matching.* An alternative weighting approach instead directly estimates $dQ/dP(x)$ via constrained optimization (Ben-Michael et al., 2021a). For example, consider the minimum variance weights that control imbalance in $\boldsymbol{x}$ between P and Q:

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

  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 that 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 $\boldsymbol{X}$, the implied ridge regression weights are:

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

  where λ 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 $\phi(\boldsymbol{x})$; 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:

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

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

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

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

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

$$\hat{\mu}^{dr}(\boldsymbol{x}^\star) = \sum_{i=1}^{N} \left(\hat{w}_i^0 + \hat{w}_i^{\text{adj}} \right) Y_i$$

$$\text{where} \quad \hat{w}_i^{\text{adj}} \equiv \hat{\omega}_i(\boldsymbol{x}^\star) - \sum_{j=1}^{n} \hat{w}_j^0 \hat{\omega}_i(\boldsymbol{x}_j).$$

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

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

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

There are many examples of combined estimators of this form: standard Augmented IPW (Chattopadhyay 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) have this form. 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 $|\mu(\boldsymbol{x}^\star) - \sum_{i=1}^n w_i Y_i|$. We begin by building intuition for our approach in three steps.

Reflection representation. Under linearity, $\mu(-\boldsymbol{x}_i) = -\mu(\boldsymbol{x}_i)$, so a negative weight $w_i < 0$ on $\boldsymbol{x}_i$ yields $w_i\mu(\boldsymbol{x}_i) = |w_i|\mu(-\boldsymbol{x}_i)$; in other words, a negative weight w_i is equivalent to applying a positive weight $|w_i|$ to the reflected point $-\boldsymbol{x}_i$. We can then construct a "reflected" estimator, denoted by ‡, which reflects points with negative weights around the origin:

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

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

Measuring parametric model violations. The difference between $\hat{\mu}(\boldsymbol{x}^\star)$ and $\hat{\mu}^\ddagger(\boldsymbol{x}^\star)$ measures the degree to which the assumed parametric model is violated. Specifically, we can write $\mu(-\boldsymbol{x}_i) = \delta(\boldsymbol{x}_i) - \mu(\boldsymbol{x}_i)$, where $\delta(\boldsymbol{x}_i) \equiv \mu(-\boldsymbol{x}_i) + \mu(\boldsymbol{x}_i)$. If μ is an odd function (e.g., μ is linear through the origin), then $\delta(\boldsymbol{x}_i) = 0$ for all i. Thus, $\delta(\boldsymbol{x}_i)$ is a point-specific measure of the extent to which the true outcome function violates the assumed parametric model; throughout, our discussion of "nonlinearity" should be understood as referring to such violations.

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

$$\begin{aligned}\hat{\mu}(\boldsymbol{x}^\star) &= \sum_{i=1}^n w_i Y_i \ = \ \sum_{i=1}^n w_i(\mu(\boldsymbol{X}_i) + \epsilon_i) \\ &= \sum_{i=1}^n w_i \mathbb{1}(w_i \geq 0)\mu(\boldsymbol{X}_i) + \\ &\qquad |w_i|\ \mathbb{1}(w_i < 0)\left(\mu(-\boldsymbol{X}_i) - \delta(\boldsymbol{X}_i)\right) + w_i\epsilon_i \\ &= \underbrace{\sum_{i=1}^n |w_i|\mu(\boldsymbol{X}_i^\ddagger)}_{\hat{\mu}^\ddagger(\boldsymbol{x}^\star)} + \underbrace{\sum_{i=1}^n w_i \mathbb{1}(w_i < 0)\delta(\boldsymbol{X}_i)}_{\text{model violation}} + \underbrace{\sum_{i=1}^n w_i\epsilon_i}_{\text{noise}}.\end{aligned}$$

Error bound. Although $\delta(\boldsymbol{X})$ is unknown, we can bound it via Hölder continuity: $|\delta(\boldsymbol{x})| = |\mu(\boldsymbol{x}) + \mu(-\boldsymbol{x})| \leq |\mu(\boldsymbol{x}) - \mu(\boldsymbol{0})| + |\mu(-\boldsymbol{x}) - \mu(\boldsymbol{0})| \leq 2a\|\boldsymbol{x}\|^\alpha$, where the last step uses $\mu(\boldsymbol{0}) = 0$, which holds after centering.* The resulting error bound is therefore

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

The first term directly depends on the imbalance between the target point $\boldsymbol{x}^\star$ and the re-weighted (reflected) training points $|\boldsymbol{w}|'\boldsymbol{X}^\ddagger$. The second term captures additional error due to model violation, corresponding to the $\delta(\boldsymbol{X})$ term above; this is the key new term that our framework regularizes. The third term is the noise.

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 μ. We do so by considering the decomposition of μ into its even and odd components, i.e., $\mu(x) = \mu_e(x) + \mu_o(x)$. By the definition of odd functions, we have $-\mu_o(x) = \mu_o(-x)$; we can then bound the worst-case risk of $\hat{\boldsymbol{w}}$ using the assumed Hölder constants a and α and isolate 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) (defined below). Given $Y_i = \mu(X_i) + \epsilon_i$ where ϵ_i are independent random variables with $\mathbb{E}[\epsilon_i] = 0$ and finite second moment $\sigma^2 = \mathbb{E}[\epsilon_i^2]$, and μ is Hölder continuous with constants a and α. If ϵ_i are sub-Gaussian† with parameter σ, then with probability at least $1 - \delta$,*

$$\begin{aligned}|\mu(x^*) - \hat{\mu}(x^*)| \leq &\left|\mu(x^*) - \sum_{i=1}^n |\hat{w}_i|\,\mu(\boldsymbol{X}_i^\ddagger)\right| \\ &+ 2\sum_{i=1}^n |\hat{w}_i|\,\mathbf{1}(\hat{w}_i < 0)\,a\|\boldsymbol{X}_i\|^\alpha + \sigma\|\hat{w}\|_2\sqrt{2\log(2/\delta)}\end{aligned} \quad (3)$$

*Replace Y_i with $Y_i - \hat{\mu}(\boldsymbol{0})$, where $\hat{\mu}(\boldsymbol{0})$ is the fitted intercept.

