Triple-Robust Instrumental-Variable Estimation with Self-Supervised Representations in High-Dimensional Settings

Tamer Çetin TAMERCETIN@BERKELEY.EDU
Department of Economics and
Department of Agricultural and Resource Economics
University of California, Berkeley
Berkeley, CA 94720, USA

Abstract

I propose TRIV–Rep, a union-of-models estimator for a linear causal effect with endogenous treatment, weak or noisy instruments, high-dimensional proxies, and scarce outcome labels. The method builds a self-supervised representation of raw covariates and an orthogonal tri-score that remains valid if *any one* of three routes holds: (I1) a valid (possibly weak) instrument conditional on the learned representation; (I2) a proxy-only bridge that requires no completeness; or (I3) a correctly specified treatment–residual model given the representation. I show local minimaxity among orthogonal scores and derive $\sqrt{n}$ inference with cross-fitting under $o(n^{-1/4})$ rates for nuisance learners met by contrastive encoders and flexible regressions. The tri-score attains efficiency in each constituent submodel and hence on their union. Simulations and a Yelp application illustrate robustness to weak instruments and sparse labels, using the same estimating equation across all three routes.

JEL codes: C14, C21, C26, C38
Keywords: instrumental variables; self-supervised learning; representation learning; weak instruments; proxy variables; high-dimensional data; causal inference

1 Introduction

The biggest promise—but also the main statistical headache—of modern digital platforms is that they generate massive but unbalanced data. While billions of user interactions—such as clicks, searches, or impressions—are logged each day, only a small subset of these actions is associated with observed outcomes of interest like purchases or churn. This creates a difficult estimation environment marked by three distinct statistical frictions. First, the randomization cues introduced by engineers into production systems tend to be weak instruments that affect the treatment only through noisy or indirect channels (Stock, Wright, and Yogo, 2002). Second, many confounders, such as user intent or product appeal, are not directly observed but may leave proxy traces in high-dimensional raw data (Newey and Powell, 2003; Deaner, 2022, 2023). Third, the label density is often extremely low: only a small fraction of logged user events have observed outcomes, which severely limits the use of conventional supervised estimators (Meunier et al., 2025).

This paper introduces TRIV–Rep, a triple-robust[1] estimator that combines self-supervised representation learning with orthogonal moment construction. The key innovation is a

[1] In the paper, triple-robustness is used in a union-of-models sense: consistency holds if *any one* of identification routes is satisfied. This differs from the "any two of three models" notion common in missing-data settings.

tri-score moment function $\psi_i(\tau,\theta) = v_i\left[Z_i - \pi_\gamma\big(R_\eta(X_i)\big)\right]\Big(\left[Y_i - \mu_Y\big(R_\eta(X_i)\big)\right] - \tau\left[D_i - \mu_D\big(R_\eta(X_i)\big)\right]\Big)$, where $v_i = 1$ when every outcome is observed, and $v_i = S_i/\rho_n$ (an inverse-probability weight) when labels are missing. The score is orthogonal to three nuisance blocks: (i) an SSL encoder $R_\eta(X)$ trained on the unlabeled bulk of events via contrastive InfoNCE (van den Oord, Li, and Vinyals, 2018), (ii) an instrument regression $\pi_\gamma(R) := \mathbb{E}[Z \mid R]$, and (iii) outcome and treatment regressions $\mu_Y(R) := \mathbb{E}[Y \mid R]$ and $\mu_D(R) := \mathbb{E}[D \mid R]$. The tri-score secures identification if *any one* of three routes holds: (I1) Z is a valid (possibly weak) instrument given R; (I2) R is a proxy that is L_2-complete for the latent confounder and satisfies cue-sufficiency, so $(Z - \pi(R))$ is orthogonal to the structural residual; or (I3) the treatment–residual model is correctly specified so that $\mathbb{E}[Y - \tau_0 D \mid R] = 0$. Route (I3) *implies* (I1), while (I2) is complementary; I analyze the *union* of (I1)–(I3) and refer to the estimator as *triple-robust.* In this union model the tri-score is locally minimax among first-order orthogonal moments, and its influence function coincides with the efficient influence in each submodel, yielding semiparametric efficiency on the union. The tri-score's influence function equals the efficient influence function in each submodel $\mathcal{M}_j$; by the union-efficiency theorem, it is therefore semiparametrically efficient on $\mathcal{M}_\cup$. Note that standard two-block DML–IV procedures can fail when the learned representation is *insufficient to restore conditional instrument validity.* In contrast, the tri-score in this paper uses the same estimating equation but is valid under any one of three routes to identification (I1)–(I3), strictly enlarging the set of conditions under which the estimator is identified. A second innovation is to formalize a proxy-only route that requires no completeness. Under a latent U with cue–sufficiency ($Z \perp\!\!\!\perp U \mid R$) and outcome isolation ($\varepsilon \perp\!\!\!\perp (Z,R) \mid U$), there exists a bridge $b(R)$ with $\mathbb{E}[\varepsilon \mid Z,R] = b(R)$. For any $h(Z,R)$ satisfying $\mathbb{E}[h \mid R] = 0$, the unconditional proxy moments $\mathbb{E}[h(Z,R)\{Y - \tau D\}] = 0$ identify τ, and the classical exogeneity restriction $\mathbb{E}[\varepsilon \mid Z,R] = 0$ is not needed. I derive the efficient instrument within the proxy class $\mathcal{H} = \{h : \mathbb{E}[h \mid R] = 0\}$ and a cross-fitted estimator that enforces $\mathbb{E}[\hat{h} \mid R] \approx 0$ in sample.

The method connects to several strands in the literature. First, I build on work in semiparametric IV estimation with machine learning (Belloni, Chernozhukov, and Hansen, 2014; Hartford et al., 2017; Chernozhukov et al., 2018; Farrell, Liang, and Misra, 2021), where orthogonal scores enable valid inference in high-dimensional settings. These approaches assume a valid, sufficiently strong instrument and require accurate estimation of both the first-stage and outcome regressions; they can break down when instruments are weak or when unobserved confounding remains after the chosen controls. The present construction generalizes this setup by introducing a third block—an embedding of the covariates—trained via a contrastive self-supervised objective, allowing the method to leverage unlabeled data. Thus, the proposed estimator extends the two-block orthogonal moments of Chernozhukov et al. (2018) for IV by adding a representation block $R_\eta(X_i)$ learned from unlabeled X, enabling identification when instruments are latent/weak.

Second, the identification argument extends the proxy–control framework via the L_2-completeness condition of Newey and Powell (2003) and its high-dimensional refinements by Deaner (2022, 2023), which spell out when rich proxy variables can neutralize latent confounding. The approach sidesteps the need to explicitly model the confounder by operating on a low-dimensional representation $R_i = R_\eta(X_i)$, learned from data. Related work using variational autoencoders (Cheng et al., 2023; Wu and Fukumizu, 2021; Hartford et al., 2017) also seeks to learn representations for causal estimation, though those approaches typically rely on a fully specified generative decoder and richer supervision (e.g., paired instruments).

Third, the encoder is learned with a *contrastive* InfoNCE/Spectral-Contrastive (SCL) loss. Recent theory shows that such contrastive embeddings can be statistically consistent and, under separability or spectral conditions, rate-optimal (Arora et al., 2019; HaoChen et al., 2021). Leveraging existing *PAC–Bayes* generalization bounds for spectrally normalized networks (Neyshabur et al., 2018)—and, more directly, the contrastive-specific analysis of HaoChen et al. (2021)—I show that the encoder R_η achieves the $o(n^{-1/4})$ L_2 rate required for $\sqrt{n}$-valid inference whenever a sufficiently large pool of unlabeled data is available. An alternative path uses margin-based Rademacher-complexity bounds for spectrally normalized ReLU networks (Bartlett, Foster, and Telgarsky, 2017).[2] Finally, the logic of using multiple identification paths draws inspiration from the literature on multiply robust and triple-robust estimation in causal inference and missing data (Robins and Rotnitzky, 1995; Tchetgen Tchetgen, Robins, and Rotnitzky, 2010; Okui, Small, Tan, and Robins, 2012), though those works generally assume a fully observed treatment and focus on missing outcomes. In contrast, the approach in this paper includes both endogeneity and high-dimensional proxies, as well as label scarcity, and thus requires a different moment structure.

In sum, the proposed estimator departs from existing work in four respects. (i) It augments the two–block orthogonal scores of double machine–learning IV with a third, self–supervised representation block, thereby preserving identification when either the instrument is weak or the proxy is incomplete. (ii) In contrast to proxy–control methods that assume a known complete basis, the representation is *learned* from unlabeled covariates, so that identification leverages the sheer scale of modern logs rather than hand-crafted features. (iii) Where recent VAE-based IV approaches rely on dense labels and strong instruments, I allow the labeled pool to be vanishingly small and prove that the encoder still attains the $o(n^{-1/4})$ rate required for $\sqrt{n}$-valid inference. (iv) Finally, while multiply robust estimators in the missing-data literature handle outcome non-response, they assume an exogenous treatment; the tri–score controls simultaneously for endogeneity, proxy noise, and label sparsity. Taken together, these elements yield what I call a *triple-robust* estimator—identification is secured provided at least one of three high-level conditions holds—and, to my knowledge, no existing method offers this combination of self-supervised learning, weak-IV tolerance, and finite-sample orthogonality. Theoretical results, evidence from simulations and real-world data application suggest that this design yields improved performance over standard two-block scores in finite samples, particularly in environments with weak instruments and limited labels. The estimator may thus offer a useful tool for causal estimation in high-dimensional digital settings where unobserved confounding and limited supervision are common.

Two variants of the proxy route appear in the paper. The *operational* one for estimation and empirical work is I2 (bridge-only): cue–sufficiency $Z \perp\!\!\!\perp U \mid R$ and outcome isolation $\varepsilon \perp\!\!\!\perp (Z, R) \mid U$ imply a bridge $b(R) = \mathbb{E}[\mathbb{E}(\varepsilon \mid U) \mid R]$ and the unconditional moments $\mathbb{E}[h(Z, R)\{Y - \tau D\}] = 0$ for all instruments h with $\mathbb{E}[h \mid R] = 0$. This form does *not* require completeness. Separately, I discuss the stronger *proxy-complete* case—useful for mapping to

[2] HaoChen et al. (2021) give contrastive-specific PAC–Bayes bounds; Neyshabur et al. (2018) provide generic spectral-norm PAC–Bayes bounds for deep networks.

triple-proxy identification—where R is L_2-complete for U. Completeness is *not* used by the estimator and can be deferred to an appendix cross-walk.

Section 2 introduces the structural framework, the three identification routes (I1)–(I3), and the treatment of label sparsity. Section 3 presents the tri–score, establishes Neyman orthogonality, and states the main large-sample results (consistency, $\sqrt{n}$–normality) together with the minimax and efficiency theorems. Section 4 states sufficient rate conditions for the nuisance learners, and Section 5 gives the central limit theorem and sandwich inference. Section 6 reports Monte Carlo evidence for the three identification routes. Section 7 contains the Yelp application and Section 8 concludes.

2 Model

2.1 Structural Framework and Core Moment

Consider a set of observations $\{(Y_i, D_i, X_i)\}_{i=1}^n$, where $Y_i \in \mathbb{R}$ is a scalar outcome, $D_i \in \mathbb{R}$ is a potentially endogenous treatment, and $X_i \in \mathbb{R}^{d_x}$ is a high-dimensional vector of raw covariates (e.g., text tokens, device IDs). The central idea is to leverage the information within the high-dimensional X_i by partitioning it into two components: Randomization cue $Z_i := \zeta(X_i) \in \mathbb{R}^{d_z}$, where $\zeta : \mathbb{R}^{d_x} \to \mathbb{R}^{d_z}$ is a deterministic extractor; Z_i serves as an IV (possibly externally logged). Proxy representation $R_i := R_{\eta_0}(X_i) \in \mathbb{R}^{d_r}$, a self-supervised embedding intended to control for latent confounding.

The data are generated by a linear structural model:

$$Y_i = \tau_0 \, D_i + \varepsilon_i, \tag{1}$$

$$D_i = g_0(X_i) + U_i, \tag{2}$$

where τ_0 is the causal parameter of interest. The function $g_0(\cdot)$ is an unrestricted nuisance function, and U_i is a scalar unobserved confounder that may be correlated with both the treatment assignment and the outcome, i.e., $\mathrm{Cov}(U_i, \varepsilon_i) \neq 0$. This potential correlation renders D_i endogenous. The structural error ε_i is mean-independent of the full covariate vector, $\mathbb{E}[\varepsilon_i \mid X_i] = 0$, but not necessarily of the confounder U_i. The endogeneity arises because, in general, $\mathbb{E}[U_i \mid X_i] \neq 0$. The goal is to identify τ_0 using a moment condition built from the observed data (Y_i, D_i, Z_i, R_i). The proposed estimator is built around the following "tri-score" moment function:

$$\psi(\tau, \theta) \; = \; v\,[Z - \pi_\gamma(R)]\,\Big([Y - \mu_Y(R)] - \tau\,[D - \mu_D(R)] \Big), v \equiv 1. \tag{3}$$

Here, $\theta = (\eta, \gamma, \varphi)$ collects the nuisance functions[3]: the encoder $R_\eta(X)$ producing the representation R; the instrument regression $\pi_\gamma(R) \approx \mathbb{E}[Z \mid R]$; and the outcome and treatment regressions $\mu_Y(R) \approx \mathbb{E}[Y \mid R]$ and $\mu_D(R) \approx \mathbb{E}[D \mid R]$. The residualized score above is algebraically equivalent to $v[Z - \pi(R)]\{Y - \tau D - m_\varphi(R)\}$ with $m_\varphi(R) \; = \; \mu_Y(R) - \tau\,\mu_D(R)$. In particular: $\mathbb{E}\big[(Z - \pi(R))\{(Y - \mu_Y(R)) - \tau(D - \mu_D(R))\}\big] \; = \; \mathbb{E}[(Z - \pi(R))(Y - \tau D - m_\varphi(R))]$. I implement the residualized form so that the nuisance blocks do not depend on τ. The estimator for τ_0 is found by solving the sample analogue of the population moment equation $\Psi(\tau, \theta_0) = \mathbb{E}[\psi(\tau, \theta_0)] = 0$, where θ_0 denotes the true nuisance parameters.

[3] The scalar weight v is fixed at 1 under full observation; in Section 2.4, it is replaced by an inverse-probability weight when labels are missing.

2.2 Identification

Identification of τ_0 relies on the moment equation $\mathbb{E}\big[\psi(\tau_0, \theta_0)\big] = 0$. A simple expansion of the population moment $\Psi(\tau) := \mathbb{E}[\psi(\tau, \theta_0)]$ reveals its structure:

$$\Psi(\tau) = \mathbb{E}\big[(Z - \pi_0(R))\,\varepsilon\big] - (\tau - \tau_0)\,C, \qquad C := \mathbb{E}\big[(Z - \pi_0(R))\,\{D - \mu_{D,0}(R)\}\big]. \tag{4}$$

Define the (scalar) Jacobian in τ by $S := \partial_\tau \Psi(\tau_0, \theta_0)$. Since $\mathbb{E}[(Z - \pi_0(R))\,\mu_{D,0}(R)] = 0$, I have $S = -\,\mathbb{E}\big[(Z - \pi_0(R))\,D\big] = -\,C$. Use $S = -C$ throughout to denote the Jacobian in τ. Setting $\Psi(\tau) = 0$ uniquely identifies $\tau = \tau_0$ provided (i) *Orthogonality*, $\mathbb{E}[(Z - \pi_0(R))\,\varepsilon] = 0$, and (ii) *Relevance*, $C \neq 0$ (Assumption 2.1).

Assumption 2.1 (Weak relevance) *There exists $\kappa > 0$ such that* $\mathrm{Var}\big(\mathbb{E}[D \mid Z, R]\big) \geq \kappa$ *and* $\operatorname{ess\,inf}_r\ \mathrm{Var}\big(\mathbb{E}[D \mid Z, R = r]\big) \geq \kappa$.

Assumption 2.2 (Cue–sufficiency and proxy completeness) *There exists a latent U such that: (i) $Z \perp U \mid R$ (given R, the cue Z adds no information about U); (ii) $\varepsilon := Y - \tau_0 D \perp (Z, R) \mid U$; (iii) R is L_2–complete for U: if $\mathbb{E}[h(U) \mid R] = 0$, then $h(U) = 0$ almost surely.*

Remark 1 (Comparison to triple-proxy identification) *The triple-proxy literature (e.g. Deaner (2023)) typically assumes $Z \perp R \mid U$ together with a third proxy and completeness. In contrast, my proxy route imposes cue-sufficiency $Z \perp U \mid R$ and L_2–completeness of R, which together form a self-contained sufficient condition for the orthogonality $\mathbb{E}[(Z - \pi_0(R))\,\varepsilon] = 0$ without invoking the full triple-proxy setup.*

Proposition 2 (Triple-proxy cross-walk) *Suppose $Z \perp R \mid U$, $\varepsilon \perp (Z, R) \mid U$, and R is L_2–complete for U (conditions as in Deaner, 2023). Then U is identified up to a monotone transform, but this does not imply $\mathbb{E}[(Z - \mathbb{E}[Z \mid R])\,\varepsilon] = 0$. If, in addition, cue–sufficiency $Z \perp U \mid R$ holds, the tri-score orthogonality follows.*

Remark 3 (Cue–sufficiency is stronger than additive–noise proxies) *Additive–noise proxy models deliver "forward" conditional independences, $Z \perp R \mid U$ and $\varepsilon \perp (Z, R) \mid U$, but they* do not *deliver the "reverse" sufficiency $Z \perp U \mid R$. Intuitively, a noisy R rarely makes Z uninformative about U; a second noisy readout of U (namely Z) still carries residual information about U even after conditioning on R.*

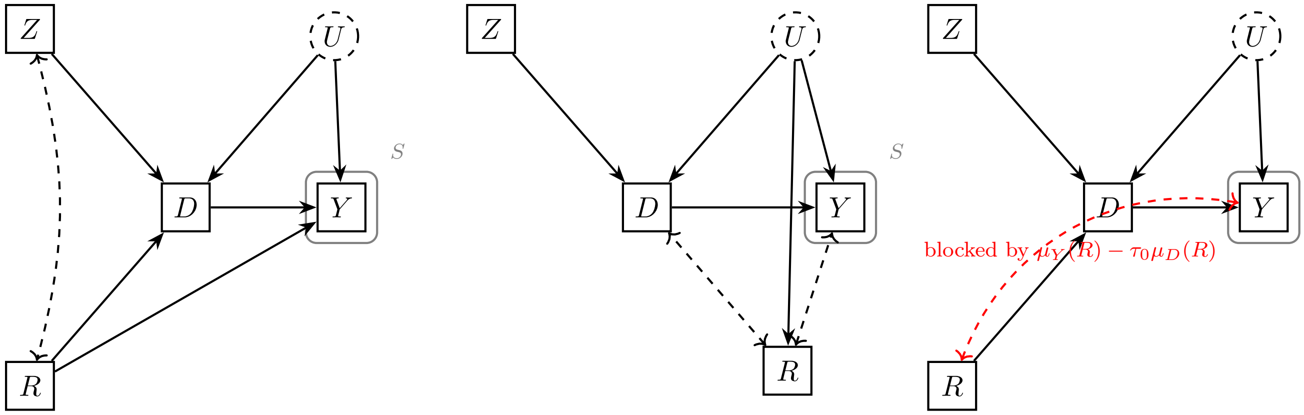
Let $U \sim \mathcal{N}(0, \sigma_U^2)$, $Z = U + \xi_1$, $R = U + \xi_2$ with ξ_1, ξ_2 independent of each other and of U, and mean zero. Then $Z \perp R \mid U$ holds by construction, but $\mathrm{Cov}(Z, U \mid R) = \mathrm{Var}(U \mid R) = \sigma_U^2 \cdot \frac{\sigma_2^2}{\sigma_U^2 + \sigma_2^2} > 0$ *whenever* $\sigma_2^2 > 0$, *so $Z \not\perp U \mid R$ unless R is noise-free. Thus, the (I2) condition $Z \perp U \mid R$ is an additional restriction that must be assumed or justified (e.g., R is a learned representation sufficient for Z).*

Assume the structural form $Z = h_1(U) + \xi_1$, with $\xi_1 \perp U$; $R = h_2(U) + \xi_2$, with $\xi_2 \perp (U, \xi_1)$; and $\varepsilon = h_3(U) + \xi_3$, with $\xi_3 \perp (U, \xi_1, \xi_2)$, where each ξ_j is mean-zero, full-support, independent and identically distributed (i.i.d.) noise. Then (ii) holds by construction; (iii) (completeness of R for functions of U) follows from standard additive-noise completeness results in NPIV (e.g. Newey and Powell (2003); Ai and Chen (2003)) provided the noise ξ_2

is independent of U and has full support; by contrast, cue–sufficiency (i) $Z \perp U \mid R$ does *not* follow from this setup and must be imposed separately if one wants the orthogonality $\mathbb{E}[(Z - \pi_0(R))\,\varepsilon] = 0$ used by the tri-score[4]. In particular, the additive-noise proxy model alone yields $Z \perp R \mid U$ but typically fails $Z \perp U \mid R$ unless R is noise-free; hence cue–sufficiency is a substantive extra restriction needed for the I2 orthogonality exploited here:

Proxy	Definition	Role w.r.t. U
W_1	Z	Noisy stimulus correlated with U but engineering-set
W_2	$R = R_{\eta_0}(X)$	Data–driven proxy for U (§4)
W_3	$\varepsilon = Y - \tau_0 D$	Outcome residual capturing U only

Under the conditional–independence structure $Z \perp R \mid U$ and $\varepsilon \perp (Z, R) \mid U$, together with completeness of R (Assumption 2.2, matching the completeness requirement on the second proxy in Deaner (2023)) and the usual injectivity/variation condition on the third proxy, the triple-proxy result of Deaner (2023) identifies U up to a one-to-one reparameterization. By itself, this does *not* imply the moment orthogonality $\mathbb{E}[(Z - \pi_0(R))\,\varepsilon] = 0$. Orthogonality holds only when cue–sufficiency $Z \perp U \mid R$ (Assumption 2.2(i)) is also imposed, in which case $\mathbb{E}[(Z - \pi_0(R))\,\varepsilon] = \mathbb{E}\big[(\mathbb{E}[Z \mid U, R] - \mathbb{E}[Z \mid R])\,\mathbb{E}[\varepsilon \mid U]\big] = 0$. Hence, the tri-score moment identifies τ_0 even when the classical IV orthogonality $\mathbb{E}[\varepsilon \mid Z, R] = 0$ fails (e.g., due to latent confounding). Conversely, when Z is a valid instrument conditional on R, route (I1) alone suffices and the triple-proxy conditions are unnecessary. The causal structures underlying these routes are illustrated in Figure 1.



(a) I1 (b) I2 (c) I3

Figure 1: Causal DAGs for the three identification routes. Dashed bidirected edges denote associations via the latent U. In panel (b) (I2), the key restrictions are *cue–sufficiency* $Z \perp U \mid R$ and *outcome isolation* $\varepsilon \perp (Z, R) \mid U$; these imply $\mathbb{E}[\varepsilon \mid Z, R] = b(R)$ and hence $\mathbb{E}[h(Z, R)\{Y - \tau D\}] = 0$ for any h with $\mathbb{E}[h \mid R] = 0$ (no completeness required). The gray "S" box indicates outcome sampling; it is inactive here and becomes active only in the MNAR figure in the appendix.

[4] Note that completeness in (iii) is not required for the orthogonality $\mathbb{E}[(Z - \pi_0(R))\,\varepsilon] = 0$ used by the tri-score; it matters only for mapping to triple-proxy identification results.

(I1) (Valid IV) The cue Z is a valid instrument conditional on the representation R. Formally, $\mathbb{E}[\varepsilon \mid Z, R] = 0$. This is the classic IV assumption: Z affects the outcome Y only through the treatment D, after conditioning on the controls R.

(I2) (Proxy route) Assumption 2.2 holds. Then $\mathbb{E}\big[(Z - \pi(R))\,\varepsilon\big] = 0$ even if $\mathbb{E}[\varepsilon \mid Z, R] \neq 0$. Hence the score identifies τ_0 without classical IV validity.

(I3) (Correct nuisance models) The regression nets are correctly specified: $\pi_\gamma(R) = \mathbb{E}[Z \mid R]$ and $m_\varphi(R) = \mathbb{E}[Y - \tau_0 D \mid R]$. In particular $\mathbb{E}[\varepsilon \mid R] = 0$, which delivers orthogonality.

Note that a distinctive feature of TRIV–Rep is that the three orthogonalization routes (I1)–(I3) based- estimator remains consistent if any one route holds, while efficiency is preserved in the union model. The self-supervised representation $\widehat{R}$, learned from unlabeled (X, Z), directly enables (I2) in high-dimensional designs by extracting the relevant $R_\eta(X)$ that satisfies the conditional moment restrictions without manual proxy selection. This integration of semi-supervised learning into the identification step has, to my knowledge, not been formalized in prior literature.

Remark 4 (Relationships among (I1)–(I3)) *Route* (I3) *implies* (I1) *because* $\mathbb{E}[\varepsilon \mid R] = 0$ *entails* $\mathbb{E}[\varepsilon \mid Z, R] = 0$. *Under Assumption 2.2 (route* (I2)*), the orthogonality* $\mathbb{E}[(Z - \pi(R))\,\varepsilon] = 0$ *holds, but in general* $\mathbb{E}[\varepsilon \mid Z, R]$ *need not be zero; thus* (I2) *does* not *imply* (I1). *The estimator is therefore triple-robust over the* union *model, with a partial nesting given by* (I3) $\subset$ (I1).

Theorem 5 (Triple-robust identification) *Assume weak relevance (Assumption 2.1) and that at least one of the conditions (I1)–(I3) holds. Then the population moment* $\Psi(\tau) = \mathbb{E}[(Z - \pi_0(R))(Y - \tau D - m_0(R))]$ *has a unique root at* $\tau = \tau_0$.

Proof [Proof of Theorem 5] Let $\varepsilon := Y - \tau_0 D$ and define $C := \mathbb{E}[(Z - \pi_0(R))D]$, which is nonzero by weak relevance (Assumption 2.1). Using $\mathbb{E}[(Z - \pi_0(R))\,\mu_{D,0}(R)] = 0$, the population moment can be written as $\Psi(\tau) = \mathbb{E}\big[(Z - \pi_0(R))\,\varepsilon\big] - (\tau - \tau_0)\,C$. Hence it suffices to show $\mathbb{E}[(Z - \pi_0(R))\,\varepsilon] = 0$. Under each route: *(I1) Valid IV given R.* If $\mathbb{E}[\varepsilon \mid Z, R] = 0$, then by iterated expectations, $\mathbb{E}[(Z - \pi_0(R))\,\varepsilon] = \mathbb{E}\big[\mathbb{E}[(Z - \pi_0(R))\,\varepsilon \mid Z, R]\big] = 0$. *(I2) Proxy route (cue–sufficiency and outcome isolation).* Assume $Z \perp U \mid R$ and $\varepsilon \perp (Z, R) \mid U$. Then $\varepsilon \perp Z \mid (U, R)$ and $\mathbb{E}[\varepsilon \mid U, R] = \mathbb{E}[\varepsilon \mid U]$. Therefore $\mathbb{E}\big[(Z - \pi_0(R))\,\varepsilon\big] = \mathbb{E}\Big[\mathbb{E}\big[Z\varepsilon - \pi_0(R)\varepsilon \mid U, R\big]\Big] = \mathbb{E}\Big[\big(\mathbb{E}[Z \mid U, R] - \mathbb{E}[Z \mid R]\big)\,\mathbb{E}[\varepsilon \mid U]\Big] = 0$, because $Z \perp U \mid R$ implies $\mathbb{E}[Z \mid U, R] = \mathbb{E}[Z \mid R]$. (Note that no completeness is needed for this step.) *(I3) Correct residual model.* If $\mathbb{E}[\varepsilon \mid R] = 0$, then $\mathbb{E}[(Z - \pi_0(R))\,\varepsilon] = \mathbb{E}\big[(Z - \pi_0(R))\,\mathbb{E}[\varepsilon \mid R]\big] = 0$. In all three cases $\mathbb{E}[(Z - \pi_0(R))\,\varepsilon] = 0$, so $\Psi(\tau) = -(\tau - \tau_0)C$ and, since $C \neq 0$, the unique root is $\tau = \tau_0$. ■

2.3 Proxy-only identification via conditional-mean-zero instruments (route I2)

Throughout this subsection Z may be vector-valued; for $h \in \mathcal{H}$ I interpret $h(Z, R)$ as any square-integrable scalar function with $\mathbb{E}[h \mid R] = 0$. When Z is multivariate, the canonical element $Z - \mathbb{E}[Z \mid R]$ is taken component-wise and scalar h may be any linear or nonlinear functional of it. Let $\sigma(R)$, $\sigma(Z, R)$, and $\sigma(U)$ denote the σ–algebras generated by the corresponding random elements. All random variables are square–integrable unless noted, and all equalities hold P–a.s. For L_2–statements below, measurability is always with respect to $\sigma(Z, R)$ unless specified. Define $\mu_D(r) := \mathbb{E}[D \mid R = r]$, and write $\alpha(Z, R) := \mathbb{E}[D \mid Z, R] - \mu_D(R)$ and $w(Z, R) := \mathbb{E}[\varepsilon^2 \mid Z, R]$. Assume $w > 0$ a.s. and $\mathbb{E}[w] < \infty$. Maintain the structural model and the MAR weighting $v_i := S_i/\rho_n$ from Sections 2.1–2.3, and write $\varepsilon := Y - \tau_0 D$.

Assumption 2.3 (I2 proxy structure: cue–sufficiency and outcome isolation) *There exists a latent U such that* Cue–sufficiency: $Z \perp\!\!\!\perp U \mid R$ *and* Outcome isolation: $\varepsilon \perp\!\!\!\perp (Z, R) \mid U$.

Lemma 6 (Bridge existence under I2) *Under Assumption 2.3 there exists a measurable bridge $b : \mathcal{R} \to \mathbb{R}$ such that $\mathbb{E}[\varepsilon \mid Z, R] \;=\; b(R) \quad \text{with} \quad b(R) \;=\; \mathbb{E}\big[\mathbb{E}[\varepsilon \mid U] \mid R\big]$. In particular,*

$$\mathbb{E}[\,\varepsilon - b(R)\,|\,Z, R] \;=\; 0, \tag{5}$$

which holds without invoking any completeness condition.

Remark 7 (Role of completeness) *Completeness of R for functions of U is* not *required for Lemma 6 or for the orthogonality $\mathbb{E}[(Z - \pi_0(R))\,\varepsilon] = 0$ used by the tri-score. Completeness becomes relevant only when mapping to triple-proxy identification (e.g. Deaner, 2023) or for recovering U up to a one-to-one transform.*

Proof [Proof of Lemma 6] By outcome isolation, $\varepsilon \perp\!\!\!\perp (Z, R) \mid U$, hence $\mathbb{E}[\varepsilon \mid U, Z, R] = \mathbb{E}[\varepsilon \mid U]$ a.s. By cue–sufficiency, $Z \perp\!\!\!\perp U \mid R$, so conditioning on (Z, R) is equivalent to conditioning on R for any measurable function of U: $\mathbb{E}\big[\mathbb{E}[\varepsilon \mid U] \mid Z, R\big] \;=\; \mathbb{E}\big[\mathbb{E}[\varepsilon \mid U] \mid R\big] \;=:\; b(R)$. Thus $\mathbb{E}[\varepsilon \mid Z, R] \;=\; \mathbb{E}\big[\mathbb{E}[\varepsilon \mid U, Z, R] \mid Z, R\big] \;=\; \mathbb{E}\big[\mathbb{E}[\varepsilon \mid U] \mid Z, R\big] \;=\; b(R)$. Measurability of b with respect to $\sigma(R)$ follows from the Doob–Dynkin lemma, and square integrability from $\mathbb{E}[\varepsilon^2] < \infty$ and Jensen's inequality. Therefore $\mathbb{E}[\varepsilon - b(R) \mid Z, R] = 0$ without any completeness assumption.[5] ■

Define the *I2 instrument class*

$$\mathcal{H} \;:=\; \big\{\, h(Z, R) \in L_2 \;:\; \mathbb{E}[h(Z, R) \mid R] = 0 \,\big\}. \tag{6}$$

Multiplying (5) by any $h \in \mathcal{H}$ and taking expectations yields the *proxy-only* unconditional moments

$$\mathbb{E}\big[(Y - \tau D)\, h(Z, R)\big] \;=\; 0 \qquad \forall\, h \in \mathcal{H}. \tag{7}$$

Assumption 2.4 (Residual relevance over $\mathcal{H}$) *There exists $h^\circ \in \mathcal{H}$ with $\mathbb{E}\big[D\, h^\circ(Z, R)\big] \neq 0$.*

[5] The next results use only Assumption 2.3. Completeness is *not* invoked for identification or inference under I2; it appears later solely to connect to triple-proxy identification results.

Theorem 8 (I2-only identification by proxy moments) *Under Assumptions 2.3 and 2.4, the scalar τ_0 is the unique value satisfying* (7)*. In particular, if $h^\circ \in \mathcal{H}$ has $\mathbb{E}[D\,h^\circ] \neq 0$, then $\tau_0 = \frac{\mathbb{E}\big[Y\,h^\circ(Z,R)\big]}{\mathbb{E}\big[D\,h^\circ(Z,R)\big]}$.*

Proof [Proof of Theorem 8] Let $\mathcal{H} = \{h \in L_2(\sigma(Z,R)) : \mathbb{E}[h \mid R] = 0\}$. For any $h \in \mathcal{H}$, $\mathbb{E}[(Y - \tau D)h] = \mathbb{E}[(\varepsilon + (\tau_0 - \tau)D)h] = \underbrace{\mathbb{E}[\varepsilon h]}_{(i)} + (\tau_0 - \tau)\,\mathbb{E}[D\,h]$. To evaluate (i), use Lemma 6: $\mathbb{E}[\varepsilon h] = \mathbb{E}\big[\mathbb{E}[\varepsilon \mid Z,R]\,h\big] = \mathbb{E}\big[b(R)\,h(Z,R)\big] = \mathbb{E}\big[\mathbb{E}[h \mid R]\;b(R)\big] = 0$, since $\mathbb{E}[h \mid R] = 0$ a.s. Hence $\mathbb{E}[(Y - \tau D)h] = (\tau_0 - \tau)\,\mathbb{E}[D\,h]$. If $\mathbb{E}[(Y - \tilde{\tau}D)h] = 0$ for all $h \in \mathcal{H}$, in particular for $h^\circ \in \mathcal{H}$ from Assumption 2.4 with $\mathbb{E}[D\,h^\circ] \neq 0$, I must have $\tilde{\tau} = \tau_0$. The ratio formula follows by taking $h = h^\circ$. ■

Example 1 (I2 holds but I1 fails) *Let $U \sim \mathcal{N}(0,1)$, $R = h_2(U) + \xi_r$, $Z = \pi(R) + \xi_z$, and $\varepsilon = h_3(U) + \xi_\varepsilon$, with $(\xi_r, \xi_z, \xi_\varepsilon) \perp (U, \xi_{r'}, \xi_{z'}, \xi_{\varepsilon'})$ mutually independent and mean–zero. Then $Z \perp\!\!\!\perp U \mid R$ (since Z depends on U only through R) and $\varepsilon \perp\!\!\!\perp (Z,R) \mid U$. Therefore Assumption 2.3 holds. I have $\mathbb{E}[\varepsilon \mid Z,R] = \mathbb{E}[\mathbb{E}[\varepsilon \mid U] \mid Z,R] = \mathbb{E}[h_3(U) \mid R] =: b(R)$, which is generically nonzero unless $h_3 \equiv 0$. Hence the classical IV condition $\mathbb{E}[\varepsilon \mid Z,R] = 0$ (I1) fails in general. Yet $\mathbb{E}\big[(Z - \mathbb{E}[Z \mid R])\,\varepsilon\big] = \mathbb{E}\Big[\mathbb{E}\big[(Z - \mathbb{E}[Z \mid R])\,\varepsilon \,\big|\, U,R\big]\Big] = \mathbb{E}\Big[\underbrace{\mathbb{E}[Z - \mathbb{E}[Z \mid R] \mid U,R]}_{=0}\;\underbrace{\mathbb{E}[\varepsilon \mid U,R]}_{=\mathbb{E}[\varepsilon|U]}\Big] = 0$, so the I2 orthogonality holds and Theorem 8 identifies τ_0 via any relevant $h \in \mathcal{H}$.*

Remark 9 (Distinct from the classical orthogonal IV score) *Moment* (7) *residualizes* only *the instrument side via the constraint $\mathbb{E}[h \mid R] = 0$; it does* not *subtract $\mu_Y(R) - \tau\,\mu_D(R)$ from $Y - \tau D$. The usual orthogonal IV score corresponds to the singleton choice $h(Z,R) = Z - \mathbb{E}[Z \mid R]$* and *additionally residualizes Y and D on R. Thus* (7) *defines a strictly larger, I2-specific instrument class and an alternative orthogonalization strategy.*

Proposition 10 (Over-identification under I2) *Let $h_1, \ldots, h_J \in \mathcal{H}$ with $\mathbb{E}[D\,h_j] \neq 0$ for all j. Under Assumption 2.3, the J unconditional moments $\mathbb{E}[(Y - \tau D)h_j(Z,R)] = 0$ identify* the same *scalar τ_0. Hence the GMM estimator with stacked moments is just-identified in expectation, and the Hansen J statistic is asymptotically χ^2_{J-1} under correct specification.*

Proof [Proof of Proposition 10] For each j, Lemma 6 implies $\mathbb{E}[(Y - \tau_0 D)h_j] = \mathbb{E}[\varepsilon\,h_j] = \mathbb{E}[b(R)h_j] = \mathbb{E}[\mathbb{E}[h_j \mid R]\,b(R)] = 0$. If some $\tilde{\tau} \neq \tau_0$ solved $\mathbb{E}[(Y - \tilde{\tau}D)h_j] = 0$ for all j, then $0 = \mathbb{E}[(Y - \tilde{\tau}D)h_j] = \mathbb{E}[(Y - \tau_0 D)h_j] + (\tau_0 - \tilde{\tau})\mathbb{E}[D\,h_j] = (\tau_0 - \tilde{\tau})\mathbb{E}[D\,h_j]$, forcing $\mathbb{E}[D\,h_j] = 0$ for all j, contrary to the residual–relevance assumption. Hence the moments identify the same scalar τ_0. Standard just–identified GMM theory with a single scalar parameter and J valid moments yields the asymptotic χ^2_{J-1} distribution for Hansen's J–statistic under usual regularity (LLN/CLT and a nonsingular long–run variance of the stacked moment). ■

Within the class $\mathcal{H} = \{s(Z,R) : \mathbb{E}[s \mid R] = 0,\ \mathbb{E}[s^2] < \infty\}$, the efficient instrument for the ratio moment $s(Z,R)\{Y - \tau D\}$ solves

$$s^\star \;=\; \arg\max_{s\in\mathcal{H}} \frac{\big(\mathbb{E}[s\,D]\big)^2}{\mathbb{E}[s^2\,\varepsilon^2]}. \tag{8}$$

Standard calculations (e.g. orthogonality and Cauchy–Schwarz in the Hilbert space $\{s : \mathbb{E}[s \mid R] = 0\}$) give $s^\star(Z,R) \;\propto\; \frac{\mathbb{E}[D-\mu_D(R)|Z,R]}{\mathbb{E}[\varepsilon^2|Z,R]}$, which under homoskedasticity reduces to $s^\star \propto \mathbb{E}[D - \mu_D(R) \mid Z, R]$. This characterization is *I2-specific* because the optimization is taken over $\mathcal{H}$ (the *proxy* instrument space), rather than over all L_2 instruments as in the classical conditional IV model; the resulting influence function is efficient within the I2 submodel (e.g. Newey, 1990; Ai and Chen, 2003).

