Identification and Inference with Machine-Learned Instruments

Fangzhou Yu*

July, 2026

Abstract

Instrumental-variables estimation increasingly pools many or high-dimensional instruments into a single machine-learned first stage, with rich controls partialled out. The resulting estimand, the partialled-out IV coefficient built from any signal of the instruments, is a signal-weighted average of the heterogeneous effects, which gives an opaque first stage a precise structural meaning. The average is convex whenever a covariance-monotonicity condition holds, and we provide a microfoundation for that condition based on vector monotonicity. With a learned signal, however, the usual debiased moment is not Neyman-orthogonal, and its first-order bias is a drift toward the learner's own signal-weighted average, so naive inference remains valid only for that learner-dependent target. We construct a heterogeneity-robust orthogonal score that restores $\sqrt{N}$ inference on the fixed, learner-invariant target at no efficiency cost, and provide a Hausman-type diagnostic and identification-robust confidence sets.

Keywords: instrumental variables; correlated random coefficients; Neyman orthogonality; local average treatment effects.

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

1 Introduction

Instrumental variables estimation increasingly runs through machine learning. To exploit many or high-dimensional instruments, a flexible first stage pools them into a single predicted signal, and a rich set of controls is partialled out nonparametrically. The resulting estimator is the partially linear IV coefficient of the "flexible IV" routines in DoubleML and ddml. It is now common in judge- and examiner-leniency, shift-share, and quarter-of-birth returns-to-schooling designs (Angrist and Frandsen, 2022), and Chen et al. (2021) analyze it as an optimal-instrument problem. This estimator is the object of the paper. Once treatment effects are heterogeneous in a way correlated with selection, two questions about it remain open. What does the estimator identify, and are its standard errors valid? We answer both.

Two strands of the literature focus on the same estimator without connecting. The weights literature shows that under unobserved effect heterogeneity, linear IV estimands can lose their causal reading. Two-stage least squares with multiple instruments can place negative weights on local effects (Mogstad et al., 2021, 2024), and so can covariate misspecification (Blandhol et al., 2022; Słoczyński, 2024). That literature offers no estimation theory for a machine-learned signal. The double/debiased machine learning literature (Chernozhukov et al., 2018) delivers $\sqrt{N}$ inference for the partially linear IV coefficient, but only under effect homogeneity or heterogeneity in observed covariates (Emmenegger and Bühlmann, 2021; Scheidegger et al., 2026). The unobserved, selection-correlated heterogeneity we treat is what makes the estimand signal-weighted and the naive moment non-orthogonal, and neither feature arises in those settings. The question is therefore what a machine-learned IV estimand identifies, and how valid inference can be conducted on it.

Our first contribution answers the identification question. Write the outcome as $Y = Y(0) + X\Delta$, with Δ the heterogeneous, unobserved effect of X that may be chosen based on Δ itself. For any measurable signal $g(Z, W)$ of instruments Z and covariates W, and in particular a cross-fitted machine-learning first stage, applied work computes the partialled-out IV estimand

$$\beta_{IV} = \frac{\mathbb{E}[\tilde{Y}\tilde{g}]}{\mathbb{E}[\tilde{X}\tilde{g}]}, \qquad \tilde{A} := A - \mathbb{E}[A \mid W], \tag{1}$$

with covariates removed by Robinson-style partialling (Robinson, 1988). This is the flexible "rich covariate" adjustment that Blandhol et al. (2022) and Słoczyński (2024) show is needed to read linear IV as weakly causal. Under conditional exogeneity and relevance alone, this

estimand is a signal-weighted average treatment effect (SWATE),

$$\beta_{IV} = \mathbb{E}_{\Delta}\big[\Delta\,\omega(\Delta)\big], \qquad \omega(\delta) = \frac{\mathbb{E}[\tilde{X}\tilde{g} \mid \Delta = \delta]}{\mathbb{E}[\tilde{X}\tilde{g}]}, \tag{2}$$

whose weights integrate to one. Whatever the algorithm outputs, β_{IV} is the corresponding signal-weighted average of effects, which gives a structural meaning to an opaque first stage. These weights are defined on the effect scale and need no single-index structure. The marginal-treatment-effect weights of Heckman and Vytlacil (2005); Heckman et al. (2006) are instead defined on the resistance scale and require such structure. And unlike the complier-effect estimators built from Imbens and Angrist (1994) and Angrist and Imbens (1995), the object we interpret is the very signal our estimation focuses on.

The outcome model $Y = Y(0) + X\Delta$ is the correlated random coefficients (CRC) model (Garen, 1984; Wooldridge, 1997; Heckman and Vytlacil, 1998; Card, 2001). It naturally nests the canonical binary-treatment case of local average treatment effect (LATE), and the SWATE becomes a weighted combination of LATEs. It also extends the problem to continuous X with linear effects. Two caveats worth noting. With continuous X, the linear effect form is a restriction adopted for tractability, while with binary X it is vacuous. The weight function $\omega(\cdot)$ is not identified, only β_{IV} and its convexity are, and we do not attempt to estimate the weights. Classical CRC results give conditions under which IV recovers the average effect. The SWATE characterizes what IV recovers absent those conditions, which reappear as the special case $\omega = 1$.

The weights integrate to one but need not be non-negative, which is the concern the multiple-instrument critique raises. Our second contribution is to address this concern. In a scalar threshold-crossing model they are automatically non-negative. For the optimal first stage $g_X = \mathbb{E}[X \mid Z, W]$ pooling combined instruments, we show that the vector monotonicity of Goff (2024)—equivalently the actual monotonicity of Mogstad et al. (2021)—together with a positive-association condition on the instruments delivers a new sufficient condition. We call it Conditional Covariance Monotonicity (CCM), under which β_{IV} is a convex combination with no single-index restriction on choice behavior. We share this monotone choice model with Goff (2024) but take a different route to convex weights: rather than the tailored complier-parameter estimator, we add a positive-association restriction on the observable distribution of Z given W and read the convex signal-weighted average off the workhorse projection. The threshold-crossing model that Vytlacil (2002) showed equivalent to instrument monotonicity is its single-index special case, and CCM is strictly weaker than monotone

selection.

Our Interpretation holds for any signal. Estimating the optimal signal is where standard theory breaks, and our third contribution addresses this issue. The naive debiased moment, implemented in the standard DML routines such as `DoubleML` package in `R` and `Python`, is not Neyman-orthogonal in g_X under heterogeneity. Its first-order bias consists of the component of the instrument's effect on outcomes that the linear signal fails to absorb, and the first-stage estimation error of g_X. Because this bias is the covariance of those two pieces, it has a geometric consequence. Rescalings and all covariate-measurable recalibrations of the signal leave the estimate unchanged, and only "shape" change aligned with that unabsorbed component moves it. This is why well-tuned learners in the standard DML routines frequently appear valid, and why that appearance cannot be relied upon.

The bias is a drift of the estimand rather than noise. The naive moment performs valid inference on the SWATE generated by the learner's own signal, $\beta_{IV}(\hat{g}_X)$. That object is a legitimate signal-weighted estimand, convex whenever CCM holds for $\hat{g}_X$. The naive procedure targets a well-defined estimand, but only a learner-dependent one. It cannot attain valid inference for the fixed target uniformly.

To recover the fixed target, we construct a CRC-robust orthogonal score, adding one new nuisance $q_Y = \mathbb{E}[Y \mid Z, W]$ for debiasing purpose. It is Neyman-orthogonal in all four nuisances and $\sqrt{N}$-consistent under a product-rate condition that tolerates a slowly learned q_Y. It is also the efficient influence function for the target, and identical to the naive score under homogeneity. Around it we build a practice toolkit. The first is a Hausman-type diagnostic, the studentized naive-versus-robust contrast. The second is a standard error result that explains why projection first stages corrects the point estimate but not the reported standard error, the semiparametric analogue of heterogeneity-invalid two-stage least squares standard errors (Kolesár, 2013; Lee, 2018). The third is an Anderson–Rubin confidence set. The optimal instrument $g_X = \mathbb{E}[X \mid Z, W]$ is classical under homogeneity (Chamberlain, 1987; Newey, 1990; Belloni et al., 2012), and our contribution is to identify the orthogonal and efficient moment under heterogeneity. The identification-robust set guards a different axis than the classical and many weak instrument literatures (Staiger and Stock, 1997; Dufour, 1997; Andrews et al., 2019; Mikusheva and Sun, 2022), namely weakness of the residual signal left after nonparametric partialling. All proofs are collected in Appendix.

The remainder of the paper is organized as follows. Section 2 defines the estimand, proving

the representation theorems, micro-founding CCM, and relating the framework to LATE. Sections 3 and 4 then develop the moving target that naive DML estimates and the CRC-robust score that recovers the fixed, learner-invariant one; and Section 5 supplies the alignment diagnostic and identification-robust inference under a weak residual signal. Sections 6–7 present the Monte Carlo and empirical evidence, and Section 8 concludes.

2 Identification and Interpretation

2.1 The CRC Model and the Partialled-out Estimand

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. We observe an independent and identically distributed sample of random vectors (Y, X, Z, W). For notational clarity, we suppress individual subscripts. Throughout, "almost surely" (a.s.) means outside a $\mathbb{P}$-null set; for a relation stated at a fixed realization of a conditioning variable—as in $\mathbb{E}[\cdot \mid \Delta = \delta] \geq 0$ or $\mathrm{Cov}(\cdot \mid W = w) \geq 0$—it means the relation holds for almost every value of that variable.

- $Y \in \mathbb{R}$ is the observed continuous outcome.
- $X \in \mathcal{X} \subseteq \mathbb{R}$ is the observed endogenous treatment. This accommodates discrete treatments (e.g., binary $\mathcal{X} = \{0, 1\}$) as well as continuous treatments (e.g., dosage, duration).
- $Z \in \mathcal{Z}$ is the observed instrumental variable, where $\mathcal{Z}$ can be a scalar, a discrete vector, or a high-dimensional feature space $\mathbb{R}^d$.
- $W \in \mathcal{W}$ is a high-dimensional vector of observed pre-treatment covariates.
- $g : \mathcal{Z} \times \mathcal{W} \to \mathbb{R}$ is a measurable signal function with finite second moments. The function $g(Z, W)$ acts as a predictive projection of the instrument (e.g., a non-parametric cross-fitted machine learner). For notational brevity, we refer to it as $g(Z)$ where the dependence on W is implicit, or simply g.

We formalize the underlying causal data-generating process within the canonical Correlated Random Coefficient (CRC) framework (Wooldridge, 1997; Heckman and Vytlacil, 1998; Angrist et al., 2000). Let $Y(0)$ denote the untreated potential outcome, and let Δ denote the

heterogeneous, unobserved causal treatment effect. The structural outcome equation is

$$Y = Y(0) + X\Delta. \tag{3}$$

Remark 1. *When X is binary, Equation (3) inherently holds under SUTVA, with $\Delta = Y(1) - Y(0)$. When X is continuous, the CRC model imposes a constant marginal response for each individual, i.e., $\partial Y_i/\partial X_i = \Delta_i$. Here X is endogenous because individuals may self-select intensity on their untreated $Y(0)$ or effects Δ (essential heterogeneity). This linearity is the substantive restriction we pay for tractability with continuous X. Angrist et al. (2000) dispense with it and obtain a double average, taken over a marginal response margin and an instrument-induced weighting, while we obtain the single effect-scale average in Theorem 1 below. With binary X the restriction is vacuous.*

Following Robinson (1988) and the Frisch–Waugh–Lovell theorem, we partial out the covariates W. Let $\tilde{A} = A - \mathbb{E}[A \mid W]$ denote the conditional residual of any random variable A. The generalized partialled-out instrumental variable (IV) estimand is defined as

$$\beta_{IV} := \frac{\mathbb{E}[\tilde{Y}\tilde{g}]}{\mathbb{E}[\tilde{X}\tilde{g}]}. \tag{4}$$

When g is a fixed instrument and only the covariate projections $\mathbb{E}[Y \mid W], \mathbb{E}[X \mid W]$ are learned, the partialled-out IV moment is already Neyman-orthogonal, so the double/debiased machine-learning theory of Chernozhukov et al. (2018) delivers valid $\sqrt{N}$ inference regardless of effect heterogeneity, with Theorem 1 supplying its structural interpretation. When the signal itself is learned, with g a cross-fitted first stage pooling many instruments, orthogonality breaks under essential heterogeneity. The identification results of this section hold for any measurable g and cover both regimes. The estimation theory of Sections 3–5 is what the second regime requires.

Having fixed the estimand, we now state the identifying assumptions and prove the two representation theorems that give β_{IV} its structural meaning.

2.2 Assumptions and discussion

Assumption 1 (Conditional Exogeneity of Instruments)**.**

$$Z \perp\!\!\!\perp \big(Y(0), \Delta\big) \mid W.$$

We implicitly assume finite second moments for all relevant variables.

Assumption 2 (Conditional Relevance)**.** *The residualized treatment $\tilde{X}$ and the residualized instrument signal $\tilde{g}$ have a non-zero population covariance, i.e., $\mathbb{E}[\tilde{X}\tilde{g}] \neq 0$.*

Because β_{IV} is invariant to rescaling the signal, and in particular to its sign, formally shown in Lemma 1 below, we adopt throughout the harmless normalization $\mathbb{E}[\tilde{X}\tilde{g}] > 0$. This is a convention rather than an additional restriction, as it fixes the denominator's sign so that the following discussion of non-negative weights and convex combination read literally.

Assumption 3 (Conditional Covariance Monotonicity, CCM)**.** *Let $F_\Delta(\cdot)$ be the marginal cumulative distribution function of Δ, with support* Supp(Δ)*. Then*

$$\mathbb{E}\big[\tilde{X}\tilde{g} \mid \Delta = \delta\big] \geq 0 \quad a.s.$$

Read within an effect stratum, CCM requires that the signal not, on average, push treatment the wrong way for units sharing a given effect. The failure case is a high effect subgroup whose treatment responds negatively to the signal, which can then receive negative SWATE weight. Because CCM constrains only a conditional covariance rather than individual response functions, it sits strictly below the monotonicity notions of the literature. We show in following sections that monotonicity in Imbens and Angrist (1994) implies it in the scalar threshold model, and vector monotonicity (Goff, 2024) together with positive association of the instruments implies it for g_X with multiple instruments.

Note that CCM is not directly testable, and $\omega(\cdot)$ is not identified. Our defense is instead micro-foundations from primitive choice behavior in Section 2.4, which restricts only the observable distribution of Z given W.

A third point concerns the denominator. As $\mathbb{E}[\tilde{X}\tilde{g}] \to 0$ the weights ω lose meaning even when every conditional covariance is well behaved, because the SWATE is a ratio whose

denominator vanishes. This failure is directly related to the weak instrument regime, and it is what motivates the identification-robust confidence sets of Section 5.

2.3 The representation theorems

Theorem 1 (Signal-weighted ATE)**.** *Suppose Assumptions 1 and 2 hold, and let g be any measurable signal. Then*

(i) the generalized partialled-out IV estimand β_{IV} admits a representation as a Signal-Weighted Average Treatment Effect (SWATE),

$$\beta_{IV} = \mathbb{E}_\Delta\big[\Delta \cdot \omega(\Delta)\big] = \int_{\mathrm{Supp}(\Delta)} \delta \cdot \omega(\delta)\, dF_\Delta(\delta), \qquad \omega(\delta) = \frac{\mathbb{E}[\tilde{X}\tilde{g} \mid \Delta = \delta]}{\mathbb{E}[\tilde{X}\tilde{g}]};$$

(ii) the weighting function integrates to one,

$$\mathbb{E}_\Delta\big[\omega(\Delta)\big] = 1;$$

(iii) if in addition Assumption 3 (CCM) holds, $\omega(\delta) \geq 0$ a.s., and consequently β_{IV} identifies a well-defined convex combination of the heterogeneous treatment effects Δ.

The representation draws only on conditional exogeneity and relevance. It asks nothing of the signal g beyond measurability, with no monotonicity, no correct specification, and no first-stage optimality. It therefore holds for any measurable g, in particular a black-box, cross-fitted machine-learning first stage. This is what equips otherwise opaque machine-learning IV estimands with an exact structural meaning.

In the scalar one-instrument case this weighting scheme is not new. Huntington-Klein (2020) shows that IV under first-stage heterogeneity identifies the first stage effect weighted average $\mathbb{E}[\gamma\beta]/\mathbb{E}[\gamma]$, with γ the individual first-stage response. This is Theorem 1 for a single raw instrument, with his γ in the role of our $\mathbb{E}[\tilde{X}\tilde{g} \mid \Delta]$. Theorem 1 generalizes it to any measurable signal $g(Z, W)$. Our weights act on the structural effects Δ, whereas the Heckman–Vytlacil identification invoked by Chen et al. (2021) places weights on marginal treatment effects, convex only under a single-index structure that Theorem 1 does not require.

Textbook just-identified IV is already covered. Taking the signal to be a fixed scalar instrument, $g = Z$, Theorem 1 gives

$$\beta_{IV} = \frac{\mathbb{E}[\tilde{Y}\tilde{Z}]}{\mathbb{E}[\tilde{X}\tilde{Z}]} = \mathbb{E}_{\Delta}\big[\Delta\,\omega(\Delta)\big], \qquad \omega(\delta) = \frac{\mathbb{E}[\tilde{X}\tilde{Z} \mid \Delta = \delta]}{\mathbb{E}[\tilde{X}\tilde{Z}]},$$

the Wald/2SLS estimand with covariates partialled out.

2.4 Covariance Monotonicity and its Micro-foundation

As mentioned in the previous section, CCM is not directly testable, and its warrant is that primitive models of treatment choice imply it. It holds automatically in the single-index threshold model $X = h(g(Z), U, W)$ with h weakly increasing in g. The map $g \mapsto \mathbb{E}[X \mid g, W, \Delta = \delta]$ is then non-decreasing and covaries non-negatively with g by Chebyshev's association inequality, so CCM holds by construction. But routing the whole instrument vector through one scalar index forces choice behavior to be effectively homogeneous across instruments, which Mogstad et al. (2021) show is restrictive precisely when Z collects multiple instruments. Their response is to relax single-index (Imbens and Angrist, 1994) monotonicity to a multidimensional monotone choice model. Mogstad et al. (2021) work with partial monotonicity, under which each instrument moves treatment in a common direction across individuals while that direction may itself depend on the levels of the other instruments. The stronger actual monotonicity, in which each instrument's direction is fixed globally, coincides with the vector monotonicity of Goff (2024). We adopt this assumption and show that under it CCM survives, given a positive dependence condition on the instruments, delivering non-negative SWATE weights without any single-index restriction.

Let the treatment be generated by an arbitrary measurable structural map

$$X = h(Z, W, V), \tag{5}$$

where $V \in \mathcal{V}$ is an unobserved, possibly infinite-dimensional vector of choice heterogeneity. This nests the single-index model $h(g(Z), U, W)$ but allows the relative responses to different coordinates of Z to vary across individuals. We take the signal to be the optimal first stage

$$g_X(Z, W) := \mathbb{E}[X \mid Z, W].$$

The same projection is used for estimation in Section 3, and write $m_X(W) := \mathbb{E}[X \mid W] = \mathbb{E}[g_X \mid W]$, so that $\tilde{g}_X = g_X - m_X$.

Assumption 4 (Extended Conditional Exogeneity)**.**

$$Z \perp\!\!\!\perp \big(Y(0), \Delta, V\big) \mid W.$$

Assumption 5 (Structural Vector Monotonicity, VM)**.** *Almost surely, the structural choice map $z \mapsto h(z, W, V)$ is coordinate-wise non-decreasing on $\mathbb{R}^d$.*

Assumption 6 (Conditional Positive Association, PA)**.** *For any two coordinate-wise non-decreasing functions $\phi, \psi : \mathbb{R}^d \to \mathbb{R}$ with finite second moments,*

$$\mathrm{Cov}\big(\phi(Z), \psi(Z) \mid W\big) \geq 0 \quad a.s.$$

Assumption 4 strengthens Assumption 1 only by naming the selection unobservable V. VM imposes the vector monotonicity of Goff (2024), which is also the nonparametric form of the actual monotonicity of Mogstad et al. (2021). It assumes each individual's choice map is coordinate-wise non-decreasing in a fixed direction. This is strictly stronger than the partial monotonicity of Mogstad et al. (2021), yet imposes no single-index restriction on how individuals weight the instruments. The fixed direction is precisely what Proposition 1 needs, since its covariance is signed by associating two coordinate-wise non-decreasing functions. Each instrument is accordingly sign-normalized so that its directional effect is non-decreasing, and PA must be oriented to match. PA is the multivariate generalization of non-negative dependence (Esary et al., 1967). It holds automatically when the coordinates of Z are conditionally independent given W, and more generally under affiliation (Milgrom and Weber, 1982) or, in the Gaussian case, whenever all conditional correlations are non-negative (Pitt, 1982). It is stronger than non-negative pairwise correlation, because for $d \geq 2$ the inequality $\mathrm{Cov}(Z_i, Z_j \mid W) \geq 0$ does not imply association.

Two practical points attach to PA. First, unlike CCM itself it is refutable. It restricts only the observable law of Z given W, so its pairwise implications (positive quadrant dependence of each (Z_i, Z_j) given W) can be assessed with the data, even if full multivariate association is harder to test. Second, association is directional, so the instruments must be sign-normalized first. Orient each coordinate of Z in the direction of its own first-stage effect on X (estimable from $\mathbb{E}[X \mid Z, W]$) so that VM reads coordinate-wise non-decreasing, and impose PA on the oriented vector. This is a design step for the analyst rather than an extra assumption on

nature.

Proposition 1 (Micro-founding CCM via Vector Monotonicity)**.** *Suppose Assumptions 4–6 hold and* $\mathbb{E}[X^2] < \infty$*. Then, for the signal* $g_X(Z, W) = \mathbb{E}[X \mid Z, W]$*, Conditional Covariance Monotonicity (Assumption 3) holds,*

$$\mathbb{E}\big[\tilde{X}\,\tilde{g}_X \mid \Delta = \delta\big] \geq 0 \quad a.s.$$

Consequently, under Assumption 2, Theorem 1 applies.

Remark 2. *Assumption 4 and VM alone do not deliver CCM. Under VM, both the type-*δ *response* $f_\delta(Z, W) := \mathbb{E}[X \mid Z, W, \Delta = \delta]$ *and the population signal* $g_X(Z, W)$ *are non-decreasing in each coordinate of* Z*. The sign of their conditional covariance and hence of the SWATE weight* $\omega(\delta)$ *is therefore governed by how the coordinates of* Z *depend on one another. Positive association makes the covariance non-negative, and negative association can make it negative. In the latter case an individual who responds mostly to one instrument can have a high treatment when the population signal, driven by a different instrument, is low, so* $\mathrm{Cov}(f_\delta, g_X \mid W) < 0$ *and* $\omega(\delta)$ *turns negative. Positive association is precisely what rules this out, and it cannot be weakened to non-negative pairwise correlation. At edge case where the instruments exactly offset, relevance itself fails, since* $\mathbb{E}[\tilde{X}\tilde{g}_X] = 0$*.*

Note that negative association is structural in fixed-pool categorical designs. In judge- or examiner-leniency settings, case-assignment balancing renders two judges' leniencies mechanically negatively associated within a court × period cell, and blocks of mutually exclusive dummies are negatively associated within block by construction. These are the configurations Remark 2 isolates, where sign-normalization cannot help and the projection can carry negative weight. Since PA is sufficient but not necessary for CCM, its failure does not by itself imply negative weights. But the route to convexity discussed in this section is then unavailable, and a convexity guarantee falls back on the single-index homogeneity this section set out to relax.

We share the monotone choice model of Goff (2024); our departure lies not in the behavioral assumption but in what we do with it. Where Goff (2024) identifies complier parameters under vector monotonicity through a tailored estimator, and Mogstad et al. (2021) characterize the sign of the raw two-stage least squares weights, we add a design restriction on the instruments and read a convex signal-weighted average directly off the workhorse projection g_X. Proposition 1 thus delivers not a new estimator but a coherence: it requires

no re-estimation of complier-specific parameters and no discrete-support requirement, and it gives validity for the very signal g_X that Section 3 learns.

2.5 Relation to the LATE Framework

The SWATE representation incorporates the LATE framework. When multiple instruments are combined, the aggregate two-stage least squares estimand can fracture into a "causal salad" of local effects with negative weights (Mogstad et al., 2021). Projecting the instruments into one scalar signal $g(Z, W)$ restores a single interpretable estimand. That estimand is convex only under CCM (Assumption 3), which the threshold-crossing structure below delivers under the extended exogeneity condition stated next.

First fix a covariate value w. Define the conditional first-stage covariance and, whenever it is positive, the corresponding within-stratum covariance ratio by

$$D(w) := \mathbb{E}[\tilde{X}\tilde{g} \mid W = w], \qquad \beta_{IV}(w) := \frac{\mathbb{E}[\tilde{Y}\tilde{g} \mid W = w]}{D(w)}.$$

Until Corollary 1, all expectations, probabilities, and distributions are taken under the conditional law given $W = w$, and the dependence of the objects below on w is suppressed. The paper's unconditional estimand β_{IV} is not an unweighted average of the ratios $\beta_{IV}(w)$; the corollary gives the exact denominator-weighted aggregation.

Consider a binary treatment $X \in \{0, 1\}$ generated by a canonical threshold-crossing model

$$X \;=\; \mathbf{1}\{\, U \le g(Z, W)\,\},$$

where U is unobserved treatment resistance. We use Assumption 4 with the selection unobservable specialized to $V = U$. Because g is measurable in (Z, W), this condition implies

$$g(Z, W) \perp\!\!\!\perp \big(Y(0), \Delta, U\big) \mid W.$$

Suppose the signal g takes L ordered values $g_1 < g_2 < \cdots < g_L$ with $\mathbb{P}(g = g_\ell) = p_\ell$, write $\bar{g} := \mathbb{E}[g] = \sum_\ell p_\ell g_\ell$, and recall $\tilde{g} = g - \bar{g}$, $\tilde{X} = X - \mathbb{E}[X]$. For $\ell = 2, \ldots, L$ define the margin-ℓ complier event

$$C_\ell \;:=\; \{\, g_{\ell-1} < U \le g_\ell \,\},$$

so that an individual is treated once the signal reaches level g_ℓ but not at $g_{\ell-1}$. Units with $U \le g_1$ are always-takers and units with $U > g_L$ are never-takers; for both groups, X is constant in g. Write $\pi_\ell(\delta) := \mathbb{P}(C_\ell \mid \Delta = \delta)$ and $\bar{\pi}_\ell := \mathbb{P}(C_\ell)$, and define

$$\mathcal{L}_+ := \{\ell \in \{2, \dots, L\} : \bar{\pi}_\ell > 0\}, \qquad \text{LATE}_\ell := \mathbb{E}[\Delta \mid C_\ell], \quad \ell \in \mathcal{L}_+,$$

where LATE_ℓ is the local average treatment effect (Imbens and Angrist, 1994) for margin-ℓ compliers. Zero-probability margins are omitted throughout. Define the covariance increments

$$S_\ell \;:=\; \mathbb{E}\big[(g - \bar{g})\,\mathbf{1}\{g \ge g_\ell\}\big] \;=\; \mathbb{P}(g \ge g_\ell)\,\big(\mathbb{E}[g \mid g \ge g_\ell] - \mathbb{E}[g]\big) \;=\; \text{Cov}\big(\mathbf{1}\{g \ge g_\ell\},\, g\big) \;\ge\; 0,$$

the inequality holding because $\mathbf{1}\{g \ge g_\ell\}$ and g are both nondecreasing in g.

Proposition 2 (Multiple-Instrument LATE Decomposition)**.** *Suppose Assumption 4 holds with $V = U$. For almost every covariate value w with $D(w) > 0$, the threshold-crossing model above implies:*

(i) the projection estimand is a convex combination of margin-specific LATEs,

$$\beta_{IV}(w) \;=\; \sum_{\ell \in \mathcal{L}_+} \lambda_\ell \,\text{LATE}_\ell, \qquad \lambda_\ell \;:=\; \frac{S_\ell\,\bar{\pi}_\ell}{\sum_{\ell' \in \mathcal{L}_+} S_{\ell'}\,\bar{\pi}_{\ell'}},$$

with $\lambda_\ell \ge 0$ and $\sum_{\ell \in \mathcal{L}_+} \lambda_\ell = 1$;

(ii) for almost every δ, the within-stratum SWATE weighting function

$$\omega_w(\delta) := \frac{\mathbb{E}[\tilde{X}\tilde{g} \mid \Delta = \delta, W = w]}{D(w)}$$

admits the closed form

$$\omega_w(\delta) \;=\; \sum_{\ell \in \mathcal{L}_+} \lambda_\ell \,\frac{dF_{\Delta \mid C_\ell}(\delta)}{dF_\Delta(\delta)}, \tag{6}$$

with the same weights λ_ℓ, where all displayed distributions are conditional on $W = w$ and $dF_{\Delta|C_\ell}/dF_\Delta$ is the complier likelihood ratio, equivalently $\mathbb{P}(C_\ell \mid \Delta = \delta)/\mathbb{P}(C_\ell)$.

