Augmented balancing weights as linear regression

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

Abstract

We provide a novel characterization of augmented balancing weights, also known as automatic debiased machine learning (AutoDML). These popular *doubly robust* or *de-biased machine learning* estimators combine outcome modeling with balancing weights — weights that achieve covariate balance directly in lieu of estimating and inverting the propensity score. When the outcome and weighting models are both linear in some (possibly infinite) basis, we show that the augmented estimator is equivalent to a single linear model with coefficients that combine the coefficients from the original outcome model coefficients and coefficients from an unpenalized ordinary least squares (OLS) fit on the same data. We see that, under certain choices of regularization parameters, the augmented estimator often collapses to the OLS estimator alone; this occurs for example in a re-analysis of the LaLonde (1986) dataset. We then extend these results to specific choices of outcome and weighting models. We first show that the augmented estimator that uses (kernel) ridge regression for both outcome and weighting models is equivalent to a single, undersmoothed (kernel) ridge regression. This holds numerically in finite samples and lays the groundwork for a novel analysis of undersmoothing and asymptotic rates of convergence. When the weighting model is instead lasso-penalized regression, we give closed-form expressions for special cases and demonstrate a "double selection" property. Our framework opens the black box on this increasingly popular class of estimators, bridges the gap between existing results on the semiparametric efficiency of undersmoothed and doubly robust estimators, and provides new insights into the performance of augmented balancing weights.

[To be read before The Royal Statistical Society at the Discussion Meeting on 'Augmented balancing weights as linear regression' to be held at the Society's 2025 annual conference in Edinburgh on Tuesday, 2 September 2025, the President, Sir John Aston, in the Chair]

1 Introduction

Combining outcome modeling and weighting, as in augmented inverse propensity score weighting (AIPW) and other doubly robust (DR) or double machine learning (DML) estimators, is a core strategy for estimating causal effects using observational data. A growing body of literature finds weights by solving a "balancing weights" optimization problem to estimate weights directly, rather than by first estimating the propensity score and then inverting. DR versions of these estimators are referred to by a number of terms, including *augmented balancing weights* (Athey et al., 2018; Hirshberg and Wager, 2021), *automatic debiased machine learning* (AutoDML; Chernozhukov et al., 2022d), and *generalized regression estimators* (GREG; Deville and Särndal, 1992); see Ben-Michael et al. (2021b) for a review. Moreover, this strategy has been applied to a wide range of linear estimands via the Riesz representation theorem (e.g., Hirshberg and Wager, 2021; Chernozhukov et al., 2022e). In this paper, we consider augmented balancing weights in which the estimators for both the outcome model and the balancing weights are based on penalized linear regressions in some possibly infinite basis; in addition to all high-dimensional linear models, this broad class includes popular nonparametric models such as kernel regression and certain forms of random forests and neural networks.

We first show that, somewhat surprisingly, augmenting any regularized linear outcome regression (the "base learner") with linear balancing weights is numerically equivalent to a single linear outcome regression applied to the target covariate profile. The resulting coefficients are an affine (and often convex) combination of the base learner model coefficients and unregularized OLS coefficients; the hyperparameter for the balancing weights estimator directly controls the regularization path defining the affine combination. In the extreme case where the weighting hyperparameter is set to zero — which we show can occur in practice — the entire procedure is equivalent to estimating a single, unregularized OLS regression.

We specialize these results to ridge and lasso regularization (ℓ_2 and ℓ_∞ balancing, respectively) and show that augmenting an outcome regression estimator with balancing weights generally corresponds to a form of *undersmoothing*. Most notably, we show that an augmented balancing weight estimator that uses (kernel) ridge regression for both outcome and weighting models — which we refer to as "double ridge" — collapses to a single, undersmoothed (kernel) ridge regression estimator.

We leverage these results to prove novel *statistical* results for double ridge estimators and to make progress towards practical hyperparameter tuning, which remains an open problem in this area. We first make explicit the connection between asymptotic results for double kernel ridge estimators (e.g., Singh, 2024) and prior results on optimal undersmoothing for a single kernel ridge outcome model (e.g., Mou et al., 2023), showing that the latter is also semiparametrically efficient. This generalizes the argument in Robins et al. (2007) that "OLS is doubly robust" to a much broader class of penalized parametric and non-parametric regression estimators. As a complementary analysis, we next adapt existing finite sample error analysis results for single ridge regression (Dobriban and Wager, 2018) to derive the finite-sample-exact bias and variance of double ridge estimators. Using these expressions, we can compute oracle hyperparameters for any given data-generating process.

Finally, we illustrate our results with several numerical examples. We first explore hyperparameter tuning for double ridge regression in an extensive simulation study on 36 data-generating processes, and compare three practical methods to the optimal hyperparameter computed using our finite sample analysis. Both asymptotic theory and our simulation results suggest equating the hyperparameters for the outcome and weighting models. We further caution against the naive application of hyperparameter tuning based solely on cross-validating the weighting model, forms of which have been suggested previously. This approach can lead to setting the weighting hyperparameter to exactly zero — and therefore recovering standard OLS — even in scenarios where OLS is far from optimal. We emphasize this point by applying our results to the

canonical LaLonde (1986) study, highlighting that researchers can inadvertently recover OLS in practice.

Broadly, our results provide important insights into the nexus of causal inference and machine learning. First, these results open the black box on the growing number of methods based on augmented balancing weights and AutoDML — methods that can sometimes be difficult to taxonomize or understand. We show that, under linearity, these estimators all share an underlying and very simple structure. Our results further highlight that estimation choices for augmented balancing weights can lead to potentially unexpected behavior. At a high level, as causal inference moves towards incorporating machine learning and automation, our work highlights how the traditional lines between weighting and regression-based approaches are becoming increasingly blurred.

Second, our results connect two approaches to "automate" semiparametric causal inference. AutoDML and related methods exploit the fact that we can estimate a Riesz representer without a closed form expression for a wide class of functionals. The estimated Riesz representer then augments a base learner by bias correcting a plug-in estimator of the functional. Older approaches, such as undersmoothing (Goldstein and Messer, 1992; Newey et al., 1998), twicing kernels (Newey et al., 2004), and sieve estimation (Newey, 1994; Shen, 1997), avoid estimation of the Riesz representer altogether, instead tuning the base learner regression fit such that an additional bias correction is not required. Achieving this optimal tuning in practice has long been a hurdle for the implementation of these methods. Subject to certain conditions, both approaches can yield estimators that are asymptotically efficient. We show that if all required tuning parameters are defined in terms of an ℓ_2-norm constraint, then these approaches can be numerically identical even in finite samples. We use these equivalences to make progress toward practical hyperparameter selection and find promising directions for new theoretical analysis.

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. In Section 3 we present our new numerical results, and in Sections 4 and 5 we cache out the implications for ℓ_2 and ℓ_∞ balancing weights specifically. Building on our numerical results, Section 6 explores both asymptotic and finite sample statistical results for kernel ridge regression. Section 7 illustrates our results with a simulation study and application to canonical data sets. Section 8 offers some other directions for future research. The appendix includes extensive additional technical discussion and extensions.

1.1 Related work

Balancing weights and AutoDML. With deep roots in survey calibration methods and the *generalized regression estimator* (GREG; see Deville and Särndal, 1992; Lumley et al., 2011; Gao et al., 2022), 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 balances features of the covariate distributions in the different treatment groups, with the aim of minimising the maximal design-conditional mean squared error of the treatment effect estimator. 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 is a special case, where "automatic" refers to the fact that we can estimate the Riesz representer without obtaining a closed form expression. Estimating the Riesz representer directly, under the assumption that it is linear in some basis, dates back at least to Robins et al. (2008); see also Robins et al. (2007). The corresponding augmented estimation framework has more recently come to be known as Automatic Debiased Machine Learning, or AutoDML; see, among others, Chernozhukov et al. (2022a), Chernozhukov et al. (2022b), Chernozhukov et al. (2022d), and Chernozhukov et al. (2022e). This approach has also been applied in a range of settings, including to corrupted data (Agarwal and Singh, 2021), to dynamic treatment regimes (Chernozhukov et al., 2022c), and to address noncompliance (Singh et al., 2022).

As we discuss in Appendix C.3, the AutoDML approach nearly always employs cross-fitting and is typically motivated by asymptotic properties rather than achieving minimax design-conditional mean squared error.

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); Chattopadhyay and Zubizarreta (2021), and between kernel ridge regression and forms of kernel weighting (Kallus, 2020; Hirshberg et al., 2019). We discuss these equivalences at length in Appendix A.5. Finally, as we discuss in Appendix D, there are close connections between balancing weights and Empirical Likelihood (Hellerstein and Imbens, 1999; Newey and Smith, 2004).

2 Problem setup and background

2.1 Setup and motivation

The core results in our paper are numeric equivalences for existing estimation procedures, and as such these results hold absent any causal assumptions or statistical model. Nonetheless, a primary motivation for this work is the task of estimating unobserved counterfactual means in causal inference, as well as estimating the broad class of linear functionals described in Chernozhukov et al. (2018b). We briefly review the corresponding setup, emphasizing that this is purely for interpretation.

2.1.1 Example: Estimating counterfactual means

Let X, Y, Z be random variables defined on $\mathcal{X}, \mathbb{R}, \mathcal{Z}$ with joint probability distribution p. To begin, consider the example of a binary treatment, $\mathcal{Z} = \{0, 1\}$ and covariates X. Define potential or counterfactual outcomes $Y(1)$ and $Y(0)$ under assignment to treatment and control, respectively. Under SUTVA (Rubin, 1980), we observe outcomes $Y = ZY(1) + (1 - Z)Y(0)$. To estimate the average treatment effect, $\mathbb{E}[Y(1) - Y(0)]$, we first estimate the means of the partially observed potential outcomes. We initially focus on estimating $\mathbb{E}[Y(1)]$; a symmetric argument holds for $\mathbb{E}[Y(0)]$.

Let $m(x, z) \coloneqq \mathbb{E}[Y \mid X = x, Z = z]$ be the *outcome model*, $e(x) \coloneqq \mathbb{P}[Z = 1 \mid X = x]$ be the *propensity score*, and $\alpha(x, z) = z/e(x)$ be the *inverse propensity score weights* (IPW). Under the additional assumptions of *conditional ignorability*, $Y(1) \perp\!\!\!\perp Z \mid X$, and *overlap*, $\mathbb{E}[\alpha(X, Z)^2] < \infty$, $\mathbb{E}[Y(1)]$ is identified by $\mathbb{E}[m(X, 1)]$, a linear functional of the observed data distribution.

