

Doubly Robust Nonparametric Local Projections

Giorgi Nikolaishvili*
Wake Forest University
nikolag@wfu.edu

This Version: April 9, 2026
([Access the Most Recent Version](#))

Abstract

Nonparametric local projections estimate impulse responses without functional form restrictions, but their consistency hinges on the accuracy of the conditional mean regression. I propose an estimator that supplements this regression with a bias correction derived from the marginal density of the shock. This estimator is doubly robust: it remains consistent when either the conditional mean or the density ratio is misspecified, and attains the semiparametric efficiency bound when the product of their errors vanishes faster than the parametric rate. This bound decomposes into the regression-only variance plus an explicit cost-of-robustness term. Monte Carlo simulations confirm the double robustness property and suggest that the efficiency cost of robustness is modest — a small premium for insurance against misspecification of the conditional mean, which in nonparametric settings cannot be ruled out.

JEL Codes: C14, C22, C32

Keywords: local projections; double robustness; semiparametric efficiency; impulse responses

*Computations were performed using the Wake Forest University High Performance Computing Facility. All errors are my own.

1 Introduction

Macroeconomic dynamics can be highly nonlinear: structural shocks may produce asymmetric effects, state-dependent propagation, and shifts in the shape of outcome distributions that linear methods cannot detect. Nonparametric local projections impose no functional form on how shocks propagate and are a natural tool for this setting (Gonçalves et al., 2024a). Yet the consistency of existing estimators hinges on the accuracy of the nonparametric regression of the outcome on the shock. If that regression is inaccurate, the resulting impulse responses inherit the error. The source of inaccuracy may be outright misspecification (the function class excludes the true g_h), slow convergence due to bandwidth-driven bias, or degraded nonparametric rates from a high-dimensional conditioning set; the regression estimator is equally vulnerable in each case. The researcher has no way to diagnose this failure.

This paper proposes a doubly robust estimator for nonparametric local projections. The estimator supplements the conditional mean regression with a bias correction derived from the density ratio of the shock, which depends only on the marginal shock density and therefore avoids the curse of dimensionality that affects the regression in higher-dimensional settings. Consistency requires only that one of the two components be correctly specified. When the product of their estimation errors vanishes faster than the parametric rate, the estimator attains the semiparametric efficiency bound and permits asymptotically valid inference via its influence function.

The paper brings together several strands of the literature. Gonçalves et al. (2024a) formulate nonlinear impulse response estimation as a potential outcomes problem and propose a nonparametric regression-based estimator; I adopt their structural model, potential outcomes definitions, and identification results as my starting point. Angrist and Kuersteiner (2011) and Angrist et al. (2018) use propensity score weighting to estimate the causal effects of discrete monetary policy interventions, providing an early precedent for reweighting methods in macroeconomic policy evaluation. The treatment effects literature progressively extended doubly robust ideas from binary treatments (Robins et al., 1994) to continuous treatments (Hirano and Imbens, 2004; Kennedy et al., 2017) and to flexible machine-learning nuisance estimation via sample splitting (Chernozhukov et al., 2018). In a complementary direction, Montiel Olea et al. (2024) show that the LP estimator is doubly robust: its bias under dynamic misspecification equals the product of errors in the outcome and first-stage specifications, so LP confidence intervals remain reliable even under detectable misspecification while VAR intervals can severely undercover. This strengthens the case for LP-based inference in the linear setting.

The present paper adapts these semiparametric tools to the nonparametric impulse response setting, where serial dependence, the continuous and unbounded nature of structural shocks, and the availability of i.i.d. shock sequences create both new challenges and new simplifications relative to cross-sectional applications.

The paper makes four contributions. First, I establish a dual identification result: the average response function can be expressed not only through the conditional mean regression of [Gonçalves et al. \(2024a\)](#), but also through a density ratio that reweights the observed shock distribution to match the counterfactual distribution under a shifted shock. Because the density ratio depends only on the marginal shock density, a univariate object regardless of the dimension of the structural model, this second route is immune to the curse of dimensionality. Having two independent paths to the same estimand is what makes doubly robust estimation possible.

Second, I show that combining these two paths yields an estimator that is consistent when either the conditional mean or the density ratio is well specified, a form of insurance that existing nonparametric local projection estimators do not provide. When both components converge at rates whose product exceeds the parametric rate, the estimator attains the semiparametric efficiency bound at the impact horizon. At longer horizons the overlapping structure of local projections inflates the asymptotic variance, but this cost is common to all LP estimators and is handled by HAC inference. For the kernel-based nuisance estimators I propose, no sample splitting is required; when the researcher prefers more flexible machine learning methods, the debiased machine learning framework of [Chernozhukov et al. \(2018\)](#) can be applied using sequential block splits suited to dependent data.

Third, I characterize the semiparametric efficiency bound and show that it decomposes into the variance of a regression-only estimator plus an augmentation term that measures the irreducible cost of not knowing the conditional mean ex ante. The augmentation term is the exact variance premium a researcher pays for robustness to misspecification of the regression surface. This decomposition gives applied researchers a concrete way to assess whether the insurance provided by double robustness is worth its price in a given application.

Fourth, the doubly robust estimator offers additional advantages for conditional average responses with higher-dimensional conditioning sets. In this setting, the joint regression of the outcome on both the shock and the conditioning variables converges slowly, and the regression-based estimator of [Gonçalves et al. \(2024a\)](#) inherits this slow convergence directly as bias. The doubly robust estimator mitigates this problem because its bias equals the product of the regression error and the density ratio error, and the

density ratio is a univariate object that converges fast regardless of the dimension of the conditioning set. A well-estimated density ratio can therefore partially compensate for a poorly estimated regression surface, suggesting that the robustness gains may be especially useful in state-dependent impulse response estimation.

Outline. Section 2 presents the structural model, identification results, and a dual representation of the average response function via density ratios that underpins the doubly robust estimator. Section 3 examines the regression-based and reweighting-based estimators side-by-side, motivating the need for a combined approach. Section 4 develops the doubly robust estimator, establishes its double robustness and asymptotic properties, and discusses estimation of the density ratio nuisance function. Section 5 characterizes the semiparametric efficiency bound and its variance decomposition. Section 6 extends the framework to conditional average responses. Section 7 reports Monte Carlo simulations, with figures collected in Appendix B. Formal assumptions and proofs are collected in Appendix A.

2 Setup and Identification

An impulse response function answers a counterfactual question: if the structural shock at time t were shifted by δ , how would outcomes at time $t + h$ change on average? Formalizing this question requires a structural model that separates the shock of interest from the rest of the system, a potential outcomes framework that defines the counterfactual, and an identification argument that links the counterfactual to observable quantities. This section develops each ingredient and shows that identification can proceed along two independent routes — one based on the conditional mean of the outcome given the shock, the other based on the density of the shock itself. The duality between these routes is the foundation for doubly robust estimation.

2.1 Structural Model

Consider a vector of observables $\mathbf{z}_t = (x_t, Y_t)'$ generated by the structural dynamic nonlinear system

$$x_t = \varphi(\mathbf{z}_{t-1}) + \varepsilon_{1t} \tag{1}$$

$$y_{it} = \psi_i(x_t, Y_{-i,t}, \mathbf{z}_{t-1}, \varepsilon_{it}), \quad i = 2, \dots, n \tag{2}$$

where $\varepsilon_t = (\varepsilon_{1t}, \dots, \varepsilon_{nt})'$ is i.i.d. with mean zero and mutually independent components. I use boldface for the full lag history of a variable. The first equation isolates a single structural shock ε_{1t} , which enters the system additively; the remaining equations allow outcomes to depend on the shock, on each other, and on the full history through unrestricted nonlinear functions ψ_i . This framework follows the structural model of [Gonçalves et al. \(2024a\)](#), which nests as special cases models with nonlinearly transformed regressors ([Gonçalves et al., 2021](#)), state-dependent coefficients ([Gonçalves et al., 2024b](#)), and nonlinear interactions between shocks and state variables.

For simplicity, I treat ε_{1t} as directly observed, setting $x_t = \varepsilon_{1t}$. This is the natural starting point when an external instrument or narrative identification strategy delivers a shock series. Identification below rests on a single feature of this specification: the mutual independence of the components of ε_t ensures that ε_{1t} is structurally independent of the other innovations.

2.2 Potential Outcomes and Estimands

The structural model defines a mapping from the shock ε_{1t} to future outcomes. To formalize the counterfactual “what if ε_{1t} had been e instead of its realized value?”, define the potential outcome

$$y_{t+h}(e) = m_h(e, U_{t+h}), \quad (3)$$

where U_{t+h} gathers every source of variation in y_{t+h} apart from the contemporaneous shock: lagged states, future shocks $\varepsilon_{1,t+1}, \dots, \varepsilon_{1,t+h}$, and non-shock innovations $\varepsilon_{2,t}, \dots, \varepsilon_{n,t+h}$. The function m_h is determined by the structural equations (1)–(2) but is left unrestricted — no functional form is imposed on how the shock propagates. This potential outcomes formulation was developed in a series of papers by [Gonçalves et al. \(2024b\)](#) and [Gonçalves et al. \(2024a\)](#), building on the earlier work of [Gonçalves et al. \(2021\)](#).

The effect of shifting the shock by δ at horizon h is captured by two estimands.

Average Response Function (ARF). The population-average effect of the shift:

$$\text{ARF}_h(\delta) \equiv E[y_{t+h}(\varepsilon_{1t} + \delta) - y_{t+h}(\varepsilon_{1t})]. \quad (4)$$

This estimand, introduced by [Gonçalves et al. \(2021\)](#) and generalized in [Gonçalves et al. \(2024a\)](#), is a nonparametric generalization of the linear impulse response: it measures the average change in the outcome when every shock realization is perturbed by δ , without

restricting the response to be linear or symmetric in δ .¹

Conditional Average Response (CAR). The effect conditional on an observable state Ω_t :

$$\text{CAR}_h(\delta, \omega) \equiv E[y_{t+h}(\varepsilon_{1t} + \delta) - y_{t+h}(\varepsilon_{1t}) \mid \Omega_t = \omega]. \quad (5)$$

The conditioning set Ω_t (e.g., a recession indicator, an asset price level, or a lagged state variable) allows the researcher to ask whether the same shock produces different effects in different states of the world. This conditional estimand was introduced by [Gonçalves et al. \(2024b\)](#), who showed that standard state-dependent local projections fail to recover it when the state of the economy is endogenous with respect to macroeconomic shocks and the shock magnitude is nonnegligible; [Gonçalves et al. \(2024a\)](#) develop a nonparametric estimator that remains valid in this setting.

2.3 Identification

Both estimands involve the potential outcome $y_{t+h}(e)$, which is observed only at the realized shock value $e = \varepsilon_{1t}$. Identification requires a link between the counterfactual and observables. The structural model provides this link through a single condition: *the shock ε_{1t} is independent of all other determinants of y_{t+h}* . Formally, $\varepsilon_{1t} \perp\!\!\!\perp U_{t+h}$, which follows from the i.i.d. assumption on ε_t and the triangular timing in equations (1)–(2): because ε_{1t} is serially independent and contemporaneously uncorrelated with $\varepsilon_{2t}, \dots, \varepsilon_{nt}$, it is independent of the collection U_{t+h} that governs the potential outcome mapping. This independence result is formalized in Lemma A.1 of [Gonçalves et al. \(2024b\)](#) and [Gonçalves et al. \(2024a\)](#), who show that it implies the potential outcomes $\{y_{t+h}(e) : e \in \mathcal{E}\}$ are independent of ε_{1t} — the analog of the unconfoundedness assumption in the treatment effects literature.

This independence condition opens two distinct routes to the same estimand.

Route 1: Conditional means. Define the conditional mean function

$$g_h(e) \equiv E[y_{t+h} \mid \varepsilon_{1t} = e]. \quad (6)$$

¹An alternative definition, following [Koop et al. \(1996\)](#), compares the potential outcomes $y_{t+h}(e + \delta)$ and $y_{t+h}(e)$ for fixed e rather than averaging over the random variable ε_{1t} . As shown by [Gonçalves et al. \(2024a\)](#), the two definitions coincide in linear models but can differ substantially in nonlinear settings. Computing the counterfactual baseline requires integrating the conditional expectation of y_{t+h} over all possible shock realizations, which is why the appropriate definition averages over the realized shock distribution.

Because $\varepsilon_{1t} \perp\!\!\!\perp U_{t+h}$, the conditional expectation of the potential outcome at any shock value e equals $g_h(e)$:

$$E[y_{t+h}(e)] = E[m_h(e, U_{t+h})] = E[y_{t+h} \mid \varepsilon_{1t} = e] = g_h(e),$$

where the first equality integrates out U_{t+h} (using independence), and the second uses the fact that $y_{t+h} = m_h(\varepsilon_{1t}, U_{t+h})$ and conditions on $\varepsilon_{1t} = e$. Substituting into the ARF definition yields the regression representation:

$$\text{ARF}_h(\delta) = E[g_h(\varepsilon_{1t} + \delta) - g_h(\varepsilon_{1t})]. \quad (\text{CM})$$

This says the impulse response is identified by the conditional mean regression of y_{t+h} on ε_{1t} : estimate g_h , evaluate it at the observed and shifted shock values, and average the difference. This identification result and the regression-based estimator it motivates are due to [Gonçalves et al. \(2024a\)](#); see their Proposition 4.1 and Algorithm 5.1.

Route 2: Density ratios. Rather than asking how the conditional mean changes along the shock axis, we can ask what the distribution of outcomes would look like if the shocks had been shifted. The idea is to leave the outcome data untouched and instead adjust how much each observation counts: if a particular shock realization would have been more common under the shifted distribution, the corresponding outcome receives more weight. Let f denote the marginal density of ε_{1t} . A change of variables gives

$$\begin{aligned} E[g_h(\varepsilon_{1t} + \delta)] &= \int g_h(e + \delta) f(e) de \\ &= \int g_h(u) f(u - \delta) du \\ &= E\left[g_h(\varepsilon_{1t}) \cdot \frac{f(\varepsilon_{1t} - \delta)}{f(\varepsilon_{1t})}\right]. \end{aligned} \quad (7)$$

Define the density ratio

$$r_\delta(e) \equiv \frac{f(e - \delta)}{f(e)}. \quad (8)$$

This ratio is well-defined provided the support of $f(\cdot - \delta)$ is contained in the support of f — an overlap condition that holds automatically when ε_{1t} has full support on \mathbb{R} (as under Gaussian or other light-tailed distributions) and is stated formally in Assumption [A.2](#). Since $E[g_h(\varepsilon_{1t})] = E[y_{t+h}]$ by iterated expectations, substituting into the ARF definition

yields the reweighting representation:

$$\text{ARF}_h(\delta) = E[y_{t+h} \cdot (r_\delta(\varepsilon_{1t}) - 1)]. \quad (\text{R})$$

The density ratio $r_\delta(e)$ reweights the observed shock distribution to match the counterfactual distribution that would prevail if every shock were shifted by δ . Shock values that become more likely under the shift receive weight $r_\delta(e) > 1$; those that become less likely receive weight $r_\delta(e) < 1$. Because f is a univariate object, the density ratio avoids the curse of dimensionality: it depends only on the marginal shock density regardless of the dimension of the structural model or the conditioning set.

Duality. The two representations place their estimation burden on different objects: the conditional mean g_h and the shock density f , respectively. Having two independent paths to the same estimand is what makes doubly robust estimation possible.

Extension to conditional responses. Both routes extend to the CAR whenever $\varepsilon_{1t} \perp \perp \Omega_t$, a condition that holds when Ω_t is a function of lagged variables since ε_{1t} is i.i.d. The regression route gives $\text{CAR}_h(\delta, \omega) = E[g_h(\varepsilon_{1t} + \delta, \omega) - g_h(\varepsilon_{1t}, \omega) \mid \Omega_t = \omega]$, where $g_h(e, \omega) \equiv E[y_{t+h} \mid \varepsilon_{1t} = e, \Omega_t = \omega]$; see [Gonçalves et al. \(2024a\)](#), Proposition 4.1(ii) and Algorithm 5.2. The reweighting route gives

$$\text{CAR}_h(\delta, \omega) = E[y_{t+h} \cdot (r_\delta(\varepsilon_{1t}) - 1) \mid \Omega_t = \omega]. \quad (\text{R}_\omega)$$

The density ratio remains the same univariate function of ε_{1t} ; conditioning on Ω_t is handled by the outer expectation. The two routes differ in dimensionality: the regression route requires a higher-dimensional regression surface $g_h(e, \omega)$, while the reweighting route keeps the density ratio univariate. This asymmetry will have important implications for estimation in higher-dimensional settings.

