Adaptive Double Machine Learning via Riesz-Risk Cross-Validation

Tamer Çetin*

January 27, 2026

Abstract

We develop an end-to-end procedure that renders double/debiased machine-learning (DML) inference adaptive to the complexity asymmetries of the first-stage nuisances. For any linear-functional target we view debiasing through the Riesz representation, whose dual object, α_0, is estimable by minimizing a quadratic Riesz risk. Exploiting this variational characterization, we propose a *decoupled* cross-validation scheme: the outcome regression g is chosen by mean-squared prediction loss, whereas the representer α is selected by held-out Riesz risk. Selection is nested inside cross-fitting, so the score remains Neyman-orthogonal. We prove: (i) a fast-rate oracle inequality for the Riesz-risk selector over growing, possibly non-nested libraries; (ii) an *adaptive orthogonality* lemma showing that data-driven architecture choice preserves second-order bias control; and (iii) asymptotic linearity at the semiparametric efficiency bound. Simulations and a 401(K) revisit show that the procedure automatically resolves bias–variance trade-offs in same-architecture DML, delivering robust $\sqrt{n}$ confidence intervals guarding against violations of the product-rate condition that can generate non-negligible $\sqrt{n}$-scale drift in standard DML implementations.

Keywords: Debiased/Double Machine Learning; Riesz Representers; Orthogonal Scores; Inference-Targeted Cross-Validation; Oracle Inequalities; Semiparametric Efficiency. **JEL Classification:** C14, C21, C52.

*University of California, Berkeley. Email: tamercetin@berkeley.edu. I thank Victor Chernozhukov for helpful comments. All errors are my own.

1 Introduction

Modern machine learning (ML) has broadened the scope of semiparametric econometrics by enabling flexible estimation of high-dimensional and nonparametric nuisance functions (Mullainathan and Spiess, 2017). The double/debiased machine learning (DML) framework of Chernozhukov et al. (2018) shows that combining cross-fitting with Neyman-orthogonal scores yields $\sqrt{n}$-consistent, asymptotically normal estimators for a wide class of causal and structural parameters even when first-stage components converge at nonparametric rates. Many such targets can be written as continuous linear functionals of a regression nuisance, $\theta_0 = \mathbb{E}[m(W, g_0)], g_0(V) = \mathbb{E}[U \mid V]$, for a known map $m(\cdot, \cdot)$ that is linear and continuous in its second argument. For this class, the efficient influence function depends not only on g_0 but also on a dual object—the *Riesz representer* α_0—that encodes the functional-specific geometry (Newey, 1994; Santos, 2011, 2012; Chernozhukov et al., 2022a,b). In particular, the canonical orthogonal score takes the form $\psi(W; \theta, g, \alpha) = m(W, g) - \theta + \alpha(W)^\top\{U - g(W)\}$, and the post-orthogonalization drift is governed by the product of $L^2(P)$ errors, $\|\hat{g} - g_0\|_{P,2} \, \|\hat{\alpha} - \alpha_0\|_{P,2}$. Consequently, accurate learning of α_0 is not an auxiliary step but a first-order determinant of inferential validity for linear-functional targets.

Despite orthogonality, empirical DML implementations necessarily require *choosing* how to learn the nuisances. Standard practice commits *ex ante* to a particular double pipeline (e.g. Lasso/Lasso, Forest/Forest, Boosting/Boosting), and then tunes within that pipeline by predictive criteria: MSE for outcome regression and log-loss (or related scores) for propensity or other components. This workflow is natural from a prediction perspective and aligns with the library-selection tradition of van der Laan et al. (2007). However, prediction is not the inferential objective. For the representer, the relevant loss is not governed by outcome prediction error, and predictive accuracy for an intermediate nuisance (e.g. a propensity score) can be weakly informative about $\|\hat{\alpha} - \alpha_0\|_{P,2}$, precisely because α_0 depends on the inverse geometry of the functional and can be sensitive to regularization and tail behavior. Recent evidence documents that hyperparameter choices that are essentially equivalent for prediction can lead to materially different bias and coverage for debiased estimators (Bach et al., 2024).

This paper isolates a distinct and practically important failure mode that is not addressed by tune-for-prediction DML: *architectural misspecification induced by symmetric pipelines*. The DML remainder is a *product* of nuisance errors, so restrictions

that force the same learner family (or the same complexity regime) for g and α can be first-order. In particular, a diagonal restriction may exclude the *only* nuisance pair in a broader library whose error product satisfies the $o_p(n^{-1/2})$ requirement, even when such a pair exists off the diagonal. This is a pre-determined architecture-selection problem rather than a tuning problem: the object is the *learning procedure itself* (learner family, feature/regularization regime, and tuning rule), and the restriction that couples the two nuisances can invalidate inference by construction.

Recent work has made major progress on learning α_0 via its variational characterization (Chernozhukov et al., 2024b). Chernozhukov et al. (2022a,b) introduce regularized Riesz regression based on the quadratic *Riesz risk*, and subsequent contributions develop flexible implementations such as RieszNet and ForestRiesz (Chernozhukov et al., 2024a; Argañaraz, 2025; Kato, 2025; Quintas-Martínez, 2025), along with related approaches to debiasing in inverse problems (Newey and Powell, 2003; Chen and Pouzo, 2012; Bennett et al., 2023) and localized settings (Kallus et al., 2024). Canonical econometric examples of such inverse/conditional-moment structures include NPIV and related models (Newey and Powell, 2003; Chen and Pouzo, 2012), and associated work on recovering linear functionals and efficiency/inference (Santos, 2011, 2012; Severini and Tripathi, 2012; Abadie, 2003). A complementary perspective arises in the balancing/weighting literature, where the representer coincides with population-optimal weights and connects debiasing to covariate balance and minimax weights (Zubizarreta, 2015; Athey et al., 2018; Hirshberg and Wager, 2021; Bruns-Smith et al., 2025). These contributions clarify that representer learning is central and that the appropriate objective is functional-specific. However, they typically analyze representer estimation *given* a chosen learner class or tune *within* a fixed architecture, leaving open the broader question that arises in standard applied DML: how to choose among qualitatively different nuisance-learning architectures in a way that is aligned with the influence function and preserves valid inference.

