Learning density ratios in causal inference using Bregman–Riesz regression

Oliver J. Hines[1] and Caleb H. Miles[1]

[1]Columbia University, New York, NY, USA

October 21, 2025

Abstract

The ratio of two probability density functions is a fundamental quantity that appears in many areas of statistics and machine learning, including causal inference, reinforcement learning, covariate shift, outlier detection, independence testing, importance sampling, and diffusion modeling. Naively estimating the numerator and denominator densities separately using, e.g., kernel density estimators, can lead to unstable performance and suffers from the curse of dimensionality as the number of covariates increases. For this reason, several methods have been developed for estimating the density ratio directly based on (a) Bregman divergences or (b) recasting the density ratio as the odds in a probabilistic classification model that predicts whether an observation is sampled from the numerator or denominator distribution. Additionally, the density ratio can be viewed as the Riesz representer of a continuous linear map, making it amenable to estimation via (c) minimization of the so-called Riesz loss, which was developed to learn the Riesz representer in the Riesz regression procedure in causal inference. In this paper we show that all three of these methods can be unified in a common framework, which we call Bregman–Riesz regression. We further show how data augmentation techniques can be used to apply density ratio learning methods to causal problems, where the numerator distribution typically represents an unobserved intervention. We show through simulations how the choice of Bregman divergence and data augmentation strategy can affect the performance of the resulting density ratio learner. A Python package is provided for researchers to apply Bregman–Riesz regression in practice using gradient boosting, neural networks, and kernel methods.

1 Introduction

The ratio of two probability density functions (density ratio) is a fundamental quantity that appears in many areas of statistics and machine learning. In this article we focus on applications in causal inference; however, density ratios also appear when: reweighting data due to sample selection biases (Huang et al., 2006); detecting outliers / covariate shift in test data (Hido et al., 2011); detecting change points in time series (Liu et al., 2013); estimating mutual information (Nguyen et al., 2007; Belghazi et al., 2018); and in generative diffusion modeling (Meng et al., 2023; Lou et al., 2024). One reason for the ubiquity of density ratios is that they can be viewed as non-negative importance weights when expressing expectations with respect to the numerator distribution P_1, as weighted expectations with respect to the denominator distribution P_0. For this reason, density ratios appear frequently in casual

inference, where P_1 represents a target intervention distribution and P_0 represents the observed data generating distribution. For instance, in policy learning / offline reinforcement learning, one is interested in computing the expected loss if a treatment/action had been assigned according to a counterfactual decision rule which may depend on the natural value of the treatment (Díaz and van der Laan, 2012; Haneuse and Rotnitzky, 2013; Kallus and Uehara, 2020; Díaz et al., 2023) or its distribution (Kennedy, 2019).

Naively learning the numerator and denominator densities separately then taking their ratio can lead to unstable estimates of the density ratio. A well-known example of this phenomenon in causal inference occurs when using inverse probability weighting (IPW) strategies to estimate the mean $\mathbb{E}_{P_0}[Y^1]$ of an outcome Y^a that would be observed under an intervention that assigns a binary treatment $A = a$ uniformly in a population. Under standard causal assumptions of consistency, positivity, and conditional exchangeability given predictors W, the target estimand is identified by $\mathbb{E}_{P_0}[Y\alpha_0(X)]$ where, for $X = (A, W)$, the quantity $\alpha_0(x) = p_1(x)/p_0(x)$ is an importance weight that represents the ratio between the density of the observed data-generating distribution $p_0(x) = p_{A,W}(a, w)$ and the density of the intervention distribution $p_1(x) = \mathbb{I}(a = 1)p_W(w)$. (We use the terms *density* and *probability mass function* interchangeably throughout.) This identification result suggests the IPW estimator

$$n^{-1}\sum_{i=1}^{n} \underbrace{\frac{\mathbb{I}(a_i = 1)}{\hat{p}(w_i)}}_{\text{importance weight}} y_i$$

where $\hat{p}(w_i)$ is an estimator of the propensity score $p_{A|W}(1 \mid w_i)$. However, the IPW estimator is known to be overly sensitive to small errors in the estimated propensity score $\hat{p}(w_i)$ when $a_i = 1$ and $p_{A|W}(1 \mid w_i)$ is close to zero, since the importance weight above can take extreme values in this case. Motivated by this and similar issues for continuous densities, several methods have been proposed to estimate density ratios directly from data, without having to learn individual component models such as the propensity score model above.

Rather than focusing on density ratios specifically, developments in the causal inference literature have instead focused on learning so-called *Riesz representers*, which are signed weighting functions corresponding to certain linear statistical estimands, with the density ratio being an example of a Riesz representer. A growing body of such literature approximates the Riesz representer as a *balancing weight*, which minimizes some measure of the worst-case imbalance over a set of test functions, while penalizing the complexity of the weights. We refer the reader to Ben-Michael et al. (2021) for a review, with early proposals including: generalized regression estimators (Deville and Särndal, 1992); kernel mean matching (Huang et al., 2006); entropy balancing (Hainmueller, 2012); covariate balance propensity scores (Imai and Ratkovic, 2014); stable balancing weights (Zubizarreta, 2015); and more recent proposals combining balancing weight estimation with outcome regression estimation, such as in: approximate residual balancing (Athey et al., 2018); augmented minimax linear estimation (Hirshberg and Wager, 2021; Dikkala et al., 2020); and recent work on duality by Bruns-Smith and Feller (2022).

Balancing weight methods successfully alleviate the problem of extreme weights in importance weighted estimators and circumvent the need to model components of the Riesz representer / density ratio. However, they also have several drawbacks: (a) when the test function set is not sufficiently rich there is no guarantee that the resulting weights provide a good approximation to the target Riesz representer; (b) when the test function set is a rich function set, the resulting minimax problem (minimize over the weights, maximize over the test functions) becomes computationally challenging (see, for example, the adversarial learner of Chernozhukov et al. (2020)); and (c) several balancing weight approaches estimate

only the weights evaluated at observations $\hat{\alpha}_i \approx \alpha_0(x_i)$ rather than learn the function α_0 itself. Such methods are not amenable to making out-of-sample predictions, meaning that model selection/validation techniques, such as cross-validation, are not possible.

To overcome these issues, we focus on methods which learn Riesz representers via empirical risk minimization, building on previous work by Chernozhukov et al. (2024) who propose *Riesz regression* as a general Riesz representer learning approach, where an empirical risk is constructed based on the mean squared error loss. A special case of Riesz regression can also be found in earlier work on score matching by Hyvärinen (2005).

Although Riesz regression has become a popular paradigm in causal inference using machine learning, several other density ratio learning methodologies have been proposed when one observes n_1 samples from the numerator distribution P_1 and n_0 samples from the denominator distribution P_0. In particular, empirical risk minimization approaches were pioneered through the Bregman divergence framework of Gutmann and Hirayama (2011) and Sugiyama et al. (2012), with Bregman divergences being a fundamental class of divergences, due to Bregman (1967), that generalize the squared Euclidean distance. Their framework includes as special cases the Kullback–Leibler Importance Estimation Procedure (KLIEP) (Sugiyama et al., 2007), which minimizes the Kullback–Leibler divergence, and unconstrained Least Squares Importance Fitting (uLSIF) (Kanamori et al., 2009), which minimizes a squared error loss using the same decomposition as in Riesz regression. Another common density ratio learning approach is to recast the density ratio as the odds from a probabilistic classifier that learns whether an observation is drawn from P_1 or P_0 (Qin, 1998; Cheng and Chu, 2004; Bickel et al., 2008). Perhaps surprisingly, however, probabilistic classification via empirical risk minimization is equivalent to learning the density ratio via an empirical Bregman divergence for a broad class of 'proper' classification losses that includes the binary cross-entropy and exponential losses (Menon and Ong, 2016).

One reason we believe that general density ratio learning methods have not been more widely adopted in causal machine learning is due to the requirement to observe samples from P_1, which in causal settings usually represents an intervention distribution that is not directly sampled from. Generating such samples requires the use of data augmentation techniques, along with careful consideration of how the intervention distribution is constructed. However, examples of such data augmentation strategies being used in causal inference do exist. For example, Díaz et al. (2023) generate augmented data for longitudinal modified treatment policies and learn their density ratio of interest via probabilistic classification. Lee and Schuler (2025) generate augmented data for policy / shift interventions, and learn their density ratio via Riesz regression. The latter work is motivated by a separate concern of Riesz regression: without data augmentation, Riesz regression is usually not amenable for use in gradient boosting algorithms because the unit loss can depend on the Riesz representer value at more than one point in the covariate space. For example, when learning the importance weights for the IPW estimator above, the loss for the ith observation depends on the values $\alpha(a_i, w_i)$ and $\alpha(1, w_i)$, for a candidate model α.

Our main contribution is to propose Bregman–Riesz regression as an extension to Riesz regression, which includes several density ratio learning methods in a unifying framework based on Bregman divergences. We show why data augmentation methods are often required when applying density ratio learning methods to causal problems, where data from the intervention distribution is not available. The need to develop estimand-specific data augmentation is arguably a barrier to the adoption of density ratio learning methods in causal inference. We address this by demonstrating several augmentation strategies for common counterfactual interventions (modified treatment policies, stabilized weights, natural mediation interventions). We also provide practical recommendations based on a numerical study of the effect of different losses and augmentation strategies when learning causal density ratios. Our experiments use neural network models, gradient-boosted tree models,

and kernel methods, with our implementations made available via a Python package at `https://github.com/CI-NYC/densityratios`.

In Section 2.1 we describe the Riesz representer setup, outlining how several causal examples related to average linear functionals and density ratio learning can be understood in a common framework. In Section 2.2 we outline the proposed Bregman–Riesz regression framework for Riesz representer learning, focusing on the least-squares loss, Kullback–Leibler divergence, Itakura–Saito distance, and a negative binomial loss that is associated with classification-based approaches for density ratio learning. In Section 2.3 we show how data augmentation methods can be used to empirically estimate Bregman–Riesz risks when learning causal density ratios using Bregman–Riesz regression. Numerical experiments are presented in Section 3, where we compare kernel density, uLSIF, KLIEP, gradient boosting, and neural network learners on several causal density ratio estimation problems. In Section 4 we discuss related works and conclude with a discussion in Section 5.

2 Methods

2.1 Problem description

Consider a random vector $X \in \mathcal{X}$ distributed according to an unknown distribution P_0. Let $\mathcal{H}$ be a Hilbert space of real-valued functions $f : \mathcal{X} \to \mathbb{R}$ with inner product $\langle f, g \rangle \equiv \mathbb{E}_{P_0}[f(X)g(X)]$ for $f, g \in \mathcal{H}$. Letting $\mathbb{H} : \mathcal{H} \to \mathbb{R}$ be a bounded linear functional, the Riesz representation theorem implies that there exists a unique function $\alpha_0 \in \mathcal{H}$, called the Riesz representation of $\mathcal{H}$, or Riesz representer for short, with the property that $\mathbb{H}(f) = \langle f, \alpha_0 \rangle$ for all $f \in \mathcal{H}$. The following are two canonical examples of this setup in the statistical literature.

Setting 1 (Density Ratio)**.** Let P_1 be an alternative distribution over $\mathcal{X}$ that is absolutely continuous with respect to P_0, i.e., the support of P_1 is contained in the support of P_0. The map $\mathbb{H}(f) = \mathbb{E}_{P_1}[f(X)]$ has a Riesz representer that is the density ratio $\alpha_0(x) = \mathrm{d}P_1(x)/\mathrm{d}P_0(x)$, also known as the Radon–Nikodym derivative in probability theory. The density ratio is more commonly written as $\alpha_0(x) = p_1(x)/p_0(x)$ where $p_1(x) = \mathrm{d}P_1(x)/\mathrm{d}\nu(x)$ and $p_0(x) = \mathrm{d}P_0(x)/\mathrm{d}\nu(x)$ are densities with respect to a dominating measure ν, e.g., the Lebesgue measure or the counting measure. In this way, the density ratio is agnostic as to whether $\mathcal{X}$ is discrete or continuous.

