Generalized Propensity Score Weighting for Continuous Treatments

Fangzhou Yu*

October, 2025

Abstract

Estimating causal effects of continuous treatments is challenging as traditional methods rely on unstable inverse density weighting. We address this challenge by extending the balancing weights framework from categorical to continuous treatments. We shift the analytical focus from derivative weights to their Riesz Representers, which allows for efficiency analysis even when derivative weights are non-smooth. We derive the optimal balancing weight that minimizes the nonparametric efficiency bound and discover that the resulting optimal estimand corresponds exactly to the projection coefficient in a Partially Linear Regression. This finding establishes a novel causal interpretation for the projection coefficient under heterogeneity and provides an efficiency-based justification for its use in empirical practice.

Keywords: *Continuous treatment effects; Propensity score weighting; Nonparametric efficiency bound*

*School of Economics, University of Sydney. fangzhou.yu@sydney.edu.au

1 Introduction

Estimating the causal effects of continuous treatments or exposures, such as drug dosage, policy intensity, or economic shock exposure, is a fundamental challenge in empirical research. The primary object of interest is often the dose-response curve, $\mathbb{E}[Y_i(d)]$, which characterizes the expected outcome when the entire population is assigned treatment level d.

A substantial body of literature addresses the estimation of this curve. Traditional regression approaches are highly sensitive to the specification of the outcome model. Alternatives utilizing the Generalized Propensity Score (GPS) (Hirano and Imbens, 2004) adjust for confounding by modeling the conditional density of the treatment. However, standard implementations require correct specification of both the GPS model and the subsequent outcome model. Recognizing these limitations, recent work has incorporated flexible, data-adaptive methods. Kreif et al. (2015) utilized machine learning to mitigate model misspecification within the GPS framework. Furthermore, Kennedy et al. (2017) developed methods for the doubly robust estimation of the nonparametric dose-response curve, offering protection if either the outcome or treatment model is correctly specified.

While estimating the full dose-response curve provides a complete picture, it is a statistically challenging task, often characterized by slower, nonparametric convergence rates. In many practical scenarios, such as settings with limited statistical power or when a summary of the main effect is needed, researchers require a scalar summary of the main effect that can be estimated at the faster parametric rate ($n^{-1/2}$). A natural summary for continuous treatments is the average slope or derivative of the dose-response curve. However, estimation of the average derivative effect (ADE) presents significant challenges. Efficient estimation typically requires an inverse density term, $1/f(D|X)$, where D is the treatment and X is the set of covariates. Estimating this density, often via kernel methods, introduces bandwidth-dependent biases and leads to highly unstable estimators when the density is near zero (Cattaneo et al., 2010; Newey and McFadden, 1994).

This challenge mirrors a well-known issue in the categorical treatment literature. The standard Average Treatment Effect (ATE) is often estimated using inverse probability weights, which are unstable when propensity scores are close to zero or one. To address this, researchers developed the "balancing weights" framework (Li et al., 2018; Li and Li, 2019). This framework recognizes that different weighting schemes target different populations. By optimizing these balancing weights to minimize the asymptotic variance of the effect estimator, one can define estimands focused on the population with the most covariate overlap, which are inherently more stable to estimate (Crump et al., 2006).

In this paper, we extend the balancing weights framework to the continuous treatment setting. We generalize the principle of optimizing balancing weights to identify the most statistically efficient way to estimate the average slope of the dose-response curve. By deriving the weights that minimize the nonparametric efficiency bound, we identify a target estimand that facilitates stable estimation while retaining a clear interpretation as a weighted average causal effect.

First, we establish causal identification of our estimands. We formally show that, under standard causal assumptions, the derivative of the observed conditional response function identifies the derivative of the potential outcome dose-response curve. Building on this, we introduce a unifying mathematical structure based on Riesz Representers that characterizes balancing weights for both categorical and continuous treatments. This unified view provides the formal foundation for generalizing optimality results from the categorical setting to the continuous one.

Second, we derive the optimal balancing weight for the average slope, identifying the estimand with the minimum nonparametric efficiency bound. This derivation generalizes the optimality results when the treatment is categorical. We show that when the treatment is binary, our optimal estimand reduces exactly to the overlap weighted average treatment effect (ATO) in Crump et al. (2006), and reduces to Li and Li (2019) when their optimization criterion of minimizing the sum of all pairwise estimators' variances is adopted. Additionally, the estimand with the optimal balancing weight corresponds to a Least Squares Estimand, $\tau_{ls} = \mathbb{E}[Cov(D, Y|X)]/Var(D|X)$, which also arises as the projection parameter in a partially linear regression (PLR). This estimand offers substantial practical advantages, as its Riesz Representer does not involve inverse density terms, avoiding the problematic kernel density estimation required by traditional ADE approaches.

Finally, we connect our optimal weighting framework to the established literature on partially linear models. We highlight that efficient estimators of the optimal estimand are readily available in the traditional nonparametric literature (Robinson, 1988), and by the modern Double Machine Learning (DML) (Chernozhukov et al., 2018). Our contribution here is the novel interpretation of this parameter as the optimally efficient average slope, providing an efficiency-based justification for its use.

The remainder of the paper is structured as follows. Section 2 introduces the balancing weights framework for continuous treatments. Section 3 analyzes the nonparametric efficiency bounds, derives the optimal balancing weight and the corresponding estimand τ^*. Section 4 discusses the estimation of the optimal estimand, and clarifies its causal interpretation. Section 5 illustrates the optimal estimand through an empirical application examining the effect of lottery winnings on labor supply. Section 6 concludes. All proofs and supplementary theoretical results are collected in the Appendices.

2 Balancing Weights for Continuous Treatments

Suppose we have n i.i.d. observations of random variables $O_i = (Y_i, D_i, X_i)$ distributed according to unknown distribution P, where $Y_i \in \mathbb{R}$ is the outcome, treatment $D_i \in \mathbb{R}$ is continuous and $X_i \in \mathbb{R}^p$ is a p-dimensional vector of covariates. We adopt the potential outcome framework. Let $Y_i(d)$ be the potential outcome for unit i under treatment level $D_i = d$.

In the literature, the focus is often the dose-response curve $d \mapsto \mathbb{E}[Y_i(d)]$ and a natural causal parameter of interest is the incremental effect $\left(\mathbb{E}[Y_i(d+\nu)] - \mathbb{E}[Y_i(d)]\right)/\nu$, which characterizes the effect of increasing the treatment level by ν on the expected potential outcome (e.g., Kreif et al., 2015). Estimating the full dose-response curve is statistically challenging. While the incremental effect for fixed ν and d is a scalar quantity, it is not an informative summary of all levels of treatment effect, especially in the presence of heterogeneity.

In this paper, we are interested in scalar parameters that summarize the effect of a continuous treatment across the whole population. Our target parameter is the weighted average causal derivative defined as

$$\dot{\tau}_w = \mathbb{E}[w(D_i, X_i) Y_i'(D_i)]$$

where $w(d, x)$ is a weight function and $Y_i'(d)$ is the derivative of the potential outcome with respect to d.

2.1 Identification

Let $\mu(d,x) = \mathbb{E}[Y_i|D_i = d, X_i = x]$ be the dose-response curve of observed outcome, and its derivative with respect to d is $\mu'(d,x)$. The conditional density of D_i given $X_i = x$ is denoted as $f(d|x)$. We assume the following identification assumption throughout the paper.

Assumption 1. *a. (SUTVA) $D_i = d$ implies $Y_i = Y_i(d)$.*

b. (Unconfoundedness) $Y_i(d) \perp D_i|X_i$ for all d.

c. (Local Overlap) $f(d|x)$ is continuous in d.

d. (Regularity and Smoothness) The potential outcome $Y_i(d)$ is bounded and continuously differentiable in d, and the derivative $Y_i'(d)$ is bounded.

e. (Bounded Weight) $\mathbb{E}[|w(D_i, X_i)|] < \infty$

The Assumptions a and b are extensions of standard identification assumptions to the continuous setting. Assumption c is a relaxed version of the standard overlap assumption, which assumes all $f(d|x)$ are positive. Here, we only require that, if some point (d_0, x_0) is observed ($f(d_0|x_0) > 0$), then there exists a neighborhood around d_0 where $f(d|x_0) > 0$. This guarantees that $\mu(d, x_0)$ is well-defined in the neighborhood of d_0. Assumption d allows us to interchange the expectation and derivative operators, which is a key step in the identification. Assumption e is trivial for economically meaningful $\dot{\tau}_w$.

The next proposition shows that the causal parameter $\dot{\tau}_w$ is identified by the derivative of observed conditional response function $\mu'(d,x)$ under Assumption 1.

Proposition 1 (Identification of $\dot{\tau}_w$)**.** *Under Assumption 1, $\dot{\tau}_w = \mathbb{E}[w(D_i, X_i)\mu'(D_i, X_i)]$.*

In the proof, we show that under Assumption d, the derivative $Y_i'(d)$ represents the limiting incremental effect as the treatment increment approaches zero, formally expressed as $\mathbb{E}[Y_i'(d)] = \lim_{\nu\to 0} \frac{\mathbb{E}[Y_i(d+\nu)] - \mathbb{E}[Y_i(d)]}{\nu}$. This relationship allows us to interpret the weighted average derivative as a causal parameter based on incremental effects. Furthermore, it enables the aggregation of derivative effects across treatment levels through integration of $\mathbb{E}[Y_i(d)|X_i = x]$ over $f(d|x)$. When $D_i \in \{0, 1\}$, the analogous identification result for binary treatment is $\mathbb{E}[w(X_i)(Y_i(1) - Y_i(0))] = \mathbb{E}[w(X_i)(\mu(1, X_i) - \mu(0, X_i))]$, which is developed in the balancing weights literature (Crump et al., 2006; Li et al., 2018).

2.2 A Class of Balancing Weights

Having established the identification result in Proposition 1, we focus on the estimand $\tau_w = \mathbb{E}[w(D_i, X_i)\mu'(D_i, X_i)]$ in the remainder of the paper. We drop the dot notation and refer to τ_w as a statistical estimand, in contrast to the causal parameter $\dot{\tau}_w$.

To analyze the properties of τ_w and develop efficient estimators, we embed this problem within a functional analysis framework. Let $P_{D,X}$ denote the joint distribution of (D_i, X_i). We define $\mathcal{H}$ as the Hilbert space $L_2(P_{D,X})$ of all measurable functions $f : \mathbb{R}^{p+1} \mapsto \mathbb{R}$ that are square integrable with respect to $P_{D,X}$. This space is equipped with inner-product $\langle f, g\rangle = \mathbb{E}_{D,X}[f(D, X)g(D, X)]$ and the induced norm $\|f\| = \langle f, f\rangle^{1/2}$. Assume that $\mu \in \mathcal{H}$.

The parameter τ_w can be viewed as a linear functional of the conditional response function μ. By the Riesz Representation theorem, there exists a unique element $\alpha_w \in \mathcal{H}$, called the Riesz Representer, such that

$$\tau_w = \langle \alpha_w, \mu\rangle = \mathbb{E}[\alpha_w(D_i, X_i)\mu(D_i, X_i)].$$

By the property of conditional expectation, this further implies that $\tau_w = \mathbb{E}[\alpha_w(D_i, X_i)Y_i]$. The function α_w acts as a weight on the observed outcome Y_i. In the literature, α_w is referred to as a balancing weight (Li et al., 2018; Li and Li, 2019), as it directly identifies the target estimand through a weighted average of the outcome.

Traditionally, the estimand is defined by first specifying the derivative weight w. Under specific regularity conditions (see Assumption 2 in Appendix A for details), Powell et al. (1989) used integration by parts to derive the corresponding balancing weight α_w:

$$\alpha_w(d, x) = -\frac{\partial w(d, x)}{\partial d} - w(d, x)\frac{\partial \log f(d|x)}{\partial d}. \tag{1}$$

This result highlights a significant challenge in the continuous treatment setting. If we start by defining w, the corresponding α_w often depends on the density $f(d|x)$ and its derivative. These are difficult to estimate nonparametrically, leading to unstable estimators (Cattaneo et al., 2010; Newey and McFadden, 1994).