†We assume mean zero sub-Gaussian noise, analogous results can be obtained with this assumption replaced by bounded noise.

where $\boldsymbol{X}_i^{\ddagger} = \text{sign}(\hat{w}_i)\boldsymbol{X}_i$ *as defined above. The first term is the imbalance of the reflected estimator* $\hat{\mu}^{\ddagger}(x^*) = \sum_i |\hat{w}_i|\mu(\boldsymbol{X}_i^{\ddagger})$*; the second is the model violation bias due to negative weights (identical to the second term of* (2)*); the third is the noise.*

The proof in Appendix C proceeds via an even-odd decomposition that shows that negative weights introduce additional bias only through the *even* component of μ; the odd part is absorbed exactly into the first term via the reflection $\boldsymbol{X}_i^{\ddagger}$. The noise term in Proposition 3.1 requires a sub-Gaussian assumption.

Since μ_e is unidentifiable from a single dataset, we construct a conservative worst-case form that does not require access to μ_e. For completeness, Proposition C.1 in Appendix C provides an empirical analog that approximates μ_e via nearest-neighbor matching when the data are approximately symmetric.

Finally, following Chattopadhyay and Zubizarreta (2023), we define *negative influence* as the fraction of total weight on units with negative weights, $\sum_i \mathbf{1}[w_i < 0]|w_i|/\sum_i |w_i|$. This is a useful summary of the extent to which the estimate relies on extrapolation.

3.2 Proposed Estimator

We now propose an estimator to learn weights $\boldsymbol{w}$ 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-negativity constraint, 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:

$$\hat{\mathbf{w}} \in \arg\min_{\mathbf{w}} \underbrace{\left\|\sum_{i=1}^n w_i \boldsymbol{X}_i - \boldsymbol{x}^\star\right\|_2^2}_{(a)\text{ imbalance}} + \underbrace{\lambda\|\mathbf{w}\|_2^2}_{(b)\text{ variance}} + \underbrace{\gamma \sum_{i=1}^n \mathbb{1}(w_i < 0)\,|w_i|\,\|\boldsymbol{X}_i\|^\alpha}_{(c)\text{ extrapolation}} \quad (4)$$

where

- Term (a): Enforces balance between the target point $\boldsymbol{x}^\star$ and the re-weighted training points $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$, recalling that $w_i\boldsymbol{X}_i = |w_i|\boldsymbol{X}_i^{\ddagger}$ for all i. We focus on $p = 2$, but this generalizes to $p = \infty$. This corresponds to the first term of (2).
- Term (b): Regularizes the dispersion of the weights $\mathbf{w}$, controlling the noise term via $\|\hat{\mathbf{w}}\|_2$ in Proposition 3.1.
- Term (c): Penalizes model violation bias: the penalty $\gamma \sum_i \mathbf{1}(w_i < 0)|w_i|\|\boldsymbol{X}_i\|^\alpha$ is proportional to the second term of (2), with γ scaling the sensitivity to parametric model violations.[‡]

Compared to the standard balancing weights problem (1), which trades off only imbalance and variance, the new objective (4) introduces term (c) to control extrapolation. When the target lies outside the convex hull of the training points, achieving balance requires some weights to be negative, which increases both $\|\boldsymbol{w}\|_2$ and reliance on parametric assumptions. For $\gamma = 0$, Equation (4) recovers unconstrained balancing weights; at the other extreme, $\gamma \to \infty$ is equivalent to a hard non-negativity constraint. Increasing γ reduces extrapolation bias and $\|\boldsymbol{w}\|_2$ but worsens imbalance in term (a).

Regularizing existing estimators. Since many causal estimators have a weighting representation (Section 2.2), we can regularize extrapolation in any baseline estimator with implied weights $\boldsymbol{w}'$ by solving

$$\hat{\mathbf{w}} \in \arg\min_{\mathbf{w}} \|\mathbf{w} - \mathbf{w}'\|_2^2 + \gamma \sum_{i=1}^n \mathbb{1}(w_i < 0)\,|w_i|\,\|\boldsymbol{X}_i\|^\alpha.$$

For example, the augmented synthetic control method (Ben-Michael et al., 2021b) first solves with non-negative weights, then augments with a ridge outcome model that implicitly introduces negative adjustment weights; the formulation above instead directly controls the degree of negativity through γ.

Convexity. Despite the indicator function in term (c), the optimization problem (4) is strongly convex and admits a unique global minimizer. To see this, note that $\mathbf{1}(w_i < 0)|w_i| = \max(0, -w_i)$, which is convex as the pointwise maximum of two affine functions. Term (a) is a squared norm of an affine function of $\boldsymbol{w}$, hence convex, and term (b) is strongly convex with parameter 2λ. More generally, term (c) can be written using an ℓ_p norm over the vector of per-unit penalties. When $p = 1$ (as written above and in Proposition 3.2), introducing slack variables $s_i \geq -w_i$, $s_i \geq 0$ reduces the problem to a quadratic program (QP), which can be solved exactly in polynomial time with standard solvers. When $p = 2$, the problem becomes a second-order cone program (SOCP), which is likewise solvable in polynomial time.