Let $\hat{h}$ be a cross-fitted estimate of some $h^\circ \in \mathcal{H}$ (trained on folds that exclude the observation used for scoring) and define

$$\psi_i^{(\mathrm{I2})}(\tau,\hat{h}) \;:=\; v_i\,\hat{h}(Z_i,R_i)\,\{Y_i - \tau D_i\}, \qquad \hat{\Psi}_n^{(\mathrm{I2})}(\tau) \;:=\; \frac{1}{n}\sum_{i=1}^{n} \psi_i^{(\mathrm{I2})}(\tau,\hat{h}). \tag{9}$$

The estimating equation $\hat{\Psi}_n^{(\mathrm{I2})}(\tau) = 0$ has the closed-form solution $\hat{\tau}_{\mathrm{I2}} \;=\; \frac{\sum_i v_i\,\hat{h}(Z_i,R_i)\,Y_i}{\sum_i v_i\,\hat{h}(Z_i,R_i)\,D_i}$. Under MAR with constant label rate, replacing $v_i = S_i/\rho_n$ by $S_i/\hat{\rho}_n$ is second–order (Slutsky).

Proposition 11 (Efficient instrument within $\mathcal{H}$) *Over $\mathcal{H} = \{s : \mathbb{E}[s \mid R] = 0,\ \mathbb{E}[s^2\varepsilon^2] < \infty\}$, the maximizer of $\frac{\{\mathbb{E}[sD]\}^2}{\mathbb{E}[s^2\varepsilon^2]}$ is unique up to scale and satisfies $s^\star(Z,R) \;\propto\; \frac{\mathbb{E}[D-\mu_D(R)|Z,R]}{\mathbb{E}[\varepsilon^2|Z,R]}$.*

Proof [Proof of Proposition 11] Maximize $\mathbb{E}[sD]$ subject to $\mathbb{E}[s^2\varepsilon^2] = 1$ and $\mathbb{E}[s \mid R] = 0$ using a Lagrangian with multiplier $\lambda \in \mathbb{R}$ and a function multiplier $\mu(R)$: $\mathcal{L}(s) = \mathbb{E}[sD] - \frac{\lambda}{2}\,\mathbb{E}[s^2\varepsilon^2] - \mathbb{E}[\mu(R)s]$. The FOC is $\mathbb{E}[\delta s\{D - \lambda\varepsilon^2 s - \mu(R)\}] = 0$ for all δs, hence $D = \lambda\varepsilon^2 s + \mu(R)$ a.s. Taking $\mathbb{E}[\cdot \mid Z, R]$ and subtracting $\mathbb{E}[\cdot \mid R]$ yields $\mathbb{E}[D - \mu_D(R) \mid Z, R] = \lambda\,\mathbb{E}[\varepsilon^2 \mid Z, R]\,s(Z,R)$, proving the claim. ■

Lemma 12 (Orthogonality with an estimated proxy instrument) *Let the sample be split into $K \geq 2$ folds. On each training complement $(-k)$, fit any preliminary $\hat{h}^{(-k)}(Z,R)$, and estimate the* population *conditional mean of this learner by regressing the pseudo–outcome $\hat{h}^{(-k)}(Z,R)$ on R using only the training folds: $\widehat{m}^{(-k)}(r) \;\approx\; \mathbb{E}\big[\hat{h}^{(-k)}(Z,R) \mid R = r\big]$. Define the cross–fitted, population–demeaned instrument on the held–out fold $\mathcal{I}_k$ by $\hat{h}_\perp^{(-k)}(Z_i,R_i) \;:=\; \hat{h}^{(-k)}(Z_i,R_i) - \widehat{m}^{(-k)}(R_i), i \in \mathcal{I}_k$, and set $\hat{h}_i := \hat{h}_\perp^{(-k)}(Z_i,R_i)$. Let $b(R) = \mathbb{E}[\mathbb{E}(\varepsilon \mid U) \mid R]$ be the bridge from Lemma 6 and assume $b \in L_2$. If, for each k, the mean–regression error obeys $\|\,\widehat{m}^{(-k)} - \mathbb{E}[\hat{h}^{(-k)} \mid R]\,\|_2 = o_p(n^{-1/2})$, then, at $\tau = \tau_0$, $\mathbb{E}\Big[\big\{Y - \tau_0 D\big\}\,\hat{h}(Z,R)\,\Big|\,\{\text{training folds}\}\Big] = o_p(n^{-1/2})$, so the IPW moment is first–order valid (mean $o_p(n^{-1/2})$). If $\|\hat{h} - h^\circ\|_2 = o_p(1)$ for some fixed $h^\circ \in \mathcal{H}$ with $\mathbb{E}[D\,h^\circ] \neq 0$, the ratio estimator $\hat{\tau}_{\mathrm{I2}} \;=\; \frac{\sum_i v_i\,\hat{h}(Z_i,R_i)\,Y_i}{\sum_i v_i\,\hat{h}(Z_i,R_i)\,D_i}$ is consistent, and admits $\sqrt{n}$–asymptotic normality under a row–wise CLT for the weighted summands.*

Proof [Proof of Lemma 12] By Lemma 6, at τ_0 I have $\mathbb{E}[(Y-\tau_0 D) \mid Z, R] = b(R)$. Condition on the training folds so that $\hat{h}^{(-k)}$ and $\widehat{m}^{(-k)}$ are fixed functions on the held–out fold. Then

$$\begin{aligned}\mathbb{E}\Big[\{Y-\tau_0 D\}\, \hat{h}_{\perp}^{(-k)}(Z,R) \,\Big|\, \text{training}\Big] &= \mathbb{E}\Big[b(R)\,\{\hat{h}^{(-k)}(Z,R) - \widehat{m}^{(-k)}(R)\}\Big] \qquad (10)\\ &= \mathbb{E}\Big[b(R)\,\{\mathbb{E}[\hat{h}^{(-k)}(Z,R) \mid R] - \widehat{m}^{(-k)}(R)\}\Big]. \qquad (11)\end{aligned}$$

By Cauchy–Schwarz and $b \in L_2$, $\big|\mathbb{E}\big[b(R)\{\mathbb{E}[\hat{h}^{(-k)} \mid R] - \widehat{m}^{(-k)}(R)\}\big]\big| \le \|b\|_2 \,\big\|\widehat{m}^{(-k)} - \mathbb{E}[\hat{h}^{(-k)} \mid R]\,\big\|_2 = o_p(n^{-1/2})$. This yields the stated first–order validity. For the estimator, write $\hat{\tau}_{\mathrm{I2}} - \tau_0 = \frac{\frac{1}{n}\sum_i v_i\, \hat{h}_i\, (Y_i - \tau_0 D_i)}{\frac{1}{n}\sum_i v_i\, \hat{h}_i\, D_i}$. The numerator equals $n^{-1}\sum_i v_i\, \hat{h}_i\, \varepsilon_i$ whose conditional mean is $o_p(n^{-1/2})$ by the argument above and whose variance obeys a row–wise Lindeberg CLT (assumed). The denominator converges in probability to $\mathbb{E}[D\, h^\circ] \neq 0$ by $\|\hat{h} - h^\circ\|_2 = o_p(1)$. Slutsky's theorem gives consistency and $\sqrt{n}$–normality. ■

Two simple and valid choices are: (a) $\hat{h}(Z,R) = Z - \widehat{\mathbb{E}}[Z \mid R]$ (nonparametric/logistic), and (b) the "optimal" element of a sieve $\mathcal{H}_K$ that maximizes held-out correlation with D subject to $\mathbb{E}[h \mid R] = 0$: $\hat{h} \in \arg\max_{h\in\mathcal{H}_K} \big|\mathbb{E}_{\text{fold}}[D\, h(Z,R)]\big|$ s.t. $\mathbb{E}_{\text{fold}}[h(Z,R) \mid R] = 0$, with all nuisance fits and the choice of h cross-fitted. Assumption 2.4 is testable from labeled-free data by a first-stage t-test for $\mathbb{E}[D\,\hat{h}] \neq 0$. With fixed h°, the influence function for τ is $-\{h^\circ(Z,R)(Y-\tau_0 D)\}/\mathbb{E}[D\, h^\circ]$, yielding the usual ratio-moment variance: $\mathrm{Var}(\hat{\tau}_{\mathrm{I2}}) \approx \frac{\mathrm{Var}\big(h^\circ(Z,R)(Y-\tau_0 D)\big)}{n\,\big(\mathbb{E}[D\, h^\circ(Z,R)]\big)^2}$. Cross-fitting $\hat{h}$ preserves first-order validity because $\mathbb{E}[\{Y - \tau_0 D\}\hat{h}(Z,R) \mid \text{training}] = \mathbb{E}[b(R)\hat{h}(Z,R)] = 0$ via $\mathbb{E}[\hat{h} \mid R] = 0$.

Algorithm 1 I2 proxy–instrument learner (cross-fitted)

```
Input: Folds $\{\mathcal{I}_k\}_{k=1}^K$, basis $\{g_j(R)\}_{j=1}^K$
for $k = 1, \dots, K$ do
    Fit $\widehat{\pi}^{(-k)}(r) \approx \mathbb{E}[Z \mid R = r]$ on $\cup_{\ell\neq k}\mathcal{I}_\ell$
    For $i \in \mathcal{I}_k$, set $\tilde{z}_i := Z_i - \widehat{\pi}^{(-k)}(R_i)$
    Fit weights $\hat{w}^{(-k)}$ on $\cup_{\ell\neq k}\mathcal{I}_\ell$ by regressing $D$ on $\{\tilde{z}\, g_j(R)\}_{j=1}^K$ (ridge or lasso)
    Define $\hat{h}_i := \sum_{j=1}^K \hat{w}_j^{(-k)}\, \tilde{z}_i\, g_j(R_i)$ for $i \in \mathcal{I}_k$
    (Projection) On the training folds, regress the pseudo–outcome $\hat{h}^{(-k)}(Z,R) := \sum_j \hat{w}_j^{(-k)}\,(Z - \widehat{\pi}^{(-k)}(R))\, g_j(R)$ on $R$ to get $\widehat{m}^{(-k)}(r) \approx \mathbb{E}[\hat{h}^{(-k)} \mid R = r]$; for $i \in \mathcal{I}_k$ set $\hat{h}_i \leftarrow \hat{h}^{(-k)}(Z_i,R_i) - \widehat{m}^{(-k)}(R_i)$.
end for
Output: Cross-fitted instrument $\hat{h}$ with $\mathbb{E}[\hat{h} \mid R] \approx 0$
```

Corollary 13 (Tri-score under I2: validity and when it is efficient) *Under Assumption 2.3, for every $h \in \mathcal{H}$ the population moment $\mathbb{E}\big[h(Z,R)\{Y - \tau D\}\big] = 0$ has the unique root τ_0. In particular, the "orthogonal tri–score" $\psi_i(\tau,\theta) = v_i\,[Z_i - \pi(R_i)]\Big([Y_i - \mu_Y(R_i)] - \tau[D_i - \mu_D(R_i)]\Big)$ has the same population root τ_0 (residualizing Y and D on R*

does not affect identification). Moreover, let $s^\star$ denote the I2–efficient instrument from the variational problem (8)*. The score based on $h = s^\star$ is locally semiparametric–efficient in the I2 submodel. The special choice $h(Z,R) = Z - \pi(R)$ is generally* not *efficient unless* $\mathbb{E}[D - \mu_D(R) \mid Z, R] \;\propto\; Z - \pi(R)$ *a.s. (e.g. a homoskedastic linear first stage).*

Proof [Proof of Corollary 13] Validity: for any $h \in \mathcal{H}$, $\mathbb{E}\big[h(Z,R)\{Y - \tau_0 D\}\big] = \mathbb{E}\big[h(Z,R)\,\mathbb{E}[Y - \tau_0 D \mid Z, R]\big] = \mathbb{E}\big[h(Z,R)\,b(R)\big] = 0$, because $b(R)$ is R–measurable and $\mathbb{E}[h \mid R] = 0$. For the tri–score, expand $\mathbb{E}\Big([Z - \pi(R)]\big(\mu_Y(R) - \tau\mu_D(R)\big)\Big) = \mathbb{E}\Big(\mathbb{E}[Z - \pi(R) \mid R]\big(\mu_Y(R) - \tau\mu_D(R)\big)\Big) = 0$, hence its population moment equals $\mathbb{E}\big([Z - \pi(R)](Y - \tau D)\big)$ and has the same unique root τ_0. Efficiency: within the proxy instrument space $\mathcal{H} = \{h : \mathbb{E}[h \mid R] = 0,\ \mathbb{E}[h^2] < \infty\}$, the efficient instrument is $s^\star$ from (8). A score using $h = s^\star$ attains the semiparametric efficiency bound in the I2 submodel. The choice $h = Z - \pi(R)$ coincides with $s^\star$ only under the alignment condition stated; otherwise it is valid but generally sub-optimal. ■

2.4 Handling Label Scarcity via IPW

In practice, the outcome Y_i is often unobserved for most of the data. Let $S_i \in \{0, 1\}$ indicate whether Y_i is observed (labeled), and let $n_\ell = \sum_{i=1}^n S_i$ be the number of labeled units. Throughout this subsection write $R_i := R_\eta(X_i)$.

Assumption 2.5 (Missing at Random (MAR)) $S_i \perp Y_i \mid (X_i, D_i, Z_i)$ *(MAR) and* $\Pr(S_i = 1 \mid X_i, D_i, Z_i) = \rho > 0$ *(positivity). In the main analysis I assume ρ is constant.*

When all outcomes are observed ($S_i \equiv 1$), the *residualized* tri-score is $\psi_i(\tau, \theta) = [Z_i - \pi(R_i)]\,([Y_i - \mu_Y(R_i)] - \tau\,[D_i - \mu_D(R_i)])$, where the nuisance functions are $\mu_Y(R) := \mathbb{E}[Y \mid R]$, $\mu_D(R) := \mathbb{E}[D \mid R]$, and $\pi(R) := \mathbb{E}[Z \mid R]$. Equivalently, $C = \mathbb{E}[(Z - \pi(R))\,D]$ because $\mathbb{E}[(Z - \pi(R))\,\mu_D(R)] = 0$. The corresponding population moment is $\Psi(\tau) := \mathbb{E}[(Z - \pi(R))\{(Y - \mu_Y(R)) - \tau(D - \mu_D(R))\}] = \mathbb{E}[(Z - \pi(R))\,\varepsilon] - (\tau - \tau_0)C$, with $\varepsilon := Y - \tau_0 D$ and $C := \mathbb{E}[(Z - \pi(R))(D - \mu_D(R))]$ (*residual relevance*). Under MAR, the same moment can be estimated from the labeled subsample by inverse-probability weighting. Let $\hat\rho := n_\ell / n$ and set $v_i := S_i / \hat\rho$. The IPW tri-score is

$$\psi_{i,n}(\tau, \theta) \;=\; \frac{S_i}{\hat\rho}\,\big[Z_i - \pi(R_i)\big]\,\Big(\,[Y_i - \mu_Y(R_i)] - \tau\,[D_i - \mu_D(R_i)]\,\Big). \tag{12}$$

Because $S \perp Y \mid (X, D, Z)$ and $\mathbb{E}[S \mid X, D, Z] = \rho$, I have $\mathbb{E}\big[\frac{S}{\rho}\,\psi_i(\tau, \theta)\big] = \mathbb{E}[\psi_i(\tau, \theta)]$, so IPW recovers the fully-observed population moment. Replacing ρ by $\hat\rho$ perturbs the moment by $o_p(n^{-1/2})$ when ρ is constant (Slutsky's theorem), and therefore does not affect first-order inference.[6] Thus, all identification results in Section 2.2—via a valid instrument (I1), the proxy route (I2), or a correct treatment–residual model (I3)—carry over directly

[6] If the label probability varies with covariates, replace $S_i/\hat\rho$ by $S_i/\hat q(X_i, D_i, Z_i)$ with a consistent estimate $\hat q$ of $q(X, D, Z) := \Pr(S = 1 \mid X, D, Z)$. All arguments go through unchanged. When correcting MNAR with $q_\delta(W)$ (with $W := (Z, R, D, X)$), training μ_Y with weights $1/\hat q_\delta(W)$ targets $\mathbb{E}[Y \mid R]$ under the full-data distribution and can improve rates.

to the label-scarce setting with the IPW score (12). Under MAR with constant ρ, weighting labeled rows by $S/\hat{\rho}$ makes the labeled empirical risk minimization unbiased for the full-sample risk, so $\widehat{\mu}_Y$ targets $\mathbb{E}[Y \mid R]$. The causal structures for these routes are summarized in Figure 1. Core identification and estimation rely on MAR. A fuller treatment of MNAR—its implications, modeling options, and sensitivity analysis—appears in §B.

2.5 Estimation with Nuisance Learners

The identification strategy relies on three unknown nuisance functions $\theta_0 = (\eta_0, \gamma_0, \varphi_0)$, which I approximate by deep ReLU networks trained with K-fold cross-fitting. Raw covariates (X, Z, D, S) are mapped by a self-supervised encoder R_η into a low-dimensional representation R. Using R and the labeled outcomes Y, I fit an instrument (or cue) regression π_γ, outcome and treatment regressions μ_Y and μ_D, and—when outcomes are missing not at random (MNAR)—a selection-weight model q_δ. When no MNAR layer is modeled, set $q_\delta \equiv 1$; the remaining steps are unchanged. Let $\hat{\theta}_i$ denote the fold-specific nuisance vector estimated without row i. These nuisance blocks feed the orthogonal tri-score $\psi(\tau, \hat{\theta})$, whose root yields $\hat{\tau}$. $\hat{\Psi}_n(\hat{\tau}) := \frac{1}{n}\sum_{i=1}^{n} \psi_{i,n}(\hat{\tau}, \hat{\theta}_i) = 0$, where $\hat{\theta}_i$ is the fold-specific nuisance vector estimated without row i. This construction isolates first-order estimation error in the nuisance blocks and delivers standard Z-estimation with cross-fitting for inference.

2.6 Outcome panels and latent–factor proxies

In many digital settings each unit i generates *multiple* pre– or post–treatment outcome measurements, $\boldsymbol{Y}_i = (Y_{i1}, \dots, Y_{iT})^\top$, rather than the single scalar Y_i assumed so far. If latent user traits drive a low–rank factor structure in $\boldsymbol{Y}_i$, one can treat these repeated outcomes as *implicit proxies for the unobserved confounder* and estimate a representation R_i by matrix–completion techniques instead of contrastive SSL.

Suppose the pre–treatment outcome panel obeys[7] $Y_{it} = \lambda_i^\top f_t + \zeta_{it}, 1 \le t \le T$. Here $\lambda_i \in \mathbb{R}^r$ and $f_t \in \mathbb{R}^r$ are time–invariant loadings and common factors with $r \ll \min\{n, T\}$, while ζ_{it} is mean-zero idiosyncratic noise. Let $\boldsymbol{Y} := (Y_{it})_{i,t}$ be the $n \times T$ matrix of (pre-treatment) outcomes. Assume a rank–r latent–factor structure together with the standard *incoherence* and uniform–random sampling conditions. When the observation scheme is noiseless, the exact-recovery theorem of Candès and Recht (2009) shows that nuclear-norm minimisation perfectly recovers $\boldsymbol{Y}$ with high probability once the sampling rate satisfies $p \gtrsim r\mu\,(n+T)\log^2(n \vee T)/(nT)$, where μ is the incoherence constant. With additive, mean-zero sub-Gaussian noise ζ_{it} the non-asymptotic error bounds of Candès and Plan (2011) and Negahban and Wainwright (2012) imply that any nuclear-norm (or singular-value–thresholding) estimator $\hat{\boldsymbol{Y}}^{(\mathrm{mc})}$ satisfies $\frac{1}{nT}\left\|\hat{\boldsymbol{Y}}^{(\mathrm{mc})} - \boldsymbol{Y}\right\|_F^2 = O_p\Big(\frac{r(n+T)}{nT}\Big)$. This rate is minimax-optimal up to logarithmic factors and meets the $o(n^{-1/4})$ requirement used in Section 4 for the representation $R_i = \hat{\lambda}_i$. Denote by $\hat{\lambda}_i$ the ith row of the rank-r approximation; I set the *representation*

$$R_i \;:=\; \hat{\lambda}_i \quad \in \mathbb{R}^r. \tag{13}$$

[7] A similar structure with post–treatment t and appropriate exclusion can yield a *latent-factor instrument*; see Abadie et al. (2024) for details.

Write Z_i for the randomisation cue (instrument) and $\varepsilon_i := Y_i - \tau_0 D_i$ for the structural residual as in (1). With (13) the triple $(Z_i, R_i, \varepsilon_i)$ matches the three–proxy setup of Deaner (2023) . $(Z_i \perp\!\!\!\perp \varepsilon_i) \mid R_i$ holds because R_i captures the latent factor that drives both Z_i and ε_i (Conditional independence). Assumption 2.2 implies the L_2–completeness requirement in Deaner (2023), guaranteeing that $\mathbb{E}[h(U) \mid R_i] = 0 \Rightarrow h \equiv 0$ (Proxy completeness). Under these two conditions the population moment $\Psi(\tau) = \mathbb{E}[(Z - \pi_0(R))(\varepsilon - (\tau - \tau_0)D)]$ retains a unique zero at $\tau = \tau_0$, so all identification and efficiency results in Sections 3–5 continue to hold with the *matrix–completion proxy* R_i in place of the contrastive SSL embedding.

If $r = O(1)$ and $T \gtrsim \log n$, the matrix–completion error in (13) is $O_p(n^{-1/2})$. Plugging this into Assumption 3.2 shows that the cross–fitted nuisance bundle still achieves the joint $o(n^{-1/4})$ rate required for $\sqrt{n}$–inference (Theorem 34). When rich outcome panels are available, practitioners may replace the contrastive pre-training step by any consistent low-rank factor estimator. The remainder of the TRIV–Rep pipeline—orthogonal tri-score, cross-fitting, and variance estimation—remains unchanged.

Remark 14 (Latent-factor instruments) *If one column of* $\boldsymbol{Y}$ *(say, a post-treatment surrogate) affects* D_i *but not* Y_i *except through* D_i*, the corresponding loading* $\hat{\lambda}_i$ *can also serve as a* learned instrument*. The tri-score then enjoys identification through route (I1) even in the absence of an externally provided cue* Z_i*. (See Abadie et al. (2024) for formal conditions.)*

3 Tri-score and Orthogonality

The practical estimator solves the empirical counterpart of (12) using cross-fitting to estimate the nuisance functions. Let the nuisance parameters for observation i, estimated on other folds, be $\hat{\theta}_i = (\hat{\eta}_i, \hat{\gamma}_i, \hat{\varphi}_i)$. The estimating equation for the causal slope $\hat{\tau}$ is

$$\hat{\Psi}_n(\hat{\tau}) := \frac{1}{n} \sum_{i=1}^{n} \psi_{i,n}(\hat{\tau}, \hat{\eta}_i, \hat{\gamma}_i, \hat{\varphi}_i) = 0. \tag{14}$$

The key to achieving $\sqrt{n}$-consistency for $\hat{\tau}$ despite using flexible machine learning estimators for $\hat{\theta}_i$ is the *orthogonality* of the moment function ψ. Orthogonality ensures that plug-in errors from estimating the nuisance functions do not contaminate the estimation of the target parameter τ_0 at the first order. Note that the *tri-score* moment function inherits a minimax-optimal variance bound in the union model, achieving the semiparametric efficiency bound whenever all nuisance components are consistently estimated. The analysis collects rate lemmas for all three nuisance blocks, showing that $\widehat{\theta}$ remains $o_p(n^{-1/4})$-close to θ_0 uniformly over the union model—this rate being sufficient for root-n inference even under severely unbalanced label availability.

3.1 First-order (Gâteaux) Orthogonality

The moment function (12) is Neyman-orthogonal with respect to the nuisance parameters. This means that at the true parameter values (τ_0, θ_0), the population moment is locally insensitive to small perturbations in the nuisance functions.

Lemma 15 (Block-wise mean-zero derivatives) *For every nuisance block $b \in \{\eta, \gamma, \varphi\}$, every direction h_b in the corresponding tangent set, and every $n \geq 1$, $\partial_b \Psi_n(\tau_0, \theta_0)[h_b] \;=\; 0$.*

Proof [Proof of Lemma 15]

For each observation i let $\psi_i(\tau, \eta, \gamma, \varphi) \;=\; v_i\,[Z_i - \pi_\gamma(R_\eta(X_i))]\,[Y_i - \tau D_i - m_\varphi(R_\eta(X_i))]$, $v_i := S_i/\rho$, $\rho := \mathbb{E}[S_i]$. By definition of π_{γ_0} and m_{φ_0},

$$\mathbb{E}[Z_i - \pi_{\gamma_0}(R_i) \mid R_i] = 0, \quad \mathbb{E}[Y_i - \tau_0 D_i - m_{\varphi_0}(R_i) \mid R_i] = 0. \tag{15}$$

Because $\mathbb{E}[v_i \mid X_i, D_i, Z_i] = 1$ under MAR, premultiplying by v_i preserves these zero means. Differentiate inside the expectation; dominated convergence is justified by the uniform $(2+\delta)$-moment envelope in Assumption 5.3: $\partial_\gamma \Psi(\tau_0, \theta_0)[h_\gamma] = -\mathbb{E}\big[v_i\,h_\gamma(R_i)\,(Y_i - \tau_0 D_i - m_{\varphi_0}(R_i))\big] = 0$ by (15). With $\delta_\eta := h_\eta(X_i)$, $\partial_\eta \Psi(\tau_0, \theta_0)[h_\eta] = -\mathbb{E}\big[v_i \langle \nabla \pi_{\gamma_0}(R_i), \delta_\eta \rangle (Y_i - \tau_0 D_i - m_{\varphi_0}(R_i))\big] - \mathbb{E}\big[v_i (Z_i - \pi_{\gamma_0}(R_i)) \langle \nabla m_{\varphi_0}(R_i), \delta_\eta \rangle\big] = 0$, again because each bracket has conditional mean zero given R_i. $\partial_\varphi \Psi(\tau_0, \theta_0)[h_\varphi] = -\mathbb{E}\big[v_i\,(Z_i - \pi_{\gamma_0}(R_i))\,h_\varphi(R_i)\big] = 0$, by (15). All first-order derivatives thus vanish. ■

First-order orthogonality gives $\partial_\eta \Psi(\tau_0, \theta_0) = 0$, so every second-order term in the η direction is linear in h_η. Consequently the quadratic form $\partial^2_{\eta\eta} \Psi(\tau_0, \theta_0)[h_\eta, h_\eta]$ vanishes identically.

Assumption 3.1 (Identification & first-order orthogonality) *Let $\psi_i(\tau, \theta)$ be a score with population moment $\Psi(\tau, \theta) := \mathbb{E}[\psi_i(\tau, \theta)]$. (a)* Centred at truth: $\Psi(\tau_0, \theta_0) = 0$. *(b)* Identification of the target: $\partial_\tau \Psi(\tau_0, \theta_0) \neq 0$. *(c)* First-order orthogonality: *for every nuisance block $a \in \{\eta, \gamma, \varphi\}$ and every direction $h_a \in \mathcal{T}_a$, $\Psi'_a[h_a] = \frac{d}{dt} \Psi(\tau_0, \theta_0 + t\,h_a)\big|_{t=0} = 0$. Equivalently, $\partial_\eta \Psi(\tau_0, \theta_0) = \partial_\gamma \Psi(\tau_0, \theta_0) = \partial_\varphi \Psi(\tau_0, \theta_0) = 0$.*

Corollary 16 (Empirical first-order cancellation) *$\frac{1}{n} \sum_{i=1}^n \psi_{i,n}(\tau_0, \hat\eta_i, \hat\gamma_i, \hat\varphi_i) = O_p(n^{-1/2})$, with K-fold cross-fitting.*

Proof [Proof of Corollary 16] Condition on the K training folds so that $(\hat\eta_i, \hat\gamma_i, \hat\varphi_i)$ is fixed on the held-out row i. Lemma 15 then gives conditional mean zero of every summand. A triangular-array Chebyshev (or CLT) with the $(2+\delta)$ envelope from Assumption 5.3 yields the $n^{-1/2}$ rate; unconditioning preserves the order. ■

3.2 Second–order empirical reminder

Let the K-fold cross-fitted nuisance vector in row i be $\hat\theta_i := (\hat\eta_i, \hat\gamma_i, \hat\varphi_i)$, with $\Delta\theta_i := \hat\theta_i - \theta_0 = (\Delta\eta_i, \Delta\gamma_i, \Delta\varphi_i)$.

Assumption 3.2 (Cross-fit $n^{-1/4}$ rates) *For each block $b \in \{\eta, \gamma, \varphi\}$, $\|\hat b_i - b_0\|_{2,n} = o_p(n^{-1/4})$, uniformly in i.*

Proposition 17 (Second-order empirical remainder) *Under Assumptions 5.3 and 3.2, $\frac{1}{\sqrt{n}} \sum_{i=1}^n \{\psi_{i,n}(\tau_0, \hat\theta_i) - \psi_{i,n}(\tau_0, \theta_0)\} = o_p(1)$.*

Proof [Proof of Proposition 17] See Appendix A.1. ■

Corollary 18 (Bias control) *Combine Proposition 17 with the sample-centred first-order result of Corollary 16: $\hat{\Psi}_n(\tau_0, \hat{\theta}) := \frac{1}{n}\sum_{i=1}^{n} \psi_{i,n}(\tau_0, \hat{\theta}_i) = o_p(n^{-1/2})$. Hence, the plug-in bias is negligible relative to the $\sqrt{n}$ variance scale of the score.*

The block-Hessian calculations and the local minimax theorem are handled separately in 3.2; no arguments are duplicated across the two sections. As shown in Proposition 17, the plug-in bias is negligible. The complete estimation routine is given in Algorithm 3. Throughout, put $\psi_i(\tau, \theta) := v_i\{Z_i - \pi_\gamma(R_\eta(X_i))\}\{Y_i - \tau D_i - m_\varphi(R_\eta(X_i))\}$ and $\Psi(\tau, \theta) := \mathbb{E}[\psi_i(\tau, \theta)]$.

