Does Residuals-on-Residuals Regression Produce Representative Estimates of Causal Effects?

Apoorva Lal*
Amazon Web Services
San Francisco, CA, USA
lal.apoorva@gmail.com

Winston Chou†
Netflix
Los Gatos, CA, USA
wchou@netflix.com

ABSTRACT

Double Machine Learning is commonly used to estimate causal effects in large observational datasets. The "residuals-on-residuals" regression estimator (RORR) is especially popular for its simplicity and computational tractability. However, when treatment effects are heterogeneous, the proper interpretation of RORR may not be well understood. We show that, for many-valued treatments with continuous dose-response functions, RORR converges to a conditional variance-weighted average of derivatives evaluated at points not in the observed dataset, which generally differs from the Average Causal Derivative (ACD). Hence, even if all units share the same dose-response function, RORR does not in general converge to an average treatment effect in the population represented by the sample. We propose an alternative estimator suitable for large datasets. We demonstrate the pitfalls of RORR and the favorable properties of the proposed estimator in both an illustrative numerical example and an application to real-world data from Netflix.

ACM Reference Format:
Apoorva Lal and Winston Chou. 2025. Does Residuals-on-Residuals Regression Produce Representative Estimates of Causal Effects?. In . ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Double Machine Learning (DML) [5, 8, 18] is fast becoming the standard for estimating causal effects in large, high-dimensional datasets under conditional ignorability assumptions, which stipulate that the treatment is as good as randomly assigned given observed covariates [13]. To strengthen such assumptions, researchers in a wide variety of fields use the DML method to condition on many covariates without strong commitments to functional form [9, 12, 15]. While DML encompasses a large family of methods, this paper focuses on the Partially Linear Model (PLM), which relates an outcome Y_i to a (continuous- or discrete-valued) treatment T_i conditional on pretreatment covariates X_i as follows:

$$Y_i = \theta T_i + g(X_i) + e_i \quad \text{and} \quad T_i = h(X_i) + u_i.$$

The PLM imposes very weak assumptions on how the treatment and outcome relate to the covariates. It also motivates an intuitive two-step DML estimator of θ, the residuals-on-residuals regression (RORR). RORR, a natural extension of the Frisch-Waugh-Lovell theorem, involves first "partialing out" the effect of X_i using flexible machine learning methods, then forming the residuals:

$$\widetilde{Y}_i = Y_i - \widehat{g}(X_i) \quad \text{and} \quad \widetilde{T}_i = T_i - \widehat{h}(X_i),$$

and lastly regressing $\widetilde{Y}_i$ on $\widetilde{T}_i$ to obtain the estimate $\hat{\theta}$ [19].

When the treatment effect θ is the same for all units in the population, it is also the Average Treatment Effect (ATE) for binary treatments, the Average Causal Derivative (ACD) for continuous treatments, and the Average Incremental Effect (AIE) for integer-valued treatments. Alternatively, we study the probability limit (plim) and interpretation of $\hat{\theta}$ when treatment effects are heterogeneous. Under such heterogeneity, $\hat{\theta}$ converges to a conditional variance-weighted average of causal effects, which puts greater weight on units whose treatment values are less predictable. For example, when X is discrete, the RORR estimand places the most weight on the treatment effects in the strata where the treatment has the most variance [1, 3, 20]. When treatments are many-valued (for example, continuous), RORR may be subject to other, more subtle biases. In this paper, we provide a general analysis of RORR with binary and many-valued treatments. We demonstrate the empirical relevance of these biases both in a stylized numerical example and with real-world data from Netflix. Lastly, we propose an alternative estimator that utilizes Augmented Inverse Propensity Weighting (AIPW) along with binning the treatment [7] and demonstrate its favorable properties theoretically and in our empirical application.

2 RESIDUALS-ON-RESIDUALS REGRESSION WITH TREATMENT EFFECT HETEROGENEITY

RORR has many favorable attributes that make it attractive in empirical applications. As mentioned above, it enables highly flexible estimation of the nuisance functions. Furthermore, RORR has the doubly-robust property, converging to the true θ if either the treatment model or the outcome model is correct. When both models are correct, it is also the efficient estimator [10]. Recent applications of RORR include studies in economics [4], ecology [11], and public health [21].

In commercial settings – characterized by large datasets, short timelines, and stakeholders with diverse technical backgrounds – RORR is also used for practical reasons. For example, the final regression step is computationally efficient, as only a small number of statistics is needed to compute the final estimate of θ. Moreover,

*Work done while at Netflix.
†Corresponding author.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

the recipe of (1) removing variation explainable by pretreatment covariates and (2) estimating the effect of the remaining exogenous variation in T_i on Y_i is intuitive and easy to explain to non-experts.

However, this simplicity can come at a cost: The interpretation of θ as an "average" treatment effect (whether the ATE, ACD, or AIE) depends on the assumption of a homogeneous treatment effect embedded in the PLM, which may not hold in applications. In this section, we discuss the interpretation of $\hat{\theta}$ under two common violations of this assumption: binary treatments with heterogeneous effects across individuals and many-valued treatments with continuous dose-response functions.

2.1 Binary Treatments with Heterogeneous Treatment Effects

We begin by studying the plim of the RORR estimator for binary treatments with heterogeneous treatment effects. Letting $T_i \in \{0, 1\}$ denote the binary treatment indicator, we consider the model:

$$\begin{aligned} Y_i &= \theta_i T_i + g(X_i) + e_i \\ T_i &= h(X_i) + u_i, \end{aligned} \tag{1}$$

where θ_i is a individual treatment effect. We further assume that the errors are conditionally exogenous and uncorrelated: $E[e_i|X_i] = 0$, $E[u_i|X_i] = 0$, and $E[e_i u_i|X_i] = 0$. This is a linear instantiation of conditional ignorability. Note that this assumption implies that $g(X_i) = E[Y_i - \theta_i T_i|X_i]$ and $h(X_i) = E[T_i|X_i]$. We also assume that θ_i is conditionally independent of T_i given X_i, $\theta_i \perp\!\!\!\perp T_i|X_i$.