Finally, we can specialize the error bound for our proposed estimator:

Proposition 3.2 (Regularized Bound)**.** *Let* $\hat{w}$ *solve Equation* (4) *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 3.1,*

[‡]In practice, $\alpha = 1$ corresponds to Lipschitz continuity; larger α assumes smoother departures from the parametric model and penalizes extrapolation less aggressively.

with probability at least $1-\delta$*:*

$$|\mu(x^*) - \hat{\mu}(x^*)| \leq \left|\mu(x^*) - \sum_{i=1}^{n} |\hat{w}_i|\, \mu(\boldsymbol{X}_i^{\ddagger})\right|$$
$$+2\sum_{i=1}^{n} |\hat{w}_i|\, \mathbf{1}(\hat{w}_i < 0)\, a\|\boldsymbol{X}_i\|^{\alpha} + \sigma\|\boldsymbol{x}^*\|_2\sqrt{\log(2/\delta)/\lambda}$$

where $\boldsymbol{X}_i^{\ddagger} = \operatorname{sign}(\hat{w}_i)\boldsymbol{X}_i$ *as in Proposition 3.1.*

The proof is provided in the appendix. The variance term $\sigma\|\boldsymbol{x}^*\|_2\sqrt{\log(2/\delta)/\lambda}$ does not depend on γ: increasing γ further constrains the feasible set and cannot inflate $\|\hat{w}\|_2$. This means regularizing extrapolation reduces the bias terms without incurring additional variance cost. The full bias-bias-variance tradeoff is demonstrated empirically in Sections 4 and 5.

3.3 Practical guidance: γ as a sensitivity parameter

We argue that γ should be treated as a sensitivity parameter rather than a tuning parameter, and encourage researchers to examine the full set of estimates it spans. We recommend the following procedure:

1. Fix λ via cross-validation on the standard balancing weights problem (i.e., with $\gamma = 0$).
2. Sweep γ over a grid from 0 to a value $\gamma_{\max}$ at which all weights become non-negative.
3. For each γ, record the point estimate, covariate balance (RMSE), and negative influence.
4. Examine how estimates change across this range to assess sensitivity to parametric assumptions.

If the goal is a single point estimate, we can instead choose γ^* in the spirit of Lepski's method, selecting the largest γ for which the change in the point estimate remains below a researcher-defined cutoff.

4 SYNTHETIC DATA STUDY

We evaluate our approach using synthetic data with both linear and nonlinear data generating processes (DGPs) where the target point lies outside the convex hull of training points (Figure 1), creating a challenging extrapolation scenario with limited sample size ($n = 10$ training units, $n/p = 5$). We also consider a high-dimensional setting ($p = 5000$, $n/p = 0.2$) using the Friedman DGP. Full descriptions and additional figures are provided in Appendix D.

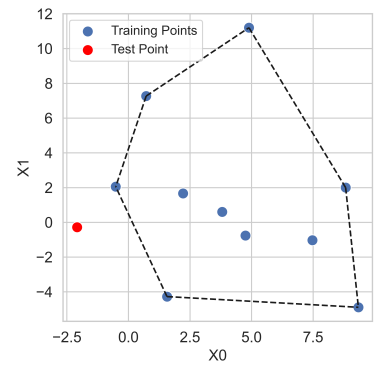

Figure 1: Convex hull of source (training) and target units. The target point lies outside the convex hull, requiring extrapolation.

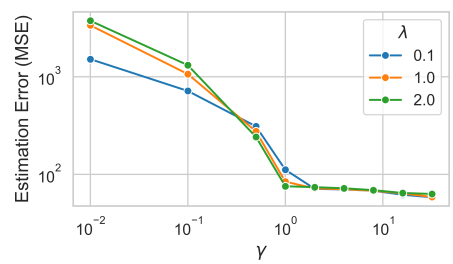

Figure 2: Estimation error (MSE) for the nonlinear DGP as a function of γ. The U-shaped curve illustrates the bias-bias-variance tradeoff: small γ allows beneficial extrapolation, while large γ incurs bias from poor balance.

We consider two DGPs:

$$\text{Linear:}\ \ \mu(\boldsymbol{X}) = \beta^{\top}\boldsymbol{X},$$
$$\text{Nonlinear:}\ \ \mu(\boldsymbol{X}) = 2X_1^2 + X_2 + X_1X_2.$$

For the linear DGP, where the parametric assumption holds, estimation error increases monotonically as we regularize extrapolation (i.e., increase γ): relying on correct parametric assumptions yields optimal estimates. However, for the nonlinear DGP (Figure 2), the quadratic and interaction terms violate the linearity assumption. The results illustrate the bias-bias-variance tradeoff predicted by our theory: small amounts of extrapolation remain beneficial due to the linear component, but excessive extrapolation leads to high error rates due to violations of the parametric model. The high-dimensional experiment further demonstrates that sweeping over γ smoothly interpolates between parametric regression-like behavior (unconstrained extrapolation) and IPW-like behavior (no extrapolation).

5 GENERALIZING MEDICATION FOR OPIOID USE DISORDER TRIAL EVIDENCE

We now apply our framework to the problem of generalizing causal estimates from a randomized trial to a target population. Appendix B states the formal identification assumptions and maps this setting onto the general framework of Section 2.2.

The Starting Treatment With Agonist Replacement Therapies (START) trial ($S_i = 1$), 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 ($S_i = 0$), highlighting a practical violation of the positivity assumption T.4.. In this study, we estimate the target average treatment effect (TATE), $\tau = \mathbb{E}[Y_i(1) - Y_i(0) \mid S_i = 0]$, 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 $Z = 1$ and buprenorphine as $Z = 0$, 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. For each treatment arm $z \in \{0, 1\}$, we estimate $\mu_z(\boldsymbol{x}^\star) = \mathbb{E}[Y_i \mid X_i = \boldsymbol{x}^\star, S_i = 1, Z_i = z]$ at the target profile $\boldsymbol{x}^\star = \mathbb{E}[X_i \mid S_i = 0]$ using the trial participants assigned to arm z as source units. The estimated TATE is then $\hat{\tau} = \hat{\mu}_1(\boldsymbol{x}^\star) - \hat{\mu}_0(\boldsymbol{x}^\star)$.