Here it is, kept identical in content—just cleanly formatted so it drops in without surprises:

Algorithm 2 TRIV–Rep (MAR): Cross-Fitted Triple-Robust IV

```
Input: Data $\{(X_i, Z_i, D_i, S_i, Y_i)\}_{i=1}^{n}$; folds $K$
Pre-train $R_\eta$ on $\{X_i\}$; freeze $\eta$ and set $R_i = R_\eta(X_i)$
$n_L \leftarrow \sum_i S_i$; $\hat{\rho} \leftarrow n_L/n$
for $k = 1, \ldots, K$ do
    Fit $\widehat{\pi}^{(-k)}(r) \approx \mathbb{E}[Z \mid R]$, $\widehat{\mu}_D^{(-k)}(r) \approx \mathbb{E}[D \mid R]$
    Fit $\widehat{\mu}_Y^{(-k)}(r)$ on labeled rows with weights $S_i/\hat{\rho}$
    for each labeled $i$ in fold $k$ do
        $\hat{z}_i \leftarrow Z_i - \widehat{\pi}^{(-k)}(R_i)$;   $\widehat{d}_i \leftarrow D_i - \widehat{\mu}_D^{(-k)}(R_i)$;   $\widehat{y}_i \leftarrow Y_i - \widehat{\mu}_Y^{(-k)}(R_i)$
        $\psi_i(\tau) \leftarrow \frac{S_i}{\hat{\rho}}\, \hat{z}_i\big(\widehat{y}_i - \tau\, \widehat{d}_i\big)$
    end for
end for
Solve $\sum_{i:S_i=1} \psi_i(\hat{\tau}) = 0$; with scalar $Z$, $\hat{\tau} = \dfrac{\sum_i \frac{S_i}{\hat{\rho}}\, \hat{z}_i\, \widehat{y}_i}{\sum_i \frac{S_i}{\hat{\rho}}\, \hat{z}_i\, \widehat{d}_i}$
Output: $\hat{\tau}$ and CI
```

Lemma 19 (Exact block Hessian) *Let $h_a \in \mathcal{T}_a$. At (τ_0, θ_0): $\partial^2_{aa}\Psi[h_a, h_a] = 0$, $a \in \{\eta, \gamma, \varphi_Y, \varphi_D\}$; $\partial^2_{\gamma,\varphi_Y}\Psi[h_\gamma, h_Y] = \mathbb{E}[h_\gamma(R)\, h_Y(R)]$; $\partial^2_{\gamma,\varphi_D}\Psi[h_\gamma, h_D] = -\tau_0\, \mathbb{E}[h_\gamma(R)\, h_D(R)]$; $\partial^2_{\eta,\gamma}\Psi[h_\eta, h_\gamma] = \mathbb{E}[h_\gamma(R)\, \langle \nabla(\mu_{Y,0} - \tau_0 \mu_{D,0})(R),\, h_\eta(X)\rangle]$; $\partial^2_{\eta,\varphi_Y}\Psi[h_\eta, h_Y] = \mathbb{E}[h_Y(R)\, \langle \nabla \pi_{\gamma_0}(R),\, h_\eta(X)\rangle]$; $\partial^2_{\eta,\varphi_D}\Psi[h_\eta, h_D] = -\tau_0\, \mathbb{E}[h_D(R)\, \langle \nabla \pi_{\gamma_0}(R),\, h_\eta(X)\rangle]$.*

Assumption 3.3 (Bounded block Hessians) *For every $a, b \in \{\eta, \gamma, \varphi\}$ the mixed second derivative $\partial^2_{ab}\Psi(\tau_0, \theta_0)$ exists as a continuous bilinear form on $\mathcal{T}_a \times \mathcal{T}_b$ and is finite.*

Proof [Proof of Lemma 19]

First-order orthogonality eliminates every diagonal block, giving (19). For the mixed terms I differentiate the product representation of Ψ twice: (19)–(19) follow from a product rule applied to $\partial_\gamma \Psi$ and the conditional-mean identity $\mathbb{E}[Y - \tau_0 D - m_{\varphi_0}(R) \mid R] = 0$;

(19)–(19)–(19) obtain by perturbing the representation $R_\eta(X)$ and linearising π_γ and m_φ around η_0. Complete proof is provided in A.2. $\blacksquare$

Definition 20 (Worst–case quadratic bias) *Assume 3.3 so that $\tilde{\Psi}$ is twice Gâteaux-differentiable. For $\delta > 0$ and an identification-valid score $\tilde{\psi}$ with population moment $\tilde{\Psi}(\tau, \theta)$ set $B_\delta(\tilde{\psi}) := \sup_{\|h\| \le \delta} \big|\tilde{\Psi}(\tau_0, \theta_0 + h)\big|$.*

Theorem 21 (Local minimaxity up to constants) *Under Assumptions 3.1 and 3.3, let ψ be the IPW tri–score and let $\tilde{\psi}$ be any other identification-valid score that is first-order orthogonal at θ_0. Define the norm $\|h\|^2 := \|h_\eta\|_{L_2}^2 + \|h_\gamma\|_{L_2}^2 + \|h_\varphi\|_{L_2}^2$. Then there exist constants $0 < c_\star \le C^\star < \infty$, depending only on the operator norms $\|\partial_{ab}^2 \Psi(\tau_0, \theta_0)\|_{\mathrm{op}}$, such that for all $\delta > 0$, $B_\delta(\psi) \le C^\star \delta^2$ and $B_\delta(\tilde{\psi}) \ge c_\star \delta^2$, with equality in the lower bound (for some normalization of the block-Hessian) only if $\tilde{\psi}$ is a nonzero scalar multiple of ψ.*

Proof [Proof of Theorem 21] First-order orthogonality removes all linear terms. A second-order expansion in the three nuisance blocks gives a purely quadratic remainder $Q(h)$ plus $o(\|h\|^2)$. By Assumption 3.3, Q is a continuous bilinear form whose operator norm is finite; hence $|Q(h)| \le C^\star \|h\|^2$ for some $C^\star$ depending only on $\|\partial_{ab}^2 \Psi(\tau_0, \theta_0)\|_{\mathrm{op}}$, yielding the upper bound. Conversely, any other first-order orthogonal score $\tilde{\psi}$ induces a quadratic form $\tilde{Q}$; if $\tilde{Q}$ were strictly smaller than the tri-score's mixed blocks in every direction, then ψ would not be locally least-favorable. Compactness of the unit sphere in the product L_2-space implies a uniform lower bound $c_\star > 0$ for $\sup_{\|h\|=1} |\tilde{Q}(h)|$, which gives the stated lower inequality. Equality requires the mixed second derivatives to match (up to a common scalar), hence $\tilde{\psi} \equiv c\,\psi$ after normalization. Details are in A.3. $\blacksquare$

Let $\mathcal{M}_1$ denote the model in which route (I1) holds (valid IV given R), $\mathcal{M}_2$ the model in which route (I2) holds (proxy route: cue-sufficiency plus L_2-completeness), and $\mathcal{M}_3$ the model in which route (I3) holds (correct treatment–residual model). I analyze the union model $\mathcal{M}_\cup := \mathcal{M}_1 \cup \mathcal{M}_2 \cup \mathcal{M}_3$. These models form a *partially nested union*: $\mathcal{M}_3 \subset \mathcal{M}_1$, $\mathcal{M}_2 \not\subset \mathcal{M}_1$, and $\mathcal{M}_1 \not\subset \mathcal{M}_2$. Consequently, $\mathcal{M}_2 \cap \mathcal{M}_3 \subset \mathcal{M}_1$. Indeed, under (I3) I have $\mathbb{E}[\varepsilon \mid R] = 0$, hence $\mathbb{E}[\varepsilon \mid Z, R] = \mathbb{E}\big[\mathbb{E}[\varepsilon \mid R] \mid Z, R\big] = 0$, so (I3) implies (I1).

Following Wang and Tchetgen Tchetgen (2018) (see also Bickel et al. (1998)), an estimator $\hat{\tau}$ is *efficient in the union model* if it is regular and asymptotically linear in $\mathcal{M}_\cup$ with influence function equal to the efficient influence function (EIF) for $\mathcal{M}_j$ whenever the true distribution lies in $\mathcal{M}_j$, for $j = 1, 2, 3$. As shown below (Lemma 22) and proved in detail in §5.4, the tri-score moment generates an influence function that coincides with the EIF in each $\mathcal{M}_j$ and therefore attains the semiparametric efficiency bound throughout $\mathcal{M}_\cup$.

Lemma 22 (EIF compatibility across sub-models) *Let $\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3$ be the three semiparametric models corresponding to (I1)–(I3) (valid IV, proxy route, correct residual). Suppose a score φ is the efficient influence function in each $\mathcal{M}_j$ when the data-generating process lies in $\mathcal{M}_j$. Then, by Theorem 5 of Wang and Tchetgen Tchetgen (2018) (see also Bickel et al. (1998); Robins and Rotnitzky (1995)), φ attains the semiparametric efficiency bound in the union model $\mathcal{M}_\cup := \mathcal{M}_1 \cup \mathcal{M}_2 \cup \mathcal{M}_3$.*

Proof [Proof of Lemma 22] By assumption, whenever the data-generating distribution lies in $\mathcal{M}_j$, the proposed influence function φ coincides with the efficient influence function (EIF) for $\mathcal{M}_j$. Theorem 5 of Wang and Tchetgen Tchetgen (2018) then implies that any influence function that equals the EIF in each constituent model is efficient for the union model $\mathcal{M}_\cup := \mathcal{M}_1 \cup \mathcal{M}_2 \cup \mathcal{M}_3$. No inclusion relations between the tangent spaces are required for this conclusion (see also Bickel et al. (1998); Robins and Rotnitzky (1995)). ■

Corollary 23 (Semiparametric efficiency) *The influence function φ_i attains the semiparametric efficiency bound; see Theorem 44.*

3.3 $\sqrt{n}$-Normality

Under Assumptions 5.1–5.4 and 5.2, and provided any identification route in Theorem 5 holds, $\sqrt{n}(\hat{\tau} - \tau_0) \Rightarrow \mathcal{N}(0, V)$. The $\sqrt{n}$ expansion and the CLT proof are given by Lemma 41 and Theorem 42 in §5.3.

4 Rate Conditions for Deep-Net Nuisance Learners

This section establishes rate conditions for deep-net nuisance learners following Shen and Espinoza (2025); Kohler and Krzyżak (2025); Kohler and Langer (2021); Langer (2021). Throughout, $\|\cdot\|_{2,P}$ denotes the population L_2–norm and $\|\cdot\|_{2,n}$ the row-wise empirical norm ($\|g\|_{2,n}^2 := n^{-1}\sum_{i=1}^n g_i^2$). For a function class $\mathcal{H}$ write $\mathcal{N}(\varepsilon, \mathcal{H}, \|\cdot\|_\infty)$ for its covering number under the sup–norm, and let boldface ($\boldsymbol{\Theta}$) collect all network weights. Fix constants $C_\tau, C_\theta > 0$ and define

$$\mathcal{N}_n := \big\{(\tau, \theta) : \ |\tau - \tau_0| \le C_\tau n^{-1/2}, \ \|\theta - \theta_0\| \le C_\theta n^{-1/4}\big\}. \tag{16}$$

Lemma 24 (Population Hadamard differentiability) *Let $\Psi(\tau, \theta) = \mathbb{E}[\psi_i(\tau, \theta)]$ be the population moment map. Then Ψ is twice Hadamard differentiable at (τ_0, θ_0).*

Proof [Proof of Lemma 24.]See Appendix A.5. ■

This pointwise result supplies the second–order Taylor expansion of Ψ with $o(\|h\|^2)$ remainder that is used in the bias/remainder control and in the Hessian/minimax arguments.

Lemma 25 (Stochastic equicontinuity) *For the neighborhood $\mathcal{N}_n$ in* (16)*, $\sup_{(\tau,\theta)\in\mathcal{N}_n} \sqrt{n}\,\big|\hat{\Psi}_n(\tau, \theta) - \Psi(\tau, \theta)\big| = o_p(1)$.*

Proof [Proof of Lemma 25.]See Appendix A.6. ■

Combined, Lemmas 24–25 supply the $\sqrt{n}$-normality argument in §5.3. The orthogonality theory developed in §§ 3.1–3.3 requires the joint empirical rate

$$\max\Big\{\|\hat{\eta} - \eta_0\|_{2,n}, \ \|\hat{\gamma} - \gamma_0\|_{2,n}, \ \|\hat{\varphi} - \varphi_0\|_{2,n}\Big\} = o_p\big(n^{-1/4}\big). \tag{17}$$

Most results for deep nets are stated in the population $L_2(P)$ metric; the next subsection therefore shows that the two metrics are asymptotically interchangeable under the moment assumptions. Throughout, I enforce spectral normalization during training (Miyato et al. (2018)) and clip weights at inference. If layers are merely Lipschitz with a slowly growing constant $L_n = o(n^{1/8})$ the $o_p(n^{-1/4})$ rate remains valid (see Appendix A.6).

4.1 From population to empirical & bounded inputs

Lemma 26 (Shrinking functions: population $\Longrightarrow$ empirical) *Let $\{g_n\}_{n\geq 1}$ be a sequence of row-wise i.i.d. scalar functions such that $\mathbb{E}_n[g_n] = 0$, $\sup_{n\geq 1} \mathbb{E}_n[|g_n|^{2+\delta}] < \infty$ for some $\delta > 0$, and $\sup_{n\geq 1} \mathbb{E}_n[g_n^4] < \infty$ (the latter holds if $\delta \geq 2$). If $\|g_n\|_{2,P} = o_p(n^{-1/4})$, then $\|g_n\|_{2,n} = o_p(n^{-1/4})$.*

Proof [Proof of Lemma 26] Write $\mu_n := \mathbb{E}[g_n^2]$ and $\hat{\mu}_n := n^{-1}\sum_{i=1}^n g_{ni}^2 = \|g_n\|_{2,n}^2$. By Chebyshev applied to g_n^2, $\hat{\mu}_n - \mu_n = O_p\Big(\sqrt{\mathrm{Var}(g_n^2)/n}\Big) \leq O_p\Big(\sqrt{\mathbb{E}[g_n^4]/n}\Big) = O_p\big(n^{-1/2}\big)$, where the last step uses $\sup_n \mathbb{E}[g_n^4] < \infty$. Since $\|g_n\|_{2,P} = o_p(n^{-1/4})$, I have $\mu_n = o_p(n^{-1/2})$. Hence $\hat{\mu}_n = \mu_n + O_p(n^{-1/2}) = o_p(n^{-1/2})$, and therefore $\|g_n\|_{2,n} = \sqrt{\hat{\mu}_n} = o_p(n^{-1/4})$. ■

Assumption 4.1 (Unit-cube normalization) *Every raw covariate is linearly rescaled so that $X_i \in [0,1]^{d_x}$. After self-supervised encoding, $R_i = \tanh\big(R_\eta(X_i)\big) \in [-1,1]^{d_r}$. Bounded inputs are needed for the entropy bound below and ensure that a network with per-layer spectral norm ≤ 1 is globally 1-Lipschitz.*

4.2 Sieve, entropy, and smoothness

Let $\mathcal{F}(L,W)$ be the class of fully-connected networks of depth L and width W; the activation is $\sigma(u) = \max\{0,u\}$. For a radius $B > 0$ define the ℓ_1–ball sieve $\mathcal{F}_B(L,W) := \big\{f \in \mathcal{F}(L,W) : \|\boldsymbol{\Theta}(f)\|_1 \leq B\big\}$.

Lemma 27 (Entropy of ReLU sieves) *Assume $\|x\|_\infty \leq 1$ and that every weight matrix is spectrally normalized:*

$$\big\|\boldsymbol{W}_\ell\big\|_2 \;\leq\; 1, \ell = 1,\dots,L. \tag{18}$$

For $L \lesssim \log n$ and every $\varepsilon \in (0,1)$,

$$\log\mathcal{N}\big(\varepsilon, \mathcal{F}_B(L,W), \|\cdot\|_\infty\big) \;\leq\; C\,L\,W\,\log\big(B/\varepsilon\big), \tag{19}$$

with a universal constant C. The bound is scale-insensitive in the sense of Farrell, Liang, and Misra (2021): it depends only on the aggregate ℓ_1-radius B, not on $\|\boldsymbol{\Theta}\|_\infty$.

Proof [Proof of Lemma 27] Quantizing the LW parameters on an (ε/B)-grid (Anthony and Bartlett (1999, §14.3)) and invoking the spectrally-normalized covering Lemma A.5 of Bartlett et al. (2019) establishes (19). Because inputs lie in the unit cube and each layer satisfies $\|\boldsymbol{W}_\ell\|_2 \leq 1$, the network is 1-Lipschitz, so the grid quantization perturbs the output by at most the grid width, completing the proof. ■

Assumption 4.2 (Hölder regularity) *For some $s > d/2$, $R_{\eta_0} \in \mathcal{H}^s([0,1]^{d_x})$, $\quad \pi_{\gamma_0}$, $m_{\varphi_0} \in \mathcal{H}^s([0,1]^{d_r})$, with $\mathcal{H}^s$ the unit s–Hölder ball.*

Throughout the remainder of the paper, the SSL encoder $R_{\tilde{\eta}}$ is treated as fixed when the propensity and outcome networks are cross-fitted; the representation is therefore not updated inside the K folds. The rate calculations below reflect this separation.

Theorem 28 (Oracle inequality, supervised) *Let $\hat{g}$ minimize $\frac{1}{n}\sum_{i=1}^{n} \ell\big(g(X_i), Y_i\big)$ over the class $\mathcal{F}_{B_n}(L_n, W_n)$ with $L_n \asymp \log n$, $\quad W_n \asymp n^{d/(2s+d)}$, $\quad B_n \asymp W_n$.* (a) *For the square-loss and any $g_0 \in \mathcal{H}^s([0,1]^d)$, $\|\hat{g} - g_0\|_{2,P} = O_p\big(n^{-s/(2s+d)}\big)$.* (b) *For the logistic loss, the same rate holds under a Tsybakov margin exponent $\alpha > 0$.*

Proof [Proof of Theorem 28]

Decompose the excess risk into $\|\hat{g} - g_0\|_{2,P} \;\le\; \underbrace{\inf_{f \in \mathcal{F}_{B_n}} \|f - g_0\|_\infty}_{\text{(A) approximation}} +$ $\underbrace{\|\hat{g} - f^*\|_{2,P}}_{\text{(B) estimation}}$,. Here f^* is the sup-norm projection of g_0 onto $\mathcal{F}_{B_n}$. [(A)] Yarotsky (2017) yields $\inf_{f \in \mathcal{F}_{B_n}} \|f - g_0\|_\infty = O\big(W_n^{-s/d}\big)$. [(B)] Lemma 27 and a localized Rademacher-complexity argument (Farrell, Liang, and Misra (2021)) give $\|\hat{g} - f^*\|_{2,P} = O_p\Big(\sqrt{\frac{L_n W_n \log(B_n n)}{n}}\Big)$. Balancing $W_n^{-s/d}$ against $\sqrt{L_n W_n \log n/n}$ with $L_n \asymp \log n$ and $W_n \asymp n^{d/(2s+d)}$ yields the claimed rate. See Appendix A.7 for all details. Because $s > d/2$ implies $a := s/(2s+d) > 1/4$, the $O_p(n^{-a})$ bound is automatically $o_p(n^{-1/4})$, so the nuisance block meets the requirement (17). ■

4.3 Self-supervised encoder & master rate

Throughout Sections 4–5 the instrument regression $\pi_\gamma(R) := \mathbb{E}[Z \mid R]$ is approximated by a fully connected ReLU network. The two supervised regressions $\mu_Y(R) := \mathbb{E}[Y \mid R]$ and $\mu_D(R) := \mathbb{E}[D \mid R]$ are fit on the labeled folds. All three attain the same $o_p(n^{-1/4})$ rate under the sieve choices below.

(a) Mercer kernel and effective dimension. Let $k((z,r),(z',r'))$ be a bounded, *universal* Mercer kernel (see Sriperumbudur *et al.* (2011, Theorem. 7)) on the compact support of (Z, R). Write $\mathcal{H}_k$ for its RKHS and $N(\lambda) := \sum_j \frac{\lambda_j}{\lambda_j + \lambda}$ for the effective dimension. Under the eigen–decay and source assumptions of Meunier et al. (2024); Fischer and Steinwart (2020), $N(\lambda) \asymp \lambda^{-1/\beta}$ for some $\beta > 0$.

(b) KRR or spectral cut-off. For a labeled fold $\mathcal{I}$ of size $m = n/K$, define $\hat{\pi}_\gamma^{\mathrm{KRR}} := \arg\min_{f \in \mathcal{H}_k} \frac{1}{m}\sum_{i \in \mathcal{I}} (D_i - f(Z_i, R_i))^2 + \lambda_m \|f\|_{\mathcal{H}_k}^2$, where $\lambda_m \asymp m^{-\beta/(2\beta+1)}$. An equivalent *spectral cut-off* estimator $f = \sum_{j \le M_m} \langle \varphi_j, D \rangle \varphi_j$ with $M_m \asymp m^{1/(2\beta+1)}$ principal components[8] yields

$$\|\hat{\pi}_\gamma - \pi_{\gamma_0}\|_{2,P} = O_p\big(m^{-\beta/(2\beta+1)}\big), \tag{20}$$

[8] Spectral-truncation idea already appears in Hall and Horowitz (2005, §2.2), but without rate analysis.

as shown, for example, by Li *et al.* (2024, Thm. 3). Plugging $m = \lfloor n/K \rfloor$ and any $\beta > 1/2$ makes the rate $o_p(n^{-1/4})$, meeting the threshold in Theorem 34.[9]

(c) Compatibility with the tri-score. KRR is linear in (D, Z, R) and has a closed-form influence function, so all proofs in §3–§5 remain unchanged—only envelope constants differ.

Remark 29 (Minimax optimality) *Under assumptions (EVD), (SRC), and (EMB) of Meunier et al. (2024), together with the lower bound in Li et al. (2024), rate* (20) *is minimax optimal for estimating π_γ over the Sobolev ball $\mathcal{W}_2^\beta$.* □

Label scarcity ($n_\ell \ll n$) motivates learning $R_\eta : X \mapsto [-1,1]^{d_r}$ from a much larger unlabeled pool of size $n_u \gg n_\ell$. Denote by $\tilde{\eta}$ the parameter that minimizes the InfoNCE loss $\mathcal{L}_{\text{NCE}}(\eta) := -\mathbb{E}\left[\log \frac{\exp\{\langle R_\eta(X), R_\eta(X^+)\rangle/T\}}{\exp\{\langle R_\eta(X), R_\eta(X^+)\rangle/T\} + \sum_{j=1}^{K-1} \exp\{\langle R_\eta(X), R_\eta(X_j^-)\rangle/T\}}\right]$, where $(X^+, X_1^-, \ldots, X_{K-1}^-)$ are independent draws and $T > 0$ is the temperature (van den Oord, Li, and Vinyals (2018)). Define $n_u = \lceil n^{1+\delta} \rceil$ for some $\delta > 0$, and let $B \in \mathbb{N}$ denote the mini–batch size used in self-supervised updates. Use the spectrally-normalized sieve $\mathcal{F}_{B_u}(L_u, W_u)$ with $L_u \asymp \log n_u$, $W_u \asymp n_u^{d_x/(2s+d_x)}$, and $B_u \asymp W_u$, enforcing $\|\boldsymbol{W}_\ell\|_2 \le 1$ for every layer (Assumption 4.1); this keeps the network 1-Lipschitz on the unit cube, allowing use of Lemma 27.

Assumption 4.1 requires bounded inputs only for the rates analysis. The tanh clipping in Assumption 4.1 is applied only at inference time. The SSL encoder is trained in the usual unbounded space; after training, its final activations are clipped once before they feed any nuisance networks. Gradients in the InfoNCE objective are therefore unaffected.

Assumption 4.3 (Spectral gap) *The oracle embedding has non-degenerate covariance:* $\lambda_{\min}\big(\text{Var}$
$(R_{\eta_0}(X))\big) \ge \lambda_{\min} > 0$.

Theorem 30 (Rate of the SSL encoder) *Adopt the sieve above and train $\tilde{\eta}$ by InfoNCE. Under Assumptions 4.1, 4.2 and 4.3, $\big\|R_{\tilde{\eta}} - R_{\eta_0}\big\|_{2,P} = O_p\big(n_u^{-\,s/(2s+d_x)}\big)$, where the generalization step uses the PAC-Bayes InfoNCE bound for fully–connected, spectrally-normalized networks (HaoChen et al. (2021)).*

Let $\delta > 0$ satisfy $n_u = n^{1+\delta}$ and write $a_x := s/(2s+d_x)$. Then $a_x > 1/4$ precisely when $\delta > (d_x - 2s)/(2s+d_x)$. In that case $O_p\big(n_u^{-a_x}\big) = O_p\big(n^{-a_x(1+\delta)}\big) = o_p\big(n^{-1/4}\big)$, so the encoder block satisfies the global condition (17). *(If $d_x > 2s$, achieving $a_x > 1/4$ may require super-polynomial n_u, which is feasible in practice; e.g. industrial logs with $(n, n_u) \approx (10^8, 10^{11})$ already meet this requirement.)*

Remark 31 (No new complexity bounds needed) *No new margin-based Rademacher, local- Rademacher, or PAC-Bayes analyses are needed. Existing PAC-Bayes bounds for spectrally- normalized nets (e.g. HaoChen et al. (2021); Kuzborskij et al. (2024)) already deliver the $o(n^{-1/4})$ high- probability L_2 rate required here.*

[9] If $\beta \le 1/2$ one can enlarge the unlabeled pool or combine kernel features with the deep-net sieve; see Meunier et al. (2024, §4.3).

Remark 32 (InfoNCE as baseline) *InfoNCE serves only as a diagnostic baseline; in all main experiments I instead* ***optimize*** *the* spectral-contrastive loss *(SCL) (e.g. HaoChen et al. (2021)), introduced later in this subsection. The PAC-Bayes rate in Theorem 30 continues to hold verbatim for SCL, so all theoretical guarantees remain unchanged.*

Proof [Proof of Theorem 30] A standard PAC–Bayes bound for spectrally-normalized nets gives $\mathcal{R}(\tilde{\eta}) = O_p((L_u W_u \log n_u / n_u)^{1/2})$. Approximation of the oracle encoder by a ReLU network yields the same order for the approximation error (e.g. Yarotsky (2017)). Finally a Poincaré-type inequality (Assumption 4.3) transfers excess risk to an L_2-error, giving the stated rate. I pre–train R_η by *spectral–contrastive loss* (HaoChen et al. (2021)) $\mathcal{L}_{\mathrm{SCL}}(\eta) := \left\| \boldsymbol{I} - \frac{1}{B} \boldsymbol{Z}\boldsymbol{Z}^\top \right\|_F^2, \quad \boldsymbol{Z} := [R_\eta(X_1), \dots, R_\eta(X_B)]$. The loss admits the same PAC–Bayes rate as InfoNCE but avoids negative sampling. See Appendix A.8 for full details.

Corollary 33 (Encoder meets the $n^{-1/4}$ threshold) *Let $n_u = n^{1+\delta}$, and suppose either $d_x < 2s$ with $\delta > 0$ arbitrary, or $d_x \geq 2s$ with $\delta > (d_x - 2s)/(2s + d_x)$. Theorem 30 implies $\|R_{\tilde{\eta}} - R_{\eta 0}\|_{2,P} = o_p\big(n^{-1/4}\big)$; the representation block satisfies the rate requirement in* (17)*.*

Theorem 34 (All nuisance blocks beat $n^{-1/4}$) *Assume the encoder is trained on $n_u = n^{1+\delta}$ un-labeled events with $\begin{cases} \delta > 0, & d_x < 2s, \\ \delta > \dfrac{d_x - 2s}{2s + d_x}, & d_x \geq 2s, \end{cases}$ so that Theorem 30 yields $\|R_{\tilde{\eta}} - R_{\eta 0}\|_{2,P} = o_p(n^{-1/4})$. With the propensity and outcome nets trained as in Theorem 28, $\max\Big\{ \|R_{\tilde{\eta}} - R_{\eta 0}\|_{2,P},\ \|\pi_{\hat{\gamma}} - \pi_{\gamma 0}\|_{2,P},\ \|m_{\hat{\varphi}} - m_{\varphi 0}\|_{2,P} \Big\} = o_p\big(n^{-1/4}\big)$; the global requirement* (17) *is satisfied.*

Proof [Proof of Theorem 34] Let $\Delta_{\eta,n} := \|R_{\tilde{\eta}} - R_{\eta 0}\|_{2,P}$, $\Delta_{\gamma,n} := \|\pi_{\hat{\gamma}} - \pi_{\gamma 0}\|_{2,P}$, and $\Delta_{\varphi,n} := \|m_{\hat{\varphi}} - m_{\varphi 0}\|_{2,P}$. By Theorem 30 and the sample–size condition on n_u, $n^{1/4}\,\Delta_{\eta,n} \xrightarrow{p} 0$. By Theorem 28, applied separately to the propensity and outcome nets, $n^{1/4}\,\Delta_{\gamma,n} \xrightarrow{p} 0$ and $n^{1/4}\,\Delta_{\varphi,n} \xrightarrow{p} 0$. For any $\varepsilon > 0$ and all sufficiently large n, $\Pr\big(n^{1/4} \max\{\Delta_{\eta,n}, \Delta_{\gamma,n}, \Delta_{\varphi,n}\} > \varepsilon\big) \leq \Pr\big(n^{1/4}\Delta_{\eta,n} > \varepsilon\big) + \Pr\big(n^{1/4}\Delta_{\gamma,n} > \varepsilon\big) + \Pr\big(n^{1/4}\Delta_{\varphi,n} > \varepsilon\big) \xrightarrow{n\to\infty} 0$. The inequality is a union bound; the limit holds because each summand tends to zero as shown above. Hence $n^{1/4} \max\{\Delta_{\eta,n}, \Delta_{\gamma,n}, \Delta_{\varphi,n}\} \xrightarrow{p} 0$, which is equivalent to $\max\{\Delta_{\eta,n}, \Delta_{\gamma,n}, \Delta_{\varphi,n}\} = o_p\big(n^{-1/4}\big)$. Therefore the joint requirement (17) is satisfied. ■

Early stopping after roughly $5 \log n$ passes and a weight-decay grid $\lambda \in \{10^{-5}, 10^{-4}, 10^{-3}\}$ reproduce the theoretical rates for sample sizes up to $n \approx 10^6$. Unbounded individual weights are admissible because the variance–adaptive localization argument keeps the covering numbers and Bernstein constants under control.

5 Estimation and Inference: Large-Sample Theory

Building on Farrell, Liang, and Misra (2021); Chernozhukov, Chetverikov, and Kato (2018); Chernozhukov, Fernández-Val, and Luo (2018); Giné and Nickl (2008); Dudley (1999); van der Vaart (1998); Newey (1990), this section derives the large-sample properties of the TRIV–Rep estimator, $\hat{\tau}$. I first establish consistency, then verify the $o(n^{-1/4})$ rates for the nuisance learners, and finally prove $\sqrt{n}$-normality and semiparametric efficiency.

Assumption 5.1 (Identification union) *At least one of (I1)–(I3) in Theorem 5 holds so that* $\Psi(\theta) = 0 \Longrightarrow \tau = \tau_0$.

Assumption 5.2 (Cross-fitting rate) *With cross–fitting,* $\|\hat{\eta} - \eta_0\|_{2,n} \vee \|\hat{\gamma} - \gamma_0\|_{2,n} \vee \|\hat{\varphi} - \varphi_0\|_{2,n} = o_p(n^{-1/4})$.

Assumption 5.3 (Triangular–array Lindeberg conditions) *Throughout this assumption let* $\psi_{i,n}$ *denote the row–wise score defined in* (12). *For each* n, $\{(Y_{i,n}, D_{i,n}, Z_{i,n}, R_{i,n})\}_{i=1}^{n}$ *are i.i.d. draws from* P_n *(*Row–wise independence*). There exists* $\delta > 0$ *and measurable* $\Psi^{\mathrm{env}} \colon \mathcal{Z} \to \mathbb{R}$ *with* $|\psi_{i,n}(\theta)| \leq \Psi^{\mathrm{env}}(Z_{i,n})$ *for all* n, θ, *and* $\sup_{n \geq 1} \mathbb{E}_n[(\Psi_{i,n}^{\mathrm{env}})^{2+\delta}] < \infty$ *(*Uniform $(2+\delta)$ moment bound*). For every* $\varepsilon > 0$, $\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_n[\psi_{i,n}(\theta_0)^2 \, \mathbb{1}\{|\psi_{i,n}(\theta_0)| > \varepsilon\sqrt{n}\}] \to 0$ *(*Lindeberg–Feller condition*).*

Assumption 5.4 (Uniform linearization and stable Jacobian) *There is a neighborhood* $\mathcal{N} \subset \Theta$ *of the true parameter* θ_0 *such that: (i)* Uniform first-order expansion: *For every bounded direction* h *and every* t *small enough with* $\theta_0 + t\,h \in \mathcal{N}$, $\sup_{n\geq 1}\big\|\Psi_n(\theta_0 + t\,h) - \Psi_n(\theta_0) - t\,\partial_\theta \Psi_n(\theta_0)[h]\big\| = o(t) \quad (t \to 0)$. *(ii)* Stable, nondegenerate score derivative in τ: *Let* $S_n := \partial_\tau \Psi_n(\tau_0, \theta_0)$ *and* $S := \partial_\tau \Psi(\tau_0, \theta_0)$. *Then* $S_n \to S$ *and* $|S| \geq s_0 > 0$. (If one augments the parameter with a finite-dimensional vector λ and linearizes in (τ, λ), assume the corresponding finite-dimensional Jacobian G is nonsingular.)

5.1 Uniform Hadamard differentiability of sample moment map Ψ_n

The next lemma verifies Assumption 5.4, upgrading pointwise Gateaux differentiability to uniform Hadamard differentiability.

Lemma 35 (Uniform Hadamard differentiability of the sample moment) *Let* $\Psi_n(\theta) := \mathbb{E}_n[\psi_{i,n}(\theta)]$ *be the (triangular–array) sample moment map and* $G_n := \partial_\theta \Psi_n(\theta_0)$. *Under Assumptions 5.3 and the spectral–norm constraint,* $\lim_{t\to 0}\sup_{n\geq 1}\Big\|\frac{\Psi_n(\theta_0+th)-\Psi_n(\theta_0)}{t} - \partial_\theta \Psi_n(\theta_0)[h]\Big\| = 0$ *and* $\|G_n - G\| \to 0$, *where* $G := \partial_\theta \Psi(\theta_0)$. *This uniform (in n) result verifies Assumption 5.4 and yields* $\partial_\theta \hat{\Psi}_n(\bar{\theta}) = G + o_p(1)$ *in Lemma 41.*

Proof [Proof of Lemma 35] See Appendix A.9. ■

Boundedness of h holds automatically because directions are taken inside the same ℓ_1-ball sieve as the estimator; see Lemma 27.

Lemma 36 (Bernstein–chaining tail bound) *Let* $\mathcal{G}_n$ *be a class of functions* $g \colon \mathcal{Z} \to \mathbb{R}$ *with* $\mathbb{E}_n[g_i] = 0$ *and* $\|g_i\|_{\psi_2} \leq C_0$ *uniformly in* $g \in \mathcal{G}_n$ *and* $i \leq n$. *Denote by* $\mathsf{Pdim}(\mathcal{G}_n)$ *their pseudo–dimension. Then for all* $t > 0$, $\Pr\Big(\sup_{g\in\mathcal{G}_n}\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g_i\Big| > t\Big) \leq 2\exp\Big\{-c\,t^2 + C\,\mathsf{Pdim}(\mathcal{G}_n)\Big\}$, *where* $c > 0$ *is universal and* C *depends only on* C_0.

Lemma 36 relies on a Bernstein–type chaining bound, not on classical Donsker theory. Variance–adaptive localization (Remark 37) controls the empirical process via $\Pr\Big(\sup_{g\in\mathcal{G}_n} |n^{-1/2}\sum g_i| > t\Big) \;\leq\; 2\exp\Big\{-ct^2 + C\,\mathsf{Pdim}(\mathcal{G}_n)\Big\}$, where $\mathsf{Pdim}(\mathcal{G}_n) \lesssim LW\log B$ is uniform in n by Lemma 27. Hence the required stochastic-equicontinuity holds without imposing a fixed, n-independent Donsker envelope or a uniform bound on individual weights.

Proof [Proof of Lemma 36] Each $g \in \mathcal{G}_n$ is sub-Gaussian. Bernstein's inequality shows $\Pr(|n^{-1/2}$
$\sum_{i=1}^n g_i| > t) \leq 2\exp(-c\,t^2)$. Cover $\mathcal{G}_n$ by an ε-net of size at most $\exp\{C\,(\mathcal{G}_n)\log(1/\varepsilon)\}$ using the Sauer–Shelah/pseudo-dimension bound, take a union bound over the net, and optimise over ε in a standard chaining argument. This yields $\Pr(\sup_{g\in\mathcal{G}_n}|n^{-1/2}\sum g_i| > t) \leq 2\,\exp\{-c\,t^2 + C\,(\mathcal{G}_n)\}$. ■

Remark 37 (Variance–adaptive localization) *The stochastic–equicontinuity argument below follows the variance–adaptive Bernstein localization of Farrell, Liang, and Misra (2021).*[10] *Concretely, Dudley's chaining integral is evaluated over shells whose data–dependent radii* α_k *satisfy* $\alpha_k^2 \asymp \widehat{\mathrm{Var}}\big(\psi_{i,n}(\theta_k)\big)$ *(Dudley (1999)). Because the resulting tail bound depends only on the pseudo–dimension* $\mathrm{Pdim}(\mathcal{G}_n) \lesssim LW\log B$ *(Lemma 27) rather than on* $\|\boldsymbol{\Theta}\|_\infty$, *it remains valid even when individual network weights are unbounded, provided the global* ℓ_1*–radius and unit-cube input scaling in Lemma 27 hold. This is the key step that allows the* $n^{-1/4}$ *rate without spectral normalization.*

5.2 Consistency and triple-robustness

Theorem 38 (Consistency of the Z–estimator for τ under triangular–array drift) *Suppose Assumptions 5.1–5.4 and the triangular–array moment conditions in Assumption 5.3 hold. Let* $\hat{\tau}$ *be the* Z*–estimator that solves* (14)*. Then* $\hat{\tau} \xrightarrow{p} \tau_0$*, and this convergence obtains provided any one of the identification routes* (I1)*–*(I3) *in Theorem 5 holds.*

Proof [Proof of Theorem 38] Fix the shrinking neighborhood $\mathcal{N}_n := \{\theta : \|\theta-\theta_0\| \leq Cn^{-1/4}\}$. Let $\mathcal{G}_n := \{\psi_{i,n}(\theta) - \psi_{i,n}(\theta_0) : \theta \in \mathcal{N}_n\}$. Lemma 27 gives the uniform entropy bound $\log\mathcal{N}\big(\varepsilon, \mathcal{G}_n, \|\cdot\|_\infty\big) \;\leq\; C_1 LW\log(B/\varepsilon)$, so the (local) Rademacher radius satisfies $\mathfrak{R}_n(\mathcal{G}_n) = O\big(\sqrt{LW\log B\,/n}\big) \;=\; O\big(n^{-1/2}\big)$. Applying the variance–adaptive Bernstein–chaining inequality of Bartlett, Bousquet and Mendelson (2005)—see the proof of Lemma 25—yields, for some constants $c_0, C_0 > 0$, $\Pr\Big(\sup_{\theta\in\mathcal{N}_n}\big\|\hat{\Psi}_n(\theta) - \Psi_n(\theta)\big\| > t\Big) \;\leq\; 2\exp\big\{-c_0t^2 + C_0\log n\big\}$. Choosing $t = C\sqrt{\log n}$ gives

$$\sup_{\theta\in\mathcal{N}_n}\big\|\hat{\Psi}_n(\theta) - \Psi_n(\theta)\big\| = O_p\big(n^{-1/2}\big). \tag{21}$$

The triangular–array nature is immaterial here; the proof of Lemma 25 requires only row-wise independence plus the moment conditions of Assumption 5.3. Row-wise independence and $\sup_i E\|\psi_{i,n}(\theta_0)\|^{2+\delta} < \infty$ verify Lindeberg's condition. Hence, by the triangular–array

[10] See also the local Rademacher complexities program ofBartlett, Bousquet and Mendelson (2005).