3 Two Estimators and Their Vulnerabilities

The identification results in Section 2 provide two independent routes to the same estimand: the conditional mean representation (CM) and the density ratio representation (R). Each route yields a natural estimator, and each has characteristic strengths and weaknesses. Examining them side by side motivates the doubly robust combination developed in Section 4.

3.1 The Regression Estimator

The regression estimator follows directly from the conditional mean representation. Estimate $g_h(e) = E[y_{t+h} \mid \varepsilon_{1t} = e]$ by local linear regression, then form:

$$\widehat{\text{ARF}}_h^{\text{reg}}(\delta) = \frac{1}{T-h} \sum_{t=1}^{T-h} [\hat{g}_h(\varepsilon_{1t} + \delta) - \hat{g}_h(\varepsilon_{1t})]. \quad (9)$$

This estimator is intuitive: it traces out the nonparametric regression surface \hat{g}_h , evaluates it at the observed and shifted shock values, and averages the difference. It is consistent when \hat{g}_h converges to g_h at a sufficient rate, and it performs well across a range of data generating processes in the simulations of [Gonçalves et al. \(2024a\)](#).

The estimator's performance, however, is entirely determined by the quality of the nonparametric regression \hat{g}_h . Several practical considerations can undermine this quality. First, bandwidth selection for the local linear smoother involves a bias-variance trade-off whose optimal resolution depends on the unknown smoothness of g_h ; data-driven bandwidth selectors can perform poorly when the regression surface has localized features such as sharp nonlinearities. Second, for conditional average responses, the regression estimator requires a multivariate regression of y_{t+h} on $(\varepsilon_{1t}, \Omega_t)$, and the curse of dimensionality rapidly degrades performance as $\dim(\Omega_t)$ increases. Third, the estimator provides no internal diagnostic for misspecification: if \hat{g}_h is a poor approximation to g_h , the resulting bias is transmitted directly to $\widehat{\text{ARF}}_h^{\text{reg}}$ with no mechanism for correction.

3.2 The Reweighting Estimator

The density ratio representation suggests a fundamentally different estimator that sidesteps the conditional mean regression entirely. Estimate the density ratio $r_\delta(e) = f(e - \delta)/f(e)$ and form:

$$\widehat{\text{ARF}}_h^{\text{rw}}(\delta) = \frac{1}{T-h} \sum_{t=1}^{T-h} y_{t+h} \cdot (\hat{r}_\delta(\varepsilon_{1t}) - 1). \quad (10)$$

Rather than asking how the expected outcome varies with the shock level, this estimator reweights the observed outcomes to reflect the counterfactual shock distribution. It requires only the univariate shock density f — a substantially simpler estimation problem than the conditional mean regression, and one whose difficulty does not increase when the conditioning set Ω_t is high-dimensional.

The reweighting estimator's appeal lies in its simplicity and its robustness to misspecification of g_h : since g_h never appears in the estimator, errors in the conditional mean function cannot affect it. However, this robustness comes at a cost. Each outcome y_{t+h} is multiplied by the weight $(\hat{r}_\delta(\varepsilon_{1t}) - 1)$, and these weights can be large when the density ratio takes extreme values — particularly in the tails of the shock distribution, where $f(e)$ is small and $f(e - \delta)/f(e)$ can diverge. The resulting variance inflation means that the reweighting estimator is noisier than the regression estimator even when both are well-specified. As I show formally in Section 5, it has higher asymptotic variance than both the regression estimator and the semiparametric efficiency bound, making it a poor choice as a standalone mean-response estimator. Its value, as we will see, lies elsewhere: as a bias-correction device when paired with regression.

3.3 The Case for Combination

The two estimators have complementary vulnerabilities: the regression estimator is efficient when \hat{g}_h is accurate but offers no protection when it is not, while the reweighting estimator is independent of g_h but pays for this with higher variance. A natural strategy is to use the regression as the primary estimator and add a bias correction that applies the density ratio weights to the regression residuals, so that the correction is negligible when \hat{g}_h is accurate and picks up the slack when it is not. This is the logic of the doubly robust estimator developed in the next section.

4 The Doubly Robust Estimator

This section constructs the doubly robust estimator, establishes its consistency under misspecification of either nuisance component, derives its asymptotic distribution, and provides practical guidance on density ratio estimation.

4.1 Construction of the DR Estimator

The combination of regression and reweighting suggested in Section 3 takes a specific form. Augment the regression estimator with a density-ratio bias correction term:

$$\widehat{\text{ARF}}_h^{DR}(\delta) = \frac{1}{T-h} \sum_{t=1}^{T-h} \left[\hat{g}_h(\varepsilon_{1t} + \delta) - \hat{g}_h(\varepsilon_{1t}) + (\hat{r}_\delta(\varepsilon_{1t}) - 1)(y_{t+h} - \hat{g}_h(\varepsilon_{1t})) \right]. \quad (\text{DR})$$

The first term, $\hat{g}_h(\varepsilon_{1t} + \delta) - \hat{g}_h(\varepsilon_{1t})$, is the regression estimator. The second term, $(\hat{r}_\delta(\varepsilon_{1t}) - 1)(y_{t+h} - \hat{g}_h(\varepsilon_{1t}))$, is the bias correction motivated in Section 3.3: it applies the density ratio weights to the regression residuals. In the treatment effects literature, this construction is known as an augmented inverse-propensity-weighted (AIPW) estimator (Robins et al., 1994).

Notice that the doubly robust (DR) estimator nests the regression estimator: when $\hat{r}_\delta = r_\delta$, the augmentation term $(r_\delta - 1)(y_{t+h} - \hat{g}_h)$ has conditional mean zero given ε_{1t} , and the DR estimator reduces to the regression estimator plus a mean-zero noise term. Conversely, if we set $\hat{g}_h = 0$ (no regression), the estimator reduces to the pure reweighting estimator $T^{-1} \sum_t y_{t+h}(\hat{r}_\delta - 1)$.

4.2 Double Robustness

Proposition 4.1 (Double robustness; informal). *The DR estimator (DR) is consistent for $ARF_h(\delta)$ if either:*

- (a) \hat{g}_h converges in probability to g_h uniformly over the relevant domain, or
- (b) \hat{r}_δ converges in probability to r_δ uniformly,

but not necessarily both, provided the misspecified component remains $O_p(1)$ in L_2 norm. The trimmed estimators proposed in this paper satisfy this boundedness condition by construction. See Appendix A for the formal statement.

Sketch. Write:

$$\widehat{ARF}_h^{DR} - ARF_h = \underbrace{\frac{1}{T-h} \sum_t \psi_t^*}_{O_p(T^{-1/2})} + \underbrace{\frac{1}{T-h} \sum_t (\hat{r}_\delta(\varepsilon_{1t}) - r_\delta(\varepsilon_{1t}))(\hat{g}_h(\varepsilon_{1t}) - g_h(\varepsilon_{1t}))}_{\text{product bias term}} + \text{lower order terms.} \quad (11)$$

The key is the product structure of the bias: it involves the product of the errors in \hat{r}_δ and \hat{g}_h . If either error is zero (i.e., one component is correctly specified), the product vanishes regardless of the other.²

²In a complementary but distinct direction, Montiel Olea et al. (2024) establish a double robustness property of standard linear local projections: the LP estimator's bias under dynamic misspecification of a finite-order VAR is proportional to the *product* of the errors in the outcome and first-stage lag specifications, so that even substantial omitted-lag misspecification does not distort inference. Their result provides a powerful rationale for preferring LP over VAR confidence intervals in the linear setting, but does not extend to the nonparametric impulse responses considered here, where the relevant misspecification is not in the lag structure but in the conditional mean function g_h itself.

Our notion of double robustness is different: the (DR) estimator is consistent when either the

4.3 Asymptotic Normality and Efficiency

The double robustness property concerns consistency alone. For \sqrt{T} -inference, we need both nuisance estimators to converge, and we need the product of their errors to be asymptotically negligible. Specifically, if the nuisance estimators satisfy

$$\|\hat{g}_h - g_h\|_2 \cdot \|\hat{r}_\delta - r_\delta\|_2 = o_p(T^{-1/2}), \quad (12)$$

then the DR estimator is \sqrt{T} -asymptotically normal:

$$\sqrt{T}(\widehat{\text{ARE}}_h^{DR}(\delta) - \text{ARE}_h(\delta)) \xrightarrow{d} N(0, \Sigma_h^*), \quad (13)$$

where Σ_h^* is the long-run variance defined in (LRV) later in the paper. At $h = 0$ the influence function ψ_t^* depends only on the i.i.d. pair (ε_{1t}, y_t) , so $\Sigma_0^* = V_0^*$ and the estimator attains the semiparametric efficiency bound exactly. For $h > 0$, the serial correlation in ψ_t^* induced by the overlapping structure of y_{t+h} inflates Σ_h^* above V_h^* ; this is a feature of the local projection design shared by all LP estimators, not specific to the DR correction. Inference requires a heteroskedasticity and autocorrelation consistent (HAC) variance estimator with bandwidth reflecting the $(h + p)$ -dependence of the influence-function process; I defer the details of the inferential procedure to future work. This is the rate double robustness property: neither component needs to converge at $T^{-1/4}$ individually — any allocation of rates whose product beats $T^{-1/2}$ suffices.³

For the nuisance estimators I propose (local linear kernel regression for \hat{g}_h and kernel density plug-in for \hat{r}_δ) the product rate condition is verified under primitive conditions in Appendix A.5 (Proposition A.3). Both are classical nonparametric estimators whose individual $L_2(P)$ rates are approximately $T^{-2/5}$ (up to logarithmic corrections from the unbounded support of the shock distribution), yielding a product rate of approximately $T^{-4/5}$ — well above the $T^{-1/2}$ threshold. The dependence between the nuisance estimates and the observations at which they are evaluated is controlled by standard stochastic equicontinuity arguments for kernel-based function classes (van der Vaart,

nonparametric regression \hat{g}_h or the density ratio \hat{r}_δ is well specified, and achieves the semiparametric efficiency bound when the product of their estimation errors is $o_p(T^{-1/2})$. Both results trace their logic to the product-of-errors structure emphasized by Chernozhukov et al. (2018), but they address different sources of fragility in different estimation frameworks.

³In Appendix A.5 (Proposition A.3), I verify the product rate condition under primitive smoothness and moment assumptions for the trimmed kernel-based estimators proposed in this paper. Each component converges at approximately $T^{-2/5}$ (up to logarithmic factors), so the product is approximately $T^{-4/5}$ — well above the $T^{-1/2}$ threshold. When the shock density is estimated nonparametrically, a mild constraint on the ratio $|\delta|/\sigma_1$ is needed to control tail instability of the plug-in density ratio estimator; see the discussion preceding Section 4.4 and Remark A.8.

1998, Chapter 19).⁴

Scope of the shift parameter. When the density ratio is estimated nonparametrically, the formal rate condition in Proposition A.3 requires $|\delta| < C_0 \sigma_1$ for a constant C_0 that depends on the tail behavior of the shock density f . The source of this restriction is intuitive: for large shifts the counterfactual density $f(\cdot - \delta)$ has poor overlap with the observed density $f(\cdot)$, so the density ratio $r_\delta(e) = f(e - \delta)/f(e)$ takes extreme values in the tails, inflating the variance of the plug-in estimator. Regularized density ratio estimators produce bounded weights by construction, avoiding the tail instability that drives the constraint. Furthermore, within the DR framework even a biased density ratio estimate contributes small product bias when \hat{g}_h is well specified, because the bias of the DR estimator is proportional to $\|\hat{r}_\delta - r_\delta\|_2 \cdot \|\hat{g}_h - g_h\|_2$; a fast-converging regression absorbs density ratio error. Formal details are in Remark A.8.

4.4 Density Ratio Estimation

The (DR) estimator requires an estimate \hat{r}_δ of the density ratio $r_\delta(e) = f(e - \delta)/f(e)$. This is a univariate estimation problem regardless of the dimension of the structural model. I discuss three approaches, in order of increasing sophistication.

Plug-in kernel density estimation. Estimate f from $\{\varepsilon_{1t}\}$ using a standard kernel density estimator \hat{f} , then form $\hat{r}_\delta(e) = \hat{f}(e - \delta)/\hat{f}(e)$.

This approach is simple and inherits well-understood asymptotic theory, including extensions to weakly dependent data (though ε_{1t} is i.i.d. by assumption, so the standard theory applies directly). Its main drawback is instability when $\hat{f}(e)$ is small (in the tails of the shock distribution, where the ratio can be noisy). In practice, a floor on $\hat{f}(e)$ (e.g., trimming observations where $\hat{f}(\varepsilon_{1t}) < c_T$ for a threshold $c_T \rightarrow 0$) provides regularization.

⁴If the researcher wishes to use more flexible estimators for \hat{g}_h or \hat{r}_δ (e.g. sieve estimators, penalized regression, or machine learning methods that do not satisfy Donsker-class conditions) the product rate condition can still be verified using sample splitting, following Chernozhukov et al. (2018). In time series settings, this requires sequential (non-overlapping block) splits rather than random cross-validation, with buffer zones between blocks to account for the serial dependence in y_{t+h} . The cost is a reduction in effective sample size, which can be substantial in the moderate samples typical of applied macroeconomics. For the kernel-based implementation that is the focus of this paper, splitting is unnecessary.

A lighter-weight alternative to sample splitting is the use of leave-one-out residuals: replacing $\hat{g}_h(\varepsilon_{1t})$ with the leave-one-out estimate $\hat{g}_{h,-t}(\varepsilon_{1t})$ when forming the augmentation term in (DR). This eliminates the own-observation bias without reducing the effective sample size and is the approach adopted in our implementation (see Section 4.5).

Direct density ratio estimation. Modern methods estimate the ratio r_δ directly without estimating f at all. The key observation is that $\{\varepsilon_{1t} - \delta\}_{t=1}^T$ constitutes a sample from the density $f(\cdot - \delta)$, while $\{\varepsilon_{1t}\}$ is a sample from $f(\cdot)$. The ratio $f(\cdot - \delta)/f(\cdot)$ can then be estimated by:

- **uLSIF** (unconstrained Least-Squares Importance Fitting): minimizes integrated squared error of the ratio approximation with Tikhonov regularization, producing smooth, bounded ratio estimates. The regularization parameter can be selected by cross-validation.
- **KLIEP** (Kullback–Leibler Importance Estimation Procedure): minimizes KL divergence subject to normalization constraints.

These methods, surveyed in [Sugiyama et al. \(2012\)](#), were developed for i.i.d. data. In our setting, the relevant samples (ε_{1t} and $\varepsilon_{1t} - \delta$) are i.i.d., so the standard theory applies. However, the formal integration of these estimators’ convergence properties into our semiparametric framework (specifically, verifying the rate conditions in [Section 4.3](#)) requires additional work. I defer this to [Appendix A](#) and focus in the simulations on the plug-in and parametric approaches, for which the asymptotic theory is more complete.

Practical recommendation. For the DR estimator, the choice of density ratio method is less critical than it might appear, because errors in \hat{r}_δ are absorbed by the double robustness property as long as \hat{g}_h is reasonably well-specified. The density ratio is a bias-correction device, not the primary carrier of the estimator. I recommend:

- **Plug-in kernel density** as the default general-purpose option. It is simple, well-understood theoretically, and the bandwidth can be selected by standard methods (e.g., Silverman’s rule of thumb or cross-validation). Trimming observations where $\hat{f}(\varepsilon_{1t})$ falls below a small threshold prevents division-by-near-zero instability.
- **uLSIF** when the researcher wants a regularized, bounded ratio estimate — particularly useful when extreme density ratio values would distort the bias correction.

Both approaches produce density ratio estimates that are bounded in finite samples, in contrast to the exact density ratio $r_\delta(e) = f(e - \delta)/f(e)$, which is generically unbounded for light-tailed distributions.

4.5 Algorithm

Algorithm 1 summarizes the DR estimator for a given shock size δ and maximum horizon H .