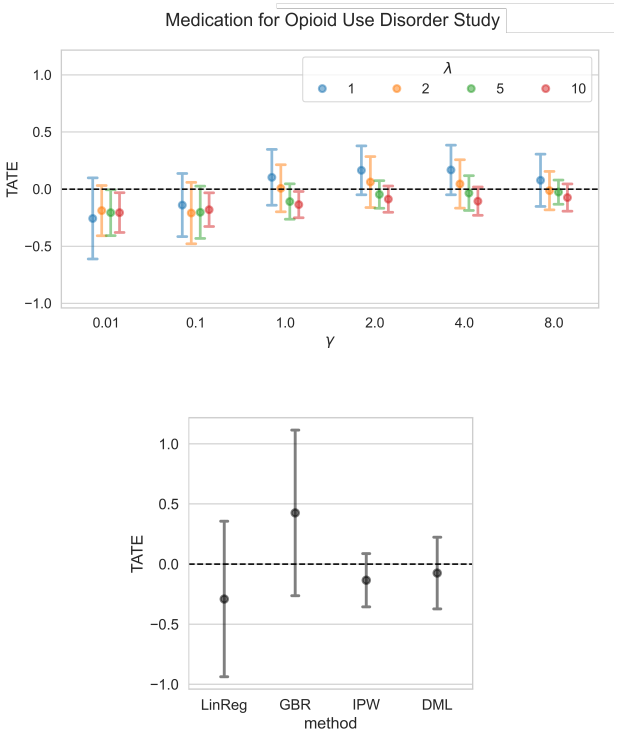

Figure 3: Target Average Treatment Effects for the Target Sample for Hispanic Females who have a history of Amphetamine and Benzodiazepine use in TEDS-A population. Each hue corresponds to a value of λ and the x-axis corresponds to different values of γ (on log scale).

We then apply our proposed framework to this problem. By varying γ 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. As regularization intensifies, however, the estimates smoothly shift towards zero and occasionally change sign from negative to positive for smaller values of λ. This sensitivity underscores the influence of assumptions on the point estimates. While increasing γ reduces negative influence (Figure 5), it worsens covariate balance, as reflected in higher RMSE values (Figure 4). Thus, our framework highlights a trade-off between minimizing reliance on parametric assumptions and achieving optimal covariate balance. Applied researchers should therefore interpret treatment effect estimates for this under-represented subgroup with caution given the sensitivity to modeling assumptions. As Parikh et al. (2025) emphasized, collecting more representative trial data is critical to credibly estimate treatment effects.

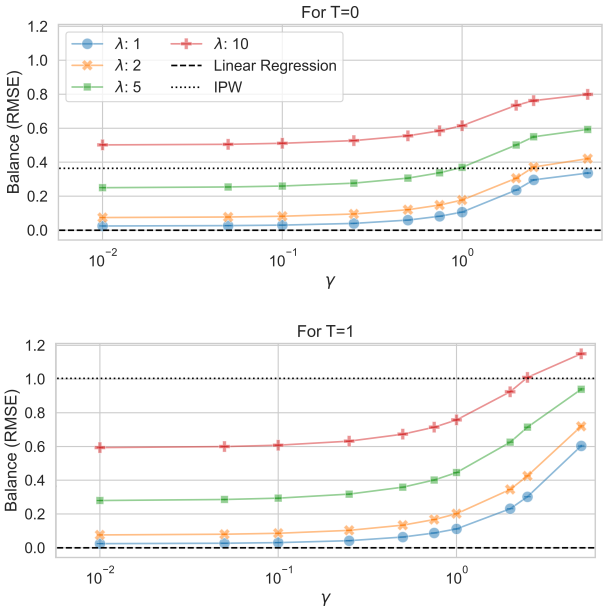

Figure 4: Balance between the trial and the target samples measured as the root mean squared error (RMSE) for different values of γ and λ.

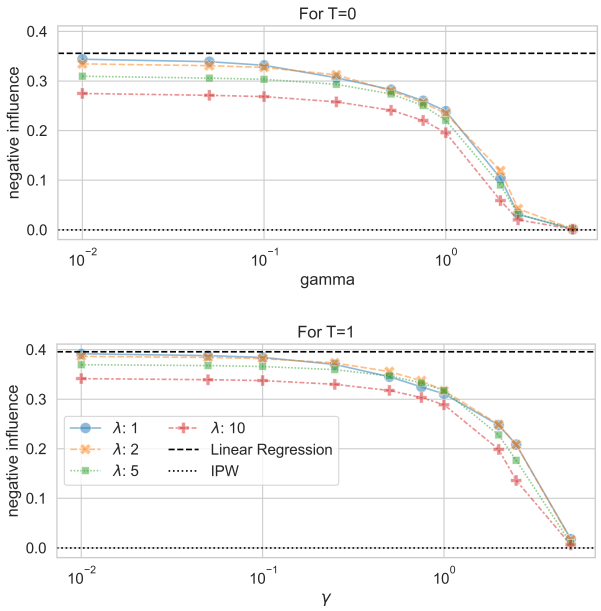

Figure 5: Negative influence, defined as the contribution of negative weights in estimation, for different values of γ and λ.

6 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 theoretical error bounds show a fundamental "bias-bias-variance" tradeoff between distributional imbalance, model misspecification, and estimator variance, decomposing extrapolation bias through a novel reflection perspective. Empirically, synthetic data experiments confirm that controlled extrapolation smoothly interpolates between fully constrained and unconstrained approaches. A real-world medication trial illustrates how sweeping over the regularization parameter provides a practical sensitivity analysis for transportability estimates under positivity violations.

Limitations and Future Work. Our approach focuses on weighting-type estimators and relies on Hölder continuity and conditional ignorability, which may not hold in practice. Operationalizing sensitivity analysis for unmeasured confounding is a critical next step; existing proposals for balancing weights (Soriano et al., 2023) do not directly apply to our framework, and adapting such methods is an important direction. More broadly, future work should extend the bias-bias-variance tradeoff analysis to more flexible estimator classes and weaker continuity assumptions. A key open question is to characterize data-generating processes under which soft-constrained extrapolation ($\gamma > 0$) provably improves MSE relative to both unconstrained ($\gamma = 0$) and fully constrained ($\gamma \to \infty$) estimators, for example when the density ratio is large near $\boldsymbol{x}^\star$. Finally, our theoretical results assume sub-Gaussian noise, though analogous bounds follow under bounded outcomes via Hoeffding-type inequalities.