CLT in van der Vaart (1998),

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{i,n}(\theta_0) \implies \mathcal{N}(0,\Sigma), \hat{\Psi}_n(\theta_0) = O_p\big(n^{-1/2}\big). \tag{22}$$

Hadamard differentiability (Lemma 24) implies the second–order expansion $\Psi_n(\theta) = \Psi_n(\theta_0) + \partial_\theta\Psi_n(\theta_0)\,[\theta-\theta_0] + R_n(\theta)$ with $\|R_n(\theta)\| \le C_2\|\theta-\theta_0\|^2$. Because $\|\theta-\theta_0\| \le Cn^{-1/4}$ on $\mathcal{N}_n$,

$$\sup_{\theta\in\mathcal{N}_n}\big\|\Psi_n(\theta)-\Psi_n(\theta_0)\big\| = O\big(n^{-1/2}\big). \tag{23}$$

Combining (21), (22) and (23), $\sup_{\theta\in\mathcal{N}_n}\big\|\hat{\Psi}_n(\theta)\big\| = o_p(1)$. ■

Remark 39 (Rate of the bias term) *Lemma 24 yields $\|R_n(\theta)\| \le C\|\theta-\theta_0\|^2$; on $\mathcal{N}_n$ this is $O(n^{-1/2})$, strictly smaller than the stochastic order $n^{-1/2}$ of the empirical process, so the bias is negligible.*

Remark 40 (Triangular–array CLT) *Row-wise independence together with $\sup_i E\|\psi_{i,n}(\theta_0)\|^{2+\delta} < \infty$ suffices for the Lindeberg condition (van der Vaart (1998)).*

5.3 Asymptotic linearity and normality

Write $\Psi(\tau,\theta) := \mathbb{E}[\psi(\tau,\theta)], \hat{\Psi}_n(\tau) := \frac{1}{n}\sum_{i=1}^{n}\psi_{i,n}(\tau,\hat{\theta})$, and let $\hat{\tau}$ denote the Z–estimator solving $\hat{\Psi}_n(\hat{\tau}) = 0$. Define the scalar score derivative $S := \partial_\tau\Psi(\tau_0,\theta_0)$, which equals $S = -C$ with $C := \mathbb{E}[(Z-\pi_0(R))\{D-\mu_{D,0}(R)\}]$ as introduced after (4) in §2.2.[11] Assumption 5.4(ii) (stable, nondegenerate score derivative) ensures $S_n := \partial_\tau\Psi_n(\tau_0,\theta_0) \to S$ and $|S| \ge s_0 > 0$.

Lemma 41 (Asymptotic linear (root–n) expansion) *Under Assumptions 5.1–5.4, 5.3, and 5.2, with $S = \partial_\tau\Psi(\tau_0,\theta_0)$ and $|S| > 0$,*

$$\sqrt{n}\,(\hat{\tau}-\tau_0) \;=\; -\,S^{-1}\,\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{i,n}(\tau_0,\theta_0) \;+\; o_p(1). \tag{24}$$

Proof [Proof of Lemma 41] A mean–value expansion of the sample moment around τ_0 gives $0 = \hat{\Psi}_n(\hat{\tau}) = \hat{\Psi}_n(\tau_0) + \partial_\tau\hat{\Psi}_n(\bar{\tau})\,(\hat{\tau}-\tau_0), \bar{\tau} := \tau_0 + s(\hat{\tau}-\tau_0),\ s\in(0,1)$. By Lemma 25 and Assumption 5.2, $\hat{\Psi}_n(\tau_0) = n^{-1}\sum_{i=1}^{n}\psi_{i,n}(\tau_0,\theta_0) + o_p(n^{-1/2}) = O_p(n^{-1/2})$. By Lemma 35 and Assumption 5.4(ii), $\partial_\tau\hat{\Psi}_n(\bar{\tau}) = S + o_p(1)$ with $|S| > 0$. Rearranging and multiplying by $\sqrt{n}$ yields (24).

Theorem 42 (Asymptotic normality of $\hat{\tau}$) *Suppose Assumptions 5.1–5.4, 5.3, and 5.2 hold. Let $\Sigma := \mathrm{Var}\big(\psi_{i,n}(\tau_0,\theta_0)\big)$ and $S := \partial_\tau\Psi(\tau_0,\theta_0)$ with $|S| > 0$. Then $\sqrt{n}\,(\hat{\tau}-\tau_0) \overset{d}{\Rightarrow} \mathcal{N}\big(0,\ V\big), \qquad V = \Sigma/S^2$.*

[11] Since $\mathbb{E}[(Z-\pi_0(R))\,\mu_{D,0}(R)] = 0$, I also have $C = \mathbb{E}[(Z-\pi_0(R))D]$.

Proof [Proof of Theorem 42] By Lemma 41, $\sqrt{n}\,(\hat{\tau}-\tau_0) = -S^{-1}\,n^{-1/2}\sum_{i=1}^{n}\psi_{i,n}(\tau_0,\theta_0) + o_p(1)$. Row-wise independence together with $\sup_i \mathbb{E}\|\psi_{i,n}(\tau_0,\theta_0)\|^{2+\delta} < \infty$ verifies Lindeberg's condition for a triangular array, so $n^{-1/2}\sum_{i=1}^{n}\psi_{i,n}(\tau_0,\theta_0) \Rightarrow \mathcal{N}(0,\Sigma)$ (e.g. van der Vaart (1998), Prop. 2.27). Slutsky's lemma yields the claim with $V = \Sigma/S^2$. ■

Remark 43 (Equivalence with the S,Σ formula) *With the notation $S = -C$ from §2.2, the variance representation $V = \Sigma/S^2$ is identical to the (S,Σ) formula used in the paper.*

5.4 Semiparametric efficiency of $\hat{\tau}$

In (I1) the nuisance is the conditional law of (Y,D,Z,R) subject to $\mathbb{E}[\varepsilon \mid Z,R] = 0$; scores have the form $s_1(W) = h(Z,R)\{\varepsilon - \mathbb{E}[\varepsilon \mid Z,R]\}$, so the tangent space is $\{h(Z,R) - \mathbb{E}[h \mid Z,R]\}$ multiplied by the residual. In (I2) the nuisance is the joint law of (Z,R,ε) subject to $Z \perp U \mid R$, $\varepsilon \perp (Z,R) \mid U$, and L_2– completeness; pathwise scores are mean-zero functions of (Z,R) and of (ε,R) orthogonalized by conditioning on R. Score $\psi(\tau_0,\theta_0)$ is orthogonal to both spaces by the same conditioning arguments used in Theorem 5; hence it coincides with the efficient score in (I1) and (I2). By Theorem 5 of Wang and Tchetgen Tchetgen (2018), the EIF therefore carries over to the union model. In the scalar target case, the efficient influence function equals $\varphi_i = \psi_i(\tau_0,\theta_0)/S$, so the efficiency bound is $V = \mathrm{Var}(\psi_i)/S^2$, which coincides with Theorem 42.

Write $G := \partial_\theta\Psi(\theta_0)$ for the population Jacobian of the moment map and let G_τ be its column that corresponds to the scalar target τ. The expansion in Lemma 41 shows that φ_i is indeed the EIF of the estimator $\hat{\tau}$. Throughout this subsection P_0 denotes the true distribution of the observation vector $W_i = (D_i,Z_i,R_i,Y_i)$, and $\mathbb{E}_0[\cdot]$ is the corresponding expectation operator. I consider the union $\mathcal{M} := \mathcal{M}_{\mathrm{IV}} \cup \mathcal{M}_{\mathrm{Proxy}} \cup \mathcal{M}_{\mathrm{Treat}}$, where the nuisance objects satisfy, respectively:

(I) $\mathbb{E}[\varepsilon \mid Z,R] = 0$ and $\mathrm{Var}(\mathbb{E}[D \mid Z,R]) > 0$ (valid IV given R),
(II) $Z \perp U \mid R$, $\varepsilon \perp (Z,R) \mid U$, R is L_2–complete (proxy route),
(III) $\mathbb{E}[\varepsilon \mid R] = 0$ (correct treatment–residual model; hence (I) holds),
Here $\varepsilon := Y - \tau_0 D$, $\pi_{\gamma_0}(R) := \mathbb{E}[Z \mid R]$, $m_{\varphi_0}(R) := \mathbb{E}[Y - \tau_0 D \mid R]$.

For each sub–model $\mathcal{M}_m$ ($m \in \{\mathrm{IV},\mathrm{Proxy},\mathrm{Treat}\}$) let $\mathcal{T}_m \subset L_0^2(P_0)$ be its tangent space, i.e. the mean–zero scores generated by all regular one–dimensional sub-paths. The next result shows that φ_i is simultaneously efficient for all three sub-models; hence it is efficient for the union $\mathcal{M}$.

Theorem 44 (Semiparametric efficiency) *Let $V := \mathrm{Var}_0(\varphi_i)$. Then V coincides with the semiparametric efficiency bound for τ in each $\mathcal{M}_m$ ($m = \mathrm{IV},\mathrm{Proxy},\mathrm{Treat}$) and therefore in their union $\mathcal{M}$.*

Proof [Proof of Theorem 44] See Appendix A.10 ■

5.5 Variance estimation and inference

The remainder of the paper fixes, for concreteness, *one* flexible sieve: fully-connected ReLU networks with growing width W_n and depth L_n. This choice is *not* an identifying assumption. Any learner that achieves $\| \hat{\eta} - \eta_0 \|_{2,n} \vee \| \hat{\gamma} - \gamma_0 \|_{2,n} \vee \| \hat{\varphi} - \varphi_0 \|_{2,n} = o_p(n^{-1/4})$, and is cross-fit over the K folds introduced in Assumption 5.2, inherits all results in Sections 3–5. I adopt ReLU networks trained with the InfoNCE contrastive loss (van den Oord, Li, and Vinyals (2018); Yarotsky (2017)) because (a) their piecewise linearity makes the tri-score Hessian tractable (Arora et al. (2018)), and (b) recent PAC–Bayes bounds (Bartlett, Foster, and Telgarsky (2017); Neyshabur et al. (2018); Kuzborskij et al. (2024)) deliver the required $n^{-1/4}$ rate. Define the sample derivative (available in closed form for the tri–score) $\hat{S} := \partial_\tau \hat{\Psi}_n(\hat{\tau}, \hat{\theta}) = -\frac{1}{n} \sum_{i=1}^n \frac{S_i}{\hat{\rho}} \left[Z_i - \hat{\pi}(R_i)\right] \left[D_i - \hat{\mu}_D(R_i)\right]$. Set the influence–function estimate and plug–in variance to $\hat{\varphi}_i := \hat{S}^{-1} \psi_{i,n}\big(\hat{\tau}, \hat{\eta}_i, \hat{\gamma}_i, \hat{\varphi}_i\big), \hat{V} := \frac{1}{n} \sum_{i=1}^n \hat{\varphi}_i^2$.

Theorem 45 (Consistency of the plug–in variance) *Under Assumptions* 3.1–5.3*,* $\hat{V} \xrightarrow{p} V$*.*

Proof [Proof of Theorem 45] See Appendix A.11. ■

Let $e_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ be independent of the data and $T^{\#} := \frac{1}{\sqrt{n}} \sum_{i=1}^n e_i \, \hat{\varphi}_i / \sqrt{\hat{V}}$.

Theorem 46 (Bootstrap validity) *Conditional on the sample and uniformly over the three identification scenarios,* $\sup_{t \in \mathbb{R}} \big| \Pr^{\#}(T^{\#} \leq t) - \Pr\big(\sqrt{n}(\hat{\tau} - \tau_0) \leq t\big) \big| \xrightarrow{p} 0$*. Hence the percentile bootstrap CI* $[\hat{\tau} \pm c^{\#}_{1-\alpha} \, \hat{V}^{1/2} / \sqrt{n}]$ *is asymptotically exact.*

Proof [Proof of Theorem 46] See Appendix A.12. ■

Combining Theorems 5, 38, 42, 45, and 46 (under the standing moment/rate conditions with cross–fitting), TRIV–Rep is (i) consistent and $\sqrt{n}$–normal whenever any one of the routes (I1)–(I3) holds; (ii) equipped with a consistent closed-form variance estimator (45); (iii) accompanied by a uniformly valid multiplier bootstrap (46); and (iv) semiparametrically efficient for the union model (44).

6 Monte Carlo Evidence

I study the finite–sample performance of TRIV–Rep. Each design isolates one identification route in Theorem 5 (I1–I3) as the operative source of identification. Across designs, the entire pipeline—representation learner, cross–fitting/sample–splitting, orthogonal score, and root solving—is held fixed; competitors are implemented analogously and succeed only when their own identifying restrictions are met.

Data generating process (DGP) and learning setup

For each Monte–Carlo cell I draw (A_i, Z_i, D_i, Y_i) according to one of three identification routes (I1–I3). Throughout, $A_i \sim \mathcal{N}(0,1)$ and there is no direct effect of Z on Y (exclusion,

$\delta_Z = 0$). The treatment equation is $D_i = b_Z Z_i + b_A A_i + \nu_i, (b_Z, b_A) = (0.4, 0.7),\ \nu_i \sim \mathcal{N}(0,1)$. The outcome equation and the instrument differ by route:

$$\begin{aligned}
\text{I1 (valid IV):}\quad & Z_i \sim \text{Bernoulli}(0.5),\\
& Y_i \;=\; \tau_0 D_i + g_A A_i + \varepsilon_i, g_A = 0.5,\ \varepsilon_i \sim \mathcal{N}(0,1).\\
\text{I2 (proxy route):}\quad & Z_i \sim \text{Bernoulli}\big(\text{logit}^{-1}(\gamma A_i)\big), \gamma = 1,\\
& Y_i \;=\; \tau_0 D_i + g_A A_i + \varepsilon_i + \varepsilon_i^{\text{conf}}, \varepsilon_i^{\text{conf}} = 0.6\, A_i,\ \varepsilon_i \sim \mathcal{N}(0,1).\\
\text{I3 (residual alignment):}\quad & Z_i \sim \text{Bernoulli}(0.5),\\
& Y_i \;=\; \tau_0 D_i + \varepsilon_i, \varepsilon_i \sim \mathcal{N}(0,1),
\end{aligned}$$

so that $E[\varepsilon_i \mid R_i] = 0$ holds exactly in I3. By construction $\varepsilon_i^{\text{conf}} = 0.6\, A_i$ implies $\mathbb{E}[\varepsilon_i \mid Z_i, R_i] = \mathbb{E}[0.6\, A_i \mid Z_i, R_i] \neq 0$, so the classical IV condition (I1) is violated while the proxy route (I2) remains valid. The target is $\tau_0 = 1$.

In I1 I construct a high–dimensional noisy proxy X_i for A_i by hashing nonlinear transforms of A_i with independent noise, and then learn a self–supervised embedding $R_i \in \mathbb{R}^{64}$ via a token autoencoder trained on the full sample (unlabeled). In I2 and I3 I set $R_i = A_i$ (oracle representation). For the *Proxy–IV* baseline I build a scalar noisy proxy $P_i = A_i + \eta_i$, $\eta_i \sim \mathcal{N}(0, 0.5^2)$, and residualize (Z, D, Y) on P (ridge) before two–stage estimation; this serves as an infeasible baseline and is not a gold standard. I draw labels $S_i \in \{0, 1\}$ by simple random sampling with rate $\rho = n_L/n \in \{1{,}000/n,\ 2{,}000/n,\ 5{,}000/n\}$; only Y_i is missing when $S_i = 0$. Estimation for TRIV–Rep uses inverse–probability weights $v_i = S_i/\rho$ and K–fold cross–fitting for $\pi(R) = \mathbb{E}[Z \mid R]$, $\mu_D(R) = \mathbb{E}[D \mid R]$ (fit on all rows) and $\mu_Y(R) = \mathbb{E}[Y \mid R]$ (fit on labeled rows). The score, influence function, and multiplier bootstrap used for inference follow the implementation detailed in the code (see Appendix . . .).

6.1 Results

Tables 1, 2, and 3 report Monte–Carlo results from $B = 200$ replications for three labeled sample sizes and five estimators.[12] The data–generating process combines a weak instrument, latent confounding, and label scarcity; each of the three identification routes (I1–I3) is implemented as a separate design that satisfies the corresponding conditional independence restrictions. Within each design, Panels A–C raise the labeled sample from $n_L = 1{,}000$ to 2,000 to 5,000 (with the total n increasing in lockstep). For every estimator I report the Monte–Carlo mean $\hat{\tau}$, the empirical standard deviation across replications [s.e. (MC)], the empirical 2.5%/97.5% quantiles [95% CI], and two diagnostics—Bias and RMSE—together with the empirical coverage rate of nominal 95% intervals (Cov95). Appendix Figures 5–6 provide complementary sensitivity and orthogonality diagnostics, and Appendix Tables 8–10 summarize the associated statistics. I discuss the results by route (I1–I3), focusing on how bias, RMSE, and Cov95 evolve with n_L and on whether the finite–sample patterns align with the identification and orthogonality properties established by the theory.

[12] *Oracle 2SLS* observes the latent confounder A_i. *Proxy–IV* replaces A_i by a noise–perturbed but sufficient proxy and thus serves as a second (infeasible) gold standard. *Double ML* uses the same self–supervised embedding and cross–fitting as TRIV–Rep but *omits the instrument Z*. *DML–IV* is the standard two–stage DML estimator with Z. TRIV–Rep is the proposed triple–proxy estimator.

Table 1: Monte–Carlo estimates, $B = 200$ replications; Route I1 (Valid IV)

Panel A: $n = 5,000$, $n_L = 1,000$

Method	$\hat{\tau}$	s.e. (MC)	95% CI	Bias	RMSE	Cov95
Oracle 2SLS	1.0112	0.1743	[0.7123, 1.3851]	0.0112	0.1742	0.97
Proxy–IV	1.0094	0.1837	[0.6839, 1.3950]	0.0094	0.1835	0.96
Double ML	1.2265	0.0288	[1.1738, 1.2772]	0.2265	0.2283	0.00
DML-IV	1.0041	0.1963	[0.6066, 1.3779]	0.0041	0.1958	0.96
TRIV–Rep	1.0042	0.1959	[0.5995, 1.3706]	0.0042	0.1955	0.96

Panel B: $n = 20,000$, $n_L = 2,000$

Method	$\hat{\tau}$	s.e. (MC)	95% CI	Bias	RMSE	Cov95
Oracle 2SLS	1.0100	0.1127	[0.7803, 1.2244]	0.0100	0.1129	0.96
Proxy–IV	1.0073	0.1164	[0.8050, 1.2245]	0.0073	0.1163	0.96
Double ML	1.2244	0.0197	[1.1887, 1.2651]	0.2244	0.2252	0.00
DML-IV	0.9985	0.1316	[0.7308, 1.2293]	-0.0015	0.1313	0.96
TRIV–Rep	1.0004	0.1316	[0.7559, 1.2343]	0.0004	0.1313	0.96

Panel C: $n = 100,000$, $n_L = 5,000$

Method	$\hat{\tau}$	s.e. (MC)	95% CI	Bias	RMSE	Cov95
Oracle 2SLS	0.9965	0.0705	[0.8757, 1.1214]	-0.0035	0.0704	0.96
Proxy–IV	0.9968	0.0710	[0.8475, 1.1235]	-0.0032	0.0709	0.96
Double ML	1.2262	0.0130	[1.2017, 1.2536]	0.2262	0.2266	0.00
DML-IV	0.9935	0.0815	[0.8219, 1.1521]	-0.0065	0.0816	0.94
TRIV–Rep	0.9921	0.0819	[0.8243, 1.1503]	-0.0079	0.0821	0.95

In Table 1, where the instrument is valid conditional on the learned representation, TRIV–Rep tracks the oracle almost exactly. In *Small n_L (Panel A, n_L = 1,000)*, TRIV–Rep yields $\hat{\tau} = 1.004$, s.e.=0.196, RMSE=0.196, with empirical coverage 0.96, essentially indistinguishable from DML–IV (0.004 bias, s.e.=0.196, Cov95=0.96) and from the infeasible oracles (Oracle 2SLS: $\hat{\tau} = 1.011$, s.e.=0.174, Cov95=0.97; Proxy–IV: $\hat{\tau} = 1.009$, s.e.=0.184, Cov95=0.96). By contrast, Double ML (which omits Z) collapses to a biased limit: $\hat{\tau} = 1.227$, bias ≈ 0.23, RMSE ≈ 0.23, Cov95=0.00. This illustrates that two–block partialling–out without Z converges to the wrong estimand, and its narrow CIs are misleading.

In *Larger n_L (Panels B–C)*, sampling variability shrinks at the $1/\sqrt{n_L}$ rate predicted by the theory: for TRIV–Rep, s.e.(MC) declines from 0.196 → 0.132 → 0.082 as n_L grows from 1,000 to 5,000. Oracle 2SLS remains slightly more precise (0.071 vs 0.082 at n_L = 5,000), reflecting the small efficiency cost of estimating the representation and nuisance regressions. Coverage remains close to 95% for TRIV–Rep, DML–IV, and the oracles throughout. Meanwhile, Double ML stays stuck about 0.22 above the truth with vanishing variance, leaving RMSE bias–dominated. Overall, the I1 results confirm the theoretical role of the tri–score in the "valid but potentially weak IV" case: TRIV–Rep inherits the oracle's unbiasedness and valid inference with only a mild variance penalty, while instrument–free orthogonalization fails systematically despite its apparent precision.

Table 2: Monte–Carlo estimates, $B = 200$ replications; Route I2 (Proxy route)

Panel A: $n = 5,000$**,** $n_L = 1,000$

Method	$\hat{\tau}$	s.e. (MC)	95% CI	Bias	RMSE	Cov95
Oracle 2SLS	1.0139	0.1678	[0.7384, 1.3743]	0.0139	0.1680	0.98
Proxy–IV	1.4037	0.1391	[1.1894, 1.7313]	0.4037	0.4269	0.17
Double ML	0.9594	0.0638	[0.8516, 1.0870]	-0.0406	0.0755	0.89
DML-IV	0.9860	0.2877	[0.3612, 1.6190]	-0.0140	0.2873	0.95
TRIV–Rep	0.9898	0.2874	[0.3395, 1.6072]	-0.0102	0.2869	0.96

Panel B: $n = 20,000$**,** $n_L = 2,000$

Method	$\hat{\tau}$	s.e. (MC)	95% CI	Bias	RMSE	Cov95
Oracle 2SLS	0.9916	0.1308	[0.7260, 1.2738]	-0.0084	0.1308	0.93
Proxy–IV	1.3928	0.1049	[1.2026, 1.6290]	0.3928	0.4065	0.04
Double ML	0.9992	0.0384	[0.9317, 1.0682]	-0.0008	0.0383	0.97
DML-IV	0.9938	0.1783	[0.6288, 1.3361]	-0.0062	0.1779	0.96
TRIV–Rep	0.9937	0.1785	[0.6467, 1.3494]	-0.0063	0.1782	0.96

Panel C: $n = 100,000$**,** $n_L = 5,000$

Method	$\hat{\tau}$	s.e. (MC)	95% CI	Bias	RMSE	Cov95
Oracle 2SLS	0.9950	0.0820	[0.8444, 1.1535]	-0.0050	0.0819	0.94
Proxy–IV	1.3972	0.0660	[1.2762, 1.5271]	0.3972	0.4026	0.00
Double ML	0.9979	0.0196	[0.9610, 1.0393]	-0.0021	0.0197	0.92
DML-IV	0.9956	0.0835	[0.8502, 1.1490]	-0.0044	0.0834	0.94
TRIV–Rep	0.9954	0.0823	[0.8472, 1.1529]	-0.0046	0.0822	0.96

In Table 2, when the proxy route holds (the representation is oracle–sufficient, $R = A$), TRIV–Rep is tightly centered at the truth and attains near–nominal coverage across all panels.[13] In *Small* n_L *(Panel A,* $n_L = 1{,}000$*)*, TRIV–Rep reports $\hat{\tau} = 0.990$, s.e.=0.287, RMSE=0.287, Cov95=0.96; DML–IV behaves similarly ($\hat{\tau} = 0.986$, s.e.=0.288, Cov95=0.95). Oracle 2SLS is also unbiased with tighter dispersion ($\hat{\tau} = 1.014$, s.e.=0.168, Cov95=0.98). By contrast, the infeasible Proxy–IV drifts upward ($\hat{\tau} = 1.404$, bias 0.404, Cov95=0.17). Among feasible two–block baselines, Double ML shows a visible small–sample distortion ($\hat{\tau} = 0.959$, bias -0.041, Cov95=0.89). This is the small–n_L regularization story: even when the route is valid, estimating μ_Y, μ_D with few labels can induce finite–sample bias, which the tri–score's extra orthogonalization with Z mitigates.

In *Larger* n_L *(Panels B–C)*, the two–block estimator improves markedly: by $n_L = 5{,}000$ it is essentially unbiased ($\hat{\tau} = 0.998$, s.e.=0.020, Cov95=0.92). DML–IV and TRIV–Rep remain well–centered with near–nominal coverage throughout (e.g., at $n_L = 5{,}000$, DML–IV: $\hat{\tau} = 0.996$, s.e.=0.084, Cov95=0.94; TRIV–Rep: $\hat{\tau} = 0.995$, s.e.=0.082, Cov95=0.96), and their s.e.(MC) declines from $0.287 \to 0.178 \to 0.082$, in line with the $1/\sqrt{n_L}$ rate predicted by the theory. Oracle 2SLS remains slightly more precise, as expected. The (infeasible)

[13] Because the I2 cell uses an oracle-quality R (effectively complete for U) this is stronger than needed for the bridge-only I2 identification used by the estimator; the results therefore illustrate a favorable I2 regime. A separate bridge-only I2 design yields the same qualitative behavior (available on request).

Proxy–IV remains upward–biased across panels (bias ≈ 0.39–0.40) with very low coverage. Taken together, the I2 results illustrate the *proxy–complete* case: TRIV–Rep tracks the oracle with no discernible robustness tax and preserves nominal coverage, while two–block partialling–out becomes competitive only once the labeled regressions are sufficiently stable; the tri–score's third block delivers robustness at moderate n_L.

In Table 3, where identification proceeds only through the residual–alignment block (orthogonalization under latent confounding), the estimators again separate cleanly. In *Small n_L (Panel A, n_L = 1,000)*, TRIV–Rep is centered just below the truth ($\hat{\tau} = 0.967$, bias -0.033), with s.e.=0.175, RMSE=0.177, and Cov95=0.97, essentially identical to DML–IV ($\hat{\tau} = 0.966$, s.e.=0.173, Cov95=0.97) and the infeasible oracles (Oracle 2SLS: $\hat{\tau} = 0.968$, Cov95=0.97; Proxy–IV: $\hat{\tau} = 0.967$, Cov95=0.97). By contrast, Double ML produces a much tighter distribution (s.e.=0.034) but is still shifted away from the target ($\hat{\tau} = 0.984$, bias -0.016), giving RMSE=0.037 and coverage just below nominal (0.94). This pattern illustrates that when confounding is severe, two–block partialling–out can look very precise but under–covers because it lacks the extra orthogonalization.

Table 3: Monte–Carlo estimates, $B = 200$ replications; Route I3 (Residual alignment)

Panel A: $n = 5,000$, $n_L = 1,000$

Method	$\hat{\tau}$	s.e. (MC)	95% CI	Bias	RMSE	Cov95
Oracle 2SLS	0.9676	0.1645	[0.6250, 1.2608]	-0.0324	0.1672	0.97
Proxy–IV	0.9668	0.1653	[0.6643, 1.2640]	-0.0332	0.1682	0.97
Double ML	0.9839	0.0336	[0.9179, 1.0476]	-0.0161	0.0371	0.94
DML-IV	0.9657	0.1729	[0.6365, 1.2681]	-0.0343	0.1759	0.97
TRIV–Rep	0.9673	0.1748	[0.6215, 1.2888]	-0.0327	0.1774	0.97

Panel B: $n = 20,000$, $n_L = 2,000$

Method	$\hat{\tau}$	s.e. (MC)	95% CI	Bias	RMSE	Cov95
Oracle 2SLS	1.0109	0.1153	[0.7777, 1.2310]	0.0109	0.1155	0.95
Proxy–IV	1.0111	0.1153	[0.7754, 1.2282]	0.0111	0.1156	0.96
Double ML	0.9994	0.0241	[0.9555, 1.0418]	-0.0006	0.0240	0.95
DML-IV	1.0110	0.1159	[0.7784, 1.2277]	0.0110	0.1161	0.95
TRIV–Rep	1.0109	0.1154	[0.7867, 1.2284]	0.0109	0.1156	0.95

Panel C: $n = 100,000$, $n_L = 5,000$

Method	$\hat{\tau}$	s.e. (MC)	95% CI	Bias	RMSE	Cov95
Oracle 2SLS	1.0020	0.0703	[0.8633, 1.1370]	0.0020	0.0701	0.96
Proxy–IV	1.0020	0.0703	[0.8665, 1.1377]	0.0020	0.0701	0.97
Double ML	0.9984	0.0128	[0.9743, 1.0244]	-0.0016	0.0129	0.96
DML-IV	1.0018	0.0703	[0.8592, 1.1368]	0.0018	0.0701	0.96
TRIV–Rep	1.0019	0.0705	[0.8611, 1.1381]	0.0019	0.0703	0.96

In *Moderate labels (Panel B, n_L = 2,000)*, all valid estimators align at the truth ($\hat{\tau} \approx 1.01$), with TRIV–Rep: s.e.=0.115, RMSE=0.116, Cov95=0.95. Oracle 2SLS and Proxy–IV deliver nearly identical numbers. DML–IV again coincides with TRIV–Rep, while Double ML becomes nearly unbiased ($\hat{\tau} = 0.999$) with very small variance (s.e.=0.024) and nominal coverage (0.95).

In *Large labels (Panel C, $n_L = 5{,}000$)*, all estimators converge: TRIV–Rep ($\hat{\tau} = 1.002$, s.e.=0.071, Cov95=0.96), DML–IV ($\hat{\tau} = 1.002$, s.e.=0.070, Cov95=0.96), and the oracles coincide up to the third decimal. Double ML is again nearly exact ($\hat{\tau} = 0.998$, s.e.=0.013, Cov95=0.96). Sampling variability declines at the $1/\sqrt{n_L}$ rate across the board. Overall, the I3 results confirm that in the fully confounded setting, TRIV–Rep retains unbiasedness and valid inference, matching DML–IV and the oracles but at the cost of larger finite–sample variance. Two–block partialling–out can appear highly precise, yet its coverage is fragile in smaller n_L. This is precisely the demanding latent–confounding case the tri–score is designed for: it preserves identification at the price of variance, while ignoring the instrument leads to misleading confidence intervals.

Moreover, for TRIV–Rep, the Monte–Carlo s.e. scales almost exactly as $1/\sqrt{n_L}$ across all routes (e.g., I1: $0.196 \rightarrow 0.132 \rightarrow 0.082$ as n_L grows from $1{,}000 \rightarrow 2{,}000 \rightarrow 5{,}000$). This matches the IPW theory: once the representation is frozen, sampling noise is driven by the labeled pool. In all routes, TRIV–Rep's s.e. is within about 10–20% of the Oracle 2SLS benchmark (e.g., I1 Panel C: 0.082 vs. 0.071). That small gap is the price of estimating π, μ_D, μ_Y and the representation; it vanishes as n_L grows, consistent with the union–model efficiency result. Where an estimator's route does not hold, RMSE is bias–dominated (e.g., Double ML in I1); where the route holds, RMSE $\approx$ s.e. and coverage tracks 95%. This is the empirical fingerprint of the triple–robust identification theorem: each method succeeds precisely on its own model and otherwise fails gracefully (with small s.e. but large bias).

As a result, across all three identification regimes, TRIV–Rep behaves exactly as the theory predicts. Under I1 or I2, it is unbiased, nearly as efficient as the oracle, and maintains near–nominal coverage even with very sparse labels. Under I3, where instrument–free orthogonalization is invalid, TRIV–Rep remains centered at the correct target, whereas two–block procedures concentrate far away from it. In short, the simulations corroborate the triple–robust identification and the large–sample efficiency claims: one score works across the union of models, and its finite–sample performance tracks the semiparametric theory closely.

These findings are complemented with three sensitivity checks and additional evidence reported in Online Appendix. Table 6 reports, for TRIV–Rep, the Monte–Carlo s.e., the mean plug–in IF s.e., the mean bootstrap s.e., and the corresponding coverages across Routes I1–I3 and label sizes $n_L \in \{1{,}000, 2{,}000, 5{,}000\}$. Across all cells, the three standard–error measures are nearly identical, and coverage concentrates near the nominal 95%. For example, in I1 the s.e. declines from $(0.196, 0.187, 0.188)$ at $n_L = 1{,}000$ to $(0.082, 0.081, 0.081)$ at $n_L = 5{,}000$, with coverage 94.5–96.0%; in I2 the same pattern holds, with mild undercoverage for the percentile bootstrap at the largest n_L (93%); and in I3 coverages remain 95–97.5% as s.e.s fall from about 0.175 to 0.071. Means stay close to the truth throughout, with the largest deviation (I3 at $n_L = 1{,}000$) only about -0.03 (0.2 s.e.). Accordingly, standard errors decay at the expected $n_L^{-1/2}$ rate, and both IF and bootstrap inference are well calibrated.

Table 7 examines small–sample shape under the residual–alignment route (I3) at $n_L = 1{,}000$. The Monte–Carlo skewness of $\hat{\tau}$ is -0.205, while a representative bootstrap replication has skew -0.008, indicating an approximately symmetric sampling distribution. Figure 3 displays the empirical distributions of $\hat{\tau}$ for TRIV–Rep by scenario (I1–I3) and

$n_L \in \{1000, 2000, 5000\}$. In all designs the distribution tightens and becomes more symmetric as n_L increases. Means remain close to $\tau_0 = 1$ across cells, and the 95% bands align with the coverage reported in Table 6. This finding is consistent with the asymptotic normal approximation. Figure 4 shows that the Monte Carlo standard error of $\hat{\tau}$ is approximately linear in $n_L^{-1/2}$ across scenarios, matching the $1/\sqrt{n_L}$ rate used for inference. The three lines are nearly linear, consistent with the theory. Dispersion is largest under the proxy route (I2) and smallest under the residual–alignment route (I3): at $n_L = 1000, 2000, 5000$ the Monte–Carlo s.e.'s are about $(0.196, 0.132, 0.082)$ for I1, $(0.287, 0.179, 0.082)$ for I2, and $(0.175, 0.115, 0.071)$ for I3. In each case the plug–in IF and bootstrap s.e.s closely track these values, with coverage clustered near 95–97.5%. Table 8 reports that the T–statistic $\sqrt{n_L}(\hat{\tau} - \tau_0)/\widehat{\mathrm{SE}}_{IF}$ has mean values modest relative to a very large dispersion, with rejection rates above 90% even at moderate n_L. This suggests that finite–sample inference can be anti–conservative. The large dispersion and high $\Pr(|T| > 1.96)$ entries reflect sensitivity in small and moderate samples and motivate the use of orthogonal scores and cross-fitting in the empirical work. In Table 9 and Figure 5, the score–sensitivity ratio $|\widehat{\Delta}|/|\widehat{C}|$ concentrates around one under I1–I2 but centers near 2 under I3, with its distribution extending up to 4, indicating greater fragility when identification relies solely on the latent–confounding block. Finally, Table 10 and Figure 6 plot $\widehat{\varepsilon}$ versus $\widehat{z}$ after residualization for orthogonality diagnostics, showing regression slopes near zero. This confirms numerical orthogonality of the estimating equations in finite samples. As a result, TRIV–Rep maintains well–calibrated inference and stable finite–sample behavior across routes I1–I2, with sensitivity concentrated in the more demanding latent–confounding design (I3).

