Riesz Representer Fitting under Bregman Divergence: A Unified Framework for Debiased Machine Learning

Masahiro Kato*

Mizuho-DL Financial Technology, Co., Ltd.
Osaka Metropolitan University

January 19, 2026

Abstract

Estimating the Riesz representer is central to debiased machine learning for causal and structural parameter estimation. We propose generalized Riesz regression, a unified framework for estimating the Riesz representer by fitting a representer model via Bregman divergence minimization. This framework includes various divergences as special cases, such as the squared distance and the Kullback–Leibler (KL) divergence, where the former recovers Riesz regression and the latter recovers tailored loss minimization. Under suitable pairs of divergence and model specification (link functions), the dual problems of the Riesz representer fitting problem correspond to covariate balancing, which we call automatic covariate balancing. Moreover, under the same specifications, the sample average of outcomes weighted by the estimated Riesz representer satisfies Neyman orthogonality even without estimating the regression function, a property we call automatic Neyman orthogonalization. This property not only reduces the estimation error of Neyman orthogonal scores but also clarifies a key distinction between debiased machine learning and targeted maximum likelihood estimation (TMLE). Our framework can also be viewed as a generalization of density ratio fitting under Bregman divergences to Riesz representer estimation, and it applies beyond density ratio estimation. We provide convergence analyses for both reproducing kernel Hilbert space (RKHS) and neural network model classes. A Python package for generalized Riesz regression is available at https://github.com/MasaKat0/grr.

1 Introduction

The Riesz representer plays a crucial role in debiased machine learning for a variety of causal and structural parameter estimation problems (Chernozhukov et al., 2022b), such as Average Treatment Effect (ATE) estimation (Imbens & Rubin, 2015), Average Marginal Effect

*Email: mkato-csecon@g.ecc.u-tokyo.ac.jp.

(AME) estimation, Average Policy Effect (APE) estimation, and covariate shift adaptation (Shimodaira, 2000; Uehara et al., 2020; Chernozhukov et al., 2025). The Riesz representer arises from the Riesz representation theorem for the parameter functional, and it also has a close connection to semiparametric efficiency bounds (Newey, 1994). In particular, by using the Riesz representer appropriately, we can obtain semiparametric estimators that are asymptotically linear with the efficient influence function, which is also referred to as the Neyman orthogonal score (Chernozhukov et al., 2018).

Straightforward approaches to approximating the Riesz representer often rely on intermediate steps. For example, in ATE estimation, the Riesz representer can be written in terms of the inverse propensity score. A straightforward approach is to estimate the propensity score and then construct the Riesz representer by taking its inverse. In covariate shift adaptation, the Riesz representer is given by the density ratio, the ratio of two probability density functions (pdfs). A straightforward approach is to estimate the two pdfs and then take their ratio. However, it is unclear whether such approaches perform well for the task of Riesz representer estimation because they are not designed to minimize the estimation error of the Riesz representer itself.

To address this issue, end-to-end approaches for Riesz representer estimation have been explored, including Riesz regression (Chernozhukov et al., 2021; Chen et al., 2014; Kanamori et al., 2009). In particular, in ATE estimation, entropy balancing weights (Hainmueller, 2012), stable balancing weights (Zubizarreta, 2015), and tailored loss minimization (Zhao, 2019) have been proposed. In covariate shift adaptation, direct density ratio estimation methods have been proposed (Sugiyama et al., 2012).

This study provides a general framework for Riesz representer estimation that accommodates these methods. We formulate Riesz representer estimation as a problem of fitting a Riesz representer model to the true Riesz representer under a Bregman divergence (Bregman, 1967). The Bregman divergence includes various discrepancy measures, such as squared distance and Kullback–Leibler (KL) divergence, as special cases. We measure the discrepancy between a Riesz representer model and the true Riesz representer using a Bregman divergence and train the model by minimizing this divergence, where the discrepancy is interpreted as a loss function. Although the true Riesz representer is unknown, we derive an objective function that does not involve the true Riesz representer and can be approximated using only observations. Therefore, we can train the Riesz representer model within an empirical risk minimization framework.

Notably, with the squared loss, the Bregman divergence minimization problem aligns with Riesz regression (Chernozhukov et al., 2021). With the KL divergence loss, the Bregman divergence minimization problem aligns with tailored loss minimization (Zhao, 2019). We note that Bruns-Smith et al. (2025) shows that stable balancing weights are dual solutions of Riesz regression, while Zhao (2019) shows that entropy balancing weights are dual solutions of tailored loss minimization. Thus, our Bregman divergence minimization formulation unifies these existing methods within a single framework.

In the following sections, we introduce our general setup (Section 2) and then propose our Riesz representer estimation method (Section 3). We refer to Riesz representer fitting under a Bregman divergence as generalized Riesz regression. We may also refer to it as Bregman–Riesz regression, direct bias correction term estimation, generalized tailored loss minimization, or generalized covariate balancing (Remark 3). However, we adopt the term

generalized Riesz regression because the choice of loss function is closely related to the choice of link function. For example, the duality relationship between Riesz regression and stable balancing weights holds when we use a linear link (a linear model) for the Riesz representer. In contrast, the duality relationship between tailored loss minimization and entropy balancing weights holds when we use a logistic link (a logistic model) for the Riesz representer. We make this correspondence explicit in Section 4. We refer to this property as automatic covariate balancing.

For our proposed framework, this paper provides convergence rate analyses, major applications, and several extensions. In Section 8, we show the convergence rate of the Riesz representer model to the true Riesz representer when using reproducing kernel Hilbert space (RKHS) regression and neural networks. In particular, we establish minimax optimality. In Section 5, we provide applications of our framework to ATE, AME, and APE estimation, as well as covariate shift adaptation. In Appendices H and I, we also introduce extensions of our framework that are developed in subsequent works. Kato (2025a) points out that nearest neighbor matching-based density ratio estimation proposed in Lin et al. (2023) is a special case of least-squares importance fitting (LSIF) for density ratio estimation (Kanamori et al., 2009). Moreover, since LSIF can be interpreted as Riesz regression, nearest neighbor matching-based ATE estimation can also be interpreted as ATE estimation via Riesz regression. Kato (2025c) proposes a Riesz representer estimation method based on score matching in diffusion models (Song et al., 2021).

Our framework is primarily built on results from Riesz regression, covariate balancing weights, and density ratio estimation. In particular, Bregman divergences have already been applied to density ratio estimation in Sugiyama et al. (2011b), where the authors also report the duality between empirical risk minimization and covariate balancing weights. Our framework generalizes these results to a broader class of applications and bridges the literature on density ratio estimation with causal inference, where Riesz regression and covariate balancing have, in parallel, studied estimation of the Riesz representer via empirical risk minimization and its dual formulation in terms of covariate balancing weights.

2 Setup

We first describe our general setup, which includes various tasks, such as ATE estimation, as special cases. We denote the pair $W \coloneqq (X, Y)$, where $Y \in \mathcal{Y}$ is an outcome and $X \in \mathcal{X}$ is a regressor vector. Here, $\mathcal{Y} \subseteq \mathbb{R}$ and $\mathcal{X} \subseteq \mathbb{R}^k$ are outcome and (k-dimensional) regressor spaces, respectively. Let P_0 be the distribution that generates W. We assume that we can observe n i.i.d. copies of W, denoted as

$$\mathcal{D} \coloneqq \{W_i\}_{i=1}^n = \{(X_i, Y_i)\}_{i=1}^n$$

We denote the regression function by $\gamma_0(x) \coloneqq \mathbb{E}_{P_0}[Y \mid X = x]$, where $\mathbb{E}_{P_0}$ denotes the expectation over P_0. We drop P_0 when the dependence is obvious.

Our goal is to estimate a parameter of interest of the form

$$\theta_0 \coloneqq \mathbb{E}[m(W, \gamma_0)],$$

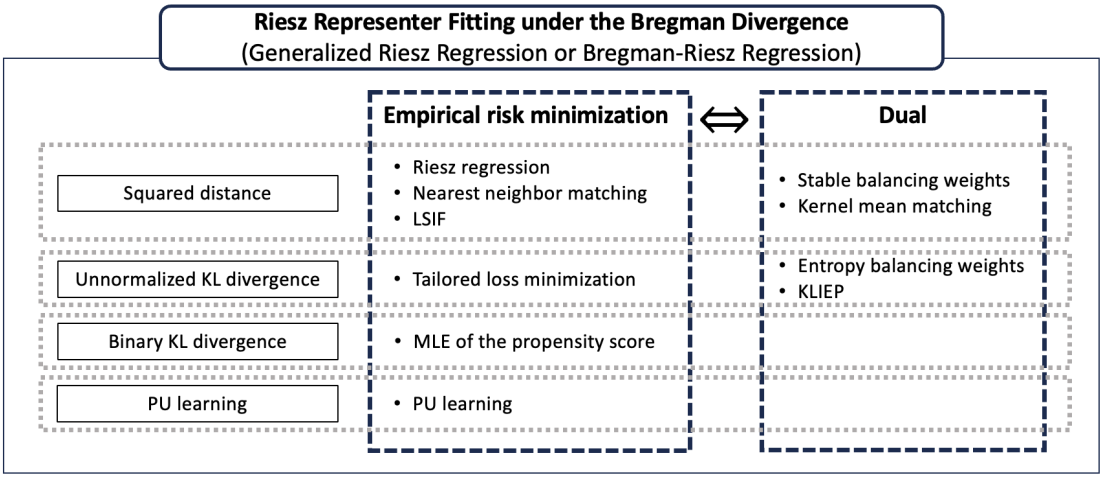

Figure 1: A unified framework for debiased machine learning via Riesz representer estimation and Bregman divergence minimization.

where $m(W, \gamma)$ is a functional that depends on data W and a regression function $\gamma \colon \mathcal{X} \to \mathcal{Y}$. Note that $m(W, \gamma)$ can receive any function $\gamma \colon \mathcal{X} \to \mathcal{Y}$, not limited to the "true" γ_0. In the following sections, when we first introduce a function (or functional) that depends on parameters of the data-generating process (DGP), we can replace it with an estimator that has the same mapping. Here, θ_0 depends on the functional m, and by changing m, we can derive various parameters as special cases, such as ATE, AME, and APE.

2.1 Riesz representer

For simplicity, we assume that the expected functional $\gamma \mapsto \mathbb{E}\big[m(W, \gamma)\big]$ is linear and continuous in γ, which implies that there is a constant $C > 0$ such that $\mathbb{E}\big[m(W, \gamma)\big]^2 \leq C\mathbb{E}\left[\gamma(X)^2\right]$ holds for all γ with $\mathbb{E}\left[\gamma(X)^2\right] < \infty$. From the Riesz representation theorem, there exists a function v_m with $\mathbb{E}\left[v_m(X)^2\right] < \infty$ such that

$$\mathbb{E}\left[m(W, \gamma)\right] = \mathbb{E}\left[v_m(X)\gamma(X)\right]$$

for all γ with $\mathbb{E}\left[\gamma(X)^2\right] < \infty$. We denote the function v_m by $\alpha_0 = v_m$, which is referred to as the Riesz representer (Chen & Liao, 2015; Chernozhukov et al., 2022b). In ATE estimation, the Riesz representer has also been referred to as the bias-correction term or the clever covariates (van der Laan, 2006; Schuler & van der Laan, 2024).

Remark. *This formulation follows Chernozhukov et al. (2022b) and can be generalized to non-linear maps $\gamma \mapsto \mathbb{E}\left[m(W, \gamma)\right]$. However, we do not present this generalization because it is not our main focus, and the linear case is sufficient for presenting our results.*

2.2 Neyman Orthogonal Scores

Let $\eta_0 = (\gamma_0, \alpha_0)$ be a pair of the nuisance parameters, where α_0 is the Riesz representer associated with the parameter functional m. The Neyman orthogonal score is defined as

$$\psi(W; \eta, \theta) \coloneqq m(W, \gamma) + \alpha(X)\big(Y - \gamma(X)\big) - \theta.$$

Here, it holds that

$$\mathbb{E}\left[\psi(W; \eta_0, \theta_0)\right] = 0,$$

which serves as the estimation equation (or moment condition) for estimating θ_0. Note that, by Neyman orthogonality, the Gateaux derivative with respect to η at η_0 vanishes as

$$\partial_\eta \mathbb{E}\left[\psi(W; \eta, \theta_0)\right]\big|_{\eta=\eta_0} = 0.$$

Thus, orthogonality ensures that first-order errors from estimating η_0 do not affect the asymptotic distribution of the final estimator $\widehat{\theta}$ of θ_0, provided cross fitting (or a Donsker condition) and mild convergence rate conditions on $\widehat{\eta}$ hold (Chernozhukov et al., 2018, 2022b).

By replacing the moment condition with its empirical analogue, we obtain an estimator $\widehat{\theta}$ of θ_0 as the value satisfying

$$\frac{1}{n}\sum_{i=1}^{n} \psi\left(W_i; \widehat{\eta}, \widehat{\theta}\right) = 0,$$

where $\widehat{\eta}$ denotes estimates of η_0 plugged into the Neyman orthogonal estimating equations. For convenience, we refer to the estimator

$$\widehat{\theta} \coloneqq \frac{1}{n}\sum_{i=1}^{n} \Big(m(W_i, \widehat{\gamma}) + \widehat{\alpha}(X_i)\big(Y_i - \widehat{\gamma}(X_i)\big)\Big)$$

as the Augmented Riesz Weighted (ARW) estimator, which corresponds to the Augmented Inverse Probability Weighting (AIPW) estimator in ATE estimation. In contrast, we refer to $\widehat{\theta} \coloneqq \frac{1}{n}\sum_{i=1}^{n} \widehat{\alpha}(X_i) Y_i$, which corresponds to the Inverse Probability Weighting (IPW) estimator in ATE estimation.

This estimation approach is traditionally called one-step bias correction, or the estimating equation approach (van der Vaart, 2002; Schuler & van der Laan, 2024). Recent work reformulates it as (automatic) debiased machine learning (Chernozhukov et al., 2018, 2022b,c, 2024). This estimator generalizes the augmented inverse probability weighting (AIPW) estimator in causal inference (Bang & Robins, 2005; Tsiatis, 2007).

2.3 Examples

We provide examples of the problems, along with the corresponding Riesz representers and Neyman orthogonal scores.

ATE estimation. Let the regressor X be $X \coloneqq (D, Z)$, where $D \in \{1, 0\}$ is a treatment indicator, and $Z \in \mathcal{Z}$ is a covariate vector, where $\mathcal{Z}$ denotes its support. Following the Neyman–Rubin framework, let $Y(1), Y(0) \in \mathcal{Y}$ be the potential outcomes for treated and control units. In ATE estimation, using the observations $\{(X_i, Y_i)\}_{i=1}^n$, our goal is to estimate the ATE, defined as

$$\theta_0^{\mathrm{ATE}} \coloneqq \mathbb{E}\big[m^{\mathrm{ATE}}\big(X, \gamma_0\big)\big], \quad m^{\mathrm{ATE}}(X, \gamma_0) \coloneqq \gamma_0(1, Z) - \gamma_0(0, Z).$$

To identify the ATE, we assume standard conditions such as unconfoundedness, positivity, and boundedness of the random variables, that is, $Y(1)$ and $Y(0)$ are independent of D given Z, there exists a universal constant $\epsilon \in (0, 1/2)$ such that $\epsilon < e_0(Z) < 1 - \epsilon$, and X, $Y(1)$, and $Y(0)$ are bounded.

In ATE estimation, the Riesz representer is given by

$$\alpha_0^{\mathrm{ATE}}(X) \coloneqq \frac{D}{e_0(Z)} - \frac{1 - D}{1 - e_0(Z)}.$$

This term is referred to by various names across different methods. In the classical semiparametric inference literature, it is called the bias-correction term (Schuler & van der Laan, 2024). In TMLE, it is called the clever covariates (van der Laan, 2006). In the debiased machine learning (DML) literature, it is called the Riesz representer (Chernozhukov et al., 2022b). The component $\frac{1}{e_0(Z)}$ may also be referred to as balancing weights (Imai & Ratkovic, 2013b; Hainmueller, 2012), the inverse propensity score (Horvitz & Thompson, 1952), or a density ratio (Sugiyama et al., 2012).

The Neyman orthogonal score is given as

$$\psi^{\mathrm{ATE}}\left(W; \eta, \theta\right) \coloneqq m^{\mathrm{ATE}}(W, \gamma) + \alpha^{\mathrm{ATE}}(X)\Big(Y - \gamma(X)\Big) - \theta.$$

Plugging in estimates of $\eta_0^{\mathrm{ATE}} \coloneqq \big(\gamma_0, \alpha_0^{\mathrm{ATE}}\big)$ and solving $\frac{1}{n}\sum_{i=1}^n \psi^{\mathrm{ATE}}\left(W_i; \widehat{\eta}, \theta\right) = 0$ for θ, we can obtain an ATE estimator as

$$\widehat{\theta}^{\mathrm{ATE}} \coloneqq \frac{1}{n}\sum_{i=1}^{n}\Big(\widehat{\alpha}^{\mathrm{ATE}}(X_i)\left(Y_i - \widehat{\gamma}(X_i)\right) + m^{\mathrm{ATE}}\left(X_i, \widehat{\gamma}\right)\Big).$$

This estimator is called the augmented inverse probability weighting (AIPW) estimator, or the doubly robust estimator.

AME estimation. Let the regressor X be $X = (D, Z)$ with a (scalar) continuous treatment D. We define the AME as

$$\theta_0^{\mathrm{AME}} \coloneqq \mathbb{E}\Big[\partial_d \gamma_0(D, Z)\Big].$$

Here, the linear functional is given by

$$m^{\mathrm{AME}}(W, \gamma) \coloneqq \partial_d \gamma(D, Z).$$

The Riesz representer that satisfies $\mathbb{E}\left[m^{\mathrm{AME}}(W,\gamma)\right] = \mathbb{E}\left[\alpha_0^{\mathrm{AME}}(X)\gamma(X)\right]$ is the (negative) score of the joint density of $X = (D, Z)$ with respect to d:

$$\alpha_0^{\mathrm{AME}}(X) = -\partial_d \log f_0(D, Z),$$

where $f_0(X)$ is the joint probability density of X.

The Neyman orthogonal score is given as

$$\psi^{\mathrm{AME}}(W;\eta,\theta) \coloneqq m^{\mathrm{AME}}(W,\gamma) + \alpha^{\mathrm{AME}}(X)\big(Y - \gamma(X)\big) - \theta.$$

Plugging in estimates of $\eta_0^{\mathrm{AME}} \coloneqq \left(\gamma_0, \alpha_0^{\mathrm{AME}}\right)$ and solving $\frac{1}{n}\sum_{i=1}^n \psi^{\mathrm{AME}}\left(W_i;\widehat{\eta},\theta\right) = 0$ for θ, we can obtain an AME estimator as

$$\widehat{\theta}^{\mathrm{AME}} \coloneqq \frac{1}{n}\sum_{i=1}^{n}\Big(\widehat{\alpha}^{\mathrm{AME}}(X_i)\left(Y_i - \widehat{\gamma}(X_i)\right) + m^{\mathrm{AME}}\left(X_i,\widehat{\gamma}\right)\Big).$$

APE estimation. We consider the average effect of a counterfactual shift in the distribution of the regressors from a known P_{-1} to another P_1, when γ_0 is invariant to the distribution of X. We refer to this average effect as the APE and define it as

$$\theta_0^{\mathrm{APE}} \coloneqq \int \gamma_0(x)\mathrm{d}\mu(x),$$

where $\mu(x) \coloneqq P_1(x) - P_{-1}(x)$. Here, the linear functional is given by

$$m^{\mathrm{APE}}(W,\gamma) \coloneqq \int \gamma(x)\mathrm{d}\mu(x).$$

For simplicity, let us assume that the distributions P_1 and P_{-1} have pdfs, which we denote by $p_1(x)$ and $p_{-1}(x)$. We also assume that there exist common supports among the marginal covariate pdf $p_0(x)$ of the DGP and the pdfs $p_1(x)$ and $p_{-1}(x)$. Then, the Riesz representer is given as

$$\alpha_0^{\mathrm{APE}}(X) \coloneqq \frac{p_1(X) - p_{-1}(X)}{p_0(X)}$$

The Neyman orthogonal score is

$$\psi^{\mathrm{APE}}(W;\eta,\theta) = m^{\mathrm{APE}}(W,\gamma) + \alpha_0^{\mathrm{APE}}(X)\big(Y - \gamma(X)\big) - \theta.$$

Plugging in estimates of $\eta_0^{\mathrm{APE}} \coloneqq \left(\gamma_0, \alpha_0^{\mathrm{APE}}\right)$ and solving $\frac{1}{n}\sum_{i=1}^n \psi^{\mathrm{APE}}\left(W_i;\widehat{\eta},\theta\right) = 0$ for θ, we can obtain an APE estimator as

$$\widehat{\theta}^{\mathrm{APE}} \coloneqq \frac{1}{n}\sum_{i=1}^{n}\Big(\widehat{\alpha}^{\mathrm{APE}}(X_i)\left(Y_i - \widehat{\gamma}(X_i)\right) + m^{\mathrm{APE}}\left(X_i,\widehat{\gamma}\right)\Big).$$

Covariate shift adaptation. Let X denote a source covariate distribution that generates labeled data (X, Y) under P_0, where the pdf of X is $p_0(x)$. Let $\widetilde{X}$ denote a target covariate distribution $P_{X,1}$ whose pdf is $p_1(x)$. Let $\{(X_i, Y_i)\}_{i \in \mathcal{I}_S}$ be i.i.d. from P_0 (source) and $\{\widetilde{X}_j\}_{j \in \mathcal{I}_T}$ be i.i.d. from $P_{X,1}$ (target), independent.

Suppose we train $\gamma_0(x) = \mathbb{E}[Y \mid X = x]$ on a source population with covariates $(X, Y) \sim P_0$, but the target parameter averages γ_0 over a shifted covariate distribution $\widetilde{X} \sim P_{X,1}$:

$$\theta_0^{\mathrm{CS}} \coloneqq \mathbb{E}\big[\gamma_0(\widetilde{X})\big] = \mathbb{E}\big[m^{\mathrm{CS}}(\widetilde{X}, \gamma_0)\big], \qquad m^{\mathrm{CS}}(W, \gamma) \coloneqq \gamma(X).$$

Let $P_{X,0}$ be the marginal distribution of X under P_0, and $p_0(x)$ and $p_1(x)$ be the pdfs. Assume that if $p_0(x) > 0$, then $p_1(x) > 0$ holds. Then, the Riesz representer is

$$\alpha_0^{\mathrm{CS}}(X) = r_0(X) = \frac{p_1(X)}{p_0(X)}.$$

The Neyman orthogonal score is given as

$$\psi^{\mathrm{CS}}(W; \eta, \theta) = \gamma(X) + \alpha^{\mathrm{CS}}(X)\big(Y - \gamma(X)\big) - \theta.$$

Plugging in estimates of $\eta_0^{\mathrm{CS}} \coloneqq \big(\gamma_0, \alpha_0^{\mathrm{CS}}\big)$ and solving $\frac{1}{n}\sum_{i=1}^n \psi^{\mathrm{CS}}\left(W_i; \widehat{\eta}, \theta\right) = 0$ for θ, we can obtain an estimator as

$$\widehat{\theta}^{\mathrm{CS}} \coloneqq \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\alpha}^{\mathrm{CS}}(X_i)\left(Y_i - \widehat{\gamma}(X_i)\right) + m^{\mathrm{CS}}\left(X_i, \widehat{\gamma}\right)\right).$$

Density ratio estimation We also note that the density ratio itself is important in various tasks such as learning with noisy labels (Liu & Tao, 2016), anomaly detection (Smola et al., 2009; Hido et al., 2008; Abe & Sugiyama, 2019; Nam & Sugiyama, 2015; Kato & Teshima, 2021), two-sample testing (Keziou & Leoni-Aubin, 2005; Sugiyama et al., 2011a), and change point detection (Kawahara & Sugiyama, 2009). Learning from positive and unlabeled data (PU learning) can also be interpreted as an application of density ratio estimation (Kato et al., 2019). Thus, density ratio estimation has been studied as an independent task in machine learning (Sugiyama et al., 2012).

Sugiyama et al. (2008) consider covariate shift adaptation using importance weighting estimated by LSIF, which are equivalent to Riesz regression in density ratio estimation (Kato, 2025b). Chernozhukov et al. (2025) and Kato et al. (2024a) investigate efficient estimation of parameters under a covariate shift from some different perspectives.

Notations and assumptions. If there are double parentheses $((\cdot))$, we omit one of them. For example, in ATE estimation, since $X = (D, Z)$, we often encounter $f(X) = f((D, Z))$ for some function f of X. In such a case, we write $f(X) = f(D, Z)$. Let $\mathbb{E}$ be the expectation over P_0 if there are no other explanations. We use the subscript $_0$ to denote parameters under P_0.

Table 1: Correspondence among Bregman divergence losses, density ratio (DR) estimation methods, and Riesz representer (RR) estimation for ATE estimation or general purposes. RR estimation for ATE estimation includes propensity score estimation and covariate balancing weights. In the table, $C \in \mathbb{R}$ denotes a constant that is determined by the problem and the loss function.

$g(\alpha)$	**DR estimation**	**RR estimation**
$(\alpha - C)^2$	LSIF (Kanamori et al., 2009) KuLSIF (Kanamori et al., 2012)	SQ-Riesz regression (**Ours**) Riesz regression (RieszNet and RieszForest) (Chernozhukov et al., 2021, 2022a) Sieve Riesz representer (Chen & Liao, 2015; Chen & Pouzo, 2015) RieszBoost (Lee & Schuler, 2025) KRRR (Singh, 2024) Nearest neighbor matching (Lin et al., 2023) Causal tree/ causal forest (Wager & Athey, 2018)
	Dual solution with a linear link function	
	Kernel mean matching (Gretton et al., 2009)	Stable balancing weights (Zubizarreta, 2015; Bruns-Smith et al., 2025) Approximate Residual Balancing (Athey et al., 2018) Covariate balancing by SVM (Tarr & Imai, 2025)
$\big(\lvert\alpha\rvert - C\big)\log\big(\lvert\alpha\rvert - C\big) - \lvert\alpha\rvert$	UKL divergence minimization (Nguyen et al., 2010)	UKL-Riesz regression (**Ours**) Tailored loss minimization ($\alpha = \beta = -1$) (Zhao, 2019) Calibrated estimation (Tan, 2019)
	Dual solution with a logistic or log link function	
	KLIEP (Sugiyama et al., 2008)	Entropy balancing weights (Hainmueller, 2012)
$(\lvert\alpha\rvert - C)\log\big(\lvert\alpha\rvert - C\big)$ $-(\lvert\alpha\rvert + C)\log(\lvert\alpha\rvert + C)$	BKL divergence minimization (Qin, 1998) TRE (Rhodes et al., 2020)	BKL-Riesz regression (**Ours**) MLE of the propensity score (Standard approach) Tailored loss minimization ($\alpha = \beta = 0$) (Zhao, 2019)
$\frac{\big(\lvert\alpha\rvert - C\big)^{1+\omega} - \big(\lvert\alpha\rvert - C\big)}{\omega} - \big(\lvert\alpha\rvert - C\big)$ for some $\omega \in (0, \infty)$	BP divergence minimization (Sugiyama et al., 2011b)	BP-Riesz regression (**Ours**)
$C\log(1 - \lvert\alpha\rvert)$ $+C\lvert\alpha\rvert\left(\log(\lvert\alpha\rvert) - \log(1 - \lvert\alpha\rvert)\right)$ for $\alpha \in (0, 1)$	PU learning (du Plessis et al., 2015) Nonnegative PU learning (Kiryo et al., 2017)	PU-Riesz regression (**Ours**)
General formulation by Bregman divergence minimization	Density-ratio matching (Sugiyama et al., 2011b) D3RE (Kato & Teshima, 2021)	Generalized Riesz regression (**Ours**)

3 Generalized Riesz Regression

Various methods for estimating the Riesz representer have been proposed. In ATE estimation, a standard approach is to estimate the propensity score $e(z) = P(D = 1 \mid X = z)$ via maximum likelihood estimation (MLE) in a logistic model and then plug the estimate into the Riesz representer $\alpha_0^{\mathrm{ATE}}(X)$. Beyond MLE, covariate balancing approaches have been proposed, where we estimate the propensity score or balancing weights, which implicitly or explicitly correspond to the inverse propensity score, by matching covariate moments. Chernozhukov et al. (2021) proposes Riesz regression as a general method for Riesz representer estimation. In density ratio estimation, related approaches include moment matching (Huang et al., 2007; Gretton et al., 2009), probabilistic classification (Qin, 1998; Cheng & Chu, 2004), density matching (Nguyen et al., 2010), and density ratio fitting (Kanamori et al., 2009). For details on related work, see Section A.

In this study, we generalize existing methods through the lens of Bregman divergence minimization. The Bregman divergence is a general discrepancy measure that includes the squared loss and the KL divergence as special cases. In this section, we propose Riesz representer fitting under a Bregman divergence, which we also refer to as generalized Riesz regression. The term generalized Riesz regression reflects the fact that the choice of loss function is closely connected to the choice of link function from the viewpoint of covariate balancing. We explain this viewpoint in Section 4 and refer to it as automatic covariate balancing. This section provides a general formulation, and we introduce applications of generalized Riesz regression in Section 5.

Remark. *We also refer to our method as direct bias correction term estimation, Bregman Riesz regression, or generalized tailored loss minimization. In an earlier draft, we used the term direct bias correction term estimation because the Riesz representer is almost equivalent to the bias correction term in one step bias correction. Bregman Riesz regression highlights that the method combines the Bregman divergence with Riesz regression. Generalized tailored loss minimization emphasizes that our generalized Riesz regression extends tailored loss minimization (Zhao, 2019) and covers a broader class of methods, including Riesz regression. As discussed in Section 6, standard covariate balancing methods implicitly assume a constant (homogeneous) ATE across x, whereas the covariate balancing property under generalized Riesz regression allows for heterogeneity. From this viewpoint, we call the method generalized covariate balancing.*

3.1 Bregman Divergence

This study fits a Riesz representer model $\alpha\colon \mathcal{X} \to \mathcal{A}$ to the true Riesz representer $\alpha_0(X)$ under a Bregman divergence, where $\mathcal{A} \subset \mathbb{R}$ is the Riesz representer space. Let $g\colon \mathcal{A} \to \mathbb{R}$ be a differentiable and strictly convex function on $\mathcal{A}$. As discussed in Section 4, this function g corresponds to the objective function in covariate balancing.

Given $x \in \mathcal{X}$, the Bregman divergence between the scalar values $\alpha_0(x)$ and $\alpha(x)$ is defined as

$$\mathrm{BD}_g^{\dagger}\big(\alpha_0(x) \mid \alpha(x)\big) \coloneqq g(\alpha_0(x)) - g(\alpha(x)) - \partial g(\alpha(x))\big(\alpha_0(x) - \alpha(x)\big),$$

where ∂g denotes the derivative of g. We then define the average Bregman divergence as

$$\mathrm{BD}_g^\dagger\big(\alpha_0 \mid \alpha\big) \coloneqq \mathbb{E}\Big[g(\alpha_0(X)) - g(\alpha(X)) - \partial g(\alpha(X))\big(\alpha_0(X) - \alpha(X)\big)\Big].$$

We define the population target as

$$\alpha^* = \arg\min_{\alpha \in \mathcal{H}} \mathrm{BD}_g^\dagger\big(\alpha_0 \mid \alpha\big),$$

where $\mathcal{H}$ denotes models for α_0. If $\alpha_0 \in \mathcal{H}$, then $\alpha^* = \alpha_0$ holds.

Although α_0 is unknown, we can define an equivalent optimization problem that does not involve α_0:

$$\alpha^* = \arg\min_{\alpha \in \mathcal{H}} \mathrm{BD}_g\big(\alpha\big),$$

where

$$\mathrm{BD}_g\big(\alpha\big) \coloneqq \mathbb{E}\Big[- g(\alpha(X)) + \partial g(\alpha(X))\alpha(X) - m\big(W, (\partial g) \circ \alpha\big)\Big].$$

Here, we use the linearity of m and the Riesz representation theorem, which imply that

$$\mathbb{E}\Big[\partial g(\alpha(X))\alpha_0(X)\Big] = \mathbb{E}\Big[m\big(W, (\partial g) \circ \alpha\big)\Big].$$

We estimate the Riesz representer α_0 by minimizing an empirical Bregman divergence:

$$\widehat{\alpha} \coloneqq \arg\min_{\alpha \in \mathcal{H}} \widehat{\mathrm{BD}}_g\big(\alpha\big) + \lambda J(\alpha), \tag{1}$$

where $J(\alpha)$ is a regularization function, and

$$\widehat{\mathrm{BD}}_g(\alpha) \coloneqq \frac{1}{n}\sum_{i=1}^n \Big(- g(\alpha(X_i)) + \partial g(\alpha(X_i))\alpha(X_i) - m\big(W_i, (\partial g) \circ \alpha\big)\Big).$$

The choice of the regularization function is important because Riesz representer estimation is known to exhibit a characteristic overfitting phenomenon, often described as train-loss hacking or the density chasm. For details, see Section C.

3.2 Special Cases of the Bregman Divergence

By choosing different g, we obtain various objectives for Riesz representer estimation, including Riesz regression. Specifically, we obtain the following divergences (loss functions) as special cases of the Bregman divergence:

- **Squared distance (squared loss):** $g^{\mathrm{SQ}}(\alpha) \coloneqq (\alpha - C)^2$ for some constant $C \in \mathbb{R}$.
- **Unnormalized KL (UKL) divergence:** $g^{\mathrm{UKL}}(\alpha) \coloneqq \big(|\alpha| - C\big)\log\big(|\alpha| - C\big) - |\alpha|$ for $\alpha \in \mathcal{A}$ and some constant $C < \inf \mathcal{A}$.
- **Binary KL (BKL) divergence:** $g^{\mathrm{BKL}}(\alpha) \coloneqq (|\alpha| - C)\log\big(|\alpha| - C\big) - (|\alpha| + C)\log(|\alpha| + C)$ for $\alpha \in \mathcal{A}$ and some constant $C < \inf \mathcal{A}$.

- **Basu's power (BP) divergence (BP-Riesz):** $g^{\mathrm{BP}}(\alpha) \coloneqq \frac{\left(|\alpha|-C\right)^{1+\omega}-\left(|\alpha|-C\right)}{\omega}-\left(|\alpha|-C\right)$ for some $\omega \in (0,\infty)$, $\alpha \in \mathcal{A}$, and some constant $C < \inf \mathcal{A}$.
- **PU learning loss:** $g^{\mathrm{PU}}(\alpha) \coloneqq \widetilde{C}\log\left(1-|\alpha|\right)+\widetilde{C}|\alpha|\left(\log\left(|\alpha|\right)-\log\left(1-|\alpha|\right)\right)$ for some $\widetilde{C} \in \mathbb{R}$, where α takes values in $(0,1)$.

See also Table 1 for a summary.

We refer to our method as SQ-Riesz when using the squared loss, UKL-Riesz when using the UKL divergence, BKL-Riesz when using the BKL divergence, BP-Riesz when using the BP divergence, and PU-Riesz when using the PU learning loss. We explain these special cases in detail in the following subsections.

3.3 SQ-Riesz Regression

Let $C \in \mathbb{R}$ be a constant. We consider the following convex function:

$$g^{\mathrm{SQ}}(\alpha) = (\alpha - C)^2.$$

This choice of convex function is motivated by the squared loss. The choice of C depends on the researcher. We propose choosing C so that the automatic covariate balancing property holds, see Section 4. The derivative of $g^{\mathrm{SQ}}(\alpha)$ with respect to α is given as

$$\partial g^{\mathrm{SQ}}(\alpha) = 2(\alpha - C).$$

Under this choice of g, the Bregman divergence objective is given as

$$\mathrm{BD}_g\big(\alpha\big) \coloneqq \mathbb{E}\Big[\alpha(X)^2 - 2m\big(W, \big(\alpha(\cdot) - C\big)\big)\Big].$$

Then, the estimation problem can be written as

$$\widehat{\alpha} \coloneqq \underset{\alpha\in\mathcal{H}}{\arg\min}\, \widehat{\mathrm{BD}}_{g^{\mathrm{SQ}}}\big(\alpha\big) + \lambda J(\alpha), \tag{2}$$

where

$$\widehat{\mathrm{BD}}_{g^{\mathrm{SQ}}}\big(\alpha\big) \coloneqq \frac{1}{n}\sum_{i=1}^{n}\left(\alpha(X_i)^2 - 2m\big(W_i, \big(\alpha(\cdot)\big)\big)\right).$$

Here, for simplicity, we drop constant terms that are irrelevant for the optimization and use the linearity of m for $2(\alpha(\cdot) - C)$[1]. This estimation method corresponds to Riesz regression in debiased machine learning (Chernozhukov et al., 2021) and least-squares importance fitting (LSIF) in density ratio estimation (Kanamori et al., 2009). Moreover, if we define $\mathcal{H}$ appropriately, we can recover nearest neighbor matching, as pointed out in Kato (2025a), which extends the argument in Lin et al. (2023).