ACKNOWLEDGMENTS

The authors would like to thank the reviewers, the area chair, and the program chair of AISTATS 2026 for their constructive input to help improve the paper. Harsh Parikh, Kara Rudolph, and Elizabeth Stuart would like to acknowledge that this work was funded by NIH NIDA R01DA056407.

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 multisite 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 trans-

ductive 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.

CHECKLIST

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [**Yes**]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [**Yes**]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [**No**]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [**Yes**]
   (b) Complete proofs of all theoretical results. [**Yes**]
   (c) Clear explanations of any assumptions. [**Yes**]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [**Yes**]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [**Yes**]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [**Yes**]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [**Yes**]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [**Yes**]
   (b) The license information of the assets, if applicable. [**Not Applicable**]
   (c) New assets either in the supplemental material or as a URL, if applicable. [**Not Applicable**]
   (d) Information about consent from data providers/curators. [**Not Applicable**]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [**Not Applicable**]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [**Not Applicable**]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [**Not Applicable**]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [**Not Applicable**]

Regularizing Extrapolation in Causal Inference: Supplementary Materials

A IMPLEMENTATION 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 a default learning rate of 0.01 for a 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 5, 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 $O(n^{3.67} * logn)$ 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 TRANSPORTABILITY ASSUMPTIONS FOR THE MOUD APPLICATION

The main text develops the framework for estimating $\mu(\boldsymbol{x}^\star) = \mathbb{E}[Y(0) \mid X = \boldsymbol{x}^\star]$ in the ATT setting. The MOUD application in Section 5 instead concerns *transportability*: generalizing treatment effect estimates from a randomized trial to a target population. Here we state the additional assumptions required and show how the general framework applies. For a comprehensive review, see Degtiar and Rose (2023).

Setup. Let $S_i \in \{0, 1\}$ indicate membership in the trial ($S_i = 1$) versus the target population ($S_i = 0$), and let $Z_i \in \{0, 1\}$ denote treatment assignment within the trial. The estimand is the *target average treatment effect* (TATE):

$$\tau = \mathbb{E}[Y_i(1) - Y_i(0) \mid S_i = 0].$$

Because neither potential outcome is observed in the target population, estimating the TATE requires identifying both $\mathbb{E}[Y_i(z) \mid S_i = 0]$ for $z \in \{0, 1\}$. We require the following assumptions:

T.1. (SUTVA) $Y_i = Y_i(z)$ when $Z_i = z$.

T.2. (Ignorability and overlap within the trial) $Z_i \perp\!\!\!\perp \{Y_i(0), Y_i(1)\} \mid X_i, S_i = 1$ and $\Pr(Z_i = z \mid X_i = \boldsymbol{x}, S_i = 1) > 0$. Both hold by design via randomization.

T.3. (Mean exchangeability over populations) For each $z \in \{0, 1\}$:

$$\mathbb{E}[Y_i(z) \mid X_i = \boldsymbol{x}, S_i = 1] = \mathbb{E}[Y_i(z) \mid X_i = \boldsymbol{x}, S_i = 0].$$

This assumes that, conditional on covariates, expected potential outcomes do not differ between the trial and target populations.

T.4. (Positivity of trial enrollment) $\Pr(S_i = 1 \mid X_i = \boldsymbol{x}) > 0$ for all $\boldsymbol{x}$ in the support of the target population.

As with the main results, a key goal is to assess sensitivity to violations of this last positivity assumption T.4..

Connection to the general framework. Under T.1.–T.3., for each treatment arm z:

$$\mathbb{E}[Y_i(z) \mid S_i = 0] = \mathbb{E}[\mu_z(X_i) \mid S_i = 0]\,,$$

where $\mu_z(\boldsymbol{x}) \equiv \mathbb{E}[Y_i \mid X_i = \boldsymbol{x},\, S_i = 1,\, Z_i = z]$ is the conditional mean outcome among trial participants assigned to treatment z. This is precisely the estimation problem in Section 2.2: the source units are trial participants in arm z (with observed outcomes), μ_z plays the role of μ, and the target covariate profile is $\boldsymbol{x}^\star = \mathbb{E}[X_i \mid S_i = 0]$. Assumption T.4. is the analog of the population overlap assumption A.2.; its violation for Hispanic women with amphetamine and benzodiazepine use history is the central challenge addressed in Section 5. The TATE is then $\tau = \hat{\mu}_1(\boldsymbol{x}^\star) - \hat{\mu}_0(\boldsymbol{x}^\star)$, where each arm is estimated separately using our proposed weights.

C PROOFS

C.1 Proof of Proposition 3.1

Proof. First taking the bound of the estimation error

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

We bound the bias term $|\mu(x^*) - \sum_{i=1}^{n} \hat{w}_i \mu(X_i)|$ by decomposing $\mu = \mu_e + \mu_o$ and establishing two preliminary identities.

Odd component. Since μ_o is odd, $\operatorname{sign}(w)\mu_o(X) = \mu_o(\operatorname{sign}(w)X)$ for any scalar w and vector X. Writing $\hat{w}_i = |\hat{w}_i| \operatorname{sign}(\hat{w}_i)$ and recalling $X_i^\ddagger = \operatorname{sign}(\hat{w}_i)X_i$,

$$\sum_{i=1}^{n} \hat{w}_i \mu_o(X_i) = \sum_{i=1}^{n} |\hat{w}_i| \operatorname{sign}(\hat{w}_i)\, \mu_o(X_i) = \sum_{i=1}^{n} |\hat{w}_i|\, \mu_o(\operatorname{sign}(\hat{w}_i)X_i) = \sum_{i=1}^{n} |\hat{w}_i|\, \mu_o(X_i^\ddagger),$$

so $\sum_{i=1}^{n} \hat{w}_i \mu_o(X_i) = \sum_{i=1}^{n} |\hat{w}_i|\, \mu_o(X_i^\ddagger)$ exactly.

