Augmented balancing weights as undersmoothed regressions*

David Bruns-Smith, UC Berkeley
Oliver Dukes, Ghent Univ.
Avi Feller, UC Berkeley
Elizabeth L. Ogburn, Johns Hopkins Univ.

February 3, 2023

PRELIMINARY DRAFT

Abstract

The augmented balancing weights framework, also known as automatic debiased machine learning, is a powerful approach to causal inference that has recently seen a flurry of attention in statistics, econometrics, and machine learning. We show in this paper that, when both the outcome and weighting models are linear in some (possibly infinite) basis, the augmented approach is equivalent to a form of undersmoothed regression, shifting regularized coefficients from the outcome model toward unregularized OLS coefficients. We characterize the general regularization path from the outcome model to OLS and then apply the results to the special cases where the weighting model is either ridge or lasso, demonstrating that the implied regularization paths have the same form as ridge or lasso penalties, respectively. When both outcome and weighting models are ridge, the combined estimator is also a form of ridge regression; when both outcome and weighting models are lasso, the combined estimator demonstrates a "double selection" property. We also show that constraining balancing weights to be non-negative is equivalent to a form of sample trimming. Finally, we extend these results to general estimands via the Riesz representer. Together, these results "open the black box" on the growing number of augmented balancing weights estimators and suggest fundamental connections between doubly robust estimators and undersmoothing.

1 Introduction

Combining outcome modeling and weighting, such as in augmented inverse propensity weighting (AIPW) and other doubly robust (DR) estimators, is a core strategy for estimating causal effects in observational causal inference and for addressing covariate shift problems in machine learning. A growing literature finds weights by solving the "balancing weights" optimization problem, which estimates the inverse propensity weights directly, rather than by first estimating the propensity score and then inverting. These estimators are referred to by a wide range of terms, including *augmented balancing weights* [Athey et al., 2018, Hirshberg and Wager, 2021] and *automatic debiased machine learning* (AutoDML) [Chernozhukov et al., 2022c]; see Ben-Michael et al. [2021b] for a review.

In this paper, we consider augmented balancing weights in which the estimators for both the outcome model and the inverse propensity weights are penalized linear models in some possibly infinite basis. We begin with a novel characterization of the augmented estimator as equivalent to applying unregularized ordinary least squares (OLS) coefficients to a *re-weighted* — rather than observed — target covariate profile. We use this characterization to show that, somewhat surprisingly, augmenting a regularized linear outcome model

*We would like to thank David Arbour, Eli Ben-Michael, Alex D'Amour, and Skip Hirshberg for useful discussion and comments. A.F. and D.B-S. were supported in part by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D200010. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education. O.D. was supported by NIH grant 579679 and by the FWO grant 1222522N. E.L.O. was supported by ONR grant N000142112820 and by the Simons Institute for Theoretical Computer Science.

with linear balancing weights is numerically equivalent to a single linear outcome model applied to the observed target covariates. The resulting coefficients are *undersmoothed* relative to the original outcome model, in the sense of shifting regularized coefficients toward unregularized OLS coefficients.[1] Moreover, the hyperparameter for the balancing weights estimator controls the *regularization path* between the original coefficients and OLS, that is, the level of undersmoothing. We exactly characterize this path in the general case and show it also depends on the proportion of covariate shift explained by the balancing weights.

We then specialize these results to estimators that augment any penalized linear outcome regression (the *base learner*) with ridge or lasso weighting models (i.e., ℓ_2 or ℓ_∞ balancing, respectively) and show that the corresponding regularization paths are exactly analogous to standard ridge and lasso penalties. When both outcome and weighting estimators are ridge regression, the augmented estimator is equivalent to a single (adaptive) ridge regression estimator with a different choice of hyperparameter; this equivalence extends to kernel ridge regression. When both outcome and weighting estimators are lasso, we obtain a familiar "double selection" result [Belloni et al., 2014]: the included covariates in the combined estimator are the union of the non-zero coefficients in the individual models.

Finally, we propose several extensions. We first show that the common practice of constraining weights to be non-negative corresponds to trimming the sample used to obtain the OLS coefficients. We then show that the numeric equivalences apply to augmented balancing weights estimation of a broad class of linear functionals, via the Riesz representation theorem [Chernozhukov et al., 2022c]. While in the main text, we consider low-dimensional linear function classes, in the appendix, we demonstrate that our results hold quite generally — they apply to *any* class of functions in a (possibly infinite-dimensional) Hilbert space.

Our novel numeric equivalence results provide important insights to two open questions in causal inference and machine learning. First, and most immediately, these results "open the black box" on the growing number of methods based on augmented balancing weights and automatic debiased machine learning that can sometimes be difficult to understand. We provide an easy-to-interpret representation of the augmented estimator as a convex combination of the coefficients of the base learner outcome regression and OLS coefficients. Our numeric results similarly highlight that estimation choices for augmented balancing weights can lead to potentially unexpected behavior. Most notably, choosing (kernel) ridge regression for both outcome and weighting models collapses to a single (kernel) ridge regression estimator. Thus, we also generalize the numeric results underlying the argument in Robins et al. [2007] that "OLS is doubly robust" to a broader class of penalized regressions.

Second, our results make explicit the connection between the use of machine learning in doubly robust methods ("double machine learning") and undersmoothed outcome regressions. Without augmentation, regularized outcome regression estimators suffer from non-negligible regularization bias [Chernozhukov et al., 2018]. One solution is to carefully undersmooth the outcome regression fit to decrease the bias [Goldstein and Messer, 1992, Newey et al., 1998, van der Laan et al., 2022]. However, undersmoothing is difficult to operationalize: the choice of hyperparameter has often proved challenging in practice and efficient undersmoothed plug-in estimators have only been shown to exist for certain models. For example, sparse methods cannot easily be undersmoothed in high-dimensional settings to yield an efficient plug-in estimator [Chernozhukov et al., 2022a]. An alternative is to instead augment the outcome model with a weighting model, such as in augmented balancing weights or AutoDML. Our result shows that, for a broad class of doubly robust estimators, this augmentation in fact explicitly undersmooths the outcome regression. This is most immediate for kernel ridge regression: we show that augmenting a kernel ridge outcome model with the corresponding kernel ridge weighting model is equivalent to undersmoothed kernel ridge regression alone, which is known to have optimal properties [Hirshberg et al., 2019, Kallus, 2020, Mou et al., 2023]. We discuss this in more detail in Section 6.

In Section 2 we introduce the problem setup, identification assumptions, and common estimation methods; we also review balancing weights and previous results linking balancing weights to outcome regression models.

[1] Our use of the term "undersmooth" here is purely numeric. For a connection to undersmoothing in the statistical sense (e.g., relative to a rate-optimal hyperparameter), see Section 6.

In Section 3 we present our new results, and in Sections 3.2 and 3.3 we cache out the implications for ℓ_2 and ℓ_∞ balancing weights specifically. Some extensions, e.g. to simplex-constrained weights and to general linear functionals, are in Section 4. Section 5 gives a numeric illustration. Section 6 discusses the connection with the semiparameterics literature and offers some other directions for future research.

1.1 Literature review

Balancing weights and AutoDML. With deep roots in survey calibration methods [Deville and Särndal, 1992], a large and growing causal inference literature uses balancing weights estimation in place of traditional inverse propensity score weighting (IPW). Ben-Michael et al. [2021b] provide a recent review; we discuss specific examples at length in Section 2.3 below. This approach typically tries to achieve minimax finite-sample balance of features of the covariate distributions in the different treatment groups. Of particular interest here are augmented balancing weights estimators that combine balancing weights with outcome regression; see, for example, Athey et al. [2018], Hirshberg and Wager [2021], Ben-Michael et al. [2021c].

A parallel literature in econometrics instead focuses on so-called *automatic* estimation of the Riesz representer, of which IPW are a special case, where "automatic" refers to the fact that the Riesz representer is estimated directly rather than as a ratio of distributions or probabilities. The corresponding augmented estimation framework is known as Automatic Debiased Machine Learning, or AutoDML; see, among others, Chernozhukov et al. [2022c], Chernozhukov et al. [2022d], Chernozhukov et al. [2022a], and Chernozhukov et al. [2022b]. This approach has also been applied in a range of settings, including to corrupted data [Agarwal and Singh, 2021] and to addressing noncompliance [Singh et al., 2022]. As we discuss below, the AutoDML approach nearly always employs cross-fitting and is typically motivated by asymptotic properties rather than achieving balance in finite samples.

Numerical equivalences for balancing weights. Many seminal papers highlight connections between weighting approaches, such as balancing weights and IPW, and outcome modeling; see Bruns-Smith and Feller [2022] for discussion. Most relevant are a series of papers that show numerical equivalences between linear regression and (exact) balancing weights, especially Robins et al. [2007], Kline [2011], and Chattopadhyay and Zubizarreta [2021], and between kernel ridge regression and forms of kernel weighting, especially Kallus [2020], Hirshberg et al. [2019]; we return to this in Section 2.4. In the context of panel data, Shen et al. [2022] establish connections between different forms of regression, which is especially relevant for our discussion of high-dimensional features in Appendix B. Finally, Lin and Han [2022] provide an interesting alternative perspective by demonstrating that a large class of outcome regression estimators can be viewed as implicitly estimating the density ratio of the covariate distributions in the two treatment groups.

2 Problem setup and background

2.1 Setup and motivation

We motivate our numerical results by focusing on the problem of estimating the mean of a missing outcome variable in a target population given observed covariates; we generalize from the mean to arbitrary linear functionals in Section 4.2. As our results are purely numeric, however, this setup is entirely for interpretation and motivation.

Let $X \in \mathcal{X}$ and $Y \in \mathbb{R}$ be a vector of covariates and an outcome, respectively. We consider two populations p and q that are distributions over (X, Y). We will refer to p as the "source" population and q as the "target" population. We consider a setting where X and Y are both observed in the source population, but in the target population we only observe X. Our goal is to estimate the expectation of the missing outcomes in the target population. Thus, the target estimand is $\mathbb{E}_q[Y]$, where we use subscripts p and q to denote the population with respect to which we take expectations. Many important examples have this structure.

Example 1 (Distribution shift)**.** *Here the goal is training a predictor under distribution shift [Gretton et al., 2012]. Let Y be the loss of a predictor, let p be the training population, and let q be a test population with a new distribution of covariates. Then the missing mean, $\mathbb{E}_q[Y]$, is the risk of the predictor in the test population.*

Example 2 (Causal inference)**.** *Here the goal is estimating the unobserved potential outcomes in an observational study. Let Y be the potential outcome under control [with appropriate restrictions, such as SUTVA; Rubin, 1980], let p be the population of individuals in the control condition, and let q be the population of individuals in the treatment condition. Then Y is observed for population p but not for population q, and the missing mean, $\mathbb{E}_q[Y]$, is the average potential outcome under control for the individuals who in fact were treated. Letting Y be the potential outcome under treatment, p the population of individuals in the treatment condition, and q the population of individuals in the control condition, $\mathbb{E}_q[Y]$ is the average potential outcome under treatment for the individuals who in fact received control.*

For both examples, the crucial assumption for identification is *conditional ignorability*: the conditional distribution of Y given X is the same in the source and target populations. This is also known as "conditional exchangeability," "selection on observables," or "no unmeasured confounding." For our purposes, we will require the mean, but not distributional, version of this assumption:

Assumption 1 (Conditional mean ignorability)**.** $\mathbb{E}_p[Y|X] = \mathbb{E}_q[Y|X]$.

Since we assume the conditional expectations are the same in the two populations, we occasionally denote the common conditional mean functional without subscripts, $\mathbb{E}[Y|X]$. Under this assumption, we can identify $\mathbb{E}_q[Y]$ with the *regression functional*, also known as the *adjustment formula* or *g-formula*:

$$\mathbb{E}_q\left[\mathbb{E}_p[Y|X]\right] = \mathbb{E}_q\left[\mathbb{E}_q[Y|X]\right] = \mathbb{E}_q[Y]. \tag{1}$$

A complementary approach instead relies on the density ratio between the source and target distribution, $\frac{dq}{dp}(X)$, also known as the Radon-Nikodym derivative, importance sampling weights, or inverse propensity score weights (IPW).[2] As we discuss in Section 4.2 below, this is also a special case of a Riesz representer [Chernozhukov et al., 2022c]. Under an additional *population overlap assumption* that $q(x)$ is absolutely continuous with respect to $p(x)$, we can identify $\mathbb{E}_q[Y]$ via the *weighting functional*, also known as the *IPW functional*:

$$\mathbb{E}_p\left[\frac{dq}{dp}(X)\, Y\right] = \mathbb{E}_p\left[\frac{dq}{dp}(X)\, \mathbb{E}_p[Y|X]\right] = \mathbb{E}_q[\mathbb{E}_p[Y|X]] = \mathbb{E}_q[Y]. \tag{2}$$

Finally, we can combine the regression and weighting functionals to create a third identifying functional, known as the *doubly robust functional* [Robins et al., 1994]:

$$\mathbb{E}_q\left[\mathbb{E}_p[Y|X]\right] + \mathbb{E}_p\left[\frac{dq}{dp}(X)\left\{Y - \mathbb{E}_p[Y|X]\right\}\right]. \tag{3}$$

This functional has the attractive property of being equal to $\mathbb{E}_q[Y]$ even if either one of $\frac{dq}{dp}(X)$ or $\mathbb{E}_p[Y|X]$ is replaced with an arbitrary function of X, hence the term "doubly robust."[3] See Chernozhukov et al. [2018], Kennedy [2022] for recent overviews of the active literature in causal inference and machine learning focused on estimating versions of Equation (3).

[2] Using Bayes Rule, we can equivalently express $\frac{dq}{dp}(X)$ via the *propensity score* $P(1_p|X)$, where 1_p is the indicator that an observation from the size-proportional mixture distribution of p and q is from population p: $\frac{dq}{dp}(X) = \frac{1-P(1_p|X)}{P(1_p|X)}$

[3] This functional is equal to $\mathbb{E}_q[Y]$ if $\mathbb{E}_p[Y|X]$ is replaced with an arbitrary well-behaved functional of X_p, because the first and last terms cancel and we are left with the weighting functional $\mathbb{E}_p[\frac{dq}{dp}(X)Y]$. It is also equal to $\mathbb{E}_q[Y]$ if $\frac{dq}{dp}(X)$ is replaced with an arbitrary well-behaved functional of X_p, because the $\mathbb{E}_p[h(X)(Y - \mathbb{E}_p[Y|X])]$ is equal to 0 for any h and therefore we are left with the regression functional $\mathbb{E}_q\left[\mathbb{E}_p[Y|X]\right]$.

The *augmented* estimators that we analyze in this paper are based on estimating this doubly robust functional. These estimators *augment* an estimator of the regression functional based on an outcome regression (or *base learner*) with appropriately weighted residuals. Alternatively, they *augment* an estimator of the weighting functional with an outcome regression-based estimator of the regression functional (subtracting off the implied estimator of $\frac{dq}{dp}(X)\mathbb{E}_p[Y|X]$).

2.2 Balancing weights: Background and general form

While the density ratio $\frac{dq}{dp}(X)$ often plays a central role in identification, particularly in doubly robust approaches, the quantity is notoriously difficult to estimate with plug-in methods. Traditional IPW estimators that first estimate the propensity score and then invert to obtain the corresponding weights can be unstable when the estimated propensity score in the denominator is small [Kang and Schafer, 2007]. This approach can also fail to control finite sample covariate imbalance in practice [see Ben-Michael et al., 2021b]. Likewise, separately estimating the densities p and q and then taking their ratio is highly unstable [Sugiyama et al., 2012].

The balancing weights framework, also known as automatic estimation of the Riesz representer, provides an alternative estimation approach [Ben-Michael et al., 2021b, Chernozhukov et al., 2022c]. Before turning to practical estimation in Section 2.3 below, we focus on two complementary motivations for this approach: (1) weighting to minimize covariate imbalance, and (2) direct estimation of the density ratio.

2.2.1 Weighting for covariate balance

A central property of the density ratio is that the corresponding weights, $w(X) = \frac{dq}{dp}(X)$, perfectly balance the mean of all functions of X between p and q. That is, $\frac{dq}{dp}(X)$ are the unique weights $w : \mathcal{X} \to \mathbb{R}$ that satisfy the *population balancing property* [Ben-Michael et al., 2021b]:

$$\mathbb{E}_p[w(X)f(X)] = \mathbb{E}_q[f(X)] \quad \text{for all measurable functions } f. \tag{4}$$

Substituting $f(X) = \mathbb{E}_p[Y|X]$ into Equation (4) recovers the weighting functional in Equation (2).

As $\frac{dq}{dp}(X)$ are the *unique* weights that satisfy this property for all measurable functions, we can further characterize $\frac{dq}{dp}(X)$ as the unique solution to the following optimization problem:

$$\min_w \sup_f \Big\{\mathbb{E}_p[w(X)f(X)] - \mathbb{E}_q[f(X)]\Big\}, \tag{5}$$

where the objective is called the "imbalance" between p and q. This optimization problem, however, does not lead to a practical estimator: the objective is unbounded for all $w \neq \frac{dq}{dp}(X)$, and, even when bounded, can give very large weights.