Algorithm 1 Doubly Robust Estimator of $\text{ARF}_h(\delta)$

Require: $\{y_t, \varepsilon_{1t} : t = 1, \dots, T\}$, shock size δ , max horizon H

1: **Step 1.** Using the full sample, estimate:

- $\hat{g}_h(e)$: local linear regression of y_{t+h} on ε_{1t} (Gaussian kernel, Fan–Gijbels ROT bandwidth)
- $\hat{r}_\delta(e)$: density ratio (see Section 4.4 for options)

2: **Step 2.** For each $t = 1, \dots, T - h$, compute:

$$\hat{\psi}_t = \hat{g}_h(\varepsilon_{1t} + \delta) - \hat{g}_h(\varepsilon_{1t}) + (\hat{r}_\delta(\varepsilon_{1t}) - 1) \cdot (y_{t+h} - \hat{g}_h(\varepsilon_{1t}))$$

3: **Step 3.** $\widehat{\text{ARF}}_h^{DR}(\delta) = \frac{1}{T-h} \sum_t \hat{\psi}_t$

Ensure: $\{\widehat{\text{ARF}}_h^{DR}(\delta) : h = 0, \dots, H\}$

In practice, the full-sample residual $y_{t+h} - \hat{g}_h(\varepsilon_{1t})$ in Step 2 can be replaced with the leave-one-out residual $y_{t+h} - \hat{g}_{h,-t}(\varepsilon_{1t})$, where $\hat{g}_{h,-t}$ denotes the local linear estimator computed from all observations except the t -th. For local linear regression with a kernel weight function, the leave-one-out fitted value is available in closed form at negligible additional cost. This substitution prevents the residual from understating the true prediction error at observation t due to overfitting, a finite-sample concern that is most pronounced when the bandwidth is small relative to the sample size. Asymptotically, $\hat{g}_{h,-t}$ and \hat{g}_h are equivalent, so the large-sample results in Sections 4.2–4.3 are unaffected.

The nuisance functions \hat{g}_h and \hat{r}_δ are estimated on the same sample used to compute the (DR) estimator. This is valid for the kernel-based estimators I propose because they satisfy the stochastic equicontinuity conditions needed to control the dependence between the nuisance estimates and the evaluation points (see Appendix A for the formal statement).

Asymptotic normality (Proposition A.2) implies that valid confidence intervals can be constructed from a HAC estimator of the long-run variance of $\hat{\psi}_t$. The influence-function process is at most $(h + p)$ -dependent, where p is the lag order of the structural model (Appendix A.4), and the HAC bandwidth must reflect this dependence structure. I defer a full inferential implementation to future work and instead focus on point estimation in

this paper.

5 Efficiency Theory

The previous section established that the DR estimator is consistent when either nuisance component is well-specified and asymptotically normal when both converge at appropriate rates. This section asks a deeper question: is the particular combination of regression and reweighting in the (DR) estimator optimal? I show that it is, by characterizing the semiparametric efficiency bound and showing that the (DR) estimator attains it. The variance decomposition that emerges also clarifies the relationship between the (DR) estimator, the regression estimator, and the pure reweighting estimator from Section 3.

5.1 Efficient Influence Function

Consider the semiparametric model in which the conditional mean g_h and the shock density f are unrestricted (subject to regularity conditions). The parameter of interest $ARF_h(\delta) = E[g_h(\varepsilon_{1t} + \delta) - g_h(\varepsilon_{1t})]$ is a functional of the joint distribution of $(\varepsilon_{1t}, y_{t+h})$.

Proposition 5.1 (Efficient influence function; informal). *Under regularity conditions (stated in Appendix A), the efficient influence function for $ARF_h(\delta)$ is:*

$$\psi_t^* = [g_h(\varepsilon_{1t} + \delta) - g_h(\varepsilon_{1t})] + [r_\delta(\varepsilon_{1t}) - 1] [y_{t+h} - g_h(\varepsilon_{1t})] - ARF_h(\delta). \quad (\text{EIF})$$

The semiparametric efficiency bound is $V_h^* = \text{Var}(\psi_t^*)$.

The efficient influence function has two components that correspond directly to the two terms in the (DR) estimator from Section 4.1. The first, $g_h(\varepsilon_{1t} + \delta) - g_h(\varepsilon_{1t})$, is the regression estimator's influence function contribution — it captures the variation due to random sampling of shocks through the conditional mean. The second, $(r_\delta - 1)(y_{t+h} - g_h(\varepsilon_{1t}))$, is the augmentation term that corrects for the estimation of g_h — the same density-ratio bias correction that appears in the (DR) estimator. If g_h were known, this term would have mean zero but nonzero variance; it reflects the efficiency cost of not knowing the conditional mean. The (DR) estimator is, in effect, the sample analog of $\psi_t^* + ARF_h(\delta)$, with population quantities replaced by estimates.

Proof sketch. The functional $ARF_h(\delta)$ depends on the time series distribution only through the bivariate marginal of $(\varepsilon_{1t}, y_{t+h})$. I derive the efficient influence function

in the *bivariate marginal model*: the semiparametric model that treats the joint density $p(e, y) = f(e) c(y | e)$ as the observed-data law, leaving both f and c unrestricted. In this model, the tangent space is the full $L_2^0(P)$ by the standard nonparametric factorization argument, exactly as in the cross-sectional continuous treatment setting of [Kennedy et al. \(2017\)](#). The pathwise derivative of $\theta = \text{ARF}_h(\delta)$ along a submodel with score $s = s_1 + s_2$ (marginal plus conditional components) can be written as $E[\psi_t^* \cdot s]$, where ψ_t^* is the stated influence function. The key steps use the change-of-variables identity $E[h(\varepsilon_{1t} + \delta)] = E[h(\varepsilon_{1t}) r_\delta(\varepsilon_{1t})]$ and the conditional mean-zero property $E[s_2 | \varepsilon_{1t}] = 0$ to separate the contributions of s_1 and s_2 . Because the structural model generates bivariate marginals that are a subset of those permitted by the marginal model, the efficiency bound V_h^* derived here is an upper bound on the efficiency bound of the structural model; an estimator that exploits structural restrictions could in principle achieve lower variance, but the DR estimator does not require such restrictions for its validity. The complete derivation is in [Appendix A.2](#).

5.2 Variance Decomposition

Because $E[y_{t+h} - g_h(\varepsilon_{1t}) | \varepsilon_{1t}] = 0$, the cross-covariance between the two components of ψ_t^* vanishes, and the efficiency bound decomposes as:

$$V_h^* = \underbrace{\text{Var}[g_h(\varepsilon_{1t} + \delta) - g_h(\varepsilon_{1t})]}_{V_h^{reg}} + \underbrace{E[(r_\delta(\varepsilon_{1t}) - 1)^2 \cdot \sigma^2(\varepsilon_{1t})]}_{V_h^{aug}} \quad (\text{VD})$$

where $\sigma^2(e) = \text{Var}(y_{t+h} | \varepsilon_{1t} = e)$ is the conditional variance of the outcome given the shock.

This decomposition has immediate implications:

1. **The regression estimator's asymptotic variance is V_h^{reg} .** When \hat{g}_h converges fast enough, the regression estimator has influence function $g_h(\varepsilon_{1t} + \delta) - g_h(\varepsilon_{1t}) - \text{ARF}$, yielding asymptotic variance V_h^{reg} . This is strictly less than the efficiency bound V_h^* whenever $V_h^{aug} > 0$ — whenever y_{t+h} has any residual variation not explained by ε_{1t} , which is generically the case.
2. **A pure reweighting estimator is generically less efficient.** An estimator based solely on the representation [\(R\)](#) has influence function $y_{t+h}(r_\delta - 1) - \text{ARF}$. Its asymptotic variance is $\text{Var}[g_h(\varepsilon_{1t})(r_\delta - 1)] + E[(r_\delta - 1)^2 \sigma^2(\varepsilon_{1t})]$, where the second term coincides with V_h^{aug} but the first term replaces $V_h^{reg} = \text{Var}[g_h(\varepsilon_{1t} + \delta) - g_h(\varepsilon_{1t})]$ with $\text{Var}[g_h(\varepsilon_{1t})(r_\delta(\varepsilon_{1t}) - 1)]$. For important special cases the reweighting variance

strictly exceeds V_h^* : when g_h is linear, $g_h(\varepsilon_{1t} + \delta) - g_h(\varepsilon_{1t})$ is constant and $V_h^{reg} = 0$, while $\text{Var}[g_h(\varepsilon_{1t})(r_\delta - 1)] > 0$, so the entire first term is pure excess variance. More broadly, the reweighting estimator magnifies outcome variation through the weights $(r_\delta - 1)$, which take extreme values in the tails of the shock distribution. This confirms that pure reweighting is not a competitive standalone estimator for the mean response; its role is as one arm of the doubly robust combination.

3. **The efficiency bound exceeds the regression estimator variance.** This does not mean the regression estimator is superefficient. The two estimators operate in different semiparametric models: the GHKP regression estimator assumes that g_h is correctly specified, placing it in a smaller model whose efficiency bound is V_h^{reg} . The bound V_h^* applies to the larger model in which neither g_h nor f is assumed known — the model in which the DR estimator operates. Neither estimator violates the efficiency bound of its own model.

The gap $V_h^{aug} = V_h^* - V_h^{reg}$ is the variance premium paid for not assuming correct specification of g_h . When the regression is well specified, the GHKP estimator achieves the lower variance V_h^{reg} and outperforms the DR estimator; when it is not, the regression estimator's actual finite-sample performance can be substantially worse than V_h^{reg} suggests. *The DR estimator hedges against this possibility at the cost of higher variance when the regression is accurate.*

Note. The decomposition (VD) describes the contemporaneous variance $V_h^* = \text{Var}(\psi_t^*)$. For $h > 0$ the influence function ψ_t^* is serially correlated at minimum of order h , because y_{t+h} depends on shocks $\varepsilon_{1,t+1}, \dots, \varepsilon_{1,t+h}$ that overlap across adjacent observations. The asymptotic variance of $\sqrt{T} \widehat{\text{ARF}}_h^{DR}(\delta)$ is therefore the long-run variance

$$\Sigma_h^* = \sum_{j=-\infty}^{\infty} \text{Cov}(\psi_t^*, \psi_{t-j}^*), \quad (\text{LRV})$$

which equals V_h^* at $h = 0$ (where ψ_t^* is a function of the i.i.d. pair (ε_{1t}, y_t)) but generally exceeds it for $h > 0$. The qualitative implications of the decomposition are unaffected: the regression estimator's long-run variance replaces V_h^{reg} with $\sum_j \text{Cov}(g_h(\varepsilon_{1t} + \delta) - g_h(\varepsilon_{1t}), g_h(\varepsilon_{1,t-j} + \delta) - g_h(\varepsilon_{1,t-j}))$, and the same HAC correction applies to all estimators equally. Inference based on (VD) directly (using the simple sample variance of $\hat{\psi}_t$) would understate estimation uncertainty at horizons $h > 0$.

5.3 Implications for Estimation Strategy

The variance decomposition quantifies the trade-off facing applied researchers. When g_h can be estimated reliably, the regression estimator achieves V_h^{reg} and the gap V_h^{aug} is the premium paid for robustness that may not be needed. When g_h is difficult to estimate, the regression estimator's actual MSE can substantially exceed V_h^{reg} , and the product-of-errors property means that even a rough density ratio estimate can substantially reduce bias. Pure reweighting is not the answer in either regime; its role is as one arm of the (DR) estimator rather than a standalone mean-response estimator.

6 Conditional Average Responses

The density ratio r_δ is a univariate object that, once estimated, provides a dimensionality reduction for estimating conditional average responses.

DR estimator for CARs. When $\varepsilon_{1t} \perp\!\!\!\perp \Omega_t$, the CAR admits both a regression representation $CAR_h(\delta, \omega) = E[g_h(\varepsilon_{1t} + \delta, \omega) - g_h(\varepsilon_{1t}, \omega)]$ and a reweighting representation (R_ω). Combining the two yields a DR estimator of the CAR whose influence function, evaluated at observation t with $\Omega_t = \omega$, takes the form

$$\hat{\psi}_t^\omega = \hat{g}_h(\varepsilon_{1t} + \delta, \omega) - \hat{g}_h(\varepsilon_{1t}, \omega) + (\hat{r}_\delta(\varepsilon_{1t}) - 1)(y_{t+h} - \hat{g}_h(\varepsilon_{1t}, \omega)). \quad (14)$$

This construction parallels the ARF estimator (DR), with $\hat{g}_h(e)$ replaced by the joint regression $\hat{g}_h(e, \omega)$. The product-of-errors logic of Proposition 4.1 carries over: the bias of the DR CAR estimator is proportional to $\|\hat{r}_\delta - r_\delta\| \cdot \|\hat{g}_h(\cdot, \omega) - g_h(\cdot, \omega)\|$, so the density ratio correction mitigates slow convergence of the joint regression.

The inferential status of (14) depends on the nature of Ω_t .

- **Discrete Ω_t .** Conditioning on $\Omega_t = \omega$ reduces to a subsample restriction. Within each cell, the problem is an unconditional ARF estimation problem, and (14) is the efficient influence function by the same argument as in Section 5.1 and Appendix A.2. Standard $\sqrt{T_\omega}$ inference applies, where T_ω is the cell size.
- **Continuous Ω_t .** Point evaluation at $\Omega_t = \omega$ is a nonparametric localization problem: the CAR is estimated by a kernel or local polynomial regression of $\hat{\psi}_t^\omega$ on Ω_t , which converges at nonparametric rates rather than \sqrt{T} . The expression (14) should be understood as the observation-level DR pseudo-outcome entering this second-stage regression, not as an efficient influence function for the pointwise

conditional parameter. A formal efficiency theory for the continuous- ω case would require a localized or smoothed target parameter; this is left to future work.

Decomposed estimation. The independence $\varepsilon_{1t} \perp\!\!\!\perp \Omega_t$ permits a decomposition of the estimation problem. Define the reweighted outcome $\tilde{y}_{t+h} = y_{t+h} \cdot (r_\delta(\varepsilon_{1t}) - 1)$. Then:

$$\text{CAR}_h(\delta, \omega) = E[\tilde{y}_{t+h} \mid \Omega_t = \omega]. \quad (15)$$

This means the CAR can be estimated by regressing the reweighted outcome on Ω_t alone. The dimension of this problem equals $\dim(\Omega_t)$ rather than $\dim(\Omega_t) + 1$. Specifically:

- **If Ω_t is discrete:** the regression on Ω_t collapses to a subsample average:

$$\widehat{\text{CAR}}_h(\delta, s) = \frac{\sum_{t: S_{t-1}=s} \tilde{y}_{t+h}}{\sum_{t: S_{t-1}=s} 1}, \quad s \in \{0, 1\}. \quad (16)$$

- **If Ω_t is continuous:** use a univariate kernel regression of \tilde{y}_{t+h} on Ω_t , requiring a single bandwidth b_ω . GHKP's approach requires a bivariate kernel regression of y_{t+h} on (ε_{1t}, r_t) , requiring two bandwidths (b_ε, b_r) .

The dimensionality trade-off. The decomposition reduces the nonparametric regression dimension from $1 + \dim(\Omega_t)$ (regression) to $\dim(\Omega_t)$ (reweighting). However, this comes at a cost: the reweighted outcome \tilde{y}_{t+h} is noisier than y_{t+h} because multiplication by $(r_\delta - 1)$ amplifies the variance, particularly for large shocks. The conditional variance of \tilde{y}_{t+h} given Ω_t is:

$$\text{Var}(\tilde{y}_{t+h} \mid \Omega_t = \omega) = E[(r_\delta(\varepsilon_{1t}) - 1)^2 \cdot y_{t+h}^2 \mid \Omega_t = \omega] - \text{CAR}_h(\delta, \omega)^2 \quad (17)$$

which can be much larger than $\text{Var}(y_{t+h} \mid \varepsilon_{1t}, \Omega_t = \omega)$. The net effect on MSE is therefore ambiguous: the dimensionality reduction improves the bias (slower curse of dimensionality), while the variance inflation worsens the variance. The gains are likely more substantial with higher-dimensional conditioning sets. I assess this trade-off via Monte Carlo simulation in Section 7.

A DR version of the CAR estimator hedges against both sources of error. When Ω_t is continuous, the second-stage regression converges at nonparametric rates in ω (e.g. the local linear rate $O(T^{-2/(2+\dim(\Omega_t))})$), not at the \sqrt{T} rate that governs the unconditional ARF estimator. The semiparametric efficiency bound and influence function characterization of Section 5 apply to the unconditional ARF only. For

discrete Ω_t , the conditioning reduces to a subsample average, \sqrt{T} inference applies within each cell, and the efficiency analysis carries over directly.