Motivated by these gaps, we propose Adaptive Riesz–DML, a novel data-driven estimator that selects nuisance learning architectures from broad libraries, in a way that respects the distinct statistical roles of g_0 and α_0. The key observation is that g_0 is a conditional-mean regression object, whereas α_0 is a dual representer defined by Hilbert-space geometry. Accordingly, we select g by cross-validated prediction loss, while selecting α by cross-validated Riesz risk, $R(\alpha) = \mathbb{E}[\|\alpha(W)\|_2^2 - 2\, m(W, \alpha)]$, $\alpha_0 = \arg\min_\alpha R(\alpha)$, and we embed both selectors in a nested cross-fitting pipeline so that

architecture choice is made on training folds and the orthogonal score is evaluated on independent held-out folds. The default implementation is structurally *decoupled*: g and α may be chosen from different learner families and complexity regimes. When g and α are constrained to share parameters, we also provide an optional composite criterion following Chernozhukov et al. (2024a).

The contribution of the adaptive DML estimator proposed in this paper is not merely the use of an alternative cross-validation loss. Rather, it resolves the *ex-ante architecture-selection problem* for orthogonal inference with a Riesz representer. Relative to existing automatic debiasing and Riesz-regression approaches, the key distinctions are as follows. (i) Existing methods typically fix a learner class and tune hyperparameters within that class; by contrast, we treat the entire nuisance-learning *procedure* as a tertiary nuisance parameter and select among heterogeneous candidates. (ii) Prediction losses (MSE or log–loss) are appropriate for the regression nuisance g_0 but do not, in general, control the representer error that governs the DML remainder. We therefore select α by a held–out *Riesz risk* criterion that directly targets α_0 through its variational definition, using a common functional–relevant yardstick that compares plugin and direct representer learners on equal footing. (iii) Standard double pipelines impose a diagonal (coupled) restriction on (g, α). Our procedure allows *decoupled* selection over the full product library, permitting g and α to lie in different learner families and complexity regimes.

To formalize the consequences of these distinctions, we first show that symmetric (coupled) architecture restrictions can be first-order suboptimal by excluding off-diagonal nuisance pairs that satisfy the product-rate condition[1]. Our theory then delivers two linked guarantees for the adaptive estimator: (i) a fast-rate oracle inequality for the Riesz-risk selector (and the corresponding prediction-risk selector), and (ii) an adaptive orthogonality property, which ensures that data-driven architecture choice creates no $\sqrt{n}$-scale drift. Under standard moment conditions and the usual product-rate requirement—assuming the library contains at least one valid pair—the resulting estimator is asymptotically linear, normal, and semiparametrically efficient.

Empirically, simulations show that Adaptive Riesz-DML tracks the best fixed

[1]When we refer to first-order bias or $\sqrt{n}$-scale drift, we mean a contribution to the population moment that does not vanish at the $n^{-1/2}$ rate and therefore contaminates the asymptotic distribution. For linear-functional scores, the population moment admits an exact second-order representation in the nuisance errors, so such drift arises only when the product-rate condition fails.

pipeline across regimes and mitigates coverage distortions driven by regularization bias and architecture misspecification in standard DML implementations. In a 401(k) application (Chernozhukov et al., 2018), the procedure resolves sensitivity across plausible baselines by selecting nonlinear propensity/weight architectures while keeping the outcome regression relatively structured. The resulting fold-level selections provide a reproducible diagnostic of *asymmetric complexity*, directly constraining researcher discretion in applied DML workflows and guarding against violations of the product-rate condition that can reintroduce first-order drift—precisely the failure mode orthogonalization is designed to rule out.

The remainder of the paper is organized as follows. Section 2 introduces the framework and the Riesz landscape. Section 3 presents the adaptive algorithm and selection criteria. Section 4 develops asymptotic theory and Section 5 provides finite-sample bounds. Sections 6 and 7 report simulations and the 401(k) application. Section 8 concludes.

2 General Framework and the Riesz Landscape

2.1 Setup and notation

Let W denote a generic observation with law P. Let $U = u(W) \in \mathbb{R}^J$ be a known vector of signal variables with $\mathbb{E}[\|U\|_2^2] < \infty$. Let $V = v(W)$ denote the feature vector used for first-stage learning and let $\mathcal{V} := \sigma(V)$ be the associated σ-field. Define the regression nuisance $g_0(V) := \mathbb{E}[U \mid \mathcal{V}] \in \mathbb{R}^J$, so that $\mathbb{E}[U - g_0(V) \mid \mathcal{V}] = 0$. When no confusion arises, we write $g_0(W)$ as shorthand for $g_0(v(W))$.

In applications, V collects the first-stage features needed for the target functional. For example, in the ATE with binary D one can take $U = Y$ and $V = (D, X)$ so that $g_0(d, x) = \mathbb{E}[Y \mid D = d, X = x]$; in a weighted average derivative one typically takes $U = Y$ and $V = X$; and in a two-period DiD with group indicator G and outcome difference ΔY one may take $U = \Delta Y$ and $V = (G, X)$. Let $\mathcal{H} := L^2(P; \mathcal{V})^J = \big\{h : \Omega \to \mathbb{R}^J : \ h \text{ is } \mathcal{V}\text{-measurable and } \mathbb{E}\|h(W)\|_2^2 < \infty\big\}$ denote the closed $\mathcal{V}$-measurable subspace of $L^2(P)^J$. We equip $\mathcal{H}$ with the inner product $\langle f, h\rangle_P := \mathbb{E}[f(W)^\top h(W)]$ and associated norm $\|h\|_{P,2} := \langle h, h\rangle_P^{1/2}$. We consider a scalar target parameter θ_0

that admits the representation

$$\theta_0 = \mathbb{E}[m(W, g_0)], \tag{1}$$

for a known map $m(\cdot, \cdot)$ that is linear in its second argument and *(strongly) continuous* on $\mathcal{H}$ in the sense that there exists $C_m < \infty$ such that

$$\|m(\cdot, h)\|_{P,2} \leq C_m \|h\|_{P,2} \qquad \text{for all } h \in \mathcal{H}. \tag{2}$$

2.2 The Riesz representer

Because $T(h) := \mathbb{E}[m(W, h)]$ is then a continuous linear functional on $\mathcal{H}$ (by Cauchy–Schwarz and (2)), the Riesz–Fréchet theorem implies that there exists a unique $\alpha_0 \in \mathcal{H}$ such that

$$\mathbb{E}[m(W, g)] = \mathbb{E}\big[\alpha_0(W)^\top g(W)\big] \quad \text{for all } g \in \mathcal{H}. \tag{3}$$

Moreover, since $g_0(V) = \mathbb{E}[U \mid V]$, we have the orthogonality

$$\mathbb{E}\big[h(W)^\top \{U - g_0(V)\}\big] = 0 \qquad \text{for all } h \in \mathcal{H}, \tag{4}$$

and in particular (4) holds for α_0 and for any estimator $\widehat{\alpha} \in \mathcal{H}$.