The balancing weights approach makes this problem more tractable by imposing simplifying assumptions about the nature of confounding or covariate shift. In particular, if we are willing to assume that $\mathbb{E}[Y|\cdot]$ lies in a model class $\mathcal{F}$, then it suffices to balance functions in that class, and we can replace the objective (5) with the imbalance over $\mathcal{F}$:

$$\text{Imbalance}_{\mathcal{F}}(w) \coloneqq \sup_{f\in\mathcal{F}} \Big\{\mathbb{E}_p[w(X)f(X)] - \mathbb{E}_q[f(X)]\Big\}.$$

For our purposes, $\mathcal{F}$ must be a convex and symmetric function class. For example, we might consider bounded functions, Lipschitz functions, or functions that are linear in some basis. Typically, the balancing weights approach also introduces a regularization hyperparameter $\delta > 0$ to ensure a unique minimum and to introduce a bias-variance tradeoff. The resulting balancing weights are the unique solution w^* to the following strictly-convex optimization problem:

$$\min_w \Big\{\text{Imbalance}_{\mathcal{F}}(w) + \delta\|w\|_2^2\Big\}. \tag{6}$$

Finite sample versions of this optimization problem can be solved efficiently, as we discuss below. If the true conditional expectation $\mathbb{E}[Y|\cdot] \in \mathcal{F}$, then the imbalance $\left\{\mathbb{E}_p[w^*(X)\mathbb{E}[Y|X]] - \mathbb{E}_q[\mathbb{E}[Y|X]]\right\}$, which captures bias in estimating $\mathbb{E}_q[Y]$, will be small. For a more formal statement, see for example Hirshberg and Wager [2021].

2.2.2 Direct estimation of the density ratio

Chernozhukov et al. [2022c] consider an alternative motivation for balancing weights: finding weights that directly minimize the mean-squared error for $\frac{dq}{dp}(X)$,[4]

$$\min_{f \in \mathcal{F}} \left\{ \mathbb{E}_p \left[\left(f(X) - \frac{dq}{dp}(X) \right)^2 \right] \right\}. \tag{7}$$

The balancing weights w^* are equal to $f^* \in \mathcal{F}$ that achieve the minimum in (7) up to a constant scaling factor that depends on δ. Ben-Michael et al. [2021b], Bruns-Smith and Feller [2022] showed that this is equivalent to the optimization problem in (6).

This gives us a second path to identification. Informally: if $\frac{dq}{dp}(\cdot) \in \mathcal{F}$, then the minimum variance weights that balance $\mathcal{F}$ are also guaranteed to balance *all other* measurable functions, including the conditional expectation $\mathbb{E}[Y|X]$, even if $\mathbb{E}[Y|\cdot] \notin \mathcal{F}$.[5] Thus, if either $\mathbb{E}[Y|\cdot] \in \mathcal{F}$ or $\frac{dq}{dp}(\cdot) \in \mathcal{F}$, then the balancing weights w^* can replace $\frac{dq}{dp}(X)$ in the weighting functional for estimating $\mathbb{E}_q[Y]$.

Remark 1 (Implied propensity score model)**.** *While not central to our discussion, we note that the balancing approach to IPW estimates an implied propensity score model, although the estimated probabilities are not necessarily constrained to be in* $[0, 1]$ *[Robins et al., 2007, Kline, 2011]. If the estimated balancing weights are* $\hat{w}$*, then the implied propensity scores are* $\hat{w}/(1+\hat{w})$*. More generally, Zhao [2019], Ben-Michael et al. [2021b] show that the dual of the balancing weights optimization problem corresponds to modified loss functions for the propensity score. See also Tan [2020].*

2.3 Linear balancing weights

In this paper, we consider finite sample solutions to balancing weights optimization problems in the special case where $\mathbb{E}[Y|X]$ is linear in some basis expansion of X. This is an extremely broad class that encompasses linear and polynomial models in arbitrary dictionaries with dimension possibly larger than the sample size, as well as non-parametric models including reproducing kernel Hilbert spaces [Gretton et al., 2012], the highly-adaptive Lasso [Benkeser and Van Der Laan, 2016], and the neural tangent kernel space of infinite-width neural networks [Jacot et al., 2018]. However, this class excludes models for $\mathbb{E}[Y|X]$ that are fundamentally non-linear in their parameters, like neural networks in general, or linear models passed through a non-linear link function. We sketch an extension to the non-linear case in Appendix D.

In the linear setting, the relevant imbalance is captured entirely by *feature mean imbalance*, as we show below. Let X_p, Y_p be n i.i.d. samples from p, and let X_q be m i.i.d. samples from q. For ease of presentation, in the main text we consider the special case where we have a feature map $\phi : \mathcal{X} \to \mathbb{R}^d$ with $\Phi_p \coloneqq \phi(X_p)$ and similarly for Φ_q. Furthermore, we assume that $d < n$ and that Φ_p has rank d. We emphasize that this is not necessary for our results — one can replace $\mathbb{R}^d$ with an infinite-dimensional Hilbert space $\mathcal{H}$ and relax the rank restriction. See Appendix B for a formal presentation of the high-dimensional setting. For the presentation in the main text we assume with no essential loss of generality that $(\Phi_p^T \Phi_p) = \text{diag}(\sigma_1^2, \sigma_2^2, ..., \sigma_d^2)$,[6] with $\sigma_j^2 > 0$ equal to the sample variance of the j^{th} column of Φ_p.

[4] See Zhao [2019] for balancing weights estimators under alternative loss functions.
[5] For a more formal statement, see Hirshberg and Wager [2021], Chernozhukov et al. [2022c].
[6] Otherwise, multiply the feature matrix by its right singular vectors.

We will write $\hat{\mathbb{E}}_p$ and $\hat{\mathbb{E}}_q$ for sample averages and define $\bar{\Phi}_q := \hat{\mathbb{E}}_q[\Phi_q]$. In what follows we write w for the $1 \times n$ vector $w(\Phi)$, to highlight the fact that we will estimate w directly rather than as an explicit function of X or Φ.

Let $\mathcal{F} = \{f(x) = \theta^T \phi(x) : \|\theta\| \leq r\}$ where $\|\cdot\|$ can be any norm on $\mathbb{R}^d$. Let $\|\cdot\|_*$ be the *dual norm* of $\|\cdot\|$; that is, $\|v\|_* := \sup_{\|u\| \leq 1} u^T v$. Many common vector norms have familiar, closed-form, dual norms, e.g. the dual norm of the ℓ_2-norm is the ℓ_2-norm; and the dual norm of the ℓ_1-norm is the ℓ_∞-norm. Then the supremum that defines the sample $\text{Imbalance}_{\mathcal{F}}(w)$ can be written as:

$$\widehat{\text{Imbalance}}_{\mathcal{F}}(w) = \|w\Phi_p - \bar{\Phi}_q\|_*.$$

Now we can write the balancing weights optimization problem in (6) equivalently as either:

$$\text{Penalized form:} \qquad \min_{w \in \mathbb{R}^n} \left\{ \|w\Phi_p - \bar{\Phi}_q\|_*^2 + \delta_1 \|w\|_2^2 \right\}$$

$$\text{Constrained form:} \qquad \min_{w \in \mathbb{R}^n} \|w\|_2^2 \quad \text{such that } \|w\Phi_p - \bar{\Phi}_q\|_* \leq \delta_2.$$

Furthermore, we can write the equivalent problem in (7) as:

$$\text{Automatic form:} \qquad \min_{\theta \in \mathbb{R}^d} \left\{ \theta^T (\Phi_p^T \Phi_p) \theta - 2\theta^T \bar{\Phi}_q + \delta_3 \|\theta\| \right\},$$

where we use the terminology "automatic" from Chernozhukov et al. [2022c]. For any parameter $\delta_2 > 0$ and corresponding constrained problem solution $\hat{w}$, there exists a parameter $\delta_3 > 0$ such that $\hat{w} = \delta_3 \Phi_p \hat{\theta}$, where $\hat{\theta}$ is the solution to the automatic form [Ben-Michael et al., 2021b].[7] As a result, for any norm $\|\cdot\|$, the penalized and constrained forms will *always* produce weights that are linear in Φ_p [Ben-Michael et al., 2021b]. Without loss of generality, we therefore use δ to denote the regularization parameter for the balancing weights problem, regardless of the specific form. Some popular choices of linear balancing weights are described below.

Example 1 (Exact balancing weights)**.** *The most common balancing weights estimation problem finds the minimum weights that exactly balance each element of* Φ.[8] *In the constrained form, exact balancing solves*

$$\min_{w \in \mathbb{R}^n} \|w\|_2^2 \quad \text{such that } w\phi_{pj} = \bar{\phi}_{qj} \quad \text{for all } j \tag{8}$$

Example 2 (ℓ_2 balancing)**.** *The ℓ_2 balancing weights problem is usually expressed via its penalized form:*

$$\min_{w \in \mathbb{R}^n} \left\{ \|w\Phi_p - \bar{\Phi}_q\|_2^2 + \delta \|w\|_2^2 \right\} \tag{9}$$

Example 3 (ℓ_∞ balancing)**.** *The constrained form of the ℓ_∞ balancing weights problem is*

$$\min_{w \in \mathbb{R}^n} \|w\|_2^2 \quad \text{such that } \|w\Phi_p - \bar{\Phi}_q\|_\infty \leq \delta \tag{10}$$

The "automatic" form of ℓ_∞ balancing is also known as the Minimum Distance Lasso *estimator for the Riesz representer [Chernozhukov et al., 2022c]. Popular variants of this problem constrain the weights to be non-negative [Zubizarreta, 2015, Athey et al., 2018].*

[7] Note also that the choice of δ implies a choice of hyperparameter r that defines the class $\mathcal{F}$ [Kallus, 2020].

[8] Many popular variants of this approach penalize alternative dispersion measures, such as the entropy of the weights [e.g., Hainmueller, 2012].

Example 4 (Kernel balancing)**.** *As a brief preview of the balancing problem in the infinite-dimensional setting, we provide an example where $\mathcal{F} = \mathcal{H}$ is a reproducing kernel Hilbert space with norm $\|\cdot\|_{\mathcal{H}}$ and kernel $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Then for any $x \in \mathcal{X}$, the representer $\mathcal{K}(x, \cdot) \in \mathcal{H}$. Define $\bar{\Phi}_q \coloneqq \frac{1}{m}\sum_{j=1}^{m} \mathcal{K}(x_{q,j}, \cdot)$ and for $w \in \mathbb{R}^n$ define the reweighted $\hat{\Phi}_q \coloneqq \frac{1}{n}\sum_{i=1}^{n} w_i \mathcal{K}(x_{p,i}, \cdot)$. The penalized balancing weights problem for $\mathcal{F} = \mathcal{H}$ is:*

$$\min_{w \in \mathbb{R}^n} \left\{ \|\hat{\Phi}_q - \bar{\Phi}_q\|_{\mathcal{H}}^2 + \delta \|w\|_2^2 \right\}. \tag{11}$$

The Hilbert space norm can be computed efficiently in finite samples using the kernel matrix. See Appendix B for details, as well as Hazlett [2020] and Kallus [2020].

Remark 2 (Intercept)**.** *An important constraint in practice is to normalize the weights, $\frac{1}{n}\sum_{i=1}^{n} w_i = 1$. This corresponds to replacing Φ_p and Φ_q with their centered forms, $\Phi_p - \bar{\Phi}_p$ and $\Phi_q - \bar{\Phi}_p$ in the dual form of the balancing weights problem. This is also equivalent to adding a column of 1s to Φ_p. Appropriately accounting for this normalization, however, unnecessarily complicates the notation. Therefore, without loss of generality, we will assume that the covariates are centered throughout, that is, $\bar{\Phi}_p = 0$.*

2.4 When outcome modeling and balancing weights are known to be equivalent

The results above motivate balancing weights from the perspective of estimating the weighting functional in Equation (2). Interestingly, for a large class of outcome models, the balancing weights problem is numerically equivalent to directly estimating the conditional expectation $\mathbb{E}_p[Y|\Phi]$ and applying the estimated coefficients to Φ_q. In particular, ℓ_2 balancing weights are numerically equivalent to ridge regression trained in the source distribution and applied to the target features. We begin with the special case of unregularized linear regression and then present the more general setting.

Linear regression. Ordinary least squares regression is equivalent to a weighting estimator that exactly balances the feature means. See Fuller [2002] for discussion in the survey sampling literature; see Robins et al. [2007], Abadie et al. [2010], Kline [2011], and Chattopadhyay et al. [2020] for relevant discussions in the causal inference literature.

In particular, let $\hat{w}_{\text{exact}}$ be the solution to the the exact balancing weights problem in Example 1 above. Let $\hat{\beta}_{\text{ols}} = (\Phi_p^T \Phi_p)^{-1} \Phi_p^T Y_p$ be the OLS coefficients from the regression of Y_p on Φ_p.[9] We then have the following numerical equivalence:

$$\begin{aligned} \hat{\mathbb{E}}_q[\Phi_q \hat{\beta}_{\text{ols}}] &= \hat{\mathbb{E}}_p[\hat{w}_{\text{exact}} \circ Y_p] \\ \hat{\mathbb{E}}_q[\Phi_q \underbrace{(\Phi_p^T \Phi_p)^{-1} \Phi_p^T Y_p}_{\hat{\beta}_{\text{ols}}}] &= \hat{\mathbb{E}}_p[\underbrace{\bar{\Phi}_q (\Phi_p^T \Phi_p)^{-1} \Phi_p^T}_{\hat{w}_{\text{exact}}} \circ Y_p], \end{aligned} \tag{12}$$

where the weights have the closed form $\hat{w}_{\text{exact}} = \bar{\Phi}_q (\Phi_p^T \Phi_p)^{-1} \Phi_p^T$.

Ridge regression. This equivalence immediately extends to ridge regression [Hirshberg et al., 2019, Kallus, 2020].[10] Let $\hat{w}_{\ell_2}^{\delta}$ be the minimizer of the ℓ_2 balancing weights problem in Example 2 above, with hyperparameter δ. Let

$$\hat{\beta}_{\text{ridge}}^{\delta} \coloneqq \operatorname*{argmin}_{\beta \in \mathbb{R}^d} \left\{ \hat{\mathbb{E}}_p[(Y_p - \Phi_p \beta)^2] + \delta \|\beta\|_2^2 \right\} \tag{13}$$

[9] If the sample Gramian is not invertible, then the same results hold using the pseudoinverse, corresponding to the minimum norm OLS coefficients.

[10] See Harshaw et al. [2019] for an interesting connection of this equivalence to experimental design. See Ben-Michael et al. [2021c] and Shen et al. [2022] for related applications in the panel data setting.

be the ridge regression coefficients from least squares regression of Y_p on Φ_p. We then have the following numerical equivalence:

$$\begin{aligned} \hat{\mathbb{E}}_q[\Phi_q \hat{\beta}^{\delta}_{\text{ridge}}] &= \hat{\mathbb{E}}_p[\hat{w}^{\delta}_{\ell_2} \circ Y_p] \\ \hat{\mathbb{E}}_q[\Phi_q \underbrace{(\Phi_p^T \Phi_p + \delta I)^{-1} \Phi_p^T Y_p}_{\hat{\beta}^{\delta}_{\text{ridge}}}] &= \hat{\mathbb{E}}_p[\underbrace{\overline{\Phi}_q (\Phi_p^T \Phi_p + \delta I)^{-1} \Phi_p^T}_{\hat{w}^{\delta}_{\ell_2}} \circ Y_p], \end{aligned} \tag{14}$$

where the weights have the closed form $\hat{w}^{\delta}_{\ell_2} = \overline{\Phi}_q (\Phi_p^T \Phi_p + \delta I)^{-1} \Phi_p^T$. Thus, the estimate from ridge regression is identical to the estimate using the ℓ_2 balancing weights. We leverage this equivalence in Section 3 below.

Kernel ridge regression. In general, the same equivalence holds in the non-parametric setting where ϕ is the feature map induced by an RKHS. In particular, let $\mathcal{F} = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq r\}$, where $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS) with kernel $\mathcal{K}$, $\|\cdot\|_{\mathcal{H}}$ denotes the norm of the RKHS, and $r > 0$. Then the equivalence above holds for $\phi(x) := \mathcal{K}(\cdot, x)$. Although ϕ is typically infinite-dimensional, the Riesz Representer Theorem shows that the least square regression and, equivalently, the balancing optimization problem have closed-form solutions. The least squares regression approach is *kernel ridge regression* and the weighting estimator is *kernel balancing weights* [see Hazlett, 2020, Kim et al., 2022]. Hirshberg et al. [2019] leverage this equivalence to analyze the asymptotic bias of kernel balancing weights. For further discussion of this equivalence see Gretton et al. [2012], Kallus [2020].

We comment briefly on the conditions necessary for guaranteeing such equivalences in general in Appendix E.