Algorithm 2 Reweighting Estimator of $\text{CAR}_h(\delta, \omega)$

Require: $\{y_t, \varepsilon_{1t}, \Omega_t : t = 1, \dots, T\}$, shock size δ , horizon h , evaluation point ω . Assumes $\varepsilon_{1t} \perp\!\!\!\perp \Omega_t$.

- 1: **Step 1.** Estimate \hat{r}_δ as in Section 4.4.
- 2: **Step 2.** Compute reweighted outcomes: $\tilde{y}_{t+h} = y_{t+h} \cdot (\hat{r}_\delta(\varepsilon_{1t}) - 1)$.
- 3: **Step 3.** Estimate $\text{CAR}_h(\delta, \omega)$ by nonparametric regression of \tilde{y}_{t+h} on Ω_t :
 - Ω_t discrete: subsample average of \tilde{y}_{t+h}
 - Ω_t continuous: univariate local linear regression

Ensure: $\widehat{\text{CAR}}_h(\delta, \omega)$

Algorithm for CAR estimation. For the DR version, replace Steps 2–3 with the full AIPW construction using $\hat{g}_h(e, \omega)$ and $\hat{r}_\delta(e)$ jointly, following the structure of Algorithm 1 adapted to include Ω_t .

7 Monte Carlo Simulations

This section uses Monte Carlo simulations to evaluate the doubly robust estimator. The first set of simulations demonstrates the formal double-robustness property: when the regression arm is structurally misspecified, the density-ratio correction eliminates the asymptotic bias that a standalone regression estimator cannot escape. The second set asks what this insurance costs when it is unnecessary, measuring the variance premium a researcher pays for robustness when the regression is already well matched to the DGP. All figures are collected in Appendix B.

7.1 Design

The data-generating process is adapted from [Gonçalves et al. \(2024a\)](#). The shock is $x_t = \varepsilon_{1t}$, where ε_{1t} and ε_{2t} are independent standard normal draws. The outcome follows

$$y_t = 0.5 y_{t-1} + 0.5 x_t + 0.3 x_{t-1} - 0.4 f(x_t) - 0.3 f(x_{t-1}) + \varepsilon_{2t}, \quad (18)$$

and the estimand is the unconditional average response function (4). I consider two functional forms for f that create complementary difficulties for nonparametric

estimation:

- *ReLU*: $f(x) = \max(x, 0)$. The kink at the origin is poorly approximated by smooth polynomial bases, so the power series estimator exhibits finite-sample bias that is slow to vanish. The local linear estimator, which imposes no global smoothness, adapts well.
- *Cubic*: $f(x) = x^3$. The high curvature induces substantial bandwidth-driven bias in the local linear estimator. The power series estimator, which nests the true specification, performs well.

Each replication discards $B = 1,000$ burn-in observations before recording a sample of length T .

The simulation varies $T \in \{250, 500, 1,000, 2,000\}$ and $\delta \in \{1, 2\}$, with $R = 5,000$ replications per design cell. The population truth is approximated by averaging over 50,000 independent paths of post-burn-in length 2,000. Each figure is a 3×4 panel whose columns correspond to the four sample sizes and whose rows display, respectively, the impulse response function, mean bias, and root mean squared error. The impulse response panels report the median and interquartile range across replications. A separate figure is produced for each combination of functional form and shock size.

Every figure compares two estimators. The first is a standalone regression estimator of g_h , whose specification varies across designs. The second is the DR estimator (DR), which pairs the same regression \hat{g}_h with the parametric Gaussian density ratio

$$\hat{r}_\delta(e) = \exp\left(\frac{\delta e}{\hat{\sigma}_1^2} - \frac{\delta^2}{2\hat{\sigma}_1^2}\right), \quad (19)$$

where $\hat{\sigma}_1^2$ is the sample variance of $\{\varepsilon_{1t}\}$. Because the shock is Gaussian by construction, this density ratio is correctly specified. Holding it fixed throughout both sets of simulations isolates the contribution of the bias-correction term from the separate question of how to estimate the density ratio nonparametrically.

7.2 Double robustness under structural misspecification

The first set of simulations illustrates the double-robustness property of Proposition 4.1 in finite samples. The regression arm is a linear local projection that omits the nonlinear term entirely,

$$y_{t+h} = \alpha_h + \psi_h \varepsilon_{1t} + \omega_{t+h},$$

estimated by OLS. This is the specification a researcher would adopt if unaware of the nonlinearity. It produces a bias that does not vanish with T : the linear projection ψ_h captures only the linear component of the ARF, missing the contribution of f . The DR estimator pairs this misspecified regression with the parametric density ratio (19).

Figures 1–4 report the results for both functional forms. The standalone linear LP (dashed) exhibits bias that persists across all sample sizes—the hallmark of structural misspecification rather than finite-sample imprecision. The DR estimator (solid) eliminates the vast majority of this bias. Because the density ratio is correctly specified, the augmentation term $(r_\delta - 1)(y_{t+h} - \hat{g}_h)$ reweights the regression residuals to recover the nonlinear signal that the linear projection discards. The bias reduction is most pronounced for $\delta = 2$, where the nonlinear component of the ARF is largest.

7.3 Cost of robustness when the regression is well specified

The second set of simulations measures the price a researcher pays for robustness that turns out to be unnecessary. The regression arm is now the well-matched nonparametric method: local linear for the ReLU DGP and power series for the cubic DGP. Tuning follows Gonçalves et al. (2024a) in both cases.

Figures 5–8 confirm that the standalone estimator (dashed) performs well: bias is small and shrinks with T at the expected nonparametric rate. The DR estimator (solid) matches this bias closely. When the regression is accurate, the augmentation term has approximate mean zero and contributes primarily variance. The gap between the two RMSE curves is the finite-sample counterpart of the augmentation variance V_h^{aug} in the efficiency-bound decomposition (VD).

Between these two polar cases lies a middle ground of correct but slow specification, where the function class contains the true g_h but the estimator converges slowly due to bandwidth choice or high dimensionality. The product-of-errors property (Proposition 4.1) implies that the DR correction provides insurance in this intermediate regime as well: even when both nuisance components contain estimation error, the bias of the DR estimator is proportional to their product, so a well-specified density ratio substantially attenuates the finite-sample bias of a slow-converging regression.

Taken together, the two sets of simulations support the use of the DR estimator as a practical default: it provides large bias reductions when the regression is misspecified and imposes only a small variance premium when it is not.

7.4 Tuning parameters

I collect the tuning choices used across both sets of simulations. The linear LP is OLS of y_{t+h} on a constant and ε_{1t} , with no tuning parameters. The local linear estimator (ReLU DGP) uses a Gaussian kernel with a [Fan and Gijbels \(1996\)](#) rule-of-thumb bandwidth based on a preliminary polynomial of order 2, following [Gonçalves et al. \(2024a\)](#). The power series estimator (cubic DGP) uses a polynomial order equal to $0.5 T^{1/3}$ rounded to the nearest integer, following [Gonçalves et al. \(2024a\)](#). The density ratio is the parametric Gaussian specification (19) throughout. All DR estimates use leave-one-out residuals as described in Section 4.5.

8 Conclusion

Nonparametric local projections allow researchers to trace out impulse response functions without restricting how shocks propagate through the economy, but their reliability hinges on the quality of a single nonparametric regression. This paper develops a doubly robust estimator that hedges against this dependence by augmenting the regression-based approach of [Gonçalves et al. \(2024a\)](#) with a bias correction based on the density ratio of the structural shock.

The estimator has three properties that together constitute the paper’s main contribution. First, it is consistent when either the conditional mean regression or the density ratio is well specified, which provides a safeguard that is absent from the existing nonparametric local projections toolkit. Second, it attains the semiparametric efficiency bound at the impact horizon; at longer horizons the asymptotic variance inherits the serial-correlation cost common to all local projection methods. The required rate condition is verified for standard kernel estimators under primitive smoothness assumptions. Third, for conditional average responses, the density ratio enables a decomposition that reduces the nonparametric regression dimension by one. Because the density ratio depends only on the marginal shock density, this dimensionality reduction applies regardless of the dimension of the conditioning set, offering a concrete path to mitigating the curse of dimensionality in state-dependent impulse response estimation.

The efficiency theory complements these estimation results. The semiparametric variance bound decomposes into the asymptotic variance of the regression estimator plus an augmentation term that measures the irreducible cost of not knowing the conditional mean function ex ante. This decomposition clarifies the trade-off facing applied researchers: the doubly robust estimator pays a variance premium quantified

exactly by the augmentation term, in exchange for robustness to misspecification of the regression surface. When the conditional mean is estimated accurately, the regression estimator of [Gonçalves et al. \(2024a\)](#) achieves lower variance and remains the natural choice; the doubly robust estimator is most valuable precisely in the settings where nonparametric regression is most difficult, such as in the presence of high-dimensional conditioning, complex nonlinearities, or limited sample sizes.

References

- Andrews, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, 59(3):817–858.
- Angrist, J. D., Jordà, O., and Kuersteiner, G. M. (2018). Semiparametric estimates of monetary policy effects: String theory revisited. *Journal of Business & Economic Statistics*, 36(3):371–387.
- Angrist, J. D. and Kuersteiner, G. M. (2011). Causal effects of monetary shocks: Semiparametric conditional independence tests with a multinomial propensity score. *Review of Economics and Statistics*, 93:725–747.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*, 21:C1–C68.
- Davidson, J. (1994). *Stochastic Limit Theory*. Oxford University Press.
- Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman & Hall.
- Gonçalves, S., Herrera, A. M., Kilian, L., and Pesavento, E. (2021). Impulse response analysis for structural dynamic models with nonlinear regressors. *Journal of Econometrics*, 225:107–130.
- Gonçalves, S., Herrera, A. M., Kilian, L., and Pesavento, E. (2024a). Nonparametric local projections. Federal Reserve Bank of Dallas Working Paper 2414.
- Gonçalves, S., Herrera, A. M., Kilian, L., and Pesavento, E. (2024b). State-dependent local projections. *Journal of Econometrics*, 244(2):105702.
- Hirano, K. and Imbens, G. W. (2004). The propensity score with continuous treatments. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, pages 73–84. Wiley.
- Information Systems and Wake Forest University (2021). WFU High Performance Computing Facility.
- Kennedy, E. H., Ma, Z., McHugh, M. D., and Small, D. S. (2017). Nonparametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society: Series B*, 79(4):1229–1245.
- Koop, G., Pesaran, M. H., and Potter, S. M. (1996). Impulse response analysis in nonlinear multivariate models. *Journal of Econometrics*, 74:119–147.
- Montiel Olea, J. L., Plagborg-Møller, M., Qian, E., and Wolf, C. K. (2024). Double robustness of local projections and some unpleasant VARithmetic. Working Paper 32495, National Bureau of Economic Research.

- Newey, W. K. (1994). Series estimation of regression functionals. *Econometric Theory*, 10:1–28.
- Newey, W. K. and West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708.
- Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89:846–866.
- Sugiyama, M., Suzuki, T., and Kanamori, T. (2012). *Density Ratio Estimation in Machine Learning*. Cambridge University Press.
- van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

A Assumptions and Proofs

This appendix collects the regularity conditions, formal statements, and proofs of the main results.

A.1 Regularity Conditions

Assumption A.1. $\varepsilon_t = (\varepsilon_{1t}, \dots, \varepsilon_{nt})'$ is i.i.d. across t with mean zero, and the components $\varepsilon_{1t}, \dots, \varepsilon_{nt}$ are mutually independent (with variances σ_i^2).

Assumption A.2. The density f of ε_{1t} is bounded, bounded away from zero on compact sets, and satisfies $E[r_\delta(\varepsilon_{1t})^2] < \infty$.

Assumption A.3. $E[y_{t+h}^2] < \infty$ and $E[y_{t+h}^2 \cdot r_\delta(\varepsilon_{1t})^2] < \infty$.

Assumption A.4. The estimators \hat{g}_h and \hat{r}_δ satisfy:

- (a) $\|\hat{g}_h - g_h\|_2 = o_p(1)$ or $\|\hat{r}_\delta - r_\delta\|_2 = o_p(1)$ (consistency of at least one component), and the other component satisfies $\|\hat{g}_h - g_h\|_2 = O_p(1)$ or $\|\hat{r}_\delta - r_\delta\|_2 = O_p(1)$ respectively (boundedness of the misspecified component). The trimmed estimators defined in Section A.5 satisfy the boundedness condition by construction;
- (b) For efficiency: $\|\hat{g}_h - g_h\|_2 \cdot \|\hat{r}_\delta - r_\delta\|_2 = o_p(T^{-1/2})$.

The following primitive conditions are sufficient for Assumption A.4(b) when using the trimmed kernel-based nuisance estimators defined in Section A.5.

Assumption A.5 (Smoothness and tail decay). (a) The density f satisfies $f \in C^2(\mathbb{R})$ with $f, f', f'' \in L_1(\mathbb{R}) \cap L_2(\mathbb{R})$ and $\sup_e |e^k f^{(j)}(e)| < \infty$ for $j, k \in \{0, 1, 2\}$. (b) The density f has sub-Gaussian tails: there exist constants $K, \lambda > 0$ such that $f(e) \leq K \exp(-\lambda e^2)$ for all $e \in \mathbb{R}$. (c) The conditional mean $g_h \in C^2(\mathbb{R})$ with $\int (g_h''(e))^2 f(e) de < \infty$.

Assumption A.6 (Moments). $E[y_{t+h}^4] < \infty$, $E[g_h(\varepsilon_{1t})^4] < \infty$, and $E[r_\delta(\varepsilon_{1t})^4] < \infty$.

Assumption A.7 (Conditional variance integrability). The conditional variance $\sigma^2(e) \equiv \text{Var}(y_{t+h} \mid \varepsilon_{1t} = e)$ is bounded: $\sup_e \sigma^2(e) < \infty$.

Assumption A.8 (Kernel and bandwidth). The kernel K is a bounded, symmetric, second-order kernel with $\int K = 1$, $\mu_2 \equiv \int u^2 K(u) du \neq 0$, $R(K) \equiv \int K^2(u) du < \infty$, and $K(u) > 0$ for all u (e.g., the Gaussian kernel). The bandwidths satisfy $b_g \asymp T^{-1/5}(\log T)^{1/10}$ and $b_f \asymp T^{-1/5}(\log T)^{1/10}$.

Remark A.1 (On Assumption A.2). The condition $E[r_\delta(\varepsilon_{1t})^2] < \infty$ holds for any distribution with sub-exponential or lighter tails, but excludes heavy-tailed distributions where the density ratio has infinite second moment. For distributions with polynomial tails (e.g., t -distributions), the condition requires the degree of freedom to be large enough relative to δ . The boundedness of r_δ itself is not assumed — this is important because r_δ is generically unbounded for light-tailed distributions.

Remark A.2 (On Assumption A.3). The condition $E[y_{t+h}^2 r_\delta^2] < \infty$ is the key moment restriction. It is satisfied when both y_{t+h} and ε_{1t} have sub-exponential or lighter tails, but requires verification for specific DGPs. This condition plays the same role as the bounded support assumption in the treatment effects literature — it ensures the reweighted moments are well-defined, but is weaker.

A.2 Efficient Influence Function

I derive the efficient influence function for $\theta \equiv \text{ARF}_h(\delta) = E[g_h(\varepsilon_{1t} + \delta) - g_h(\varepsilon_{1t})]$ stated in Proposition 5.1. The derivation proceeds in three steps: I characterize the tangent space of the semiparametric model, compute the pathwise derivative of θ , and verify that ψ_t^* is the Riesz representer of this derivative in the tangent space.

Step 1. Tangent space. The functional θ depends on the distribution of the full time series $\{z_t\}$ only through the bivariate marginal of $(\varepsilon_{1t}, y_{t+h})$. I work in the *bivariate marginal model*: the semiparametric model that treats the joint density of this pair, $p(e, y) = f(e)c(y | e)$, as the observed-data law, leaving both the marginal density f and the conditional density $c(\cdot | e)$ unrestricted (subject to Assumptions A.2–A.3). See Remark A.3 below for the relationship between this model and the full structural model.

Consider a smooth one-dimensional submodel $\{P_\eta : \eta \in (-\varepsilon, \varepsilon)\}$ through the true distribution $P = P_0$, with score

$$s(e, y) = \left. \frac{\partial}{\partial \eta} \log p_\eta(e, y) \right|_{\eta=0} = s_1(e) + s_2(e, y), \quad (20)$$

where $s_1(e) = \left. \frac{\partial}{\partial \eta} \log f_\eta(e) \right|_{\eta=0}$ is the marginal score and $s_2(e, y) = \left. \frac{\partial}{\partial \eta} \log c_\eta(y | e) \right|_{\eta=0}$ is the conditional score. These satisfy $E[s_1(\varepsilon_{1t})] = 0$ and $E[s_2(\varepsilon_{1t}, y_{t+h}) | \varepsilon_{1t}] = 0$ almost surely.