Corollary 1 (Unconditional Decomposition)**.** *Under the threshold-crossing model, suppose Assumptions 4 and 2 hold, with $V = U$ in Assumption 4 and the positive-denominator normalization, and suppose g has finite conditional support for almost every covariate value.*

Restoring the dependence on w in Proposition 2,

$$\begin{aligned}\beta_{IV} &= \frac{\displaystyle\int_{\{w:\, D(w)>0\}} D(w)\beta_{IV}(w)\, dF_W(w)}{\mathbb{E}[D(W)]} \\ &= \int_{\{w:\, D(w)>0\}} \sum_{\ell\in\mathcal{L}_+(w)} \frac{D(w)\lambda_\ell(w)}{\mathbb{E}[D(W)]}\mathrm{LATE}_\ell(w)\, dF_W(w).\end{aligned} \tag{7}$$

The displayed joint (W, ℓ) weights are non-negative and integrate to one, so β_{IV} is a convex mixture of the covariate- and margin-specific local effects.

Corollary 1 makes the covariate aggregation precise. A stratum is weighted by its conditional first-stage covariance $D(w)$ as well as its population frequency. A stratum with $D(w) = 0$ contributes zero to both unconditional moments, even though its conditional ratio is undefined.

Unlike the aggregate two-stage least squares estimand, whose multiple-instrument weights can be negative (Mogstad et al., 2021), the scalar-signal projection has non-negative weights both within covariate strata and unconditionally. The non-negative covariance increments S_ℓ generate the threshold-model CCM kernel, so collapsing the instruments into g turns the "causal salad" back into a convex combination of margin-specific local effects. The λ_ℓ are the ordered-instrument analogue of Angrist and Imbens (1995). Where Mogstad et al. (2021) obtain a signed decomposition of the raw 2SLS estimand and Goff (2024) identify complier parameters under vector monotonicity via a tailored estimator, we characterize what the workhorse projection estimand is, and Section 3 shows how to make its inference robust.

Part (ii) shows where the within-stratum weight comes from. Equation (6) expresses ω_w as a λ-convex average of complier likelihood ratios, and the ratio $dF_{\Delta|C_\ell}/dF_\Delta$ measures how over- or under-represented an effect δ is among margin-ℓ compliers in that stratum, so $\omega_w(\delta) > 1$ when δ is over-represented. An effect δ^* occurring only among always- or never-takers has $\pi_\ell(\delta^*) = 0$, hence $\omega_w(\delta^*) = 0$. This is the CRC generalization of LATE's always-/never-taker zero-weighting, and the formal counterpart of Huntington-Klein (2020)'s scalar observation that units with $\gamma_i \approx 0$ receive vanishing weight.

3 The Moving Target: What Naive DML Estimates

Everything to this point concerns the population signal. We now show the fact that the optimal signal $g_X = \mathbb{E}[X \mid Z, W]$ must itself be learned, and that learning it breaks the standard moment.

The sample analogue of the covariance ratio is not robust to the first-order estimation error of a machine-learned signal. Under essential heterogeneity the naive moment is not Neyman-orthogonal, so plugging in a slowly converging $\hat{g}$ contaminates $\hat{\beta}_{IV}$ with regularization bias. This section formalizes that obstruction and quantifies it, showing that the naive plug-in validly estimates a moving target that drifts with the learner. Section 4 then constructs a locally robust score that restores $\sqrt{N}$-consistent and asymptotically normal estimation of the fixed target.

The closest precursor is Chen et al. (2021), who learn the optimal instrument for a linear IV model by predicting treatment from instruments and covariates with cross-fitted machine learning. They establish its asymptotic normality and Chamberlain efficiency, interpreted under threshold selection as a convex average of marginal treatment effects. Their argument that the learned signal's error has no first-order effect rests on mean-independence of the structural error from the exogenous variables, which essential heterogeneity breaks. The effective residual is correlated with the instrument through selection, so the first-order error of a learned g_X enters the moment and the moment is not Neyman-orthogonal for any fixed target. Their normality is validity for the learner-dependent estimand the plug-in converges to, whereas inference on the fixed β_0 requires the orthogonal score below. The estimand is thus mostly harmless for what it estimates, though not for how one infers it.

We adopt one notational convention for the whole section. The symbol Δ (standalone) always denotes the structural effect of (3), whereas a nuisance estimation error always carries a nuisance operand, as in $\Delta l_Y := \hat{l}_Y - l_Y$ and likewise $\Delta m_X, \Delta q_Y, \Delta g_X$, collected as $\Delta\eta := \hat{\eta} - \eta_0$. The direction $\delta_g \in L_2(\sigma(Z, W))$ denotes a perturbation of the signal.

3.1 The Learned Signal and the Target

To render the estimand operational with off-the-shelf machine learning, we fix the generic signal to the structurally optimal, predictively estimable quantity

$$g(Z, W) = g_X(Z, W) := \mathbb{E}[X \mid Z, W]. \tag{8}$$

This is the projection of the (possibly high-dimensional) instrument Z onto the treatment, which is what a cross-fitted first-stage learner targets. The four nuisance functions are collected in the vector $\eta = (l_Y, m_X, q_Y, g_X)$,

$$\begin{aligned} l_Y(W) &:= \mathbb{E}[Y \mid W], \qquad & m_X(W) &:= \mathbb{E}[X \mid W], \\ q_Y(Z, W) &:= \mathbb{E}[Y \mid Z, W], \qquad & g_X(Z, W) &:= \mathbb{E}[X \mid Z, W]. \end{aligned}$$

Note that $\mathbb{E}[g_X \mid W] = m_X$, and the residualized signal is $\tilde{g} = g_X - m_X$. The estimand (4) can be written as

$$\beta_0 \;:=\; \beta_{IV}(g_X) \;=\; \frac{\mathbb{E}\big[(Y - l_Y)(g_X - m_X)\big]}{\mathbb{E}\big[(X - m_X)(g_X - m_X)\big]} \;=\; \frac{\mathbb{E}\big[(Y - l_Y)(g_X - m_X)\big]}{\mathbb{E}\big[(g_X - m_X)^2\big]}, \tag{9}$$

where the last equality uses $\mathbb{E}[(X - m_X)(g_X - m_X)] = \mathbb{E}[(g_X - m_X)^2]$, which holds because $\mathbb{E}[X - m_X \mid Z, W] = g_X - m_X$. As a specialization of β_{IV}, the estimand β_0 inherits the SWATE interpretation of Theorem 1. It is the signal-weighted average of the heterogeneous effects Δ, with the signal taken to be the optimal first stage g_X.

Three properties recommend g_X as the signal. First, it is the maximal-relevance choice as the projection maximizes covariance with $\tilde{X}$. Second, β_{IV} is invariant to rescaling and to W-measurable recalibration of the signal (Lemma 1). Third, g_X is the classical optimal instrument under homoskedastic homogeneity (Chamberlain, 1987; Newey, 1990).

Definition 1 (Heterogeneity along the Signal)**.** *The structural effects are correlated with selection in a manner that the instrument shifts,*

$$\mathbb{P}\Big(\mathbb{E}[X\Delta \mid Z, W] - \mathbb{E}[X\Delta \mid W] \;\neq\; \beta_0\,(g_X - m_X)\Big) > 0.$$

To read this definition, define the heterogeneity residual

$$\mathcal{H}(Z,W) := \big(\mathbb{E}[X\Delta \mid Z,W] - \mathbb{E}[X\Delta \mid W]\big) - \beta_0\,(g_X - m_X). \tag{10}$$

Definition 1 is exactly the statement $\mathbb{P}(\mathcal{H} \neq 0) > 0$. When effects are constant, i.e., $\Delta = \beta_0$, one has $\mathbb{E}[X\Delta \mid Z,W] = \beta_0\, g_X$ and $\mathbb{E}[X\Delta \mid W] = \beta_0\, m_X$, so $\mathcal{H} = 0$. Under essential heterogeneity, $\mathcal{H}$ is the component of the instrument's effect on $\mathbb{E}[X\Delta \mid Z,W]$ that the linear signal $\beta_0(g_X - m_X)$ fails to absorb. We therefore read Definition 1 throughout as "heterogeneity along the signal."

The four nuisances η are unknown and are estimated from the sample by K-fold cross-fitting (Chernozhukov et al., 2018). Partition the observations into $K \geq 2$ folds $\mathcal{I}_1, \ldots, \mathcal{I}_K$ of equal size $n_k = N/K$, with K fixed as $N \to \infty$, and write $\mathcal{D}_{-k}$ for the complementary subsample of observations outside $\mathcal{I}_k$. On each fold we fit the nuisance vector $\hat{\eta}_k = (\hat{l}_{Y,k}, \hat{m}_{X,k}, \hat{q}_{Y,k}, \hat{g}_{X,k})$ using only $\mathcal{D}_{-k}$, by any machine learner (such as lasso, random forests, gradient boosting, or an ensemble), and evaluate it on the held-out fold $\mathcal{I}_k$. Its j-th coordinate is written $\hat{\eta}_{j,k}$. Because $\hat{\eta}_k$ is trained without the observations at which it is used, it is independent of them and may be held fixed conditional on $\mathcal{D}_{-k}$, so overfitting does not enter the moment. When no fold index is displayed, $\hat{\eta}$ refers to this out-of-fold estimator, and fold by fold nuisance error is denoted as $\Delta\eta_k := \hat{\eta}_k - \eta_0$. The realized precision of the first stage is measured by the fold-wise L_2 errors

$$\rho_{j,k} := \|\hat{\eta}_{j,k} - \eta_{0,j}\|_{\mathbb{P},2} \quad (j \in \{l_Y, m_X, q_Y, g_X\}), \qquad \rho_N := \max_{j,k} \rho_{j,k}.$$

The next two assumptions place regularity and rate conditions on this construction.

Assumption 7 (Boundedness and Moments)**.** *(i)* $\mathbb{E}[Y^4] + \mathbb{E}[X^4] \leq C_4 < \infty$ *and* $\mathbb{E}[\Delta^2] + \mathbb{E}[X^2\Delta^2] < \infty$. *(ii) There is* $\bar{C} < \infty$ *such that* $\max_j \|\eta_{0,j}\|_\infty \leq \bar{C}$ *and, a.s.,* $\max_j \|\hat{\eta}_{j,k}\|_\infty \leq \bar{C}$ *for every fold* k.

Part (ii) is innocuous. Truncating each fitted nuisance at $\bar{C} \geq \max_j \|\eta_{0,j}\|_\infty$ preserves cross-fitting and weakly reduces every L_2 error, so it can always be imposed by the analyst.

Assumption 8 (Product Rate Conditions)**.** *(i)* $\rho_N = o_{\mathbb{P}}(1)$*; and (ii)* $\max_k \big\{\rho_{m,k}^2 + \rho_{g,k}^2\big\} = o_{\mathbb{P}}(N^{-1/2})$ *and* $\max_k \big(\rho_{l,k} + \rho_{q,k}\big)\big(\rho_{m,k} + \rho_{g,k}\big) = o_{\mathbb{P}}(N^{-1/2})$.

Assumption 8(ii) holds in particular under the single-rate $\|\hat{\eta}_{j,k} - \eta_{0,j}\|_{\mathbb{P},2} = o_{\mathbb{P}}(N^{-1/4})$ for each nuisance, which is attainable for lasso, random forests, and gradient boosting under

standard sparsity or smoothness. The product form relaxes this single-rate condition. The outcome-side projections l_Y, q_Y may converge arbitrarily slowly provided the treatment-side projections m_X, g_X compensate, which is valuable because $q_Y = \mathbb{E}[Y \mid Z, W]$ is typically the hardest nuisance to learn.

The standard estimator in DML routines such as `DoubleML` package in `R` and `Python` sets the empirical average of the canonical partialled-out moment

$$\psi_{\text{naive}}(O; \beta, \eta) := \big(Y - l_Y - \beta(X - m_X)\big)(g_X - m_X) \tag{11}$$

to zero. This moment identifies β_0, since its expectation vanishes at (β_0, η_0) by construction, but it is not robust to estimation error in the signal g_X.

With a fixed instrument and only the covariate nuisances (l_Y, m_X) estimated, the partially linear IV score of Chernozhukov et al. (2018) is Neyman-orthogonal irrespective of heterogeneity. The obstruction below is specific to the learned-signal variant, where $g_X = \mathbb{E}[X \mid Z, W]$ is itself estimated, and it arises only under heterogeneity along the signal in Definition 1.

Proposition 3 (Non-Orthogonality of the Naive Score)**.** *Under Assumption 1, view the naive moment as a functional of the signal, $g \mapsto \mathbb{E}[\psi_{\text{naive}}(O; \beta_0, (l_Y, m_X, q_Y, g))]$. Its pathwise (Gâteaux) derivative at $g = g_X$, in any direction $\delta_g \in L_2(\sigma(Z, W))$, is*

$$\begin{aligned}\partial_{g_X} \mathbb{E}\big[\psi_{\text{naive}}(O; \beta_0, \eta_0)\big][\delta_g] &:= \frac{d}{dt}\Big|_{t=0} \mathbb{E}\big[\psi_{\text{naive}}\big(O; \beta_0, (l_Y, m_X, q_Y, g_X + t\,\delta_g)\big)\big] \\ &= \mathbb{E}\big[\mathcal{H}(Z, W)\,\delta_g(Z, W)\big].\end{aligned}$$

Consequently, under essential heterogeneity, the score is not Neyman-orthogonal with respect to g_X. Taking $\delta_g = \mathcal{H}$ then gives $\mathbb{E}[\mathcal{H}^2] > 0$.

A black-box first-stage learner $\hat{g}_X$ converges at a non-parametric rate slower than $N^{-1/2}$, and Proposition 3 shows its error enters the moment at first order. The first-order influence of the estimated signal is $\mathbb{E}[\mathcal{H}\,\Delta g_X]$. Because it is a covariance, only the component of the first-stage error aligned with $\mathcal{H}$ can move the estimate. In the following section, we will show that this first-order consequence for the estimator is a drift of the estimand rather than noise.

3.2 What Naive DML Estimates

This section characterizes the first-order bias identified in Proposition 3. We first describe the geometry of $\mathcal{H}$, which governs the first-stage errors that the naive moment can and cannot detect. We then show that the bias is a drift of the estimand (Theorem 2), so that the naive plug-in performs valid inference on a moving target. Throughout, each perturbation direction inherits the measurability of the nuisance it perturbs: W-measurable for l_Y and m_X, and (Z, W)-measurable for q_Y and g_X. We write $\|\cdot\| := \|\cdot\|_{\mathbb{P},2}$.

Lemma 1 (Orthogonal Structure of $\mathcal{H}$)**.** *Let Assumptions 1 and 2 hold. With $g = g_X$,*

(i) $\mathbb{E}[\mathcal{H} \mid W] = 0$ *a.s.;*

(ii) $\mathbb{E}[\mathcal{H}\,\tilde{g}_X] = 0$*;*

(iii) the set $\mathcal{G} := \{\, c\,\tilde{g}_X + d \;:\; c \in \mathbb{R},\ d \in L_2(\sigma(W)) \,\} \subset L_2(\sigma(Z, W))$ *is a closed linear subspace,* $\mathcal{H} \perp \mathcal{G}$*, and for every direction* $\delta_g \in L_2(\sigma(Z, W))$*,*

$$\mathbb{E}[\mathcal{H}\,\delta_g] \;=\; \mathbb{E}\big[\mathcal{H}\,(I - \Pi_{\mathcal{G}})\delta_g\big], \qquad \big|\mathbb{E}[\mathcal{H}\,\delta_g]\big| \;\le\; \|\mathcal{H}\| \cdot \big\|(I - \Pi_{\mathcal{G}})\delta_g\big\|,$$

where $\Pi_{\mathcal{G}}$ *denotes orthogonal projection onto* $\mathcal{G}$*;*

(iv) for any $g \in L_2(\sigma(Z, W))$ *with* $\mathbb{E}[\tilde{X}\tilde{g}] \neq 0$*, every* $a \in \mathbb{R} \setminus \{0\}$ *and every* $d \in L_2(\sigma(W))$*, the signal* $g' := a\,g + d$ *satisfies* $\mathbb{E}[\tilde{X}\tilde{g}'] = a\,\mathbb{E}[\tilde{X}\tilde{g}] \neq 0$ *and* $\beta_{IV}(g') = \beta_{IV}(g)$*. The SWATE estimand is invariant to rescalings of the signal and to arbitrary W-measurable recalibrations.*

Part (iii) shows the operator of Proposition 3 eliminates every perturbation in $\mathcal{G}$, a rescaled signal plus a W-measurable shift, leaving only the residual shape component $(I - \Pi_{\mathcal{G}})\delta_g$ to move the naive moment at first order. The obstruction is broader than essential heterogeneity. Decomposing $\mathcal{H} = (\bar{\Delta}_W - \beta_0)\tilde{g}_X + \tilde{\kappa}$ into an observable-heterogeneity term ($\bar{\Delta}_W := \mathbb{E}[\Delta \mid W]$) and a selection-on-gains term ($\tilde{\kappa} := \kappa - \mathbb{E}[\kappa \mid W]$, $\kappa := \mathrm{Cov}(X, \Delta \mid Z, W)$), Lemma A3 in Appendix A shows observable heterogeneity alone makes $\mathcal{H} \neq 0$.

We now formalize the drifting estimand of naive DML. We write the naive moment as

$$\mathbb{E}\big[\psi_{\text{naive}}(O; \beta_0, \eta)\big] \;=\; \mathbb{E}[\mathcal{H}\,\Delta g_X] \;+\; \mathbb{E}\big[(-\Delta l_Y + \beta_0\,\Delta m_X)(\Delta g_X - \Delta m_X)\big]. \tag{12}$$

The leading term is the drift loading, and the second is a product of nuisance errors, hence higher order.

Define $A_k := \mathbb{E}\big[(Y - \hat{l}_{Y,k})(\hat{g}_{X,k} - \hat{m}_{X,k}) \mid \mathcal{D}_{-k}\big]$ and $B_k := \mathbb{E}\big[(X - \hat{m}_{X,k})(\hat{g}_{X,k} - \hat{m}_{X,k}) \mid \mathcal{D}_{-k}\big]$,

$$\beta_k^* := \frac{A_k}{B_k}, \qquad \beta_N^* := \frac{\sum_{k=1}^K n_k A_k}{\sum_{k=1}^K n_k B_k} \;=\; \sum_{k=1}^K w_k\, \beta_k^*, \qquad w_k := \frac{n_k B_k}{\sum_{j=1}^K n_j B_j}. \tag{13}$$

Let $\bar{D}_N := \sum_{k=1}^K (n_k/N)\, \mathbb{E}\big[\mathcal{H}\, \Delta g_{X,k} \mid \mathcal{D}_{-k}\big]$ denote the average realized drift loading, where $\Delta g_{X,k} := \hat{g}_{X,k} - g_X$.

Theorem 2 (The Moving Target: Learner-Induced SWATE)**.** *Let Assumptions 1 and 2 hold with $g = g_X$. For parts (iii)–(iv) let Assumptions 7 and 8 also hold and let $\hat{\beta}_{\text{naive}}$ solve the cross-fitted empirical analogue of (11). Then*

(i) for every $\delta_g \in L_2(\sigma(Z, W))$ with $J_0 + \mathbb{E}[\tilde{g}_X\, \delta_g] \neq 0$,

$$\beta_{IV}(g_X + \delta_g) - \beta_0 \;=\; \frac{\mathbb{E}[\mathcal{H}\, \delta_g]}{J_0 + \mathbb{E}[\tilde{g}_X\, \delta_g]}; \tag{14}$$

(ii) β_{IV} is Fréchet differentiable at g_X with derivative $\delta_g \mapsto J_0^{-1}\, \mathbb{E}[\mathcal{H}\, \delta_g]$ and a second-order remainder of order $\|\delta_g\|^2$. Combining with Lemma 1(iii), $\beta_{IV}(g_X + \delta_g) - \beta_0 = \mathbb{E}[\mathcal{H}\, (I - \Pi_{\mathcal{G}})\delta_g]/(J_0 + \mathbb{E}[\tilde{g}_X\, \delta_g])$, so the numerator and hence the Fréchet derivative depend only on the residual shape $(I - \Pi_{\mathcal{G}})\delta_g$. The drift also depends on the component parallel to $\tilde{g}_X$ through the denominator, and it is zero for every $\delta_g \in \mathcal{G}$ for which the perturbed signal remains relevant;

(iii) $|\bar{D}_N| \leq \|\mathcal{H}\|\, \max_k \rho_{g,k} = o_{\mathbb{P}}(N^{-1/4})$, and

$$\beta_N^* \;=\; \beta_0 + J_0^{-1}\, \bar{D}_N + o_{\mathbb{P}}(N^{-1/2}), \qquad \beta_N^* \;=\; \sum_{k=1}^K w_k\, \beta_{IV}\big(\hat{g}_{X,k}\big) + o_{\mathbb{P}}(N^{-1/2}),$$

the naive procedure targets the B_k-weighted SWATE of the learner's own signals;

(iv) with $U := Y - l_Y - \beta_0(X - m_X)$, $V := g_X - m_X$, and $\Omega_0^{\text{nv}} := \mathbb{E}[U^2 V^2]$, suppose $\Omega_0^{\text{nv}} > 0$. Then

$$\sqrt{N}\,\big(\hat{\beta}_{\text{naive}} - \beta_N^*\big) \;\xrightarrow{d}\; \mathcal{N}\big(0,\; J_0^{-2}\, \Omega_0^{\text{nv}}\big),$$

The asymptotic variance is consistently estimable by plug-ins $\hat{J}_{\text{nv}} \xrightarrow{\mathbb{P}} J_0$ and $\hat{\Omega}^{\text{nv}} \xrightarrow{\mathbb{P}} \Omega_0^{\text{nv}}$, so the nominal Wald interval satisfies $\mathbb{P}\big(\beta_N^* \in [\hat{\beta}_{\text{naive}} \pm z_{1-\alpha/2}\, \hat{J}_{\text{nv}}^{-1} \sqrt{\hat{\Omega}^{\text{nv}}/N}]\big) \to 1 - \alpha$.

The Fréchet derivative in (ii) equals the Gâteaux derivative of the naive moment divided by J_0, so the "regularization bias" is, to first order, an estimand drift. The naive plug-in tracks $\beta_{IV}(\hat{g}_X)$, the SWATE of the learner's own signal, itself legitimate and convex only if CCM holds for $\hat{g}_X$. Combining parts (iii) and (iv), the naive DML estimator has the first-order expansion

$$\hat{\beta}_{\text{naive}} - \beta_0 = \frac{1}{N}\sum_{i=1}^{N} J_0^{-1}\,\psi_{\text{naive}}(O_i;\beta_0,\eta_0) \;+\; J_0^{-1}\,\mathbb{E}\big[\mathcal{H}(Z,W)\,\Delta g_X(Z,W)\big] \;+\; o_{\mathbb{P}}\big(N^{-1/2}\big),$$

whose regularization bias term $J_0^{-1}\mathbb{E}[\mathcal{H}\,\Delta g_X]$ is $o_{\mathbb{P}}(N^{-1/4})$ but not, in general, $o_{\mathbb{P}}(N^{-1/2})$. Hence $\sqrt{N}(\hat{\beta}_{\text{naive}} - \beta_0)$ generally has no mean-zero limit and conventional intervals for β_0 are invalid.

Corollary 2 (Coverage of the Fixed Target)**.** *In the setting of Theorem 2, let* $\sigma_{\text{nv}}^2 := J_0^{-2}\Omega_0^{\text{nv}}$, $\mu_N := \sqrt{N}\,J_0^{-1}\bar{D}_N/\sigma_{\text{nv}}$, *and let* CI_N *denote the nominal* $(1-\alpha)$ *naive interval of part (iv). Then*

(a) if $\mu_N \xrightarrow{\mathbb{P}} 0$*:* $\mathbb{P}(\beta_0 \in \text{CI}_N) \to 1-\alpha$*;*

(b) if $\mu_N \xrightarrow{\mathbb{P}} \mu \neq 0$ *finite:* $\mathbb{P}(\beta_0 \in \text{CI}_N) \to \Phi(z_{1-\alpha/2}-\mu) - \Phi(-z_{1-\alpha/2}-\mu) < 1-\alpha$*;*

(c) if $|\mu_N| \xrightarrow{\mathbb{P}} \infty$*:* $\mathbb{P}(\beta_0 \in \text{CI}_N) \to 0$.

In particular, $\sqrt{N}(\hat{\beta}_{\text{naive}} - \beta_0)$ *is asymptotically centered and the naive interval is asymptotically valid for* β_0 *if and only if* $\sqrt{N}\,\bar{D}_N \xrightarrow{\mathbb{P}} 0$*, or equivalently, the realized* $\mathbb{E}[\mathcal{H}\,\Delta g_X]$ *is* $o_{\mathbb{P}}(N^{-1/2})$.

Remark 3. *The leading systematic error of a well-tuned regularized learner (e.g., ridge, lasso, boosting, depth-limited forests) is attenuation plus recalibration,* $\hat{g}_X \approx (1-s_N)\,g_X + d_N(W)$*, which, provided the rescaled signal remains relevant, lies in* $\mathcal{G}$ *and hence moves neither the naive moment (Lemma 1) nor the estimand (Lemma 1(iv), Theorem 2(ii)), exactly rather than only to first order. This is why a well-tuned first stage often leaves the naive estimator approximately valid even when* $\mathcal{H} \neq 0$*. The phenomenon cannot be relied upon, however, since misspecified forms, coarse bases, sparsity rules that drop curvature, and fit quality that varies across the strata where* $\bar{\Delta}_W$ *varies (Lemma A3) all inject shape error, a distortion of how* g_X *varies with* Z *within a stratum rather than merely its scale, aligned with* $\mathcal{H}$*. It also does not license a pre-test. Conditioning the reported inference on a test of* $\sqrt{N}\,\bar{D}_N \xrightarrow{\mathbb{P}} 0$ *re-introduces the size distortion the test was meant to avoid and forfeits*

any uniform coverage guarantee (Andrews et al., 2019). We therefore report the CRC-robust estimator as the default and the diagnostic we develop in Section 5 should be used as an interpretive statistic, instead of a model selection tool.

4 The Fixed Target: CRC-Robust Score

We now build the score whose target is the fixed, learner-invariant β_0, which is an orthogonalized moment that eliminates the first-order influence of the learned signal, restores $\sqrt{N}$ inference, attains the efficiency bound, and costs nothing under homogeneity.

4.1 The CRC-Robust Orthogonal Score

Neyman-orthogonality is restored by augmenting the numerator and denominator moments with correction terms that eliminate the first-order influence of every nuisance. Define

$$\psi_Y(O;\eta) := (Y - l_Y)(g_X - m_X) + (X - g_X)(q_Y - l_Y), \tag{15}$$

$$\psi_X(O;\eta) := (X - m_X)(g_X - m_X) + (X - g_X)(g_X - m_X), \tag{16}$$

and the CRC-robust score

$$\psi_{\mathrm{CRC}}(O;\beta,\eta) := \psi_Y(O;\eta) - \beta\,\psi_X(O;\eta). \tag{17}$$

Both augmentation terms carry the first-stage residual $(X - g_X)$, whose conditional mean given (Z, W) is zero. They therefore leave the estimand unchanged, so that $\mathbb{E}[\psi_{\mathrm{CRC}}(O;\beta_0,\eta_0)] = 0$ still identifies β_0, while cancelling the pathwise derivative of the moment with respect to the signal g_X and the new outcome projection q_Y.

Lemma 2 (Score Difference Identity)**.** *For every $\beta \in \mathbb{R}$ and every admissible nuisance value $\eta = (l, m, q, g)$,*

$$\psi_{\mathrm{CRC}}(O;\beta,\eta) - \psi_{\mathrm{naive}}(O;\beta,\eta) \;=\; (X - g)\,\big[(q - l) - \beta\,(g - m)\big].$$

At (β_0, η_0) the correction equals $(X - g_X)\,\mathcal{H}(Z, W)$, and, for every β,

$$\mathbb{E}\Big[(X - g_X)\big\{(q_Y - l_Y) - \beta\,(g_X - m_X)\big\}\Big] \;=\; 0.$$

Moreover $(X - g_X)\,\mathcal{H} = 0$ *a.s. if and only if* $\sigma_X^2(Z,W)\,\mathcal{H}^2 = 0$ *a.s., where* $\sigma_X^2(Z,W) := \mathrm{Var}(X \mid Z, W)$.

The orthogonalization thus re-injects the first-stage residual weighted by the unabsorbed heterogeneity. Three consequences follow from the identity. First, the correction is mean zero at every β, so both scores identify β_0. Second, when $\sigma_X^2 > 0$ it vanishes when the essential heterogeneity of Definition 1 fails, which explains why the naive score is orthogonal under homogeneity. Third, the correction is the influence contribution of the estimated signal that the naive sandwich variance omits (Proposition A1 in Appendix A).

4.2 Asymptotic Distribution and Efficiency

Write $\beta_0 = \theta_Y / J_0$ with $\theta_Y = \mathbb{E}[(Y - l_Y)(g_X - m_X)]$ and $J_0 = \mathbb{E}[(g_X - m_X)^2]$. The CRC-robust score restores $\sqrt{N}$ inference on the fixed target, attains the efficiency bound, and reduces to the naive score under homogeneity. Orthogonality, together with cross-fitting at the rates of Assumption 8, reduces the impact of nuisance estimation to an asymptotically negligible second-order remainder.