3 New insights into the relationship between balancing weights and outcome models

We now present our main results on numerical equivalences for linear balancing weights, both for weighting alone and for augmented estimators with a linear outcome regression. We start with results for general linear balancing weights and then specialize to ℓ_2 and ℓ_∞ balancing weights. In Section 4, we extend these results to account for high-dimensional covariates and constrained weights.

3.1 General linear balancing weights

We first demonstrate that any linear balancing weights estimator is equivalent to applying OLS to the re-weighted features. We then demonstrate that augmenting any linear balancing weights estimator with a linear outcome regression estimator results in an estimator that is equivalent to a single (adaptive) linear regression fit, the coefficients of which are a weighted combination of estimated OLS coefficients and the coefficients of the linear outcome model specified in the augmented estimator.

3.1.1 Weighting alone

Our first result is that estimating $\mathbb{E}_q[Y]$ with *any* linear balancing weights is equivalent to estimating an OLS model for $\mathbb{E}[Y|\Phi]$ in the source population and then applying those coefficients to the re-weighted target population features.

Proposition 3.1. *Let $\hat{w}^{\delta} := \hat{\theta}^{\delta} \Phi_p^T$, $\hat{\theta}^{\delta} \in \mathbb{R}^d$, be any linear balancing weights, with corresponding weighted features $\hat{\Phi}^{\delta}_q := \hat{w}^{\delta} \Phi_p$. Let $\hat{\beta}_{ols} = (\Phi_p^T \Phi_p)^{\dagger} \Phi_p^T Y_p$ be the OLS coefficients of the regression of Y_p on Φ_p. Then:*

$$\begin{aligned} \hat{\mathbb{E}}_p\left[\hat{w}^{\delta} \circ Y_p\right] &= \hat{\Phi}^{\delta}_q \hat{\beta}_{ols} \\ &= \left(\bar{\Phi}_p + \widehat{\Delta}^{\delta}\right) \hat{\beta}_{ols}, \end{aligned}$$

where $\widehat{\Delta}^\delta = \hat{\Phi}_q^\delta - \bar{\Phi}_p$[11] *is the mean feature shift implied by the balancing weights and where superscript* δ *indicates possible dependence on a hyperparameter.*

Here we have written the OLS coefficients using the pseudo-inverse †. For clarity in the main text, we focus on the full rank setting, where $(\Phi_p^T\Phi_p)^\dagger = (\Phi_p^T\Phi_p)^{-1}$; we provide a proof for the general setting in Appendix B.3. When the weights achieve exact balance, so that $\hat{\Phi}_q^\delta = \bar{\Phi}_q$, Proposition 3.1 recovers the expression in Equation (12). In Section 4, we extend Proposition 3.1 to balancing weights that incorporate a non-negativity constraint and to general linear functional estimands.

We emphasize that this a new and quite general result. As we discussed in Section 2.4, it has been shown previously that for exact balancing weights, $\hat{\mathbb{E}}_p[\hat{w}_{\text{exact}} Y_p] = \bar{\Phi}_q\hat{\beta}_{\text{ols}}$. However, Proposition 3.1 holds for any weights of the form $w = \theta\Phi_p^T$ with arbitrary $\theta \in \mathbb{R}^d$.

The key idea behind Proposition 3.1 begins with the simple unregularized regression prediction for $\mathbb{E}_q[Y]$, $\bar{\Phi}_q\hat{\beta}_{\text{ols}}$. Regularized regression models navigate a bias-variance trade-off by regularizing estimated coefficients $\hat{\beta}_{\text{reg}}$ relative to $\hat{\beta}_{\text{ols}}$, leading to $\bar{\Phi}_q\hat{\beta}_{\text{reg}}$. The balancing weights approach keeps $\hat{\beta}_{\text{ols}}$ fixed and instead regularizes the target feature distribution by penalizing the implied covariate shift, $\widehat{\Delta}^\delta = \hat{\Phi}_q^\delta - \bar{\Phi}_p$. In Sections 3.2 and 3.3, we make this regularization explicit for ℓ_2 and ℓ_∞ balancing, respectively.

3.1.2 Augmented balancing weights

A natural question is whether it is desirable to regularize *both* the coefficients and the covariate shift? It turns out, as our next result shows, that augmented balancing weights estimators achieve their desirable theoretical properties by doing just this.

Let $\hat{\beta}_{\text{reg}}^\lambda$ be the coefficients of any regularized linear model for the relationship between Y_p and Φ_p where the superscript λ indicates dependence on a hyperparameter — for example estimated by regularized least squares. Consider augmenting $\hat{\mathbb{E}}_p\left[\hat{w}^\delta \circ Y_p\right]$ with $\hat{\beta}_{\text{reg}}^\lambda$ using the doubly robust functional in Equation (3). The augmented estimator is:[12]

$$\hat{\mathbb{E}}_q[\Phi_q\hat{\beta}_{\text{reg}}^\lambda] + \hat{\mathbb{E}}_p[\hat{w}^\delta \circ (Y_p - \Phi_p\hat{\beta}_{\text{reg}}^\lambda)] = \hat{\mathbb{E}}_p[\hat{w}^\delta \circ Y_p] + \hat{\mathbb{E}}_q\left[\left(\Phi_q - \hat{\Phi}_q^\delta\right)\hat{\beta}_{\text{reg}}^\lambda\right]. \tag{15}$$

Many recently proposed estimators have this form; see e.g. Athey et al. [2018], Hirshberg and Wager [2021], Ben-Michael et al. [2021b], Chernozhukov et al. [2022c].

We apply Proposition 3.1 to the first term of the right-hand side of (15) to yield the following result. Note that as this result is purely numerical, it applies to arbitrary vectors $\hat{\beta}_{\text{reg}}^\lambda \in \mathbb{R}^d$, but substantively we think of $\hat{\beta}_{\text{reg}}^\lambda$ as the estimated coefficients from an outcome model.

Proposition 3.2. *For any* $\hat{\beta}_{reg}^\lambda \in \mathbb{R}^d$*, and any linear balancing weights estimator with estimated coefficients* $\hat{\theta}^\delta \in \mathbb{R}^d$*, and with* $\hat{w}^\delta := \hat{\theta}^\delta\Phi_p^T$ *and* $\hat{\Phi}_q^\delta := \hat{w}^\delta\Phi_p$*, the resulting augmented estimator*

$$\begin{aligned}
&\hat{\mathbb{E}}_p[\hat{w}^\delta \circ Y_p] + \hat{\mathbb{E}}_q\left[\left(\Phi_q - \hat{\Phi}_q^\delta\right)\hat{\beta}_{reg}^\lambda\right] \\
&= \hat{\mathbb{E}}_q\left[\hat{\Phi}_q^\delta\hat{\beta}_{ols} + \left(\Phi_q - \hat{\Phi}_q^\delta\right)\hat{\beta}_{reg}^\lambda\right] \\
&= \hat{\mathbb{E}}_q[\Phi_q\hat{\beta}_{aug}],
\end{aligned}$$

[11] We have assumed without loss of generality that $\bar{\Phi}_p = 0$, but we sometimes choose to use $\hat{\Delta}$ notation in order to demonstrate the role of mean feature shift in various expressions.

[12] If the weighting model and outcome model have different bases, our result applies to a shared basis by either combining the dictionaries as in Chernozhukov et al. [2022c] or by applying an appropriate projection as in Hirshberg and Wager [2021].

where the jth element of $\hat{\beta}_{aug}$ is:

$$\hat{\beta}_{aug,j} \coloneqq \left(1 - a_j^\delta\right) \hat{\beta}^\lambda_{reg,j} + a_j^\delta \hat{\beta}_{ols,j}$$

$$a_j^\delta \coloneqq \frac{\widehat{\Delta}_j^\delta}{\Delta_j},$$

where $\Delta_j = \overline{\Phi}_{q,j} - \overline{\Phi}_{p,j}$ is the observed mean feature shift for feature j; and $\widehat{\Delta}_j^\delta = \hat{\Phi}^\delta_{q,j} - \overline{\Phi}_{p,j}$ is the feature shift for feature j implied by the balancing weights model.

This is our central numerical result for augmented balancing weights: when both the outcome and weighting models are linear, the augmented estimator is equivalent to a linear model applied to the target features Φ_q, with coefficients that are element-wise affine combinations of the base learner coefficients, $\hat{\beta}^\lambda_{\text{reg}}$, and the coefficients $\hat{\beta}_{\text{ols}}$ from an OLS regression of Y_p on Φ_p.[13] As $\hat{\beta}^\lambda_{\text{reg}}$ will typically be an outcome model that has been regularized to be smoother than OLS, we say that the augmented coefficients *undersmooth* relative to the regularized base learner.[14]

The regularization parameter for the balancing weights problem, δ, parameterizes the regularization path between $\hat{\beta}^\lambda_{\text{reg}}$ and $\hat{\beta}_{\text{ols}}$, where the extent of un-regularizing varies across features. In general, as $\delta \to 0$ the balancing weights problem prioritizes minimizing balance over controlling variance, and $\widehat{\Delta}_j \to \Delta_j$ for all j.[15] Conversely, as $\delta \to \infty$, the balancing weights problem prioritizes controlling variance, leading to uniform weights and $\widehat{\Delta}_j \to 0$. To connect this to the level of undersmoothing for each feature, if $\widehat{\Delta}_j^\delta / \Delta_j \to 1$, then we can fully "de-bias" the base learner coefficient $\hat{\beta}^\lambda_{\text{reg}}$, in the sense of shifting the coefficient entirely to OLS, $\hat{\beta}_{\text{aug}} \to \hat{\beta}_{\text{ols}}$. Conversely, if $\widehat{\Delta}_j^\delta / \Delta_j \approx 0$, then the weighting model does very little and $\hat{\beta}_{\text{aug}} \approx \hat{\beta}^\lambda_{\text{reg}}$.

To build intuition for this numerical result, consider the following simple example. Imagine that a researcher fits an elastic net outcome regression of Y_p against Φ_p, with the hyperparameter chosen via cross-validation to minimize the prediction MSE. Concerned about regularization bias, the researcher augments the elastic net with balancing weights with parameter δ. Proposition 3.2 shows that the augmented estimator that combines these weights with the regularized elastic net outcome model is equivalent to a new outcome model $\hat{\beta}_{\text{aug}}$ applied to Φ_q that *is always closer to OLS* than the original elastic net coefficients. The smaller the parameter δ, the closer the augmented coefficients are brought toward OLS. As the OLS coefficients (mechanically) have no regularization bias, this is the numerical sense in which the augmentation procedure de-biases the base learner; we return to this idea in Section 6.

For a fixed value of δ, the scaling of a_j^δ with Δ_j depends on the choice of balancing norm. For augmented ℓ_2 balancing weights, a_j^δ is independent of the level of covariate shift. For augmented ℓ_∞ balancing weights, as Δ_j increases, a_j^δ also increases; for this setting, when covariate shift increases, the augmented coefficients are closer to the OLS coefficients. We describe these in detail in the following sections.

Remark 3 (Sample splitting)**.** *Sample splitting is a common technique in the AutoDML literature especially, where we only apply the outcome and weighting models to data points not used for estimation. Since Proposition 3.2 holds for arbitrary vectors $\hat{\beta}^\lambda_{reg}$ and $\hat{\theta}^\delta$, the result goes through immediately. See Appendix A for an extended discussion. Interestingly, we show that augmentation partially undoes sample splitting by shrinking the out-of-sample outcome model fit toward the in-sample OLS coefficients.*

[13] Typically, and for all the examples in this paper, $a_j^\delta \in [0, 1]$ and so the coefficients are *convex* combinations. However, in general there do exist balancing weights problems for which the end points are 0 and 1 but some intermediate δ may be outside of that interval, resulting in an affine combination.

[14] However, the results hold for arbitrary vector $\hat{\beta}^\lambda_{\text{reg}}$. So in this sense by "undersmoothing" we really mean "closer to unregularized OLS." See Section 6 for further connections to undersmoothing in a more narrow statistical sense, e.g. relative to a rate-optimal parameter.

[15] Recall that we assume that we have centered covariates, such that $\overline{\Phi}_{p,j} = 0$ for all j. Thus, $\Delta_j = \overline{\Phi}_{q,j}$ and $\widehat{\Delta}_j^\delta = \hat{\Phi}^\delta_{q,j}$. So $\widehat{\Delta}_j^\delta \to \Delta_j$ is equivalent to $\hat{\Phi}_q^\delta \to \overline{\Phi}_{q,j}$.

Remark 4 (Infinite dimensional setting)**.** *While we emphasize the linear, low-dimensional setting where $\Phi_p^T \Phi_p$ is invertible, Proposition 3.2 holds far more broadly. The result remains true when the function class $\mathcal{F}$ is a subset of* any *Hilbert space. This includes the high dimensional setting where $d > n$ and the infinite dimensional setting. See Appendix B for a formal statement.*

3.1.3 OLS and exact balancing weights

Proposition 3.2 is a general equivalence that allows for arbitrary (linear) outcome and weighting models. In our results below, we use this characterization to derive the regularization paths for ℓ_2 or ℓ_∞ balancing weights and when $\hat{\beta}^{\lambda}_{\text{reg}}$ are estimated via ridge- and lasso-penalized linear regression. Before turning to these, we briefly review two edge cases. In Appendix A, we show that these equivalences only hold approximately in the setting with sample splitting.

OLS outcome model. Consider the special case of fitting an *unregularized* outcome regression model, i.e. $\hat{\beta} = \hat{\beta}_{\text{ols}}$. Then Proposition 3.2 reproduces the result, originally due to Robins et al. [2007], that "OLS is doubly robust" [see also Kline, 2011]. This is because $\hat{\beta}_{\text{aug}} = \hat{\beta}_{\text{ols}}$ for arbitrary linear weights $\theta \in \mathbb{R}^d$. Thus, OLS augmented by linear balancing weights collapses to OLS alone; equivalently, we can view OLS alone as an augmented estimator that combines a linear regression outcome model with linear balancing weights.

Exact balancing weights. A similar result holds for *unregularized* balancing weights, i.e., exact balancing weights. Let $\hat{w}_{\text{exact}}$ be the solution to a balancing weights problem in Section 2.3 with hyperparameter $\delta = 0$, and let $\beta \in \mathbb{R}^d$ be arbitrary coefficients. Then from the balance condition, $\hat{\Phi}_q = \overline{\Phi}_q$, $a_j^0 = 1$ for all j, and we have that $\hat{\beta}_{\text{aug}} = \hat{\beta}_{\text{ols}}$. Thus, the augmented exact balancing weights estimator also collapses to the OLS regression estimator. Equivalently, by Equation (12), the augmented exact balancing weights estimator collapses to the *unaugmented* exact balancing weights estimator. Zhao and Percival [2017] use a very similar result to argue that entropy balancing, a form of exact balancing weights, is doubly robust.

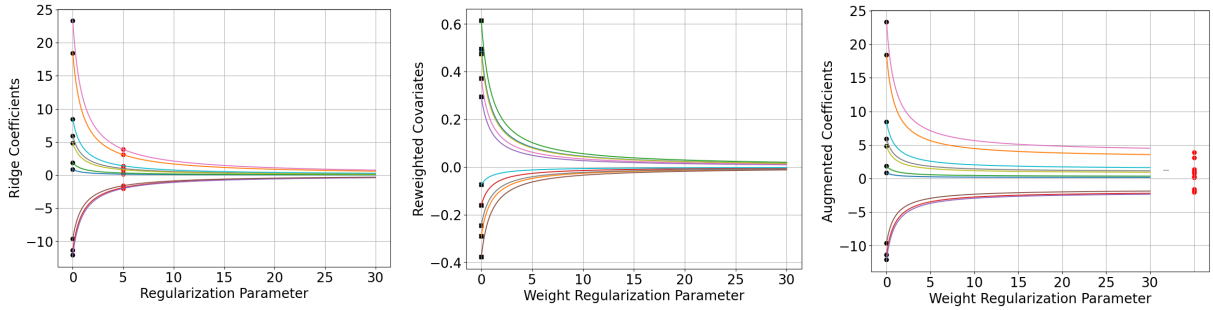
3.2 Augmented ℓ_2 Balancing Weights

We now specialize the general results above to ℓ_2 balancing weights, which are commonly used in the context of kernel balancing [Gretton et al., 2012, Hirshberg et al., 2019, Kallus, 2020, Ben-Michael et al., 2021a] and for panel data methods [Abadie et al., 2010, Ben-Michael et al., 2021c]. In this case, the regularization path between the coefficients of an arbitrary outcome model used to augment the ℓ_2 balancing weights and the coefficients of an unpenalized OLS regression follows typical ridge regression shrinkage, with a smooth decay. When the outcome model used to augment ℓ_2 balancing weights is itself a ridge regression, $\hat{\beta}_{\text{aug}}$ are themselves (adaptive) ridge coefficients, albeit undersmoothed relative to the base learner; we call this "double ridge" estimation. These results extend immediately to the RKHS setting of "double kernel ridge" estimation, combining kernel balancing weights and kernel ridge regression.