The tangent space \mathcal{T} of the bivariate marginal model is the set of all such scores:

$$\mathcal{T} = \{s_1(\varepsilon_{1t}) + s_2(\varepsilon_{1t}, y_{t+h}) : s_1 \in L_2^0(f), s_2 \in L_2(p), E[s_2 | \varepsilon_{1t}] = 0\}. \quad (21)$$

This is the full space $L_2^0(P)$: any mean-zero square-integrable function $h(e, y)$ can be decomposed as $h(e, y) = E[h(\varepsilon_{1t}, y_{t+h}) | \varepsilon_{1t} = e] + (h(e, y) - E[h(\varepsilon_{1t}, y_{t+h}) | \varepsilon_{1t} = e])$, where the first term has mean zero (by iterated expectations) and the second has conditional mean zero given ε_{1t} (Lemma A.1).

Lemma A.1 (Tangent space of the bivariate marginal model). *In the bivariate marginal model that treats $p(e, y) = f(e)c(y | e)$ as the observed-data density with both f and c unrestricted (subject to Assumptions A.2–A.3), the tangent space equals (21), i.e. the full $L_2^0(P)$.*

Proof. In a nonparametric model with density $p(e, y) = f(e)c(y | e)$ where f and c are both unrestricted, the marginal and conditional components can be perturbed independently. A marginal perturbation $f_\eta(e) = f(e)(1 + \eta s_1(e))$ with $s_1 \in L_2^0(f)$ is a valid submodel for any such s_1 , holding c fixed. A conditional perturbation $c_\eta(y |$

$e) = c(y | e)(1 + \eta s_2(e, y))$ with $E[s_2 | \varepsilon_{1t} = e] = 0$ is a valid submodel for any such s_2 , holding f fixed. Every mean-zero $h \in L_2(P)$ decomposes as $h(e, y) = E[h | \varepsilon_{1t} = e] + (h(e, y) - E[h | \varepsilon_{1t} = e])$, where the first term lies in $L_2^0(f)$ and the second has conditional mean zero, so $\mathcal{T} = L_2^0(P)$. \square

Remark A.3 (Relationship to the structural model). The bivariate marginal model is (weakly) larger than the set of bivariate marginals generated by the structural model (1)–(2), because stationarity and the specific functional form of the structural equations may constrain which pairs (f, c) are jointly achievable. A larger model has a (weakly) larger tangent space, hence a (weakly) higher efficiency bound. The bound $V_h^* = \text{Var}(\psi_t^*)$ derived here is therefore an upper bound on the efficiency bound that would obtain if one exploited the structural restrictions. The DR estimator attains V_h^* without requiring knowledge of those restrictions, which is the appropriate benchmark for an estimator designed to be robust to misspecification of the conditional mean.

If the structural functions ψ_i are unrestricted and the shock distribution is unrestricted (the maintained assumptions of this paper), the two models coincide in practice: any bivariate marginal (f, c) satisfying the regularity conditions can be generated by some configuration of the structural model. The formal verification of this claim would require showing that the resolvent of the transition operator associated with the structural model is bounded, ensuring that perturbations of the structural functions induce only second-order changes in the stationary distribution. I defer this to future work and note that it would strengthen the efficiency claim from an upper bound to an exact bound.

Step 2. Pathwise derivative. Under the submodel P_η , the marginal density and conditional mean become f_η and $g_{h,\eta}$. The functional is

$$\theta(\eta) = \int g_{h,\eta}(e + \delta) f_\eta(e) de - \int g_{h,\eta}(e) f_\eta(e) de.$$

Differentiating at $\eta = 0$, using $f_\eta(e) = f(e)(1 + \eta s_1(e)) + o(\eta)$ and the product rule:

$$\theta'(0) = \underbrace{\int g'_{h,\eta}(e + \delta) f(e) de - \int g'_{h,\eta}(e) f(e) de}_{(I)} + \underbrace{E[s_1(\varepsilon_{1t})(g_h(\varepsilon_{1t} + \delta) - g_h(\varepsilon_{1t}))]}_{(II)}, \quad (22)$$

where $g'_{h,\eta}(e) = \frac{d}{d\eta} \int y c_\eta(y | e) dy|_{\eta=0}$. Since $c_\eta(y | e) = c(y | e)(1 + \eta s_2(e, y)) + o(\eta)$ and $E[s_2 | \varepsilon_{1t} = e] = 0$, this gives:

$$g'_{h,\eta}(e) = \int y s_2(e, y) c(y | e) dy = E[y_{t+h} s_2(e, y_{t+h}) | \varepsilon_{1t} = e]. \quad (23)$$

Term (I). Substituting $u = e + \delta$ in the first integral:

$$(I) = \int g'_{h,\eta}(u) f(u - \delta) du - \int g'_{h,\eta}(e) f(e) de$$

$$\begin{aligned}
&= \int g'_{h,\eta}(e) [f(e - \delta) - f(e)] de \\
&= E[g'_{h,\eta}(\varepsilon_{1t}) (r_\delta(\varepsilon_{1t}) - 1)].
\end{aligned} \tag{24}$$

Now apply (23) and the tower property:

$$\begin{aligned}
(I) &= E\left[E[y_{t+h} s_2(\varepsilon_{1t}, y_{t+h}) \mid \varepsilon_{1t}] \cdot (r_\delta(\varepsilon_{1t}) - 1)\right] \\
&= E[y_{t+h} s_2(\varepsilon_{1t}, y_{t+h}) (r_\delta(\varepsilon_{1t}) - 1)].
\end{aligned} \tag{25}$$

Combining (25) and (II):

$$\theta'(0) = E[s_1(\varepsilon_{1t})(g_h(\varepsilon_{1t} + \delta) - g_h(\varepsilon_{1t}))] + E[s_2(\varepsilon_{1t}, y_{t+h}) (r_\delta(\varepsilon_{1t}) - 1) y_{t+h}]. \tag{26}$$

Step 3. Verification of the Riesz representation. I verify that $\psi_t^* = [g_h(\varepsilon_{1t} + \delta) - g_h(\varepsilon_{1t})] + (r_\delta(\varepsilon_{1t}) - 1)(y_{t+h} - g_h(\varepsilon_{1t})) - \theta$ satisfies $\theta'(0) = E[\psi_t^* \cdot s(\varepsilon_{1t}, y_{t+h})]$ for every score $s = s_1 + s_2 \in \mathcal{T}$. Since $\mathcal{T} = L_2^0(P)$, this identifies ψ_t^* as the unique element of \mathcal{T} representing the pathwise derivative, hence the efficient influence function.

Expand $E[\psi_t^* \cdot s]$ into three terms using the decomposition of ψ_t^* :

(a) *Regression term.*

$$\begin{aligned}
&E[(g_h(\varepsilon_{1t} + \delta) - g_h(\varepsilon_{1t})) (s_1(\varepsilon_{1t}) + s_2(\varepsilon_{1t}, y_{t+h}))] \\
&= E[(g_h(\varepsilon_{1t} + \delta) - g_h(\varepsilon_{1t})) s_1(\varepsilon_{1t})] + E[(g_h(\varepsilon_{1t} + \delta) - g_h(\varepsilon_{1t})) \underbrace{E[s_2 \mid \varepsilon_{1t}]}_{=0}] \\
&= E[(g_h(\varepsilon_{1t} + \delta) - g_h(\varepsilon_{1t})) s_1(\varepsilon_{1t})].
\end{aligned} \tag{27}$$

This matches term (II) in (26).

(b) *Augmentation term.*

$$\begin{aligned}
&E[(r_\delta(\varepsilon_{1t}) - 1)(y_{t+h} - g_h(\varepsilon_{1t})) (s_1(\varepsilon_{1t}) + s_2(\varepsilon_{1t}, y_{t+h}))] \\
&= E[(r_\delta - 1) \underbrace{E[y_{t+h} - g_h(\varepsilon_{1t}) \mid \varepsilon_{1t}]}_{=0} \cdot s_1(\varepsilon_{1t})] + E[(r_\delta(\varepsilon_{1t}) - 1)(y_{t+h} - g_h(\varepsilon_{1t})) s_2(\varepsilon_{1t}, y_{t+h})] \\
&= E[(r_\delta(\varepsilon_{1t}) - 1) y_{t+h} s_2(\varepsilon_{1t}, y_{t+h})] - E[(r_\delta(\varepsilon_{1t}) - 1) g_h(\varepsilon_{1t}) \underbrace{E[s_2 \mid \varepsilon_{1t}]}_{=0}] \\
&= E[(r_\delta(\varepsilon_{1t}) - 1) y_{t+h} s_2(\varepsilon_{1t}, y_{t+h})].
\end{aligned} \tag{28}$$

This matches term (I) in (26).

(c) *Centering term.* $E[\theta \cdot s] = \theta \cdot E[s] = 0$ because $s \in L_2^0(P)$.

Adding (27)–(28) recovers (26), confirming that

$$\theta'(0) = E[\psi_t^* \cdot s(\varepsilon_{1t}, y_{t+h})] \quad \text{for all } s \in \mathcal{T}. \tag{29}$$

Since the tangent space equals $L_2^0(P)$, the function ψ_t^* is the efficient influence function.

The semiparametric efficiency bound is $V_h^* = \text{Var}(\psi_t^*)$. \square

Remark A.4 (The orthogonal decomposition). The verification reveals why the two components of ψ_t^* are orthogonal (as claimed in Section 5.2). The regression term $g_h(\varepsilon_{1t} + \delta) - g_h(\varepsilon_{1t})$ is measurable with respect to ε_{1t} alone, while the augmentation term $(r_\delta - 1)(y_{t+h} - g_h(\varepsilon_{1t}))$ has conditional mean zero given ε_{1t} . Their covariance therefore vanishes by the tower property, yielding the variance decomposition (VD).

A.3 Consistency of the DR Estimator

Proposition A.1. *Under Assumptions A.1–A.4(a) and stochastic equicontinuity of the nuisance estimator classes, $\widehat{\text{ARF}}_h^{\text{DR}}(\delta) \xrightarrow{p} \text{ARF}_h(\delta)$ as $T \rightarrow \infty$. In particular, the boundedness condition on the misspecified component (Assumption A.4(a)) is satisfied automatically by the trimmed estimators of Section A.5.*

Proof. Decompose the estimation error as in Section 4.2:

$$\widehat{\text{ARF}}_h^{\text{DR}} - \text{ARF}_h = \underbrace{\frac{1}{T-h} \sum_t \psi_t^*}_{\rightarrow 0 \text{ by LLN}} + \underbrace{\frac{1}{T-h} \sum_t \Delta \hat{r}_\delta(\varepsilon_{1t}) \Delta \hat{g}_h(\varepsilon_{1t})}_{\text{product bias}}$$

where $\Delta \hat{r}_\delta = \hat{r}_\delta - r_\delta$, $\Delta \hat{g}_h = \hat{g}_h - g_h$, and R_T collects remainder terms. The first term converges to zero in probability: ψ_t^* has mean zero and finite variance (by Assumptions A.2–A.3), and a law of large numbers applies (using i.i.d. structure of ε_{1t} and mixing of y_{t+h}). The product bias term satisfies $|T^{-1} \sum_t \Delta \hat{r}_\delta \Delta \hat{g}_h| \leq [T^{-1} \sum_t \Delta \hat{r}_\delta^2]^{1/2} [T^{-1} \sum_t \Delta \hat{g}_h^2]^{1/2}$ by Cauchy–Schwarz. Under Assumption A.4(a), at least one factor is $o_p(1)$, so the product vanishes regardless of the other. The remainder R_T consists of terms involving $\Delta \hat{g}_h(\varepsilon_{1t} + \delta) - \Delta \hat{g}_h(\varepsilon_{1t})$ weighted by $\Delta \hat{r}_\delta(\varepsilon_{1t})$, which are controlled by the same stochastic equicontinuity argument. \square

A.4 Asymptotic Normality

Proposition A.2. *Under Assumptions A.1–A.4(b), stochastic equicontinuity of the nuisance estimator classes, and a mixing condition on $\{y_{t+h}\}$ sufficient for a CLT (e.g. α -mixing with summable mixing coefficients):*

$$\sqrt{T}(\widehat{\text{ARF}}_h^{\text{DR}}(\delta) - \text{ARF}_h(\delta)) \xrightarrow{d} N(0, \Sigma_h^*) \quad (30)$$

where the long-run variance is

$$\Sigma_h^* = \sum_{j=-\infty}^{\infty} \text{Cov}(\psi_t^*, \psi_{t-j}^*).$$

At $h = 0$, ψ_t^* is a measurable function of the i.i.d. pair (ε_{1t}, y_t) , so $\Sigma_0^* = V_0^* = \text{Var}(\psi_t^*)$. For $h > 0$, ψ_t^* inherits the serial dependence of y_{t+h} ; in particular, if the structural model implies

that y_{t+h} depends on $\varepsilon_{1,t+1}, \dots, \varepsilon_{1,t+h}$, then $\{\psi_t^*\}$ is at most $(h+p)$ -dependent (where p is the lag order of the structural model) and at least h -dependent, and $\Sigma_h^* \geq V_h^*$ with strict inequality generically.

The long-run variance is consistently estimated by the Newey–West HAC estimator

$$\hat{\Sigma}_h = \hat{\gamma}_0 + \sum_{j=1}^{B_T} \kappa\left(\frac{j}{B_T}\right) (\hat{\gamma}_j + \hat{\gamma}'_j), \quad \hat{\gamma}_j = \frac{1}{T-h} \sum_{t=j+1}^{T-h} \tilde{\psi}_t \tilde{\psi}_{t-j}, \quad (31)$$

where $\tilde{\psi}_t = \hat{\psi}_t - \widehat{\text{ARF}}_h^{\text{DR}}(\delta)$, κ is the Bartlett kernel, and $B_T \rightarrow \infty$ with $B_T/T \rightarrow 0$. A practical default is $B_T = \lceil 1.5h \rceil$.

Proof. Under the rate condition A.4(b), verified under primitive conditions in Proposition A.3, the product bias term is $o_p(T^{-1/2})$, so:

$$\sqrt{T}(\widehat{\text{ARF}}_h^{\text{DR}} - \text{ARF}_h) = \frac{1}{\sqrt{T}} \sum_{t=1}^T \psi_t^* + o_p(1).$$

At $h = 0$, the summands ψ_t^* are i.i.d. (as functions of (ε_{1t}, y_t)) with mean zero and variance $V_0^* < \infty$, so the Lindeberg–Lévy CLT gives $N(0, V_0^*)$ directly. For $h > 0$, the serial dependence in y_{t+h} induces dependence in ψ_t^* ; specifically, ψ_t^* is at most $(h+p)$ -dependent when the structural model has lag order p (and at least h -dependent, since y_{t+h} shares shocks $\varepsilon_{1,t+1}, \dots, \varepsilon_{1,t+h}$ with adjacent observations). Under the stated mixing condition, a CLT for dependent data (e.g., Davidson, 1994, Theorem 27.4) delivers the result with the long-run variance Σ_h^* . Consistency of the HAC estimator follows from standard results for kernel-based long-run variance estimation under mixing (Newey and West, 1987; Andrews, 1991). \square

Remark A.5 (Boundedness of the misspecified component). The Cauchy–Schwarz step in the proof requires that the L_2 norm of the misspecified nuisance component is $O_p(1)$; consistency of the well-specified component alone does not control the product if the other diverges. For the trimmed estimators defined in Section A.5, this condition holds automatically: the density ratio is capped at M_T and the regression is evaluated only where $\hat{f}(\varepsilon_{1t}) \geq a_T$, so both $[T^{-1} \sum_t \Delta \hat{p}^2]^{1/2}$ and $[T^{-1} \sum_t \Delta \hat{g}^2]^{1/2}$ are $O_p(1)$ even when the underlying nuisance is misspecified. For user-supplied nuisance estimators that are not trimmed or otherwise bounded, the double-robustness guarantee requires the additional assumption that the misspecified component is $O_p(1)$ in L_2 norm.

A.5 Verification of the Product Rate Condition

I verify Assumption A.4(b) under the primitive conditions in Assumptions A.5–A.8. The argument proceeds in three parts: we bound the $L_2(P)$ rate for a trimmed regression estimator, then for a trimmed density ratio estimator, and finally verify the product condition.