Theorem 3 (Asymptotic Distribution and Efficiency)**.** *Let Assumptions 1, 2, 7, and 8 hold, and let* $\hat{\beta}_{\mathrm{CRC}}$ *be the root of the* K*-fold cross-fitted score equation*

$$\frac{1}{N}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k}\psi_{\mathrm{CRC}}\big(O_i;\hat{\beta},\hat{\eta}_k\big) = 0,$$

where $\hat{\eta}_k$ *is trained on the observations outside fold* $\mathcal{I}_k$*. Then*

(i) $\mathbb{E}[\psi_{\mathrm{CRC}}(O;\beta_0,\eta_0)] = 0$*, and it is Neyman-orthogonal with respect to the nuisance vector* $\eta = (l_Y, m_X, q_Y, g_X)$ *at* η_0*, i.e.,* $\partial_\eta\,\mathbb{E}[\psi_{\mathrm{CRC}}(O;\beta_0,\eta_0)][\eta - \eta_0] = 0$*;*

(ii) $\sqrt{N}\,(\hat{\beta}_{\mathrm{CRC}} - \beta_0) \xrightarrow{d} \mathcal{N}\big(0,\ J_0^{-2}\,\Omega_0\big)$*, where* $\Omega_0 = \mathbb{E}\big[\psi_{\mathrm{CRC}}(O;\beta_0,\eta_0)^2\big]$ *with the asymptotic variance consistently estimated by* $\hat{J}^{-2}\hat{\Omega}$*, where* $\hat{J} = \frac{1}{N}\sum_i(\hat{g}_{X,i} - \hat{m}_{X,i})^2$ *and* $\hat{\Omega} = \frac{1}{N}\sum_i \psi_{\mathrm{CRC}}(O_i;\hat{\beta},\hat{\eta})^2$*;*

(iii) the Efficient Influence Function (EIF) of β_0 *is*

$$\phi_{\beta_0}(O) \;=\; J_0^{-1}\,\psi_{\mathrm{CRC}}(O;\beta_0,\eta_0) \;=\; J_0^{-1}\big(UV + e\,\mathcal{H}\big),$$

with $U = Y - l_Y - \beta_0(X - m_X)$, $V = g_X - m_X$, $e = X - g_X$. *Consequently the semiparametric efficiency bound for* β_0 *is* $J_0^{-2}\,\Omega_0$, *which is attained by* $\hat{\beta}_{\mathrm{CRC}}$;

(iv) if effects are homogeneous, $\psi_{\mathrm{CRC}}(\cdot;\beta_0,\eta_0) = \psi_{\mathrm{naive}}(\cdot;\beta_0,\eta_0) = UV$ *and* $\Omega_0 = \Omega_0^{\mathrm{nv}}$.

The estimation of $\hat{\beta}_{\mathrm{CRC}}$ is the standard DML problem for a user-defined linear score, which can be directly handled by the `DoubleML` libraries in R and Python.[1] The standard library routine returns the same estimator and standard error of Theorem 3.

4.3 Worst-case bias over a realization ball

DML's validity is uniform, and it must hold across the whole ball of nuisance realizations a learner of a given accuracy could return, not just at one. Cross-fitting controls the magnitude of the first-stage error but not its direction, and the practitioner never observes which realization the learner draws, so a guarantee is only as good as its worst point in the ball. Coverage that holds on average, or at a favorable realization, need not extend to the least favorable one. Lemma A4 makes the comparison of the two scores over such sets exact.

Proposition 4 (Worst-case first-order bias over a realization ball)**.** *Let Assumptions 1 and 2 and the heterogeneity condition* $\mathbb{E}[\mathcal{H}^2] > 0$ *(Definition 1) hold, fix radii* $r = (r_l, r_m, r_q, r_g) \in (0,1]^4$, *fix* $\bar{C} \geq \max_j \|\eta_{0,j}\|_\infty + \|\mathcal{H}\|_\infty / \|\mathcal{H}\|_{\mathbb{P},2}$, *and define the realization set*

$$\mathcal{T}(r) := \Big\{\eta \text{ admissible} : \|\eta_j\|_\infty \leq \bar{C},\ \ \|\eta_j - \eta_{0,j}\|_{\mathbb{P},2} \leq r_j,\ j \in \{l_Y, m_X, q_Y, g_X\}\Big\}.$$

Then

(i) $r_g\,\|\mathcal{H}\| \ \leq \ \sup_{\eta\in\mathcal{T}(r)} \big|\mathbb{E}[\psi_{\mathrm{naive}}(O;\beta_0,\eta)]\big| \ \leq \ r_g\,\|\mathcal{H}\| + (r_l + |\beta_0|\,r_m)(r_g + r_m)$, *with the lower*

[1] Here we provide code for using CRC-robust score in R and Python:

```
## For R: nuisances lY=E[Y|W], mX=E[X|W], qY=E[Y|Z,W], gX=E[X|Z,W]  (cross-fitted)
psi_X <- (df$x-mX)*(gX-mX) + (df$x-gX)*(gX-mX)       # denominator moment
psi_Y <- (df$y-lY)*(gX-mX) + (df$x-gX)*(qY-lY)       # numerator moment
crc <- function(y,z,d,l_hat,m_hat,r_hat,g_hat,smpls) list(psi_a=-psi_X, psi_b=psi_Y)
```

```
## For Python: nuisances l=lY=E[Y|W], m=gX=E[X|Z,W], r=mX=E[X|W], g=qY=E[Y|Z,W] (cross-fitted)
def crc(y, z, d, l, m, r, g, smpls):
    lY, gX, mX, qY = l, m, r, g
    psiX = (d-mX)*(gX-mX) + (d-gX)*(gX-mX)            # denominator moment
    psiY = (y-lY)*(gX-mX) + (d-gX)*(qY-lY)            # numerator moment
    return -psiX, psiY                                 # (psi_a, psi_b); pass score=crc
```

bound attained at $\eta = \big(l_Y,\, m_X,\, q_Y,\, g_X + r_g\,\mathcal{H}/\|\mathcal{H}\|\big)$;

(ii) $\sup_{\eta\in\mathcal{T}(r)} \big|\mathbb{E}[\psi_{\mathrm{CRC}}(O;\beta_0,\eta)]\big| \;\le\; r_l r_m + r_g r_q + |\beta_0|\big(r_m^2 + r_g^2\big)$.

Hence over any realization set of nonparametric radius ($\sqrt{N}\,r_g \to \infty$)*, the naive moment's worst-case first-order bias is of the order* r_g*, while that of the CRC-robust moment is of the order of squared radii. Uniform validity over* $\mathcal{T}(r_N)$ *holds for the CRC-robust score under Assumption 8 and fails for the naive score whenever* $\mathcal{H} \neq 0$.

Proposition 4 is more than a bound. An explicit deterministic sequence $\hat{g}_X^{(N)} = g_X + N^{-\gamma}\mathcal{H}$ with $\gamma \in (1/4, 1/2)$ drives naive coverage of β_0 to zero while the CRC-robust interval stays nominal (Corollary A1, Appendix A).

5 Diagnostics and inference under a weak residual signal

There are two remaining tasks for practice. First, telling whether the naive number is on target and inferring β_0 when partialling out W leaves little residual signal. This section supplies an alignment diagnostic and gives an identification-robust confidence set that stays valid when the instruments are weak.

5.1 The Alignment Diagnostic

Corollary 2 makes the practitioner's question precise. Is the realized drift negligible, or is the naive estimator answering the question about β_0? Lemma 2 supplies a test. With $\hat{\mathcal{H}}_k(Z,W) := (\hat{q}_{Y,k} - \hat{l}_{Y,k}) - \hat{\beta}_{\mathrm{CRC}}\,(\hat{g}_{X,k} - \hat{m}_{X,k})$, define

$$\widehat{\mathcal{A}}_N := \frac{1}{N}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k}\big(X_i - \hat{g}_{X,k,i}\big)\,\hat{\mathcal{H}}_{k,i}, \quad \widehat{\Xi}_N := \frac{1}{N}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k}\big(X_i - \hat{g}_{X,k,i}\big)^2\,\hat{\mathcal{H}}_{k,i}^2, \quad T_N := \frac{\sqrt{N}\,\widehat{\mathcal{A}}_N}{\widehat{\Xi}_N^{1/2}}.$$

Proposition 5 (Alignment Diagnostic)**.** *Let Assumptions 1, 2, 7, and 8 hold, and suppose* $\Xi_0 := \mathbb{E}\big[(X - g_X)^2\,\mathcal{H}^2\big] > 0$. *Then*

(i) $\sqrt{N}\,\widehat{\mathcal{A}}_N = \dfrac{1}{\sqrt{N}}\sum_{i=1}^{N}(X_i - g_{X,i})\,\mathcal{H}_i \;-\; \sqrt{N}\,\bar{D}_N \;+\; o_{\mathbb{P}}(1)$, *and* $\widehat{\Xi}_N \xrightarrow{\mathbb{P}} \Xi_0$;

(ii) *under* $H_0 : \sqrt{N}\,\bar{D}_N \xrightarrow{\mathbb{P}} 0$ *and* $T_N \xrightarrow{d} \mathcal{N}(0,1)$, *so the test rejecting when* $|T_N| > z_{1-\alpha/2}$ *has asymptotic size* α;

(iii) *if* $\sqrt{N}\,|\bar{D}_N| \xrightarrow{\mathbb{P}} \infty$, *then* $|T_N| \xrightarrow{\mathbb{P}} \infty$, *and the test is consistent;*

(iv) $\sqrt{N}\,\widehat{\mathcal{A}}_N \;=\; J_0\,\sqrt{N}\,\big(\hat{\beta}_{\mathrm{CRC}} - \hat{\beta}_{\mathrm{naive}}\big) + o_{\mathbb{P}}(1)$, *so* T_N *is asymptotically the studentized contrast of the two estimators.*

A rejection says that the learner's first-stage error is aligned with $\mathcal{H}$, so the naive number answers a different, learner-dependent question. Because the CRC-robust estimator is valid regardless, the test has power against the naive estimator's failure mode, in the specification test pattern of Hausman (1978). Note that it does not test $\mathcal{H} = 0$. Unlike the CRC tests of Heckman et al. (2010), T_N tests the realized alignment of the first-stage error with $\mathcal{H}$, while $\widehat{\Xi}_N$ points to detectable heterogeneity along the signal. We recommend reporting the pair $(T_N, \widehat{\Xi}_N)$. As mentioned in the pre-testing discussion of Section 3.2, the pair is an interpretive report rather than a selector between estimators.

Remark 4. *If* $\mathcal{H} = 0$ *then* $\Xi_0 = 0$, $\sqrt{N}\,\widehat{\mathcal{A}}_N = o_{\mathbb{P}}(1)$, *and indeed* $\hat{\beta}_{\mathrm{naive}} - \hat{\beta}_{\mathrm{CRC}} = o_{\mathbb{P}}(N^{-1/2})$, *so the two estimators are first-order equivalent and no correction is needed. We therefore recommend reporting* $\widehat{\Xi}_N$ *alongside* T_N, *since a negligible* $\widehat{\Xi}_N$ *indicates no detectable heterogeneity along the signal. Likewise, first stages that "self-correct" in the sense of Assumption A1 (Appendix A) with* $a = \mathcal{H}$ *drive the non-centrality of* T_N *to zero (Proposition A1(iii)), since for them the naive point estimator is first-order equivalent to the CRC-robust one. What such first stages do not repair is the reported standard error.*

5.2 Identification-Robust Sets

A distinct failure concerns the standard error rather than the point estimate. Even when the naive point estimator is centered at β_0, its plug-in sandwich variance, which treats $\hat{g}_X$ as known, is generally inconsistent, because the drift loading carries a mean-zero $O_{\mathbb{P}}(N^{-1/2})$ component inherited from the training folds. Under a first-stage asymptotic linearity condition, the naive estimator is $\sqrt{N}$-normal around β_0 with variance $\Omega_S = \Omega_0^{\mathrm{nv}} + 2\,\mathbb{E}[UVea] + \mathbb{E}[e^2a^2]$, of which the sandwich captures only Ω_0^{nv} (Proposition A1). A least-squares first

stage whose dictionary spans g_X and $\mathcal{H}$ is the leading case, satisfying Assumption A1 with $a = \mathcal{H}$ (Proposition A2). Such projection first stages correct the point estimate implicitly, in the tradition of Newey (1994), yet leave the standard error inconsistent. Outside the projection family (such as trees, boosting, heavily penalized fits) even the centering is unprotected. This is the semiparametric analogue of heterogeneity-invalid two-stage least squares standard errors (Kolesár, 2013; Evdokimov and Kolesár, 2018; Lee, 2018).

Finally, because ψ_{CRC} is affine in β, inference can avoid the Jacobian. This is valuable when partialling W out of a strongly W-predictable instrument leaves little residual signal ($J_0 \approx 0$), the regime in which the Wald interval of Theorem 3 degrades. Define

$$\bar{\psi}(\beta) := \frac{1}{N}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k}\psi_{\mathrm{CRC}}\big(O_i;\beta,\hat{\eta}_k\big), \qquad \hat{\sigma}^2(\beta) := \frac{1}{N}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k}\psi_{\mathrm{CRC}}\big(O_i;\beta,\hat{\eta}_k\big)^2, \tag{18}$$

$\mathrm{AR}_N(\beta) := \sqrt{N}\,\bar{\psi}(\beta)/\hat{\sigma}(\beta)$, and the confidence set $\mathcal{C}_{1-\alpha} := \{\beta \in \mathbb{R} : |\mathrm{AR}_N(\beta)| \le z_{1-\alpha/2}\}$.

Proposition 6 (Validity with Weak Signal)**.** *Let Assumptions 1, 2, 7, and 8 hold, and suppose* $\Omega_0 = \mathbb{E}\big[\psi_{\mathrm{CRC}}(O;\beta_0,\eta_0)^2\big] > 0$. *Then*

$$\mathbb{P}\big(\beta_0 \in \mathcal{C}_{1-\alpha}\big) \longrightarrow 1-\alpha.$$

Remark 5. *Writing* $\psi_{Y,i}, \psi_{X,i}$ *for the plug-in values of the moments* (15)–(16)*, let* $S_Y := \frac{1}{N}\sum_i \psi_{Y,i}$, $S_X := \frac{1}{N}\sum_i \psi_{X,i}$, $M_{YY} := \frac{1}{N}\sum_i \psi_{Y,i}^2$, $M_{XY} := \frac{1}{N}\sum_i \psi_{Y,i}\psi_{X,i}$, $M_{XX} := \frac{1}{N}\sum_i \psi_{X,i}^2$. *Since* $\bar{\psi}(\beta) = S_Y - \beta S_X$ *and* $\hat{\sigma}^2(\beta) = M_{YY} - 2\beta M_{XY} + \beta^2 M_{XX}$*, the set is the solution of a quadratic inequality,*

$$\mathcal{C}_{1-\alpha} = \big\{\beta : A\beta^2 - 2B\beta + C \le 0\big\},$$

with

$$A = NS_X^2 - z^2 M_{XX}, \quad B = NS_XS_Y - z^2 M_{XY}, \quad C = NS_Y^2 - z^2 M_{YY},$$

with $z = z_{1-\alpha/2}$*. The set is an interval when* $A > 0$*, and the complement of an interval or the whole line when* $A \le 0$ *(Anderson and Rubin, 1949; Fieller, 1954). Under strong identification,* J_0 *is bounded away from zero,* $\mathcal{C}_{1-\alpha}$ *is asymptotically equivalent to the Wald interval of Theorem 3.*

6 Monte Carlo Evidence

This section examines the finite-sample performance of the naive and CRC-robust estimators and of the accompanying diagnostics. The experiments have three aims: to show that the drift of Theorem 2 distorts naive inference at conventional sample sizes; to assess the coverage and efficiency of the CRC-robust estimator, including its behavior under homogeneity; and to show the finite-sample size and power of the diagnostics of Section 5 in designs where the target is known.

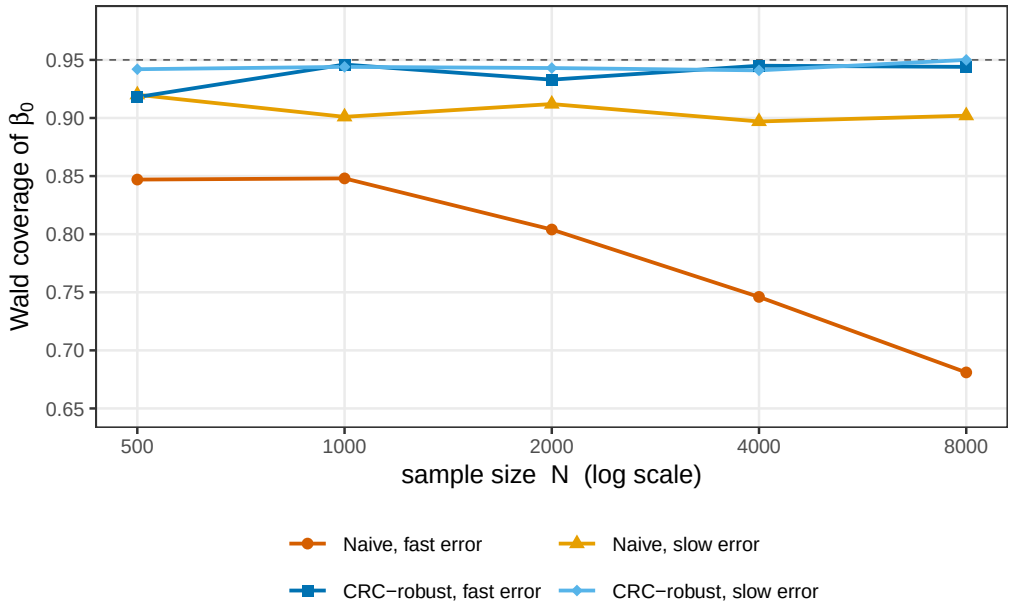
The data-generating process is a threshold-crossing CRC model, in which a binary treatment is selected on its own gain

$$\begin{aligned} \nu &= \alpha_0 + Z^\top \alpha_Z + W^\top \alpha_W, & X &= \mathbf{1}\{U \le \nu\}, \\ \Delta &= a_0 + a_W\,(W^\top \gamma) + b\,U + \sigma_\xi\,\xi, & Y(0) &= c_0 + W^\top \gamma_Y + \sigma_e\,e, \end{aligned} \tag{19}$$

with $Y = Y(0) + X\Delta$ as in equation (3). The shocks (U, ξ, e) are standard normal, $Z \sim \mathcal{N}(0, I_{p_Z})$, $W \sim \mathcal{N}(0, I_{p_W})$, and all five are mutually independent, so conditional exogeneity (Assumption 1) holds by construction. Because the selection shock U enters both the treatment equation and the return, the coefficient b measures essential heterogeneity in the sense of Definition 1, with $b = 0$ the homogeneous benchmark, and a_W measures heterogeneity in observables through $\mathbb{E}[\Delta \mid W]$. All four nuisance functions are available in closed form. In particular, $g_X = \mathbb{E}[X \mid Z, W] = \Phi(\nu)$.

We consider two main configurations. The first, the oracle design, has $p_Z = 50$ instruments whose first-stage coefficient vector α_Z is approximately sparse, $p_W = 10$ covariates, strong essential heterogeneity ($b = 2.4$), and no heterogeneity in observables ($a_W = 0$). Its nuisances are evaluated at their closed-form values rather than estimated, which separates the behavior of the estimators from first-stage estimation error. An adversarial variant perturbs the signal by an error of known size and direction. The second, the learnable design, is lower-dimensional and less noisy ($p_Z = 8$, $p_W = 4$, $a_W = 1.2$, $b = 0.6$), and its nuisances are cross-fitted with the lasso and gradient boosting. The configuration is chosen so that the product-rate condition of Assumption 8 is plausible for both learners. The two configurations imply the same target, $\beta_0 \approx 0.50$. For each sample size $N \in \{500, 1000, 2000, 4000, 8000\}$ we generate $R = 1000$ samples. Throughout, coverage refers to the frequency with which a nominal 95% confidence interval covers the fixed target β_0, and the interval is the Wald interval unless stated otherwise. Two auxiliary designs, constructed to evaluate the diagnostics

Figure 1: Coverage of the fixed target under a contaminated signal



Notes: The figure plots the empirical coverage of nominal 95% Wald intervals for the fixed target β_0 against the sample size N (log scale) in the oracle design with $b = 2.4$. The signal is contaminated along the heterogeneity residual, $\hat{g}_X = g_X + c_N \mathcal{H}$, with a fast error ($c_N = 2\,N^{-1/3}$) and a slow error ($c_N = 2\sqrt{\log p_Z / N}$). The naive interval realizes the two under-covering points of Corollary 2. The CRC-robust interval remains near the nominal level under both errors (Theorem 3).

of Section 5, are described together with the corresponding results below.

Figure 1 reports the behavior of naive inference when the first-stage error is aligned with the heterogeneity residual. Starting from the oracle design, we replace the signal with $\hat{g}_X = g_X + c_N \mathcal{H}$ under two schemes for c_N. The fast error, $c_N = 2\,N^{-1/3}$, is of the form in Corollary A1 and drives μ_N in Corollary 2 to infinity. The slow error, $c_N = 2\sqrt{\log p_Z / N}$, holds μ_N at a nonzero constant. The two schemes realize branches (c) and (b) of Corollary 2, respectively. Under the fast error, coverage of the naive Wald interval falls from 0.847 at $N = 500$ to 0.681 at $N = 8000$, the scaled bias $\sqrt{N}(\hat{\beta}_{\text{naive}} - \beta_0)$ grows from 4.69 to 7.08, and the alignment test of Section 5.1 rejects in every replication. This pattern is consistent with the drift of Theorem 2. Under the slow error, coverage is near 0.90 and does not improve as N grows. The CRC-robust interval has coverage between 0.918 and 0.950 in both cases (Theorem 3).

Table 1 shows the result of the CRC-robust estimator. In the first panel effects are homogeneous, so the naive and CRC-robust scores coincide and the correction has no efficiency loss.

Table 1: CRC-robust inference: no cost under homogeneity, efficiency under heterogeneity

		Sample size N				
Design	Statistic	500	1000	2000	4000	8000
Homogeneous, $b = 0$	Cov $\hat{\beta}_{\text{naive}}$	0.947	0.954	0.957	0.949	0.940
	Cov $\hat{\beta}_{\text{CRC}}$	0.944	0.957	0.958	0.949	0.939
Oracle, $b = 2.4$	$N\widehat{\text{Var}}$ ($\to 20.63$)	21.17	20.67	19.30	19.46	20.75
Learnable (cross-fitted)	Lasso $N\widehat{\text{Var}}$ ($\to 2.86$)	3.16	2.63	3.00	2.64	2.75
	Boosting $N\widehat{\text{Var}}$ ($\to 2.86$)	5.98	4.91	4.16	3.53	3.74
	Lasso Cov β_0	0.941	0.956	0.949	0.959	0.952
	Boosting Cov β_0	0.867	0.905	0.925	0.949	0.947

Notes: Coverage (Cov) is the frequency of a nominal 95% Wald interval covering the fixed target $\beta_0 \approx 0.50$. $N\widehat{\text{Var}}$ denotes $N\times$ the mean squared error of $\hat{\beta}_{\text{CRC}}$ about β_0 across replications. b indexes the strength of essential heterogeneity. In the oracle design the nuisances are evaluated at their closed-form values; in the learnable design they are cross-fitted with the lasso and with gradient boosting.

The simulated coverage rates of the two estimators are both close to the nominal rate. The second panel reports N times the mean squared error of $\hat{\beta}_{\text{CRC}}$ about β_0, denoted $N\widehat{\text{Var}}$, in the oracle design. The semiparametric bound $J_0^{-2}\Omega_0$ equals 20.63, and the simulated values lie between 19.30 and 21.17 across the sample-size grid, in line with the efficiency statement of Theorem 3(iii). The third panel repeats the exercise in the learnable design, where the bound equals 2.86 and the nuisances are cross-fitted. With lasso nuisances the scaled variance is near the bound from moderate N onward, between 2.63 and 3.16 across the grid, and coverage is close to nominal at every sample size, between 0.941 and 0.959. Because the design has a low-dimensional sparse linear index, the lasso recovers the nuisances at a near-parametric rate, so the second-order remainder that orthogonality leaves is already negligible at $N = 500$ and the Wald interval is both centered and correctly scaled. With gradient boosting the scaled variance approaches the bound from above, from 5.98 to 3.74, and coverage rises from 0.867 at $N = 500$ to 0.947 at $N = 8000$, the pattern implied by Theorem 3(iii) when a nuisance converges more slowly while Assumption 8 holds.

Table 2 shows the inference when partialling out W leaves little residual signal, a failure mode that does not operate through bias in the point estimate. We scale down J_0 from 0.075 to 0.0013. The identification-robust set of Section 5.2 has coverage 0.944 and 0.947 at

Table 2: Identification-robust inference under a weak residual signal

J_0	$\text{Cov}_{\text{AR/Fieller}}$	Cov_{Wald}	Frac. unbounded
0.075 (strong)	0.944	0.949	0.00
0.0013 (weak)	0.947	1.000	0.62

Notes: Scalar-instrument threshold model in which the instrument loading is scaled down so that the residual-signal strength J_0 falls. $\text{Cov}_{\text{AR/Fieller}}$ is the coverage of the identification-robust set of Proposition 6, Cov_{Wald} the coverage of the plug-in Wald interval, and Frac. unbounded the fraction of replications in which the identification-robust set is unbounded (Remark 5). Based on $R = 1000$ replications; seed 20260413.

the two values of J_0, whereas the plug-in Wald interval has coverage 0.949 under the strong signal and 1.000 under the weak one. In the weak-signal design the AR set is unbounded in 62% of replications, which is the expected behavior of an identification-robust set when J_0 is near zero (Proposition 6, Remark 5).

7 Empirical Illustration

We revisit the quarter-of-birth (QOB) design for the returns to schooling, the natural experiment of Angrist and Krueger (1991) and the machine-learning-first-stage setting of Angrist and Frandsen (2022). The instruments are numerous and individually weak. The first stage is exactly the high-dimensional problem of pooling many weak instruments into a single signal, and the returns to schooling are an example of essential heterogeneity, since individuals plausibly select schooling on their own unobserved return. The design is therefore a test of whether the machine-learned-signal estimand drifts with the first-stage learner.

7.1 Data and design

We use the Angrist and Krueger (1991) data from the 1980 U.S. Census[2], restricted to men born 1930–1939, giving $N = 329{,}509$. The outcome Y is log weekly wage and the treatment X is years of completed schooling. Following Angrist and Frandsen (2022), the excluded instruments Z are the three quarter-of-birth indicators fully interacted with year and state

[2]The data is downloaded from the Angrist Data Archive, https://economics.mit.edu/people/faculty/josh-angrist/angrist-data-archive

of birth, in total 186 raw instruments. The covariates W partialled out are a saturated set of year-of-birth and state-of-birth indicators together with race, marital status, and SMSA residence. The covariate projections $l_Y = \mathbb{E}[Y \mid W]$ and $m_X = \mathbb{E}[X \mid W]$ are estimated by saturated least squares on the discrete W values, which is nonparametric and consistent because W takes finitely many values. The signal $g_X = \mathbb{E}[X \mid Z, W]$ and the outcome projection $q_Y = \mathbb{E}[Y \mid Z, W]$ are the two functions estimated by the machine learner. All four nuisances are 5-fold cross-fitted, and the procedure is repeated over 11 independent random partitions of the sample; we report the median of the 11 estimates with the median-adjusted variance, following the median aggregation of Chernozhukov et al. (2018).

We estimate the signal with four first stages: an ordinary least-squares projection on the full instrument set (the "optimal instrument" 2SLS of Chen et al. (2021)), the lasso, random forest, and gradient boosting. For each we compute $\hat{\beta}_{\text{naive}}$ and $\hat{\beta}_{\text{CRC}}$, the alignment diagnostic pair $(T_N, \widehat{\Xi}_N)$, and the identification-robust set $\mathcal{C}_{0.95}$.

7.2 Results

Table 3 and Figure 2 report the results. The naive $\hat{\beta}_{\text{naive}}$ estimates a learner-dependent moving SWATE $\beta_{IV}(\hat{g}_X)$, while $\hat{\beta}_{\text{CRC}}$ estimates the fixed, learner-invariant β_0, so the gap between the two estimates measures the drift of the naive estimand learner by learner. The alignment diagnostic T_N tests, for each learner, whether the naive estimate targets β_0 or a drifted alternative.

The diagnostic behaves as Proposition 5(iv) predicts, which shows T_N to be the studentized contrast between the naive and robust estimates. It rejects alignment for gradient boosting ($T_N = 10.6$), whose naive and robust estimates differ by 0.044, and does not reject for the lasso ($T_N = 0.0$), whose two estimates coincide near 0.11. The random forest lies just inside the band ($T_N = 1.6$). The heterogeneity measure $\widehat{\Xi}_N$ is largest for the random forest, so heterogeneity along the signal is most detectable for the tree learners.

