Proceedings of Machine Learning Research TBD:1–25, 2025 4th Conference on Causal Learning and Reasoning

Automatic debiasing of neural networks via moment-constrained learning

Christian L. Hines CHINES@TURING.AC.UK
The Alan Turing Institute
London, UK

Oliver J. Hines OLIVER.HINES@QUANTCO.COM
QuantCo
Berlin, Germany

Editors: Biwei Huang and Mathias Drton

Abstract

Causal and nonparametric estimands in economics and biostatistics can often be viewed as the mean of a linear functional applied to an unknown outcome regression function. Naively learning the regression function and taking a sample mean of the target functional results in biased estimators, and a rich debiasing literature has developed where one additionally learns the so-called Riesz representer (RR) of the target estimand (targeted learning, double ML, automatic debiasing etc.). Learning the RR via its derived functional form can be challenging, e.g. due to extreme inverse probability weights or the need to learn conditional density functions. Such challenges have motivated recent advances in automatic debiasing (AD), where the RR is learned directly via minimization of a bespoke loss. We propose moment-constrained learning as a new RR learning approach that addresses some shortcomings in AD, constraining the predicted moments and improving the robustness of RR estimates to optimization hyperparamters. Though our approach is not tied to a particular class of learner, we illustrate it using neural networks, and evaluate on the problems of average treatment/derivative effect estimation using semi-synthetic data. Our numerical experiments show improved performance versus state of the art benchmarks.

1. Introduction

Several problems in causal inference, economics, and biostatistics can be viewed as inferring the average moment estimand $\Psi \equiv \mathbb{E}[m(\mu, W)]$, where $W = (Y, Z)$ consists of an outcome Y and inputs $Z = (A, X)$ often comprised of a treatment A and covariates X, with $\mu(z) \equiv \mathbb{E}[Y|Z = z]$, and $m(f, W)$ is a known functional that is linear in f. Examples of this setup include average treatment, policy and derivative effects, as outlined below. For such estimands, naively plugging in regression estimates $\hat{\mu}$ and taking the sample mean of $m(\hat{\mu}, W)$, given i.i.d. observations of W, generally leads to biased estimates which converge to Ψ at less than the parametric $\sqrt{n}$ rate.

These biases arise because the bias-variance trade off of the regression estimator is controlled by a generic loss (e.g. mean squared error, cross-entropy) that does not adequately control for biases in the downstream estimation task. In particular, the true regression function μ satisfies $\mathbb{E}[\alpha(Z)\{Y - \mu(Z)\}] = 0$ for any function α, but the same is not true of

© 2025 C.L. Hines & O.J. Hines.

the empirical mean $\mathbb{E}_n[\alpha(Z)\{Y - \hat{\mu}(Z)\}]$, which may converge to zero slower than the $\sqrt{n}$ rate. Biases for average moment estimands take this form for an estimand-specific function α. Specifically, the 'plug-in bias' is characterized by the Riesz representer (RR) α, which is an unknown function such that $\Psi = \langle \mu, \alpha \rangle$, where $\langle f, g \rangle \equiv \mathbb{E}[f(Z)g(Z)]$ denotes an inner product over a Hilbert space $\mathcal{H}$ equipped with norm $||f|| \equiv \langle f, f \rangle^{1/2} < \infty$ and it is assumed that $\mu \in \mathcal{H}$. Existence of a unique $\alpha \in \mathcal{H}$ follows by Riesz's representation theorem since $f \mapsto h(f) \equiv \mathbb{E}[m(f, W)]$ is a bounded linear map.

Example 1: Average treatment effect (ATE). *For a binary treatment $A \in \{0, 1\}$, the ATE is the difference in mean outcome when an intervention assigns treatment versus no-treatment uniformly across the population (Rosenbaum and Rubin, 1983). Under standard causal assumptions, the ATE is identified by $\Psi = \mathbb{E}[\mu(1, X) - \mu(0, X)]$, which is an average moment estimand for the moment functional $m(f, W) \equiv f(1, X) - f(0, X)$. Letting $p(x) \equiv \mathbb{E}[A|X = x]$, and assuming $p(x) \in (0, 1)$, the ATE has the RR $\alpha(z) = \{a - p(x)\}/[p(x)\{1 - p(x)\}]$.*

Example 2: Average policy effect (APE) *Using the setup from Example 1, the APE considers the mean outcome when an intervention assigns treatment according to a known treatment policy $x \mapsto \pi(x) \in \{0, 1\}$ (Dudik et al., 2011; van der Laan and Luedtke, 2014; Athey and Wager, 2021). The APE is identified by $\Psi = \mathbb{E}[\pi(X)\{\mu(1, X) - \mu(0, X)\} + \mu(0, X)]$, which is an average moment estimand for the moment functional $m(f, W) \equiv \pi(X)\{f(1, X) - f(0, X)\} + f(0, X)$. The APE has the RR $\alpha(z) = [\pi(x)\{a - p(x)\} + p(x)\{1 - a\}]/[p(x)\{1 - p(x)\}]$.*

Example 3: Average derivative effect (ADE) *For a continuous treatment $A \in \mathbb{R}$, the ADE, $\Psi = \mathbb{E}[\mu'(A, X)]$ is average derivative of the conditional response function, where superscript prime denotes the derivative w.r.t. a, and we assume that μ' exists (Härdle and Stoker, 1989; Newey and Stoker, 1993; Imbens and Newey, 2009; Rothenhäusler and Yu, 2019). The ADE is an average moment estimand for the moment functional $m(f, W) \equiv f'(A, X)$. Letting $p(a|x)$ denote the conditional density of A given X, and assuming $p(a|x) > 0$, and $p(a|x) = 0$ for a on the boundary of the support of A, then the ADE has the RR $\alpha(z) = p'(a|x)/p(a|x)$.*

Example 4: Incremental policy effect (IPE) *Using the setup from Example 3, the IPE (Athey and Wager, 2021) is $\Psi = \mathbb{E}[\pi(X)\mu'(A, X)]$, where $x \mapsto \pi(x) \in [-1, 1]$ is a known policy function. The IPE is an average moment estimand for the moment functional $m(f, W) \equiv \pi(X)f'(A, X)$ and has the RR $\alpha(z) = \pi(x)p'(a|x)/p(a|x)$.*

Following semiparametric efficiency results (Robins et al., 1994; Newey, 1994), a rich literature has developed in recent years that compensates for plug-in biases either by estimating the RR then shifting the naive estimator (double machine learning (Chernozhukov et al., 2018)), or retrospectively modifying the estimates $\hat{\mu}$ such that the estimated plug-in bias is negligible (targeted learning (Zheng and van der Laan, 2011)). Both approaches are celebrated for constructing efficient estimators that converge at $\sqrt{n}$ rate even when learners for the conditional mean outcome and the RR converge at a slower e.g. $n^{1/4}$ rate.

As exemplified above, the RR can be a complicated function of the data distribution making learning the RR using its derived form challenging. For instance, the RR of the ATE and APE can be estimated using a learner $\hat{p}$ of the propensity score p, but the resulting estimates may be overly sensitive to the error $\hat{p}(x) - p(x)$ when $\hat{p}(x)$ is close to 0 or 1, since

p appears in the denominator of the RR. Similarly, RR estimators for the ADE typically use kernel estimators and are overly sensitive to the bandwidth (Cattaneo et al., 2013).