There are three broad strategies for estimating $\mathbb{E}[Y(1)]$. First, the identifying functional above suggests estimating the outcome model, $m(x, 1)$ among those units with $Z = 1$, and plugging this into the *regression functional*, $\mathbb{E}[m(X, 1)]$. Second, the equality $\mathbb{E}[m(X, 1)] = \mathbb{E}[Z/e(X)Y] = \mathbb{E}[\alpha(X, Z)Y]$ suggests estimating the inverse propensity score weights, $\alpha(x, z) = z/e(x)$, and plugging these into the *weighting functional*. Finally, we can combine these two via the *doubly robust functional* (Robins et al., 1994):

$$\mathbb{E}[m(X, 1) + \alpha(X, Z)(Y - m(X, 1))].$$

This functional has the attractive property of being equal to $\mathbb{E}[m(X, 1)]$ even if either one of α or m is replaced with an arbitrary function of X and Z, hence the term "doubly robust." Doubly robust estimators have been studied extensively in semiparametric theory; note that $m(X, 1) + \alpha(X, Z)(Y - m(X, Z)) - \psi(m)$ coincides with the efficient influence function for $\psi(m)$ under a nonparametric model. See Chernozhukov et al. (2018a) and Kennedy (2022) for overviews of the active literature in causal inference and machine learning focused on estimating versions of this functional.

2.1.2 General class of functionals via the Riesz representer

Our results apply well beyond the example above. In particular, they apply to any functional of the form

$$\psi(m) = \mathbb{E}[h(X, Z, m)], \tag{1}$$

where Z a random variable with support $\mathcal{Z}$; and h is a real-valued, mean-squared continuous linear functional of m (Chernozhukov et al., 2018b; Hirshberg and Wager, 2021; Chernozhukov et al., 2022d). Following Chernozhukov et al. (2022d,e), we can generalize the weighting functional to this general class of estimands via the *Riesz representer*, which is a function $\alpha(X, Z) \in L_2(p)$ such that, for all square-integrable functions $f \in L_2(p)$:

$$\mathbb{E}[h(X, Z, f)] = \mathbb{E}[\alpha(X, Z) f(X, Z)]. \tag{2}$$

As in the counterfactual mean example, we can identify the more general target functional in (2) via the outcome regression functional in (1), via the Riesz representer functional in (2) with $f = m$, or via the doubly robust functional

$$\mathbb{E}[h(X, Z, m) + \alpha(X, Z)(Y - m(X, Z))]. \tag{3}$$

Estimators of this DR functional are *augmented* in the sense that they augment the "plug-in," "outcome regression," or "base learner" estimator of $\mathbb{E}[h(X, Z, m)]$ with appropriately weighted residuals; or, equivalently, augment the weighting estimator with an appropriate outcome regression. This is the class of estimators to which our results apply. As before, $h(X, Z, m) + \alpha(X, Z)(Y - m(X, Z)) - \psi(m)$ coincides with the efficient influence function for $\psi(m)$ under a nonparametric model. In future work we will explore whether we can extend our results to a different class of functionals that admit DR functional forms, first introduced by Robins et al. (2008), and to the superset of such functionals characterized by Rotnitzky et al. (2021).

2.2 Balancing weights: Background and general form

The core idea behind balancing weights is to estimate the Riesz representer directly — rather than via an analytic functional form (e.g., by estimating the propensity score and inverting it). As a result, balancing weights do not require a known analytic form for the Riesz representer (Chernozhukov et al., 2022e), are often more stable (Zubizarreta, 2015), and can offer improved control of finite sample covariate imbalance (Zhao, 2019). We briefly describe two primary motivations for this approach.

First, a central property of the Riesz representer is that the corresponding weights, $w(X, Z) = \alpha(X, Z)$, are the unique weights that satisfy the *population balance property* property in Equation (2) for all square-integrable functions $f \in L_2(p)$. For our target estimand $\psi(m)$ we only need to satisfy the condition in Equation (2) for the special case of $f = m$. If we are willing to assume that m lies in a model class $\mathcal{F} \subset L_2(p)$, then it suffices to balance functions in that class. This is achieved by minimizing the imbalance over $\mathcal{F}$:

$$\text{Imbalance}_{\mathcal{F}}(w) := \sup_{f \in \mathcal{F}} \Big\{ \mathbb{E}[w(X, Z) f(X, Z)] - \mathbb{E}[h(X, Z, f)] \Big\}. \tag{4}$$

As we discuss next, balancing weights minimize a (penalized) sample analog of Equation (4).

Alternatively, Chernozhukov et al. (2022d) consider finding weights f that minimize the mean-squared error for $\alpha(X, Z)$:

$$\min_{f \in \mathcal{F}} \left\{ \mathbb{E}\left[\left(f(X, Z) - \alpha(X, Z) \right)^2 \right] \right\}. \tag{5}$$

Automatic estimation of the Riesz representer, also known as *Riesz regression* (Chernozhukov et al., 2024), minimizes a sample analog of Equation (5). When $\mathcal{F}$ is convex, then up to choice of hyperparameters (see (6) below), the solutions to Equations (4) and (5) are equivalent.

2.3 Linear balancing weights

In this paper, we consider the special case in which outcome models are linear in some basis expansion of X and Z. This is an extremely broad class that encompasses linear and polynomial models of arbitrary functions of X and Z and with dimension possibly larger than the sample size, as well as non-parametric models such as reproducing kernel Hilbert spaces (RKHSs; Gretton et al., 2012), the Highly-Adaptive Lasso (Benkeser and Van Der Laan, 2016), the neural tangent kernel space of infinite-width neural networks (Jacot et al., 2018), and "honest" random forests (Agarwal et al., 2022). However, this class excludes models for m that are fundamentally non-linear in their parameters, like general neural networks or generalized linear models with a non-linear link function. We sketch a preliminary extension of our results to arbitrary nonlinear balancing weights in Appendix D.

Under linearity, the imbalance over all $f \in \mathcal{F}$ has a simple closed form. Because our results concern numeric equivalences, we will focus on the finite sample version of the linear balancing weights problem. Let $\mathcal{F} = \{f(x,z) = \theta^\top \phi(x,z) : \|\theta\| \leq 1\}$ where $\|\cdot\|$ can be any norm on $\mathbb{R}^d$. The general setup constrains $\|\theta\| \leq r$; we set $r = 1$ without loss of generality, which simplifies exposition below. Let $\|\cdot\|_*$ be the *dual norm* of $\|\cdot\|$; that is, $\|v\|_* \coloneqq \sup_{\|u\|\leq 1} u^\top v$. 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. Let X_p, Y_p, Z_p be n i.i.d. samples from the distribution p of the observed data. Define the feature map $\phi : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}^d$ and let $\phi_j : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ denote the mapping for the jth feature. Define $\Phi_p \coloneqq \phi(X_p, Z_p)$ and let $\Phi_q \coloneqq h(X_p, Z_p, \phi)$ denote the *target features*. We will write $\hat{\mathbb{E}}$ for sample averages; define $\bar{\Phi}_p \coloneqq \hat{\mathbb{E}}[\Phi_p]$ and $\bar{\Phi}_q \coloneqq \hat{\mathbb{E}}[\Phi_q]$. For exposition, 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 ($d > n$) setting.

In what follows we write w for the $1 \times n$ vector $w(\Phi_p)$, to highlight the fact that we will estimate w directly rather than as an explicit function of X or Φ_p. Using the derivation above, we can directly calculate the finite sample imbalance as:

$$\widehat{\text{Imbalance}}_{\mathcal{F}}(w) = \|\tfrac{1}{n} w\Phi_p - \bar{\Phi}_q\|_*.$$

Now we can write the penalized sample analog of balancing weights optimization problem in (4) equivalently as either:

$$\text{Penalized form:} \qquad \min_{w \in \mathbb{R}^n} \left\{ \|\tfrac{1}{n} w\Phi_p - \bar{\Phi}_q\|_*^2 + \delta_1 \|w\|_2^2 \right\}$$

$$\text{Constrained form:} \qquad \min_{w \in \mathbb{R}^n} \|w\|_2^2 \quad \text{such that } \|\tfrac{1}{n} w\Phi_p - \bar{\Phi}_q\|_* \leq \delta_2.$$

Furthermore, we can write the equivalent problem in (5) as:

$$\text{Riesz regression form:} \qquad \min_{\theta \in \mathbb{R}^d} \left\{ \tfrac{1}{n} \theta^\top (\Phi_p^\top \Phi_p)\theta - \tfrac{1}{n} 2\theta^\top \bar{\Phi}_q + \delta_3 \|\theta\| \right\}, \tag{6}$$

where we use the terminology "Riesz regression" from Chernozhukov et al. (2024). 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 Riesz regression form. As a result, for any norm $\|\cdot\|$, the penalized and constrained forms will always produce weights that are linear in Φ_p (see Ben-Michael et al., 2021b, Section 9). Therefore, since the three problems are equivalent, we typically use a generic δ to denote the regularization parameter, and will specify the particular form only if necessary. In Appendix A.2 we illustrate several concrete examples for this problem and in Appendix D we consider alternative dispersion parameters and discuss popular forms of balancing that constrain the weights to be non-negative.

Remark 1 (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 versions,* $\Phi_p - \bar{\Phi}_p$ *and* $\Phi_q - \bar{\Phi}_p$, *in the dual form of the balancing weights problem. This is also equivalent to adding a column of* 1*s to* Φ_p. *Appropriately accounting for this normalization, however, unnecessarily complicates the notation. Therefore, without loss of generality, we will assume that the features are centered throughout, that is,* $\bar{\Phi}_p = 0$.

Remark 2 (Equivalence with kernel ridge regression). *For the special case of* ℓ_2 *balancing (as in Appendix A.2) the balancing weights problem is numerically equivalent to directly estimating the conditional expectation* $\mathbb{E}[Y_p|\Phi_p]$ *via (kernel) ridge regression and applying the estimated coefficients to* $\bar{\Phi}_q$. *Moreover, the solution to the balancing weights problem has a closed form that is always linear in* $\bar{\Phi}_q$; *we provide further details in Appendix A.5. For exact balance with* $\delta = 0$, *the balancing weights problem is equivalent to fitting unregularized OLS; see, for example, Robins et al. (2007), Kline (2011), and Chattopadhyay et al. (2020).*

3 Novel equivalence results for (augmented) balancing weights and outcome regression models

Our first main result demonstrates that *any* linear balancing weights estimator is equivalent to applying OLS to the re-weighted features. Our second result provides a novel analysis of augmented balancing weights, demonstrating that augmenting any linear balancing weights estimator with a linear outcome regression estimator is equivalent to a plug-in estimator of a new linear model with coefficients that are a weighted combination of estimated OLS coefficients and the coefficients of the original linear outcome model.

3.1 Weighting alone

Our first result is that estimating the target estimand $\psi(m)$ with any linear balancing weights is equivalent to fitting OLS for the regression of Y_p on Φ_p and then applying those coefficients to the re-weighted target feature profile. The key idea for this result begins with the simple unregularized regression prediction for $\psi(m)$, $\bar{\Phi}_q\hat{\beta}_{\text{ols}}$.