The CRC-robust estimates are more dispersed across learners than the naive ones, inflated by the OLS-projection case, which is also the most instructive case. Its naive estimate, 0.083 (0.024), coincides with the classic Angrist and Krueger (1991) two-stage least squares estimate. But partialling the saturated W out of the 186 weak and mechanically collinear QOB interactions leaves little residual signal with $\widehat{J}_0 = 0.012$, and the identification-robust

Table 3: Returns to schooling: naive and CRC-robust estimates

First stage	$\hat{\beta}_{\text{naive}}$	$\hat{\beta}_{\text{CRC}}$	T_N	$\widehat{\Xi}_N$	$\widehat{J}_0$	$\mathcal{C}_{0.95}$
OLS projection	0.083 (0.024)	−0.008 (0.044)	−4.7	0.003	0.012	$(-\infty, \infty)$
Lasso	0.113 (0.123)	0.108 (0.012)	0.0	0.004	0.017	[0.084, 0.134]
Random forest	0.114 (0.014)	0.132 (0.004)	1.6	0.046	0.143	[0.118, 0.147]
Gradient boosting	0.114 (0.006)	0.158 (0.007)	10.6	0.031	0.097	[0.137, 0.183]

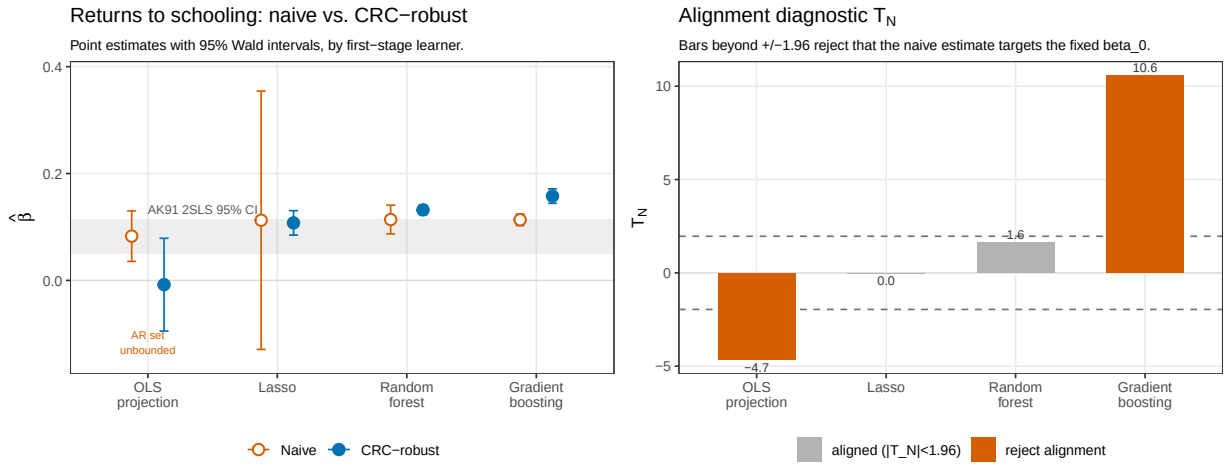
Notes: The table compares naive and CRC-robust estimates across first-stage learners in the quarter-of-birth design. Point estimates are medians over 11 cross-fitting splits, with median-adjusted standard errors in parentheses. T_N and $\widehat{\Xi}_N$ are the alignment diagnostic of Proposition 5, $\widehat{J}_0$ is the cross-fitted sample estimate of the residual-signal strength $J_0 = \mathbb{E}[(g_X - m_X)^2]$, and $\mathcal{C}_{0.95}$ is the AR identification-robust set of Proposition 6.

set is the whole real line. This is the weak residual signal regime of Proposition 6 and Remark 5, and it is the semiparametric counterpart of the Bound et al. (1995) critique of the QOB design: the unbounded $\mathcal{C}_{0.95}$ reports that β_0 is not identified by unregularized pooling of these instruments, which the naive Wald interval does not reveal. The alignment statistic indicates the same degeneracy, with its large value $T_N = -4.7$ reflecting the distance between the naive and robust estimates (0.083 against −0.008).

Regularizing the first stage restores a bounded target. The lasso, random forest, and gradient boosting all yield bounded sets, though only the two tree learners deliver a strong residual signal. The lasso is an intermediate case. Its residual signal $\widehat{J}_0 = 0.017$ is barely above the OLS projection's, and its naive Wald interval is too wide to be informative, yet the CRC-robust score returns a tight estimate, 0.108 (0.012), inside the bounded set [0.084, 0.134]. The orthogonal score and its identification-robust set thus recover an interpretable, convex signal-weighted average of returns to schooling from a first stage whose naive interval would be uninformative.

The three reports give the empirical recommendation. We report $\hat{\beta}_{\text{CRC}}$ as the estimate, a signal-weighted average of the heterogeneous returns to schooling that is convex under covariance monotonicity (Theorem 1). We read $(T_N, \widehat{\Xi}_N)$ as interpretation rather than as a rule for selecting a preferred learner. A rejecting T_N is an informative warning, but a non-rejecting T_N does not certify naive validity, which cannot be verified without estimating $\mathcal{H}$. Where the residual signal is weak, we read the unbounded $\mathcal{C}_{0.95}$ as a statement about the interpretability and validity of the estimand.

Figure 2: Naive and CRC-robust estimates and the alignment diagnostic T_N



Notes: Left: naive $\hat{\beta}_{\text{naive}}$ (hollow circles) and CRC-robust $\hat{\beta}_{\text{CRC}}$ (filled) with 95% Wald intervals, quarter-of-birth returns to schooling for four first-stage learners. The shaded band is the classic Angrist and Krueger (1991) 2SLS 95% confidence interval. Right: the alignment diagnostic T_N. Bars outside the ± 1.96 band reject the hypothesis that the naive estimate targets.

8 Conclusion

We have given the machine-learned-signal IV estimand a structural meaning and a valid inference theory under correlated random coefficients. The partialled-out estimand built from any signal of the instruments is a signal-weighted average of the heterogeneous effects, convex under a covariance-monotonicity condition, which we micro-found from vector monotonicity and positive association. When the signal is learned, the standard debiased moment is not Neyman-orthogonal for the fixed target, and its first-order bias is a drift of the estimand. Therefore, naive DML validly answers a moving, learner-dependent question while inference on the fixed, learner-invariant target requires the heterogeneity-robust orthogonal score we construct, which is the efficient influence function for the target.

Our recommended estimator is $\hat{\beta}_{\text{CRC}}$, which stays valid where the naive one is inconsistent even when the point estimate is centered. Alongside it we report the pair $(T_N, \widehat{\Xi}_N)$, read as interpretation and never as a selector, since pre-testing distorts coverage. When the residual signal is weak, we add the identification-robust set $\mathcal{C}_{1-\alpha}$, whose unbounded realizations indicate that the estimand's interpretability is questionable. Together, the estimator and its diagnostics allow a flexible machine-learned first stage to be used for instrumental-variables estimation while the target remains a fixed, interpretable parameter on which valid inference

is available. As such first stages become routine, our results give the resulting estimands a structural meaning and a way to tell whether a reported estimate answers this fixed question or instead tracks the learner that produced it.

References

Anderson, T. W. and H. Rubin (1949): "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations," *Annals of Mathematical Statistics*, 20, 46–63.

Andrews, I., J. H. Stock, and L. Sun (2019): "Weak Instruments in Instrumental Variables Regression: Theory and Practice," *Annual Review of Economics*, 11, 727–753.

Angrist, J. D. and B. Frandsen (2022): "Machine Labor," *Journal of Labor Economics*, 40, S97–S140.

Angrist, J. D., K. Graddy, and G. W. Imbens (2000): "The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish," *Review of Economic Studies*, 67, 499–527.

Angrist, J. D. and G. W. Imbens (1995): "Two-stage least squares estimation of average causal effects in models with variable treatment intensity," *Journal of the American Statistical Association*, 90, 431–442.

Angrist, J. D. and A. B. Krueger (1991): "Does Compulsory School Attendance Affect Schooling and Earnings?" *The Quarterly Journal of Economics*, 106, 979–1014.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012): "Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain," *Econometrica*, 80, 2369–2429.

Bickel, P. J., C. A. J. Klaassen, Y. Ritov, and J. A. Wellner (1993): *Efficient and Adaptive Estimation for Semiparametric Models*, Baltimore: Johns Hopkins University Press.

Blandhol, C., J. Bonney, M. Mogstad, and A. Torgovitsky (2022): "When is TSLS Actually LATE?" NBER Working Paper 29709, National Bureau of Economic Research, forthcoming, Review of Economic Studies.

Bound, J., D. A. Jaeger, and R. M. Baker (1995): "Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable Is Weak," *Journal of the American Statistical Association*, 90, 443–450.

Card, D. (2001): "Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems," *Econometrica*, 69, 1127–1160.

Chamberlain, G. (1987): "Asymptotic Efficiency in Estimation with Conditional Moment Restrictions," *Journal of Econometrics*, 34, 305–334.

Chen, J., D. L. Chen, and G. Lewis (2021): "Mostly Harmless Machine Learning: Learning Optimal Instruments in Linear IV Models," ArXiv:2011.06158; NeurIPS 2020 Workshop on Machine Learning for Economic Policy.

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

Dufour, J.-M. (1997): "Some Impossibility Theorems in Econometrics with Applications to Structural and Dynamic Models," *Econometrica*, 65, 1365–1387.

Emmenegger, C. and P. Bühlmann (2021): "Regularizing Double Machine Learning in Partially Linear Endogenous Models," *Electronic Journal of Statistics*, 15, 6461–6543.

Esary, J. D., F. Proschan, and D. W. Walkup (1967): "Association of random variables, with applications," *The Annals of Mathematical Statistics*, 38, 1466–1474.

Evdokimov, K. S. and M. Kolesár (2018): "Inference in Instrumental Variables Analysis with Heterogeneous Treatment Effects," Working paper, Princeton University.

Fieller, E. C. (1954): "Some Problems in Interval Estimation," *Journal of the Royal Statistical Society. Series B*, 16, 175–185.

Garen, J. (1984): "The Returns to Schooling: A Selectivity Bias Approach with a Continuous Choice Variable," *Econometrica*, 52, 1199–1218.

Goff, L. (2024): "A Vector Monotonicity Assumption for Multiple Instruments," *Journal of Econometrics*, 241, 105735.

Hausman, J. A. (1978): "Specification Tests in Econometrics," *Econometrica*, 46, 1251–1271.

Heckman, J. J., D. Schmierer, and S. Urzua (2010): "Testing the Correlated Random Coefficient Model," *Journal of Econometrics*, 158, 177–203.

Heckman, J. J., S. Urzua, and E. Vytlacil (2006): "Understanding Instrumental Variables in Models with Essential Heterogeneity," *The Review of Economics and Statistics*, 88, 389–432.

Heckman, J. J. and E. Vytlacil (2005): "Structural equations, treatment effects, and econometric policy evaluation," *Econometrica*, 73, 669–738.

Heckman, J. J. and E. J. Vytlacil (1998): "Instrumental variables methods for the correlated random coefficient model: Estimating the average rate of return to schooling when the return is correlated with schooling," *Journal of Human Resources*, 33, 974–987.

Huntington-Klein, N. (2020): "Instruments with Heterogeneous Effects: Bias, Monotonicity, and Localness," *Journal of Causal Inference*, 8, 182–208.

Imbens, G. W. and J. D. Angrist (1994): "Identification and estimation of local average treatment effects," *Econometrica*, 62, 467–475.

Kolesár, M. (2013): "Estimation in an Instrumental Variables Model with Treatment Effect Heterogeneity," Working paper, Princeton University.

Lee, S. (2018): "A Consistent Variance Estimator for 2SLS When Instruments Identify Different LATEs," *Journal of Business & Economic Statistics*, 36, 400–410.

Mikusheva, A. and L. Sun (2022): "Inference with Many Weak Instruments," *The Review of Economic Studies*, 89, 2663–2686.

Milgrom, P. R. and R. J. Weber (1982): "A theory of auctions and competitive bidding," *Econometrica*, 50, 1089–1122.

Mogstad, M., A. Torgovitsky, and C. R. Walters (2021): "The causal interpretation of two-stage least squares with multiple instrumental variables," *American Economic Review*, 111, 3663–3698.

——— (2024): "Policy Evaluation with Multiple Instrumental Variables," *Journal of Econometrics*, 243, 105718.

Newey, W. K. (1990): "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics*, 5, 99–135.

——— (1994): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62, 1349–1382.

Pitt, L. D. (1982): "Positively correlated normal variables are associated," *The Annals of Probability*, 10, 496–499.

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

Scheidegger, C., Z. Guo, and P. Bühlmann (2026): "Inference for Heterogeneous Treatment Effects with Efficient Instruments and Machine Learning," *Electronic Journal of Statistics*, 20, 718–770.

Słoczyński, T. (2024): "When Should We (Not) Interpret Linear IV Estimands as LATE?" *Review of Economic Studies*, forthcoming.

Staiger, D. and J. H. Stock (1997): "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65, 557–586.

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

Vytlacil, E. (2002): "Independence, Monotonicity, and Latent Index Models: An Equivalence Result," *Econometrica*, 70, 331–341.

Wooldridge, J. M. (1997): "On two stage least squares estimation of the average treatment effect in a random coefficient model," *Economics Letters*, 56, 129–133.

A Proofs and supplementary results

Proofs for Section 2.4: Covariance Monotonicity and its Micro-foundation

Proof of Proposition 1. Step 1 (within-stratum covariance). Conditioning on W and using $m_X(W) = \mathbb{E}[g_X \mid W]$,

$$\mathbb{E}[\tilde{X}\,\tilde{g}_X \mid \Delta = \delta] = \mathbb{E}_{W|\Delta=\delta}\Big[\mathbb{E}\big[(X - m_X)(g_X - m_X) \mid W, \Delta = \delta\big]\Big].$$

Since g_X is a function of (Z, W) and $Z \perp\!\!\!\perp \Delta \mid W$ (a consequence of Assumption 4), we have $\mathbb{E}[g_X \mid W, \Delta = \delta] = \mathbb{E}[g_X \mid W] = m_X(W)$. The terms involving m_X then collapse and the inner expectation reduces to $\mathrm{Cov}(X, g_X \mid W, \Delta = \delta)$, so

$$\mathbb{E}[\tilde{X}\,\tilde{g}_X \mid \Delta = \delta] = \mathbb{E}_{W|\Delta=\delta}\big[\,\mathrm{Cov}(X, g_X \mid W, \Delta = \delta)\big].$$

Step 2 (conditional response functions). Define $f_\delta(z, w) := \mathbb{E}[X \mid Z = z, W = w, \Delta = \delta]$. By Assumption 4, $Z \perp\!\!\!\perp (\Delta, V) \mid W$, and weak union gives $Z \perp\!\!\!\perp V \mid (W, \Delta)$. Hence the conditional law $F_{V|W,\Delta}(\cdot \mid w, \delta)$ does not depend on z, and

$$f_\delta(z, w) = \int_{\mathcal{V}} h(z, w, v)\, dF_{V|W,\Delta}(v \mid w, \delta).$$

Because $h(\cdot, w, v)$ is increasing in each coordinate under VM and the integrating measure is invariant to z, $f_\delta(\cdot, w)$ is increasing in each coordinate. The same argument with $F_{V|W}$, using $Z \perp\!\!\!\perp V \mid W$, gives this property for $g_X(z, w) = \int_{\mathcal{V}} h(z, w, v)\, dF_{V|W}(v \mid w)$.

Step 3 (association). By the law of total covariance and the (Z, W)-measurability of $g_X(Z, W)$,

$$\mathrm{Cov}(X, g_X \mid W, \Delta = \delta) = \mathrm{Cov}\left(f_\delta(Z, W),\, g_X(Z, W) \mid W, \Delta = \delta\right).$$

The integrand f_δ retains its dependence on δ through the law of V; only the distribution of Z entering the covariance matters, and since $Z \perp\!\!\!\perp \Delta \mid W$ that distribution is $F_{Z|W}$. Both $f_\delta(\cdot, w)$ and $g_X(\cdot, w)$ are coordinate-wise non-decreasing (Step 2), so PA yields

$$\mathrm{Cov}\left(f_\delta(Z, W),\, g_X(Z, W) \mid W = w\right) \geq 0 \quad \text{a.s.}$$

Step 4 (integration over W). The outer expectation in Step 1 preserves the inequality, so $\mathbb{E}[\tilde{X}\,\tilde{g}_X \mid \Delta = \delta] \geq 0$ a.s. Theorem 1 then gives non-negative weights and the convex-combination conclusion. □

Supplement to Section 2.4: A non-monotone example

At the level of propensity rationalizability, Conditional Covariance Monotonicity is strictly weaker than the monotone selection of Imbens and Angrist (1994). Three instrument support points are the minimum needed for this separation.

Example A1 (CCM does not require monotone selection)**.** *Work within a single stratum of W (so $\tilde{A} = A - \mathbb{E}[A]$). Let the instrument have three points $Z \in \{0, 1, 2\}$, uniform ($\mathbb{P}(Z = z) = \frac{1}{3}$), and let the heterogeneous effect take two values $\Delta \in \{a, b\}$ with $\mathbb{P}(\Delta = a) = \mathbb{P}(\Delta = b) = \frac{1}{2}$, drawn independently of Z (conditional exogeneity, W trivial). Treatment $X \in \{0, 1\}$ is Bernoulli given (Z, Δ) with propensity $p(z, \delta) := \mathbb{P}(X = 1 \mid Z = z, \Delta = \delta)$:*

	$z = 0$	$z = 1$	$z = 2$	
$p(\cdot, a)$	0.1	0.5	0.9	(*increasing*)
$p(\cdot, b)$	0.2	0.8	0.3	(*single-peaked*).

Since selection depends on Δ, essential heterogeneity is present. The signal is $g_X(z) = \mathbb{E}[X \mid Z = z] = \frac{1}{2}\{p(z, a) + p(z, b)\} = (0.15,\, 0.65,\, 0.60)$, with $\mu := \mathbb{E}[X] = \frac{7}{15}$ and centered values $g_X(z) - \mu = (-\frac{19}{60},\, \frac{11}{60},\, \frac{8}{60})$.

(i) CCM holds. Because $Z \perp\!\!\!\perp \Delta$, the within-stratum object reduces to a covariance across Z,

$$\mathbb{E}[\tilde{X}\,\tilde{g}_X \mid \Delta = \delta] = \sum_z \tfrac{1}{3}\big(p(z, \delta) - \mu\big)\big(g_X(z) - \mu\big) = \mathrm{Cov}_Z\big(p(\cdot, \delta),\, g_X\big),$$

and direct computation gives $\mathbb{E}[\tilde{X}\tilde{g}_X \mid \Delta = a] = \frac{3}{50} = 0.060 > 0$ and $\mathbb{E}[\tilde{X}\tilde{g}_X \mid \Delta = b] = \frac{37}{900} \approx 0.041 > 0$. Thus CCM holds, and relevance follows from $\mathbb{E}[\tilde{X}\tilde{g}_X] = \frac{91}{1800} = \mathrm{Var}(g_X) > 0$. The density weights are $\omega(a) = \frac{108}{91}$ and $\omega(b) = \frac{74}{91}$; after multiplying by the effect probabilities, the convex masses are $\frac{54}{91}$ and $\frac{37}{91}$. Hence $\beta_{IV} = \frac{54}{91}\,a + \frac{37}{91}\,b$ (Theorem 1).

(ii) No monotone single-index rule rationalizes the data. Any threshold rule $X = \mathbf{1}\{U \leq \nu(Z)\}$ with $(U, \Delta) \perp\!\!\!\perp Z$ (or the continuous analogue $X = h(\nu(Z), U)$, h non-decreasing)

yields $p(z,\delta) = F_{U|\Delta=\delta}(\nu(z))$, weakly increasing in $\nu(z)$ for every δ; it thus requires a single ordering of $\{0,1,2\}$ making $z \mapsto p(z,\delta)$ weakly monotone for $\delta = a$ and $\delta = b$ simultaneously. None exists: $p(\cdot, a)$ is strict, so only the orders $0<1<2$ and $2<1<0$ render it monotone, and under both the single-peaked $p(\cdot, b)$ is non-monotone (a tie in ν is excluded, as it would force $p(z,a) = p(z',a)$). The two effect strata require incompatible orderings, so no monotone-selection model in the sense of Imbens and Angrist (1994) can rationalize the propensity table. Yet the covariances stay non-negative: g_X reorders Z by predicted treatment ($0<2<1$), type b is monotone in that ordering while type a is monotone in the natural one, and each stratum's propensity still covaries positively with the aggregate signal.

Remark A1. *The separation in Example A1 cannot occur with a binary instrument at the level of propensity rationalizability. If $Z \in \{0,1\}$ then $\mathbb{E}[\tilde{X}\tilde{g}_X \mid \Delta = \delta] = \pi_0\pi_1\,(p(1,\delta) - p(0,\delta))\,D$ and $\mathbb{E}[\tilde{X}\tilde{g}_X] = \pi_0\pi_1 D^2$ with $D := g_X(1) - g_X(0)$. CCM and relevance ($D \neq 0$) then force $(p(1,\delta) - p(0,\delta))D \geq 0$ for all δ. After relabeling Z if necessary, $p(1,\delta) \geq p(0,\delta)$ for every δ, so the conditional propensities admit a common-order monotone threshold rationalization. This is an existence statement: without an additional restriction on the potential-treatment coupling, it does not imply that the true coupling contains no defiers. Hence $|\mathrm{Supp}(Z)| \geq 3$ is required for CCM to be incompatible with every such common-order monotone rationalization, and the example is minimal in that sense.*

Proofs for Section 2.3: The representation theorem

Proof of Theorem 1, part (i). Step 1 (numerator). Substituting the CRC equation $Y = Y(0) + X\Delta$ into $\mathbb{E}[\tilde{Y}\tilde{g}]$ gives

$$\tilde{Y} = Y - \mathbb{E}[Y \mid W] = \big(Y(0) - \mathbb{E}[Y(0) \mid W]\big) + \big(X\Delta - \mathbb{E}[X\Delta \mid W]\big).$$

Thus, the numerator becomes

$$\mathbb{E}[\tilde{Y}\tilde{g}] = \mathbb{E}\big[\big(Y(0) - \mathbb{E}[Y(0) \mid W]\big)\tilde{g}\big] + \mathbb{E}\big[\big(X\Delta - \mathbb{E}[X\Delta \mid W]\big)\tilde{g}\big].$$

Assumption 1 gives $Z \perp\!\!\!\perp Y(0) \mid W$. Because $\tilde{g}$ is a function of (Z, W) and $\mathbb{E}[\tilde{g} \mid W] = 0$, iterated expectations set the first term to zero. They also give $\mathbb{E}\big[\mathbb{E}[X\Delta \mid W]\,\tilde{g}\big] = \mathbb{E}\big[\mathbb{E}[X\Delta \mid W]\,\mathbb{E}[\tilde{g} \mid W]\big] = 0$. Therefore

$$\mathbb{E}[\tilde{Y}\tilde{g}] = \mathbb{E}[X\Delta\tilde{g}].$$

Step 2 (residualizing X). Substitute $X = \tilde{X} + \mathbb{E}[X \mid W]$:

$$\mathbb{E}[X\Delta\tilde{g}] = \mathbb{E}\big[(\tilde{X} + \mathbb{E}[X \mid W])\Delta\tilde{g}\big] = \mathbb{E}[\tilde{X}\Delta\tilde{g}] + \mathbb{E}\big[\mathbb{E}[X \mid W]\,\Delta\tilde{g}\big].$$

Assumption 1 also gives $Z \perp\!\!\!\perp \Delta \mid W$, so $\mathbb{E}[\Delta\tilde{g} \mid W] = \mathbb{E}[\Delta \mid W]\,\mathbb{E}[\tilde{g} \mid W] = 0$. The second term is therefore zero, and

$$\mathbb{E}[\tilde{Y}\tilde{g}] = \mathbb{E}[\tilde{X}\Delta\tilde{g}].$$

Step 3 (conditioning on Δ). Iterated expectations give

$$\mathbb{E}[\tilde{X}\Delta\tilde{g}] = \mathbb{E}_\Delta\big[\mathbb{E}[\tilde{X}\Delta\tilde{g} \mid \Delta]\big] = \mathbb{E}_\Delta\big[\Delta \cdot \mathbb{E}[\tilde{X}\tilde{g} \mid \Delta]\big].$$

Step 4 (forming the ratio). Dividing by the denominator gives

$$\beta_{IV} = \frac{\mathbb{E}_\Delta\big[\Delta \cdot \mathbb{E}[\tilde{X}\tilde{g} \mid \Delta]\big]}{\mathbb{E}[\tilde{X}\tilde{g}]} = \mathbb{E}_\Delta\left[\Delta \cdot \left(\frac{\mathbb{E}[\tilde{X}\tilde{g} \mid \Delta]}{\mathbb{E}[\tilde{X}\tilde{g}]}\right)\right] \equiv \mathbb{E}_\Delta\big[\Delta \cdot \omega(\Delta)\big]. \qquad \square$$

Proof of Theorem 1, parts (ii)–(iii). Taking the unconditional expectation of the conditional numerator in part (i) gives

$$\mathbb{E}_\Delta\big[\omega(\Delta)\big] = \frac{\mathbb{E}_\Delta\big[\mathbb{E}[\tilde{X}\tilde{g} \mid \Delta]\big]}{\mathbb{E}[\tilde{X}\tilde{g}]} = \frac{\mathbb{E}[\tilde{X}\tilde{g}]}{\mathbb{E}[\tilde{X}\tilde{g}]} = 1.$$

Assumption 3 gives $\mathbb{E}[\tilde{X}\tilde{g} \mid \Delta = \delta] \geq 0$ a.s., while Assumption 2 and the maintained sign normalization make $\mathbb{E}[\tilde{X}\tilde{g}]$ strictly positive. Hence $\omega(\delta) \geq 0$ a.s. Together with part (i), this makes β_{IV} a convex combination of heterogeneous treatment effects. $\square$

Proofs for Section 2.5: Relation to the LATE Framework

Lemma A1 (Margin decomposition of the residual kernel)**.** *Fix a covariate value w outside the relevant null set and suppress w from the notation. Assumption 4 with $V = U$ then gives $g \perp\!\!\!\perp (U, \Delta)$ in this conditional law. Under the threshold-crossing model, for almost every*

$\delta \in \mathrm{Supp}(\Delta)$,

$$\mathbb{E}\big[\tilde{X}\,\tilde{g} \mid \Delta = \delta\big] \;=\; \sum_{\ell=2}^{L} S_\ell\,\pi_\ell(\delta), \qquad \textit{and in particular} \qquad \mathbb{E}\big[\tilde{X}\,\tilde{g}\big] \;=\; \sum_{\ell=2}^{L} S_\ell\,\bar{\pi}_\ell.$$

Proof of Lemma A1. Joint independence $g \perp\!\!\!\perp (U, \Delta)$ implies both $\mathbb{E}[g \mid \Delta = \delta] = \bar{g}$ and $U \perp\!\!\!\perp g \mid \Delta$. Hence centering X does not change its conditional covariance with $\tilde{g}$, and

$$\mathbb{E}[\tilde{X}\,\tilde{g} \mid \Delta = \delta] = \mathrm{Cov}\left(X, g \mid \Delta = \delta\right) = \mathbb{E}\big[(g - \bar{g})\,\mathbb{P}(U \le g \mid \Delta = \delta)\big].$$

Let $F^\delta(t) := \mathbb{P}(U \le t \mid \Delta = \delta)$ denote the conditional c.d.f. of U. Then

$$\mathbb{E}\big[\tilde{X}\,\tilde{g} \mid \Delta = \delta\big] \;=\; \sum_{j=1}^{L} p_j\,(g_j - \bar{g})\,F^\delta(g_j).$$

Write $F^\delta(g_j) = \sum_{\ell \le j} \pi_\ell(\delta)$ with $\pi_1(\delta) := \mathbb{P}(U \le g_1 \mid \Delta = \delta)$ and $\pi_\ell(\delta) = F^\delta(g_\ell) - F^\delta(g_{\ell-1})$ for $\ell \ge 2$. Substituting and exchanging the order of summation (Abel/summation-by-parts),

$$\sum_{j=1}^{L} p_j(g_j - \bar{g}) \sum_{\ell \le j} \pi_\ell(\delta) \;=\; \sum_{\ell=1}^{L} \pi_\ell(\delta) \underbrace{\sum_{j \ge \ell} p_j(g_j - \bar{g})}_{= S_\ell} \;=\; \sum_{\ell=1}^{L} S_\ell\,\pi_\ell(\delta).$$

The $\ell = 1$ term vanishes because $S_1 = \sum_{j \ge 1} p_j(g_j - \bar{g}) = \mathbb{E}[g - \bar{g}] = 0$; equivalently, always-takers ($U \le g_1$) have $X \equiv 1$ and contribute nothing to a covariance with $\tilde{g}$. Hence $\mathbb{E}[\tilde{X}\,\tilde{g} \mid \Delta = \delta] = \sum_{\ell=2}^{L} S_\ell\,\pi_\ell(\delta)$. Taking $\mathbb{E}_\Delta$ and using $\mathbb{E}[\pi_\ell(\Delta)] = \bar{\pi}_\ell$ (law of iterated expectations) gives

$$D(w) = \mathbb{E}[\tilde{X}\,\tilde{g} \mid W = w] = \sum_{\ell=2}^{L} S_\ell\,\bar{\pi}_\ell.$$

Proposition 2 assumes this within-stratum denominator is strictly positive when it forms $\beta_{IV}(w)$. □

Proof of Proposition 2, part (i). Work in the conditional law given $W = w$. Assumption 4 with $V = U$ supplies the signal exogeneity required by Theorem 1, while $D(w) > 0$ supplies relevance in this stratum. Applying that theorem conditionally and inserting Lemma A1

gives

$$\mathbb{E}_\Delta\big[\Delta \cdot \mathbb{E}[\tilde{X}\,\tilde{g} \mid \Delta]\big] = \sum_{\ell \in \mathcal{L}_+} S_\ell\, \mathbb{E}\big[\Delta\, \pi_\ell(\Delta)\big] = \sum_{\ell \in \mathcal{L}_+} S_\ell \int \delta\, \mathbb{P}(C_\ell \mid \Delta = \delta)\, dF_\Delta(\delta)$$
$$= \sum_{\ell \in \mathcal{L}_+} S_\ell\, \bar{\pi}_\ell\, \mathrm{LATE}_\ell,$$

where the last equality follows from $\mathbb{E}[\Delta\, \pi_\ell(\Delta)] = \bar{\pi}_\ell\, \mathrm{LATE}_\ell$ by iterated expectations, equivalently by the complier-density Bayes identity in part (ii). Dividing by $D(w) = \sum_{\ell' \in \mathcal{L}_+} S_{\ell'} \bar{\pi}_{\ell'}$ gives $\beta_{IV}(w) = \sum_\ell \lambda_\ell\, \mathrm{LATE}_\ell$. Because $S_\ell \geq 0$ and $\bar{\pi}_\ell = \mathbb{P}(C_\ell) > 0$, the weights are nonnegative and sum to one. □