[1]The original Bregman divergence objective using $g(\alpha) = (\alpha - C)^2$ in (1) is given as

$$\begin{aligned}\widehat{\mathrm{BD}}_{g^{\mathrm{SQ}}}\big(\alpha\big) &= \mathbb{E}\left[-(\alpha(X) - C)^2 + 2(\alpha(X) - C)\alpha(X) - 2m\big(W, (\alpha(\cdot) - C)\big)\right] \\ &= \mathbb{E}\left[\alpha(X)^2 - C^2 - 2m\big(W, \big(\alpha(\cdot) - C\big)\big)\right].\end{aligned}$$

3.4 UKL-Riesz Regression

Next, we consider a KL-divergence-motivated convex function. Let $C < \inf_x |\alpha(x)|$ be a constant. We define

$$g^{\mathrm{UKL}}(\alpha) = (|\alpha| - C)\log(|\alpha| - C) - |\alpha|.$$

The choice of C depends on the researcher. We propose choosing C so that the automatic covariate balancing property holds, see Section 4. The derivative of $g^{\mathrm{UKL}}(\alpha)$ with respect to α is given as

$$\partial g^{\mathrm{UKL}}(\alpha) = \mathrm{sign}(\alpha)\log(|\alpha| - C).$$

Under this choice of g, the Bregman divergence objective is given as follows[2]:

$$\mathrm{BD}_{g^{\mathrm{UKL}}}(\alpha) \coloneqq \mathbb{E}\Big[C\log(|\alpha(X)| - C) + |\alpha(X)| - m\Big(W, \mathrm{sign}\left(\alpha(\cdot)\right)\log(|\alpha(\cdot)| - C)\Big)\Big].$$

We estimate α_0 by minimizing the empirical objective:

$$\widehat{\alpha} \coloneqq \arg\min_{\alpha\in\mathcal{H}} \widehat{\mathrm{BD}}_{g^{\mathrm{UKL}}}(\alpha) + \lambda J(\alpha),$$

where

$$\widehat{\mathrm{BD}}_{g^{\mathrm{UKL}}}(\alpha) = \frac{1}{n}\sum_{i=1}^{n}\Big(C\log(|\alpha(X_i)| - C) + |\alpha(X_i)| - m\Big(W_i, \mathrm{sign}\left(\alpha(\cdot)\right)\log(|\alpha(\cdot)| - C)\Big)\Big).$$

In the next subsection, we also introduce the BKL divergence as a KL-divergence-motivated divergence, but the UKL divergence more closely corresponds to the standard KL divergence. The equivalent formulation is known as KLIEP in density ratio estimation. Note that KLIEP is a constrained formulation that is equivalent to UKL divergence minimization, and this equivalence is also known as Silverman's trick (Silverman, 1978; Kato et al., 2023). This constrained formulation can also be interpreted as a dual formulation. In ATE estimation, tailored loss minimization corresponds to UKL minimization, whose dual yields entropy balancing weights (Hainmueller, 2012).

3.5 BKL-Riesz Regression

We introduce the BKL divergence and BKL-Riesz, which are motivated by the KL divergence and logistic regression. Let $C < \inf_x |\alpha(x)|$ be a constant. We define

$$g^{\mathrm{BKL}}(\alpha) \coloneqq (|\alpha| - C)\log\big(|\alpha| - C\big) - (|\alpha| + C)\log(|\alpha| + C).$$

[2]This Bregman divergence objective is derived as follows:

$$\widehat{\mathrm{BD}}_{g^{\mathrm{UKL}}}(\alpha) = \mathbb{E}\Big[-(|\alpha(X)| - C))\log(|\alpha(X)| - C) + |\alpha(X)| + \mathrm{sign}\left(\alpha(X)\right)\alpha(X)\log(|\alpha(X)| - C) - m\Big(W, \mathrm{sign}\left(\alpha(\cdot)\right)\log(|\alpha(\cdot)| - C)\Big)\Big].$$

The choice of C depends on the researcher. We propose choosing C so that the automatic covariate balancing property holds, see Section 4.

Under this choice of g, the Bregman divergence objective is given as follows[3]:

$$\mathrm{BD}_{g^{\mathrm{BKL}}}(\alpha) \coloneqq \mathbb{E}\left[C\log\left(\frac{|\alpha(X)|-C}{|\alpha(X)|+C}\right) - m\left(W, \mathrm{sign}\left(\alpha(\cdot)\right)\log\left(\frac{|\alpha(\cdot)|-C}{|\alpha(\cdot)|+C}\right)\right)\right].$$

We estimate α_0 by minimizing the empirical objective:

$$\widehat{\alpha} \coloneqq \arg\min_{\alpha\in\mathcal{H}} \widehat{\mathrm{BD}}_{g^{\mathrm{BKL}}}(\alpha) + \lambda J(\alpha),$$

where

$$\widehat{\mathrm{BD}}_{g^{\mathrm{BKL}}}(\alpha) = \frac{1}{n}\sum_{i=1}^{n}\left(C\log\left(\frac{|\alpha(X_i)|-C}{|\alpha(X_i)|+C}\right) - m\left(W_i, \mathrm{sign}\left(\alpha(\cdot)\right)\log\left(\frac{|\alpha(\cdot)|-C}{|\alpha(\cdot)|+C}\right)\right)\right).$$

In ATE estimation, this formulation corresponds to MLE for a logistic model of the propensity score. In density ratio estimation, this formulation corresponds to a logistic regression approach, where we classify two datasets using a logistic model and then take the ratio to obtain a density ratio estimator. For details, see Section 5.

3.6 BP-Riesz Regression

Basu's power (BP) divergence bridges the squared loss and KL divergence (Basu et al., 1998). Let $C < \inf_x |\alpha(x)|$ be a constant. Based on the BP divergence, we introduce the following function:

$$g^{\mathrm{BP}}(\alpha) \coloneqq \frac{\left(|\alpha|-C\right)^{1+\omega} - \left(|\alpha|-C\right)}{\omega} - |\alpha|.$$

The derivative is given as

$$\partial g^{\mathrm{BP}}(\alpha) = \left(1+\frac{1}{\omega}\right)\mathrm{sign}(\alpha)\Big(\left(|\alpha|-C\right)^{\omega} - 1\Big).$$

Using this function in the Bregman divergence yields a BP-motivated loss and the corresponding objective for BP-Riesz regression. The choice of C depends on the researcher. We propose choosing C so that the automatic covariate balancing property holds, see Section 4.

[3]This Bregman divergence objective is derived as follows:

$$\begin{aligned}
&\mathrm{BD}_{g^{\mathrm{BKL}}}(\alpha)\\
&= \mathbb{E}\Big[-\left(|\alpha(X)|-C\right)\log\left(|\alpha(X)|-C\right) - \left(|\alpha(X)|+C\right)\log(|\alpha(X)|+C) + \mathrm{sign}\left(\alpha(X)\right)\alpha(X)\log\left(\frac{|\alpha(X)|-C}{|\alpha(X)|+C}\right)\\
&\qquad\qquad - m\left(W, \mathrm{sign}\left(\alpha(\cdot)\right)\log\left(\frac{|\alpha(\cdot)|-C}{|\alpha(\cdot)|+C}\right)\right)\Big].
\end{aligned}$$

Under this choice of g, the Bregman divergence objective is given as follows[4]:

$$\mathrm{BD}_{g^{\mathrm{BP}}}(\alpha) := \mathbb{E}\Bigg[\frac{C\big(\big(|\alpha(X)|-C\big)^{\omega}-1\big)}{\omega} + C|\alpha(X)|\big(|\alpha(X)|-C\big)^{\omega} - m\left(W, \left(1+\frac{1}{\omega}\right)\mathrm{sign}(\alpha(\cdot))\Big(\big(|\alpha(\cdot)|-C\big)^{\omega}-1\Big)\right)\Bigg].$$

We estimate α_0 by minimizing the empirical objective:

$$\widehat{\alpha} := \arg\min_{\alpha\in\mathcal{H}} \widehat{\mathrm{BD}}_{g^{\mathrm{BP}}}(\alpha) + \lambda J(\alpha),$$

where

$$\widehat{\mathrm{BD}}_{g^{\mathrm{BP}}}(\alpha) := \frac{1}{n}\sum_{i=1}^{n}\left(\frac{C\big(\big(|\alpha(X_i)|-C\big)^{\omega}-1\big)}{\omega} + C|\alpha(X_i)|\big(|\alpha(X_i)|-C\big)^{\omega} - m\left(W_i, \left(1+\frac{1}{\omega}\right)\mathrm{sign}(\alpha(\cdot))\Big(\big(|\alpha(\cdot)|-C\big)^{\omega}-1\Big)\right)\right).$$

Basu's power divergence bridges the squared loss and the (U)KL divergence. When $\omega = 1$, BP-Riesz regression reduces to SQ-Riesz regression, while when $\omega \to 0$, BP-Riesz regression reduces to UKL-Riesz regression. This follows because

$$\lim_{\omega\to 0}\frac{\big(|\alpha|-C\big)^{\omega}-1}{\omega} = \log\big(|\alpha|-C\big).$$

BP-Riesz regression plays an important role in robust estimation of the Riesz representer. UKL-Riesz regression implicitly assumes exponential or sigmoid models for the Riesz representer. If the model is misspecified, the estimation accuracy can deteriorate. As Sugiyama et al. (2012) notes, SQ-Riesz regression is more robust to outliers, while UKL-Riesz regression can perform well under correct specification. BP-Riesz regression provides an intermediate objective between these two extremes. In addition, BP-Riesz regression is useful for understanding the automatic covariate balancing property.

3.7 PU-Riesz Regression

We introduce PU learning loss and PU-Riesz, which are motivated by PU learning. Let $C < \inf_x |\alpha(x)|$ be some constant. We define g^{PU} as

$$g^{\mathrm{PU}}(\alpha) := \widetilde{C}\log(1-|\alpha|) + \widetilde{C}|\alpha|\Big(\log(|\alpha|) - \log(1-|\alpha|)\Big)$$

[4]This Bregman divergence objective is derived from

$$\mathrm{BD}_{g^{\mathrm{BP}}}(\alpha) := \mathbb{E}\Bigg[-\frac{\big(|\alpha(X)|-C\big)^{1+\omega}-\big(|\alpha(X)|-C\big)}{\omega} + |\alpha(X)| + (1+1/\omega)|\alpha(X)|\Big(\big(|\alpha(X)|-C\big)^{\omega}-1\Big) - m\left(W, (1+1/\omega)\,\mathrm{sign}(\alpha)\Big(\big(|\alpha(\cdot)|-C\big)^{\omega}-1\Big)\right)\Bigg].$$

for some $\widetilde{C} \in \mathbb{R}$, and we restrict α to take values in $(0, 1)$. The choice of $\widetilde{C}$ depends on the researcher. It corresponds to the class prior in PU learning and plays a role that differs from the parameter C in the other loss functions. The derivative of $g^{\mathrm{PU}}(\alpha)$ with respect to α is given as

$$\begin{aligned}\partial g^{\mathrm{PU}}(\alpha) &= -\frac{\widetilde{C}\,\mathrm{sign}(\alpha)}{1-|\alpha|} + \widetilde{C}\,\mathrm{sign}(\alpha)\left(\log\left(|\alpha|\right) - \log\left(1-|\alpha|\right) + \frac{1}{1-|\alpha|}\right) \\ &= \widetilde{C}\,\mathrm{sign}(\alpha)\Big(\log\left(|\alpha|\right) - \log\left(1-|\alpha|\right)\Big).\end{aligned}$$

Under this choice of g, the Bregman divergence objective is given as follows:

$$\mathrm{BD}_{g^{\mathrm{PU}}}\left(\alpha\right) \coloneqq \mathbb{E}\left[-\widetilde{C}\log\left(1-|\alpha(X)|\right) - m\left(W, \widetilde{C}\,\mathrm{sign}(\alpha)\Big(\log\left(|\alpha(\cdot)|\right) - \log\left(1-|\alpha(\cdot)|\right)\Big)\right)\right].$$

Then, we estimate α_0 by minimizing the empirical objective:

$$\widehat{\alpha} \coloneqq \arg\min_{\alpha\in\mathcal{H}} \widehat{\mathrm{BD}}_{g^{\mathrm{PU}}}\left(\alpha\right) + \lambda J(\alpha),$$

where

$$\widehat{\mathrm{BD}}_{g^{\mathrm{PU}}}\left(\alpha\right) = \frac{1}{n}\sum_{i=1}^{n}\left(-\widetilde{C}\log\left(1-|\alpha(X_i)|\right) - m\left(W_i, \widetilde{C}\,\mathrm{sign}(\alpha)\Big(\log\left(|\alpha(\cdot)|\right) - \log\left(1-|\alpha(\cdot)|\right)\Big)\right)\right).$$

PU learning is a classical problem. For example, Lancaster & Imbens (1996) studies this problem under a stratified sampling scheme (Wooldridge, 2001). du Plessis et al. (2015) rediscovers this formulation and calls it unbiased PU learning. Kato et al. (2019) points out the relationship between PU learning and density ratio estimation, and Kato & Teshima (2021) shows that PU learning is a special case of density ratio model fitting under a Bregman divergence. Our results further generalize these results. Note that PU learning in these settings and our setting is called case-control PU learning. There is also another formulation called censoring PU learning (Elkan & Noto, 2008). Kato et al. (2025) considers ATE estimation in a PU learning setup and applies our method in their study.

4 Automatic Covariate Balancing

Under specific choices of the Riesz representer model and the Bregman divergence, we can guarantee an automatic covariate balancing property. The key tool is the duality between Bregman divergence minimization and covariate balancing methods. This automatic covariate balancing property yields the automatic Neyman orthogonalization property discussed in Section 6, which in turn automatically guarantees Neyman orthogonality for IPW-type estimators. We note that the covariate balancing property does not hold if we use cross fitting.

4.1 Generalized Linear Models

Throughout this section, we consider a Riesz representer model of the form

$$\alpha(X) = \zeta^{-1}\Big(X, \boldsymbol{\phi}(X)^\top \boldsymbol{\beta}\Big),$$

where ζ^{-1} is the inverse of a link function and $\boldsymbol{\phi}\colon \mathcal{X} \to \mathbb{R}^p$ is a basis function. A link function ζ connects the Riesz representer α to a linear, or linear in parameters, index,

$$\boldsymbol{\phi}(X)^\top \boldsymbol{\beta}.$$

This introduction of a linear index is motivated by generalized linear models. While standard generalized linear models assume a link of the form $\alpha(X) = \zeta^{-1}\big(\boldsymbol{\phi}(X)^\top \boldsymbol{\beta}\big)$, we allow the transformation to depend on X as $\alpha(X) = \zeta^{-1}\big(X, \boldsymbol{\phi}(X)^\top \boldsymbol{\beta}\big)$, as described below. Given X, we define ζ so that

$$\zeta\big(X, \alpha(X)\big) = \boldsymbol{\phi}(X)^\top \boldsymbol{\beta}.$$

Examples. For example, we can approximate the Riesz representer by

$$\alpha_{\boldsymbol{\beta}}(X) \coloneqq \boldsymbol{\phi}(X)^\top \boldsymbol{\beta},$$

which corresponds to using a linear link function for ζ^{-1}. This linear specification can be applied in many settings, including ATE estimation and density ratio estimation.

We can improve estimation accuracy by incorporating additional modeling assumptions. For example, in ATE estimation, we can approximate the propensity score $e_0(Z) = P(D = 1 \mid Z)$ by a logistic model,

$$e_{\boldsymbol{\beta}}(Z) \coloneqq \frac{1}{1 + \exp\Big(-\boldsymbol{\phi}(Z)^\top \boldsymbol{\beta}\Big)},$$

where $\boldsymbol{\phi}\colon \mathcal{Z} \to \mathbb{R}^p$ is a basis function, and $\boldsymbol{\beta}$ is the corresponding parameter. Note that in this case, we consider a basis function that receives Z not X, or we can interpret that $\boldsymbol{\phi}(D, Z)$ only depends on Z and is independent of D. Plugging this propensity score model into the Riesz representer for ATE, we can approximate the Riesz representer by

$$\alpha^{\mathrm{ATE}}_{\boldsymbol{\beta}}(X) \coloneqq \frac{D}{e_{\boldsymbol{\beta}}(Z)} - \frac{1 - D}{1 - e_{\boldsymbol{\beta}}(Z)}.$$

In such cases, we define ζ^{-1} so that

$$\begin{aligned}
\alpha_{\boldsymbol{\beta}}(X) &= \zeta^{-1}\left(X, \boldsymbol{\phi}(Z)^\top \boldsymbol{\beta}\right) \\
&= \frac{D}{e_{\boldsymbol{\beta}}(Z)} - \frac{1 - D}{1 - e_{\boldsymbol{\beta}}(Z)} \\
&= D\Big(1 + \exp\Big(-\boldsymbol{\phi}(Z)^\top \boldsymbol{\beta}\Big)\Big) - (1 - D)\Big(1 + \exp\Big(\boldsymbol{\phi}(Z)^\top \boldsymbol{\beta}\Big)\Big). \qquad (3)
\end{aligned}$$

Note that we can also model the Riesz representer as

$$\alpha_{\boldsymbol{\beta}}(X) = \zeta^{-1}\left(X, \boldsymbol{\phi}(X)^\top \boldsymbol{\beta}\right)$$

$$= D\Big(1+\exp\Big(-\boldsymbol{\phi}(Z)^\top\boldsymbol{\beta}\Big)\Big) - (1-D)\Big(1+\exp\Big(\boldsymbol{\phi}(Z)^\top\boldsymbol{\beta}\Big)\Big), \tag{4}$$

with including D in the basis function. The choice of basis functions depend on the heterogeneity of $\gamma_0(X)$. As we discuss in Section 6, if $\gamma_0(x)$ is constant for all x, (3) may be more appropriate. In contrast, if $\gamma_0(x)$ varies across x, (4) may be more appropriate.

Similarly, in covariate shift adaptation, we can model the density ratio as

$$\alpha^{\mathrm{CS}}_{\boldsymbol{\beta}}(X) = \exp\Big(-\boldsymbol{\phi}(X)^\top\boldsymbol{\beta}\Big).$$

Here, the link function is given by

$$\zeta\left(X, \alpha^{\mathrm{CS}}_{\boldsymbol{\beta}}(X)\right) = -\boldsymbol{\phi}(X)^\top\boldsymbol{\beta}.$$

4.2 Automatic Covariate Balancing

This section presents a covariate balancing property for the basis function $\boldsymbol{\phi}(X)$ when the Riesz representer model has the form $\alpha(X) = \zeta^{-1}\big(X, \boldsymbol{\phi}(X)^\top\boldsymbol{\beta}\big)$. Related properties have been reported in Ben-Michael et al. (2021). More specifically, for Riesz regression, Bruns-Smith & Feller (2022) and Bruns-Smith et al. (2025) discuss the duality between Riesz regression and stable balancing weights (Zubizarreta, 2015). Zhao (2019) also reports the duality between tailored loss minimization and entropy balancing weights. Tarr & Imai (2025) investigates such a duality in covariate balancing method using the support vector machine (SVM). Thus, the duality has garnered attention in covariate balancing (Tan, 2019; Orihara et al., 2025). In density ratio estimation, such a duality is also well known, as discussed in Sugiyama et al. (2007) and Sugiyama et al. (2011b). In the density ratio estimation literature, the resulting balancing property is typically interpreted as moment matching. This study generalizes these results and provides new findings on the balancing property. For the proof, see Appendix B.

Theorem 4.1 (Automatic covariate balancing)**.** *Assume that there exists a function* $\widetilde{g}\colon \mathcal{X}\times\mathbb{R}\to\mathbb{R}$ *such that*

$$\partial g\big(\alpha_{\boldsymbol{\beta}}(X_i)\big) = \sum_{j=1}^{p}\beta_j\widetilde{g}\left(X_i, \phi_j(X_i)\right);$$

that is, $\partial g\big(\alpha_{\boldsymbol{\beta}}(X_i)\big)$ *is linear in* $\widetilde{g}\left(X_i, \phi_1(X_i)\right), \ldots, \widetilde{g}\left(X_i, \phi_p(X_i)\right)$*. Consider a Riesz representer estimator* $\widehat{\alpha} = \alpha_{\widehat{\boldsymbol{\beta}}}$ *trained by generalized Riesz regression (empirical risk minimization) with* ℓ_a*-penalty as*

$$\widehat{\boldsymbol{\beta}} \coloneqq \operatorname*{arg\,min}_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{\frac{1}{n}\sum_{i=1}^{n}\Big(-g(\alpha_{\boldsymbol{\beta}}(X_i)) + \alpha_{\boldsymbol{\beta}}(X_i)\partial g\big(\alpha_{\boldsymbol{\beta}}(X_i)\big) - m\big(W_i, (\partial g)\circ\alpha_{\boldsymbol{\beta}}\big)\Big) + \frac{1}{a}\lambda\|\boldsymbol{\beta}\|_a^a\right\}.$$

In this case $\widehat{\alpha}$ *satisfies*

$$\left|\frac{1}{n}\sum_{i=1}^{n}\Big(\widehat{\alpha}(X_i)\widetilde{g}\big(X_i, \phi_j(X_i)\big) - m(W_i, \widetilde{g}\left(X_i, \phi_j(X_i)\right))\Big)\right| \le \lambda\left|\widehat{\beta}_j\right|^{a-1}.$$

Generalized Riesz regression returns a Riesz representer model that minimizes the given loss among models satisfying the above inequality.

In addition, if $\partial g\big(\alpha_{\boldsymbol{\beta}}(X_i)\big) = \sum_{j=1}^p \beta_j \phi_j(X_i)$ holds, the following result holds.

Corollary 4.2. *Assume the conditions in Theorem 4.1. Then, if* $\partial g\big(\alpha_{\boldsymbol{\beta}}(X_i)\big) = \sum_{j=1}^p \beta_j \phi_j(X_i)$ *holds, then the inequality can be written as*

$$\left| \frac{1}{n} \sum_{i=1}^n \Big(\widehat{\alpha}(X_i)\phi_j(X_i) - m(W_i, \phi_j) \Big) \right| \leq \lambda \left| \widehat{\beta}_j \right|^{a-1}.$$

These results correspond to covariate balancing in ATE estimation and moment matching in density ratio estimation.

Examples. For simplicity, we consider ℓ_1-penalty in generalized Riesz regression. Then, in ATE estimation, if we use a linear link and SQ-Riesz regression, we have

$$\left| \frac{1}{n} \sum_{i=1}^n \widehat{\alpha}(X_i)\phi_j(D_i, Z_i) - \frac{1}{n} \sum_{i=1}^n \Big(\phi_j(1, Z_i) - \phi_j(0, Z_i) \Big) \right| < \lambda \quad (\text{for } j = 1, 2, \ldots, p).$$

If we use a logistic link and UKL-Riesz regression, we have

$$\left| \frac{1}{n} \sum_{i=1}^n \Big(\mathbb{1}[D_i = 1]\widehat{\alpha}(X_i)\phi_j(Z_i) - \mathbb{1}[D_i = 0]\widehat{\alpha}(X_i)\phi_j(Z_i) \Big) \right| \leq \lambda \quad (\text{for } j = 1, 2, \ldots, p).$$

In density ratio estimation, if we use a linear link and SQ-Riesz regression, we have

$$\left| \frac{1}{n} \sum_{i=1}^n \widehat{\alpha}(X_i)\phi_j(X_i) - \frac{1}{n} \sum_{i=1}^n \phi_j(X_i) \right| \leq \lambda \quad (\text{for } j = 1, 2, \ldots, p).$$

If we use a logistic link and UKL-Riesz regression, we have

$$\left| \frac{1}{n} \sum_{i=1}^n \widehat{\alpha}(X_i)\phi_j(X_i) - 1 \right| \leq \lambda \quad (\text{for } j = 1, 2, \ldots, p).$$

We explain the details of the automatic covariate balancing property in the following subsections.

4.3 Choice of Loss and Link Functions

For the automatic covariate balancing property, the linearity assumption

$$\partial g\big(\alpha_{\boldsymbol{\beta}}(X_i)\big) = \sum_{j=1}^p \beta_j \phi_j(X_i)$$

or $\partial g\big(\alpha_{\boldsymbol{\beta}}(X_i)\big) = \sum_{j=1}^p \beta_j \widetilde{g}\left(X_i, \phi_j(X_i)\right)$ are crucial. These assumptions depend on the choice of the function g in the Bregman divergence, which corresponds to the loss function, and on the choice of the link function.

For example, for SQ-Riesz regression, UKL-Riesz regression, and BP-Riesz regression, the following choice of Riesz representer models (link functions) guarantee the linearity assumption:

- **Squared distance (SQ-Riesz):** We use the identity link function,

$$\begin{aligned}\alpha_{\boldsymbol{\beta}}(X) &= \zeta^{-1}\Big(X, \boldsymbol{\phi}(X)^\top \boldsymbol{\beta}\Big) \\ &= \boldsymbol{\phi}(X)^\top \boldsymbol{\beta}/2 + C.\end{aligned}$$

- **UKL divergence loss (UKL-Riesz):** We use a log or logistic link function,

$$\begin{aligned}\alpha_{\boldsymbol{\beta}}(X) &= \zeta^{-1}\Big(X, \boldsymbol{\phi}(X)^\top \boldsymbol{\beta}\Big) \\ &= \xi(X)\Big(C + \exp\big(\boldsymbol{\phi}(X)^\top \boldsymbol{\beta}\big)\Big) - (1-\xi(X))\Big(C + \exp\big(-\boldsymbol{\phi}(X)^\top \boldsymbol{\beta}\big)\Big).\end{aligned}$$

- **BP divergence loss (BP-Riesz):**

$$\begin{aligned}\alpha_{\boldsymbol{\beta}}(X) &= \zeta^{-1}\Big(X, \boldsymbol{\phi}(X)^\top \boldsymbol{\beta}\Big) \\ &= \xi(X)\left(C + \left(1 + \frac{\boldsymbol{\phi}(X)^\top \boldsymbol{\beta}}{k}\right)^{1/\omega}\right) - (1-\xi(X))\left(C + \left(1 - \frac{\boldsymbol{\phi}(X)^\top \boldsymbol{\beta}}{k}\right)^{1/\omega}\right),\end{aligned}$$

  for $k \coloneqq 1 + 1/\omega$.

Here, $\xi\colon \mathcal{X} \to \{1, 0\}$ is chosen by the researcher. Note that these are not the only possible choices, and we can still choose link functions that satisfy the linearity assumption. For example, we can multiply $\alpha(X) = \boldsymbol{\phi}(X)^\top \boldsymbol{\beta} + 1$ by a function $\kappa\colon \mathcal{X} \to \mathbb{R}$. Such extensions and other cases are straightforward, so we omit them.

For BKL-Riesz regression, we can also obtain automatic covariate balancing under specific link functions. However, since the link function is somewhat complicated and is not commonly used in practice, this approach is not practical. Therefore, we only remark on this point in Remark 4.3. Note that this result implies that, in ATE estimation, the standard MLE for logistic regression models (sigmoid function + BKL-Riesz regression) does not yield covariate balancing. Zhao (2019)'s arguments about the choice of tailored loss function parameters are related to this point. If the parameter of interest is the Optimally Weighted ATE (OWATE), the MLE of logistic regression (sigmoid function + BKL-Riesz regression = tailored loss minimization with $\alpha = \beta = 0$) also attains covariate balancing. See Section 7 for the arguments.

Justification. We justify the above choices of the link functions by showing that under them, the automatic covariate balancing property holds. The derivative of g is given as

- **Squared distance (SQ-Riesz):** $\partial g^{\mathrm{SQ}}(\alpha) = 2(\alpha - C)$.
- **UKL divergence loss (UKL-Riesz):** $\partial g^{\mathrm{UKL}}(\alpha) = \operatorname{sign}(\alpha) \log\big(|\alpha| - C\big)$ for $\alpha < -C$ and $\alpha > C$. We introduce the following branchwise maps:

$$\begin{aligned}\partial g_+^{\mathrm{UKL}}(\alpha) &\coloneqq \log(\alpha - C), \qquad \alpha \in (C, \infty), \\ \partial g_-^{\mathrm{UKL}}(\alpha) &\coloneqq -\log(-\alpha - C), \qquad \alpha \in (-\infty, -C),\end{aligned}$$

  so that $\partial g^{\mathrm{UKL}}(\alpha) = \partial g_+^{\mathrm{UKL}}(\alpha)$ for $\alpha > C$ and $\partial g^{\mathrm{UKL}}(\alpha) = \partial g_-^{\mathrm{UKL}}(\alpha)$ for $\alpha < -C$.

- **BP divergence loss (BP-Riesz):** $\partial g^{\mathrm{BP}}(\alpha) = \operatorname{sign}(\alpha)\Big(1+\frac{1}{\omega}\Big)\Big(\big(|\alpha|-C\big)^{\omega}-1\Big)$, $\operatorname{dom}\big(\partial g^{\mathrm{BP}}\big) = \{\alpha\in\mathbb{R} : |\alpha|\geq C\}$ for $\omega>0$, $\alpha<-C$, and $\alpha>C$. We introduce the following branchwise maps:

$$\partial g_{+}^{\mathrm{BP}}(\alpha) \coloneqq \Big(1+\frac{1}{\omega}\Big)\Big((\alpha-C)^{\omega}-1\Big), \qquad \alpha\in[C,\infty),$$
$$\partial g_{-}^{\mathrm{BP}}(\alpha) \coloneqq -\Big(1+\frac{1}{\omega}\Big)\Big((-\alpha-C)^{\omega}-1\Big), \qquad \alpha\in(-\infty,-C],$$

  so that $\partial g^{\mathrm{BP}}(\alpha) = \partial g_{+}^{\mathrm{BP}}(\alpha)$ for $\alpha\geq C$ and $\partial g^{\mathrm{BP}}(\alpha) = \partial g_{-}^{\mathrm{BP}}(\alpha)$ for $\alpha\leq -C$.

We introduce $\partial g_{\pm}^{\mathrm{UKL}}$ and $\partial g_{\pm}^{\mathrm{BP}}$ (and related functions) to make the inverse mapping one-to-one on each branch.

The inverse of these derivatives gives intuition for how to choose loss and link functions. The inverse functions are

- **Squared distance (SQ-Riesz):**

$$\big(\partial g^{\mathrm{SQ}}\big)^{-1}(v) \coloneqq (v+C)/2 \qquad (v\in\mathbb{R}).$$

- **UKL divergence loss (UKL-Riesz):**

$$\big(\partial g_{+}^{\mathrm{UKL}}\big)^{-1}(v) = C+\exp(v), \qquad (v\in\mathbb{R}),$$
$$\big(\partial g_{-}^{\mathrm{UKL}}\big)^{-1}(v) = -C-\exp(-v), \qquad (v\in\mathbb{R}).$$

- **BP divergence loss (BP-Riesz):**

$$\big(\partial g_{+}^{\mathrm{BP}}\big)^{-1}(v) = C+\Big(1+\frac{v}{k}\Big)^{1/\omega}, \qquad (v\geq -k),$$
$$\big(\partial g_{-}^{\mathrm{BP}}\big)^{-1}(v) = -C-\Big(1-\frac{v}{k}\Big)^{1/\omega}, \qquad (v\leq k),$$

  for $k\coloneqq 1+1/\omega$.

These inverse functions suggest the above link functions.

Remark (Automatic covariate balancing in BKL-Riesz regression)**.** *Consider the BKL generator*

$$g^{\mathrm{BKL}}(\alpha) \coloneqq (|\alpha|-C)\log(|\alpha|-C)-(|\alpha|+C)\log(|\alpha|+C), \qquad C>0,$$

with domain $\{|\alpha|>C\}$*. Its derivative is*

$$\partial g^{\mathrm{BKL}}(\alpha) = \operatorname{sign}(\alpha)\log\left(\frac{|\alpha|-C}{|\alpha|+C}\right).$$

Then a corresponding link function is given by

$$\zeta^{-1}(X,\boldsymbol{\phi}(X)^{\top}\boldsymbol{\beta}) = C\left(\xi(X)\frac{1+\exp(\boldsymbol{\phi}(X)^{\top}\boldsymbol{\beta})}{1-\exp\left(\boldsymbol{\phi}(X)^{\top}\boldsymbol{\beta}\right)}-(1-\xi(X))\frac{1+\exp(\boldsymbol{\phi}(X)^{\top}\boldsymbol{\beta})}{\exp\left(\boldsymbol{\phi}(X)^{\top}\boldsymbol{\beta}\right)-1}\right).$$

Consider the generalized Riesz regression problem for $\boldsymbol{\beta}$ (as defined in Section 3), and let $\widehat{\boldsymbol{\beta}}$ be any solution. If we use ℓ_1-penalty, the KKT conditions yield, for each $j = 1, \ldots, p$,

$$\left| \frac{1}{n} \sum_{i=1}^{n} \Big(\widehat{\alpha}(X_i)\phi_j(X_i) - m(W_i, \phi_j) \Big) \right| \leq \lambda.$$

In particular, when $\lambda = 0$ the fitted BKL-Riesz representer $\widehat{\alpha} = \alpha_{\widehat{\boldsymbol{\beta}}}$ satisfies the corresponding exact sample balancing conditions.

4.4 SQ-Riesz regression with a Linear Link Function

We first introduce the combination of the squared loss and a linear model. Consider the linear model

$$\alpha_{\boldsymbol{\beta}}(X) = \boldsymbol{\phi}(X)^\top \boldsymbol{\beta},$$

where $\boldsymbol{\phi} \colon \mathcal{X} \to \mathbb{R}^p$ is a basis function.

For this model, using the squared distance (SQ-Riesz regression, or standard Riesz regression) yields automatic covariate balancing. Specifically, under a linear model, if we use ℓ_1-penalty in SQ-Riesz regression, the dual formulation implies that SQ-Riesz regression is equivalent to solving

$$\begin{aligned} \min_{\boldsymbol{\alpha} \in \mathbb{R}^n} \quad & \frac{1}{n} \sum_{i=1}^{n} (\alpha_i - C)^2 \\ \text{subject to} \quad & \left| \frac{1}{n} \sum_{i=1}^{n} \Big(\alpha_i \phi_j(X_i) - m(W_i, \phi_j) \Big) \right| \leq \lambda \qquad j = 1, \ldots, p. \end{aligned}$$

This optimization problem matches that used to obtain stable balancing weights (Zubizarreta, 2015). When $\lambda = 0$, it enforces the covariate balancing condition

$$\sum_{i=1}^{n} \widehat{\alpha}_i \phi_j(D_i, Z_i) - \left(\sum_{i=1}^{n} \Big(\phi_j(1, Z_i) - \phi_j(0, Z_i) \Big) \right) = 0, \text{for } j = 1, 2, \ldots, p,$$

where $\widehat{\alpha}_i = \boldsymbol{\phi}(X_i)^\top \widehat{\boldsymbol{\beta}}$.

An advantage of linear models is that we can express the entire ATE estimation problem using a single linear model, as shown by Bruns-Smith et al. (2025).

4.5 UKL-Riesz Regression with a Logistic or Log Link Function

We next introduce the combination of the UKL divergence loss and a log link function. Consider the log link model

$$\alpha_{\boldsymbol{\beta}}(X) = \xi(X) \Big(C + \exp\big(\boldsymbol{\phi}(X)^\top \boldsymbol{\beta} \big) \Big) - (1 - \xi(X)) \Big(C + \exp\big(-\boldsymbol{\phi}(X)^\top \boldsymbol{\beta} \big) \Big).$$

We call this a log link function because $\alpha_{\boldsymbol{\beta}}(X) = \zeta^{-1}(X, \boldsymbol{\phi}(X)^\top \boldsymbol{\beta})$ can be inverted in $\boldsymbol{\phi}(X)^\top \boldsymbol{\beta}$ by taking logarithms on each branch given $\xi(X)$. For example, in ATE estimation, we set

$\xi(X) = D$ and $C = 1$. In density ratio estimation, we set $\xi(X) = 1$ and $C = 0$, as discussed below.

For this model, using the UKL divergence (UKL-Riesz regression) yields automatic covariate balancing. Specifically, under this specification, if we use ℓ_1-penalty in UKL-Riesz regression, the dual formulation implies that UKL-Riesz regression is equivalent to solving

$$\begin{aligned} \min_{\boldsymbol{\alpha}\in\mathbb{R}^n} \quad & \frac{1}{n}\sum_{i=1}^{n} \operatorname{sign}(\alpha_i)\log\left(|\alpha_i| - C\right) \\ \text{subject to} \quad & \left|\frac{1}{n}\sum_{i=1}^{n}\left(\alpha_i\Big(\xi(X_i)\phi_j(X_i) - (1-\xi(X_i))\phi_j(X_i)\Big) - m\Big(W_i, \phi_j(\cdot)\Big)\right)\right| \leq \lambda \qquad j = 1, \ldots, p. \end{aligned}$$

This optimization problem matches that used to obtain entropy balancing weights (Hainmueller, 2012). When $\lambda = 0$, it enforces the covariate balancing condition

$$\frac{1}{n}\sum_{i=1}^{n}\left(\alpha_i\Big(\xi(X_i)\phi_j(X_i) - (1-\xi(X_i))\phi_j(X_i)\Big) - m\Big(W_i, \phi_j(\cdot)\Big)\right) = 0,$$

where $\widehat{\alpha}_i = \alpha_{\widehat{\boldsymbol{\beta}}}(X_i)$.

An advantage of a log link function is that it naturally imposes modeling assumptions. For example, when we assume a logistic model for the propensity score, the induced Riesz representer takes this log link form.

Special case of the log link function. As a special case, we can model the Riesz representer as

$$\alpha_{\boldsymbol{\beta}}(X) = \exp\left(\boldsymbol{\phi}(X)^\top \boldsymbol{\beta}\right),$$

which corresponds to $\xi(X) = 1$ and $C = 0$. Such a specification appears in density ratio estimation, as discussed below. For this model, using the UKL divergence (UKL-Riesz regression) yields automatic covariate balancing. Specifically, under this specification, the dual formulation implies that UKL-Riesz regression is equivalent to solving

$$\begin{aligned} \min_{\boldsymbol{\alpha}\in\mathbb{R}^n} \quad & \frac{1}{n}\sum_{i=1}^{n} \log\left(\alpha_i\right) \\ \text{subject to} \quad & \left|\frac{1}{n}\sum_{i=1}^{n}\left(\alpha_i\phi_j(X_i) - m\Big(W_i, \phi_j(\cdot)\Big)\right)\right| \leq \lambda \qquad \text{for} \quad j = 1, \ldots, p. \end{aligned}$$

This optimization problem matches that used to obtain entropy balancing weights in the density ratio setting. When $\lambda = 0$, it enforces the covariate balancing condition

$$\frac{1}{n}\sum_{i=1}^{n}\left(\alpha_i\phi_j(X_i) - m\Big(W_i, \phi_j(\cdot)\Big)\right) = 0,$$

where $\widehat{\alpha}_i = \alpha_{\widehat{\boldsymbol{\beta}}}(X_i)$.