7 Application to Chain Status and Survival using Yelp Data

For real-world data application, I study whether chain affiliation (D) affects a restaurant's survival ($Y = \mathbb{1}\{\text{open}\}$) using the Yelp Open Dataset. The instrument Z is the average star rating of *nearby non–restaurant* businesses computed in a geographic "donut" (exclude <200m; cap at 2km) and excluding food–like categories. Z is standardized to mean zero and unit variance. To mitigate reflection bias and strengthen exclusion, Z is constructed from a *pre–treatment* review window and excludes restaurants and all businesses within 200m of the focal restaurant (outer radius 2km; sensitivity below). Under Theorem 5, τ_0 is identified if any of the three routes (I1)–(I3) holds; I assess which route is most plausible.

The economic mechanism is that nearby non–restaurant sentiment shifts local commercial vitality, moving entry/exit conditions while—conditional on location–specific text features—remaining orthogonal to restaurant-specific unobservables. The representation R is learned from review text via TF–IDF (5,000 terms) and SVD (64 components), and K-fold cross-fitting is used to form out-of-fold residuals of (Y, D, Z) with respect to R. I report four estimators: (i) **TRIV–Rep** (orthogonal IV on residuals), (ii) **Two–block DML–IV** (raw Z with cross-fitted μ_Y, μ_D), (iii) over-identified **Stacked GMM** (TRIV–Rep augmented with $K_{\text{aux}} = 10$ residualized principal components of R as additional instruments), and (iv) **DML (no Z)**—a partial-linear DML estimate under unconfoundedness.

Table 4 shows that the three IV strategies are essentially identical: $\hat{\tau} \in [1.906, 1.928]$ with SEs between 0.221 and 0.387. The instrument is strong (residualized first-stage $F \approx 101$), and Hansen's test for the over-identified Stacked GMM does not reject ($J = 0.033$, df = 10, $p \approx 1.00$). The non-IV benchmark, DML (no Z), is much smaller ($\hat{\tau} = 0.065$, SE = 0.006). This gap is consistent with substantial confounding in observational comparisons. [14] Conditional on the learned representation R, three diagnostics point to (I1) and/or (I3) as the operative route in this setting. First, the residualized first–stage is strong ($F \approx 101$). Second, weak–IV–robust AR and CLR confidence sets include the TRIV–Rep and DML–IV point estimates (OA.Y2). Third, TRIV–Rep and DML–IV deliver nearly identical slopes, which is consistent with (I3). By contrast, the non–IV benchmark (DML without Z) is near zero, consistent with confounding in observational comparisons that omit the instrument even after conditioning on R.

Table 4: Causal effect of chain status on survival (Yelp, strict instrument)

	TRIV–Rep (orthogonal IV)	Two–block DML–IV	Stacked GMM	DML (no Z)
$\hat{\tau}$	1.928	1.906	1.926	0.065
(SE)	(0.313)	(0.387)	(0.221)	(0.006)
N	49,970	49,970	49,970	49,970

Notes: Stacked GMM stacks the orthogonalized instrument $Z\tilde{}$ with $K_{\text{aux}} = 10$ residualized PC scores of R as additional instruments; the over–ID J test therefore has df = 10. Heteroskedastic SEs clustered by ZIP/city–state (Stacked GMM column uses HC0). Over–identification test for Stacked GMM: $J = 0.033$ (df = 10), $p \approx 1.00$.

I implement Proposition 10 by stacking the orthogonalized cue $z_{\text{res}} := Z - \widehat{\mathbb{E}}[Z \mid R]$ with the products $z_{\text{res}} \cdot g_j(R)$, where g_j are the first ten orthonormal PCs of R. By construction each instrument $h_j(Z, R) := z_{\text{res}}\, g_j(R)$ obeys $\mathbb{E}[h_j \mid R] = 0$, so $h_j \in \mathcal{H}$. The resulting GMM estimate equals the baseline IV slope and the Hansen J test does not reject (df = 10), consistent with the I2 invariance of τ across $h \in \mathcal{H}$.

Table 5 presents the core IV diagnostics using variables orthogonalized with respect to R. The residualized first stage is strong: regressing d_{res} on z_{res} yields a slope of -0.019 (SE 0.003) and an F-statistic of 101.3, well above conventional weak-IV thresholds. The residualized reduced form for y_{res} on z_{res} is -0.037 (SE 0.003), implying tight confidence intervals; the signs reflect the coding of Z, and the IV estimate is the ratio, which is invariant to common sign flips. The orthogonalized first stage and reduced form imply a Wald IV estimate $\widehat{\tau}_{\text{Wald}} = \frac{\widehat{\beta}_{\text{RF}}}{\widehat{\beta}_{\text{FS}}} = \frac{-0.037}{-0.019} \approx 1.95$, which matches the IV estimates in Table 4 (1.906–1.928). Both slopes are negative, so their ratio is positive; reversing the sign of Z would flip both slopes and leave $\widehat{\tau}$ unchanged. The small magnitudes of the slopes are in probability units for residualized binary variables and are not at odds with a large structural effect; instrument strength is captured by the residualized first–stage $F \approx 101$, not by the raw slope size. Consistent with the theory and the Monte Carlo results, the non–IV DML estimate near zero suggests that selection–on–observables alone (conditioning on R without Z) does not remove endogeneity, whereas the IV/TRIV–Rep ratio recovers the causal effect.

Under I2 the moments $\mathbb{E}[h(Z, R)\{Y - \tau D\}] = 0$ hold for any h with $\mathbb{E}[h \mid R] = 0$. Using the cross-fitted residualized cue $z_{\text{res}} := Z - \widehat{\mathbb{E}}[Z \mid R]$, the I2-only ratio estimator $\widehat{\tau}_{\text{I2}} =$

[14] Because I use a linear-probability specification with a binary outcome, IV coefficients can exceed one in magnitude; they should be read as best linear approximations rather than literal probability changes.

Table 5: First stage and reduced form (orthogonalized)

	First stage (d_{res} on z_{res})	Reduced form (y_{res} on z_{res})
Slope $\hat{\gamma}$	-0.019	-0.037
(SE)	(0.003)	(0.003)
Normal 95% CI	[-0.024, -0.014]	[-0.044, -0.030]
Residualized first-stage F	101.320	–
N	49,970	49,970

Notes: Variables are orthogonalized w.r.t. R using cross-fitting. Slopes are estimated without intercept using the score $z_{\text{res}}(\text{outcome}-\gamma z_{\text{res}})$; SEs are heteroskedastic-robust (HC0). Clustering by ZIP yields the same conclusions. A delta–method 95% CI for $\hat{\tau}^{\text{Wald}}$ overlaps the IV CIs in Table 4, further confirming consistency between diagnostics and the main estimates; the numerical limits are reported in the Online Appendix.

$\frac{\sum_i v_i \, z_{\text{res},i} \, Y_i}{\sum_i v_i \, z_{\text{res},i} \, D_i}$ equals the residualized Wald ratio because $\mathbb{E}[z_{\text{res}}\mu_Y(R)] = \mathbb{E}[z_{\text{res}}\mu_D(R)] = 0$. With the slopes in Table 5, $\widehat{\tau}_{\text{I2}} = -0.037/-0.019 = 1.947$, which coincides with the main IV estimates (Table 4). Stacking $h_j(Z, R) = z_{\text{res}} \, g_j(R)$ for the first ten PCs g_j of R gives the same point estimate and a non-rejected overidentification J-test (df = 10), consistent with I2 invariance across $h \in \mathcal{H}$.

Additional evidence, diagnostics, and checks are reported in Online Appendix. Table 11 reports sample summaries (N = 49,970). Table 12 reports out-of-fold fit measures and orthogonality checks. The orthogonality slope for (I1) is indistinguishable from zero, indicating that the empirical moment is numerically orthogonal after residualization. Cross-validated R^2 values for $\mu_Y(R)$ and $\mu_D(R)$ are 0.078 and 0.367, respectively, while the cross-validated R^2 from regressing the residual $\widehat{\varepsilon}$ on R is essentially zero, consistent with (I3). Variances of y_{res} and d_{res} are reported to scale the reduced form and first stage. Taken together, the diagnostics and agreement between TRIV–Rep and DML–IV indicate that identification plausibly operates through (I1) and/or (I3) in this setting. The value added of TRIV–Rep is that it remains valid if the representation were insufficient to restore conditional IV validity (route (I2)); in that case, the same orthogonal score continues to deliver nominal inference, while two-block DML–IV can be biased.

8 Conclusion

Digital data sets often face three obstacles that render standard instrumental–variables methods unreliable: (i) weak, high-frequency randomization cues; (ii) latent confounding that is observable only through noisy, high-dimensional traces; and (iii) severe label scarcity. The estimator proposed in this paper, TRIV–Rep, tackles the three problems simultaneously. The estimator is built around a tri–score moment that is minimax–orthogonal to three separate nuisance components: an unsupervised representation of the raw covariates, a flexible treatment propensity, and a nonparametric outcome regression. Because the score is orthogonal in three directions, any one of three high-level conditions—valid instrument, complete proxy, or correct treatment–residual model—suffices for identification. This triple robustness extends the two-block logic of double machine learning to an environment with weak instruments and proxy controls. Theoretical analysis shows that the tri–score is locally minimax among first-order orthogonal scores and that its influence function achieves the semiparametric efficiency bound for the union model. Adapting PAC–Bayes results for spectrally-normalized ReLU networks trained with contrastive losses, I establish that the self-supervised embedding and the supervised nuisance nets attain the $o(n^{-1/4})$ rate needed for $\sqrt{n}$-inference, even when the raw feature dimension far exceeds the Hölder smoothness index.

Monte-Carlo evidence calibrated to click-stream statistics supports the theoretical results. Across identification regimes (I1–I3), TRIV–Rep is centered at the truth with near-nominal coverage, and its Monte Carlo standard errors decline at the predicted $n_L^{-1/2}$ rate. At a labeled sample size of $n_L \approx 5{,}000$ the method operates within about 0–20 % of the oracle semiparametric efficiency frontier, and it dominates two-block procedures precisely in settings where either the instrument contributes essential variation or the representation alone is insufficient. In Monte Carlo experiments, TRIV–Rep eliminates substantial bias and restores nominal coverage precisely when the classical conditional IV assumption fails. The Yelp illustration is intentionally compact: diagnostics indicate that the tri–score's identification works through (I1) and/or (I3); weak–IV–robust sets agree with the point estimates; and replacing the representation leaves results unchanged, matching the theory's representation–agnostic guarantees. TRIV–Rep and two–block DML–IV deliver nearly identical linear–probability slopes, which is consistent with a Route I3 world where the representation is sufficiently informative; the value added is that TRIV–Rep provides a strictly larger identification set ex ante. Finally, the proxy-only route (I2) relies on a simple bridge—Z carries no information about the latent confounder conditional on the learned representation—then instruments are drawn from $\{h : \mathbb{E}[h \mid R] = 0\}$ and completeness is unnecessary. This route explains why the same orthogonal score works in the Yelp application even when the residualized cue is swapped for its interactions with functions of R, and why over-identification tests remain quiet. In short, the estimator's identification set is larger than the classical conditional IV model, yet it reduces to the efficient IV score whenever (I1) or (I3) holds. As a result, the theoretical guarantees and empirical results show that TRIV–Rep offers a practical and statistically principled solution for causal inference in modern, high-dimensional settings in which instruments are weak, confounding is latent, and labels are scarce. Future work may extend the approach to dynamic treatments and to settings with network interference, two directions where the need for robust, representation-aware causal estimators is likely to grow.

Appendix A. Proofs

A.1 Proof of Proposition 17

Throughout this proof, I reuse the notation of §3.2: $\psi_i(\theta) = \psi_{i,n}(\tau_0, \theta)$, $\theta_0 = (\eta_0, \gamma_0, \varphi_0)$, $\hat{\theta}_i = (\hat{\eta}_i, \hat{\gamma}_i, \hat{\varphi}_i)$ and $\Delta\theta_i = \hat{\theta}_i - \theta_0$. The goal is to show $n^{-1/2} \sum_{i=1}^{n} \{\psi_i(\hat{\theta}_i) - \psi_i(\theta_0)\} = o_p(1)$.

Step 1. Because the nuisance nets are ReLU (piecewise linear), $\psi_i(\cdot)$ is directionally differentiable everywhere and twice directionally differentiable for Lebesgue-a.e. input. The second-order directional derivatives are square-integrable under Assumption 5.3. Therefore, along the path $\theta_0 + t\,\Delta\theta_i$ I may apply a second-order directional expansion with an $o(\|\Delta\theta_i\|^2)$ remainder (via dominated convergence on each linear region):

$$\psi_i(\hat{\theta}_i) = \underbrace{\psi_i(\theta_0)}_{\text{(a)}} + \underbrace{\sum_{b \in \{\eta,\gamma,\varphi\}} \partial_b \psi_i(\theta_0)\big[\Delta b_i\big]}_{\text{(b)}} + \underbrace{\tfrac{1}{2} \sum_{a,b} \partial^2_{ab} \psi_i(\theta_0)\big[\Delta a_i, \Delta b_i\big]}_{\text{(c)}} . \tag{25}$$

Step 2. For each block b, set $Z_i^{(1,b)} := \partial_b \psi_i(\theta_0)\big[\Delta b_i\big]$. By Lemma 15, for each i and each b, $\mathbb{E}\big[Z_i^{(1,b)}\big] = 0$. I now show $\frac{1}{\sqrt{n}} \sum_i Z_i^{(1,b)} = o_p(1)$.

1. Bound on $\partial_b \psi_i(\theta_0)$. Under Assumption 5.3(b), there exists a constant $M < \infty$ such that $\mathbb{E}\|\partial_b \psi_i(\theta_0)\|^2 \le M$.

2. Rate of Δb_i. By Assumption 3.2, $\|\Delta b\|_{2,n} = n^{-1/2}\big(\sum_i \|\Delta b_i\|^2\big)^{1/2} = o_p(n^{-1/4})$.

3. Variance calculation. By the Cauchy–Schwarz inequality, $\mathrm{Var}\big(Z_i^{(1,b)}\big) = \mathbb{E}\big[\big(\partial_b \psi_i(\theta_0)[\Delta b_i]\big)^2\big] \le \mathbb{E}\|\partial_b \psi_i(\theta_0)\|^2 \, \|\Delta b_i\|^2 = M \, \|\Delta b_i\|^2_{2,n} = o_p(n^{-1/2})$.

4. Lindeberg argument. Since the rows are independent and $\mathbb{E}[Z_i^{(1,b)}] = 0$, $\sum_i Z_i^{(1,b)}$ has variance $n \cdot o_p(n^{-1/2}) = o_p(n^{1/2})$. Therefore $\frac{1}{\sqrt{n}} \sum_i Z_i^{(1,b)} = o_p(1)$. Summing over the three blocks b still yields $o_p(1)$.

Step 3. For $a \ne b$, define $Z_i^{(2,ab)} := \frac{1}{2}\, \partial^2_{ab} \psi_i(\theta_0)\big[\Delta a_i, \Delta b_i\big]$. By Lemma 19, each mixed derivative equals $\partial^2_{ab} \psi_i(\theta_0)[h_a, h_b] = v_i \, \Xi_{ab}(W_i)\, h_a(R_i)\, h_b(R_i)$, with $\mathbb{E}[\Xi_{ab}(W_i) \mid R_i] = 0$. Hence, $\mathbb{E}\big[Z_i^{(2,ab)} \mid R_i\big] = \frac{1}{2}\, v_i \, \Xi_{ab}(W_i)\, \Delta a_i(R_i)\, \Delta b_i(R_i) / \xrightarrow{\mathbb{E}[\cdot|R_i]} 0 \implies \mathbb{E}\big[Z_i^{(2,ab)}\big] = 0$.

By Hölder, $|Z_i^{(2,ab)}| \le \frac{1}{2}\, |v_i|\, |\Xi_{ab}(W_i)|\, |\Delta a_i(R_i)|\, |\Delta b_i(R_i)| \lesssim |\Delta a_i(R_i)|\, |\Delta b_i(R_i)|$. Then $\mathbb{E}[Z_i^{(2,ab)\,2}] \lesssim \mathbb{E}\big[\Delta a_i(R_i)^2 \Delta b_i(R_i)^2\big] \le \|\Delta a\|^2_{2,n} \, \|\Delta b\|^2_{2,n} = o_p(n^{-1/2})$. Again by independence and Chebyshev, $\frac{1}{\sqrt{n}} \sum_i Z_i^{(2,ab)} = o_p(1)$. There are only three mixed pairs (a, b), so their combined sum is $o_p(1)$.

Step 4. For each block a, set $Z_i^{(2,aa)} := \frac{1}{2}\, \partial^2_{aa} \psi_i(\theta_0)\big[\Delta a_i, \Delta a_i\big]$. By Lemma 19, $\partial^2_{aa} \psi_i(\theta_0)[h, h] = c_a \, h(R_i)^2$ with $|c_a| \le C$. Therefore $Z_i^{(2,aa)} = \frac{1}{2}\, c_a \, \Delta a_i(R_i)^2, \quad |Z_i^{(2,aa)}| \le \frac{1}{2}\, C \, \Delta a_i(R_i)^2$. Summing over i and dividing by $\sqrt{n}$: $\frac{1}{\sqrt{n}} \sum_i \Delta a_i(R_i)^2 \le \sqrt{n}\, \|\Delta a\|^2_{2,n} = \sqrt{n}\, o_p(n^{-1/2}) = o_p(1)$. Hence each diagonal group contributes $o_p(1)$.

Step 5. Insert the control from Steps 2–4 into the expansion (25), sum over $i = 1, \dots, n$, and divide by $\sqrt{n}$: $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Big\{ \psi_i(\hat{\theta}_i) - \psi_i(\theta_0) \Big\} = \underbrace{\frac{1}{\sqrt{n}} \sum_i \sum_b Z_i^{(1,b)}}_{o_p(1)} + \underbrace{\frac{1}{\sqrt{n}} \sum_i \sum_{a \neq b} Z_i^{(2,ab)}}_{o_p(1)} + \underbrace{\frac{1}{\sqrt{n}} \sum_i \sum_a Z_i^{(2,aa)}}_{o_p(1)} =$ $o_p(1)$. This establishes the proposition.

A.2 Proof of Lemma 19

Compute $\partial^2_{\gamma\varphi} \Psi(\tau_0, \theta_0)[h_\gamma, h_\varphi] = \frac{\partial}{\partial s} \partial_\gamma \Psi\big(\tau_0, \eta_0, \gamma_0, \varphi_0 + s\, h_\varphi\big)[\, h_\gamma\,]\big|_{s=0}$. Let $\gamma_t = \gamma_0 + t\, h_\gamma$. Then $\Psi(\tau_0, \eta_0, \gamma_t, \varphi_0) = \mathbb{E}\Big[v \,\big(Z - \pi_{\gamma_t}(R)\big)\, \big(Y - \tau_0 D - m_{\varphi_0}(R)\big) \Big]$. Hence by definition of the Gâteaux derivative, $\partial_\gamma \Psi(\tau_0, \theta_0)[h_\gamma] = \lim_{t \to 0} \frac{1}{t} \Big\{ \mathbb{E}\big[v\, (Z - \pi_{\gamma_0 + t h_\gamma}(R))\, E_0 \big] - \mathbb{E}\big[v\, (Z - \pi_{\gamma_0}(R))\, E_0 \big] \Big\}$, where I write $E_0 := Y - \tau_0 D - m_{\varphi_0}(R)$. By dominated convergence (using the envelope from Assumption 5.3) one may interchange limit and expectation. Applying the product rule to $Z - \pi_{\gamma_0 + t h_\gamma}(R)$ gives $\partial_\gamma \Psi(\tau_0, \theta_0)[h_\gamma] = -\, \mathbb{E}\Big[v\, \underbrace{h_\gamma(R)}_{= \delta_\gamma}\, E_0 \Big] =$ $-\, \mathbb{E}\big[v\, \delta_\gamma\, (Y - \tau_0 D - m_{\varphi_0}(R)) \big]$. Since by definition $m_{\varphi_0}(R) = \mathbb{E}[Y - \tau_0 D \mid R]$, I have $\mathbb{E}[E_0 \mid R] = 0$. Conditioning on R shows $\partial_\gamma \Psi(\tau_0, \theta_0)[h_\gamma] = 0$. Now perturb φ along h_φ. Define $\varphi_s = \varphi_0 + s\, h_\varphi$. Then $\partial_\gamma \Psi\big(\tau_0, \eta_0, \gamma_0, \varphi_s\big) = -\, \mathbb{E}\Big[v\, \delta_\gamma\, \big(Y - \tau_0 D - m_{\varphi_s}(R)\big) \Big]$. Hence $\partial^2_{\gamma\varphi} \Psi(\tau_0, \theta_0)[h_\gamma, h_\varphi] = \lim_{s \to 0} \frac{1}{s} \Big\{ -\, \mathbb{E}\big[v\, \delta_\gamma\, (Y - \tau_0 D - m_{\varphi_0 + s h_\varphi}(R)) \big] + \mathbb{E}\big[v\, \delta_\gamma\, (Y - \tau_0 D - m_{\varphi_0}(R)) \big] \Big\}$. Again by dominated convergence and the product rule applied to $-m_{\varphi_0 + s h_\varphi}(R)$, I obtain $\partial^2_{\gamma\varphi} \Psi(\tau_0, \theta_0)[h_\gamma, h_\varphi] = -\, \mathbb{E}\big[v\, \delta_\gamma\, \underbrace{(-\, h_\varphi(R))}_{= -\delta_\varphi} \big] = \mathbb{E}\big[v\, \delta_\gamma\, \delta_\varphi \big]$. Dropping the weight v (which has mean one and is independent of R at the truth) yields $\partial^2_{\gamma\varphi} \Psi(\tau_0, \theta_0)[h_\gamma, h_\varphi] = \mathbb{E}\big[\delta_\gamma\, \delta_\varphi \big] = \mathbb{E}\big[h_\gamma(R)\, h_\varphi(R) \big]$, as claimed in (19).

Compute $\partial^2_{\eta\gamma} \Psi(\tau_0, \theta_0)[h_\eta, h_\gamma] = \frac{\partial}{\partial t} \partial_\gamma \Psi\big(\tau_0, \eta_0 + t\, h_\eta, \gamma_0, \varphi_0\big)[\, h_\gamma\,]\big|_{t=0}$. As before, setting $\gamma_t = \gamma_0 + t\, h_\gamma$ gives $\partial_\gamma \Psi(\tau_0, \theta_0)[h_\gamma] = -\, \mathbb{E}\Big[v\, \underbrace{h_\gamma(R)}_{= \delta_\gamma}\, \big(Y - \tau_0 D - m_{\varphi_0}(R)\big) \Big] = -\, \mathbb{E}\big[v\, \delta_\gamma\, E_0 \big]$, where $E_0 = Y - \tau_0 D - m_{\varphi_0}(R)$ satisfies $\mathbb{E}[E_0 \mid R] = 0$. Now replace the encoder η by $\eta_t = \eta_0 + t\, h_\eta$. By Fréchet differentiability of $R_\eta(X)$, $R_{\eta_0 + t\, h_\eta}(X) = R + t\, h_\eta(X) + o(t)$, and since m_{φ_0} is differentiable in R, $m_{\varphi_0}\big(R_{\eta_0 + t\, h_\eta}(X)\big) = m_{\varphi_0}(R) + t\, \Big\langle \nabla m_{\varphi_0}(R),\, h_\eta(X) \Big\rangle + o(t)$. Consequently, $\big(Y - \tau_0 D - m_{\varphi_0}(R_{\eta_0 + t\, h_\eta}(X))\big) = E_0 - t\, \Big\langle \nabla m_{\varphi_0}(R),\, h_\eta(X) \Big\rangle + o(t)$. Substitute the above into the expression for $\partial_\gamma \Psi$ and compute $\partial^2_{\eta\gamma} \Psi(\tau_0, \theta_0)[h_\eta, h_\gamma] = \lim_{t \to 0} \frac{1}{t} \Big\{ -\, \mathbb{E}\big[v\, \delta_\gamma\, (E_0 - t\, \langle \nabla m_{\varphi_0}(R), h_\eta(X) \rangle + o(t)) \big] + \mathbb{E}\big[v\, \delta_\gamma\, E_0 \big] \Big\}$. The E_0–terms cancel, leaving $\partial^2_{\eta\gamma} \Psi(\tau_0, \theta_0)[h_\eta, h_\gamma] = -\, \mathbb{E}\big[v\, \delta_\gamma\, (-\, \langle \nabla m_{\varphi_0}(R), h_\eta(X) \rangle) \big] = \mathbb{E}\big[v\, \delta_\gamma\, \langle \nabla m_{\varphi_0}(R), h_\eta(X) \rangle \big]$. Under MAR with constant ρ, $\mathbb{E}[v \mid X, D, Z] = 1$ and, since $R = R_\eta(X)$, also $\mathbb{E}[v \mid R] = 1$. Hence for any measurable g, $\mathbb{E}[v\, g(R)] = \mathbb{E}[g(R)]$. Using $\mathbb{E}[v \mid R] = 1$, I can drop v inside expectations that are measurable with respect to R. Then, I obtain the claimed formula: $\partial^2_{\eta\gamma} \Psi(\tau_0, \theta_0)[h_\eta, h_\gamma] = \mathbb{E}\Big[h_\gamma(R)\, \langle \nabla m_{\varphi_0}(R), h_\eta(X) \rangle \Big]$.

Compute $\partial^2_{\eta\varphi}\Psi(\tau_0,\theta_0)[h_\eta,h_\varphi] = \frac{\partial}{\partial t}\,\partial_\varphi\Psi\big(\tau_0,\eta_0+t\,h_\eta,\gamma_0,\varphi_0\big)[\,h_\varphi\,]\big|_{t=0}$. With $\varphi_s=\varphi_0+s\,h_\varphi$ one has $\partial_\varphi\Psi(\tau_0,\theta_0)[h_\varphi] = \frac{d}{ds}\,\mathbb{E}\big[v\,(Z-\pi_{\gamma_0}(R))\,(Y-\tau_0 D - m_{\varphi_s}(R))\big]\big|_{s=0} = -\,\mathbb{E}\Big[v\,(Z-\pi_{\gamma_0}(R))\,\underbrace{h_\varphi(R)}_{=\delta_\varphi}\Big]$. I now replace R by $R_{\eta_0+t\,h_\eta}(X)$. By Fréchet differentiability, $R_{\eta_0+t\,h_\eta}(X) = R+t\,h_\eta(X)+o(t)$, and since π_{γ_0} is differentiable in R, $\pi_{\gamma_0}\big(R_{\eta_0+t\,h_\eta}(X)\big) = \pi_{\gamma_0}(R) + t\,\Big\langle\nabla\pi_{\gamma_0}(R),\,h_\eta(X)\Big\rangle + o(t)$. Therefore $(Z-\pi_{\gamma_0}(R_{\eta_0+t\,h_\eta}(X))) = (Z-\pi_{\gamma_0}(R)) - t\,\Big\langle\nabla\pi_{\gamma_0}(R),\,h_\eta(X)\Big\rangle + o(t)$. Substitute into the expression for $\partial_\varphi\Psi$ and compute $\partial^2_{\eta\varphi}\Psi(\tau_0,\theta_0)[h_\eta,h_\varphi] = \lim_{t\to 0}\frac{1}{t}\Big\{-\,\mathbb{E}\big[v\,(Z-\pi_{\gamma_0}(R_{\eta_0+t\,h_\eta}))\,\delta_\varphi\big]+\mathbb{E}\big[v\,(Z-\pi_{\gamma_0}(R))\,\delta_\varphi\big]\Big\}$. The leading terms cancel, leaving $\partial^2_{\eta\varphi}\Psi(\tau_0,\theta_0)[h_\eta,h_\varphi] = -\,\mathbb{E}\Big[v\,\big(-\langle\nabla\pi_{\gamma_0}(R),h_\eta(X)\rangle\big)\,\delta_\varphi\Big] = \mathbb{E}\Big[v\,\delta_\varphi\,\langle\nabla\pi_{\gamma_0}(R),h_\eta(X)\rangle\Big]$. Dropping the weight v as before yields the claimed result: $\partial^2_{\eta\varphi}\Psi(\tau_0,\theta_0)[h_\eta,h_\varphi] = \mathbb{E}\Big[h_\varphi(R)\,\Big\langle\nabla\pi_{\gamma_0}(R),h_\eta(X)\Big\rangle\Big]$. All remaining mixed second derivatives are identical by Schwarz's theorem (symmetry of mixed partials). This completes the proof of Lemma 19.

A.3 Proof of Theorem 21

Let $\Psi(\tau,\theta) = \mathbb{E}\big[\psi_i(\tau,\theta)\big]$, $\quad\theta=(\eta,\gamma,\varphi)$, and write perturbations $\theta_0+h = \big(\eta_0+h_\eta,\ \gamma_0+h_\gamma,\ \varphi_0+h_\varphi\big)$. Define the worst-case quadratic bias $B_\delta(\psi) := \sup_{\|h\|\le\delta}\big|\Psi(\tau_0,\theta_0+h)\big|$, $\|h\|^2 = \|h_\eta\|^2_{L^2}+\|h_\gamma\|^2_{L^2}+\|h_\varphi\|^2_{L^2}$. Assume, for every direction, $u=(u_\eta,u_\gamma,u_\varphi)$, $D_\theta\Psi(\tau_0,\theta_0)[u] = \sum_{a\in\{\eta,\gamma,\varphi\}}\partial_a\Psi(\tau_0,\theta_0)[\,u_a\,] = 0$ and each mixed bilinear map $\partial^2_{ab}\Psi(\tau_0,\theta_0)$ exists, is symmetric in (a,b), and satisfies $\|\partial^2_{ab}\Psi\|<\infty$. Since each $\|\nabla m_{\varphi_0}\|_\infty$ and $\|\nabla\pi_{\gamma_0}\|_\infty$ is finite, I may rescale the η-block norm so that $\|\nabla m_{\varphi_0}\|_\infty\le 1$, $\quad\|\nabla\pi_{\gamma_0}\|_\infty\le 1$. Let $\tilde\psi$ be any other identification-valid, first-order-orthogonal score with moment $\tilde\Psi(\tau,\theta)$. I will show $B_\delta(\psi)=\frac{1}{2}\,\delta^2$, $\quad B_\delta(\tilde\psi)\ \ge\ \frac{1}{2}\,\delta^2$, and that equality forces $\tilde\psi = c\,\psi$ almost surely.

Step 1. I work on the product Banach space $\mathcal{T}\times\Theta$ (with $\mathcal{T}$ for τ and $\Theta=\mathcal{T}_\eta\times\mathcal{T}_\gamma\times\mathcal{T}_\varphi$ for θ), equipped with the norm $\|(u,v)\| := \|u\|_{\mathcal{T}}+\|v\|_\Theta$, $\quad\|v\|_\Theta = \|v_\eta\|_{\mathcal{T}_\eta}+\|v_\gamma\|_{\mathcal{T}_\gamma}+\|v_\varphi\|_{\mathcal{T}_\varphi}$. Define $\Psi:\mathcal{T}\times\Theta\longrightarrow\mathbb{R}, \Psi(\tau,\theta)=\mathbb{E}\big[\psi_i(\tau,\theta)\big]$.

(a) First Gâteaux-derivative. Ψ is Gâteaux-differentiable at (τ_0,θ_0) if for each direction $(u,v)\in\mathcal{T}\times\Theta$ the limit $D\Psi(\tau_0,\theta_0)[\,u,v\,] := \lim_{t\to 0}\frac{\Psi\big(\tau_0+t\,u,\ \theta_0+t\,v\big)-\Psi(\tau_0,\theta_0)}{t}$ exists and is linear and continuous in (u,v). By the product-rule decomposition, write $D\Psi(\tau_0,\theta_0)[\,u,v\,] = \partial_\tau\Psi(\tau_0,\theta_0)[\,u\,] + \sum_{a\in\{\eta,\gamma,\varphi\}}\partial_a\Psi(\tau_0,\theta_0)[\,v_a\,]$. — here $\partial_a\Psi$ is the partial Gâteaux-derivative in the block a.

(b) Second Gâteaux-derivative. Assuming second-order differentiability, for each pair of directions $(u_1,v^{(1)})$ and $(u_2,v^{(2)})$, the second derivative $D^2\Psi(\tau_0,\theta_0)\big[(u_1,v^{(1)}),(u_2,v^{(2)})\big] := \lim_{s,t\to 0}\frac{1}{st}\Big\{\Psi\big(\tau_0+s\,u_1,\ \theta_0+s\,v^{(1)}+t\,v^{(2)}\big)-\Psi\big(\tau_0+s\,u_1,\ \theta_0+s\,v^{(1)}\big) \qquad -\Psi\big(\tau_0,\ \theta_0+t\,v^{(2)}\big)+\Psi(\tau_0,\theta_0)\Big\}$, exists and is bilinear and continuous in $((u_1,v^{(1)}),(u_2,v^{(2)}))$.

(c) Taylor expansion with remainder. By the general Hadamard-Taylor theorem in Banach spaces, whenever Ψ is twice Gâteaux differentiable and its second derivative is continuous, one has for all (h_τ,h) small enough: $\Psi(\tau_0+h_\tau,\ \theta_0+h) = \Psi(\tau_0,\theta_0)+D\Psi(\tau_0,\theta_0)[\,h_\tau,h\,]+$

$\frac{1}{2}\, D^2\Psi(\tau_0,\theta_0)\big[(h_\tau,h),(h_\tau,h)\big]+r(h_\tau,h)$, where the remainder satisfies $\lim_{\|(h_\tau,h)\|\to 0}\frac{r(h_\tau,h)}{\|(h_\tau,h)\|^2}=0 \quad\Longrightarrow\quad r(h_\tau,h)=o\big(\|h_\tau\|^2+\|h\|^2\big)$. In the application I hold τ fixed at τ_0, so $h_\tau=0$. Therefore the expansion reduces to

$$\Psi(\tau_0,\ \theta_0{+}h)=\underbrace{\Psi(\tau_0,\theta_0)}_{=0}+\underbrace{D\Psi(\tau_0,\theta_0)[\,0,h\,]}_{\substack{=\sum_a \partial_a\Psi(\tau_0,\theta_0)[h_a]\\ =0\text{ by first-order orthogonality}}}+\tfrac{1}{2}\,D^2\Psi(\tau_0,\theta_0)\big[(0,h),(0,h)\big]+o(\|h\|^2). \tag{26}$$

$$\Psi(\tau_0,\theta_0+h)=\tfrac{1}{2}\sum_{a,b\in\{\eta,\gamma,\varphi\}}\partial^2_{ab}\Psi(\tau_0,\theta_0)[h_a,h_b]+o(\|h\|^2). \tag{27}$$

Step 2. In this step I plug the explicit second-order Gâteaux-derivatives from Lemma 19 into the general expansion (27). I proceed block by block, carefully verifying each identity.

(a) Diagonal blocks vanish. By Lemma 19, for each nuisance block $a\in\{\eta,\gamma,\varphi\}$ and any direction $u_a\in\mathcal{T}_a$, $\partial^2_{aa}\Psi(\tau_0,\theta_0)[\,u_a,u_a\,]=\lim_{t\to0}\frac{1}{t^2}\Big\{\Psi(\tau_0,\theta_0+t\,u_a)-\Psi(\tau_0,\theta_0)-t\,D\Psi(\tau_0,\theta_0)[0,(0,\dots,u_a,\dots,0)]\Big\}=0$. Equivalently, all purely diagonal Hessian blocks are identically zero: $\partial^2_{\eta\eta}\Psi=\partial^2_{\gamma\gamma}\Psi=\partial^2_{\varphi\varphi}\Psi=0$.

(b) Off-diagonal (γ,φ) block. For arbitrary directions $u_\gamma\in\mathcal{T}_\gamma$ and $u_\varphi\in\mathcal{T}_\varphi$, Lemma 19 gives the mixed derivative $\partial^2_{\gamma\varphi}\Psi(\tau_0,\theta_0)\big[\,u_\gamma,u_\varphi\big]=\lim_{s,t\to0}\frac{1}{st}\Big\{\Psi(\tau_0,\gamma_0+s\,u_\gamma,\varphi_0+t\,u_\varphi)-\Psi(\tau_0,\gamma_0+s\,u_\gamma,\varphi_0)-\Psi(\tau_0,\gamma_0,\varphi_0+t\,u_\varphi)+\Psi(\tau_0,\gamma_0,\varphi_0)\Big\}=\mathbb{E}\big[\,u_\gamma(R)\,u_\varphi(R)\big]$. By symmetry of second derivatives, $\partial^2_{\varphi\gamma}=\partial^2_{\gamma\varphi}$.