Proposition 3.1. *Let* $\hat{w}^\delta := \hat{\theta}^\delta\Phi_p^\top$, $\hat{\theta}^\delta \in \mathbb{R}^d$, *be any linear balancing weights, with corresponding weighted features* $\hat{\Phi}_q^\delta := \frac{1}{n}\hat{w}^\delta\Phi_p$. *Let* $\hat{\beta}_{ols} = (\Phi_p^\top\Phi_p)^\dagger\Phi_p^\top Y_p$ *be the OLS coefficients of the regression of* Y_p *on* Φ_p. *Then:*

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

where $\widehat{\Delta}^\delta = \hat{\Phi}_q^\delta - \bar{\Phi}_p$ *is the mean feature shift implied by the balancing weights and where superscript* δ *indicates possible dependence on a hyperparameter. We have assumed without loss of generality that* $\bar{\Phi}_p = 0$, *but we sometimes use* $\hat{\Delta}$ *notation to demonstrate the role of mean feature shift in various expressions. We use the symbol* $\circ$ *to denote element-wise multiplication.*

Note that 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 $(\Phi_p^\top\Phi_p)^\dagger = (\Phi_p^\top\Phi_p)^{-1}$; we provide a proof for the general setting in Appendix B.3. In Appendix D, we extend Proposition 3.1 to non-linear balancing weights, including those with a non-negativity constraint.

We can interpret this result via a contrast with standard regularization. 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 instead keeps $\hat{\beta}_{\text{ols}}$ fixed and regularizes the target feature distribution by penalizing the implied feature shift, $\widehat{\Delta}^\delta = \hat{\Phi}_q^\delta - \bar{\Phi}_p$.

We emphasize that this is a new and quite general result. As we discuss in Appendix A.5, it has been shown previously that for exact balancing weights, $\hat{\mathbb{E}}[\hat{w}_{\text{exact}}Y_p] = \bar{\Phi}_q\hat{\beta}_{\text{ols}}$. However, Proposition 3.1 holds for any

weights of the form $w = \theta\Phi_p^\top$ with arbitrary $\theta \in \mathbb{R}^d$. In Sections 4 and 5, we consider the particular form of $\hat{\Phi}_q^\delta$ for ℓ_2 and ℓ_∞ balancing, respectively.

3.2 Augmented balancing weights

We can immediately extend this to augmented balancing weights, which regularize *both* the coefficients and the feature shift. Let $\hat{\beta}^\lambda_{\text{reg}}$ 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 (e.g., estimated by regularized least squares). We consider augmenting $\hat{\mathbb{E}}\left[\hat{w}^\delta \circ Y_p\right]$ with $\hat{\beta}^\lambda_{\text{reg}}$ using the doubly robust functional representation in Equation (3). The augmented estimator is:

$$\hat{\mathbb{E}}[\Phi_q\hat{\beta}^\lambda_{\text{reg}}] + \hat{\mathbb{E}}[\hat{w}^\delta \circ (Y_p - \Phi_p\hat{\beta}^\lambda_{\text{reg}})] = \hat{\mathbb{E}}[\hat{w}^\delta \circ Y_p] + \hat{\mathbb{E}}\left[\left(\Phi_q - \hat{\Phi}_q^\delta\right)\hat{\beta}^\lambda_{\text{reg}}\right]. \tag{7}$$

Many recently proposed estimators have this form; see e.g., Athey et al. (2018); Ben-Michael et al. (2021b). 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. (2022d) or by applying an appropriate projection as in Hirshberg and Wager (2021).

We apply Proposition 3.1 to the first term of the right-hand side of (7) to yield the following result. As this result is purely numerical, it applies to arbitrary vectors $\hat{\beta}^\lambda_{\text{reg}} \in \mathbb{R}^d$, but substantively we think of $\hat{\beta}^\lambda_{\text{reg}}$ as the estimated coefficients from an outcome model.

Proposition 3.2. *For any $\hat{\beta}^\lambda_{reg} \in \mathbb{R}^d$, and any linear balancing weights estimator with estimated coefficients $\hat{\theta}^\delta \in \mathbb{R}^d$, and with $\hat{w}^\delta \coloneqq \hat{\theta}^\delta\Phi_p^\top$ and $\hat{\Phi}_q^\delta \coloneqq \frac{1}{n}\hat{w}^\delta\Phi_p$, the resulting augmented estimator*

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

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

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

where $\Delta_j = \overline{\Phi}_{q,j} - \overline{\Phi}_{p,j}$ is the observed mean feature shift for feature j; and $\widehat{\Delta}_j^\delta = \hat{\Phi}^\delta_{q,j} - \overline{\Phi}_{p,j}$ is the feature shift for feature j implied by the balancing weights model. Finally, $a^\delta \in [0,1]^d$ when the covariance matrix is diagonal, $(\Phi_p^\top\Phi_p) = diag(\sigma_1^2, \sigma_2^2, ..., \sigma_d^2)$, with $\sigma_j^2 > 0$.

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. (The coefficients are additionally *convex* combinations of $\hat{\beta}^\lambda_{\text{reg}}$ and $\hat{\beta}_{\text{ols}}$ when the covariance matrix is diagonal.) In Sections 4 and 5 below, we analyze some of the properties of the augmented estimator for ℓ_2 and ℓ_∞ balancing weights problems respectively.

The regularization parameter for the balancing weights problem, δ, parameterizes the path between $\hat{\beta}^\lambda_{\text{reg}}$ and $\hat{\beta}_{\text{ols}}$. To see this, consider the cases where $\delta \to 0$ and $\delta \to \infty$. As $\delta \to 0$ the balancing weights problem prioritizes minimizing balance over controlling variance, and $\widehat{\Delta}_j^\delta \to \Delta_j$ for all j. (Recall that we assume $\overline{\Phi}_{p,j} = 0$ 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 $\hat{\Phi}_q^\delta \to \overline{\Phi}_{q,j}$.) In this case,

$a_j^\delta = \widehat{\Delta}_j^\delta / \Delta_j \to 1$, and the weights fully "de-bias" the original outcome model by recovering unregularized regression, $\hat{\beta}_{\text{aug}} \to \hat{\beta}_{\text{ols}}$. In Section 7.2, we will see that when chosen by cross-validation, δ sometimes equals exactly 0 in applied problems; thus even when $\hat{\beta}^\lambda_{\text{reg}}$ is a sophisticated regularized estimator, the final augmented point estimate can nonetheless be numerically equivalent to the simple OLS plug-in estimate. Conversely, as $\delta \to \infty$, the balancing weights problem prioritizes controlling variance, leading to uniform weights and $\widehat{\Delta}_j \to 0$. In this case, $a_j^\delta = \widehat{\Delta}_j^\delta / \Delta_j \to 0$, the weighting model does very little, and $\hat{\beta}_{\text{aug}} \to \hat{\beta}^\lambda_{\text{reg}}$.

It is also instructive to consider two other extremes: unregularized outcome model and unregularized balancing weights. First, consider the special case of fitting an unregularized linear regression outcome model, i.e., $\hat{\beta}^\lambda_{\text{reg}} = \hat{\beta}_{\text{ols}}$. 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 $\hat{\theta}^\delta \in \mathbb{R}^d$. Thus, OLS augmented by *any* choice of linear balancing weights collapses to OLS alone. Equivalently, we can view OLS alone as an augmented estimator that combines an OLS base learner with linear 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 $\hat{\beta}^\lambda_{\text{reg}} \in \mathbb{R}^d$ be arbitrary coefficients. Then from the balance condition, $\hat{\Phi}_q = \overline{\Phi}_q$, $a_j^\delta = 1$ 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, 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.

Finally, before we turn to new results for ℓ_2 and ℓ_∞ balancing, we briefly comment on several points that are discussed in more detail in the Appendix.

Remark 3 (Sample splitting)**.** *Sample splitting is a common technique in the AutoDML literature especially, in which we only apply the outcome and weighting models to data points not used for estimation; see, for example, Newey and Robins (2018); Chernozhukov et al. (2022d). Since Proposition 3.2 holds for arbitrary vectors $\hat{\beta}^\lambda_{reg}$ and $\hat{\theta}^\delta$, the results still hold under cross-fitting. See Appendix C for an extended discussion.*

Remark 4 (Infinite dimensional setting)**.** *While we emphasize the linear, low-dimensional setting where $\Phi_p^\top \Phi_p$ 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.*

Remark 5 (Nonlinear balancing weights)**.** *A rich tradition in survey statistics (e.g., Deville and Särndal, 1992), machine learning (e.g., Menon and Ong, 2016), and causal inference (e.g., Vermeulen and Vansteelandt, 2015; Zhao, 2019; Tan, 2020) focuses on* non-linear *balancing weights, such as when the weights correspond to a specific* link function $g(\cdot)$ *applied to the linear predictor, $\hat{w} = g(\hat{\theta}\Phi_p^\top)$, or, equivalently, when the balancing weights problem penalizes an alternative dispersion penalty. In Appendix D, we briefly consider extending Proposition 3.1 to nonlinear weights and show that the nonlinearity introduces an additional approximation error. A more thorough extension is a promising direction for future research.*

Remark 6 (Non-negative weights)**.** *A common modification of the (minimum variance) balancing weights problem is to constrain the estimated weights to be non-negative or on the simplex; examples include Stable Balancing Weights (Zubizarreta, 2015) and the Synthetic Control Method (Abadie et al., 2010), as well as their augmented analogues (Athey et al., 2018; Ben-Michael et al., 2021c). 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 sometimes aid interpretability (Robins et al., 2007). In Appendix D.2, we extend Proposition 3.1 and show that restricting weights to be non-negative is equivalent to sample trimming. In particular, let $\hat{w}_+^\delta$ be the estimated non-negative weights and $\hat{\beta}^+_{ols}$ be the OLS coefficient of the regression of Y_p on Φ_p, but restricted to units with positive weight. Then, Proposition*

3.1 continues to hold, but with $\hat{\beta}^{+}_{ols}$ in place of the unrestricted $\hat{\beta}_{ols}$: $\hat{\mathbb{E}}\left[\hat{w}^{\delta}_{+} \circ Y_p\right] = \hat{\Phi}^{\delta}_{q}\hat{\beta}^{+}_{ols}$. See Arbour and Feller (2024) for additional discussion of the simplex constraint.

Remark 7 (Bilinear form)**.** *As pointed out by a reviewer, (many of) the functionals we consider can be written as a bilinear form $\alpha^T\Sigma\beta$ where β is the coefficient for the outcome model, α is the coefficient for the Riesz representer and Σ is the some weighted population Gram matrix (Robins et al., 2008); for $E[Y(1)]$, it would be $E[Z\phi(X)\phi(X)^T]$. Proposition 3.2 suggests that β can be estimated using the methods we discuss here, and moreover that the aggregation weights would then be entangled with Σ or α. Understanding whether this could be used to then motivate new estimators is an interesting topic for future work.*

4 Augmented ℓ_2 Balancing Weights

In this section, we study ℓ_2 balancing weights estimators, 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). We first show that the regularization path a^{δ}_{j} from Proposition 3.2 follows typical ridge regression shrinkage, with a smooth decay. Moreover, augmenting with ℓ_2 balancing weights is equivalent to boosting with ridge regression, and always overfits relative to the unaugmented outcome model alone. We then show that when the outcome model used to augment ℓ_2 balancing weights is also a ridge regression (which we refer to as "double ridge"), the augmented estimator is itself equivalent to a single, generalized ridge regression, albeit undersmoothed relative to the base learner. These results extend immediately to the RKHS setting of "double kernel ridge" estimation, combining kernel balancing weights and kernel ridge regression. In Section 6, we show the implications of these numeric results for undersmoothing in the statistical sense.