This challenge motivates a shift in perspective. Instead of starting with w and deriving the complex α_w, we propose defining the estimand by specifying the balancing weight $\alpha \in \mathcal{H}$

directly. This allows us to select forms of α that facilitate efficient estimation, potentially avoiding the density estimation issues in Equation (1).

The critical question then becomes: What properties must α satisfy to ensure that the resulting estimand $\mathbb{E}[\alpha(D_i, X_i)\mu(D_i, X_i)]$ represents a meaningful weighted average derivative, free from confounding? We identify two essential constraints. First, the estimand must correctly measure the scale of the treatment effect. This requires the normalization constraint $\mathbb{E}[\alpha(D_i, X_i)D_i] = 1$, which ensures the estimand is a properly weighted average, analogous to $\mathbb{E}[w(D_i, X_i)] = 1$ for the derivative weights[1]. Second, the estimand must control for confounding. This requires that the weight α must be orthogonal to the covariates. We formalize this as the orthogonality condition $\mathbb{E}[\alpha(D_i, X_i)|X_i] = 0$.

This orthogonality condition is a direct generalization of the concept of "balancing weights" established in the categorical treatment literature (Li et al., 2018). In the binary setting, balancing weights (w_1 for treated, w_0 for control) are defined by their ability to equalize the weighted distributions of covariates across groups. The Riesz Representer α represents the resulting contrast $\alpha = D_i w_1 - (1 - D_i)w_0$. The distributional balance can be expressed as the orthogonal condition $\mathbb{E}[\alpha|X_i] = 0$[2]. While achieving full distributional balance across a continuum of treatment levels is complex, the orthogonality condition provides a unified and tractable definition of balance for both categorical and continuous treatments.

We therefore define the class of generalized balancing weights as the set of Riesz Representers satisfying these two essential properties:

$$\mathcal{A} = \{\alpha \in \mathcal{H} | \mathbb{E}[\alpha(D_i, X_i)D_i] = 1, \mathbb{E}[\alpha(D_i, X_i)|X_i] = 0\}. \tag{2}$$

This class $\mathcal{A}$ provides a unified framework. For instance, when $D_i \in \{0, 1\}$ with propensity score $e(x) = P(D_i = 1|X_i = x)$, the inverse propensity weight used to estimate the ATE, $\alpha_{w(x)=1} = D_i/e(X_i) - (1 - D_i)/(1 - e(X_i))$, satisfies both conditions.

The next proposition demonstrates that the class $\mathcal{A}$ fully characterizes the set of weighted average derivative estimands τ_w with normalized weights.

Proposition 2. *Let $F(d|x)$ be the conditional CDF of D_i given $X_i = x$. For any balancing weight $\alpha \in \mathcal{A}$, the estimand $\mathbb{E}[\alpha(D_i, X_i)Y_i]$ is a normalized weighted average derivative τ_w*

[1] This is formally proved in the proof of Proposition 2 in Appendix A.

[2] For binary treatment, the orthogonality condition $\mathbb{E}[\alpha|X_i] = 0$ holds under the balancing property $w_1(x) \propto w(x)/e(x)$, $w_0(x) \propto w(x)/(1 - e(x))$ in Li et al. (2018), where $e(x)$ is the propensity score.

(i.e., $\mathbb{E}[w(D_i, X_i)] = 1$*), with the implied derivative weight:*

$$w(d,x) = -\frac{F(d|x)}{f(d|x)} \cdot \mathbb{E}[\alpha(D_i, X_i)|D_i \leq d, X_i = x]. \tag{3}$$

Proposition 2, combined with the identification result in Proposition 1, confirms that any estimand defined by $\alpha \in \mathcal{A}$ identifies a normalized weighted average causal derivative $\dot{\tau}_w$. However, normalized $w(d,x)$ is often not enough for τ_w to be interpreted as a causal parameter. Following the recent discussion in the literature of difference-in-differences, another requirement is that $w(d,x)$ is non-negative (e.g., Goodman-Bacon, 2021; Sun and Abraham, 2021). We provide a sufficient and necessary condition for $\alpha \in \mathcal{A}$ to identify non-negative $w(d,x)$.

Lemma 1. *For* w *and* α *in Equation (3),*

$$w(d,x) \geq 0 \textit{ if and only if } \mathbb{E}[\alpha(D_i, X_i)|D_i \leq d, X_i = x] \leq 0 \textit{ for all } d, x.$$

For the average $\mathbb{E}[\alpha(D_i, X_i)|D_i \leq d, X_i = x]$ to be non-positive when d is near the lower boundary of the support, $\alpha(d,x)$ must generally be negative. For the overall average $\mathbb{E}[\alpha(D_i, X_i)|X_i]$ to return to zero when d is near the upper boundary of the support, $\alpha(d,x)$ must generally be positive. Therefore, the condition in Lemma 1 enforces a structural requirement that $\alpha(d,x)$ must transition from negative to positive as d increases.

One illustrative example is the ATE when $D_i \in \{0,1\}$, with the corresponding balancing weight $\alpha_{w(x)=1} = (D_i - e(X_i))/\left(e(X_i)(1-e(X_i))\right)$. Under the standard overlap assumption, the denominator is positive, while the numerator takes negative values when $D_i = 0$ and positive values when $D_i = 1$, naturally satisfying the sign transition requirement. For the average derivative in the continuous setting, Powell et al. (1989) derived the corresponding balancing weight $\alpha_{w(d,x)=1} = -f'(d|x)/f(d|x)$. For this case, we can verify that $\mathbb{E}[\alpha(D_i, X_i)|D_i \leq d, X_i = x] = -f(d|x)/F(d|x)$ is non-positive since both the density and cumulative distribution function are positive.

In this paper, we study the class of balancing weights in $\mathcal{A}$ rather than directly imposing the condition in Lemma 1. This is because allowing for negative weights facilitates the decomposition of existing estimands in a causal setting. Notably, the two-way fixed effect estimand has been shown to exhibit negative weights under staggered treatment adoption, and recent literature has examined the causal weighting scheme of OLS and 2SLS in the presence of covariates or multiple treatments (Blandhol et al., 2022; Goldsmith-Pinkham et al., 2022; Borusyak and Hull, 2024). Our balancing weight framework provides a complementary analytical tool to these works.

3 Efficiency Bounds and Optimal Balancing Weights

3.1 Optimization Criterion

The class of balancing weights $\mathcal{A}$ is large, and Proposition 2 confirms that any $\alpha \in \mathcal{A}$ identifies a normalized effect τ_w. This raises a crucial question: which balancing weight should be selected? Following the established approach in the balancing weights literature (Crump et al., 2006; Li and Li, 2019), we seek the balancing weight that is optimal, in the sense of minimizing the nonparametric efficiency bound.

The efficiency bound characterizes the lowest possible asymptotic variance achievable by any regular asymptotically linear (RAL) estimator. For an estimand defined by a balancing weight α, this bound is given by the variance of its influence function. When the derivative weight $w(d,x)$ (and thus the balancing weight α) is known, the influence function of τ_w has been derived by Newey and Stoker (1993)

$$\psi(O_i) = \underbrace{\alpha_w(D_i, X_i)(Y_i - \mu(D_i, X_i))}_{\psi_A(O_i)} + \underbrace{w(D_i, X_i)\mu'(D_i, X_i) - \tau_w}_{\psi_B(O_i)}. \tag{4}$$

Given the influence function and an efficient estimator $\hat{\tau}_w$, we have

$$\sqrt{n}(\hat{\tau}_w - \tau_w) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(O_i) + o_p(1).$$

The asymptotic variance of $\hat{\tau}_w$ is then given by

$$V_\psi = \mathbb{E}[\psi^2(O_i)] = \mathbb{E}[\psi_A^2(O_i)] + \mathbb{E}[\psi_B^2(O_i)]$$

as $\mathbb{E}[\psi_A(O_i)\psi_B(O_i)] = 0$, which is a consequence of the orthogonality between the residual term $Y_i - \mu(D_i, X_i)$ and the conditional effects. $\mathbb{E}[\psi_A^2(O_i)]$ represents the variance contribution from the noise in the outcome amplified by the magnitude of the balancing weights α_w. $\mathbb{E}[\psi_B^2(O_i)]$ is a measure of effect heterogeneity, which represents the variance of the weighted individual derivative effects, $w(D_i, X_i)\mu'(D_i, X_i)$, around the population average τ_w.

While minimizing the total variance might seem intuitive, selecting w by minimizing the component $\mathbb{E}[\psi_B^2(O_i)]$ is conceptually and practically problematic. First, $\mathbb{E}[\psi_B^2(O_i)]$ is a function of τ_w, the true population derivative effect. If the goal were to select weight by minimizing the total variance, the definition of the optimal weight would depend on the unknown value of the estimand itself. This conceptual circularity is also documented in Crump et al. (2006) for binary treatment setting. Second, since $\mathbb{E}[\psi_B^2(O_i)]$ measures effect heterogeneity, if we optimized this quantity, we would be prioritizing weights that make $w(D_i, X_i)\mu'(D_i, X_i)$ as constant as possible. This fundamentally changes the objective of seeking for statistical efficiency.

To address this issue and optimize for statistical efficiency, we follow the strategy in Crump et al. (2006) and Li and Li (2019) to redefine the objective as minimizing the asymptotic variance of the estimator relative to the sample analogue of τ_w. Specifically, we define the sample analogue of τ_w as

$$\tau_{w,S} = \frac{1}{n}\sum_{i=1}^{n} w(D_i, X_i)\mu'(D_i, X_i).$$

and we have

$$\sqrt{n}(\hat{\tau}_w - \tau_{w,S}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \psi_A(O_i) + o_p(1).$$

The asymptotic variance of $\sqrt{n}(\hat{\tau}_w - \tau_{w,S})$ is the expectation of the square of the leading term

$$V_S = \mathbb{E}[\psi_A^2(O_i)] = \mathbb{E}[\alpha_w^2(D_i, X_i)(Y_i - \mu(D_i, X_i))^2],$$

which is therefore the efficiency bound for the class of τ_w conditional on the realized distribution of the sample.

3.2 Optimal Balancing Weight

We select the optimal balancing weight as the one that minimizes V_S. The resulting optimal estimand τ^* has the lowest efficiency bound among the class of weighted average derivative effects τ_w.

Theorem 1 (Optimal Balancing Weight). *Assume Y_i is homoscedastic, i.e., $Var(Y_i|D_i, X_i)$ is a constant. The optimal balancing weight that minimizes the efficiency bound V_S, subject to $\alpha \in \mathcal{A}$ is given by*

$$\alpha^*(d, x) = \frac{d - \mathbb{E}[D_i|X_i = x]}{\mathbb{E}[Var(D_i|X_i)]}. \tag{5}$$

and thus, the optimal estimand is given by

$$\tau^* = \frac{\mathbb{E}[Cov(D_i, Y_i|X_i)]}{\mathbb{E}[Var(D_i|X_i)]}. \tag{6}$$

Theorem 1 establishes the optimal balancing weight in the continuous setting. Importantly, the optimization framework used to derive this result provides a unified perspective that incorporates existing results for categorical treatments. When the treatment is binary ($D_i \in \{0, 1\}$), the optimization criterion remains the minimization of V_S. The general form of the optimal balancing weight α^* derived in Theorem 1 still applies. By noting that $\mathbb{E}[D_i|X_i] = e(X_i)$ (the propensity score) and $Var(D_i|X_i) = e(X_i)(1 - e(X_i))$, the optimal balancing weight simplifies to:

$$\alpha^*(D_i, X_i) = \frac{D_i - e(X_i)}{\mathbb{E}[e(X_i)(1 - e(X_i))]}.$$

This corresponds exactly to the estimand targeted by the overlap weights, confirming the optimality results established in Theorem 5.1 of Crump et al. (2006) and Li et al. (2018). The interpretation aligns precisely with the intuition behind overlap weights in the binary setting: the estimand focuses on the population where the conditional variance of the treatment, $Var(D_i|X_i)$, is highest, leading to the most stable estimation.