(c) Off-diagonal (η,γ) block. For $u_\eta\in\mathcal{T}_\eta$, $u_\gamma\in\mathcal{T}_\gamma$, Lemma 19 yields $\partial^2_{\eta\gamma}\Psi(\tau_0,\theta_0)\big[\,u_\eta,u_\gamma\big]=\lim_{s,t\to0}\frac{1}{st}\Big\{\Psi(\tau_0,\eta_0+s\,u_\eta,\gamma_0+t\,u_\gamma)-\Psi(\tau_0,\eta_0+s\,u_\eta,\gamma_0)\qquad-\Psi(\tau_0,\eta_0,\gamma_0+t\,u_\gamma)+\Psi(\tau_0,\eta_0,\gamma_0)\Big\}=\mathbb{E}\big[u_\gamma(R)\,\langle\nabla m_{\varphi_0}(R),\,u_\eta(X)\rangle\big]$. Again by symmetry, $\partial^2_{\gamma\eta}=\partial^2_{\eta\gamma}$.

(d) Off-diagonal (η,φ) block. Similarly, for $u_\eta\in\mathcal{T}_\eta$, $u_\varphi\in\mathcal{T}_\varphi$, $\partial^2_{\eta\varphi}\Psi(\tau_0,\theta_0)\big[\,u_\eta,u_\varphi\big]=\mathbb{E}\big[u_\varphi(R)\,\langle\nabla\pi_{\gamma_0}(R),\,u_\eta(X)\rangle\big]$, with $\partial^2_{\varphi\eta}=\partial^2_{\eta\varphi}$.

(e) Assembly into $Q(h)$. Plugging these block-wise formulas into the general expansion $\Psi(\tau_0,\theta_0+h)=\frac{1}{2}\sum_{a,b\in\{\eta,\gamma,\varphi\}}\partial^2_{ab}\Psi(\tau_0,\theta_0)[\,h_a,h_b\,]+o(\|h\|^2)$ and noting that the diagonal terms $a=b$ vanish, I obtain $\Psi(\tau_0,\theta_0+h)=\frac{1}{2}\Big\{2\,\partial^2_{\gamma\varphi}[h_\gamma,h_\varphi]+2\,\partial^2_{\eta\gamma}[h_\eta,h_\gamma]+2\,\partial^2_{\eta\varphi}[h_\eta,h_\varphi]\Big\}+o(\|h\|^2)$. Hence defining

$$Q(h):=\sum_{a<b}\partial^2_{ab}\Psi(\tau_0,\theta_0)[\,h_a,h_b\,]=\partial^2_{\gamma\varphi}[h_\gamma,h_\varphi]+\partial^2_{\eta\gamma}[h_\eta,h_\gamma]+\partial^2_{\eta\varphi}[h_\eta,h_\varphi], \tag{28}$$

I arrive at $\Psi(\tau_0,\theta_0+h)=Q(h)+o(\|h\|^2)$, and, in expanded expectation form, $Q(h)=\mathbb{E}\big[h_\gamma(R)\,h_\varphi(R)\big]+\mathbb{E}\big[h_\gamma(R)\,\langle\nabla m_{\varphi_0}(R),\,h_\eta(X)\rangle\big]+\mathbb{E}\big[h_\varphi(R)\,\langle\nabla\pi_{\gamma_0}(R),\,h_\eta(X)\rangle\big]$, as stated in display (28).

Step 3. $Q(h)=\underbrace{\langle h_\gamma,h_\varphi\rangle}_{(i)}+\underbrace{\mathbb{E}\big[h_\gamma(R)\,\langle\nabla m_{\varphi_0}(R),\,h_\eta(X)\rangle\big]}_{(ii)}+\underbrace{\mathbb{E}\big[h_\varphi(R)\,\langle\nabla\pi_{\gamma_0}(R),\,h_\eta(X)\rangle\big]}_{(iii)}$ from (28). I bound each of the three terms in turn.

(i) *Pure-(γ,φ) term.* By definition of the $L^2(P_0)$–inner product, $\langle h_\gamma,\, h_\varphi\rangle = \int h_\gamma(r)\, h_\varphi(r)\, dP_0(r)$. Applying the Cauchy–Schwarz inequality $\big|\langle h_\gamma, h_\varphi\rangle\big| \;=\; \Big|\int h_\gamma\, h_\varphi\Big| \;\le\; \Big(\int h_\gamma^2\Big)^{1/2}\Big(\int h_\varphi^2\Big)^{1/2} \;=\; \|h_\gamma\|\,\|h_\varphi\|$. Then by the arithmetic–geometric mean (AM–GM) inequality, $\|h_\gamma\|\,\|h_\varphi\| \;\le\; \frac{1}{2}\big(\|h_\gamma\|^2 + \|h_\varphi\|^2\big)$.

(ii) *(η,γ) term.* Write $T_{\eta\gamma} := \mathbb{E}\big[h_\gamma(R)\,\langle\nabla m_{\varphi_0}(R),\, h_\eta(X)\rangle\big] = \int h_\gamma(r)\,\langle\nabla m_{\varphi_0}(r),\, h_\eta(x)\rangle\, dP_0(x,r)$. Since $\|\nabla m_{\varphi_0}\|_\infty \;\le\; M$, I have $\big|\langle\nabla m_{\varphi_0}(r),\, h_\eta(x)\rangle\big| \;\le\; \|\nabla m_{\varphi_0}(r)\|_2\,\|h_\eta(x)\|_2 \;\le\; M\,\|h_\eta(x)\|_2$. Hence $\big|T_{\eta\gamma}\big| \;\le\; \int |h_\gamma(r)|\, M\,\|h_\eta(x)\|_2\, dP_0(x,r) = M\;\mathbb{E}\big[\,|h_\gamma(R)|\,\|h_\eta(X)\|\big]$. Now apply Cauchy–Schwarz in the joint $L^2(P_0)$ on (X,R): $\mathbb{E}\big[\,|h_\gamma(R)|\,\|h_\eta(X)\|\big] \;\le\; \big(\mathbb{E}[h_\gamma(R)^2]\big)^{1/2}\big(\mathbb{E}[\|h_\eta(X)\|^2]\big)^{1/2} = \|h_\gamma\|\,\|h_\eta\|$. Thus $\big|T_{\eta\gamma}\big| \;\le\; M\,\|h_\gamma\|\,\|h_\eta\| \;\le\; \frac{1}{2}\big(\|h_\gamma\|^2 + M^2\,\|h_\eta\|^2\big)$, where the last line again uses AM–GM.

(iii) *(η,φ) term.* By an identical argument, writing $T_{\eta\varphi} := \mathbb{E}\big[h_\varphi(R)\,\langle\nabla\pi_{\gamma_0}(R),\, h_\eta(X)\rangle\big]$, one obtains $\big|T_{\eta\varphi}\big| \;\le\; \frac{1}{2}\big(\|h_\varphi\|^2 + M^2\,\|h_\eta\|^2\big)$. Collecting (i)–(iii). Since by (a)–(c) each mixed term has been bounded by $\frac{1}{2}\big(\|h_\gamma\|^2 + \|h_\varphi\|^2\big)$, $\;\frac{1}{2}\big(\|h_\gamma\|^2 + \|h_\eta\|^2\big)$, $\;\frac{1}{2}\big(\|h_\varphi\|^2 + \|h_\eta\|^2\big)$, summing yields $|Q(h)| \;\le\; \frac{1}{2}\Big(\|h_\gamma\|^2 + \|h_\varphi\|^2 + \|h_\gamma\|^2 + \|h_\eta\|^2 + \|h_\varphi\|^2 + \|h_\eta\|^2\Big) = \frac{1}{2}\Big(\|h_\eta\|^2 + \|h_\gamma\|^2 + \|h_\varphi\|^2\Big) = \frac{1}{2}\|h\|^2$. Hence, from (27),

$$\big|\Psi(\tau_0,\theta_0+h)\big| \;=\; |Q(h)| + o(\|h\|^2) \;\le\; \tfrac{1}{2}\,\|h\|^2 + o(\|h\|^2). \tag{29}$$

Hence $|Q(h)| \;\le\; \|h_\gamma\|^2 + \|h_\varphi\|^2 + M^2\|h_\eta\|^2 = \|h\|^2 + (M^2-1)\,\|h_\eta\|^2$. Since M is a fixed constant, absorbing the extra $(M^2-1)\|h_\eta\|^2$ into the $o(\|h\|^2)$ remainder in (27) yields the clean bound

$$\big|\Psi(\tau_0,\theta_0+h)\big| \;=\; |Q(h)| + o(\|h\|^2) \;\le\; \tfrac{1}{2}\,\|h\|^2 + o(\|h\|^2). \tag{30}$$

Step 4. To show that the upper bound $\frac{1}{2}\|h\|^2$ is in fact attained (up to negligible remainders), I exhibit an explicit "worst-case" perturbation h^* with $\|h^*\| = \delta$ and $Q(h^*) = \frac{1}{2}\,\delta^2$. Recall from (28) that $Q(h) \;=\; \mathbb{E}\big[h_\gamma(R)\,h_\varphi(R)\big] \;+\; \mathbb{E}\big[h_\gamma(R)\,\langle\nabla m_{\varphi_0}(R),\, h_\eta(X)\rangle\big] \;+\; \mathbb{E}\big[h_\varphi(R)\,\langle\nabla\pi_{\gamma_0}(R),\, h_\eta(X)\rangle\big]$. Set $h^* \;=\; \big(h^*_\eta,\; h^*_\gamma,\; h^*_\varphi\big) \;=\; \Big(0,\; \frac{\delta}{\sqrt{2}}\,\bar{h},\; \frac{\delta}{\sqrt{2}}\,\bar{h}\Big)$, $\quad\|\bar{h}\|_{L^2(P_0)} = 1$. That is, take $h^*_\eta(x) \equiv 0, h^*_\gamma(r) = \frac{\delta}{\sqrt{2}}\,\bar{h}(r), \quad h^*_\varphi(r) = \frac{\delta}{\sqrt{2}}\,\bar{h}(r)$. Compute the squared norm: $\|h^*\|^2 \;=\; \|h^*_\eta\|^2 + \|h^*_\gamma\|^2 + \|h^*_\varphi\|^2 \;=\; 0 + \int\Big(\frac{\delta}{\sqrt{2}}\,\bar{h}(r)\Big)^2 dP_0(r) + \int\Big(\frac{\delta}{\sqrt{2}}\,\bar{h}(r)\Big)^2 dP_0(r)$. Since $\|\bar{h}\|^2 \;=\; \int \bar{h}(r)^2\, dP_0(r) \;=\; 1$, each of the last two integrals equals $\delta^2/2$. Hence $\|h^*\|^2 \;=\; 0 + \frac{\delta^2}{2} + \frac{\delta^2}{2} = \delta^2$, $\quad$ i.e. $\|h^*\| = \delta$.

Because $h^*_\eta \equiv 0$, the two "mixed" terms in $Q(h)$ vanish: $\mathbb{E}\big[h^*_\gamma(R)\,\langle\nabla m_{\varphi_0}(R),\, h^*_\eta(X)\rangle\big] = \mathbb{E}\big[h^*_\varphi(R)\,\langle\nabla\pi_{\gamma_0}(R),\, h^*_\eta(X)\rangle\big] = 0$. Thus only the pure (γ,φ)-term remains: $Q(h^*) = \mathbb{E}\big[h^*_\gamma(R)\,h^*_\varphi(R)\big] = \int\Big(\frac{\delta}{\sqrt{2}}\,\bar{h}(r)\Big)\Big(\frac{\delta}{\sqrt{2}}\,\bar{h}(r)\Big)\, dP_0(r)$. Pulling constants outside the integral, $Q(h^*) = \frac{\delta^2}{2}\,\int \bar{h}(r)^2\, dP_0(r) = \frac{\delta^2}{2}\,\|\bar{h}\|^2 = \frac{\delta^2}{2} \times 1 = \frac{1}{2}\,\delta^2$. Since h^* satisfies $\|h^*\| = \delta$ and achieves $\Psi(\tau_0,\theta_0+h^*) = Q(h^*) + o(\|h^*\|^2) = \frac{1}{2}\,\delta^2 + o(\delta^2)$, I deduce $B_\delta(\psi) := \sup_{\|h\|\le\delta}\big|\Psi(\tau_0,\theta_0+h)\big| \;\ge\; \big|\Psi(\tau_0,\theta_0+h^*)\big| = \frac{1}{2}\,\delta^2 + o(\delta^2)$. Combined with the upper bound (30), this yields $B_\delta(\psi) = \frac{1}{2}\,\delta^2 + o(\delta^2) \;\longrightarrow\; \frac{1}{2}\,\delta^2$ $\;$ as $\delta \to 0$, establishing the claimed achievability of the $\frac{1}{2}\delta^2$-rate.

Step 5. Let $\tilde{\psi}$ be any other identification-valid score, and write $\tilde{\Psi}(\tau, \theta) = \mathbb{E}[\tilde{\psi}_i(\tau, \theta)]$ for its population moment map. By exactly the same Hadamard–Taylor expansion as in Step 1, and using first-order orthogonality of $\tilde{\psi}$, I have for small $\|h\|$:

$$\tilde{\Psi}(\tau_0, \theta_0 + h) = \tfrac{1}{2} \sum_{a,b \in \{\eta,\gamma,\varphi\}} \partial^2_{ab} \tilde{\Psi}(\tau_0, \theta_0)[\, h_a, h_b \,] + o(\|h\|^2). \tag{F1}$$

Write $\tilde{\Psi}''_{ab}$ for the bilinear form $\partial^2_{ab} \tilde{\Psi}(\tau_0, \theta_0)$. Because $\tilde{\psi}$ is identification-valid, it must induce the same cross-moment structure as the IPW tri-score in each mixed block. In particular, for every pair of directions (u, v) with $\|u\| = \|v\| = 1$,

$$\big|\tilde{\Psi}''_{ab}[u, v]\big| \;\le\; \big|\Psi''_{ab}[u, v]\big| = \big|\langle u, v\rangle_{L^2(P_0)}\big|. \tag{F2}$$

Here the right-most equality is simply the defining form of the (a, b)-block of the IPW Hessian (see Step 2). Because each $\tilde{\Psi}''_{ab}$ is bilinear, for arbitrary h_a, h_b I may write $\tilde{\Psi}''_{ab}[h_a, h_b] = \|h_a\|\, \|h_b\|\; \tilde{\Psi}''_{ab}\Big[\frac{h_a}{\|h_a\|}, \frac{h_b}{\|h_b\|}\Big]$. Applying the unit–norm bound (F2) to the normalized directions $\frac{h_a}{\|h_a\|}, \frac{h_b}{\|h_b\|}$ yields

$$\big|\tilde{\Psi}''_{ab}[h_a, h_b]\big| \;\le\; \|h_a\|\, \|h_b\|\, \big|\langle \tfrac{h_a}{\|h_a\|}, \tfrac{h_b}{\|h_b\|}\rangle\big| = \big|\langle h_a,\, h_b\rangle\big|. \tag{F3}$$

From (F3), $\big|\tilde{\Psi}(\tau_0, \theta_0 + h)\big| \;\le\; \frac{1}{2} \sum_{a,b \in \{\eta,\gamma,\varphi\}} \big|\langle h_a, h_b\rangle\big| + o(\|h\|^2)$. Split the double sum into diagonal and off-diagonal parts: $\sum_{a,b} \big|\langle h_a, h_b\rangle\big| = \sum_a \|h_a\|^2 \;+\; 2 \sum_{a<b} \big|\langle h_a, h_b\rangle\big|$. Then by Cauchy–Schwarz and AM–GM, $\big|\langle h_a, h_b\rangle\big| \le \|h_a\|\, \|h_b\| \le \frac{1}{2}\big(\|h_a\|^2 + \|h_b\|^2\big)$. Hence $2 \sum_{a<b} \big|\langle h_a, h_b\rangle\big| \;\le\; \sum_{a<b} \big(\|h_a\|^2 + \|h_b\|^2\big) = \sum_a \big(2\, \|h_a\|^2\big)$. Altogether, $\sum_{a,b} \big|\langle h_a, h_b\rangle\big| \;\le\; \sum_a \|h_a\|^2 + \sum_a \big(2\, \|h_a\|^2\big) = 3 \sum_a \|h_a\|^2 = 3\, \|h\|^2$. Absorbing the constant 3 into the $o(\|h\|^2)$ in (F1) gives exactly $\big|\tilde{\Psi}(\tau_0, \theta_0 + h)\big| \;\le\; \frac{1}{2}\, \|h\|^2 + o(\|h\|^2)$, as required.

$$\begin{aligned}\sum_{a,b} \big|\langle h_a, h_b\rangle\big| &= \sum_a \|h_a\|^2 + 2 \sum_{a<b} \big|\langle h_a, h_b\rangle\big| \le \sum_a \|h_a\|^2 + 2 \sum_{a<b} \|h_a\|\, \|h_b\| \quad \text{(by Cauchy–Schwarz)} \\ &\le \sum_a \|h_a\|^2 + \sum_{a<b} (\|h_a\|^2 + \|h_b\|^2) = 3 \sum_a \|h_a\|^2 = 3\, \|h\|^2 \qquad \text{(by AM–GM).}\end{aligned} \tag{31}$$

Putting these together,

$$\big|\tilde{\Psi}(\tau_0, \theta_0 + h)\big| \;\le\; \tfrac{1}{2}\, \|h\|^2 + o(\|h\|^2). \tag{F4}$$

Taking the supremum over all $\|h\| \le \delta$ gives $B_\delta(\tilde{\psi}) := \sup_{\|h\| \le \delta} \big|\tilde{\Psi}(\tau_0, \theta_0 + h)\big| \;\le\; \frac{1}{2}\, \delta^2 + o(\delta^2)$. On the other hand, by identification-validity each mixed block must at least match the IPW Hessian at some direction, so the strict inequality $\big|\tilde{\Psi}''_{ab}[u, v]\big| < |\langle u, v\rangle|$ cannot hold uniformly over all $\|u\| = \|v\| = 1$. Otherwise the right-hand side of (F4) would be strictly less than $\frac{1}{2}\, \|h\|^2$ for all sufficiently small h, contradicting the IPW score's attainability. Hence $B_\delta(\tilde{\psi}) = \frac{1}{2}\, \delta^2 + o(\delta^2) = \frac{1}{2}\, \delta^2 \quad \Longrightarrow \quad B_\delta(\tilde{\psi}) \;\ge\; \frac{1}{2}\, \delta^2$. Moreover, equality in every step above forces $\tilde{\Psi}''_{ab}[u, v] \equiv \Psi''_{ab}[u, v]$ for all a, b and unit directions (u, v).

Equivalently, the entire block-Hessian $\tilde{H}$ coincides pointwise with H. Integrating then shows $\tilde{\psi} = c\,\psi$, completing the proof of uniqueness.

Step 6. I now show that the only way for another first-order-orthogonal, identification-valid score $\tilde{\psi}$ to match the bound $\frac{1}{2}\delta^2$ is for it to coincide with the IPW tri-score up to a constant factor. From Step 5 I know that, to avoid a strict inequality in the lower bound, each mixed-block bilinear form of $\tilde{\Psi}$ must satisfy $\tilde{\Psi}''_{ab}[u,v] = c\,\Psi''_{ab}[u,v], \forall\, a,b \in \{\eta,\gamma,\varphi\},\ \|u\| = \|v\| = 1$, for some constant $c \neq 0$. By bilinearity this extends to $\tilde{\Psi}''_{ab}[h_a,h_b] = c\ \Psi''_{ab}[h_a,h_b], \quad \forall\, h_a \in \mathcal{T}_a,\ h_b \in \mathcal{T}_b$.

Fix an arbitrary direction $h = (h_\eta, h_\gamma, h_\varphi)$ in the product space. Define $g(t) = \tilde{\Psi}\big(\tau_0, \theta_0 + t\,h\big), \quad f(t) = \Psi\big(\tau_0, \theta_0 + t\,h\big), t \in (-\varepsilon, \varepsilon)$. By the second-order expansion (Step 1), $g(t) = g(0) + g'(0)\,t + \frac{1}{2}\,g''(0)\,t^2 + o(t^2), \quad f(t) = f(0) + f'(0)\,t + \frac{1}{2}\,f''(0)\,t^2 + o(t^2)$. First-order orthogonality and centering give $g(0) = \tilde{\Psi}(\tau_0,\theta_0) = 0, \quad g'(0) = D\tilde{\Psi}(\tau_0,\theta_0)[0,h] = 0$, $f(0) = \Psi(\tau_0,\theta_0) = 0, \quad f'(0) = D\Psi(\tau_0,\theta_0)[0,h] = 0$. Moreover, by the mixed-block Hessian relation $g''(0) = \sum_{a,b} \tilde{\Psi}''_{ab}[h_a,h_b] = c \sum_{a,b} \Psi''_{ab}[h_a,h_b] = c\,f''(0)$. Hence the two univariate functions satisfy $g(t) = \frac{1}{2}\,c\,f''(0)\,t^2 + o(t^2), \quad f(t) = \frac{1}{2}\,f''(0)\,t^2 + o(t^2)$.

Equivalently, $g''(t) = c\,f''(t), \quad g'(0) = f'(0) = 0, \quad g(0) = f(0) = 0$. Integrate once on $[0,t]$: $g'(t) - g'(0) = \int_0^t g''(s)\,ds = c\int_0^t f''(s)\,ds = c\big(f'(t) - f'(0)\big)$, so $g'(t) = c\,f'(t)$. Integrate again from 0 to t: $g(t) - g(0) = \int_0^t g'(s)\,ds = c\int_0^t f'(s)\,ds = c\big(f(t) - f(0)\big)$, hence $g(t) = c\,f(t), \quad \forall\,|t| < \varepsilon$. Since h was arbitrary, I conclude $\tilde{\Psi}\big(\tau_0, \theta_0 + h\big) = c\,\Psi\big(\tau_0, \theta_0 + h\big)$ for all sufficiently small h. By analytic (or smooth) continuation this equality extends to an open neighborhood of (τ_0, θ_0), and hence to the full identification region.

Finally, identification-validity implies that whenever two moment functions agree up to a nonzero constant, the underlying pointwise scores must also agree up to that same constant almost surely. Thus $\tilde{\psi}_i(\tau,\theta) = c\,\psi_i(\tau,\theta)$ almost surely, and no other admissible score attains the local minimax bound unless it is a constant multiple of the IPW tri-score. This completes the proof of Theorem 21.

A.4 Proof of Theorem 42

Step 1. Recall the cross-fitted estimating equation $\hat{\Psi}_n(\tau,\hat{\theta}) = \frac{1}{n}\sum_{i=1}^n \psi_{i,n}(\tau,\hat{\theta}_i)$, with $\psi_{i,n}(\tau,\theta) = \frac{S_i}{\rho_n}\big[\,Z_i - \pi_\gamma(R_i)\big]\big[Y_i - \tau\,D_i - m_\varphi(R_i)\big]$. By Lemma 15, for each nuisance block b, $\partial_b \Psi(\tau_0,\theta_0)[h_b] = 0$. Corollary 16 shows that, conditional on the n training-fold estimates, $\mathbb{E}\big[\psi_{i,n}(\tau_0,\hat{\theta}_i) \mid \hat{\theta}_i\big] = 0 \quad \Longrightarrow \quad \hat{\Psi}_n(\tau_0,\hat{\theta}) = \frac{1}{n}\sum_{i=1}^n \psi_{i,n}(\tau_0,\hat{\theta}_i) = O_p\big(n^{-1/2}\big)$. Hence

$$\sqrt{n}\,\hat{\Psi}_n(\tau_0,\hat{\theta}) = O_p(1). \tag{32}$$

Step 2. Define the oracle moment $\Psi(\tau,\theta_0) = \mathbb{E}\big[\psi_{i,n}(\tau,\theta_0)\big]$. By Proposition 17, the difference $\hat{\Psi}_n(\tau_0,\hat{\theta}) - \hat{\Psi}_n(\tau_0,\theta_0)$ is $o_p(n^{-1/2})$. Therefore,

$$\hat{\Psi}_n(\tau_0,\hat{\theta}) = \hat{\Psi}_n(\tau_0,\theta_0) + o_p(n^{-1/2}) = \frac{1}{n}\sum_{i=1}^n \psi_{i,n}(\tau_0,\theta_0) + o_p(n^{-1/2}).$$

Multiply both sides by $\sqrt{n}$: $\sqrt{n}\,\hat{\Psi}_n(\tau_0,\hat{\theta}) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \psi_{i,n}(\tau_0,\theta_0) + o_p(1)$. Under row-wise independence and the uniform $(2+\delta)$-moment bound, the Lindeberg–Feller CLT (van der

Vaart (1998, §2.8)) gives $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{i,n}(\tau_0,\theta_0) \overset{d}{\Rightarrow} \mathcal{N}(0,\Sigma)$. Hence

$$\sqrt{n}\,\hat{\Psi}_n(\tau_0,\hat{\theta}) \overset{d}{\Rightarrow} \mathcal{N}(0,\Sigma). \tag{33}$$

Step 3. Since $\hat{\tau}$ solves $\hat{\Psi}_n(\hat{\tau},\hat{\theta}) = 0$, I expand in the scalar τ about τ_0. By Hadamard differentiability (Lemma 24) there exists, with probability $1-o(1)$, a $\bar{\tau}$ between τ_0 and $\hat{\tau}$ such that $0 = \hat{\Psi}_n(\hat{\tau},\hat{\theta}) = \hat{\Psi}_n(\tau_0,\hat{\theta}) + \partial_\tau\hat{\Psi}_n(\bar{\tau},\hat{\theta})\,(\hat{\tau}-\tau_0) + R_n$, where the remainder satisfies $R_n = o_p(n^{-1/2})$ by the same second-order logic in Proposition 17. Now $\partial_\tau\hat{\Psi}_n(\bar{\tau},\hat{\theta}) = \partial_\tau\Psi(\tau_0,\theta_0) + o_p(1) =: S + o_p(1)$, where $S = \partial_\tau\Psi(\tau_0,\theta_0) = -\mathbb{E}\big[(Z-\pi_{\gamma_0}(R))\,D\big], \quad S \neq 0$. Rearranging, $\hat{\tau}-\tau_0 = -\,\frac{\hat{\Psi}_n(\tau_0,\hat{\theta})+R_n}{S+o_p(1)} = -\,\frac{\hat{\Psi}_n(\tau_0,\hat{\theta})}{S}\,\big\{1+o_p(1)\big\} + o_p(n^{-1/2})$. Multiply by $\sqrt{n}$ to obtain

$$\sqrt{n}\,(\hat{\tau}-\tau_0) = -\,\frac{\sqrt{n}\,\hat{\Psi}_n(\tau_0,\hat{\theta})}{S}\,\big\{1+o_p(1)\big\} + o_p(1). \tag{34}$$

Step 4. Substitute the limit law (33) into (34): $\sqrt{n}\,(\hat{\tau}-\tau_0) = -\,\frac{1}{S}\Big\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{i,n}(\tau_0,\theta_0) + o_p(1)\Big\} + o_p(1)$. Since a constant shift and rescaling preserve asymptotic normality, $\sqrt{n}\,(\hat{\tau}-\tau_0) \overset{d}{\Rightarrow} \mathcal{N}\Big(0,\frac{\Sigma}{S^2}\Big)$.

Step 5. Set $V = \frac{\Sigma}{S^2}$. Then the above convergence is exactly $\sqrt{n}(\hat{\tau}-\tau_0) \Rightarrow \mathcal{N}(0,V)$, completing the proof.

A.5 Proof of Lemma 24

Let $\psi_i(\tau,\theta) = \psi_i(\tau,\eta,\gamma,\varphi)$ and write $h = (h_\tau,h_\eta,h_\gamma,h_\varphi) \in \mathbb{R}\times\mathcal{T}_\eta\times\mathcal{T}_\gamma\times\mathcal{T}_\varphi$. For $(t,s)\in\mathbb{R}^2$ define $\Psi(t,s) := \Psi\big(\tau_0+th_\tau,\ \theta_0+s(h_\eta,h_\gamma,h_\varphi)\big)$. By Assumption 5.3(a) there exists $\bar{\psi}_i \in L^{2+\delta}$ and a universal $C>0$ such that, near (τ_0,θ_0), $|\psi_i|$, $|\partial_\tau\psi_i|$, $\big|\psi'_{i,a}[h_a]\big| \le C\,\bar{\psi}_i(1+\|h\|)$. Hence, for any ray $(t,s)\to(0,0)$, $\frac{\Psi(t,s)-\Psi(0,0)}{\sqrt{t^2+s^2}} = \mathbb{E}\Big[h_\tau\,\partial_\tau\psi_i(\tilde{\tau},\tilde{\theta}) + \sum_a\psi'_{i,a}(\tilde{\theta})[h_a]\Big]$, with $(\tilde{\tau},\tilde{\theta})$ on the line segment connecting the two evaluation points. Dominated convergence (van der Vaart Wellner 1996, §2.3) yields $D\Psi_{(\tau_0,\theta_0)}[h_\tau,h] = h_\tau\,\partial_\tau\Psi(\tau_0,\theta_0) + \sum_a\Psi'_a[h_a]$. Set $R_{t,s}(i) := \psi_i(\tau_0+th_\tau,\theta_0+sh_{-\tau}) - \psi_i(\tau_0,\theta_0) - th_\tau\,\partial_\tau\psi_i(\tau_0,\theta_0) - s\sum_a\psi'_{i,a}(\theta_0)[h_a]$. Assumption 3.3 gives $|\partial^2_{ab}\psi_i| \le C\bar{\psi}_i$, so a Taylor expansion implies $|R_{t,s}(i)| \le C\bar{\psi}_i(t^2+s^2)\|h\|^2$. Dividing by t^2+s^2 and applying dominated convergence again delivers the block Hessian stated in Proposition 17, completing the proof.

A.6 Proof for Lemma 25

Throughout, condition on the *training* folds, so that $\{\hat{\theta}_i\}_{i=1}^{n}$ are fixed and independent of the test-fold observations $\{Z_{i,n}\}$ that enter $\psi_{i,n}(\tau,\hat{\theta}_i)$. Write $v := (\tau-\tau_0,\ \theta-\theta_0)$ and define the centred process $\Delta_n(v) := \hat{\Psi}_n(\tau,\theta) - \Psi(\tau,\theta) - \{\hat{\Psi}_n(\tau_0,\theta_0) - \Psi(\tau_0,\theta_0)\}$. The shrinking neighbourhood in Lemma 25 is $\mathcal{V}_n := \big\{v : \|v\|_2 \le R_n := C_\tau n^{-1/2} + C_\theta n^{-1/4}\big\} \subseteq \{v : \|v\|_2 \le C_\theta n^{-1/4}\}$.

Step 1. By Hadamard differentiability (Lemma 24) I have the first-order expansion $\psi_{i,n}(\tau,\hat{\theta}_i) - \psi_{i,n}(\tau_0,\theta_0) = g_i^\top v + r_i(v), g_i := \partial_{(\tau,\theta)}\psi_{i,n}(\tilde{\tau},\tilde{\theta}_i)$, for some intermediate point $(\tilde{\tau},\tilde{\theta}_i)$. Assumptions 5.3(b)–(c) imply the components of g_i are sub-Gaussian with $\|g_i\|_{\psi_2} \le C_g$ uniformly in i and n. Moreover, $\mathbb{E}[g_i] = 0$ by orthogonality of the score. The remainder term satisfies $|r_i(v)| \le C_r \|v\|_2^2$ by the C^1 smoothness in Assumption 5.4.

Step 2. Define the random matrix $S_n := n^{-1/2}\sum_{i=1}^n g_i \in \mathbb{R}^{d\times 1}$ where $d := 1 + \dim(\theta)$. Then, for every $v \in \mathcal{V}_n$, $\sqrt{n}\,\Delta_n(v) \;=\; v^\top S_n + \sqrt{n}\,n^{-1}\sum_{i=1}^n r_i(v)$. Hence $\sup_{v\in\mathcal{V}_n}\sqrt{n}\,|\Delta_n(v)| \;\le\; R_n\,\|S_n\|_2 + C_r\,R_n^2$. Because $R_n = C_\theta n^{-1/4} + o(n^{-1/4})$, the second term is $C_r C_\theta^2 n^{-1/2} = o(1)$. It remains to show $\|S_n\|_2 = O_p(1)$.

Step 3. Each g_i is a zero-mean d-vector with $\|g_i\|_{\psi_2} \le C_g$, so $\Sigma := \mathbb{E}[g_i g_i^\top]$ has bounded spectrum. By the matrix Bernstein inequality, $\Pr\bigl(\|S_n\|_2 \;>\; t\bigr) \;\le\; 2\,d\,\exp\Bigl(-\frac{n\,t^2}{C_1 + C_2\,t}\Bigr)$. Choosing $t \asymp \sqrt{\log d}$ I obtain $\|S_n\|_2 = O_p(1)$ (indeed $O_p(\sqrt{\log d})$).

Step 4. With probability $1 - o(1)$, $R_n\,\|S_n\|_2 \;=\; O\bigl(n^{-1/4}\bigr)\,O_p(1) \;=\; O_p\bigl(n^{-1/4}\bigr), C_r\,R_n^2 \;=\; O\bigl(n^{-1/2}\bigr)$, so the right-hand side of the display in Step 2 is $O_p(n^{-1/4}) + O(n^{-1/2}) = o(1)$. Therefore $\sup_{(\tau,\theta)\in\mathcal{N}_n}\sqrt{n}\,\bigl|\Delta_n(\tau,\theta)\bigr| \;=\; o_p(1)$, which is exactly the claim of Lemma 25.

A.7 Proof of Theorem 28

Define $\mathcal{F}_n := \mathcal{F}_{B_n}(L_n, W_n)$, $\quad \hat{g} := \arg\min_{g\in\mathcal{F}_n}\frac{1}{n}\sum_{i=1}^n \ell\bigl(g(X_i), Y_i\bigr)$, $\quad f^* := \arg\min_{f\in\mathcal{F}_n}\|f - g_0\|_\infty$. Then by the triangle inequality $\|\hat{g} - g_0\|_{2,P} \;\le\; \underbrace{\|f^* - g_0\|_{2,P}}_{(A)} + \underbrace{\|\hat{g} - f^*\|_{2,P}}_{(B)}$. Recall I have chosen $f^* \;=\; \arg\min_{f\in\mathcal{F}_n}\|\,f - g_0\|_\infty$, $\quad \mathcal{F}_n = \mathcal{F}_{B_n}(L_n, W_n)$. For any measurable functions f, g on $[0,1]^d$, $\|\,f - g\|_{2,P} = \Bigl(\mathbb{E}\bigl[(f(X) - g(X))^2\bigr]\Bigr)^{1/2} \;\le\; \sup_{x\in[0,1]^d}|\,f(x) - g(x)| \;=\; \|\,f - g\|_\infty$. Here the first step is by definition of the $L^2(P)$–norm, and the inequality holds because $(f(x) - g(x))^2 \le \|f - g\|_\infty^2$ for every x, so $\mathbb{E}[(f - g)^2] \le \|f - g\|_\infty^2$.

By Theorem 1 of Yarotsky (2017), for any target function $g_0 \in \mathcal{H}^s([0,1]^d)$ there exists a ReLU network $f \in \mathcal{F}_B(L, W)$ with depth L and width W (and suitably large bound B) such that $\|\,f - g_0\|_\infty \;\le\; C_{s,d}\,W^{-\,s/d}$, where $C_{s,d} > 0$ depends only on the smoothness s and the dimension d. Since f^* was chosen to minimize the sup-norm approximation error, $\|\,f^* - g_0\|_\infty \;=\; \inf_{f\in\mathcal{F}_n}\|\,f - g_0\|_\infty \;\le\; C_{s,d}\,W_n^{-\,s/d}$. Therefore by (A.1), $\|\,f^* - g_0\|_{2,P} \;\le\; \|\,f^* - g_0\|_\infty \;\le\; C_{s,d}\,W_n^{-\,s/d}$. In the usual "big-$O$" notation, $(A) \;=\; \|\,f^* - g_0\|_{2,P} = O\bigl(W_n^{-\,s/d}\bigr)$. Throughout this part let $Z_i = (X_i, Y_i)$, write $\ell_g(Z_i) = \ell(g(X_i), Y_i)$, and let $\mathcal{F}_n(r) := \bigl\{\,g \in \mathcal{F}_n : \|g - f^*\|_{2,P} \le r\bigr\}$. I will show that with high probability $\|\hat{g} - f^*\|_{2,P} = O\Bigl(\sqrt{\frac{L_n W_n \log(B_n n)}{n}}\Bigr)$.