Even component. Since μ_e is even, $\mu_e(X_i^\ddagger) = \mu_e(\operatorname{sign}(\hat{w}_i)X_i) = \mu_e(X_i)$ for all i. Applying Hölder continuity and the centering condition $\mu_e(\mathbf{0}) = 0$ (which holds after centering as described in the main text):

$$|\mu_e(X_i)| = |\mu_e(X_i) - \mu_e(\mathbf{0})| \leq a\|X_i\|^\alpha.$$

Combining. The odd-component analysis gives $\sum_i \hat{w}_i \mu_o(X_i) = \sum_i |\hat{w}_i| \mu_o(X_i^\ddagger)$, and writing $\hat{w}_i = |\hat{w}_i| - 2|\hat{w}_i|\mathbf{1}(\hat{w}_i < 0)$ for the even component gives

$$\begin{aligned}
\sum_i \hat{w}_i \mu(X_i) &= \sum_i |\hat{w}_i| \mu_e(X_i) - 2\sum_i |\hat{w}_i| \mathbf{1}(\hat{w}_i < 0)\mu_e(X_i) + \sum_i |\hat{w}_i| \mu_o(X_i^\ddagger) \\
&= \sum_{i=1}^{n} |\hat{w}_i| \mu(X_i^\ddagger) - 2\sum_{i=1}^{n} |\hat{w}_i| \mathbf{1}(\hat{w}_i < 0)\mu_e(X_i),
\end{aligned}$$

where the last step uses $\mu_e(X_i^\ddagger) = \mu_e(X_i)$. The triangle inequality then gives

$$|\mu(x^*) - \hat{\mu}(x^*)| \leq \left|\mu(x^*) - \sum_{i=1}^{n} |\hat{w}_i|\, \mu(X_i^\ddagger)\right| + 2\sum_{i=1}^{n} |\hat{w}_i|\, \mathbf{1}(\hat{w}_i < 0)\, |\mu_e(X_i)| + \left|\sum_{i=1}^{n} \hat{w}_i \epsilon_i\right|.$$

Applying $|\mu_e(X_i)| \le a\|X_i\|^\alpha$ from the even-component analysis above:

$$|\mu(x^*) - \hat{\mu}(x^*)| \le \left|\mu(x^*) - \sum_{i=1}^n |\hat{w}_i|\, \mu(X_i^\ddagger)\right| + 2\sum_{i=1}^n |\hat{w}_i|\, \mathbf{1}(\hat{w}_i < 0)\, a\|X_i\|^\alpha + \left|\sum_{i=1}^n \hat{w}_i \epsilon_i\right|.$$

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,

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

Solving for t, setting the right hand side to δ:

$$t^2 = 2\sigma^2\|\hat{w}\|_2^2 \log\left(\frac{2}{\delta}\right),$$

so $t = \sigma\|\hat{w}\|_2\sqrt{2\log(2/\delta)}$, and with probability at least $1-\delta$,

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

The desired statement follows by combining the bias and noise bounds. □

C.2 Proposition C.1 and Proof

Proposition C.1 (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 4. Given $Y_i = \mu(X_i) + \epsilon_i$ where ϵ_i are independent sub-Gaussian random variables with parameter σ, μ is Hölder continuous with constants a and α, and $\sum_{i=1}^n \hat{w}_i = 1$. Let $I_{paired} = \{i : -X_i \in \{X_1, \dots, X_n\}\}$, $I_{nn} = \{1, \dots, n\} \setminus I_{paired}$, and for each $i \in I_{nn}$, define $j^*(i) = \arg\min_{j\ne i} \|X_j - (-X_i)\|$. Then with probability at least $1-\delta$, $|\mu(x^*) - \hat{\mu}(x^*)|$ is at most*

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

where $\mu_e(x^*)$ is bounded using the closest observation $j^* = \arg\min_j \|X_j - x^*\|$:

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

Proof. Our approach will be to construct an observable bound by replacing $\mu_e(X_i)$ and $\mu_e(x^*)$ with empirical estimates. Starting from the bias-variance decomposition and applying the even-odd decomposition (as in the proof of Proposition 3.1), the estimation error is bounded by

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

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

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

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:

$$\left|\sum_{i=1}^n \hat{w}_i \epsilon_i\right| \leq \sigma \|\hat{w}\|_2 \sqrt{2\log(6/\delta)},$$

$$\left|\sum_{i \in I_{\text{paired/nn}}} \hat{w}_i \frac{\epsilon_i + \epsilon_j}{2}\right| \leq \frac{\sigma}{\sqrt{2}} \left\|\hat{w}_{I_{\text{paired/nn}}}\right\|_2 \sqrt{2\log(6/\delta)}$$

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. □

C.3 Proof of proposition 3.2

Proof. By Proposition 3.1, with probability at least $1 - \delta$:

$$|\mu(x^*) - \hat{\mu}(x^*)| \leq \left|\mu(x^*) - \sum_{i=1}^n |\hat{w}_i| \, \mu(\boldsymbol{X}_i^\ddagger)\right|$$
$$+ 2\sum_{i=1}^n |\hat{w}_i| \, \mathbf{1}(\hat{w}_i < 0) \, a\|\boldsymbol{X}_i\|^\alpha + \sigma\|\hat{w}\|_2 \sqrt{2\log(2/\delta)} \quad (5)$$

The improvement in the noise bound follows from bounding $\|\hat{w}\|_2$ via the first-order optimality conditions. For each i:

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

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:

$$2\left\|\sum_{j=1}^{n}\hat{w}_j X_j\right\|_2^2 - 2(x^*)^T\sum_{i=1}^{n}\hat{w}_i X_i + 2\lambda\|\hat{w}\|_2^2 + \gamma\sum_{i=1}^{n}|\hat{w}_i|\mathbf{1}(\hat{w}_i < 0)\|X_i\|^\alpha = 0 \tag{7}$$

where the sign of the γ term is positive because for $\hat{w}_i < 0$: $\xi_i = -1$ (subgradient of $\max(0, -w_i)$), so $\hat{w}_i \cdot \gamma\|X_i\|^\alpha \cdot \xi_i = (-|\hat{w}_i|) \cdot \gamma\|X_i\|^\alpha \cdot (-1) = +|\hat{w}_i|\gamma\|X_i\|^\alpha$.