Setting 2 (Average Linear Functional)**.** Let $\mathbb{H}(f) = \mathbb{E}_{P_0}[m(f, X, Y)]$, where $m : \mathcal{H} \times \mathcal{X} \times \mathbb{R} \to \mathbb{R}$ is a known functional that is linear in its first argument, and $Y \in \mathbb{R}$ is an outcome with $(Y, X) \sim P_0$. The Riesz representer is the unique function $\alpha_0 \in \mathcal{H}$ such that $\mathbb{E}_{P_0}[m(f, X, Y)] = \langle f, \alpha_0 \rangle$. This setting has been studied by Hirshberg and Wager (2021), Chernozhukov et al. (2022), and Chernozhukov et al. (2024), with primary interest in the debiased estimation of $\mathbb{H}(\mu_0)$, where $\mu_0(x) = \mathbb{E}_{P_0}[Y \mid X = x]$ and it is assumed that $\mu_0 \in \mathcal{H}$.

One key difference between these two settings is that in Setting 1, the density ratio $\alpha_0(x) \geq 0$ is a nonnegative importance weight, but in Setting 2, the Riesz representer is generally signed depending on the linear functional m. In the current work we focus on the density ratio setting; however, the two settings are not mutually exclusive, as we illustrate through the causal inference examples below. In fact, many of the Riesz representer examples that motivated the consideration of Setting 2, can be viewed as density ratios.

Example 1 (Outcome Regression)**.** To show the generality of the Riesz representer setup, consider Setting 2, with $m^{(\mathrm{OR})}(f, x, y) = f(x)y$. In this setting the Riesz representer is the

outcome regression function $\alpha_0^{(\mathrm{OR})} = \mu_0$ (with $\mu_0(x) = \mathbb{E}_{P_0}[Y \mid X = x]$ as above), since for all $f \in \mathcal{H}$

$$\mathbb{E}_{P_0}[m^{(\mathrm{OR})}(f, X, Y)] = \mathbb{E}_{P_0}[f(X)Y] = \langle f, \alpha_0^{(\mathrm{OR})} \rangle.$$

Thus, one can view outcome regression as learning the Riesz representer for the bounded linear functional $f \mapsto \mathbb{E}[f(X)Y]$. In some sense, therefore, Riesz representer learning generalizes outcome regression to a broader class of statistical functionals.

Example 2 (Average Treatment Effect)**.** Let $X = (A, W)$ consist of a binary treatment $A \in \{0, 1\}$ and a vector of covariates $W \in \mathbb{R}^p$, and let Y^a denote the outcome that would be observed if an intervention assigned treatment $A = a$. The average treatment effect $\mathbb{E}[Y^1 - Y^0]$ is the difference in the mean outcome when treatment / no treatment is applied uniformly in the population (Rosenbaum and Rubin, 1983). Under standard causal assumptions, the average treatment effect is identified by $\mathbb{E}_{P_0}[m^{(\mathrm{ATE})}(\mu_0, X, Y)]$ where $m^{(\mathrm{APE})}(f, x, y) = f(1, w) - f(0, w)$ is a known linear functional. Hence, the average treatment effect is an example of Setting 2, and can be expressed as $\langle \mu_0, \alpha_0^{(\mathrm{ATE})} \rangle$ where

$$\alpha_0^{(\mathrm{ATE})}(x) \equiv \frac{\mathbb{I}\{a = 1\} p_W(w)}{p_{A,W}(a, w)} - \frac{\mathbb{I}\{a = 0\} p_W(w)}{p_{A,W}(a, w)}$$

is a signed Riesz representer. The Riesz representer $\alpha_0^{(\mathrm{ATE})}$ also appears in pseudo-outcome learners for the conditional average treatment effect, such as in the DR-learner of Kennedy (2023).

Example 3 (Average Policy Effect)**.** Using the setup from Example 2, the average policy effect $\mathbb{E}[Y^{\pi(W)}]$ is the mean outcome when treatment is determined by a known decision rule $\pi : \mathbb{R}^p \to \{0, 1\}$ that is a function of covariates (Dudik et al., 2011; van der Laan and Luedtke, 2014; Athey and Wager, 2021). Under standard causal assumptions, the average policy effect is identified by $\langle \mu_0, \alpha_0^{(\mathrm{APE})} \rangle$ where

$$\alpha_0^{(\mathrm{APE})}(x) \equiv \frac{\mathbb{I}\{a = \pi(w)\} p_W(w)}{p_{A,W}(a, w)}$$

is a density ratio (Riesz representer from Setting 1). The average policy effect can also be expressed as $\mathbb{E}_{P_0}[m^{(\mathrm{APE})}(\mu_0, X, Y)]$ where $m^{(\mathrm{APE})}(f, x, y) = \pi(w) f(1, w) + \{1 - \pi(w)\} f(0, w)$ is a known linear functional for all $f(a, w)$ in $\mathcal{H}$. Hence, the average policy effect is also an example of Setting 2. Comparing the Riesz representer above with that in Example 2, we see that $\alpha_0^{(\mathrm{ATE})}$ represents the difference in the density ratios corresponding to the average policy effects $\mathbb{E}[Y^1]$ and $\mathbb{E}[Y^0]$.

Example 4 (Average Shift Effect)**.** Let $X = (A, W)$ consist of a continuous treatment $A \in \mathbb{R}$ and covariates $W \in \mathbb{R}^p$. Letting Y^a denote the potential outcome, as in Example 3, the average shift effect $\mathbb{E}[Y^{A+\delta}]$ is the mean outcome when the natural value of treatment is uniformly shifted by a constant $\delta \in \mathbb{R}$ (Díaz and van der Laan, 2012; Díaz et al., 2023). Under standard causal assumptions, this quantity is identified by $\langle \mu_0, \alpha_0^{(\mathrm{ASE})} \rangle$ where

$$\alpha_0^{(\mathrm{ASE})}(x) \equiv \frac{p_{A,W}(a - \delta, w)}{p_{A,W}(a, w)}$$

is a density ratio (Riesz representer from Setting 1). The average shift effect can also be expressed as $\mathbb{E}_{P_0}[m^{(\mathrm{ASE})}(\mu_0, X, Y)]$ where $m^{(\mathrm{ASE})}(f, x, y) = f(a + \delta, w)$ is a known linear functional for all $f(a, w)$ in $\mathcal{H}$. Hence, the average shift effect is also an example of Setting 2.

Example 5 (Average Derivative Effect)**.** Using the setup from Example 4, the average derivative effect $\lim_{\delta \to 0} \mathbb{E}[Y^{A+\delta} - Y]/\delta$ is the limiting difference between the average shift

effect and no intervention per unit change in treatment (Rothenhäusler and Yu, 2020). Under causal assumptions, and assuming that μ_0 is differentiable with respect to treatment, the average derivative effect is identified by $\mathbb{E}_{P_0}[m^{(\mathrm{ADE})}(\mu_0, X, Y)]$, where $m^{(\mathrm{ADE})}(f, x, y) = \partial_a f(a, x)$ and ∂_a denotes the partial derivative with respect to a (Hardle and Stoker, 1989; Newey and Stoker, 1993; Imbens and Newey, 2009). Under regularity conditions regarding the boundary of the support of treatment, the average derivative effect is also identified by $\langle \mu_0, \alpha_0^{(\mathrm{ADE})} \rangle$, where

$$\alpha_0^{(\mathrm{ADE})}(x) \equiv -\partial_a \log\{p_X(a, w)\}$$

is a Riesz representer from Setting 2. Although this Riesz representer is not a density ratio, it can be viewed as a limiting scaled difference between an average shift effect and no intervention (density ratio of one)

$$\alpha_0^{(\mathrm{ADE})}(x) = \lim_{\delta \to 0} \frac{1}{\delta} \left[\frac{p_{A,W}(a - \delta, w)}{p_{A,W}(a, w)} - 1 \right].$$

Moreover, in Appendix A we show how a vectorized version of this Riesz representer forms the basis of score matching methods in generative diffusion modeling, and can be estimated by the Riesz representer learning methods in Section 2.2 known as *score matching* (Hyvärinen, 2005; Lyu, 2009; Song and Ermon, 2019).

Example 6 (Average Dose Response Curve)**.** Using the setup of Example 4, the dose response curve $\psi(a) \equiv \mathbb{E}_{P_0}[Y^a]$ is the mean outcome when an intervention uniformly assigns treatment $A = a$ (Robins and Gill, 2001; Kennedy et al., 2017), and the average dose response curve $\mathbb{E}_{P_0}[\psi(A)]$ is its mean. Under standard causal assumptions, the average dose response curve is identified by $\langle \mu_0, \alpha_0^{(\mathrm{SW})} \rangle$, where

$$\alpha_0^{(\mathrm{SW})}(x) \equiv \frac{p_A(a) p_W(w)}{p_{A,W}(a, w)}$$

is a density ratio (Riesz representer from Setting 1), known as the stabilized weight (Robins et al., 2000; Ai et al., 2021). The stabilized weight also appears in pseudo-outcome based learners of the dose response curve itself (Kennedy et al., 2017) and in the definition of the mutual information $-\mathbb{E}_{P_0}[\log\{\alpha_0^{(\mathrm{SW})}(X)\}]$ between A and W (Nguyen et al., 2007; Belghazi et al., 2018).

Example 7 (Natural Mediation Effect)**.** Let $X = (M, A, W)$ consist of a mediator $M \in \mathbb{R}^q$, a binary treatment $A \in \{0, 1\}$, and covariates $W \in \mathbb{R}^p$. Also, let $Y^{a,m}$ denote the outcome that would be observed if treatment and mediator were set to $A = a$ and $M = m$, and let M^a denote the mediator that would be observed if treatment were set to $A = a$. Natural mediation effects are defined via the quantity $\mathbb{E}[Y^{a', M^{a\dagger}}]$, where a' and $a^\dagger$ are two levels of treatment (Robins and Greenland, 1992; Pearl, 2001). Under causal assumptions given in Pearl (2001), the natural mediation effect is identified by $\langle \mu_0, \alpha_0^{(\mathrm{NME})} \rangle$ where

$$\alpha_0^{(\mathrm{NME})}(x) \equiv \frac{p_{M|A,W}(m \mid a^\dagger, w) \mathbb{I}\{a = a'\} p_W(w)}{p_{M,A,W}(m, a, w)}$$

is a density ratio (Riesz representer from Setting 1).

Remark 1. The absolute continuity condition $P_1 \ll P_0$ in the density ratio setting (Setting 1) is also referred to as an overlap or positivity assumption. This condition can often be quite strong and unlikely to hold, especially when $\mathcal{X}$ is high-dimensional (D'Amour et al., 2021). In particular, even when population overlap is satisfied, practical overlap violations in finite samples may remain. That is, there may be regions of the covariate space $\mathcal{X}$ in which there are very few or no observations from P_0, but many observations from P_1. In the density ratio learning literature, this is also referred to as the density chasm problem (Rhodes et al., 2020). See additional discussion on overlap in causal inference by Crump et al. (2009), Khan and Tamer (2010), Petersen et al. (2012), and Susmann et al. (2025).

Remark 2. One of the key reasons for Riesz representer learning in causal inference is to obtain efficient and debiased point estimates for the quantity $\mathbb{H}(\mu_0)$ in the average linear functional setting (Setting 2). Here we give a very brief account of this theory and refer to Rotnitzky et al. (2021) and Hirshberg and Wager (2021) for more details. Given a point-wise consistent regression estimator $\hat{\mu}$ for μ_0, a naive estimator for $\mathbb{H}(\mu_0)$ can be constructed as $\mathbb{H}_n(\hat{\mu})$, where $\mathbb{H}_n(f) \equiv \mathbb{E}_n[m(f, X, Y)]$ and $\mathbb{E}_n[\cdot] \equiv n^{-1}\sum_{i=1}^n[\cdot]_i$ denotes expectation under an empirical measure given n samples from P_0. However, this naive estimator is generally biased since, under limited regularity conditions (e.g., when $\hat{\mu}$ is learned on an independent sample) one obtains the asymptotic result

$$\sqrt{n}\left\{\mathbb{H}_n(\hat{\mu}) - \mathbb{H}(\mu_0)\right\} + \sqrt{n}\mathbb{E}_n[\alpha_0(X)\{Y - \hat{\mu}(X)\}] \xrightarrow{d} \mathcal{N}\left(0, v_0\right),$$

where $v_0 = \mathrm{var}_{P_0}[m(\mu_0, Y, X) + \alpha_0(X)\{Y - \mu_0(X)\}]$. In this result, the term $\sqrt{n}\mathbb{E}_n[\alpha_0(X)\{Y - \hat{\mu}(X)\}]$ generally does not converge to zero and therefore represents an asymptotic bias in the naive estimator. This bias can be estimated as $\mathbb{E}_n[\hat{\alpha}(X)\{Y - \hat{\mu}(X)\}]$ using a learner $\hat{\alpha}$ of the Riesz representer, and can be viewed as an importance weighted estimate of $\mathbb{H}(\mu_0 - \hat{\mu}) = \mathbb{E}_{P_0}[\alpha_0(X)\{Y - \hat{\mu}(X)\}]$. Estimators which correct for this bias are referred to as debiased (because they correct for the bias), efficient (because they have variance v_0 equal to the semiparametric efficiency bound), and rate double robust (because they require a mixed bias condition $\sqrt{n}\langle\hat{\mu} - \mu_0, \hat{\alpha} - \alpha_0\rangle \to 0$, which allows for a trade-off in the learning rates of $\hat{\mu}$ and $\hat{\alpha}$).