While the following results hold for arbitrary covariance matrices, in the main text we simplify the presentation by assuming that $\Phi_p^\top\Phi_p$ is diagonal; that is, $(\Phi_p^\top\Phi_p) = \text{diag}(\sigma_1^2, \sigma_2^2, ..., \sigma_d^2)$, with $\sigma_j^2 > 0$. We show that this is without loss of generality for ℓ_2 balancing in Appendix E.

4.1 General linear outcome model

Following Remark 2 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 closed form 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.

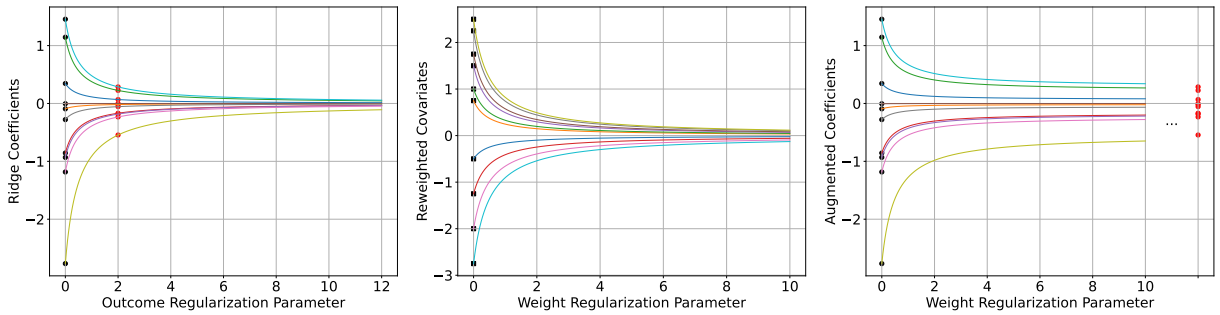
Proposition 4.1. *Let $\hat{w}^{\delta}_{\ell_2}$ be (penalized) linear balancing weights with regularization parameter δ and $\mathcal{F} = \{f(x) = \theta^\top\phi(x) : \|\theta\|_2 \leq 1\}$. Then $\frac{1}{n}\hat{w}^{\delta}_{\ell_2} = \overline{\Phi}_q(\Phi_p^\top\Phi_p + \delta I)^{-1}\Phi_p^\top$. Therefore, the augmented ℓ_2 balancing weights estimator with outcome model $\hat{\beta}^{\lambda}_{reg} \in \mathbb{R}^d$ has the form*

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

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

$$\begin{aligned}\hat{\beta}_{\ell_2,j} &\coloneqq \left(1 - a^{\delta}_{j}\right)\hat{\beta}^{\lambda}_{reg,j} + a^{\delta}_{j}\hat{\beta}_{ols,j} \\ a^{\delta}_{j} &\coloneqq \frac{\sigma_j^2}{\sigma_j^2 + \delta}.\end{aligned} \tag{8}$$

In this case, the a^{δ}_{j} are exactly equal to the standard regularization path of ridge regression. To see this,



(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 $\lambda = 2$ illustrate the coefficients at a plausible hyperparameter value, $\hat{\beta}^{2}_{\text{reg}}$. 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, $\hat{\beta}_{\ell_2}$ 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 $\hat{\beta}^{2}_{\text{reg}}$. All three regularization paths have essentially identical qualitative behavior.

recall that ridge regression with penalty δ shrinks the $\hat{\beta}_{\text{ols}}$ coefficients as follows:

$$\hat{\beta}^{\delta}_{\text{ridge},j} = \left(\frac{\sigma_j^2}{\sigma_j^2 + \delta}\right)\hat{\beta}_{\text{ols},j} = a_j^{\delta}\hat{\beta}_{\text{ols},j}. \tag{9}$$

This is identical to the expression in (8) but with $\hat{\beta}^{\lambda}_{\text{reg}}$ set to 0: Ridge regression shrinks $\hat{\beta}_{\text{ols}}$ towards 0 with regularization path a_j^{δ}, while ℓ_2 augmenting shrinks $\hat{\beta}_{\text{ols}}$ towards $\hat{\beta}^{\lambda}_{\text{reg}}$ with the same regularization path.

As an illustration, the right panel of Figure 1 shows $\hat{\beta}_{\ell_2}$ (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 $a_j^0 = 1$ 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.

We can also view $\hat{\beta}_{\ell_2}$ as the output of a single iteration of a ridge boosting procedure, fit using Y_p and Φ_p alone. See Bühlmann and Yu (2003) and Park et al. (2009) for detailed discussion; Newey et al. (2004) make a similar connection in the context of twicing kernels.

Proposition 4.2. *Let $\check{Y}_p = Y_p - \Phi_p\hat{\beta}^{\lambda}_{reg}$ be the residuals from the base learner. Let $\hat{\beta}^{\delta}_{boost}$ be the coefficients from the ridge regression of $\check{Y}_p$ on Φ_p with hyperparameter δ. Then, $\hat{\beta}_{\ell_2} = \hat{\beta}^{\lambda}_{reg} + \hat{\beta}^{\delta}_{boost}$, and $\|Y_p - \Phi_p\hat{\beta}_{\ell_2}\|_2^2 \leq \|Y_p - \Phi_p\hat{\beta}^{\lambda}_{reg}\|_2^2$.*

So for a fixed δ, the augmented ℓ_2 balancing estimator is equivalent to estimating a new outcome model coefficient estimator $\hat{\beta}_{\ell_2}$ that *overfits* relative to $\hat{\beta}^{\lambda}_{\text{reg}}$ (in the sense of having smaller in-sample training error), and then applying that model to Φ_q.

Surprisingly — and in contrast to the general result in Proposition 3.2 — the augmented coefficients $\hat{\beta}_{\ell_2}$ are the same for *every* target covariate profile Φ_q. To see this, note that Proposition 4.1 shows that ℓ_2 balancing weights are always linear in $\overline{\Phi}_q$. Therefore, the corresponding regularization path a_j^{δ} does not depend on the target profile Φ_q; it depends only on δ and the source distribution variances σ_j^2. This property is closely related to *universal adaptability* in the computer science literature on multi-group fairness (Kim et al.,

2022). The particular Φ_q may nonetheless impact the choice of δ in hyperparameter selection, e.g., via cross-validating imbalance, which in turn influences the degree of overfitting; we do find this to be the case theoretically in Section 6.2.

4.2 Ridge regression outcome model

Proposition 4.1 holds for arbitrary linear outcome model coefficient estimators $\hat{\beta}^{\lambda}_{\text{reg}} \in \mathbb{R}^d$; we now state the corresponding result for a "double ridge" estimator, where the base learner outcome model is itself fit via ridge regression. The key takeaway is that the implied augmented coefficients are *undersmoothed* relative to the base learner ridge coefficients.

For this section, we will consider the following generalized ridge regression, sometimes known as "adaptive" ridge regression (Grandvalet, 1998). Let $\Lambda \in \mathbb{R}^{d\times d}$ be a diagonal matrix with jth diagonal entry $\lambda_j \geq 0$. Then the generalized ridge coefficients are:

$$\begin{aligned}\hat{\beta}^{\Lambda}_{\text{ridge}} &\coloneqq \underset{\beta\in\mathbb{R}^d}{\text{argmin}} \|\Phi_p\beta - Y_p\|_2^2 + \beta^\top\Lambda\beta \\ &= (\Phi_p^\top\Phi_p + \Lambda)^{-1}\Phi_p^\top Y_p.\end{aligned}$$

Standard ridge regression is the special case where the λ_j all take the same value and so $\Lambda = \lambda I$. As above, the generalized ridge coefficients can be rewritten as shrinking the OLS coefficients:

$$\hat{\beta}^{\Lambda}_{\text{ridge},j} = \left(\frac{\sigma_j^2}{\sigma_j^2+\lambda_j}\right)\hat{\beta}_{\text{ols},j}. \tag{10}$$

We now demonstrate that the augmented ℓ_2 balancing weights estimator with base learner $\hat{\beta}^{\Lambda}_{\text{ridge}}$ is equivalent to a plug-in estimator using generalized ridge with *smaller* hyperparameters, $\hat{\beta}^{\Gamma}_{\text{ridge}}$, where Γ is a diagonal matrix with jth diagonal entry $\gamma_j \in [0, \lambda_j]$.

Proposition 4.3. *Let $\hat{\beta}^{\Lambda}_{ridge}$ denote the coefficients of a generalized ridge regression of Y_p on Φ_p with hyperparameters Λ, and let $\hat{w}^{\delta}_{\ell_2}$ denote ℓ_2 balancing weights with hyperparameter δ defined in Section 2.3. Define the diagonal matrix Γ with jth diagonal entry:*

$$\gamma_j \coloneqq \frac{\delta\lambda_j}{\sigma_j^2+\lambda_j+\delta} \leq \lambda_j.$$

Then:

$$\hat{\mathbb{E}}[\Phi_q\hat{\beta}^{\Lambda}_{ridge}] + \hat{\mathbb{E}}[\hat{w}^{\delta}_{\ell_2}(Y_p - \Phi_p\hat{\beta}^{\Lambda}_{ridge})] = \hat{\mathbb{E}}[\Phi_q\hat{\beta}^{\Gamma}_{ridge}].$$

Furthermore, $\hat{\beta}^{\Gamma}_{ridge}$ are standard ridge regression coefficients (i.e., γ_j is a constant for all j) when $\lambda_j = \lambda$ and $\sigma_j = \sigma$ for all j.

The same result holds for kernel ridge regression; see Appendix B.4.

In this setting, augmenting with balancing weights is equivalent to undersmoothing the original outcome model fit. In particular, we can use the expansion in Equation (10) to see the undersmoothing in $\hat{\beta}^{\Gamma}_{\text{ridge}}$ explicitly:

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

where the first term is the shrinkage from the original generalized ridge model alone, and the second term is due to augmenting with ℓ_2 balancing weights. Importantly, the second term is in $[1, \frac{\sigma_j^2+\lambda_j}{\sigma_j^2}]$ and therefore partially reverses the shrinkage of the original estimate. In Section 6.1, we connect this to undersmoothing in the statistical sense.

5 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., 2022d). In the main text, we consider the special case where the covariance matrix $\Phi_p^\top \Phi_p$ is diagonal; that is, $(\Phi_p^\top \Phi_p) = \text{diag}(\sigma_1^2, \sigma_2^2, ..., \sigma_d^2)$, with $\sigma_j^2 > 0$. Unlike with ℓ_2 balancing, this is no longer without loss of generality. We discuss this general case in Appendix E.3.