Importantly, the double ridge formulation clearly highlights how augmenting with balancing weights undersmooths the original ridge regression model. Conversely, we can always write a single, undersmoothed ridge regression outcome model as a "double ridge" estimator for a range of possible hyperparameters. This gives additional interpretation to a set of recent theoretical results on undersmoothed kernel ridge regression [Hirshberg et al., 2019, Mou et al., 2023].

3.2.1 General linear outcome model

As we show in Section 2.3 above, ℓ_2 balancing weights, including kernel balancing weights, have a closed form that is always linear in $\overline{\Phi}_q$. Our next result applies this equivalence to Proposition 3.2 to derive the regularization path that results from augmenting an arbitrary linear outcome model with ℓ_2 balancing weights. Although this is an immediate consequence of Proposition 3.2, the resulting form of the augmented estimator has unique structure that warrants a new result.



(a) Outcome model (b) Weighting model (c) Augmented model

Figure 1: Regularization paths for "double ridge" augmented ℓ_2 balancing weights. Panel (a) shows the coefficients $\hat{\beta}^{\lambda}_{\text{reg}}$ of a ridge regression of Y_p on Φ_p with hyperparameter λ. The black dots on the left are the OLS coefficients, with $\lambda = 0$. The red dots at $\lambda = 5$ illustrate the coefficients at a plausible hyperparameter value, $\hat{\beta}^{5}_{\text{reg}}$. Panel (b) shows re-weighted covariates, $\hat{\Phi}^{\delta}_{q}$, for the ℓ_2 balancing weights problem with hyperparameter δ; the black dots show exact balance, which corresponds to OLS. As δ increases, the weights converge to uniform weights and $\hat{\Phi}^{\delta}_{q}$ converges to $\overline{\Phi}_p$, which we have centered at zero. Panel (c) shows the augmented coefficients, $\hat{\beta}_{\text{aug}}$ as a function of the weight regularization parameter δ. The black dots on the left are the OLS coefficients. As $\delta \to \infty$, the coefficients converge to $\hat{\beta}^{5}_{\text{reg}}$. All three regularization paths have essentially identical qualitative behavior.

Proposition 3.3. *Let $\hat{w}^{\delta}_{\ell_2}$ be linear balancing weights with regularization parameter δ and $\mathcal{F} = \{f(x) = \theta^T \phi(x) : \|\theta\|_2 \leq 1\}$. Then $\hat{w}^{\delta}_{\ell_2} = \Phi_p(\Phi_p^T \Phi_p + \delta I)^{-1} \overline{\Phi}_q$. Therefore, the augmented ℓ_2 balancing weights estimator with outcome model $\hat{\beta}^{\lambda}_{reg} \in \mathbb{R}^d$ has the form*

$$\hat{\mathbb{E}}_q[\Phi_q \hat{\beta}^{\lambda}_{reg}] + \hat{\mathbb{E}}_p[\hat{w}^{\delta}_{\ell_2}(Y_p - \Phi_p \hat{\beta}^{\lambda}_{reg})] = \hat{\mathbb{E}}_q[\Phi_q \hat{\beta}_{\ell_2}],$$

where the jth coefficient of $\hat{\beta}_{\ell_2}$ is given by

$$\begin{aligned} \hat{\beta}_{\ell_2,j} &:= \left(1 - a^{\delta}_j\right) \hat{\beta}^{\lambda}_{reg,j} + a^{\delta}_j \hat{\beta}_{ols,j} \\ a^{\delta}_j &:= \frac{\sigma_j^2}{\sigma_j^2 + \delta}. \end{aligned} \tag{16}$$

So in this case, the augmented estimator simplifies to a linear model with coefficients that are a convex combination of the original model $\hat{\beta}^{\lambda}_{\text{reg}}$ and the OLS coefficients. The right panel of Figure 1 shows $\hat{\beta}_{\text{aug}}$ (on the y-axis) for ten covariates, with δ increasing from 0 (on the x-axis). The dots on the left pick out $\hat{\beta}_{\text{ols}}$; when $\delta = 0$, then $a^0_j = 1$ and $\hat{\beta}_{\ell_2} = \hat{\beta}_{\text{ols}}$. The limit on the right shows $\hat{\beta}^{\lambda}_{\text{reg}}$. The smooth regularization path is characteristic of ridge regression shrinkage.

To see the analytic connection to the regularization path of standard ridge regression, recall that a ridge regression with penalty δ has coefficients

$$\hat{\beta}^{\delta}_{\text{ridge},j} = \left(\frac{\sigma_j^2}{\sigma_j^2 + \delta}\right) \hat{\beta}_{\text{ols},j} = a^{\delta}_j \hat{\beta}_{\text{ols},j}. \tag{17}$$

This is identical to the expression in (16) but with $\hat{\beta}^{\lambda}_{\text{reg}}$ set to 0. Ridge regression shrinks $\hat{\beta}_{\text{ols}}$ towards 0 with regularization path a^{δ}_j, while double ridge shrinks $\hat{\beta}_{\text{ols}}$ towards $\hat{\beta}^{\lambda}_{\text{reg}}$ with the same regularization path.

Surprisingly — and in contrast to the general result in Proposition 3.2 — the regularization path does not depend on the degree of covariate shift. In particular, Proposition 3.3 shows that ℓ_2 balancing weights are

always linear in $\overline{\Phi}_q$, and as a result, the corresponding regularization path a_j^δ does not depend on the target covariates Φ_q; it depends only on δ and the distribution of the source covariates Φ_p (through the covariate sample variances σ_j^2). For balancing weights in any other norm, the solution will *not* generally be linear in $\overline{\Phi}_q$ and therefore, the augmented estimator shrinks the OLS coefficients toward the base learner coefficients in a way that depends on $\overline{\Phi}_q$. Honestly, we find this quite unintuitive and have no idea why this is the case.

3.2.2 Ridge regression outcome model

Proposition 3.3 holds for arbitrary linear outcome models; we now state the corresponding result for the double ridge estimator, in which the base learner outcome regression is itself ridge regression. The resulting augmented estimator is an undersmoothed, adaptive ridge regression, collapsing to non-adaptive ridge when the sample covariate variances are uniform (as in the orthonormal setting).[16]

Proposition 3.4. *Let $\hat{\beta}^\lambda_{ridge}$ denote the coefficients of a ridge regression of Y_p on Φ_p with hyperparameter λ, and let $\hat{w}^\delta_{\ell_2}$ denote ℓ_2 balancing weights with hyperparameter δ defined in Section 2.3. Define the vector η with jth component:*

$$\eta_j \coloneqq \frac{\delta\lambda}{\sigma_j^2 + \lambda + \delta}.$$

Then:

$$\hat{\mathbb{E}}_q[\Phi_q \hat{\beta}^\lambda_{ridge}] + \hat{\mathbb{E}}_p[\hat{w}^\delta_{\ell_2}(Y_p - \Phi_p \hat{\beta}^\lambda_{ridge})] = \hat{\mathbb{E}}_q[\Phi_q \hat{\beta}_{\ell_2}]$$

where

$$\hat{\beta}_{\ell_2,j} \coloneqq \left(\frac{\sigma_j^2}{\sigma_j^2 + \eta_j}\right) \hat{\beta}_{ols,j}.$$

When $\sigma_j^2 = \sigma^2$ for all j, such as for the orthonormal design $\Phi_p^T \Phi_p = \sigma^2 I$, then $\hat{\beta}_{\ell_2} = \hat{\beta}^\eta_{ridge}$ with hyperparameter:

$$\eta = \frac{\delta\lambda}{\sigma^2 + \delta + \lambda}.$$

Thus, the double ridge coefficients are nearly identical to single ridge regression coefficients in (17), and are equivalent to single ridge regression coefficients for an orthonormal design.

This result highlights that augmenting with balancing weights is equivalent to undersmoothing the original outcome model fit. In particular, we can re-write the shrinkage in Proposition 3.4 as:

$$\frac{\sigma_j^2}{\sigma_j^2 + \eta_j} = \underbrace{\left(\frac{\sigma_j^2}{\sigma_j^2 + \lambda}\right)}_{\text{outcome model}} \underbrace{\left(\frac{\sigma_j^2 + \lambda + \delta}{\sigma_j^2 + \delta}\right)}_{\text{augmentation}},$$

where the first term is the shrinkage from ridge regression outcome modeling alone, and the second term is due to augmenting with ℓ_2 balancing weights. Importantly, the second term is in $[1, \frac{\sigma_j^2+\lambda}{\sigma_j^2}]$ and therefore partially reverses the shrinkage of the original estimate.

Note that Proposition 3.4 recovers the "OLS is doubly robust" result from Robins et al. [2007], discussed in Section 3: if either $\delta = 0$ or $\lambda = 0$, then $\eta_j = 0$ and "double ridge" collapses to a single linear regression. In other words, if the outcome model is fit using OLS or if the weights achieve exact balance, then the combined estimator is itself OLS.

[16] Following Grandvalet [1998] and the subsequent literature, we use the term *adaptive* to refer to ridge regression models that have different shrinkage penalties for different covariates.

Remark 5 (Connection to boosting)**.** *In this setting, the augmented estimator is equivalent to applying one stage of "boosting" with ridge regression to a ridge regression base learner. See Bühlmann and Yu [2003], Park et al. [2009]. A similar connection is made in Newey et al. [2004] in the context of twicing kernels.*

3.3 Augmented ℓ_∞ balancing weights

In this section, we study ℓ_∞ balancing weights estimators, which are widely used in the balancing weights literature [Zubizarreta, 2015, Athey et al., 2018] and in the AutoDML literature [Chernozhukov et al., 2022c], where the approach is referred to as the Minimum Distance Lasso estimator of the Riesz representer. We show that ℓ_∞ balancing is equivalent to applying a soft-thresholding operator to the feature shift from $\overline{\Phi}_p$ to $\overline{\Phi}_q$. Similarly, the regularization path for augmented ℓ_∞ balancing weights parallels a Lasso regularization path. When the outcome model is also fit via the Lasso, we use the resulting representation to demonstrate a familiar "double selection" phenomenon [Belloni et al., 2014].

The Lasso and ℓ_∞ balancing problems will lack a closed form solution when $\Phi_p^T \Phi_p$ is not diagonal. While we present the diagonal case in the main text, we also provide results for the correlated case in Appendix C.

3.3.1 Weighting alone

We begin by defining the soft-thresholding operator and showing that the ℓ_∞ balancing problem has a closed form solution.

Definition (Soft-thresholding operator)**.** *For $t > 0$, define the soft-thresholding operator,*

$$\mathcal{T}_t(z) \coloneqq \begin{cases} 0 & \text{if } |z| < t \\ z - t & \text{if } z > t \\ z + t & \text{if } z < -t \end{cases}.$$

Proposition 3.5 (ℓ_∞ Balancing)**.** *The solution, $w^\delta_{\ell_\infty}$, to the ℓ_∞ optimization problem (10) is:*

$$\begin{aligned} w^\delta_{\ell_\infty} &= \Phi_p (\Phi_p^T \Phi_p)^{-1} \left[\overline{\Phi}_p + \mathcal{T}_\delta(\overline{\Phi}_q - \overline{\Phi}_p) \right] \\ &= \Phi_p (\Phi_p^T \Phi_p)^{-1} \left[\overline{\Phi}_p + \mathcal{T}_\delta(\Delta) \right] \end{aligned}$$

where $\Delta = \overline{\Phi}_q - \overline{\Phi}_p$[17] *and with corresponding reweighted features, $\hat{\Phi}^\delta_q = \overline{\Phi}_p + \mathcal{T}_\delta(\overline{\Phi}_q - \overline{\Phi}_p)$.*

For intuition, compare the (un-augmented) ℓ_∞ balancing weights estimator to the Lasso outcome model estimator with orthonormal design [Hastie et al., 2009]:

$$\begin{aligned} \hat{\mathbb{E}}_p[w^\delta_{\ell_\infty} Y_p] &= \mathcal{T}_\delta(\overline{\Phi}_q)^T \hat{\beta}_{\text{ols}} \\ \hat{\mathbb{E}}_q[\Phi_q \hat{\beta}^\lambda_{\text{lasso}}] &= \overline{\Phi}_q^T \mathcal{T}_\lambda(\hat{\beta}_{\text{ols}}), \end{aligned}$$

where we simplify $\hat{\Phi}^\delta_q$ here to emphasize the connections between the methods. Whereas Lasso performs soft-thresholding on the OLS coefficients (regularizing the outcome regression), ℓ_∞ balancing performs soft-thresholding on the target features.

3.3.2 General linear outcome model

We can then plug the closed-form solution for the weights into Proposition 3.2.

[17] While we include $\overline{\Phi}_p$ here to emphasize the dependence on covariate shift, recall that we assume the features are centered, such that $\overline{\Phi}_p = 0$.

Proposition 3.6. *Let $\hat{w}^{\delta}_{\ell_\infty}$ be defined as above. Then the augmented ℓ_∞ balancing weights estimator with outcome model fit $\hat{\beta}^{\lambda}_{reg} \in \mathbb{R}^d$ has the form,*

$$\hat{\mathbb{E}}_q[\Phi_q \hat{\beta}^{\lambda}_{reg}] + \hat{\mathbb{E}}_p[\hat{w}^{\delta}_{\ell_2}(Y_p - \Phi_p \hat{\beta}^{\lambda}_{reg})] = \hat{\mathbb{E}}_q[\Phi_q \hat{\beta}_{\ell_\infty}],$$

where the jth coefficient of $\hat{\beta}_{\ell_\infty}$ equals:

$$\hat{\beta}_{\ell_\infty,j} = \begin{cases} \hat{\beta}^{\lambda}_{reg,j} & \text{if } |\Delta_j| < \delta \\ \left|\frac{\delta}{\Delta_j}\right| \hat{\beta}^{\lambda}_{reg,j} + \left(1 - \left|\frac{\delta}{\Delta_j}\right|\right) \hat{\beta}_{ols,j} & \text{otherwise} \end{cases},$$

where $\Delta_j = \overline{\Phi}_{q,j} - \overline{\Phi}_{p,j}$.

To obtain this result from the general form of $a^{\delta}_j = \widehat{\Delta}_j / \Delta_j$ in Proposition 3.2, notice that the implied covariate shift, $\widehat{\Delta}_j = \hat{\Phi}^{\delta}_{q,j} - \overline{\Phi}_{p,j} = \mathcal{T}_\delta(\overline{\Phi}_{q,j} - \overline{\Phi}_{p,j})$ is:

$$\widehat{\Delta}_j = \begin{cases} 0 & \text{if } |\Delta_j| < \delta \\ \Delta_j - \delta & \text{if } \Delta_j > \delta \\ \Delta_j + \delta & \text{if } \Delta_j < -\delta \end{cases}.$$

Thus, for instance, $\frac{\widehat{\Delta}_j}{\Delta} = \frac{\Delta_j - \delta}{\Delta_j} = 1 - \frac{\delta}{\Delta_j}$ when $\Delta_j > \delta$.

Again, Proposition 3.6 explicitly highlights the central role of undersmoothing, in the sense of shifting the outcome model coefficients toward OLS. Importantly, this shift is "sparse" in that the outcome regression coefficients remain unchanged when Δ_j is small.

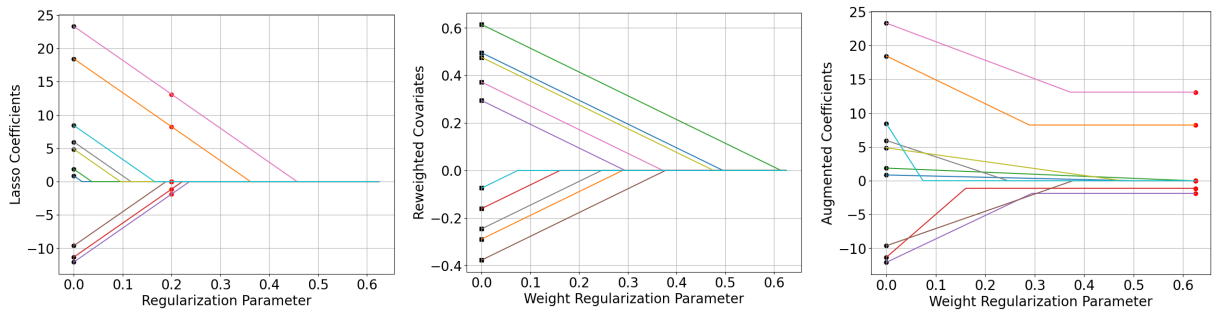
Figure 2 summarizes these results, and their implications for the augmented estimator. As with Figure 1, we generate simple simulated data with $d = 10$. In the left panel, we plot the coefficients from Lasso regression of Y_p on Φ_p as a function of the Lasso regularization parameter. The regularization path begins with the black dots which represent the OLS coefficients. Each Lasso coefficient — represented by a colored line — then shrinks linearly to exactly zero, due to the soft-thresholding operator. The middle panel plots the coefficients of the model for the Riesz representer for ℓ_∞ balancing weights between Φ_p and Φ_q solved in the constrained form. The black dots represent $\overline{\Phi}_q$, corresponding to exact balance. Then as the weight regularization parameter increases, the coefficients shrink linearly to exactly zero, just as in Lasso. The right panel plots results for the augmented estimator that combines a baseline outcome model fit $\hat{\beta}^{\lambda}_{\text{reg}}$ with ℓ_∞ balancing weights. The lines correspond to $\hat{\beta}_{\text{aug}}$ as defined in Proposition 3.2. The regularization path begins at the black dots, where $\hat{\beta}_{\text{aug}} = \hat{\beta}_{\text{ols}}$, and eventually converges to $\hat{\beta}^{\lambda}_{\text{reg}}$, showing the usual soft-thresholding behavior. The order at which the coefficients go to zero reflects the size of $\overline{\Phi}_q$, because the regularization path depends on the weight coefficients from the middle panel. Thus, the augmented estimator shrinks $\hat{\beta}_{\text{ols}}$ toward $\hat{\beta}^{\lambda}_{\text{reg}}$ but via a soft-thresholding operator applied to the feature shift, Δ_j.