When considering multiple treatments, the objective shifts from a single estimand to all pairwise contrasts. As demonstrated in the proof of Theorem 1 in Appendix A, if we adapt the optimization criterion to minimize the sum of the asymptotic variances for all pairwise comparisons as proposed by Li and Li (2019), our framework successfully recovers their optimal solution. Under homoscedasticity, this optimal weight corresponds to the harmonic mean of the generalized propensity scores, known as the Generalized Overlap Weights. Thus, the proposed framework unifies the derivation of optimal balancing weights across different treatment types.

3.3 When Weights are Unknown

Theorem 1 characterizes the optimal balancing weight by minimizing the efficiency bound V_S, assuming the balancing weight α is known. In practice, weights often depend on the underlying distribution P and must be estimated. We now develop a generalized framework to derive the influence function for τ_w when the balancing weights depend on estimated nuisance parameters.

In the continuous treatment setting, the relationship between the balancing weight α and the derivative weight w involves the conditional density $f(d|x)$. Since the density is generally not a smooth functional of P, traditional approaches that rely on the smoothness of w (e.g., Crump et al., 2006) are often inapplicable. We overcome this challenge by focusing on the smoothness of the balancing weight α. Even if w is non-smooth, the estimand $\tau_w = \mathbb{E}[\alpha\mu]$ can remain regular provided that α itself depends smoothly on P through conditional or marginal expectations.

Theorem 2 (Influence Function with Estimated Weights)**.** *Let $\tau_w = \mathbb{E}[\alpha(D_i, X_i)\mu(D_i, X_i)]$ be the target estimand. Assume the balancing weight $\alpha(d, x)$ depends smoothly on the distribution P through a set of nuisance parameters. For illustration, we consider:*

$$\begin{aligned} \eta_X(X_i) &= \mathbb{E}[H_X(O_i)|X_i] \\ \theta &= \mathbb{E}[G(O_i; \eta_X)] \end{aligned}$$

Assume $\alpha(d, x) = A(d, x; \eta_X(x), \theta)$ for a known, differentiable function A.

The influence function for τ_w is given by:

$$\phi(O_i) = \psi(O_i) + \phi_{adj}(O_i),$$

where $\psi(O_i) = \alpha(Y_i - \mu) + (\alpha\mu - \tau_w)$ is the influence function in Equation (4), and the *adjustment term is* $\phi_{adj} = \phi_{adj,X} + \phi_{adj,\theta}$*, where*

$$\begin{aligned}\phi_{adj,X}(O_i) &= R_X(X_i)(H_X(O_i) - \eta_X(X_i)) \\ \phi_{adj,\theta}(O_i) &= R_\theta \cdot \phi_\theta(O_i) \\ R_X(X) &= \mathbb{E}[\partial_{\eta_X} A \cdot \mu | X] \\ R_\theta &= \mathbb{E}[\partial_\theta A \cdot \mu]\end{aligned}$$

and $\phi_\theta(O_i)$ *is the influence function of* θ*.*

Theorem 2 provides a framework for analyzing efficiency when the weight is unknown and highlights several key insights. First, the theorem establishes that if the balancing weight α is smooth, the estimand τ_w is regular. This is crucial because the corresponding derivative weight w might be non-smooth. For example, the optimal balancing weight α^* is smooth, ensuring the regularity of τ^*, while the corresponding w as in Equation (3) is not smooth as it depends on $f(d|x)$.

Second, the efficiency bound is $V_\phi = \mathbb{E}[\phi^2]$. The adjustment term ϕ_{adj} is generally correlated with the fixed-weight influence function ψ. Therefore, the efficiency bound decomposes as:

$$V_\phi = V_\psi + V_{adj} + 2C_{\psi,adj},$$

where $V_\psi = \mathbb{E}[\psi^2]$, $V_{adj} = \mathbb{E}[\phi_{adj}^2]$, and $C_{\psi,adj} = \mathbb{E}[\psi\phi_{adj}]$ is the covariance term. In the categorical setting, specific structures often ensure that the adjustment Riesz Representers are orthogonal to ψ, leading to $C_{\psi,adj} = 0$. This results in $V_\phi = V_\psi + V_{adj}$, implying that estimating weights increases the variance ($V_\phi \geq V_\psi$)[3]. In the continuous setting, however, the covariance $C_{\psi,adj}$ can be non-zero.

Lastly, the generalized efficiency bound V_ϕ depends explicitly on the outcome model μ and the unknown estimand τ_w through the adjustment terms and the covariance $C_{\psi,adj}$. Attempting to minimize V_ϕ would lead to a circular definition of the optimal weight and prioritize populations based on heterogeneity. This reinforces the strategy in Section 3.1. To identify the estimand that maximizes statistical precision, the optimization should focus solely on the noise component V_S.

[3]See Appendix B for this special case, which generalizes Theorem 5.1 of Crump et al. (2006)

4 Estimation and Discussion of the Optimal Estimand

The optimal balancing weight derived in Theorem 1 identifies a target estimand, τ^*, which possesses the minimum nonparametric efficiency bound component V_S among the class of weighted average derivative effects. We now discuss the estimation of this optimal estimand, its connection to established methods, and its interpretation.

4.1 The Influence Function of Optimal Estimand

Recall the optimal estimand from Equation (6):

$$\tau^* = \frac{\mathbb{E}[Cov(D_i, Y_i|X_i)]}{\mathbb{E}[Var(D_i|X_i)]}.$$

This estimand offers a significant practical advantage over the standard ADE. As shown in Equation (1), the Riesz Representer for a general weighted average derivative effect typically depends on the conditional density $f(d|x)$ and its derivative, leading to unstable estimators (Cattaneo et al., 2010; Newey and McFadden, 1994). In contrast, the optimal balancing weight $\alpha^*(d, x)$ in Equation (5) does not depend on the density $f(d|x)$.

Furthermore, τ^* corresponds exactly to the projection parameter τ in a partially linear regression (PLR):

$$Y_i = \tau D_i + g(X_i) + \epsilon_i, \tag{7}$$

for some function $g(x)$. This connection allows us to leverage robust estimation techniques developed in the semiparametric literature (e.g., Robinson, 1988; Chernozhukov et al., 2018).

To see this connection, we utilize a "partialling out" representation. Let us define the generalized propensity score $e(x) = \mathbb{E}[D_i|X_i = x]$ and the conditional outcome mean $\rho(x) = \mathbb{E}[Y_i|X_i = x]$. Let $K = \mathbb{E}[Var(D_i|X_i)]$. We can rewrite the estimand τ^* in terms of the residuals from the projections of Y_i and D_i onto X_i:

$$\tau^* = \frac{\mathbb{E}[(D_i - e(X_i))(Y_i - \rho(X_i))]}{K}. \tag{8}$$

We apply Theorem 2 to derive the influence function ϕ^* of τ^*. This involves calculating the influence function assuming fixed weights, ψ, and adding adjustment terms (ϕ_{adj}) that account for the sensitivity of α^* to the estimation of $e(X_i)$ and K. The detailed derivation is provided in Appendix B. The resulting influence function is:

$$\phi^*(O_i) = \frac{(D_i - e(X_i))(Y_i - \rho(X_i)) - \tau^*(D_i - e(X_i))^2}{K}. \tag{9}$$

This result confirms that the influence function derived from Theorem 2 exactly matches the existing influence function in the PLR literature.

4.2 Estimation and Interpretation of Optimal Estimand

The representation in Equation (8) immediately suggests an estimation strategy based on orthogonalization, popularized as Double Machine Learning (DML) (Chernozhukov et al., 2018). We construct an estimator $\hat{\tau}^*$ by replacing the population expectations with sample analogues and the unknown nuisance functions $e(x)$ and $\rho(x)$ with their nonparametric estimates $\hat{e}(x)$ and $\hat{\rho}(x)$

$$\hat{\tau}^* = \frac{\sum_{i=1}^n (D_i - \hat{e}(X_i))(Y_i - \hat{\rho}(X_i))}{\sum_{i=1}^n (D_i - \hat{e}(X_i))^2}. \tag{10}$$

When the treatment is binary, this estimator coincides with the efficient estimator for the overlap weighted average treatment effect (ATO) discussed in Crump et al. (2006).

Crucially, ϕ^* satisfies the Neyman orthogonality condition with respect to the nuisance parameters $e(x)$ and $\rho(x)$. This orthogonality arises because the Gâteaux derivatives of $\mathbb{E}[\phi^*(O_i)]$ with respect to these nuisance functions are zero when evaluated at the true values.

The significance of Neyman orthogonality is twofold. First, it implies that the estimation error in the nuisance functions has only a second-order effect on the estimation of τ^*. This local robustness permits the use of flexible machine learning methods for $\hat{e}$ and $\hat{\rho}$, provided they converge sufficiently fast, e.g., at a rate of $n^{-1/4}$.

Second, Neyman orthogonality is crucial for achieving efficiency. As demonstrated by the derivation above using Theorem 2, the influence function ϕ^* includes adjustment terms $\phi_{adj,e}$ and $\phi_{adj,K}$ that account for the estimation uncertainty of the nuisance parameters in

α^*. The DML estimator in Equation (10) is constructed based on the PLR representation, whose score function embodies these adjustments automatically. The Neyman orthogonality of this score ensures that the DML estimator achieves the efficiency bound $V = \mathbb{E}[(\phi^*(O_i))^2]$ without requiring estimation of the complex, non-smooth derivative weights w^*.

While Neyman orthogonality mitigates the bias from estimating nuisance functions, achieving $\sqrt{n}$-consistency and asymptotic normality requires careful implementation when using flexible machine learning estimators. Traditional approaches often relied on restrictive Donsker conditions to control the complexity of the estimated functions (Robins et al., 1994; Newey and McFadden, 1994), which may not hold for methods like random forests or deep learning. To address this limitation, the DML framework utilizes cross-fitting (Chernozhukov et al., 2018). By estimating the nuisance functions on an independent auxiliary sample and constructing the final estimator on the main sample, cross-fitting bypasses the need for the Donsker condition.

We summarize the conditions required for the root-n consistency and asymptotic normality of $\hat{\tau}^*$, unifying insights from this literature.

Proposition 3. *Assume that the following conditions hold:*

a. $\|e - \hat{e}\| = o_P(n^{-a/4})$ *and* $\|\rho - \hat{\rho}\| = o_P(n^{-b/4})$ *where* $a \geqslant 1$, $b \geqslant 0$ *and* $a + b \geqslant 2$. *Furthermore,* $\|e - \hat{e}\|_4 = o_P(1)$ *and* $\|\rho - \hat{\rho}\|_4 = o_P(1)$.

b. *There exists a constant* $C_1 < \infty$ *such that* $\mathbb{E}[(D_i - e(X_i))^2] > C_1 > 0$ *and* $n^{-1}\sum_{i=1}^{n}(D_i - \hat{e}(X_i))^2 > C_1$ *with probability approaching 1.*

c. *There exists a constant* $C_2 < \infty$ *such that* $Var(D_i|X_i) < C_2$ *and* $Var(Y_i|X_i) < C_2$.

And assume that at least one of d or e holds:

d. *The quantities* $(D_i - \hat{e}(X_i))(Y_i - \hat{\rho}(X_i))$ *and* $(D_i - \hat{e}(X_i))^2$ *fall within a P-Donsker class with probability approaching* 1.

e. *The sample used to estimate* $\hat{e}(x)$ *and* $\hat{\rho}(x)$ *is independent of the sample used to construct* $\hat{\tau}^*$.