2.2 Bregman–Riesz Regression

In this Section we outline Bregman–Riesz regression as any method that minimizes an empirical approximation of a Bregman–Riesz Risk (BRR), which we define presently. The BRR generalizes the Bregman divergence frameworks from the density ratio learning methods of Gutmann and Hirayama (2011) and Sugiyama et al. (2012) to Riesz representer learning. Bregman divergences (Bregman, 1967; Lafferty, 1999; Grunwald and Dawid, 2004) are a fundamental class of convex divergences that generalize the squared Euclidean distance. Letting $F : \mathbb{R} \to \mathbb{R}$ be a known, strictly convex function with derivative F', the Bregman divergence is defined for any pair of functions $g, g_0 \in \mathcal{H}$ as

$$D_F(g_0 \| g) \equiv \mathbb{E}_{P_0}\left[F\{g_0(X)\} - F\{g(X)\} - F'\{g(X)\}\{g_0(X) - g(X)\}\right]. \tag{1}$$

This divergence can be interpreted as the mean of a loss, where the individual loss represents the residual of the Taylor first-order approximation $F\{g_0(x)\} \approx F\{g(x)\} + F'\{g(x)\}\{g_0(x) - g(x)\}$. The Bregman divergence has the appealing property that $D_F(g_0\|g)$ is convex in g_0 and $D_F(g_0\|g) \geq 0$ with equality if and only if $g = g_0$ almost surely. This makes Bregman divergences useful in machine learning and in optimization problems based on empirical risk minimization.

To provide some intuition for the choice of convex generating function F, we consider the case where the second derivative F'' exists. Following similar interpretations by Menon and Ong (2016) and Reid and Williamson (2009), we express the Bregman divergence using the *cost-sensitive point-wise regret*

$$r(t \mid v_0, v) \equiv |v_0 - t|\mathbb{I}\left\{\min(v, v_0) < t < \max(v, v_0)\right\},$$

which for a true value $v_0 \in \mathbb{R}$ and candidate value $v \in \mathbb{R}$ returns the absolute error of an alternative value $t \in \mathbb{R}$ when t is between v and v_0, and returns zero otherwise. Using this quantity we write the Bregman divergence as a weighted integral over all possible values with value weights F'':

$$D_F(g_0\|g) = \int_{-\infty}^{\infty} \mathbb{E}_{P_0}\left[r\{t \mid g_0(X), g(X)\}\right] F''(t)\mathrm{d}t \tag{2}$$

See Appendix B for a derivation of this result. This integral form of the Bregman divergence is insightful because it suggests that the Bregman divergence gives greater emphasis towards accurately modeling $g_0(X)$ for values where F'' is large. The weight $F''(t)$ can be viewed in a similar way to a variance function $V(t) \propto 1/F''(t)$ that defines the population deviance in the theory of generalized linear models (see Appendix C for details). Additionally, we see that the Bregman divergence is invariant to the addition of an affine function to the convex generating function F, since such modifications do not affect F''.

One of the key advantages of basing Riesz representer learning around Bregman divergences is that the divergence $D_F(\alpha_0 \| \alpha)$ implies a risk $\mathcal{R}_F(\alpha)$ that depends only implicitly on the unknown Riesz representer α_0 through the bounded linear functional $\mathbb{H}$. We refer to this risk as the *Bregman–Riesz Risk (BRR)*, which is defined as

$$\mathcal{R}_F(\alpha) \equiv \mathbb{E}_{P_0}\left[F'\{\alpha(X)\}\alpha(X) - F\{\alpha(X)\}\right] - \mathbb{H}\left(F' \circ \alpha\right),$$

and is minimized at α_0. For proof, consider that

$$\begin{aligned}
\alpha_0 &= \arg\min_{\alpha \in \mathcal{H}} D_F(\alpha_0 \| \alpha) \\
&= \arg\min_{\alpha \in \mathcal{H}} \mathbb{E}_{P_0}\left[-F\{\alpha(X)\} + F'\{\alpha(X)\}\alpha(X)\right] - \mathbb{E}_{P_0}\left[F'\{\alpha(X)\}\alpha_0(X)\right] \\
&= \arg\min_{\alpha \in \mathcal{H}} \mathcal{R}_F(\alpha)
\end{aligned}$$

where, in the second line above the constant $\mathbb{E}_{P_0}\left[F\{\alpha_0(X)\}\right]$ is discarded, and the third line follows from $\mathbb{E}_{P_0}\left[F'\{\alpha(X)\}\alpha_0(X)\right] = \mathbb{H}\left(F' \circ \alpha\right)$. In the density ratio and average linear functional settings (Settings 1 and 2), the BRR, respectively, reduces to

$$\mathcal{R}_F(\alpha) = \mathbb{E}_{P_0}\left[F'\{\alpha(X)\}\alpha(X) - F\{\alpha(X)\}\right] - \mathbb{E}_{P_1}\left[F'\{\alpha(X)\}\right], \tag{3}$$

$$\mathcal{R}_F(\alpha) = \mathbb{E}_{P_0}\left[F'\{\alpha(X)\}\alpha(X) - F\{\alpha(X)\} - m\left(F' \circ \alpha, X, Y\right)\right]. \tag{4}$$

Neither of these expressions depend on the unknown α_0, making them useful for learning α_0 via empirical risk minimization. We show the form of the BRR for various choices of convex differentiable functions F in Table 1. For Bregman divergences using other convex generator functions see Menon and Ong (2016, Table 1) and Banerjee et al. (2004, Table 1).

Remark 3. Viewing the BRR in (4) as the mean of a loss, the resulting loss generally depends on the value of α at multiple points of the input space $\mathcal{X}$ through the term $m\left(F' \circ \alpha, x, y\right)$. For example, for the average treatment effect (Example 2), the individual level loss in (4) is

$$F'\{\alpha(a,w)\}\alpha(a,w) - F\{\alpha(a,w)\} - F'\{\alpha(1,w)\} + F'\{\alpha(0,w)\},$$

which depends on the values $\alpha(1,w)$ and $\alpha(0,w)$, since F' cannot be constant by the requirement that F is strictly convex. This makes Riesz representer learning via empirical loss minimization challenging for some algorithms, such as gradient boosting (Lee and Schuler, 2025).

Table 1: Summary of Bregman–Riesz Risks considered in this paper, as defined by their generating function F which is convex on the given domain. The second derivative F'' relates to the cost weight in the integral Bregman divergence expression in (2).

Name	Domain	$F(t)$	$F''(t)$
Least squares	$\mathbb{R}$	t^2	2
Kullback–Leibler	$\mathbb{R}_{>0}$	$t\log(t) - t$	t^{-1}
Negative binomial	$\mathbb{R}_{>0}$	$t\log(t) - (1+t)\log(1+t)$	$t^{-1}(1+t)^{-1}$
Itakura–Saito	$\mathbb{R}_{>0}$	$-\log(t) - 1$	t^{-2}

Divergence 1 (Least squares)**.** For $F(t) = t^2$ the divergence in (1) becomes the mean squared error $D_F(\alpha_0\|\alpha) = \mathbb{E}_{P_0}[\{\alpha_0(X) - \alpha(X)\}^2]$ and the BRR reduces to

$$\mathcal{R}_F(\alpha) = \mathbb{E}_{P_0}\left[\alpha^2(X)\right] - 2\mathbb{H}(\alpha).$$

The least-squares divergence was studied in the density ratio setting by Kanamori et al. (2009), who propose Least-Squares Importance Fitting (LSIF) and unconstrained LSIF (uLSIF) procedures based on empirical minimization of the least-squares BRR when $\alpha(x)$ is a sieve model (see Section 3). In the average linear functional setting, Chernozhukov et al. (2024) propose neural network Riesz representer learners based on empirical minimization of the least-squares BRR in a procedure referred to as Riesz regression. An earlier special case of Riesz regression is the score matching proposal of Hyvärinen (2005), see Appendix A for details.

Divergence 2 (Kullback–Leibler)**.** When $\alpha_0(X) > 0$ and $\alpha(X) > 0$ almost surely then one can let $F(t) = t\log(t) - t$, in which case the divergence in (1) is the Kullback–Leibler (KL) divergence

$$D_F(\alpha_0\|\alpha) = \mathbb{E}_{P_0}\left[\alpha_0(X)\log\left\{\frac{\alpha_0(X)}{\alpha(X)}\right\} + \alpha(X) - \alpha_0(X)\right].$$

This version of the KL divergence is *unnormalized*, with the more conventional normalized version obtained when $\mathbb{E}_{P_0}[\alpha(X)] = \mathbb{E}_{P_0}[\alpha_0(X)]$ and hence the term $\mathbb{E}_{P_0}[\alpha(X) - \alpha_0(X)]$ can be discarded. Note that, in the density ratio setting with denominator density $p_0(x)$, $\mathbb{E}_{P_0}[\alpha_0(X)] = 1$ a priori because the numerator density integrates to 1. The BRR corresponding to the KL divergence is

$$\mathcal{R}_F(\alpha) = \mathbb{E}_{P_0}[\alpha(X)] - \mathbb{H}(\log\circ\alpha).$$

This KL BRR was studied in the density ratio setting by Sugiyama et al. (2007), who propose the KL Importance Estimation Procedure (KLIEP) based on minimizing an empirical approximation of $-\mathbb{H}(\log\circ\alpha)$ subject to the constraint that (empirically) $\mathbb{E}_{P_0}[\alpha(X)] = 1$ and $\alpha(x)$ is a sieve model (see Section 3). Their original derivation considers the KL divergence between the numerator density $p_1(x) = \alpha_0(x)p_0(x)$ and $\tilde{p}_1(x) \equiv \alpha(x)p_0(x)$, which is a density when $\mathbb{E}_{P_0}[\alpha(X)] = 1$.

Divergence 3 (Negative binomial)**.** When $\alpha_0(X) > 0$ and $\alpha(X) > 0$ almost surely one can let $F(t) = t\log(t) - (1+t)\log(1+t)$ in which case the divergence in (1) and BRR are

$$D_F(\alpha_0\|\alpha) = \mathbb{E}_{P_0}\left[\alpha_0(X)\log\left\{\frac{1+1/\alpha(X)}{1+1/\alpha_0(X)}\right\} + \log\left\{\frac{1+\alpha(X)}{1+\alpha_0(X)}\right\}\right]$$
$$\mathcal{R}_F(\alpha) = \mathbb{E}_{P_0}[\log\{1+\alpha(X)\}] + \mathbb{H}[\log\{1+1/\alpha(\cdot)\}].$$

In Section 2.3 we show how, in the density ratio setting, the negative binomial divergence is connected to the binary cross-entropy loss used in classification. We call this divergence 'negative binomial' due to its connection with negative binomial generalized linear model deviance (see Appendix C). The connection between density ratio learning and probabilistic classification has previously been established by Qin (1998) for parametric models, and Menon and Ong (2016) for Bregman divergences. The generating function F for the negative binomial divergence has the property $F(t) = tF(1/t)$, which we argue relates to a kind of symmetry in the density ratio setting when the roles of P_0 and P_1 are reversed (see Appendix D).

Divergence 4 (Itakura–Saito)**.** When $\alpha_0(X) > 0$ and $\alpha(X) > 0$ almost surely, one can let $F(t) = -\log(t) - 1$, in which case the divergence in (1) and BRR are

$$D_F(\alpha_0\|\alpha) = \mathbb{E}_{P_0}\left[\frac{\alpha_0(X)}{\alpha(X)} - \log\left\{\frac{\alpha_0(X)}{\alpha(X)}\right\} - 1\right]$$
$$\mathcal{R}_F(\alpha) = \mathbb{E}_{P_0}[\log\{\alpha(X)\}] + \mathbb{H}\left[\frac{1}{\alpha(\cdot)}\right].$$

This divergence is known as the Itakura–Saito distance and has previously been used for matrix factorization and audio analysis (Banerjee et al., 2004). In the density ratio setting, we observe that the Itakura–Saito distance is equivalent to the unnormalized Kullback–Leibler divergence (Divergence 2) when the roles of P_0 and P_1 are reversed (see Appendix D). To our knowledge, this divergence has not previously been used for density ratio learning, but we include it because of its connection with the Kullback–Leibler divergence.