Exponential models, or log link functions, are closely related to density ratio modeling. When two probability densities $p(x)$ and $q(x)$ belong to exponential families, the density ratio

can also be expressed in an exponential form. Exponential models also impose nonnegativity of the density ratio without sacrificing smoothness. For example, if we model a density ratio $r(x)$ by a linear model, the linear model can violate the nonnegativity condition, since $r(x) > 0$ must hold.

4.6 BP-Riesz Regression and a Power Link Function

We next introduce a specification that pairs the BP divergence loss with a link function that interpolates between the linear link used for SQ-Riesz regression and the log link used for UKL-Riesz regression. This specification is useful both as a robustness device and as a way to understand how automatic covariate balancing varies continuously with the choice of loss and link functions.

Let $\omega \in (0, \infty)$ and define $k := 1 + 1/\omega$. Consider the following model for the Riesz representer:

$$\alpha_{\boldsymbol{\beta}}(X) = \xi(X)\left(C + \left(1 + \frac{\boldsymbol{\phi}(X)^\top \boldsymbol{\beta}}{k}\right)^{1/\omega}\right) - (1 - \xi(X))\left(C + \left(1 - \frac{\boldsymbol{\phi}(X)^\top \boldsymbol{\beta}}{k}\right)^{1/\omega}\right), \quad (5)$$

where $\boldsymbol{\phi}\colon \mathcal{X} \to \mathbb{R}^p$ is a basis function and $\xi\colon \mathcal{X} \to \{0, 1\}$ selects the branch. We call a link function in (5) a power link function.

The choice of (ξ, C) is application dependent. For example, in ATE estimation we typically set $\xi(X) = D$ and $C = 1$, while in density ratio estimation we often set $\xi(X) = 1$ and $C = 0$ so that $\alpha_{\boldsymbol{\beta}}(X)$ is nonnegative by construction.

Under (5), the dual characterization implies that BP-Riesz regression returns the minimum BP-loss solution among approximately balancing models. In particular, if we use ℓ_1-penalty in BP-Riesz regression, BP-Riesz regression is equivalent to solving a constrained problem of the form

$$\begin{aligned} \min_{\boldsymbol{\alpha} \in \mathbb{R}^n} \quad & \frac{1}{n}\sum_{i=1}^n g^{\mathrm{BP}}(\alpha_i) \\ \text{subject to} \quad & \left|\frac{1}{n}\sum_{i=1}^n \Big(\alpha_i \phi_j(X_i) - m(W_i, \phi_j)\Big)\right| \le \lambda \qquad \text{for} \quad j = 1, \dots, p, \end{aligned} \quad (6)$$

with α_i restricted to the domain of g^{BP}, that is, $|\alpha_i| \ge C$. When $\lambda = 0$, the constraint (6) enforces exact balancing:

$$\frac{1}{n}\sum_{i=1}^n \widehat{\alpha}_i \phi_j(X_i) = \frac{1}{n}\sum_{i=1}^n m(W_i, \phi_j), \qquad \text{for} \quad j = 1, \dots, p,$$

where $\widehat{\alpha}_i = \alpha_{\widehat{\boldsymbol{\beta}}}(X_i)$.

Relationship to the linear and log links. The power link (5) provides a continuous bridge between the specifications in the previous subsections. As $\omega \to 0$, we have $k = 1 + 1/\omega \to \infty$ and

$$\left(1 + \frac{t}{k}\right)^{1/\omega} = \Big(1 + \omega t + o(\omega)\Big)^{1/\omega} \to \exp(t),$$

so (5) reduces to the log link form used for UKL-Riesz regression. At $\omega = 1$, the BP loss reduces to the squared loss, and the link becomes an affine transformation of the linear index around the origin, which connects BP-Riesz regression to SQ-Riesz regression up to reparameterization.

This interpolation perspective is also consistent with the robustness interpretation of the BP divergence (Basu et al., 1998; Sugiyama et al., 2012). Smaller ω makes the objective closer to a KL-type criterion, which can be efficient under correct specification, while larger ω yields behavior closer to squared loss and is typically more robust to misspecification and extreme weights.

5 Applications

This section provides applications of generalized Riesz regression: ATE estimation, AME estimation, and covariate shift adaptation (density ratio estimation).

5.1 ATE Estimation.

In ATE estimation, the linear functional is

$$m^{\mathrm{ATE}}(W, \gamma) \coloneqq \gamma(1, Z) - \gamma(0, Z),$$

and the Riesz representer is

$$\alpha_0^{\mathrm{ATE}}(X) = \frac{D}{e_0(Z)} - \frac{1 - D}{1 - e_0(Z)},$$

where $e_0(Z) = P(D = 1 \mid Z)$ is the propensity score. Let $r_0(1, Z) \coloneqq \frac{1}{e_0(Z)}$ and $r_0(0, Z) \coloneqq \frac{1}{1-e_0(Z)}$ be the inverse propensity score, also called the density ratio. We estimate α_0^{ATE} by minimizing the empirical Bregman divergence objective $\widehat{\mathrm{BD}}_g(\alpha)$ introduced in Section 3, with $m = m^{\mathrm{ATE}}$, an application-specific choice of g, and a model class for α.

SQ-Riesz Regression. We take the squared loss,

$$g^{\mathrm{SQ}}(\alpha) = \alpha^2,$$

and minimize the corresponding empirical Bregman objective. By substituting g^{SQ} into (1) and using $m^{\mathrm{ATE}}(W, \gamma) = \gamma(1, Z) - \gamma(0, Z)$, we obtain, up to an additive constant that does not depend on α,

$$\mathrm{BD}_{g^{\mathrm{SQ}}}(\alpha) = \mathbb{E}\left[\alpha(D, Z)^2 - 2\big(\alpha(1, Z) - \alpha(0, Z)\big)\right].$$

Thus, SQ-Riesz regression estimates α_0^{ATE} by

$$\widehat{\alpha} \coloneqq \arg\min_{\alpha \in \mathcal{H}} \widehat{\mathrm{BD}}_{g^{\mathrm{SQ}}}(\alpha) + \lambda J(\alpha),$$

where

$$\widehat{\mathrm{BD}}_{g^{\mathrm{SQ}}}(\alpha) \coloneqq \frac{1}{n}\sum_{i=1}^{n}\big(\alpha(D_i,Z_i)^2 - 2\big(\alpha(1,Z_i)-\alpha(0,Z_i)\big)\big).$$

This coincides with Riesz regression in Chernozhukov et al. (2021) and corresponds to LSIF in density ratio estimation (Kanamori et al., 2009). With appropriate choices of $\mathcal{H}$, it also recovers nearest neighbor matching-based constructions, as discussed in Kato (2025a).

Remark (Linear link function)**.** *A recommended Riesz representer modeling is*

$$\alpha_{\boldsymbol{\beta}}(X) = \boldsymbol{\phi}(X)^\top \boldsymbol{\beta},$$

where $\boldsymbol{\phi}\colon \mathcal{X} \to \mathbb{R}^p$ is a basis function. Under this model, minimizing the SQ-Riesz regression yields an estimator that satisfies an automatic covariate balancing property, as discussed in Section 4 and Zhao (2019).

Concretely, letting $\widehat{\alpha} = \alpha_{\widehat{\boldsymbol{\beta}}}$, we estimate $\boldsymbol{\beta}$ by

$$\widehat{\boldsymbol{\beta}} \coloneqq \arg\min_{\boldsymbol{\beta}} \frac{1}{n}\sum_{i=1}^{n}\left(\left(\boldsymbol{\phi}(X_i)^\top\boldsymbol{\beta}\right)^2 - 2\left(\boldsymbol{\phi}(1,Z_i)^\top - \boldsymbol{\phi}(0,Z_i)^\top\right)\boldsymbol{\beta}\right) + \frac{1}{a}\lambda\|\boldsymbol{\beta}\|_a^a.$$

By duality, if $a = 1$, SQ-Riesz regression is equivalent to the following covariate balancing problem:

$$\begin{aligned}&\min_{\alpha\in\mathbb{R}^n} \sum_{i=1}^{n}\alpha_i^2\\ &\text{s.t.}\quad \left|\sum_{i=1}^{n}\alpha_i\phi_j(X_i) + \phi_j(1,Z_i) - \phi_j(0,Z_i)\right| \le \lambda \quad \text{for } j = 1,2,\dots,p,\end{aligned}$$

where the solution $\widehat{w}_i$ corresponds to the estimator of $\alpha_0(X_i)$ if $D_i = 1$ and that of $-\alpha_0(X_i)$ if $D_i = 0$; that is,

$$\widehat{w}_i = \begin{cases}\widehat{\alpha}(1,Z_i) & \text{if } D_i = 1,\\ -\widehat{\alpha}(0,Z_i) & \text{if } D_i = 0.\end{cases}.$$

UKL-Riesz Regression. Consider Riesz representer models α such that $\alpha(1,x) > 1$ and $\alpha(0,x) < -1$ for all x. We next use the UKL divergence loss with $C = 1$,

$$g^{\mathrm{UKL}}(\alpha) = (|\alpha| - 1)\log(|\alpha| - 1) - |\alpha|.$$

By substituting g^{UKL} into (1) and using $m^{\mathrm{ATE}}(W,\gamma) = \gamma(1,Z) - \gamma(0,Z)$, we obtain

$$\begin{aligned}\mathrm{BD}_{g^{\mathrm{UKL}}}(\alpha) \coloneqq \mathbb{E}\Big[&\log(|\alpha(X)| - 1) + |\alpha(X)|\\ &- \Big(\operatorname{sign}\big(\alpha(1,Z)\big)\log(|\alpha(1,Z)| - 1) - \operatorname{sign}\big(\alpha(0,Z)\big)\log(|\alpha(0,Z)| - 1)\Big)\Big].\end{aligned}$$

Thus, UKL-Riesz regression estimates α_0^{ATE} by

$$\widehat{\alpha} \coloneqq \arg\min_{\alpha\in\mathcal{H}} \widehat{\mathrm{BD}}_{g^{\mathrm{UKL}}}(\alpha) + \lambda J(\alpha),$$

where

$$\mathrm{BD}_{g^{\mathrm{UKL}}}(\alpha) \coloneqq \frac{1}{n}\sum_{i=1}^{n}\Big(\log\left(|\alpha(X_i)|-1\right)+|\alpha(X_i)| \\ -\Big(\operatorname{sign}\left(\alpha(1,Z_i)\right)\log\left(|\alpha(1,Z_i)|-1\right)-\operatorname{sign}\left(\alpha(0,Z_i)\right)\log\left(|\alpha(0,Z_i)|-1\right)\Big)\Big).$$

This coincides with the tailored loss minimization with $\alpha=\beta=-1$ in Zhao (2019) and corresponds to KLIEP in density ratio estimation (Sugiyama et al., 2008).

Remark (Log link function)**.** *A recommended Riesz representer modeling is*

$$\alpha_{\boldsymbol{\beta}}(X)=\mathbb{1}[D=1]\Big(1+\exp\big(-\boldsymbol{\phi}(X)^\top\boldsymbol{\beta}\big)\Big)-\mathbb{1}[D=0]\Big(1+\exp\big(\boldsymbol{\phi}(X)^\top\boldsymbol{\beta}\big)\Big),$$

where $\boldsymbol{\phi}\colon\mathcal{X}\to\mathbb{R}^p$ is a basis function. Under this model, minimizing the UKL-Riesz regression yields an estimator that satisfies an automatic covariate balancing property, as discussed in Section 4 and Zhao (2019).

Concretely, letting $\widehat{\alpha}=\alpha_{\widehat{\boldsymbol{\beta}}}$, we estimate $\boldsymbol{\beta}$ by

$$\widehat{\boldsymbol{\beta}}\coloneqq\arg\min_{\boldsymbol{\beta}}\frac{1}{n}\sum_{i=1}^{n}\Bigg(\mathbb{1}[D_i=1]\left(-\boldsymbol{\phi}(1,Z_i)^\top\boldsymbol{\beta}+1+\exp\big(-\boldsymbol{\phi}(1,Z_i)^\top\boldsymbol{\beta}\big)\right) \\ +\mathbb{1}[D_i=0]\left(\boldsymbol{\phi}(0,Z_i)^\top\boldsymbol{\beta}+1+\exp\left(\boldsymbol{\phi}(0,Z_i)^\top\boldsymbol{\beta}\right)\right) \\ -\Big(\boldsymbol{\phi}(1,Z_i)^\top-\boldsymbol{\phi}(0,Z_i)^\top\Big)\boldsymbol{\beta}\Bigg)+\frac{1}{a}\lambda\|\boldsymbol{\beta}\|_a^a.$$

By duality, if $a=1$, UKL-Riesz regression is equivalent to the following covariate balancing problem:

$$\min_{\boldsymbol{w}\in(1,\infty)^n}\ \sum_{i=1}^{n}(w_i-1)\log(w_i-1)$$
$$s.t.\quad\left|\sum_{i=1}^{n}\Big(\mathbb{1}[D_i=1]w_i\phi_j(X_i)-\mathbb{1}[D_i=0]w_i\phi_j(X_i)\Big)-\Big(\phi_j(1,Z_i)-\phi_j(0,Z_i)\Big)\right|\le\lambda$$
$$\text{for } j=1,2,\dots,p,$$

where the solution $\widehat{w}_i$ corresponds to the estimator of $\alpha_0(X_i)$ if $D_i=1$ and that of $-\alpha_0(X_i)$ if $D_i=0$; that is,

$$\widehat{w}_i=\begin{cases}\widehat{\alpha}(1,Z_i) & \text{if } D_i=1,\\ -\widehat{\alpha}(0,Z_i) & \text{if } D_i=0.\end{cases}.$$

This modeling enforces the correct signs and nonnegativity of the Riesz representer.

Remark (Propensity score modeling)**.** *We can interpret that the Riesz representer model is based on a propensity score model:*

$$\alpha_{\boldsymbol{\beta}}(X)=\mathbb{1}[D=1]r_{\boldsymbol{\beta}}(1,Z)-\mathbb{1}[D=0]r_{\boldsymbol{\beta}}(0,Z),$$

where

$$r_{\boldsymbol{\beta}}(1,Z)=\frac{1}{e_{\boldsymbol{\beta}}(Z)},\qquad r_{\boldsymbol{\beta}}(0,Z)=\frac{1}{1-e_{\boldsymbol{\beta}}(Z)},$$
$$e_{\boldsymbol{\beta}}(Z)\coloneqq\frac{1}{1+\exp\big(-\boldsymbol{\phi}(Z)^\top\boldsymbol{\beta}\big)},$$

and $\boldsymbol{\phi}\colon\mathcal{Z}\to\mathbb{R}^p$ *is a basis function. Under this model, minimizing the UKL flavored empirical Bregman objective yields an estimator that satisfies an automatic covariate balancing property, as discussed in Section 4 and Zhao (2019).*

Concretely, letting $\widehat{\alpha}=\alpha_{\widehat{\boldsymbol{\beta}}}$*, we estimate* $\boldsymbol{\beta}$ *by*

$$\widehat{\boldsymbol{\beta}}\coloneqq\arg\min_{\boldsymbol{\beta}}\frac{1}{n}\sum_{i=1}^{n}\Bigg(\mathbb{1}[D_i=1]\left(-\log\left(\frac{1}{r_{\boldsymbol{\beta}}(1,Z_i)-1}\right)+r_{\boldsymbol{\beta}}(1,Z_i)\right)$$
$$+\mathbb{1}[D_i=0]\left(-\log\left(\frac{1}{r_{\boldsymbol{\beta}}(0,Z_i)-1}\right)+r_{\boldsymbol{\beta}}(0,Z_i)\right)\Bigg)+\frac{1}{a}\lambda\|\boldsymbol{\beta}\|_a^a.$$

By duality, if $a=1$*, this KL divergence objective is equivalent to a covariate balancing program:*

$$\min_{\boldsymbol{w}\in(1,\infty)^n}\ \sum_{i=1}^{n}(w_i-1)\log(w_i-1)$$
$$s.t.\quad\left|\sum_{i=1}^{n}\Big(\mathbb{1}[D_i=1]w_i\phi_j(Z_i)-\mathbb{1}[D_i=0]w_i\phi_j(Z_i)\Big)\right|\le\lambda\quad\text{for } j=1,2,\dots,p,$$

where the solution $\widehat{w}_i$ *corresponds to the estimator of* $\alpha_0(X_i)$ *if* $D_i=1$ *and that of* $-\alpha_0(X_i)$ *if* $D_i=0$*; that is,*

$$\widehat{w}_i=\begin{cases}\widehat{r}(1,Z_i) & \text{if } D_i=1,\\ \widehat{r}(0,Z_i) & \text{if } D_i=0.\end{cases},$$

and $\widehat{r}$ *is an estimator of the density ratio* r_0*. This constrained optimization matches entropy balancing (Hainmueller, 2012). In particular, when* $\lambda=0$ *we obtain exact balance,*

$$\sum_{i=1}^{n}\Big(\mathbb{1}[D_i=1]\widehat{w}_i\phi_j(Z_i)-\mathbb{1}[D_i=0]\widehat{w}_i\phi_j(Z_i)\Big)=0\quad\text{for } j=1,2,\dots,p.$$

This specification has the practical advantage that $\phi_j(Z)$ *can be chosen independently of* D*, which reduces the dimension of the model.*

BP-Riesz Regression. BP-Riesz regression uses Basu's power divergence with $C=1$ and $\omega\in(0,\infty)$:

$$g^{\mathrm{BP}}(\alpha)\coloneqq\frac{\big(|\alpha|-1\big)^{1+\omega}-\big(|\alpha|-1\big)}{\omega}-|\alpha|.$$

Plugging g^{BP} into (1) and using m^{ATE} yields the empirical objective

$$\widehat{\mathrm{BD}}_{g^{\mathrm{BP}}}(\alpha) \coloneqq \frac{1}{n}\sum_{i=1}^{n}\Bigg(\frac{\big(|\alpha(D_i,Z_i)|-1\big)^{\omega}-1}{\omega}+|\alpha(D_i,Z_i)|\big(|\alpha(D_i,Z_i)|-1\big)^{\omega} \\ -\upsilon\Big(\big(|\alpha(1,X_i)|-1\big)^{\omega}-\big(|\alpha(0,X_i)|-1\big)^{\omega}\Big)\Bigg),$$

where $\upsilon \coloneqq 1+1/\omega$. We then estimate α_0^{ATE} by

$$\widehat{\alpha} \coloneqq \operatorname*{arg\,min}_{\alpha\in\mathcal{H}} \widehat{\mathrm{BD}}_{g^{\mathrm{BP}}}(\alpha)+\lambda J(\alpha).$$

Remark (Power link function)**.** *A convenient parametric specification that is consistent with Section 4 models α through a power link function for the inverse propensity components:*

$$\alpha_{\boldsymbol{\beta}}(X)=\mathbb{1}[D=1]r_{\boldsymbol{\beta}}(1,Z)-\mathbb{1}[D=0]r_{\boldsymbol{\beta}}(0,Z),$$

with

$$r_{\boldsymbol{\beta}}(1,Z)\coloneqq 1+\left(1+\frac{\boldsymbol{\phi}(1,Z)^{\top}\boldsymbol{\beta}}{\upsilon}\right)^{1/\omega},\qquad r_{\boldsymbol{\beta}}(0,Z)\coloneqq 1+\left(1-\frac{\boldsymbol{\phi}(0,Z)^{\top}\boldsymbol{\beta}}{\upsilon}\right)^{1/\omega},$$

on the domain where the above powers are well defined. This specification interpolates between the squared distance and UKL divergence: $\omega=1$ recovers SQ-Riesz regression, and the limit $\omega\to 0$ recovers UKL-Riesz regression, as discussed in Section 3. In applications, BP-Riesz regression can mitigate sensitivity to extreme inverse propensity weights while retaining the covariate balancing behavior implied by the dual characterization.

BKL-Riesz Regression. Consider Riesz representer models α such that $\alpha(1,x)>1$ and $\alpha(0,x)<-1$ for all x. BKL-Riesz regression uses BKL divergence with $C=1$:

$$g^{\mathrm{BKL}}(\alpha)\coloneqq(|\alpha|-1)\log\big(|\alpha|-1\big)-(|\alpha|+1)\log(|\alpha|+1).$$

By plugging g^{BKL} into (1) and using m^{ATE}, we have the following empirical objective function:

$$\widehat{\mathrm{BD}}_{g^{\mathrm{BKL}}}(\alpha)\coloneqq\frac{1}{n}\sum_{i=1}^{n}\Bigg(\log\left(\frac{|\alpha(X_i)|-1}{|\alpha(X_i)|+1}\right) \\ -\left(\log\left(\frac{\alpha(1,Z_i)-1}{\alpha(1,Z_i)+1}\right)+\log\left(\frac{-\alpha(0,Z_i)-1}{-\alpha(0,Z_i)+1}\right)\right)\Bigg).$$

We then estimate α_0^{ATE} by

$$\widehat{\alpha}\coloneqq\operatorname*{arg\,min}_{\alpha\in\mathcal{H}}\widehat{\mathrm{BD}}_{g^{\mathrm{BKL}}}(\alpha)+\lambda J(\alpha).$$

Remark (MLE of the propensity score)**.** *BKL-Riesz regression corresponds to estimating the propensity score by regularized logistic likelihood, which is the standard MLE approach in ATE estimation. Let*

$$e_{\boldsymbol{\beta}}(Z) \coloneqq \frac{1}{1+\exp\big(-\boldsymbol{\phi}(Z)^\top \boldsymbol{\beta}\big)},$$

and define the Riesz representer model obtained by plugging in $e_{\boldsymbol{\beta}}$*,*

$$\alpha_{\boldsymbol{\beta}}(X) \coloneqq \frac{D}{e_{\boldsymbol{\beta}}(Z)} - \frac{1-D}{1-e_{\boldsymbol{\beta}}(Z)}.$$

Under the BKL choice in Section 3, minimizing the corresponding empirical Bregman divergence specializes to minimizing the Bernoulli negative log-likelihood:

$$\widehat{\boldsymbol{\beta}} \coloneqq \arg\min_{\boldsymbol{\beta}} -\frac{1}{n}\sum_{i=1}^{n}\Big(D_i \log e_{\boldsymbol{\beta}}(Z_i) + (1-D_i)\log\big(1-e_{\boldsymbol{\beta}}(Z_i)\big)\Big) + \lambda\|\boldsymbol{\beta}\|_2^2,$$

and we set $\widehat{e}(Z) \coloneqq e_{\widehat{\boldsymbol{\beta}}}(Z)$ *and*

$$\widehat{\alpha}(X) \coloneqq \frac{D}{\widehat{e}(Z)} - \frac{1-D}{1-\widehat{e}(Z)}.$$

This viewpoint aligns with the interpretation of BKL-Riesz as a probabilistic classification approach to density ratio estimation (Qin, 1998; Cheng & Chu, 2004), here applied to treatment assignment modeling. It also provides a baseline for comparison with the direct objectives in SQ-Riesz, UKL-Riesz, and BP-Riesz regression.

5.2 AME Estimation

We consider the AME setup described in Section 2. Let $X = (D, Z)$, where D is a scalar continuous regressor and Z is a vector of covariates. The target parameter is

$$\theta_0^{\mathrm{AME}} \coloneqq \mathbb{E}\Big[\partial_d \gamma_0(D,Z)\Big], \qquad m^{\mathrm{AME}}(W,\gamma) \coloneqq \partial_d \gamma(D,Z).$$

Assume that X admits a density f_0 that is continuously differentiable and that an integration by parts argument is valid, for example, $\gamma(x)f_0(x)$ vanishes on the boundary of the support in the d direction. Then

$$\mathbb{E}\Big[\partial_d \gamma(X)\Big] = \mathbb{E}\Big[\alpha_0^{\mathrm{AME}}(X)\gamma(X)\Big], \qquad \alpha_0^{\mathrm{AME}}(X) = -\partial_d \log f_0(X),$$

so the AME Riesz representer is the negative score of the marginal density of X with respect to d. Since $\partial_d \log f_0(D,Z) = \partial_d \log f_0(D \mid Z)$, we can equivalently view α_0^{AME} as the negative score of the conditional density of D given Z.

To estimate α_0^{AME}, we apply generalized Riesz regression with $m = m^{\mathrm{AME}}$. The population objective in Section 3 becomes

$$\mathrm{BD}_g^{\mathrm{AME}}(\alpha) \coloneqq \mathbb{E}\Big[-g\big(\alpha(X)\big) + \partial g\big(\alpha(X)\big)\alpha(X) - \partial_d\Big(\partial g\big(\alpha(X)\big)\Big)\Big],$$

and we minimize its empirical analogue over a differentiable model class $\mathcal{H}$ (so that $\partial_d \alpha(X)$ and $\partial_d\{\partial g(\alpha(X))\}$ are well defined), possibly with regularization.

SQ-Riesz Regression. Let $g^{\mathrm{SQ}}(\alpha) = (\alpha - C)^2$ for an arbitrary constant $C \in \mathbb{R}$, so that $\partial g^{\mathrm{SQ}}(\alpha) = 2(\alpha - C)$. Substituting into $\mathrm{BD}_g^{\mathrm{AME}}$ yields

$$\begin{aligned}\mathrm{BD}_{g^{\mathrm{SQ}}}^{\mathrm{AME}}(\alpha) &= \mathbb{E}\Big[-\big(\alpha(X) - C\big)^2 + 2\big(\alpha(X) - C\big)\alpha(X) - \partial_d\Big(2\big(\alpha(X) - C\big)\Big)\Big] \\ &= \mathbb{E}\Big[\alpha(X)^2 - 2\partial_d\alpha(X)\Big] + \text{const},\end{aligned}$$

where the constant does not depend on α. Hence SQ-Riesz regression targets α_0^{AME} in L_2. This objective is also a score matching style criterion for estimating the score, written here in terms of the negative score α_0^{AME}. The empirical estimator is

$$\widehat{\alpha} := \arg\min_{\alpha \in \mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}\Big(\alpha(X_i)^2 - 2\partial_d\alpha(X_i)\Big) + \lambda J(\alpha).$$

This method corresponds to Riesz regression for AME, as discussed in Chernozhukov et al. (2021).

UKL-Riesz Regression. To obtain a KL motivated loss that allows signed α, we use the signed KL type convex function

$$g^{\mathrm{UKL}}(\alpha) = |\alpha|\log|\alpha| - |\alpha|, \qquad \partial g^{\mathrm{UKL}}(\alpha) = \mathrm{sign}(\alpha)\log|\alpha|,$$

on a domain that excludes $\alpha = 0$. Plugging into $\mathrm{BD}_g^{\mathrm{AME}}$ gives

$$\begin{aligned}\mathrm{BD}_{g^{\mathrm{UKL}}}^{\mathrm{AME}}(\alpha) &= \mathbb{E}\Big[-|\alpha(X)|\log|\alpha(X)| + |\alpha(X)| + \mathrm{sign}\big(\alpha(X)\big)\log|\alpha(X)|\alpha(X) - \partial_d\Big(\mathrm{sign}\big(\alpha(X)\big)\log|\alpha(X)|\Big)\Big] \\ &= \mathbb{E}\Big[|\alpha(X)| - \partial_d\Big(\mathrm{sign}\big(\alpha(X)\big)\log|\alpha(X)|\Big)\Big] + \text{const}.\end{aligned}$$

Accordingly, we estimate α_0^{AME} by minimizing the empirical version over $\mathcal{H}$:

$$\widehat{\alpha} := \arg\min_{\alpha \in \mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}\Big(|\alpha(X_i)| - \partial_d\Big(\mathrm{sign}\big(\alpha(X_i)\big)\log|\alpha(X_i)|\Big)\Big) + \lambda J(\alpha).$$

In practice, one can use the shifted UKL loss in Section 3 to avoid the singularity at zero and combine it with a branchwise link specification as in Section 4.

BP-Riesz Regression. BP-Riesz regression interpolates between squared distance and UKL divergence. For simplicity, we present the unshifted form with $C = 0$:

$$g^{\mathrm{BP}}(\alpha) := \frac{|\alpha|^{1+\gamma} - |\alpha|}{\gamma} - |\alpha|, \qquad \partial g^{\mathrm{BP}}(\alpha) = \left(1 + \frac{1}{\gamma}\right)\mathrm{sign}(\alpha)\Big(|\alpha|^{\gamma} - 1\Big), \qquad \gamma \in (0, \infty).$$

Let $k := 1 + 1/\gamma$. Then $\mathrm{BD}_g^{\mathrm{AME}}$ simplifies to

$$\mathrm{BD}_{g^{\mathrm{BP}}}^{\mathrm{AME}}(\alpha) = \mathbb{E}\Big[|\alpha(X)|^{1+\gamma} - \partial_d\Big(k\,\mathrm{sign}\big(\alpha(X)\big)\Big(|\alpha(X)|^{\gamma} - 1\Big)\Big)\Big] + \text{const}.$$

We estimate α_0^{AME} by minimizing the empirical objective:

$$\widehat{\alpha} \coloneqq \arg\min_{\alpha\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}\Big(|\alpha(X_i)|^{1+\gamma} - \partial_d\Big(k\,\mathrm{sign}\,\big(\alpha(X_i)\big)\Big(|\alpha(X_i)|^{\gamma}-1\Big)\Big)\Big) + \lambda J(\alpha).$$

As in Section 3, $\gamma = 1$ recovers the squared loss behavior (up to scaling), while $\gamma \to 0$ approaches a KL type criterion through the identity $\lim_{\gamma\to 0}(|\alpha|^{\gamma}-1)/\gamma = \log|\alpha|$.

BKL-Riesz Regression. Finally, we can use the BKL loss from Section 3 to obtain a logistic motivated criterion. Let $C > 0$ and define

$$g^{\mathrm{BKL}}(\alpha) \coloneqq (|\alpha|-C)\log\big(|\alpha|-C\big) - (|\alpha|+C)\log\big(|\alpha|+C\big), \qquad \partial g^{\mathrm{BKL}}(\alpha) = \mathrm{sign}(\alpha)\log\left(\frac{|\alpha|-C}{|\alpha|+C}\right).$$

Then the AME objective is

$$\mathrm{BD}^{\mathrm{AME}}_{g^{\mathrm{BKL}}}(\alpha) = \mathbb{E}\left[C\log\left(\frac{|\alpha(X)|-C}{|\alpha(X)|+C}\right) - \partial_d\left(\mathrm{sign}\,\big(\alpha(X)\big)\log\left(\frac{|\alpha(X)|-C}{|\alpha(X)|+C}\right)\right)\right] + \mathrm{const},$$

and the estimator minimizes its empirical counterpart over $\mathcal{H}$ with regularization. As in the ATE case, this loss is naturally paired with a logistic style link for the magnitude of α, while sign changes can be handled via the branchwise constructions in Section 4.

Once we obtain $\widehat{\alpha}^{\mathrm{AME}}$ and an outcome regression estimator $\widehat{\gamma}$, we plug them into the Neyman orthogonal score in Section 2 to form an estimator of θ_0^{AME}.

5.3 Covariate Shift Adaptation (Density Ratio Estimation)

We consider the covariate shift setting in Section 2. Let X be the source covariate distribution that generates labeled observations $\{(X_i, Y_i)\}_{i=1}^n$, and let $\widetilde{X}$ be the target covariate distribution that generates unlabeled observations $\{\widetilde{X}_j\}_{j=1}^m$, independent of the source sample. Let $p_0(x)$ and $p_1(x)$ be the pdfs of X and $\widetilde{X}$, respectively. We assume that $p_0(x), p_1(x) > 0$ for all $x \in \mathcal{X}$. The Riesz representer for covariate shift adaptation is the density ratio

$$\alpha_0^{\mathrm{CS}}(X) = r_0(X) \coloneqq \frac{p_1(X)}{p_0(X)}.$$

We estimate r_0 directly by density ratio fitting under a Bregman divergence, avoiding separate density estimation for $p_0(X)$ and $p_1(X)$.

Let $g\colon \mathbb{R}_+ \to \mathbb{R}$ be differentiable and strictly convex. The Bregman divergence between r_0 and a candidate ratio model α is

$$\mathrm{BD}_g^{\dagger}\big(r_0 \mid \alpha\big) \coloneqq \mathbb{E}_X\Big(g\big(r_0(X)\big) - g\big(\alpha(X)\big) - \partial g\big(\alpha(X)\big)\big(r_0(X)-\alpha(X)\big)\Big).$$

Dropping the constant $\mathbb{E}_X\big[g(r_0(X))\big]$ and using the identity $\mathbb{E}_X\big[r_0(X)h(X)\big] = \mathbb{E}_{\widetilde{X}}\big[h(X)\big]$, we obtain the equivalent population objective

$$\mathrm{BD}_g^{\mathrm{CS}}(\alpha) \coloneqq \mathbb{E}_X\big[\partial g(\alpha(X))\alpha(X) - g(\alpha(X))\big] - \mathbb{E}_{\widetilde{X}}\big[\partial g(\alpha(X))\big].$$

Given samples $\{X_i\}_{i\in\mathcal{I}_S}$ and $\{\widetilde{X}_j\}_{j\in\mathcal{I}_T}$, the empirical objective is

$$\widehat{\mathrm{BD}}_g^{\mathrm{CS}}(\alpha) \coloneqq \frac{1}{n}\sum_{i=1}^{n}\Big(\partial g\big(\alpha(X_i)\big)\alpha(X_i) - g\big(\alpha(X_i)\big)\Big) - \frac{1}{m}\sum_{j=1}^{m}\partial g\big(\alpha(\widetilde{X}_j)\big). \tag{7}$$

We estimate the density ratio by

$$\widehat{\alpha} \coloneqq \arg\min_{\alpha\in\mathcal{H}} \widehat{\mathrm{BD}}_g^{\mathrm{CS}}(\alpha) + \lambda J(\alpha),$$

where $\mathcal{H}$ is a model class and J is a regularizer. A convenient way to enforce $\alpha(x) \geq 0$ is to use a link specification such as $\alpha(x) = \exp\big(f(x)\big)$ with a flexible regression model f.

SQ-Riesz Regression. For the squared loss, take

$$g^{\mathrm{SQ}}(\alpha) = (\alpha-1)^2, \qquad \partial g^{\mathrm{SQ}}(\alpha) = 2(\alpha-1).$$

Substituting into (7) and dropping constants that do not depend on α, we obtain

$$\widehat{\mathrm{BD}}_{g^{\mathrm{SQ}}}^{\mathrm{CS}}(\alpha) = \frac{1}{n}\sum_{i=1}^{n}\alpha(X_i)^2 - \frac{2}{m}\sum_{j=1}^{m}\alpha(\widetilde{X}_j).$$

This is the classical least-squares importance fitting (LSIF) objective in density ratio estimation (Kanamori et al., 2009). In the debiased machine learning literature, the same squared loss criterion is also used as Riesz regression for covariate shift adaptation (Chernozhukov et al., 2025). Related extensions include doubly robust covariate shift adaptation schemes that combine density ratio estimation with regression adjustment (Kato et al., 2024a).

UKL-Riesz Regression. For a KL motivated objective on $\mathbb{R}_+$, take

$$g^{\mathrm{UKL}}(\alpha) = \alpha\log\alpha - \alpha, \qquad \partial g^{\mathrm{UKL}}(\alpha) = \log\alpha.$$

Then (7) becomes, up to an additive constant that does not depend on α,

$$\widehat{\mathrm{BD}}_{g^{\mathrm{UKL}}}^{\mathrm{CS}}(\alpha) = \frac{1}{n}\sum_{i=1}^{n}\alpha(X_i) - \frac{1}{m}\sum_{j=1}^{m}\log\alpha(\widetilde{X}_j).$$

A standard implementation imposes the normalization constraint $\frac{1}{n}\sum_{i=1}^{n}\alpha(X_i) = 1$, in which case minimizing $\widehat{\mathrm{BD}}_{g^{\mathrm{UKL}}}^{\mathrm{CS}}$ is equivalent to maximizing the target log-likelihood $\frac{1}{m}\sum_{j=1}^{m}\log\alpha(\widetilde{X}_j)$ subject to normalization and nonnegativity, which yields KLIEP style procedures (Sugiyama et al., 2008). This constrained view is also useful for understanding the dual characterization and the associated moment matching property.

BP-Riesz Regression. BP-Riesz regression interpolates between squared loss and KL type objectives. For $\gamma \in (0, \infty)$, consider the BP choice on $\mathbb{R}_+$ with $C = 0$,

$$g^{\mathrm{BP}}(\alpha) \coloneqq \frac{\alpha^{1+\gamma} - \alpha}{\gamma} - \alpha, \qquad \partial g^{\mathrm{BP}}(\alpha) = \left(1 + \frac{1}{\gamma}\right)\left(\alpha^{\gamma} - 1\right).$$

A useful simplification is that $\partial g^{\mathrm{BP}}(\alpha)\alpha - g^{\mathrm{BP}}(\alpha) = \alpha^{1+\gamma}$, so (7) reduces, up to constants, to

$$\widehat{\mathrm{BD}}^{\mathrm{CS}}_{g^{\mathrm{BP}}}(\alpha) = \frac{1}{n}\sum_{i=1}^{n} \alpha(X_i)^{1+\gamma} - \left(1 + \frac{1}{\gamma}\right)\frac{1}{m}\sum_{j=1}^{m} \alpha(\widetilde{X}_j)^{\gamma}.$$

When $\gamma = 1$, this objective coincides with the SQ-Riesz objective, up to scaling and constants. As $\gamma \to 0$, it approaches a KL flavored criterion via the expansion $\alpha^{\gamma} = 1 + \gamma \log \alpha + o(\gamma)$, providing a continuous bridge between LSIF and KLIEP, and offering a robustness device against extreme ratios (Basu et al., 1998; Sugiyama et al., 2012).