We consider the plim of the OLS regression of $Y_i - g(X_i)$ on $T_i - h(X_i)$ with observations $O_i = (Y_i, T_i, X_i)$, $i = 1, \ldots N$. We assume that the observations $O_1, \ldots, O_N \sim O$ are independent and identically distributed. While g and h must be estimated in practice, for ease of exposition, we assume consistent estimators for these and focus on the (true) limiting g and h.[1]

First observe that:

$$\begin{aligned} \hat{\theta} &\xrightarrow{p} \frac{E[(T_i - h(X_i))(Y_i - g(X_i))]}{E[(T_i - h(X_i))^2]} \\ &= \frac{E[\theta_i (T_i^2 - T_i h(X_i))]}{E[(T_i - h(X_i))^2]}. \end{aligned} \tag{2}$$

Using the fact that T_i is binary and applying the law of iterated expectations, we can rewrite this as:

$$\frac{E[\theta_i (T_i^2 - T_i h(X_i))]}{E[(T_i - h(X_i))^2]} = \frac{E[\theta_i E[T_i(1 - h(X_i))|X_i]]}{E[(T_i - h(X_i))^2]} \tag{3}$$

$$= \frac{E[\theta_i (T_i - h(X_i))^2]}{E[(T_i - h(X_i))^2]}, \tag{4}$$

where Equation (3) follows from θ_i being conditionally independent of T_i. This demonstrates the known result that, with a binary treatment, linear regression converges to a conditional variance-weighted average of individual treatment effects [1, 3].

For an intuitive restatement, denote the conditional variance weights by $\omega_i := \frac{(T_i - h(X_i))^2}{E[(T_i - h(X_i))^2]}$ and note that $E[\omega_i] = 1$ by construction. Then the bias of the RORR plim (which we will denote by $\tilde{\theta}$) with respect to the ATE can be written as:

$$\begin{aligned} \tilde{\theta} - E[\theta_i] &= E[\omega_i \theta_i] - E[\theta_i] && (5) \\ &= \mathrm{Cov}(\omega_i, \theta_i). && (6) \end{aligned}$$

In other words, the bias of RORR for binary treatments when treatment effects are heterogeneous is equal to the covariance of the individual treatment effects and the conditional variance of the treatment residual. This covariance will not equal zero except in special cases (e.g., the treatment is assigned uniformly at random) and therefore $\tilde{\theta} \neq E[\theta_i]$ in general.[2]

2.2 Many-Valued Treatments

We now turn our attention to many-valued (e.g., continuous or integer-valued) treatments with continuous dose-response functions. Although previous research has studied the effect of treatment effect heterogeneity on the interpretation of linear treatment effect estimators, it has mainly done so in the context of binary treatments and/or linear treatment effects [3]. However, in many applications, treatments are continuous and/or have nonlinear effects on the outcome (for example, they may have diminishing returns). Such nonlinearity is an important form of treatment effect heterogeneity that has received less attention in previous work, with notable exceptions [e.g., 2, 22].

Specifically, we study the model:

$$\begin{aligned} Y_i &= f(T_i) + g(X_i) + e_i \\ T_i &= h(X_i) + u_i, \end{aligned} \tag{7}$$

where f is a continuously differentiable function of a many-valued treatment T_i. For reasons that will become apparent, we assume that f is well-defined on the convex hull of T_i, even if T_i itself is not continuous. As before, we assume conditional exogeneity, consistent estimators for g and h, and iid observations.

Under these assumptions, the RORR estimate converges in probability to:

$$\hat{\theta} \xrightarrow{p} \frac{E[(T_i - h(X_i)) f(T_i)]}{E[(T_i - h(X_i))^2]}. \tag{8}$$

Since $h(X_i)$ is a constant given X_i and applying the law of iterated expectations, we can rewrite the above as:

$$\begin{aligned} &\frac{E[(T_i - h(X_i)) f(T_i)]}{E[(T_i - h(X_i))^2]} \\ &= \frac{E[E[(T_i - h(X_i))(f(T_i) - f(h(X_i)))|X_i]]}{E[(T_i - h(X_i))^2]}. \end{aligned} \tag{9}$$

By the mean value theorem, there exists a T_i^* between T_i and $h(X_i)$ for every X_i such that:

$$\begin{aligned} &\frac{E[(T_i - h(X_i))(f(T_i) - f(h(X_i)))|X_i]}{E[(T_i - h(X_i))^2]} \\ &= \frac{E[(T_i - h(X_i))^2 f'(T_i^*)]}{E[(T_i - h(X_i))^2]} \\ &= \frac{E[\omega_i f'(T_i^*)]}{E[\omega_i]}, \end{aligned} \tag{10}$$

[1] Researchers typically use extremely flexible function classes for g and h (e.g., gradient boosted trees or deep neural networks) that are able to approximate the true nuisance functions arbitrarily closely.

[2] A corollary is that ranking treatments based on their PLM coefficient is not the same as ranking them based on their ATEs [17].

showing that, as in the binary case, $\hat{\theta}$ also converges to a conditional variance-weighted average of causal effects.[3] However, unlike in the binary treatment case, the quantity being averaged cannot be interpreted as the causal effect of increasing the treatment in the population represented by the sample. This is because the point T_i^* is not actually the treatment dose received by i, but rather a convex combination of the received treatment T_i and its conditional mean $h(X_i)$. As such, T_i^* may not be an observed treatment level. If T_i is not continuous, it may not even be a realizable treatment value.

Proposition 1 derives the conditions under which $\tilde{\theta}$ converges to the ACD.

Proposition 1. *Let $O_i = (Y_i, T_i, X_i)$, $i = 1, \dots N$ be iid draws from some distribution obeying the model (7). We assume f and h are consistently estimated and that $E[e_i|X_i] = 0$, $E[u_i|X_i] = 0$, and $E[e_i u_i|X_i] = 0$. We also assume that f is everywhere differentiable and that f' is Lipschitz with constant L. Lastly, we assume that $E[(T_i - h(X_i))^2] > 0$. Then the residuals-on-residuals regression (RORR) plim $\tilde{\theta}$ equals the Average Causal Derivative (ACD) $E[f'(T_i)]$ if f is affine ($L = 0$).*

Proof. With iid observations, conditional ignorability, and f and h consistently estimated, $\tilde{\theta}$ is as given in (10). We can decompose the bias of $\tilde{\theta}$ relative to the ACD into two pieces by adding and subtracting:

$$\begin{aligned} &\frac{E[(T_i - h(X_i))^2 f'(T_i^*)]}{E[(T_i - h(X_i))^2]} - E[f'(T_i)] \qquad (11)\\ = &\underbrace{\frac{E[(T_i - h(X_i))^2 f'(T_i^*)]}{E[(T_i - h(X_i))^2]} - \frac{E[(T_i - h(X_i))^2 f'(T_i)]}{E[(T_i - h(X_i))^2]}}_{:=A} \\ &+ \underbrace{\frac{E[(T_i - h(X_i))^2 f'(T_i)]}{E[(T_i - h(X_i))^2]} - E[f'(T_i)]}_{:=B}. \end{aligned}$$