We refer to α_0 as the *Riesz representer*. It quantifies the Gateaux derivative of θ_0 with respect to perturbations of g_0: if g_0 is replaced by $g_0 + th$ for some $h \in \mathcal{H}$, then $\frac{d}{dt}\big|_{t=0} \mathbb{E}[m(W, g_0 + th)] = \mathbb{E}\big[\alpha_0(W)^\top h(W)\big]$. In DML, the representer enters the orthogonal estimating equation via the score,

$$\psi(W; \theta, g, \alpha) = m(W, g) - \theta + \alpha(W)^\top \{U - g(W)\}. \tag{5}$$

which satisfies $\partial_g \mathbb{E}[\psi(W; \theta_0, g_0, \alpha_0)] = 0$. The score implies that small errors in the estimated nuisance functions $\hat{g}$ do not have a first–order effect on the estimator of θ_0.

2.3 Variational characterization and the Riesz risk

Our approach leverages the variational characterization of the Riesz representer on the Hilbert space $\mathcal{H} = L^2(P; \mathcal{V})^J$ of $\mathcal{V}$-measurable functions. Define the *Riesz risk* for $\alpha \in \mathcal{H}$ by

$$R(\alpha) := \mathbb{E}\big[\|\alpha(W)\|_2^2 - 2\, m(W, \alpha)\big]\,. \tag{6}$$

Using the Riesz identity $\mathbb{E}[m(W,\alpha)] = \mathbb{E}[\alpha_0(W)^\top \alpha(W)]$ (valid for $\alpha \in \mathcal{H}$), rewrite $R(\alpha) = \mathbb{E}[\|\alpha(W) - \alpha_0(W)\|_2^2] - \mathbb{E}[\|\alpha_0(W)\|_2^2]$, so that $R(\alpha) - R(\alpha_0) = \|\alpha - \alpha_0\|_{P,2}^2$.

Proposition 1 (Variational definition established by Chernozhukov et al. (2024b))**.** *The Riesz representer α_0 is the unique minimizer of $R(\alpha)$ over $\alpha \in \mathcal{H}$.*

Proof. See Appendix A.1. □

This characterization makes α_0 estimable via empirical risk minimization. Unlike the outcome regression g_0, which is typically estimated by minimizing the squared loss on Y, the Riesz representer must be estimated by minimizing the empirical analog of $R(\alpha)$; this distinction underlies our cross–validation strategy.

2.4 Examples

In examples where the functional $m(W, g)$ involves counterfactual evaluations such as $g(1, X)$ and $g(0, X)$, we interpret g as a vector-valued function collecting the required components, e.g. $g(X) = (g_1(X), g_0(X))$ with $g_d(X) = \mathbb{E}[Y \mid D = d, X]$. We then write $g(d, X)$ as shorthand for the corresponding component $g_d(X)$. Analogously, when $m(W, \alpha)$ requires $\alpha(1, X)$ and $\alpha(0, X)$, we view α as a function of (d, X) (or as a two-component vector) so these evaluations are well-defined.

When the target functional requires counterfactual evaluations (e.g. $g(1, X)$ and $g(0, X)$), we view g and α as functions on an augmented domain $\mathcal{S} \times \mathcal{X}$, where $\mathcal{S}$ indexes the relevant state (e.g. treatment status). For a realized observation W with state component $S \in \mathcal{S}$, we write $g(W) := g(S, X)$ and $\alpha(W) := \alpha(S, X)$, while $g(s, X)$ and $\alpha(s, X)$ denote counterfactual evaluations at state $s \in \mathcal{S}$. This convention keeps the definition in (1) compatible with ATE/DiD-style functionals without changing the underlying DML score.

Remark 1 (Binary treatments: vector vs. augmented-domain notation)**.** For binary $D \in \{0, 1\}$, it is equivalent to view the regression nuisance as (i) a vector-valued map $x \mapsto (g(1, x), g(0, x)) \in \mathbb{R}^2$ or (ii) a scalar map $(d, x) \mapsto g(d, x)$ on $\mathcal{S} \times \mathcal{X}$ with $\mathcal{S} = \{0, 1\}$. These are isomorphic representations of the same element of the V-measurable Hilbert space: one can identify $g(d, x) = e_d^\top g(x)$. In the examples, we write $g(W) = g(D, X)$ and use $g(1, X)$ and $g(0, X)$ as shorthand for the two slices of this function. Importantly, the Hilbert geometry (inner product and norm) is always inherited from $L^2(P; \mathcal{V})$ with $\mathcal{V} = \sigma(V)$ (and $V = (D, X)$ in the binary-treatment

ATE example); the vector notation is purely representational and does *not* switch to an $L^2(P_X)^2$ inner product based on the marginal distribution of X.

Average treatment effect (ATE). For a binary treatment $D \in \{0,1\}$ the target is $\theta_0 = \mathbb{E}[g_0(1,X) - g_0(0,X)]$ with $g_0(d,x) = \mathbb{E}[Y \mid D = d, X = x]$. Then $m(W,g) = g(1,X) - g(0,X)$ and the Riesz representer is the stabilized inverse propensity weight:

$$\alpha_0(W) = \frac{D}{p_0(X)} - \frac{1-D}{1-p_0(X)} \tag{7}$$

where $p_0(X) = \Pr(D = 1 \mid X)$. Minimizing $R(\alpha)$ in this context corresponds to balancing covariates across treatment groups[2].