BKL-Riesz Regression. BKL-Riesz regression corresponds to probabilistic classification based density ratio estimation, which estimates the log density ratio by fitting a classifier that discriminates target covariates from source covariates (Qin, 1998; Cheng & Chu, 2004). Let $S \in \{0, 1\}$ denote a domain indicator, where $S = 1$ for target and $S = 0$ for source, and let $\pi \coloneqq P(S = 1)$ denote the mixture class prior. Under Bayes' rule, the density ratio satisfies

$$r_0(x) = \frac{p_1(x)}{p_0(x)} = \frac{1 - \pi}{\pi}\frac{P(S = 1 \mid X = x)}{P(S = 0 \mid X = x)}.$$

We model $P(S = 1 \mid X = x)$ by a logistic specification

$$p_{\boldsymbol{\beta}}(S = 1 \mid X = x) \coloneqq \frac{1}{1 + \exp\left(-\boldsymbol{\phi}(x)^{\top}\boldsymbol{\beta}\right)},$$

and estimate $\boldsymbol{\beta}$ by regularized Bernoulli likelihood on the pooled sample:

$$\widehat{\boldsymbol{\beta}} \coloneqq \arg\min_{\boldsymbol{\beta}} -\frac{1}{n+m}\left(\sum_{i=1}^{n} \log\left(1 - p_{\boldsymbol{\beta}}(S = 1 \mid X_i)\right) + \sum_{j=1}^{m} \log p_{\boldsymbol{\beta}}(S = 1 \mid \widetilde{X}_j)\right) + \lambda\|\boldsymbol{\beta}\|_2^2.$$

With $\widehat{\pi} \coloneqq \frac{m}{n+m}$, we then set

$$\widehat{\alpha}(x) \coloneqq \frac{1 - \widehat{\pi}}{\widehat{\pi}}\frac{p_{\widehat{\boldsymbol{\beta}}}(S = 1 \mid X = x)}{1 - p_{\widehat{\boldsymbol{\beta}}}(S = 1 \mid X = x)} = \frac{1 - \widehat{\pi}}{\widehat{\pi}}\exp\left(\boldsymbol{\phi}(x)^{\top}\widehat{\boldsymbol{\beta}}\right).$$

This construction enforces nonnegativity by design and connects density ratio estimation to standard classification tools.

Remark (From density ratio estimation to covariate shift adaptation)**.** *Once we obtain* $\widehat{\alpha}$ *and an outcome regression estimator* $\widehat{\gamma}$*, we plug them into the covariate shift Neyman score*

in Section 2. In particular, a doubly robust estimator that accommodates separate source and target samples is

$$\widehat{\theta}^{CS} \coloneqq \frac{1}{m}\sum_{j=1}^{m}\widehat{\gamma}(\widetilde{X}_j) + \frac{1}{n}\sum_{i=1}^{n}\widehat{\alpha}(X_i)\Big(Y_i - \widehat{\gamma}(X_i)\Big).$$

The corresponding IPW estimator is obtained by dropping the regression adjustment term and using $\widehat{\theta}^{CS}_{IPW} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\alpha}(X_i)Y_i$. *Cross fitting can be applied by estimating* $\widehat{\alpha}$ *and* $\widehat{\gamma}$ *on auxiliary folds and evaluating the above scores on held out folds.*

6 Automatic Neyman Orthogonalization

The advantage of (automatic) covariate balancing lies in the fact that it automatically implies Neyman orthogonality. Specifically, we consider an estimator given by

$$\widehat{\theta} \coloneqq \frac{1}{n}\sum_{i=1}^{n}\widehat{\alpha}(X_i)Y_i,$$

where $\widehat{\alpha}$ is an estimator of the Riesz representer obtained from generalized Riesz regression. Recall that we refer to this estimator as a Riesz Weighted (RW) estimator, which corresponds to the IPW estimator in ATE estimation (Horvitz & Thompson, 1952) and covariate shift adaptation with importance weighting (Shimodaira, 2000). We show that this estimator satisfies Neyman orthogonality if the regression function γ_0 belongs to the linear space spanned by $\boldsymbol{\phi}(X)$, which is used for Riesz representer estimation. We refer to this property as automatic Neyman orthogonalization.

6.1 Automatic Neyman Orthogonalization

The following result holds for a Riesz representer model $\widehat{\alpha}$ trained by generalized Riesz regression. As discussed below, this result implies that if exact balancing holds,

- the RW estimator is automatically Neyman orthogonal;
- the choice of loss and link functions does not affect the final estimator.

Theorem 6.1. *Consider generalized Riesz regression with the model*

$$\alpha(X) = \zeta^{-1}\Big(X, \boldsymbol{\phi}(X)^\top\boldsymbol{\beta}\Big),$$

where ζ^{-1} *is the inverse of a link function and* $\boldsymbol{\phi}\colon \mathcal{X} \to \mathbb{R}^p$ *is a vector of basis functions. Suppose that* $\partial g\big(\alpha_{\boldsymbol{\beta}}(X_i)\big) = \sum_{j=1}^{p}\beta_j\phi_j(X_i)$ *holds. If* γ_0 *belongs to the linear span of* $\boldsymbol{\phi}(X)$*, then it holds that*

$$\begin{aligned}\widehat{\theta} &\coloneqq \frac{1}{n}\sum_{i=1}^{n}\widehat{\alpha}(X_i)Y_i \\ &= \frac{1}{n}\sum_{i=1}^{n}\Big(\widehat{\alpha}(X_i)\big(Y_i - \gamma_0(X_i)\big) + m(W_i, \gamma_0)\Big).\end{aligned}$$

If $\widehat{\alpha}$ *satisfies the Donsker condition and is consistent, then* $\widehat{\theta}$ *is asymptotically efficient.*

Theorem 6.1 is shown as follows. Recall that the sample average of the Neyman orthogonal score is given by

$$\frac{1}{n}\sum_{i=1}^{n}\psi(W;\widetilde{\eta},\theta)=\frac{1}{n}\sum_{i=1}^{n}\Big(\widehat{\alpha}(X_i)\big(Y_i-\gamma_0(X_i)\big)+m(W_i,\gamma_0)-\theta\Big),$$

where $\widetilde{\eta}\coloneqq(\gamma_0,\widehat{\alpha})$.

In Section 4, we model the Riesz representer as $\alpha(X)=\zeta^{-1}\Big(X,\boldsymbol{\phi}(X)^\top\boldsymbol{\beta}\Big)$. Then, a Riesz representer estimator $\widehat{\alpha}$ trained by generalized Riesz regression with $\lambda=0$ satisfies

$$\frac{1}{n}\sum_{i=1}^{n}\Big(\widehat{\alpha}(X_i)\phi_j(X_i)-m(W_i,\phi_j)\Big)=0$$

for all $j=1,2,\ldots,p$ by the automatic covariate balancing property (Corollary 4.2).

If $\gamma_0(X)$ is given as $\gamma_0(X)=\boldsymbol{\phi}(X)^\top\boldsymbol{\rho}_0$ for some $\boldsymbol{\rho}_0\in\mathbb{R}^p$, then we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\psi(W;\widetilde{\eta},\theta)\\
&=\frac{1}{n}\sum_{i=1}^{n}\Big(\widehat{\alpha}(X_i)\big(Y_i-\gamma_0(X_i)\big)+m(W_i,\gamma_0)-\theta\Big)\\
&=\frac{1}{n}\sum_{i=1}^{n}\Big(\widehat{\alpha}(X_i)\big(Y_i-\boldsymbol{\phi}(X_i)^\top\boldsymbol{\rho}_0\big)+\boldsymbol{\rho}_0^\top m(W_i,\boldsymbol{\phi}(X))-\theta\Big)\\
&=\frac{1}{n}\sum_{i=1}^{n}\Big(\widehat{\alpha}(X_i)Y_i-\theta\Big),
\end{aligned}$$

where we used

$$\sum_{j=1}^{p}\rho_{0,j}\frac{1}{n}\sum_{i=1}^{n}\Big(\widehat{\alpha}(X_i)\phi_j(X_i)-m(W_i,\phi_j)\Big)=0$$

from the automatic covariate balancing property.

Such a property has been reported by Wong & Chan (2017) and Zhao (2019) in ATE estimation. Although they require a constant treatment effect, our results generalize them by allowing heterogeneous treatment effects. We emphasize that (i) the Riesz representer estimator $\widehat{\alpha}$ does not require a specific convergence rate; (ii) we do not use cross-fitting, and we assume the Donsker condition (if we use cross-fitting, the covariate balancing collapses).

Riesz weighted estimator = OLS = doubly robust. Theorem 6.1 is closely related to Proposition 3.1 in Bruns-Smith et al. (2025), which shows the following result. Let an OLS estimator of $\gamma_0(X)$ be $\widehat{\gamma}_{\mathrm{OLS}}(X)\coloneqq\boldsymbol{\phi}(X)^\top\widehat{\boldsymbol{\rho}}_{\mathrm{OLS}}$, where

$$\widehat{\boldsymbol{\rho}}_{\mathrm{OLS}}\coloneqq\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\phi}(X_i)\boldsymbol{\phi}(X_i)^\top\right)^{\dagger}\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\phi}(X_i)Y_i,$$

where † denotes the pseudo-inverse and $\boldsymbol{\phi}(X)$ is the basis function also used in the Riesz representer estimator. Then, for $\widehat{\alpha}(X) = \boldsymbol{\phi}(X)^\top \widehat{\boldsymbol{\beta}}$, it holds that

$$\begin{aligned}
\widehat{\theta} &\coloneqq \frac{1}{n}\sum_{i=1}^{n} \widehat{\alpha}(X_i) Y_i \\
&= \frac{1}{n}\sum_{i=1}^{n} \widehat{\alpha}(X_i)\boldsymbol{\phi}(X_i)^\top \left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\phi}(X_i)\boldsymbol{\phi}(X_i)^\top\right)^{\dagger} \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\phi}(X_i) Y_i \\
&= \frac{1}{n}\sum_{i=1}^{n} \widehat{\alpha}(X_i)\boldsymbol{\phi}(X_i)^\top \widehat{\boldsymbol{\rho}}_{\mathrm{OLS}} \\
&= \frac{1}{n}\sum_{i=1}^{n} \widehat{\alpha}(X_i)\widehat{\gamma}_{\mathrm{OLS}}(X_i).
\end{aligned}$$

Thus, the Riesz weighted estimator can be written in terms of an OLS estimator. Furthermore, if exact balancing holds ($\lambda = 0$), we have

$$\begin{aligned}
\widehat{\theta} &\coloneqq \frac{1}{n}\sum_{i=1}^{n} \widehat{\alpha}(X_i) Y_i \\
&= \frac{1}{n}\sum_{i=1}^{n} \widehat{\alpha}(X_i)\boldsymbol{\phi}(X_i)^\top \widehat{\boldsymbol{\rho}}_{\mathrm{OLS}} \\
&= \frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} m(W_i, \phi_1) \\ m(W_i, \phi_2) \\ \vdots \\ m(W_i, \phi_p) \end{pmatrix}^\top \widehat{\boldsymbol{\rho}}_{\mathrm{OLS}},
\end{aligned}$$

where we used

$$\frac{1}{n}\sum_{i=1}^{n} \Big(\widehat{\alpha}(X_i)\phi_j(X_i) - m(W_i, \phi_j)\Big) = 0$$

for all $j = 1, 2, \ldots, p$.

This result essentially implies that the Riesz weighted estimator with exact balancing ($\lambda = 0$) is a sample average of an OLS estimator, projected onto the parameter-of-interest space by the known parameter function m. Here, recall that a Riesz weighted estimator is automatically Neyman orthogonal, which corresponds to double robustness in many settings. Therefore, this property corresponds to the result that "Ordinary Least Squares (OLS) is doubly robust," as discussed in Robins et al. (2007); that is, the OLS estimator

This result also implies the perhaps surprising fact that the choice of loss and link functions does not affect the final estimator if exact covariate balancing holds. Even though this is convenient because it allows us to omit the Riesz representer estimation procedure, this property does not guarantee good performance of the final estimator. In fact, the OLS estimator can suffer from severe overfitting. In such cases, adding regularization may improve performance. Adding regularization implies inexact covariate balancing.

General case with inexact covariate balancing. Bruns-Smith & Feller (2022) further generalizes this result from the augmented covariate balancing viewpoint to the case with inexact covariate balancing. Such cases occur when using a regularization parameter $\lambda \neq 0$ or cross-fitting.

For this case, Bruns-Smith et al. (2025) show that the AWR estimator can be written as the sample average of two regression function estimators. Recall that the AWR estimator is

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n}\Big(m(W_i, \widehat{\gamma}) + \widehat{\alpha}(X_i)\Big(Y_i - \widehat{\gamma}(X_i)\Big)\Big).$$

Then, if the Riesz representer estimator $\widehat{\alpha}(X)$ can be written as $\widehat{\alpha}(X) = \boldsymbol{\phi}(X)^\top \widehat{\boldsymbol{\beta}}$ for some $\widehat{\boldsymbol{\beta}}$, and the regression function estimator $\widehat{\gamma}_{\text{aug}}(X)$ is written as $\widehat{\gamma}_{\text{aug}}(X) = \boldsymbol{\phi}(X)^\top \widehat{\boldsymbol{\rho}}$ for some $\widehat{\boldsymbol{\rho}}$, then we have

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} m(W_i, \phi_1) \\ m(W_i, \phi_2) \\ \vdots \\ m(W_i, \phi_p) \end{pmatrix}^\top \widehat{\boldsymbol{\rho}}^\dagger,$$

where

$$\begin{aligned} \widehat{\rho}_j^\dagger &:= \Big(1 - w_j\Big)\widehat{\rho}_{\text{aug},j} + w_j \widehat{\rho}_{\text{OLS},j}, \\ w_j &:= \frac{\frac{1}{n}\sum_{i=1}^{n}\Big(m(W_i, \phi_j) - \phi_j(X_i)\Big)}{\frac{1}{n}\sum_{i=1}^{n}\Big(\widehat{\alpha}(X_i)\phi_j(X_i) - \phi_j(X_i)\Big)}. \end{aligned}$$

Special case with inexact covariate balancing. Under some special cases, we can simplify the final estimator. For example, if we use an ℓ_2-penalty for both generalized Riesz regression and regression function estimation of γ_0, the resulting final AWR estimator becomes a ridge estimator of γ_0 (Bruns-Smith et al., 2025). For example, consider kernel ridge regression (KRR). Let K be the kernel matrix and let δ be the ridge regularization for the outcome fit. If we use ℓ_2-penalized generalized Riesz regression with regularization parameter λ, then in the resulting AWR estimator

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n}\Big(m(W_i, \widehat{\gamma}) + \widehat{\alpha}(X_i)\Big(Y_i - \widehat{\gamma}(X_i)\Big)\Big) = \frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} m(W_i, \phi_1) \\ m(W_i, \phi_2) \\ \vdots \\ m(W_i, \phi_p) \end{pmatrix}^\top \widehat{\boldsymbol{\rho}}^\dagger,$$

$\widehat{\boldsymbol{\rho}}^\dagger$ can be written as a single KRR estimator with an effective regularization parameter $\widetilde{\delta}$:

$$\widehat{\boldsymbol{\rho}}^\dagger = \Big(K + \widetilde{\delta} I\Big)^{-1}\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \qquad \widetilde{\delta} := \frac{\delta\lambda}{\sigma^2 + \delta + \lambda},$$

where σ^2 denotes the (simplifying) common eigenvalue scale used in the diagonalized representation. This equivalence implies that augmentation can be interpreted as a data-dependent undersmoothing rule for the outcome regression. Motivated by this result, Singh (2024) analyzes kernel balancing.

6.2 Automatic Neyman Error Minimization

Covariate balancing can also be interpreted as minimizing the estimation error of the Neyman orthogonal score. Given a nuisance parameter estimator $\widehat{\eta}$ and the true nuisance parameter η_0, the estimation error is given by

$$\begin{aligned}&\psi(W;\widehat{\eta},\theta) - \psi(W;\eta_0,\theta)\\ &= \Big(m(W,\widehat{\gamma}) + \widehat{\alpha}(X)\big(Y - \widehat{\gamma}(X)\big)\Big) - \Big(m(W,\gamma_0) + \alpha_0(X)\big(Y - \gamma_0(X)\big)\Big).\end{aligned}$$

Here, $\alpha_0(X)\big(Y - \gamma_0(X)\big)$ has mean zero and is ignorable in expectation. Therefore, we consider

$$\text{NeymanError} \coloneqq \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\alpha}(X_i)\big(Y_i - \widehat{\gamma}(X_i)\big) + m(W_i,\widehat{\gamma}) - m(W_i,\gamma_0)\right).$$

Suppose that we estimate α_0 by generalized Riesz regression using the basis function $\boldsymbol{\phi}(X)$, and that γ_0 belongs to the linear space spanned by $\boldsymbol{\phi}(X)$. Recall that we have

$$\sum_{j=1}^{p}\rho_{0,j}\frac{1}{n}\sum_{i=1}^{n}\Big(\widehat{\alpha}(X_i)\phi_j(X_i) - m(W_i,\phi_j)\Big) = 0.$$

If $\widehat{\gamma}$ equals γ_0 or $\widehat{\gamma}$ is estimated as $\boldsymbol{\phi}(X)^\top\boldsymbol{\rho}$, then by Neyman orthogonality, we have

$$\text{NeymanError} = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\alpha}(X_i)Y_i - m(W_i,\gamma_0)\right).$$

Furthermore, since $Y = \gamma_0(X) + \varepsilon$ for an error term ε such that $\mathbb{E}[\varepsilon \mid X] = 0$, we have

$$\text{NeymanError} = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\alpha}(X_i)\gamma_0(X_i) - m(W_i,\gamma_0)\right) + U,$$

where U is a term whose expectation is zero. Here, we again have

$$\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\alpha}(X_i)\gamma_0(X_i) - m(W_i,\gamma_0)\right) = 0$$

from the automatic covariate balancing property. Thus, automatic covariate balancing and automatic Neyman orthogonalization also imply minimization of the Neyman error.

This property is discussed in Zhao (2019) for the case where UKL-Riesz regression with a basis function $\boldsymbol{\phi}\colon \mathcal{Z} \to \mathbb{R}^p$ is used in ATE estimation. Our result generalizes the finding in that work to more general models, losses, and parameters of interest.

Cross fitting and the Neyman error If we use cross fitting, the exact automatic covariate balancing property may not hold. On the other hand, exact automatic covariate balancing typically requires the Donsker condition to obtain an efficient estimator of the parameter of interest. We need to carefully address this trade-off.

Basis functions depending only on Z In ATE estimation with standard propensity score modeling, we typically use basis functions depending on Z, that is, $\boldsymbol{\phi}\colon \mathcal{Z} \to \mathbb{R}^p$. Under this choice, $m(W, \phi_j(\cdot)) = 0$. Therefore, automatic covariate balancing guarantees only $\frac{1}{n}\sum_{i=1}^n (\widehat{\alpha}(X_i)\gamma_0(X_i)) = 0$. This result eliminates part of the Neyman error only when the conditional ATE given x is invariant across x, that is, when the treatment effect is homogeneous. From the regression-function point of view, this requires that $\gamma_0(X)$ lies in the linear space spanned by $\boldsymbol{\phi}(Z)$. If we aim to guarantee automatic Neyman orthogonalization under treatment effect heterogeneity, we should use $\boldsymbol{\phi}\colon \mathcal{X} \to \mathbb{R}^p$, even though the Riesz representer corresponds to inverse propensity scores $1/e_0(Z)$ and $1/(1-e_0(Z))$.

6.3 Comparison with TMLE

TMLE is another promising approach in debiased machine learning (van der Laan, 2006). TMLE adds a perturbation to an initial estimate of γ_0. TMLE works to eliminate the sample average of the $(\star)$ part in the following Neyman error:

$$\text{NeymanError} \coloneqq \frac{1}{n}\sum_{i=1}^n \left(\underbrace{\widehat{\alpha}(X_i)(Y_i - \widehat{\gamma}(X_i))}_{=(\star)} + m(W_i, \widehat{\gamma}) - m(W_i, \gamma_0) \right).$$

In contrast, generalized Riesz regression works to eliminate the sample average of the $(\star\star)$ part in the following Neyman error:

$$\text{NeymanError} \coloneqq \frac{1}{n}\sum_{i=1}^n \left(\widehat{\alpha}(X_i)(Y_i - \underbrace{\widehat{\gamma}(X_i)) + m(W_i, \widehat{\gamma})}_{=(\star\star)} - m(W_i, \gamma_0) \right).$$

That is, TMLE and generalized Riesz regression differ as follows:

- TMLE puts the difficulty of efficient estimation of θ_0 into the estimation of γ_0 by eliminating the influence of $\widehat{\alpha}$.
- Generalized Riesz regression puts the difficulty of efficient estimation of θ_0 into the estimation of α_0 by eliminating the influence of $\widehat{\gamma}$.

Remark (TMLE)**.** *We introduce TMLE in more detail. Let $\widehat{\gamma}^{(0)}$ be an initial estimate of γ_0. Then, given estimates $\widehat{\gamma}^{(0)}$ and $\widehat{\alpha}$, TMLE updates the regression function as*

$$\widehat{\gamma}^{(1)}(x) \coloneqq \widehat{\gamma}^{(0)}(x) + \frac{\sum_{i=1}^n \widehat{\alpha}(X_i)\big(Y_i - \widehat{\gamma}^{(0)}(X_i)\big)}{\sum_{i=1}^n \widehat{\alpha}(X_i)^2}\widehat{\alpha}(x).$$

It then estimates the parameter of interest as

$$\widehat{\theta}^{TMLE} := \frac{1}{n}\sum_{i=1}^{n} m(W_i, \widehat{\gamma}^{(1)}).$$

Note that this update is derived as the solution in ϵ of

$$\sum_{i=1}^{n} \widehat{\alpha}(X_i)\left(Y_i - \left(\widehat{\gamma}^{(0)}(X_i) + \epsilon\widehat{\alpha}(X_i)\right)\right) = 0,$$

which is given by

$$\widehat{\epsilon} := \frac{\sum_{i=1}^{n} \widehat{\alpha}(X_i)\Big(Y_i - \widehat{\gamma}^{(0)}(X_i)\Big)}{\sum_{i=1}^{n} \widehat{\alpha}(X_i)^2}.$$

Then, we derive $\widehat{\gamma}^{(1)}(x)$ as

$$\widehat{\gamma}^{(1)}(x) = \widehat{\gamma}^{(0)}(x) + \widehat{\epsilon}\widehat{\alpha}(x).$$

Under this update, the sample average of $(\star)$ in the Neyman error becomes zero.

6.4 Modeling of Regression Function and Riesz Representer

We have focused on estimating the Riesz representer α_0. However, the regression function $\gamma_0(x) = \mathbb{E}[Y \mid X = x]$ remains the other essential nuisance. We highlight the relationship between (i) fitting α_0 directly and (ii) fitting γ_0 in a way that implicitly stabilizes, or even eliminates, the need for separate Riesz representer estimation. For simplicity, in this section, we focus on RKHS-based modeling of regression and Riesz representer functions and discuss how they are related in the Riesz Weighted (RW) estimator, and the Augmented WR (AWR) estimator. As a result, we confirm that regression function modeling and Riesz representer estimation are complementary.

Bias in the AWR estimator Given estimates $\widehat{\gamma}$ and $\widehat{\alpha}$, the AWR estimator is given as

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n}\Big(m(W_i, \widehat{\gamma}) + \widehat{\alpha}(X_i)\Big(Y_i - \widehat{\gamma}(X_i)\Big)\Big).$$

A standard second-order expansion implies that the leading remainder term is governed by the product

$$\|\widehat{\gamma} - \gamma_0\|_2 \cdot \|\widehat{\alpha} - \alpha_0\|_2,$$

under cross-fitting or Donsker-type conditions. Consequently, the estimator can be efficient even when each nuisance converges slower than $n^{-1/2}$, provided

$$\|\widehat{\gamma} - \gamma_0\|_2 \cdot \|\widehat{\alpha} - \alpha_0\|_2 = o_p(n^{-1/2}).$$

This decomposition suggests two extremes: (i) Riesz-representer-centric strategies that target accurate $\widehat{\alpha}$, and (ii) regression-function-centric strategies that make $\widehat{\gamma}$ sufficiently accurate for the functional of interest, often via undersmoothing.

Regression-function-centric view Our final estimation target is not the regression function γ_0, but $\theta_0 = \mathbb{E}[m(W, \gamma_0)]$. Therefore, a recurring theme is that prediction-optimal regularization may be too aggressive for inference on θ_0, and one often needs undersmoothing (smaller regularization) to reduce functional bias.

Representer-centric view Complementing regression-function-centric undersmoothing, Singh (2024) studies kernel ridge estimation of the Riesz representer (Kernel Ridge Riesz Regression; KRRR) and analyzes its generalization error in population L_2. Singh (2024) shows that Riesz representer estimation error can be characterized by the counterfactual effective dimension, which is controlled by the usual effective dimension under a semiparametric continuity condition.

A second message in Singh (2024) is robustness under misspecification in orthogonal constructions. For an orthogonal estimator built from a regression function approximation γ_G and an Riesz representer approximation α_A, a double-robust type statement persists: if either $\gamma_G = \gamma_0$ or $\alpha_A = \alpha_0$, then the resulting debiased estimator is consistent, up to stochastic terms controlled by the convergence rates of $\widehat{\gamma}$ and $\widehat{\alpha}$.

Remark (Minimax rate and definition of common support)**.** *Mou et al. (2023) provides a complementary minimax viewpoint for estimating weighted linear functionals from observational data, including regimes where strict overlap fails and semiparametric efficiency bounds may be infinite. Two aspects are especially relevant for RR-based debiasing. First, the functional difficulty is governed by a modulus of continuity. Second, for RKHS classes, Mou et al. (2023) shows that this lower bound can be achieved up to constants by computationally simple outcome regression estimators that do not require knowledge of the behavioral policy π. This underscores that the geometry of the function class and the induced Riesz representer govern the attainable risk.*

7 Choice of Basis, Link, and Loss Functions

We have shown that generalized Riesz regression includes a broad class of objective functions (loss functions) for Riesz representer estimation. This section summarizes and discusses how we select basis functions, link functions, loss functions, and the final estimator of the parameter of interest. These elements should be chosen based on the following perspectives:

- **Loss functions** should be chosen based on the estimand of the final estimation (the parameter of interest) and the data sensitivity of Riesz representer estimation.
- **Link functions** should be compatible with the loss functions to preserve the automatic covariate balancing property (Section 4).
- **Basis functions** should be chosen based on the relationship between the Riesz representer and outcome models. If outcome models lie in the linear span of the basis functions, we can automatically obtain Neyman orthogonality of the Riesz weighted estimator (Section 6).

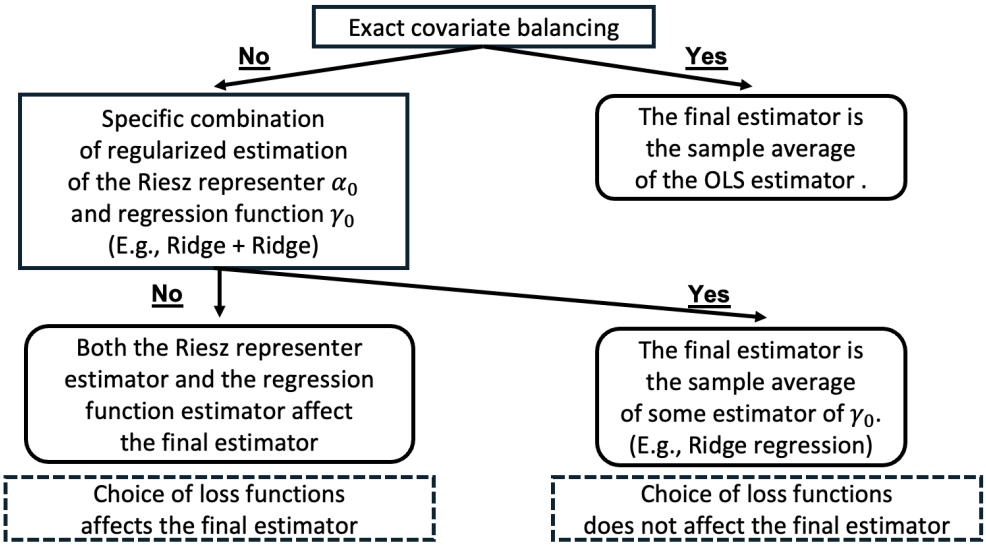

Figure 2: The case where the choice of loss function in generalized Riesz regression affects the final estimator of θ_0.

- **Final estimators** There are three main choices of final estimators: the Riesz weighted (RW) estimator, the augmented Riesz weighted (ARW) estimator, and TMLE. Under exact balancing, the RW and ARW estimators are equal. Under inexact balancing, they behave differently.

They are closely related and cannot be discussed separately.

Inexact balancing and specific combinations of regularization As discussed above and shown in Bruns-Smith et al. (2025), in some situations, the choice of loss function in generalized Riesz regression does not affect the final estimator. If exact balancing holds, the final estimator is the sample average of the OLS estimator of γ_0 under any loss function used in generalized Riesz regression. Therefore, in this case, there is no need to carefully choose the loss function for generalized Riesz regression. Moreover, under specific combinations of Riesz representer estimation for α_0 and regression function estimation for γ_0, the resulting final estimator can be simplified. For example, if we use an ℓ_2-penalty (ridge) for both estimators, the final estimator is the sample average of the ridge estimator of γ_0 (Singh, 2024). In other cases, the choice of loss function affects the final estimator (Figure 2).

Automatic covariate balancing determines the choice of loss and link functions. First, from the viewpoint of constructing a Neyman orthogonal final estimator, we aim to use the automatic covariate balancing property. As discussed in Section 4, the combination of loss and link functions is determined accordingly.

Loss-link pair is determined from the sensitivity viewpoint. Second, the choice of the loss-link pair is determined by sensitivity to the data. For example, in ATE estimation, the link function affects the sensitivity of the Riesz representer estimator to the data as follows:

- SQ-Riesz + linear link. This choice makes Riesz representer estimation robust to outliers.
- UKL-Riesz + log link. This choice introduces model specification for the Riesz representer. Riesz representer estimation becomes accurate if the model specification is correct. However, since the model includes an exponential function, it is easily affected by outliers.
- BP-Riesz + power link. This choice is intermediate between the above choices.

See also Menon & Ong (2016) and Zellinger (2025) for related discussions in density ratio estimation.

Regularization and final estimator. If we do not use cross-fitting, $\lambda = 0$, and exact balancing is feasible, the RW estimator and the ARW estimator are equivalent. If $\lambda > 0$, the ARW estimator and RW estimator differ. From the viewpoint of Neyman orthogonality, we should use the ARW estimator or the TMLE estimator.

ARW estimator and TMLE estimator. The ARW estimator shifts the difficulty of semiparametric inference to Riesz representer estimation, while the TMLE estimator shifts the difficulty of semiparametric inference to regression function estimation (Section 6.3).

Choice of basis functions. Ideally, as discussed in Section 6, the regression function γ_0 lies in the linear span of $\boldsymbol{\phi}(X)$. Under certain conditions, we can mitigate the overlapping assumption if we are only interested in the minimax rate, as discussed in Section 6.4.

Remark (Proper scoring rule based on Beta family)**.** *As discussed in Zhao (2019), if we restrict the loss functions to the Beta family, the parameter of interest corresponding to BKL-Riesz regression is the Optimally Weighted ATE (OWATE), not the ATE, which is defined as*

$$\theta_0^{OWATE} := \mathbb{E}\Big[e_0(Z)\Big(1 - e_0(Z)\Big)(Y(1) - Y(0))\Big].$$

This argument assumes sigmoid-function-based propensity modeling, that is, the log link function. If we use a more complicated link function, we can still attain covariate balancing (Remark 4.3). Therefore, if we carefully choose a link-function and loss-function pair, we can rebut the claim by Zhao (2019). However, even if covariate balancing can be attained, such a choice is not practical, so we do not discuss this approach in detail.

Remark (Choice of loss functions and exact balancing)**.** *If we do not use cross-fitting and exact covariate balancing is feasible, the choice of loss function does not affect the final estimator of the parameter of interest. Moreover, as discussed in Section 6, the final estimator becomes equivalent to the OLS estimator.*

8 Convergence Rate Analysis

This section provides an estimation error analysis for generalized Riesz regression. We model the Riesz representer α_0 by

$$\alpha_f(X) = \zeta^{-1}\Big(X, f(X)\Big),$$

where ζ is a continuously differentiable and globally Lipschitz link function, and f is a base model. Note that unlike Section 4, we do not restrict f to be a linear model. For example, in addition to linear models $\boldsymbol{\phi}(X)^\top \boldsymbol{\beta}$, we can use random forests, neural networks, and other models for f. In this section, we consider the case where we use RKHS methods and neural networks for f.

Throughout this section, we assume that the Riesz representer is bounded.

Assumption 8.1. *There exists a constant $C > 0$ independent of n such that $|\alpha(x)| < C$ for all $x \in \mathcal{X}$.*

This boundedness assumption holds in the standard ATE setting, which assumes common support of the treated and control groups and bounded outcomes. In many other applications, this assumption also holds. If we wish to allow unbounded support, we can develop such an extension by imposing appropriate tail conditions. For example, the density ratio between two Normal distributions may violate this assumption. In such cases, Zheng et al. (2022) presents a convergence rate analysis, and we can follow their approach. In practical data analysis, it is often reasonable to treat the Riesz representer as bounded.

8.1 RKHS

First, we study the case with RKHS regression. Let $\mathcal{F}^{\mathrm{RKHS}}$ be a class of RKHS functions, and define

$$\widehat{f}^{\mathrm{RKHS}} \coloneqq \operatorname*{arg\,min}_{f \in \mathcal{F}^{\mathrm{RKHS}}} \widehat{\mathrm{BD}}_g(\alpha_f) + \lambda \|f\|_{\mathcal{F}}^2,$$

where $\|\cdot\|_{\mathcal{F}}^2$ is the RKHS norm. Then, we define an estimator as

$$\widehat{\alpha}^{\mathrm{RKHS}}(x) \coloneqq \alpha_{\widehat{f}^{\mathrm{RKHS}}}(x) \coloneqq \zeta^{-1}\left(x, \widehat{f}^{\mathrm{RKHS}}(x)\right).$$

We analyze the estimation error by employing the results in Kanamori et al. (2012), which studies RKHS based LSIF for DRE. We define the following localized class of RKHS functions as a technical device: $\mathcal{F}_M^{\mathrm{RKHS}} \coloneqq \left\{f \in \mathcal{F}^{\mathrm{RKHS}} \colon I(f) \leq M\right\}$ for some norm $I(f)$ of f. We also define $\mathcal{H}^{\mathrm{RKHS}} \coloneqq \left\{\zeta^{-1}(\cdot, f(\cdot)) \colon f \in \mathcal{F}^{\mathrm{RKHS}}\right\}$. We then impose the following assumption on this localized class.

Assumption 8.2. *There exist constants $0 < \tau < 2$, $0 \leq \boldsymbol{\beta} \leq 1$, $c_0 > 0$, and $A > 0$ such that for all $M \geq 1$, it holds that $H_B(\delta, \mathcal{F}_M^{RKHS}, P_0) \leq A\left(\frac{M}{\delta}\right)^\tau$, where $H_B(\delta, \mathcal{F}_M^{RKHS}, P_0)$ is the bracketing entropy with radius $\delta > 0$ for the function class $\mathcal{F}_M^{RKHS}$ and the distribution P_0.*

For details on bracketing entropy, see Appendix F and Definition 2.2 in van de Geer (2000).

Under these preparations, we establish an estimation error bound.

Theorem 8.1 (L_2-norm estimation error bound)**.** *Suppose that g is μ-strongly convex and there exists a constant $C > 0$ such that $|g''(t)| \leq C \quad \forall t \in \mathbb{R}$. Assume also that $\zeta^{-1}(0)$ is finite. Suppose that Assumptions 8.1 and 8.2 hold. Set the regularization parameter $\lambda = \lambda_n$ so that $\lim_{n\to\infty} \lambda_n = 0$ and $\lambda_n^{-1} = O(n^{1-\delta})$ $(n \to \infty)$. If $\alpha_0 \in \mathcal{H}^{RKHS}$, then we have*

$$\left\|\widehat{\alpha}^{RKHS}(X) - \alpha_0(X)\right\|_{L_2(P_0)}^2 = O_{P_0}\left(\lambda^{1/2}\right).$$

The proof is provided in Appendix F, following the approach of Kanamori et al. (2012). The parameter τ is determined by the function class to which f_0 belongs.

8.2 Neural Networks

Second, we provide an estimation error analysis when we use neural networks for $\mathcal{H}$. Our analysis is mostly based on Kato & Teshima (2021) and Zheng et al. (2022). We define Feedforward neural networks (FNNs) as follows:

Definition 8.1 (FNNs. From Zheng et al. (2022))**.** *Let $\mathcal{D}$, $\mathcal{W}$, $\mathcal{U}$, and $\mathcal{S} \in (0, \infty)$ be parameters that can depend on n. Let $\mathcal{F}^{FNN} := \mathcal{F}^{FNN}_{M,\mathcal{D},\mathcal{W},\mathcal{U},\mathcal{S}}$ be a class of ReLU activated FNNs with parameter $\boldsymbol{\beta}$, depth $\mathcal{D}$, width $\mathcal{W}$, size $\mathcal{S}$, and number of neurons $\mathcal{U}$, and satisfying the following conditions: (i) the number of hidden layers is $\mathcal{D}$; (ii) the maximum width of the hidden layers is $\mathcal{W}$; (iii) the number of neurons in $e_{\boldsymbol{\beta}}$ is $\mathcal{U}$; (iv) the total number of parameters in $e_{\boldsymbol{\beta}}$ is $\mathcal{S}$.*

For the model $\mathcal{F}^{\mathrm{FNN}}$, we define $\widehat{f}^{\mathrm{FNN}} := \arg\min_{f\in\mathcal{F}^{\mathrm{FNN}}} \widehat{\mathrm{BD}}_g(\alpha_f)$. Then, we define an estimator as

$$\widehat{\alpha}^{\mathrm{FNN}}(x) := \zeta^{-1}\left(x, \widehat{f}^{\mathrm{FNN}}(x)\right).$$

For this estimator, we can prove an estimation error bound. We make the following assumption.