3.3.3 Lasso outcome model

In the case where $\hat{\beta}^{\lambda}_{\text{reg}}$ is itself fit via Lasso — as studied in Chernozhukov et al. [2022c] — then we recover a familiar double selection phenomenon [Belloni et al., 2014].

Proposition 3.7 (Double Selection)**.** *Let $\hat{\beta}^{\lambda}_{lasso}$ denote the coefficients of Lasso regression of Φ_p on Y_p with regularization parameter λ. Denote the indices of the non-zero coefficients as I_λ. Let $\hat{w}^{\delta}_{\ell_\infty}$ be ℓ_∞ balancing weights with parameter δ as in Proposition 3.5. Let I_δ denote the non-zero entries of the reweighted covariates $\hat{\Phi}_q$. Then the indices of the non-zero entries of the augmented coefficients $\hat{\beta}_{aug\ell_\infty}$ are:*

$$I_{aug} = I_\lambda \cup I_\delta.$$



(a) Outcome model (b) Weighting model (c) Augmented model

Figure 2: Regularization paths for "double lasso" augmented ℓ_∞ balancing weights. Panel (a) shows the coefficients $\hat{\beta}^{\lambda}_{\text{reg}}$ of a lasso regression of Y_p on Φ_p with hyperparameter λ. The black dots on the left are the OLS coefficients, with $\lambda = 0$. The red dots at $\lambda = 0.2$ illustrate the coefficients at a plausible hyperparameter value, $\hat{\beta}^{0.2}_{\text{reg}}$. Panel (b) shows re-weighted covariates, $\hat{\Phi}^{\delta}_{q}$, for the ℓ_∞ balancing weights problem with hyperparameter δ; the black dots show exact balance, which corresponds to OLS. As δ increases, the weights converge to uniform weights and $\hat{\Phi}^{\delta}_{q}$ converges to $\overline{\Phi}_p$, which we have centered at zero. Panel (c) shows the augmented coefficients, $\hat{\beta}_{\text{aug}}$ as a function of the weight regularization parameter δ. The black dots on the left are the OLS coefficients. As $\delta \to \infty$, the coefficients converge to $\hat{\beta}^{0.2}_{\text{reg}}$. All three regularization paths show the typical Lasso "soft thresholding" behavior. The regularization path for the augmented estimator also shows "double selection" behavior.

The lasso coefficients will have a sparsity pattern generated by soft-thresholding the OLS coefficients. Then the augmented estimator shrinks from OLS toward $\hat{\beta}^{\lambda}_{\text{reg}}$ by soft-thresholding the target covariates. As a result, wherever the Lasso covariates are non-zero *or* the weight coefficients are non-zero, the final augmented coefficients are non-zero. The "included covariates" for the final estimator are the union of the covariates included in either individual model.

4 Extensions

Our results thus far focus on using augmented linear balancing weights to estimate the missing mean. We now discuss two key extensions. First, we show that constraining the weights to be non-negative is equivalent to a form of sample trimming. Second, we show that our numeric results immediately extend to general linear functionals via the Riesz representer.

4.1 Constraining weights to be non-negative

A common modification of the balancing weights problem is to constrain the estimated weights to be non-negative.[18] Such weights have a number of attractive practical properties: they limit extrapolation; they ensure that the final weighting estimator is sample bounded; and they are typically sparse, which can which can sometimes aid interpretability [Robins et al., 2007, Abadie et al., 2010]. Examples of constrained ℓ_∞ weights include Stable Balancing Weights and extensions [Zubizarreta, 2015, Athey et al., 2018, Wang and Zubizarreta, 2020]; examples of constrained ℓ_2 weights include popular variants of the synthetic control method [Abadie et al., 2010, Ben-Michael et al., 2021c].

Using the dual form of the problem, Ben-Michael et al. [2021b] show that linear balancing weights with a non-negativity constraint have the form $\hat{w} = \{\Phi_p \hat{\theta}^{\delta}\}_+$, where $\{x\}_+ = \max(x, 0)$ and where the coefficients

[18] Recall that we already impose the constraint that the weights sum to 1. So imposing non-negativity is equivalent to constraining the weights to be on the simplex.

$\hat{\theta}^\delta$ are generally different from the corresponding coefficients in the unconstrained model.[19] We can apply this insight to extend Proposition 3.1 to non-negative weights.

Proposition 4.1. *Let $\hat{w}_+^\delta \coloneqq \{\Phi_p \hat{\theta}^\delta\}_+$, with $\hat{\theta}^\delta \in \mathbb{R}^d$ and where $\{x\}_+ = max(x, 0)$, be any linear balancing weights with a non-negativity constraint, with corresponding weighted covariates $\hat{\Phi}_q^\delta \coloneqq \hat{w}_+ \Phi_p$. Let Φ_{p+}, and Y_{p+} denote the respective quantities restricted to those data points where $\hat{w}_+^\delta > 0$. Let $\hat{\beta}_{ols}^+ \coloneqq (\Phi_{p+}^T \Phi_{p+})^{-1} \Phi_{p+}^T Y_{p+}$ be the OLS coefficients of the regression of Y_{p+} on Φ_{p+}. Then:*

$$\hat{\mathbb{E}}_p \left[\hat{w}_+ \circ Y_p\right] = \hat{\Phi}_q^\delta \hat{\beta}_{ols}^+.$$

Importantly, this has exactly the same form as Proposition 3.1, except that the usual OLS coefficients from the entire population p, $\hat{\beta}_{\text{ols}}$, are replaced with the OLS coefficients from only those units with positive weight, $\hat{w}_+^\delta > 0$, $\hat{\beta}_{\text{ols}}^+$. We can therefore view the non-negativity constraint as a form of sample trimming. In particular, we can think of the data points where $\hat{w}_+^\delta = 0$ as a set of outliers — too dissimilar from the target population Φ_q — that we trim before applying OLS. But the key is that the definition of "outlier" depends the choice of δ and $\mathcal{F}$, and even the target covariates Φ_q: in general, changing δ, $\mathcal{F}$, or Φ_q will change the set where $\hat{w}_+^\delta > 0$. Defining and characterizing how this set changes is a promising avenue for future research.

Proposition 4.1 simplifies further when the balancing weights achieve exact balance; see Rubinstein et al. [2021, Proposition 10] for a discussion of this special case. When $\hat{\Phi}_q = \Phi_q$, the balancing weight estimator with the non-negativity constraint is equivalent to $\hat{\mathbb{E}}_q[\Phi_q \hat{\beta}_{\text{ols}}^+]$. Thus, linear balancing weights with a simplex constraint is equivalent to trimming the control group and applying standard OLS (note that the trimming does not affect the target covariate profile, Φ_q).

Finally, we can extend the results in Proposition 3.2 for augmented balancing weights to incorporate a non-negativity constraint.

Proposition 4.2. *For $\hat{w}_+^\delta, \hat{\theta}^\delta, \hat{\Phi}_q^\delta, \hat{\beta}_{ols}^+$ as defined in Proposition 4.1 and any regularized linear outcome regression of Y_p on Φ_p with estimated coefficients $\hat{\beta}_{reg}^\lambda$, the resulting augmented estimator*

$$\begin{aligned}
&\hat{\mathbb{E}}_p[\hat{w}^\delta \circ Y_p] + \hat{\mathbb{E}}_q\left[\left(\Phi_q - \hat{\Phi}_q^\delta\right)\hat{\beta}_{reg}^\lambda\right] \\
&= \hat{\mathbb{E}}_q\left[\hat{\Phi}_q^\delta \hat{\beta}_{ols}^+ + \left(\Phi_q - \hat{\Phi}_q^\delta\right)\hat{\beta}_{reg}^\lambda\right] \\
&= \hat{\mathbb{E}}_q[\Phi_q \hat{\beta}_{aug}^+],
\end{aligned}$$

where the jth element of $\hat{\beta}_{aug}^+$ is:

$$\begin{aligned}
\hat{\beta}_{aug,j}^+ &\coloneqq \left(1 - a_j^\delta\right)\hat{\beta}_{reg,j}^\lambda + a_j^\delta \hat{\beta}_{ols,j}^+ \\
a_j^\delta &\coloneqq \frac{\widehat{\Delta}_j^\delta}{\Delta_j},
\end{aligned}$$

where $\Delta_j = \overline{\Phi}_{q,j} - \overline{\Phi}_{p,j}$ is the observed mean feature shift for feature j, and $\widehat{\Delta}_j^\delta = \hat{\Phi}_{q,j}^\delta - \overline{\Phi}_{p,j}$ is the mean feature shift for feature j implied by the balancing weights model.

Many popular augmented balancing weights estimators have the form of Proposition 3.2, including Athey et al. [2018] and Ben-Michael et al. [2021c]. Understanding the implications of this connection is an interesting direction for future work.

[19] This "positive part link" representation is unique to the minimum variance weights. Other dispersion penalties, such as the entropy of the weights, imply a different "link function." See Zhao [2019] and Ben-Michael et al. [2021b].

4.2 General linear functionals

The results above apply to the target estimand $\mathbb{E}_q[Y]$ and to the setting with two populations, p and q. However, the numerical results generalize to other linear functionals and arbitrary sets of populations as in Chernozhukov et al. [2018], Hirshberg and Wager [2021]. Let $\mathcal{T}$ be some arbitrary set indexing populations and let T be a random variable ranging over $\mathcal{T}$. Assume that we observe tuples (T, X, Y) from a training distribution p and define $\mu(t, x) \coloneqq \mathbb{E}_p[Y|T = t, X = x]$.[20] Let our estimand be $\mathbb{E}_p[m(T, X, \mu)]$ where m is a (real-valued) mean-squared continuous linear functional of μ.[21]

Following Chernozhukov et al. [2022c], there exists a function $\alpha(T, X)$ called the Riesz representer such that:

$$\mathbb{E}_p[m(T, X, \mu)] = \mathbb{E}_p[\alpha(T, X)\mu(T, X)] = \mathbb{E}_p[\alpha(T, X)Y]. \tag{18}$$

We can generalize our results from the rest of the paper to work for any $m(\mu, t, x)$ and any conditional expectation $\mu(t, x)$. To do so, first define the finite signed measure q such that $dq(t, x)/dp(t, x) = \alpha(t, x)$. Such a finite signed measure is guaranteed to exist because for any mean-squared continuous m, $\mathbb{E}_p[\alpha(T, X)^2] \leq \infty$.

Now we need to verify that the results described in the paper hold when q is a finite signed measure instead of a probability measure. As demonstrated above, the regression functional and IPW functionals hold. Furthermore, in the population-level, the solutions to the penalized and constrained balancing weights problem and their equivalent solutions to the automatic form all exist as demonstrated in Bruns-Smith and Feller [2022], where q is assumed to be a finite signed measure throughout.

Given the new definition of q, $\text{Imbalance}_{\mathcal{F}}(w)$ can be written:

$$\begin{aligned}
\text{Imbalance}_{\mathcal{F}}(w) &\coloneqq \sup_{f \in \mathcal{F}} \Big\{ \mathbb{E}_p[w(T, X)f(T, X)] - \mathbb{E}_q[f(T, X)] \Big\} \\
&= \sup_{f \in \mathcal{F}} \Big\{ \mathbb{E}_p[w(T, X)f(T, X)] - \mathbb{E}_p[\alpha(T, X)f(T, X)] \Big\} \\
&= \sup_{f \in \mathcal{F}} \Big\{ \mathbb{E}_p[w(T, X)f(T, X)] - \mathbb{E}_p[m(T, X, f)] \Big\}.
\end{aligned}$$

So then, perhaps surprisingly, we can consider complicated sets of populations $\mathcal{T}$ and various linear functionals, but always reduce the problem to a balancing weights problem between two finite signed measures p and q. If both $f \in \mathcal{F}$ and $m(t, x, f)$ are linear, then *all* of our numerical results above go through exactly, including the un-regularization toward $\hat{\beta}_{\text{ols}}$ implemented by the augmented estimator.

An important difference in this more general setting is that the basis for p and q will not generally be the same. Define $\phi : \mathcal{T} \times \mathcal{X} \to \mathbb{R}^d$ and let $\phi_i : \mathcal{T} \times \mathcal{X} \to \mathbb{R}$ denote the mapping for the ith feature. Let $\mathcal{F}$ be linear functions in the features $\phi(X)$. Then for any $f = \phi(t, x)^T\beta \in \mathcal{F}$:

$$\begin{aligned}
\mathbb{E}_p[m(T, X, f)] &= \mathbb{E}_p[f(T, X)\alpha(T, X)] \\
&= \mathbb{E}_p[\beta^T\phi(T, X)\alpha(T, X)] \\
&= \beta^T\mathbb{E}_p[\phi(T, X)\alpha(T, X)] \\
&= \mathbb{E}_p[\beta^T m(T, X, \phi)],
\end{aligned}$$

where $\mathbb{E}_p[m(T, X, \phi)]$ is short-hand for the vector with ith entry $\mathbb{E}_p[m(T, X, \phi_i)]$. So in expectation, if f is linear in parameters of length d then it is sufficient to represent $m(T, X, f)$ as linear in parameters of length

[20] Note that for estimation in this setting to have a causal interpretation, we would need to define potential outcomes $Y(t)$ for $t \in \mathcal{T}$ and a corresponding ignorability assumption that $\mathbb{E}_p[Y(t)|T, X]$ are equal for all t. But this additional setup is not necessary for any of the statements in this section.

[21] While we consider real-valued linear functionals, see the discussion in Singh et al. [2020] for estimands that are function-valued linear functionals. Their presentation combines nicely with our approach and an extension to this more complicated setting is a promising avenue for future work.

d but in a transformed basis, $m(T, X, \phi)$. Thus if X_p, T_p are n i.i.d. samples from p, then $\Phi_p = \phi(T_p, X_p)$ whereas $\Phi_q = m(T_p, X_p, \phi)$. However, once we have defined these quantities Φ_p and Φ_q then we have successfully reduced the general linear functional setting to be numerically identical to the binary setting of the rest of the text, and all of our numerical results go through without alteration.

4.2.1 The Average Derivative Effect

We provide an illustrative example for the average derivative effect. In this setting $T \in \mathbb{R}$ denotes some continuous treatment and

$$m(t, x, \mu) \coloneqq \frac{\partial \mu}{\partial t}(t, x).$$

Assume that $\mathcal{F}$ are linear functions in a basis $\phi(t, x) \in \mathbb{R}^d$ with coefficients that lie within a $\|\cdot\|$ norm ball with dual norm $\|\cdot\|_*$. Then we take inspiration from Klosin [2021], Klosin and Vilgalys [2022] and use the fact that for any $f(x, t) = \phi(t, x)^T \beta$:

$$m(t, x, \phi) = \frac{\partial \phi}{\partial t}(t, x),$$

and therefore, from the discussion above, $\mathbb{E}_p[m(T, X, f)] = \mathbb{E}_p[m(T, X, \phi)^T \beta]$.