The first piece (A) is the difference between the RORR plim and the variance-weighted average causal derivative evaluated over the treatment distribution actually observed in the sample. Note that we can bound the absolute value of this term as a function of L and the distribution of T_i. First, rewrite A as:

$$\begin{aligned} &\frac{E[(T_i - h(X_i))^2 f'(T_i^*)]}{E[(T_i - h(X_i))^2]} - \frac{E[(T_i - h(X_i))^2 f'(T_i)]}{E[(T_i - h(X_i))^2]} \qquad (12)\\ = &\frac{E[(T_i - h(X_i))^2 (f'(T_i^*) - f'(T_i))]}{E[(T_i - h(X_i))^2]}. \end{aligned}$$

Under the assumption that f' is Lipschitz, there exists a constant $L \geq 0$ such that, for all u, v:

$$|f'(u) - f'(v)| \leq L|u - v|. \qquad (13)$$

Since T_i^* lies between T_i and $h(X_i)$, we have that:

$$|f'(T_i^*) - f'(T_i)| \leq L|T_i^* - T_i| \leq L|T_i - h(X_i)|. \qquad (14)$$

Multiplying both sides by $(T_i - h(X_i))^2$ and taking expectations yields:

$$\begin{aligned} &E[(T_i - h(X_i))^2 |f'(T_i^*) - f'(T_i)|] \qquad (15)\\ &\leq LE[(T_i - h(X_i))^2 |T_i - h(X_i)|] = LE[|T_i - h(X_i)|^3]. \end{aligned}$$

Dividing both sides by $E[(T_i - h(X_i))^2]$ yields the following bound:

$$|A| \leq L \underbrace{\frac{E[|T_i - h(X_i)|^3]}{E[(T_i - h(X_i))^2]}}_{:=\kappa}. \qquad (16)$$

Note that $|T_i - h(X_i)|^3$ is always weakly positive. Therefore, κ only equals zero if $T_i = h(X_i)$ almost surely. However, $E[(T_i - h(X_i))^2] > 0$ by assumption, so T_i cannot equal $h(X_i)$ almost surely. Therefore, the right-hand side of (16) equals zero if and only if $L = 0$.

The second term B is the familiar bias between the variance-weighted average derivative and the ACD. This has a similar interpretation as in the binary treatment case. That is, letting ω_i once again denote the conditional variance weight:

$$E[\omega_i f'(T_i)] - E[f'(T_i)] = \mathrm{Cov}(\omega_i, f'(T_i)), \qquad (17)$$

which is the continuous analog of Equation 6. If f is affine, $f'(T_i)$ is a constant, so $\mathrm{Cov}(\omega_i, f'(T_i)) = 0$, Therefore, the absolute bias of $\tilde{\theta}$ is equal to $|A + \mathrm{Cov}(\omega_i, f'(T_i))| \leq |A| + |\mathrm{Cov}(\omega_i, f'(T_i))| \leq L\kappa + |\mathrm{Cov}(\omega_i, f'(T_i))| = 0$ when f is affine, which obtains the result. □

Corollary 1. *Proposition 1 establishes that f being affine is sufficient for $\tilde{\theta} = E[f'(T_i)]$. If we further assume that $L > 0 \implies A \neq 0$ and $A \neq -B$, where A and B are defined as in (11), then $\tilde{\theta} = E[f'(T_i)]$ if and only if f is affine.*

In other words, $\tilde{\theta}$ will be closer to the conditional variance-weighted average of f' when f is close to affine. In turn, the conditional variance-weighted average of f' is equal to $E[f'(T_i)]$ when $\mathrm{Cov}(\omega_i, f'(T_i)) = 0$, which holds trivially if f is affine. Therefore, both biases vanish when f is affine (and thus the PLM is correctly specified). However, if f is not affine, then $\tilde{\theta} \neq E[f'(T_i)]$ except in contrived cases (e.g., both biases exactly offset).

3 NUMERICAL EXAMPLE

To help build intuition, this section presents a stylized numerical example.[4] Although we make simplistic assumptions to facilitate closed-form analysis, our choices are also intended to reflect qualitative aspects of real-world data. In particular, we assume that:

(1) While $E[Y_i|T_i, X_i]$ is increasing in T_i, it also exhibits diminishing returns. That is, letting $f(T_i)$ be defined as in Equation (7), $f'(T) > 0$ and $f''(T) < 0$.
(2) T_i is an overdispersed count variable, such that even a correct model for $E[T_i|X_i]$ has heteroskedastic errors.

Let X_i be a Categorical variable that takes on values $j = 1, \dots, J$ with probabilities $\pi_1, \dots, \pi_J$ and T_i be a Poisson distributed conditional on X_i with parameters λ_j. We further assume that $f(T_i) = \log(T_i + 1)$. This allows us to derive the following expression for

[3] Not coincidentally, this representation of the OLS estimator closely resembles the representation of the Wald estimator with a binary instrument and continuous treatment as an first-stage effect-weighted average of derivatives at the mean values T_i^* [2].

[4] Code to reproduce all figures and tables in this section are available from https://github.com/winston-chou/rorr.

the conditional expected derivative of Y_i with respect to T_i given X_i (see Appendix A.1):

$$E[f'(T_i)|X = j] = \frac{1 - \exp(-\lambda_j)}{\lambda_j}. \tag{18}$$

The ACD is then just $\sum_j \pi_j \frac{1-\exp(-\lambda_j)}{\lambda_j}$.

We can also derive the RORR plim analytically as:

$$\hat{\theta} \xrightarrow{p} \tilde{\theta} = \frac{\sum_j^J \pi_j E[(T_i - \lambda_j)^2 f'(T_i^*)|X_i = j]}{\sum_j^J \pi_j \lambda_j}, \tag{19}$$

where, as before, T_i^* is a point between T_i and λ_j. Note the two biases relative to the ACD: First, rather than evaluate f' at T_i, we evaluate it at T_i^*. Second, we additionally weight each $f'(T_i^*)$ by its squared deviation from the mean.

Figure 1 illustrates the resulting bias via simulation from this data-generating process. First, we plot $f(T_i) = \log(T_i + 1)$ in top panel of Figure 1, as well as tangent lines with slopes equal to $E[f'(T_i)]$ in blue and to $E[f'(\omega_i T_i^*)]$ in red, where $\omega_i T_i^*$ is the "effective" treatment analyzed by RORR. The key takeaway is that RORR targets a quantity other (and smaller) than the ACD.[5] The subsequent panels give intuition for this result: After weighting by ω_i and transforming T_i to T_i^*, the effective treatment distribution is much more right-skewed than the observed treatment distribution. This means that we tend to evaluate the slope of f at higher values of T_i. This leads to negative bias because $f''(T_i) < 0$.