Assumption 8.3. *There exists a constant $0 < M < \infty$ such that $\|f_0\|_\infty < M$, and $\|f\|_\infty \leq M$ for any $f \in \mathcal{F}^{FNN}$.*

Let $\mathrm{Pdim}(\mathcal{F}^{\mathrm{FNN}})$ be the pseudodimension of $\mathcal{F}^{\mathrm{FNN}}$. For the definition, see Anthony & Bartlett (1999) and Definition 3 in Zheng et al. (2022). Then, we prove the following estimation error bound:

Theorem 8.2 (Estimation error bound for neural networks)**.** *Suppose that g is μ-strongly convex and there exists a constant $C > 0$ such that $|g''(t)| \leq C \quad \forall t \in \mathbb{R}$. Assume also that $\zeta^{-1}(0)$ is finite. Suppose that Assumption 8.3 holds. For f_0 such that*

$$\alpha_0(x) = \zeta^{-1}(x, f_0(x)),$$

also assume $f_0 \in \Sigma(\nu, M, [0,1]^d)$ with $\nu = k + a$, where $k \in \mathbb{N}^+$ and $a \in (0, 1]$, and $\mathcal{F}^{FNN}$ has width $\mathcal{W}$ and depth $\mathcal{D}$ such that $\mathcal{W} = 38\big(\lfloor\nu\rfloor + 1\big)^2 d^{\lfloor\nu\rfloor+1}$ and $\mathcal{D} = 21\big(\lfloor\nu\rfloor + 1\big)^2 \lceil n^{\frac{d}{2(d+2\nu)}} \log_2\left(8n^{\frac{d}{2(d+2\nu)}}\right)\rceil$. Then, for $M \geq 1$ and $n \leq Pdim(\mathcal{F}^{FNN})$, it holds that

$$\left\|\widehat{\alpha}^{FNN}(X) - \alpha_0(X)\right\|_{L_2(P_0)}^2 = C_0\big(\lfloor\nu\rfloor + 1\big)^9 d^{2\lfloor\nu\rfloor+(\nu\wedge 3)} n^{-\frac{2\nu}{d+2\nu}} \log^3 n,$$

where $C_0 > 0$ is a constant independent of n.

The proof is provided in Appendix G, following the approach of Zheng et al. (2022). This result directly implies the minimax optimality of the proposed method when f_0 belongs to a Hölder class.

8.3 Construction of an Efficient Estimator

This section provides how we construct an efficient estimator for the parameter of interest θ_0 using generalized Riesz regression. As we discussed in Section 2, we construct an estimator $\widehat{\theta}$ of θ_0 as

$$\frac{1}{n}\sum_{i=1}^{n}\psi\left(W_i;\widehat{\eta},\widehat{\theta}\right)=0,$$

where $\psi(W;\eta,\theta)$ is the Neyman orthogonal score is defined as

$$\psi(W;\eta,\theta)\coloneqq m(W,\gamma)+\alpha(X)\big(Y-\gamma(X)\big)-\theta$$

for $\eta=(\alpha,\gamma)$ $(\alpha,\gamma\colon\mathcal{X}\to\mathbb{R})$. As introduced in Section 2, we refer to this estimator as the AIPW estimator. We prove that under certain conditions, the proposed estimator is asymptotically normal.

We first make the following assumption.

Assumption 8.4 (Donsker condition or cross fitting)**.** *Either of the followings holds: (i) the hypothesis classes $\mathcal{H}$ and $\mathcal{M}$ belong to the Donsker class, or (ii) $\widehat{\gamma}$ and $\widehat{\alpha}$ are estimated via cross fitting.*

For example, the Donsker condition holds when the bracketing entropy of $\mathcal{H}$ is finite. In contrast, it is violated in high-dimensional regression or series regression settings where the model complexity diverges as $n\to\infty$. For neural networks, the assumption holds if both the number of layers and the width are finite. However, if these quantities grow with the sample size, the assumption is no longer valid.

Even if the Donsker condition does not hold, we can still establish asymptotic normality by employing sample splitting (Klaassen, 1987). There are various ways to implement sample splitting, and one of the most well-known is cross fitting, used in debiased machine learning (Chernozhukov et al., 2018). In debiased machine learning, the dataset is split into several folds, and the nuisance parameters are estimated using only a subset of the folds. This ensures that in $m\big(W_i,\widehat{\gamma}\big)+\widehat{\alpha}(X_i)\left(Y_i-\widehat{\gamma}(X_i)\right)$, the observations (X_i,Y_i) are not used to construct $\widehat{\gamma}$ and $\widehat{\alpha}$. For more details, see Chernozhukov et al. (2018).

Assumption 8.5 (Convergence rate)**.** $\left\|\widehat{\alpha}-\alpha_0\right\|_2=o_p(1)$, $\left\|\widehat{\gamma}-\gamma_0\right\|_2=o_p(1)$, *and* $\left\|\widehat{\alpha}-\alpha_0\right\|_2\left\|\widehat{\gamma}-\gamma_0\right\|_2=o_p(1/\sqrt{n})$.

Under these assumptions, we show the asymptotic normality of $\widehat{\theta}$. We omit the proof. For details, see Chernozhukov et al. (2018) or Schuler & van der Laan (2024), for example.

Theorem 8.3 (Asymptotic normality)**.** *Suppose that Assumptions 8.1, and 8.4–8.5 hold. Then, the AIPW estimator converges in distribution to a normal distribution as*

$$\sqrt{n}\Big(\widehat{\theta}-\theta_0\Big)\xrightarrow{\mathrm{d}}\mathcal{N}(0,V^*),$$

Table 2: Experimental results using the synthetic dataset. We report the mean squared error (MSE) of the ATE estimator and the empirical coverage ratio (CR) of nominal 95% Wald-type confidence intervals over 100 Monte Carlo replications. The column block "True" uses oracle nuisance functions (the true propensity score and true outcome regressions) and is therefore infeasible. "SQ-Riesz" and "UKL-Riesz" estimate the ATE Riesz representer by generalized Riesz regression with the squared-loss and unnormalized-KL objectives, respectively; "BKL-Riesz = MLE" corresponds to estimating the propensity score by BKL objective (logistic MLE) and plugging it into the ATE Riesz representer. For SQ-Riesz we compare two link specifications (Linear and Logit). For UKL-Riesz we compare two feature sets: $\boldsymbol{\phi}(Z)$ uses only covariates Z, while $\boldsymbol{\phi}(X)$ uses the full regressor $X = (D, Z)$ (allowing treatment-dependent features). For each Riesz-representer fit we report three estimators: the direct method (DM), inverse probability weighting (IPW), and augmented IPW (AIPW). DM depends only on the outcome regression; small differences across DM columns arise from recomputing the outcome regression within each run of the AIPW construction. Values of CR close to 0.95 indicate well-calibrated uncertainty quantification.

	True			SQ-Riesz (Linear)			SQ-Riesz (Logit)			UKL-Riesz ($\boldsymbol{\phi}(Z)$)			UKL-Riesz ($\boldsymbol{\phi}(X)$)			BKL-Riesz (MLE)		
	DM	IPW	AIPW	DM	IPW	AIPW	DM	IPW	AIPW	DM	IPW	AIPW	DM	IPW	AIPW	DM	IPW	AIPW
MSE	0.00	1.44	0.01	0.39	0.49	0.19	0.39	1.38	0.08	0.38	1.50	0.10	0.40	1.52	0.10	0.39	3.79	0.23
CR	1.00	0.84	0.98	0.06	0.98	0.87	0.12	0.80	0.89	0.08	0.73	0.77	0.06	0.68	0.81	0.06	0.32	0.60

where V^* is the efficiency bound defined as

$$V^* := \mathbb{E}\left[\psi(W; \eta_0, \theta_0)^2\right].$$

.

Here, V^* matches the efficiency bound given as the variance of the efficient influence function (van der Vaart, 1998; Hahn, 1998). Thus, this estimator is efficient.

9 Experiments

We evaluate generalized Riesz regression as a building block for debiased machine learning, focusing on average treatment effect (ATE) estimation. Across all experiments, we compare three ways of estimating the ATE Riesz representer (bias-correction term) introduced in Section 3: SQ-Riesz (squared-loss objective), UKL-Riesz (unnormalized-KL objective), and BKL-Riesz (binary-KL objective). In the ATE setting, BKL-Riesz coincides with estimating the propensity score by Bernoulli likelihood (logistic MLE) and then plugging it into the closed-form ATE Riesz representer; we therefore refer to it as "BKL-Riesz = MLE."

Given an estimate of the outcome regression $\widehat{\gamma}$ and an estimate of the Riesz representer $\widehat{\alpha}$, we report three ATE estimators:

- **DM**: the plug-in direct method based only on $\widehat{\gamma}$,
- **IPW**: the IPW estimator based only on $\widehat{\alpha}$,

- **AIPW**: the Neyman-orthogonal (doubly robust) estimator combining $\widehat{\gamma}$ and $\widehat{\alpha}$ as in Section 2.

We quantify accuracy by the mean squared error (MSE) of the ATE estimate and quantify uncertainty by the empirical coverage ratio (CR) of nominal 95

9.1 Experiments with synthetic dataset

Design The covariates are three-dimensional, $K = 3$, and we fix the sample size at $n = 3000$. In each Monte Carlo replication, we generate covariates $Z_i \in \mathbb{R}^3$ from a multivariate normal distribution $\mathcal{N}(0, I_3)$ and construct a nonlinear propensity score model with polynomial and interaction terms as

$$e_0(Z_i) = \frac{1}{1 + \exp\big(-h(Z_i)\big)},$$

where

$$h(Z_i) = \sum_{j=1}^{3} a_j Z_{i,j} + \sum_{j=1}^{3} b_j Z_{i,j}^2 + c_1 Z_{i,1} Z_{i,2} + c_2 Z_{i,2} Z_{i,3} + c_3 Z_{i,1} Z_{i,3}.$$

The coefficients a_j, b_j, and c_j are independently drawn from $\mathcal{N}(0, 0.5)$. Given these propensity scores, the treatment assignment D_i is sampled accordingly. We then generate the outcome as

$$Y_i = 1.0 + \left(\sum_{j=1}^{3} Z_{i,j} \widetilde{a}_j\right)^2 + 1/\left(1 + \exp\left(-\left(\sum_{j=1}^{3} Z_{i,j}^2 \widetilde{b}_j\right)\right)\right) + \tau_0 D_i + \varepsilon_i,$$

where $\varepsilon_i \sim \mathcal{N}(0, 1)$ and $\tau_0 = 5.0$.

Estimators and implementation We estimate the Riesz representer using the following variants, matched to Table 2.

- **SQ-Riesz (Linear)** and **SQ-Riesz (Logit)**: squared-loss generalized Riesz regression with two different link specifications for the Riesz-representer model.
- **UKL-Riesz ($\boldsymbol{\phi}(Z)$)** and **UKL-Riesz ($\boldsymbol{\phi}(X)$)**: UKL generalized Riesz regression with a log-type link, comparing two feature sets. Here $X = (D, Z)$ and $\boldsymbol{\phi}(Z)$ uses only Z, while $\boldsymbol{\phi}(X)$ uses the full regressor (allowing treatment-dependent features).
- **BKL-Riesz (MLE)**: propensity-score MLE (Bernoulli likelihood) followed by plugging $\widehat{e}(Z)$ into the ATE Riesz representer.

For the Riesz representer and regression models, we separately use a neural network with one hidden layer consisting of 100 nodes. To avoid relying on the Donsker condition, we estimate all nuisance functions using two-fold cross fitting. In each replication, we split the sample into two folds, estimate the nuisance functions on one fold, evaluate the corresponding

scores on the other fold, and then swap the roles of the folds. The final estimators aggregate the two cross-fitted scores.

This experiment does not guarantee automatic Neyman orthogonalization, since we use cross fitting and do not use the same basis functions for outcome modeling. However, this implementation is standard in debiased machine learning; therefore, we adopt it.

We repeat the simulation 100 times. The "True" columns in Table 2 report infeasible oracle performance using the true nuisance functions.

Results Table 2 highlights three robust patterns. First, oracle baselines separate estimation error from intrinsic variance. The oracle AIPW estimator is close to the efficiency benchmark (MSE = 0.01) and achieves near-nominal coverage (CR = 0.98). In contrast, even with the true propensity score, oracle IPW remains noisy (MSE = 1.44) and undercovers (CR = 0.84), reflecting the well-known finite-sample instability of pure weighting in challenging overlap regimes.

Second, the plug-in DM estimator is not reliable for inference in this design. Across feasible implementations, DM has moderate MSE (about 0.38–0.40) but extremely poor coverage (CR = 0.06–0.12). This indicates that the outcome regression learner, while not catastrophically inaccurate in MSE, does not deliver a reliable uncertainty estimate when used without orthogonalization, and the resulting Wald intervals are severely miscalibrated.

Third, how we fit the Riesz representer matters substantially for IPW, and AIPW mitigates (but does not eliminate) this sensitivity. For IPW, SQ-Riesz (Linear) is the best-performing option in Table 2 (MSE = 0.49) and yields near-nominal coverage (CR = 0.98). In contrast, IPW based on UKL-Riesz has larger MSE (about 1.50) and noticeably lower coverage (CR = 0.68–0.73), while BKL-Riesz (= MLE) performs worst (MSE = 3.79, CR = 0.32), consistent with propensity-score MLE producing more extreme effective weights in this design.

The AIPW estimator is uniformly more stable than IPW and DM in terms of MSE, but calibration still depends on the Riesz-representer fit. SQ-Riesz (Logit) attains the best AIPW MSE (0.08) with CR 0.89. UKL-Riesz achieves similarly small AIPW MSE (0.10) but exhibits undercoverage (CR = 0.77–0.81). BKL-Riesz (= MLE) improves substantially over its IPW counterpart (AIPW MSE = 0.23), yet its coverage remains poor (CR = 0.60). Overall, directly fitting the Riesz representer via generalized Riesz regression can materially improve finite-sample performance relative to the MLE plug-in baseline, and the combination of objective and link specification plays a first-order role, especially for IPW and for the calibration of AIPW intervals.

9.2 Experiments with semi synthetic datasets

We next evaluate the same family of estimators on the semi-synthetic IHDP benchmark, following Chernozhukov et al. (2022a). We use the standard setting "A" in the `npci` package and generate 1000 replications. Each replication contains $n = 747$ observations with a binary treatment, an outcome, and $p = 25$ covariates. The estimand is the ATE.

We report DM, IPW, and AIPW for each Riesz-representer estimator (SQ-Riesz, UKL-Riesz, and BKL-Riesz (= MLE)). We consider two nuisance-learner families:

Table 3: Experimental results using the semi-synthetic IHDP dataset. We report the mean squared error (MSE) and the empirical coverage ratio (CR) of nominal 95% confidence intervals over 1000 replications for the direct method (DM), inverse probability weighting (IPW), and augmented IPW (AIPW) estimators. Nuisance functions are estimated either by a neural network with one hidden layer of size 100 or by an RKHS regression with 100 Gaussian basis functions. The columns correspond to SQ-Riesz regression (SQ-Riesz), UKL-Riesz regression (UKL-Riesz), and BKL-Riesz regression (BKL-Riesz (MLE)). BKL-Riesz (MLE) implies BKL-Riesz regression is essentially equivalent to MLE of logistic models for the propensity score.

	Neural network								
	SQ-Riesz			UKL-Riesz			BKL-Riesz (MLE)		
	DM	IPW	AIPW	DM	IPW	AIPW	DM	IPW	AIPW
MSE	1.52	6.82	0.31	1.55	2.84	0.32	1.58	3.00	0.43
CR	0.03	0.41	1.00	0.03	0.73	0.94	0.01	0.61	0.90
	RKHS								
	SQ-Riesz			UKL-Riesz			BKL-Riesz (MLE)		
	DM	IPW	AIPW	DM	IPW	AIPW	DM	IPW	AIPW
MSE	19.98	3.56	19.97	2.59	1.78	4.45	2.48	1.22	2.32
CR	0.00	0.00	0.00	0.48	0.93	0.88	0.39	0.81	0.84

- a feedforward neural network with one hidden layer of 100 units,
- an RKHS learner with 100 Gaussian basis functions (with tuning by cross validation).

For each configuration, we compute the MSE of the ATE estimate and the empirical coverage ratio (CR) of nominal 95% Wald-type confidence intervals across the 1000 replications; CR close to 0.95 indicates well-calibrated uncertainty quantification. Results appear in Table 3.

Two findings stand out. With neural networks, AIPW is consistently accurate (MSE around 0.31–0.43) and well calibrated for UKL-Riesz and BKL-Riesz (CR = 0.94 and 0.90), while SQ-Riesz yields overly conservative AIPW intervals (CR = 1.00). In contrast, DM has very low coverage (CR near zero) and IPW exhibits large error, especially for SQ-Riesz (MSE = 6.82), reinforcing that orthogonalization is essential in this benchmark.

With RKHS, performance becomes much more sensitive to the particular objective: SQ-Riesz deteriorates sharply (MSE around 20 with CR = 0 for both DM and AIPW), whereas UKL-Riesz and BKL-Riesz remain substantially more stable. In particular, UKL-Riesz attains strong IPW calibration under RKHS (CR = 0.93) with comparatively low MSE (1.78), while BKL-Riesz provides a competitive alternative (IPW MSE = 1.22 with CR = 0.81). These results underscore that, in finite samples, the interaction between the Riesz-representer objective and the nuisance-function learner can be decisive, and that UKL-type objectives can offer noticeably more robust behavior than squared-loss fitting in this semi-synthetic setting.

10 Conclusion

This paper develops a unified perspective on estimating the Riesz representer, namely, the bias correction term that appears in Neyman orthogonal scores for a broad class of causal and structural parameters. We formulate Riesz representer estimation as fitting a model to the unknown representer under a Bregman divergence, which yields an empirical risk minimization objective that depends only on observed data. This generalized Riesz regression recovers Riesz regression and least squares importance fitting under squared loss, it recovers KL based tailored loss minimization and its dual entropy balancing weights under a KL type loss, and it connects logistic likelihood based propensity modeling with classification based density ratio estimation through a binary KL criterion. By pairing the loss with an appropriate link function, we make explicit a dual characterization that delivers automatic covariate balancing or moment matching, which clarifies when popular balancing schemes arise as primal or dual solutions. We provide convergence rate results for kernel methods and neural networks, including minimax optimality under standard smoothness classes, and we show how the framework instantiates in ATE, AME, APE, and covariate shift adaptation. Our experiments suggest that directly estimating the bias correction term can be competitive with common propensity score based baselines and can be stable across divergence choices when combined with cross fitting. Overall, the proposed framework bridges density ratio estimation and causal inference, and it offers a single set of tools for designing, analyzing, and implementing Riesz representer estimators, while motivating extensions such as nearest neighbor and score matching based constructions.

References

Alberto Abadie and Guido W. Imbens. Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267, 2006. 85

Masahiro Abe and Masashi Sugiyama. Anomaly detection by deep direct density ratio estimation, 2019. openreview. 8, 67

Daniel Ackerberg, Xiaohong Chen, Jinyong Hahn, and Zhipeng Liao. Asymptotic efficiency of semiparametric two-step gmm. *The Review of Economic Studies*, 81(3):919–943, 2014. 64

Chunrong Ai and Xiaohong Chen. The semiparametric efficiency bound for models of sequential moment restrictions containing unknown functions. *Journal of Econometrics*, 170 (2):442–457, 2012. 64

Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999. 46

Susan Athey, Guido W. Imbens, and Stefan Wager. Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 80(4):597–623, 2018. 9

Heejung Bang and James M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005. 5

Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497 – 1537, 2005. 70

Ayanendranath Basu, Ian R. Harris, Nils L. Hjort, and M. C. Jones. Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85(3):549–559, 1998. 14, 25, 34

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference for high-dimensional sparse econometric models, 2011. 64

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, 28 (2):29–50, 2014. 64

Alexandre Belloni, Victor Chernozhukov, and Ying Wei. Post-selection inference for generalized linear models with many controls. *Journal of Business & Economic Statistics*, 34 (4):606–619, 2016. 64

Eli Ben-Michael, Avi Feller, David A. Hirshberg, and José R. Zubizarreta. The balancing act in causal inference, 2021. arXiv: 2110.14831. 18, 66

P.J. Bickel, C.A.J. Klaassen, Y. Ritov, and J.A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins series in the mathematical sciences. Springer New York, 1998. 64

L. M. Bregman. 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, 1967. ISSN 0041-5553. 2

David Bruns-Smith, Oliver Dukes, Avi Feller, and Elizabeth L Ogburn. Augmented balancing weights as linear regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 04 2025. 2, 9, 18, 22, 36, 38, 43, 66

David A. Bruns-Smith and Avi Feller. Outcome assumptions and duality theory for balancing weights. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 11037–11055, 2022. 18, 38, 66

Kwun Chuen Gary Chan, Sheung Chi Phillip Yam, and Zheng Zhang. Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. *Journal of the Royal Statistical Society*, 2016. 66

Wei Chen, Shigui Li, Jiacheng Li, Junmei Yang, John Paisley, and Delu Zeng. Dequantified diffusion-schrödinger bridge for density ratio estimation. In *International Conference on Machine Learning (ICML)*, 2025. 88, 89

Xiaohong Chen and Zhipeng Liao. Sieve m inference on irregular parameters. *Journal of Econometrics*, 182(1):70–86, 2014. 65

Xiaohong Chen and Zhipeng Liao. Sieve semiparametric two-step gmm under weak dependence. *Journal of Econometrics*, 189(1):163–186, 2015. 4, 9, 65, 90, 95

Xiaohong Chen and Demian Pouzo. Sieve wald and qlr inferences on semi/nonparametric conditional moment models. *Econometrica*, 83(3):1013–1079, 2015. 9, 65, 90, 95

Xiaohong Chen, Han Hong, and Alessandro Tarozzi. Semiparametric efficiency in GMM models with auxiliary data. *The Annals of Statistics*, 36(2):808 – 843, 2008. 64

Xiaohong Chen, Zhipeng Liao, and Yixiao Sun. Sieve inference on possibly misspecified semi-nonparametric time series models. *Journal of Econometrics*, 178:639–658, 2014. 2, 65

kuang-Fu Cheng and C.K. Chu. Semiparametric density estimation under a two-sample density ratio model. *Bernoulli*, 10, 2004. 10, 30, 34, 66

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 2018. 2, 5, 47, 64

Victor Chernozhukov, Whitney K. Newey, Victor Quintas-Martinez, and Vasilis Syrgkanis. Automatic debiased machine learning via riesz regression, 2021. arXiv:2104.14737. 2, 9, 10, 12, 26, 31, 66

Victor Chernozhukov, Whitney Newey, Víctor M Quintas-Martínez, and Vasilis Syrgkanis. RieszNet and ForestRiesz: Automatic debiased machine learning with neural nets and random forests. In *International Conference on Machine Learning (ICML)*, 2022a. 9, 50, 66

Victor Chernozhukov, Whitney K. Newey, and Rahul Singh. Automatic debiased machine learning of causal and structural effects. *Econometrica*, 90(3):967–1027, 2022b. 1, 4, 5, 6, 64, 65

Victor Chernozhukov, Whitney K Newey, and Rahul Singh. Debiased machine learning of global and local parameters using regularized riesz representers. *The Econometrics Journal*, 25(3):576–601, 04 2022c. 5

Victor Chernozhukov, Christian Hansen, Nathan Kallus, Martin Spindler, and Vasilis Syrgkanis. *Applied Causal Inference Powered by ML and AI*. CausalML-book.org, 2024. URL https://arxiv.org/abs/2403.02467. arXiv:2403.02467. 5

Victor Chernozhukov, Michael Newey, Whitney K Newey, Rahul Singh, and Vasilis Srygkanis. Automatic debiased machine learning for covariate shifts, 2025. arXiv: 2307.04527. 2, 8, 33

Kristy Choi, Chenlin Meng, Yang Song, and Stefano Ermon. Density ratio estimation via infinitesimal classification. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022. 69, 88, 89

Marthinus Christoffel du Plessis, Gang. Niu, and Masashi Sugiyama. Convex formulation for learning from positive and unlabeled data. In *International Conference on Machine Learning (ICML)*, 2015. 9, 16

Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. In *International Conference on Knowledge Discovery and Data Mining (KDD)*, 2008. 16

Tongtong Fang, Nan Lu, Gang Niu, and Masashi Sugiyama. Rethinking importance weighting for deep learning under distribution shift. In *NeurIPS*, 2020. 67

A. Gretton, A. J. Smola, J. Huang, Marcel Schmittfull, K. M. Borgwardt, and B. Schölkopf. Covariate shift by kernel mean matching. *Dataset Shift in Machine Learning, 131-160 (2009)*, 01 2009. 9, 10, 66

Jinyong Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66(2):315–331, 1998. 48

Jens Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20(1):25–46, 2012. 2, 6, 9, 13, 23, 28, 66

Jaroslav Hájek. A characterization of limiting distributions of regular estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 14(4):323–330, Dec 1970. 64

Chad Hazlett. Kernel balancing: a flexible non-parametric weighting procedure for estimating causal effects. *Statistica Sinica*, 30:1155–1189, 2020. 66

Shohei Hido, Yuta Tsuboi, Hisashi Kashima, Masashi Sugiyama, and Takafumi Kanamori. Inlier-based outlier detection via direct density ratio estimation. In *ICDM*, 2008. 8, 67

Rei Higuchi and Taiji Suzuki. Direct density ratio optimization: A statistically consistent approach to aligning large language models, 2025. arXiv: 2505.07558. 67

Daniel G. Horvitz and Donovan J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260): 663–685, 1952. 6, 35

Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex J. Smola. Correcting sample selection bias by unlabeled data. In *NeurIPS*, pp. 601–608. MIT Press, 2007. 10, 66

I. A. Ibragimov and R. Z. Khas'minskii. *Statistical Estimation: Asymptotic Theory*, volume 255 of *Grundlehren der mathematischen Wissenschaften*. Springer-Verlag, New York, 1981. doi: 10.1007/978-1-4899-0027-2. 64

Kosuke Imai and Marc Ratkovic. Covariate balancing propensity score. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1):243–263, 07 2013a. ISSN 1369-7412. 66

Kosuke Imai and Marc Ratkovic. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443 – 470, 2013b. 6, 66

Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015. 1

Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 10(Jul.):1391–1445, 2009. 2, 3, 9, 10, 12, 26, 33, 66, 81

Takafumi Kanamori, Taiji Suzuki, and Masashi Sugiyama. f -divergence estimation and two-sample homogeneity test under semiparametric density-ratio models. *IEEE Transactions on Information Theory*, 58, 2010. 67

Takafumi Kanamori, Taiji Suzuki, and Masashi Sugiyama. Statistical analysis of kernel-based least-squares density-ratio estimation. *Mach. Learn.*, 86(3):335–367, March 2012. ISSN 0885-6125. 9, 45, 46, 73, 74

Masahiro Kato. Nearest neighbor matching as least squares density ratio estimation and riesz regression, 2025a. arXiv: 2510.24433. 3, 12, 26, 82

Masahiro Kato. Riesz regression as direct density ratio estimation, 2025b. arXiv: 2511.04568. 8, 80

Masahiro Kato. Scorematchingriesz: Auto-dml with infinitesimal classification, 2025c. arXiv: 2512.20523. 3, 86, 87, 89

Masahiro Kato and Takeshi Teshima. Non-negative bregman divergence minimization for deep direct density ratio estimation. In *International Conference on Machine Learning (ICML)*, 2021. 8, 9, 16, 46, 66, 69, 71, 73, 77

Masahiro Kato, Takeshi Teshima, and Junya Honda. Learning from positive and unlabeled data with a selection bias. In *International Conference on Learning Representations (ICLR)*, 2019. 8, 16, 66

Masahiro Kato, Masaaki Imaizumi, and Kentaro Minami. Unified perspective on probability divergence via the density-ratio likelihood: Bridging kl-divergence and integral probability metrics. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 5271–5298, 2023. 13

Masahiro Kato, Kota Matsui, and Ryo Inokuchi. Double debiased covariate shift adaptation robust to density-ratio estimation, 2024a. arXiv: 2310.16638. 8, 33, 66

Masahiro Kato, Akihiro Oga, Wataru Komatsubara, and Ryo Inokuchi. Active adaptive experimental design for treatment effect estimation with covariate choice. In *International Conference on Machine Learning (ICML)*, 2024b. 67

Masahiro Kato, Fumiaki Kozai, and Ryo Inokuchi. Puate: Semiparametric efficient average treatment effect estimation from treated (positive) and unlabeled units, 2025. arXiv:2501.19345. 16, 67

Yoshinobu Kawahara and Masashi Sugiyama. Change-point detection in time-series data by direct density-ratio estimation. In *International Conference on Data Mining (ICDM)*, 2009. 8, 67

Amor Keziou and Samuela Leoni-Aubin. Test of homogeneity in semiparametric two-sample density ratio models. *Comptes Rendus Mathematique*, 340:905–910, 06 2005. 8, 67

Ryuichi Kiryo, Gang Niu, Marthinus Christoffel du Plessis, and Masashi Sugiyama. Positive-unlabeled learning with non-negative risk estimator. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 9, 66, 69

Chris A. J. Klaassen. Consistent estimation of the influence function of locally asymptotically linear estimators. *Annals of Statistics*, 15, 1987. 47, 64

Tony Lancaster and Guido Imbens. Case-control studies with contaminated controls. *Journal of Econometrics*, 71(1):145–160, 1996. 16

L Le Cam. Limits of experiments. In *Theory of Statistics*, pp. 245–282. University of California Press, 1972. 64

Lucien Le Cam. *Asymptotic Methods in Statistical Decision Theory (Springer Series in Statistics)*. Springer, 1986. 64

Kaitlyn J. Lee and Alejandro Schuler. Rieszboost: Gradient boosting for riesz regression, 2025. arXiv: 2501.04871. 9, 66

B. Ya. Levit. On the efficiency of a class of non-parametric estimates. *Theory of Probability & Its Applications*, 20(4):723–740, 1976. 64

Fan Li, Kari Lock Morgan, and Alan M. Zaslavsky. Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521), 2017. 66

Zhexiao Lin, Peng Ding, and Fang Han. Estimation based on nearest neighbor matching: from density ratio to average treatment effect. *Econometrica*, 91(6):2187–2217, 2023. 3, 9, 12, 66, 82, 83, 84, 85

Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 447–461, 2016. 8, 67

Aditya Menon and Cheng Soon Ong. Linking losses for density ratio and class-probability estimation. In *International Conference on Machine Learning (ICML)*, 2016. 44

Wenlong Mou, Peng Ding, Martin J. Wainwright, and Peter L. Bartlett. Kernel-based off-policy estimation without overlap: Instance optimality beyond semiparametric efficiency, 2023. arXiv: 2301.06240. 42

Hyunha Nam and Masashi Sugiyama. Direct density ratio estimation with convolutional neural networks with application in outlier detection. *IEICE Transactions on Information and Systems*, E98.D(5):1073–1079, 2015. 8

Whitney K. Newey. The asymptotic variance of semiparametric estimators. *Econometrica*, 62(6), 1994. 2, 64

XuanLong Nguyen, Martin Wainwright, and Michael Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE*, 2010. 9, 10, 66

Shunichiro Orihara, Tomotaka Momozaki, and Tomoyuki Nakagawa. Generalized bayesian inference for causal effects using the covariate balancing procedure. *Biometrical Journal*, 67(6):e70085, 2025. 18

J. Pfanzagl and W. Wefelmeyer. *Contributions to a General Asymptotic Statistical Theory*. Lecture notes in statistics. Springer-Verlag, 1982. 64

Jing Qin. Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika*, 85(3):619–630, 1998. 9, 10, 30, 34, 66

B. Rhodes, K. Xu, and M.U. Gutmann. Telescoping density-ratio estimation. In *NeurIPS*, 2020. 9, 66, 69

James Robins, Mariela Sued, Quanhong Lei-Gomez, and Andrea Rotnitzky. Comment: Performance of Double-Robust Estimators When "Inverse Probability" Weights Are Highly Variable. *Statistical Science*, 22(4):544 – 559, 2007. 37

P. M. Robinson. Root-n-consistent semiparametric regression. *Econometrica*, 56(4):931–954, 1988. 64

Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983. 66

Alejandro Schuler and Mark van der Laan. Introduction to modern causal inference, 2024. URL https://alejandroschuler.github.io/mci/introduction-to-modern-causal-inference.html. 4, 5, 6, 47

Alejandro Schuler, Michael Baiocchi, Robert Tibshirani, and Nigam Shah. A comparison of methods for model selection when estimating individual treatment effects, 2018. 64

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000. 2, 35, 66

Bernard Silverman. Density ratios, empirical likelihood and cot death. *Applied Statistics*, 27, 1978. 13

Rahul Singh. Kernel ridge riesz representers: Generalization, mis-specification, and the counterfactual effective dimension, 2024. arXiv: 2102.11076. 9, 39, 42, 43

Alex Smola, Le Song, and Choon Hui Teo. Relative novelty detection. In *AISTATS*, 2009. 8, 67

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021. 3, 86, 87

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution, 2020. arXiv: 1907.05600. 86

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8 (35):985–1005, 2007. URL http://jmlr.org/papers/v8/sugiyama07a.html. 18

Masashi Sugiyama, Taiji Suzuki, Shinichi Nakajima, Hisashi Kashima, Paul von Bünau, and Motoaki Kawanabe. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60(4):699–746, 2008. 8, 9, 27, 33

Masashi Sugiyama, Taiji Suzuki, Yuta Itoh, Takafumi Kanamori, and Manabu Kimura. Least-squares two-sample test. *Neural networks*, 24:735–51, 04 2011a. 8, 67

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. Density ratio matching under the bregman divergence: A unified framework of density ratio estimation. *Annals of the Institute of Statistical Mathematics*, 64, 10 2011b. 3, 9, 18, 67

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density Ratio Estimation in Machine Learning*. Cambridge University Press, 2012. 2, 6, 8, 15, 25, 34

Zhiqiang Tan. Regularized calibrated estimation of propensity scores with model misspecification and high-dimensional data. *Biometrika*, 107(1):137–158, 2019. 9, 18

Alexander Tarr and Kosuke Imai. Estimating average treatment effects with support vector machines. *Statistics in Medicine*, 44(5), 2025. 9, 18, 66

Riku Togashi, Masahiro Kato, Mayu Otani, and Shin'ichi Satoh. Density-ratio based personalised ranking from implicit feedback. In *The World Wide Web Conference*, 2021. 67

A. Tsiatis. *Semiparametric Theory and Missing Data*. Springer Series in Statistics. Springer New York, 2007. 5

Masatoshi Uehara, Masahiro Kato, and Shota Yasui. Off-policy evaluation and learning for external validity under a covariate shift. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020. 2, 66, 67

Sara van de Geer. *Empirical Processes in M-Estimation*, volume 6. Cambridge University Press, 2000. 45, 70, 73

van der Laan. Targeted maximum likelihood learning, 2006. U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 213. https://biostats.bepress.com/ucbbiostat/paper213/. 4, 6, 40, 64

Mark J. van der Laan and Sherri Rose. *Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies*. Springer Publishing Company, Incorporated, 1st edition, 2018. ISBN 3319653032. 64

Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998. 48, 64

Aad W. van der Vaart. Semiparametric statistics, 2002. URL https://sites.stat.washington.edu/jaw/COURSES/EPWG/stflour.pdf. 5, 64

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018. 9, 85

Raymond K W Wong and Kwun Chuen Gary Chan. Kernel-based covariate functional balancing for observational studies. *Biometrika*, 105(1):199–213, 12 2017. 36, 66

Jeffrey M. Wooldridge. Asymptotic properties of weighted m-estimation for standard stratified samples. *Econometric Theory*, 2001. 16

Werner Zellinger. Binary losses for density ratio estimation. In *International Conference on Learning Representations (ICLR)*, 2025. 44

Qingyuan Zhao. Covariate balancing propensity score by tailored loss functions. *The Annals of Statistics*, 47(2):965 – 993, 2019. 2, 9, 10, 18, 20, 26, 27, 28, 36, 39, 44, 66, 95, 97, 98

Siming Zheng, Guohao Shen, Yuling Jiao, Yuanyuan Lin, and Jian Huang. An error analysis of deep density-ratio estimation with bregman divergence, 2022. URL https://openreview.net/forum?id=dfOBSd3tF9p. 45, 46, 47, 77, 78

José R. Zubizarreta. Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511):910–922, 2015. 2, 9, 18, 22, 66

Contents

A Related Work

A.1 Asymptotically Efficient Estimation

Early history. The construction of asymptotically efficient estimators is a classical problem in statistics, machine learning, economics, epidemiology, and related fields. In this problem, we consider semiparametric models with a low-dimensional parameter of interest and additional nuisance parameters. Our interest lies in obtaining $\sqrt{n}$-consistent and asymptotically normal estimators of the parameter of interest. The difficulty stems from the estimation error of nuisance parameters, whose convergence rates are typically slower than, or at best comparable to, $\sqrt{n}$. Reducing the influence of nuisance estimation error has been investigated in many studies, including Levit (1976), Ibragimov & Khas'minskii (1981), Pfanzagl & Wefelmeyer (1982), Klaassen (1987), Robinson (1988), Newey (1994), van der Vaart (1998), Bickel et al. (1998), Ai & Chen (2012), and Chernozhukov et al. (2018).