Remark 2 (Overlap, clipping, and interpretation)**.** In ATE-type problems, boundedness of the Riesz representer $\alpha_0(W) = D/p_0(X) - (1-D)/(1-p_0(X))$ is equivalent to a strong overlap condition: $p_0(X) \in [\underline{p}, 1-\underline{p}]$ a.s. for some $\underline{p} > 0$. Our main theory is stated under such boundedness conditions, which ensure the orthogonal score has well-behaved moments and that representer estimation error can be controlled in $L^2(P)$. In practice, researchers often apply *stabilization* when overlap is fragile, e.g. clipping estimated propensities $\widehat{p}(X)$ to $[\tau, 1-\tau]$ or trimming observations with $\widehat{p}(X) \notin [\tau, 1-\tau]$. Two regimes are conceptually distinct: (i) (*bounded-overlap regime*) if $p_0(X)$ is bounded away from $\{0,1\}$ and τ is smaller than the overlap margin, clipping/trimming is asymptotically inactive and the procedure targets the original estimand θ_0. (ii) (*limited-overlap regime*) if $p_0(X)$ approaches $\{0,1\}$ with non-negligible probability, then boundedness assumptions fail and root-n inference for θ_0 may be delicate or impossible without further restrictions. In this regime, clipping/trimming should be interpreted as targeting a stabilized estimand (e.g. a trimmed ATE on the overlap region) and used as a sensitivity analysis that reveals dependence on extreme implied weights.

[2]In the ATE example, the representer coincides with the familiar inverse-propensity-weight form. It can be interpreted as the population-optimal balancing weight associated with the linear functional $m(W,h) = h(1,X) - h(0,X)$ through the Riesz identity $\mathbb{E}[m(W,h)] = \mathbb{E}[\alpha_0(W)\,h(W)]$. In the orthogonal score, taking $\alpha = \alpha_0$ delivers Neyman orthogonality: the Gateaux derivative of $P\psi(W;\theta_0,g,\alpha_0)$ with respect to g vanishes at $g = g_0$, so the remaining drift is governed by the second-order product $\|\hat{g} - g_0\|_{P,2}\,\|\hat{\alpha} - \alpha_0\|_{P,2}$. Consequently, instability or misspecification in the propensity/weight model affects inference primarily through representer error (and hence the product remainder), rather than through a first-order bias term in g.

Weighted average derivative (WAD). A weighted average derivative parameter takes the form $\theta_0 = \mathbb{E}\big[\omega(X)^\top \nabla_x g_0(X)\big], g_0(x) = \mathbb{E}[Y \mid X = x]$, for a known weight function ω. This is a *derivative-type* functional and, in general, is *not* continuous on $L^2(P)$; consequently, it need not admit an $L^2(P)$ Riesz representer. To incorporate such targets, one can instead work on a Sobolev-type Hilbert space that controls derivatives (e.g. an H^1 space with a norm involving both L^2 size and derivative size), redefine the Riesz representer and the representer risk with respect to that inner product, and then apply the same cross-validated selection logic on that space. Because this requires additional functional-analytic setup and notation, we focus the main text on targets that are continuous on $\mathcal{H} = L^2(P; \mathcal{V})^J$ (ATE, DiD, etc.).

Remark 3 (Beyond the L^2 setting)**.** The arguments in Sections 2–5extend to any separable Hilbert space $\mathcal{H}$ on which $h \mapsto \mathbb{E}[m(W, h)]$ is linear and continuous and for which the cross-fitted score is Neyman-orthogonal. In such cases one replaces the $L^2(P)$ inner products, norms, and Riesz risk by their $\mathcal{H}$ counterparts.

Difference–in–differences (DiD). Let $G \in \{0, 1\}$ denote the treated-group indicator and let ΔY be the two-period change in the outcome. Define $p := \Pr(G = 1)$ and $p(X) := \Pr(G = 1 \mid X)$, and let $\mu_g(X) := \mathbb{E}[\Delta Y \mid G = g, X]$. The ATT can be written as the linear functional $\theta_0 = \mathbb{E}\Big[\frac{G}{p}\{\mu_1(X) - \mu_0(X)\}\Big]$. In this representation the true regression nuisance is $g_0(W) = \mu_G(X)$, while the functional $m(W, g)$ is defined for a generic g by $m(W, g) = \frac{G}{p}\{g(1, X) - g(0, X)\}$. The corresponding Riesz representer is $\alpha_0(W) = \frac{G}{p} - \frac{1-G}{p}\,\frac{p(X)}{1-p(X)}$. Limited overlap or misspecification of the odds ratio $p(X)/(1 - p(X))$ is a common source of bias in DiD; our framework permits data-driven selection of the propensity architecture that governs this representer.

3 The adaptive Riesz-DML algorithm

This section describes Adaptive Riesz-DML and the nested cross-validation criteria used to select g by prediction risk and α by held-out Riesz risk. When the best $n^{-1/4}$-rate approximation regime for g_0 differs from that for α_0, restrictions that force the same learner class for both nuisances can exclude the oracle pair, causing the DML remainder $\sqrt{n}\|\hat{g} - g_0\|_{P,2}\|\hat{\alpha} - \alpha_0\|_{P,2}$ to fail to vanish. Section 3.1 formalizes coupling vs. decoupling and states the result; Appendix A.2 gives a formal rate-based statement

and proof. Section 6.3.3 provides a simulation where both diagonal architectures fail while the off-diagonal pair succeeds.

3.1 The architecture-selection: coupling vs. decoupling

Orthogonal-score inference depends on *both* the regression nuisance g_0 and the Riesz representer α_0. Since the leading drift is proportional to $\|\hat{g} - g_0\|_{P,2} \|\hat{\alpha} - \alpha_0\|_{P,2}$, architecture restrictions that tie the two nuisance learners together can be first-order when the best approximation regimes for g_0 and α_0 differ.

Definition 1 (Decoupled vs. coupled nuisance architectures)**.** Let Φ_n^g and Φ_n^α denote libraries of candidate learning procedures for the regression nuisance g and the Riesz representer α. A DML implementation is *decoupled* if it selects $(\widehat{g}, \widehat{\alpha})$ over the full product library $\Phi_n^g \times \Phi_n^\alpha$. It is *coupled* if it restricts the search to a proper subset $\mathcal{S}_n \subsetneq \Phi_n^g \times \Phi_n^\alpha$. A prominent special case—and the focus of our negative results and benchmarks—is a *(mapped-)diagonal ("symmetric") coupling* of the form $\mathcal{S}_n^{\text{diag}} = \{(\varphi, T_n(\varphi)) : \varphi \in \Phi_n^g\}$, for some (possibly n-dependent) map $T_n : \Phi_n^g \to \Phi_n^\alpha$ (including the diagonal $T_n(\varphi) = \varphi$ when the two libraries coincide).