Proof of Proposition 2, part (ii). By the conditional version of the weighting function in Theorem 1 and Lemma A1,

$$\omega_w(\delta) \;=\; \frac{\mathbb{E}[\tilde{X}\,\tilde{g} \mid \Delta = \delta, W = w]}{D(w)} \;=\; \frac{\sum_{\ell \in \mathcal{L}_+} S_\ell\, \pi_\ell(\delta)}{\sum_{\ell' \in \mathcal{L}_+} S_{\ell'}\, \bar{\pi}_{\ell'}}.$$

Apply Bayes' rule to each $\ell \in \mathcal{L}_+$:

$$\pi_\ell(\delta)\, dF_\Delta(\delta) = \mathbb{P}(C_\ell \mid \Delta = \delta)\, dF_\Delta(\delta) = \bar{\pi}_\ell\, dF_{\Delta|C_\ell}(\delta),$$
$$\Longrightarrow \quad \frac{\pi_\ell(\delta)}{\bar{\pi}_\ell} = \frac{dF_{\Delta|C_\ell}(\delta)}{dF_\Delta(\delta)} \quad \text{a.e.}$$

Substituting $\pi_\ell(\delta) = \bar{\pi}_\ell\, dF_{\Delta|C_\ell}/dF_\Delta$ into the numerator,

$$\omega_w(\delta) \;=\; \frac{\sum_{\ell \in \mathcal{L}_+} S_\ell\, \bar{\pi}_\ell \left(dF_{\Delta|C_\ell}(\delta)/dF_\Delta(\delta)\right)}{\sum_{\ell' \in \mathcal{L}_+} S_{\ell'}\, \bar{\pi}_{\ell'}} \;=\; \sum_{\ell \in \mathcal{L}_+} \lambda_\ell\, \frac{dF_{\Delta|C_\ell}(\delta)}{dF_\Delta(\delta)},$$

which is (6). This identity is taken under the conditional law of (Δ, U, g) given $W = w$. Each likelihood ratio integrates to one under F_Δ, so $\mathbb{E}_\Delta[\omega_w(\Delta)] = \sum_\ell \lambda_\ell = 1$, as required by Theorem 1. □

Proof of Corollary 1. For $D(w) > 0$, Proposition 2 gives

$$\mathbb{E}[\tilde{Y}\tilde{g} \mid W = w] = D(w)\beta_{IV}(w) = D(w) \sum_{\ell \in \mathcal{L}_+(w)} \lambda_\ell(w) \mathrm{LATE}_\ell(w).$$

The margin calculation in Lemma A1 also shows that both conditional moments vanish when

$D(w) = 0$. Iterated expectations therefore give (7). Non-negativity follows from $D(w) \geq 0$ and $\lambda_\ell(w) \geq 0$, while normalization follows from

$$\int_{\{w:\, D(w)>0\}} \sum_{\ell \in \mathcal{L}_+(w)} \frac{D(w)\lambda_\ell(w)}{\mathbb{E}[D(W)]}\, dF_W(w) = 1.$$

□

Proofs for Section 3: The Moving Target

Throughout this part, $\eta_0 = (l_Y, m_X, q_Y, g_X)$ denotes the true nuisance functions of Section 3.1, $\mathcal{H} = \mathcal{H}(Z, W)$ is the heterogeneity residual in (10), and perturbation directions inherit the measurability of the corresponding nuisance. Thus directions for l_Y, m_X are functions of W, whereas directions for q_Y, g_X are functions of (Z, W). The CRC equation and Assumption 1 imply the following identity, in which the $Y(0)$ term cancels:

$$q_Y - l_Y = \mathbb{E}[Y \mid Z, W] - \mathbb{E}[Y \mid W] = \mathbb{E}[X\Delta \mid Z, W] - \mathbb{E}[X\Delta \mid W]. \tag{A1}$$

Proof of Proposition 3. The moment depends on the signal coordinate only through the factor $(g_X - m_X)$. Perturbing $g_X \mapsto g_X + t\,\delta_g$ and differentiating at $t = 0$,

$$\partial_{g_X}\, \mathbb{E}\big[\psi_{\text{naive}}(O; \beta_0, \eta_0)\big][\delta_g] = \mathbb{E}\big[\big(Y - l_Y - \beta_0(X - m_X)\big)\, \delta_g\big].$$

Since δ_g is (Z, W)-measurable, condition the first factor on (Z, W), using $\mathbb{E}[Y \mid Z, W] = q_Y$ and $\mathbb{E}[X \mid Z, W] = g_X$:

$$= \mathbb{E}\big[\big(q_Y - l_Y - \beta_0(g_X - m_X)\big)\, \delta_g\big].$$

By (A1) the bracket equals $\mathcal{H}$, so the derivative is $\mathbb{E}[\mathcal{H}\,\delta_g]$. By the heterogeneity condition (Definition 1), $\mathbb{P}(\mathcal{H} \neq 0) > 0$; taking $\delta_g = \mathcal{H}$ gives $\mathbb{E}[\mathcal{H}^2] > 0$, so the derivative does not vanish in all directions and Neyman-orthogonality fails. □

Derivation of the expansion following Theorem 2. Write $U = Y - l_Y - \beta_0(X - m_X)$ and $V = g_X - m_X$. By a standard cross-fitted Z-estimation expansion, $\hat{\beta}_{\text{naive}} - \beta_0 = J_0^{-1}\, \frac{1}{N} \sum_i \psi_{\text{naive}}(O_i; \beta_0, \hat{\eta}) + o_{\mathbb{P}}(N^{-1/2})$, the Jacobian being

$$J_0 = -\,\partial_\beta\, \mathbb{E}[\psi_{\text{naive}}(O; \beta_0, \eta_0)] = \mathbb{E}[(X - m_X)(g_X - m_X)] = \mathbb{E}[(g_X - m_X)^2], \tag{A2}$$

where the last equality uses $\mathbb{E}[X - m_X \mid Z, W] = g_X - m_X$ under the specialization $g = g_X$. Because $\mathbb{E}[\psi_{\text{naive}}(O; \beta_0, \eta_0)] = 0$, cross-fitting gives

$$\tfrac{1}{N}\sum_i \psi_{\text{naive}}(O_i; \beta_0, \hat{\eta}) = \tfrac{1}{N}\sum_i \psi_{\text{naive}}(O_i; \beta_0, \eta_0) + \mathbb{E}\big[\psi_{\text{naive}}(O; \beta_0, \hat{\eta})\big] + o_{\mathbb{P}}(N^{-1/2}),$$

with the middle term evaluated at the independent, out-of-fold $\hat{\eta} = \eta_0 + \Delta\eta$. Substituting into (11),

$$\psi_{\text{naive}}(O; \beta_0, \hat{\eta}) = \big(U - \Delta l_Y + \beta_0\, \Delta m_X\big)\big(V + \Delta g_X - \Delta m_X\big).$$

Take expectations term by term. The constant term $\mathbb{E}[UV]$ is zero by the definition of β_0. Terms linear in the W-measurable errors vanish because $\mathbb{E}[V \mid W] = 0$ and $\mathbb{E}[U \mid W] = 0$. The remaining linear term is

$$\mathbb{E}[U\, \Delta g_X] = \mathbb{E}\big[\Delta g_X\, \mathbb{E}[U \mid Z, W]\big] = \mathbb{E}\big[\Delta g_X \left(q_Y - l_Y - \beta_0(g_X - m_X)\right)\big] = \mathbb{E}[\mathcal{H}\, \Delta g_X].$$

Cauchy–Schwarz bounds the quadratic term $\mathbb{E}\big[(-\Delta l_Y + \beta_0 \Delta m_X)(\Delta g_X - \Delta m_X)\big]$ by sums of products of $L_2(\mathbb{P})$ nuisance-error norms. Assumption 8(ii) makes each product $o_{\mathbb{P}}(N^{-1/2})$. Thus $\mathbb{E}[\psi_{\text{naive}}(O; \beta_0, \hat{\eta})] = \mathbb{E}[\mathcal{H}\, \Delta g_X] + o_{\mathbb{P}}(N^{-1/2})$, and multiplication by J_0^{-1} gives the stated expansion. If $\Delta \equiv \beta_0$, then (A1) gives $q_Y - l_Y = \beta_0(g_X - m_X)$, so $\mathcal{H} \equiv 0$ and the bias is zero. □

Proofs for Section 4: The Fixed Target

Proof of Theorem 3, part (i). Identification. At η_0 each augmentation term carries the factor $(X - g_X)$, with $\mathbb{E}[X - g_X \mid Z, W] = 0$, multiplied by a (Z, W)-measurable factor ($q_Y - l_Y$ in ψ_Y; $g_X - m_X$ in ψ_X). Conditioning on (Z, W), both vanish in expectation, leaving $\mathbb{E}[\psi_Y(O; \eta_0)] = \mathbb{E}[(Y - l_Y)(g_X - m_X)]$ and $\mathbb{E}[\psi_X(O; \eta_0)] = \mathbb{E}[(X - m_X)(g_X - m_X)]$, whose ratio is β_0 by (9). Hence $\mathbb{E}[\psi_{\text{CRC}}(O; \beta_0, \eta_0)] = 0$.

Orthogonality. Differentiate $\mathbb{E}[\psi_{\text{CRC}}] = \mathbb{E}[\psi_Y] - \beta_0\, \mathbb{E}[\psi_X]$ in each coordinate at η_0, in a direction h of the appropriate measurability.

(i) Direction l_Y (appears only in ψ_Y):

$$\partial_{l_Y}\mathbb{E}[\psi_Y][h] = \mathbb{E}\big[-h(g_X - m_X) - h(X - g_X)\big] = \mathbb{E}[-h(X - m_X)] = \mathbb{E}\big[-h\, \mathbb{E}[X - m_X \mid W]\big] = 0.$$

(ii) Direction q_Y (appears only in ψ_Y):

$$\partial_{q_Y}\mathbb{E}[\psi_Y][h] = \mathbb{E}[h(X - g_X)] = \mathbb{E}\big[h\,\mathbb{E}[X - g_X \mid Z, W]\big] = 0.$$

(iii) Direction m_X. In ψ_Y, $\partial_{m_X}\mathbb{E}[\psi_Y][h] = \mathbb{E}[-h(Y - l_Y)] = \mathbb{E}[-h\,\mathbb{E}[Y - l_Y \mid W]] = 0$. In ψ_X the parameter appears three times, and

$$\partial_{m_X}\mathbb{E}[\psi_X][h] = \mathbb{E}\big[-h(g_X - m_X) - h(X - m_X) - h(X - g_X)\big] = \mathbb{E}[-2h(X - m_X)] = 0,$$

again by $\mathbb{E}[X - m_X \mid W] = 0$. The total derivative is $0 - \beta_0 \cdot 0 = 0$.

(iv) Direction g_X. In ψ_Y,

$$\partial_{g_X}\mathbb{E}[\psi_Y][h] = \mathbb{E}\big[h(Y - l_Y) - h(q_Y - l_Y)\big] = \mathbb{E}[h(Y - q_Y)] = \mathbb{E}\big[h\,\mathbb{E}[Y - q_Y \mid Z, W]\big] = 0.$$

In ψ_X (three occurrences),

$$\partial_{g_X}\mathbb{E}[\psi_X][h] = \mathbb{E}\big[h(X - m_X) - h(g_X - m_X) + h(X - g_X)\big] = \mathbb{E}[2h(X - g_X)] = 0.$$

The total derivative is again $0 - \beta_0 \cdot 0 = 0$. All four coordinate derivatives vanish, so ψ_{CRC} is Neyman-orthogonal at η_0. □

Proof of Theorem 3, part (ii). Since ψ_{CRC} is affine in β, the Jacobian is exact: $-\partial_\beta\mathbb{E}[\psi_{\mathrm{CRC}}(O; \beta_0, \eta_0)] = \mathbb{E}[\psi_X(O; \eta_0)] = \mathbb{E}[(g_X - m_X)^2] = J_0$, strictly positive by Assumption 2. The cross-fitted Z-estimator admits the expansion

$$\sqrt{N}\,(\hat{\beta} - \beta_0) = J_0^{-1}\,\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\psi_{\mathrm{CRC}}(O_i; \beta_0, \eta_0) + \sqrt{N}\,J_0^{-1}R_N + o_{\mathbb{P}}(1),$$

where $R_N = \mathbb{E}[\psi_{\mathrm{CRC}}(O; \beta_0, \hat{\eta}) - \psi_{\mathrm{CRC}}(O; \beta_0, \eta_0)]$. The coefficient is $+J_0^{-1}$ because the first-order condition and $\partial_\beta\mathbb{E}[\psi_{\mathrm{CRC}}] = -J_0$ give $\hat{\beta} - \beta_0 = J_0^{-1}\frac{1}{N}\sum_i \psi_{\mathrm{CRC}}(O_i; \beta_0, \hat{\eta}) + o_{\mathbb{P}}(N^{-1/2})$. Each of ψ_Y and ψ_X is quadratic in the nuisances, so its Taylor expansion stops at second order. Theorem 3(i) removes the linear terms. After the products $\pm\Delta l_Y\,\Delta g_X$ cancel, the remainder is

$$R_Y = \mathbb{E}\big[\Delta l_Y\,\Delta m_X - \Delta g_X\,\Delta q_Y\big], \qquad R_X = \mathbb{E}\big[(\Delta m_X)^2 - (\Delta g_X)^2\big], \qquad R_N = R_Y - \beta_0 R_X.$$

By Cauchy–Schwarz,

$$|R_N| \le \|\Delta l_Y\|_{\mathbb{P},2}\|\Delta m_X\|_{\mathbb{P},2} + \|\Delta g_X\|_{\mathbb{P},2}\|\Delta q_Y\|_{\mathbb{P},2} + |\beta_0|\big(\|\Delta m_X\|^2_{\mathbb{P},2} + \|\Delta g_X\|^2_{\mathbb{P},2}\big),$$

and each product is $o_{\mathbb{P}}(N^{-1/2})$ by Assumption 8(ii), so $\sqrt{N}\,R_N \xrightarrow{\mathbb{P}} 0$. The leading term is a sample average of the mean-zero, finite-variance score $\psi_{\mathrm{CRC}}(O;\beta_0,\eta_0)$; the Lindeberg–Lévy central limit theorem and Slutsky's theorem yield $\sqrt{N}(\hat\beta - \beta_0) \xrightarrow{d} \mathcal{N}(0,\, J_0^{-2}\Omega_0)$ with $\Omega_0 = \mathbb{E}[\psi_{\mathrm{CRC}}(O;\beta_0,\eta_0)^2]$. □

Proof of Theorem 3, part (iii). Throughout, expectations are under the true P; write $V = g_X - m_X$, $e = X - g_X$, and note the identities

$$\mathbb{E}[V \mid W] = \mathbb{E}[g_X \mid W] - m_X = 0, \qquad \mathbb{E}[e \mid Z, W] = 0, \qquad \mathbb{E}[Y - l_Y \mid W] = 0. \tag{A3}$$

Step 0 (tangent space and uniqueness). The efficiency calculation uses the unrestricted observed-law model in which the four conditional expectations are functionals and finite moments are the only restrictions. Scores of regular one-dimensional submodels therefore have closed linear span $L_2^0(P)$. Any gradient ϕ satisfying $\frac{d}{dt}\beta(P_t)|_{t=0} = \mathbb{E}[\phi\, s]$ for every score s is unique. Indeed, if ϕ_1 and ϕ_2 are gradients, then $\mathbb{E}[(\phi_1 - \phi_2)s] = 0$ for every $s \in L_2^0(P)$; taking $s = \phi_1 - \phi_2$ gives $\phi_1 = \phi_2$ a.s. This unique gradient is the efficient influence function (Bickel et al., 1993; van der Vaart, 1998). The CRC and exogeneity assumptions supply the causal interpretation of the observed-law functional.

Step 1 (pathwise derivatives). For $A \in L_2$ and a coordinate σ-field $\mathcal{C}$, along a submodel with bounded score s,

$$\tfrac{d}{dt}\mathbb{E}_t[A \mid \mathcal{C}]\big|_{t=0} = \mathbb{E}\big[(A - \mathbb{E}[A \mid \mathcal{C}])\, s \mid \mathcal{C}\big], \tag{A4}$$

so, writing $\dot{(\cdot)}$ for the t-derivative at 0, $\dot g_X = \mathbb{E}[e\, s \mid Z, W]$, $\dot m_X = \mathbb{E}[(X - m_X)s \mid W]$, $\dot l_Y = \mathbb{E}[(Y - l_Y)s \mid W]$; and $\frac{d}{dt}\mathbb{E}_t[h_t]|_{t=0} = \mathbb{E}[\dot h] + \mathbb{E}[h\, s]$.

Step 2 (EIF of θ_Y). Differentiating $\theta_Y(P_t) = \mathbb{E}_t[(Y - l_{Y,t})(g_{X,t} - m_{X,t})]$ in its four channels,

$$\dot\theta_Y = \mathbb{E}[(Y - l_Y)V\, s] - \mathbb{E}[\dot l_Y V] + \mathbb{E}[(Y - l_Y)\dot g_X] - \mathbb{E}[(Y - l_Y)\dot m_X].$$

By (A3) and W-measurability, $\mathbb{E}[\dot l_Y V] = 0$ and $\mathbb{E}[(Y - l_Y)\dot m_X] = 0$. As $\dot g_X$ is (Z, W)-measurable, projecting the outer factor gives $\mathbb{E}[(Y - l_Y)\dot g_X] = \mathbb{E}[(q_Y - l_Y)\,\mathbb{E}[es \mid Z, W]] =$

$\mathbb{E}[(q_Y - l_Y)e\,s]$. Thus q_Y enters when differentiation of the embedded $g_X = \mathbb{E}[X \mid Z, W]$ reprojects Y onto (Z, W). Hence $\mathrm{EIF}(\theta_Y) = (Y - l_Y)V + (q_Y - l_Y)e - \theta_Y$, which is mean zero since $\mathbb{E}[(q_Y - l_Y)e] = 0$, and equals $\psi_Y(O; \eta_0) - \theta_Y$.

Step 3 (EIF of J_0). From $J_0(P_t) = \mathbb{E}_t[V_t^2]$, $\dot{J}_0 = \mathbb{E}[V^2 s] + 2\mathbb{E}[V\dot{g}_X] - 2\mathbb{E}[V\dot{m}_X]$; the last term is zero by (A3), and $\mathbb{E}[V\dot{g}_X] = \mathbb{E}[Ve\,s]$ since V is (Z, W)-measurable, so $\mathrm{EIF}(J_0) = V^2 + 2Ve - J_0 = V(2X - g_X - m_X) - J_0 = \psi_X(O; \eta_0) - J_0$.

Step 4 (ratio and identification). As $J_0 > 0$, the map $(\theta, J) \mapsto \theta/J$ is Hadamard differentiable, so $\beta_0 = \theta_Y/J_0$ has gradient $\phi_{\beta_0} = J_0^{-1}(\mathrm{EIF}(\theta_Y) - \beta_0\,\mathrm{EIF}(J_0))$, and the constant $\theta_Y - \beta_0 J_0 = \mathbb{E}[UV] = 0$ cancels. Expanding $UV + e\mathcal{H}$ from the definitions ($U = Y - l_Y - \beta_0(X - m_X)$, $\mathcal{H} = (q_Y - l_Y) - \beta_0 V$),

$$UV + e\mathcal{H} = (Y - l_Y)V + (q_Y - l_Y)e - \beta_0 V(2X - g_X - m_X) = J_0\,\phi_{\beta_0},$$

so $\phi_{\beta_0} = J_0^{-1}(UV + e\mathcal{H}) = J_0^{-1}\psi_{\mathrm{CRC}}(O; \beta_0, \eta_0)$. By Step 0 this is the efficient influence function; the bound is $\mathrm{Var}(\phi_{\beta_0}) = J_0^{-2}\mathbb{E}[(UV + e\mathcal{H})^2] = J_0^{-2}\Omega_0$, attained by $\hat{\beta}_{\mathrm{CRC}}$ (Theorem 3). □

Proof of Theorem 3, part (iv). If effects are homogeneous, $\Delta \equiv \beta_0$ and hence $\mathcal{H} = 0$. Lemma 2 then gives $\psi_{\mathrm{CRC}}(\cdot; \beta_0, \eta_0) = \psi_{\mathrm{naive}}(\cdot; \beta_0, \eta_0) = UV$, so $\Omega_0 = \Omega_0^{\mathrm{nv}}$. □

Conventions for the proofs of Sections 3–5

Unless stated otherwise, the following conventions use Assumption 7.

Population residuals at the truth. Write

$$U := Y - l_Y - \beta_0(X - m_X), \qquad V := g_X - m_X, \qquad e := X - g_X.$$

By Assumption 7(i) and Cauchy–Schwarz, $U, V, e \in L_2(\mathbb{P})$, and $UV, ea \in L_2(\mathbb{P})$ for bounded a. Three conditional-mean identities recur. By $l_Y = \mathbb{E}[Y \mid W]$, $m_X = \mathbb{E}[X \mid W]$ and the tower property,

$$\mathbb{E}[U \mid W] = \mathbb{E}[Y \mid W] - l_Y - \beta_0(\mathbb{E}[X \mid W] - m_X) = 0. \tag{A5}$$

By the identity (A1) and the definition (10), $q_Y - l_Y = \mathbb{E}[X\Delta \mid Z, W] - \mathbb{E}[X\Delta \mid W]$ and $\mathcal{H} = (q_Y - l_Y) - \beta_0 V$; since $\mathbb{E}[Y \mid Z, W] = q_Y$, $\mathbb{E}[X \mid Z, W] = g_X$ and hence $\mathbb{E}[X - m_X \mid Z, W] = g_X - m_X = V$,

$$\mathbb{E}[U \mid Z, W] = (q_Y - l_Y) - \beta_0 V = \mathcal{H}(Z, W). \tag{A6}$$

Taking a further expectation and using (A5) gives $\mathbb{E}[\mathcal{H} \mid W] = 0$ (Lemma 1(i)). Finally, by definition of $g_X = \mathbb{E}[X \mid Z, W]$, for every (Z, W)-measurable $f \in L_2$,

$$\mathbb{E}[e\, f \mid Z, W] = 0, \qquad \text{hence} \quad \mathbb{E}[e\, f] = 0. \tag{A7}$$

Cross-fitting notation. For a score coordinate $\phi(O; \beta, \eta)$ set $\Delta\eta_{j,k} := \hat{\eta}_{j,k} - \eta_{0,j}$, $\rho_{j,k} := \|\Delta\eta_{j,k}\|_{\mathbb{P},2}$, $\rho_N := \max_{j,k} \rho_{j,k}$, as in the cross-fitting construction of Section 3.1. Conditional expectations $\mathbb{E}[\cdot \mid \mathcal{D}_{-k}]$ hold the $\mathcal{D}_{-k}$-measurable estimate $\hat{\eta}_k$ fixed and integrate a fresh $O \perp \mathcal{D}_{-k}$; thus A_k, B_k, β_k^* of (13) and the loadings $\mathbb{E}[\mathcal{H}\, \Delta g_{X,k} \mid \mathcal{D}_{-k}]$ are $\mathcal{D}_{-k}$-measurable random variables. Finite constants depending only on $(\bar{C}, C_4, |\beta_0|)$ are written L, c, C and may change from line to line. We abbreviate $\|\cdot\| := \|\cdot\|_{\mathbb{P},2}$.

Two moment inequalities. For (Z, W)- or W-measurable f with $\|f\|_\infty < \infty$ and $S \in L_4$,

$$\mathbb{E}[f^2 S^2] \le \|f\|_\infty^2\, \mathbb{E}[S^2], \qquad \big|\mathbb{E}[f\, S^2]\big| \le \|f\|\, \sqrt{\mathbb{E}[S^4]}. \tag{A8}$$

For bounded b ($\|b\|_\infty < \infty$), the interpolation bound is $\|b\|_{\mathbb{P},4}^2 = (\mathbb{E}[b^4])^{1/2} \le (\|b\|_\infty^2\, \mathbb{E}[b^2])^{1/2} = \|b\|_\infty\, \|b\|$.

Lemma A2 (Modulus of continuity of the score coordinates)**.** *Let Assumption 7 hold, and consider* $\phi \in \{\psi_{\text{naive}}, \psi_Y, \psi_X, \psi_{\text{CRC}}, \phi_{\text{diag}}, j\}$, *where* $\phi_{\text{diag}}(O; \beta, \eta) := (X - g)\{(q - l) - \beta(g - m)\}$ *and* $j(O; \eta) := (X - m)(g - m)$. *There is* $L = L(\bar{C}, C_4, |\beta_0|) < \infty$ *such that, for all admissible* $(\beta, \eta), (\beta', \eta')$ *with nuisance coordinates bounded by* $\bar{C}$ *in sup-norm and* $|\beta|, |\beta'| \le \bar{C}$,

$$\big\|\phi(\,\cdot\,; \beta', \eta') - \phi(\,\cdot\,; \beta, \eta)\big\| \;\le\; L\left(|\beta' - \beta| \;+\; \sum_j \|\eta_j' - \eta_j\|^{1/2}\right).$$

In particular, if $|\beta' - \beta| \to 0$ *and* $\max_j \|\eta_j' - \eta_j\| \to 0$, *then* $\|\phi(\,\cdot\,; \beta', \eta') - \phi(\,\cdot\,; \beta, \eta)\| \to 0$; *and if* ϕ *does not depend on* β *(the kernel* j*), the* $|\beta' - \beta|$ *term is absent.*

Proof. Every listed kernel is a finite sum of bilinear terms $F = F_1 F_2$ in which each factor

is affine in (β, η); the coefficients are either constants, bounded nuisance coordinates, or the components $Y, X \in L_4$ (Assumption 7(i)). It suffices to bound the L_2 increment of a representative term of each of the two shapes that occur: (a) both factors bounded, and (b) one factor an L_4 component (Y or X) and the other bounded. We use the telescoping identity $F' - F = (F_1' - F_1)F_2' + F_1(F_2' - F_2)$.

Shape (a): $F = b_1 b_2$, where both b_1 and b_2 are bounded affine factors. Each increment $b_i' - b_i$ is a bounded nuisance difference. By the first inequality in (A8), $\|(b_1' - b_1)b_2'\| \le \|b_2'\|_\infty \|b_1' - b_1\| \le \bar{C}' \|b_1' - b_1\|$, and similarly for the second summand, where $\bar{C}'$ is a constant multiple of $\bar{C}$. Hence such a term contributes at most $C \sum_j \|\eta_j' - \eta_j\|$. On the relevant regime this is at most $C \sum_j \|\eta_j' - \eta_j\|^{1/2}$ because deviations are bounded by $2\bar{C}$ and $t \le (2\bar{C})^{1/2} t^{1/2}$.

Shape (b): $F = A \cdot b$ with $A \in \{Y, X\}$ (up to bounded shifts, themselves of shape (a)) and b a bounded affine factor. The increment is $A(b' - b)$ plus shape-(a) remainders. Now

$$\|A(b'-b)\|^2 = \mathbb{E}[A^2(b'-b)^2] \overset{\text{C–S}}{\le} \|A\|_{\mathbb{P},4}^2 \|b'-b\|_{\mathbb{P},4}^2 \le \|A\|_{\mathbb{P},4}^2 \|b'-b\|_\infty \|b'-b\| \le C \|b'-b\|,$$

using the interpolation bound and $\|b' - b\|_\infty \le 2\bar{C}$, $\|A\|_{\mathbb{P},4}^2 \le \sqrt{C_4}$. As $b' - b$ is an affine function of the nuisance differences, $\|b' - b\| \le \sum_j \|\eta_j' - \eta_j\|$, so this term contributes $\le C\big(\sum_j \|\eta_j' - \eta_j\|\big)^{1/2} \le C \sum_j \|\eta_j' - \eta_j\|^{1/2}$.

The β-increment. Each kernel is affine in β with slope a bounded $\times L_4$ bilinear form $r(O; \eta)$ (e.g. $-j$ for ψ_{naive} and $-(X - g)(g - m)$ for ϕ_{diag}), so $\|\phi(\cdot; \beta', \eta') - \phi(\cdot; \beta, \eta')\| = |\beta' - \beta| \, \|r(\cdot; \eta')\| \le C |\beta' - \beta|$, since $\|r(\cdot; \eta')\| \le C$ by the L_4 bound and boundedness of the nuisances.

Summing the finitely many terms and combining the β-increment (bounding $\phi(\beta', \eta') - \phi(\beta, \eta)$ by $\|\phi(\beta', \eta') - \phi(\beta, \eta')\| + \|\phi(\beta, \eta') - \phi(\beta, \eta)\|$) gives the stated bound with a suitable L. The two convergence claims follow from this bound. □

Proofs for Section 3.2

Proof of Lemma 1, parts (i)–(iii). Part (i) is (A5)–(A6) in the Conventions: $\mathbb{E}[\mathcal{H} \mid W] = \mathbb{E}[\mathbb{E}[U \mid Z, W] \mid W] = \mathbb{E}[U \mid W] = 0$.