2.3 Data Augmentation

Here we show how data augmentation methods, similar to those of Lee and Schuler (2025) and Díaz et al. (2023), can be used to approximate the BRR from Section 2.2 empirically when the target Riesz representer is a density ratio. We also show how data augmentation methods connect Bregman–Riesz regression to approaches based on probabilistic classification (Qin, 1998; Cheng and Chu, 2004; Bickel et al., 2008; Menon and Ong, 2016). Letting $\Delta \in \{0, 1\}$ be a binary random variable, we define the augmented distribution $(X, \Delta) \sim Q$ such that $P_Q(\Delta = 0) = P_Q(\Delta = 1) = 1/2$, $X \sim P_1$ when $\Delta = 1$, and $X \sim P_0$ when $\Delta = 0$. In other words, Q is a mixture distribution formed of the numerator and the denominator distributions, each with equal weight. This augmented distribution is useful for two reasons. First, Bayes' Theorem implies that the density ratio is equivalent to the odds

$$\alpha_0(x) = \frac{P_Q(\Delta = 1 \mid X = x)}{P_Q(\Delta = 0 \mid X = x)} = \frac{q_0(x)}{1 - q_0(x)}, \tag{5}$$

where $q_0(x) \equiv P_Q(\Delta = 1 \mid X = x)$. This formulation means that α_0 can be learned using probabilistic classification methods which aim to learn q_0 given samples from Q. This approach is used in causal inference, for example, by Díaz et al. (2023) to learn the density ratio corresponding to modified treatment policy interventions. Second, the augmented distribution allows us to express the BRR in (3) as

$$\mathcal{R}_F(\alpha) = 2\mathbb{E}_Q\left[(1 - \Delta)\left[F'\{\alpha(X)\}\alpha(X) - F\{\alpha(X)\}\right] - \Delta F'\{\alpha(X)\}\right] \tag{6}$$

meaning that α_0 can be learned by minimizing an empirical BRR approximation obtained using samples from Q, which we discuss in more detail shortly. Although these two approaches may seem different, the BRR in (6) can be used as a probabilistic classifier by making the substitution $\alpha(x) = q(x)/\{1 - q(x)\}$, with $q_0(x) = \alpha_0(x)/\{1 + \alpha_0(x)\}$ by (5). For example, when F is the negative-binomial generating function $t \mapsto t\log(t) - (1+t)\log(1+t)$ in Table 1, then this substitution implies that $\alpha_0 = \arg\min_{\alpha \in \mathcal{H}} \mathcal{R}_F(\alpha)$ is equivalent to $q_0/(1 - q_0)$ where

$$q_0 = \arg\min_q 2\mathbb{E}_Q\left[-(1 - \Delta)\log\{1 - q(X)\} - \Delta\log\{q(X)\}\right],$$

which is the binary cross-entropy, also known as the mean logistic loss, up to a constant factor of 2. Similar classifiers can be obtained for other choices of F, such as in Table 1.

One of the main challenges when using the BRR in (6) to learn the density ratio directly, is how to estimate expectations over the augmented distribution Q. The density ratio learning literature typically considers the setting where one has access to n_1 samples from the numerator distribution and n_0 samples from the denominator distribution, in which case one can approximate the mean of a function f as

$$\mathbb{E}_Q\left[f(\Delta, X)\right] \approx \sum_{i=1}^{n} \omega_i f(\delta_i, x_i)$$

where $n = n_0 + n_1$; the index i iterates through the samples from P_0 with $\delta_i = 0$, then the samples from P_1 with $\delta_i = 1$; and the normalized sample weight $\omega_i = \delta_i/(2n_1) + (1-\delta_i)/(2n_0)$

ensures that $\mathbb{E}_Q[\Delta] = \sum_{i=1}^n \omega_i \delta_i = 1/2$. This weight reduces to $\omega_i = 1/n$ when $n_0 = n_1$. Moreover, we will consider some examples where the observations from P_1 each have an unnormalized weight γ_i, in which case one obtains the more general form

$$\omega_i = \frac{\delta_i \gamma_i}{2 \sum_{j=1}^n \delta_i \gamma_i} + \frac{1 - \delta_i}{2 n_0}. \tag{7}$$

In the causal inference examples from Section 2.1, P_1 represents a counterfactual intervention distribution from which one does not have samples, while P_0 represents the observed data distribution from which one does have samples. This means that in order to approximate expectations of the form $\mathbb{E}_{P_1}[f(1, X)]$ (and hence of the form $\mathbb{E}_Q[f(\Delta, X)]$), additional techniques are required to generate n_1 representative (possibly weighted) samples from P_1, depending on the form of the intervention distribution. Supposing such an augmentation is possible, the resulting causal density ratio learning algorithm is summarized in Algorithm 1. This algorithm is agnostic to the form of the model class $\mathcal{M}$ and regularizer Λ_n which can represent one of a broad class of machine learning algorithms (neural networks, gradient boosted trees, lasso, etc.).

Algorithm 1 (Bregman–Riesz regression for causal density ratio learning)

1. Choose a convex generating function F.
2. Given n_0 samples from P_0 use an augmentation strategy to generate n_1 (possibly) weighted samples from P_1 with weights γ_i.
3. Learn the density ratio as $\hat{\alpha} = \arg\min_{\alpha \in \mathcal{M}} \{\mathcal{R}_n(\alpha) + \Lambda_n(\alpha)\}$, where

$$\mathcal{R}_n(\alpha) = \sum_{i=1}^n \omega_i \left[(1 - \delta_i) \left[F'\{\alpha(x_i)\}\alpha(x_i) - F\{\alpha(x_i)\} \right] - \delta_i F'\{\alpha(x_i)\} \right],$$

the ω_i's are defined as above in (7), $\mathcal{M}$ is a user-specified function class, and $\Lambda_n : \mathcal{M} \to \mathbb{R}$ is an optional penalization term.

We view the requirement for data augmentation as one of the main practical barriers to using density ratio learning methods in causal inference. We address this issue by demonstrating possible augmentation schemes in several examples. These examples use a combination of three main strategies: (a) decomposing the density ratio learning problem into a series of simpler learning problems, (b) augmenting the observed data, and (c) reweighting the augmented data.

Augmentation Example 1 (Modified Treatment Policy)**.** In Examples 3 and 4, the intervention assigns a new treatment $\tilde{a} = \tilde{a}(a, w)$ that depends on the natural value of treatment a, and/or covariates w (Díaz and van der Laan, 2012; Haneuse and Rotnitzky, 2013). Specifically, for the average policy effect $\tilde{a} = \pi(w)$ and for the average shift effect $\tilde{a} = a + \delta$. Thus, given observations $\{a_i, w_i\}_{i=1}^{n_0}$ from P_0, one can generate $n_1 = n_0$ observations from P_1 as $\{\tilde{a}(a_i, w_i), w_i\}_{i=1}^{n_1}$. This technique is used by Lee and Schuler (2025) for the average policy/shift effect, and by Díaz et al. (2023) for the average shift effect and for similar modified treatment policy interventions, generalizing to the setting of longitudinal treatments.

Augmentation Example 2 (Binary Modified Treatment Policy)**.** In Example 3, the density ratio $\alpha^{(APE)}$ is known to be zero when the observed treatment never agrees with the intervention policy $\pi(w)$. For this reason, the Bregman divergences that are defined only on positive values may not be appropriate for learning $\alpha^{(APE)}$ directly using the strategy in Augmentation Example 1. However, one can factorize the density ratio as

$$\alpha^{(APE)}(x) = \frac{p_W(w)}{p_{W|A}\{w \mid \pi(w)\}} \left\{ \frac{\mathbb{I}\{a = \pi(w)\}}{p_A\{\pi(w)\}} \right\}$$

where the term in curly braces is a known function up to the normalizing constant $p_A\{\pi(w)\} = \mathbb{E}_{P_0}[\mathbb{I}\{A = \pi(W)\}]$. This factorization suggests learning $\alpha^{(APE)}$ by estimating this normalizing constant and separately learning the density ratio $\alpha_0^{(1)}(w) = p_W(w)/p_{W|A}\{w \mid \pi(w)\}$ corresponding to the term outside the curly braces above. Given samples $\{a_i, w_i\}_{i=1}^{n_0}$ from P_0, an augmentation dataset for learning $\alpha_0^{(1)}$ can be constructed since $\{w_i\}_{i=1}^{n_0}$ represent samples from the numerator density $p_W(w)$, and for the denominator density $p_{W|A}\{w \mid \pi(w)\}$ one can use the $m \leq n_0$ samples $\{w_i\}_{i=1}^{m}$ where $a_i = \pi(w_i)$.

Augmentation Example 3 (Average Treatment Effect)**.** In Example 2, the Riesz representer $\alpha^{(ATE)}$ is not a density ratio, but does admit a similar factorization to $\alpha^{(APE)}$ in Augmentation Example 2 above. Specifically,

$$\alpha_0^{(\mathrm{ATE})}(x) = \frac{p_A(a)p_W(w)}{p_{A,W}(a,w)} \left\{ \frac{\mathbb{I}\{a=1\}}{p_A(1)} - \frac{\mathbb{I}\{a=0\}}{p_A(0)} \right\},$$

where we recognize that the term outside the curly braces is the stabilized weight $\alpha^{(SW)}$ of Example 6, except with a binary treatment A. Similar to Augmentation Example 2, this factorization suggests a procedure where one first estimates the constant $p_A(1)$ as $\hat{p}_A(1)$, and hence $p_A(0)$ as $\hat{p}_A(0) = 1 - \hat{p}_A(1)$, and then separately learns the density ratio $\alpha^{(SW)}$. Given samples $\{a_i, w_i\}_{i=1}^{n_0}$ from P_0, one can generate $n_1 = 2n_0$ weighted samples from the intervention distribution $p_1(x) = p_A(a)p_W(w)$ as $\{1, w_i\}_{i=1}^{n_0}$ and $\{0, w_i\}_{i=n_0+1}^{n_1}$ where the ith sample has weight $\gamma_i = \hat{p}_A(a_i)$.

Augmentation Example 4 (Stabilized Weight)**.** For the stabilized weight (Example 6), the intervention distribution P_1 effectively samples treatment A and covariates W independently under their respective marginal distributions under P_0, i.e., $p_1(a, w) = p_A(a)p_W(w)$. Therefore, given observations $\{a_i, w_i\}_{i=1}^{n_0}$ from P_0, the pair (a_i, w_j) represents a random draw from P_1. Thus, one might consider using all possible $n_1 = n_0^2$ such draws. However, the resulting augmented dataset of size $n = n_0(n_0+1)$ may be computationally impractical even for modest n_0 due to long training times or memory costs. Even if it were feasible to use all possible $n_1 = n_0^2$ samples of the form (a_i, w_j), it may be preferable to discard samples where $i = j$, resulting in $n_1 = n_0(n_0 - 1)$ samples. This distinction is equivalent to that of estimating $\mathbb{E}_{P_1}[\cdot]$ using a V-statistic versus a U-statistic, with the latter being the minimum variance unbiased estimator (van der Vaart, 1998, Chapter 12), (Zhou et al., 2021). In order to reduce the size n_1 of the augmented data, we consider the following schemes for drawing samples from P_1.

- **Sampling with (without) replacement**: One generates n_1 draws $\{\tilde{a}_i, \tilde{w}_i\}_{i=1}^{n_1}$ where $\{\tilde{a}\}_{i=1}^{n_1}$ is obtained by sampling from $\{a_i\}_{i=1}^{n_0}$ with (without) replacement and, $\{\tilde{w}\}_{i=1}^{n_1}$ is obtained by sampling from $\{w_i\}_{i=1}^{n_0}$ with (without) replacement and independently of $\{\tilde{a}\}_{i=1}^{n_1}$.
- **Permutation**: One generates $n_1 = n_0$ draws $\{\tilde{a}_i, w_i\}_{i=1}^{n_0}$ where $\{\tilde{a}\}_{i=1}^{n_0}$ is obtained by sampling from $\{a_i\}_{i=1}^{n_0}$ without replacement, i.e., by permuting the treatments.
- **Derangement**: A permutation is used that has no fixed points, also called a derangement.
- **m-Permutations/m-Derangements**: The permutation/derangement method is repeated m times to obtain $n_1 = mn_0$ draws.
- **Train-time Sampling/Permutation/Derangement**: For learning algorithms based on stochastic gradient descent, the sampling/permutation/derangement methods can be repeated for each learning iteration or batch sample. See, e.g., Belghazi et al. (2018, Algorithm 1) for a similar proposal where minibatches are permuted in neural network training.