Then, $\hat{\tau}^*$ is a RAL estimator of τ^* with influence function ϕ^*. Hence,

$$\sqrt{n}(\hat{\tau}^* - \tau^*) \xrightarrow{d} N(0, V),$$

where $V = \mathbb{E}[(\phi^*(O_i))^2]$.

While regression methods for continuous treatment effects are widely used in empirical research, there is limited discussion on the causal identification of the estimands. A key contribution of this paper is providing a formal causal interpretation for the PLR projection parameter τ^* or τ in Equation (7). In the literature, τ is often interpreted as a causal effect under the strong assumption of homogeneity, i.e., PLR is correctly specified and $\mu'(d,x) = \tau$ as in Chernozhukov et al. (2018). Our results demonstrate that τ retains a clear causal interpretation even in the presence of heterogeneous treatment effects. We do not assume Equation (7) is the true data generating process. By combining the identification result in Proposition 1 and the optimality result in Theorem 1, we established that τ identifies a specific weighted average derivative effect with its balancing weight given in Equation (5). This finding provides a robust, efficiency-based justification for the common practice of using PLR to estimate the effect of continuous treatments.

5 Application: The Income Effect on Labor Supply

We illustrate our proposed optimal estimand τ^* by re-examining the effect of unearned income on labor supply, utilizing the dataset from a survey of Massachusetts lottery winners analyzed by Imbens et al. (2001) and Hirano and Imbens (2004) (hereafter HI04). The objective is to estimate the Marginal Propensity to Earn (MPE) out of unearned income, which corresponds to the average derivative of labor earnings with respect to the lottery prize.

The dataset consists of 237 individuals who won the Megabucks lottery in the mid-1980s. The continuous treatment D_i is the annualized value of the lottery prize, and the outcome Y_i is the average labor earnings recorded by the Social Security Administration approximately six years after winning. The covariates X_i include demographic characteristics (age, gender, education) and six years of pre-lottery earnings.

While the lottery prize itself is randomly assigned, the study design introduced potential confounding due to substantial nonresponse (around 50%). HI04 demonstrated that non-response was correlated with the prize amount: winners of larger prizes were less likely to participate in the survey. This induced selection bias in the observed sample. For instance, men and individuals with higher pre-lottery earnings tended to have won larger prizes among the respondents. Following the literature, we maintain the assumption that conditional on the extensive set of covariates X_i, the treatment assignment is unconfounded.

HI04 addressed this setting by introducing the Generalized Propensity Score (GPS) methodology to estimate the entire dose-response curve, $\mu(d) = \mathbb{E}[Y_i(d)]$. Their analysis revealed significant heterogeneity in the MPE, $\mu'(d)$. They estimated that the MPE ranged from approximately -0.10 for small prizes (\$10,000) to -0.02 for larger prizes (\$100,000), suggesting the negative income effect is substantially stronger at lower levels of unearned income.

We compare two estimands for summarizing the entire dose-response curve: the standard ADE, denoted by τ_1, and the optimally efficient estimand τ^*. We estimate the ADE using the GPS methodology, adopting the parametric specifications of log-normal linear GPS model employed by HI04. The estimate $\hat{\tau}_1$ is the sample average of the derivatives $\hat{\mu}'(D_i)$. Standard errors are obtained via 1000 bootstrap replications.[4]

We estimate τ^* using the estimator defined in Equation (10). We implement two versions. The GPS version estimates $e(X_i) = \mathbb{E}[D_i|X_i]$ based on the log-normal GPS specification from HI04 and estimates $\hat{\rho}(X_i)$ using OLS, assuming a linear relationship between earnings and covariates. In the DML version, $e(X_i)$ and $\rho(X_i)$ are estimated using Random Forests[5], implemented with 5-fold cross-fitting to ensure condition e in Proposition 3. For both versions, standard errors are derived from the asymptotic distribution established in Proposition 3.

Table 1 presents the estimation results. The GPS estimate of the ADE is -0.0308. The GPS estimate of the optimal estimand τ^* is -0.0402, while the DML estimate using Random Forest is -0.0253. All estimates suggest a negative income effect, although the magnitudes vary slightly depending on the weighting scheme and nuisance function estimation.

[4]We use the data provided in the GitHub repository of Imbens and Xu (2024). We failed to reproduce the results of HI04, as the summary statistics of pre-lottery earnings are slightly different from those reported in HI04. Our replication results are provided in Appendix C.

[5]We use `DoubleML` package in `R` to implement the DML estimator, and `ranger` package for Random Forest estimation of the nuisance functions, with 500 trees and maximum depth of 5.

Table 1: Estimates of the Marginal Propensity to Earn (MPE) out of Lottery Prizes

Estimand	Method	Estimate	Std. Error	95% CI
ADE (τ_1)	GPS	-0.0308	(0.0270)	[-0.1069, -0.0001]
Optimal (τ^*)	GPS	-0.0402	(0.0125)	[-0.0648, -0.0156]
Optimal (τ^*)	DML	-0.0253	(0.0107)	[-0.0463, -0.0044]

The key advantage of our framework is highlighted by comparing the statistical efficiency of the estimators. The GPS estimator for the optimal estimand achieves a substantially tighter confidence interval compared to the GPS estimator for the ADE, representing around 50% improvement in efficiency. This efficiency gain is the direct result of the optimality established in Theorem 1. Flexible estimation of the nuisance functions via Random Forest further improves the efficiency.

Furthermore, τ^* provides a transparent interpretation of the population it targets. It is a weighted average derivative where the weights are proportional to the conditional variance of the treatment, $Var(D_i|X_i)$. This optimally prioritizes individuals for whom the prize amount is most variable given their characteristics. To interpret this weighting in the context of the lottery data, we consider the structure of $Var(D_i|X_i)$. HI04 modeled the log of the prize as homoscedastic: $\log(D_i)|X_i \sim N(\beta' X_i, \sigma^2)$. Under this log-normal structure, the conditional variance of the prize is heteroscedastic and proportional to $\exp(2\beta' X_i)$. This implies the variance is higher for individuals whose characteristics are associated with larger expected prizes.

Since HI04 found that men and individuals with higher pre-lottery earnings tended to have won larger prizes in the sample, due to the nonresponse bias, the optimal estimand τ^* places greater weight on these groups. Our estimate primarily reflects the income effect for this specific subpopulation. This application demonstrates how our framework provides an efficiency-based causal estimand, clarifying the population for which the causal effect is being summarized, even in the presence of significant heterogeneity.

6 Conclusion

Estimating the causal effects of continuous treatments requires summarizing the dose-response curve in a way that is both interpretable and statistically stable. This endeavor faces two challenges. First, traditional summaries, such as the average derivative effect, are often statistically unstable because their estimation relies on the non-smooth conditional density function. Second, while the literature offered a path to optimality in categorical settings, generalizing it to continuous treatments presented a significant theoretical hurdle. Standard efficiency theories rely on the smoothness of derivative weights w, yet this condition frequently fails in continuous settings.

This paper overcomes this challenge by shifting the analytical focus from the derivative weights to their Riesz Representers α, recognized as the balancing weights. Our key insight is that the estimand remains regular if the balancing weight α is smooth, even if the corresponding derivative weight w is not. This provides a foundation for efficiency analysis in the continuous setting and unifies optimality results across different treatment types.

Our central finding is the derivation of the optimal estimand, τ^*, which minimizes the nonparametric efficiency bound. Crucially, we show that τ^* corresponds exactly to the projection coefficient in a Partially Linear Regression. This connection encourages the use of regression tools, such as Double Machine Learning, for estimating continuous treatment effects. While the PLR coefficient is widely used in empirical works, its causal interpretation under heterogeneity has often been ambiguous in the literature. We bridge this gap by providing a clear causal interpretation: τ^* weights the population based on the conditional variance of the treatment, focusing on the subset where the effect is most reliably estimated.

References

Blandhol, C., J. Bonney, M. Mogstad, and A. Torgovitsky (2022): "When is TSLS actually late?" Tech. rep., National Bureau of Economic Research.

Borusyak, K. and P. Hull (2024): "Negative weights are no concern in design-based specifications," in *AEA Papers and Proceedings*, American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, vol. 114, 597–600.

Cattaneo, M. D., R. K. Crump, and M. Jansson (2010): "Robust data-driven inference for density-weighted average derivatives," *Journal of the American Statistical Association*, 105, 1070–1083.

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

Crump, R. K., V. J. Hotz, G. Imbens, and O. Mitnik (2006): "Moving the goalposts: Addressing limited overlap in the estimation of average treatment effects by changing the estimand," .

Goldsmith-Pinkham, P., P. Hull, and M. Kolesár (2022): "Contamination bias in linear regressions," Tech. rep., National Bureau of Economic Research.

Goodman-Bacon, A. (2021): "Difference-in-differences with variation in treatment timing," *Journal of Econometrics*, 225, 254–277.

Hirano, K. and G. Imbens (2004): "The propensity score with continuous treatments," .

Imbens, G. W., D. B. Rubin, and B. I. Sacerdote (2001): "Estimating the effect of unearned income on labor earnings, savings, and consumption: Evidence from a survey of lottery players," *American economic review*, 91, 778–794.

Imbens, G. W. and Y. Xu (2024): "Comparing Experimental and Nonexperimental Methods: What Lessons Have We Learned Four Decades After LaLonde (1986)?" *Available at SSRN 4849285*.

Kennedy, E. H., S. Balakrishnan, and M. G'sell (2020): "Sharp instruments for classifying compliers and generalizing causal effects," *The Annals of Statistics*.

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

Kreif, N., R. Grieve, I. Díaz, and D. Harrison (2015): "Evaluation of the effect of a continuous treatment: a machine learning approach with an application to treatment for traumatic brain injury," *Health economics*, 24, 1213–1228.

Li, F. and F. Li (2019): "Propensity score weighting for causal inference with multiple treatments," *Annals of Applied Statistics*.

Li, F., K. L. Morgan, and A. M. Zaslavsky (2018): "Balancing covariates via propensity score weighting," *Journal of the American Statistical Association*, 113, 390–400.

Newey, W. K. and D. McFadden (1994): "Large sample estimation and hypothesis testing," *Handbook of econometrics*, 4, 2111–2245.

Newey, W. K. and T. M. Stoker (1993): "Efficiency of weighted average derivative estimators and index models," *Econometrica: Journal of the Econometric Society*, 1199–1223.

Powell, J. L., J. H. Stock, and T. M. Stoker (1989): "Semiparametric estimation of index coefficients," *Econometrica: Journal of the Econometric Society*, 1403–1430.

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

Robinson, P. M. (1988): "Root-N-consistent semiparametric regression," *Econometrica: Journal of the Econometric Society*, 931–954.

Sun, L. and S. Abraham (2021): "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects," *Journal of Econometrics*, 225, 175–199.

Van der Vaart, A. (1998): "Functional delta method," *Asymptotic Statistics*, 291–303.

A Appendix: Proofs

Note: For notational simplicity, we omit individual subscripts in this appendix, writing X instead of X_i, Y instead of Y_i, D instead of D_i, etc.

Assumption 2 (Regularity Condition for RR (Powell et al., 1989))**.** *a. The weighted conditional density $w(d,x)f(d|x)$ is differentiable with respect to d.*

b. $w(\underline{d},x)f(\underline{d}|x) = w(\bar{d},x)f(\bar{d}|x) = 0$ where $\underline{d}$ and $\bar{d}$ are the lower and upper bounds of the support of D.

c. $f(d|x) = 0$ implies $w(d,x) = 0$.