For diagonal covariance, we first show that ℓ_∞ balancing has a closed form: it is equivalent to applying a soft-thresholding operator to the feature shift from $\overline{\Phi}_p$ to $\overline{\Phi}_q$. We then write the resulting augmented estimator as applying coefficients $\hat{\beta}_{\ell_\infty}$ to Φ_q and show that $\hat{\beta}_{\ell_\infty}$ is a sparse, element-wise convex combination of the base learner coefficients and OLS coefficients. 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), where $\hat{\beta}_{\ell_\infty}$ inherits the non-zero coefficients of both the base learner and the weighting model. This is a form of undersmoothing in the ℓ_0 "norm," in the sense that $\hat{\beta}_{\ell_\infty}$ always has at least as many non-zero coefficients as the base learner, $\hat{\beta}_{\text{reg}}$.

5.1 Weighting alone

We first define the soft-thresholding operator and show 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 5.1 (ℓ_∞ Balancing)**.** *If $\Phi_p^\top \Phi_p$ is diagonal, the solution $w_{\ell_\infty}^\delta$ to the ℓ_∞ optimization problem (3) is:*

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

where $\Delta = \overline{\Phi}_q - \overline{\Phi}_p$, where we include $\overline{\Phi}_p$ (equal to 0 by assumption) to emphasize the dependence on feature shift, and with corresponding reweighted features, $\hat{\Phi}_q^\delta = \overline{\Phi}_p + \mathcal{T}_\delta(\overline{\Phi}_q - \overline{\Phi}_p)$.

For intuition, compare the (un-augmented) ℓ_∞ balancing weights estimator to the lasso-based coefficient estimates (Hastie et al., 2009):

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

where we simplify $\hat{\Phi}_q^\delta$ 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 implied feature shift to the target features.

5.2 General linear outcome model

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

Proposition 5.2. *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,*

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

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

The augmented coefficients $\hat{\beta}_{\ell_\infty}$ are an element-wise convex combination of $\hat{\beta}^{\lambda}_{\text{reg}}$ and $\hat{\beta}_{\text{ols}}$. For features where the mean feature shift Δ_j is small (relative to δ), $\hat{\beta}_{\ell_\infty}$ is equivalent to the base learner coefficient $\hat{\beta}^{\lambda}_{\text{reg}}$. The remaining coefficients are interpolated linearly toward the $\hat{\beta}_{\text{ols}}$ coefficients.

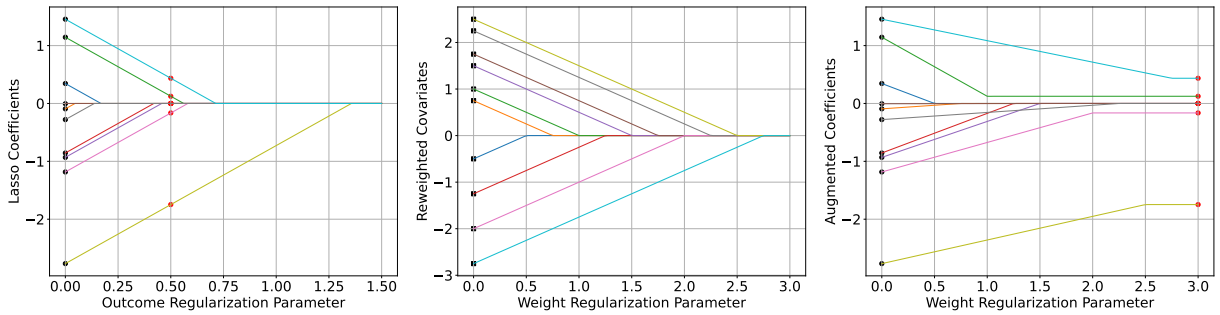
Figure 2 summarizes these results and their implications for the augmented estimator. As with Figure 1, we generate simple simulated data with $d = 10$. In the left panel, we plot the coefficients from lasso regression of Y_p on Φ_p as a function of the lasso regularization parameter. The regularization 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 reweighted covariates using 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 reweighted covariates shrink linearly to exactly zero, just as in lasso. The right panel plots coefficients for the augmented estimator that combines a baseline outcome model fit $\hat{\beta}^{\lambda}_{\text{reg}}$ with ℓ_∞ balancing weights. The lines correspond to $\hat{\beta}_{\ell_\infty}$ as defined in Proposition 5.2. The regularization path begins at the black dots, where $\hat{\beta}_{\ell_\infty} = \hat{\beta}_{\text{ols}}$, 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.

5.3 Lasso outcome model

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

Proposition 5.3 (Double Selection)**.** *Let $\hat{\beta}^{\lambda}_{lasso}$ denote the coefficients of lasso regression of Y_p on Φ_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 5.1. Let I_δ denote the non-zero entries of the reweighted covariates $\hat{\Phi}_q$. Assume that $\hat{\beta}_{ols}$ is dense. Then the indices of the non-zero entries of the augmented coefficients $\hat{\beta}_{\ell_\infty}$ are $I_{aug} = I_\lambda \cup I_\delta$.*

The lasso coefficients have a sparsity pattern generated by soft-thresholding the OLS coefficients. The augmented estimator then shrinks from OLS toward $\hat{\beta}^{\lambda}_{\text{reg}}$ by soft-thresholding the implied feature shift to the target features. As a result, wherever the lasso coefficients are non-zero *or* the weight coefficients are non-zero, the final augmented coefficients are also non-zero. The "included coefficients" for the final estimator are then the union of the coefficients included in either individual model. Therefore, augmenting a lasso outcome model with ℓ_∞ balancing also exhibits a form of undersmoothing in the ℓ_0 "norm", $\|\hat{\beta}_{\ell_\infty}\|_0$, in the sense that there are always at least as many non-zero coefficients as for the unaugmented lasso outcome model. However, this will not correspond to undersmoothing the base learner in the traditional sense, because in general there will not exist a lasso hyperparameter λ that will produce sparsity pattern I_{aug}.



(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 $\lambda = 0.5$ illustrate the coefficients at a plausible hyperparameter value, $\hat{\beta}^{0.5}_{\text{reg}}$. 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, $\hat{\beta}_{\ell_\infty}$ 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 $\hat{\beta}^{0.5}_{\text{reg}}$. All three regularization paths show the typical lasso "soft thresholding" behavior. The regularization path for the augmented estimator also shows "double selection" behavior.

As noted by, for example, Tang et al. (2023), the double selection estimator may suffer from imprecision due to adjustment for covariates that are associated with treatment but not outcome. One could in principle remove covariates that are only predictive of the treatment, but this can jeopardize statistical inference. See Moosavi et al. (2023) for further discussion on this trade-off.

6 Kernel Ridge Regression: Asymptotic and Finite Sample Analysis

The results above are *numerical*: they hold without any statistical or causal assumptions. However, the connection between augmented estimators and outcome models also presents *statistical* insights that we discuss here. In particular, we leverage the numerical result that double (kernel) ridge regression — which uses ridge regression for fitting both the outcome and weighting models — is equivalent to a single, undersmoothed outcome ridge regression plug-in estimator.

First, we consider an asymptotic analysis in Section 6.1: we use this equivalence to make explicit the connection between asymptotic results for augmented balancing weights with kernel ridge regression and prior results on optimal undersmoothing of a kernel ridge plug-in estimator. As a result, optimally undersmoothed kernel ridge regression inherits guarantees from augmented ridge regression. An implication is that we can generalize the insight from Robins et al. (2007) that "OLS is doubly robust" to a wider class of non-parametric estimators. This equivalence also suggests an appropriate hyperparameter scheme when the outcome regression is an element of an RKHS.

Second, we consider a finite sample analysis in Section 6.2: we use this equivalence to derive the finite-sample design-conditional mean squared error of augmented kernel ridge regression. We then use this expression to characterize finite-sample-optimal hyperparameter tuning. We turn to hyperparameter tuning in practice in the next section.

6.1 Asymptotic Results

We now use our results in Proposition 4.3 to make explicit the connection between two otherwise distinct sets of asymptotic results. First, Wong and Chan (2018) and Singh (2024) argue that double kernel ridge regression can deliver $\sqrt{n}$-consistent estimation of functionals in certain scenarios. Wong and Chan (2018) also proposes an optimally undersmoothed ℓ_2 balancing weights estimator. Separately, Hirshberg et al. (2019) and Mou et al. (2023) propose optimally undersmoothed (single) kernel ridge outcome regression. Since, as we have shown in Proposition 4.3 (see also Remark 2), these three procedures are equivalent, we can connect these results and show that plug-in estimators based on optimally undersmoothed kernel ridge regression or ℓ_2 balancing weights can be $\sqrt{n}$-consistent. Moveover, results on RKHSs suggest a simple heuristic for hyperparameter choice. We give the high-level argument here and defer additional technical details to Appendix L.

Assume that the outcome model, $m(x, z) \coloneqq \mathbb{E}[Y \mid X = x, Z = z]$, belongs to an RKHS $\mathcal{H}$ with kernel k, and that we observe n iid samples of (x_i, y_i, z_i) from p. Define $K \in \mathbb{R}^{n \times n}$ to be the kernel matrix with i, j-th entry $K_{ij} = k((x_i, z_i), (x_j, z_j))$. Let σ_j^2 denote the eigenvalues of K. We assume that $\sigma_j^2 = \sigma^2 > 0$ is constant for all j; we can relax this at the cost of additional complexity. The "single kernel ridge" regression outcome regression estimator with parameter λ has coefficient estimates:

$$\hat{\beta}^{\lambda}_{\text{ridge}} = (K + \lambda I)^{-1} y.$$

Applying Proposition 4.3, the augmented "double kernel ridge" estimator with hyperparameter δ is equivalent to a plug-in estimate for a new kernel ridge model:

$$\hat{\beta}_{\text{aug}} = (K + \gamma I)^{-1} y, \qquad \text{with } \gamma = \frac{\lambda \delta}{\sigma^2 + \lambda + \delta}.$$

We can now use modern rate results for kernel ridge regression to explicitly link double kernel ridge and efficiently undersmoothed kernel ridge. First, if we choose hyperparameter schedule λ_n for kernel ridge regression as in Fischer and Steinwart (2020), we obtain a corresponding convergence rate for the outcome model (see Appendix L for specifics). Second, we can use the hyperparameter schedule δ_n and Theorem 1 from Singh (2024) to establish a convergence rate for the Riesz representer. Finally by Theorem 4.2 of Chernozhukov et al. (2022e), the augmented estimator that combines these two kernel ridge nuisance estimates is $\sqrt{n}$-consistent. Note that while in this discussion we assume well-specification, i.e. $m \in \mathcal{H}$, these rates are *model-agnostic*; similar results hold when $m \notin \mathcal{H}$ (see Singh, 2024).