Given the absence of clear practical recommendations regarding the sampling scheme for stabilized weight learning, we compare several methods through a numerical study in Section 3. Specifically, we compare sampling with replacement ($n_1 = mn_0$) versus m-permutations and m-derangements for $m \in \{1, 2, 5, 10\}$. For neural network based models we also compare train-time sampling with replacement/permutation/derangement, each with $n_1 = n_0$.

Augmentation Example 5 (Natural Mediation Effect)**.** Consider the density ratio $\alpha_0^{(\mathrm{NME})}(x)$ from Example 7. Other proposals for learning $\alpha_0^{(\mathrm{NME})}(x)$, such as those of Wu and Benkeser (2024), are based on decomposing the density ratio into products of odds functions from probabilistic classifiers. Similarly, we exploit the decomposition of the intervention distribution as

$$\begin{aligned} p_1(m, a, w) &= p_{M|A,W}(m \mid a^\dagger, w)\mathbb{I}\{a = a'\}p_W(w) \\ &= p_{M,W|A}(m, w \mid a^\dagger)\mathbb{I}\{a = a'\}\left\{\frac{p_W(w)}{p_{W|A}(w \mid a^\dagger)}\right\}, \end{aligned}$$

which suggests a two stage procedure: first learn the density ratio $\alpha_0^{(1)}(w) = p_W(w)/p_{W|A}(w \mid a^\dagger)$; then generate weighted samples from P_1 as (m_i, a', w_i) where i indexes the $n_1 = \sum_{i=1}^{n_0} \mathbb{I}(a_i = a^\dagger)$ observations from P_0 where $a_i = a^\dagger$, and each sample has weight $\gamma_i = \hat{\alpha}^{(1)}(w_i)$ where $\hat{\alpha}^{(1)}$ is the learned density ratio from the first stage.

Remark 4. The problem of drawing samples from an intervention distribution P_1 is closely related to the problem of drawing counterfactual outcomes under an intervention. The latter problem is of interest in the generative modeling literature, with recent proposals combining transport map learning and importance weight learning in a double robust manner; see Luedtke and Fukumizu (2025) and references therein. Moreover, in longitudinal treatment settings, interventions can be considered with reference to the *natural value of treatment*, which is the treatment value that would be observed at time t if an intervention were carried out until time $t-1$, but discontinued thereafter. Such values could, in some sense, be viewed as outcomes of the intervention (Richardson and Robins, 2013; Young et al., 2014; Díaz et al., 2023).

3 Numerical Experiments

In our numerical experiments we consider estimating the average policy effect (Example 3) with policy $\pi(w) = 1$, the average shift effect (Example 4) with shift $\delta = 0.1$, and the average dose response curve (Example 6) using the importance weighted estimator

$$\frac{1}{n_{\mathrm{eval}}} \sum_{i=1}^{n_{\mathrm{eval}}} y_i \hat{\alpha}(x_i),$$

where $\hat{\alpha}$ is the relevant learned density ratio, based on $n_0 = 2,000$ training observations and $n_{\mathrm{eval}} = 10,000$ evaluation observations. We use a large independent evaluation sample so that our evaluation metrics provide a close approximation to their population analogues. Our main metrics of interest are the absolute bias in the importance weighted estimator versus an oracle where the density ratio is known (i.e., where $\hat{\alpha}$ is replaced with α_0 above) and the mean absolute error (MAE) and the root mean squared error (RMSE) in the learned density ratio.

Remark 5. By the triangle inequality and Hölder's inequality, the bias in the importance

weighted estimator, the MAE, and the RMSE are related as

$$\widehat{\text{Bias}} \equiv \frac{1}{n_{\text{eval}}} \sum_{i=1}^{n_{\text{eval}}} y_i \{\hat{\alpha}(x_i) - \alpha_0(x_i)\}$$
$$|\widehat{\text{Bias}}| \leq \|y\|_{\infty,\text{eval}} \underbrace{\|\hat{\alpha}(X) - \alpha_0(X)\|_{1,\text{eval}}}_{\text{MAE}}$$
$$|\widehat{\text{Bias}}| \leq \|y\|_{2,\text{eval}} \underbrace{\|\hat{\alpha}(X) - \alpha_0(X)\|_{2,\text{eval}}}_{\text{RMSE}}$$

where the empirical p-norm is defined as

$$\|\cdot\|_{p,\text{eval}} \equiv \left(\frac{1}{n_{\text{eval}}} \sum_{i=1}^{n_{\text{eval}}} |(\cdot)_i|^p \right)^{1/p}$$

and $p \to \infty$ recovers the supremum norm. These bounds are insightful, since the empirical bias in the importance weighted estimator depends on the specific outcome generating process $Y \sim P_{Y|X}$ in our simulation setup, but the MAE and RMSE do not. Therefore, the MAE and RMSE can also be used to bound other outcome generating processes that we did not consider.

We compare the following density ratio learners, described in more detail below: a Nadaraya–Watson kernel (NWK) estimator; the uLSIF and KLIEP algorithms; multilayer perceptron (MLP) neural networks and gradient boosted machines (GBMs) for the log density ratio, each trained using one of the divergences from Section 2.2. Additionally, for the binary treatment experiments, we consider the standard approach of directly learning the propensity score using a probabilistic classifier based on the binary cross-entropy loss. These propensity score methods are denoted MLP-PS and GBM-PS because they use the same MLP and GBM algorithms as the Bregman–Riesz regression methods. Full implementation details are provided in Appendix E.

The NWK estimator is based on the kernel density estimator

$$\hat{p}_X(x) = \frac{1}{n_0} \sum_{i=1}^{n_0} K_X(x, x_i)$$

where $K_X(\cdot,\cdot)$ denotes a Gaussian kernel with dimensions of X. Using this kernel density estimator, we obtain the NWK estimators

$$\hat{\alpha}^{(\text{ASE})}(x) = \frac{\hat{p}_X(a-\delta, w)}{\hat{p}_X(a, w)}$$
$$\hat{\alpha}^{(\text{SW})}(x) = \frac{\hat{p}_A(a)\hat{p}_W(w)}{\hat{p}_X(a, w)}.$$

The uLSIF and KLIEP algorithms are sieve-based estimators of the form

$$\hat{\alpha}(x) = \sum_{i=1}^{p} \hat{\theta}_i K_X(x, x_i)$$

where $\{x_i\}_{i=1}^p$ represents p random kernel centers, sampled from the numerator observations, and $\hat{\theta}_i \geq 0$ are estimated coefficients. The KLIEP algorithm estimates coefficients by minimizing the unnormalized Kullback–Leibler BRR using gradient descent with feasibility satisfaction. The uLSIF algorithm first minimizes the least-squares BRR without constraints, then clips and rescales the resulting coefficients to satisfy the constraint that $\hat{\theta}_i \geq 0$.

The MLP and GBM algorithms learn the log density ratio $\hat{f}$, which implies the density ratio $\hat{\alpha}(x) = \exp\{\hat{f}(x)\}$. For the MLPs, $\hat{f}(x) = f(x \mid \hat{\theta})$ is the output of an MLP neural network with weights and biases $\hat{\theta}$ learned by stochastic gradient descent. For the GBMs, $\hat{f}(x) = \eta\hat{f}_1(x) + \cdots + \eta\hat{f}_K(x)$ represents a sum of K decision trees $\hat{f}_i(x)$ with constant learning rate η. For both MLP and GBM algorithms, we use early stopping where the n_0 observations are split into an 80:20 train-to-validation ratio, with each data split augmented separately. The validation set is also used to tune additional hyperparameters. Both algorithms were trained to optimize an empirical BRR from Section 2.2, which we abbreviate as: least squares (-LS), Kullback–Leibler (-KL), negative binomial (-NB), and Itakura–Saito (-IS).

We consider a data generating process with covariates $X = (A, W)$, where $W \in \mathbb{R}^{20}$ is a vector of predictors with each component $W_i \sim \mathcal{N}(0, 1)$ being independent unit normal and outcome $Y \sim \mathcal{N}(A + AW_1 + W_1W_2 + W_3, 1)$. For binary treatment experiments, we let A be Bernoulli distributed with mean $p_{A|W}(1 \mid w) = \sigma\{|w_1| + (1 - 0.5w_2)w_3\}$ where $\sigma(u) = 1/(1 + \exp(-u))$ is the sigmoid function, and for continuous treatment experiments, we let $A \sim \mathcal{N}(cW_1, 1)$, with $c = 0.5$. For this data generating process, the average policy effect is 1, the average shift effect is $\delta + c = 0.6$, the average dose response curve is 0, and the density ratios are $\alpha_0^{(\text{APE})} = a/p_{A|W}(1 \mid w)$,

$$\alpha_0^{(\text{ASE})}(a, w) = \exp\left(\delta(a - cw_1) - \frac{\delta^2}{2}\right),$$
$$\alpha_0^{(\text{SW})}(a, w) = \exp\left(\frac{(a - cw_1)^2}{2} - \frac{a^2}{2(1 + c^2)} - \frac{\log(1 + c^2)}{2}\right).$$

The range between the 2.5 and 97.5 percentiles is $(0.0, 2.84)$ for $\alpha_0^{(\text{APE})}(X)$, $(0.82, 1.21)$ for $\alpha_0^{(\text{ASE})}(X)$, and $(0.34, 2.37)$ for $\alpha_0^{(\text{SW})}(X)$.

3.1 Results

For the average policy effect, we consider the augmentation scheme described in Augmentation Example 2, which we compare against the standard approach where the propensity score is learned directly (MLP-PS and GBM-PS). The results in Table 2 show that density ratio learning using Bregman–Riesz regression appears to perform better than propensity score-based methods in terms of our primary metrics. Additionally, we see that the choice of the Bregman divergence has some impact on learning performance, with the least squares divergence generally performing worse than the other divergences. This phenomenon is possibly explained by appealing to the weight expressions F'' in Table 1, with the least squares divergence assigning a constant weight to all values of the density ratio, versus the other divergences where the weight decreases as the density ratio increases. Consequently, it is possible that the least-squares divergence is overfitting to large density ratio values, degrading overall performance. Conversely, the negative binomial and Itakura–Saito divergences, which give relatively less weight to large density ratio values, appear to perform better for both the MLP and GBM learners.

For the average shift effect, we consider the modified treatment policy augmentation scheme presented in Augmentation Example 1. The results in Table 3, show that the KDE gives the best performance in terms of absolute bias; however, the MLP neural networks outperform all other algorithms in terms of MAE and RMSE, suggesting that MLP learners might generalize better to other outcome distributions not considered in this study. The choice of Bregman divergence appears to have no clear effect on MLP and GBM performance in this experiment, which could be because the true density ratios $\alpha_0^{(\text{ASE})}(X)$ are concentrated relatively close to 1, compared with $\alpha_0^{(\text{APE})}(X)$ and $\alpha_0^{(\text{SW})}(X)$, and hence overfitting to

larger density ratio values is less of a concern.

For the average dose response curve, we compare sampling with replacement ($n_1 = mn_0$) versus m-permutations and m-derangements for $m \in \{1, 2, 5, 10\}$ as described in Augmentation Example 4. For the MLP, we also compare train-time sampling (MLP-TT) with replacement/permutation/derangement, each with $n_1 = n_0$. For the MLP-TT results, the multiplier m affects only the sampling scheme in the validation split of the training data.

The results in Figure 1 show that the least squares divergence generally performs poorly for stabilized weight learning, possibly due to the least squares divergence overfitting to large density ratio values, as described above. Due to the noisy performance of the least squares divergence, we consider only the negative binomial, Itakura–Saito, and Kullback–Leibler divergences when drawing conclusions about the effect of the sampling scheme on learning stabilized weights. We see that the m-permutation and m-derangement methods give broadly similar results for all learners, and that for the MLP and GBM learners, m-permutation and m-derangement both tend to outperform the sampling with replacement method, though this improvement attenuates as the multiplier m increases. The MLP-TT learners have the best overall performance in this study, with no clear difference between m-permutation, m-derangement, and sampling with replacement. Moreover, the multiplier m has limited effect on the MLP-TT learners, which is to be expected since m affects only the validation split and not the training split for the MLP-TT algorithms.