(ii) Using (A6), $\tilde{g}_X = g_X - \mathbb{E}[g_X \mid W]$, and that $\tilde{g}_X$ is (Z, W)-measurable,

$$\mathbb{E}[\mathcal{H}\,\tilde{g}_X] = \mathbb{E}\big[\mathbb{E}[U \mid Z, W]\,\tilde{g}_X\big] = \mathbb{E}[U\,\tilde{g}_X].$$

Now $V = g_X - m_X = \tilde{g}_X + (\mathbb{E}[g_X \mid W] - m_X)$, and the second summand is W-measurable, so by (A5), $\mathbb{E}[U(\mathbb{E}[g_X \mid W] - m_X)] = \mathbb{E}[(\mathbb{E}[g_X \mid W] - m_X)\mathbb{E}[U \mid W]] = 0$; hence $\mathbb{E}[U\tilde{g}_X] = \mathbb{E}[UV]$. By the definition (9) of β_0 as the least-squares coefficient of $Y - l_Y$ on V,

$$\mathbb{E}[UV] = \mathbb{E}\big[(Y - l_Y)V\big] - \beta_0\,\mathbb{E}\big[(X - m_X)V\big] = \beta_0 J_0 - \beta_0 J_0 = 0,$$

because $\mathbb{E}[(Y - l_Y)V] = \beta_0\,\mathbb{E}[V^2] = \beta_0 J_0$ by (9) and $\mathbb{E}[(X - m_X)V] = \mathbb{E}[V \cdot V] + \mathbb{E}[(X - g_X)V]$; the last term is $\mathbb{E}[eV] = 0$ by (A7) since V is (Z, W)-measurable, so $\mathbb{E}[(X - m_X)V] = \mathbb{E}[V^2] = J_0$. Thus $\mathbb{E}[\mathcal{H}\tilde{g}_X] = \mathbb{E}[UV] = 0$.

(iii) $\mathcal{G} = \mathbb{R}\tilde{g}_X \oplus L_2(\sigma(W))$ is the sum of a one-dimensional subspace and the closed subspace $L_2(\sigma(W)) \subset L_2(\sigma(Z, W))$. The sum is L_2-orthogonal: for $d \in L_2(\sigma(W))$, $\mathbb{E}[\tilde{g}_X d] = \mathbb{E}[d\,\mathbb{E}[\tilde{g}_X \mid W]] = 0$ since $\mathbb{E}[\tilde{g}_X \mid W] = 0$. An orthogonal sum of a finite-dimensional subspace and a closed subspace is closed, so $\mathcal{G}$ is closed and $\Pi_{\mathcal{G}}$ is well defined. By (i), $\mathcal{H} \perp L_2(\sigma(W))$ (take $d = \mathbb{E}[\mathcal{H} \mid W]$-projections: $\mathbb{E}[\mathcal{H}d] = \mathbb{E}[d\mathbb{E}[\mathcal{H} \mid W]] = 0$); by (ii), $\mathcal{H} \perp \tilde{g}_X$. Hence $\mathcal{H} \perp \mathcal{G}$, so $\mathbb{E}[\mathcal{H}\Pi_{\mathcal{G}}\delta_g] = 0$ and $\mathbb{E}[\mathcal{H}\delta_g] = \mathbb{E}[\mathcal{H}(I - \Pi_{\mathcal{G}})\delta_g]$; the bound is Cauchy–Schwarz. □

The next lemma decomposes the heterogeneity residual into the observable and selection-on-gains components used in Section 3.2.

Lemma A3 (Sources of the Heterogeneity Residual)**.** *Let Assumptions 1 and 2 hold and suppose* $\mathbb{E}[\Delta^2] + \mathbb{E}[X^2\Delta^2] < \infty$. *Define* $\bar{\Delta}_W := \mathbb{E}[\Delta \mid W]$ *and the selection-on-gains kernel* $\kappa(Z, W) := \mathrm{Cov}(X, \Delta \mid Z, W)$, *with* $\tilde{\kappa} := \kappa - \mathbb{E}[\kappa \mid W]$. *Then:*

(i) $\mathcal{H} \;=\; \big(\bar{\Delta}_W - \beta_0\big)\,\tilde{g}_X \;+\; \tilde{\kappa}$;

(ii) if $\kappa = 0$ *a.s. (in particular if* $X \perp\!\!\!\perp \Delta \mid (Z, W)$*), then, with* $v(W) := \mathbb{E}[\tilde{g}_X^2 \mid W]$,

$$\beta_0 \;=\; \frac{\mathbb{E}\big[v(W)\,\bar{\Delta}_W\big]}{\mathbb{E}[v(W)]}, \qquad \mathbb{E}[\mathcal{H}^2] \;=\; \mathbb{E}\big[v(W)\,\big(\bar{\Delta}_W - \beta_0\big)^2\big];$$

(iii) consequently, the condition of Definition 1 holds, and the naive score is non-orthogonal, whenever the conditional average effect $\bar{\Delta}_W$ *varies across covariate strata that the instrument moves* ($v(W) > 0$)*, even in the complete absence of selection on gains.*

Proof of Lemma A3. (i) By Assumption 1 ($Z \perp\!\!\!\perp (Y(0), \Delta) \mid W$), for any (Z, W)-measurable conditioning, $\mathbb{E}[\Delta \mid Z, W] = \mathbb{E}[\Delta \mid W] = \bar{\Delta}_W$. Decompose the reduced-form residual using (A1):

$$q_Y - l_Y = \mathbb{E}[X\Delta \mid Z, W] - \mathbb{E}[X\Delta \mid W].$$

Write $\mathbb{E}[X\Delta \mid Z, W] = \mathbb{E}[X \mid Z, W]\,\mathbb{E}[\Delta \mid Z, W] + \mathrm{Cov}(X, \Delta \mid Z, W) = g_X\,\bar{\Delta}_W + \kappa$, and $\mathbb{E}[X\Delta \mid W] = \mathbb{E}[X \mid W]\bar{\Delta}_W + \mathrm{Cov}(X, \Delta \mid W) = m_X\bar{\Delta}_W + \mathbb{E}[\kappa \mid W] + \mathrm{Cov}(\mathbb{E}[X \mid Z, W], \bar{\Delta}_W \mid W)$. Because $\bar{\Delta}_W$ is W-measurable, $\mathrm{Cov}(X, \Delta \mid W) = \mathbb{E}[\mathrm{Cov}(X, \Delta \mid Z, W) \mid W] + \mathrm{Cov}(g_X, \bar{\Delta}_W \mid W) = \mathbb{E}[\kappa \mid W] + \bar{\Delta}_W \cdot 0$; more directly, subtracting,

$$q_Y - l_Y = (g_X - m_X)\,\bar{\Delta}_W + \big(\kappa - \mathbb{E}[\kappa \mid W]\big) = \bar{\Delta}_W\,V + \tilde{\kappa}.$$

Then $\mathcal{H} = (q_Y - l_Y) - \beta_0 V = (\bar{\Delta}_W - \beta_0)V + \tilde{\kappa}$. Since $m_X = \mathbb{E}[X \mid W] = \mathbb{E}[g_X \mid W]$ by the tower property, $V = g_X - m_X = g_X - \mathbb{E}[g_X \mid W] = \tilde{g}_X$ exactly, so $\mathcal{H} = (\bar{\Delta}_W - \beta_0)\tilde{g}_X + \tilde{\kappa}$, the stated identity.

(ii) If $\kappa \equiv 0$ then $\tilde{\kappa} = 0$ and $\mathcal{H} = (\bar{\Delta}_W - \beta_0)\tilde{g}_X$. By Lemma 1(ii), $0 = \mathbb{E}[\mathcal{H}\tilde{g}_X] = \mathbb{E}[(\bar{\Delta}_W - \beta_0)\tilde{g}_X^2] = \mathbb{E}[(\bar{\Delta}_W - \beta_0)v(W)]$, using $\mathbb{E}[\tilde{g}_X^2 \mid W] = v(W)$ and W-measurability of $\bar{\Delta}_W - \beta_0$. Solving gives $\beta_0 = \mathbb{E}[v(W)\bar{\Delta}_W]/\mathbb{E}[v(W)]$. Substitution into $\mathcal{H}$ gives $\mathbb{E}[\mathcal{H}^2] = \mathbb{E}[(\bar{\Delta}_W - \beta_0)^2\tilde{g}_X^2] = \mathbb{E}[v(W)(\bar{\Delta}_W - \beta_0)^2]$.

(iii) By (ii), $\mathbb{E}[\mathcal{H}^2] = \mathbb{E}[v(W)(\bar{\Delta}_W - \beta_0)^2] > 0$ whenever $\bar{\Delta}_W$ is non-constant on $\{v(W) > 0\}$. Definition 1 states $\mathbb{P}(\mathcal{H} \neq 0) > 0$, equivalently $\mathbb{E}[\mathcal{H}^2] > 0$, and this can therefore occur with $\kappa \equiv 0$. □

Proof of Lemma 1, part (iv). For $g' = ag + d$ with d W-measurable, $\widetilde{g'} = g' - \mathbb{E}[g' \mid W] = a(g - \mathbb{E}[g \mid W]) + (d - d) = a\tilde{g}$. Hence $\mathbb{E}[\tilde{X}\widetilde{g'}] = a\mathbb{E}[\tilde{X}\tilde{g}]$ and $\mathbb{E}[\tilde{Y}\widetilde{g'}] = a\mathbb{E}[\tilde{Y}\tilde{g}]$, so $\beta_{IV}(g') = \mathbb{E}[\tilde{Y}\widetilde{g'}]/\mathbb{E}[\tilde{X}\widetilde{g'}] = (a\mathbb{E}[\tilde{Y}\tilde{g}])/(a\mathbb{E}[\tilde{X}\tilde{g}]) = \beta_{IV}(g)$, the factor $a \neq 0$ cancelling. In particular $\mathbb{E}[\tilde{X}\widetilde{g'}] = a\mathbb{E}[\tilde{X}\tilde{g}] \neq 0$. □

Proof of Theorem 2, parts (i)–(ii). Write $\delta := \delta_g$ and recall $\beta_{IV}(g) = \mathbb{E}[\tilde{Y}\tilde{g}]/\mathbb{E}[\tilde{X}\tilde{g}]$ with $\tilde{\cdot}$ the W-demeaning. For $g_X + \delta$, linearity of demeaning gives $\widetilde{g_X + \delta} = \tilde{g}_X + \tilde{\delta}$, and, arguing as in (A7) (numerator/denominator inner products depend on δ only through its (Z, W)-measurable self),

$$\mathbb{E}[\tilde{X}\,\widetilde{(g_X + \delta)}] = \mathbb{E}[\tilde{X}\tilde{g}_X] + \mathbb{E}[\tilde{X}\tilde{\delta}] = J_0 + \mathbb{E}[\tilde{g}_X\delta],$$

since $\mathbb{E}[\tilde{X}\tilde{g}_X] = \mathbb{E}[(X - m_X)\tilde{g}_X] = \mathbb{E}[V\tilde{g}_X] + \mathbb{E}[e\tilde{g}_X] = \mathbb{E}[\tilde{g}_X^2] = J_0$ (the cross term $\mathbb{E}[e\tilde{g}_X] = 0$

by (A7), and $\mathbb{E}[V\tilde{g}_X] = \mathbb{E}[\tilde{g}_X^2] = J_0$ as in the proof of Lemma 1(ii)); and $\mathbb{E}[\tilde{X}\tilde{\delta}] = \mathbb{E}[\tilde{g}_X\delta]$ likewise. For the numerator, $\mathbb{E}[\tilde{Y}\,(g_X+\delta)] = \mathbb{E}[\tilde{Y}\tilde{g}_X] + \mathbb{E}[\tilde{Y}\tilde{\delta}]$. Now $\mathbb{E}[\tilde{Y}\tilde{g}_X] = \beta_0 J_0$ by definition of β_0 (equivalently $\beta_{IV}(g_X) = \beta_0$), and

$$\mathbb{E}[\tilde{Y}\tilde{\delta}] = \mathbb{E}[(Y-l_Y)\delta] = \mathbb{E}[q_Y\delta - l_Y\delta] = \mathbb{E}[(q_Y-l_Y)\delta] = \mathbb{E}[(\mathcal{H}+\beta_0 V)\delta] = \mathbb{E}[\mathcal{H}\delta] + \beta_0\mathbb{E}[\tilde{g}_X\delta],$$

using $\mathbb{E}[(Y-l_Y)\delta] = \mathbb{E}[(\mathbb{E}[Y \mid Z,W]-l_Y)\delta] = \mathbb{E}[(q_Y-l_Y)\delta]$ for (Z,W)-measurable δ, $\mathcal{H} = (q_Y-l_Y) - \beta_0 V$, and $\mathbb{E}[V\delta] = \mathbb{E}[\tilde{g}_X\delta]$ (same demeaning argument). Therefore

$$\beta_{IV}(g_X+\delta) = \frac{\beta_0 J_0 + \mathbb{E}[\mathcal{H}\delta] + \beta_0\mathbb{E}[\tilde{g}_X\delta]}{J_0 + \mathbb{E}[\tilde{g}_X\delta]} = \beta_0 + \frac{\mathbb{E}[\mathcal{H}\delta]}{J_0 + \mathbb{E}[\tilde{g}_X\delta]},$$

which is (14), proving (i).

(ii) Let $D := J_0 + \mathbb{E}[\tilde{g}_X\delta]$. By Cauchy–Schwarz $|\mathbb{E}[\tilde{g}_X\delta]| \le \|\tilde{g}_X\|\,\|\delta\|$, so if $\|\delta\| < J_0/\|\tilde{g}_X\|$ then $D \ge J_0 - \|\tilde{g}_X\|\|\delta\| > 0$. Subtracting the linear term,

$$\beta_{IV}(g_X+\delta) - \beta_0 - J_0^{-1}\mathbb{E}[\mathcal{H}\delta] = \mathbb{E}[\mathcal{H}\delta]\Big(\frac{1}{D} - \frac{1}{J_0}\Big) = -\frac{\mathbb{E}[\mathcal{H}\delta]\,\mathbb{E}[\tilde{g}_X\delta]}{J_0\,D}.$$

Bounding $|\mathbb{E}[\mathcal{H}\delta]| \le \|\mathcal{H}\|\|\delta\|$, $|\mathbb{E}[\tilde{g}_X\delta]| \le \|\tilde{g}_X\|\|\delta\|$ and $D \ge J_0 - \|\tilde{g}_X\|\|\delta\|$ gives, whenever $\|\delta\| < J_0/\|\tilde{g}_X\|$, the second-order remainder bound

$$\Big|\beta_{IV}(g_X+\delta) - \beta_0 - J_0^{-1}\mathbb{E}[\mathcal{H}\delta]\Big| \;\le\; \frac{\|\mathcal{H}\|\,\|\tilde{g}_X\|}{J_0\big(J_0 - \|\tilde{g}_X\|\,\|\delta\|\big)}\,\|\delta\|^2; \tag{A9}$$

since this is $O(\|\delta\|^2)$, β_{IV} is Fréchet differentiable at g_X with derivative $\delta \mapsto J_0^{-1}\mathbb{E}[\mathcal{H}\delta]$ (a bounded linear functional by Cauchy–Schwarz).

The null-space form in (ii) follows by substituting Lemma 1(iii), $\mathbb{E}[\mathcal{H}\delta] = \mathbb{E}[\mathcal{H}(I-\Pi_{\mathcal{G}})\delta]$, into (14). More explicitly, write $\delta = r + c\tilde{g}_X + d$, where $r := (I-\Pi_{\mathcal{G}})\delta$, $c := \mathbb{E}[\tilde{g}_X\delta]/J_0$, and $d \in L_2(\sigma(W))$. Then

$$\beta_{IV}(g_X+\delta) - \beta_0 = \frac{\mathbb{E}[\mathcal{H}r]}{(1+c)J_0}, \qquad c \neq -1.$$

Thus the numerator and the Fréchet derivative depend only on r, while the exact finite drift also depends on the retained signal scale $1+c$. For $\delta \in \mathcal{G}$, $r = 0$, so the drift vanishes whenever $c \neq -1$; when $c = -1$, the perturbed denominator is zero and $\beta_{IV}(g_X+\delta)$ is undefined. □

The exact second-order expansions (Lemma A4)

Lemma A4 (Exact Second-Order Moment Expansions)**.** *Let Assumptions 1 and 2 hold with $g = g_X$. For any admissible nuisance value $\eta = (l, m, q, g)$ with square-integrable coordinates of the correct measurability, writing $\Delta\eta := \eta - \eta_0$ componentwise,*

$$\begin{aligned}\mathbb{E}\big[\psi_{\text{naive}}(O;\beta_0,\eta)\big] &= \mathbb{E}[\mathcal{H}\,\Delta g_X] + \mathbb{E}\big[(-\Delta l_Y + \beta_0\,\Delta m_X)(\Delta g_X - \Delta m_X)\big],\\ \mathbb{E}\big[\psi_{\text{CRC}}(O;\beta_0,\eta)\big] &= \mathbb{E}\big[\Delta l_Y\,\Delta m_X - \Delta g_X\,\Delta q_Y\big] - \beta_0\,\mathbb{E}\big[(\Delta m_X)^2 - (\Delta g_X)^2\big],\end{aligned}$$

Both identities hold for every admissible η, not only near η_0. The naive expansion contains the linear term $\mathbb{E}[\mathcal{H}\,\Delta g_X]$ from Proposition 3; the CRC expansion is quadratic in $\eta - \eta_0$, as Theorem 3(i) requires.

Proof of Lemma A4. Write $\Delta l := \Delta l_Y$, $\Delta m := \Delta m_X$, $\Delta q := \Delta q_Y$, $\Delta g := \Delta g_X$ for the coordinates of $\eta - \eta_0$, and recall U, V, e and the residuals $R := Y - l_Y$ (with $\mathbb{E}[R \mid W] = 0$, $\mathbb{E}[R \mid Z, W] = q_Y - l_Y =: P$) and $\mathcal{H} = P - \beta_0 V$.

Naive score. At β_0,

$$Y - l - \beta_0(X - m) = U - \Delta l + \beta_0\Delta m, \qquad g - m = V + \Delta g - \Delta m,$$

so $\psi_{\text{naive}}(O;\beta_0,\eta) = (U - \Delta l + \beta_0\Delta m)(V + \Delta g - \Delta m)$. Take expectations and expand into four inner products:

- $\mathbb{E}[UV] = 0$ (proof of Lemma 1(ii));
- $\mathbb{E}[U(\Delta g - \Delta m)] = \mathbb{E}[\mathcal{H}\,\Delta g]$: indeed $\mathbb{E}[U\Delta g] = \mathbb{E}[\mathbb{E}[U \mid Z, W]\Delta g] = \mathbb{E}[\mathcal{H}\Delta g]$ by (A6), and $\mathbb{E}[U\Delta m] = \mathbb{E}[\mathbb{E}[U \mid W]\Delta m] = 0$ by (A5);
- $\mathbb{E}[(-\Delta l + \beta_0\Delta m)V] = \mathbb{E}[(-\Delta l + \beta_0\Delta m)\,\mathbb{E}[V \mid W]] = 0$, since $\mathbb{E}[V \mid W] = \mathbb{E}[g_X \mid W] - m_X = 0$ (as $m_X = \mathbb{E}[X \mid W] = \mathbb{E}[\mathbb{E}[X \mid Z, W] \mid W] = \mathbb{E}[g_X \mid W]$) and $-\Delta l + \beta_0\Delta m$ is W-measurable;
- the remaining product $\mathbb{E}[(-\Delta l + \beta_0\Delta m)(\Delta g - \Delta m)]$ is retained.

Summing the four terms proves the first identity.

CRC score. We compute $\mathbb{E}[\psi_Y(O;\eta)]$ and $\mathbb{E}[\psi_X(O;\eta)]$ separately (neither depends on β). Using $Y-l = R-\Delta l$, $g-m = V+\Delta g-\Delta m$, $X-g = e-\Delta g$, $q-l = P+\Delta q-\Delta l$:

$$\begin{aligned}\mathbb{E}[(Y-l)(g-m)] &= \mathbb{E}[(R-\Delta l)(V+\Delta g-\Delta m)]\\ &= \underbrace{\mathbb{E}[RV]}_{=\beta_0 J_0} + \underbrace{\mathbb{E}[R(\Delta g-\Delta m)]}_{=\mathbb{E}[P\Delta g]} - \underbrace{\mathbb{E}[\Delta l\, V]}_{=0} - \mathbb{E}[\Delta l(\Delta g-\Delta m)]\\ &= \beta_0 J_0 + \mathbb{E}[P\Delta g] - \mathbb{E}[\Delta l\Delta g] + \mathbb{E}[\Delta l\Delta m],\end{aligned}$$

where $\mathbb{E}[RV] = \mathbb{E}[(Y-l_Y)V] = \beta_0 J_0$ by (9); $\mathbb{E}[R\Delta g] = \mathbb{E}[\mathbb{E}[R \mid Z,W]\Delta g] = \mathbb{E}[P\Delta g]$ and $\mathbb{E}[R\Delta m] = \mathbb{E}[\mathbb{E}[R \mid W]\Delta m] = 0$; and $\mathbb{E}[\Delta l V] = \mathbb{E}[\Delta l\,\mathbb{E}[V \mid W]] = 0$. Next,

$$\begin{aligned}\mathbb{E}[(X-g)(q-l)] &= \mathbb{E}[(e-\Delta g)(P+\Delta q-\Delta l)]\\ &= \underbrace{\mathbb{E}[eP]}_{=0} + \underbrace{\mathbb{E}[e(\Delta q-\Delta l)]}_{=0} - \mathbb{E}[\Delta g\, P] - \mathbb{E}[\Delta g(\Delta q-\Delta l)]\\ &= -\mathbb{E}[P\Delta g] - \mathbb{E}[\Delta g\Delta q] + \mathbb{E}[\Delta g\Delta l],\end{aligned}$$

the first two terms vanishing by (A7) ($P, \Delta q, \Delta l$ are (Z,W)-measurable). Adding, the $\mathbb{E}[P\Delta g]$ terms cancel and $-\mathbb{E}[\Delta l\Delta g] + \mathbb{E}[\Delta g\Delta l] = 0$, leaving

$$\mathbb{E}[\psi_Y(O;\eta)] = \beta_0 J_0 + \mathbb{E}[\Delta l\Delta m - \Delta g\Delta q].$$

For $\psi_X = (g-m)(2X-m-g)$, use $2X-m-g = 2e+V-\Delta m-\Delta g$ (since $2X-m_X-g_X = 2e+V$):

$$\begin{aligned}\mathbb{E}[\psi_X(O;\eta)] &= \mathbb{E}[(V+\Delta g-\Delta m)(2e+V-\Delta m-\Delta g)]\\ &= 2\,\underbrace{\mathbb{E}[e(V+\Delta g-\Delta m)]}_{=0} + \mathbb{E}[(V+\Delta g-\Delta m)(V-\Delta m-\Delta g)]\\ &= \mathbb{E}[V^2] - 2\,\underbrace{\mathbb{E}[V\Delta m]}_{=0} - \mathbb{E}[(\Delta g)^2] + \mathbb{E}[(\Delta m)^2] = J_0 + \mathbb{E}[(\Delta m)^2 - (\Delta g)^2],\end{aligned}$$

where $\mathbb{E}[e(\cdot)] = 0$ by (A7) ($V, \Delta g, \Delta m$ are (Z,W)-measurable), the cross terms $-\mathbb{E}[V\Delta g] + \mathbb{E}[\Delta g V]$ and $-\mathbb{E}[\Delta m V] - \mathbb{E}[V\Delta m]$ combine as shown, and $\mathbb{E}[V\Delta m] = \mathbb{E}[\Delta m\,\mathbb{E}[V \mid W]] = 0$. Therefore

$$\mathbb{E}[\psi_{\mathrm{CRC}}(O;\beta_0,\eta)] = \mathbb{E}[\psi_Y] - \beta_0\mathbb{E}[\psi_X] = \mathbb{E}[\Delta l\Delta m - \Delta g\Delta q] - \beta_0\,\mathbb{E}[(\Delta m)^2 - (\Delta g)^2],$$

which proves the second identity. □

Proof of Theorem 2, parts (iii)–(iv). Write $\psi := \psi_{\text{naive}}$ and let $j(O;\eta) := (X-m)(g-m)$ be its negative β-slope. Affinity gives $\psi(O;\beta,\eta) = \psi(O;\beta',\eta) - (\beta-\beta')\, j(O;\eta)$ for all β, β'. Also, since $m_X = \mathbb{E}[g_X \mid W]$,

$$V = g_X - m_X = g_X - \mathbb{E}[g_X \mid W] = \tilde{g}_X, \tag{A10}$$

so $\mathbb{E}[V\Delta g_k] = \mathbb{E}[\tilde{g}_X \Delta g_k]$. The estimator solves $N^{-1}\sum_k \sum_{i\in\mathcal{I}_k} \psi(O_i;\hat{\beta}_{\text{naive}},\hat{\eta}_k) = 0$; by affinity,

$$\hat{\beta}_{\text{naive}} = \frac{\hat{A}}{\hat{J}_{\text{nv}}}, \quad \hat{A} := \frac{1}{N}\sum_k \sum_{i\in\mathcal{I}_k} (Y_i - \hat{l}_{k,i})(\hat{g}_{k,i} - \hat{m}_{k,i}), \quad \hat{J}_{\text{nv}} := \frac{1}{N}\sum_k \sum_{i\in\mathcal{I}_k} (X_i - \hat{m}_{k,i})(\hat{g}_{k,i} - \hat{m}_{k,i}). \tag{A11}$$

Step 1 (pseudo-true values by fold and after aggregation). Condition on $\mathcal{D}_{-k}$ and apply Lemma A4 to a new observation. This is valid because $\hat{\eta}_k$ is $\mathcal{D}_{-k}$-measurable and the identity holds for every fixed admissible η. To shorten the fold notation, set $\Delta l_k := \hat{l}_{Y,k} - l_Y$, $\Delta m_k := \hat{m}_{X,k} - m_X$, $\Delta q_k := \hat{q}_{Y,k} - q_Y$, and $\Delta g_k := \hat{g}_{X,k} - g_X$. Write $D_k := \mathbb{E}[\mathcal{H}\,\Delta g_k \mid \mathcal{D}_{-k}]$ and $r_{2,k} := \mathbb{E}[(-\Delta l_k + \beta_0 \Delta m_k)(\Delta g_k - \Delta m_k) \mid \mathcal{D}_{-k}]$. The analogous calculation for j in the proof of Lemma A4 gives

$$B_k = J_0 + s_k, \qquad s_k := \mathbb{E}[\tilde{g}_X \Delta g_k \mid \mathcal{D}_{-k}] - \mathbb{E}[\Delta m_k(\Delta g_k - \Delta m_k) \mid \mathcal{D}_{-k}],$$

and the conditional identities are

$$A_k - \beta_0 B_k = \mathbb{E}[\psi(O;\beta_0,\hat{\eta}_k) \mid \mathcal{D}_{-k}] = D_k + r_{2,k}, \qquad \beta_k^* - \beta_0 = \frac{D_k + r_{2,k}}{J_0 + s_k}.$$

By Cauchy–Schwarz, $|D_k| \le \|\mathcal{H}\|\,\rho_{g,k}$, $|s_k| \le \|\tilde{g}_X\|\,\rho_{g,k} + \rho_{m,k}(\rho_{g,k}+\rho_{m,k})$, and $|r_{2,k}| \le (\rho_{l,k} + |\beta_0|\rho_{m,k})(\rho_{g,k}+\rho_{m,k})$. Assumption 8 gives $\max_k |s_k| = O_{\mathbb{P}}(\rho_N) = o_{\mathbb{P}}(1)$. The arithmetic–geometric inequality $\rho_m\rho_g \le \frac{1}{2}(\rho_m^2 + \rho_g^2)$ and the bound $\rho_{l,k}(\rho_{g,k}+\rho_{m,k}) \le (\rho_{l,k}+\rho_{q,k})(\rho_{m,k}+\rho_{g,k})$ give

$$\max_k |r_{2,k}| \le (1+|\beta_0|)\Big[(\rho_{l,k}+\rho_{q,k})(\rho_{m,k}+\rho_{g,k}) + \tfrac{3}{2}(\rho_{m,k}^2 + \rho_{g,k}^2)\Big]_{\max k} = o_{\mathbb{P}}(N^{-1/2}). \tag{A12}$$

On the event $\mathcal{E}_N := \{\max_k |s_k| \le J_0/2\}$, which has $\mathbb{P}(\mathcal{E}_N) \to 1$, each $B_k \ge J_0/2 > 0$, so β_k^* and the weights $w_k = n_k B_k / \sum_j n_j B_j$ are well defined, and

$$\beta_k^* - \beta_0 - J_0^{-1} D_k = \frac{r_{2,k}}{J_0 + s_k} - \frac{D_k\, s_k}{J_0(J_0 + s_k)}.$$

On $\mathcal{E}_N$ the first term is $\le (2/J_0)|r_{2,k}| = o_{\mathbb{P}}(N^{-1/2})$; the second is, using $|D_k s_k| \le \|\mathcal{H}\|\rho_{g,k}(\|\tilde{g}_X\|\rho_{g,k} + \rho_{m,k}\rho_{g,k} + \rho_{m,k}^2) \le C(\rho_{g,k}^2 + \rho_{m,k}^2)$ (leading term $\|\mathcal{H}\|\|\tilde{g}_X\|\rho_{g,k}^2$, remainder higher order), $\le (2/J_0^2)|D_k s_k| = O_{\mathbb{P}}(\rho_{m,k}^2 + \rho_{g,k}^2) = o_{\mathbb{P}}(N^{-1/2})$. Hence

$$\max_k \left|\beta_k^* - \beta_0 - J_0^{-1} D_k\right| = o_{\mathbb{P}}(N^{-1/2}), \qquad \max_k |\beta_k^* - \beta_0| = O_{\mathbb{P}}(\rho_{g,N}) + o_{\mathbb{P}}(N^{-1/2}), \tag{A13}$$

with $\rho_{g,N} := \max_k \rho_{g,k} = o_{\mathbb{P}}(N^{-1/4})$. Since $\sum_k w_k = 1$, $\beta_N^* - \beta_0 = \sum_k w_k(\beta_k^* - \beta_0) = J_0^{-1}\sum_k w_k D_k + o_{\mathbb{P}}(N^{-1/2})$. With equal folds $n_k = N/K$, $w_k = B_k/\sum_j B_j$ and $w_k - \frac{1}{K} = (K s_k - \sum_j s_j)/(K\sum_j B_j) = O_{\mathbb{P}}(\rho_N)$ on $\mathcal{E}_N$; as $|D_k| \le \|\mathcal{H}\|\rho_{g,k}$ and $(w_k - \frac{1}{K})$ depends only on $\{s_j\}$ (hence only on ρ_g, ρ_m), $\sum_k (w_k - \frac{1}{K})D_k = O_{\mathbb{P}}(\rho_{g,N}^2 + \rho_{m,N}\rho_{g,N}) = o_{\mathbb{P}}(N^{-1/2})$. Therefore

$$\beta_N^* - \beta_0 = J_0^{-1}\bar{D}_N + o_{\mathbb{P}}(N^{-1/2}), \qquad |\bar{D}_N| \le \|\mathcal{H}\|\,\rho_{g,N} = o_{\mathbb{P}}(N^{-1/4}), \tag{A14}$$

which is the first display of part (iii). For the second display of (iii), apply Theorem 2(ii) to $\hat{g}_k$ (true l_Y, m_X, q_Y, only g perturbed): $\beta_{IV}(\hat{g}_k) - \beta_0 = J_0^{-1}D_k + O_{\mathbb{P}}(\rho_{g,k}^2) = J_0^{-1}D_k + o_{\mathbb{P}}(N^{-1/2})$, so $\beta_{IV}(\hat{g}_k)$ and β_k^* share the expansion (A13); taking w-weighted averages, $\sum_k w_k \beta_{IV}(\hat{g}_k) = \beta_0 + J_0^{-1}\bar{D}_N + o_{\mathbb{P}}(N^{-1/2}) = \beta_N^* + o_{\mathbb{P}}(N^{-1/2})$.