To overcome such issues, recent work has sought to learn the RR directly from the data, without using knowledge of its functional form. Initial approaches for binary treatments used balancing weights rather than propensity scores to estimate plug-in biases (Zubizarreta, 2015; Athey et al., 2018). These approaches have been generalized through the adversarial RR learner of (Chernozhukov et al., 2020) that builds on the (also adversarial) augmented minimax linear estimator (Hirshberg and Wager, 2021) and similar estimators for conditional moment models Dikkala et al. (2020). More recently, automatic debiasing (AD) (Chernozhukov et al., 2021, 2022a) has been proposed to bypass the need to solve a computationally challenging adversarial learning problem by constructing a simple loss that is equivalent to minimizing the mean squared error in the RR. AD generalizes similar approaches using approximately sparse linear regression Chernozhukov et al. (2022b) and reproducing kernel Hilbert spaces Singh et al. (2023).

Despite the success of AD, there are several areas for improvement which we address in our work. First, the AD loss is unbounded, and includes a negative average moment term that can lead to extreme moment predictions in the final RR estimator. In practice, early stopping using an external validation set is recommend to avoid such issues, however the resulting learners may be overly sensitive to e.g. early stopping and learning rate hyperparameters. Second, oftentimes the RR admits known inner products which are ignored by the AD loss. For instance, we know *a priori* that for the ADE/ATE $h(a) = \mathbb{E}[A\alpha(Z)] = 1$. Methods which estimate the RR using its derived form approximately encode such identities, but this is not the case when the RR is learned by AD.

Contributions: We propose average moment estimators based on a new decomposition of the RR in terms of the moment-constrained function $\beta_\perp(z)$. Specifically, $\beta_\perp$ minimizes the mean squared error in predicting a known function $\beta(z)$ subject to $h(\beta_\perp) = 0$, where β is chosen such that one knows *a priori* that $h(\beta) \neq 0$. We propose an approach to learning $\beta_\perp$ and derive debiased estimators of Ψ based on initial ML estimates $\hat{\mu}$ and $\hat{\beta}_\perp$.

The advantage of learning the RR via $\beta_\perp$ rather than the AD loss is that the resulting RR estimates better control for extreme out-of-sample RR predictions since constants of proportionality in the RR are estimated using the estimation sample rather than the training sample. Moreover, our proposed estimator for $\beta_\perp$ is robust to overfitting issues that may arise when using the AD loss and thus is less sensitive to the tuning of optimization/model hyperparameters. Our proposal remains 'automatic' in the sense of not requiring the functional form of the RR to be derived. However, unlike AD, which only requires knowledge of the moment function m, we additionally require a known function β with $h(\beta) \neq 0$. The need to construct such a function may be viewed as a limitation of our proposal, though we contend that doing so is straightforward, as we demonstrate for Examples 1 to 4.

Though our approach is not tied to a particular machine learning method, we evaluate our estimators on two semi-synthetic datasets using multi-tasking neural networks, making comparisons with RieszNet (Chernozhukov et al., 2022a) for ADE/ATE estimation, and DragonNet (Shi et al., 2019), Reproducing Kernel Hilbert Space (RKHS) Embedding (Singh et al., 2023), Neural Net (NN) Embedding (Xu and Gretton, 2023) for ATE estimation. To ensure a fair comparison we re-implement RieszNet and DragonNet learners and estimators, with reproduction code available at `https://github.com/crimbs/madnet`.

2. Estimation

2.1. Debiased estimation

Given a sample of n i.i.d. observations, estimators of Ψ are typically based on the empirical distribution $\mathbb{E}_n[.] = n^{-1}\sum_{i=1}^{n}(.)_i$, the regression function $\hat{\mu}$, and the estimated RR function $\hat{\alpha}$. An estimator $\hat{\psi}$ of Ψ is said to be regular asymptotically linear (RAL) if its error behaves like an empirical process, i.e. $\sqrt{n}(\hat{\psi} - \Psi) = G_n[\varphi(W)] + o_p(1)$ for some finite variance function $\varphi(W)$, where $G_n[.] \equiv \sqrt{n}(\mathbb{E}_n[.] - \mathbb{E}[.])$ is the empirical process operator. RAL estimators are unbiased since, by the central limit theorem, $\sqrt{n}(\hat{\psi} - \Psi)$ converges (in distribution) to a mean zero normal random variable with variance $\text{var}[\varphi(W)]$. Results from nonparametric efficiency theory (Newey, 1994) imply that $\text{var}[\varphi(W)]$ is minimized when $\varphi(W) = m(\mu, W) + \alpha(Z)\{Y - \mu(Z)\}$ is the uncentered influence curve (Hampel, 1974; Ichimura and Newey, 2022) of Ψ, also called the pseudo-outcome (Kennedy, 2023; Hines et al., 2023) (see Hines et al. (2022); Kennedy (2022) for pedagogical reviews). Thus, one can construct standard errors for $\hat{\psi}$ by approximating φ with some $\hat{\varphi}$ and taking a sample variance. To consider specific estimators we use the identity

$$\sqrt{n}(\hat{\psi} - \Psi) = G_n\left[\varphi(W)\right] - \underbrace{\sqrt{n}\mathbb{E}_n\left[\hat{\varphi}(W) - \hat{\psi}\right]}_{\text{plug-in bias}} + \underbrace{\sqrt{n}\mathbb{E}[\hat{\varphi}(W) - \Psi]}_{\text{first-order remainder}} + \underbrace{G_n\left[\hat{\varphi}(W) - \varphi(W)\right]}_{\text{second-order remainder}} \tag{1}$$

which can be shown to hold (by canceling terms on the right hand side) for any $\hat{\psi}$ and pair of measurable functions $\varphi, \hat{\varphi}$. The second-order remainder above is usually not a concern and is $o_p(1)$ under weak assumptions, e.g. when $\mathbb{E}[\{\varphi(W) - \hat{\varphi}(W)\}^2] = o_p(1)$ and $\hat{\varphi}$ is obtained from an independent sample. In practice, the latter assumption motivates estimators which apply some form of sample-splitting/cross-fitting to estimate $\hat{\varphi}$ and evaluate the estimator (Zheng and van der Laan, 2011; Chernozhukov et al., 2018).

Letting $h_n(f) \equiv \mathbb{E}_n[m(f, W)]$, a naive estimator is $\hat{\Psi}^{(\text{Direct})} \equiv h_n(\hat{\mu})$, which does not use the RR estimates $\hat{\alpha}$. To examine the bias properties of the naive estimator, consider (1) when $\hat{\psi} = \hat{\Psi}^{(\text{Direct})}$ and $\hat{\varphi}(W) = m(\hat{\mu}, W) + \hat{\alpha}(Z)\{Y - \hat{\mu}(Z)\}$. Following (Chernozhukov et al., 2020), the first-order remainder reduces to the 'mixed bias', $-\sqrt{n}\langle\hat{\mu} - \mu, \hat{\alpha} - \alpha\rangle$, the square of which is bound by Cauchy-Schwarz as $n\langle\hat{\mu} - \mu, \hat{\alpha} - \alpha\rangle^2 \leq n||\hat{\mu} - \mu||^2||\hat{\alpha} - \alpha||^2$. The first-order remainder will therefore be $o_p(1)$ when $\hat{\mu}$ or $\hat{\alpha}$ converge to their true counterparts at sufficiently fast rates with sample size. Moreover, one can trade off accuracy in $\hat{\mu}$ and $\hat{\alpha}$, a property known as rate double robustness.