Proposition 2 (Why symmetric restrictions can be first-order (informal))**.** *If one candidate nuisance architecture learns g_0 well but approximates α_0 poorly, while another candidate learns α_0 well but incurs larger error for g_0, then restricting attention to diagonal (same-architecture) pairs can violate the DML product-rate condition even though an off-diagonal pair in the full product library satisfies it.*

Intuition. Orthogonality eliminates first-order bias in each nuisance separately, but the remaining drift is proportional to $\|\hat{g} - g_0\|_{P,2} \|\hat{\alpha} - \alpha_0\|_{P,2}$. A diagonal restriction forces a trade-off between these two errors and may exclude the only pair whose product is $o(n^{-1/2})$. A formal rate-based statement and proof appear in Appendix A.2.

Remark 4 (Concrete interpretation: asymmetric complexity)**.** Empirically, g_0 may be well-approximated by a structured, low-variance class (e.g. sparse linear regression), while α_0 depends on a substantially more nonlinear object. Decoupled selection over $\Phi_n^g \times \Phi_n^\alpha$ avoids excluding the valid off-diagonal pair that satisfies the product-rate condition when these effective complexities differ.

3.2 Library of learners and tertiary parameters

Having established the necessity of decoupling, we construct a library of candidate learners. For the regression function g_0, we construct candidate estimators $\widehat{\mathcal{G}} = \{\hat{g}_\varphi : \varphi \in \Phi_n^g\}$; for the Riesz representer α_0, we construct candidates $\widehat{\mathcal{A}} = \{\hat{\alpha}_\kappa : \kappa \in \Phi_n^\alpha\}$. We reserve φ for regression-library indices and κ for representer-library indices. The libraries Φ_n^g and Φ_n^α may contain diverse learners (e.g., Lasso, random forests, gradient boosting, neural networks) with varying hyperparameters. To ensure uniform convergence of the cross–validated estimator, we assume $|\Phi_n|$ grows at most polynomially with n so that $\log|\Phi_n| = O(\log n)$. This baseline design corresponds to a *structurally decoupled* implementation in the sense of Definition 1; Section 3.4 discusses a deliberate coupled (composite) criterion when g and α are constrained to share representation layers.

Remark 5 (Library construction and rate verification)**.** The library Φ_n^α indexes candidate procedures $\widehat{\alpha}_\kappa$, including both plugin representers (mapping estimated nuisances $\widehat{\eta}$ to α via $\mathcal{T}$) and direct Riesz learners (minimizing empirical Riesz risk). Crucially, Assumption 2 requires only that the library contains *some* pair $(\hat{g}_{\varphi^*}, \hat{\alpha}_{\kappa^*})$ satisfying the product-rate condition $\|\hat{g}_{\varphi^*} - g_0\|_{P,2}\,\|\hat{\alpha}_{\kappa^*} - \alpha_0\|_{P,2} = o_p(n^{-1/2})$. This is verifiable for both constructions: (i) For direct learners, the identity $R(\alpha) - R(\alpha_0) = \|\alpha - \alpha_0\|_{P,2}^2$ ensures that standard quadratic-risk rates (e.g., for Lasso or Forests) control the representer error[3]. (ii) For plugin learners, Lipschitz continuity of the map $\eta \mapsto \alpha(\eta)$ (e.g., under overlap) ensures $\|\widehat{\alpha} - \alpha_0\|_{P,2} \lesssim \|\widehat{\eta} - \eta_0\|_{P,2}$. Note that the oracle indices for g and α need not coincide; the adaptive procedure selects the best combination from the product library.

3.3 Riesz–corrected cross–validation

For the regression nuisance $g_0(V) = \mathbb{E}[U \mid \mathcal{V}]$, the population prediction risk satisfies the standard decomposition $\mathbb{E}\big[\|U - g(V)\|_2^2\big] = \mathbb{E}\big[\|U - g_0(V)\|_2^2\big] + \|g - g_0\|_{P,2}^2$. Consequently, minimizing population MSE is equivalent to minimizing the estimation error $\|g - g_0\|_{P,2}$. This motivates selecting g by standard cross-validated MSE. In contrast, the representer α_0 is characterized by the Riesz identity (3) rather than a prediction

[3]For example, $o_p(n^{-1/4})$ rates are available under approximate sparsity (Bickel et al., 2009; Belloni and Chernozhukov, 2011), honesty conditions for forests (Wager and Walther, 2016; Athey et al., 2019), or smoothness for neural networks (Chen and White, 1999; Schmidt-Hieber, 2020).

problem, so predictive performance for U does not, in general, control $\|\widehat{\alpha} - \alpha_0\|_{P,2}$.

We propose *Riesz–corrected cross–validation* (RC–CV), which selects the candidate minimizing an empirical analog of the Riesz risk. Split the sample into K folds. Denote fold k by I_k and its complement by I_k^c. For each candidate $\kappa \in \Phi_n^\alpha$, let $\hat{\alpha}_\kappa^{(-k)}$ be the estimator fit on I_k^c. Define the (quadratic) Riesz loss $\ell(W, \alpha) := \|\alpha(W)\|_2^2 - 2\, m(W, \alpha)$, where $\|\cdot\|_2$ denotes the Euclidean norm in $\mathbb{R}^J$ (in the scalar case $J = 1$, $\|\alpha(W)\|_2^2 = \alpha(W)^2$).

$$\widehat{R}_{CV}(\kappa) = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{|I_k|} \sum_{i \in I_k} \ell\big(W_i, \hat{\alpha}_\kappa^{(-k)}\big)\,. \tag{8}$$

Our adaptive Riesz estimator chooses

$$\hat{\kappa}_\alpha = \arg\min_{\kappa \in \Phi_n^\alpha} \widehat{R}_{CV}(\kappa), \tag{9}$$

relying on the fact that the true α_0 uniquely minimizes the population risk.[4] This selection uses only cross–fitted observations, preserving independence between training and evaluation.