Applying our numerical results, the augmented estimator is also a kernel ridge estimator with a new (undersmoothed) hyperparameter schedule, γ_n. We will now show that γ_n recovers existing rates for undersmoothed ridge estimators. In particular, we will consider the special case where the hyperparameter schedule $\delta_n = \lambda_n$ satisfies the conditions for Theorem 1 and Assumption 2 of Singh (2024). This is a non-trivial assumption — i.e. that the smoothness of the Riesz representer RKHS matches that of the outcome model — but the idea is motivated by the concept of the "minimal" Riesz representer from Lemma S3.1 in Chernozhukov et al. (2022a). For two functions of n, f_n and g_n, let $f_n \asymp g_n$ denote that $f_n = O(g_n)$ and $g_n = O(f_n)$. The resulting augmented hyperparameter is then $\gamma_n \asymp \lambda_n^2$.

When the RKHS is finite dimensional, the choice $\lambda_n = \delta_n = n^{-1/2}$ is optimal for controlling the prediction error for both the outcome and weighting models (Caponnetto and De Vito, 2007; Singh, 2024). The augmented estimator is then equivalent to a single ridge regression with hyperparameter $\gamma_n \asymp n^{-1}$, which recovers the rate of Hirshberg et al. (2019); Mou et al. (2023).

When the RKHS is infinite-dimensional, when $\lambda_n = \delta_n = n^{-1/2}$, then $\gamma_n \asymp n^{-1}$, again matching the rate in Hirshberg et al. (2019); Mou et al. (2023). This provides further motivation to fix $\delta_n = \lambda_n$. However, depending on the smoothness and effective dimension of the RKHS, λ_n can take on a range of values, resulting possibly faster or slower rates than n^{-1}; we give concrete examples in the Appendix. This somewhat contrasts with the results in Hirshberg et al. (2019); Mou et al. (2023), and might be an interesting direction for future

analysis. Inspired by these results, in the next section, we will assess the performance of setting $\delta = \lambda$ for hyperparameter tuning in practice. In fully generality, when λ_n and δ_n differ, we will end up with a product rate, again contrasting with existing work. In this sense, Proposition 4.3 generalizes the standard undersmoothing arguments, which typically change the regularization schedule from $n^{-1/2}$ to n^{-1}.

Remark 8 (Single-model double robustness)**.** *Another interesting implication of the equivalence of these two procedures is that the single kernel ridge procedure is doubly robust, much the same way OLS is. Because estimating the coefficients from an OLS regression of Y onto features of (Z, X) is equivalent to a balancing weights or an IPW estimator based on a model for the inverse weights that is linear in the same features, this procedure is consistent whenever* either *the weights or the outcome model is truly linear—that is, whenever either of these two linear models is correctly specified (Robins et al., 2007). Similarly, the single kernel ridge procedure is doubly robust in that it is consistent if either the true outcome regression or the inverse propensity score is consistently estimated. However, valid inference in the case where the inverse weight model but* not *the outcome model is truly linear will typically require different tuning parameter selection.*

6.2 Finite Sample Mean-Squared Error

We now use our numerical equivalences to write out the exact finite-sample mean squared error of the augmented kernel ridge estimator: by re-writing the augmented balancing weights estimator as a single outcome model, we can immediately leverage existing results from Dobriban and Wager (2018).

Following their setup, we define the diagonal matrix $\hat{\Sigma} \coloneqq \frac{1}{n}\Phi_p^\top \Phi_p$; if $\hat{\Sigma}$ is not diagonal, we can apply the rotation in Appendix E.2. We consider ridge regression with rescaled hyperparameter λ and solution $(\hat{\Sigma} + \lambda I)^{-1}\Phi_p Y_p/n$; this is equivalent to standard ridge regression above with hyperparameter $n\lambda$, and also accommodates kernel ridge regression with appropriate choice of Φ_p. Assume that $Y_p = \Phi_p\beta_0 + \epsilon$ with $\beta_0 \in \mathbb{R}^d$, and where $\epsilon \in \mathbb{R}^n$ are iid with mean zero and variance σ^2. Then the exact, design-conditional, squared bias and variance of the ridge regression prediction applied to a new iid sample $(\Phi_{\text{new}}, Y_{\text{new}}) \sim p$ are:

$$B_{\text{p}}^2(\lambda) = \lambda^2 \beta_0^\top (\hat{\Sigma} + \lambda I)^{-1}\mathbb{E}[\Phi_p^\top \Phi_p](\hat{\Sigma} + \lambda I)^{-1}\beta_0$$
$$V_{\text{p}}(\lambda) = \frac{\sigma^2}{n}\text{tr}\left[\hat{\Sigma}(\hat{\Sigma} + \lambda I)^{-1}\mathbb{E}[\Phi_p^\top \Phi_p](\hat{\Sigma} + \lambda I)^{-1}\right].$$

Applying Proposition 4.3, we can simlarly derive the squared bias and variance of an augmented ridge estimator for our linear functional estimand; we denote these quantities B_q^2 and V_q respectively. We express the bias and variance in terms of the two hyperparameters, λ and δ:

Proposition 6.1. *Let σ_j^2 denote the eigenvalues of $\hat{\Sigma}$ and define $\Gamma_{\lambda,\delta}$ to be the diagonal matrix with non-zero entries $\gamma_j \coloneqq \frac{\delta\lambda}{\sigma_j^2 + \delta + \lambda}$. Then,*

$$B_q^2(\lambda, \delta) = \beta_0^\top (\hat{\Sigma} + \Gamma_{\lambda,\delta})^{-1}\Gamma_{\lambda,\delta}\mathbb{E}[\Phi_q]^\top \mathbb{E}[\Phi_q]\Gamma_{\lambda,\delta}(\hat{\Sigma} + \Gamma_{\lambda,\delta})^{-1}\beta_0$$
$$V_q(\lambda, \delta) = \frac{\sigma^2}{n} tr\left[\hat{\Sigma}(\hat{\Sigma} + \Gamma_{\lambda,\delta})^{-1}\mathbb{E}[\Phi_q]^\top \mathbb{E}[\Phi_q](\hat{\Sigma} + \Gamma_{\lambda,\delta})^{-1}\right].$$

In the next section, we compare — numerically and via simulation — existing hyperparameter selection schemes to the optimal trade-off between B_q^2 and V_q. However, first we note that the analysis above opens up exciting new avenues for both theoretical and methodological work. One could theoretically analyze the mean squared error to understand how the optimal δ scales with the problem parameters; for example, by using proportionate asymptotics from random matrix theory as in the high-dimensional ridge regression literature (Hastie et al., 2022; Patil et al., 2024). Furthermore, while Proposition 6.1 requires the linear model to be well-specified, in the mis-specified setting, we could adapt the model-agnostic decomposition from Proposition 1 of Patil et al. (2024). Finally, our analysis here suggests a novel, hyperparameter selection scheme based on plugging in the unknown quantities in Proposition 6.1. We leave this to future work.

7 Numerical illustrations and hyperparameter tuning

This section illustrates our results in practice. We first explore hyperparameter tuning for double ridge regression, comparing practical methods to the optimal hyperparameter computed using our results from Proposition 6.1. Following our asymptotic results in Section 6.1, we recommend equating the weighting and outcome model hyperparameters in practice. We then apply both double ridge and lasso-augmented ℓ_∞-balancing to two versions of the canonical LaLonde (1986) application. An important theme throughout is that some approaches for hyperparameter selection can choose $\delta = 0$, which collapses the augmented estimate to OLS alone — even in settings where this is far from optimal. Overall, we take this as a warning that existing hyperparameter tuning schemes can be potentially misleading when applied naively.

7.1 Hyperparameter tuning for ridge-augmented ℓ_2 balancing

We begin with practical hyperparameter tuning for the special case of double ridge, building on the MSE expression in Section 6.2. There is an active literature on selecting hyperparameters for augmented balancing weights estimators and double machine learning estimators more broadly (Kallus, 2020; Wang and Zubizarreta, 2020; Ben-Michael et al., 2021b; Bach et al., 2024). We contribute to this literature by comparing practical hyperparameter tuning schemes with an oracle hyperparameter tuning scheme based on Proposition 6.1.

Reflecting empirical practice, we focus here on choosing hyperparameters sequentially: we first select the outcome model hyperparameter λ (e.g. by cross-validation) and then select the weighting model hyperparameter δ. Ultimately, we find strong performance for both *CV imbalance* and *CV outcome* hyperparameters, as defined below. We especially recommend the latter as a reasonable starting point in practice. In addition to theoretical support from our asymptotic analysis, the outcome model hyperparameter scheme does not require any additional algorithm or code after having fit the initial outcome model.

7.1.1 Oracle and practical hyperparameter tuning

Oracle hyperparameter. To compute oracle hyperparameters, we first compute the prediction-MSE-optimal λ using the standard ridge regression MSE expression, and then we use Proposition 6.1 to compute the corresponding optimal δ for the linear functional estimand:

$$\lambda^* := \text{argmax}_\lambda\{B_p^2(\lambda) + V_p(\lambda)\}$$
$$\delta^* := \text{argmax}_\delta\{B_q^2(\lambda^*, \delta) + V_q(\lambda^*, \delta)\}.$$

While there is not a closed form for δ^*, we can nonetheless directly compute this optimal hyperparameter and characterize its behavior under a range of scenarios. We draw several conclusions about optimal δ^* for a wide range of DGPs of the form $Y_p = \Phi_p\beta_0 + \epsilon$. First, δ^* is generally increasing in the noise, σ^2: larger σ^2 typically implies larger δ^*. Second, δ^* generally depends on the target mean, $\mathbb{E}[\Phi_q]$; that is, two DGPs that are identical except for $\mathbb{E}[\Phi_q]$ can have different values of δ^*. The optimal hyperparameter, however, does *not* depend on the magnitude of the shift in the target mean: replacing $\mathbb{E}[\Phi_q]$ with $c\mathbb{E}[\Phi_q]$ for $c \neq 0$, scales both the bias and variance by c^2, leaving δ^* unchanged.

Practical hyperparameter. We compare the oracle hyperparameter with three implementable practical proposals. In all cases, we first pick λ by cross-validating the mean squared error of a ridge outcome model.

- *CV imbalance.* Choose δ by cross-validating the estimated imbalance, $\|\frac{1}{n}\hat{w}\Phi_p - \bar{\Phi}_q\|_2^2$, adapting a proposal from Wang and Zubizarreta (2020).
- *CV Riesz loss.* Choose δ by cross-validating the Riesz loss in Equation (6), adapting a proposal from Chernozhukov et al. (2022d); this is the dual form of cross-validating the estimated imbalance.
- *CV outcome.* Choose δ to be equal to the cross-validated ridge outcome λ, as inspired by the asymptotic theory in Mou et al. (2023); Singh (2024).

	# of DGPs		Relative MSE			
Method	Best	Worst	Median	Best	Worst	Prop.($\delta = 0$)
CV Outcome	9	2	0.57	0.096	2×10^5	0
CV Imbalance	27	2	0.41	0.043	2×10^5	0
CV Riesz Loss	0	32	9,268	0.330	3×10^7	0.56

Table 1: Mean-squared error (relative to the oracle) for three hyperparameter selection methods for *double ridge regression* from a numerical investigation of 36 data generating processes (30 synthetic and 6 semi-synthetic). The final column is the proportion of draws where the hyperparameter $\delta = 0$.