The plug-in bias $\sqrt{n}\mathbb{E}_n[\hat{\varphi}(W) - \hat{\Psi}^{(\text{Direct})}] = \sqrt{n}\mathbb{E}_n[\hat{\alpha}(Z)\{Y - \hat{\mu}(Z)\}]$, however, is generally not $o_p(1)$, and hence $\hat{\Psi}^{(\text{Direct})}$ is not RAL. The main challenge in obtaining RAL estimators is therefore removing plug-in biases and there are two main strategies for doing so. One-step debiased estimators simply add the plug-in bias to both sides of (1), resulting in the (double robust) RAL estimator $\hat{\Psi}^{(\text{DR})} \equiv \mathbb{E}_n[\hat{\varphi}(W)]$, or equivalently

$$\hat{\Psi}^{(\text{DR})} = \underbrace{h_n(\hat{\mu})}_{\text{direct estimator}} + \underbrace{\mathbb{E}_n[\hat{\alpha}(Z)\{Y - \hat{\mu}(Z)\}]}_{\text{bias correction}}. \tag{2}$$

Targeted maximum likelihood estimators (TMLEs) are direct estimators of the form $\hat{\Psi}^{(\text{TMLE})} \equiv h_n(\hat{\mu}^*)$ where $\hat{\mu}$ is replaced with a 'targeted' alternative $\hat{\mu}^*$ that solves

$$\mathbb{E}_n[\hat{\alpha}(Z)\{Y - \hat{\mu}^*(Z)\}] = 0. \tag{3}$$

Like one-step debiased estimators, TMLEs are double robust, with the mixed bias condition $\langle \hat{\mu}^* - \mu, \hat{\alpha} - \alpha \rangle = o_p(n^{-1/2})$. Targeting can be achieved in many ways, for example by first defining the linear parametric submodel $\hat{\mu}_t(z) = \hat{\mu}(z) + t\hat{\alpha}(z)$ for a univariate indexing parameter $t \in \mathbb{R}$. Then obtaining an optimal $t^* = \arg\min_{t\in\mathbb{R}} \mathbb{E}_n[\{Y - \hat{\mu}_t(Z)\}^2]$ which improves the fit of the outcome learner in the estimation sample, and ensures that $\hat{\mu}^* = \hat{\mu}_{t^*}$ is a solution to (3). For the linear parametric submodel $t^* = \mathbb{E}_n[\hat{\alpha}(Z)\{Y - \hat{\mu}(Z)\}]/\mathbb{E}_n[\hat{\alpha}^2(Z)]$ and $\hat{\Psi}^{(\text{TMLE})} = h_n(\hat{\mu}) + t^* h_n(\hat{\alpha})$, or equivalently

$$\hat{\Psi}^{(\text{TMLE})} = \underbrace{h_n(\hat{\mu})}_{\text{direct estimator}} + \underbrace{\left(\frac{h_n(\hat{\alpha})}{\mathbb{E}_n[\hat{\alpha}^2(Z)]}\right)}_{\text{scale factor}} \underbrace{\mathbb{E}_n[\hat{\alpha}(Z)\{Y - \hat{\mu}(Z)\}]}_{\text{bias correction}}. \tag{4}$$

Comparing $\hat{\Psi}^{(\text{TMLE})}$ with $\hat{\Psi}^{(\text{DR})}$ in (2), we see that the TMLE introduces a scale factor $h_n(\hat{\alpha})/\mathbb{E}_n[\hat{\alpha}^2(Z)]$, which is an empirical approximation to the population value $h(\alpha)/||\alpha||^2 = 1$, thereby rescaling the bias correction of the DR estimator. Variations of the TMLE method often include canonical generalized linear model (GLM) link functions in the parametric submodel definition, and maximize the associated GLM log-likelihood (hence the name TMLE), see e.g. (van der Laan and Gruber, 2016) for submodel proposals. GLM variations of this type may be used e.g. when Y is binary and a cross-entropy outcome loss is preferred. Finally, to motivate new RR learning methodologies, we remark that (3) and (4) are invariant to constants of proportionality in $\hat{\alpha}$, thus one might consider RR learners that are agnostic to such constants.

2.2. Debiased estimation with moment constraints

Our main contribution is to propose average moment estimators based on the identity

$$\alpha(z) = \frac{h(\beta)}{||\beta - \beta_\perp||^2}\{\beta(z) - \beta_\perp(z)\} \tag{5}$$

where $\beta \in \mathcal{H}$ is a known function with $h(\beta) \neq 0$ and $\perp$ denotes projection on to orthogonal complement set $\mathcal{C}^\perp \equiv \{f \in \mathcal{H} \mid h(f) = 0\}$. Specifically, $\beta_\perp = \arg\min_{f\in\mathcal{C}^\perp} ||\beta - f||$. A geometric illustration of this result is provided in Figure 1.

Proof of (5): *Note that $\mathcal{C}^\perp = \{f \in \mathcal{H} \mid \langle f, \alpha \rangle = 0\}$. By Hilbert's projection theorem, $\beta_\perp$ exists, with*

$$\beta_\perp(z) \equiv \beta(z) - \frac{\langle \beta, \alpha \rangle}{||\alpha||^2}\alpha(z) \quad \Longleftrightarrow \quad \alpha(z) = \frac{||\alpha||^2}{h(\beta)}\{\beta(z) - \beta_\perp(z)\}$$

where we use $\langle \beta, \alpha \rangle = h(\beta)$. Taking the norm of both sides and solving for $||\alpha|| \neq 0$ gives $||\alpha|| = |h(\beta)|/||\beta - \beta_\perp||$ which completes the proof.

The identity in (5) offers new avenues for debiased estimation of Ψ via learning $\beta_\perp$ and μ. One limitation of this proposal, however, is that we require construction of a function

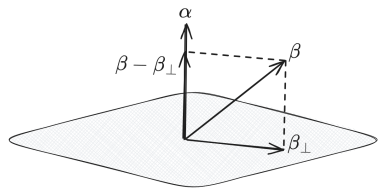

Figure 1: Illustration of moment-constrained functions. The plane represents the space of zero average moment functions, i.e. f such that $h(f) = \langle f, \alpha \rangle = 0$. The non-zero function $\beta - \beta_\perp$ is orthogonal to the plane, and thus is a scalar multiple of α.

$\beta \in \mathcal{H}$ such that $h(\beta) \neq 0$ *a priori* (and $||\beta|| < \infty$ since $\beta \in \mathcal{H}$). Due to the variety of possible moment functionals $m(f, W)$, we do not offer a general algorithm for constructing such functions, however, we have found in practice that there are usually candidates for β, where $m(\beta, W)$ is a known constant, and hence $h(\beta)$ is known *a priori*. For instance, for the ADE we set $m(\beta, W) = \beta'(A, X) = 1$ and integrate to obtain $\beta(z) = \beta(a, x) = a$ with $h(\beta) = 1$ *a priori*. This choice is not unique, however, since $\beta(z) = \exp(a)$ is also a valid choice for the ADE, with $h(\beta) = \mathbb{E}[\exp(A)] \neq 0$. As for the other estimands in Examples 1 to 4: for the ATE $h(\beta) = 1$ when $\beta(z) = a$; for the APE, $h(\beta) = 1$ when $\beta(z) = a + 1 - \pi(x)$; for the IPE, $h(\beta) = 1$ when $\beta(z) = a/\pi(x)$, or if there is concern that $\pi(X)$ can be zero, one can instead let $\beta(z) = a/\pi(x)$ when $\pi(x) \neq 0$ and $\beta(z) = 0$ otherwise, then estimate $h(\beta) = \Pr[\pi(X) \neq 0]$. For full generality we develop estimators in the setting where $h(\beta)$ must be estimated, but our results simplify slightly when $h(\beta)$ is known.