Trimmed estimators. Define the trimming threshold $a_T = T^{-\alpha}$ for $\alpha \in (0, 1/4)$, and let \hat{f} be a kernel density estimator of f with bandwidth b_f . The *trimmed regression estimator* is $\hat{g}_h^T(e) = \hat{g}_h(e) \cdot \mathbb{I}(\hat{f}(e) \geq a_T)$, and the *trimmed density ratio estimator* is

$$\hat{r}_\delta^T(e) = \frac{\hat{f}(e - \delta)}{\hat{f}(e)} \cdot \mathbb{I}(\hat{f}(e) \geq a_T, |\hat{f}(e - \delta)/\hat{f}(e)| \leq M_T) + 1 \cdot \mathbb{I}(\text{otherwise}),$$

where $M_T = T^\beta$ for some $\beta > 0$ truncates extreme ratio values (the default value 1 ensures $\hat{r}_\delta^T - 1 = 0$ in the trimmed region, contributing nothing to the DR augmentation term).

Proposition A.3 (Primitive rate verification). *Under Assumptions A.1–A.8, with trimming parameters $\alpha \in (0, 1/4)$ and $\beta > 0$, there exists a constant $C_0 > 0$ depending on f such that if $|\delta| < C_0 \sigma_1$ (where $\sigma_1^2 = \text{Var}(\varepsilon_{1t})$), then:*

- (a) $\|\hat{g}_h^T - g_h\|_2 = O_p(T^{-2/5}(\log T)^{1/4})$;
- (b) $\|\hat{r}_\delta^T - r_\delta\|_2 = O_p(T^{-2/5+C|\delta|/(2\sigma_1^2)}(\log T)^{1/4})$ for a constant $C > 0$;
- (c) $\|\hat{g}_h^T - g_h\|_2 \cdot \|\hat{r}_\delta^T - r_\delta\|_2 = O_p(T^{-4/5+C|\delta|/(2\sigma_1^2)}(\log T)^{1/2}) = o_p(T^{-1/2})$.

When the shock density f is known up to a finite-dimensional parameter θ (estimated at \sqrt{T} rate), part (b) improves to $\|\hat{r}_\delta - r_\delta\|_2 = O_p(T^{-1/2})$ and the constraint on $|\delta|/\sigma_1$ is eliminated.

Proof. Part (a): Regression rate. Decompose $\|\hat{g}_h^T - g_h\|_2^2 = I_1 + I_2$ where

$$I_1 = E[(\hat{g}_h(\varepsilon_{1t}) - g_h(\varepsilon_{1t}))^2 \cdot \mathbb{I}(\hat{f}(\varepsilon_{1t}) \geq a_T)], \quad I_2 = E[g_h(\varepsilon_{1t})^2 \cdot \mathbb{I}(\hat{f}(\varepsilon_{1t}) < a_T)].$$

Tail term I_2 . By uniform consistency of \hat{f} on compact sets, $P(\hat{f}(\varepsilon_{1t}) < a_T, f(\varepsilon_{1t}) \geq 2a_T) \rightarrow 0$. Define the effective support boundary c_T by $f(c_T) = 2a_T$, so $c_T = \Theta(\sqrt{\log(1/a_T)}) = \Theta(\sqrt{\alpha \log T})$. Then $I_2 \leq E[g_h(\varepsilon_{1t})^2 \cdot \mathbb{I}(|\varepsilon_{1t}| > c_T)] + o_p(1) \rightarrow 0$ by dominated convergence under Assumption A.6. By Cauchy–Schwarz and sub-Gaussian tail bounds: $I_2 = O_p(T^{-c\alpha})$ for a constant $c > 0$.

Interior term I_1 . On $\{f(\varepsilon_{1t}) \geq a_T/2\} \supseteq \{\hat{f}(\varepsilon_{1t}) \geq a_T, \|\hat{f} - f\|_\infty < a_T/2\}$ (which holds with probability approaching 1), the standard pointwise MSE formula for the local linear estimator gives:

$$E[(\hat{g}_h(e) - g_h(e))^2] = \underbrace{\frac{b_g^4 \mu_2^2}{4} (g_h''(e))^2}_{\text{squared bias}} + \underbrace{\frac{R(K) \sigma^2(e)}{T b_g f(e)}}_{\text{variance}} + \text{h.o.t.}$$

The serial correlation in $u_{t+h} = y_{t+h} - g_h(\varepsilon_{1t})$ contributes $O(h/(T^2 b_g f(e)))$ to the variance, which is negligible for fixed h since the kernel weights depend only on the i.i.d. sequence $\{\varepsilon_{1s}\}$.

Integrating over $\{|e| \leq c_T\}$ against $f(e)$:

$$E[I_1] \leq \frac{b_g^4 \mu_2^2}{4} \int (g_h''(e))^2 f(e) de + \frac{R(K)}{T b_g} \int_{|e| \leq c_T} \sigma^2(e) de + o(1)$$

$$= O(b_g^4) + \frac{R(K)\bar{\sigma}^2}{Tb_g} \cdot 2c_T + o(1), \quad (32)$$

where $\bar{\sigma}^2 = \sup_e \sigma^2(e) < \infty$ by Assumption A.7, and the integral of the squared bias uses Assumption A.5(b).

With $b_g \asymp T^{-1/5}(\log T)^{1/10}$ and $c_T = O(\sqrt{\log T})$:

$$\|\hat{g}_h^T - g_h\|_2^2 = O_p(T^{-4/5}(\log T)^{2/5} + T^{-4/5}(\log T)^{2/5}) = O_p(T^{-4/5}(\log T)^{1/2}),$$

yielding $\|\hat{g}_h^T - g_h\|_2 = O_p(T^{-2/5}(\log T)^{1/4})$.

Part (b): Density ratio rate. Decompose $\|\hat{r}_\delta^T - r_\delta\|_2^2 = J_1 + J_2 + J_3$, separating the interior (both density and ratio untrimmed), the density tail, and the ratio tail. The tail terms J_2 and J_3 are handled identically to I_2 : by Assumption A.6, $J_2 + J_3 = O_p(T^{-c\alpha} + T^{-\beta})$.

For the interior J_1 , linearize on $\{f(e) \geq a_T/2, r_\delta(e) \leq 2M_T\}$:

$$\hat{r}_\delta(e) - r_\delta(e) \approx \frac{\Delta \hat{f}(e - \delta) - r_\delta(e) \Delta \hat{f}(e)}{f(e)},$$

where $\Delta \hat{f}(e) = \hat{f}(e) - f(e)$. Squaring and integrating against $f(e)$:

$$J_1 \leq 2 \int_{|e| \leq c_T} \frac{E[\Delta \hat{f}(e - \delta)^2]}{f(e)} de + 2 \int_{|e| \leq c_T} r_\delta(e)^2 \frac{E[\Delta \hat{f}(e)^2]}{f(e)} de. \quad (33)$$

For the standard kernel density estimator of i.i.d. data, $E[\Delta \hat{f}(u)^2] = O(b_f^4(f''(u))^2 + f(u)/(Tb_f))$.

Consider the variance contribution to the second integral in (33):

$$\frac{2}{Tb_f} \int_{|e| \leq c_T} r_\delta(e)^2 de.$$

On the restricted domain $|e| \leq c_T = O(\sqrt{\log T})$, the density ratio satisfies $\sup_{|e| \leq c_T} r_\delta(e)^2 = O(T^{C|\delta|/\sigma_1^2})$ for a constant $C > 0$ determined by f . Hence,

$$\int_{|e| \leq c_T} r_\delta(e)^2 de \leq 2c_T \cdot \sup_{|e| \leq c_T} r_\delta(e)^2 = O(\sqrt{\log T} \cdot T^{C|\delta|/\sigma_1^2}).$$

The variance contribution is therefore $O(T^{C|\delta|/\sigma_1^2 - 4/5}(\log T)^{1/2})$, which vanishes provided $C|\delta|/\sigma_1^2 < 4/5$. The bias contribution is lower order under Assumption A.5(a). The first integral in (33) is handled by the substitution $u = e - \delta$, yielding analogous bounds. Collecting terms:

$$\|\hat{r}_\delta^T - r_\delta\|_2 = O_p(T^{(C|\delta|/\sigma_1^2 - 4/5)/2}(\log T)^{1/4}).$$

Part (c): *Product rate.* Multiplying parts (a) and (b):

$$\|\hat{g}_h^T - g_h\|_2 \cdot \|\hat{r}_\delta^T - r_\delta\|_2 = O_p(T^{C|\delta|/(2\sigma_1^2)-4/5}(\log T)^{1/2}).$$

This is $o_p(T^{-1/2})$ provided $C|\delta|/(2\sigma_1^2) < 3/10$, i.e., $|\delta| < C_0 \sigma_1$ for $C_0 = 3\sigma_1/(5C)$.

Lemma A.2 (From population norms to sample averages). *Under Assumptions A.1–A.8, with the trimmed nuisance estimators \hat{g}_h^T and \hat{r}_δ^T defined in Section A.5:*

$$\frac{1}{T} \sum_{t=1}^T \Delta \hat{r}_\delta(\varepsilon_{1t}) \Delta \hat{g}_h(\varepsilon_{1t}) = E[\Delta \hat{r}_\delta(\varepsilon_{1t}) \Delta \hat{g}_h(\varepsilon_{1t})] + o_p(T^{-1/2}),$$

where $\Delta \hat{r}_\delta = \hat{r}_\delta^T - r_\delta$ and $\Delta \hat{g}_h = \hat{g}_h^T - g_h$. In particular, the product bias in the DR estimator is controlled by the $L_2(P)$ rates established in Proposition A.3.

Proof. The argument has two steps: a leave-one-out replacement that removes the dependence between each nuisance estimate and its own evaluation point, and a law of large numbers for the resulting sum.

Step 1: Leave-one-out approximation for \hat{g}_h . The local linear estimator with kernel K and bandwidth b_g admits the closed-form leave-one-out representation

$$\hat{g}_h(\varepsilon_{1t}) = \hat{g}_{h,-t}(\varepsilon_{1t}) + \frac{w_{tt}(y_{t+h} - \hat{g}_{h,-t}(\varepsilon_{1t}))}{1 - w_{tt}},$$

where w_{tt} is the diagonal element of the local linear smoother matrix at point ε_{1t} . For a positive kernel (Assumption A.8), the smoother weights satisfy $0 \leq w_{tt} \leq C/(Tb_g f(\varepsilon_{1t}))$ for a constant C depending only on K . On the trimmed region $\{\hat{f}(\varepsilon_{1t}) \geq a_T\}$, we have $f(\varepsilon_{1t}) \geq a_T/2$ with probability approaching one (by uniform consistency of \hat{f}), so

$$w_{tt} = O\left(\frac{1}{Tb_g a_T}\right) = O\left(\frac{T^\alpha}{T^{4/5}(\log T)^{1/10}}\right) = o(1),$$

since $\alpha < 1/4 < 4/5$. The leave-one-out residual $y_{t+h} - \hat{g}_{h,-t}(\varepsilon_{1t})$ is $O_p(1)$ on the trimmed region (by Assumption A.7 and the consistency of $\hat{g}_{h,-t}$), so

$$\hat{g}_h(\varepsilon_{1t}) - \hat{g}_{h,-t}(\varepsilon_{1t}) = O_p\left(\frac{1}{Tb_g a_T}\right) \tag{34}$$

uniformly over the trimmed region.

Step 1': Leave-one-out approximation for \hat{r}_δ . The trimmed density ratio \hat{r}_δ^T depends on ε_{1t} through the kernel density estimator \hat{f} . The own-observation contribution to $\hat{f}(\varepsilon_{1t})$ is $K(0)/(Tb_f)$, so the leave-one-out density satisfies

$$\hat{f}(\varepsilon_{1t}) = \hat{f}_{-t}(\varepsilon_{1t}) + \frac{K(0)}{Tb_f}.$$

On the trimmed region where $\hat{f}(\varepsilon_{1t}) \geq a_T$, a first-order expansion of the ratio gives

$$\hat{r}_\delta(\varepsilon_{1t}) - \hat{r}_{\delta,-t}(\varepsilon_{1t}) = O_p\left(\frac{1}{Tb_f a_T}\right), \quad (35)$$

which is $o(T^{-1/2})$ under the stated bandwidth and trimming choices, by the same calculation as above.

Step 2: Law of large numbers. Replace each same-sample evaluation with its leave-one-out counterpart:

$$\frac{1}{T} \sum_t \Delta \hat{r}_\delta(\varepsilon_{1t}) \Delta \hat{g}_h(\varepsilon_{1t}) = \frac{1}{T} \sum_t \Delta \hat{r}_{\delta,-t}(\varepsilon_{1t}) \Delta \hat{g}_{h,-t}(\varepsilon_{1t}) + o_p(T^{-1/2}),$$

where the remainder is controlled by (34), (35), and Cauchy–Schwarz. In the leave-one-out sum, $\hat{g}_{h,-t}(\varepsilon_{1t})$ and $\hat{r}_{\delta,-t}(\varepsilon_{1t})$ are independent of $(\varepsilon_{1t}, y_{t+h})$ conditional on the remaining observations, because the local linear and kernel density estimators exclude observation t . The summands therefore behave as a triangular array with the same conditional mean as the population quantity. Since the trimmed nuisance functions are bounded ($|\hat{r}_\delta^T| \leq M_T$, $|\hat{g}_h^T| \leq C c_T$ on the trimmed region by Assumption A.7), the summands have uniformly bounded second moments. A law of large numbers for triangular arrays (e.g., Newey, 1994, Lemma A.2) then gives

$$\frac{1}{T} \sum_t \Delta \hat{r}_{\delta,-t}(\varepsilon_{1t}) \Delta \hat{g}_{h,-t}(\varepsilon_{1t}) = E[\Delta \hat{r}_\delta(\varepsilon_{1t}) \Delta \hat{g}_h(\varepsilon_{1t})] + o_p(T^{-1/2}),$$

where the replacement of leave-one-out with full-sample quantities inside the expectation introduces only an $O(1/(Tb_g a_T))$ error, which is $o(T^{-1/2})$. Combining with the previous display completes the proof. \square

Remark A.6 (Cross-fitting for non-kernel nuisance estimators). The leave-one-out argument in Lemma A.2 exploits the closed-form structure of local linear regression and kernel density estimation: the own-observation influence is explicit, bounded, and vanishing. Sieve estimators, penalized regression, random forests, and neural networks lack this closed-form structure, so the leave-one-out approximation is not available and same-sample evaluation can introduce a non-negligible bias. In such cases, cross-fitting is required: partition the sample into K folds (using sequential non-overlapping blocks with buffer zones of width h to respect the serial dependence in y_{t+h}), estimate the nuisance functions on the complement of each fold, and evaluate them on the fold itself. The product rate condition (Assumption A.4(b)) is then verified fold-by-fold using the independence between the nuisance estimates and the evaluation points, following Chernozhukov et al. (2018).

Known shock density. When f belongs to a parametric family $\{f_\theta : \theta \in \Theta\}$ with θ estimated at \sqrt{T} rate, a Taylor expansion of $r_\delta(\cdot; \hat{\theta})$ around θ_0 gives $\|\hat{r}_\delta - r_\delta\|_2 = O_p(T^{-1/2})$, eliminating the δ/σ_1 constraint and reducing the product rate condition to $\|\hat{g}_h^T - g_h\|_2 =$

$o_p(1)$. □

Remark A.7 (Role of the sub-Gaussian tail condition). Assumption A.5(b) is used in two places in the proof of Proposition A.3. First, it ensures that the effective support boundary satisfies $c_T = \Theta(\sqrt{\log T})$: solving $f(c_T) = 2a_T$ with $a_T = T^{-\alpha}$ and $f(e) \asymp \exp(-\lambda e^2)$ gives $c_T = \Theta(\sqrt{\alpha \log T / \lambda})$. Second, it controls the density ratio on the effective support: for sub-Gaussian f ,

$$\log r_\delta(e) = \log f(e - \delta) - \log f(e) \leq \lambda(2|e||\delta| + \delta^2),$$

so $\sup_{|e| \leq c_T} r_\delta(e)^2 = O(\exp(4\lambda|\delta|c_T)) = O(T^{C|\delta|/\sigma_1^2})$ with $C = 4\lambda\sqrt{\alpha/\lambda} = 4\sqrt{\alpha\lambda}$, as claimed in the proof.