Remark 6 (Riesz risk estimation)**.** The Riesz risk $R(\alpha)$ in (6) is a population quantity and cannot, in general, be evaluated on the same data used to construct $\hat{\alpha}$ without inducing training-sample bias. We therefore estimate $R(\alpha)$ using held-out observations. For each $\kappa \in \Phi_n^\alpha$ and inner fold k, let $\hat{\alpha}_\kappa^{(-k)}$ denote the estimator trained on I_k^c, and define the fold-specific risk estimate $\widehat{R}_k(\kappa) := \frac{1}{|I_k|} \sum_{i \in I_k} \ell\Big(W_i, \hat{\alpha}_\kappa^{(-k)}\Big)$, $\widehat{R}_{\mathrm{CV}}(\kappa) := \frac{1}{K} \sum_{k=1}^{K} \widehat{R}_k(\kappa)$. Let $\mathcal{F}_k := \sigma(\{W_i : i \in I_k^c\})$ be the training σ–field for fold k. Since $\hat{\alpha}_\kappa^{(-k)}$ is $\mathcal{F}_k$–measurable and $\{W_i : i \in I_k\}$ are i.i.d. and independent of $\mathcal{F}_k$, we have, for each fixed κ, $\mathbb{E}\Big[\widehat{R}_k(\kappa) \mid \mathcal{F}_k\Big] = R\Big(\hat{\alpha}_\kappa^{(-k)}\Big)$, $\mathbb{E}\Big[\widehat{R}_{\mathrm{CV}}(\kappa) \mid \{\mathcal{F}_k\}_{k=1}^K\Big] = \frac{1}{K} \sum_{k=1}^{K} R\Big(\hat{\alpha}_\kappa^{(-k)}\Big)$. Thus $\widehat{R}_{\mathrm{CV}}(\kappa)$ provides a conditionally unbiased estimate of the *out-of-sample* Riesz risk of the training procedure associated with κ.

The selector $\hat{\kappa}_\alpha \in \arg\min_{\kappa \in \Phi_n^\alpha} \widehat{R}_{\mathrm{CV}}(\kappa)$ is data dependent, and the minimum $\widehat{R}_{\mathrm{CV}}(\hat{\kappa}_\alpha)$ is generally optimistically biased relative to the corresponding oracle risk. Our theory does not rely on unbiasedness of the minimum. Instead, Theorem 1 controls the selection effect through a uniform deviation bound over the finite library,

[4]When the argmin over a finite grid is not unique, we use deterministic tie-breaking under a fixed enumeration (e.g., select the smallest index). This ensures measurability and has no effect on probability limits.

yielding an oracle inequality for $R(\hat{\alpha}_{\hat{\kappa}_\alpha})$.

3.3.1 Computing the Riesz loss on held-out data

The cross-validated criterion $\widehat{R}_{\mathrm{CV}}(\kappa)$ in (8) requires evaluating $\ell(W_i, \widehat{\alpha})$ on validation observations. This evaluation is fully mechanical once $\widehat{\alpha}$ is available as a function, because $m(W, \cdot)$ is known and linear. For a given target, fix the evaluation operator $h \mapsto m(W, h)$. Given a fitted representer candidate $\widehat{\alpha} : \mathcal{S} \times \mathcal{X} \to \mathbb{R}^J$, the held-out loss is $\ell(W, \widehat{\alpha}) = \|\widehat{\alpha}(W)\|_2^2 - 2\, m(W, \widehat{\alpha}), \text{with} \widehat{\alpha}(W) = \widehat{\alpha}(S, X)$. If $m(W, \widehat{\alpha})$ involves counterfactual evaluations (e.g. $\widehat{\alpha}(1, X)$ and $\widehat{\alpha}(0, X)$), these are computed by evaluating the learned function at the required states; they do not require observing those states. For the ATE with $\mathcal{S} = \{0, 1\}$ and $m(W, h) = h(1, X) - h(0, X)$, the Riesz loss becomes $\ell(W, \alpha) = \alpha(D, X)^2 - 2\{\alpha(1, X) - \alpha(0, X)\}, R(\alpha) = \mathbb{E}[\ell(W, \alpha)]$. On a validation fold, we compute $\alpha(D_i, X_i)$ using the realized D_i, and compute $\alpha(1, X_i)$ and $\alpha(0, X_i)$ by direct functional evaluation at both states.

3.4 Composite selection for shared representations (Practical Extension)

While our theoretical analysis focuses on the structurally decoupled regime (Definition 1), modern deep learning architectures often employ *shared representations*, where $\hat{g}$ and $\hat{\alpha}$ are heads of a common neural network trunk (Chernozhukov et al., 2024a). In such settings, independent selection is not feasible. We therefore propose a heuristic *composite criterion*:

$$\mathcal{J}_{CV}(\varphi) = \widehat{\mathrm{MSE}}_{CV}\big(\hat{g}_\varphi\big) + \lambda_n\, \widehat{R}_{CV}\big(\hat{\alpha}_\varphi\big). \tag{10}$$

Minimizing this joint criterion balances the bias–variance trade-off for the functional θ_0 rather than for the nuisance functions individually.

Remark 7 (Theoretical Scope)**.** The oracle inequality (Theorem 1) and asymptotic normality (Theorem 3) are established for the *decoupled* selector (Algorithm 1), where $\hat{\alpha}$ is chosen purely to minimize Riesz risk. We treat the composite criterion as a practical extension for parameter-sharing constraints; its formal theoretical properties require analyzing the joint optimization landscape and are left for future work.

3.5 Adaptive estimator via cross–fitting

In Algorithm 1, the final estimator is formalized with nested cross–fitting to ensure the selection of learners and the evaluation of the score are statistically independent.