Example ATE: *Denoting $\beta(z) = \beta(a, x) = a$, the known RR and (5) imply*

$$\beta_\perp(a, x) = p(x) + \{a - p(x)\} \left(1 - \frac{1}{p(x)\{1 - p(x)\}} \mathbb{E} \left[\frac{1}{p(X)\{1 - p(X)\}} \right]^{-1} \right).$$

It is insightful to compare $\beta_\perp$, which minimizes $\mathbb{E}[\{A - f(A, X)\}^2]$ given $\mathbb{E}[f(1, X)] = \mathbb{E}[f(0, X)]$, with $p(x)$, which minimizes the same mean squared error, under the stronger constraint $f(1, X) = f(0, X)$. We notice that $p(x)$ lies on the interval $(0, 1)$, but the same is not true of $\beta_\perp$, which has weaker restrictions on its outputs: $\beta_\perp(1, x) < 1$ and $\beta_\perp(0, x) > 0$. Also p and $\beta_\perp$ are related by the identity $p(x) = \mathbb{E}[\beta_\perp(A, X)|X = x]$.

Example ADE: *Denoting $\beta(z) = \beta(a, x) = a$, the known RR and (5) imply*

$$\beta_\perp(a, x) = a - \mathbb{E} \left[\left(\frac{p'(A|X)}{p(A|X)} \right)^2 \right]^{-1} \frac{p'(a|x)}{p(a|x)},$$

which minimizes $\mathbb{E}[\{A - f(A, X)\}^2]$ given $\mathbb{E}[f'(A, X)] = 0$. As in the previous example, we remark that $\mathbb{E}[\beta_\perp(A, X)|X = x] = \mathbb{E}[A|X = x]$.

In Section 2.3 below, we propose methods for learning $\beta_\perp$, however, we first show how the standard debiased estimators from Section 2.1 look when debiasing is achieved using initial estimates $\hat{\beta}_\perp$ instead of $\hat{\alpha}$. Specifically, we consider

$$\hat{\alpha}(z) = \frac{h_n(\beta - \hat{\beta}_\perp)\{\beta(z) - \hat{\beta}_\perp(z)\}}{\mathbb{E}_n[\{\beta(Z) - \hat{\beta}_\perp(Z)\}^2]}.$$

Under this parameterization, the DR estimator in (2) and TMLE in (4) both become

$$\hat{\Psi}^{(\perp,\mathrm{DR})} \equiv \underbrace{h_n(\hat{\mu})}_{\text{direct estimator}} + \underbrace{\left(\frac{h_n(\beta - \hat{\beta}_\perp)}{\mathbb{E}_n[\{\beta(Z) - \hat{\beta}_\perp(Z)\}^2]}\right)}_{\text{scale factor}} \underbrace{\mathbb{E}_n[\{\beta(Z) - \hat{\beta}_\perp(Z)\}\{Y - \hat{\mu}(Z)\}]}_{\text{unscaled bias correction}}, \quad (6)$$

which is of the same form as (4) but with $\beta - \hat{\beta}_\perp$ replacing $\hat{\alpha}$. Moreover, since the the TMLE score equation in (3) only requires estimating the RR up to constants of proportionality, TMLEs can be derived using parametric submodels where $\hat{\alpha}$ is replaced with $\beta - \hat{\beta}_\perp$, e.g. by using the linear submodel $\hat{\mu}_t(z) = \hat{\mu}(z) + t\{\beta(z) - \hat{\beta}_\perp(z)\}$, which gives $h_n(\hat{\mu}^*) = \hat{\Psi}^{(\perp,\mathrm{DR})}$. Theorem 1 gives general conditions under which a TMLE $h_n(\hat{\mu}^*)$ based on $\hat{\beta}_\perp$ is RAL, with $\hat{\Psi}^{(\perp,\mathrm{DR})}$ being a special case. Essentially, this theorem controls the first-order remainder of (1) using the mixed-bias condition $\langle \hat{\mu}^* - \mu, \hat{\beta}_\perp - \beta_\perp \rangle = o_p(n^{-1/2})$, and uses empirical process assumptions to control the second-order remainder of (1).

Theorem 1 *Let $\hat{\beta}_\perp$ be an estimator for $\beta_\perp$, and let $\hat{\mu}^*$ be an estimator for μ that is targeted such that $\mathbb{E}_n[\{\beta(Z) - \hat{\beta}_\perp(Z)\}\{Y - \hat{\mu}^*(Z)\}] = 0$. Assume that each of the following terms are $o_p(1)$: $\sqrt{n}\langle \hat{\mu}^* - \mu, \hat{\beta}_\perp - \beta_\perp \rangle$, $G_n[m(\hat{\mu}^* - \mu, W)]$, $G_n\left[\{\hat{\beta}_\perp(Z) - \beta_\perp(Z)\}\{\hat{\mu}^*(Z) - \mu(Z)\}\right]$, $G_n\left[\{\beta(Z) - \beta_\perp(Z)\}\{\hat{\mu}^*(Z) - \mu(Z)\}\right]$, and $G_n\left[\{\hat{\beta}_\perp(Z) - \beta_\perp(Z)\}\{Y - \mu(Z)\}\right]$. Then $h_n(\hat{\mu}^*)$ is a RAL estimator of Ψ with uncentered influence curve $\varphi(W)$, and hence $\sqrt{n}(h_n(\hat{\mu}^*) - \Psi)$ converges in distribution to a mean-zero normal random variable with variance* $\mathrm{var}[\varphi(W)]$. *Proof in Supplement B.3.*

2.3. Moment-constrained learning

We propose learners for the moment-constrained function $\beta_\perp$ using the property that $\beta_\perp$ is the function $f \in \mathcal{H}$ that solves

$$\begin{aligned} \text{minimize:} \quad & ||\beta - f||^2 \\ \text{subject to:} \quad & h(f) = 0 \end{aligned} \quad (7)$$

Similar constrained learning problems have been studied in the context of ML with fairness constraints (Nabi et al., 2024). E.g. Zafar et al. (2017); Akhtar et al. (2021) minimize a classification loss, while ensuring that predictions are uncorrelated with specific sensitive attributes (race, sex etc.). The primal problem in (7) is characterized by the Lagrangian

$$\mathcal{L}(f, \lambda) \equiv \mathbb{E}\left[\{\beta(Z) - f(Z)\}^2 + \lambda m(f, W)\right] \quad (8)$$

where $\lambda \in \mathbb{R}$ is a Lagrange multiplier, and a solution is obtained by finding f^* and λ^* such that $\mathcal{L}(f^*, \lambda^*) = \max_{\lambda \in \mathbb{R}} \min_{f \in \mathcal{H}} \mathcal{L}(f, \lambda)$. Naively, therefore, one might learn f by performing gradient descent over parameters indexing f and gradient ascent on λ, as in the basic differential multiplier method (BDMM) of Platt and Barr (1987).