Rearranging: $2\lambda\|\hat{w}\|_2^2 = 2(x^*)^T\sum_i \hat{w}_i X_i - 2\left\|\sum_j \hat{w}_j X_j\right\|_2^2 - \gamma\sum_i |\hat{w}_i|\mathbf{1}(\hat{w}_i < 0)\|X_i\|^\alpha$. Applying Young's inequality $(x^*)^T\sum_i \hat{w}_i X_i \leq \frac{1}{2}\|x^*\|_2^2 + \frac{1}{2}\|\sum_i \hat{w}_i X_i\|_2^2$:

$$\begin{aligned}2\lambda\|\hat{w}\|_2^2 &\leq \|x^*\|_2^2 + \left\|\textstyle\sum_i \hat{w}_i X_i\right\|_2^2 - 2\left\|\textstyle\sum_j \hat{w}_j X_j\right\|_2^2 - \gamma\textstyle\sum_i |\hat{w}_i|\mathbf{1}(\hat{w}_i < 0)\|X_i\|^\alpha \\ &= \|x^*\|_2^2 - \left\|\textstyle\sum_i \hat{w}_i X_i\right\|_2^2 - \gamma\textstyle\sum_i |\hat{w}_i|\mathbf{1}(\hat{w}_i < 0)\|X_i\|^\alpha \;\leq\; \|x^*\|_2^2\end{aligned}$$

since both subtracted terms are non-negative. Therefore:

$$\|\hat{w}\|_2 \leq \frac{\|x^*\|_2}{\sqrt{2\lambda}} \tag{8}$$

This bound notably does not depend on γ or the support radius R: increasing γ further constrains the feasible set and cannot inflate $\|\hat{w}\|_2$. Substituting into Proposition 3.1 completes the proof. □

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 $\boldsymbol{x}^\star$ is outside the convex hull of the training points $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$. For linear DGP, the outcome Y is a linear function of $\boldsymbol{X}$ while for nonlinear DGP, the outcome is a quadratic function of $\boldsymbol{X}$. The two DGPs are as follows:

$$\begin{aligned}\text{Linear DGP:}\mu(\boldsymbol{X}) &= \beta^T\boldsymbol{X} \\ \text{Nonlinear DGP:}\mu(\boldsymbol{X}) &= 2X_1^2 + X_2 + X_1X_2 + \epsilon\end{aligned}$$

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 1 (in main text). 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 (γ) decreases the negative influence while increasing the balance RMSE, as shown in Figures 6(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 6(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 7(a) is in congruency with the above-discussed expectation – thus highlighting the 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 p/n is large. We generate synthetic data via Friedman DGP which is commonly used across machine learning and

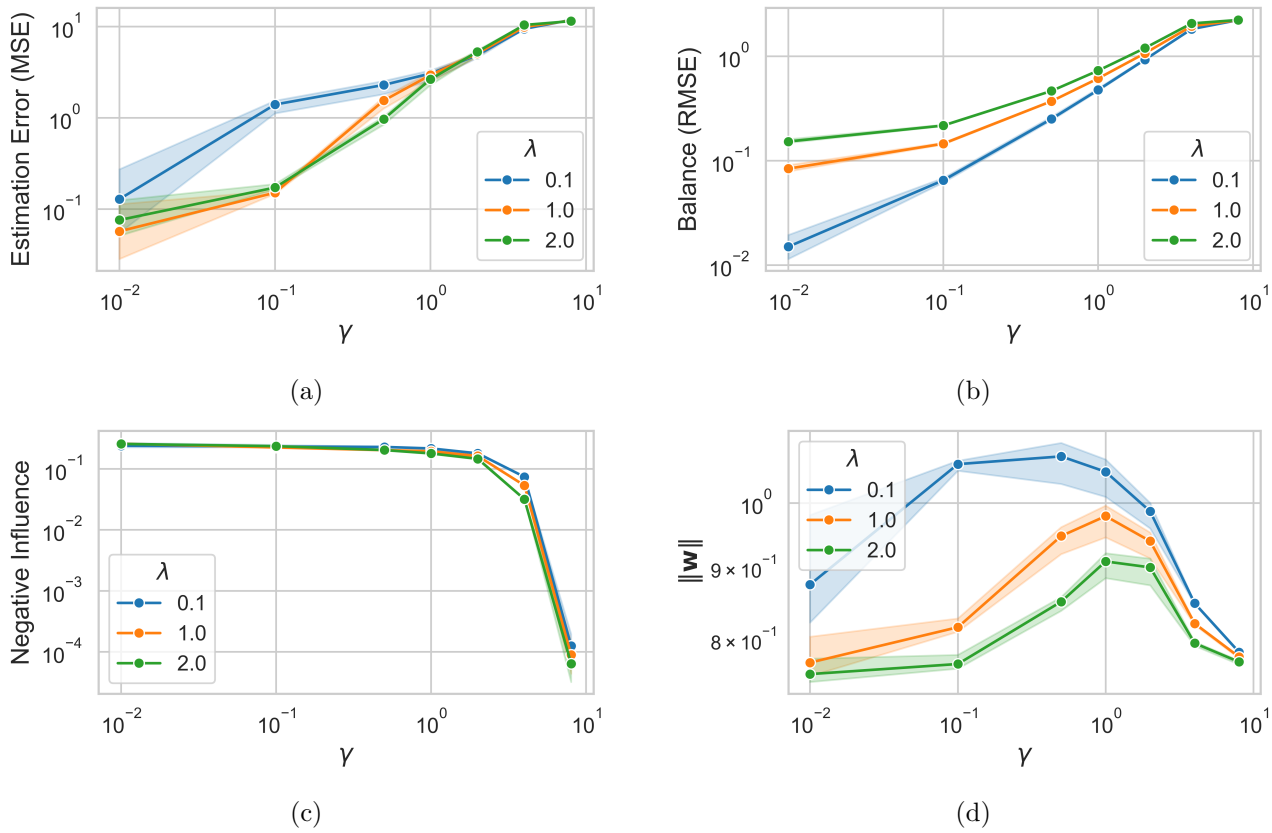

Figure 6: Results on synthetic data generated using linear DGP. (a) Estimation error measured as mean squared error, (b) balance between the weighted source and target populations, (c) extent of extrapolation measured as negative influence – contribution on units with negative weights, (d) L2 norm $\mathbf{w}$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