Let X_p, T_p, Y_p be n i.i.d. samples from p. Define $\Phi_p = \phi(T_p, X_p) \in \mathbb{R}^{n \times d}$ and $\Phi_q = \frac{\partial \phi}{\partial t}(T_p, X_p)$. Then:

$$\text{Imbalance}_{\mathcal{F}}(w) = \|w\Phi_p - \bar{\Phi}_q\|_*$$

and the rest of our numerical results apply exactly. For example, let $\hat{w}^\delta_{\mathcal{F}}$ be weights with hyperparameter δ that solve the balancing weights problem between Φ_p and $\bar{\Phi}_q$, exactly as in Section 2.3. Define $\hat{\bar{\Phi}}_q \coloneqq \hat{w}^\delta_{\mathcal{F}} \Phi_p$. Then Proposition 3.1 gives us:

$$\hat{\mathbb{E}}_p \left[\hat{w}^\delta_{\mathcal{F}} \circ Y_p\right] = \hat{\bar{\Phi}}^\delta_q \hat{\beta}_{\text{ols}}.$$

Likewise, consider a regularized outcome regression of Y_p on Φ_p, $\hat{\beta}^\lambda_{\text{reg}}$. Then applying Proposition 3.2, the augmented estimator equals (expanding the definitions of Φ_p and Φ_q for clarity)

$$\hat{\mathbb{E}}_p \left[\frac{\partial \phi}{\partial t}(T_p, X_p)\hat{\beta}^\lambda_{\text{reg}}\right] + \hat{\mathbb{E}}_p \left[\hat{w}^\delta_{\mathcal{F}}\left(Y_p - \phi(T_p, X_p)\hat{\beta}^\lambda_{\text{reg}}\right)\right] = \hat{\mathbb{E}}_p \left[\frac{\partial \phi}{\partial t}(T_p, X_p)\hat{\beta}_{\text{aug}}\right],$$
$$\hat{\beta}_{\text{aug},j} \coloneqq (1 - a^\delta_j)\hat{\beta}^\lambda_{\text{reg},j} + a^\delta_j \hat{\beta}_{\text{ols},j}.$$

The intuitive idea here, is that estimating the average derivative effect is still an example of a binary, $p \to q$ covariate shift even though T is continuous. We fit an outcome model in the training population $\phi(T_p, X_p)$ but then apply it to a (potentially very different) test population $\frac{\partial \phi}{\partial t}(T_p, X_p)$. And so nothing fundamental changes about the results — augmentation unregularizes that outcome model fit back toward OLS exactly like it would when training in a control population and testing in a binary treated population.

5 Numerical illustration

6 Discussion

We have shown that augmenting a base learner linear outcome regression with linear balancing weights results in a new regression that shrinks the coefficients from the base learner toward OLS. Our focus in this paper has been on interpreting balancing weights as a form of linear regression. The converse is also valid, however: we could instead focus on how many outcome regression-based plug-in estimators are, in fact, a form of balancing weights [see Lin and Han, 2022, for connections between outcome modeling and density ratio estimation].

While our results are numerical and finite-sample, they have implications for the asymptotic analysis of plug-in and augmented estimators where nuisance parameters are estimated using regularization, such as new ways to choose hyperparameters in asymptotic regimes and new conditions under which a base learner outcome regression may be efficient. The technical details are the subject of our future research, but we comment here on some high-level insights related to the asymptotic properties of augmented estimators.

Connection between doubly robust estimation and undersmoothing. Our results suggest a way to understand DML augmented estimators in general through the lens of undersmoothing.

Consider the setting where we want to estimate $\mathbb{E}_q[Y]$ under enough sparsity or smoothness constraints that $\frac{dq}{dp}(X)$ and $\mathbb{E}_p[Y|X]$ are estimable at rates sufficiently fast that their product is $o_p(n^{-1/2})$. As we discuss in the introduction, even though the outcome regression estimator is consistent for $\mathbb{E}_q[Y]$, a plug-in estimator using the regression fit will lead to non-negligible regularization bias.

In general, augmented or doubly robust estimators are the result of subtracting the (estimated) first-order bias of a plug-in estimator of the regression functional from that estimator [Kennedy, 2022]. This first-order bias could be due to misspecification of the outcome regression, but when the outcome regression is consistent but subject to slow convergence, the first-order bias is the regularization bias. Heuristically, regularization bias arises because features that do not strongly predict Y may still engender nonnegligible confounding when both outcome and weight models are taken together; these features are penalized in the outcome regression, leading to confounding bias. Augmenting the outcome regression with a weight model counteracts this bias [Chernozhukov et al., 2018, Kennedy, 2022].

Undersmoothing, i.e., choosing a sub-optimal hyperparameter that prioritizes bias over MSE, has been proposed as an alternative way to drive down regularization bias [Newey and Robins, 2018, McGrath and Mukherjee, 2022, Mou et al., 2023, van der Laan et al., 2022]. The literature on undersmoothing typically focuses on outcome regression and ignores weights altogether. For a weight-agnostic strategy to yield desirable convergence rates, undersmoothing must capture all covariates that could possibly result in confounding bias for *any* weight model. In other words, undersmoothing without reference to the weight model must protect against the confounding that would be implied by worst-case, adversarial weight models.

The numeric results in our paper suggest that augmenting is an alternative approach to undersmoothing that can directly target the confounding surface that matters. This demonstrates, as a proof of concept, that knowledge about the weight surface, even if nonparametric (e.g. smoothness or sparsity assumptions), can reduce the burden on undersmoothing. Moreover, we anticipate that researchers will be able to mix-and-match various assumptions on undersmoothing versus augmentation depending on the specifics of the application.

Double robustness of single-stage estimators. A related line of inquiry generalizes the result from Robins et al. [2007] that "OLS is doubly robust." As we show in Section 3.2, we can always re-write a single kernel ridge regression as an augmented balancing weights estimator that augments a kernel ridge regression outcome model with kernel balancing weights. Thus, under a suitable rate-double robustness regime [Rotnitzky et al., 2021], we can also analyze kernel ridge regression (or, equivalently, kernel balancing weights) as a doubly robust estimator. This insight complements previous papers on the efficiency of balancing weights approaches [Graham et al., 2012, Chan et al., 2016, Fan et al., 2016], especially those that incorporate regularization [Wang and Zubizarreta, 2019, 2020]. In particular, our results suggest that there may be a wider range of hyperparameter values than previously known for which a (kernel) ridge plug-in or ℓ_2-balancing weights estimator attains efficiency, which would likely depend on the range of hyperparameter values for which the product bias goes to zero. This does not carry over straightforwardly to ℓ_∞-balancing weights, suggesting possible tradeoffs between the two main families of balancing weight norms.

Additional directions. We anticipate that connecting our numeric results to underlying statistical models will lead to better guidance for deploying augmented balancing weights in practice. First, hyperparameter

tuning is a major implementation challenge, with a range of recommendations [e.g., Kallus, 2020, Wang and Zubizarreta, 2020, Chernozhukov et al., 2022c]. We expect that the connection to theoretically optimal results for undersmoothing [Mou et al., 2023] can help guide hyperparameter choice in these cases.

Second, many common panel data estimators are forms of augmented balancing weight estimation [Abadie et al., 2010, Ben-Michael et al., 2021c, Arkhangelsky et al., 2021]. We plan to use the numeric results here, especially the results for simplex-constrained weights in Section 4.1, to better understand connections between methods and to inform inference.

Third, we conjecture that these results may provide new insights into the estimation of causal effects in the proximal causal inference framework [Tchetgen et al., 2020]. This framework uses proxy variables to estimate causal effects in the presence of unmeasured confounding. Estimation has been complicated by the fact that, in the absence of strong parametic assumptions, estimators of proximal causal effects are solutions to ill-posed Fredholm integral equations. Ghassami et al. [2022] recently proposed the first tractable nonparametric estimator in this setting. The authors use a version of double kernel ridge regression—where the weighting and outcome models have different bases—to estimate the solution to the required Fredholm integral equations. Our result applies immediately to standard augmented estimators with different bases for the outcome and weighting models, either via a union basis [Chernozhukov et al., 2022c] or by applying an appropriate projection as in Hirshberg and Wager [2021], and we believe that our results can also be formally extended to proximal causal effect estimators like that in Ghassami et al. [2022]. This suggests, for example, that undersmoothed outcome regression could be consistent and achieve reasonable rates for proximal causal effects under the weak assumptions of Ghassami et al. [2022].

Finally, despite the growing popularity of balancing weights methods, traditional IPW remains the workhorse approach for estimating the density ratio. Understanding the implications of our equivalence results for model-based propensity score estimation, such as via logistic regression, is an important direction for future work. In Appendix D, we explore results for *nonlinear* balancing weights, which could provide a useful bridge to traditional IPW approaches.

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.

A. Agarwal and R. Singh. Causal inference with corrupted data: Measurement error, missing values, discretization, and differential privacy. *arXiv preprint arXiv:2107.02780*, 2021.

D. Arkhangelsky, S. Athey, D. A. Hirshberg, G. W. Imbens, and S. Wager. Synthetic difference-in-differences. *American Economic Review*, 111(12):4088–4118, 2021.

S. Athey, G. W. Imbens, and S. Wager. Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):597–623, 2018.

P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

A. Belloni, V. Chernozhukov, and C. Hansen. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650, 2014.

E. Ben-Michael, A. Feller, and E. Hartman. Multilevel calibration weighting for survey data. *arXiv preprint arXiv:2102.09052*, 2021a.

E. Ben-Michael, A. Feller, D. A. Hirshberg, and J. R. Zubizarreta. The balancing act in causal inference. *arXiv preprint arXiv:2110.14831*, 2021b.

E. Ben-Michael, A. Feller, and J. Rothstein. The augmented synthetic control method. *Journal of the American Statistical Association*, 116(536):1789–1803, 2021c.

D. Benkeser and M. Van Der Laan. The highly adaptive lasso estimator. In *2016 IEEE international conference on data science and advanced analytics (DSAA)*, pages 689–696. IEEE, 2016.

D. A. Bruns-Smith and A. Feller. Outcome assumptions and duality theory for balancing weights. In *International Conference on Artificial Intelligence and Statistics*, pages 11037–11055. PMLR, 2022.

P. Bühlmann and B. Yu. Boosting with the l 2 loss: regression and classification. *Journal of the American Statistical Association*, 98(462):324–339, 2003.

K. C. G. Chan, S. C. P. Yam, and Z. Zhang. Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(3):673–700, 2016.

A. Chattopadhyay and J. R. Zubizarreta. On the implied weights of linear regression for causal inference. *arXiv preprint arXiv:2104.06581*, 2021.

A. Chattopadhyay, C. H. Hase, and J. R. Zubizarreta. Balancing vs modeling approaches to weighting in practice. *Statistics in Medicine*, 39(24):3227–3254, 2020.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

V. Chernozhukov, J. C. Escanciano, H. Ichimura, W. K. Newey, and J. M. Robins. Locally robust semiparametric estimation. *Econometrica*, 90(4):1501–1535, 2022a.

V. Chernozhukov, W. Newey, V. M. Quintas-Martınez, and V. Syrgkanis. Riesznet and forestriesz: Automatic debiased machine learning with neural nets and random forests. In *International Conference on Machine Learning*, pages 3901–3914. PMLR, 2022b.

V. Chernozhukov, W. K. Newey, and R. Singh. Automatic debiased machine learning of causal and structural effects. *Econometrica*, 90(3):967–1027, 2022c.

V. Chernozhukov, W. K. Newey, and R. Singh. Debiased machine learning of global and local parameters using regularized riesz representers. *The Econometrics Journal*, 2022d.

J.-C. Deville and C.-E. Särndal. Calibration estimators in survey sampling. *Journal of the American statistical Association*, 87(418):376–382, 1992.

J. Fan, K. Imai, H. Liu, Y. Ning, X. Yang, et al. Improving covariate balancing propensity score: A doubly robust and efficient approach. *URL: https://imai. fas. harvard. edu/research/CBPStheory. html*, 2016.

W. A. Fuller. Regression estimation for survey samples. *Survey Methodology*, 28(1):5–24, 2002.

A. Ghassami, A. Ying, I. Shpitser, and E. T. Tchetgen. Minimax kernel machine learning for a class of doubly robust functionals with application to proximal causal inference. In *International Conference on Artificial Intelligence and Statistics*, pages 7210–7239. PMLR, 2022.

L. Goldstein and K. Messer. Optimal plug-in estimators for nonparametric functional estimation. *The annals of statistics*, pages 1306–1328, 1992.

B. S. Graham, C. C. de Xavier Pinto, and D. Egel. Inverse probability tilting for moment condition models with missing data. *The Review of Economic Studies*, 79(3):1053–1079, 2012.

Y. Grandvalet. Least absolute shrinkage is equivalent to quadratic penalization. In *International Conference on Artificial Neural Networks*, pages 201–206. Springer, 1998.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

J. Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political analysis*, 20(1):25–46, 2012.

C. Harshaw, F. Sävje, D. Spielman, and P. Zhang. Balancing covariates in randomized experiments with the gram–schmidt walk design. *arXiv preprint arXiv:1911.03071*, 2019.

T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

C. Hazlett. Kernel balancing. *Statistica Sinica*, 30(3):1155–1189, 2020.

D. A. Hirshberg and S. Wager. Augmented minimax linear estimation. *The Annals of Statistics*, 49(6): 3206–3227, 2021.

D. A. Hirshberg, A. Maleki, and J. R. Zubizarreta. Minimax linear estimation of the retargeted mean. *arXiv preprint arXiv:1901.10296*, 2019.

A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

N. Kallus. Generalized optimal matching methods for causal inference. *J. Mach. Learn. Res.*, 21:62–1, 2020.

J. D. Kang and J. L. Schafer. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, 22(4):523–539, 2007.

E. H. Kennedy. Semiparametric doubly robust targeted double machine learning: a review. *arXiv preprint arXiv:2203.06469*, 2022.

K. Kim, B. A. Niknam, and J. R. Zubizarreta. Scalable kernel balancing weights in a nationwide observational study of hospital profit status and heart attack outcomes. 2022.

P. Kline. Oaxaca-blinder as a reweighting estimator. *American Economic Review*, 101(3):532–37, 2011.

S. Klosin. Automatic double machine learning for continuous treatment effects. *arXiv preprint arXiv:2104.10334*, 2021.

S. Klosin and M. Vilgalys. Estimating continuous treatment effects in panel data using machine learning with an agricultural application. *arXiv preprint arXiv:2207.08789*, 2022.

R. R. Lin, H. Z. Zhang, and J. Zhang. On reproducing kernel banach spaces: Generic definitions and unified framework of constructions. *Acta Mathematica Sinica, English Series*, 38(8):1459–1483, 2022.

Z. Lin and F. Han. On regression-adjusted imputation estimators of the average treatment effect. *arXiv preprint arXiv:2212.05424*, 2022.

S. McGrath and R. Mukherjee. On undersmoothing and sample splitting for estimating a doubly robust functional. *arXiv preprint arXiv:2212.14857*, 2022.

W. Mou, P. Ding, M. J. Wainwright, and P. L. Bartlett. Kernel-based off-policy estimation without overlap: Instance optimality beyond semiparametric efficiency, 2023. URL https://arxiv.org/abs/2301.06240.

W. K. Newey and J. R. Robins. Cross-fitting and fast remainder rates for semiparametric estimation. *arXiv preprint arXiv:1801.09138*, 2018.

W. K. Newey, F. Hsieh, and J. Robins. Undersmoothing and bias corrected functional estimation. 1998.

W. K. Newey, F. Hsieh, and J. M. Robins. Twicing kernels and a small bias property of semiparametric estimators. *Econometrica*, 72(3):947–962, 2004.

B. Park, Y. Lee, and S. Ha. l_2 boosting in kernel regression. *Bernoulli*, 15(3):599–613, 2009.

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.

J. M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994.

A. Rotnitzky, E. Smucler, and J. M. Robins. Characterization of parameters with a mixed bias property. *Biometrika*, 108(1):231–238, 2021.

D. B. Rubin. Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American statistical association*, 75(371):591–593, 1980.

M. Rubinstein, A. Haviland, and D. Choi. Balancing weights for region-level analysis: the effect of medicaid expansion on the uninsurance rate among states that did not expand medicaid. *arXiv preprint arXiv:2105.02381*, 2021.

D. Shen, P. Ding, J. Sekhon, and B. Yu. A tale of two panel data regressions. *arXiv preprint arXiv:2207.14481*, 2022.

R. Singh, L. Xu, and A. Gretton. Kernel methods for causal functions: Dose, heterogeneous, and incremental response curves. *arXiv preprint arXiv:2010.04855*, 2020.

R. Singh, L. Sun, et al. Double robustness for complier parameters and a semiparametric test for complier characteristics. Technical report, 2022.

M. Sugiyama, T. Suzuki, and T. Kanamori. Density-ratio matching under the bregman divergence: a unified framework of density-ratio estimation. *Annals of the Institute of Statistical Mathematics*, 64(5):1009–1044, 2012.

Z. Tan. Regularized calibrated estimation of propensity scores with model misspecification and high-dimensional data. *Biometrika*, 107(1):137–158, 2020.

E. J. T. Tchetgen, A. Ying, Y. Cui, X. Shi, and W. Miao. An introduction to proximal causal learning. *arXiv preprint arXiv:2009.10982*, 2020.

M. J. van der Laan, D. Benkeser, and W. Cai. Efficient estimation of pathwise differentiable target parameters with the undersmoothed highly adaptive lasso. *The International Journal of Biostatistics*, 2022.

Y. Wang and J. R. Zubizarreta. Large sample properties of matching for balance. *arXiv preprint arXiv:1905.11386*, 2019.

Y. Wang and J. R. Zubizarreta. Minimal dispersion approximately balancing weights: asymptotic properties and practical considerations. *Biometrika*, 107(1):93–105, 2020.

Q. Zhao. Covariate balancing propensity score by tailored loss functions. *The Annals of Statistics*, 47(2): 965–993, 2019.