In Table 1, we report the estimated empirical RORR from simulations at varying sample sizes. For comparison, we also report the empirical ACD (calculated as the sample mean of $\frac{1}{T_i+1}$) and the true ACD computed using (18). Note that, because T_i is integer-valued in this example, the more appropriate causal estimand is arguably the Average Incremental Effect (AIE), defined as:

$$E[Y_i(T_i + 1) - Y_i(T_i)] = \sum_{t=0}^{\infty} (f(t + 1) - f(t))p(t), \tag{20}$$

where p is the mass function of T_i. However, because the RORR plim is a weighted average of derivatives, we focus on the ACD in Table 1 and propose a consistent estimator of the AIE in Section 4. As Table 1 shows, the RORR plim is negatively biased for the ACD. This is due to the fact that it places more weight on the derivative of the dose-response curve at larger values of the treatment. In the following section, we propose a consistent estimator for the AIE with integer-valued treatments and the ACD with continuous treatments.

4 COARSENED AUGMENTED IPW ESTIMATOR

A common benchmark for estimating the Average Causal Derivative (ACD) with continuous treatments is the Generalized Propensity Score (GPS) [14]. However, GPS requires estimating the conditional density of the treatment, which can suffer from slow rates and instability without simplifying parametric assumptions [16].

[5] An analogous result in the welfare economics literature is that OLS up-weights the slopes of higher-income groups in regressions of consumption on income, leading to attenuation [22].

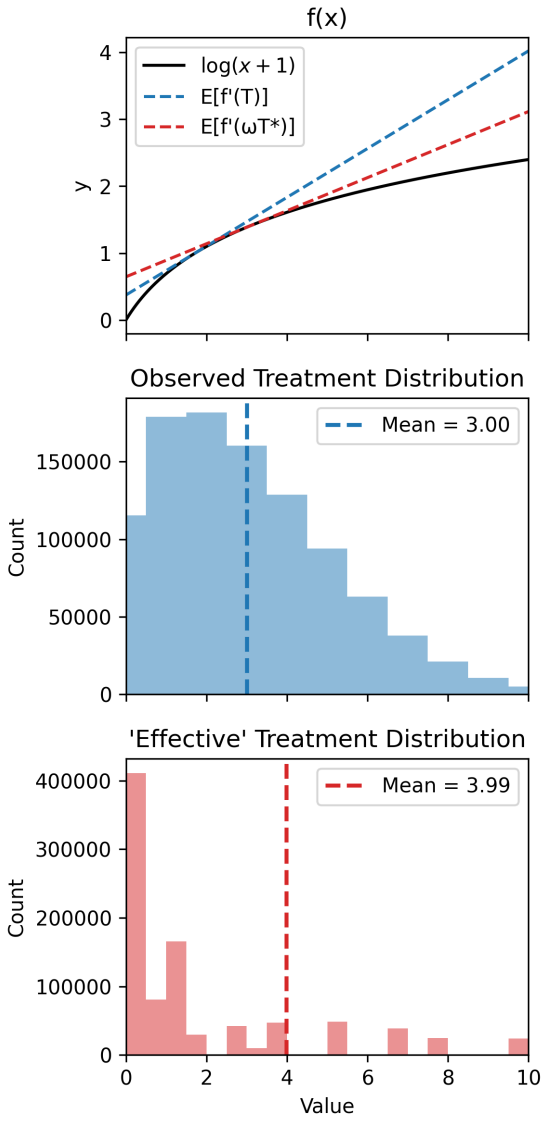

Figure 1: Bias of Residuals-on-Residuals Regression

Table 1: Simulation Results for RORR and ACD

Sample Size	Empirical RORR	RORR 95% CI	RORR Plim
10,000	0.250	(0.237, 0.262)	0.247
100,000	0.248	(0.244, 0.252)	0.247
1,000,000	0.248	(0.247, 0.249)	0.247
Sample Size	**Empirical ACD**	**ACD 95% CI**	**True ACD**
10,000	0.362	(0.357, 0.367)	0.365
100,000	0.364	(0.362, 0.366)	0.365
1,000,000	0.365	(0.364, 0.365)	0.365

As an alternative to both RORR and GPS, we propose a simple coarsened Augmented Inverse Propensity Weighting (AIPW) estimator, which uses the AIPW estimator of counterfactual means as building blocks [7]. This estimator proceeds by first partitioning the support of T_i into K disjoint segments (or bins) $\{S_1, S_2, \ldots, S_K\}$ with $S_k = [t_{k-1}, t_k)$. For each bin S_k, we estimate the bin-level propensity score $p_k(X_i) := \Pr(T_i \in S_k \mid X_i)$, for example by fitting a multiclass classification model. Denote this estimate by $\hat{p}_k(X_i)$. We also estimate a flexible outcome regression for $m_k(X_i) := \mathbb{E}[Y_i \mid T_i \in S_k, X_i]$, which we denote by $\hat{m}_k(X_i)$.

Next, we form the usual AIPW estimator for the marginal mean in bin S_k:

$$\hat{\psi}_k := \frac{1}{N}\sum_{i=1}^{N}\left(\left[\frac{\mathbf{1}(T_i \in S_k)}{\hat{p}_k(X_i)}(Y_i - \hat{m}_k(X_i))\right] + \hat{m}_k(X_i)\right).$$

Like the RORR, this estimator is doubly robust, meaning that if either $\hat{p}_k$ or $\hat{m}$ is consistently estimated, then $\hat{\psi}_k$ is also consistent. Note that the plim of $\hat{\psi}_k$ has a somewhat subtle interpretation due to averaging over the bin. Under standard assumptions, it can be interpreted as the population-average potential outcome if units are treated under the conditional distribution of T_i given X_i and $T_i \in S_k$.

We further define weights w_k for each bin that are proportional to their empirical proportion:

$$w_k = \begin{cases} \frac{\widehat{\Pr}(T_i \in S_k)}{\sum_{\ell=1}^{K-1}\widehat{\Pr}(T_i \in S_\ell)} & \text{for } k < K \\ 0 & \text{for } k = K. \end{cases}$$

Lastly, the overall ACD estimate is given by:

$$\hat{\psi} := \sum_{k=1}^{K-1} w_k \left(\frac{\hat{\psi}_{k+1} - \hat{\psi}_k}{\bar{t}_{k+1} - \bar{t}_k}\right),$$

where $\bar{t}_k = \frac{t_{k+1}+t_k}{2}$ is the midpoint of S_k. Heuristically, $\hat{\psi}$ approximates f' using a piecewise linear function and computes its weighted average using the empirical distribution of the lower segment. Appendix B proves the consistency of this estimator for the ACD as both N and K tend to infinity.