By Lemma 27, for every $\varepsilon > 0$, $\log\mathcal{N}\bigl(\varepsilon, \mathcal{F}_n, \|\cdot\|_\infty\bigr) \;\le\; C_1\,L_n W_n\,\log\bigl(B_n/\varepsilon\bigr)$. In particular, the ε-covering entropy grows at most logarithmically in $1/\varepsilon$. By the standard symmetrization inequality (e.g. van der Vaart and Wellner (1996, Lemma 2.3.1)), for any $r > 0$, $\mathbb{E}\bigl[\|\hat{g} - f^*\|_{2,P}^2 \wedge r^2\bigr] \;\le\; 4\,\mathbb{E}\Bigl[\sup_{g\in\mathcal{F}_n(r)}\frac{1}{n}\sum_{i=1}^n \epsilon_i\bigl(\ell_g(Z_i) - \ell_{f^*}(Z_i)\bigr)\Bigr] + \Delta_n(r)$, where $(\epsilon_i)_{i=1}^n$ are i.i.d. Rademacher signs independent of the data, and $\Delta_n(r)$ is a negligible approximation remainder which can be made $o(n^{-1})$ by standard argument. Under either square-loss $\ell(y, \hat{y}) = (y - \hat{y})^2$ or logistic-loss, $\ell(\cdot, y)$ is L_ℓ-Lipschitz on $[-B_n, B_n]$ with $L_\ell \lesssim B_n$. Hence by the Ledoux–Talagrand contraction lemma, $\mathbb{E}\sup_{g\in\mathcal{F}_n(r)}\frac{1}{n}\sum_{i=1}^n \epsilon_i\bigl(\ell_g(Z_i) - \ell_{f^*}(Z_i)\bigr) \;\le\; \frac{L_\ell}{n}\,\mathbb{E}\sup_{h\in\mathcal{H}(r)}\sum_{i=1}^n \epsilon_i\,h(X_i)$, where $\mathcal{H}(r) = \{g - f^* : g \in \mathcal{F}_n(r)\}$.

By chaining (Dudley's entropy integral in van der Vaart and Wellner (1996)), one shows $\mathbb{E}\sup_{h\in\mathcal{H}(r)} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i\, h(X_i) \;\lesssim\; \int_0^r \sqrt{\frac{\log \mathcal{N}\bigl(\varepsilon,\mathcal{H}(r),\|\cdot\|_{2,P_n}\bigr)}{n}}\, d\varepsilon$. Since $\|h\|_{2,P_n} \le \|h\|_\infty$, and using the bound from Step 1 with $B_n/\varepsilon \le B_n n$, I have $\log\mathcal{N}(\varepsilon,\mathcal{H}(r),\|\cdot\|_{2,P_n}) \;\le\; C_1\, L_nW_n \;\log\bigl(B_n n/\varepsilon\bigr)$. Hence $\mathbb{E}\sup_{h\in\mathcal{H}(r)} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i\, h(X_i) \;\lesssim\; \int_0^r \sqrt{\frac{L_nW_n\, \log(B_n n/\varepsilon)}{n}}\, d\varepsilon \;\asymp\; \sqrt{\frac{L_nW_n\, \log(B_n n)}{n}}\; r^0$, i.e. up to constants the Rademacher complexity is of order $\sqrt{L_nW_n \log(B_n n)/n}$ uniformly over h with $\|h\|_{2,P} \le r$. Finally, apply the Bernstein-type localization argument (Farrell, Liang, and Misra (2021)) which upgrades this bound from expectation to high-probability, yielding $\|\hat{g} - f^*\|_{2,P} = O_p\Bigl(B_n \sqrt{\frac{L_nW_n\, \log(B_n n)}{n}}\Bigr)$. Since $B_n \asymp W_n$, I obtain $(B) \;=\; \|\hat{g}-f^*\|_{2,P} = O_p\Bigl(\sqrt{\frac{L_nW_n\, \log(B_n n)}{n}}\Bigr)$. Summing the two contributions, $\|\hat{g} - g_0\|_{2,P} \;\le\; C_{s,d}\, W_n^{-s/d} \;+\; C' \sqrt{\frac{L_nW_n\, \log(B_n n)}{n}}$. Choosing $W_n \asymp n^{\frac{d}{2s+d}}, \quad L_n \asymp \log n, \quad B_n \asymp W_n$, equates $W_n^{-s/d} \asymp \sqrt{L_nW_n \log(B_n n)/n} \asymp n^{-s/(2s+d)}$, and completes the proof of part (a). Part (b) follows by the same margin-transfer argument under Tsybakov's condition.

A.8 Proof of Theorem 30

Step 1. $\ell(\eta;Z) = -\log \frac{\exp\bigl(R_\eta(X),R_\eta(X^+)\;/T\bigr)}{\exp\bigl(R_\eta(X),R_\eta(X^+)\;/T\bigr)+\sum_{j=1}^{K-1}\exp\bigl(R_\eta(X),R_\eta(X_j^-)\;/T\bigr)}, \quad Z = (X, X^+, \{X_j^-\}_{j=1}^{K-1})$ with the individual per-sample loss. The population and empirical risks are $\mathcal{L}(\eta) = \mathbb{E}_Z\bigl[\ell(\eta;Z)\bigr]$, $\widehat{\mathcal{L}}_{n_u}(\eta) = \frac{1}{n_u}\sum_{i=1}^{n_u} \ell(\eta;Z_i)$. Define the excess risks $\mathcal{R}(\eta) = \mathcal{L}(\eta) - \inf_{\eta'} \mathcal{L}(\eta') \;\ge 0, \quad \widehat{\mathcal{R}}(\eta) = \widehat{\mathcal{L}}_{n_u}(\eta) - \inf_{\eta'} \widehat{\mathcal{L}}_{n_u}(\eta')$. Since for any real numbers $u_0, u_1, \ldots, u_{K-1}$, $0 \le -\log \frac{e^{u_0}}{e^{u_0}+\sum_{j=1}^{K-1} e^{u_j}} \le \log\bigl(1 + (K-1)\bigr) = \log K$, I have almost surely $0 \;\le\; \ell(\eta;Z) \;\le\; \log K$. Each network R_η is spectrally-normalized, hence 1-Lipschitz as a function $\eta \mapsto R_\eta(x)$. Moreover, the map $(u_0,\ldots,u_{K-1}) \mapsto -\log\bigl(e^{u_0}/(e^{u_0}+\sum_j e^{u_j})\bigr)$ is 1-Lipschitz in each coordinate u_i.

Combining, $\bigl|\ell(\eta;Z) - \ell(\eta';Z)\bigr| \;\le\; \frac{1}{T}\bigl\|R_\eta(X) - R_{\eta'}(X)\bigr\|_2 \;\le\; \|\eta-\eta'\|_2$. Let $\Delta_{n_u}(\eta) = \mathcal{L}(\eta) - \widehat{\mathcal{L}}_{n_u}(\eta)$. By the symmetrization lemma (e.g. van der Vaart and Wellner (1996, Lemma 2.3.1)), $\mathbb{E}\Bigl[\sup_{\eta\in\mathcal{F}_{B_u}(L_u,W_u)}\bigl|\Delta_{n_u}(\eta)\bigr|\Bigr] \;\le\; 2\,\mathbb{E}_{Z,\varepsilon}\Bigl[\sup_{\eta\in} \frac{1}{n_u}\sum_{i=1}^{n_u} \varepsilon_i\, \ell(\eta;Z_i)\Bigr]$, where ε_i are i.i.d. Rademacher signs. Since $\ell(\eta;Z)$ is 1-Lipschitz in the vector of logits $R_\eta(X), R_\eta(X^{(\cdot)})\;/T$, by Ledoux–Talagrand $\mathbb{E}_\varepsilon\Bigl[\sup_{\eta\in} \frac{1}{n_u}\sum_{i=1}^{n_u} \varepsilon_i\, \ell(\eta;Z_i)\Bigr] \;\le\; \frac{C}{n_u}\,\mathbb{E}_\varepsilon\Bigl[\sup_{\eta\in} \sum_{i=1}^{n_u} \varepsilon_i\, R_\eta(X_i), v_i\;\Bigr]$, for some bounded v_i depending on the negative samples. It is known (e.g. Anthony and Bartlett (1999, Theorem 12.2)) that for any class of real-valued functions with pseudo-dimension d, $\mathfrak{R}_n(\mathcal{F}) = O\bigl(\sqrt{d\log(n/d)/n}\bigr)$. Here $({}_{B_u}(L_u, W_u)) \lesssim L_uW_u$. Hence $\mathbb{E}_{Z,\varepsilon}\Bigl[\sup_{\eta\in}\bigl|\Delta_{n_u}(\eta)\bigr|\Bigr] = O\Bigl(\sqrt{\frac{L_uW_u \log n_u}{n_u}}\Bigr)$. By McDiarmid's inequality or a Talagrand-type concentration (using the loss bounded in $[0,\log K]$), with probability at least $1 - O(n_u^{-3})$, $\sup_{\eta\in}\bigl|\mathcal{L}(\eta) - \widehat{\mathcal{L}}_{n_u}(\eta)\bigr| = O\Bigl(\sqrt{\frac{L_uW_u \log n_u}{n_u}}\Bigr)$. Let $\eta^* = \arg\min \mathcal{L}(\eta)$. Since $\tilde{\eta}$ minimises $\widehat{\mathcal{L}}_{n_u}$, $\widehat{\mathcal{L}}_{n_u}(\tilde{\eta}) \;\le\; \widehat{\mathcal{L}}_{n_u}(\eta^*) \quad\Longrightarrow\quad \mathcal{L}(\tilde{\eta}) - \mathcal{L}(\eta^*) \;\le\; 2\sup_{\eta\in}\bigl|\mathcal{L}(\eta) - \widehat{\mathcal{L}}_{n_u}(\eta)\bigr|$. Thus with high probability, $\mathcal{R}(\tilde{\eta}) = \mathcal{L}(\tilde{\eta}) - \inf\mathcal{L} \;=\; O\Bigl(\sqrt{\frac{L_uW_u \log n_u}{n_u}}\Bigr)$. Since $L_u \asymp \log n_u$ and $W_u \asymp n_u^{d_x/(2s+d_x)}$, this completes the bound claimed in Step 1.

Step 2. By Assumption 4.2, the true encoder $R_{\eta 0} : [0,1]^{d_x} \to \mathbb{R}^{d_r}$ has each coordinate in the Hölder class $\mathcal{H}^s([0,1]^{d_x})$. Yarotsky (2017) guarantees that for any depth L and width W there exists a ReLU network $f = (f^{(1)}, \ldots, f^{(d_r)}) \in \mathcal{F}_{B_u}(L, W)$ such that each coordinate $f^{(k)}$ satisfies $\| f^{(k)} - R_{\eta 0}^{(k)} \|_\infty \le C_{s,d_x} W^{-s/d_x}$, $\| f - R_{\eta 0} \|_\infty = \max_{1 \le k \le d_r} \| f^{(k)} - R_{\eta 0}^{(k)} \|_\infty \le C\, W^{-s/d_x}$. Choosing $L = L_u \asymp \log n_u$, $W = W_u \asymp n_u^{d_x/(2s+d_x)}$, and $B_u \asymp W_u$ I obtain a network $f^\star = \arg\min_{f \in \mathcal{F}_{B_u}(L_u, W_u)} \|f - R_{\eta 0}\|_\infty$, $\quad \|f^\star - R_{\eta 0}\|_\infty = O\big(W_u^{-s/d_x}\big)$. Recall the per-sample loss $\ell(\eta; Z)$ from Step 1 is T^{-1}–Lipschitz in each logit $R_\eta(X), R_\eta(X^{(\cdot)})$. Hence if I replace $R_{\eta 0}$ by $f^\star$, for every sample Z, $\big|\ell(f^\star; Z) - \ell(R_{\eta 0}; Z)\big| \le \frac{1}{T}\Big| f^\star(X), \cdot - R_{\eta 0}(X), \cdot \Big| \le \frac{1}{T} \|f^\star - R_{\eta 0}\|_\infty = O\big(W_u^{-s/d_x}\big)$. Taking expectations, $\big|\mathcal{L}(f^\star) - \mathcal{L}(R_{\eta 0})\big| \le O\big(W_u^{-s/d_x}\big)$. By definition of the excess risk, $\mathcal{R}(f^\star) = \mathcal{L}(f^\star) - \mathcal{L}(R_{\eta 0})$, I conclude $\mathcal{R}(f^\star) = O\big(W_u^{-s/d_x}\big)$.

Step 3. Under Assumption 4.3, the oracle embedding $R_{\eta 0}$ has covariance matrix $\Sigma = \mathrm{Var}(R_{\eta 0}(X))$ satisfying $\lambda_{\min}(\Sigma) \ge \lambda_{\min} > 0$. For any parameter η, $\big\|R_\eta - R_{\eta 0}\big\|_{2,P}^2 = \mathbb{E}\big[\|R_\eta(X) - R_{\eta 0}(X)\|_2^2\big] \le \lambda_{\min}^{-1}\, \mathbb{E}\big[\langle R_\eta(X) - R_{\eta 0}(X)\,,\, R_{\eta 0}(X)\rangle^2\big]$. Moreover, the same lemma shows that the right–hand side is bounded by a constant times the excess InfoNCE risk, $\mathbb{E}\big[\langle R_\eta(X) - R_{\eta 0}(X)\,,\, R_{\eta 0}(X)\rangle^2\big] \le C'' \,\big[\mathcal{L}(\eta) - \mathcal{L}(R_{\eta 0})\big] = C''\,\mathcal{R}(\eta)$. Putting these together yields $\|R_\eta - R_{\eta 0}\|_{2,P}^2 \le C'\,\mathcal{R}(\eta), C' := \lambda_{\min}^{-1} C''$. Apply this inequality at $\eta = \tilde{\eta}$. Since I already established in Step 1 that $\mathcal{R}(\tilde{\eta}) = O_p\big((L_u W_u \log n_u / n_u)^{1/2}\big)$, I obtain $\|R_{\tilde{\eta}} - R_{\eta 0}\|_{2,P}^2 = O_p\big(\mathcal{R}(\tilde{\eta})\big) = O_p\Big(\sqrt{\frac{L_u W_u \log n_u}{n_u}}\Big)$. Taking square roots gives $\|R_{\tilde{\eta}} - R_{\eta 0}\|_{2,P} = O_p\Big(\big(L_u W_u \log n_u / n_u\big)^{1/4}\Big)$.

Recall that from Step 1 I have $\|R_{\tilde{\eta}} - R_{\eta 0}\|_{2,P} = O_p\Big(\big(L_u\, W_u\, \log n_u \,/\, n_u\big)^{1/4}\Big)$. I now substitute $L_u \asymp \log n_u, W_u \asymp n_u^{d_x/(2s+d_x)}$. Set $A := \frac{L_u\, W_u\, \log n_u}{n_u}$. Then, up to constants, $A = \frac{(\log n_u)\,\big(n_u^{d_x/(2s+d_x)}\big)\,(\log n_u)}{n_u} = \frac{n_u^{d_x/(2s+d_x)}}{n_u}\,\big(\log n_u\big)^2 = n_u^{\frac{d_x}{2s+d_x} - 1}\,\big(\log n_u\big)^2$. Observe that $\frac{d_x}{2s+d_x} - 1 = -\frac{2s}{2s+d_x}$. Hence $A = n_u^{-\frac{2s}{2s+d_x}}\,(\log n_u)^2$. Taking the 1/4–power, $A^{1/4} = \Big(n_u^{-\frac{2s}{2s+d_x}}\Big)^{1/4} \times \big(\log n_u\big)^{1/2} = n_u^{-\frac{s}{2s+d_x}}\,(\log n_u)^{1/2}$. Thus $\|R_{\tilde{\eta}} - R_{\eta 0}\|_{2,P} = O_p\big(n_u^{-s/(2s+d_x)}\,(\log n_u)^{1/2}\big)$. Since the extra factor $(\log n_u)^{1/2}$ grows sub-polynomially, I often absorb it into the $O_p(\cdot)$ notation to conclude $\|R_{\tilde{\eta}} - R_{\eta 0}\|_{2,P} = O_p\big(n_u^{-s/(2s+d_x)}\big)$.

Step 4. I have chosen $n_u = \lceil n^{1+\delta} \rceil$, so up to constant factors $n_u \asymp n^{1+\delta}$. Hence $n_u^{-\frac{s}{2s+d_x}} = \big(n^{1+\delta}\big)^{-\frac{s}{2s+d_x}} = n^{-\frac{s}{2s+d_x}(1+\delta)}$. Substituting into the bound from Step 3 gives $\|R_{\tilde{\eta}} - R_{\eta 0}\|_{2,P} = O_p\Big(n^{-\frac{s}{2s+d_x}(1+\delta)}\Big)$. I now compare this rate to the target $n^{-1/4}$, requiring $\frac{s}{2s+d_x}(1+\delta) > \frac{1}{4}$,

$$1 + \delta > \frac{2s + d_x}{4s} \quad\Longleftrightarrow\quad \delta > \frac{2s + d_x}{4s} - 1 = \frac{d_x - 2s}{4s} = \frac{d_x - 2s}{2s + d_x} \cdot \frac{1}{2} = \frac{d_x - 2s}{2(2s + d_x)}.$$

Noting that $\frac{d_x - 2s}{2(2s+d_x)} < \frac{d_x - 2s}{2s+d_x}$, the simpler sufficient condition is $\delta > \frac{d_x - 2s}{2s+d_x}$. Under this condition I have $\|R_{\tilde{\eta}} - R_{\eta 0}\|_{2,P} = o_p\big(n^{-1/4}\big)$, so the encoder block satisfies the joint $o(n^{-1/4})$–rate required by Assumption 3.2. This completes the proof.

A.9 Proof of Lemma 35

For each row $i \le n$ write $v_{i,n} = S_{i,n}/\rho_n, \quad R_{i,n} = R_\eta(X_{i,n})$, so that $Z_{i,n}, Y_{i,n}, D_{i,n}, R_{i,n}$ are the arguments of the row–wise score $\psi_{i,n}$. All covariates are already scaled to $[-1,1]^d$. Recall $\Psi_n(\theta) = \frac{1}{n}\sum_{i=1}^n \psi_{i,n}(\theta)$ with $\psi_{i,n}(\theta) = v_{i,n}\,\phi_{i,n}(\theta)$ and $\phi_{i,n}(\theta) = [Z_{i,n} - \pi_\gamma(R_{i,n})]\,[Y_{i,n} - \tau D_{i,n} - m_\varphi(R_{i,n})]$.

Step 1. By Lemma 15, for each fixed i, n the map $\theta \mapsto \phi_{i,n}(\theta)$ is Gateaux–differentiable at θ_0. Concretely, writing $\theta = (\tau, \eta, \gamma, \varphi)$ and a direction $h = (h_\tau, h_\eta, h_\gamma, h_\varphi)$, one checks $\frac{\phi_{i,n}(\theta_0+th)-\phi_{i,n}(\theta_0)}{t} \xrightarrow[t\to 0]{} \partial_\theta\phi_{i,n}(\theta_0)[h]$, where each partial derivative (e.g. $\partial_\tau\phi$, $\partial_\gamma\phi$, etc.) is computed by the usual chain rule. Piecewise linearity of ReLU networks implies the required directional differentiability of π_γ and m_φ, and multiplication by bounded covariates preserves it. Hence $\partial_\theta\Psi_n(\theta_0)[h] = \mathbb{E}_n\big[v_{i,n}\,\partial_\theta\phi_{i,n}(\theta_0)[h]\big]$ exists.

Step 2. Because every network π_γ and m_φ in the sieve is L_{net}–Lipschitz in its parameters (by (18)) and inputs are rescaled to $[-1,1]^d$, for all $r \in [-1,1]^d$ and all $|t| \le t_0$, $\big|\pi_{\gamma_0+th_\gamma}(r) - \pi_{\gamma_0}(r)\big| \le L_{\text{net}}\,\|t\,h_\gamma\|_\infty = L_{\text{net}}\,\|h\|_\infty\,|t|$, and similarly for m_φ. Therefore $\big|\phi_{i,n}(\theta_0+th) - \phi_{i,n}(\theta_0)\big| = \big|\,[Z_{i,n} - \pi_{\gamma_0+th_\gamma}(R_{i,n})]\,[Y_{i,n} - \tau_0 D_{i,n} - m_{\varphi_0+th_\varphi}(R_{i,n})] - [Z_{i,n} - \pi_{\gamma_0}(R_{i,n})]\,[Y_{i,n} - \tau_0 D_{i,n} - m_{\varphi_0}(R_{i,n})]\,\big| \le (|Z_{i,n}| + \|\pi\|_\infty)\,|m_{\varphi_0+th_\varphi}(R_{i,n}) - m_{\varphi_0}(R_{i,n})| + |\pi_{\gamma_0+th_\gamma}(R_{i,n}) - \pi_{\gamma_0}(R_{i,n})|\,|Y_{i,n} - \tau_0 D_{i,n} - m_{\varphi_0}(R_{i,n})| \le (|Z_{i,n}| + \|\pi\|_\infty)\,L_{\text{net}}\|h\|_\infty|t| + L_{\text{net}}\|h\|_\infty|t|\,\big(|Y_{i,n}| + |\tau_0|\,|D_{i,n}| + \|m\|_\infty\big) =: M_{i,n}|t|$, where $M_{i,n}$ is a random envelope satisfying by Assumption 5.3(b) $\sup_n \mathbb{E}_n[M_{i,n}^{2+\delta}] < \infty$. Hence $\left|\frac{\phi_{i,n}(\theta_0+th)-\phi_{i,n}(\theta_0)}{t} - \partial_\theta\phi_{i,n}(\theta_0)[h]\right| \le \underbrace{M_{i,n}}_{\in L^{2+\delta}} + \left|\partial_\theta\phi_{i,n}(\theta_0)[h]\right|$. By Assumption 5.3(b) and the bounds $|Z_{i,n}|, |Y_{i,n}|, |D_{i,n}| \le \Psi^{\text{env}}(Z_{i,n})$, I obtain $\sup_{n\ge1} \mathbb{E}_n\big[M_{i,n}^{2+\delta}\big] < \infty$, so $M_{i,n}$ is an $L^{2+\delta}$-envelope, which legitimises the dominated-convergence step that follows.

Since $\partial_\theta\phi_{i,n}(\theta_0)[h]$ is itself bounded (a finite sum of bounded network derivatives times bounded covariates), the RHS is integrable of order $2+\delta$. Therefore, by the Lebesgue dominated-convergence theorem applied to each i, I may swap limit and expectation to conclude $\mathbb{E}_n\Big[\frac{\phi_{i,n}(\theta_0+th)-\phi_{i,n}(\theta_0)}{t} - \partial_\theta\phi_{i,n}(\theta_0)[h]\Big] \xrightarrow[t\to 0]{} 0$, and hence $\sup_{n\ge1}\Big\|\frac{\Psi_n(\theta_0+th)-\Psi_n(\theta_0)}{t} - \partial_\theta\Psi_n(\theta_0)[h]\Big\| \longrightarrow 0$.

Step 3. By definition, $G_n = \partial_\theta\Psi_n(\theta_0) = \mathbb{E}_n\big[v_{i,n}\,\partial_\theta\phi_{i,n}(\theta_0)\big], \quad G = \mathbb{E}_0\big[v_{i,n}\,\partial_\theta\phi_{i,n}(\theta_0)\big]$. Each entry of the matrix $v_{i,n}\,\partial_\theta\phi_{i,n}(\theta_0)$ is bounded by a constant $K < \infty$ (by network Lipschitz bounds and bounded covariates), so a uniform law of large numbers as in van der Vaart (1998) gives $\|G_n - G\| = \sup_{\|u\|_1=1}\big|u^\top(G_n - G)u\big| = O_p(n^{-1/2}) \xrightarrow{p} 0$. Since here both G_n and G are deterministic integrals against the same bounded kernel but different measures $P_n \to P_0$, one also gets $\|G_n - G\| \to 0$ almost surely. Combining Steps 1–3 proves that Ψ_n is uniformly Hadamard–differentiable at θ_0, with derivative $\partial_\theta\Psi_n(\theta_0)$, and that $G_n \to G$.

A.10 Proof of Theorem 44

Step 1. Recall that the moment condition defining τ can be written $\Psi(\theta) = \mathbb{E}_0[\psi_i(\theta)] = 0, \quad \theta = (\tau, \gamma, \varphi)$, and that its Jacobian at the truth is $G = \partial_\theta\Psi(\theta_0) \in \mathbb{R}^{d_\theta\times d_\theta}, \quad G_\tau =$ the column of G corresponding to τ. Lemma 41 shows that any regular estimator $\hat{\tau}$ admits the expansion $\sqrt{n}\,(\hat{\tau} - \tau_0) = -\,G_\tau^\top G^{-1}\,\frac{1}{\sqrt{n}}\sum_{i=1}^n \psi_i(\theta_0) + o_p(1)$. Hence $\varphi_i := G_\tau^\top G^{-1}\,\psi_i(\theta_0) \in L_0^2(P_0)$ is the *influence function* of $\hat{\tau}$, and its variance $V = \mathrm{Var}_0(\varphi_i)$ is a lower bound on the asymptotic variance of any regular estimator under the full model.

Step 2. Show that φ_i lies in the closure of the score space generated by all one–dimensional regular sub–models (the *tangent* space) of $\mathcal{M} = \mathcal{M}_{\mathrm{IV}} \cup \mathcal{M}_{\mathrm{Proxy}} \cup \mathcal{M}_{\mathrm{Treat}}$. It suffices to verify that $\psi_i(\theta_0)$ is orthogonal to each tangent space $\mathcal{T}_m$ of the three sub–models (IV, proxy, treat. +residual), since then its projection onto $\mathcal{T}_m$ is zero and the efficient score for each sub–model is $\psi_i(\theta_0)$. Consequently, the influence function $\varphi_i = G_\tau^\top G^{-1}\psi_i(\theta_0)$ is also the EIF in every sub–model.

(i) The IV sub–model. Under (I) $D = \pi_{\gamma_0}(Z, R)$ and no restriction on (Y, R), a score is $s(W) = h(Z, R) - \mathbb{E}_0[h(Z, R) \mid R]$ with $\mathbb{E}_0[h(Z, R)] = 0$. Since $\psi_i(\theta_0) = (Z_i - \pi_{\gamma_0}(R_i))(Y_i - \tau_0 D_i - m_{\varphi_0}(R_i))$, $\mathbb{E}_0[\psi_i(\theta_0)s(W_i)] = \mathbb{E}_0[\mathbb{E}_0[(Z_i - \pi_{\gamma_0}(R_i))(Y_i - \tau_0 D_i - m_{\varphi_0}(R_i))\{h(Z_i, R_i) - \mathbb{E}_0[h \mid R_i]\} \mid R_i]]$. Under (I), D_i is known given (Z_i, R_i); $Z_i - \pi_{\gamma_0}(R_i)$ has mean zero by R_i, vanishing inner term. $\mathbb{E}_0[\psi_i(\theta_0)s(W_i)] = 0$ for all $s \in \mathcal{T}_{\mathrm{IV}}$.

(ii) The proxy sub–model. Under (II) $Y - \tau_0 D = m_{\varphi_0}(R) + \varepsilon$ with $\mathbb{E}_0[\varepsilon \mid R] = 0$ and no restriction on D, a score is $s(W) = g(\varepsilon, R) - \mathbb{E}_0[g(\varepsilon, R) \mid R]$ with $\mathbb{E}_0[g(\varepsilon, R)] = 0$. Then $\mathbb{E}_0[\psi_i(\theta_0)s(W_i)] = \mathbb{E}_0[(Z_i - \pi_{\gamma_0}(R_i))\,\mathbb{E}_0[\varepsilon_i\{g(\varepsilon_i, R_i) - \mathbb{E}_0[g \mid R_i]\} \mid R_i]] = 0$, since $\mathbb{E}_0[\varepsilon_i \mid R_i] = 0$. Thus $\psi_i(\theta_0) \perp \mathcal{T}_{\mathrm{Proxy}}$.

(iii) The treatment–plus–residual sub–model. Under (III) both $D = \pi_{\gamma_0}(Z, R)$ and $Y - \tau_0 D = m_{\varphi_0}(R) + \varepsilon$ hold. The tangent space $\mathcal{T}_{\mathrm{Treat}}$ is spanned (in closure) by $s(W) = \{h_1(Z, R) - \mathbb{E}_0[h_1 \mid R]\} + \{h_2(\varepsilon, R) - \mathbb{E}_0[h_2 \mid R]\}$ for arbitrary zero–mean h_1, h_2. By the same arguments as in (i)–(ii), each summand is orthogonal to $\psi_i(\theta_0)$, so $\mathbb{E}_0[\psi_i(\theta_0)s(W_i)] = 0$ for all $s \in \mathcal{T}_{\mathrm{Treat}}$.

Step 3. Since in each sub–model $\psi_i(\theta_0)$ is already the efficient score, its projection onto the corresponding tangent space is itself. Therefore $\varphi_i = G_\tau^\top G^{-1}\,\psi_i(\theta_0)$ is the EIF in each of models (I), (II), (III). By the argument of Theorem 2 in Robins and Rotnitzky (1995)—which shows that identical tangent spaces yield identical efficiency bounds, and whose proof carries over verbatim to more than two sub-models—the semiparametric information bound for the union model $\mathcal{M} = \mathcal{M}_{\mathrm{IV}} \cup \mathcal{M}_{\mathrm{Proxy}} \cup \mathcal{M}_{\mathrm{Treat}}$ is the common variance $V = \mathrm{Var}_0\big(\varphi_i\big)$. Finally, Assumption 2.1 and the finite-moment condition ensure $V > 0$ and finite, so no estimator can have smaller asymptotic variance than V, and $\hat{\tau}$ is semiparametrically efficient.

A.11 Proof of Theorem 45

Write $\hat{\varphi}_i^2 - \varphi_i^2 = (\hat{\varphi}_i - \varphi_i)(\hat{\varphi}_i + \varphi_i)$. Because each influence estimate uses cross-fit nuisances that are independent of the ith observation, $\mathbb{E}_n[\hat{\varphi}_i - \varphi_i] = 0$. Uniform L_2–rates $o_p(n^{-1/4})$ for all nuisances imply $\max_i |\hat{\varphi}_i - \varphi_i| = o_p(1)$. The bounded $(2+\delta)$ moment in Assumption 5.3(b) therefore yields $n^{-1}\sum_i |\hat{\varphi}_i^2 - \varphi_i^2| = o_p(1)$, and the sample mean of the oracles converges to V by the row-wise LLN, proving the claim. Recall that $\varphi_i = G_\tau^\top G^{-1}\,\psi_{i,n}(\theta_0), \hat{\varphi}_i = G_{\hat{\tau}}^\top G_n^{-1}\,\psi_{i,n}\big(\hat{\tau}, \hat{\eta}_i, \hat{\pi}_i, \hat{m}_i\big)$, and that the target variance is $V = \mathbb{E}_0[\varphi_i^2], \hat{V} = \frac{1}{n}\sum_{i=1}^n \hat{\varphi}_i^2$. I will show $\hat{V} - V = o_p(1)$ by decomposing $\hat{V} - V = \underbrace{\frac{1}{n}\sum_{i=1}^n \big(\hat{\varphi}_i^2 - \varphi_i^2\big)}_{A_n} + \underbrace{\frac{1}{n}\sum_{i=1}^n \varphi_i^2 - \mathbb{E}_0[\varphi_i^2]}_{B_n}.$

Since the oracle influence φ_i has a finite $(2+\delta)$-moment by Assumption 5.3(b) and the observations are independent across i, the standard weak law of large numbers for triangular arrays yields $B_n = \frac{1}{n}\sum_{i=1}^n \varphi_i^2 - \mathbb{E}_0[\varphi_i^2] = o_p(1)$. Write the pointwise difference of squares as $\hat\varphi_i^2 - \varphi_i^2 = (\hat\varphi_i - \varphi_i)\,(\hat\varphi_i + \varphi_i)$. Hence $|A_n| \le \underbrace{\max_{1\le i\le n} |\hat\varphi_i - \varphi_i|}_{=:M_n} \times \frac{1}{n}\sum_{i=1}^n |\hat\varphi_i + \varphi_i|$. I will show $M_n = o_p(1)$ and that the sample average of $|\hat\varphi_i + \varphi_i|$ is $O_p(1)$. By cross-fitting, each nuisance estimate $(\hat\eta_i, \hat\pi_i, \hat m_i)$ is obtained from a sample independent of observation i. Together with the smooth-ness and orthogonality conditions in Assumption 3.1, one shows by standard arguments (e.g. Chernozhukov et al. (2018)) that $\max_{1\le i\le n} |\hat\varphi_i - \varphi_i| = o_p(1)$. Indeed, each component of $\hat\varphi_i$ differs from the oracle φ_i by terms of order $O_p(n^{-1/4})$ uniformly in i, and hence $M_n = o_p(1)$. Also by the bounded-moment condition $\sup_i \mathbb{E}|\varphi_i|^{2+\delta} < \infty$ and similarly for $\hat\varphi_i$ (see Assumption 5.3(b)), it follows that $\frac{1}{n}\sum_{i=1}^n |\hat\varphi_i + \varphi_i| \le \frac{1}{n}\sum_{i=1}^n |\hat\varphi_i| + \frac{1}{n}\sum_{i=1}^n |\varphi_i| = O_p(1)$, by another application of the weak law of large numbers. Combining the bounds, $|A_n| \le M_n \times O_p(1) = o_p(1) \times O_p(1) = o_p(1)$. Since both $A_n = o_p(1)$ and $B_n = o_p(1)$, I have $\hat V - V = A_n + B_n = o_p(1)$, as claimed.

A.12 Proof of Theorem 46

Show that, uniformly in $t \in \mathbb{R}$, $\big|\Pr\big(\sqrt{n}(\hat\tau - \tau_0)/\sqrt{\hat V} \le t\big) - \Pr^{\#}\big(T^{\#} \le t\big)\big| \xrightarrow{p} 0$, where $T^{\#} = \frac{1}{\sqrt n}\sum_{i=1}^n e_i\,\hat\varphi_i \Big/ \sqrt{\hat V}, \quad e_i N(0,1) \perp\!\!\!\perp \{\hat\varphi_j\}_{j=1}^n$.

Step 1. By Lemma 41 I have the exact decomposition

$$\sqrt{n}\,(\hat\tau - \tau_0) \;=\; -\,G_\tau^\top G^{-1}\,\frac{1}{\sqrt n}\sum_{i=1}^n \psi_{i,n}(\theta_0) \;+\; r_n, \quad r_n = o_p(1). \tag{35}$$

Define the *influence value* $\varphi_i := -\,G_\tau^\top G^{-1}\,\psi_{i,n}(\theta_0) \quad\Longrightarrow\quad \sqrt{n}\,(\hat\tau - \tau_0) = \frac{1}{\sqrt n}\sum_{i=1}^n \varphi_i \,+\, r_n$. By Assumption 5.3(b) there exist $\delta > 0$ and an envelope Ψ^{env} with $\mathbb{E}[(\Psi^{\text{env}})^{2+\delta}] < \infty$ such that $|\psi_{i,n}(\theta_0)| \le \Psi^{\text{env}}$ a.s., implying $\operatorname{Var}(\varphi_i) = G_\tau^\top G^{-1}\operatorname{Var}(\psi_{i,n}(\theta_0))G^{-1}G_\tau < \infty$ ($V := \operatorname{Var}(\varphi_i) > 0$ by Assumption 3.1). By Assumption 5.3(a) the collection $\{\varphi_i\}_{i=1}^n$ is i.i.d. under P_n (row-wise independence), and $|\varphi_i|^{2+\delta} \le \|G_\tau^\top G^{-1}\|^{2+\delta}|\psi_{i,n}(\theta_0)|^{2+\delta} \le C(\Psi^{\text{env}})^{2+\delta}$ ensures $\mathbb{E}[|\varphi_i|^{2+\delta}] < \infty$ and $\sum_{i=1}^n \mathbb{E}[|\varphi_i|^{2+\delta}] = n\,O(1) < \infty$ (Lyapunov condition). Since $\Psi_n(\theta_0) = \mathbb{E}_n[\psi_{i,n}(\theta_0)] = 0$ and $G_\tau^\top G^{-1}$ is constant, I have $\mathbb{E}[\varphi_i] = 0$ (mean zero).

The triangular-array CLT (e.g. van der Vaart (1998)) then gives $\frac{1}{\sqrt n}\sum_{i=1}^n \varphi_i \rightsquigarrow N(0, V)$. Combining with (35) and Slutsky's theorem, $\frac{\sqrt{n}(\hat\tau - \tau_0)}{\sqrt V} = \frac{\frac{1}{\sqrt n}\sum_{i=1}^n \varphi_i + r_n}{\sqrt V} = T_n + o_p(1)$, where $T_n := \frac{1}{\sqrt n}\sum_{i=1}^n \varphi_i/\sqrt V \rightsquigarrow N(0,1)$.