For heavier-tailed distributions (e.g., t -distributions with moderate degrees of freedom), c_T grows polynomially in T rather than logarithmically, inflating the density ratio bound and tightening the constraint on $|\delta|/\sigma_1$. The double-robustness property (Proposition A.1) and the form of the efficient influence function (Proposition 5.1) are unaffected, since they do not depend on the specific rates in Proposition A.3. When the shock density belongs to a known parametric family, the sub-Gaussian condition is not needed at all (see the final paragraph of the proof).

Remark A.8 (The δ/σ_1 constraint). The constraint $|\delta| < C_0\sigma_1$ arises from nonparametric estimation of the density ratio in the tails: for large shifts, the counterfactual density $f(\cdot - \delta)$ has poor overlap with $f(\cdot)$, inflating the variance of the reweighted estimator. Three observations mitigate this limitation: (i) When the shock density belongs to a known parametric family (estimated at \sqrt{T} rate), the constraint disappears entirely. (ii) Regularized density ratio estimators such as uLSIF produce bounded weights by construction, avoiding tail instability. (iii) Within the DR framework, even a biased density ratio estimate contributes small product bias when \hat{g}_h is well-specified, because the bias is proportional to $\|\Delta\hat{r}\| \cdot \|\Delta\hat{g}\|$.

A.6 On the GHKP Estimator's Efficiency

Proposition A.4 (Informal). *Under the conditions of GHKP Proposition 5.1, the GHKP estimator has asymptotic variance $V_h^{reg} = \text{Var}[g_h(\varepsilon_{1t} + \delta) - g_h(\varepsilon_{1t})]$, which satisfies $V_h^{reg} \leq V_h^*$ with equality if and only if $\sigma^2(\varepsilon_{1t}) = 0$ a.s. (i.e., y_{t+h} is a deterministic function of ε_{1t} , which is generically false).*

As discussed in Section 5.2, this inequality reflects the different maintained assumptions of the two models, not superefficiency of the GHKP estimator. Whether the variance premium V_h^{aug} is worth paying depends on the difficulty of estimating g_h in the specific application.

B Monte Carlo Figures

Computations were performed using the Wake Forest University High Performance Computing Facility ([Information Systems and Wake Forest University, 2021](#)).

B.1 Double robustness under structural misspecification

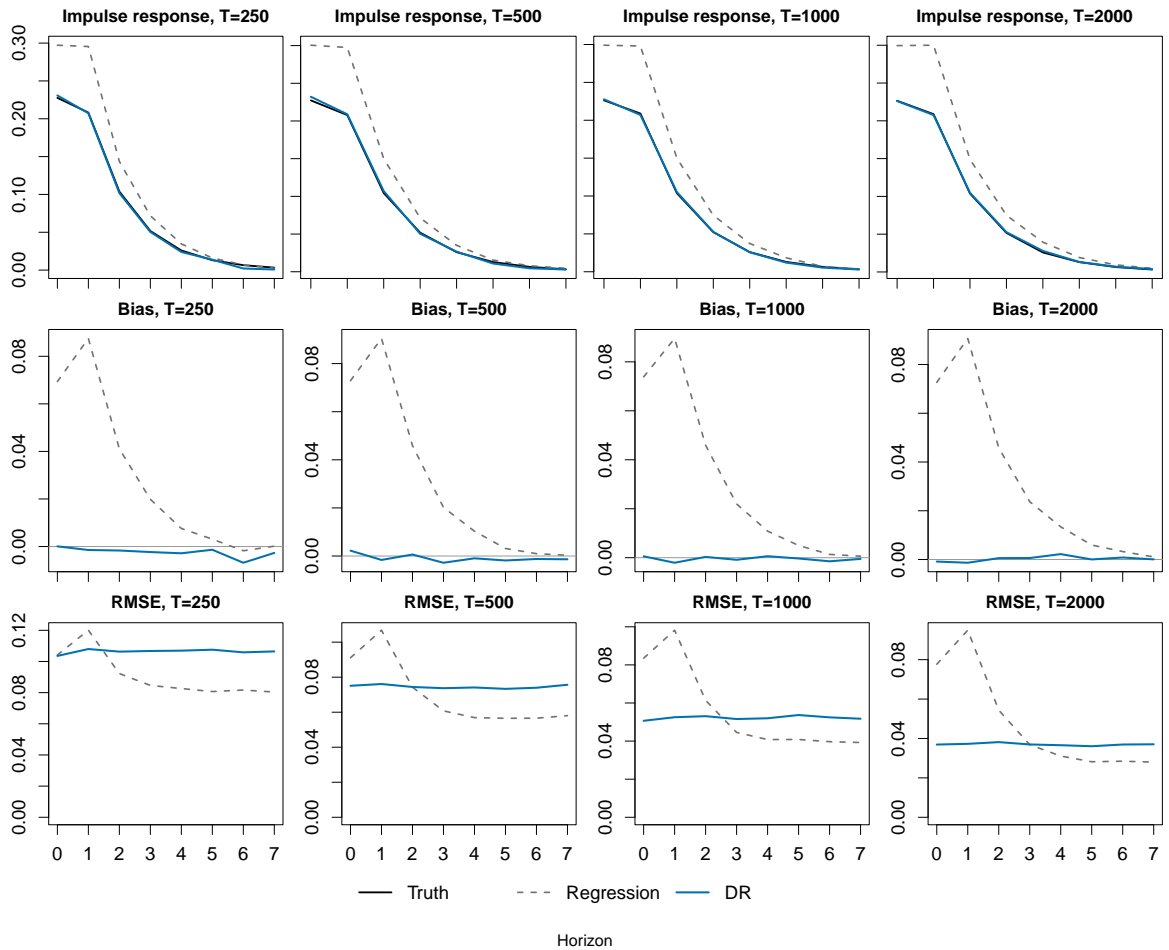


Figure 1: Misspecified regression. DGP with ReLU nonlinearity $f(x) = \max(x, 0)$ and shock size $\delta = 1$. Each column corresponds to a sample size T ; rows show the estimated ARF, mean bias, and RMSE. Dashed: linear LP omitting the nonlinear term; solid blue: DR estimator pairing the same misspecified regression with a parametric density ratio

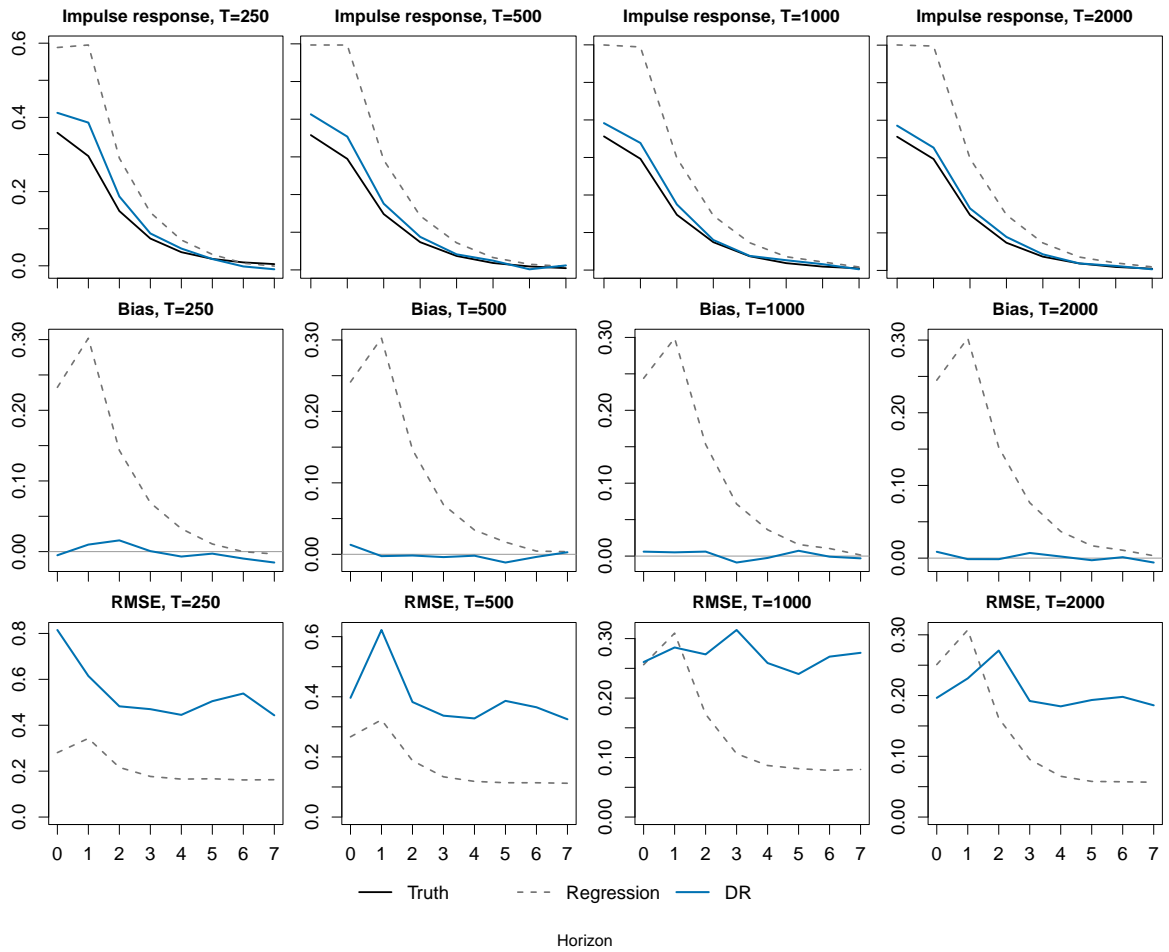


Figure 2: Misspecified regression. DGP with ReLU nonlinearity $f(x) = \max(x, 0)$ and shock size $\delta = 2$. Each column corresponds to a sample size T ; rows show the estimated ARF, mean bias, and RMSE. Dashed: linear LP omitting the nonlinear term; solid blue: DR estimator pairing the same misspecified regression with a parametric density ratio.

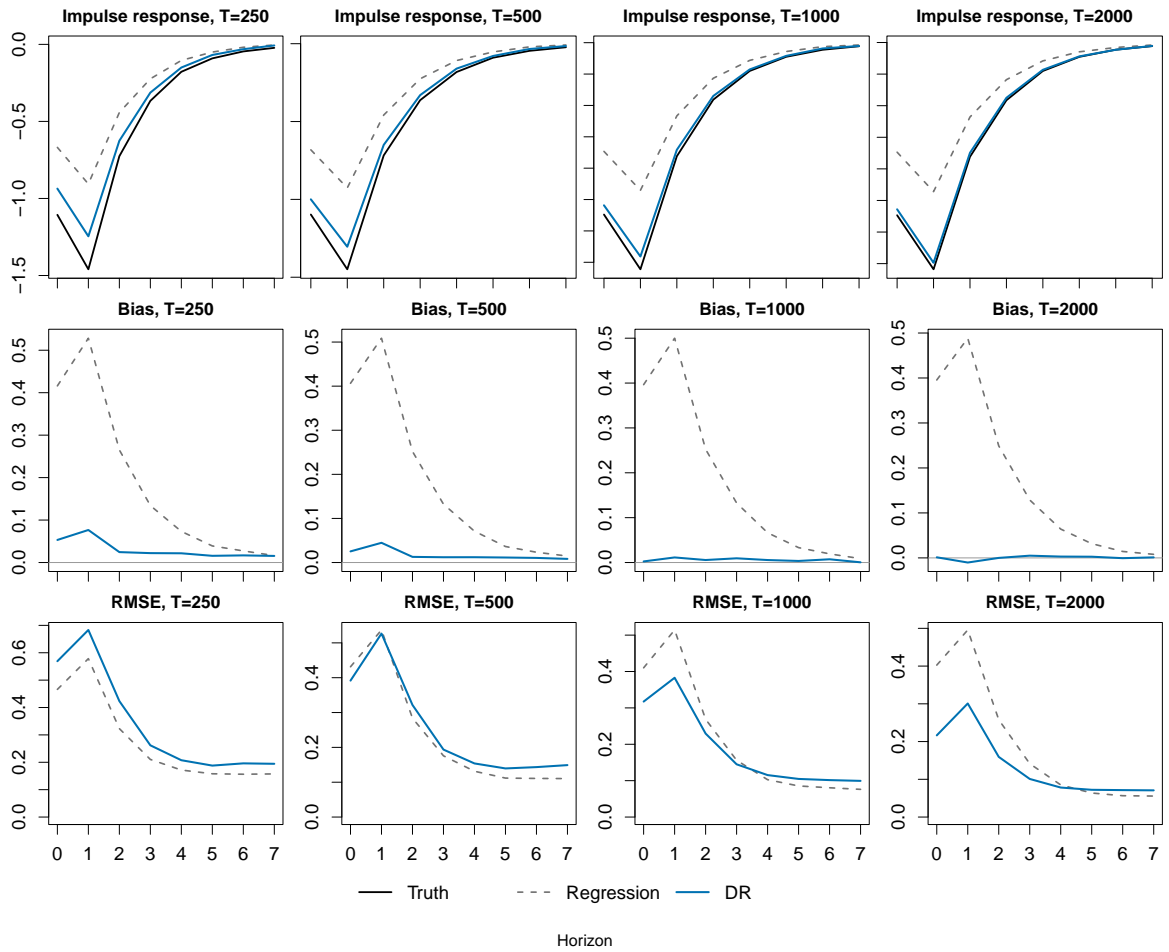


Figure 3: Misspecified regression. DGP with cubic nonlinearity $f(x) = x^3$ and shock size $\delta = 1$. Each column corresponds to a sample size T ; rows show the estimated ARF, mean bias, and RMSE. Dashed: linear LP omitting the nonlinear term; solid blue: DR estimator pairing the same misspecified regression with a parametric density ratio.

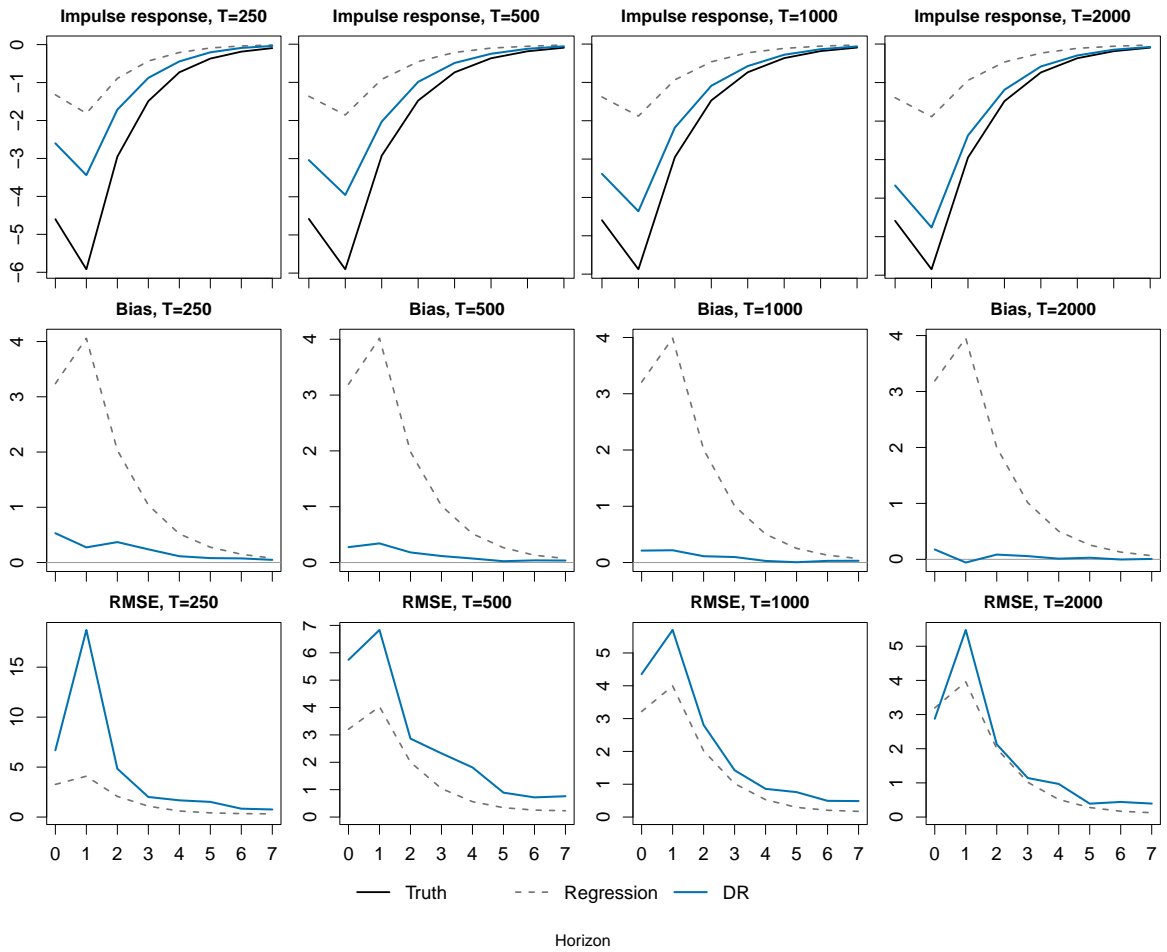


Figure 4: Misspecified regression. DGP with cubic nonlinearity $f(x) = x^3$ and shock size $\delta = 2$. Each column corresponds to a sample size T ; rows show the estimated ARF, mean bias, and RMSE. Dashed: linear LP omitting the nonlinear term; solid blue: DR estimator pairing the same misspecified regression with a parametric density ratio.