Proof of Proposition 1. Conditioning on the treatment and covariates, we apply the law of iterated expectation to write

$$\begin{aligned}\dot{\tau}_w &= \mathbb{E}[w(D,X)Y'(D)] \\ &= \mathbb{E}[w(D,X)\mathbb{E}[Y'(D)|D,X]]\end{aligned}$$

Let (d_0, x_0) be a point in the support of the data, i.e., $f(d_0, x_0) > 0$. We aim to identify $\mathbb{E}[Y'(d_0)|D = d_0, X = x_0]$. By Assumption 1c, since $f(d_0|x_0) > 0$ and $f(d|x_0)$ is continuous in d, there exists an open neighborhood $\mathcal{N}(d_0)$ around d_0 such that $f(d|x_0) > 0$ for all $d \in \mathcal{N}(d_0)$. Within $\mathcal{N}(d_0)$, the regression function $\mu(d, x_0)$ is well-defined. We can relate it to the conditional dose-response function $\mathbb{E}[Y(d)|X = x_0]$. Let $m(d, x_0) = \mathbb{E}[Y(d)|X = x_0]$.

For any $d \in \mathcal{N}(d_0)$,

$$\begin{aligned}\mu(d, x_0) &= \mathbb{E}[Y|D = d, X = x_0] \\ &= \mathbb{E}[Y(d)|D = d, X = x_0] \\ &= \mathbb{E}[Y(d)|X = x_0] \\ &= m(d, x_0).\end{aligned}$$

Since $\mu(d, x_0) = m(d, x_0)$, their derivatives with respect to d evaluated at d_0 are equal:$\partial_d\mu(d, x_0)|_{d=d_0} = \partial_d m(d, x_0)|_{d=d_0}$, where ∂_d denotes the partial derivative with respect to d.

By Assumption 1*d*, we invoke the Dominated Convergence Theorem to interchange the derivative and expectation:

$$\partial_d m(d, x_0)|_{d=d_0} = \mathbb{E}[Y'(d_0)|X = x_0].$$

Applying Assumption 1*b* again, we have

$$\mathbb{E}[Y'(d_0)|X = x_0] = \mathbb{E}[Y'(d_0)|D = d_0, X = x_0].$$

Combining the above equalities, we obtain that for any (d_0, x_0) in the support of the data,

$$\mu'(d_0, x_0) = \mathbb{E}[Y'(d_0)|D = d_0, X = x_0].$$

Substituting back into $\dot{\tau}_w$, we have

$$\dot{\tau}_w = \mathbb{E}[w(D, X)\mu'(D, X)].$$

This completes the proof.

Additionally, note that by the Mean Value Theorem, there exists a value $d' \in [d, d + \nu]$ such that

$$\frac{Y(d + \nu) - Y(d)}{\nu} = Y'(d').$$

Then by the Dominated Convergence Theorem, we have

$$\begin{aligned}
\lim_{\nu \to 0} \frac{\mathbb{E}[Y(d + \nu)] - \mathbb{E}[Y(d)]}{\nu} &= \lim_{\nu \to 0} \mathbb{E}[Y'(d')] \\
&= \mathbb{E}[\lim_{\nu \to 0} Y'(d')] \\
&= \mathbb{E}[Y'(d)].
\end{aligned}$$

This demonstrates that our proposed estimand can be interpreted as the limit of the incremental effect as the increment ν approaches zero. □

Proof of Proposition 2. We show two results. First, the estimand defined by $\alpha \in \mathcal{A}$ corresponds to a weighted average derivative with the specified weight $w(d, x)$. Second, this weight is normalized, i.e., $\mathbb{E}[w(D, X)] = 1$.

Relating α and w

We need to verify that for any sufficiently smooth and bounded function $\mu \in \mathcal{H}$, the following identity holds:

$$\mathbb{E}[\alpha(D,X)\mu(D,X)] = \mathbb{E}[w(D,X)\mu'(D,X)]. \tag{A.1}$$

We analyze this identity conditionally on $X = x$. Let $w(d,x)$ be defined as in Equation (3):

$$w(d,x) = -\frac{F(d|x)}{f(d|x)} \cdot \mathbb{E}[\alpha(D,X)|D \leq d, X = x].$$

Let $H(d,x) = w(d,x)f(d|x)$. We can rewrite $H(d,x)$ using the definition of conditional expectation:

$$\begin{aligned} H(d,x) &= -F(d|x) \cdot \mathbb{E}[\alpha(D,X)|D \leq d, X = x] \\ &= -\mathbb{E}[\alpha(D,X)\mathbb{I}(D \leq d)|X = x] \\ &= -\int_{\underline{d}}^{d} \alpha(t,x)f(t|x)\mathrm{d}t. \end{aligned}$$

We evaluate the right-hand side of Equation (A.1) conditional on $X = x$ using integration by parts:

$$\begin{aligned} \mathbb{E}[w(D,X)\mu'(D,X)|X = x] &= \int_{\underline{d}}^{\bar{d}} w(d,x)\mu'(d,x)f(d|x)\mathrm{d}d \\ &= \int_{\underline{d}}^{\bar{d}} \mu'(d,x)H(d,x)\mathrm{d}d \\ &= [\mu(d,x)H(d,x)]_{\underline{d}}^{\bar{d}} - \int_{\underline{d}}^{\bar{d}} \mu(d,x)\frac{\partial H(d,x)}{\partial d}\mathrm{d}d. \end{aligned}$$

We must verify that the boundary terms vanish. At the lower bound $\underline{d}$

$$H(\underline{d},x) = -\int_{\underline{d}}^{\underline{d}} \alpha(t,x)f(t|x)\mathrm{d}t = 0.$$

At the upper bound $\bar{d}$

$$H(\bar{d},x) = -\int_{\underline{d}}^{\bar{d}} \alpha(t,x)f(t|x)\mathrm{d}t = -\mathbb{E}[\alpha(D,X)|X = x].$$

Since $\alpha \in \mathcal{A}$, Equation (2) ensures $\mathbb{E}[\alpha(D,X)|X = x] = 0$. Thus, $H(\bar{d},x) = 0$. The boundary terms vanish (since μ is bounded by Assumption 1d).[6]

Now we analyze the derivative of $H(d,x)$. By the Fundamental Theorem of Calculus:

$$\frac{\partial H(d,x)}{\partial d} = \frac{\partial}{\partial d}\left(-\int_{\underline{d}}^{d} \alpha(t,x) f(t|x) \mathrm{d}t\right) = -\alpha(d,x) f(d|x).$$

Substituting this back into the integration by parts formula:

$$\begin{aligned}
\mathbb{E}[w(D,X)\mu'(D,X)|X=x] &= 0 - \int_{\underline{d}}^{\bar{d}} \mu(d,x)[-\alpha(d,x) f(d|x)] \mathrm{d}d \\
&= \int_{\underline{d}}^{\bar{d}} \mu(d,x)\alpha(d,x) f(d|x) \mathrm{d}d \\
&= \mathbb{E}[\alpha(D,X)\mu(D,X)|X=x].
\end{aligned}$$

This confirms the identity in Equation (A.1).

Normalization of $w(d,x)$

We must show that $\mathbb{E}[w(D,X)] = 1$. We utilize the identity established in Part 1, which holds for any smooth function $\mu(d,x)$. We choose the test function $\mu(d,x) = d$. Then $\mu'(d,x) = 1$.

Substituting this into the identity $\mathbb{E}[w\mu'|X] = \mathbb{E}[\alpha\mu|X]$:

$$\mathbb{E}[w(D,X) \cdot 1|X] = \mathbb{E}[\alpha(D,X) \cdot D|X].$$

Taking the expectation over X using the Law of Iterated Expectations:

$$\begin{aligned}
\mathbb{E}[w(D,X)] &= \mathbb{E}[\mathbb{E}[w(D,X)|X]] \\
&= \mathbb{E}[\mathbb{E}[\alpha(D,X)D|X]] \\
&= \mathbb{E}[\alpha(D,X)D].
\end{aligned}$$

Since $\alpha \in \mathcal{A}$, the normalization constraint (Equation (2)) ensures $\mathbb{E}[\alpha(D,X)D] = 1$. Therefore, $\mathbb{E}[w(D,X)] = 1$. $\square$

[6]If the support is unbounded (e.g., $\mathbb{R}$), we assume standard regularity conditions such that $\lim_{d\to\pm\infty} \mu(d,x)H(d,x) = 0$.

Proof of Lemma 1. From Equation (3), since $F(d|x) \geq 0$ and $f(d|x) \geq 0$ for all d, x, it is straightforward to show that $w(d, x) \geq 0$ if and only if

$$\mathbb{E}[\alpha(D, X)|D \leq d, X = x] \leq 0 \text{ for all } d, x.$$

□

Proof of Theorem 1. The proof proceeds in two parts. First, we derive the optimal balancing weight for the continuous treatment setting using functional optimization. Second, we demonstrate how this optimization framework generalizes to recover the optimal balancing weights for categorical (binary and multiple) treatments when the appropriate optimization criteria for those settings are applied.

We seek the optimal balancing weight $\alpha^* \in \mathcal{A}$ that minimizes the efficiency bound $V_S = \mathbb{E}[\alpha^2(D, X)(Y - \mu(D, X))^2]$. Under the homoskedasticity assumption, $Var(Y|D, X) = \sigma^2$, the objective simplifies to minimizing the L_2 norm of the balancing weight:

$$V_S = \sigma^2 \mathbb{E}[\alpha^2(D, X)].$$

We aim to solve the following constrained optimization problem in the Hilbert space $\mathcal{H} = L_2(P_{D,X})$:

$$\begin{aligned} \min_{\alpha \in \mathcal{H}} \quad & \mathbb{E}[\alpha^2(D, X)] \\ \text{subject to} \quad & (C1) : \mathbb{E}[\alpha(D, X)D] = 1, \\ & (C2) : \mathbb{E}[\alpha(D, X)|X] = 0 \end{aligned}$$

We use the method of Lagrange multipliers to handle these constraints. Let $\lambda \in \mathbb{R}$ be the multiplier for the scalar constraint (C1), and let $\eta(X) \in L_2(P_X)$ be the multiplier function corresponding to the functional constraint (C2). The Lagrangian is:

$$\begin{aligned} L(\alpha, \lambda, \eta) &= \mathbb{E}[\alpha^2] - 2\lambda(\mathbb{E}[\alpha D] - 1) - \mathbb{E}[2\eta(X)\mathbb{E}[\alpha|X]] \\ &= \mathbb{E}[\alpha^2 - 2\lambda\alpha D - 2\eta(X)\alpha] + 2\lambda. \end{aligned}$$

The factor of 2 is introduced for analytical convenience. To find the optimum, we compute the Gâteaux derivative of L with respect to α in an arbitrary direction $h \in \mathcal{H}$ and set it to zero:

$$\begin{aligned}\nabla_\alpha L(\alpha; h) &= \lim_{\epsilon \to 0} \frac{L(\alpha + \epsilon h, \lambda, \eta) - L(\alpha, \lambda, \eta)}{\epsilon} \\ &= \frac{d}{d\epsilon} \mathbb{E}[(\alpha + \epsilon h)^2 - 2\lambda(\alpha + \epsilon h)D - 2\eta(X)(\alpha + \epsilon h)]\Big|_{\epsilon=0} \\ &= \mathbb{E}[2\alpha h - 2\lambda D h - 2\eta(X)h] \\ &= 2\mathbb{E}[h(\alpha - \lambda D - \eta(X))].\end{aligned}$$

For the derivative to be zero for all $h \in \mathcal{H}$, the term in the parenthesis must be zero almost surely. This yields the first-order condition:

$$\alpha(D, X) = \lambda D + \eta(X).$$

We now use the constraints to solve for the multipliers λ and $\eta(X)$. Applying constraint (C2), $\mathbb{E}[\alpha|X] = 0$:

$$\mathbb{E}[\lambda D + \eta(X)|X] = \lambda \mathbb{E}[D|X] + \eta(X) = 0.$$

Thus, the functional multiplier is $\eta(X) = -\lambda \mathbb{E}[D|X]$. Substituting this back into the expression for α:

$$\alpha(D, X) = \lambda D - \lambda \mathbb{E}[D|X] = \lambda(D - \mathbb{E}[D|X]).$$

Finally, applying the normalization constraint (C1), $\mathbb{E}[\alpha D] = 1$:

$$\begin{aligned}1 &= \mathbb{E}[\lambda(D - \mathbb{E}[D|X])D] \\ &= \lambda \mathbb{E}[D^2 - D\mathbb{E}[D|X]] \\ &= \lambda \mathbb{E}[\mathbb{E}[D^2|X] - \mathbb{E}[D|X]^2] \\ &= \lambda \mathbb{E}[Var(D|X)].\end{aligned}$$

Assuming $\mathbb{E}[Var(D|X)] > 0$, we solve for λ:

$$\lambda = \frac{1}{\mathbb{E}[Var(D|X)]}.$$

The optimal balancing weight is therefore:

$$\alpha^*(d, x) = \frac{d - \mathbb{E}[D|X = x]}{\mathbb{E}[Var(D|X)]}.$$

The corresponding optimal estimand is:

$$\tau^* = \mathbb{E}[\alpha^*(D,X)Y] = \frac{\mathbb{E}[(D-\mathbb{E}[D|X])Y]}{\mathbb{E}[Var(D|X)]} = \frac{\mathbb{E}[Cov(D,Y|X)]}{\mathbb{E}[Var(D|X)]}.$$

Generalization to Categorical Treatments

We now demonstrate how this optimization framework incorporates existing results for categorical treatments $D \in \{1,\ldots,J\}$. In this setting, the focus is typically on pairwise contrasts defined relative to a target population specified by a tilting function $w(X)$, normalized such that $\mathbb{E}[w(X)] = 1$.