Algorithm 1 Adaptive Riesz–DML with Nested Cross–Validation (outer cross-fitting, *decoupled* inner selection)

Input: data $\{W_i\}_{i=1}^n$, libraries Φ_n^g and Φ_n^α, outer folds $L \geq 2$, inner folds $K \geq 2$.
Notation: $V_i = v(W_i)$ denotes first-stage features; Z_i denotes any additional components of W_i entering $m(W, g)$.
(Decoupling.) For each outer fold ℓ, select $\hat{\varphi}_{g,\ell}$ and $\hat{\kappa}_{\alpha,\ell}$ independently over Φ_n^g and Φ_n^α.
Partition $\{1, \ldots, n\}$ into disjoint outer folds $(I_\ell)_{\ell=1}^L$.
for $\ell = 1$ to L **do**
  Set outer training indices $I_\ell^c := \{1, \ldots, n\} \setminus I_\ell$.
  Partition I_ℓ^c into disjoint inner folds $(I_{\ell,k})_{k=1}^K$.
  Inner CV for g: For each $\varphi \in \Phi_n^g$ and each k, train $\hat{g}_{\ell,\varphi}^{(-k)}$ on $I_\ell^c \setminus I_{\ell,k}$ and compute $\widehat{\mathrm{MSE}}_{\ell,k}(\varphi) := \frac{1}{|I_{\ell,k}|} \sum_{i \in I_{\ell,k}} \|U_i - \hat{g}_{\ell,\varphi}^{(-k)}(V_i)\|_2^2$. Define $\widehat{\mathrm{MSE}}_{CV,\ell}(\varphi) := \frac{1}{K} \sum_{k=1}^K \widehat{\mathrm{MSE}}_{\ell,k}(\varphi)$ and select $\hat{\varphi}_{g,\ell} \in \arg\min_{\varphi \in \Phi_n^g} \widehat{\mathrm{MSE}}_{CV,\ell}(\varphi)$ (deterministic tie-breaking).
  Inner CV for α: For each $\kappa \in \Phi_n^\alpha$ and each k, train $\hat{\alpha}_{\ell,\kappa}^{(-k)}$ on $I_\ell^c \setminus I_{\ell,k}$ and compute $\widehat{R}_{\ell,k}(\kappa) := \frac{1}{|I_{\ell,k}|} \sum_{i \in I_{\ell,k}} \ell\big(W_i, \hat{\alpha}_{\ell,\kappa}^{(-k)}\big)$, $\ell(W, \alpha) := \|\alpha(W)\|_2^2 - 2m(W, \alpha)$. Define $\widehat{R}_{CV,\ell}(\kappa) := \frac{1}{K} \sum_{k=1}^K \widehat{R}_{\ell,k}(\kappa)$ and select $\hat{\kappa}_{\alpha,\ell} \in \arg\min_{\kappa \in \Phi_n^\alpha} \widehat{R}_{CV,\ell}(\kappa)$ (deterministic tie-breaking).
  Refit on outer training data: Fit $\hat{g}_\ell$ on I_ℓ^c using $\hat{\varphi}_{g,\ell}$ and fit $\hat{\alpha}_\ell$ on I_ℓ^c using $\hat{\kappa}_{\alpha,\ell}$.
  Outer-fold score evaluation: For each $i \in I_\ell$, set $\hat{\phi}_i := m(W_i, \hat{g}_\ell) + \hat{\alpha}_\ell(W_i)^\top \{U_i - \hat{g}_\ell(V_i)\}$.
end for
Estimator (closed form): $\hat{\theta}_{\mathrm{Adaptive}} := \frac{1}{n} \sum_{i=1}^n \hat{\phi}_i$.
return $\hat{\theta}_{\mathrm{Adaptive}}$.

In practice, inner-CV criteria can be statistically indistinguishable across similar nonlinear learners. We therefore use an ε-minimizer: we retain all candidates within one standard error of the minimum inner-CV criterion and select the simplest/stablest architecture within this set, with deterministic tie-breaking.

Remark 8 (Fold notation)**.** Throughout the paper, L denotes the number of *outer* cross–fitting folds used for score evaluation, while K denotes the number of *inner* folds

used for cross–validation during nuisance selection. In Sections where generic fold indices k appear in empirical-process decompositions, these refer to outer evaluation folds unless stated otherwise.

Remark 9 (Computational Complexity)**.** The nested cross-validation procedure scales as $O(L{\cdot}K{\cdot}|\Phi_n|{\cdot}\mathrm{Cost}_{\mathrm{train}})$. For large libraries or sample sizes ($N > 10^4$), full grid search can be prohibitive. In such regimes, we recommend two modifications that preserve the theoretical guarantees: (i) setting the inner fold count to $K = 2$, which suffices for unbiased risk estimation; or (ii) employing *successive halving* algorithms (e.g., `HalvingGridSearchCV`) to discard unpromising candidates early. In our simulations ($N = 1000$), full grid search was feasible; for the 401(k) application ($N \approx 10{,}000$), we utilized a discrete grid of 5–10 hyperparameters per learner class.

3.6 A stable selection variant: approximate minimizers

In empirical work we sometimes replace the exact argmin in Algorithm 1 by a stability rule (Cory-Wright and Gómez, 2025): we first form the set of candidates whose inner-CV criterion is within one estimated standard error of the minimum, and then select the simplest/stablest model within this set (ties broken deterministically). We refer to this as an ε-*minimizer* rule.

Definition 2 (ε_n-minimizer selector)**.** Let $\widehat{R}_{\mathrm{CV}}(\kappa)$ denote the inner-CV criterion. Given a tolerance sequence $\varepsilon_n \geq 0$, define the near-minimizer set $\widehat{\mathcal{N}}(\varepsilon_n) := \Big\{\kappa \in \Phi_n^\alpha : \widehat{R}_{\mathrm{CV}}(\kappa) \leq \min_{\kappa' \in \Phi_n^\alpha} \widehat{R}_{\mathrm{CV}}(\kappa') + \varepsilon_n\Big\}$. An ε_n-minimizer selector returns any $\widehat{\kappa} \in \widehat{\mathcal{N}}(\varepsilon_n)$ (e.g., the sparsest model).

Proposition 3 (Oracle bound under approximate minimization)**.** *Under the conditions of Theorem 1, suppose the selector returns $\hat{\kappa}_\alpha \in \widehat{\mathcal{N}}(\varepsilon_n)$ with $\varepsilon_n = O(n^{-1/2})$. Then $R(\hat{\alpha}_{\hat{\kappa}_\alpha}) - R(\alpha_0) \leq (1 + o_p(1)) \inf_{\kappa \in \Phi_n^\alpha}\{R(\hat{\alpha}_\kappa) - R(\alpha_0)\} + O_p\Big(\frac{\log|\Phi_n^\alpha|}{n} + \varepsilon_n + \delta_n\Big)$, where δ_n is the refit-stability error term from Theorem 1.*

Proof. The proof follows immediately from Theorem 1 by noting that the empirical excess risk is bounded by the minimum plus ε_n. Provided ε_n decays as fast as the parametric rate, the product-rate condition for DML remains satisfied. We use this variant in the empirical application to reduce fold-to-fold switching among near-equivalent nonlinear candidates.