Step 2 (exact affine identity for the estimator). By affinity and $\sum_k \sum_i \psi(O_i; \hat{\beta}_{\text{naive}}, \hat{\eta}_k) = 0$,

$$\hat{\beta}_{\text{naive}} - \beta_N^* = \hat{J}_{\text{nv}}^{-1} \frac{1}{N} \sum_k \sum_{i \in \mathcal{I}_k} \psi(O_i; \beta_N^*, \hat{\eta}_k), \tag{A15}$$

on $\{\hat{J}_{\text{nv}} \neq 0\}$.

Step 3 (recentering fold by fold). Using affinity again, $\psi(O_i; \beta_N^*, \hat{\eta}_k) = \psi(O_i; \beta_k^*, \hat{\eta}_k) - (\beta_N^* - \beta_k^*)j(O_i; \hat{\eta}_k)$. Let $\bar{j}_k := \frac{1}{n_k}\sum_{i \in \mathcal{I}_k} j(O_i; \hat{\eta}_k)$, so $\mathbb{E}[\bar{j}_k \mid \mathcal{D}_{-k}] = B_k$ and, since $\|j(\cdot; \hat{\eta}_k)\| = O_{\mathbb{P}}(1)$ (finite $\|j(\cdot; \eta_0)\|$ by the L_4 bound on X, plus Lemma A2), $\text{Var}(\bar{j}_k \mid \mathcal{D}_{-k}) = O_{\mathbb{P}}(n_k^{-1})$, whence $\bar{j}_k - B_k = O_{\mathbb{P}}(N^{-1/2})$ by conditional Chebyshev. Then

$$\frac{1}{N}\sum_k \sum_i \psi(O_i; \beta_N^*, \hat{\eta}_k) = \frac{1}{N}\sum_k \sum_i \psi(O_i; \beta_k^*, \hat{\eta}_k) - \sum_k \frac{n_k}{N}(\beta_N^* - \beta_k^*)\bar{j}_k,$$

and the last sum equals $\sum_k \frac{n_k}{N}(\beta_N^* - \beta_k^*)B_k + \sum_k \frac{n_k}{N}(\beta_N^* - \beta_k^*)(\bar{j}_k - B_k)$. The first piece is $\frac{1}{N}\sum_k n_k B_k(\beta_N^* - \beta_k^*) = 0$, because $\sum_k n_k B_k \beta_k^* = \sum_k n_k A_k = (\sum_j n_j B_j)\beta_N^*$ by the definition of β_N^*. The second piece is bounded by $\max_k |\beta_N^* - \beta_k^*| \cdot \max_k |\bar{j}_k - B_k| = (O_{\mathbb{P}}(\rho_{g,N}) +$

$o_{\mathbb{P}}(N^{-1/2})) \cdot O_{\mathbb{P}}(N^{-1/2}) = o_{\mathbb{P}}(N^{-1/2})$, using (A13) and $\rho_{g,N} = o_{\mathbb{P}}(1)$. Hence

$$\frac{1}{N}\sum_{k}\sum_{i}\psi(O_i;\beta_N^*,\hat{\eta}_k) = \frac{1}{N}\sum_{k}\sum_{i}\psi(O_i;\beta_k^*,\hat{\eta}_k) + o_{\mathbb{P}}(N^{-1/2}). \tag{A16}$$

Step 4 (stochastic equicontinuity). For $i \in \mathcal{I}_k$ set $D_{ik} := \psi(O_i;\beta_k^*,\hat{\eta}_k) - \psi(O_i;\beta_0,\eta_0)$; note $\psi(O_i;\beta_0,\eta_0) = U_iV_i$. Conditional on $\mathcal{D}_{-k}$ the $\{O_i\}_{i\in\mathcal{I}_k}$ are i.i.d. and $(\beta_k^*,\hat{\eta}_k)$ is fixed, so $\mathbb{E}[D_{ik} \mid \mathcal{D}_{-k}] = \mathbb{E}[\psi(O;\beta_k^*,\hat{\eta}_k) \mid \mathcal{D}_{-k}] - \mathbb{E}[UV] = (A_k - \beta_k^* B_k) - 0 = 0$ by definition of β_k^*. By Lemma A2, $\mathbb{E}[D_{ik}^2 \mid \mathcal{D}_{-k}] = \|\psi(\cdot;\beta_k^*,\hat{\eta}_k) - \psi(\cdot;\beta_0,\eta_0)\|^2 \le L^2(|\beta_k^* - \beta_0| + \sum_j \rho_{j,k}^{1/2})^2 =: \delta_{N,k}^2 = o_{\mathbb{P}}(1)$. Consequently $\mathrm{Var}\big(N^{-1/2}\sum_{i\in\mathcal{I}_k} D_{ik} \mid \mathcal{D}_{-k}\big) = \frac{n_k}{N}\mathrm{Var}(D_{ik} \mid \mathcal{D}_{-k}) \le \frac{1}{K}\delta_{N,k}^2 = o_{\mathbb{P}}(1)$, and conditional Chebyshev followed by bounded convergence gives $N^{-1/2}\sum_{i\in\mathcal{I}_k} D_{ik} = o_{\mathbb{P}}(1)$; summing over the K fixed folds,

$$\frac{1}{N}\sum_{k}\sum_{i\in\mathcal{I}_k}\psi(O_i;\beta_k^*,\hat{\eta}_k) = \frac{1}{N}\sum_{i=1}^{N} U_iV_i + o_{\mathbb{P}}(N^{-1/2}). \tag{A17}$$

Step 5 (central limit theorem and Jacobian). Chaining (A15)–(A17),

$$\sqrt{N}\big(\hat{\beta}_{\text{naive}} - \beta_N^*\big) = \hat{J}_{\text{nv}}^{-1}\Big(N^{-1/2}\sum_{i=1}^{N} U_iV_i + o_{\mathbb{P}}(1)\Big).$$

The summands U_iV_i are i.i.d. with $\mathbb{E}[UV] = 0$ and variance Ω_0^{nv}. Assumption 7(i) gives finiteness because $\mathbb{E}[U^2V^2] \le \|U\|_{\mathbb{P},4}^2\|V\|_{\mathbb{P},4}^2 < \infty$, and part (iv) imposes $\Omega_0^{\text{nv}} > 0$. The Lindeberg–Lévy theorem therefore gives $N^{-1/2}\sum_i U_iV_i \xrightarrow{d} \mathcal{N}(0,\Omega_0^{\text{nv}})$. For the Jacobian, conditional on $\mathcal{D}_{-k}$, $\mathbb{E}[\bar{j}_k \mid \mathcal{D}_{-k}] = B_k = J_0 + o_{\mathbb{P}}(1)$ and $\mathrm{Var}(\bar{j}_k \mid \mathcal{D}_{-k}) = o_{\mathbb{P}}(1)$. Hence $\bar{j}_k = J_0 + o_{\mathbb{P}}(1)$ and $\hat{J}_{\text{nv}} = \sum_k \frac{1}{K}\bar{j}_k \xrightarrow{\mathbb{P}} J_0$. Slutsky gives $\sqrt{N}(\hat{\beta}_{\text{naive}} - \beta_N^*) \xrightarrow{d} \mathcal{N}(0, J_0^{-2}\Omega_0^{\text{nv}})$, the claim in part (iv).

Step 6 (variance estimator; part (iv)). With $\hat{\psi}_i := \psi(O_i;\hat{\beta}_{\text{naive}},\hat{\eta}_k)$ and $\psi_i^0 := U_iV_i$, Cauchy–Schwarz gives $\big|\frac{1}{N}\sum_i \hat{\psi}_i^2 - \frac{1}{N}\sum_i(\psi_i^0)^2\big| \le (\frac{1}{N}\sum_i(\hat{\psi}_i - \psi_i^0)^2)^{1/2}(\frac{1}{N}\sum_i(\hat{\psi}_i + \psi_i^0)^2)^{1/2}$. For the first factor, split $\hat{\psi}_i - \psi_i^0 = -(\hat{\beta}_{\text{naive}} - \beta_k^*)j(O_i;\hat{\eta}_k) + D_{ik}$. The contribution of the first term is $\sum_k \frac{n_k}{N}(\hat{\beta}_{\text{naive}} - \beta_k^*)^2 \cdot \frac{1}{n_k}\sum_{i\in\mathcal{I}_k} j(O_i;\hat{\eta}_k)^2$; here $\hat{\beta}_{\text{naive}} - \beta_k^* = (\hat{\beta}_{\text{naive}} - \beta_N^*) + (\beta_N^* - \beta_k^*) = O_{\mathbb{P}}(N^{-1/2}) + O_{\mathbb{P}}(\rho_{g,N}) = o_{\mathbb{P}}(1)$ and $\frac{1}{n_k}\sum_{i\in\mathcal{I}_k} j(O_i;\hat{\eta}_k)^2 = O_{\mathbb{P}}(1)$, so the contribution is $o_{\mathbb{P}}(1)$. The contribution of D_{ik} is $\frac{1}{N}\sum_k\sum_i D_{ik}^2$, whose conditional mean is $\sum_k \frac{n_k}{N}\mathbb{E}[D_{ik}^2 \mid \mathcal{D}_{-k}] \le \frac{1}{K}\sum_k \delta_{N,k}^2 = o_{\mathbb{P}}(1)$, hence $o_{\mathbb{P}}(1)$ by conditional Markov. Thus the first factor is $o_{\mathbb{P}}(1)$; the second factor is $O_{\mathbb{P}}(1)$ because $\frac{1}{N}\sum_i(\psi_i^0)^2 \xrightarrow{\mathbb{P}} \Omega_0^{\text{nv}}$ (law of large numbers) and the cross term

is controlled by the first factor. Therefore the plug-in variance estimator

$$\hat{\Omega}^{\mathrm{nv}} := \frac{1}{N}\sum_i \hat{\psi}_i^2 = \frac{1}{N}\sum_k \sum_{i\in\mathcal{I}_k} \psi_{\mathrm{naive}}(O_i;\hat{\beta}_{\mathrm{naive}},\hat{\eta}_k)^2 \tag{A18}$$

satisfies $\hat{\Omega}^{\mathrm{nv}} \xrightarrow{\mathbb{P}} \Omega_0^{\mathrm{nv}}$. With $\hat{J}_{\mathrm{nv}} \xrightarrow{\mathbb{P}} J_0$, the studentized statistic $\sqrt{N}(\hat{\beta}_{\mathrm{naive}}-\beta_N^*)/(\hat{J}_{\mathrm{nv}}^{-1}\sqrt{\hat{\Omega}^{\mathrm{nv}}}) \xrightarrow{d} \mathcal{N}(0,1)$, so the nominal interval covers β_N^* with probability tending to $1-\alpha$. □

Proof of Corollary 2. By Theorem 2(iii)–(iv) and (A14),

$$\frac{\sqrt{N}(\hat{\beta}_{\mathrm{naive}}-\beta_0)}{\sigma_{\mathrm{nv}}} = \frac{\sqrt{N}(\hat{\beta}_{\mathrm{naive}}-\beta_N^*)}{\sigma_{\mathrm{nv}}} + \frac{\sqrt{N}\,J_0^{-1}\bar{D}_N}{\sigma_{\mathrm{nv}}} + o_{\mathbb{P}}(1) =: Z_N + \mu_N + o_{\mathbb{P}}(1), \qquad Z_N \xrightarrow{d} \mathcal{N}(0,1).$$

Let $\hat{\sigma}_{\mathrm{nv}}^2 := \hat{J}_{\mathrm{nv}}^{-2}\hat{\Omega}^{\mathrm{nv}} \xrightarrow{\mathbb{P}} \sigma_{\mathrm{nv}}^2$ (Step 6). Coverage of the nominal interval for β_0 equals $\mathbb{P}(|T_N^0| \le z_{1-\alpha/2})$ with $T_N^0 := \sqrt{N}(\hat{\beta}_{\mathrm{naive}}-\beta_0)/\hat{\sigma}_{\mathrm{nv}} = (Z_N+\mu_N)(\sigma_{\mathrm{nv}}/\hat{\sigma}_{\mathrm{nv}}) + o_{\mathbb{P}}(1)$.

(a) If $\mu_N \xrightarrow{\mathbb{P}} 0$, then $T_N^0 \xrightarrow{d} \mathcal{N}(0,1)$ by Slutsky, and coverage $\to \mathbb{P}(|\mathcal{N}(0,1)| \le z_{1-\alpha/2}) = 1-\alpha$.

(b) If $\mu_N \xrightarrow{\mathbb{P}} \mu$ finite, then $\sigma_{\mathrm{nv}}/\hat{\sigma}_{\mathrm{nv}} \xrightarrow{\mathbb{P}} 1$ and $T_N^0 = Z_N + \mu_N + o_{\mathbb{P}}(1) \xrightarrow{d} \mathcal{N}(\mu,1)$, so coverage $\to \mathbb{P}(-z_{1-\alpha/2} \le \mathcal{N}(\mu,1) \le z_{1-\alpha/2}) = \Phi(z_{1-\alpha/2}-\mu) - \Phi(-z_{1-\alpha/2}-\mu) =: m(\mu)$. Writing $m(\mu) = \int_{-z_{1-\alpha/2}}^{z_{1-\alpha/2}} \phi(t-\mu)\,dt$, the mass a standard normal places on a fixed-width interval centered at $-\mu$ is, by symmetry and unimodality of ϕ, strictly maximized at $\mu = 0$; hence $m(\mu) < m(0) = 1-\alpha$ for every $\mu \neq 0$.

(c) If $|\mu_N| \xrightarrow{\mathbb{P}} \infty$, then $|T_N^0| \ge (|\mu_N| - |Z_N| - o_{\mathbb{P}}(1)) \cdot (\sigma_{\mathrm{nv}}/\hat{\sigma}_{\mathrm{nv}}) \xrightarrow{\mathbb{P}} \infty$ since $Z_N = O_{\mathbb{P}}(1)$, so coverage $\to 0$.

Finally, $\mu_N \xrightarrow{\mathbb{P}} 0 \iff \sqrt{N}\bar{D}_N \xrightarrow{\mathbb{P}} 0$ because $J_0, \sigma_{\mathrm{nv}}$ are fixed and positive. This condition makes the average realized loading $\bar{D}_N = \sum_k (n_k/N)\mathbb{E}[\mathcal{H}\Delta g_k \mid \mathcal{D}_{-k}]$ equal to $o_{\mathbb{P}}(N^{-1/2})$. □

Empirical-process supplement for Theorems 2 and 3. If every $\rho_{j,k} = o_{\mathbb{P}}(N^{-1/4})$, then $\rho_{m,k}^2 + \rho_{g,k}^2 = o_{\mathbb{P}}(N^{-1/2})$ and $(\rho_{l,k}+\rho_{q,k})(\rho_{m,k}+\rho_{g,k}) = o_{\mathbb{P}}(N^{-1/2})$. Thus the single rate implies Assumption 8.

Expansion following Theorem 2. Combining parts (iii)–(iv), (A14), and Step 5 gives $\hat{\beta}_{\mathrm{naive}} - \beta_0 = (\hat{\beta}_{\mathrm{naive}} - \beta_N^*) + (\beta_N^* - \beta_0) = J_0^{-1}N^{-1}\sum_i U_i V_i + J_0^{-1}\bar{D}_N + o_{\mathbb{P}}(N^{-1/2})$. Here $\bar{D}_N =$

$\sum_k (n_k/N)\mathbb{E}[\mathcal{H}\Delta g_k \mid \mathcal{D}_{-k}]$ is the cross-fitted counterpart of $\mathbb{E}[\mathcal{H}\Delta g_X]$ and $\psi_{\text{naive}}(O_i;\beta_0,\eta_0) = U_i V_i$. This gives the expansion following the theorem, with Step 4 supplying the empirical-process argument.

Fixed-target estimator. Apply Steps 1–6 with ψ_{CRC} in place of ψ_{naive}, the negative β-slope $j_{\text{CRC}}(O;\eta) := (X-m)(g-m) + (X-g)(g-m) = \psi_X(O;\eta)$, and the fold pseudo-true values $\beta_k^{\text{crc},*}$. Lemma A4 gives $\mathbb{E}[\psi_{\text{CRC}}(O;\beta_0,\hat\eta_k) \mid \mathcal{D}_{-k}] = R_{Y,k} - \beta_0 R_{X,k}$, where $R_{Y,k} = \mathbb{E}[\Delta l_k \Delta m_k - \Delta g_k \Delta q_k \mid \mathcal{D}_{-k}]$ and $R_{X,k} = \mathbb{E}[(\Delta m_k)^2 - (\Delta g_k)^2 \mid \mathcal{D}_{-k}]$. Hence $\beta_k^{\text{crc},*} - \beta_0 = (R_{Y,k} - \beta_0 R_{X,k})/(J_0 + R_{X,k})$. Cauchy–Schwarz gives $|R_{Y,k}| \le \rho_{l,k}\rho_{m,k} + \rho_{g,k}\rho_{q,k}$ and $|R_{X,k}| \le \rho_{m,k}^2 + \rho_{g,k}^2$. The bound $\rho_{l,k}\rho_{m,k} + \rho_{g,k}\rho_{q,k} \le (\rho_{l,k} + \rho_{q,k})(\rho_{m,k} + \rho_{g,k})$ and Assumption 8 make both remainders $o_{\mathbb{P}}(N^{-1/2})$. Therefore $\beta_k^{\text{crc},*} - \beta_0 = o_{\mathbb{P}}(N^{-1/2})$ for every fold and $\beta_N^{\text{crc},*} - \beta_0 = o_{\mathbb{P}}(N^{-1/2})$. There is no linear drift because ψ_{CRC} is Neyman-orthogonal. Steps 2–6 then give $\sqrt{N}(\hat\beta_{\text{CRC}} - \beta_0) \xrightarrow{d} \mathcal{N}(0, J_0^{-2}\Omega_0)$, where $\Omega_0 = \mathbb{E}[\psi_{\text{CRC}}(O;\beta_0,\eta_0)^2]$, and $\hat\Omega \xrightarrow{\mathbb{P}} \Omega_0$. The empirical CRC score slope used in this argument converges to J_0. The estimator $\hat J = N^{-1}\sum_i (\hat g_{X,i} - \hat m_{X,i})^2$ stated in Theorem 3(ii) has the same limit by the cross-fitted law of large numbers and L_2 consistency of $\hat g_X - \hat m_X$. Finally, $R_N = R_Y - \beta_0 R_X$ is the population counterpart of $R_{Y,k} - \beta_0 R_{X,k}$. □

Proofs for Section 4.3: Worst-case bias over a realization ball

Proof of Proposition 4. (i) Upper bound. By Lemma A4 and Cauchy–Schwarz, for $\eta \in \mathcal{T}(r)$,

$$\big|\mathbb{E}[\psi_{\text{naive}}(O;\beta_0,\eta)]\big| \le |\mathbb{E}[\mathcal{H}\Delta g]| + |\mathbb{E}[(-\Delta l + \beta_0 \Delta m)(\Delta g - \Delta m)]| \le \|\mathcal{H}\|\, r_g + (r_l + |\beta_0| r_m)(r_g + r_m),$$

using $\|\Delta g\| \le r_g$, $\|-\Delta l + \beta_0 \Delta m\| \le r_l + |\beta_0| r_m$ and $\|\Delta g - \Delta m\| \le r_g + r_m$.

Lower bound. Take $\eta^\star = (l_Y, m_X, q_Y, g_X + r_g \mathcal{H}/\|\mathcal{H}\|)$, so $\Delta l = \Delta m = \Delta q = 0$ and $\Delta g = r_g \mathcal{H}/\|\mathcal{H}\|$. Then $\|\Delta g\| = r_g \le r_g$ and $\|g_X + \Delta g\|_\infty \le \|g_X\|_\infty + r_g \|\mathcal{H}\|_\infty/\|\mathcal{H}\| \le \|g_X\|_\infty + \|\mathcal{H}\|_\infty/\|\mathcal{H}\| \le \bar C$ (as $r_g \le 1$), so $\eta^\star \in \mathcal{T}(r)$. By Lemma A4, the second term vanishes and $\mathbb{E}[\psi_{\text{naive}}(O;\beta_0,\eta^\star)] = \mathbb{E}[\mathcal{H}\Delta g] = (r_g/\|\mathcal{H}\|)\mathbb{E}[\mathcal{H}^2] = r_g \|\mathcal{H}\|$. Hence the supremum is at least $r_g\|\mathcal{H}\|$.

(ii) By Lemma A4 and Cauchy–Schwarz, $|\mathbb{E}[\psi_{\text{CRC}}(O;\beta_0,\eta)]| \le |\mathbb{E}[\Delta l \Delta m]| + |\mathbb{E}[\Delta g \Delta q]| + |\beta_0|(|\mathbb{E}[(\Delta m)^2]| + |\mathbb{E}[(\Delta g)^2]|) \le r_l r_m + r_g r_q + |\beta_0|(r_m^2 + r_g^2)$.

Finally, along a sequence with $\sqrt{N}r_g \to \infty$, the naive bound $r_g\|\mathcal{H}\|$ satisfies $\sqrt{N}r_g\|\mathcal{H}\| \to \infty$ whenever $\|\mathcal{H}\| > 0$ (Definition 1), while under Assumption 8 the CRC bound is $o(N^{-1/2})$. □

Corollary A1, referenced from Section 4.3, realizes the worst case with an explicit adversarial sequence.

Corollary A1 (An Admissible Adversarial Sequence)**.** *Fix $\gamma \in (1/4, 1/2)$ and define the deterministic first-stage sequence $\hat{g}_X^{(N)} := g_X + c_N\,\mathcal{H}$ with $c_N := N^{-\gamma}$, setting $(\hat{l}_Y, \hat{m}_X, \hat{q}_Y) := (l_Y, m_X, q_Y)$. This sequence satisfies Assumptions 7–8 after enlarging $\bar{C}$. Cross-fitting is automatic because the sequence is data-independent. Along it, let Assumptions 1, 2, and 7 hold, impose the heterogeneity condition $\mathbb{E}[\mathcal{H}^2] > 0$ from Definition 1, and suppose $\Omega_0^{\mathrm{nv}} > 0$ and $\Omega_0 > 0$. Then:*

(i) $\beta_N^* - \beta_0 = c_N\,\mathbb{E}[\mathcal{H}^2]/J_0$ *exactly, for every* N*;*

(ii) $\sqrt{N}\,\big|\hat{\beta}_{\mathrm{naive}} - \beta_0\big| \xrightarrow{\mathbb{P}} \infty$ *and the coverage of the naive interval for* β_0 *tends to* 0*;*

(iii) $\sqrt{N}\,\big(\hat{\beta}_{\mathrm{CRC}} - \beta_0\big) \xrightarrow{d} \mathcal{N}\big(0, J_0^{-2}\Omega_0\big)$ *and the CRC-robust interval's coverage tends to* $1 - \alpha$*.*

Proof of Corollary A1. The sequence $\hat{g}_X^{(N)} = g_X + c_N\mathcal{H}$ has $\rho_{g,k} = \|c_N\mathcal{H}\| = N^{-\gamma}\|\mathcal{H}\| = o(N^{-1/4})$ because $\gamma > 1/4$, and $\rho_{l,k} = \rho_{m,k} = \rho_{q,k} = 0$, so Assumption 8 holds. Boundedness holds after enlarging $\bar{C}$ to $\|g_X\|_\infty + \|\mathcal{H}\|_\infty$. Since the sequence is deterministic, all conditional expectations $\mathbb{E}[\cdot \mid \mathcal{D}_{-k}]$ reduce to ordinary expectations.

(i) With $\Delta l = \Delta m = \Delta q = 0$ and $\Delta g = c_N\mathcal{H}$, the quantities of Step 1 of the proof of Theorem 2 are, for every fold, $r_{2,k} = 0$, $s_k = \mathbb{E}[\tilde{g}_X\, c_N\mathcal{H}] = c_N\mathbb{E}[\mathcal{H}\tilde{g}_X] = 0$ by Lemma 1(ii), and $D_k = \mathbb{E}[\mathcal{H}\, c_N\mathcal{H}] = c_N\mathbb{E}[\mathcal{H}^2]$. Hence $B_k = J_0$ and $\beta_k^* - \beta_0 = D_k/J_0 = c_N\mathbb{E}[\mathcal{H}^2]/J_0$ identically across folds, so $\beta_N^* - \beta_0 = c_N\mathbb{E}[\mathcal{H}^2]/J_0$ exactly for every N.

(ii) By Theorem 2(iv), $\hat{\beta}_{\mathrm{naive}} - \beta_N^* = O_{\mathbb{P}}(N^{-1/2})$, so $\sqrt{N}(\hat{\beta}_{\mathrm{naive}} - \beta_0) = O_{\mathbb{P}}(1) + N^{1/2-\gamma}\mathbb{E}[\mathcal{H}^2]/J_0$; since $\gamma < 1/2$ and $\mathbb{E}[\mathcal{H}^2] > 0$, this diverges in probability. Because $\mu_N = N^{1/2-\gamma}J_0^{-1}\mathbb{E}[\mathcal{H}^2]/\sigma_{\mathrm{nv}} \to \infty$, Corollary 2(c) gives naive coverage $\to 0$.