Q. Zhao and D. Percival. Entropy balancing is doubly robust. *Journal of Causal Inference*, 5(1), 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 Sample Splitting and Cross-fitting

We briefly discuss the application of our numerical results in the setting with sample splitting. In standard balancing weights, we fit the weighting model θ^δ using data from Φ_p, but then also apply the weighting model at Φ_p. We can break any possible dependences by only applying parameters on samples that were not used for estimation, such as via cross-fitting, which is a core technique in AutoDML [Chernozhukov et al., 2022c].

Let the n samples from population p be split into S partitions or "splits" and assume for simplicity that each split has size $n' := n/S$. Denote the split s covariates $\Phi_{p,s}$ and outcomes $Y_{p,s}$. Let $\Phi_{p,-s}$ and $Y_{p,-s}$ denote covariates and outcomes that are not in split s. As a simple example, consider a cross-fit, unaugmented ℓ_2 balancing weights with parameter $\delta = 0$. We first solve the balancing problem *out-of-sample* by solving for the coefficients as in the example (12):

$$\begin{aligned}\hat{\theta}^0_{-s} &:= \text{argmin}_{\theta \in \mathbb{R}^d} \left\{ \|\theta \Phi^T_{p,-s} \Phi_{p,-s} - \bar{\Phi}_q\|_2^2 + \delta \|\theta \Phi^T_{p,-s}\|_2^2 \right\} \\ &= \bar{\Phi}_q (\Phi^T_{p,-s} \Phi_{p,-s})^{-1}.\end{aligned}$$

Note that we have re-written the balancing problem to be in terms of the coefficients θ instead of the weights $w = \theta \Phi_{p,-s}$, in order to emphasize that the goal is to apply this weighting model to the split s samples to obtain weights $\hat{\theta}^0_{-s} \Phi^T_{p,s}$. The split s balancing weights estimate is then $\hat{\theta}^0_{-s} \Phi^T_{p,s} Y_{p,s}$ and the final cross-fit estimator averages over these splits:

$$\frac{1}{S} \sum_{s=1}^{S} \hat{\theta}^0_{-s} \Phi^T_{p,s} Y_{p,s}.$$

Note that the coefficients $\hat{\theta}^0_{-s}$ enforce exact balance — but only for data outside split s. In general, these weights will not achieve exact balance for split s. That is:

$$\begin{aligned}\|\hat{\theta}^0_{-s} \Phi^T_{p,-s} \Phi_{p,-s} - \bar{\Phi}_q\|_2 &= 0, \\ \|\hat{\theta}^0_{-s} \Phi^T_{p,-s} \Phi_{p,s} - \bar{\Phi}_q\|_2 &\neq 0.\end{aligned}$$

For the augmented estimator, we would also fit an outcome model using data from outside split s. For example, we could fit OLS:

$$\hat{\beta}_{\text{ols},-s} := (\Phi^T_{p,-s} \Phi_{p,-s})^\dagger \Phi^T_{p,-s} Y_{p,-s}.$$

The augmented estimator combining $\hat{\theta}^0_{-s}$ with $\hat{\beta}_{\text{ols},-s}$ would give:

$$\frac{1}{S} \sum_{s=1}^{S} \left(\bar{\Phi}_q \hat{\beta}_{\text{ols},-s} + \hat{\theta}^0_{-s} \Phi^T_{p,s} (Y_{p,s} - \Phi_{p,s} \hat{\beta}_{\text{ols},-s}) \right).$$

A.1 Proposition 3.2 with Sample Splitting

For Proposition 3.2 with sample splitting, the numerical result is identical, but the substantive point of interest is that we always unregularize toward the *in-sample* OLS coefficients. That is, the augmented estimator always chooses a linear model that overfits to the sample more than the base-learner.

Let the coefficients $\hat{\beta}^\lambda_{-s}$ and $\hat{\theta}^\delta_{-s}$ be some fixed set of vectors; in practice they will be models fit using samples not in s. Define $\hat{\Phi}^\delta_{q,s} := \hat{w}^\delta_s \Phi_{p,s}$, and $\hat{\beta}_{\text{ols},s} := (\Phi^T_{p,s} \Phi_{p,s})^\dagger \Phi^T_{p,s} Y_{p,s}$. Then the augmented estimator in the sth partition follows immediately from Proposition 3.2 in the main text:

$$\hat{\mathbb{E}}_q[\Phi_q \hat{\beta}_{\text{aug},s}],$$

where the jth element of $\hat{\beta}_{\text{aug},s}$ is:

$$\hat{\beta}_{\text{aug},s,j} \coloneqq \left(1 - a^{\delta}_{j,s}\right) \hat{\beta}^{\lambda}_{-s,j} + a^{\delta}_{j,s} \hat{\beta}_{\text{ols},s,j}$$

$$a^{\delta}_{j,s} \coloneqq \frac{\widehat{\Delta}^{\delta}_{j,s}}{\Delta_{j,s}},$$

where $\Delta_{j,s} = \overline{\Phi}_{q,j} - \overline{\Phi}_{p,s,j}$ and $\widehat{\Delta}^{\delta}_{j,s} = \hat{\Phi}^{\delta}_{q,s,j} - \overline{\Phi}_{p,j}$.

A.2 Unregularized Outcome Model

Whereas Proposition 3.2 is unchanged by sample splitting, some of the equivalence results are affected by sample splitting. For example, we know from Robins et al. [2007] that when the base learner is unregularized, i.e., $\hat{\beta}^{\lambda}_{\text{reg}} = \hat{\beta}_{\text{ols}}$, then the entire estimator collapses to OLS alone. With sample splitting, this is only true if $\hat{\beta}^{\lambda}_{\text{reg}} = \hat{\beta}_{\text{ols},s}$. With cross-fitting, however, the outcome model would typically be estimated using only data from outside split s.

For example, consider $\hat{\beta}_{\text{ols},-s}$ introduced above. Plugging this in to the result in Appendix A.1 yields:

$$\hat{\beta}_{\text{aug},s,j} \coloneqq \left(1 - a^{\delta}_{j,s}\right) \hat{\beta}^{\lambda}_{\text{ols},-s,j} + a^{\delta}_{j,s} \hat{\beta}_{\text{ols},s,j}.$$

When the OLS coefficients are fit out of sample, this prevents overfitting to the ℓth split. Augmentation actually shifts the out-of-split OLS coefficents back toward the in-split OLS coefficients — effectively another form of "unregularizing". As the sample size in each split goes to infinity, then both $\hat{\beta}_{\text{ols},s}$ and $\hat{\beta}_{\text{ols},-s}$ converge to the same population OLS coefficients and the augmented coefficients converge to standard OLS for any weighting model $\theta^{\delta} \in \mathbb{R}^d$.

A.3 Unregularized Weight Model

Consider the opposite case where $\hat{\beta}^{\lambda}_{\text{reg}}$ is arbitrary and the weight model θ^{0}_{-s} achieves exact balance between $\Phi_{p,-\ell}$ and Φ_q, as defined above. Then, as suggested in Appendix A.1, $\hat{w}^0_s = \theta^0_{-\ell} \Phi_{p,\ell}$, and $\hat{\Phi}^0_{q,s} \coloneqq \hat{w}^0_s \Phi_{p,\ell} \neq \bar{\Phi}_q$ in general. Instead, we only have an approximation:

$$a^{\delta}_{j,s} \coloneqq \frac{\hat{\Phi}^0_{q,s,j} - \overline{\Phi}_{p,s,j}}{\overline{\Phi}_{q,j} - \overline{\Phi}_{p,s,j}} \approx 1,$$

where the approximation becomes equality as the sample size in each split goes to infinity. As a result, $\hat{\beta}_{\text{aug},s} \approx \hat{\beta}_{\text{ols},s}$, the in-split OLS coefficients, where again these coefficients are equal as the sample size in each split goes to infinity.

A.4 "Double Ridge"

Similarly, whereas Proposition 3.4 reduced to a single ridge outcome model, with sampling splitting we instead obtain an affine combination of in-split and out-of-split ridge regressions. Let $\hat{\beta}^{\lambda}_{-s}$ and $\hat{\theta}^{\delta}_{-s}$ denote ridge and ℓ_2 balancing coefficients respectively fit outside of the ℓth split. Assume that $(\Phi^T_{p,s}\Phi_{p,s}) = \text{diag}(\sigma^2_{1,s}, ..., \sigma^2_{d,s})$ and similarly for $-s$. Then the augmented estimator in the sth split equals $\hat{\mathbb{E}}_q[\Phi_q \hat{\beta}_{\text{aug},s}]$, where

$$\hat{\beta}_{\text{aug},s,j} \coloneqq \left(\frac{\sigma^2_{j,-s} - \sigma^2_{j,s} + \delta}{\sigma^2_{j,-s} + \delta}\right) \hat{\beta}^{\lambda}_{-s,j} + \left(\frac{\sigma^2_{j,s} + \delta}{\sigma^2_{j,-s} + \delta}\right) \hat{\beta}^{\delta}_{s,j}.$$

B Details for when $d > n$

None of the results in the main text essentially require that $d < n$ or that Φ_p have rank d. In this section, we formally state our results in the high-dimensional setting. In all that follows we will assume that $d > n$ and we will assume Φ_p^T has rank n.[22] For $d = \infty$, we replace $\mathbb{R}^d$ with any infinite-dimensional Hilbert space $\mathcal{H}$ and we require the norm defining $\mathcal{F}$ to be the norm of the Hilbert space. In this case (with a slight abuse of notation) it should be understood that $\Phi_p \in \mathcal{H}^n$.

B.1 Balancing weights when $d > n$

In the main text, recall that there are three equivalent versions of the balancing weights problem: the penalized, constrained, and automatic form with hyperparameters $\delta_1, \delta_2, \delta_3 \geq 0$ respectively. When $\Phi_p^T \Phi_p$ is no longer invertible, a unique solution may fail to exist for certain values of these hyperparameters. We provide the relevant technical caveats here.

We begin by mentioning that for $\delta_1 > 0$, the penalized form of the balancing weights optimization problem is strictly convex, and therefore a unique solution exists, regardless of whether $d > n$. However, when $\delta_1 = 0$, there could potentially be infinite many solutions. In this setting, we choose the one with the minimum norm:

$$\begin{aligned} &\min_{w \in \mathbb{R}^n} \|w\|_2^2 \\ &\text{such that } \|w\Phi_p - \overline{\Phi}_q\|_*^2 = \min_v \|v\Phi_p - \overline{\Phi}_q\|_*^2. \end{aligned} \tag{19}$$

If we define $\delta_{\min} \coloneqq \min_v \|v\Phi_p - \overline{\Phi}_q\|_*^2$, we see that the minimum norm solution in Equation (19) corresponds to a solution to the constrained form of balancing weights with $\delta_2 = \delta_{\min}$. Importantly, no solution exists for $\delta_2 < \delta_{\min}$, and we must make the additional restriction that $\delta_2 \geq \delta_{\min}$. In particular, no solution exists for $\delta_2 = 0$ and we cannot achieve exact balance; that is, for all w, $w\Phi_p \neq \bar{\Phi}_q$.

As in the penalized form, the automatic form is strictly convex and a unique solution exists for $\delta_3 > 0$. When $\delta_3 = 0$ we choose the minimum norm solution: by duality this will be equivalent to the minimum norm solution to the penalized problem [see Bruns-Smith and Feller, 2022].

Note that for $d = \infty$, each "row" of Φ_p is a vector in a Hilbert space $\mathcal{H}$. To solve the balancing weights problem computationally, we need a closed-form solution to the Hilbert space norm $\|\cdot\|_{\mathcal{H}}$. For example, this is a tractable computation when $\mathcal{H}$ is an RKHS.

B.2 Equivalences from Section 2.4 when $d > n$

As we mention at the end of Section 2.4, the equivalences between OLS and balancing weights hold when $d > n$. We now state these results formally. Let $\delta \geq 0$ be the hyperparameter for the penalized form of balancing weights — as stated above, this is important to state explicitly, as the constrained form will not have a solution for all values of its hyperparameter. For hyperparameter $\delta > 0$, the solutions to ℓ_2 balancing weights and ridge regression are identical as in Equation (13) with no alterations; ridge regression works by default when $d > n$. On the other hand, when $\delta = 0$, there exist infinitely many solutions to the normal equations that define the solution to the OLS optimization problem. Since $(\Phi_p^T \Phi_p)$ is not invertible, Equation (13) does not apply directly. Instead, we introduce the minimum norm solution to OLS:

$$\begin{aligned} &\min_{\beta \in \mathbb{R}^d} \|\beta\|_2^2 \\ &\text{such that } \|\Phi_p\beta - Y_p\|_2^2 = \min_{\beta'} \|\Phi_p\beta' - Y_p\|_2^2. \end{aligned}$$

[22] Alternatively, we could follow Bartlett et al. [2020] and assume that, almost surely, the projection of Φ_p on the space orthogonal to any eigenvector of $\mathbb{E}[\Phi_p \Phi_p^T]$ spans a space of dimension n. But as our results are numerical this has no real advantage.

See Bartlett et al. [2020] for an extensive discussion of this optimization problem and its statistical properties as an OLS estimator. For $d > n$, the minimum norm solution is:

$$\hat{\beta}_{\text{OLS}} \coloneqq (\Phi_p^T \Phi_p)^\dagger \Phi_p^T Y_p = \Phi_p^T (\Phi_p \Phi_p^T)^{-1} Y_p,$$

where $A^\dagger$ denotes the pseudoinverse of a matrix A. Note that the definition holds in general.[23] In the low-dimensional setting in the main text, $(\Phi_p^T \Phi_p)$ is invertible, and so $(\Phi_p^T \Phi_p)^\dagger = (\Phi_p^T \Phi_p)^{-1}$. The second equality holds only when $d > n$.

A version of Equation (12) holds between the minimum norm ℓ_2 balancing weights and minimum norm OLS estimators. Because the minimum norm ℓ_2 balancing weights do not achieve exact balance, we change the notation from $\hat{w}_{\text{exact}}$ to $\hat{w}^0_{\ell_2}$. In this setting, $\|\cdot\|_* = \|\cdot\|_2$ and the minimum-norm balancing weights problem in Equation (19) is also a minimum norm linear regression, but of $\bar{\Phi}_q \in \mathbb{R}^d$ on $\Phi_p^T \in \mathbb{R}^{d \times n}$:

$$\hat{w}^0_{\ell_2} = \Phi_p^T (\Phi_p^T \Phi_p)^\dagger \bar{\Phi}_q = (\Phi_p \Phi_p^T)^{-1} \Phi_p \bar{\Phi}_q.$$

Therefore, Equation (12) holds by replacing the inverse with the pseudo-inverse:

$$\begin{aligned}
\hat{\mathbb{E}}_q[\Phi_q \hat{\beta}_{\text{ols}}] &= \hat{\mathbb{E}}_p[\hat{w}^0_{\ell_2} \circ Y_p] \\
\hat{\mathbb{E}}_q[\Phi_q \underbrace{(\Phi_p^T \Phi_p)^\dagger \Phi_p^T Y_p}_{\hat{\beta}_{\text{ols}}}] &= \hat{\mathbb{E}}_p[\underbrace{\bar{\Phi}_q (\Phi_p^T \Phi_p)^\dagger \Phi_p^T}_{\hat{w}^0_{\ell_2}} \circ Y_p], \\
\hat{\mathbb{E}}_q[\Phi_q \underbrace{\Phi_p^T (\Phi_p \Phi_p^T)^{-1} Y_p}_{\hat{\beta}_{\text{ols}}}] &= \hat{\mathbb{E}}_p[\underbrace{\bar{\Phi}_q \Phi_p^T (\Phi_p \Phi_p^T)^{-1}}_{\hat{w}^0_{\ell_2}} \circ Y_p].
\end{aligned}$$

B.3 Propositions 3.1 and 3.2 when $d > n$

The results in Propositions 3.1 and 3.2 apply to the setting where $d > n$ without any further alteration using the pseudo-inverse.

Proof of Proposition 3.1.

$$Y_p^T \Phi_p \hat{\theta}^\delta = Y_p^T \Phi_p \Phi_p^\dagger \Phi_p \hat{\theta}^\delta = Y_p^T \Phi_p (\Phi_p^T \Phi_p)^\dagger \Phi_p^T \Phi_p \hat{\theta}^\delta = \hat{\beta}_{\text{ols}} \hat{\Phi}_q,$$

where the first two equalities follow from the pseudoinverse identities $A = AA^\dagger A$ and $A^\dagger = (A^T A)^\dagger A^T$ for any matrix A. □

Likewise Proposition 3.2 holds exactly for $\hat{\beta}_{\text{ols}}$ defined with the pseudoinverse.