Table 2 shows the results of applying this estimator to the simulated data from Section 3. Because the treatment is integer-valued, we can simply set the bins to each observed treatment value. As the table shows, this yields a consistent estimator for the AIE.[6]

Table 2: Simulation Results for Coarsened AIPW and AIE

Sample Size	Empirical AIE	AIE 95% CI	True AIE
10,000	0.277	(0.247, 0.308)	0.295
100,000	0.291	(0.282, 0.300)	0.295
1,000,000	0.295	(0.290, 0.300)	0.295

[6]Note that the AIE is less than the ACD because $f(t+1) - f(t) = \log\left(1 + \frac{1}{t+1}\right) \leq \frac{1}{t+1}$ for all $t \geq 0$.

5 EMPIRICAL APPLICATION

We now demonstrate the empirical relevance of our theoretical analysis using real-world data from Netflix. Although we are limited in what we can share for confidentiality reasons, the main thrust of this section is to show that the theoretical biases discussed above can (and, in our experience, often do) appear in real-world data.

In this particular application, we sought to understand how the use of a feature, which we will call Feature A, affects future visits to Netflix. To answer this question, we drew a random sample of 2,971,128 members and counted the number of times they used Feature A over a 28 day window. We then divided this number by the member's count of visits to Netflix over the same period to define our continuous treatment, `Feature A Usage Rate`. Next, we defined our outcome as the count of each member's visits to Netflix in the next 28 day window. As covariates, we included the count of times each member used Feature A and the count of times each member visited Netflix in the seven, 14, and 28 days preceding the treatment period.[7]

We divided our dataset into roughly equal-sized training, validation, and test datasets consisting of ≈980,000 units each. To estimate the nuisance parameters in the PLM, we fit gradient boosted regression trees to the treatment and outcome variables observed in the training dataset, using the validation dataset to tune the number of boosting rounds. Lastly, we regressed the outcome residuals on the treatment residuals in the test dataset to obtain the RORR treatment estimate, which is shown in Table 3. As the table shows, the RORR estimate of the effect of Feature A on subsequent visits is small, negative, and statistically significant.[8] This finding contradicted our prior belief that Feature A would increase visits to Netflix.

Table 3: RORR and AIPW Estimates of the Effect of Feature A Usage on Netflix Visits (N = 980,139)

	RORR	Std. Err.	95% CI
Feature A Usage Rate	-0.0038	0.001	(-0.005, -0.002)
	AIPW	Std. Err.	95% CI
Feature A Usage Rate	5.343	0.010	(5.324, 5.362)

Our coarsened AIPW estimator helps provide intuition for this puzzling result. To fit this estimator, we first coarsened the treatment into five bins and then fit a multiclass classifier using gradient boosting to the resulting bins. We assigned zero values (i.e., no usage of Feature A during the treatment period) to the first bin and then divided the remaining non-zero values into quartiles. For the outcome regression, we reused the same function used to fit the RORR.

Figure 2 is a standard diagnostic that plots the difference in the standardized pretreatment value of the outcome in each bin and bin 1 before and after inverse propensity score weighting. As the figure shows, IPW significantly reduces pretreatment differences

[7]We complemented these six covariates with an additional 25 covariates; most of these measured the usage of other Netflix features in the 28 days preceding to the treatment period.

[8]Note that treatment effects are reported after standardizing the treatment and outcome by their respective standard deviations.

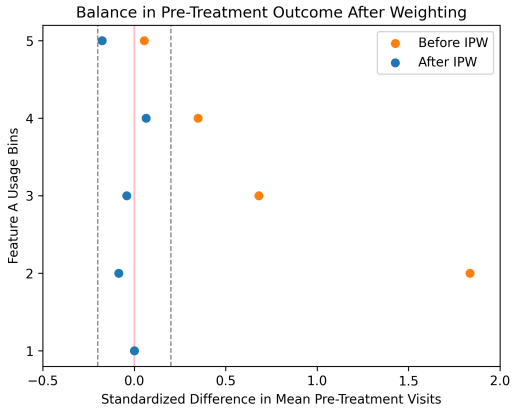

Figure 2: Balance in Pre-Treatment Outcomes After Inverse Propensity Score Weighting

in the outcome variable, making the bins more comparable to each other and strengthening the credibility of conditional ignorability.

We plot the main results in Figure 3, whose panels show, from top to bottom, the counterfactual mean of the post-treatment outcome in each treatment bin; the estimated treatment effect associated with incrementing each bin; and lastly the proportion of the dataset in each bin, which is clearly concentrated in the zero-usage bin.

As Figure 3 shows, AIPW estimates a large *positive* treatment effect of moving from the zero-usage bin (bin 1) to the next bin (bin 2). Moreover, because Feature A usage is zero-inflated, bin 1 is the most representative bin. Therefore, as shown in Table 3, the coarsened AIPW estimate is positive, statistically significant, and substantially larger in magnitude than the RORR estimate. This discrepancy arises because the coarsened AIPW estimator explicitly weights the treatment effects to be representative of the treatment distribution, whereas RORR up-weights units with higher values of the treatment, where the dose-response curve is downward-sloping. The discrepancy is highly relevant for decision making: Although RORR indicates that Feature A has a negative treatment effect on the outcome, the AIPW results show that increasing Feature A usage would have a positive effect for the vast majority of members. Indeed, *all* nonzero Feature A usage bins have a higher conditional means than the zero-usage bin, indicating that any Feature A usage is preferable to none.

6 CONCLUSION

Although DML estimators are becoming increasingly popular in both academic and commercial research, researchers must – as ever – carefully evaluate their suitability for specific applications. Focusing on the special case of residuals-on-residuals regression (RORR), this paper studies the proper interpretation of RORR when treatment effects are heterogeneous and/or when treatments are many-valued with nonlinear dose-response functions. We show that, in the latter case, RORR converges to a conditional variance-weighted average of causal derivatives, with the added complication that these derivatives are evaluated on a "pseudo-treatment" distribution that differs from the treatment distribution seen in the data. As our empirical application shows, the subtle biases of RORR relative to any "average" treatment effect can have significant consequences for decision-making. To address these biases, we propose a coarsened AIPW estimator and show that it yields more representative estimates of causal effects.