In our numerical experiments, we consider the setting where $f = f_w$ is the output of a multilayer perceptron (MLP) with weights w. We observe that application of BDMM to a sample analogue of $\mathcal{L}(f_w, \lambda)$ leads to empirical constraint violations that oscillate

around zero, as the number of ascent/descent iterations increases (shown in Figure 4 of the supplement). Similar behavior is documented elsewhere for adversarial function learners (Schäfer and Anandkumar, 2019; Mokhtari et al., 2020). Instead, stable constraint violations were achieved by effectively replacing the constraint in (7) with the equivalent constraint $|h(f)| \leq 0$, yielding the Lagrangian

$$\tilde{\mathcal{L}}(f, \tilde{\lambda}) \equiv \mathbb{E}\left[\{\beta(Z) - f(Z)\}^2\right] + \tilde{\lambda}|h(f)| \tag{9}$$

with empirical MLP analogue $\tilde{\mathcal{L}}_n(f_w, \tilde{\lambda}) \equiv \mathbb{E}_n\left[\{\beta(Z) - f_w(Z)\}^2\right] + \tilde{\lambda}|h_n(f_w)|$. In this formulation, $\tilde{\lambda} \geq 0$ penalizes the sample average moment of f_w in a similar way to the smoothing parameters in conventional Lasso/ridge regression. The key difference between these classical methods, however, is that the penalty $|h_n(f_w)|$ depends on the observed data, and not only on the weights w. In practice we set $\tilde{\lambda}$ to a constant value during training, and minimize over w using gradient descent, though one might consider alternative methods e.g. where $\tilde{\lambda}$ increases monotonically with the number of descent iterations (epochs).

2.4. Comparison with Automatic debiasing (AD)

AD (Chernozhukov et al., 2021, 2022a) is an RR learning method based on the identity

$$\begin{aligned}
\alpha &= \arg\min_{\hat{\alpha} \in \mathcal{H}} ||\alpha - \hat{\alpha}||^2 \\
&= \arg\min_{\hat{\alpha} \in \mathcal{H}} ||\hat{\alpha}||^2 - 2\langle \hat{\alpha}, \alpha \rangle \\
&= \arg\min_{\hat{\alpha} \in \mathcal{H}} \mathbb{E}\left[\hat{\alpha}(Z)^2 - 2m(\hat{\alpha}, W)\right].
\end{aligned}$$

The AD RR learner minimizes a sample analogue of this expectation. We connect AD to our proposal as follows. Consider that $\hat{\alpha} \in \mathcal{H}$ can be written as $\hat{\alpha}(z) = 2\lambda^{-1}\{\beta(z) - f(z)\}$ where $\lambda \neq 0$ is constant and $f \in \mathcal{H}$. Thus, the AD population minimization becomes

$$\begin{aligned}
\beta_\lambda &\equiv \arg\min_{f \in \mathcal{H}} \mathbb{E}\left[4\lambda^{-2}\{\beta(Z) - f(Z)\}^2 - 4\lambda^{-1} m(\beta - f, W)\right] \\
&= \arg\min_{f \in \mathcal{H}} \mathbb{E}\left[\{\beta(Z) - f(Z)\}^2 + \lambda m(f, W)\right].
\end{aligned}$$

with $\alpha(z) = 2\lambda^{-1}\{\beta(z) - \beta_\lambda(z)\}$. The objective above is the Lagrangian $\mathcal{L}(f, \lambda)$ in (8). In this construction λ is unrestricted, therefore provided that $\lambda_\perp \equiv \arg\max_{\lambda \in \mathbb{R}} \min_{f \in \mathcal{H}} \mathcal{L}(f, \lambda)$ is finite and non-zero, one can write $\alpha(z) = 2\lambda_\perp^{-1}\{\beta(z) - \beta_{\lambda_\perp}(z)\}$. Appealing to the primal problem in (7), we see that $\beta_{\lambda_\perp} = \beta_\perp$ and, $\lambda_\perp = 2||\beta - \beta_\perp||^2/h(\beta)$ as in (5).

Moment-constrained learning, therefore, reinterprets the AD loss as a Lagrangian when one has access to a function β such that $h(\beta) \neq 0$, and one is agnostic to constants of proportionality in the RR. When estimating Ψ, these constants are estimated using the estimation sample rather than by the RR learner directly. By connecting the RR to the primal problem, one is able to leverage constrained learning methods that may have better empirical performance, e.g. using the Lagrangian $\tilde{\mathcal{L}}$ in (9). Moreover, this connection hints at future theoretical study of RRs and AD via constrained function learning theory.

3. MADNet: Moment-constrained Automatic Debiasing Networks

Multi-headed MLPs, as illustrated through the schematic in Figure 2, are emerging as a popular architecture for estimating average moment estimands using deep learning. Initial efforts focused on the binary treatment setting, such as the multi-headed MLP outcome learner TARNet (Treatment Agnostic Representation Network) (Shalit et al., 2017). TARNet takes inputs X and produces two scalar outputs representing $\mu(1, X)$ and $\mu(0, X)$ respectively. During training, an outcome prediction error (e.g. mean-squared error) is minimized, with predictions of $\mu(A, X)$ obtained from one of the scalar outputs according to whether an observation is treated/untreated. The resulting outcome learner can be used to obtain plug-in estimates for e.g. the ATE/APE, optionally with debiasing by a separate RR learner as described in Section 2.1.

DragonNet (Shi et al., 2019) also focuses on binary treatments, extending TARNet by introducing a third scalar output MLP of zero depth, which is used to estimate the propensity score $p(X)$, and hence the RR of the ATE/APE. The authors reason that the propensity score MLP should have zero depth so that the shared MLP learns representations of X that are predictive of the RR, since, for ATE estimation it is sufficient to learn the outcome conditional on A and $p(X)$ only. This approach is generalized by Chernozhukov et al. (2022a, Lemma 3.1), where it is shown that: for estimation of Ψ it is sufficient to learn the outcome conditional on the RR only, i.e $E[Y|\alpha(Z) = \alpha(z)]$. Similarly, for estimation of Ψ, we show (Supplement B.4) that it is sufficient to learn the outcome conditional on $\beta(Z) - \beta_\perp(Z)$ only. Moreover, we show (Supplement B.5) that $\mu(Z) = \mu_\perp(Z) + \Psi\{\beta(Z) - \beta_\perp(Z)\}/h(\beta)$ where $\mu_\perp$ is the projection of μ on $\mathcal{C}^\perp$. This result further highlights the role of the RR when learning the outcome for average moment estimation.

RieszNet is similar in structure to DragonNet, except with input $Z = (A, X)$ rather than X. Both also use a multi-tasking loss to learn μ and α simultaneously. We propose MADNet which uses the same network structure as RieszNet and a multi-tasking loss to learn μ and $\beta_\perp$ simultaneously. Specifically we consider the loss $\tilde{\mathcal{L}}_n(f_{w,1}, \tilde{\lambda}) + \rho\,\text{REGLoss}_n(f_{w,2})$ where $\rho \geq 0$ is a hyperparameter, REGLoss_n is a regression loss, e.g. the mean-squared error in the outcome prediction $\text{REGLoss}_n(f) = \mathbb{E}_n[\{Y - f(Z)\}^2]$, and $f_w(z) = (f_{w,1}(z), f_{w,2}(z))$ represents two outputs from a multi-headed MLP. Note that e.g. $f_{w,1}$ depends only on the weights of the shared MLP and the first non-shared MLP, but we write it as a function of all multi-headed MLP weights w for convenience. Like RieszNet, for ATE/ADE estimation we replace $f_{w,2}(z)$ in the regression loss with $\tilde{a} f_{w,2}(z) + (1 - \tilde{a}) f_{w,3}(z)$, where $\tilde{a}$ represents the min-max normalized treatment a scaled on to the interval $[0, 1]$, i.e. $\tilde{a}_i = \{a_i - \min_n(a_i)\}/\{\max_n(a_i) - \min_n(a_i)\}$.