Based on this numerical study we recommend the negative-binomial or Itakura–Saito divergences in settings where overfitting to large density ratio values may be a concern. One expects large density ratio values when there is poor overlap between the numerator and denominator distributions—a situation where Riesz representer and density ratio learning is known to be more challenging (D'Amour et al., 2021). For learning stabilized weights, we conclude that train-time augmentation with even a small validation set ($m = 1$) gives improved performance over using fixed sampling schemes. When train-time augmentation is not possible, however, we recommend m-permutations or m-derangements sampling with a modest multiplier ($m \geq 2$).

Table 2: Results for estimation of the average policy effect ($\pi(w) = 1$) and corresponding density ratio estimates obtained as the median over 100 dataset replicates. Learners are sorted by absolute bias (AB) in the importance weighted estimator.

Learner	AB	MAE	RMSE
MLP-PS	0.822	0.671	0.983
GBM-PS	0.63	0.585	0.901
uLSIF	0.269	0.276	0.705
KLIEP	0.249	0.267	0.703
GBM-LS	0.244	0.265	0.617
GBM-KL	0.188	0.224	0.585
GBM-NB	0.121	0.171	**0.539**
MLP-LS	0.0778	0.257	0.674
GBM-IS	0.0764	**0.158**	0.568
MLP-KL	0.0448	0.221	0.599
MLP-NB	0.0387	0.209	0.592
MLP-IS	**0.0343**	0.207	0.578

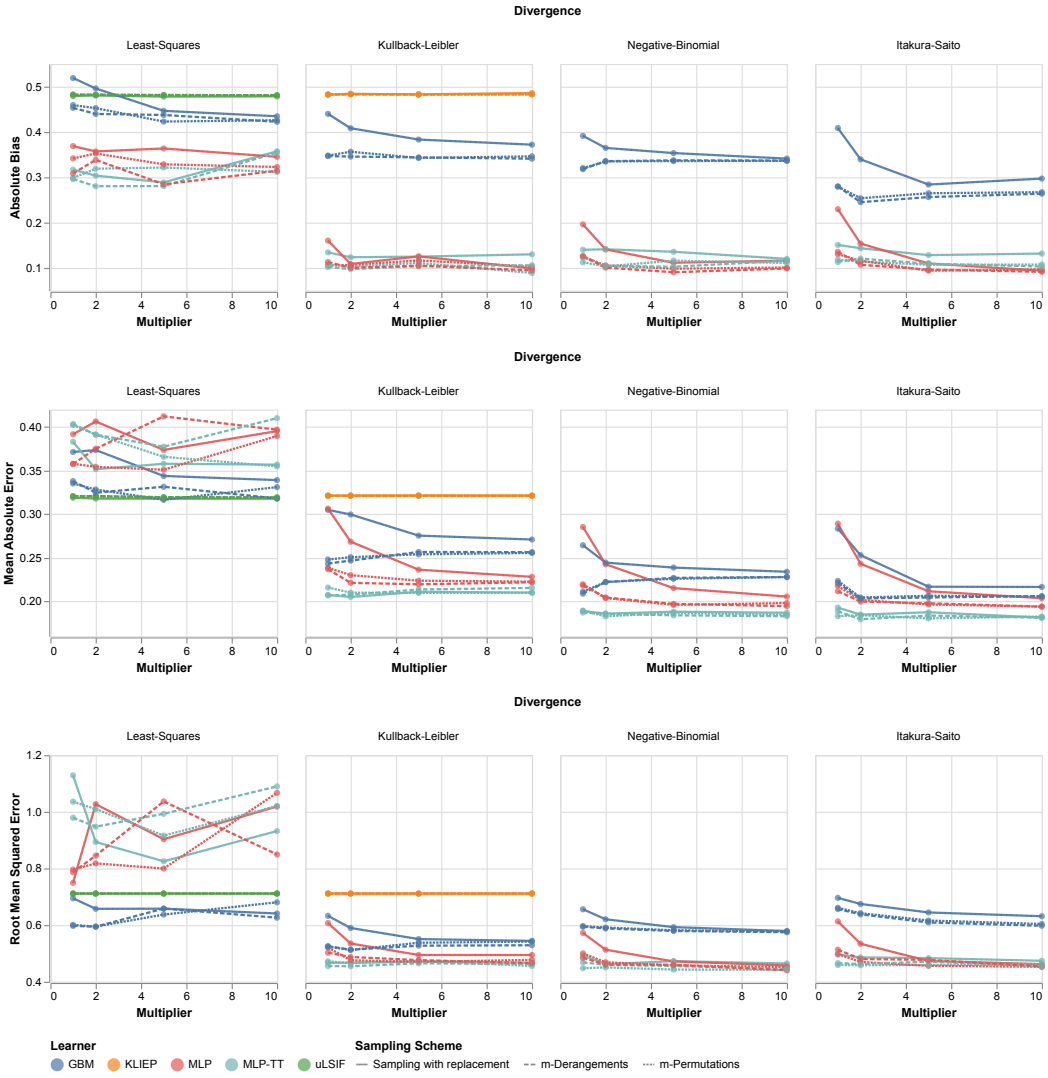

Figure 1: Absolute bias of the importance weighted average dose response curve estimator and MAE and RMSE in the corresponding stabilized weights versus multiplier m for each learner and sampling scheme. Each point represents a median over 100 dataset replicates. Results for the KDE, which do not depend on the sampling scheme, are not shown because they far exceed the axis limits in the MAE and RMSE plots (absolute bias: 0.390, MAE: 1.35, RMSE: 16.3).

Table 3: Results for estimation of the average shift effect ($\delta = 0.1$) and corresponding density ratio estimates, obtained as the median over 100 dataset replicates. Learners are sorted by absolute bias (AB) in the importance weighted estimator.

Learner	AB	MAE	RMSE
uLSIF	0.0955	0.0798	0.0997
GBM-IS	0.0933	0.0773	0.0978
KLIEP	0.0910	0.0738	0.0933
GBM-KL	0.0845	0.0778	0.0985
GBM-NB	0.0826	0.0772	0.0978
GBM-LS	0.0619	0.0766	0.0969
MLP-IS	0.0151	0.0381	0.0483
MLP-KL	0.0146	0.0379	0.0480
MLP-NB	0.0135	**0.0376**	**0.0480**
MLP-LS	0.0130	0.0383	0.0488
KDE	**0.0113**	0.0596	0.0782

4 Related work

Time Score Matching: One of the fundamental challenges in density ratio learning is dealing with poor overlap between the numerator and denominator distributions, known as the *density chasm* problem. To overcome this, Rhodes et al. (2020) propose a telescopic density ratio estimator, which decomposes the density ratio into a product of simpler interpolating density ratios. This approach is extended in the infinitesimal limit by Choi et al. (2022), who characterize the log density ratio as the integral

$$\log\{\alpha_0(x)\} = \log\left\{\frac{p_1(x)}{p_0(x)}\right\} = \int_0^1 \frac{\partial}{\partial t}\log\{p_t(x)\}\mathrm{d}t$$

where $p_t(x)$ is a user-defined path between $p_0(x)$ and $p_1(x)$ with univariate parameter $t \in [0, 1]$. Rather than learn the density ratio, they propose learning the *time score* function $\frac{\partial}{\partial t}\log\{p_t(x)\}$ and then using numerical integration to approximate the log density ratio. Variants of this method, with additional conditioning and path proposals, are studied by (Yu et al., 2025), with further connections to flow matching and optimal transport techniques (see references therein).

Other density ratio learning algorithms: In the current work we provide an overview of several foundational density ratio learning approaches; however, there is a rich literature of recent methods which we are unable to cover in detail. In particular we highlight the following papers for further study. Sugiyama et al. (2011) propose learning the density ratio in a linear subspace of $\mathcal{X}$ via dimensional reduction. Izbicki et al. (2014) propose a spectral algorithm that expands the density ratio in terms of the eigenfunctions of a kernel-based operator. Choi et al. (2021) use a generative neural network model to learn the density ratio in a latent space where the numerator and denominator distributions have improved overlap. Nguyen et al. (2024) extend the regularization theory for density ratio estimators in reproducing kernel Hilbert spaces. Wang et al. (2025) propose learning density ratios using a modified projection pursuit algorithm. Huk et al. (2025) describe connections between copula-based density ratio models and probabilistic classification.

Calibration: In addition to Riesz regression, recent work has proposed using the Riesz regression loss to calibrate Riesz representer estimates in order to reduce bias in estimators that rely on cross-fitted Riesz representer estimates (van der Laan et al., 2025; van der Laan and Alaa, 2025). For instance, consider that the data sample is split into a Riesz representer training sample and an evaluation sample. The uncalibrated procedure learns $\hat{\alpha}$

in the training sample, then evaluates the estimator in the evaluation sample using $\hat{\alpha}$. The calibrated procedure still learns $\hat{\alpha}$ in the training sample, but instead evaluates the estimator in the evaluation sample using $\hat{\alpha}^* = \hat{f} \circ \hat{\alpha}$, where $\hat{f} : \mathbb{R} \to \mathbb{R}$ is an isotonic (monotonic) calibration function that is learned in the evaluation sample by Riesz regression. In principle, different Riesz losses could be considered for the calibration step, such as those described in the current paper; however, more work is required to verify that the theoretical benefits of calibration are maintained.

5 Discussion

We propose Bregman–Riesz regression as a framework that unifies recent advances in Riesz regression by Chernozhukov et al. (2024) with existing proposals for learning the density ratio using Bregman divergences (Gutmann and Hirayama, 2011; Sugiyama et al., 2012). Compared with the conventional density ratio learning setting, learning density ratios in causal inference is uniquely challenging because one does not usually have access to samples from both the numerator and denominator densities, since the numerator typically represents a target intervention distribution of interest. We show how data augmentation techniques can be used to overcome this issue, making density ratio learning methods immediately applicable to many causal inference problems, with density ratio learning itself being a highly active area of research in the machine learning literature. Our discussion of data augmentation also clarifies the need for augmentation observed by Díaz et al. (2023) and Lee and Schuler (2025) when learning Riesz representers. Furthermore, we investigate several data augmentation strategies for learning the stabilized weight density ratio, providing concrete recommendations based on numerical experiments.

Through our numerical experiments, we also show that the choice of Bregman divergence can have a large impact on the performance of density ratio learning algorithms, especially when faced with large true density ratio values as would be expected when there is poor overlap between the numerator and denominator densities. In such cases, the least squares divergence appears to perform poorly compared with the negative binomial, Kullback–Leibler, and Itakura–Saito divergences, which we reason is due to differences in how each divergence assigns weight over the density ratio space. Future work in this area may consider whether an optimal divergence can be learned from data, for example, by iteratively learning the density ratio and then approximating the distribution of the learned values. Analogous procedures for linear regression have recently been proposed by Feng et al. (2025), where the distribution of regression residuals is used to inform a subsequent regression loss.

An additional direction for future work is to consider how one might tailor Bregman–Riesz regression to better learn Riesz representers in causal inference that are not density ratios. For example, the Riesz representer of the average treatment effect (Example 2) represents the signed difference between two density ratios. Since this Riesz representer can be negative, it is unsuitable for direct learning using Bregman divergences that are restricted to positive values (such as the Kullback—Leibler divergence). Perhaps, however, it would be possible to exploit the 'difference of density ratios' structure to construct suitable alternatives.

References

Ai, C., Linton, O., Motegi, K., and Zhang, Z. (2021). A unified framework for efficient estimation of general treatment models. *Quantitative Economics*, 12(3):779–816.

Athey, S., Imbens, G. W., and Wager, S. (2018). 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.

Athey, S. and Wager, S. (2021). Policy Learning With Observational Data. *Econometrica*, 89(1):133–161.

Banerjee, A., Merugu, S., Dhillon, I., and Ghosh, J. (2004). Clustering with Bregman Divergences. In *Proceedings of the 2004 SIAM International Conference on Data Mining*, pages 234–245. Society for Industrial and Applied Mathematics.

Belghazi, M. I., Baratin, A., Rajeswar, S., Ozair, S., Bengio, Y., Courville, A., and Hjelm, R. D. (2018). Mutual Information Neural Estimation. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 531–540. PMLR.