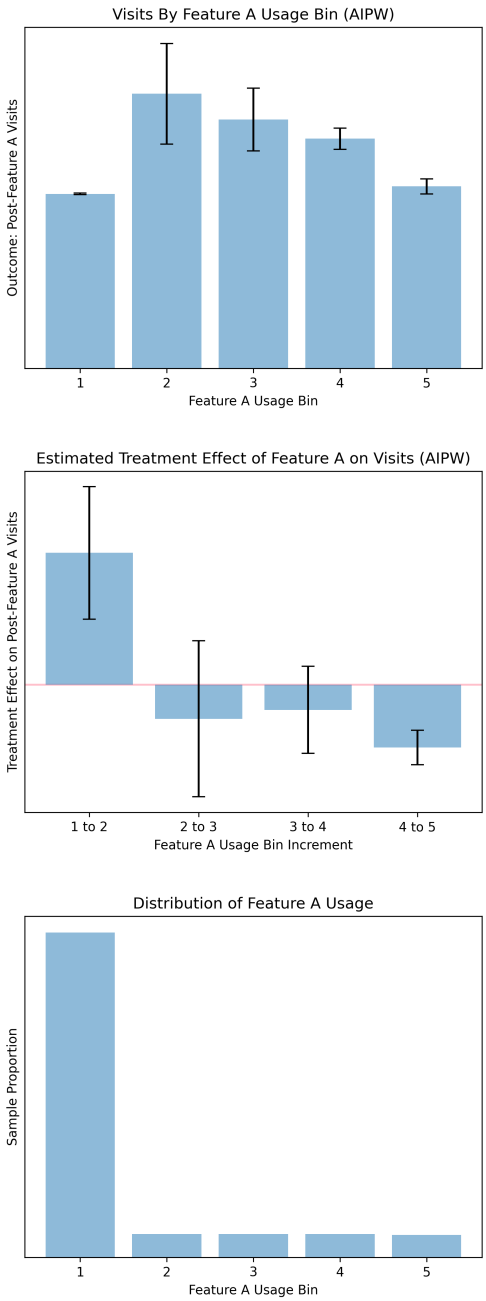

Figure 3: Treatment Effects of Feature A Usage on Netflix Visits After AIPW Weighting. Note that we remove y-axis labels to preserve confidentiality.

REFERENCES

[1] Joshua D Angrist. [n. d.]. Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants. 66 ([n. d.]), 249–288. Issue 2. http://www.jstor.org/stable/2998558

[2] Joshua D Angrist and Alan B Krueger. [n. d.]. *Chapter 23 - Empirical Strategies in Labor Economics*. Vol. 3. Elsevier, 1277–1366. http://www.sciencedirect.com/science/article/pii/S1573446399030047

[3] Peter M Aronow and Cyrus Samii. [n. d.]. Does Regression Produce Representative Estimates of Causal Effects? 60 ([n. d.]), 250–267. Issue 1. https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12185

[4] Anna Baiardi and Andrea A. Naghi. 2024. The Effect of Plough Agriculture on Gender Roles: A Machine Learning Approach. *Journal of Applied Econometrics* 39, 7 (2024), 1396–1402. https://doi.org/10.1002/jae.3083

[5] Peter J Bickel, Chris AJ Klaassen, Peter J Bickel, Ya'acov Ritov, J Klaassen, Jon A Wellner, and YA'Acov Ritov. 1993. *Efficient and adaptive estimation for semiparametric models*. Vol. 4. Johns Hopkins University Press Baltimore.

[6] Richard L. Burden and J. Douglas Faires. 2011. Numerical Analysis.

[7] Matias D Cattaneo. 2010. Efficient semiparametric estimation of multi-valued treatment effects under ignorability. 155 (2010), 138–154. Issue 2. https://www.sciencedirect.com/science/article/pii/S030440760900236X

[8] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. [n. d.]. Double/debiased machine learning for treatment and structural parameters. 21 ([n. d.]), C1–C68. Issue 1. http://doi.wiley.com/10.1111/ectj.12097

[9] Victor Chernozhukov, Hiroyuki Kasahara, and Paul Schrimpf. 2020. Causal impact of masks, policies, behavior on early covid-19 pandemic in the US. *Journal of econometrics* 220, 1 (2020), 23.

[10] Oliver Dukes, Stijn Vansteelandt, and David Whitney. 2024. On Doubly Robust Inference for Double Machine Learning in Semiparametric Regression. *Journal of Machine Learning Research* 25, 279 (2024), 1–46.

[11] Daniel Fink, Alison Johnston, Matt Strimas-Mackey, Tom Auer, Wesley M. Hochachka, Shawn Ligocki, Lauren O. Jaromczyk, Orin Robinson, Chris Wood, Steve Kelling, and Amanda D. Rodewald. 2023. A Double Machine Learning Trend Model for Citizen Science Data. *Methods in Ecology and Evolution* 14 (2023), 2435–2448. https://doi.org/10.1111/2041-210X.14186

[12] David Holtz, Michael Zhao, Seth G Benzell, Cathy Y Cao, Mohammad Amin Rahimian, Jeremy Yang, Jennifer Allen, Avinash Collis, Alex Moehring, Tara Sowrirajan, et al. 2020. Interdependence and the cost of uncoordinated responses to COVID-19. *Proceedings of the National Academy of Sciences* 117, 33 (2020), 19837–19843.

[13] G Imbens. [n. d.]. Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review. 86 ([n. d.]), 4–29. Issue 1. https://doi.org/10.1162/003465304323023651 doi: 10.1162/003465304323023651.

[14] G Imbens. 2000. The role of the propensity score in estimating dose-response functions. 87 (2000), 706–710. Issue 3. https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/87.3.706

[15] Jake Alton Jares and Neil Malhotra. 2023. Policy impact and voter mobilization: Evidence from farmers' trade war experiences. *American Political Science Review* (2023), 1–23.

[16] Edward H Kennedy, Zongming Ma, Matthew D McHugh, and Dylan 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, 4 (2017), 1229–1245.

[17] Apoorva Lal. 2024. Does Regression Produce Representative Causal Rankings? arXiv:2411.02675 [econ.EM] https://arxiv.org/abs/2411.02675

[18] James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. 1994. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association* 89, 427 (1994), 846–866.

[19] P M Robinson. [n. d.]. Root-N-Consistent Semiparametric Regression. 56 ([n. d.]), 931–954. Issue 4. http://www.jstor.org/stable/1912705

[20] Tymon Słoczyński. [n. d.]. Interpreting OLS Estimands when Treatment Effects are Heterogeneous. ([n. d.]).