Let $e_j(x) = P(D = j|X = x)$ be the generalized propensity score and $v_j(x) = Var(Y|D = j, X = x)$ be the conditional outcome variance. The weighted pairwise contrast between treatment j and k is:

$$\tau_{jk}(w) = \mathbb{E}[w(X)(\mu(j,X) - \mu(k,X))].$$

The Riesz Representer (balancing weight) for $\tau_{jk}(w)$ is:

$$\alpha_{jk}(D,X;w) = w(X)\left(\frac{\mathbb{I}(D=j)}{e_j(X)} - \frac{\mathbb{I}(D=k)}{e_k(X)}\right).$$

The corresponding efficiency bound $V_{S,jk}(w)$ is $\mathbb{E}[\alpha_{jk}^2(D,X;w)(Y-\mu(D,X))^2]$. We calculate the expectation:

$$\begin{aligned} V_{S,jk}(w) &= \mathbb{E}\left[\sum_{l=1}^{J} e_l(X)\alpha_{jk}^2(l,X;w)v_l(X)\right] \\ &= \mathbb{E}\left[w^2(X)\left(e_j(X)\frac{v_j(X)}{e_j^2(X)} + e_k(X)\frac{v_k(X)}{e_k^2(X)}\right)\right] \\ &= \mathbb{E}\left[w^2(X)\left(\frac{v_j(X)}{e_j(X)} + \frac{v_k(X)}{e_k(X)}\right)\right]. \end{aligned}$$

Case 2a: Multiple Treatments ($J \geq 3$). When comparing multiple nominal treatments, Li and Li (2019) proposed minimizing the total asymptotic variance of all pairwise contrasts.

We adopt this criterion: Minimize $\sum_{j<k} V_{S,jk}(w)$ subject to $\mathbb{E}[w(X)] = 1$.

$$\sum_{1 \le k < j \le J} V_{S,jk}(w) = \mathbb{E}\left[w^2(X) \sum_{1 \le k < j \le J} \left(\frac{v_j(X)}{e_j(X)} + \frac{v_k(X)}{e_k(X)} \right) \right].$$

In the summation over all unique pairs, each term $v_j(X)/e_j(X)$ appears exactly $J-1$ times. Thus,

$$\sum_{1 \le k < j \le J} V_{S,jk}(w) = (J-1)\mathbb{E}\left[w^2(X) \sum_{j=1}^{J} \frac{v_j(X)}{e_j(X)} \right].$$

Let $K(X) = \sum_{j=1}^{J} v_j(X)/e_j(X)$. We minimize $\mathbb{E}[w^2(X)K(X)]$ subject to $\mathbb{E}[w(X)] = 1$. By the Cauchy-Schwarz inequality or using a Lagrangian approach, the optimal weight is $w^*(x) \propto 1/K(x)$.

$$w^*(x) \propto \left(\sum_{j=1}^{J} \frac{v_j(x)}{e_j(x)} \right)^{-1}.$$

Under homoscedasticity ($v_j(x) = v$), the optimal weight is the harmonic mean of the generalized propensity scores, $w^*(x) \propto (\sum_{j=1}^{J} 1/e_j(x))^{-1}$, recovering the Generalized Overlap Weights of Li and Li (2019).

Case 2b: Binary Treatment ($J = 2$). If $D \in \{1, 2\}$, there is only one contrast, τ_{12}. The optimization criterion naturally reduces to minimizing $V_{S,12}(w)$. Following the derivation in Case 2a, the optimal weight is:

$$w^*(x) \propto \left(\frac{v_1(x)}{e_1(x)} + \frac{v_2(x)}{e_2(x)} \right)^{-1}.$$

Under homoscedasticity, $w^*(x) \propto (1/e_1(x) + 1/e_2(x))^{-1} = e_1(x)e_2(x)$. Since $e_1(x) + e_2(x) = 1$, this recovers the overlap weights proposed by Crump et al. (2006) and Li et al. (2018). $\square$

Proof of Theorem 2. We use the method of pathwise differentiation to derive the influence function. Let P_ϵ be a regular parametric submodel passing through the true distribution P_0 at $\epsilon = 0$, with score function $S(O) = \frac{\partial}{\partial \epsilon} \log p_\epsilon(O)|_{\epsilon=0}$. We seek the EIF $\phi(O) \in L_2(P_0)$ such that the pathwise derivative $PD(\tau_\alpha) = \frac{\partial \tau_\alpha(\epsilon)}{\partial \epsilon}|_{\epsilon=0}$ satisfies $PD(\tau_\alpha) = \mathbb{E}[\phi(O)S(O)]$.

The estimand under P_ϵ is $\tau_\alpha(\epsilon) = \mathbb{E}_\epsilon[\alpha_\epsilon(D, X)\mu_\epsilon(D, X)]$. The pathwise derivative evaluated at $\epsilon = 0$ is denoted with a tilde (e.g., $\tilde{\alpha}$). The total pathwise derivative is derived from the definition of the expectation:

$$PD(\tau_\alpha) = \frac{\partial}{\partial \epsilon} \int \alpha_\epsilon(d,x)\mu_\epsilon(d,x)p_\epsilon(o)\mathrm{d}o\Big|_{\epsilon=0}$$
$$= \underbrace{\mathbb{E}[\tilde{\alpha}\mu]}_{PD_{adj}} + \underbrace{\mathbb{E}[\alpha\tilde{\mu}] + \mathbb{E}[\alpha\mu S]}_{PD_\psi}.$$

The terms PD_ψ correspond to the variation in the estimand assuming the balancing weights α are fixed. We analyze $\mathbb{E}[\alpha\tilde{\mu}]$. The pathwise derivative of the outcome regression $\mu(d,x) = \mathbb{E}[Y|D=d, X=x]$ is:

$$\tilde{\mu}(d,x) = \mathbb{E}[YS_{O|D,X}|D=d, X=x],$$

where $S_{O|D,X} = S(O) - \mathbb{E}[S(O)|D,X]$ is the conditional score given (D,X).

$$\mathbb{E}[\alpha\tilde{\mu}] = \mathbb{E}[\alpha(D,X)\mathbb{E}[YS_{O|D,X}|D,X]]$$
$$= \mathbb{E}[\alpha(D,X)YS_{O|D,X}] \quad \text{(Law of Iterated Expectations, LIE).}$$

Since $\mathbb{E}[S_{O|D,X}|D,X] = 0$, we have $\mathbb{E}[\alpha\mu S_{O|D,X}] = 0$. Thus,

$$\mathbb{E}[\alpha\tilde{\mu}] = \mathbb{E}[\alpha(D,X)(Y - \mu(D,X))S_{O|D,X}].$$

Let $\psi_A(O) = \alpha(Y - \mu)$. Since $\mathbb{E}[\psi_A(O)|D,X] = 0$, we can replace the conditional score $S_{O|D,X}$ with the full score $S(O)$. To see this, note that $\mathbb{E}[\psi_A S] = \mathbb{E}[\psi_A S_{O|D,X}] + \mathbb{E}[\psi_A \mathbb{E}[S|D,X]]$. The second term is $\mathbb{E}[\mathbb{E}[\psi_A|D,X]\mathbb{E}[S|D,X]] = 0$.

$$\mathbb{E}[\alpha\tilde{\mu}] = \mathbb{E}[\psi_A(O)S(O)].$$

We analyze $\mathbb{E}[\alpha\mu S]$. This term accounts for the change in the expectation measure. Since $\mathbb{E}[S] = 0$:

$$\mathbb{E}[\alpha\mu S] = \mathbb{E}[(\alpha\mu - \tau_\alpha)S].$$

Combining these, we recover the influence function assuming fixed weights $\psi(O) = \psi_A(O) + (\alpha\mu - \tau_\alpha)$:

$$PD_\psi = \mathbb{E}[\psi(O)S(O)].$$

The term $PD_{adj} = \mathbb{E}[\tilde{\alpha}\mu]$ captures the sensitivity of the estimand to the estimation of the balancing weights. We must find $\phi_{adj}(O)$ such that $\mathbb{E}[\phi_{adj}(O)S(O)] = PD_{adj}$.

The balancing weight is $\alpha(d,x) = A(d,x;\eta_X(x),\theta)$. By the chain rule, the pathwise derivative of α is:

$$\tilde{\alpha} = \partial_{\eta_X} A \cdot \tilde{\eta}_X + \partial_\theta A \cdot \tilde{\theta}.$$

The adjustment term decomposes as $PD_{adj} = PD_{adj,X} + PD_{adj,\theta}$. We first consider the contribution from $\eta_X(X) = \mathbb{E}[H_X(O)|X]$, which is $PD_{adj,X} = \mathbb{E}[\partial_{\eta_X} A \cdot \mu \cdot \tilde{\eta}_X]$. The pathwise derivative of the conditional expectation $\eta_X(X)$ is:

$$\tilde{\eta}_X(X) = \mathbb{E}[H_X(O) S_{O|X}|X],$$

where $S_{O|X} = S(O) - \mathbb{E}[S(O)|X]$ is the conditional score given X. Rewrite $PD_{adj,X}$ by conditioning on X:

$$PD_{adj,X} = \mathbb{E}_X \left[\mathbb{E}_{D|X}[\partial_{\eta_X} A \cdot \mu | X] \cdot \tilde{\eta}_X(X) \right].$$

Let $R_X(X) = \mathbb{E}[\partial_{\eta_X} A \cdot \mu|X]$ be the Riesz Representer defined in the theorem.

$$\begin{aligned} PD_{adj,X} &= \mathbb{E}_X \left[R_X(X) \cdot \mathbb{E}[H_X(O) S_{O|X}|X] \right] \\ &= \mathbb{E}[R_X(X) H_X(O) S_{O|X}]. \end{aligned}$$

We now identify the influence function $\phi_{adj,X}$. Let $\phi_{adj,X}(O) = R_X(X)(H_X(O) - \eta_X(X))$. We verify this solution. Since $\mathbb{E}[\phi_{adj,X}(O)|X] = 0$, we have $\mathbb{E}[\phi_{adj,X} S] = \mathbb{E}[\phi_{adj,X} S_{O|X}]$.

$$\mathbb{E}[\phi_{adj,X} S_{O|X}] = \mathbb{E}[R_X(X)(H_X(O) - \eta_X(X)) S_{O|X}].$$

Since $\mathbb{E}[S_{O|X}|X] = 0$, we have $\mathbb{E}[R_X(X)\eta_X(X) S_{O|X}] = 0$. Thus, $\mathbb{E}[\phi_{adj,X} S_{O|X}] = \mathbb{E}[R_X(X) H_X(O) S_{O|X}]$, which matches $PD_{adj,X}$.