In the construction of asymptotically efficient estimators, we aim to develop estimators that are $\sqrt{n}$-consistent and asymptotically normal, with asymptotic variances that attain the asymptotic efficiency bounds. Asymptotic efficiency bounds are called H'ajek–Le Cam efficiency bounds, or semiparametric efficiency bounds when we consider semiparametric models (Hájek, 1970; Le Cam, 1972, 1986). They share the same motivation as the Cram'er–Rao lower bound. While the Cram'er–Rao lower bound provides a lower bound for unbiased estimators, asymptotic efficiency bounds provide lower bounds for asymptotically unbiased estimators, called regular estimators. It is known that efficient estimators are regular and asymptotically linear (RAL) with the efficient influence function. Therefore, the construction of asymptotically efficient estimators is equivalent to the construction of RAL estimators (van der Vaart, 1998).

There are three main approaches to constructing efficient estimators, one-step bias correction, estimating equation methods, and TMLE (Schuler et al., 2018; van der Vaart, 2002; van der Laan, 2006; van der Laan & Rose, 2018). In many cases, these approaches yield estimators that are asymptotically equivalent. However, their finite sample behavior may differ. Another related line of work is post-selection inference with high-dimensional control variables (Belloni et al., 2011, 2014, 2016).

Debiased machine learning and Riesz representer. Debiased/double machine learning (DML) provides a general recipe for constructing asymptotically linear and semiparametrically efficient estimators by combining flexible first-step learning with Neyman-orthogonal scores (Chernozhukov et al., 2018). In classical semiparametric theory, such orthogonalization is naturally expressed through the efficient influence function (EIF), obtained by projecting the pathwise derivative onto the nuisance tangent space (Newey, 1994). Related influence-function/projection-based bias corrections also appear in earlier semiparametric two-step and sieve inference work (e.g., (Chen et al., 2008; Ackerberg et al., 2014)).

For many targets, the orthogonal score admits an augmentation (bias-correction) form. In particular, for linear (and local) functionals of a regression-type nuisance, Chernozhukov et al. (2022b) make explicit that one can write an orthogonal score as a plug-in term plus a correction that multiplies the regression residual by the functional's Riesz representer; they treat the Riesz representer itself as an additional nuisance parameter and propose to estimate it

from data via regularized *Riesz representer regression*, combined with cross-fitting, yielding adaptive inference for regular and nonregular local functionals (Chernozhukov et al., 2022b). This places the adjustment term, often called the one-step bias-correction term or the clever covariate, into the same estimation pipeline as other nuisances.

Closely related Riesz-representation-based characterizations and feasible approximations of this adjustment term have been developed in the sieve/semiparametric inference literature by Chen and coauthors. In semiparametric conditional moment restriction settings, Chen & Pouzo (2015) characterize the pathwise derivative of a target functional as a linear functional on an (infinite-dimensional) Hilbert space and show that a *population* Riesz representer exists if and only if this derivative is bounded; when it is unbounded (an irregular functional), the population representer may fail to exist (Chen & Pouzo, 2015). Importantly, on any finite-dimensional sieve space the derivative is automatically bounded, so a *sieve* Riesz representer is always well-defined; it can be used to construct implementable "sieve influence functions" and variance estimators (Chen & Pouzo, 2015). Building on this perspective, Chen et al. (2014) emphasize that even when the population Riesz representer is difficult to compute (or does not exist on the infinite-dimensional space), the sieve Riesz representer can always be computed explicitly, enabling a unified treatment of regular and irregular functionals (Chen et al., 2014). Moreover, Chen et al. (2014) relate regularity to the behavior of the sieve Riesz representer norm as sieve dimension increases, providing a convenient diagnostic of whether root-n inference is attainable (Chen et al., 2014). Finally, Chen & Liao (2015) show that while the population representer may not admit a closed-form solution, its sieve analogue often does and can be computed via finite-dimensional linear algebra (generalized inverse formulas), yielding practical influence-function-based inference and variance estimation procedures (Chen & Liao, 2015). For the relationship our generalized Riesz regression and series Riesz representer, see Appendix J.

From this viewpoint, the "Riesz representer regression" terminology of Chernozhukov et al. (2022b) can be interpreted as a modern, high-dimensional regularized analogue of the sieve Riesz representer constructions in Chen & Pouzo (2015); Chen et al. (2014); Chen & Liao (2014, 2015): both lines of work use Riesz representation to express and estimate the orthogonal-score adjustment term, but Chernozhukov et al. (2022b) focus on learning the representer in large ML dictionaries via regularization and cross-fitting, complementing the closed-form series/sieve calculations emphasized in the sieve literature (Chen et al., 2014; Chen & Liao, 2015).

A.2 ATE Estimation

Randomized controlled trials are the gold standard for causal inference, where treatments are assigned while maintaining balance between treatment groups. However, they are not always feasible, and we aim to estimate causal effects from observational data, where imbalance between treatment groups often arises. To correct this imbalance, propensity scores or balancing weights have been proposed.

In ATE estimation, the Riesz representer corresponds to inverse propensity weights. Accurate estimation of the propensity score is therefore central to ATE estimation. A standard choice is maximum likelihood estimation, but many alternative approaches have been studied. Riesz regression provides an end-to-end approach to estimating the Riesz representer

and can be applied to tasks beyond ATE estimation (Chernozhukov et al., 2021, 2022a; Lee & Schuler, 2025). Another promising approach is covariate balancing. The propensity score is also known as the coarsest balancing score (Rosenbaum & Rubin, 1983), and propensity score estimation via covariate balancing has been extensively studied (Li et al., 2017; Imai & Ratkovic, 2013a; Hainmueller, 2012; Zubizarreta, 2015; Tarr & Imai, 2025; Chan et al., 2016; Wong & Chan, 2017). As discussed in this study and related works (Bruns-Smith & Feller, 2022; Bruns-Smith et al., 2025; Ben-Michael et al., 2021; Zhao, 2019), Riesz regression and covariate balancing are dual to each other, in the sense that they correspond to essentially the same optimization problem.

Covariate balancing. Covariate balancing is a popular approach for propensity score or balancing weight estimation. The propensity score is a balancing score, and based on this property, existing studies propose estimating the propensity score or the weights so that weighted covariate moments match between treated and control groups. Imai & Ratkovic (2013b) proposes estimating the propensity score by matching first moments, and Hazlett (2020) extends this idea to higher-moment matching by mapping covariates into a high-dimensional space via basis functions. On the other hand, methods that do not directly specify a propensity score model have also been proposed. Such methods are called empirical balancing and include entropy balancing (Hainmueller, 2012) and stable weights (Zubizarreta, 2015). These two approaches may appear different, but Zhao (2019) and Bruns-Smith et al. (2025) show that they are essentially equivalent through a duality relationship.

A.3 Density Ratio Estimation

A parallel line of work is density ratio estimation, which has been extensively studied in machine learning. We refer to the ratio between two densities as the density ratio. The density ratio is a useful tool in semiparametric inference, as used in covariate shift adaptation (Shimodaira, 2000; Kato et al., 2024a), and we show that this framework can be generalized to a wider class of applications, including ATE estimation.

While the density ratio can be estimated by separately estimating each density, such an approach may magnify estimation errors by compounding the errors from two separate estimators. To address this issue, end-to-end, direct density ratio estimation methods have been studied, including moment matching (Huang et al., 2007; Gretton et al., 2009), probabilistic classification (Qin, 1998; Cheng & Chu, 2004), density matching (Nguyen et al., 2010), density ratio fitting (Kanamori et al., 2009), and PU learning (Kato et al., 2019). It is also known that when complicated models such as neural networks are used for this task, the loss function can diverge in finite samples (Kiryo et al., 2017). Therefore, density ratio estimation methods with neural networks have been investigated (Kato & Teshima, 2021; Rhodes et al., 2020).

As discussed in this study and in existing work such as Uehara et al. (2020) and Lin et al. (2023), density ratio estimation is closely related to propensity score estimation. In particular, this study shows that the formulations of Riesz regression and LSIF in density ratio estimation are essentially the same. While Riesz regression applies to more general problems, the LSIF literature provides a range of theoretical and empirical results. One important ex-

tension is to generalize LSIF via Bregman divergence minimization (Sugiyama et al., 2011b), and this study is strongly inspired by that work.

Note that density ratios are used not only for semiparametric analysis but also in tasks such as learning with noisy labels (Liu & Tao, 2016; Fang et al., 2020), anomaly detection (Smola et al., 2009; Hido et al., 2008; Abe & Sugiyama, 2019), two-sample testing (Keziou & Leoni-Aubin, 2005; Kanamori et al., 2010; Sugiyama et al., 2011a), change-point detection (Kawahara & Sugiyama, 2009), causal inference (Uehara et al., 2020), and recommendation systems (Togashi et al., 2021). In causal inference, Uehara et al. (2020) investigates efficient ATE estimation and policy learning under covariate shift. Kato et al. (2024b) applies this approach to adaptive experimental design, and Kato et al. (2025) extends the framework to a PU learning setup. Density ratio estimation is discussed from the viewpoint of large language models (LLMs) by Higuchi & Suzuki (2025).

B Proof of the Automatic Covariate Balancing Property

For simplicity, we only consider the case with ℓ_1-penalty.

B.1 Constrained Optimization Problem

From the Riesz representation theorem, the following equation holds:

$$\mathbb{E}\big[m(W,(\partial g)\circ\alpha)\big]=\mathbb{E}\big[\alpha_0(X)\partial g\big(\alpha(X)\big)\big].$$

Therefore, we can consider an algorithm that estimates α_0 so that its estimator $\widehat{\alpha}$ satisfies

$$\mathbb{E}\Big[m(W,(\partial g)\circ\widehat{\alpha})\Big]\approx\mathbb{E}\Big[\widehat{\alpha}(X)\partial g\big(\widehat{\alpha}(X)\big)\Big],$$

where we replace α_0 with $\widehat{\alpha}$. Then, it holds that

$$\begin{aligned}\min_{\boldsymbol{\alpha}\in\mathbb{R}^n}\quad & \frac{1}{n}\sum_{i=1}^n g(\alpha_i)\\ \text{subject to}\quad & \left|\frac{1}{n}\sum_{i=1}^n\Big(\alpha_i\partial g\big(\alpha(X_i)\big)-m(W_i,(\partial g)\circ\phi_j)\Big)\right|\le\lambda\qquad j=1,\dots,p.\end{aligned}$$

Linearity for the basis functions. Next, we consider the case where

$$\partial g\big(\alpha_{\boldsymbol{\beta}}(X_i)\big)=\sum_{j=1}^p\boldsymbol{\beta}_j\widetilde{g}\left(X_i,\phi_j(X_i)\right).$$

We consider the following constrained optimization problem:

$$\begin{aligned}\min_{\boldsymbol{\alpha}\in\mathbb{R}^n}\quad & \frac{1}{n}\sum_{i=1}^n g(\alpha_i)\\ \text{subject to}\quad & \left|\frac{1}{n}\sum_{i=1}^n\Big(\alpha_i\widetilde{g}\left(X_i,\phi_j(X_i)\right)-m(W_i,(\partial g)\circ\phi_j)\Big)\right|\le\lambda\qquad j=1,\dots,p.\end{aligned}$$

B.2 Dual Formulation

Using Lagrange multipliers $\boldsymbol{\beta}_j \in \mathbb{R}$ $(j = 1, 2, \ldots, p)$, the constrained optimization problem can be written as

$$\min_{\boldsymbol{\alpha}\in\mathcal{H}^n} \sup_{\boldsymbol{\beta}\in\mathbb{R}^p} \quad \left\{ \frac{1}{n}\sum_{i=1}^{n} g(\alpha_i) + \sum_{j=1}^{p} \boldsymbol{\beta}_j \left(\frac{1}{n}\sum_{i=1}^{n} \Big(m(W_i, (\partial g)\circ\phi_j) - \alpha_i \widetilde{g}\left(X_i, \phi_j(X_i)\right) \Big) - \operatorname{sign}(\boldsymbol{\beta}_j)\lambda \right) \right\}.$$

The dual problem of the above constrained problem is

$$\max_{\boldsymbol{\beta}\in\mathbb{R}^p} \inf_{\boldsymbol{\alpha}\in\mathcal{H}^n} \quad \left\{ \frac{1}{n}\sum_{i=1}^{n} g(\alpha_i) + \sum_{j=1}^{p} \left(\boldsymbol{\beta}_j \frac{1}{n}\sum_{i=1}^{n} \Big(m(W_i, (\partial g)\circ\phi_j) - \alpha_i \widetilde{g}\left(X_i, \phi_j(X_i)\right) \Big) - |\boldsymbol{\beta}_j|\lambda \right) \right\}.$$

Let $\alpha_{\boldsymbol{\beta}}(X_i) = \boldsymbol{\phi}(X_i)^\top \boldsymbol{\beta}$. Recall that the empirical Bregman divergence objective is given by

$$\widehat{\mathrm{BD}}_g(\alpha_{\boldsymbol{\beta}}) \coloneqq \frac{1}{n}\sum_{i=1}^{n} \Big(-g(\alpha_{\boldsymbol{\beta}}(X_i)) + \partial g(\alpha_{\boldsymbol{\beta}}(X_i))\alpha_{\boldsymbol{\beta}}(X_i) - m\big(W_i, \partial g(\alpha_{\boldsymbol{\beta}}(X_i))\big)\Big).$$

Let $\alpha_i = \alpha_{\boldsymbol{\beta}}(X_i)$. Then the objective can be written as

$$\max_{\boldsymbol{\beta}\in\mathbb{R}^p} \inf_{\boldsymbol{\alpha}\in\mathcal{H}^n} \quad \frac{1}{n}\sum_{i=1}^{n} \left(g(\alpha_i) + \sum_{j=1}^{p} \left(\boldsymbol{\beta}_j \frac{1}{n}\sum_{i=1}^{n} \Big(m(W_i, (\partial g)\circ\phi_j) - \alpha_i \widetilde{g}\left(X_i, \phi_j(X_i)\right) \Big) - |\boldsymbol{\beta}_j|\lambda \right) \right).$$

From $\partial g\big(\alpha_{\boldsymbol{\beta}}(X_i)\big) = \sum_{j=1}^{p} \boldsymbol{\beta}_j \widetilde{g}\left(X_i, \phi_j(X_i)\right)$, we have

$$\max_{\boldsymbol{\beta}\in\mathbb{R}^p} \inf_{\boldsymbol{\alpha}\in\mathcal{H}^n} \quad \left\{ \frac{1}{n}\sum_{i=1}^{n} \Big(g(\alpha_i) - \alpha_i \partial g\big(\alpha_{\boldsymbol{\beta}}(X_i)\big) + m\big(W_i, (\partial g)\circ\alpha_{\boldsymbol{\beta}}\big) \Big) + \lambda\|\boldsymbol{\beta}\|_1 \right\}.$$

Consider the problem

$$\inf_{\boldsymbol{\alpha}\in\mathcal{H}^n} \quad \left\{ \frac{1}{n}\sum_{i=1}^{n} \Big(g(\alpha_i) - \alpha_i \partial g\big(\alpha_{\boldsymbol{\beta}}(X_i)\big) + m\big(W_i, (\partial g)\circ\alpha_{\boldsymbol{\beta}}\big) \Big) + \lambda\|\boldsymbol{\beta}\|_1 \right\}.$$

Since g is twice differentiable and strictly convex for a given domain, the infimum is attained when

$$\alpha_i = \alpha_{\boldsymbol{\beta}}(X_i) = \boldsymbol{\phi}(X_i)^\top \boldsymbol{\beta}, \qquad i = 1, \ldots, n.$$

Substituting $\alpha_i = \alpha_{\boldsymbol{\beta}}(X_i) = \boldsymbol{\phi}(X_i)^\top \boldsymbol{\beta}$, we obtain

$$\max_{\boldsymbol{\beta}\in\mathbb{R}^p} \quad \left\{ \frac{1}{n}\sum_{i=1}^{n} \Big(g(\alpha_{\boldsymbol{\beta}}(X_i)) - \alpha_i \partial g\big(\alpha_{\boldsymbol{\beta}}(X_i)\big) + m\big(W_i, (\partial g)\circ\alpha_{\boldsymbol{\beta}}\big) \Big) - \lambda\|\boldsymbol{\beta}\|_1 \right\}.$$

This is equivalent to

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \quad \left\{ \frac{1}{n}\sum_{i=1}^{n} \Big(-g(\alpha_i) + \alpha_i \partial g\big(\alpha_{\boldsymbol{\beta}}(X_i)\big) - m\big(W_i, (\partial g)\circ\alpha_{\boldsymbol{\beta}}\big) \Big) + \lambda\|\boldsymbol{\beta}\|_1 \right\}.$$

C Overfitting Problems

Density ratio estimation often suffers from a characteristic form of overfitting. Kato & Teshima (2021) refers to this issue as *train-loss hacking* and shows that the empirical objective can be artificially reduced by inflating $r(X^{(\mathrm{nu})})$ at the training points. Rhodes et al. (2020) highlights a related mechanism: when p_{nu} and p_{de} are far apart, for example when $\mathrm{KL}(p_{\mathrm{nu}}\|p_{\mathrm{de}})$ is on the order of tens of nats, the estimation problem enters a large-gap regime that exacerbates overfitting. They refer to this phenomenon as the *density chasm*. Although the two papers emphasize different viewpoints, both point to the same underlying difficulty, finite samples provide weak control of the ratio in regions where the two distributions have little overlap.

Non-negative Bregman divergence Kato & Teshima (2021) proposes a modification of the Bregman divergence objective that isolates the problematic component and applies a non-negative correction under a mild boundedness condition on r_0. Specifically, choose $0 < C < 1/R$ with $R \coloneqq \sup r_0$. The population objective decomposes, up to an additive constant, into a non-negative term plus a bounded residual. At the sample level, the method replaces the non-negative component with its positive part $[\cdot]_+$. This yields an objective that curbs train-loss hacking while remaining within the Bregman-divergence framework (Kiryo et al., 2017; Kato & Teshima, 2021).

Telescoping density ratio estimation Rhodes et al. (2020) proposes *telescoping density ratio estimation*, which targets overfitting in large-gap regimes by introducing intermediate waymark distributions $p_0 = p_{\mathrm{nu}}, p_1, \ldots, p_m = p_{\mathrm{de}}$. The method estimates local ratios p_k/p_{k+1} and combines them through the identity

$$\frac{p_0(x)}{p_m(x)} = \prod_{k=0}^{m-1} \frac{p_k(x)}{p_{k+1}(x)}.$$

Each local ratio corresponds to a smaller distributional gap, which makes perfect classification harder and typically makes the ratio estimation problem more stable at finite sample sizes. As a result, telescoping can improve robustness and generalization in practice.

Telescoping density ratio estimation is also closely connected to score matching. When the number of intermediate ratios tends to infinity, the log density ratio can be expressed as an integral of time scores along a continuum of bridge distributions, and can be approximated by aggregating these score functions (Choi et al., 2022). Building on this idea, Choi et al. (2022) proposes density ratio estimation via infinitesimal classification. See Appendix I.4 for details.

D Preliminary for the Convergence Rate Analysis

This section introduces notions that are useful for the theoretical analysis.

D.1 Rademacher complexity

Let $\sigma_1, \ldots, \sigma_n$ be n independent Rademacher random variables; that is, independent random variables for which $P(\sigma_i = 1) = P(\sigma_i = -1) = 1/2$. Let us define

$$\mathfrak{R}_n f \coloneqq \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(W_i).$$

Additionally, given a class $\mathcal{F}$, we define

$$\mathfrak{R}_n \mathcal{F} \coloneqq \sup_{f \in \mathcal{F}} \mathfrak{R}_n f.$$

Then, we define the Rademacher average as $\mathbb{E}[\mathfrak{R}_n \mathcal{F}]$ and the empirical Rademacher average as $\mathbb{E}_\sigma[\mathfrak{R}_n \mathcal{F} \mid X_1, \ldots, X_n]$.

D.2 Local Rademacher complexity bound

Let $\mathcal{F}$ be a class of functions that map $\mathcal{X}$ into $[a, b]$. For $f \in \mathcal{F}$, let us define

$$Pf \coloneqq \mathbb{E}[f(W)],$$
$$P_n f \coloneqq \frac{1}{n} \sum_{i=1}^{n} f(W_i).$$

We introduce the following result about the Rademacher complexity.

Proposition D.1 (From Theorem 2.1 in Bartlett et al. (2005))**.** *Let $\mathcal{F}$ be a class of functions that map $\mathcal{X}$ into $[a, b]$. Assume that there is some $r > 0$ such that for every $f \in \mathcal{F}$, $Var(f(W)) \leq r$. Then, for every $z > 0$, with probability at least $1 - \exp(-z)$, it holds that*

$$\sup_{f \in \mathcal{F}} \left(Pf - P_n f \right) \leq \inf_{\alpha > 0} \left\{ 2(1+\alpha)\mathbb{E}[\mathfrak{R}_n f] + \sqrt{\frac{2rx}{n}} + (b-a)\left(\frac{1}{3} + \frac{1}{\alpha}\right)\frac{z}{n} \right\}.$$

D.3 Bracketing entropy

We define the bracketing entropy. For a more detailed definition, see Definition 2.2 in van de Geer (2000).

Definition D.1. *Bracketing entropy. Given a class of functions $\mathcal{F}$, the logarithm of the smallest number of balls in a norm $\|\cdot\|_{2,P}$ of radius $\delta > 0$ needed to cover $\mathcal{F}$ is called the δ-entropy with bracketing of $\mathcal{F}$ under the $L_2(P)$ metric, denoted by $H_B(\delta, \mathcal{F}, P)$.*

D.4 Talagrand's concentration inequality

We introduce Talagrand's lemma.

Proposition D.2 (Talagrand's Lemma)**.** *Let $\phi \colon \mathbb{R} \to \mathbb{R}$ be a Lipschitz continuous function with a Lipschitz constant $L > 0$. Then, it holds that*

$$\mathfrak{R}_n(\phi \circ \mathcal{F}) \leq L \mathfrak{R}_n(\mathcal{F}).$$

E Basic inequalities

E.1 Strong convexity

Lemma E.1 (L_2 distance bound from Lemma 4 in Kato & Teshima (2021))**.** *If* $\inf_{\alpha\in(-\infty),\infty} g''(\alpha) > 0$*, then there exists* $\mu > 0$ *such that for all* $\alpha \in \mathcal{H}$*,*

$$\|\alpha - \alpha_0\|_2^2 \leq \frac{2}{\mu}\Big(BD_g(\alpha) - BD_g(\alpha_0)\Big)$$

holds.

From the strong convexity and Lemma E.1, we have

$$\frac{\mu}{2}\|\widehat{\alpha} - \alpha_0\|_2^2 \leq \mathrm{BD}_g(\widehat{\alpha}) - \mathrm{BD}_g\big(\alpha_0\big).$$

Recall that we have defined an estimator $\widehat{r}$ as follows:

$$\widehat{\alpha} \coloneqq \arg\min_{\alpha\in\mathcal{H}} \widehat{\mathrm{BD}}_g(\alpha) + \lambda J(\alpha),$$

where $J\,(h)$ is some regularization term.

E.2 Preliminary

Proposition E.2. *The estimator* $\widehat{r}$ *satisfies the following inequality:*

$$\widehat{BD}_g(\widehat{\alpha}) + \lambda J(\widehat{\alpha}) \leq \widehat{BD}_g(\alpha^*) + \lambda J(\alpha^*),$$

where recall that

$$\widehat{BD}_g(\alpha) \coloneqq \frac{1}{n}\sum_{i=1}^{n}\Big(- g(\alpha(X_i)) + \partial g(\alpha(X_i))\alpha(X_i) - \partial g(\alpha(1, X_i)) - \partial g(\alpha(0, X_i))\Big).$$

Let $Z \in \mathcal{Z}$ be a random variable with a space $\mathcal{Z}$, and $\{Z_i\}_{i=1}^n$ be its realizations. For a function $f\colon \mathcal{Z} \to \mathbb{R}$ and X following P, let us denote the sample mean as

$$\widehat{\mathbb{E}}[f(Z)] \coloneqq \frac{1}{n}\sum_{i=1}^{n} f(Z_i).$$

We also denote $\widehat{\mathbb{E}}[f(Z)] - \mathbb{E}[f(Z)] = (\widehat{\mathbb{E}} - \mathbb{E})f(Z)$

E.3 Risk bound

Recall that

$$\widehat{\mathrm{BD}}_g(\alpha) = \frac{1}{n}\sum_{i=1}^{n}\Big(- g(\alpha(X_i)) + \partial g(\alpha(X_i))\alpha(X_i) - \partial g(\alpha(1, X_i)) - \partial g(\alpha(0, X_i))\Big).$$

Let us define

$$L(h, D, X) \coloneqq -g(\alpha(X)) + \partial g(\alpha(X))\alpha(X) - \partial g(\alpha(1, X)) - \partial g(\alpha(0, X)),$$

and we can write

$$\widehat{\mathrm{BD}}_g(\alpha) = \widehat{\mathbb{E}}\big[L(h, D, X)\big]$$

Then, from Proposition E.2, we have

$$\widehat{\mathbb{E}}\big[L(\alpha^*, D, X)\big] - \widehat{\mathbb{E}}\big[L(\widehat{\alpha}, D, X)\big] + \lambda J(\widehat{\alpha}) - \lambda J(\alpha^*) \geq 0.$$

Throughout the proof, we use the following basic inequalities that hold for $\widehat{\alpha}$.

Proposition E.3. *The estimator $\widehat{r}$ satisfies the following inequality:*

$$\begin{aligned}
&\frac{\mu}{2} \|\widehat{\alpha}(X) - \alpha_0(X)\|^2_{L_2(P_0)} \\
&\leq \left(\mathbb{E} - \widehat{\mathbb{E}}\right) [L(\widehat{\alpha}, D, X) - L(\alpha_0, D, X)] + \widehat{\mathbb{E}} [L(\alpha^*, D, X) - L(\alpha_0, D, X)] + \lambda J(r_0) - \lambda J(\widehat{r}).
\end{aligned}$$

Proof of Proposition E.2 is trivial. We prove Proposition E.3 below.

Proof. From the strong convexity and Lemma E.1, we have

$$\frac{\mu}{2}\|\widehat{\alpha} - \alpha_0\|_2^2 \leq \mathrm{BD}_g(\widehat{\alpha}) - \mathrm{BD}_g\big(\alpha_0\big) = \mathbb{E}\left[L(\widehat{\alpha}, D, X) - L(\alpha_0, D, X)\right].$$

From Proposition E.2, we have

$$\begin{aligned}
&\frac{\mu}{2} \|\widehat{\alpha}(X) - \alpha_0(X)\|^2_{L_2(P_0)} \\
&\leq \mathbb{E}\left[L(\widehat{\alpha}, D, X) - L(\alpha_0, D, X)\right] \\
&= \mathbb{E}\left[L(\widehat{\alpha}, D, X) - L(\alpha_0, D, X)\right] \\
&\qquad - \widehat{\mathbb{E}}\left[L(\widehat{\alpha}, D, X) - L(\alpha_0, D, X)\right] \\
&\qquad + \widehat{\mathbb{E}}\left[L(\widehat{\alpha}, D, X) - L(\alpha_0, D, X)\right] \\
&\leq \mathbb{E}\left[L(\widehat{\alpha}, D, X) - L(\alpha_0, D, X)\right] \\
&\qquad - \widehat{\mathbb{E}}\left[L(\widehat{\alpha}, D, X) - L(\alpha_0, D, X)\right] \\
&\qquad + \widehat{\mathbb{E}}\left[L(\widehat{\alpha}, D, X) - L(\alpha_0, D, X)\right] \\
&\qquad - \widehat{\mathbb{E}}\left[L(\widehat{\alpha}, D, X) - L(\alpha^*, D, X)\right] + \lambda J(\widehat{\alpha}) - \lambda J(\alpha_0).
\end{aligned}$$

□

F Proof of Theorem 8.1

We show Theorem 8.1 by bounding

$$\left(\mathbb{E} - \widehat{\mathbb{E}}\right) \left[L(\widehat{\alpha}, D, X) - L(\alpha_0, D, X)\right], \tag{8}$$

in Proposition E.3. We can bound this term by using the empirical-process arguments.

Note that since $\alpha_0 \in \mathcal{H}$, it holds that $\alpha^* = \alpha_0$, which implies that

F.1 Preliminary

We introduce the following propositions from van de Geer (2000), Kanamori et al. (2012) and Kato & Teshima (2021).

Definition F.1 (Derived function class and bracketing entropy (from Definition 4 in Kato & Teshima (2021))). *Given a real-valued function class $\mathcal{F}$, define $\ell \circ \mathcal{F} \coloneqq \{\ell \circ f \colon f \in \mathcal{F}\}$. By extension, we define $I : \ell \circ \mathcal{H} \to [1, \infty)$ by $I(\ell \circ h) = I(\alpha)$ and $\ell \circ \mathcal{H}_M \coloneqq \{\ell \circ \alpha \colon \alpha \in \mathcal{H}_M\}$. Note that, as a result, $\ell \circ \mathcal{H}_M$ coincides with $\{\ell \circ \alpha \in \ell \circ \mathcal{H} : I(\ell \circ h) \leq M\}$.*

Proposition F.1. *Let $\ell : \mathbb{R} \to \mathbb{R}$ be a v-Lipschitz continuous function. Let $H_B\big(\delta, \mathcal{F}, \|\cdot\|_{L_2(P_0)}\big)$ denote the bracketing entropy of $\mathcal{F}$ with respect to a distribution P. Then, for any distribution P, any $\gamma > 0$, any $M \geq 1$, and any $\delta > 0$, we have*

$$H_B\big(\delta, \ell \circ \mathcal{H}, \|\cdot\|_{L_2(P_0)}\big) \leq \frac{(s+1)(2v)^\gamma}{\gamma} \left(\frac{M}{\delta}\right)^\gamma.$$

Moreover, there exists $M > 0$ such that for any $M \geq 1$ and any distribution P,

$$\sup_{\ell\circ\alpha \in \ell\circ\mathcal{H}_M} \|\ell \circ \alpha - \ell \circ \alpha^*\|_{L_2(P_0)} \leq c_0 v M,$$
$$\sup_{\substack{\ell\circ\alpha \in \ell\circ\mathcal{H}_M \\ \|\ell\circ\alpha - \ell\circ\alpha^*\|_{L_2(P_0)} \leq \delta}} \|\ell \circ \alpha - \ell \circ \alpha^*\|_\infty \leq c_0 v M, \quad \text{for all } \delta > 0.$$

Proposition F.2 (Lemma 5.13 in van de Geer (2000), Proposition 1 in Kanamori et al. (2012)). *Let $\mathcal{F} \subset L^2(P)$ be a function class and the map $I(f)$ be a complexity measure of $f \in \mathcal{F}$, where I is a non-negative function on $\mathcal{F}$ and $I(f_0) < \infty$ for a fixed $f_0 \in \mathcal{F}$. We now define $\mathcal{F}_M = \{f \in \mathcal{F} : I(f) \leq M\}$ satisfying $\mathcal{F} = \bigcup_{M \geq 1} \mathcal{F}_M$. Suppose that there exist $c_0 > 0$ and $0 < \gamma < 2$ such that*

$$\sup_{f \in \mathcal{F}_M} \|f - f_0\| \leq c_0 M, \qquad \sup_{\substack{f \in \mathcal{F}_M \\ \|f - f_0\|_{L^2(P)} \leq \delta}} \|f - f_0\|_\infty \leq c_0 M, \quad \text{for all } \delta > 0,$$

and that $H_B(\delta, \mathcal{F}_M, P) = O\left((M/\delta)^\gamma\right)$. Then, we have

$$\sup_{f \in \mathcal{F}} \frac{\left|\int (f - f_0) d(P - P_n)\right|}{D(f)} = O_p(1), \ (n \to \infty),$$

where $D(f)$ is defined by

$$D(f) = \max \frac{\|f - f_0\|_{L^2(P)}^{1-\gamma/2} I(f)^{\gamma/2}}{\sqrt{n}} \frac{I(f)}{n^{2/(2+\gamma)}}.$$

Proposition F.3. *Let $g \colon \mathcal{K} \to \mathbb{R}$ be twice continuously differentiable and strictly convex for the space $\mathcal{K}$ of α_0, and suppose that there exists $M > 0$ such that*

$$|g''(t)| \leq M \quad \text{for all } t \in \mathbb{R}.$$

Let $\zeta^{-1}\colon \mathbb{R} \to \mathbb{R}$ be continuously differentiable and globally Lipschitz, that is, there exists $L_\zeta > 0$ such that

$$|\zeta^{-1}(s) - \zeta^{-1}(t)| \le L_\zeta |s - t| \quad \text{for all } s, t \in \mathbb{R}.$$

Assume also that $\zeta^{-1}(0)$ is finite, and define

$$a_0 \coloneqq |\zeta^{-1}(0)|, \qquad a_1 \coloneqq L_\zeta,$$

so that

$$|\zeta^{-1}(u)| \le a_0 + a_1 |u| \quad \text{for all } u \in \mathbb{R}.$$

Let h be a bounded real-valued function on the domain of (D, X), and write

$$\|h\|_\infty \coloneqq \sup_{d,x} |\alpha(d, x)|.$$

Let L be a linear functional acting on bounded functions, such that for some constant $C_L > 0$,

$$|L(f)| \le C_L\big(1 + \|f\|_\infty\big) \quad \text{for all bounded } f.$$

Define

$$\begin{aligned} L(\zeta^{-1} \circ f) = g\big(\zeta^{-1} \circ f(D, X)\big) + \partial g\big(\zeta^{-1} \circ f(D, X)\big)\zeta^{-1} \circ \alpha(D, X) \\ - \partial g\big(\zeta^{-1} \circ f(1, X)\big) - \partial g\big(\zeta^{-1} \circ f(0, X)\big). \end{aligned}$$

Then there exists a constant $C > 0$ (depending only on g, ζ^{-1} and C_L) such that

$$|L(\zeta^{-1} \circ f)| \le C\big(1 + \|f\|_\infty^2\big).$$

F.2 Upper bound using the empirical-process arguments

From Propositions F.1–F.3, we obtain the following result.

Proposition F.4. *Under the conditions of Theorem 8.1, for any $0 < \gamma < 2$, we have*

$$\begin{aligned} & d\left(\mathbb{E} - \widehat{\mathbb{E}}\right)[L(\widehat{\alpha}, D, X) - L(\alpha_0, D, X)] \\ & = O_p\left(\max\left\{\frac{\|\widehat{\alpha} - \alpha^*\|_{L^2(P_0)}^{1-\gamma/2}\left(1 + \|\widehat{\alpha}\|_{\mathcal{H}}\right)^{1+\gamma/2}}{\sqrt{n}}, \frac{\left(1 + \|\widehat{\alpha}\|_{\mathcal{H}}\right)^2}{n^{2/(2+\gamma)}}\right\}\right), \end{aligned}$$

as $n \to \infty$.

F.3 Proof of Theorem 8.1

We prove Theorem 8.1 following the arguments in Kanamori et al. (2012).

Proof. From Proposition E.3 and $\alpha_0 \in \alpha^{\mathrm{RKHS}}$, we have

$$\begin{aligned}
&\|\widehat{\alpha}(X) - \alpha_0(X)\|_{L_2(P_0)}^2 + \lambda \|\widehat{\alpha}\|_{\mathcal{H}}^2 \\
&\leq \left(\mathbb{E} - \widehat{\mathbb{E}}\right) [L(\widehat{\alpha}, D, X) - L(\alpha_0, D, X)] + \lambda \|f_0\|_{\mathcal{H}}^2.
\end{aligned}$$

From Proposition F.4, we have

$$\begin{aligned}
&\|\widehat{\alpha}(X) - \alpha_0(X)\|_{L_2(P_0)}^2 + \lambda \|\widehat{f}\|_{\mathcal{H}}^2 \\
&= O_p\left(\max\left\{ \frac{\|\widehat{\alpha} - \alpha_0\|_{L^2(P_0)}^{1-\gamma/2} \left(1 + \left\|\widehat{f}\right\|_{\mathcal{H}}\right)^{1+\gamma/2}}{\sqrt{n}}, \frac{(1 + \|\widehat{\alpha}\|_{\mathcal{H}})^2}{n^{2/(2+\gamma)}} \right\} \right) + \lambda \|r_0\|_{\mathcal{H}}^2.
\end{aligned}$$

We consider the following three possibilities:

$$\|\widehat{\alpha}(X) - \alpha_0(X)\|_{L_2(P_0)}^2 + \lambda \|\widehat{f}\|_{\mathcal{H}}^2 = O_p(\lambda), \tag{9}$$

$$\|\widehat{\alpha}(X) - \alpha_0(X)\|_{L_2(P_0)}^2 + \lambda \|\widehat{f}\|_{\mathcal{H}}^2 = O_p\left(\frac{\|\widehat{f} - f_0|_{L^2(P_0)}^{1-\gamma/2} \left(1 + \left\|\widehat{f}\right\|_{\mathcal{H}}\right)^{1+\gamma/2}}{\sqrt{n}} \right), \tag{10}$$

$$\|\widehat{\alpha}(X) - \alpha_0(X)\|_{L_2(P_0)}^2 + \lambda \|\widehat{f}\|_{\mathcal{H}}^2 = O_p\left(\frac{\left(1 + \left\|\widehat{f}\right\|_{\mathcal{H}}\right)^2}{n^{2/(2+\gamma)}} \right). \tag{11}$$

The above inequalities are analyzed as follows:

Case (9). We have

$$\begin{aligned}
&\|\widehat{\alpha}(X) - \alpha_0(X)\|_{L_2(P_0)}^2 = O_p(\lambda), \\
&\lambda \|\widehat{f}\|_{\mathcal{H}}^2 = O_p(\lambda).
\end{aligned}$$

Therefore, we have $\|\widehat{\alpha}(X) - \alpha_0(X)\|_{P_0} = O_p(\lambda^{1/2})$ and $\|\widehat{r}\|_{\mathcal{H}} = O_p(1)$.