Before presenting simulation results, we provide a preliminary analytic discussion, comparing these practical schemes to the behavior of the oracle δ^*. For the first two proposals: just like the oracle, both depend on the target mean $\mathbb{E}[\Phi_q]$ and are invariant to re-scaling. However, these two approaches are mechanically independent of the outcomes Y_p, unlike the oracle δ^* which, in general, depends on the variance of the outcomes. By constrast, the last proposal depends on the outcomes Y_p but is mechanically independent of $\mathbb{E}[\Phi_q]$.

This suggests that any one of these tuning parameter approaches cannot perform well across all DGPs. In future work, if we pursue a theoretical analysis of the oracle hyperparameter, e.g. in a proportionate asymptotics framework, we may be able predict when either the outcomes or the covariate shift is more important. In this work we begin by demonstrating that no one tuning scheme does uniformly best in simulations.

7.1.2 Simulation study

To assess the behavior of these hyperparameter tuning schemes, we conduct a simulation study using 36 distinct data-generating processes, 30 synthetic and 6 semi-synthetic; see Appendix H for a detailed discussion. For each DGP, we directly compute the oracle hyperparameter using the results in Section 6.2. We then compute values from the three practical hyperparameter tuning methods discussed above. The mean squared error that we consider is design-conditional, and so we draw samples of the covariates for each DGP only once.

Table 1 presents a summary of the MSE for the three methods across the 36 DGPs. Overall, we find that the *CV outcome* approach of choosing $\delta = \lambda$ and the *CV imbalance* approach both perform well in practice: one of these two achieves the lowest MSE all 36 DGPs, with CV imbalance performing slightly better on average. By contrast, selecting δ via CV for the Riesz loss has numerical stability problems that compromises performance. The performance for the *outcome* and *balance* approaches, on the other hand, seem to degrade gracefully and rarely perform catastrophically. Taken together, these preliminary findings suggest researchers should begin with these two tuning methods as defaults.

Recovering the OLS point estimate. As we discuss above (see, e.g., Figure 1), when $\delta = 0$ the point estimate for the augmented balancing weights estimator is numerically identical to the OLS point estimate. Thus, when a hyperparameter tuning procedure chooses $\delta = 0$ in practice, researchers are simply estimating the equivalent of OLS — even if they are unaware they are doing so. This is especially problematic in settings where OLS is far from optimal (though see Kobak et al., 2020; Hastie et al., 2022, for counterexamples). In our synthetic and semi-synthetic DGPs, $\delta = 0$ is never optimal, and is usually associated with a very large error driven by extreme variance — see for example, Figure I.11 in the Appendix. Thus the fact that hyperparameter tuning procedures can return $\delta = 0$ in these DGPs represents a pathological case.

In our simulation study, we find that, when cross validating the Riesz loss, over half of all draws returned $\delta = 0$. By contrast, none of the other methods returned $\delta = 0$ in the synthetic DGPs, though, as we

discuss below, we do observe exact zeros for δ occasionally when cross-validating imbalance in the standard LaLonde dataset. This highlights the potential numeric instability of hyperparameter tuning via CV for the Riesz loss, at least in the settings we consider here. We further suggest that in these cases, practitioners assess the sensitivity of the $\delta = 0$ results to the particular tuning procedure used or to the random choice of cross-validation splits.

7.2 Application to LaLonde (1986)

We now illustrate our equivalence and hyperparameter tuning results on real-world datasets. Following Chernozhukov et al. (2022d), we focus on the canonical LaLonde (1986) data set evaluating a job training program in the National Supported Work (NSW) Demonstration. The primary outcome of interest is annual earnings in 1978 dollars.

For these illustrations, we estimate the Average Treatment Effect on the Treated (ATT), $\mathbb{E}[Y(1)-Y(0) \mid Z = 1]$. We recover the missing conditional mean $\mathbb{E}[Y(0) \mid Z = 1]$ using the setup from Example 3 in Appendix A, where the source and target populations are the control and treated units respectively. Thus Φ_p and Φ_q correspond to the feature expansion $\phi(X)$ applied to the covariates in the control group and treated group respectively. We consider two different features expansions of the original covariates: (1) a "short" set of 11 covariates used in Dehejia and Wahba (1999);[1] and (2) an expanded, "long" set of 171 interacted features used in Farrell (2015).

Our goal is to explicate how augmented estimators under different hyperparameter tuning schemes undersmooth in practice in both low and high-dimensional settings. In some cases, the augmented estimator collapses to exactly OLS as we document above. Appendix I contains extensive additional analyses, including dataset summaries, additional results from the Infant Health Development Program (IHDP), and sensitivity of these numerical results to cross-fitting.

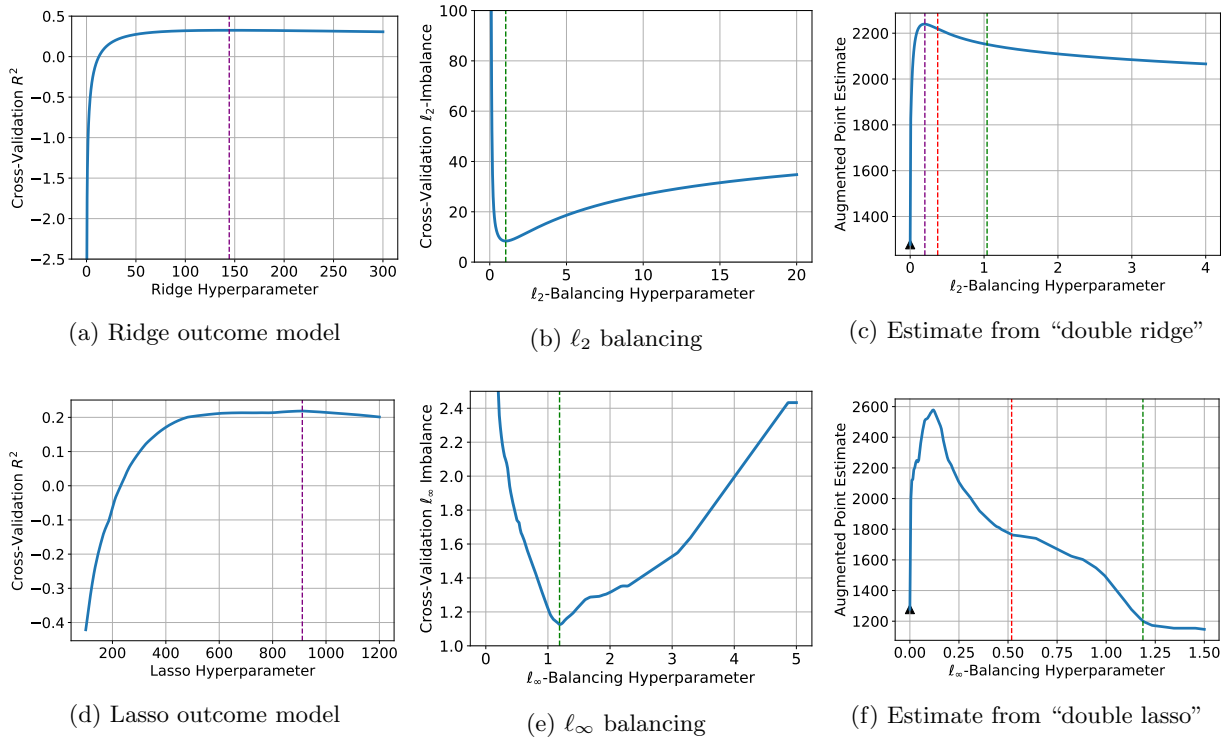
7.2.1 High-dimensional setting

Following Chernozhukov et al. (2022d), we first consider the expanded set of 171 features for LaLonde (1986) used in Farrell (2015). Figure 3 shows estimates for ridge-augmented ℓ_2 balancing (top row) and lasso-augmented ℓ_∞ balancing (bottom row). We explicitly characterize these results in terms of undersmoothing in Appendix I.4. The left two panels of each row show the cross-validation curves for the outcome regression and balancing weights, respectively. The right panels show the point estimate as a function of the weighting hyperparamter δ, holding the outcome model hyperparameter λ fixed; the black triangle represents the OLS plug-in point estimate. For context, the corresponding experimental estimate is \$1,794 (see Dehejia and Wahba, 1999). The green and red dotted lines correspond to hyperparameters chosen by cross-validating balance and the Riesz loss, respectively. For the double ridge estimate, the purple line corresponds to $\delta = \hat{\lambda}$, the outcome hyperparameter selected via cross validation.

Figure 3 highlights that both the imbalance and the point estimate are highly nonlinear close to zero. Thus, even small departures from OLS (at $\delta = 0$) lead to large changes in the point estimate — in Appendix I.5 we give some suggestive evidence that the variance blows up relative to the bias in this range. We can also assess the sensitivity of the point estimate to the hyperparameter selection scheme. In this case, choosing δ via CV balance leads to meaningfully larger choices than via other methods.

Finally, the selected δ is always strictly greater than zero for this high-dimensional dataset. However, we find this is sensitive to small perturbations in the problem parameters. For example, when we perturb $\mathbb{E}[\Phi_q]$ by adding a small value to all the even elements, then the cross-validated ℓ_2 Riesz loss chooses $\delta = 0$ in 38% of draws of the cross-validation splits. As suggested by Appendix I.5 and our simulation results, this is likely to result in extremely large mean squared error.

[1] These are: age, years of education, Black indicator, Hispanic indicator, married indicator, 1974 earnings, 1975 earnings, age squared, years of education squared, 1974 earnings squared, and 1975 earnings squared.