Next, we consider the contribution from $\theta = \mathbb{E}[G(O;\eta_X)]$, which is $PD_{adj,\theta} = \mathbb{E}[\partial_\theta A \cdot \mu \cdot \tilde{\theta}]$. Let $R_\theta = \mathbb{E}[\partial_\theta A \cdot \mu]$ be the Riesz Representer defined in the theorem. Since R_θ is a constant and $\tilde{\theta}$ is a scalar:

$$PD_{adj,\theta} = R_\theta \cdot \tilde{\theta}.$$

By definition of the influence function $\phi_\theta(O)$ for θ, we have $\tilde{\theta} = \mathbb{E}[\phi_\theta(O) S(O)]$.

$$PD_{adj,\theta} = R_\theta \cdot \mathbb{E}[\phi_\theta(O) S(O)] = \mathbb{E}[R_\theta \phi_\theta(O) S(O)].$$

The corresponding influence function is $\phi_{adj,\theta}(O) = R_\theta \phi_\theta(O)$.

The total pathwise derivative is $PD(\tau_\alpha) = PD_\psi + PD_{adj,X} + PD_{adj,\theta}$. By the Riesz Representation Theorem, the influence function is:

$$\phi(O) = \psi(O) + \phi_{adj,X}(O) + \phi_{adj,\theta}(O).$$

This completes the proof. $\square$

Proof of Proposition 3. We employ standard empirical process notation. P and P_n be linear operators such that for some function $g(O)$, $P[g(O)] = \mathbb{E}[g(O)]$ and $P_n[g(O)] = n^{-1}\sum_{i=1}^n g(O_i)$. We use $\|\cdot\|_p$ to denote the $L_p(P)$ norm. Let $\rho_0(x)$ and $e_0(x)$ denote the true nuisance functions, and $\hat{\rho}(x), \hat{e}(x)$ their estimates.

We define the moment functions for the numerator and denominator:

$$\begin{aligned} m_\gamma(O;\rho,e) &= (Y-\rho(X))(D-e(X)), \\ m_\pi(O;e) &= (D-e(X))^2. \end{aligned}$$

The population parameters are $\gamma = P[m_\gamma(O;\rho_0,e_0)]$ and $\pi = P[m_\pi(O;e_0)]$. The estimators are $\hat{\gamma} = P_n[m_\gamma(O;\hat{\rho},\hat{e})]$ and $\hat{\pi} = P_n[m_\pi(O;\hat{e})]$. The influence functions are:

$$\begin{aligned} \phi_\gamma(O) &= m_\gamma(O;\rho_0,e_0) - \gamma, \\ \phi_\pi(O) &= m_\pi(O;e_0) - \pi. \end{aligned}$$

We aim to show that $\hat{\gamma}$ and $\hat{\pi}$ are Regular Asymptotically Linear (RAL):

$$\begin{aligned} \hat{\gamma} - \gamma &= P_n[\phi_\gamma(O)] + o_P(n^{-1/2}), \\ \hat{\pi} - \pi &= P_n[\phi_\pi(O)] + o_P(n^{-1/2}). \end{aligned}$$

If this holds, by the Delta method for the ratio functional, and since $\pi > 0$ by Assumption (b), $\hat{\tau}^* = \hat{\gamma}/\hat{\pi}$ is RAL for τ^*:

$$\begin{aligned} \hat{\tau}^* - \tau^* &= P_n\left[\frac{\phi_\gamma(O) - (\gamma/\pi)\phi_\pi(O)}{\pi}\right] + o_P(n^{-1/2}) \\ &= P_n[\phi^*(O)] + o_P(n^{-1/2}), \end{aligned}$$

where $\phi^*(O)$ is the influence function defined in Equation (9). Asymptotic normality follows from the Central Limit Theorem.

We decompose the estimation error:

$$\begin{aligned}\hat{\gamma} - \gamma &= P_n[m_\gamma(\hat{\rho}, \hat{e})] - P[m_\gamma(\rho_0, e_0)] \\ &= (P_n - P)[m_\gamma(\rho_0, e_0)] + E_n + R_n \\ &= P_n[\phi_\gamma(O)] + E_n + R_n\end{aligned}$$

where the Empirical Process term E_n and the Remainder term R_n are:

$$\begin{aligned}E_n &= (P_n - P)[m_\gamma(\hat{\rho}, \hat{e}) - m_\gamma(\rho_0, e_0)], \\ R_n &= P[m_\gamma(\hat{\rho}, \hat{e}) - m_\gamma(\rho_0, e_0)].\end{aligned}$$

We must show $E_n = o_P(n^{-1/2})$ and $R_n = o_P(n^{-1/2})$.

The Remainder Term R_n (Neyman Orthogonality)

We analyze R_n using the Law of Iterated Expectations (LIE), conditioning on X.

$$R_n = P[(Y - \hat{\rho})(D - \hat{e}) - (Y - \rho_0)(D - e_0)].$$

We evaluate the conditional expectation of the difference. Since $\mathbb{E}[Y|X] = \rho_0$ and $\mathbb{E}[D|X] = e_0$:

$$\begin{aligned}\mathbb{E}[(Y - \hat{\rho})(D - \hat{e})|X] &= \mathbb{E}[YD|X] - \rho_0\hat{e} - \hat{\rho}e_0 + \hat{\rho}\hat{e}. \\ \mathbb{E}[(Y - \rho_0)(D - e_0)|X] &= \mathbb{E}[YD|X] - \rho_0 e_0.\end{aligned}$$

Subtracting the two yields the conditional expectation of the difference:

$$\mathbb{E}[m_\gamma(\hat{\rho}, \hat{e}) - m_\gamma(\rho_0, e_0)|X] = -\rho_0\hat{e} - \hat{\rho}e_0 + \hat{\rho}\hat{e} + \rho_0 e_0 = (\rho_0 - \hat{\rho})(e_0 - \hat{e}).$$

Thus, $R_n = P[(\rho_0 - \hat{\rho})(e_0 - \hat{e})]$. This cancellation of first-order terms demonstrates the Neyman orthogonality of the score function.

We bound R_n using the Cauchy-Schwarz inequality:

$$|R_n| \leq \sqrt{P[(\rho_0 - \hat{\rho})^2]}\sqrt{P[(e_0 - \hat{e})^2]} = \|\rho_0 - \hat{\rho}\|_2 \cdot \|e_0 - \hat{e}\|_2.$$

By the L_2 rates in Assumption (a), $|R_n| = o_P(n^{-b/4}) \cdot o_P(n^{-a/4}) = o_P(n^{-(a+b)/4})$. Since $a + b \geq 2$, $R_n = o_P(n^{-1/2})$.

The Empirical Process Term E_n

We need to show $E_n = o_P(n^{-1/2})$. This requires (i) $L_2(P)$ consistency of the score difference, and (ii) complexity control (Assumption d or e).

We verify $P[(m_\gamma(\hat{\rho}, \hat{e}) - m_\gamma(\rho_0, e_0))^2] = o_P(1)$. Let $\Delta m_\gamma = m_\gamma(\hat{\rho}, \hat{e}) - m_\gamma(\rho_0, e_0)$. Let $\Delta\rho = \hat{\rho} - \rho_0$ and $\Delta e = \hat{e} - e_0$.

$$\Delta m_\gamma = -(Y - \rho_0)\Delta e - (D - e_0)\Delta\rho + \Delta\rho\Delta e.$$

By the Minkowski inequality, it suffices to show that the L_2 norm squared of each term converges to zero.

Term 1: $P[(Y - \rho_0)^2 \Delta e^2] = P[Var(Y|X)\Delta e^2]$. By Assumption (c) (Bounded moments), $Var(Y|X) < C_2$. This is bounded by $C_2\|\Delta e\|_2^2$. By Assumption (a) ($a \geq 1$), this is $o_P(1)$.

Term 2: Similarly bounded by $C_2\|\Delta\rho\|_2^2$. By Assumption (a) ($b \geq 0$), this is $o_P(1)$.

Term 3: $P[\Delta\rho^2\Delta e^2]$. By the Cauchy-Schwarz inequality:

$$P[\Delta\rho^2\Delta e^2] \leq \sqrt{P[\Delta\rho^4]}\sqrt{P[\Delta e^4]} = \|\Delta\rho\|_4^2\|\Delta e\|_4^2.$$

By the L_4 consistency requirement in the revised Assumption (a), this is $o_P(1) \cdot o_P(1) = o_P(1)$.

Thus, $L_2(P)$ consistency holds.

If Assumption (d) (Donsker condition) holds, then by standard empirical process results (e.g., Lemma 19.24 of Van der Vaart (1998)), $L_2(P)$ consistency implies $E_n = o_P(n^{-1/2})$.

If Assumption (e) (Cross-fitting) holds, $\hat{\rho}, \hat{e}$ are estimated on an independent sample. Conditional on the estimation sample, E_n is a centered empirical process evaluated at a function whose $L_2(P)$ norm converges to zero in probability. This implies $E_n = o_P(n^{-1/2})$ (see, e.g., Lemma 2 of Kennedy et al. (2020)).

The Denominator $\hat{\pi}$

The analysis for $\hat{\pi}$ is analogous. The remainder term is:

$$R_n^\pi = P[m_\pi(\hat{e}) - m_\pi(e_0)] = P[(D - \hat{e})^2 - (D - e_0)^2].$$

Applying LIE, we find:

$$R_n^\pi = P[(e_0 - \hat{e})^2] = \|e_0 - \hat{e}\|_2^2.$$

By Assumption (a), since $a \geq 1$, $R_n^\pi = o_P(n^{-a/2}) = o_P(n^{-1/2})$.

The empirical process term E_n^π requires $L_2(P)$-consistency of $\Delta m_\pi = m_\pi(\hat{e}) - m_\pi(e_0) = -2(D - e_0)\Delta e + \Delta e^2$. The analysis of the first term is similar to E_n. The second term requires $\|\Delta e^2\|_2 = \|\Delta e\|_4^2 = o_P(1)$, which holds by L_4-consistency of $\hat{e}$ (Assumption a). Given this, $E_n^\pi = o_P(n^{-1/2})$ under (d) or (e).

We have established that $\hat{\gamma}$ and $\hat{\pi}$ are RAL. By the Delta method, $\hat{\tau}^*$ is RAL for τ^* with the influence function ϕ^*. This completes the proof. $\square$

B Appendix: Extra Results

Generalization of Theorem 5.1 of Crump et al. (2006)

The generalized efficiency bound in Theorem 2 decomposes as $V_\phi = V_\psi + V_{adj} + 2C_{\psi,adj}$. In certain settings, specific structural and orthogonality conditions ensure that the covariance term $C_{\psi,adj}$ is zero. This leads to a simpler decomposition where estimating the weights necessarily increases the variance ($V_\phi \geq V_\psi$). This specialized case is common in the categorical treatment literature.

We present a corollary that highlights the conditions required for this simplification, generalizing the findings of Theorem 5.1 in Crump et al. (2006).

Corollary 1 (Efficiency Decomposition under Orthogonality)**.** *Assume that the weight function $w(d,x) = \lambda(\eta(d,x))$, where $\lambda(\cdot)$ is known and $\eta(d,x) = \mathbb{E}[H(O)|D=d, X=x]$ for some measurable function $H(O) \in L_2(P)$. Furthermore, assume the following orthogonality condition holds:*

$$\mathbb{E}[(Y - \mu(D,X))(H(O) - \eta(D,X))|D,X] = 0.$$

This condition implies that the noise in the outcome is uncorrelated with the noise in the nuisance parameter estimation, conditional on (D,X).