Case (10). We have

$$\begin{aligned}
&\|\widehat{\alpha}(X) - \alpha_0(X)\|_{L_2(P_0)}^2 = O_p\left(\frac{\|\widehat{f} - f_0)\|_{L^2(P_0)}^{1-\gamma/2} \left(1 + \left\|\widehat{f}\right\|_{\mathcal{F}}\right)^{1+\gamma/2}}{\sqrt{n}} \right), \\
&\lambda \|\widehat{f}\|_{\mathcal{H}}^2 = O_p\left(\frac{\|\widehat{f} - f_0)\|_{L^2(P_0)}^{1-\gamma/2} \left(1 + \left\|\widehat{f}\right\|_{\mathcal{F}}\right)^{1+\gamma/2}}{\sqrt{n}} \right).
\end{aligned}$$

From the first inequality, we have

$$\|\widehat{\alpha}(X) - \alpha_0(X)\|_{P_0} = \sum_{d\in\{1,0\}} O_p\left(\frac{\left(1+\left\|\widehat{f}\right\|_{\mathcal{F}}\right)^{1+\gamma/2}}{n^{1/(2+\gamma)}}\right).$$

By using this result, from the second inequality, we have

$$\begin{aligned}
\lambda\|\widehat{f}\|_{\mathcal{H}}^2 &= O_p\left(\frac{\|\widehat{f}-f_0)\|_{L^2(P_0)}^{1-\gamma/2}\left(1+\left\|\widehat{f}\right\|_{\mathcal{F}}\right)^{1+\gamma/2}}{\sqrt{n}}\right) \\
&= O_p\left(\left(\frac{1+\left\|\widehat{f}\right\|_{\mathcal{F}}}{n^{1/(2+\gamma)}}\right)^{1-\gamma/2}\frac{\left(1+\left\|\widehat{f}\right\|_{\mathcal{F}}\right)^{1+\gamma/2}}{\sqrt{n}}\right) \\
&= O_p\left(\frac{\left(1+\left\|\widehat{f}\right\|_{\mathcal{F}}\right)^{2}}{n^{2/(2+\gamma)}}\right).
\end{aligned}$$

This implies that

$$\|\widehat{f}\|_{\mathcal{H}} = O_p\left(\frac{\left(1+\left\|\widehat{f}\right\|_{\mathcal{F}}\right)^{2}}{\lambda^{1/2}n^{2/(2+\gamma)}}\right) = o_p(1).$$

Therefore, the following inequity is obtained.

$$\|\widehat{\alpha}(X) - \alpha_0(X)\|_{P_0} = O_p\left(\frac{1}{n^{1/(2+\gamma)}}\right) = O_p(\lambda^{1/2}).$$

Case 11. We have

$$\begin{aligned}
\|\widehat{\alpha}(X) - \alpha_0(X)\|_{L_2(P_0)}^2 &= O_p\left(\frac{\left(1+\left\|\widehat{f}\right\|_{\mathcal{F}}\right)^{2}}{n^{2/(2+\gamma)}}\right), \\
\lambda\|\widehat{f}\|_{\mathcal{H}}^2 &= O_p\left(\frac{\left(1+\left\|\widehat{f}\right\|_{\mathcal{F}}\right)^{2}}{n^{2/(2+\gamma)}}\right).
\end{aligned}$$

As well as the argument in (10), we have $\|\widehat{r}\|_{\mathcal{H}} = o_p(1)$. Therefore, we have

$$\|\widehat{\alpha}(X) - \alpha_0(X)\|_{P_0} = O_p\left(\frac{1}{n^{1/(2+\gamma)}}\right) = O_p(\lambda^{1/2}).$$

□

G Proof of Theorem 8.2

Our proof procedure mainly follows those in Kato & Teshima (2021) and Zheng et al. (2022). In particular, we are inspired by the proof in Zheng et al. (2022).

We prove Theorem 8.2 by proving the following lemma:

Lemma G.1. *Suppose that Assumption 8.3 holds. For any $n \geq Pdim(\mathcal{F}^{FNN})$, there exists a constant $C > 0$ depending on (μ, σ, M) such that for any $\gamma > 0$, with probability at least $1 - \exp(-\gamma)$, it holds that*

$$\left\| \widehat{f} - f_0 \right\|_2 \leq C \left(\sqrt{\frac{Pdim(\mathcal{F}^{FNN}) \log(n)}{n}} + \left\| f^* - f_0 \right\|_2 + \sqrt{\frac{\gamma}{n}} \right).$$

As shown in Zheng et al. (2022), we can bound $\mathrm{Pdim}(\mathcal{F}^{\mathrm{FNN}}) \log(n)$ by specifying neural networks and obtain Theorem 8.2.

G.1 Proof of Lemma G.1

We prove Lemma G.1 by bounding (8) in Proposition E.3.

To bound (8), we show several auxiliary results. Define

$$\begin{aligned}
\widehat{\mathcal{F}}^{f^*,u} &:= \{ f \in \mathcal{F}^{\mathrm{FNN}} \colon \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - f^*(X_i))^2 \leq u \}, \\
\overline{\mathcal{G}}^{f^*,u} &:= \left\{ (f - f^*) \colon f \in \widehat{\mathcal{F}}^{f^*,u} \right\}, \\
\kappa_n^u(u) &:= \mathbb{E}_\sigma \left[\mathfrak{R}_n \overline{\mathcal{G}}^{f^*,u} \right], \\
u^\dagger &:= \inf \left\{ u \geq 0 \colon \kappa_n^u(s) \leq s^2 \quad \forall s \geq u \right\}.
\end{aligned}$$

Here, we show the following two lemmas:

Lemma G.2 (Corresponding to (26) in Zheng et al. (2022))**.** *Suppose that the conditions in Lemma G.1 hold. Then, for any $z > 0$, with probability $1 - \exp(-z)$ it holds that*

$$\begin{aligned}
&\widehat{\mathbb{E}} \left[L(\widehat{\alpha}, D, X) - L(\alpha_0, D, X) \right] \\
&\leq C \left(\| f^*(X) - f_0(X) \|_2^2 + \| f^*(X) - f_0(X) \|_2 \sqrt{\frac{z}{n}} + \frac{16 M z}{3n} \right).
\end{aligned}$$

Lemma G.3 (Corresponding to (29) in Zheng et al. (2022))**.** *Suppose that the conditions in Lemma G.1 hold. If there exists $u_0 > 0$ such that*

$$\| \widehat{f}(X) - f^*(X) \|_2 \leq u_0,$$

then it holds that

$$\begin{aligned}
&\left(\mathbb{E} - \widehat{\mathbb{E}} \right) \left[L(\widehat{\alpha}, D, X) - L(\alpha_0, D, X) \right] \\
&\leq C \left(\mathbb{E}_\sigma \left[\mathfrak{R}_n \overline{\mathcal{G}}^{f^*,u_0} \right] + u_0 \sqrt{\frac{z}{n}} + \frac{M z}{n} \right).
\end{aligned}$$

Additionally, we use the following three propositions directly from Zheng et al. (2022).

Proposition G.4 (From (32) in Zheng et al. (2022)). *Let $u > 0$ be a positive value such that*

$$\|f - f_0\|_2 \leq u$$

for all $f \in \mathcal{F}$. Then, for every $z > 0$, with probability at least $1 - 2\exp(-z)$, it holds that

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}\big(f(X_i) - f_0(X_i)\big)^2} \leq 2u.$$

Proposition G.5 (Corresponding to (36) in Step 3 of Zheng et al. (2022)). *Suppose that the conditions in Lemma G.1 hold. Then, there exists a universal constant $C > 0$ such that*

$$u^\dagger \leq CM\sqrt{\frac{Pdim(\mathcal{F}^{FNN})\log(n)}{n}}.$$

Proposition G.6 (Upper bound of the Rademacher complexity). *Suppose that the conditions in Lemma G.1 hold. If $n \geq Pdim(\mathcal{F}^{FNN})$, $u_0 \geq 1/n$, and $n \geq (2eM)^2$, we have*

$$\mathbb{E}_\sigma\left[\mathfrak{R}_n\overline{\mathcal{G}}^{f^*,u_0}\right] \leq Cr_0\sqrt{\frac{Pdim(\mathcal{F}^{FNN})\log n}{n}}.$$

Then, we prove Lemma G.1 as follows:

Proof of Lemma G.1. If there exists $u_0 > 0$ such that

$$\|\widehat{f}(X) - f^*(X)\|_2 \leq u_0,$$

then from (8) and Lemmas G.2 and G.3, for every $z > 0$, there exists a constant $C > 0$ independent n such that

$$\begin{aligned}&\|\widehat{\alpha}(X) - \alpha_0(X)\|^2_{L_2(P_0)}\\ &\leq C\left(\|f^* - f_0\|_2\sqrt{\frac{z}{n}} + \frac{16Mz}{3n} + u_0\sqrt{\frac{\mathrm{Pdim}(\mathcal{F}^{\mathrm{FNN}})\log n}{n}} + u_0\sqrt{\frac{z}{n}} + \frac{Mz}{n}\right). \qquad (12)\end{aligned}$$

This result implies that if $\sqrt{\mathrm{Pdim}(\mathcal{F}^{\mathrm{FNN}})}$, then there exists n_0 such that for all $n > n_0$, there exists $u_1 < u_0$ such that

$$\|\widehat{\alpha}(X) - \alpha_0(X)\|^2_{L_2(P_0)} \leq u_1.$$

For any $z > 0$, define $\overline{u}$ as

$$\overline{u}_z \geq \max\left\{\sqrt{\log(n)/n}, 4\sqrt{3}M\sqrt{z/n}, u^\dagger\right\}.$$

Define a subspace of $\mathcal{F}^{\mathrm{FNN}}$ as

$$\mathcal{S}^{\mathrm{FNN}}(f_0, \overline{u}_z := \left\{f \in \mathcal{F}^{\mathrm{FNN}} \colon \|f - f_0\| \leq \overline{u}_z\right\}.$$

Define

$$\ell := \lfloor \log_2(2M/\sqrt{\log(n)/n}) \rfloor.$$

Using the definition of subspaces, we divide $\mathcal{F}^{\mathrm{FNN}}$ into the following $\ell + 1$ subspaces:

$$\begin{aligned}
\overline{\mathcal{S}}_0^{\mathrm{FNN}} :=& \mathcal{S}^{\mathrm{FNN}}(f_0, \overline{u}), \\
\overline{\mathcal{S}}_1^{\mathrm{FNN}} :=& \mathcal{S}^{\mathrm{FNN}}(f_0, \overline{u}) \backslash \mathcal{S}^{\mathrm{FNN}}(f_0, \overline{u}), \\
&\vdots \\
\overline{\mathcal{S}}_\ell^{\mathrm{FNN}} :=& \mathcal{S}^{\mathrm{FNN}}(f_0, 2^\ell \overline{u}) \backslash \mathcal{S}^{\mathrm{FNN}}(f_0, 2^{\ell-1} \overline{u}).
\end{aligned}$$

Since $\overline{u}_z > u^\dagger$, from the definition of $u^\dagger$, we have

$$\overline{u}_z^2 \leq \kappa_n^u(\overline{u}).$$

If there exists $j \leq \ell$ such that $\widehat{f} \in \overline{\mathcal{S}}_j^{\mathrm{FNN}}$, then from (12), for every $z > 0$, with probability at least $1 - 8\exp(-z)$, there exists a constant $C > 0$ independent of n such that

$$\begin{aligned}
&\|\widehat{\alpha}(X) - \alpha_0(X)\|_2^2 \\
&\leq C \left(2^{\ell-1} \overline{u} \left(\sqrt{\frac{\mathrm{Pdim}(\mathcal{F}^{\mathrm{FNN}}) \log(n)}{n}} + \sqrt{\frac{z}{n}} \right) + \|f^* - f_0\|_2^2 + \|f^* - f_0\|_2 \sqrt{\frac{z}{n}} + \frac{Mz}{n} \right).
\end{aligned} \tag{13}$$

Additionally, if

$$C \left(\sqrt{\frac{\mathrm{Pdim}(\mathcal{F}^{\mathrm{FNN}}) \log(n)}{n}} + \sqrt{\frac{z}{n}} \right) \leq \frac{1}{8} 2^j \overline{u}, \tag{14}$$

$$C \left(\|f^* - f_0\|_2^2 + \|f^* - f_0\|_2 \sqrt{\frac{z}{n}} + \frac{Mz}{n} \right) \leq \frac{1}{8} 2^{2j} \overline{u}^2 \tag{15}$$

hold, then

$$\|\widehat{\alpha}(X) - \alpha_0(X)\|_2 \leq 2^{j-1} \overline{u}. \tag{16}$$

Here, to obtain (16), we used $\overline{u} \geq \max\left\{ \sqrt{\log(n)/n}, 4\sqrt{3}M\sqrt{z/n}, u^\dagger \right\}$, (13), (14), and (15).

From Proposition G.5, it holds that

$$u^\dagger \leq CM \sqrt{\frac{\mathrm{Pdim}(\mathcal{F}^{\mathrm{FNN}}) \log(n)}{n}}.$$

Therefore, we can choose $\overline{u}$ as

$$\overline{u} := C \left(\sqrt{\frac{\mathrm{Pdim}(\mathcal{F}^{\mathrm{FNN}}) \log(n)}{n}} + \sqrt{\log(n)/n} + 4\sqrt{3}M\sqrt{z/n} \right),$$

where $C > 0$ is a constant independent of n. □

G.2 Proof of Lemma G.2

From Proposition D.1, we have

$$\begin{aligned}
&\widehat{\mathbb{E}}\left[L(\widehat{\alpha}, D, X) - L(\alpha_0, D, X)\right] \\
&\leq \mathbb{E}\left[L(\widehat{\alpha}, D, X) - L(\alpha_0, D, X)\right] + \sqrt{2}C\|f^*(X) - f_0(X)\|\sqrt{\frac{z}{n}} + \frac{16C_1Mz}{3n}.
\end{aligned}$$

This is a direct consequence of Proposition D.1. Note that α^* and α_0 are fixed, and it is enough to apply the standard law of large numbers; that is, we do not have to consider the uniform law of large numbers. However, we can still apply Proposition D.1, which is a general than the standard law of large numbers, with ignoring the Rademacher complexity part.

We have

$$\begin{aligned}
&\widehat{\mathbb{E}}\left[L(\widehat{\alpha}, D, X) - L(\alpha_0, D, X)\right] \\
&\leq \mathbb{E}\left[L(\widehat{\alpha}, D, X) - L(\alpha_0, D, X)\right] \\
&\quad + \sqrt{2}C_1\|f^* - f_0\|\sqrt{\frac{z}{n}} + \frac{16C_2Mz}{3n} + \sqrt{2}C_2\|f^* - f_0\|\sqrt{\frac{z}{n}} + \frac{16C_2Mz}{3n} \\
&\leq C\left(\|f^* - f_0\|_2^2 + \|f^* - f_0\|\sqrt{\frac{z}{n}} + \frac{16CMz}{3n}\right).
\end{aligned}$$

G.3 Proof of Lemma G.3

Let $g \coloneqq (f - f^*)^2$. From the definition of FNNs, we have

$$g \leq 4M^2$$

Additionally, we assumed that $\|\widehat{f} - f^*\|_2 \leq u_0$ holds. Then, it holds that $\mathrm{Var}_{P_0}(g) \leq 4M^2u_0^2$.

Here, we note that the followings hold for all f (r):

$$L(\alpha) - L(\alpha^*) \leq C\Big|f(X) - f^*(X)\Big|,$$

where $C > 0$ is some constant

Then, from Proposition D.1, for every $z > 0$, with probability at least $1 - \exp(-z)$, it holds that

$$\begin{aligned}
&\left(\mathbb{E} - \widehat{\mathbb{E}}\right)\left[L(\widehat{\alpha}, D, X) - L(\alpha_0, D, X)\right] \\
&\leq C\left(\mathbb{E}_\sigma\left[\mathfrak{R}_n\overline{\mathcal{G}}^{f^*, u_0}\right] + r_0\sqrt{\frac{z}{n}} + \frac{Mz}{n}\right).
\end{aligned}$$

H Riesz Regression and Density Ratios

As explained in the main text and in Kato (2025b), the Riesz representer is closely connected to density ratio estimation. In particular, for ATE, the Riesz representer can be expressed in terms of two density ratios relative to the marginal covariate distribution, which leads to a decomposition of the squared loss objective into two LSIF problems.

Riesz representer and density ratio. Let p_Z denote the marginal density of Z and $p_{Z|D=d}$ the conditional density of Z given $D = d$. Let $\kappa_d \coloneqq P_0(D = d)$. By Bayes' rule,

$$p_{Z|D=d}(z) = \frac{p_Z(z)P_0(D = d \mid Z = z)}{P_0(D = d)} = \frac{p_Z(z)e_0(z)^d(1 - e_0(z))^{1-d}}{\kappa_d},$$

where $e_0(z) = P_0(D = 1 \mid Z = z)$.

Define the density ratios with respect to the marginal distribution of Z by

$$r_1(z) \coloneqq \frac{p_Z(z)}{p_{Z|D=1}(z)}, \qquad r_0(z) \coloneqq \frac{p_Z(z)}{p_{Z|D=0}(z)}.$$

From the expression above,

$$r_1(z) = \frac{\kappa_1}{e_0(z)}, \qquad r_0(z) = \frac{\kappa_0}{1 - e_0(z)}.$$

Therefore, the ATE Riesz representer can be written as

$$\alpha_0^{\mathrm{ATE}}(D, Z) = \frac{\mathbb{1}[D = 1]}{e_0(Z)} - \frac{\mathbb{1}[D = 0]}{1 - e_0(Z)} = \mathbb{1}[D = 1]\frac{r_1(Z)}{\kappa_1} - \mathbb{1}[D = 0]\frac{r_0(Z)}{\kappa_0}.$$

Equivalently, estimating α_0^{ATE} reduces to estimating the pair (r_1, r_0), which compare the marginal covariate distribution to the treated and control covariate distributions.

Squared loss objective and decomposition into two LSIF problems. We next connect this representation to LSIF, a density ratio estimation method proposed in Kanamori et al. (2009). Let $g^{\mathrm{SQ}}(u) \coloneqq (u - 1)^2$ be the squared loss. The corresponding population squared loss Bregman objective can be written as

$$\mathrm{BD}_{g^{\mathrm{SQ}}}(\alpha) = \mathbb{E}\Big[- 2\big(\alpha(1, Z) - \alpha(0, Z)\big) + \alpha(D, Z)^2\Big],$$

where $\alpha(d, Z)$ denotes the value of the representer evaluated at treatment status d and covariates Z. Under the parameterization

$$\alpha(D, Z) = \mathbb{1}[D = 1]\frac{r_1(Z)}{\kappa_1} - \mathbb{1}[D = 0]\frac{r_0(Z)}{\kappa_0},$$

we have $\alpha(1, Z) = r_1(Z)/\kappa_1$ and $\alpha(0, Z) = -r_0(Z)/\kappa_0$, hence $\alpha(1, Z) - \alpha(0, Z) = r_1(Z)/\kappa_1 + r_0(Z)/\kappa_0$. Substituting this into $\mathrm{BD}_{g^{\mathrm{SQ}}}(\alpha)$ and using the law of total expectation yields

$$\mathrm{BD}_{g^{\mathrm{SQ}}}(\alpha) = -2\mathbb{E}\left[\frac{r_1(Z)}{\kappa_1} + \frac{r_0(Z)}{\kappa_0}\right] + \mathbb{E}\left[\alpha(D, Z)^2\right]. \tag{17}$$

Moreover,

$$\mathbb{E}\big[\alpha(D, Z)^2\big] = \kappa_1\mathbb{E}\left[\left(\frac{r_1(Z)}{\kappa_1}\right)^2 \,\middle|\, D = 1\right] + \kappa_0\mathbb{E}\left[\left(\frac{r_0(Z)}{\kappa_0}\right)^2 \,\middle|\, D = 0\right].$$

Rewriting (17) in terms of expectations with respect to p_Z and $p_{Z|D=d}$ and dropping constants gives

$$\mathrm{BD}_{g^{\mathrm{SQ}}}(\alpha) \coloneqq -2\mathbb{E}_Z\left[r_1(Z)\right] + \mathbb{E}_{Z|D=1}\left[r_1(Z)^2\right] - 2\mathbb{E}_Z\left[r_0(Z)\right] + \mathbb{E}_{Z|D=0}\left[r_0(Z)^2\right]. \tag{18}$$

Minimizing this objective is exactly LSIF, and in our setting it coincides with SQ-Riesz regression for ATE estimation.

Furthermore, if $r_1(\cdot)$ and $r_0(\cdot)$ are treated as independent functions, minimizing $\mathrm{BD}_{g^{\mathrm{SQ}}}(\alpha)$ over (r_1, r_0) separates into two independent LSIF type problems

$$r_1^* = \arg\min_{r_1} \left\{-2\mathbb{E}_Z[r_1(Z)] + \mathbb{E}_{Z|D=1}[r_1(Z)^2]\right\},$$
$$r_0^* = \arg\min_{r_0} \left\{-2\mathbb{E}_Z[r_0(Z)] + \mathbb{E}_{Z|D=0}[r_0(Z)^2]\right\},$$

where $\mathbb{E}_Z$ and $\mathbb{E}_{Z|D=d}$ denote expectations under $P_0(Z)$ and $P_0(Z \mid D = d)$. At the sample level, with $\mathcal{G}_1$ and $\mathcal{G}_0$ defined as in the Introduction, the empirical LSIF objectives are

$$\widehat{R}_1(r_1) \coloneqq -\frac{2}{n}\sum_{i=1}^{n} r_1(Z_i) + \frac{1}{|\mathcal{G}_1|}\sum_{i\in\mathcal{G}_1} r_1(Z_i)^2,$$
$$\widehat{R}_0(r_0) \coloneqq -\frac{2}{n}\sum_{i=1}^{n} r_0(Z_i) + \frac{1}{|\mathcal{G}_0|}\sum_{i\in\mathcal{G}_0} r_0(Z_i)^2.$$

I Extensions

I.1 Nearest Neighbor Matching

Following this study, Kato (2025a) shows that nearest neighbor matching for ATE estimation can be interpreted as a special case of SQ-Riesz regression, that is, Riesz regression or LSIF. The key step is to express the ATE Riesz representer $\alpha_0^{\mathrm{ATE}}(D, Z)$ in terms of density ratios with respect to the marginal covariate distribution, and to approximate these density ratios via nearest neighbor cells, following the density ratio interpretation in Lin et al. (2023).

NN matching ATE estimator. Let

$$J_M(i) \subset \{1, \ldots, n\}$$

be the index set of the M nearest neighbors of X_i among the units with $D_j = 1 - D_i$. We define estimators $Y(d)$ as

$$\widehat{Y}_i(0) \coloneqq \begin{cases} Y_i, & \text{if } D_i = 0 \\ \frac{1}{M}\sum_{j\in J_M(i)} Y_j, & \text{if } D_i = 1 \end{cases},$$
$$\widehat{Y}_i(1) \coloneqq \begin{cases} \frac{1}{M}\sum_{j\in J_M(i)} Y_j, & \text{if } D_i = 0 \\ Y_i, & \text{if } D_i = 1 \end{cases}.$$

Then, the NN matching ATE estimator is given by

$$\widehat{\theta}_M \coloneqq \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{Y}_i(1) - \widehat{Y}_i(0)\right).$$

Introduce the *matched-times count* (the number of times unit i is used as a match by units in the opposite group) as

$$K_M(i) \coloneqq \sum_{j=1,\ D_j=1-D_i}^{n} \mathbb{1}\big[i \in J_M(j)\big].$$

Then, $\widehat{\theta}_M$ can be written as follows:

$$\widehat{\theta}_M = \frac{1}{n}\left(\sum_{i:D_i=1}\left(1+\frac{K_M(i)}{M}\right)Y_i - \sum_{i:D_i=0}\left(1+\frac{K_M(i)}{M}\right)Y_i\right) = \frac{1}{n}\sum_{i=1}^{n}(2D_i-1)\left(1+\frac{K_M(i)}{M}\right)Y_i.$$

Nearest neighbor matching as density ratio estimation. Lin et al. (2023) first shows that nearest neighbor matching can be interpreted as a method for density ratio estimation. Let $X, Z \in \mathcal{X}$ be independent whose pdfs are $p_1(x)$ and $p_0(z)$. We assume that $p_1(x), p_0(x) > 0$ for all $x \in \mathcal{X}$. We observe i.i.d. samples $\{X_i\}_{i=1}^{N_0}$ and $\{Z_j\}_{j=1}^{N_1}$ and aim to estimate the density ratio

$$r_0^\dagger(x) \coloneqq \frac{p_1(x)}{p_0(x)}.$$

For $M \in \{1, \ldots, N_0\}$ and $z \in \mathbb{R}^d$, let $\mathcal{X}_{(M)}(z)$ be the M-th nearest neighbor of z in $\{X_i\}_{i=1}^{N_0}$ under a given metric $\|\cdot\|$. Define the *catchment area* of x as

$$A_M(x) \coloneqq \big\{z : \|x - z\| \le \|\mathcal{X}_{(M)}(z) - z\|\big\},$$

and the *matched-times count* as

$$K_M(x) \coloneqq \sum_{j=1}^{N_1} \mathbb{1}\big(Z_j \in A_M(x)\big).$$

Lin et al. (2023) proposes the one-step estimator

$$\widehat{r}_M^\dagger(x) = \frac{N_0}{N_1}\frac{K_M(x)}{M},$$

which corresponds to nearest neighbor matching in ATE estimation.

Using this result, Lin et al. (2023) explains that nearest neighbor matching corresponds to the estimation of the density ratio r_0 defined above. They also show that their method is computationally efficient and rate-optimal for Lipschitz densities.

Nearest neighbor matching as LSIF We next show that the density ratio estimator of Lin et al. (2023) is a variant of LSIF. Therefore, since we have already discussed that Riesz regression and LSIF are essentially the same, NN matching can also be interpreted as a special case of Riesz regression.

Let us consider the following density ratio model:

$$r(1,z) = \phi(z)\beta,$$

where $\phi(\cdot)$ is a basis function defined as

$$\phi(z) \coloneqq \phi_c(z) = \mathbb{1}\big[z \in A_M(c)\big],$$

and recall that $A_M(c) \coloneqq \big\{z : \|c - z\| \leq \|\mathcal{X}_{(M)}(z) - z\|\big\}$.

For $z = c = X_i$, we define an estimator of the density ratio as

$$\widehat{r}(1,c) = \phi(c)\widehat{\beta}$$

with the estimated parameter defined as

$$\widehat{\beta} \coloneqq \underset{\beta\in\mathbb{R}}{\arg\min} \left\{ \frac{1}{2\sum_{i=1}^n \mathbb{1}[D_i = 1]} \Big(\phi_c(Z_i)\beta\Big)^2 - \frac{1}{n}\phi_c(Z_i)\beta \right\},$$

This estimation corresponds to LSIF with the kernel function.

This estimator is equivalent to

$$\widehat{\beta} \coloneqq \underset{\beta\in\mathbb{R}}{\arg\min} \left\{ \frac{1}{2}\beta^\top \widehat{H}\beta - \beta^\top \widehat{h} + \frac{\lambda}{2}\|\beta\|_2^2 \right\} = \Big(\widehat{H} + \lambda I\Big)^{-1}\widehat{h},$$

where

$$\begin{aligned}
\widehat{H} &\coloneqq \frac{1}{n}\sum_{i=1}^n \mathbb{1}\left[D_i = 1\right]\phi_c(Z_i)^2 = \frac{M}{n}, \\
\widehat{h} &\coloneqq \frac{1}{n}\sum_{i=1}^n \phi_c(Z_i) = \frac{1}{n}\big(M + K_M(i)\big), \\
\phi_c(c) &= 1.
\end{aligned}$$

Here, we have

$$\begin{aligned}
\widehat{H} &= \frac{1}{n}\sum_{i=1}^n \mathbb{1}\left[D_i = 1\right]\phi_c(Z_i) = \frac{M}{n}, \\
\widehat{h} &= \frac{1}{n}\sum_{i=1}^n \phi_c(Z_i) = \frac{1}{n}\sum_{i=1}^n \big(\mathbb{1}\left[D_i = 1\right]\phi_c(Z_i) + \mathbb{1}\left[D_i = 0\right]\phi_c(X_i)\big) = \frac{1}{n}\big(M + K_M(i)\big), \\
\phi_c(c) &= 1,
\end{aligned}$$

where we recall that

$$K_M(i) \coloneqq \sum_{j=1,\ D_j=1-D_i}^n \mathbb{1}\left[i \in J_M(j)\right].$$

Therefore, when $\lambda = 0$, the estimator $\widehat{r}_1(c)$ is given by

$$\widehat{r}(1,c) = \widehat{r}(1,Z_i) = 1 + \frac{K_M(i)}{M}.$$

Similarly, we can estimate $\widehat{r}(0,c) = \phi_c(c)\widehat{\beta}$ by solving an empirical version of the following problem:

$$\widehat{\beta} := \arg\min_{\beta \in \mathbb{R}} \left\{ \frac{1}{2\sum_{i=1}^n \mathbb{1}[D_i = 0]} \Big(\phi_c(Z_i)\beta\Big)^2 - \frac{1}{n}\phi_c(Z_i)\beta \right\},$$

Then, the estimator is given by

$$\widehat{r}(0,c) = \widehat{r}(0,Z_i) = 1 + \frac{K_M(i)}{M}.$$

Using these estimators, we construct the following inverse propensity score estimator for the ATE:

$$\widehat{\theta}_M = \frac{1}{n}\left(\sum_{i:D_i=1} \left(1 + \frac{K_M(i)}{M}\right) Y_i - \sum_{i:D_i=0} \left(1 + \frac{K_M(i)}{M}\right) Y_i \right).$$

This estimator is equivalent to an ATE estimator proposed in Lin et al. (2023), which is shown to be equal to the NN matching estimator of Abadie & Imbens (2006).

Thus, NN matching estimator is a special case of SQ-Riesz regression (LSIF) with a particular choice of a basis function.

I.2 Causal Tree / Causal Forest

Causal trees and causal forests estimate the conditional average treatment effect (CATE) by constructing a partition of the covariate space and estimating a local ATE within each cell, as in Wager & Athey (2018). We emphasize that this procedure implicitly constructs an estimator of the corresponding Riesz representer. In particular, once a partition is fixed, the leafwise CATE estimator can be rewritten as an inverse probability weighting type estimator, with weights that coincide with a leafwise Riesz representer estimator.

Leafwise CATE as a Riesz representer plug in. Let $\Pi = \{\ell\}$ be a partition of the covariate space $\mathcal{Z}$ produced by a causal tree, and let $\ell(z) \in \Pi$ denote the leaf containing $z \in \mathcal{Z}$. For a leaf ℓ, define $n_\ell := \sum_{i=1}^n \mathbb{1}[Z_i \in \ell]$, $n_{1,\ell} := \sum_{i=1}^n \mathbb{1}[D_i = 1, Z_i \in \ell]$, and $n_{0,\ell} := \sum_{i=1}^n \mathbb{1}[D_i = 0, Z_i \in \ell]$. The CATE estimator obtained by a causal tree is the leafwise difference in means

$$\widehat{\theta}(z) := \frac{1}{n_{1,\ell(z)}} \sum_{i:D_i=1, Z_i \in \ell(z)} Y_i - \frac{1}{n_{0,\ell(z)}} \sum_{i:D_i=0, Z_i \in \ell(z)} Y_i.$$

This estimator admits the weighted representation

$$\widehat{\theta}(z) = \frac{1}{n}\sum_{i=1}^n \widehat{\alpha}(D_i, Z_i; z) Y_i,$$

where

$$\widehat{\alpha}(D, Z; z) \coloneqq \mathbb{1}[Z \in \ell(z)] \left(\frac{D}{\widehat{\pi}_{1,\ell(z)}} - \frac{1-D}{\widehat{\pi}_{0,\ell(z)}} \right) \times \frac{1}{\widehat{p}_{\ell(z)}}, \qquad \widehat{\pi}_{d,\ell} \coloneqq \frac{n_{d,\ell}}{n_\ell}, \qquad \widehat{p}_\ell \coloneqq \frac{n_\ell}{n}.$$

Hence $\widehat{\theta}(z)$ is an inverse probability weighting type estimator with a weight function $\widehat{\alpha}(\cdot, \cdot; z)$. This weight function is a plug-in estimator of the leafwise Riesz representer for the local ATE

$$\theta(\ell) \coloneqq \mathbb{E}\big[Y(1) - Y(0) \mid Z \in \ell\big],$$

because the corresponding population representer takes the same form, with $(\widehat{\pi}_{d,\ell}, \widehat{p}_\ell)$ replaced by their population counterparts. Therefore, conditional on the partition, causal trees estimate the CATE by implicitly estimating a Riesz representer that is constant on each leaf.

Connection to SQ-Riesz regression and adaptive nearest neighbors. The expression above shows that a causal tree is a histogram-type estimator of the Riesz representer, where the feature dictionary is given by leaf indicators $\{\mathbb{1}[Z \in \ell]\}_{\ell \in \Pi}$. This is directly analogous to the nearest neighbor histogram model in the previous subsection, except that the partition is learned from the data rather than fixed a priori. From this viewpoint, the splitting criterion in a causal tree can be interpreted as choosing an adaptive partition that reduces the error of the induced leafwise Riesz representer approximation, and hence reduces the error of the resulting local CATE estimator.

A causal forest averages many such trees, built on subsamples and random feature choices, and therefore produces weights that average the leafwise Riesz representer estimators across trees. Equivalently, causal forests produce an adaptive nearest neighbor type representation for CATE, where the neighborhood structure is learned via the random partitions. This clarifies why causal trees and causal forests fit naturally into the same squared loss Bregman divergence, namely SQ-Riesz, perspective as nearest neighbor matching, with the main difference being that causal forests learn the partition adaptively to target CATE estimation accuracy.

I.3 AME Estimation by Score Matching

A subsequent work Kato (2025c) shows that, for derivative-type linear functionals, the Riesz representer can be estimated via score matching. This principle also underlies score-based diffusion models (Song & Ermon, 2020; Song et al., 2021). This viewpoint is useful for AME and APE estimation, and for mitigating overfitting in flexible Riesz representer models, because score matching objectives introduce smoothing through derivatives or noise perturbations.

Score matching identity for AME. Recall the AME example in Section 2, where

$$m^{\mathrm{AME}}(W, \gamma) = \partial_d \gamma(D, Z), \qquad \alpha_0^{\mathrm{AME}}(D, Z) = -\partial_d \log f_0(D, Z),$$

with f_0 denoting the joint density of $X = (D, Z)$. Let $s_{0,d}(x) \coloneqq \partial_d \log f_0(x)$ be the dth component of the score. Consider a sufficiently smooth candidate function $\alpha(x)$ such that integration by parts is valid and boundary terms vanish. Then,

$$\mathbb{E}\left[\partial_d \alpha(X)\right] = \int \partial_d \alpha(x) f_0(x) \mathrm{d}x = -\int \alpha(x) \partial_d f_0(x) \mathrm{d}x = -\mathbb{E}\left[\alpha(X) s_{0,d}(X)\right].$$

Therefore, the squared loss Bregman objective for AME can be rewritten as

$$\mathbb{E}\left[\alpha(X)^2 - 2\partial_d \alpha(X)\right] = \mathbb{E}\left[\alpha(X)^2 + 2\alpha(X) s_{0,d}(X)\right] = \mathbb{E}\left[\left(\alpha(X) + s_{0,d}(X)\right)^2\right] - \mathbb{E}\left[s_{0,d}(X)^2\right].$$

The last term is constant in α. Hence minimizing $\mathbb{E}\left[\alpha(X)^2 - 2\partial_d \alpha(X)\right]$ is equivalent to minimizing $\mathbb{E}\left[\left(\alpha(X) - \alpha_0^{\mathrm{AME}}(X)\right)^2\right]$, and the population minimizer is $\alpha_0^{\mathrm{AME}} = -s_{0,d}$. This is a coordinatewise form of the classical score matching principle and shows that, for derivative-type m, our squared loss Bregman risk coincides with an L_2 score matching risk for the Riesz representer.

Denoising score matching via diffusion. In high dimensions, directly learning the score $x \mapsto \nabla_x \log f_0(x)$ can be unstable. Score-based diffusion models address this issue by learning scores of noise-perturbed distributions via denoising score matching (Song et al., 2021). Let T be a noise index, continuous or discrete, and generate noisy covariates by

$$X_T \coloneqq X + \sigma(T) Z, \qquad Z \sim \mathcal{N}(0, I),$$

independent of $X \sim f_0$. Let p_T denote the density of X_T. A time-dependent score model $s_\theta(\cdot, T)$ is trained by minimizing the denoising objective

$$\mathbb{E}\left[\left\|\sigma(T) s_\theta(X_T, T) + Z\right\|^2\right],$$

which is equivalent, up to an additive constant, to matching $s_\theta(\cdot, T)$ to the true score $\nabla_x \log p_T(x)$ under an L_2 risk. Once s_θ is trained, we can recover an estimator of the original score $\nabla_x \log f_0(x)$ by evaluating at small noise levels and then extracting the relevant component to estimate

$$\alpha_0^{\mathrm{AME}}(x) = -\partial_d \log f_0(x).$$

Operationally, this replaces the derivative term $\partial_d \alpha(X)$ in the score matching objective with a denoising criterion that learns a smoothed score field. This smoothing can mitigate overfitting in high-capacity models and can be combined with flexible neural architectures through automatic differentiation.

I.4 Riesz Representer Estimation via Infinitesimal Classification

Next, following Kato (2025c), we introduce Riesz representer estimation via infinitesimal classification, which also reduces to score matching. This approach applies to a broader range of applications, not only to AME estimation.

Density ratio estimation via infinitesimal classification We first review density ratio estimation via infinitesimal classification, proposed in Choi et al. (2022). Let $p_0(x)$ and $p_1(x)$ be two probability density functions such that $p_0(x) > 0$ holds for all $x \in \mathcal{X}$. For $x \in \mathcal{X}$, the density ratio is defined as

$$r_0(x) := \frac{p_0(x)}{p_1(x)}.$$

We aim to estimate r_0.