(a) Ridge outcome model

(b) ℓ_2 balancing

(c) Estimate from "double ridge"

(d) Lasso outcome model

(e) ℓ_∞ balancing

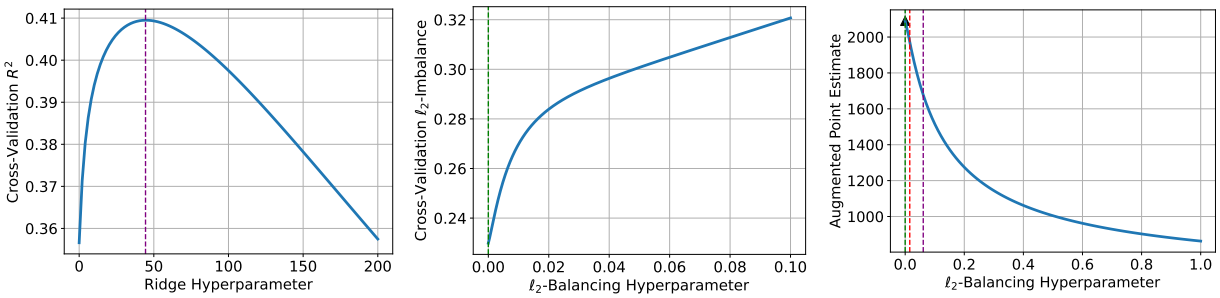
(f) Estimate from "double lasso"

Figure 3: Augmented balancing weights estimates for the LaLonde (1986) data set with the expanded set of 171 features used in Farrell (2015); the top row shows ridge-augmented ℓ_2 balancing, and the bottom row shows lasso-augmented ℓ_∞ balancing. Panels (a) and (d) show the 3-fold cross-validated R^2 for the ridge- and lasso-penalized regression of Y_p on Φ_p among control units across the hyperparameter λ; the purple dotted lines show the CV-optimal value for each. Panel (b) and (e) show the 3-fold cross-validated imbalance for ℓ_2 and ℓ_∞ balancing weights across the hyperparameter δ; the green dotted lines show the CV-optimal value for each. Panels (c) and (f) show the point estimates for the augmented estimators across the weighting hyperparameter δ; the black triangles correspond to the OLS point estimate; the green and red dotted lines correspond to the cross-validated balance and Riesz loss respectively; the purple line corresponds to the cross-validated ridge hyperparameter (for $\delta = \hat{\lambda}$).

7.2.2 Low-dimensional setting: Recovering OLS

Finally, we apply double ridge to the "short" version of the LaLonde (1986) data set with 11 features. Figure 4 shows the cross-validation curves for the outcome and weighting models, as well as the point estimate as a function of the balance hyperparameter, with the OLS estimate given by the black triangle. As above, the green, red, and purple dotted lines correspond to hyperparameters chosen by cross-validating balance, cross-validating the Riesz loss, and choosing $\delta = \lambda$ respectively.

Unlike for the "long" dataset in Figure 3, Figure 4 does not display as stark a nonlinearity around zero. Importantly, however, setting δ by cross-validating imbalance can yield $\delta = 0$, which reduces the augmented estimator to exactly the estimate from a simple OLS regression — even though the base learner ridge outcome model is heavily regularized. By contrast, our preferred hyperparameter tuning scheme of choosing $\delta = \lambda$ results in an estimate that is roughly \$400 dollars smaller than the OLS estimate. The choice of $\delta = 0$ is sensitive to the specific cross-validation splits used, further emphasizing that this is likely anomalous behavior. See Section 7.1.2 for further discussion.



(a) Ridge outcome model (b) ℓ_2 balancing (c) Augmented estimate

Figure 4: Ridge-augmented ℓ_2 balancing weights ("double ridge") for LaLonde (1986) with the original 11 covariates. Panel (a) shows the 3-fold cross-validated R^2 for the Ridge-penalized regression of Y_p on Φ_p among control units across the hyperparameter λ; the purple dotted line shows the CV-optimal value, $\hat{\lambda}$. Panel (b) shows the 3-fold cross-validated imbalance for ℓ_2 balancing weights across the hyperparameter δ; the green dotted line shows the CV-optimal value, which is $\delta = 0$ or exact balance. Panel (c) shows the point estimate for the augmented estimator across the weighting hyperparameter δ; the black triangle corresponds to the OLS point estimate, the green dotted line corresponds to cross-validated balance, the red dotted line corresponds to cross-validated Riesz loss, and the purple dotted line corresponds to the ridge outcome hyperparameter.

8 Discussion

We have shown that augmenting a plug-in regression estimator with linear balancing weights results in a new plug-in estimator with coefficients that are shrunk towards — in some cases all the way to — the estimates from OLS fit on the same observations. We generalize this equivalence for different choices of outcome and weighting regressions. In the asymptotic setting, we draw the explicit connection between augmented estimators and undersmoothing for the special case of kernel ridge regression. Then we derive the design-conditional finite sample MSE for the double ridge estimator, and use it to solve numerically for oracle hyperparameters. We compare the oracle hyperparameters with three practical tuning schemes and then illustrate our results on the canonical LaLonde (1986) data set. In the Appendix, we also explore many extensions, including to nonlinear weights and to high-dimensional features.

There are many promising avenues for future research. The fundamental connection between doubly robust estimation and undersmoothing opens up several theoretical directions. While we focus on the special case of kernel ridge regression in Section 6.1, we anticipate that these connections will hold more broadly. Similarly, while our focus in this paper has been on interpreting balancing weights as a form of linear regression, the converse is also valid: 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.

We also anticipate that the MSE we derive in Section 6.2 is a starting place for future theoretical analysis that can inform practice. We demonstrate in our simulation study that existing hyperparameter selection methods cannot perform uniformly well over all DGPs. We expect that analyzing the optimal hyperparameters — e.g. in a proportionate asymptotics regime — can help devise new tuning schemes and inform which tuning method will work best on the dataset at hand.

We further conjecture that these results may provide new insights into the estimation of causal effects in the proximal causal inference framework (Tchetgen Tchetgen et al., 2020). This framework uses proxy variables to identify 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) and Kallus et al. (2021) recently proposed

tractable nonparametric estimators in this setting. They use an "adversarial" version of double kernel ridge regression — allowing the weighting and outcome models to have different bases — to estimate the solution to the required Fredholm integral equations. Our results apply immediately to standard augmented estimators with different bases for the outcome and weighting models, either via a union basis (Chernozhukov et al., 2022d) or by applying an appropriate projection as in Hirshberg and Wager (2021), and extending these results to proximal causal effect estimators might help in constructing new proximal balancing weights, matching, or regression estimators with attractive asymptotic properties.

Finally, 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 Appendix D.2, to better understand connections between methods and to inform inference.

Acknowledgements

We would like to thank David Arbour, Eli Ben-Michael, Andreas Buja, Alex D'Amour, Skip Hirshberg, Guido Imbens, Apoorva Lal, Mark van der Laan, Whitney Newey, Rahul Singh, Jann Spiess, Eric Tchetgen Tchetgen, and Qingyuan Zhao 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 grants N000142112820 and N000142412701 and by the Simons Institute for Theoretical Computer Science.

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.

A. Agarwal, Y. S. Tan, O. Ronen, C. Singh, and B. Yu. Hierarchical shrinkage: Improving the accuracy and interpretability of tree-based models. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 111–135. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/agarwal22b.html.

D. Arbour and A. Feller. The role of simplex constraints in regularizing treatment effect estimates. 2024.

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. Bach, O. Schacht, V. Chernozhukov, S. Klaassen, and M. Spindler. Hyperparameter tuning for causal inference with double machine learning: A simulation study. *arXiv preprint arXiv:2402.04674*, 2024.

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.

A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7:331–368, 2007.

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. *The Econometrics Journal*, 21(1):C1–C68, 01 2018a.

V. Chernozhukov, W. K. Newey, and R. Singh. Learning l2-continuous regression functionals via regularized riesz representers. *arXiv preprint arXiv:1809.05224*, 8, 2018b.

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-Martinez, 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. Newey, R. Singh, and V. Syrgkanis. Automatic debiased machine learning for dynamic treatment effects and general nested functionals. *arXiv preprint arXiv:2203.13887*, 2022c.

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

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

V. Chernozhukov, W. K. Newey, V. Quintas-Martinez, and V. Syrgkanis. Automatic debiased machine learning via riesz regression, 2024.

R. H. Dehejia and S. Wahba. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American statistical Association*, 94(448):1053–1062, 1999.

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

E. Dobriban and S. Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279, 2018.

M. H. Farrell. Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, 189(1):1–23, 2015.

S. Fischer and I. Steinwart. Sobolev norm learning rates for regularized least-squares algorithms. *The Journal of Machine Learning Research*, 21(1):8464–8501, 2020.

C. Gao, S. Yang, and J. K. Kim. Soft calibration for selection bias problems under mixed-effects models. *arXiv preprint arXiv:2206.01084*, 2022.

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.

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.

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.

T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Annals of statistics*, 50(2):949, 2022.

J. K. Hellerstein and G. W. Imbens. Imposing moment restrictions from auxiliary data by weighting. *Review of Economics and Statistics*, 81(1):1–14, 1999.

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.

N. Kallus, X. Mao, and M. Uehara. Causal inference under unmeasured confounding with negative controls: A minimax learning approach. *arXiv preprint arXiv:2103.14029*, 2021.

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

M. P. Kim, C. Kern, S. Goldwasser, F. Kreuter, and O. Reingold. Universal adaptability: Target-independent inference that competes with propensity scoring. *Proceedings of the National Academy of Sciences*, 119(4):e2108097119, 2022.

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

D. Kobak, J. Lomond, and B. Sanchez. The optimal ridge penalty for real-world high-dimensional data can be zero or negative due to the implicit ridge regularization. *The Journal of Machine Learning Research*, 21(1):6863–6878, 2020.

R. J. LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, pages 604–620, 1986.

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

T. Lumley, P. A. Shaw, and J. Y. Dai. Connections between survey calibration estimators and semiparametric models for incomplete data. *International Statistical Review*, 79(2):200–220, 2011.

A. Menon and C. S. Ong. Linking losses for density ratio and class-probability estimation. In *International Conference on Machine Learning*, pages 304–313. PMLR, 2016.

N. Moosavi, J. Häggström, and X. de Luna. The costs and benefits of uniformly valid causal inference with high-dimensional nuisance parameters. *Statistical Science*, 38(1):1–12, 2023.

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. The asymptotic variance of semiparametric estimators. *Econometrica: Journal of the Econometric Society*, pages 1349–1382, 1994.

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 and R. J. Smith. Higher order properties of gmm and generalized empirical likelihood estimators. *Econometrica*, 72(1):219–255, 2004.

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.

P. Patil, J.-H. Du, and R. J. Tibshirani. Optimal ridge regularization for out-of-distribution prediction. *arXiv preprint arXiv:2404.01233*, 2024.

J. Robins, M. Sued, Q. Lei-Gomez, and A. Rotnitzky. Comment: Performance of double-robust estimators when" inverse probability" weights are highly variable. *Statistical Science*, 22(4):544–559, 2007.

J. Robins, L. Li, E. Tchetgen, A. van der Vaart, et al. Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and statistics: essays in honor of David A. Freedman*, volume 2, pages 335–422. Institute of Mathematical Statistics, 2008.

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.

X. Shen. On methods of sieves and penalization. *The Annals of Statistics*, 25(6):2555–2591, 1997.

R. Singh. Kernel ridge riesz representers: Generalization, mis-specification, and the counterfactual effective dimension. *arXiv preprint arXiv:2102.11076v4*, 2024.

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

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

D. Tang, D. Kong, W. Pan, and L. Wang. Ultra-high dimensional variable selection for doubly robust causal inference. *Biometrics*, 79(2):903–914, 2023.

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

K. Vermeulen and S. Vansteelandt. Bias-reduced doubly robust estimation. *Journal of the American Statistical Association*, 110(511):1024–1036, 2015.

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

R. K. Wong and K. C. G. Chan. Kernel-based covariate functional balancing for observational studies. *Biometrika*, 105(1): 199–213, 2018.

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.