Ben-Michael, E., Feller, A., Hirshberg, D. A., and Zubizarreta, J. R. (2021). The Balancing Act in Causal Inference. arXiv:2110.14831 [stat].

Bickel, S., Bogojeska, J., Lengauer, T., and Scheffer, T. (2008). Multi-task learning for HIV therapy screening. In *Proceedings of the 25th international conference on Machine learning - ICML '08*, pages 56–63, Helsinki, Finland. ACM Press.

Bregman, L. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200–217.

Bruns-Smith, D. and Feller, A. (2022). Outcome Assumptions and Duality Theory for Balancing Weights. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151, pages 11037–11055. PMLR.

Cheng, K. and Chu, C. (2004). Semiparametric density estimation under a two-sample density ratio model. *Bernoulli*, 10(4).

Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K., and Robins, J. M. (2022). Locally Robust Semiparametric Estimation. *Econometrica*, 90(4):1501–1535.

Chernozhukov, V., Newey, W. K., Quintas-Martinez, V., and Syrgkanis, V. (2024). Automatic Debiased Machine Learning via Riesz Regression. arXiv:2104.14737 [math].

Chernozhukov, V., Newey, W. K., Singh, R., and Syrgkanis, V. (2020). Adversarial Estimation of Riesz Representers. *arXiv (2101.00009)*.

Choi, K., Liao, M., and Ermon, S. (2021). Featurized Density Ratio Estimation. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161, pages 172–182. PMLR.

Choi, K., Meng, C., Song, Y., and Ermon, S. (2022). Density Ratio Estimation via Infinitesimal Classification. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151, pages 2552–2573. PMLR.

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association*, 87(418):376–382.

Dikkala, N., Lewis, G., Mackey, L., and Syrgkanis, V. (2020). Minimax estimation of conditional moment models. *Advances in Neural Information Processing Systems*, 2020-December(NeurIPS).

Dudik, M., Langford, J., and Li, H. (2011). Doubly robust policy evaluation and learning. *Proceedings of the 28th International Conference on Machine Learning, ICML 2011*, pages 1097–1104.

Dunn, P. K. and Smyth, G. K. (2018). *Generalized Linear Models With Examples in R*. Springer Texts in Statistics. Springer New York, New York, NY.

Díaz, I. and van der Laan, M. (2012). Population Intervention Causal Effects Based on Stochastic Interventions. *Biometrics*, 68(2):541–549.

Díaz, I., Williams, N., Hoffman, K. L., and Schenck, E. J. (2023). Nonparametric Causal Effects Based on Longitudinal Modified Treatment Policies. *Journal of the American Statistical Association*, 118(542):846–857.

D'Amour, A., Ding, P., Feller, A., Lei, L., and Sekhon, J. (2021). Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 221(2):644–654.

Feng, O. Y., Kao, Y.-C., Xu, M., and Samworth, R. J. (2025). Optimal convex M-estimation via score matching. arXiv:2403.16688 [math].

Grunwald, P. D. and Dawid, A. P. (2004). Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. *The Annals of Statistics*, 32(4). arXiv:math/0410076.

Gutmann, M. U. and Hirayama, J.-i. (2011). Bregman divergence as general framework to estimate unnormalized statistical models. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*. AUAI Press.

Hainmueller, J. (2012). Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies. *Political Analysis*, 20(1):25–46.

Haneuse, S. and Rotnitzky, A. (2013). Estimation of the effect of interventions that modify the received treatment. *Statistics in Medicine*, 32(30):5260–5277.

Hardle, W. and Stoker, T. M. (1989). Investigating Smooth Multiple Regression by the Method of Average Derivatives. *Journal of the American Statistical Association*, 84(408):986.

Hido, S., Tsuboi, Y., Kashima, H., Sugiyama, M., and Kanamori, T. (2011). Statistical outlier detection using direct density ratio estimation. *Knowledge and Information Systems*, 26(2):309–336.

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

Huang, J., Gretton, A., Borgwardt, K., Schölkopf, B., and Smola, A. (2006). Correcting sample selection bias by unlabeled data. In Schölkopf, B., Platt, J., and Hoffman, T., editors, *Advances in Neural Information Processing Systems*, volume 19. MIT Press.

Huk, D., Steel, M., and Dutta, R. (2025). Your copula is a classifier in disguise: classification-based copula density estimation. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258, pages 3790–3798. PMLR.

Hyvärinen, A. (2005). Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(24):695–709.

Imai, K. and Ratkovic, M. (2014). Covariate Balancing Propensity Score. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1):243–263.

Imbens, G. W. and Newey, W. K. (2009). Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity. *Econometrica*, 77(5):1481–1512.

Izbicki, R., Lee, A. B., and Schafer, C. M. (2014). High-Dimensional Density Ratio Estimation with Extensions to Approximate Likelihood Computation. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33, pages 420–429, Reykjavik, Iceland. PMLR.

Kallus, N. and Uehara, M. (2020). Double Reinforcement Learning for Efficient Off-Policy Evaluation in Markov Decision Processes. arXiv:1908.08526 [cs].

Kanamori, T., Hido, S., and Sugiyama, M. (2009). A Least-squares Approach to Direct Importance Estimation. *Journal of Machine Learning Research*, 10(48):1391–1445.

Kennedy, E. H. (2019). Nonparametric Causal Effects Based on Incremental Propensity Score Interventions. *Journal of the American Statistical Association*, 114(526):645–656.

Kennedy, E. H. (2023). Towards optimal doubly robust estimation of heterogeneous causal effects. arXiv:2004.14497 [math].

Kennedy, E. H., Ma, Z., McHugh, M. D., and Small, D. S. (2017). Non-parametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 79(4):1229–1245.

Khan, S. and Tamer, E. (2010). Irregular Identification, Support Conditions, and Inverse Weight Estimation. *Econometrica*, 78(6):2021–2042.

Lafferty, J. (1999). Additive models, boosting, and inference for generalized divergences. In *Proceedings of the twelfth annual conference on Computational learning theory*, pages 125–133, Santa Cruz California USA. ACM.

Lee, K. J. and Schuler, A. (2025). RieszBoost: Gradient Boosting for Riesz Regression. arXiv:2501.04871 [stat].

Liu, S., Yamada, M., Collier, N., and Sugiyama, M. (2013). Change-point detection in time-series data by relative density-ratio estimation. *Neural Networks*, 43:72–83.

Lou, A., Meng, C., and Ermon, S. (2024). Discrete Diffusion Modeling by Estimating the Ratios of the Data Distribution. arXiv:2310.16834 [stat].

Luedtke, A. and Fukumizu, K. (2025). DoubleGen: Debiased Generative Modeling of Counterfactuals. arXiv:2509.16842 [stat].

Lyu, S. (2009). Interpretation and generalization of score matching. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, pages 359–366, Montreal, Quebec, Canada. AUAI Press.

Meng, C., Choi, K., Song, J., and Ermon, S. (2023). Concrete Score Matching: Generalized Score Matching for Discrete Data. arXiv:2211.00802 [cs].

Menon, A. K. and Ong, C. S. (2016). Linking losses for density ratio and class-probability estimation. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pages 304–313. PMLR.

Morris, C. N. (1983). Natural Exponential Families with Quadratic Variance Functions: Statistical Theory. *The Annals of Statistics*, 11(2).

Newey, W. K. and Stoker, T. M. (1993). Efficiency of Weighted Average Derivative Estimators and Index Models. *Econometrica*, 61(5):1199.

Nguyen, D. H., Zellinger, W., and Pereverzyev, S. (2024). On Regularized Radon–Nikodym Differentiation. *Journal of Machine Learning Research*, 25(266):1–24.

Nguyen, X., Wainwright, M. J., and Jordan, M. I. (2007). Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization. In *Advances in Neural Information Processing Systems*, volume 20, pages 1089–1096.

Pearl, J. (2001). Direct and Indirect Effects. In *Proceedings of the Seventeenth Conference on Uncertainy in Artificial Intelligence*, UAI'01, pages 411–420, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Petersen, M. L., Porter, K. E., Gruber, S., Wang, Y., and van der Laan, M. J. (2012). Diagnosing and responding to violations in the positivity assumption. *Statistical Methods in Medical Research*, 21(1):31–54.

Qin, J. (1998). Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika*, 85(3):619–630.

Reid, M. D. and Williamson, R. C. (2009). Surrogate regret bounds for proper losses. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 897–904, Montreal Quebec Canada. ACM.

Rhodes, B., Xu, K., and Gutmann, M. U. (2020). Telescoping Density-Ratio Estimation. In *Advances in Neural Information Processing Systems*, volume 33, pages 4905–4916. Curran Associates, Inc.

Richardson, T. S. and Robins, J. M. (2013). Single World Intervention Graphs (SWIGs): A Unification of the Counterfactual and Graphical Approaches to Causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, 128.

Robins, J. M. and Gill, R. D. (2001). Causal Inference for Complex Longitudinal Data: The Continuous Case. *Annals of Statistics*, 29(6):1785–1811.

Robins, J. M. and Greenland, S. (1992). Identifiability and Exchangeability for Direct and Indirect Effects:. *Epidemiology*, 3(2):143–155.

Robins, J. M., Hernán, M. Á., and Brumback, B. (2000). Marginal Structural Models and Causal Inference in Epidemiology:. *Epidemiology*, 11(5):550–560.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rothenhäusler, D. and Yu, B. (2020). Incremental causal effects. arXiv:1907.13258 [stat].

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

Song, Y. and Ermon, S. (2019). Generative Modeling by Estimating Gradients of the Data Distribution. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, volume 32, pages 11918–11930. Curran Associates, Inc.

Sugiyama, M., Nakajima, S., Kashima, H., von Bünau, P., and Kawanabe, M. (2007). Direct Importance Estimation with Model Selection and Its Application to Covariate Shift Adaptation. In *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc.

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

Sugiyama, M., Yamada, M., von Bünau, P., Suzuki, T., Kanamori, T., and Kawanabe, M. (2011). Direct density-ratio estimation with dimensionality reduction via least-squares hetero-distributional subspace search. *Neural Networks*, 24(2):183–198.

Susmann, H. P., McClean, A., and Díaz, I. (2025). Non-overlap Average Treatment Effect Bounds. arXiv:2509.20206 [stat].

van der Laan, L. and Alaa, A. (2025). Generalized Venn and Venn-Abers Calibration with Applications in Conformal Prediction. arXiv:2502.05676 [stat].

van der Laan, L., Luedtke, A., and Carone, M. (2025). Doubly robust inference via calibration. arXiv:2411.02771 [stat].

van der Laan, M. J. and Luedtke, A. R. (2014). Targeted Learning of the Mean Outcome under an Optimal Dynamic Treatment Rule. *Journal of Causal Inference*, 3(1):61–95.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, 1 edition.

Wang, M., Huang, W., Gong, M., and Zhang, Z. (2025). Projection Pursuit Density Ratio Estimation. arXiv:2506.00866 [stat].

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss—Newton method. *Biometrika*, 61(3):439–447.

Wu, W. and Benkeser, D. (2024). A Density Ratio Super Learner. arXiv:2408.04796 [stat].

Young, J. G., Hernán, M. A., and Robins, J. M. (2014). Identification, Estimation and Approximation of Risk under Interventions that Depend on the Natural Value of Treatment Using Observational Data. *Epidemiologic Methods*, 3(1):1–19.

Yu, H., Klami, A., Hyvärinen, A., Korba, A., and Chehab, O. (2025). Density Ratio Estimation with Conditional Probability Paths. In *Proceedings of the 42nd International Conference on Machine Learning*, pages 781–788. arXiv:2502.02300 [cs].

Zhou, Z., Mentch, L., and Hooker, G. (2021). V -statistics and Variance Estimation. *Journal of Machine Learning Research*, 22(287):1–48.

Zubizarreta, J. R. (2015). Stable Weights that Balance Covariates for Estimation With Incomplete Outcome Data. *Journal of the American Statistical Association*, 110(511):910–922.

A Score Matching

Here we show how the Riesz representer setting described in Section 2.1 generalizes when the Hilbert space $\mathcal{H}$ is made up of vector functions. Such an extension forms the basis of *score matching* techniques used in score-based generative diffusion models (Hyvärinen, 2005; Lyu, 2009; Song and Ermon, 2019).