We define a continuum of bridge densities $\{p_t\}_{t\in[0,1]}$ through a simple sampling procedure. Let $p_t(x)$ be the probability density function of the random variable

$$X_t = \beta_t^{(1)} X_0 + \beta_t^{(2)} X_1,$$

where $\beta_\cdot^{(1)}, \beta_\cdot^{(2)} \colon [0,1] \to [0,1]$ are C^2 and monotonic, and satisfy the boundary conditions $\beta_0^{(1)} = 1$, $\beta_0^{(2)} = 0$, $\beta_1^{(1)} = 0$, and $\beta_1^{(2)} = 1$. Using $\frac{p_{(t-1)/T}(x)}{p_{t/T}(x)}$ as an intermediate density ratio, we decompose the density ratio into a product of density ratios as

$$r_0(x) = \prod_{t=1}^{T} \frac{p_{(t-1)/T}(x)}{p_{t/T}(x)}.$$

We can choose $\beta_t^{(1)}$ and $\beta_t^{(2)}$ so that the density ratio can be trained stably. For example, DRE-∞ proposes using $\beta_t^{(1)} = 1 - t$ and $\beta_t^{(2)} = t$ in some applications.

In practice, when optimizing objectives that integrate over t, we sample t jointly with (X_0, X_1). Specifically, for each stochastic gradient step we draw a mini batch $\{(X_{0,i}, X_{1,i})\}_{i=1}^B$ with $X_{0,i} \sim p_0$ and $X_{1,i} \sim p_1$ independently, and we draw times $\{t_i\}_{i=1}^B$ i.i.d. from a reference density $q(t)$ on $[0,1]$. We then form $X_{t_i,i} = \beta^{(1)}(t_i)X_{0,i} + \beta^{(2)}(t_i)X_{1,i}$ and approximate time integrals using importance weights. For example, an integral term of the form $\int_0^1 \mathbb{E}_{X_t\sim p_t}[h(X_t,t)]\mathrm{d}t$ is estimated by

$$\int_0^1 \mathbb{E}_{X_t\sim p_t}[h(X_t,t)]\mathrm{d}t \approx \frac{1}{B}\sum_{i=1}^{B} \frac{h(X_{t_i,i},t_i)}{q(t_i)}.$$

Endpoint expectations, such as $\mathbb{E}_{X_0\sim p_0}[\cdot]$ and $\mathbb{E}_{X_1\sim p_1}[\cdot]$, are approximated by sample averages over $\{X_{0,i}\}$ and $\{X_{1,i}\}$, respectively. All derivatives with respect to t that appear in the objective, such as $\partial_t(\lambda(t)s_\beta(X_t,t))$, can be computed by automatic differentiation through the explicit dependence of X_t on t via $\beta^{(1)}(t)$ and $\beta^{(2)}(t)$.

By taking the logarithm, we have

$$\log\left(r_0(x)\right) = \sum_{t=1}^{T} \log \frac{p_{(t-1)/T}(x)}{p_{t/T}(x)}.$$

Then, as $T \to \infty$, the following holds (Choi et al., 2022; Chen et al., 2025):

$$\log r_0(x) = \log\left(\frac{p_0(x)}{p_1(x)}\right) = \sum_{t=1}^{T} \log\left(\frac{p_{(t-1)/T}(x)}{p_{t/T}(x)}\right) = \int_1^0 \partial_t \log p_t(x)\mathrm{d}t.$$

Let $s_\beta^{\mathrm{time}}(x,t)$ be a time score model that approximates the time score $\partial_t \log p_t(x)$. We train $s_\beta^{\mathrm{time}}(x,t)$ by minimizing the following time score matching loss (Choi et al., 2022; Chen et al., 2025):

$$\mathcal{R}^\dagger\left(s_\beta^{\mathrm{time}}\right) \coloneqq \int_0^1 \mathbb{E}_{X_t\sim p_t(x)}\left[\lambda(t)\Big(\partial_t \log p_t(X_t) - s_\beta^{\mathrm{time}}(X_t,t)\Big)^2\right] \mathrm{d}t,$$

where $\lambda\colon [0,1]\to\mathbb{R}_+$ is a positive weighting function. Although $\log p_t(x)$ is unknown in practice, the following alternative objective has been proposed, which is equivalent to $\mathcal{R}^\dagger\left(s_\beta^{\mathrm{time}}\right)$ up to a constant term that is irrelevant for optimization:

$$\begin{aligned}\mathcal{R}\left(s_\beta^{\mathrm{time}}\right) \coloneqq{}& \mathbb{E}_{X_0\sim p_0(x)}\left[\lambda(0)s_\beta^{\mathrm{time}}(X_0,0)\right] - \mathbb{E}_{X_1\sim p_1(x)}\left[\lambda(1)s_\beta^{\mathrm{time}}(X_1,1)\right]\\ &+ \int_0^1 \mathbb{E}_{X_t\sim p_t(x)}\left[\partial_t\left(\lambda(t)s_\beta^{\mathrm{time}}(X_t,t)\right) + \frac{1}{2}\lambda(t)s_\beta^{\mathrm{time}}(X_t,t)^2\right] \mathrm{d}t,\end{aligned}$$

To generate a sample from p_t, we proceed as follows. First, draw two independent endpoint samples

$$X_0\sim p_0, \qquad X_1\sim p_1,$$

independently across draws and independent of each other. Second, for a given time $t\in[0,1]$, construct the bridge sample by the deterministic map

$$X_t \coloneqq \beta^{(1)}(t)X_0 + \beta^{(2)}(t)X_1.$$

We define p_t as the probability law of X_t induced by this procedure, that is, p_t is the pushforward of the product measure $p_0\otimes p_1$ through the map $(x_0,x_1)\mapsto \beta^{(1)}(t)x_0+\beta^{(2)}(t)x_1$. With this definition, expectations under p_t can be evaluated by Monte Carlo as

$$\mathbb{E}_{X_t\sim p_t}\big[f(X_t,t)\big] = \mathbb{E}\big[f(\beta^{(1)}(t)X_0+\beta^{(2)}(t)X_1,t)\big],$$

where the outer expectation is taken over $(X_0,X_1)\sim p_0\otimes p_1$.

Riesz representer estimation via infinitesimal classification Kato (2025c) extends density ratio estimation via infinitesimal classification to Riesz representer estimation. In this subsection, we introduce an example of the method for APE estimation. For implementations in other applications, see Kato (2025c).

In APE estimation, the Riesz representer is given by

$$\alpha^{\mathrm{APE}}(X) \coloneqq \frac{p_1(X)-p_{-1}(X)}{p_0(X)}.$$

By using intermediate density ratios, we have

$$\begin{aligned}\frac{p_1(x)}{p_0(x)} &= \prod_{t=1}^{T}\frac{p_{t/T}(x)}{p_{(t-1)/T}(x)},\\ \frac{p_{-1}(x)}{p_0(x)} &= \prod_{t=1}^{T}\frac{p_{-t/T}(x)}{p_{-(t-1)/T}(x)}.\end{aligned}$$

Then, we can approximate the density ratio as

$$\log \frac{p_1(x)}{p_0(x)} = \sum_{t=1}^{T} \log \frac{p_{t/T}(x)}{p_{(t-1)/T}(x)} \to \int_0^1 \partial_t \log p_t(x) \mathrm{d}t \quad (T \to \infty),$$
$$\log \frac{p_{-1}(x)}{p_0(x)} = \sum_{t=1}^{T} \log \frac{p_{-t/T}(x)}{p_{-(t-1)/T}(x)} \to \int_0^{-1} \partial_t \log p_t(x) \mathrm{d}t \quad (T \to \infty).$$

We define a random variable X_t as

$$X_t \coloneqq \begin{cases} \beta_t^{(1)} X_1 + \beta_t^{(2)} X_0 & \text{if} \quad t \geq 0 \\ \beta_t^{(1)} X_{-1} + \beta_t^{(2)} X_0 & \text{if} \quad t < 0 \end{cases},$$

where $\beta_t^{(1)}, \beta_t^{(2)} \colon [-1, 1] \to [0, 1]$ are of class C^2 and monotonic, with $\beta_t^{(1)}$ increasing and $\beta_t^{(2)}$ decreasing for $t \geq 0$, $\beta_t^{(1)}$ decreasing and $\beta_t^{(2)}$ increasing for $t < 0$, and satisfying the boundary conditions: $\beta_0^{(1)} = 0$, $\beta_0^{(2)} = 1$, $\beta_{-1}^{(1)} = 1$, $\beta_{-1}^{(2)} = 0$, $\beta_1^{(1)} = 1$, and $\beta_1^{(2)} = 0$.

Let $p_t(x)$ be the probability density function. Let $s_\beta^{\mathrm{time}}(x, t)$ be a time score model that approximates the time score $\partial_t \log p_t(x)$. We train the score model by minimizing

$$\mathcal{R}^{\mathrm{APE}\dagger}\left(s_\beta^{\mathrm{time}}\right) \coloneqq \int_{-1}^{1} \mathbb{E}_{X_t \sim p_t(x)} \left[\lambda(t) \Big(\partial_t \log p_t(X_t) - s_\beta^{\mathrm{time}}(X_t, t) \Big)^2 \right] \mathrm{d}t,$$

where $\lambda \colon [-1, 1] \to \mathbb{R}_+$ is a positive weighting function. Since $\partial_t \log p_t(x)$ is unknown, we minimize the following risk:

$$\begin{aligned} \mathcal{R}^{\mathrm{APE}}\left(s_\beta^{\mathrm{time}}\right) \coloneqq{}& \mathbb{E}_{X_{-1} \sim p_{-1}(x)} \left[\lambda(-1) s_\beta^{\mathrm{time}}(X_{-1}, -1) \right] - \mathbb{E}_{X_1 \sim p_1(x)} \left[\lambda(1) s_\beta^{\mathrm{time}}(X_1, 1) \right] \\ &+ \int_{-1}^{1} \mathbb{E}_{X_t \sim p_t(x)} \left[\partial_t \left(\lambda(t) s_\beta^{\mathrm{time}}(X_t, t) \right) + \frac{1}{2} \lambda(t) s_\beta^{\mathrm{time}}(X_t, t)^2 \right] \mathrm{d}t. \end{aligned}$$

J KKT Conditions as Bregman Projections

In this section, we show that how the first-order (KKT) conditions in our generalized Riesz regression coincide with the characterization of a (sieve) Riesz representer as the solution to a linear equation in a Hilbert space discussed in Chen & Liao (2015) and Chen & Pouzo (2015), which show that the Riesz representer can be formulated via linear equation in semiparametric generalized method of moments (GMM) and efficiency analysis.

J.1 Riesz Representer as a Linear Equation in a Hilbert Space

Let $\mathcal{H} \coloneqq L_2(P_X)$ with inner product $\langle f, g \rangle \coloneqq \mathbb{E}[f(X)g(X)]$. For the linear map $\gamma \mapsto \mathbb{E}[m(W, \gamma)]$ (Section 2), the Riesz representation theorem yields $\alpha_0 \in \mathcal{H}$ such that

$$\mathbb{E}[m(W, \gamma)] = \langle \alpha_0, \gamma \rangle \qquad \forall \gamma \in \mathcal{H}. \tag{19}$$

If we restrict to a finite-dimensional sieve space $\mathcal{H}_p := \operatorname{span}\{\phi_1, \dots, \phi_p\}$, the sieve Riesz representer $\alpha_p \in \mathcal{H}_p$ is the unique element satisfying

$$\langle \alpha_p, \phi_j \rangle = \mathbb{E}[m(W, \phi_j)] \qquad j = 1, \dots, p. \tag{20}$$

Writing $\alpha_p(x) = \boldsymbol{\phi}(x)^\top \boldsymbol{\beta}$ with $\boldsymbol{\phi} := (\phi_1, \dots, \phi_p)^\top$, (20) becomes the linear system

$$\underbrace{\mathbb{E}[\boldsymbol{\phi}(X)\boldsymbol{\phi}(X)^\top]}_{=:G} \boldsymbol{\beta} = \underbrace{\mathbb{E}[m(W, \boldsymbol{\phi})]}_{=:b}, \tag{21}$$

which is the familiar "Gram matrix × coefficients = RHS" equation emphasized in sieve Riesz-representer constructions.

J.2 Bregman objectives, dual variables, and a common projection geometry

Recall the pointwise Bregman divergence

$$\mathrm{BD}_g^\dagger(\alpha_0(x) \mid \alpha(x)) := g(\alpha_0(x)) - g(\alpha(x)) - \partial g(\alpha(x))\big(\alpha_0(x) - \alpha(x)\big),$$

and the population target $\alpha^* := \arg\min_{\alpha \in \mathcal{H}} \mathbb{E}[\mathrm{BD}_g^\dagger(\alpha_0(X) \mid \alpha(X))]$. A standard first-order characterization of Bregman projections is the following condition: if $\mathcal{H}$ is convex and α^* is an interior minimizer, then

$$\Big\langle \partial g(\alpha_0) - \partial g(\alpha^*),\, \alpha - \alpha^* \Big\rangle \le 0 \qquad \forall \alpha \in \mathcal{H}, \tag{22}$$

with ≤ 0 replaced by $= 0$ along feasible smooth directions; with a KKT form for general constraints. For the derivation, see Remark J.2. Equation (22) makes clear that all losses share the same underlying $L_2(P_X)$ inner product geometry; what changes across losses is the *dual coordinate* $\partial g(\alpha)$ that appears in the orthogonality.

A particularly convenient reparameterization uses the convex conjugate g^* and the dual variable

$$u(x) := \partial g(\alpha(x)). \tag{23}$$

Whenever g is strictly convex and differentiable on its domain, the Fenchel–Young identity implies $g^*(u) = \alpha u - g(\alpha)$ when $u = \partial g(\alpha)$, and hence the (population) objective in Section 3 can be written as

$$\mathrm{BD}_g(\alpha) = \mathbb{E}\big[g^*(u(X))\big] - \mathbb{E}\big[m(W, u)\big] \qquad \text{with } u = \partial g \circ \alpha, \tag{24}$$

up to an additive constant independent of α. This dual form is useful because its score is simple: $\partial g^*(u) = (\partial g)^{-1}(u) = \alpha$.

Finite-dimensional models and KKT. Consider a model class specified in *dual* coordinates as

$$u_{\boldsymbol{\beta}}(X) = \boldsymbol{\phi}(X)^\top \boldsymbol{\beta}, \qquad \alpha_{\boldsymbol{\beta}}(X) = (\partial g)^{-1}\big(u_{\boldsymbol{\beta}}(X)\big), \tag{25}$$

possibly with a branch indicator $\xi(X) \in \{0,1\}$ to enforce sign restrictions (as in Section 4). Let the empirical objective be the penalized M-estimation problem

$$\widehat{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \Big\{ \widehat{\mathbb{E}}\big[g^*(u_{\boldsymbol{\beta}}(X))\big] - \widehat{\mathbb{E}}\big[m(W, u_{\boldsymbol{\beta}})\big] + \tfrac{\lambda}{a}\|\boldsymbol{\beta}\|_a^a \Big\}.$$

Using $(g^*)'(u) = \alpha$ and $\partial u_{\boldsymbol{\beta}}/\partial\beta_j = \phi_j$, the KKT conditions take the unified form

$$\widehat{\mathbb{E}}\Big[\widehat{\alpha}(X)\phi_j(X) - m(W,\phi_j)\Big] \in \lambda\,\partial\Big(\tfrac{1}{a}|\beta_j|^a\Big) \qquad j = 1,\dots,p, \tag{26}$$

where $\widehat{\alpha} := \alpha_{\widehat{\boldsymbol{\beta}}}$ and $\partial(\cdot)$ denotes the (sub)gradient. In particular, when $\lambda = 0$, (26) reduces exactly to the sieve Riesz equations (20):

$$\widehat{\mathbb{E}}\Big[\widehat{\alpha}(X)\phi_j(X)\Big] = \widehat{\mathbb{E}}\Big[m(W,\phi_j)\Big], \qquad j = 1,\dots,p. \tag{27}$$

Thus, *independently of the choice of* g, once we model $\partial g(\alpha)$ linearly in the basis $\boldsymbol{\phi}$, the KKT conditions say that generalized Riesz regression returns (approximately) the *sieve Riesz representer* characterized by the linear equations (20)–(21). The role of g is to select, among (approximately) balancing solutions, the one that is a Bregman projection (hence a minimum-g solution).

Remark (Derivation of (22))**.** *This remark derives (22) as a first-order optimality (KKT) condition. Throughout, equip* $L_2(P_X)$ *with the inner product* $\langle f, h\rangle := \mathbb{E}[f(X)h(X)]$.

Recall that for each point $\alpha, \alpha_0 \in \mathbb{R}$*, the Bregman divergence is given as*

$$\mathrm{BD}_g^\dagger(\alpha_0 \mid \alpha) := g(\alpha_0) - g(\alpha) - \partial g(\alpha)\big(\alpha_0 - \alpha\big).$$

Fix α*. By convexity of* g*,* $u \mapsto \mathrm{BD}_g^\dagger(\alpha_0 \mid \alpha)$ *is convex and*

$$\frac{\partial}{\partial\alpha}\mathrm{BD}_g^\dagger(\alpha_0 \mid \alpha) = \partial g(\alpha_0) - \partial g(\alpha).$$

For functions $\alpha, \alpha_0 \colon \mathcal{X} \to \mathcal{A}$*, consider the following problem*

$$\alpha^* \in \arg\min_{\alpha\in\mathcal{H}} \mathbb{E}\left[\mathrm{BD}_g^\dagger(\alpha_0(X) \mid \alpha(X))\right],$$

where $\mathcal{H}$ *is a convex subset of* $L_2(P_X)$*. Dropping constants that do not depend on* α *yields*

$$\mathrm{BD}_g(\alpha) = \mathbb{E}\big[g(\alpha(X)) - \partial g(\alpha_0(X))\,\alpha(X)\big]$$

Hence the (Fréchet/Gâteaux) gradient of $\mathrm{BD}_g(\alpha)$ *at* α *in the* $L_2(P_X)$ *geometry is*

$$\nabla\mathrm{BD}_g(\alpha) = \partial g(\alpha) - \partial g(\alpha_0),$$

in the sense that for any direction $h \in L_2(P_X)$*,*

$$\frac{d}{dt}\mathrm{BD}_g(\alpha + th)\Big|_{t=0} = \big\langle \partial g(\alpha) - \partial g(\alpha_0),\, h\big\rangle,$$

whenever differentiation and expectation can be interchanged (e.g., under dominated convergence and mild integrability conditions).

Assume $\mathcal{H}$ is convex and α^* is an interior minimizer in $\mathcal{H}$. For any $\alpha \in \mathcal{H}$ and $t \in [0,1]$, define the feasible path

$$\alpha_t := \alpha^* + t(\alpha - \alpha^*) \in \mathcal{H}.$$

Since α^* minimizes Q over $\mathcal{H}$, the one-sided directional derivative along $\alpha - \alpha^*$ must be nonnegative:

$$0 \le \frac{d}{dt}\mathrm{BD}_g(\alpha_t)\Big|_{t=0+}.$$

Compute:

$$\frac{d}{dt}\mathrm{BD}_g(\alpha_t)\Big|_{t=0} = \Big\langle \partial g(\alpha^*) - \partial g(\alpha_0),\, \alpha - \alpha^* \Big\rangle.$$

Therefore,

$$0 \le \Big\langle \partial g(\alpha^*) - \partial g(\alpha_0),\, \alpha - \alpha^* \Big\rangle \quad \Longleftrightarrow \quad \Big\langle \partial g(\alpha_0) - \partial g(\alpha^*),\, \alpha - \alpha^* \Big\rangle \le 0,$$

which is exactly (22).

If α^* is an interior point of $\mathcal{H}$, then for sufficiently small $|t|$ we have $\alpha^* + th \in \mathcal{H}$ for any admissible direction h. Applying the previous argument to both $t \downarrow 0$ and $t \uparrow 0$ forces

$$\frac{d}{dt}\mathrm{BD}_g(\alpha^* + th)\Big|_{t=0} = 0 \quad \textit{for any (smooth) feasible direction } h,$$

which corresponds to the "$= 0$ along feasible smooth directions" statement.

Define the normal cone of $\mathcal{H}$ at α^* by

$$N_{\mathcal{H}}(\alpha^*) := \Big\{ v \in L_2(P_X) \colon\ \langle v, \alpha - \alpha^* \rangle \le 0 \ \ \forall \alpha \in \mathcal{H} \Big\}.$$

Then (22) is equivalent to the normal-cone inclusion

$$\partial g(\alpha_0) - \partial g(\alpha^*) \in N_{\mathcal{H}}(\alpha^*),$$

which is the standard KKT characterization for minimizing a convex functional over a convex set.

J.3 (A) Squared loss + linear link (SQ-Riesz) as an L_2 projection

Take $g^{\mathrm{SQ}}(\alpha) = (\alpha - C)^2$ so that $\partial g(\alpha) = 2(\alpha - C)$ and $(\partial g)^{-1}(u) = (u + C)/2$. Under the dual linear specification $u_{\boldsymbol{\beta}}(X) = \boldsymbol{\phi}(X)^\top \boldsymbol{\beta}$, the primal model is the affine (linear-link) form

$$\alpha_{\boldsymbol{\beta}}(X) = \frac{\boldsymbol{\phi}(X)^\top \boldsymbol{\beta} + C}{2}. \tag{28}$$

With $\lambda = 0$, the KKT equations (27) become the usual normal equations

$$\mathbb{E}\big[\boldsymbol{\phi}(X)\boldsymbol{\phi}(X)^\top\big]\,\boldsymbol{\beta} = 2\,\mathbb{E}[m(W, \boldsymbol{\phi})] - C\,\mathbb{E}[\boldsymbol{\phi}(X)]\,, \tag{29}$$

which is exactly the "Riesz representer = linear system" form (21). Geometrically, because $\partial g(\alpha)$ is affine, Bregman orthogonality (22) reduces to the standard $L_2(P_X)$ projection property:

$$\langle \alpha_0 - \alpha^*,\, \delta\alpha \rangle = 0 \quad \text{for all feasible directions } \delta\alpha \in T_{\mathcal{H}}(\alpha^*).$$

Hence SQ-Riesz with a linear link is literally an L_2-projection of α_0 onto the linear sieve space.

J.4 (B) KL-type losses + exponential/logit links (UKL/BKL)

For KL-type losses, the same projection geometry holds, but in the *dual* coordinate $u = \partial g(\alpha)$.

UKL-Riesz regression with exponential/log link. Consider the branchwise UKL generator (shifted to avoid singularities) on the domain $|\alpha| > C$:

$$g^{\mathrm{UKL}}(\alpha) = (|\alpha| - C)\log(|\alpha| - C) - |\alpha|, \qquad \partial g(\alpha) = \operatorname{sign}(\alpha)\log(|\alpha| - C).$$

Fix a branch indicator $\xi(X) \in \{0, 1\}$ so that the sign of $\alpha_{\boldsymbol{\beta}}(X)$ is predetermined (e.g., $\xi(X) = D$ in ATE), and impose the dual linear model

$$u_{\boldsymbol{\beta}}(X) = \partial g\big(\alpha_{\boldsymbol{\beta}}(X)\big) = \boldsymbol{\phi}(X)^\top \boldsymbol{\beta}. \tag{30}$$

Inverting ∂g on each branch yields the familiar exponential (log-link) form

$$\alpha_{\boldsymbol{\beta}}(X) = \xi(X)\Big(C + \exp\big(\boldsymbol{\phi}(X)^\top \boldsymbol{\beta}\big)\Big) - \big(1 - \xi(X)\big)\Big(C + \exp\big(-\boldsymbol{\phi}(X)^\top \boldsymbol{\beta}\big)\Big). \tag{31}$$

Despite the nonlinearity in $\boldsymbol{\beta}$, the KKT conditions remain linear *in the test functions*: for $\lambda = 0$ they are exactly the sieve Riesz equations (27). Hence UKL-Riesz returns the Bregman (information) projection solution *subject to* the same Riesz linear equations that define the representer on the sieve.

BKLL-Riesz regression with logit link. For the BKL generator (again on $|\alpha| > C$),

$$g^{\mathrm{BKL}}(\alpha) = (|\alpha| - C)\log(|\alpha| - C) - (|\alpha| + C)\log(|\alpha| + C), \qquad \partial g(\alpha) = \operatorname{sign}(\alpha)\log\Big(\tfrac{|\alpha| - C}{|\alpha| + C}\Big),$$

impose the same dual linear model $u_{\boldsymbol{\beta}}(X) = \boldsymbol{\phi}(X)^\top \boldsymbol{\beta}$ (with a sign branch fixed by $\xi(X)$). Inverting ∂g yields a logit/tanh-type link for the magnitude $|\alpha_{\boldsymbol{\beta}}|$ (and sign controlled by ξ), and the KKT conditions are again (26)–(27). In applications such as ATE, this specialization recovers regularized logistic likelihood (propensity-score MLE) as a particular Bregman–Riesz choice, while still fitting into the same "Bregman projection under an L_2 inner product" template through (22).

J.5 Summary

Both (A) SQ-Riesz + linear link and (B) UKL/BKL + exponential/logit links can be written as the same object:

- Under the dual linear specification $\partial g(\alpha_{\boldsymbol{\beta}}) = \boldsymbol{\phi}^\top \boldsymbol{\beta}$, the KKT conditions reduce to the same sieve Riesz equations (27), i.e., the same "Riesz representer = linear equation" characterization in the formulations of Chen & Liao (2015); Chen & Pouzo (2015).
- The choice of g (squared vs. KL-type) changes which solution is selected among (approximately) balancing solutions: SQ-Riesz, UKL-Riesz, and BKL-Riesz regression.

K Why a Sigmoid Propensity Model Implies UKL-Riesz

In this section, using our automatic covariate balancing result (Section 4), we explain why Zhao (2019)'s "estimand-driven loss selection" implies that, once we commit to a sigmoid (logistic) model for the propensity score, the compatible generalized Riesz regression for estimating the ATE Riesz representer is the UKL-type loss (UKL-Riesz), and using other losses without changing the link breaks the covariate balancing characterization.

K.1 Compatibility beteen Loss choice and Covariate Balancing for the Target Estimand

Zhao (2019) emphasizes that many causal estimands can be written as (or are closely related to) weighted averages of outcomes (our RW estimator), and that the loss used to estimate the weights/propensity score should be chosen so that the resulting fitted weights satisfy the covariate balancing conditions relevant for the estimand. In particular, in ATE estimation, different choices of loss paired with a logistic propensity model correspond to different target weightings (and hence different estimands), and only specific losses deliver covariate balancing for the ATE under the logistic specification.

Our generalized Riesz regression framework makes this principle explicit: automatic covariate balancing arises only when the loss generator g and the link function are paired so that $\partial g(\alpha_{\boldsymbol{\beta}}(X))$ is linear in the features used in the index (Theorem 4.1 and Corollary 4.2).

K.2 Sigmoid Propensity Modeling and a Log Link Function

Consider the usual logistic (sigmoid) propensity score model

$$e_{\boldsymbol{\beta}}(Z) \coloneqq \Lambda\big(\eta_{\boldsymbol{\beta}}(Z)\big), \qquad \eta_{\boldsymbol{\beta}}(Z) \coloneqq \boldsymbol{\phi}(Z)^\top \boldsymbol{\beta}, \qquad \Lambda(t) \coloneqq \frac{1}{1+\exp(-t)}.$$

Then the inverse-propensity components satisfy

$$r_{\boldsymbol{\beta}}(1, Z) \coloneqq \frac{1}{e_{\boldsymbol{\beta}}(Z)} = 1 + \exp\big(-\eta_{\boldsymbol{\beta}}(Z)\big),$$
$$r_{\boldsymbol{\beta}}(0, Z) \coloneqq \frac{1}{1-e_{\boldsymbol{\beta}}(Z)} = 1 + \exp\big(\eta_{\boldsymbol{\beta}}(Z)\big).$$

Therefore, the induced ATE Riesz representer model

$$\alpha_{\boldsymbol{\beta}}^{\mathrm{ATE}}(D,Z) \coloneqq \frac{D}{e_{\boldsymbol{\beta}}(Z)} - \frac{1-D}{1-e_{\boldsymbol{\beta}}(Z)}$$

can be written as the branchwise exponential form

$$\alpha_{\boldsymbol{\beta}}^{\mathrm{ATE}}(D,Z) = D\Big(1+\exp\big(-\eta_{\boldsymbol{\beta}}(Z)\big)\Big) - (1-D)\Big(1+\exp\big(\eta_{\boldsymbol{\beta}}(Z)\big)\Big). \tag{32}$$

This is exactly the "log-link" Riesz representer specification described in Section 4 with $(\xi, C) = (D, 1)$. In particular, (32) implies the sign and domain restrictions

$$\alpha_{\boldsymbol{\beta}}^{\mathrm{ATE}}(1,z) > 1, \qquad \alpha_{\boldsymbol{\beta}}^{\mathrm{ATE}}(0,z) < -1,$$

so the natural shifted domain $|\alpha| > 1$ is compatible with the shifted UKL/BKL generators used in Section 3.

K.3 Automatiuc Covariate Balancng under UKL-Riesz Regression

The automatic covariate balancing theorem (Theorem 4.1) requires that

$$\partial g\big(\alpha_{\boldsymbol{\beta}}(X)\big) \;\text{ is linear in }\; \boldsymbol{\phi}(X)^\top\boldsymbol{\beta},$$

in the sense that it can be written as a linear combination of fixed feature transforms independent of $\boldsymbol{\beta}$.

For ATE with the sigmoid-induced model (32), consider the shifted UKL generator with $C = 1$,

$$g^{\mathrm{UKL}}(\alpha) \coloneqq (|\alpha|-1)\log(|\alpha|-1) - |\alpha|, \qquad \partial g^{\mathrm{UKL}}(\alpha) = \mathrm{sign}(\alpha)\log(|\alpha|-1).$$

Evaluate $\partial g^{\mathrm{UKL}}$ at $\alpha_{\boldsymbol{\beta}}^{\mathrm{ATE}}(D,Z)$. Let $\eta = \eta_{\boldsymbol{\beta}}(Z)$.

Treated branch ($D = 1$). Then $\alpha_{\boldsymbol{\beta}}^{\mathrm{ATE}}(1,Z) = 1+\exp(-\eta)$, so $|\alpha| - 1 = \exp(-\eta)$ and $\mathrm{sign}(\alpha) = +1$, hence

$$\partial g^{\mathrm{UKL}}\Big(\alpha_{\boldsymbol{\beta}}^{\mathrm{ATE}}(1,Z)\Big) = \log\Big(\exp(-\eta)\Big) = -\eta.$$

Control branch ($D = 0$). Then $\alpha_{\boldsymbol{\beta}}^{\mathrm{ATE}}(0,Z) = -(1+\exp(\eta))$, so $|\alpha| - 1 = \exp(\eta)$ and $\mathrm{sign}(\alpha) = -1$, hence

$$\partial g^{\mathrm{UKL}}\Big(\alpha_{\boldsymbol{\beta}}^{\mathrm{ATE}}(0,Z)\Big) = -\log\Big(\exp(\eta)\Big) = -\eta.$$

Key identity. Combining both branches yields the same linear index:

$$\partial g^{\mathrm{UKL}}\Big(\alpha_{\boldsymbol{\beta}}^{\mathrm{ATE}}(D,Z)\Big) = -\eta_{\boldsymbol{\beta}}(Z) = -\boldsymbol{\phi}(Z)^\top\boldsymbol{\beta}. \tag{33}$$

Thus $\partial g^{\mathrm{UKL}}(\alpha_{\boldsymbol{\beta}}^{\mathrm{ATE}}(X))$ is *exactly linear* in the basis $\boldsymbol{\phi}(Z)$. Therefore, the conditions of Theorem 4.1 (and Corollary 4.2) are met for the original covariate features used in the propensity index.

K.4 Resulting Automatic Covariate Balancing

Take ℓ_1-penalized generalized Riesz regression for $\boldsymbol{\beta}$ as in Theorem 4.1 and let $\widehat{\alpha} = \alpha_{\widehat{\boldsymbol{\beta}}}$. Because (33) makes $\partial g(\alpha_{\boldsymbol{\beta}})$ linear in $\boldsymbol{\phi}(Z)^\top \boldsymbol{\beta}$, the KKT conditions imply approximate balancing of the corresponding moments.

To see the standard ATE interpretation, suppose ϕ_j depends only on Z (as in standard propensity modeling), so that

$$m^{\mathrm{ATE}}(W, \phi_j) = \phi_j(1, Z) - \phi_j(0, Z) = 0.$$

Then Corollary 4.2 yields (up to the penalty tolerance)

$$\left| \frac{1}{n} \sum_{i=1}^{n} \widehat{\alpha}(D_i, Z_i)\, \phi_j(Z_i) \right| \leq \lambda \qquad (j = 1, \ldots, p), \tag{34}$$

which is equivalent to the familiar "treated vs. control" balancing condition

$$\frac{1}{n} \sum_{i=1}^{n} \frac{D_i}{\widehat{e}(Z_i)} \phi_j(Z_i) \;\approx\; \frac{1}{n} \sum_{i=1}^{n} \frac{1 - D_i}{1 - \widehat{e}(Z_i)} \phi_j(Z_i),$$

because $\widehat{\alpha}(D, Z) = D/\widehat{e}(Z) - (1 - D)/(1 - \widehat{e}(Z))$. This is precisely the covariate balancing behavior that motivates the ATE-targeted tailored loss choice in Zhao (2019), and it is also consistent with the dual characterization leading to entropy balancing weights (Table 1).

K.5 Why other losses fail to deliver automatic covariate balancing under the *same* sigmoid propensity model

The key requirement behind automatic covariate balancing is *loss–link compatibility*: the link must be (up to branchwise constants) the inverse map of ∂g. When the propensity is parameterized by a sigmoid, the induced Riesz representer (32) is of log-link form, which matches the inverse map of the UKL derivative (Section 4). If we keep the sigmoid model but replace the loss, this compatibility is broken and $\partial g(\alpha_{\boldsymbol{\beta}})$ is no longer linear in the index.

We illustrate this mismatch for two prominent alternatives.

Squared loss (SQ-Riesz) + sigmoid propensity. With $g^{\mathrm{SQ}}(\alpha) = (\alpha - 1)^2$ we have $\partial g^{\mathrm{SQ}}(\alpha) = 2(\alpha - 1)$. Under (32),

$$\partial g^{\mathrm{SQ}}\Big(\alpha^{\mathrm{ATE}}_{\boldsymbol{\beta}}(1, Z)\Big) = 2 \exp\big(-\eta_{\boldsymbol{\beta}}(Z)\big), \qquad \partial g^{\mathrm{SQ}}\Big(\alpha^{\mathrm{ATE}}_{\boldsymbol{\beta}}(0, Z)\Big) = -2\Big(2 + \exp\big(\eta_{\boldsymbol{\beta}}(Z)\big)\Big).$$

These expressions are *not linear* in $\eta_{\boldsymbol{\beta}}(Z) = \boldsymbol{\phi}(Z)^\top \boldsymbol{\beta}$, so the linearity condition in Theorem 4.1 fails. Hence the SQ-Riesz objective does *not* yield the ATE-style balancing equations (34) when we insist on a sigmoid propensity model. (Equivalently: to obtain balancing with squared loss, we must change the link to the linear link discussed in Section 4.)

Logistic MLE (BKL-Riesz) + sigmoid propensity. BKL-Riesz corresponds to Bernoulli likelihood (Section 3). Its generator satisfies

$$\partial g^{\mathrm{BKL}}(\alpha) = \operatorname{sign}(\alpha) \log\left(\frac{|\alpha| - 1}{|\alpha| + 1}\right).$$

Under (32), on the treated branch $|\alpha| - 1 = \exp(-\eta)$ but $|\alpha| + 1 = 2 + \exp(-\eta)$, so

$$\partial g^{\mathrm{BKL}}\left(\alpha_{\boldsymbol{\beta}}^{\mathrm{ATE}}(1, Z)\right) = \log\left(\frac{\exp(-\eta_{\boldsymbol{\beta}}(Z))}{2 + \exp(-\eta_{\boldsymbol{\beta}}(Z))}\right) = -\eta_{\boldsymbol{\beta}}(Z) - \log\left(2 + \exp(-\eta_{\boldsymbol{\beta}}(Z))\right),$$

which is not linear in $\eta_{\boldsymbol{\beta}}(Z)$. Therefore BKL-Riesz (logistic MLE) does not satisfy the automatic covariate balancing conditions for the ATE under the sigmoid specification. This aligns with Zhao (2019)'s discussion: within their tailored-loss family, the logistic likelihood corresponds to a different weighting/estimand than the ATE (see also Remark 7 in the main text).

BP-Riesz and other divergences. The same point applies more broadly: if we keep the sigmoid propensity link (32), then for $\omega \neq 0$ the BP derivative involves powers $(|\alpha| - 1)^{\omega} = \exp(\pm\omega\eta)$ and is not linear in η. Thus BP-Riesz does not yield automatic balancing under the sigmoid link unless one also changes the link to the compatible power link in Section 4.

K.6 Summary

The above calculations show that, under the sigmoid propensity score model, the induced ATE Riesz representer has the branchwise exponential (log-link) form (32), and *only* the UKL generator makes $\partial g(\alpha_{\boldsymbol{\beta}})$ exactly linear in the logistic index $\boldsymbol{\phi}(Z)^{\top}\boldsymbol{\beta}$ (equation (33)). Consequently, by Theorem 4.1, UKL-Riesz is the loss that yields the ATE-relevant automatic covariate balancing equations (34) under sigmoid propensity modeling.

In contrast, using SQ-Riesz, BKL-Riesz (logistic MLE), or BP-Riesz *without changing the link* breaks the loss–link compatibility, so the automatic balancing characterization no longer applies to the original covariate features. Therefore, following Zhao (2019)'s estimand-consistent principle, if the manuscript adopts a sigmoid approximation for the propensity score, then UKL-Riesz is the appropriate generalized Riesz regression objective for ATE-oriented covariate balancing.