B.4 The RKHS Setting

The results for $d = \infty$ can be computed efficiently for reproducing kernel Hilbert spaces. Let $\mathcal{H}$ be a possibly-infinite-dimensional RKHS with kernel $\mathcal{K}$ and induced feature map via the representer theorem, $\phi : \mathcal{X} \to \mathcal{H}$ with $\phi(x) = \mathcal{K}(\cdot, x)$. Let $\|\cdot\|_{\mathcal{H}}$ denote the norm of $\mathcal{H}$. Let K_p be the matrix with entries $\mathcal{K}(x_i, x_j)$, where $x_i, x_j \in \mathcal{X}$ are the ith and jth entries of X_p. Then $\Phi_p \Phi_p^T = K_p$ is invertible.

We will write out the versions of the main results for $\mathcal{F} = \mathcal{H}$ to demonstrate how to compute the corresponding results for RKHSs even though $d = \infty$. Denote the solution to the regularized least squares problem in $\mathcal{H}$ with $\lambda \geq 0$:

$$\hat{f}^\delta \coloneqq \text{argmin}_{f \in \mathcal{H}} \|f(X_p) - Y_p\|_2^2 + \lambda \|f\|_{\mathcal{H}}^2.$$

[23] For example, when $\Phi_p \in \mathcal{H}^n$ for an infinite-dimensional Hilbert space $\mathcal{H}$, $(\Phi_p^T \Phi_p)^\dagger$ is guaranteed to exist, since it is bounded and has closed range.

This is equivalent to the following problem by the representer theorem:

$$\begin{aligned}\hat{\alpha}^{\delta}_{\mathcal{H}} &\coloneqq \operatorname{argmin}_{\alpha\in\mathbb{R}^n} \|K\alpha - Y_p\|_2^2 + \lambda\alpha K\alpha \\ &= (K_p + \lambda I)^{-1} Y_p.\end{aligned}$$

Let $K_{x,p} \in \mathbb{R}^n$ be the row vector with entries $\mathcal{K}(x, x_i)$ where x is an arbitrary element of $\mathcal{X}$ and x_i is the ith entry of X_p. Likewise let $K_{q,p}$ be the matrix with (i,j)th entry $\mathcal{K}(x_{qi}, x_{pj})$ where x_{qi} is the ith entry of X_q and x_{qj} is the jth entry of X_p. Then for any element $x \in \mathcal{X}$, $\hat{f}^{\delta}(x) = K_{x,p}\hat{\alpha}^{\delta}_{\mathcal{H}}$ and $\hat{f}^{\delta}(X_q) = K_{q,p}\hat{\alpha}^{\delta}_{\mathcal{H}}$.

Furthermore, define $\bar{K}_{q,p} \coloneqq \hat{\mathbb{E}}_q[K_{p,q}] \in \mathbb{R}^n$. Then similarly, for any solution $\hat{w}^{\delta}_{\mathcal{H}}$ to the penalized form of balancing weights with function class $\mathcal{F}$ and hyperparameter $\delta \geq 0$:

$$\hat{w}^{\delta}_{\mathcal{H}} = (K_p + \lambda I)^{-1} \bar{K}_{q,p}. \tag{20}$$

The proof follows from the closed-form of $\text{Imbalance}_{\mathcal{H}}(w)$, known as the Maximum Mean Discrepancy (MMD) [Gretton et al., 2012]; see, e.g., Hirshberg et al. [2019], Kallus [2020], Bruns-Smith and Feller [2022].

With these preliminaries, we immediately have the following equivalence from Hirshberg et al. [2019], which generalizes Section 2.4 to the RKHS case:

$$\begin{aligned}\hat{\mathbb{E}}_q[K_{q,p}\hat{\alpha}^{\delta}_{\mathcal{H}}] &= \hat{\mathbb{E}}_p[\hat{w}^{\delta}_{\mathcal{H}} \circ Y_p] \\ \hat{\mathbb{E}}_q[K_{q,p}\underbrace{(K_p + \delta I)Y_p}_{\hat{\alpha}^{\delta}_{\mathcal{H}}}] &= \hat{\mathbb{E}}_p[\underbrace{\bar{K}_{q,p}(K_p + \delta I)^{-1}}_{\hat{w}^{\delta}_{\ell_2}} \circ Y_p].\end{aligned} \tag{21}$$

Likewise, we have the following form for Proposition 3.1. Define $\hat{K}_{q,p} \coloneqq {\hat{w}^{\delta}_{\mathcal{H}}}^T K_p$. Then, for any $\delta \geq 0$:

$$\hat{\mathbb{E}}_p[\hat{w}^{\delta}_{\mathcal{H}} \circ Y_p] = \hat{\mathbb{E}}_q[\hat{K}_{q,p}\hat{\alpha}^{0}_{\mathcal{H}}].$$

The resulting expression for Proposition 3.2 is:

$$\hat{\mathbb{E}}_p[\hat{w}^{\delta}_{\mathcal{H}} \circ Y_p] + \hat{\mathbb{E}}_q\left[\left(K_{q,p} - \hat{K}^{\delta}_{q,p}\right)\hat{\alpha}^{\lambda}_{\mathcal{H}}\right] = \hat{\mathbb{E}}_q[K_{q,p}\hat{\alpha}_{\text{aug}}],$$

where the jth element of $\hat{\alpha}_{\text{aug}}$ is:

$$\begin{aligned}\hat{\alpha}_{\text{aug},j} &\coloneqq \left(1 - a^{\delta}_j\right)\hat{\alpha}^{\lambda}_{\mathcal{H},j} + a^{\delta}_j\hat{\alpha}^{0}_{\mathcal{H},j} \\ a^{\delta}_j &\coloneqq \frac{\widehat{\Delta}^{\delta}_j}{\Delta_j},\end{aligned}$$

where $\Delta_j = \bar{K}_{q,p,j} - \bar{K}_{p,j}$ and $\widehat{\Delta}^{\delta}_j = \hat{K}^{\delta}_{q,p,j} - \bar{K}_{p,j}$ with $\bar{K}_{p,j} \coloneqq \hat{\mathbb{E}}_p[K_p]$.

Identical versions for the RKHS setting apply to Section 3.2. These follow directly from the expressions above so we will omit repeating them explicitly. Importantly, equivalent versions for ℓ_∞ balancing in Section 3.3 do *not* follow immediately because an infinite dimensional vector space equipped with the ℓ_1 norm does not form a Hilbert space. We conjecture that such extensions could be constructed using the Reproducing Kernel Banach Space literature [Lin et al., 2022].

C Augmented ℓ_∞ Balancing With Correlated Features

We now consider the case where $\Phi_p^T\Phi_p$ is invertible but not diagonal. Here, both the Lasso and ℓ_∞ balancing weights lack a closed-form solution. Qualitatively, the Lasso coefficients still have a regularization path similar to the diagonal case. The coefficients start at OLS and shrink exactly to zero at different rates. However, due to the correlation between the features, each coefficient no longer linearly (or even monotonically) moves

toward zero. In ℓ_∞ balancing weights, we see identical behavior for the coefficients $\hat{\theta}$. These coefficients are sparse, with regularization paths that move toward zero (but not monotonically) exactly as in Lasso.

In the diagonal setting, the regularization path for the augmented estimator, a_j^δ was proportionate (element-wise) to $\hat{\theta}$. Therefore, because $\hat{\theta}$ was sparse, the augmented estimator coefficients defined in Proposition 3.2 was also sparse. For general design, the impact of $\hat{\theta}$ on the augmented coefficients is mediated by the empirical covariance matrix.

Lemma C.1. *Let $\hat{w}$ be the solution to balancing weights with hyperparameter δ for $\mathcal{F} = \{f(x) = \theta^T \phi(x) : \|\theta\| \leq r\}$, where $\|\cdot\|$ is any norm. Let $\hat{\theta}$ be the corresponding solution to the dual form of balancing weights, so that $\hat{w} = \Phi_p \hat{\theta}$. Then,*

$$\hat{\Phi}_q = \hat{\theta}^T \Sigma,$$

and

$$a_j^\delta = \frac{\hat{\theta}^T \Sigma_j}{\overline{\Phi}_{q,j}}.$$

In the case of ℓ_∞ balancing, since $\hat{\theta}$ is sparse, a_j^δ is a sparse combination of the elements of the jth column of the empirical covariance matrix Σ. But unless that column of Σ is itself sparse (typically a measure zero event), the resulting a_j^δ will not inherit sparsity from $\hat{\theta}$.

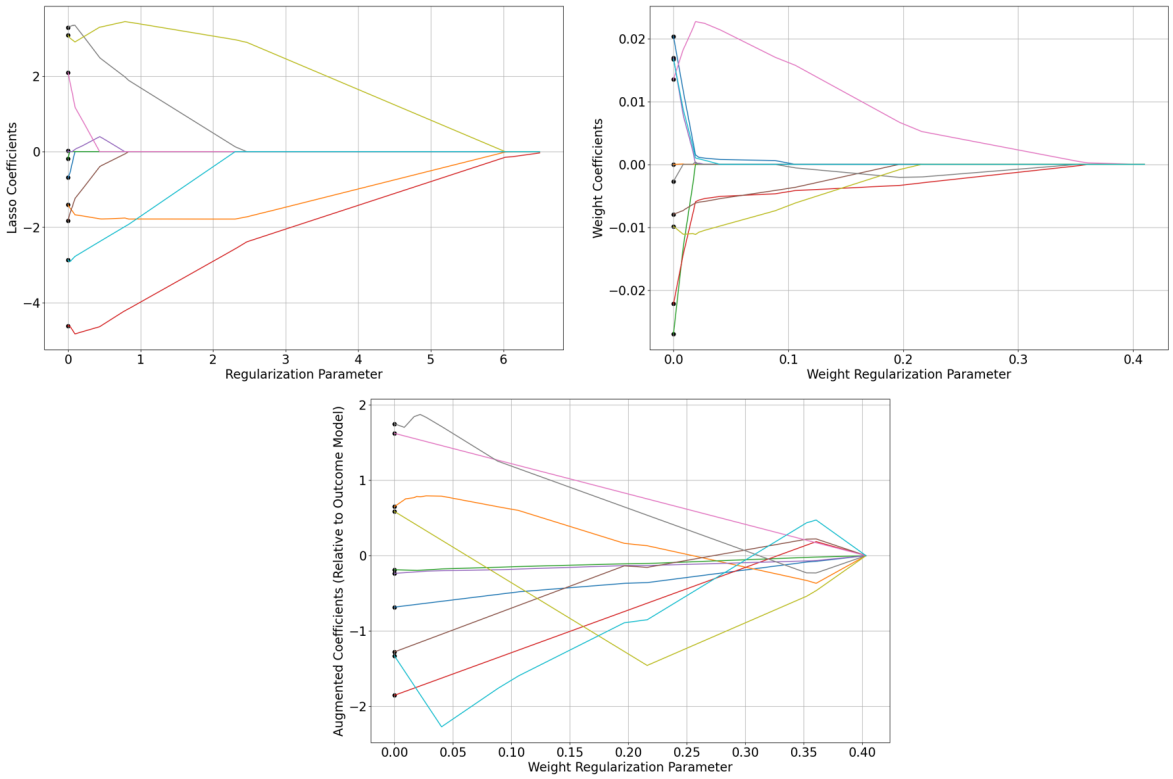


Figure 3: Regularization paths for ℓ_∞ balancing weights in the general design setting. For comparison, in the top left are the Lasso coefficients as function of the Lasso hyperparameter. In the top right are the ℓ_∞ weight coefficients as a function of the balancing weights hyperparameter. And at the bottom are the ℓ augmented coefficients relative to the outcome model, $\hat{\beta}_{\text{aug}} - \hat{\beta}^\lambda_{\text{reg}}$.

Figure 3 illustrates the regularization paths for simluated data where Φ_p is not orthonormal. The top left panel plots Lasso coefficients which go exactly to zero but not monotonically. The top right panel plots the weight coefficients for ℓ_∞ balancing, which have qualitatively similar behavior, but regularizing $\overline{\Phi}_q$ instead of $\hat{\beta}_{\text{ols}}$. The bottom panel plots the augmented coefficients relative to a baseline outcome model, $\hat{\beta}_{\text{aug}} - \hat{\beta}^{\lambda}_{\text{reg}}$. Unlike in Figure 2, the regularization path is not sparse and some individual coefficients display a high level of non-monotonicity. The regularization path follows the sparse path of the weight coefficients, but via inner product with the columns of the empirical covariance matrix. To see this visually, note that the augmented regularization path is piece-wise linear, where the knots occur exactly when one of the weight coefficients goes to zero.

D Beyond Linear: Differentiable Functions

In this section, we provide a very preliminary sketch of how the results might be extended to the case where $\mathcal{F}$ is non-linear but still differentiable in its parameters.

Let $\mathcal{F} = \{f(X, \theta) : \theta \in \mathbb{R}^d, \nabla_\theta f(X, \theta) \text{ exists}\}$. Then just like Proposition 3.1 relates any w that are linear in X to the OLS coefficients, we can relate any $w \in \mathcal{F}$ to the least squares regressor in the function class $\mathcal{F}$.

First, let θ_{LS} be the unregularized least squares regressor (where we choose the least norm θ to break ties):

$$\theta_{\text{LS}} \coloneqq \min_\theta \|Y_p - f(\theta, X_p)\|_2^2$$

We have the first-order condition:

$$\nabla_\theta f(\theta_{\text{LS}}, X_p)^T (Y_p - f(\theta_{\text{LS}}, X_p)) = 0$$

Now we can get a version of Proposition 2.1 for $w \in \mathcal{F}$ by considering the following taylor expansion:

$$w(X_p) \coloneqq f(\theta_w, X_p) \approx f(\theta_{\text{LS}}, X_p) + \nabla_\theta f(\theta_{\text{LS}}, X_p)(\theta_w - \theta_{\text{LS}})$$

In which case, applying the first-order condition above, we get:

$$w(X_p)^T y \approx \underbrace{w(X_p)^T f(\theta_{\text{LS}}, X_p)}_{\text{identical to 3.1}} + \underbrace{f(\theta_{\text{LS}}, X_p)^T (Y_p - f(\theta_{\text{LS}}, X_p))}_{\text{this term is zero in linear case}}.$$

E Necessary and Sufficient Conditions for Equivalence

In this section, we comment briefly on when, for an outcome model β and balancing weights w, does $\beta^T \overline{\Phi}_q = w^T y_p$. For any point estimate $\hat{\psi}$ there always exists $\beta \in \mathbb{R}^d$ and $w \in \mathbb{R}^n$ such that $\hat{\psi} = \beta^T \overline{\Phi}_q = w^T Y_p$. However, β and w are not guaranteed to have any desirable properties - β will not generally be a least squares regression model and w will not satisfy a balance property.

If we instead consider an outcome model $\hat{\beta}$ that is the solution to an empirical risk minimization problem with loss $\mathcal{L}$ and regularizer ρ, and weights $\hat{w}$ that solve the balancing problem with imbalance $\mathcal{B}$ and regularizer ζ; then we can ask, for which choices $\mathcal{L}, \mathcal{B}, \rho$, and ζ does $\hat{\beta}^T \overline{\Phi}_q = \hat{w}^T Y_p$ for all Φ_p, Φ_q, and Y_p?

We can start to answer this question by considering the necessary and sufficient conditions of the two optimization problems. Assume that $\mathcal{L}(\beta, \Phi_p, Y_p) + \rho(\beta)$ is strictly convex in β. In this case, there is a unique minimum and the KKT conditions are necessary and sufficient. For the optimization problem to be equivalent to a weighting estimator, the KKT conditions must imply $\hat{\beta} = SY_p$ for some $S \in \mathbb{R}^{d \times n}$. In other words, the estimator must be a linear smoother. Likewise, if $\mathcal{B}(w, \Phi_p, \Phi_q) + \rho(w)$ is strictly convex in w, then

the KKT conditions must imply $\hat{w} = \overline{\Phi}_q^T A$ for some $A \in \mathbb{R}^{d \times n}$. When $S = A$, then, $\hat{\beta}^T \overline{\Phi}_q = \hat{w}^T Y_p = \overline{\Phi}_q^T S Y_p$. Note that because the optimization problems do not share the inputs $\overline{\Phi}_q$ and Y_p, for this condition to be satisfied, S and A must be only a function of Φ_p.

When $\mathcal{L}$ is the least squares loss, then the broad class of (kernel) ridge regression and (kernel) ℓ_2 balancing weights are the only cases where the two procedures give equivalent estimators. In particular, for $\hat{\beta}$ to be linear in Y_p, ρ must be quadratic in β. The resulting optimization problem has a closed-form solution which precisely pins down the form of A in the balance weights problem. So the exact equivalence of (kernel) ridge regression and (kernel) ℓ_2 balancing weights is in some sense a knife-edge special case. This contrasts with our characterization in Proposition 3.1, which applies to any linear balancing weights.

Nonetheless, Section 5 of Lin and Han [2022] demonstrates that random forests — a class of linear smoothers in the sense that the final predictions can always be written $w^T Y_p$ for some w — are actually a consistent estimator of the Riesz representer when viewed as a weighting problem. Exactly what kind of finite sample balancing property (if any) these weights guarantee is an interesting avenue for future research that could open new connections between outcome models and estimators for the Riesz representer.

F Proofs