The setup extends as follows: consider a random variable $X \in \mathcal{X}$ distributed according to an unknown distribution P_0. Let $\mathcal{H}$ be a Hilbert space of real-valued functions $f : \mathcal{X} \to \mathbb{R}^d$ with inner product $\langle f, g \rangle \equiv \mathbb{E}_{P_0}[f(X)^\top g(X)]$ for $f, g \in \mathcal{H}$. Letting $\mathbb{H} : \mathcal{H} \to \mathbb{R}$ be a continuous and linear map, Riesz representation theorem implies that there exists a unique Riesz representer $\alpha_0 \in \mathcal{H}$ such that $\mathbb{H}(f) = \langle f, \alpha_0 \rangle$ for all $f \in \mathcal{H}$. Moreover, the Bregman divergence in (1) generalizes as

$$D_F(\alpha_0 \| \alpha) = \mathbb{E}_{P_0}\left[F\{\alpha_0(X)\} - F\{\alpha(X)\} - \nabla F\{\alpha(X)\}^\top \{\alpha_0(X) - \alpha(X)\}\right]$$

where $F : \mathbb{R}^d \to \mathbb{R}$ is a convex differentiable function with Jacobian ∇F. Discarding the constant term $\mathbb{E}_{P_0}\left[F\{\alpha_0(X)\}\right]$, one obtains the BRR

$$\mathcal{R}_F(\alpha) = \mathbb{E}_{P_0}\left[\nabla F\{\alpha(X)\}^\top \alpha(X) - F\{\alpha(X)\}\right] - \mathbb{H}(\nabla F \circ \alpha).$$

For score matching, consider the setting where X is a continuous random variable over $\mathcal{X} \subseteq \mathbb{R}^d$, and we define the linear map $\mathbb{H}(f) = -\mathbb{E}_{P_0}[\mathrm{Tr}\{\nabla f(X)\}]$ where ∇f is the $d \times d$ Jacobian matrix of f and Tr denotes the matrix trace. Note that this linear map is of the average linear functional type described in Setting 2 of Section 2.1, and that the trace of the Jacobian is also called the *divergence* in vector calculus. Under mild regularity conditions, including that p_X is differentiable, one can write $\mathbb{H}(f) = \langle f, \alpha_0^{(\mathrm{SM})} \rangle$ where $\alpha_0^{(\mathrm{SM})}(x) = \nabla \log\{p_X(x)\}$.

When $F(t) = \|t\|^2$ is the squared Euclidean distance, with $\nabla F(t) = 2t$, then one obtains the least-squares divergence and BRR

$$\begin{aligned} D_F(\alpha_0 \| \alpha) &= \mathbb{E}_{P_0}\left[\|\alpha(X) - \alpha_0(X)\|^2\right] \\ \mathcal{R}_F(\alpha) &= \mathbb{E}_{P_0}\left[\|\alpha(X)\|^2 + 2\,\mathrm{Tr}\{\nabla\alpha(X)\}\right]. \end{aligned}$$

The score matching procedure minimizes an empirical approximation of this least-squares BRR.

B Integral form of the Bregman divergence

Claim:

$$\mathbb{E}_{P_0}\left[F\{g_0(X)\} - F\{g(X)\} - F'\{g(X)\}\{g_0(X) - g(X)\}\right] = \int_{-\infty}^{\infty} \mathbb{E}_{P_0}\left[r\{t \mid g_0(X), g(X)\}\right] F''(t)\mathrm{d}t$$

Proof: Start by noting that

$$\begin{aligned} \int_{-\infty}^{\infty} r\{t \mid v_0, v\} F''(t)\mathrm{d}t &= \int_{\min(v_0,v)}^{\max(v_0,v)} |v_0 - t| F''(t)\mathrm{d}t \\ &= \int_{v}^{v_0} (v_0 - t) F''(t)\mathrm{d}t \\ &= F(v_0) - F(v) - F'(v)(v_0 - v) \end{aligned}$$

where the last line follows by integration by parts. Hence,

$$\begin{aligned}&\mathbb{E}_{P_0}\left[F\{g_0(X)\} - F\{g(X)\} - F'\{g(X)\}\{g_0(X) - g(X)\}\right]\\ =&\mathbb{E}_{P_0}\left[\int_{-\infty}^{\infty} r\{t \mid g_0(X), g(X)\}F''(t)\mathrm{d}t\right]\end{aligned}$$

Changing the order of integration and expectation completes the proof.

C Generalized linear model deviance and the Bregman divergence

In this Appendix we review theory for the generalized linear model deviance, comparing it to the Bregman divergence in the main text. Our review is based on Dunn and Smyth (2018, Chapter 5) and Wedderburn (1974). Let $Y \in \mathbb{R}$ be distributed according to a one-parameter exponential family distribution with canonical parameter $\theta_0 \in \mathbb{R}$ and density

$$p(y \mid \theta_0) = a(y, \phi) \exp\left\{\frac{y\theta_0 - \kappa(\theta_0)}{\phi}\right\}$$

where $\kappa(\theta)$ is a known cumulant function, $a(y, \phi)$ is a normalizing function, and the dispersion parameter $\phi > 0$ is known. For this distribution, $\mathbb{E}[Y] = \kappa'(\theta_0)$ and $\mathrm{var}(Y) = \phi\kappa''(\theta_0)$. Since $\kappa''(\theta) > 0$ is a variance, $k'(\theta)$ is strictly increasing with an inverse, called the link function, which we denote g. Hence, the variance function is defined as $V(\mu) = \kappa''\{g(\mu)\}$ with $\mathrm{var}(Y) = \phi V(\mu_0)$ where $\mu_0 \equiv \mathbb{E}[Y]$.

The unit deviance can be defined as

$$d(y, \mu) \equiv 2\int_{\mu}^{y} \frac{y - t}{V(t)}\mathrm{d}t.$$

Making the substitution $u = g(t)$ with $\tilde{\theta} \equiv g(y)$ and $\theta \equiv g(\mu)$ one obtains the more familiar likelihood ratio expression

$$\begin{aligned}d(y, \mu) &= 2\int_{\theta}^{\tilde{\theta}} y - \kappa'(u)\mathrm{d}u\\ &= 2\left\{y(\tilde{\theta} - \theta) - \kappa(\tilde{\theta}) + \kappa(\theta)\right\}\\ &= 2\left\{\log[p(y \mid \tilde{\theta})] - \log[p(y \mid \theta)]\right\}.\end{aligned}$$

For some examples of variance functions and their implied unit deviances see Dunn and Smyth (2018, Table 5.1). Using the the unit deviance, one can reparameterize the density function as

$$p(y \mid \mu_0) = b(y, \phi) \exp\left\{\frac{-d(y, \mu_0)}{2\phi}\right\}$$

where b is a different normalizing function.

Additionally, the unit deviance can be used to define the population deviance, which we

rewrite using the cost sensitive point-wise regret

$$\begin{aligned}\mathbb{E}[d(Y,\mu)] &= \mathbb{E}\left[2\int_{\mu}^{Y}\frac{Y-t}{V(t)}\mathrm{d}t\right] \\ &= \mathbb{E}\left[\int_{-\infty}^{\infty} r(t \mid Y,\mu)\frac{2}{V(t)}\mathrm{d}t\right] \\ &= \int_{-\infty}^{\infty}\mathbb{E}\left[r(t \mid Y,\mu)\right]\frac{2}{V(t)}\mathrm{d}t.\end{aligned}$$

This generalized linear model deviance is of the same form as the integral expression for the Bregman divergence in (2), with the weight $F''(t) = 2/V(t)$. Thus, the Bregman convex generating function F implies a weighting scheme for model errors, analogously to the variance function $V(\mu)$ in generalized linear model regression. In particular, according to Morris (1983), the negative binomial distribution with the number of successes fixed to one has variance function $V(\mu) = \mu + \mu^2$ and hence, in our main manuscript, we refer to the divergence with $F''(t) = t^{-1}(1+t)^{-1}$ as the 'negative-binomial' divergence.

D Reciprocally Symmetric divergences

In the density ratio setting, one can reverse the roles of P_0 and P_1, provided that $P_0 \ll P_1$, in which case the resulting density ratio is $1/\alpha_0(x)$ and the Bregman divergence becomes

$$D_F^{(\mathrm{rec})}\left(g_0 \| g\right) \equiv \mathbb{E}_{P_1}\left[F\left\{g_0(X)\right\} - F\left\{g(X)\right\} - F'\left\{g(X)\right\}\left\{g_0(X) - g(X)\right\}\right].$$

It is relatively straightforward to check that for a function $x \mapsto 1/\alpha(x)$

$$D_F^{(\mathrm{rec})}\left(\frac{1}{\alpha_0}\middle\|\frac{1}{\alpha}\right) = D_{F^{(\mathrm{rec})}}\left(\alpha_0 \| \alpha\right)$$

where $F^{(\mathrm{rec})}(t) \equiv tF(1/t)$, which is convex for $t > 0$. For example, for the convex generating function $F(t) = -\log(t) - 1$ one obtains $F^{(\mathrm{rec})}(t) = t\log(t) - t$. We see, therefore, that the Itakura–Saito distance is equivalent to the unnormalized Kullback–Leibler divergence when the roles of P_0 and P_1 are reversed.

When $F(t) = tF(1/t)$ we say that the divergence is *reciprocally symmetric*. When F is twice differentiable, this condition can be expressed as $F''(t) = t^{-3}F''(1/t)$, which relates to the integral representation of the Bregman divergence in (2). The negative binomial is a reciprocally symmetric divergence. In Table 4 we illustrate the form of $F^{(\mathrm{rec})}$ for several examples.

Table 4: Reciprocal convex generator functions $F^{(\mathrm{rec})}$ constructed from convex generating functions F

$F(t)$	$F^{(\mathrm{rec})}(t)$
t^2	$1/t$
$t\log(t) - t$	$-\log(t) - 1$
$-\log(t) - 1$	$t\log(t) - t$
$t\log(t) - (1+t)\log(1+t)$	$t\log(t) - (1+t)\log(1+t)$

E Numerical Study Implementation Details

Here we detail implementation details for the numerical study presented in Section 3. Full replication code is available at `https://github.com/CI-NYC/densityratios`.

For Kernel Density Estimation we use the `jax.scipy.stats.gaussian_kde` function available in version `0.7.2` of JAX (`http://github.com/google/jax`). To determine the bandwidth we use the default 'scott' rule of thumb.

For the KLIEP and uLSIF algorithms we re-implement the reference implementations from `https://github.com/srome/pykliep` and `https://github.com/hoxo-m/densratio_py` respectively. Our implementation adds a convergence criteria to the reference KLIEP algorithm (coefficients change by less than 10^{-6} in consecutive iterations), and use JAX function compilation to reduce runtime. The KLIEP and uLSIF implementations use the hyperparameter search algorithms described in the original papers, respectively: Sugiyama et al. (2007, Figure 1) and Kanamori et al. (2009, Figure 2). We search over the hyperparameter values described in Table 5.

For the GBM learners we use version `4.6.0` of LightGBM (`https://github.com/microsoft/LightGBM`). The number of trees (maximum 1000) is determined by early stopping, via loss stagnation in a validation set (minimum improvement of 0, patience 10). The validation set is also used to tune hyperparameters according to the search grid described in Table 5. Specifically, the LightGBM training algorithm is run for each hyperparameter combination and the learner with the lowest validation loss is selected.

For the MLP learners we use a custom implementation, built using version `2.8.0` of Pytorch (`https://github.com/pytorch/pytorch`). MLP weights and biases are trained via stochastic gradient descent (batch size 125, ADAM optimizer, learning rate 10^{-4}). The number of training epochs (maximum 1000) is determined by early stopping, via loss stagnation in a validation set (minimum improvement of 0, patience 5). The validation set is also used to tune hyperparameters according to the search grid described in Table 5. Specifically, our training algorithm is run for each hyperparameter combination and the learner with the lowest validation loss is selected.

Table 5: Hyperparameter search values for each learning algorithm.

Learner	Hyperparameter	Grid
KLIEP/uLSIF	basis dimension	$\{100, 200, 500\}$
KLIEP/uLSIF	bandwidth	$\{1, 2, 5, 10\}$
uLSIF	smoothing parameter	$\{0.0, 0.1, 0.5, 1.0\}$
GBM	maximum tree depth	$\{10, 20, 50\}$
GBM	learning rate	$\{10^{-3}, 10^{-4}\}$
GBM	bagging fraction	$\{0.8, 1.0\}$
MLP	depth	$\{2, 3\}$
MLP	width	$\{20, 50\}$