Convergence rates: The standard theory in Section 2.1 controls first-order remainders by requiring estimators to converge to their true counterparts at sufficiently fast rates. For neural network learners, recent convergence rate results have been obtained using the theory of critical radii (Wainwright, 2019; Foster and Syrgkanis, 2023; Chernozhukov et al., 2020). In particular, results are provided for MLPs with Rectified Linear Unit (ReLU) activation functions (Farrell et al., 2021), describing L_2 convergence rates in terms of the number of training observations and the network width/depth. Similar results exist for AD learners with moments satisfying the mean-squared continuity property $\mathbb{E}[\{m(u, W) - m(v, W)\}^2] \leq M||u - v||^2$ for $M \geq 1$ and $u, v \in \mathcal{A}_n$, where $\mathcal{A}_n$ is a function set described by Chernozhukov

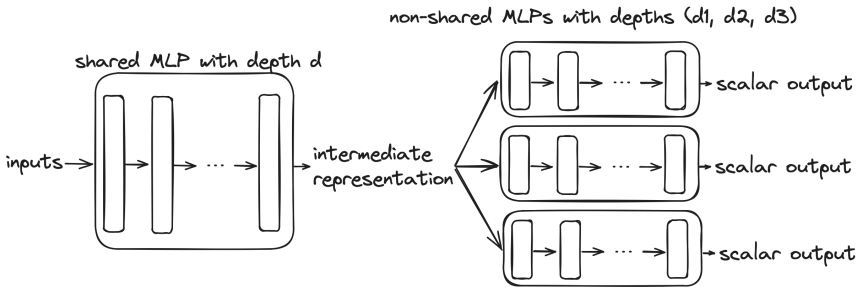

Figure 2: Multi-headed MLP architecture with three outputs. Typically the intermediate representation has the same width as the internal layers of the shared MLP, and the non-shared MLPs have internal layers with half the width of the shared MLP. During training, a single loss is used based on all scalar outputs, and MLP weights are learned using back-propagation over the entire multi-headed MLP.

et al. (2021). We expect similar theoretical guarantees to hold in the current context due to the connection of AD with moment-constrained learning discussed above.

Sample-splitting: To obtain valid inference from a single sample, the standard theory in Section 2.1 relies on cross-fitting to control the second-order remainder terms in (1). Cross-fitting is used to bypass Donsker class assumptions (Bickel, 1982; Zheng and van der Laan, 2011; Chernozhukov et al., 2018), which restrict the complexity of the initial estimators and are usually not satisfied by ML algorithms. Recent work has sought to bypass Donsker conditions by instead relying on entropic arguments (van de Geer et al., 2014) or leave-one-out stability (Chen et al., 2022). In practice DragonNet and RieszNet do not use sample splitting due to the associated computational burden. Instead, implementations of both algorithms split the data into training and validation sets, with the validation set used to control early stopping of the training algorithm. For estimation, the full dataset is used (training + validation). We also use this strategy for moment-constrained learning.

4. Numerical experiments

We consider ATE and ADE estimation in the following semi-synthetic data scenarios.

IHDP: (Infant Health and Development Program). IHDP is a randomized experiment on the effects of home visits by specialists (binary treatment, A) on infant cognition scores (Y), given 25 baseline covariates. The data consists of n = 747 infants. Synthetic outcomes are drawn from a normal distribution given (A, X), as described by (Hill, 2011). We consider 1000 synthetic IHDP datasets in total.

BHP: (Blundell, Horowitz and Parey) (Blundell et al., 2017). BHP consists of 3,640 household level observations from the 2001 (U.S.) National Household Travel Survey, with the goal of estimating the price elasticity of gasoline consumption given 18 confounding variables. Price elasticity can be defined through the ADE of log price (A) on the log quantity of gasoline sold (Y). Following (Chernozhukov et al., 2022a), we draw synthetic treatments from a normal distribution, with conditional mean and variance obtained from random forest predictions of the mean and variance of the true log price. Conditionally

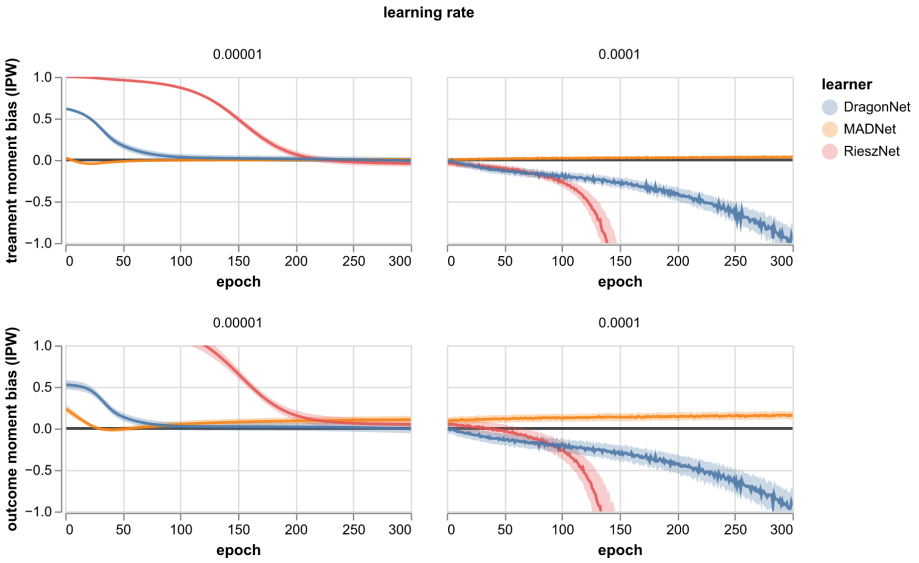

Figure 3: Mean and standard error of $\mathbb{E}_n[A\hat{\alpha}(Z)] - 1$ (top row) and of $\hat{\Psi}^{(\text{IPW})} - h_n(\mu)$ (bottom row) using 20 datasets of IHDP data where predictions are made on a 20% validation set and the outcome is scaled by its standard deviation.

normal synthetic outcomes are then generated, given (A, X), with a mean function that is cubic in treatment. The results in Table 1 are evaluated over 200 random seeds.

The first experimental research question is: to what extent do RR predictions from moment-constrained auto-debiasing learners satisfy oracle RR properties that are known *a priori*? To answer this, we consider Inverse Probability Weighted (IPW) estimators of $\mathbb{E}[A\alpha(Z)] = 1$, and Ψ. The corresponding IPW estimators $\mathbb{E}_n[A\hat{\alpha}(Z)]$ and $\hat{\Psi}^{(\text{IPW})} \equiv \mathbb{E}_n[Y\hat{\alpha}(Z)]$ do not depend on the outcome model, thus are a convenient way of comparing RR learners. Figure 3 shows, using IHDP data, how the mean error evolves over gradient descent iterations (epochs) for two different learning rates. These plots show that the MADNet RR estimator converges rapidly to a stable optimum, and is therefore more robust to changes in learning rate and early stopping hyperparameters.

Next we compare absolute errors of MADNet estimators versus several alternatives: DragonNet (Shi et al., 2019), Reproducing Kernel Hilbert Space (RKHS) Embedding (Singh et al., 2023), Neural Net (NN) Embedding (Xu and Gretton, 2023), and RieszNet (Chernozhukov et al., 2022a), with only the latter applying to ADE estimation in the BHP scenario. The main proposal of our work is the Double Robust MADNet estimator, though to examine the effect of RR based bias correction, we have also provide results for the Direct (outcome model only without RR bias corrections), and IPW (RR model only) estimators.