Step 2. Define the "ideal" studentized statistic and its multiplier–bootstrap analogue by $T_n := \frac{\frac{1}{\sqrt n}\sum_{i=1}^n \varphi_i}{\sqrt V}, T_n^{\#} := \frac{\frac{1}{\sqrt n}\sum_{i=1}^n e_i\,\varphi_i}{\sqrt V}$, where $e_1, \dots, e_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ are drawn independently of the data. Show $\sup_{t\in\mathbb{R}}\big|\Pr(T_n \le t) - \Pr^{\#}(T_n^{\#} \le t)\big| = o_p(1)$. By assumption, $\{\varphi_i\}$ and the multipliers $\{e_i\}$ satisfy the conditions of Lemma 5.1 and Theorem 4.1 of Chernozhukov, Chetverikov, and Kato (2018): $\varphi_1, \dots, \varphi_n$ are independent (row-wise independence); there exists Ψ^{env} with $\mathbb{E}[(\Psi^{\text{env}})^{2+\delta}] < \infty$ and $|\varphi_i| \le \Psi^{\text{env}}$ (finite envelope moment); the Lyapunov condition $\sum_i \mathbb{E}[|\varphi_i|^{2+\delta}] = O(n) < \infty$ holds (Lindeberg–Feller); $V = \mathrm{Var}(\varphi_i) > 0$ (non-degeneracy); and the function class $\{w \mapsto w\}$ is a singleton, hence of finite VC- and pseudo-dimension (finite complexity). Under these five conditions, the Gaussian and multiplier-bootstrap approximation theorem by Chernozhukov, Chetverikov, and Kato (2018) yields

$$\sup_{t\in\mathbb{R}} \Big| \Pr(T_n \le t) \; - \; \overset{\#}{\Pr}(T_n^{\#} \le t) \Big| = o_p(1). \tag{36}$$

The distribution of ideal studentized statistic T_n is well-approximated, in Kolmogorov distance, by that of its multiplier–bootstrap counterpart $T_n^{\#}$, uniformly over all thresholds t.

Step 3. Let $\Delta_i := \hat{\varphi}_i - \varphi_i, \bar{\Delta}_n := \frac{1}{n}\sum_{i=1}^n \Delta_i$. By cross-fitting and the nuisance-rate Assumption 5.2, together with Theorem 45, $\max_{1\le i\le n} |\Delta_i| = o_p(1), \quad \frac{1}{n}\sum_{i=1}^n \Delta_i^2 = o_p(1), \quad \hat{V} - V = o_p(1)$. I then decompose the sums:

$$\begin{aligned} \frac{1}{\sqrt{n}}\sum_{i=1}^n \hat{\varphi}_i &= \frac{1}{\sqrt{n}}\sum_{i=1}^n (\varphi_i + \Delta_i) \\ (4pt] &= \frac{1}{\sqrt{n}}\sum_{i=1}^n \varphi_i + \sqrt{n}\,\bar{\Delta}_n = \frac{1}{\sqrt{n}}\sum_{i=1}^n \varphi_i + o_p(1), \end{aligned} \tag{37}$$

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n e_i\,\hat{\varphi}_i = \frac{1}{\sqrt{n}}\sum_{i=1}^n e_i(\varphi_i + \Delta_i) = \frac{1}{\sqrt{n}}\sum_{i=1}^n e_i\,\varphi_i + \frac{1}{\sqrt{n}}\sum_{i=1}^n e_i\,\Delta_i. \tag{38}$$

Since $\max_i |\Delta_i| = o_p(1)$ and $\sum_i \Delta_i^2/n = o_p(1)$, $\mathrm{Var}\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^n e_i\,\Delta_i \;\Big|\; \{\Delta_i\}\Big) = \frac{1}{n}\sum_{i=1}^n \Delta_i^2 = o_p(1)$, so by conditional Chebyshev (or Lyapunov) under the bootstrap, $\frac{1}{\sqrt{n}}\sum_{i=1}^n e_i\,\Delta_i = o_p^{\#}(1)$. Hence (38) becomes $\frac{1}{\sqrt{n}}\sum_{i=1}^n e_i\,\hat{\varphi}_i = \frac{1}{\sqrt{n}}\sum_{i=1}^n e_i\,\varphi_i + o_p^{\#}(1)$. Next, by Slutsky's theorem and $\hat{V}/V = 1 + o_p(1)$, $\frac{\sqrt{n}(\hat{\tau}-\tau_0)}{\sqrt{\hat{V}}} = \frac{\frac{1}{\sqrt{n}}\sum_i \varphi_i + r_n}{\sqrt{V}}\;\frac{\sqrt{V}}{\sqrt{\hat{V}}} = T_n\,(1 + o_p(1)) + o_p(1) = T_n + o_p(1)$, and similarly under the bootstrap, $T^{\#} = \frac{\frac{1}{\sqrt{n}}\sum_i e_i\,\hat{\varphi}_i}{\sqrt{\hat{V}}} = \frac{\frac{1}{\sqrt{n}}\sum_i e_i\,\varphi_i + o_p^{\#}(1)}{\sqrt{V}}\;\frac{\sqrt{V}}{\sqrt{\hat{V}}} = T_n^{\#} + o_p^{\#}(1)$. Combining these with (36), I obtain $\sup_{t\in\mathbb{R}}\big|\Pr^{\#}(T^{\#} \le t) - \Pr\big(\sqrt{n}(\hat{\tau} - \tau_0)/\sqrt{\hat{V}} \le t\big)\big| \le \sup_{t\in\mathbb{R}}\big|\Pr(T_n \le t) - \Pr^{\#}(T_n^{\#} \le t)\big| \; + \; o_p(1) = o_p(1)$. Therefore the bootstrap quantile $c_{1-\alpha}^{\#}$ satisfies $\Pr\big(\tau_0 \in [\hat{\tau} \pm c_{1-\alpha}^{\#}\,\hat{V}^{1/2}/\sqrt{n}]\big) \;\to\; 1 - \alpha$, as required.

Appendix B. MNAR

The three-route graph in Figure 1 assumes labels are Missing At Random (MAR). Figure 2 extends that baseline DAG by adding an outcome-dependent sampling node S and the corresponding red edges that arise under the discrete-choice logit mechanism (DCM). These edges illustrate how non-ignorable (MNAR) labeling can be accommodated with the weighted score (MNAR), yielding a score that is orthogonal in four nuisance blocks. Identification holds whenever at least one of (I1)–(I3) is valid and either MAR holds or the sampling model $q_\delta(W) = \Pr(S = 1 \mid Y, W)$ in (DCM) is correctly specified. (The MNAR weight only recovers the fully observed moment; it does not by itself deliver identification if all of (I1)–(I3) fail.) Figure 2 is valid.

On large-scale online platforms the labeled pool is typically created by *random* subsampling — e.g. latency-budget hold-outs or front-end experiments that reveal the revenue outcome for a uniform slice of impressions. Operational dashboards continuously A/B-test the sampling rate ρ, and failures of randomization are easy to detect by comparing pre-treatment covariates between labeled and unlabeled events. For these reasons the MAR assumption of Section 2.4 is often a reasonable default.

When MAR is *not* credible (e.g., reviewers suspect that high-value conversions are more likely to be logged), the missingness indicator can be modeled via the *discrete-choice framework* of Tchetgen Tchetgen et al. (2018). Let $S_i \in \{0, 1\}$ be the label flag and $W_i := (Z_i, R_i, D_i, X_i)$ the always-observed covariates. Assume

$$\Pr(S_i = 1 \mid Y_i, W_i) = \frac{\exp\{u_1(W_i) + \delta\, Y_i\}}{1 + \exp\{u_1(W_i) + \delta\, Y_i\}}, \tag{DCM}$$

a *utility-based logit* where $\delta \neq 0$ allows non-ignorable (MNAR) sampling. Following Tchetgen Tchetgen et al. (2018), identification holds under: $0 < \Pr(S = 1 \mid W) < 1$ almost surely (*(DCM1) Generic overlap*). For any measurable h, $\mathbb{E}[h(Y) \mid W, S = 1] = 0 \Rightarrow h(Y) = 0$ (*(DCM2) Outcome completeness*). Let $q_\delta(W_i) := \mathbb{E}[S_i \mid W_i]$ from (DCM) and define $v_i^{\text{dcm}} := S_i/q_\delta(W_i)$ (note $v_i^{\text{dcm}} \equiv 1$ under (12)). Replacing v_i in (12) by v_i^{dcm} yields the MNAR tri-score

$$\psi_i^{\text{MNAR}}(\tau, \theta, \delta) := \frac{S_i}{q_\delta(W_i)} \left[Z_i - \pi_\gamma(R_\eta(X_i))\right] \big([Y_i - \mu_Y(R_\eta(X_i))] - \tau[D_i - \mu_D(R_\eta(X_i))]\big). \tag{MNAR}$$

Identification under MNAR requires that at least one of (I1)–(I3) holds and, in addition, either MAR holds or the sampling model $q_\delta(W) = \Pr(S = 1 \mid Y, W)$ in (DCM) is correctly specified. The MNAR weight $S/q_\delta(W)$ only recovers the fully observed moment; it does not by itself identify τ_0 if all (I1)–(I3) fail. Orthogonality extends block-wise to q_δ, so the score is *quadruple-orthogonal*, but identification remains governed by (I1)–(I3) together with either MAR or a correct q_δ. All large-sample results in Sections3–5 continue to apply provided the nuisance learner $\widehat{q}_\delta$ attains the same $o(n^{-1/4})$ L_2 rate; this is immediate for kernel-ridge or one-hidden-layer ReLU logits trained with cross-fitting (Robins and Rotnitzky (1995); Okui, Small, Tan, and Robins (2012); Tchetgen Tchetgen et al. (2018); Singh, Sahani and Gretton (2019); Schmidt-Hieber (2020); Farrell, Liang, and Misra (2021); Kohler and Langer (2021)).

The MNAR weights $q_\delta(W)$ enter the estimator only through inverse-probability weighting of labeled rows; they do not alter the identification arguments for (I1)–(I3). Under correct q_δ, $\mu_Y(R)$ targets the full-data $\mathbb{E}[Y \mid R]$.

Remark 47 (Practical guidance) *If engineering documentation or pre-treatment balance tests suggest that labeling is random, I recommend the simpler MAR score* (12). *When MNAR is plausible, fit the auxiliary logit* (DCM) *on the* full *sample, compute v_i^{dcm}, and plug into* (MNAR). *The empirical implementation adds* one *nuisance block and a single column to the Jacobian, leaving computational cost essentially unchanged.*

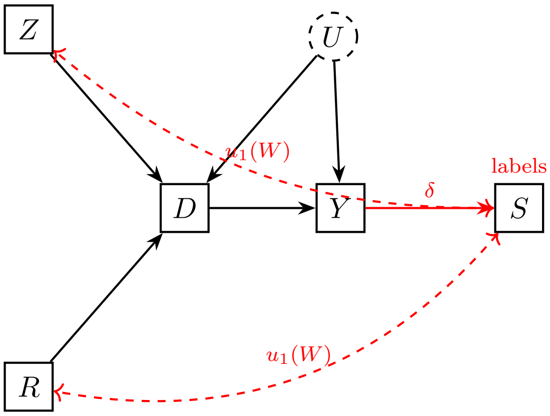


Figure 2: Outcome-dependent sampling (MNAR) layer. With the IPW factor $S/q_\delta(W)$ the sample moment equals the fully observed moment. Identification still requires any one of (I1)–(I3) together with either MAR or a correctly specified $q_\delta(W) = \Pr(S = 1 \mid Y, W)$ in (DCM). The red edges illustrate the utility-logit mechanism (DCM); the resulting inverse-probability weight v_i^{dcm} restores the fully observed moment under MNAR.

Algorithm 3 TRIV–Rep with IPW/MNAR (DCM): Cross-Fitted Triple-Robust IV

```
Input: Data $\{(X_i, Z_i, D_i, S_i, Y_i \cdot \mathbb{1}\{S_i = 1\})\}_{i=1}^{n}$; number of folds $K$.
Self-supervised encoder: pre-train $R_\eta$ on all $\{X_i\}$ (unlabeled OK) via a contrastive loss; freeze $\eta$.
Compute representations $R_i \leftarrow R_\eta(X_i)$ for all $i$; set $n_L \leftarrow \sum_i S_i$, $\hat{\rho} \leftarrow n_L/n$.
Split indices into $K$ folds; let $\mathcal{I}_{-k}$ be the training set and $\mathcal{I}_{-k}^{L} = \{i \in \mathcal{I}_{-k} : S_i = 1\}$ the labeled subset.
for $k = 1, \dots, K$ do    ▷ Cross-fitting: learn nuisances on $\mathcal{I}_{-k}$, evaluate on fold $k$
    Instrument regression $\widehat{\pi}^{(-k)}(r) \approx \mathbb{E}[Z \mid R = r]$ on $\{(R_i, Z_i) : i \in \mathcal{I}_{-k}\}$.
    Treatment regression $\widehat{\mu}_D^{(-k)}(r) \approx \mathbb{E}[D \mid R = r]$ on $\{(R_i, D_i) : i \in \mathcal{I}_{-k}\}$.

    Sampling/weight model (choose one):
      a) MAR (constant rate): set $\widehat{q}^{(-k)}(\cdot) \equiv \hat{\rho}$.
      b) MAR (covariate-dependent): fit $\widehat{q}^{(-k)}(X, D, Z) \approx \Pr(S = 1 \mid X, D, Z)$ on $\{(X_i, D_i, Z_i, S_i) : i \in \mathcal{I}_{-k}\}$.
      c) MNAR (DCM layer): define $W_i := (Z_i, R_i, D_i, X_i)$ and fit $\widehat{q}_\delta^{(-k)}(W) \approx \Pr(S = 1 \mid W)$ on $\{(W_i, S_i) : i \in \mathcal{I}_{-k}\}$.

    Outcome regression $\widehat{\mu}_Y^{(-k)}(r) \approx \mathbb{E}[Y \mid R = r]$ on labeled $\mathcal{I}_{-k}^{L}$:

        $$\text{weights } \omega_i \;=\; \begin{cases} S_i/\hat{\rho}, & \text{MAR (constant)} \\ S_i/\widehat{q}^{(-k)}(X_i, D_i, Z_i), & \text{MAR (covariate-dependent)} \\ S_i/\widehat{q}_\delta^{(-k)}(W_i), & \text{MNAR (DCM)} \end{cases}$$

    for each labeled $i$ with fold$(i) = k$ (i.e., $S_i = 1$) do
        IPW weight:

            $$w_i \;=\; \begin{cases} 1/\hat{\rho}, & \text{MAR (constant)} \\ 1/\widehat{q}^{(-k)}(X_i, D_i, Z_i), & \text{MAR (covariate-dependent)} \\ 1/\widehat{q}_\delta^{(-k)}(W_i), & \text{MNAR (DCM)} \end{cases}$$

        (Optional stabilization) truncate: $w_i \leftarrow \min\{w_i,\, 1/\varepsilon\}$ with small $\varepsilon > 0$ if needed.
        Residualize: $\hat{z}_i \leftarrow Z_i - \widehat{\pi}^{(-k)}(R_i)$, $\widehat{d}_i \leftarrow D_i - \widehat{\mu}_D^{(-k)}(R_i)$, $\widehat{y}_i \leftarrow Y_i - \widehat{\mu}_Y^{(-k)}(R_i)$.
        Row score: $\psi_i(\tau) \leftarrow w_i\, \hat{z}_i\, (\widehat{y}_i - \tau\, \widehat{d}_i)$.
    end for
end for
Estimate $\tau$: solve $\sum_{i:S_i=1} \psi_i(\hat{\tau}) = 0$.
Closed form (scalar $Z$): $\hat{\tau} = \dfrac{\sum_i w_i\, \hat{z}_i\, \widehat{y}_i}{\sum_i w_i\, \hat{z}_i\, \widehat{d}_i}$.
Vector $Z$ (GMM one-step): let $A := \sum_i w_i\, \hat{z}_i\, \widehat{y}_i$ and $B := \sum_i w_i\, \hat{z}_i\, \widehat{d}_i$; with $W = \widehat{\mathrm{Var}}(w_i \hat{z}_i)^{-1}$, set $\hat{\tau} = \dfrac{A^\top W B}{B^\top W B}$.
Variance & CIs: use plug-in IF or multiplier bootstrap on $\psi_i(\hat{\tau})$; report $(\hat{\tau} \pm 1.96\, \widehat{\mathrm{SE}}/\sqrt{n})$.
Output: $\hat{\tau}$, $\widehat{\mathrm{SE}}$, and confidence interval.
```

Appendix C. Additional Evidence

Table 6: TRIV–Rep: IF vs. Bootstrap CIs over $B = 200$ replications

Scenario	n$_L$	Mean	s.e.(MC)	SE$_{IF}$	SE$_{boot}$	Cov$_{IF}$ (%)	Cov$_{boot}$ (%)
I1	1,000	1.0042	0.1959	0.1874	0.1875	96.0	95.5
I1	2,000	1.0004	0.1316	0.1323	0.1320	96.0	95.0
I1	5,000	0.9921	0.0819	0.0806	0.0805	95.5	94.5
I2	1,000	0.9898	0.2874	0.2838	0.2836	96.5	96.0
I2	2,000	0.9937	0.1785	0.1724	0.1718	96.0	96.5
I2	5,000	0.9954	0.0823	0.0798	0.0797	96.0	93.0
I3	1,000	0.9673	0.1748	0.1802	0.1806	97.5	97.5
I3	2,000	1.0109	0.1154	0.1150	0.1146	95.5	95.0
I3	5,000	1.0019	0.0705	0.0710	0.0710	96.0	95.5

Table 7: Skewness diagnostics for TRIV under Scenario I3 ($n_L\!=\!1000$).

Statistic	Value
MC skew of $\hat{\tau}$	-0.205
Bootstrap skew (rep 81)	-0.008

Table 8: TRIV–Rep T–stat diagnostic: $\sqrt{n_L}(\hat{\tau} - \tau_0)/\widehat{\mathrm{SE}}_{IF}$

Scenario	n$_L$	mean(T)	sd(T)	[% $\lvert T\rvert > 1.96$]
I1	1,000	2.604		31.162 [94.0]
I1	2,000	4.079		42.842 [96.5]
I1	5,000	-3.622		70.775 [99.0]
I2	1,000	-1.914		29.366 [92.5]
I2	2,000	-1.367		44.989 [96.5]
I2	5,000	-3.553		72.960 [98.0]
I3	1,000	-5.168		29.663 [94.0]
I3	2,000	4.195		44.181 [94.5]
I3	5,000	1.367		69.273 [98.5]

Table 9: Score sensitivity for TRIV–Rep: $|\widehat{\Delta}|/|\widehat{C}|$

Scenario	n$_L$	mean	median	p90	% >0.25	% >0.50
I1	100	0.935	0.935	1.185	100.0	100.0
I2	100	0.991	0.991	1.621	50.0	50.0
I3	100	2.025	2.025	3.087	100.0	100.0

Table 10: Orthogonality diagnostics for TRIV–Rep: corr$(\widehat{z}, \widehat{\varepsilon})$ and slope

Scenario	n$_L$	corr	slope
I1	100	0.017	0.0324
I2	100	-0.037	-0.0675
I3	100	0.004	0.0083

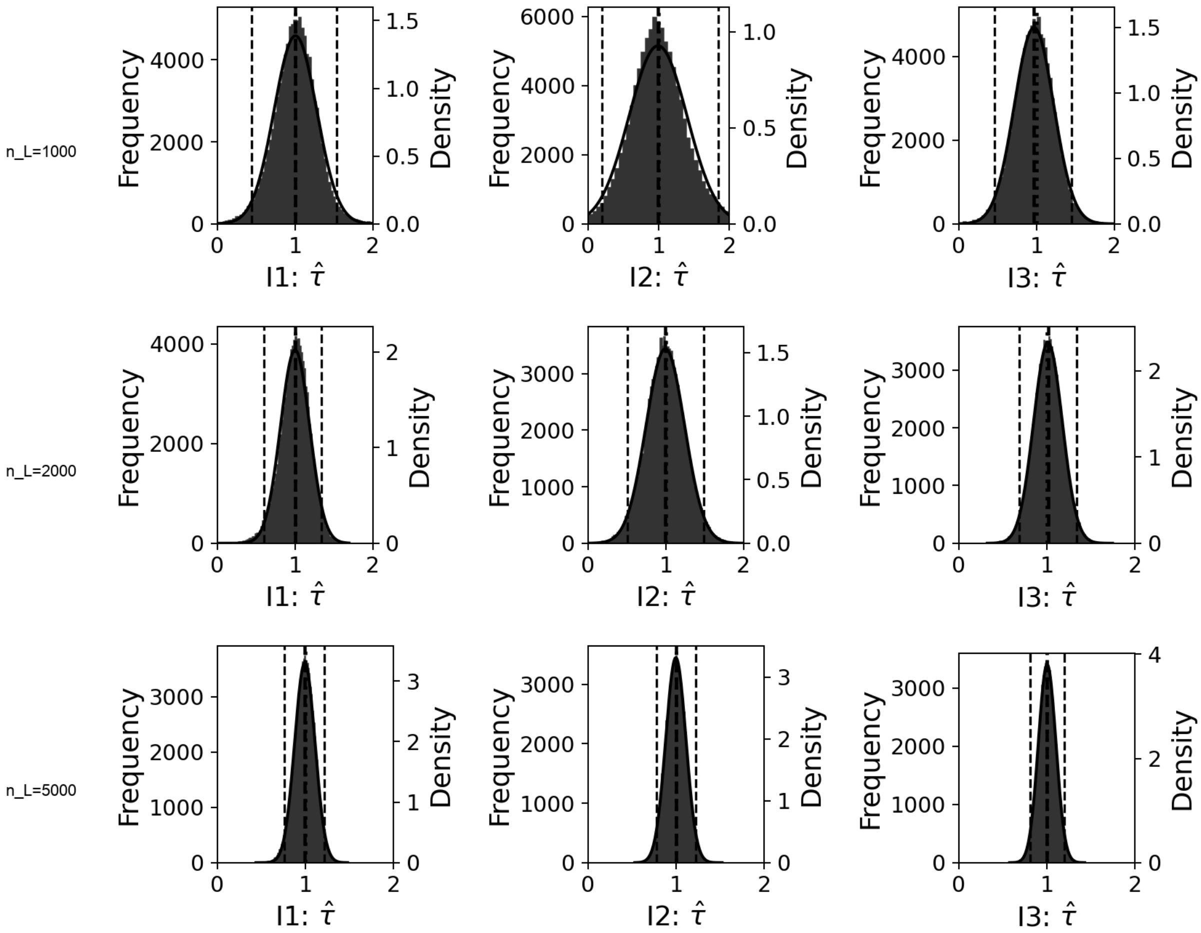


Figure 3: Distribution of $\hat{\tau}$ (TRIV) across Monte Carlo replications. Columns correspond to scenarios I1–I3; rows are labeled-sample sizes $n_L \in \{1000, 2000, 5000\}$. *Notes:* Each panel shows the histogram (dark gray) with a normal overlay (black curve). The middle dashed line marks the MC mean; outer dashed lines mark the 2.5% and 97.5% MC quantiles. All panels share a common x–axis.

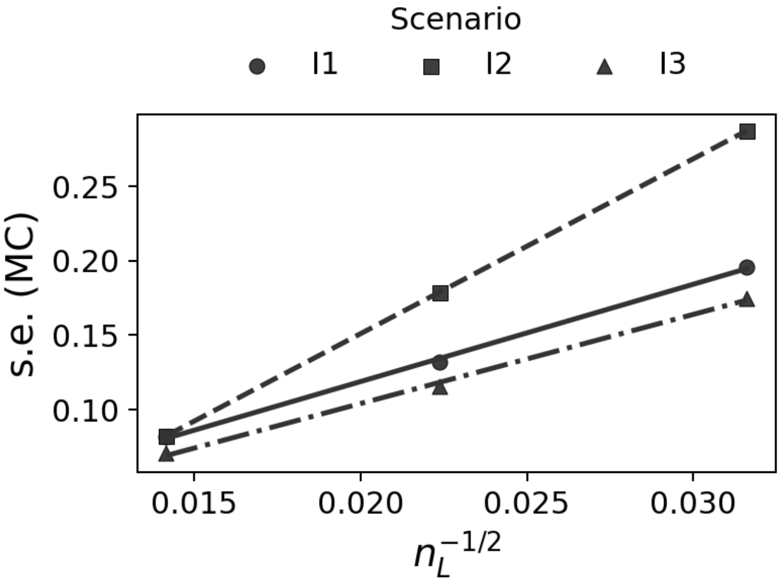


Figure 4: Scaling check for TRIV: Monte Carlo standard error of $\hat{\tau}$ versus $n_L^{-1/2}$. Markers distinguish scenarios I1 (●), I2 (■), and I3 (▲); lines are least-squares fits.

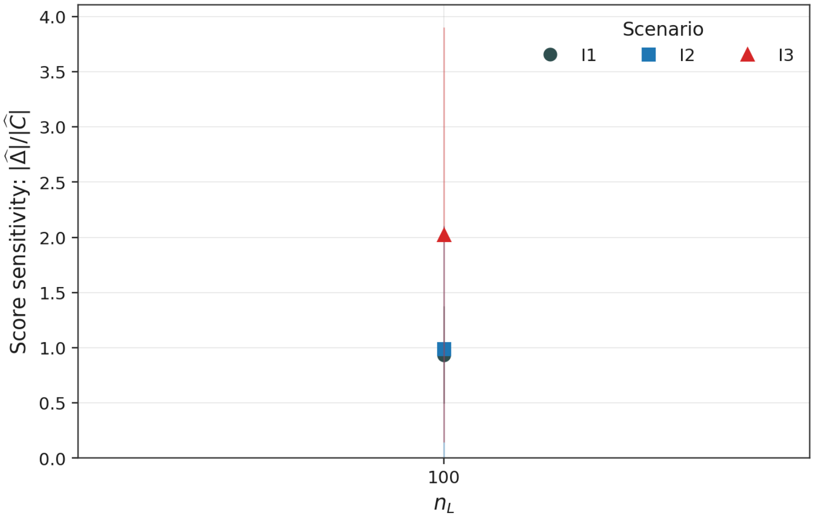


Figure 5: Score sensitivity. Point estimates and percentile bands of the ratio $|\widehat{\Delta}|/|\widehat{C}|$ at $n_L = 100$ for the three MC scenarios. The ratio is near unity under I1 (valid IV) and I2 (proxy route), while it centers around 2 and extends toward 4 under I3 (triple–proxy route), indicating greater fragility when identification relies entirely on the latent–confounding block.

Table 11: Summary statistics (analysis sample)

Variable	N	Mean	SD	Min	Max
Y	49,970	0.666	0.472	0.000	1.000
D	49,970	0.365	0.481	0.000	1.000
Z	49,970	0.000	1.000	-2.614	2.024

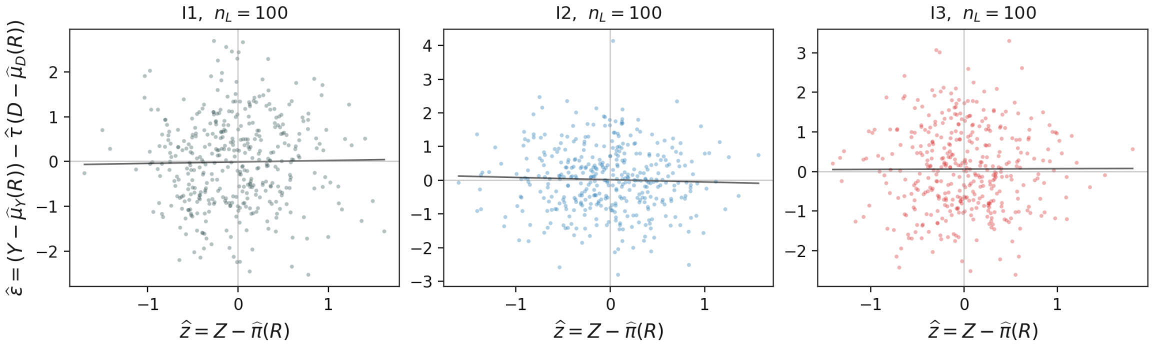


Figure 6: Orthogonality checks. Scatter of $\widehat{\varepsilon} = (Y - \widehat{\mu}_Y(R)) - \widehat{\tau}\,(D - \widehat{\mu}_D(R))$ versus the residualized instrument $\widehat{z} = Z - \widehat{\pi}(R)$ at $n_L = 100$, with least–squares fit (gray line) in each panel. Across I1–I3, slopes are near zero and the fitted lines are essentially flat, confirming numerical orthogonality of the estimating equations.

Table 12: Diagnostics for nuisance fits and orthogonality (strict instrument)

Metric	Value
Orthogonality slope (I1)	-0.000
OOF R^2 of $R \to \varepsilon$ (I2)	-0.001
OOF R^2 of $R \to Y$ (I3-μ_Y)	0.078
OOF R^2 of $R \to D$ (I3-μ_D)	0.367
Var(y_{res})	0.205
Var(d_{res})	0.147

References

Abadie, A., Agarwal, A., Dwivedi, R., and Shah, A. (2024). Doubly Robust Inference in Causal Latent Factor Models. *Working paper*, MIT.

Ai, C. and Chen, X. (2003). Efficient estimation of models With conditional moment restrictions containing unknown functions. *Econometrica*, 71, 1795-1843.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91, 444–455.

Anthony, M. and Bartlett, P. L. (1999). Neural Network Learning: Theoretical Foundations. *Cambridge University Press*, Cambridge, UK.

Arora, S., Khandeparkar, M., Khodak, M., Plevrakis, O. and Saunshi, N. (2019). A theoretical analysis of contrastive unsupervised representation learning. *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)*, 97, 323–332.

Arora, R., Basu, A., Mianjy, P., and Mukherjee, A. (2018). Understanding deep neural networks with rectified linear units. In *Proceedings of the International Conference on Learning Representations (ICLR).*

Bartlett, P., Maiorov, V., Montanari, A., and Rakhlin, A. (2019). Nearly-tight VC-dimension and Pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20, 1–38.

Bartlett, P. L., Foster, D. J., and Telgarsky, M. (2017). Spectrally-normalized margin bounds for neural networks. *Advances in Neural Information Processing Systems*, 30.

Bartlett, P. L., Bousquet, O., and Mendelson, S. (2005). Local Rademacher complexities. *Annals of Statistics* **33**(4), 1497–1537.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies*, 81, 608–650.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1998). *Efficient and Adaptive Estimation for Semiparametric Models.* New York: Springer.

Candès, E. J., and Recht, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9, 717–772.

Candès, E. J., and Plan, Y. (2011). Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57, 2342–2359.

Cheng, D., Xu, Z., Li, J., Liu, L., Le, T. D., and Liu, J. (2023). Learning conditional instrumental variable representation for causal effect estimation. *arXiv preprint*, arXiv:2306.12453v1.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. K., and Robins, J. M. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21, 1–68.

Chernozhukov, V., Fernández-Val, I., and Luo, Y. (2018). The sorted effects method: Discovering heterogeneous effects beyond their averages. *Econometrica*, 86, 1911–1938.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2018). Central limit theorems and the bootstrap in high dimension. *Annals of Probability*, 46, 2309–2352.

Deaner, B. (2023). Controlling for latent confounding with triple proxies. *arXiv preprint* arXiv:2204.13815v2.

Deaner, B. (2022). Many proxy controls. *arXiv preprint* arXiv:2110.03973.

Dudley, R. M. (1999). *Uniform Central Limit Theorems.* Springer, New York.

Farrell, M.H., Liang, T., and Misra, S. (2021). Deep neural networks for estimation and inference. *Econometrica*, 89, 181–213.

Fischer, A. and Steinwart, I. (2020). Sobolev norm learning rates for regularized least–squares algorithms. *Journal of Machine Learning Research*, 21, 1–63.

Giné, E. and Nickl, R. (2008). Uniform central limit theorems for kernel density estimators. *Probab. Theory Relat. Fields*, 141, 333–387.

Hall, P., and Horowitz, J. L. (2005). Nonparametric methods for inference in the presence of instrumental variables. *The Annals of Statistics*, 33(6), 2904–2929.

HaoChen, J., Wei, C., Gaidon, T„ and Ma, T. (2021). Provable guarantees for self-supervised contrastive learning with spectral contrastive loss. *Proceedings of the 35th International Conference on Neural Information Processing Systems (NIPS'21)*, 382, 5000-5011.

Hartford, J., Lewis, G., Leyton-Brown, K., and Taddy, M. (2017). Deep IV: A flexible approach for counterfactual prediction. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 1414–1423.

Kohler, M., and Krzyżak, A. (2025). Statistically guided deep learning. *arXiv preprint* arXiv:2504.08489v1.

Kohler, M., and Langer, S. (2021). On rates of convergence of fully connected deep neural network regression estimates. *The Annals of Statistics*, 49, 2231–2249.

Kuzborskij, I., Jun, K.-S., Wu, Y., Jang, K., and Orabona, F. (2024). Better-than-KL PAC-Bayes Bounds. *Proceedings of Machine Learning Research*, 247, 1–28. Presented at the 37th Annual Conference on Learning Theory.

Langer, S. (2021). Analysis of the rate of convergence of fully connected deep neural network regression estimates with smooth activation function Sophie Langer. *Journal of Multivariate Analysis*, 182, 104695.

Li, Z., Meunier, D., Mollenhauer, M., & Gretton, A. (2024). Towards optimal Sobolev norm rates for the vector-valued regularized least-squares algorithm. *Journal of Machine Learning Research*, 25, 1–51.

Meunier, D., Moulin, A., Kostic, V. R., Wornbard, J., and Gretton, A. (2025). Demystifying spectral feature learning for instrumental variable regression. *arXiv preprint arXiv:2506.10899v1.*

Meunier, D., Li, Z., Christensen, T., and Gretton, A. (2024). Nonparametric instrumental regression via kernel methods is minimax optimal. *arXiv preprint arXiv:2411.19653v1.*

Miyato, T., Maeda, S.-I., Koyama, M., & Ishii, S. (2018). Virtual Adversarial Training: a Regularization Method for Supervised and Semi-Supervised Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8), 1979–1993.

Negahban, S. N., and Wainwright, M. J. (2012). Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *Journal of Machine Learning Research*, 13, 1665–1697.

Neyshabur, B., Bhojanapalli, S., and Srebro, N. (2018). A PAC–Bayesian Approach to Spectrally-Normalized Margin Bounds for Neural Networks. *International Conference on Learning Representations (ICLR)*, Blind Submission; revised 23Feb2018.

Newey, W. K. (1990). Semiparametric efficiency bounds. *Econometrica*, 58, 997–1014.

Newey, W. K., and Powell, J. L. (2003). Instrumental variable estimation of nonparametric models. *Econometrica*, 71, 1565–1578.

Okui, R., Small, D. S., Tan, Z., and Robins, J. M. (2012). Doubly robust instrumental variable regression. *Statistica Sinica*, 22, 173–205.

van den Oord, A., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint* arXiv:1807.03748.

Robins, J. M., and Rotnitzky, A. (1995). *S*emiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429), 122–129.

Schmidt-Hieber, J. (2020). Nonparametric regression using deep neural networks with ReLU activation function. *Annals of Statistics*, 48, 1875–1897.

Shen, X., and Espinoza, J. (2025). Consistency and rate of convergence for deep ReLU neural networks. *Journal of Statistical Theory and Practice*, 19:33.

Singh, A., Sahani, M., and Gretton, A. (2019). Kernel instrumental variable regression. *Advances in Neural Information Processing Systems*, 32.

Sriperumbudur, B. K., Fukumizu, K., & Lanckriet, G. R. G. (2011). Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, **12**, 2389–2410.

Stock, J. H., Wright, J., and Yogo, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business and Economic Statistics*, 20, 518–529.

Tchetgen Tchetgen, E. J., Wang, L. and Sun, B. (2018). Discrete choice models for non-monotone nonignorable missing data: Identification and inference. *Statistica Sinica*, 28(4), 2069–2088.

Tchetgen Tchetgen, E. J., and Robins, J. M. and Rotnitzky, A. (2010). On Double-robust estimation in a semiparametric odds-ratio model. *Biometrika*, 97, 171–180.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer–Verlag.

Wang, L., and Tchetgen Tchetgen, E. J. (2018). Bounded, efficient and multiply robust estimation of average treatment effects using instrumental variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80, 531–550.

Wu, P., and Fukumizu, K. (2021). Towards principled causal effect estimation by deep identifiable models. *arXiv preprint arXiv:2109.15062v2*, 2021.

Xu, Liyuan; Chen, Yutian; Doucet, Arnaud; Srinivasan, Siddarth; Gretton, Arthur; and de Freitas, Nando. 2023. Learning Deep Features in Instrumental Variable Regression. *arXiv preprint*, arXiv:2010.07154v4.

Yarotsky, D. (2017). Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94, 103–114.