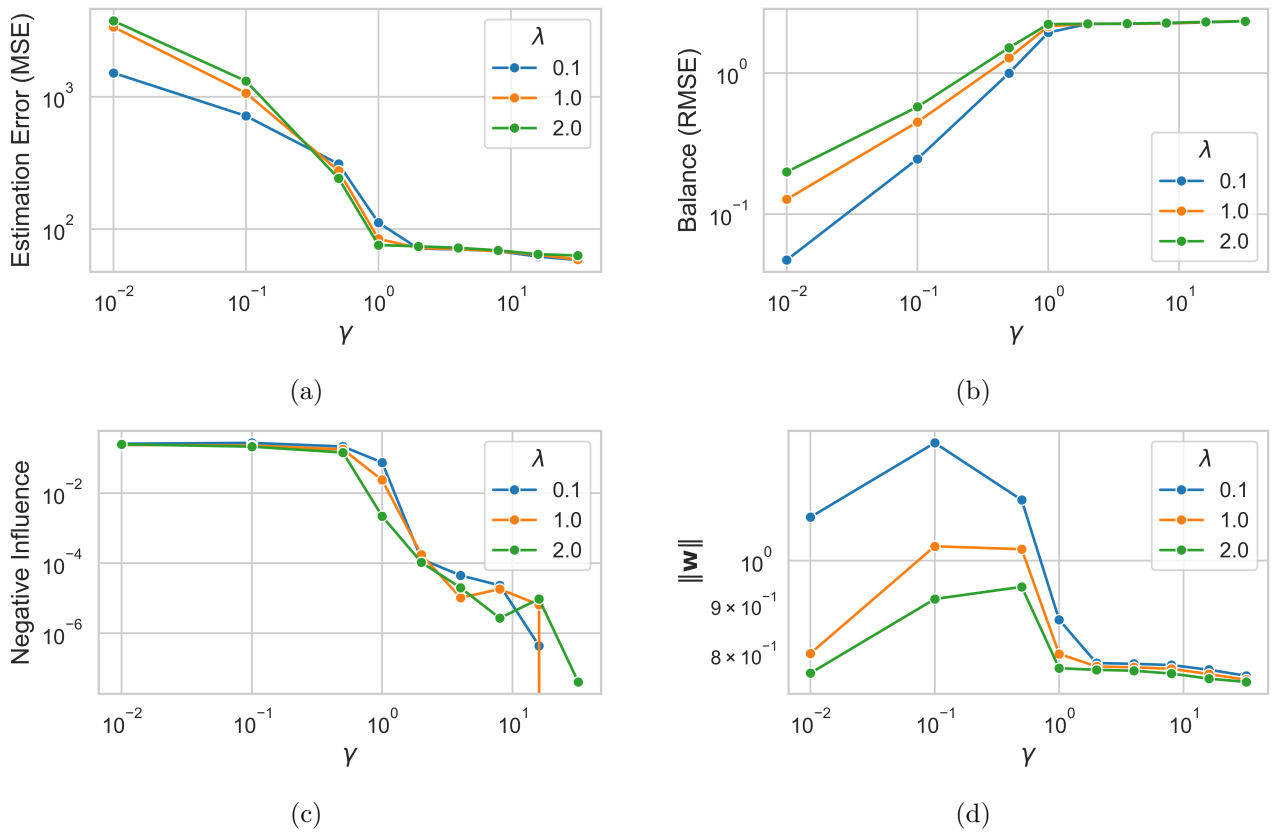

Figure 7: Results on synthetic data generated using non-linear DGP. (a) Estimation error measured as mean squared error, (b) balance between the weighted source and target populations, (c) extent of extrapolation measured as negative influence – contribution on units with negative weights, (d) L2 norm $\mathbf{w}$ capturing asymptotic variance.

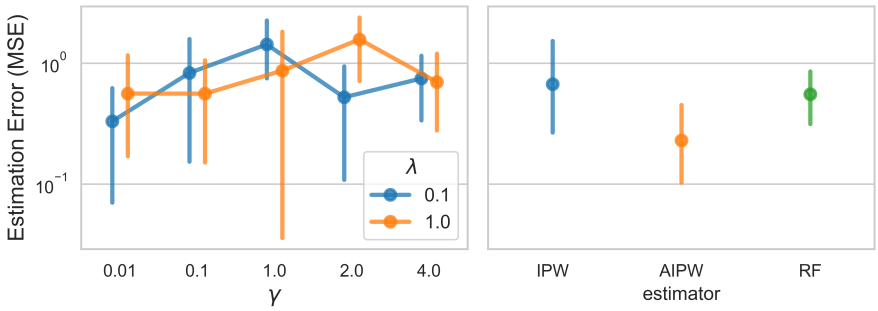

Figure 8: Results on synthetic data generated using High-dimensional Friedman's DGP.

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 8 shows the mean squared error in estimating ATT and compares it with IPW, AIPW and RF. Our results show that sweeping over the range of γ (i.e., regularizing extrapolation) smoothly 'interpolates' between AIPW and IPW estimators (the former allows extrapolation using a linear regression outcome model, while the latter relies only on interpolation).