Results in Table 1 show that the MADNet DR estimator has improved empirical performance (reduced mean absolute error) versus all alternatives in both scenarios. Interestingly, for the IHDP scenario, this improvement in performance is observed despite the MADNet Direct estimator performing poorly compared to its RieszNet counterpart. Moreover, the MADNet DR estimator effectively combines two estimators with poor performance (Direct

and IPW) to create an estimator with good performance. One possible explanation for this could be that MADNet results in estimators for which the mixed bias term $\langle \hat{\mu}-\mu, \hat{\beta}_\perp - \beta_\perp \rangle$ is small, i.e. errors in the outcome and RR learners are uncorrelated. Further work is needed, however, to examine this effect. Similar results are obtained when the multi-headed architectures of MADNet and RieszNet are ablated and replaced with simple feedforward NNs (Supplement Table 2). Furthermore, MADNet results with a reduced constraint penalization ($\tilde{\lambda} = 1$) show slightly worse performance, highlighting the importance of satisfying the constraint in our procedure (Supplement Table 3).

Detailed implementation notes are provided in Supplement A alongside DragonNet/RieszNet results obtained using our implementation (Table 2). We highlight, however, three main differences in our implementation: (i) our implementation is built on a JAX + Equinox computational stack (Bradbury et al., 2018; Kidger and Garcia, 2021); (ii) for ADE estimation we use automatic differentiation of NN outputs, rather than a finite difference approximation; (iii) the original RieszNet uses a complicated learning rate / early stopping scheme but we use a slightly simpler scheme. Overall we found that replication via re-implementation of the RieszNet results was challenging, possibly due to the stability issues in Figure 3.

Table 1: Absolute error (mean ± standard error) of the ATE and ADE estimates for both semi-synthetic data scenarios. For the RieszNet IHDP benchmark we report values obtained by running `RieszNet_IHDP.ipynb` from https://github.com/victor5as/RieszLearning without modification. MAEs for the BHP scenario are not reported by Chernozhukov et al. (2022a), and we instead use values from our RieszNet re-implementation. DragonNet and RKHS/NN embedding values are retrieved from Xu and Gretton (2023, Table 1).

Estimator	IHDP	BHP	Citation
DragonNet (DR)	0.146 ± 0.010	–	(Shi et al., 2019)
RKHS Embedding	0.166 ± 0.003	–	(Singh et al., 2023)
NN Embedding	0.117 ± 0.002	–	(Xu and Gretton, 2023)
RieszNet (Direct)	0.128 ± 0.004	0.692 ± 0.040	(Chernozhukov et al., 2022a)
RieszNet (IPW)	0.789 ± 0.036	0.449 ± 0.025	(Chernozhukov et al., 2022a)
RieszNet (DR)	0.114 ± 0.003	0.428 ± 0.023	(Chernozhukov et al., 2022a)
MADNet (Direct)	0.504 ± 0.016	0.471 ± 0.026	Proposed
MADNet (IPW)	0.719 ± 0.039	0.474 ± 0.026	Proposed
MADNet (DR)	**0.094 ± 0.002**	**0.391 ± 0.019**	Proposed

5. Conclusion

We present a new algorithm for estimating average moment estimands. Our approach leverages functions for which the average moment is known *a priori* to be non-zero, though the need to construct such functions may also be viewed as a limitation. Nonetheless, we contend that constructing such functions is significantly simpler than deriving the functional form of the RR, which is required for non-automatic RR learning methods. Our proposal

is therefore 'automatic' in the sense of not requiring complicated estimand-specific machinery. Moreover, rather than learning the full RR, as in conventional AD, we instead learn a moment-constrained function that is sufficient for debiasing the naive (direct) average moment estimator. From a practical perspective, the need to learn a constrained function requires additional techniques and hyperparameter tuning to ensure that constraints are approximately satisfied. In our work, we propose a Lagrange-type penalization method for moment-constrained learning and apply this method using multi-tasking neural networks, though experimenting with other approaches to constrained function learning represents an important direction for future study.

There are several other directions which one might extend our work. First, our set up considers estimands that represent the average moment of a regression functions, but extensions of AD to so-called generalized regressions (e.g. quantile functions) could also be considered (Chernozhukov et al., 2021). Second, we consider neural network learners, but similar extensions for gradient boosted trees / random forests should also be possible. Finally, our approach to moment-constrained function learning may be applied to other problems with stochastic constraints e.g. those related to fairness of ML predictions.

Acknowledgments

We acknowledge insightful conversations on this topic with Alejandro Schuler and Stijn Vansteelandt.

References

Zeeshan Akhtar, Amrit Bedi, and Ketan Rajawat. Conservative Stochastic Optimization with Expectation Constraints. *IEEE Transactions on Signal Processing*, 69:3190–3205, 2021. ISSN 19410476. doi: 10.1109/TSP.2021.3082467.

Susan Athey and Stefan Wager. Policy Learning With Observational Data. *Econometrica*, 89(1):133–161, 2021. ISSN 0012-9682. doi: 10.3982/ecta15732.

Susan Athey, Guido W. Imbens, and Stefan 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. ISSN 14679868. doi: 10.1111/rssb.12268.

P. J. Bickel. On Adaptive Estimation. *The Annals of Statistics*, 10(3), sep 1982. ISSN 0090-5364. doi: 10.1214/aos/1176345863. URL https://projecteuclid.org/journals/annals-of-statistics/volume-10/issue-3/On-Adaptive-Estimation/10.1214/aos/1176345863.full.

Richard Blundell, Joel Horowitz, and Matthias Parey. Nonparametric estimation of a nonseparable demand function under the slutsky inequality restriction. *Review of Economics and Statistics*, 99(2):291–304, 2017. ISSN 15309142. doi: 10.1162/REST_a_00636.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs (v0.3.13), 2018. URL http://github.com/google/jax.

Matias D. Cattaneo, Richard K. Crump, and Michael Jansson. Generalized jackknife estimators of weighted average derivatives. *Journal of the American Statistical Association*, 108(504):1243–1256, 2013. ISSN 1537274X. doi: 10.1080/01621459.2012.745810.

Qizhao Chen, Vasilis Syrgkanis, and Morgane Austern. Debiased Machine Learning without Sample-Splitting for Stable Estimators. *Advances in Neural Information Processing Systems*, 35:1–48, 2022. ISSN 10495258.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney K. Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*, 21(1):C1–C68, 2018. ISSN 1368423X. doi: 10.1111/ectj.12097.

Victor Chernozhukov, Whitney K. Newey, Rahul Singh, and Vasilis Syrgkanis. Adversarial Estimation of Riesz Representers. *arXiv (2101.00009)*, 2020. URL http://arxiv.org/abs/2101.00009.

Victor Chernozhukov, Whitney K. Newey, Victor Quintas-Martinez, and Vasilis Syrgkanis. Automatic Debiased Machine Learning via Neural Nets for Generalized Linear Regression. *arXiv (2104.14737)*, 2021. URL http://arxiv.org/abs/2104.14737.

Victor Chernozhukov, Whitney K. Newey, Víctor Quintas-Martínez, and Vasilis Syrgkanis. RieszNet and ForestRiesz: Automatic Debiased Machine Learning with Neural Nets and Random Forests. *Proceedings of Machine Learning Research*, 162:3901–3914, 2022a. ISSN 26403498.

Victor Chernozhukov, Whitney K. Newey, and Rahul Singh. Debiased machine learning of global and local parameters using regularized Riesz representers. *The Econometrics Journal*, 25(3):576–601, sep 2022b. ISSN 1368-4221. doi: 10.1093/ectj/utac002. URL https://academic.oup.com/ectj/article/25/3/576/6572833.