[21] Xinyu Wei, Mingwang Cheng, Kaifeng Duan, and Xiangxing Kong. 2024. Effects of Big Data on $PM_{2.5}$: A Study Based on Double Machine Learning. *Land* 13, 3 (2024), 327. https://doi.org/10.3390/land13030327

[22] Shlomo Yitzhaki. [n. d.]. On Using Linear Regressions in Welfare Economics. 14 ([n. d.]), 478–486. Issue 4. http://www.jstor.org/stable/1392256

A DERIVATIONS

A.1 Derivation of Equation 18

Note that, in our stylized example, T_i is Poisson(λ_j) conditional on $X_i = j$. Therefore, its probability mass function is given by:

$$P(T_i = t|X_i = j) = e^{-\lambda_j}\frac{\lambda_j^t}{t!}, t = 0, 1, 2, \ldots, \tag{21}$$

and the conditional expectation of the causal derivative $\frac{1}{T_i+1}$ given $X_i = j$ is:

$$\sum_{t=0}^{\infty}\frac{1}{t+1}e^{-\lambda_j}\frac{\lambda_j^t}{t!} = \sum_{t=0}^{\infty}e^{-\lambda_j}\frac{\lambda_j^t}{(t+1)!}. \tag{22}$$

Define $s := t + 1$. Then we can rewrite the above as:

$$\sum_{s=1}^{\infty}e^{-\lambda_j}\frac{\lambda_j^{s-1}}{s!} = \frac{e^{-\lambda_j}}{\lambda_j}\sum_{s=1}^{\infty}\frac{\lambda_j^s}{s!} \tag{23}$$

$$= \frac{(e^{\lambda_j}-1)}{\lambda_j e^{\lambda_j}} \tag{24}$$

$$= \frac{1-e^{-\lambda_j}}{\lambda_j}. \tag{25}$$

where the second-to-last equality uses the identity $\sum_{s=0}^{\infty}\frac{x^s}{s!} = e^x$.

A.2 Derivation of T_i^*

Within a given stratum $X_i = j$, $h(X_i) = E[T_i|X_i] = \lambda_j$ by assumption. We want to derive the point T_i^* between T_i and λ_j at which we are evaluating the derivative $f'(t) = 1/(t+1)$. We will condition on X_i throughout.

By the mean value theorem, there is some T_i^* between T_i and λ_j for which:

$$f(T_i) - f(\lambda_j) = f'(T_i^*)(T_i - \lambda_j).$$

Plugging in the definition of f and its derivative, we write this as:

$$\log(T_i+1) - \log(\lambda_j+1) = \frac{1}{T_i^*+1}(T_i - \lambda_j).$$

Solving for T_i^* yields:

$$T_i^* = \frac{T_i - \lambda_j}{\log\frac{T_i+1}{\lambda_j+1}} - 1.$$

Note that, when $T_i = \lambda_j$, this quantity is undefined (in which case we just set $T_i^* = T_i$).

In our simulations, we estimate the theoretical plim of RORR by taking many draws of T_i, plugging them into this formula, and using the resulting stratum means to estimate Equation 19

B PROOF OF CONSISTENCY OF COARSENED AIPW

We assume the PLM for continuous treatments, but introduce the potential outcomes notation:

$$Y_i(t) = f(t) + g(X_i) + e_i \tag{26}$$

$$T_i = h(X_i) + u_i. \tag{27}$$

We assume that f is continuously differentiable, $E[e_i|X_i] = 0$, $E[u_i|X_i] = 0$, and $E[e_iu_i|X_i] = 0$. Note that the last of these assumptions implies that $Y_i(t)$ is conditionally independent of T_i given X_i. We further require positivity of treatment for all x and t:

$$p(t|x) \geq \epsilon > 0, \tag{28}$$

where $p(t|x)$ is the continuous conditional density of T_i given X_i. This is the continuous analog of the overlap assumption under binary or discrete treatments, which requires that every unit has a positive probability of treatment given covariates.

Partition the domain of T_i into K equal-spaced segments, denoted by $S_k, k = 1, \ldots, K$, with $S_k = [t_k, t_{k+1})$ and $t_{k+1} - t_k = CK^{-1}$ for all k, where C is some positive constant that spans the domain of T_i. Define $p_k(X_i) = \Pr(T_i \in S_k|X_i)$ and $m_k(X_i) = E[Y_i|T_i \in S_k, X_i]$ and let one or both $\hat{p}_k$ and $\hat{m}_k$ be consistent estimators.

The coarsened AIPW estimator for a given segment S_k is defined as:

$$\hat{\psi}_k := \frac{1}{N}\sum_{i=1}^{N}\frac{1(T_i \in S_k)}{\hat{p}_k(X_i)}(Y_i - \hat{m}_k) + \hat{m}_k. \tag{29}$$

Further define the approximate weights:

$$w_k = \begin{cases}\frac{\Pr(T_i \in S_k)}{\Pr(T_i \notin S_K)} & \text{for } k < K \\ 0 & \text{for } k = K,\end{cases} \tag{30}$$

and denote the midpoint of bin S_k by $\bar{t}_k = \frac{t_{k+1}-t_k}{2}$.

Then coarsened AIPW estimator of the Average Causal Derivative is defined as

$$\hat{\psi} := \sum_{k=1}^{K-1} w_k\left(\frac{\hat{\psi}_{k+1} - \hat{\psi}_k}{\bar{t}_{k+1} - \bar{t}_k}\right).$$

Theorem 1. *$\hat{\psi}$ converges in probability to the Average Causal Derivative $E[f'(T_i)]$ as both the number of observations N and the number of segments $K \to \infty$.*

Proof. By consistency of $\hat{p}_k$ or $\hat{m}_k$ and conditional independence of $Y_i(t)$ and X_i,

$$\hat{\psi}_k \xrightarrow{p} \int\int Y_i(t)\frac{p(t|x)1(t \in S_k)}{\Pr(t \in S_k|x)}p(x)dtdx. \tag{31}$$

Subtracting $\hat{\psi}_k$ from $\hat{\psi}_{k+1}$ and plugging in the definition of Y_i yields:

$$\begin{aligned}\hat{\psi}_{k+1} - \hat{\psi}_k &\xrightarrow{p} \int\left(\int f(t)\frac{p(t|x)1(t \in S_{k+1})}{\Pr(t \in S_{k+1}|x)}dt \right. \\ &\left. - \int f(t)\frac{p(t|x)1(t \in S_k)}{\Pr(t \in S_k|x)}dt\right)p(x)dx \\ &:= \Delta_k.\end{aligned} \tag{32}$$

Δ_k has a subtle interpretation: It is the "effect" of moving from segment S_k to S_{k+1} when units are treated according to the conditional treatment distribution in each segment. We now show that Δ_k can be interpreted as a very specific ATE. Given continuity of f, p, and positivity, the mean value theorem for integrals states that there exists a $\tilde{t}_k \in S_k$ such that

$$\int f(t)\frac{p(t|x)1(t \in S_k)}{\Pr(t \in S_k|x)}dt = f(\tilde{t}_k).$$

Therefore, we can rewrite the above as:

$$\Delta_k = \int (f(\tilde{t}_{k+1}) - f(\tilde{t}_k))p(x)dx, \tag{33}$$

showing that Δ_k is the ATE of fixing T_i to $\tilde{t}_{k+1}$ compared to $\tilde{t}_k$.