Under these conditions, the covariance between the fixed-weight influence function ψ (Equation (4)) and the adjustment term ϕ_{adj} is zero ($C_{\psi,adj} = 0$). The adjustment term is:

$$\phi_{adj}(O) = \frac{1}{K}\lambda'(\eta(Z))(\mu'(Z) - \tau_w)(H(O) - \eta(D,X)). \tag{B.1}$$

The efficiency bound decomposes additively: $V_\phi = V_\psi + V_{adj}$, where

$$V_{adj} = \frac{1}{K^2}\mathbb{E}\left[(\mu'(D,X) - \tau_w)^2(\lambda'(\eta(D,X)))^2 Var(H(O)|D,X)\right].$$

Corollary 1 explains why the simple additive decomposition holds in the binary setting. In the binary case, weights typically depend on the propensity score $e(X)$. The required structure and the orthogonality condition are satisfied, leading to the result $V_\phi \geq V_\psi$. The generalized Theorem 2 is necessary because these restrictive conditions may not hold generally in the continuous setting, particularly when balancing weights depend on nuisance functions defined only over X (like $e(X)$ in τ^*).

Proof of Corollary 1. The corollary follows from the generalized framework of Theorem 2 by demonstrating that the covariance term $C_{\psi,adj} = \mathbb{E}[\psi(O)\phi_{adj}(O)]$ is zero under the stated assumptions.

Let $Z = (D, X)$. The influence function assuming fixed weights (Equation (4)) is:

$$\psi(O) = \alpha_w(Z)(Y - \mu(Z)) + (w(Z)\mu'(Z) - \tau_w).$$

Under the assumption that the weights depend smoothly on $\eta(Z) = \mathbb{E}[H(O)|Z]$, the adjustment term derived via pathwise differentiation (accounting for the normalization $K = \mathbb{E}[w]$) is:

$$\phi_{adj}(O) = \frac{1}{K}\lambda'(\eta(Z))(\mu'(Z) - \tau_w)(H(O) - \eta(Z)).$$

We analyze the covariance $C_{\psi,adj}$ using the Law of Iterated Expectations, conditioning on Z. Let $\epsilon = Y - \mu(Z)$ (outcome noise) and $\nu = H(O) - \eta(Z)$ (nuisance noise). We can write $\psi(O) = \alpha_w(Z)\epsilon + (w\mu' - \tau_w)$ and $\phi_{adj}(O) = C(Z)\nu$, where $C(Z)$ represents the terms in ϕ_{adj} that are functions of Z.

The conditional expectation is:

$$\begin{aligned}\mathbb{E}[\psi(O)\phi_{adj}(O)|Z] &= \mathbb{E}\left[\left(\alpha_w(Z)\epsilon + (w\mu' - \tau_w)\right)\left(C(Z)\nu\right)|Z\right] \\ &= C(Z)\left(\alpha_w(Z)\mathbb{E}[\epsilon\nu|Z] + (w\mu' - \tau_w)\mathbb{E}[\nu|Z]\right).\end{aligned}$$

By the definition of $\eta(Z) = \mathbb{E}[H(O)|Z]$, the nuisance noise has mean zero: $\mathbb{E}[\nu|Z] = 0$. By the orthogonality assumption stated in the corollary, the noise terms are uncorrelated: $\mathbb{E}[\epsilon\nu|Z] = \mathbb{E}[(Y - \mu(Z))(H(O) - \eta(Z))|Z] = 0$.

Therefore, the conditional expectation $\mathbb{E}[\psi(O)\phi_{adj}(O)|Z] = 0$. Consequently, the unconditional expectation (the covariance) $C_{\psi,adj} = 0$. The efficiency bound $V_\phi = \mathbb{E}[(\psi + \phi_{adj})^2]$ simplifies to the additive decomposition:

$$V_\phi = V_\psi + V_{adj}.$$

The expression for V_{adj} is obtained by calculating $\mathbb{E}[\phi_{adj}^2(O)]$, utilizing $\mathbb{E}[\nu^2|Z] = Var(H(O)|Z)$.

□

Influence Function Derivation by Theorem 2

We derive the influence function for the optimal estimand τ^* using the generalized framework of Theorem 2. The optimal balancing weight is $\alpha^*(D, X) = (D - e(X))/K$. This depends smoothly on the nuisance parameters:

$$\eta_X = e(X) = \mathbb{E}[D|X] \quad \text{with } H_X(O) = D$$
$$\theta = K = \mathbb{E}[(D - e(X))^2]$$

We also use the notation $\rho(X) = \mathbb{E}[\mu(D, X)|X] = \mathbb{E}[Y|X]$.

We apply Theorem 2 to calculate the adjustment term $\phi_{adj} = \phi_{adj,e} + \phi_{adj,K}$.

The derivative of α^* with respect to e is $\partial_e\alpha^* = -1/K$. The Riesz Representer is $R_X(X) = \mathbb{E}[\partial_e\alpha^* \cdot \mu|X] = \mathbb{E}[-1/K \cdot \mu|X] = -\rho(X)/K$. The adjustment term is:

$$\phi_{adj,e}(O) = R_X(X)(D - e(X)) = -\frac{\rho(X)}{K}(D - e(X)).$$

The derivative of α^* with respect to K is $\partial_K\alpha^* = -(D-e(X))/K^2$. The Riesz Representer is $R_\theta = \mathbb{E}[\partial_K\alpha^* \cdot \mu] = \mathbb{E}[-(D - e(X))\mu]/K^2$. Since $\mathbb{E}[(D - e(X))\mu] = \mathbb{E}[Cov(D, Y|X)] = \tau^*K$, we have $R_\theta = -(\tau^*K)/K^2 = -\tau^*/K$.

We also need the influence function of K, $\phi_K(O)$. We analyze the sensitivity of $K = \mathbb{E}[(D - e(X))^2]$ to $e(X)$. The derivative of the argument with respect to e is $-2(D - e(X))$. The conditional expectation of this derivative is $\mathbb{E}[-2(D - e(X))|X] = 0$. This Neyman orthogonality means K is locally insensitive to $e(X)$, simplifying its influence function to:

$$\phi_K(O) = (D - e(X))^2 - K.$$

The adjustment term for K is:

$$\phi_{adj,K}(O) = R_\theta \cdot \phi_K(O) = -\frac{\tau^*}{K}((D - e(X))^2 - K).$$

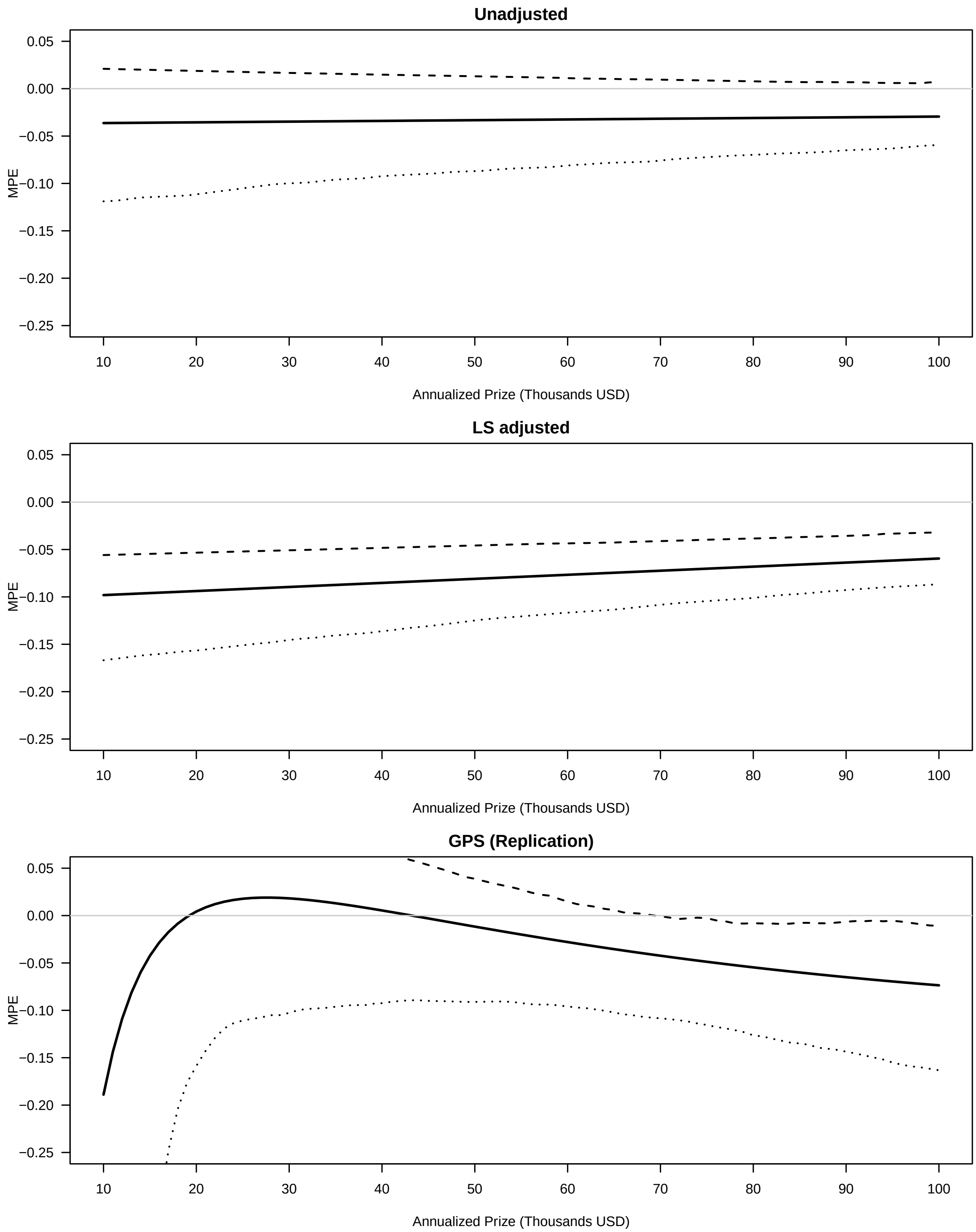
The influence function is $\phi^* = \psi_{\alpha^*} + \phi_{adj,e} + \phi_{adj,K}$. The fixed-weight influence function is $\psi_{\alpha^*} = \alpha^*(Y - \mu) + (\alpha^*\mu - \tau^*) = \alpha^* Y - \tau^*$.

$$
\begin{aligned}
\phi^*(O) &= \psi_{\alpha^*} + \phi_{adj,e}(O) + \phi_{adj,K}(O) \\
&= \left[\frac{D - e(X)}{K} Y - \tau^*\right] + \left[-\frac{\rho(X)}{K}(D - e(X))\right] + \left[-\frac{\tau^*}{K}(D - e(X))^2 + \tau^*\right] \\
&= \frac{(D - e(X))(Y - \rho(X)) - \tau^*(D - e(X))^2}{K}.
\end{aligned}
$$

C Appendix: Replication of HI04

Appendix Table C1: Summary Statistics for Lottery Winners Sample

Variable	Mean	SD
Outcome and Treatment		
Earnings (Y)	10.318	13.163
Annualized Prize (D)	55.196	61.803
Demographics		
Age at Winning	46.945	13.797
Years High School	3.603	1.071
Years College	1.367	1.601
Male	0.578	0.495
Tickets Bought	4.570	3.282
Working at Winning	0.802	0.400
Year Won	1986.059	1.294
Pre-Treatment Earnings		
Earnings year (-6)	11.965	11.790
Earnings year (-5)	12.115	11.992
Earnings year (-4)	12.037	12.081
Earnings year (-3)	12.820	12.654
Earnings year (-2)	13.479	12.965
Earnings year (-1)	14.468	13.624
Observations	237	



Appendix Figure C1: Replication of Figure 7.1 in HI04