(iii) For the CRC score, $R_Y = \mathbb{E}[\Delta l\Delta m - \Delta g\Delta q] = 0$ and $R_X = \mathbb{E}[(\Delta m)^2 - (\Delta g)^2] = -c_N^2\mathbb{E}[\mathcal{H}^2]$, so the CRC pseudo-true value satisfies $\beta_N^{\mathrm{crc},*} - \beta_0 = (0 - \beta_0 R_X)/(J_0 + R_X) =$

$\beta_0 c_N^2 \mathbb{E}[\mathcal{H}^2]/(J_0 - c_N^2 \mathbb{E}[\mathcal{H}^2])$. As $2\gamma > 1/2$, $\sqrt{N} c_N^2 = N^{1/2-2\gamma} \to 0$, so $\sqrt{N}(\beta_N^{\mathrm{crc},*} - \beta_0) \to 0$. Theorem 3(ii) then gives $\sqrt{N}(\hat{\beta}_{\mathrm{CRC}} - \beta_0) \xrightarrow{d} \mathcal{N}(0, J_0^{-2}\Omega_0)$ and asymptotic coverage $1-\alpha$. □

Proof for Section 4.1: The score-difference identity

Proof of Lemma 2. From (11), $\psi_{\mathrm{naive}}(O;\beta,\eta) = (Y-l)(g-m) - \beta(X-m)(g-m)$; from (15)–(17), $\psi_{\mathrm{CRC}}(O;\beta,\eta) = (Y-l)(g-m) + (X-g)(q-l) - \beta[(X-m)(g-m) + (X-g)(g-m)]$. Subtracting,

$$\psi_{\mathrm{CRC}} - \psi_{\mathrm{naive}} = (X-g)(q-l) - \beta(X-g)(g-m) = (X-g)\big[(q-l) - \beta(g-m)\big].$$

At (β_0, η_0) this is $e\big[(q_Y - l_Y) - \beta_0(g_X - m_X)\big] = e\,\mathcal{H}$ by definition of $\mathcal{H}$. For the zero-mean property, the bracket $f := (q_Y - l_Y) - \beta(g_X - m_X)$ is (Z,W)-measurable, so $\mathbb{E}[(X-g_X)f] = \mathbb{E}[ef] = 0$ by (A7), for every β. Finally $\mathbb{E}[(e\mathcal{H})^2] = \mathbb{E}[e^2\mathcal{H}^2] = \mathbb{E}[\mathbb{E}[e^2 \mid Z,W]\mathcal{H}^2] = \mathbb{E}[\sigma_X^2(Z,W)\mathcal{H}^2]$, so $e\mathcal{H} = 0$ a.s. iff $\sigma_X^2\mathcal{H}^2 = 0$ a.s. □

Proof for Section 5.1: The Alignment Diagnostic

Proof of Proposition 5. Write $e_i := X_i - g_{X,i}$, $\Delta g_k := \hat{g}_{X,k} - g_X$, and $\Delta_{H,k} := \hat{\mathcal{H}}_k^0 - \mathcal{H}$, where $\hat{\mathcal{H}}_k^0 := (\hat{q}_{Y,k} - \hat{l}_{Y,k}) - \beta_0(\hat{g}_{X,k} - \hat{m}_{X,k})$ is the plug-in of $\mathcal{H}$ at the true β_0; thus $\Delta_{H,k} = (\Delta q_k - \Delta l_k) - \beta_0(\Delta g_k - \Delta m_k)$ is an affine combination of nuisance errors with $\|\Delta_{H,k}\| \le \rho_{q,k} + \rho_{l,k} + |\beta_0|(\rho_{g,k} + \rho_{m,k})$, and $\mathcal{H}, \Delta_{H,k}$ are bounded by a constant $C(\bar{C}, |\beta_0|)$ (Assumption 7(ii)). Since $\hat{\mathcal{H}}_k = \hat{\mathcal{H}}_k^0 - (\hat{\beta}_{\mathrm{CRC}} - \beta_0)(\hat{g}_{X,k} - \hat{m}_{X,k})$,

$$\widehat{\mathcal{A}}_N = \underbrace{\frac{1}{N}\sum_k \sum_{i\in\mathcal{I}_k} (X_i - \hat{g}_{X,k,i})\hat{\mathcal{H}}_{k,i}^0}_{=:\,\mathrm{A}} - (\hat{\beta}_{\mathrm{CRC}} - \beta_0) \underbrace{\frac{1}{N}\sum_k \sum_{i\in\mathcal{I}_k} (X_i - \hat{g}_{X,k,i})(\hat{g}_{X,k,i} - \hat{m}_{X,k,i})}_{=:\,\mathrm{B}}. \tag{A19}$$

Term B (the β-leakage). The conditional mean of a summand is $\mathbb{E}[(e-\Delta g_k)(V + \Delta g_k - \Delta m_k) \mid \mathcal{D}_{-k}] = -\mathbb{E}[\Delta g_k V] - \mathbb{E}[(\Delta g_k)^2] + \mathbb{E}[\Delta g_k \Delta m_k] = O_{\mathbb{P}}(\rho_{g,k})$ (the e-terms vanish by (A7)); the empirical fluctuation about it is $O_{\mathbb{P}}(N^{-1/2})$ by bounded conditional variance (finite via the L_4 bound on X). Hence $\mathrm{B} = O_{\mathbb{P}}(\rho_{g,N}) + O_{\mathbb{P}}(N^{-1/2}) = o_{\mathbb{P}}(1)$, and with $\hat{\beta}_{\mathrm{CRC}} - \beta_0 = O_{\mathbb{P}}(N^{-1/2})$ (Theorem 3), the second term of (A19) is $O_{\mathbb{P}}(N^{-1/2}) \cdot o_{\mathbb{P}}(1) = o_{\mathbb{P}}(N^{-1/2})$.

Term A. Substituting $X_i - \hat{g}_{X,k,i} = e_i - \Delta g_{k,i}$ and $\hat{\mathcal{H}}^0_{k,i} = \mathcal{H}_i + \Delta_{H,k,i}$ and expanding,

$$\mathrm{A} = \underbrace{\frac{1}{N}\sum_i e_i\mathcal{H}_i}_{\mathrm{A}_1} + \underbrace{\frac{1}{N}\sum_k\sum_i e_i\Delta_{H,k,i}}_{\mathrm{A}_2} - \underbrace{\frac{1}{N}\sum_k\sum_i \Delta g_{k,i}\mathcal{H}_i}_{\mathrm{A}_3} - \underbrace{\frac{1}{N}\sum_k\sum_i \Delta g_{k,i}\Delta_{H,k,i}}_{\mathrm{A}_4}.$$

We multiply each by $\sqrt{N}$. A_1: $\sqrt{N}\,\mathrm{A}_1 = N^{-1/2}\sum_i e_i\mathcal{H}_i$, the stated leading term ($\mathbb{E}[e\mathcal{H}] = 0$ by (A7)). A_3: for $i \in \mathcal{I}_k$, $\mathbb{E}[\Delta g_{k,i}\mathcal{H}_i \mid \mathcal{D}_{-k}] = \mathbb{E}[\Delta g_k\mathcal{H}] = D_k$, and the within-fold fluctuation has conditional variance $\le n_k^{-1}\|\mathcal{H}\|^2_\infty\rho^2_{g,k}$, so $\frac{1}{n_k}\sum_{i\in\mathcal{I}_k}\Delta g_{k,i}\mathcal{H}_i = D_k + O_{\mathbb{P}}(\rho_{g,k}N^{-1/2})$; averaging, $\sqrt{N}\,\mathrm{A}_3 = \sqrt{N}\bar{D}_N + O_{\mathbb{P}}(\rho_{g,N}) = \sqrt{N}\bar{D}_N + o_{\mathbb{P}}(1)$. A_2: $\Delta_{H,k}$ is (Z,W)-measurable, so $\mathbb{E}[e_i\Delta_{H,k,i} \mid \mathcal{D}_{-k}] = \mathbb{E}[\mathbb{E}[e \mid Z,W]\Delta_{H,k}] = 0$ by (A7); the conditional variance of $N^{-1/2}\sum_{i\in\mathcal{I}_k} e_i\Delta_{H,k,i}$ is $\le \frac{1}{K}\mathbb{E}[e^2\Delta^2_{H,k} \mid \mathcal{D}_{-k}] \le \frac{1}{K}\|e\|^2_{\mathbb{P},4}\|\Delta_{H,k}\|_\infty\|\Delta_{H,k}\| = o_{\mathbb{P}}(1)$ (interpolation bound), so $\sqrt{N}\,\mathrm{A}_2 = o_{\mathbb{P}}(1)$ by conditional Chebyshev. A_4: $\mathbb{E}[\Delta g_k\Delta_{H,k} \mid \mathcal{D}_{-k}] \le \rho_{g,k}\|\Delta_{H,k}\| = O_{\mathbb{P}}\big(\rho_{g,k}(\rho_{l,k}+\rho_{q,k}) + \rho^2_{g,k} + \rho_{g,k}\rho_{m,k}\big) = o_{\mathbb{P}}(N^{-1/2})$ under Assumption 8 (the first summand is dominated by $(\rho_l+\rho_q)(\rho_m+\rho_g)$, the rest by $\frac{3}{2}(\rho_g^2+\rho_m^2)$), and the fluctuation has conditional variance $\le \frac{1}{K}\|\Delta_{H,k}\|^2_\infty\rho^2_{g,k} = o_{\mathbb{P}}(1)$, so $\sqrt{N}\,\mathrm{A}_4 = o_{\mathbb{P}}(1)$. Collecting,

$$\sqrt{N}\,\mathrm{A} = N^{-1/2}\sum_i e_i\mathcal{H}_i - \sqrt{N}\bar{D}_N + o_{\mathbb{P}}(1),$$

and combining with the $o_{\mathbb{P}}(1)$ contribution of Term B gives part (i)'s first display. For $\widehat{\Xi}_N$: it equals $\frac{1}{N}\sum_i \phi_{\mathrm{diag}}(O_i;\hat{\beta}_{\mathrm{CRC}},\hat{\eta}_k)^2$ with $\phi_{\mathrm{diag}}(O_i;\beta_0,\eta_0) = e_i\mathcal{H}_i$; by the argument of Step 6 of Theorem 2 applied to ϕ_{diag} (which Lemma A2 covers, splitting off the $\hat{\beta}_{\mathrm{CRC}}$-dependence through affinity as there), $\widehat{\Xi}_N \xrightarrow{\mathbb{P}} \mathbb{E}[(e\mathcal{H})^2] = \Xi_0$.

(ii) Under H_0, part (i) gives $\sqrt{N}\widehat{\mathcal{A}}_N = N^{-1/2}\sum_i e_i\mathcal{H}_i + o_{\mathbb{P}}(1) \xrightarrow{d} \mathcal{N}(0,\Xi_0)$ by Lindeberg–Lévy ($\mathbb{E}[e\mathcal{H}] = 0$, $\mathrm{Var}(e\mathcal{H}) = \Xi_0 \in (0,\infty)$); with $\widehat{\Xi}_N \xrightarrow{\mathbb{P}} \Xi_0 > 0$, Slutsky yields $T_N \xrightarrow{d} \mathcal{N}(0,1)$, so the size is α.

(iii) If $\sqrt{N}|\bar{D}_N| \to \infty$, the term $-\sqrt{N}\bar{D}_N$ dominates the $O_{\mathbb{P}}(1)$ term $N^{-1/2}\sum_i e_i\mathcal{H}_i$, so $|\sqrt{N}\widehat{\mathcal{A}}_N| \xrightarrow{\mathbb{P}} \infty$; as $\widehat{\Xi}_N \xrightarrow{\mathbb{P}} \Xi_0 < \infty$, $|T_N| \xrightarrow{\mathbb{P}} \infty$.

(iv) By the CRC influence expansion (Theorem 3) and Lemma 2 ($\psi_{\mathrm{CRC}}(O;\beta_0,\eta_0) = UV + e\mathcal{H}$), $\sqrt{N}(\hat{\beta}_{\mathrm{CRC}} - \beta_0) = J_0^{-1}N^{-1/2}\sum_i(U_iV_i + e_i\mathcal{H}_i) + o_{\mathbb{P}}(1)$; by the naive expansion of Theorem 2, $\sqrt{N}(\hat{\beta}_{\mathrm{naive}} - \beta_0) = J_0^{-1}[N^{-1/2}\sum_i U_iV_i + \sqrt{N}\bar{D}_N] + o_{\mathbb{P}}(1)$. Subtracting and multiplying

by J_0,

$$J_0\sqrt{N}(\hat{\beta}_{\mathrm{CRC}} - \hat{\beta}_{\mathrm{naive}}) = N^{-1/2}\sum_i e_i\mathcal{H}_i - \sqrt{N}\bar{D}_N + o_{\mathbb{P}}(1) = \sqrt{N}\widehat{\mathcal{A}}_N + o_{\mathbb{P}}(1)$$

by part (i). This is the Hausman form; $T_N = J_0\sqrt{N}(\hat{\beta}_{\mathrm{CRC}} - \hat{\beta}_{\mathrm{naive}})/\widehat{\Xi}_N^{1/2} + o_{\mathbb{P}}(1)$, the studentized contrast.

(Remark 4.) If $\mathcal{H} \equiv 0$ then $\Xi_0 = 0$ and $\bar{D}_N = 0$; part (i) gives $\sqrt{N}\widehat{\mathcal{A}}_N = o_{\mathbb{P}}(1)$ and part (iv) gives $\hat{\beta}_{\mathrm{naive}} - \hat{\beta}_{\mathrm{CRC}} = o_{\mathbb{P}}(N^{-1/2})$. □

Supplementary results for Section 5.2: The standard error channel

A first stage can center the naive point estimator at β_0 while leaving its reported standard error inconsistent. The next assumption and proposition state this distinction.

Assumption A1 (First-Stage Asymptotic Linearity along $\mathcal{H}$)**.** *There exists a bounded, (Z, W)-measurable function a with $\mathbb{E}[a^2] < \infty$ such that, for each fold k,*

$$\mathbb{E}\big[\mathcal{H}\,\Delta g_{X,k} \mid \mathcal{D}_{-k}\big] \;=\; \frac{1}{N - n_k}\sum_{j \notin \mathcal{I}_k}\big(X_j - g_X(Z_j, W_j)\big)\,a(Z_j, W_j) + \varepsilon_{k,N}, \qquad \sqrt{N}\,\max_k|\varepsilon_{k,N}| \xrightarrow{\mathbb{P}} 0. \tag{A20}$$

Proposition A1 (The Standard-Error Channel)**.** *Let Assumptions 1, 2, 7, 8, and A1 hold, write $e := X - g_X$, and suppose $\Omega_S := \mathbb{E}\big[(UV + e\,a)^2\big] > 0$. Then:*

(i) $\sqrt{N}\,\big(\hat{\beta}_{\mathrm{naive}} - \beta_0\big) \xrightarrow{d} \mathcal{N}\big(0,\; J_0^{-2}\,\Omega_S\big)$, *with* $\Omega_S = \Omega_0^{\mathrm{nv}} + 2\,\mathbb{E}[UVea] + \mathbb{E}[e^2a^2]$;

(ii) the naive plug-in variance estimator satisfies $\hat{J}_{\mathrm{nv}}^{-2}\hat{\Omega}^{\mathrm{nv}} \xrightarrow{\mathbb{P}} J_0^{-2}\,\Omega_0^{\mathrm{nv}}$, *so the true asymptotic coverage of the nominal $(1-\alpha)$ naive interval for β_0 is* $2\,\Phi\big(z_{1-\alpha/2}\sqrt{\Omega_0^{\mathrm{nv}}/\Omega_S}\big) - 1$. *If* $\mathrm{Cov}(U, X \mid Z, W) = 0$ *a.s., then* $\mathbb{E}[UVea] = 0$, *hence* $\Omega_S = \Omega_0^{\mathrm{nv}} + \mathbb{E}[e^2a^2] > \Omega_0^{\mathrm{nv}}$ *whenever* $\mathbb{E}[e^2a^2] > 0$, *and the naive interval strictly under-covers even though its centering is correct;*

(iii) if $a = \mathcal{H}$, then $\Omega_S = \Omega_0$, the CRC-robust variance of Theorem 3, and $\hat{\beta}_{\mathrm{naive}} - \hat{\beta}_{\mathrm{CRC}} = o_{\mathbb{P}}(N^{-1/2})$, *so the first stage orthogonalizes implicitly, yet the naive sandwich still reports* $J_0^{-2}\Omega_0^{\mathrm{nv}}$.

Parts (ii)–(iii) concern inference on the fixed target β_0 and therefore do not contradict Theorem 2(iv), which concerns inference around the moving target β_N^*. The naive sandwich consistently estimates the variance around β_N^* but omits the first-stage term needed for inference on β_0. When $\mathrm{Cov}(U, X \mid Z, W) = 0$, this omission produces strict undercoverage whenever $\mathbb{E}[e^2a^2] > 0$. Proposition A2 gives a concrete first stage satisfying Assumption A1: projection estimators can supply the point-estimate correction implicitly (Newey, 1994), while the naive sandwich remains inconsistent for fixed-target inference unless its probability limit equals the true variance. Outside that projection structure, even correct centering is not guaranteed.

Proof of Proposition A1. (i) With equal folds $n_k = N/K$, Assumption A1 gives $\frac{n_k}{N}\mathbb{E}[\mathcal{H}\Delta g_k \mid \mathcal{D}_{-k}] = \frac{1}{N(K-1)}\sum_{j\notin\mathcal{I}_k} e_j a_j + \frac{1}{K}\varepsilon_{k,N}$, because $\frac{n_k}{N}\cdot\frac{1}{N-n_k} = \frac{1}{N(K-1)}$. Summing over k and using that each index j is omitted from exactly its own fold, hence appears in $\sum_{j\notin\mathcal{I}_k}$ for $K-1$ of the K folds, $\sum_k\sum_{j\notin\mathcal{I}_k} e_j a_j = (K-1)\sum_{j=1}^N e_j a_j$, so

$$\bar{D}_N = \sum_k \frac{n_k}{N}\mathbb{E}[\mathcal{H}\Delta g_k \mid \mathcal{D}_{-k}] = \frac{1}{N}\sum_{j=1}^N e_j a_j + \frac{1}{K}\sum_k \varepsilon_{k,N},$$

while

$$\sqrt{N}\left|\tfrac{1}{K}\sum_k \varepsilon_{k,N}\right| \le \sqrt{N}\max_k|\varepsilon_{k,N}| = o_{\mathbb{P}}(1),$$

whence $\sqrt{N}\bar{D}_N = N^{-1/2}\sum_j e_j a_j + o_{\mathbb{P}}(1)$. Substituting into the naive expansion of Theorem 2,

$$\sqrt{N}(\hat{\beta}_{\text{naive}} - \beta_0) = J_0^{-1}\left[N^{-1/2}\sum_i U_i V_i + \sqrt{N}\bar{D}_N\right] + o_{\mathbb{P}}(1) = J_0^{-1}N^{-1/2}\sum_i (U_iV_i + e_i a_i) + o_{\mathbb{P}}(1).$$

The summands $U_iV_i + e_ia_i$ are i.i.d. with mean $\mathbb{E}[UV] + \mathbb{E}[ea] = 0$ (the second by (A7), a being (Z,W)-measurable) and variance $\Omega_S = \mathbb{E}[(UV+ea)^2] = \Omega_0^{\text{nv}} + 2\mathbb{E}[UVea] + \mathbb{E}[e^2a^2]$, so Lindeberg–Lévy and Slutsky give $\sqrt{N}(\hat{\beta}_{\text{naive}} - \beta_0) \xrightarrow{d} \mathcal{N}(0, J_0^{-2}\Omega_S)$.

(ii) Step 6 of the proof of Theorem 2 shows $\hat{J}_{\text{nv}}^{-2}\hat{\Omega}^{\text{nv}} \xrightarrow{\mathbb{P}} J_0^{-2}\Omega_0^{\text{nv}}$ using only Assumptions 7–8 (it does not use Assumption A1), so the reported variance is $J_0^{-2}\Omega_0^{\text{nv}}$ while the true one is $J_0^{-2}\Omega_S$. The nominal interval has half-width $z_{1-\alpha/2}\sqrt{J_0^{-2}\Omega_0^{\text{nv}}/N}(1+o_{\mathbb{P}}(1))$ and $\hat{\beta}_{\text{naive}} - \beta_0 = \sqrt{J_0^{-2}\Omega_S/N}\,Z(1+o_{\mathbb{P}}(1))$ with $Z \xrightarrow{d} \mathcal{N}(0,1)$; hence coverage $\to \mathbb{P}(|Z| \le z_{1-\alpha/2}\sqrt{\Omega_0^{\text{nv}}/\Omega_S}) = 2\Phi(z_{1-\alpha/2}\sqrt{\Omega_0^{\text{nv}}/\Omega_S}) - 1$. For the sign, compute $\mathbb{E}[UVea] = \mathbb{E}[Va\,\mathbb{E}[Ue \mid Z,W]]$ (as V, a are (Z,W)-measurable), and $\mathbb{E}[Ue \mid Z,W] = \mathbb{E}[UX \mid Z,W] - g_X\mathbb{E}[U \mid Z,W] = \mathrm{Cov}(U, X \mid Z,W) + \mathcal{H}g_X - g_X\mathcal{H} = \mathrm{Cov}(U, X \mid Z, W)$ by (A6). Thus $\mathbb{E}[UVea] = \mathbb{E}[Va\ \mathrm{Cov}(U, X \mid Z, W)]$,

which vanishes when $\mathrm{Cov}(U, X \mid Z, W) = 0$ a.s.; then $\Omega_S = \Omega_0^{\mathrm{nv}} + \mathbb{E}[e^2 a^2] > \Omega_0^{\mathrm{nv}}$ whenever $\mathbb{E}[e^2 a^2] > 0$, and $2\Phi(z_{1-\alpha/2}\sqrt{\Omega_0^{\mathrm{nv}}/\Omega_S}) - 1 < 1 - \alpha$.

(iii) If $a = \mathcal{H}$, then $\Omega_S = \mathbb{E}[(UV + e\mathcal{H})^2] = \mathbb{E}[\psi_{\mathrm{CRC}}(O; \beta_0, \eta_0)^2] = \Omega_0$ by Lemma 2. The influence function of $\hat{\beta}_{\mathrm{naive}}$ in (i) is then $J_0^{-1}(UV + e\mathcal{H})$, identical to that of $\hat{\beta}_{\mathrm{CRC}}$ (Theorem 3 with Lemma 2), so $\hat{\beta}_{\mathrm{naive}} - \hat{\beta}_{\mathrm{CRC}} = o_{\mathbb{P}}(N^{-1/2})$. By part (ii), however, the naive sandwich converges to $J_0^{-2}\Omega_0^{\mathrm{nv}}$, which need not equal $J_0^{-2}\Omega_0$. Lemma 2 gives the exact condition:

$$\Omega_0 - \Omega_0^{\mathrm{nv}} = 2\mathbb{E}[UVe\mathcal{H}] + \mathbb{E}[e^2\mathcal{H}^2].$$

Thus the naive sandwich is inconsistent precisely when the right-hand side is nonzero. If $\mathrm{Cov}(U, X \mid Z, W) = 0$, the cross term vanishes, and inconsistency follows whenever $\mathbb{E}[e^2\mathcal{H}^2] > 0$. □

Proposition A2 (Series Least Squares Verifies Assumption A1)**.** *Let $p(Z, W) \in \mathbb{R}^{d_p}$ be a fixed, bounded dictionary with $Q := \mathbb{E}[p\,p']$ nonsingular; suppose $g_X = p'\theta_g$ and $\mathcal{H} = p'\theta_{\mathcal{H}}$ for some $\theta_g, \theta_{\mathcal{H}} \in \mathbb{R}^{d_p}$; and let $\hat{g}_{X,k} = p'\hat{\theta}_k$, where $\hat{\theta}_k$ is the ordinary least squares coefficient of X on p computed on $\mathcal{D}_{-k}$. Then Assumption A1 holds with $a = \mathcal{H}$, and $\rho_{g,k} = O_{\mathbb{P}}(N^{-1/2})$.*

Proof of Proposition A2. Let $\hat{Q}_k := \frac{1}{N-n_k}\sum_{j \notin \mathcal{I}_k} p_j p_j'$ and note the fold-k OLS coefficient is $\hat{\theta}_k = \hat{Q}_k^{-1}\frac{1}{N-n_k}\sum_{j \notin \mathcal{I}_k} p_j X_j$. Since $X = p'\theta_g + e$ with $\mathbb{E}[pe] = \mathbb{E}[p\,\mathbb{E}[e \mid Z, W]] = 0$ by (A7),

$$\hat{\theta}_k - \theta_g = \hat{Q}_k^{-1}\frac{1}{N - n_k}\sum_{j \notin \mathcal{I}_k} p_j e_j.$$

As $\Delta g_k = p'(\hat{\theta}_k - \theta_g)$ and $\mathcal{H} = p'\theta_{\mathcal{H}}$, and $\hat{\theta}_k$ is $\mathcal{D}_{-k}$-measurable,

$$\mathbb{E}[\mathcal{H}\Delta g_k \mid \mathcal{D}_{-k}] = \theta_{\mathcal{H}}'\,\mathbb{E}[pp']\,(\hat{\theta}_k - \theta_g) = \theta_{\mathcal{H}}' Q \hat{Q}_k^{-1}\frac{1}{N - n_k}\sum_{j \notin \mathcal{I}_k} p_j e_j.$$

Split $\theta_{\mathcal{H}}' Q\hat{Q}_k^{-1} = \theta_{\mathcal{H}}' + \theta_{\mathcal{H}}'(Q\hat{Q}_k^{-1} - I)$. The first piece gives $\theta_{\mathcal{H}}'\frac{1}{N-n_k}\sum_{j \notin \mathcal{I}_k} p_j e_j = \frac{1}{N-n_k}\sum_{j \notin \mathcal{I}_k}(p_j'\theta_{\mathcal{H}})e_j = \frac{1}{N-n_k}\sum_{j \notin \mathcal{I}_k}(X_j - g_{X,j})\mathcal{H}_j$, which is the leading term of Assumption A1 with $a = \mathcal{H}$. The second piece is $\varepsilon_{k,N} = \theta_{\mathcal{H}}'(Q\hat{Q}_k^{-1} - I)\frac{1}{N-n_k}\sum_{j \notin \mathcal{I}_k} p_j e_j$; using $Q\hat{Q}_k^{-1} - I = -Q\hat{Q}_k^{-1}(\hat{Q}_k - Q)Q^{-1}$ with $\hat{Q}_k - Q = O_{\mathbb{P}}(N^{-1/2})$ (bounded dictionary) and $\frac{1}{N-n_k}\sum_{j \notin \mathcal{I}_k} p_j e_j = O_{\mathbb{P}}(N^{-1/2})$ (mean zero, finite variance since p bounded and $e \in L_2$), we get $\varepsilon_{k,N} = O_{\mathbb{P}}(N^{-1})$, so $\sqrt{N}\max_k |\varepsilon_{k,N}| = o_{\mathbb{P}}(1)$. Thus Assumption A1 holds with $a = \mathcal{H}$. Finally $\rho_{g,k}^2 = \|p'(\hat{\theta}_k - \theta_g)\|^2 = (\hat{\theta}_k - \theta_g)'Q(\hat{\theta}_k - \theta_g) \le \lambda_{\max}(Q)\|\hat{\theta}_k - \theta_g\|^2 = O_{\mathbb{P}}(N^{-1})$, i.e. $\rho_{g,k} = O_{\mathbb{P}}(N^{-1/2})$. □

Proof for Section 5.2: Identification-Robust Sets

Proof of Proposition 6. Decompose $\sqrt{N}\bar{\psi}(\beta_0) = N^{-1/2}\sum_i \psi_{\mathrm{CRC}}(O_i;\beta_0,\eta_0) + N^{-1/2}\sum_k \sum_{i\in\mathcal{I}_k} E_{ik}$ with $E_{ik} := \psi_{\mathrm{CRC}}(O_i;\beta_0,\hat{\eta}_k) - \psi_{\mathrm{CRC}}(O_i;\beta_0,\eta_0)$. By Lemma A4 (conditional form) and $\mathbb{E}[\psi_{\mathrm{CRC}}(O;\beta_0,\eta_0)] = 0$,

$$\mathbb{E}[E_{ik} \mid \mathcal{D}_{-k}] = R_{Y,k} - \beta_0 R_{X,k}, \qquad |R_{Y,k} - \beta_0 R_{X,k}| \le \rho_{l,k}\rho_{m,k} + \rho_{g,k}\rho_{q,k} + |\beta_0|(\rho_{m,k}^2 + \rho_{g,k}^2),$$

which is $o_{\mathbb{P}}(N^{-1/2})$ under Assumption 8; hence the fold-mean contribution $N^{-1/2}\sum_k n_k(R_{Y,k} - \beta_0 R_{X,k}) = \sqrt{N}\sum_k \frac{n_k}{N}(R_{Y,k} - \beta_0 R_{X,k}) = o_{\mathbb{P}}(1)$. The centered part $N^{-1/2}\sum_{i\in\mathcal{I}_k}[E_{ik} - \mathbb{E}(E_{ik} \mid \mathcal{D}_{-k})]$ has conditional variance $\le \frac{1}{K}\mathbb{E}[E_{ik}^2 \mid \mathcal{D}_{-k}] = \frac{1}{K}\|\psi_{\mathrm{CRC}}(\cdot;\beta_0,\hat{\eta}_k) - \psi_{\mathrm{CRC}}(\cdot;\beta_0,\eta_0)\|^2 \le \frac{L^2}{K}(\sum_j \rho_{j,k}^{1/2})^2 = o_{\mathbb{P}}(1)$ by Lemma A2 (with $\beta' = \beta = \beta_0$), so it is $o_{\mathbb{P}}(1)$ by conditional Chebyshev. Therefore $\sqrt{N}\bar{\psi}(\beta_0) = N^{-1/2}\sum_i \psi_{\mathrm{CRC}}(O_i;\beta_0,\eta_0) + o_{\mathbb{P}}(1) \xrightarrow{d} \mathcal{N}(0,\Omega_0)$ by Lindeberg–Lévy. For the denominator, the same Step-6 argument (now with β_0 fixed, so only the nuisance-continuity part is needed) gives $\hat{\sigma}^2(\beta_0) \xrightarrow{\mathbb{P}} \mathbb{E}[\psi_{\mathrm{CRC}}(O;\beta_0,\eta_0)^2] = \Omega_0 > 0$. Slutsky yields $\mathrm{AR}_N(\beta_0) \xrightarrow{d} \mathcal{N}(0,1)$, so $\mathbb{P}(\beta_0 \in \mathcal{C}_{1-\alpha}) = \mathbb{P}(|\mathrm{AR}_N(\beta_0)| \le z_{1-\alpha/2}) \to 1-\alpha$. No step divides by J_0 or uses a lower bound on it.

(Remark 5.) With $\bar{\psi}(\beta) = S_Y - \beta S_X$ and $\hat{\sigma}^2(\beta) = M_{YY} - 2\beta M_{XY} + \beta^2 M_{XX}$, the defining inequality $N\bar{\psi}(\beta)^2 \le z^2\hat{\sigma}^2(\beta)$ becomes

$$A\beta^2 - 2B\beta + C \le 0.$$

Let $D := B^2 - AC$. If $A > 0$, the set is empty when $D < 0$ and is the closed interval with endpoints $(B \pm \sqrt{D})/A$ when $D \ge 0$. If $A < 0$, it is the whole line when $D \le 0$ and the complement of the open interval between the two roots when $D > 0$. If $A = 0$ and $B \ne 0$, it is a half-line with boundary $C/(2B)$; if $A = B = 0$, it is the whole line when $C \le 0$ and empty when $C > 0$. These are the Fieller–Anderson–Rubin cases. □