B.2 Cost of robustness when the regression is well-specified

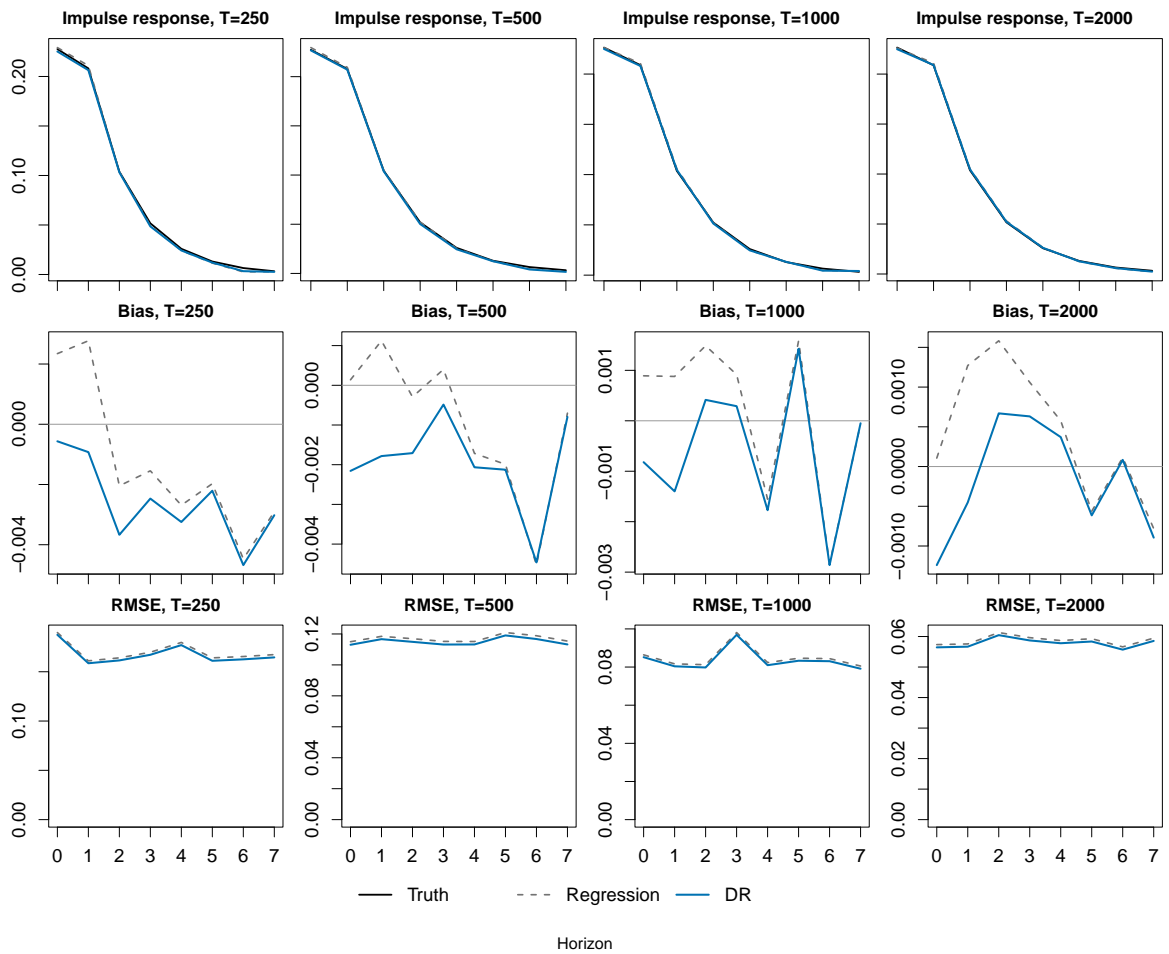


Figure 5: Well-specified regression. DGP with ReLU nonlinearity $f(x) = \max(x, 0)$ and shock size $\delta = 1$. Each column corresponds to a sample size T ; rows show the estimated ARF, mean bias, and RMSE. Dashed: local linear LP; solid blue: DR estimator pairing the same well-specified regression with a parametric density ratio.

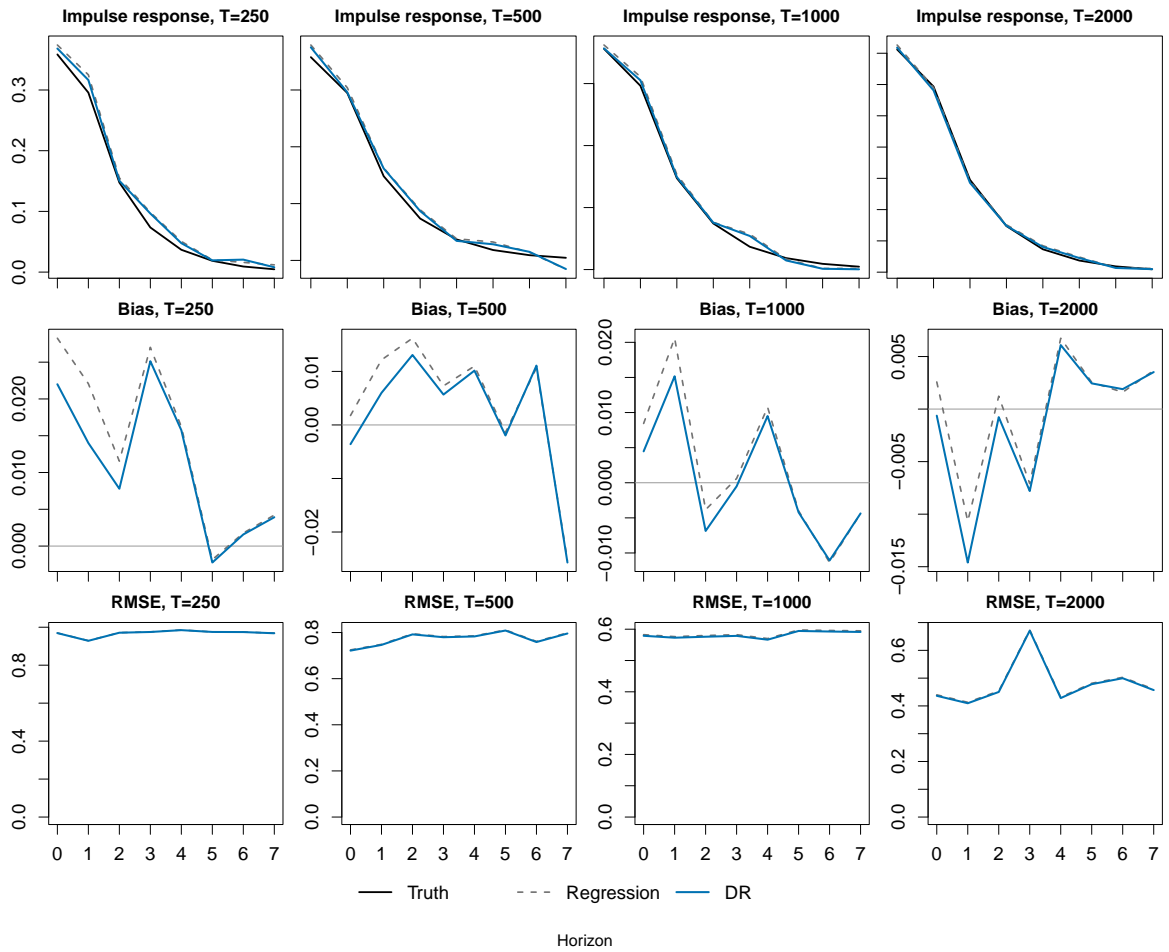


Figure 6: Well-specified regression. DGP with ReLU nonlinearity $f(x) = \max(x, 0)$ and shock size $\delta = 2$. Each column corresponds to a sample size T ; rows show the estimated ARF, mean bias, and RMSE. Dashed: local linear LP; solid blue: DR estimator pairing the same well-specified regression with a parametric density ratio.

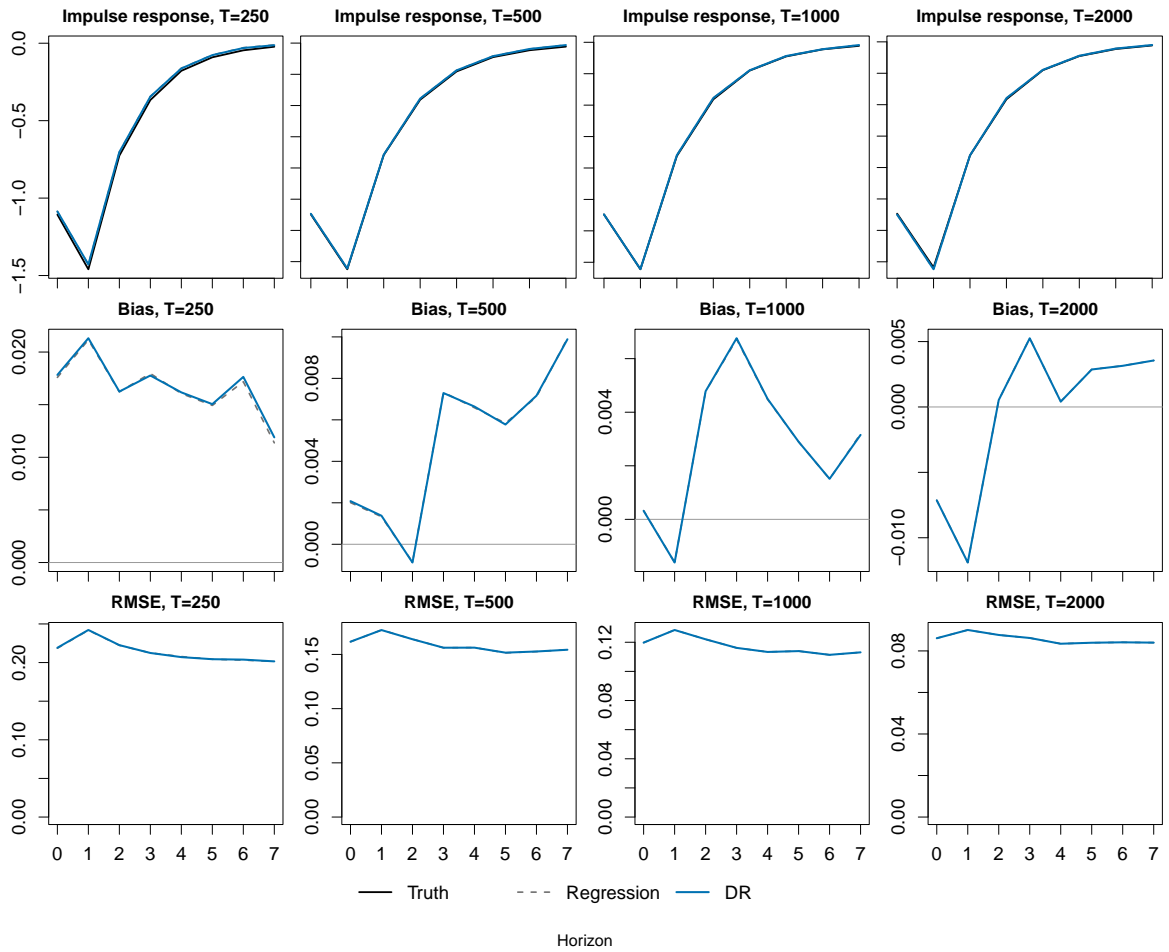


Figure 7: Well-specified regression. DGP with cubic nonlinearity $f(x) = x^3$ and shock size $\delta = 1$. Each column corresponds to a sample size T ; rows show the estimated ARF, mean bias, and RMSE. Dashed: power series LP; solid blue: DR estimator pairing the same well-specified regression with a parametric density ratio.

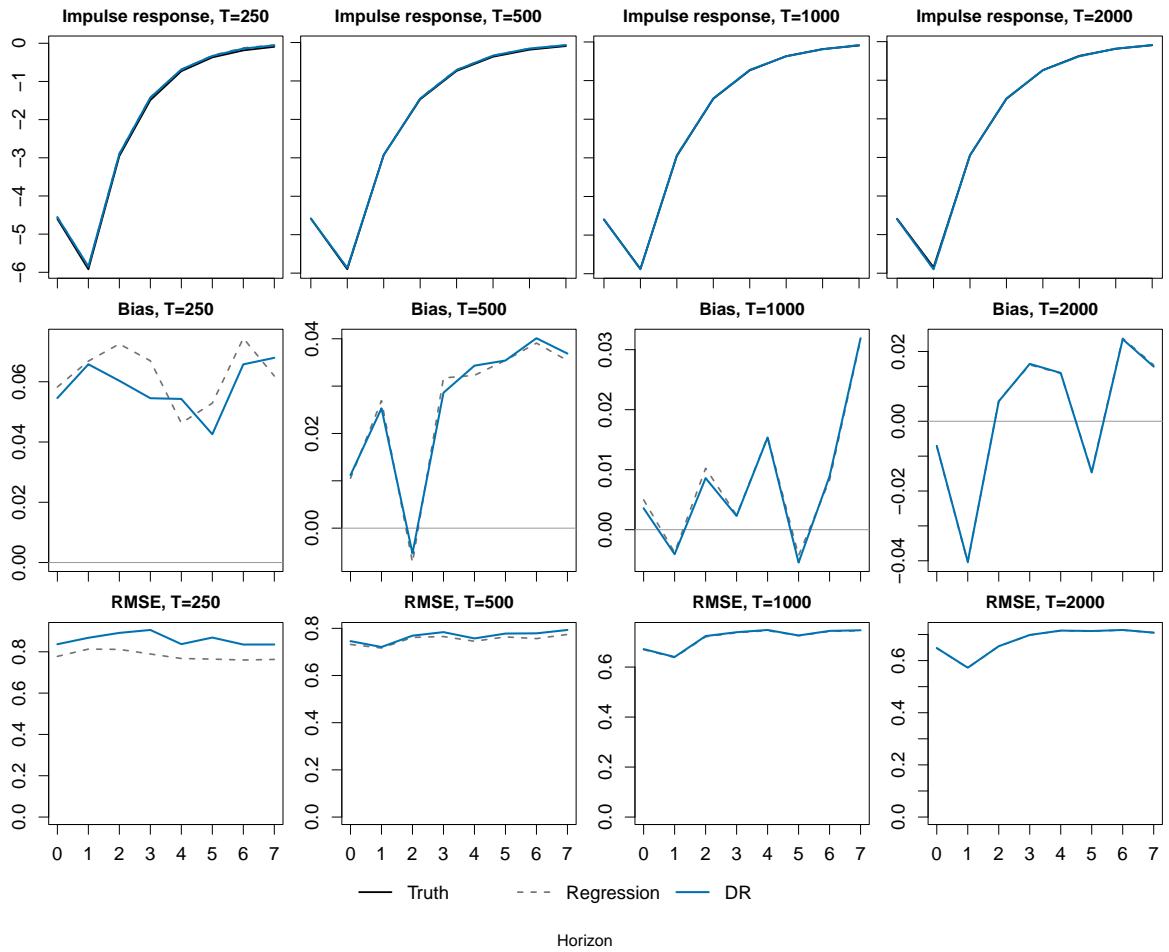


Figure 8: Well-specified regression. DGP with cubic nonlinearity $f(x) = x^3$ and shock size $\delta = 2$. Each column corresponds to a sample size T ; rows show the estimated ARF, mean bias, and RMSE. Dashed: power series LP; solid blue: DR estimator pairing the same well-specified regression with a parametric density ratio.