By another application of the mean value theorem, there exists a t_k^* between $\tilde{t}_k$ and $\tilde{t}_{k+1}$ such that:

$$\begin{aligned} & \int (f(\tilde{t}_{k+1}) - f(\tilde{t}_k))p(x)dx \\ = & \int f'(t_k^*)(\tilde{t}_{k+1} - \tilde{t}_k)p(x)dx \\ = & E[f'(t_k^*)(\tilde{t}_{k+1} - \tilde{t}_k)]. \end{aligned} \tag{34}$$

Note that the absolute difference between any of $\tilde{t}_k$, t_k^*, and $\bar{t}_k$ is bounded by $|2CK^{-1}|$. Therefore, as $K \to \infty$, these converge to the same point. Consequently:

$$\sum_{k=1}^{K-1} w_k \left(\frac{\hat{\psi}_{k+1} - \hat{\psi}_k}{\bar{t}_{k+1} - \bar{t}_k} \right) \xrightarrow{p} \int f'(t) \frac{t_{k+1} - t_k}{t_{k+1} - t_k} p(t)dt = E[f'(t)]. \tag{35}$$

□

C CHOOSING THE NUMBER OF SEGMENTS

Choosing the number of segments K in which to bin the treatment involves the usual considerations of bias and variance. In many practical applications, interpretability is also a goal. In this section, we provide a semiformal justification, relying on many simplifying assumptions, for choosing $K = O(N^{1/7})$. This implies fewer than ten bins for medium to large datasets (i.e., between 100,000 to 1,000,000 units).

In our practical work at Netflix, we find that a relatively small number of bins – for example, dividing users into low, medium, and high usage segments – is sufficient to detect meaningful heterogeneity in the dose-response function. Although using a small number of bins increases bias, it reduces variance, adds robustness to slow nuisance estimation rates, and increases interpretability.

The MSE of the coarsened AIPW estimator is:

$$MSE(\hat{\psi}) = \underbrace{(E[\hat{\psi}] - E[f'(t)])^2}_{\text{Bias}^2} + \text{Var}(\hat{\psi}). \tag{36}$$

We begin by decomposing the Bias term into three components:

$$\begin{aligned} & E[\hat{\psi}] - E[f'(t)] \\ & = \underbrace{E\left[\sum_{k=1}^{K-1} w_k \left(\frac{\hat{\psi}_{k+1} - \hat{\psi}_k}{\bar{t}_{k+1} - \bar{t}_k} \right)\right] - \sum_{k=1}^{K-1} w_k \left(\frac{\Delta_k}{\bar{t}_{k+1} - \bar{t}_k} \right)}_{:=a} \\ & \quad + \underbrace{\sum_{k=1}^{K-1} w_k \left(\frac{\Delta_k}{\bar{t}_{k+1} - \bar{t}_k} \right) - \sum_{k=1}^{K-1} w_k f'(t_k)}_{:=b} \\ & \quad + \underbrace{\sum_{k=1}^{K-1} w_k f'(t_k) - E[f'(t)]}_{:=c}. \end{aligned} \tag{37}$$

In other words, a is statistical estimation error, b is the bias incurred by approximating f by a piecewise linear function, and c is the bias incurred by approximating an integral by a series of rectangles.

For simplicity, we will assume that T_i is uniformly distributed, such that the number of units in each bin $N_k = N/K$, $\bar{t}_{k+1} - \bar{t}_k$ is a constant $\ell \propto K^{-1}$, and $w_k = (K-1)^{-1}$ for all k. We will also assume that f is at least thrice continuously differentiable. Also for simplicity, we ignore cross-fitting and assume that the nuisance functions are estimated on a separate dataset of equal size (i.e., also consisting of N units partitioned into K bins). We assume the estimated nuisances satisfy standard regularity conditions, specifically $\delta < \hat{p}(x) < 1 - \delta$ for some $\delta \in (0, 1)$ and $\hat{m}(x, k)^2 < \infty$ almost surely and that $Y_i^2 < \infty$, so that $\text{Var}(\psi_i) = O(1)$.

Suppose that $\hat{\psi}_k - m_k = o_p\left((N/K)^{-1/2}\right)$. Then, under mild regularity conditions, $a = \sum_{k=1}^{K} O\left((N/K)^{-1/2}\right) = O(K^{3/2}N^{-1/2})$. The error of the midpoint approximation b when f is thrice continuously differentiable is known to be $O(\ell^2) = O(K^{-2})$ [6, p. 177]. The error of the Riemann sum approximation is also known to be $c = O(K^{-2})$ [6, p. 207]. Therefore, the error of the Bias^2 term is:

$$(a + b + c)^2 = O(K^3N^{-1}) + O(K^{-4}). \tag{38}$$

The variance is:

$$\begin{aligned} \text{Var}(\hat{\psi}) &= \text{Var}\left(\sum_{k=1}^{K-1} w_k \left(\frac{\hat{\psi}_{k+1} - \hat{\psi}_k}{b} \right)\right) \\ &= K \,\text{Var}(\hat{\psi}_K - \hat{\psi}_1) \\ &= O(K^2N^{-1}). \end{aligned} \tag{39}$$

Putting these together, we obtain

$$MSE(\hat{\psi}) = O(K^3N^{-1}) + O(K^{-4}) + O(K^2N^{-1}).$$

The K^* that minimizes this is $O(N^{1/7})$.

More generally, suppose that ψ_k is $n^{-1/d}$ consistent for $d > 0$. Then the Bias^2 term is $O(K^{2(d+1)/d}N^{-2/d}) + O(K^{-4})$, while the variance remains $O(K^2N^{-1})$. If $d < 2$, the variance term dominates and the optimal $K^* = O\left(N^{1/6}\right)$. If $d \geq 2$, the bias term dominates, and the optimal $K^* = O\left(N^{1/(3d+1)}\right)$.