Nishanth Dikkala, Greg Lewis, Lester Mackey, and Vasilis Syrgkanis. Minimax estimation of conditional moment models. *Advances in Neural Information Processing Systems*, 2020-December(NeurIPS), 2020. ISSN 10495258.

Miroslav Dudik, John Langford, and Hong Li. Doubly robust policy evaluation and learning. *Proceedings of the 28th International Conference on Machine Learning, ICML 2011*, pages 1097–1104, 2011.

Max H. Farrell, Tengyuan Liang, and Sanjog Misra. Deep Neural Networks for Estimation and Inference. *Econometrica*, 89(1):181–213, 2021. ISSN 0012-9682. doi: 10.3982/ecta16901.

Dylan J. Foster and Vasilis Syrgkanis. Orthogonal Statistical Learning. *Annals of Statistics*, 51(3):879–908, 2023. ISSN 21688966. doi: 10.1214/23-AOS2258.

Frank R Hampel. The Influence Curve and its Role in Robust Estimation. *Journal of the American Statistical Association*, 69(346):383–393, 1974.

Wolfgang Härdle and Thomas M. Stoker. Investigating smooth multiple regression by the method of average derivatives. *Journal of the American Statistical Association*, 84(408): 986–995, 1989. ISSN 1537274X. doi: 10.1080/01621459.1989.10478863.

Jennifer L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011. ISSN 10618600. doi: 10.1198/jcgs.2010.08162.

Oliver J. Hines, Oliver Dukes, Karla Diaz-Ordaz, and Stijn Vansteelandt. Demystifying Statistical Learning Based on Efficient Influence Functions. *The American Statistician*, 76(3):292–304, jul 2022. ISSN 0003-1305. doi: 10.1080/00031305.2021.2021984. URL https://doi.org/10.1080/00031305.2021.2021984https://www.tandfonline.com/doi/full/10.1080/00031305.2021.2021984.

Oliver J. Hines, Karla Diaz-Ordaz, and Stijn Vansteelandt. Variable importance measures for heterogeneous causal effects. *arXiv (2204.06030)*, 2023.

David A. Hirshberg and Stefan Wager. Augmented minimax linear estimation. *The Annals of Statistics*, 49(6), dec 2021. ISSN 0090-5364. doi: 10.1214/21-AOS2080. URL https://projecteuclid.org/journals/annals-of-statistics/volume-49/issue-6/Augmented-minimax-linear-estimation/10.1214/21-AOS2080.full.

Hidehiko Ichimura and Whitney K. Newey. The influence function of semiparametric estimators. *Quantitative Economics*, 13(1):29–61, 2022. ISSN 1759-7323. doi: 10.3982/qe826.

Guido W. Imbens and Whitney K. Newey. Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity. *Econometrica*, 77(5):1481–1512, 2009. ISSN 0012-9682. doi: 10.3982/ECTA7108. URL http://doi.wiley.com/10.3982/ECTA7108.

Edward H. Kennedy. Semiparametric doubly robust targeted double machine learning: a review. *arXiv (2203.06469)*, 2022. URL http://arxiv.org/abs/2203.06469.

Edward H. Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, 17(2):3008–3049, 2023. ISSN 19357524. doi: 10.1214/23-EJS2157.

Patrick Kidger and Cristian Garcia. Equinox: neural networks in JAX via callable PyTrees and filtered transformations. *Differentiable Programming workshop at Neural Information Processing Systems 2021*, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization, January 2019.

Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. A Unified Analysis of Extra-gradient and Optimistic Gradient Methods for Saddle Point Problems: Proximal Point Approach. *Proceedings of Machine Learning Research*, 108:1497–1507, 2020. ISSN 26403498.

Razieh Nabi, Nima S Hejazi, Mark J van der Laan, and David Benkeser. Statistical learning for constrained functional parameters in infinite-dimensional models with applications in fair machine learning. *arXiv (2404 . 09847)*, 2024.

Whitney K. Newey. The Asymptotic Variance of Semiparametric Estimators. *Econometrica*, 62(6):1349, nov 1994. ISSN 00129682. doi: 10.2307/2951752. URL https://www.jstor.org/stable/2951752?origin=crossref.

Whitney K. Newey and Thomas M. Stoker. Efficiency of Weighted Average Derivative Estimators and Index Models. *Econometrica*, 61(5):1199, sep 1993. ISSN 00129682. doi: 10.2307/2951498. URL https://www.jstor.org/stable/2951498?origin=crossref.

John Platt and Alan Barr. Constrained differential optimization. In D. Anderson, editor, *Neural Information Processing Systems*, volume 0. American Institute of Physics, 1987. URL https://proceedings.neurips.cc/paper_files/paper/1987/file/a87ff679a2f3e71d9181a67b7542122c-Paper.pdf.

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of Regression Coefficients When Some Regressors Are Not Always Observed. *Journal of the American Statistical Association*, 89(427):846, sep 1994. ISSN 01621459. doi: 10.2307/2290910. URL https://www.jstor.org/stable/2290910?origin=crossref.

Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983. ISSN 0006-3444. doi: 10.1093/biomet/70.1.41. URL https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/70.1.41.

Dominik Rothenhäusler and Bin Yu. Incremental causal effects. *arXiv (1907.13258)*, 2019. URL http://arxiv.org/abs/1907.13258.

Florian Schäfer and Anima Anandkumar. Competitive gradient descent. *Advances in Neural Information Processing Systems*, 32(NeurIPS), 2019. ISSN 10495258.

Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: Generalization bounds and algorithms. *34th International Conference on Machine Learning, ICML 2017*, 6:4709–4718, 2017.

Claudia Shi, David M. Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. *Advances in Neural Information Processing Systems*, 32:1–14, 2019. ISSN 10495258.

R Singh, L Xu, and A Gretton. Kernel methods for causal functions: dose, heterogeneous and incremental response curves. *Biometrika*, 2023. ISSN 0006-3444. doi: 10.1093/biomet/asad042.

Sara van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*, 42(3):1166–1202, 2014. ISSN 00905364. doi: 10.1214/14-AOS1221.

Mark J. van der Laan and Susan Gruber. One-Step Targeted Minimum Loss-based Estimation Based on Universal Least Favorable One-Dimensional Submodels. *International Journal of Biostatistics*, 12(1):351–378, 2016. ISSN 15574679. doi: 10.1515/ijb-2015-0054.

Mark J. van der Laan and Alexander R. Luedtke. Targeted Learning of the Mean Outcome under an Optimal Dynamic Treatment Rule. *Journal of Causal Inference*, 3(1):61–95, 2014. ISSN 2193-3677. doi: 10.1515/jci-2013-0022.

Martin J. Wainwright. *High-Dimensional Statistics*. Cambridge University Press, feb 2019. ISBN 9781108627771. doi: 10.1017/9781108627771. URL https://www.cambridge.org/core/product/identifier/9781108627771/type/book.

Liyuan Xu and Arthur Gretton. A neural mean embedding approach for back-door and front-door adjustment. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=rLguqxYvYHB.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness constraints: Mechanisms for fair classification. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017*, 54, 2017.

Wenjing Zheng and Mark J. van der Laan. Cross-validated targeted minimum-loss-based estimation. In *Targeted Learning*, pages 459–474. Springer New York, New York, NY, 2011. ISBN 978-1-4419-9781-4 978-1-4419-9782-1.

José R. Zubizarreta. Stable Weights that Balance Covariates for Estimation With Incomplete Outcome Data. *Journal of the American Statistical Association*, 110(511):910–922, 2015. ISSN 1537274X. doi: 10.1080/01621459.2015.1023805.