4 Asymptotic theory

Let $(W_i)_{i=1}^n$ be i.i.d. with law P. Let $\hat{\theta}_{\text{Adaptive}}$ denote the nested cross–fitted estimator in (13), obtained by selecting $(\hat{g}, \hat{\alpha})$ from finite libraries Φ_n^g and Φ_n^α using Algorithm 1. This section establishes: (i) an oracle inequality for the Riesz–risk selector of α_0; (ii) preservation of Neyman–orthogonality under data–driven selection; and (iii) asymptotic normality, efficiency, and valid studentized inference. Throughout, $\|\cdot\|_{P,2}$ denotes the $L^2(P)$ norm and $\mathbb{E}$ denotes expectation under P.

4.1 Assumptions

Assumption 1 (Moments, continuity, and bounded representer)**.** There exists $q > 4$ such that $\mathbb{E}[\|U\|_2^q] < \infty$, $\mathbb{E}[|\psi(W; \theta_0, g_0, \alpha_0)|^q] < \infty$, and $\mathbb{E}\Big[\Big(\mathbb{E}[\|U - g_0(V)\|_2^2 \mid V]\Big)^{q/2}\Big] < \infty$. The map $h \mapsto m(\cdot, h)$ is linear and satisfies the $L^2(P)$ continuity bound

$$\|m(\cdot, h)\|_{P,2} \le C_m \|h\|_{P,2} \qquad \text{for all } h \in \mathcal{H}, \tag{11}$$

for some constant $C_m < \infty$. Finally, assume a *bounded representer* condition: $\|\alpha_0\|_\infty \le B_0$ for some finite B_0. In estimation, we apply componentwise clipping to all representer candidates so that for some deterministic $B \ge B_0$, $\sup_{\kappa \in \Phi_n^\alpha} \|\widehat{\alpha}_\kappa\|_\infty \le B$ a.s. In addition, assume that m is bounded on the relevant sup-norm ball: there exists $C_{m,\infty} < \infty$ such that

$$\|m(\cdot, h)\|_\infty \le C_{m,\infty} \|h\|_\infty \qquad \text{for all } h \in \mathcal{H}, \tag{12}$$

so that, in particular, $|m(W, \widehat{\alpha}_\kappa)| \le C_{m,\infty} B$ a.s. for all $\kappa \in \Phi_n^\alpha$.

The conditional moment condition $\mathbb{E}\Big[\Big(\mathbb{E}[\|U - g_0(V)\|_2^2 \mid V]\Big)^{q/2}\Big] < \infty$ is implied by $\mathbb{E}[\|U\|_2^q] < \infty$ (by Jensen's inequality) since $g_0(V) = \mathbb{E}[U \mid V]$ and $q > 4$. We state it separately because this conditional quantity is used directly in the variance and concentration arguments.

Assumption 2 (Finite libraries and oracle elements)**.** The candidate sets Φ_n^g and Φ_n^α are finite and satisfy $\log|\Phi_n^g| + \log|\Phi_n^\alpha| = O(\log n)$. For each $\varphi \in \Phi_n^g$ (resp. $\kappa \in \Phi_n^\alpha$), the estimator $\hat{g}_\varphi$ (resp. $\hat{\alpha}_\kappa$) is measurable with respect to the training sample and $\sup_\varphi \|\hat{g}_\varphi\|_\infty + \sup_\kappa \|\hat{\alpha}_\kappa\|_\infty = O_p(1)$. There exist (possibly n–dependent) indices $\varphi_{g,n}^* \in$

Φ_n^g and $\kappa_{\alpha,n}^* \in \Phi_n^\alpha$ such that $\|\hat{g}_{\varphi_{g,n}^*} - g_0\|_{P,2} \, \|\hat{\alpha}_{\kappa_{\alpha,n}^*} - \alpha_0\|_{P,2} = o_p(n^{-1/2})$. Whenever an argmin over a finite library is not unique, ties are broken deterministically.

In practice, tuning parameters often range over a continuum (e.g. $\lambda \in [\lambda_{\min}, \lambda_{\max}]$). Our analysis matches the standard implementation in which one evaluates a finite grid of candidates and selects by cross-validation. To satisfy $\log|\Phi_n| = O(\log n)$, one may use geometrically spaced grids, e.g. $\Phi_n = \{\lambda_{\max}\rho^j : j = 0, 1, \ldots, J_n\}$ with $\rho \in (0,1)$ fixed and $J_n = O(\log n)$.

Assumption 3 (Second–order regularity of the score)**.** The map $(g, \alpha) \mapsto \mathbb{E}[\psi(W; \theta_0, g, \alpha)]$ is twice Gateaux differentiable in a neighborhood of (g_0, α_0), with continuous first derivatives and uniformly bounded mixed second derivative.

Remark 10 (Role of Assumption 3)**.** In the linear-functional setting studied here, the key conditional moment identity in Theorem 2 is exact and does not rely on a Taylor expansion in (g, α). Accordingly, the proof of Theorem 3 uses only the moment/boundedness conditions, cross-fitting, and the explicit consistency and product-rate conditions on the nuisances. Assumption 3 is therefore not needed for Theorem 3 in this paper; it is included only to facilitate comparison with the broader DML literature and for potential extensions beyond the specific linear score considered here.

Assumption 4 (Bernstein condition for the Riesz loss)**.** Let $\ell(W, \alpha)$ and $R(\alpha)$ be defined as in (6). There exists $\nu > 0$ such that for any (possibly random) α measurable with respect to an independent training sample, $\mathbb{E}\Big[\big(\ell(W,\alpha) - \ell(W,\alpha_0)\big)^2 \,\Big|\, \alpha\Big] \leq \nu\big(R(\alpha) - R(\alpha_0)\big)$. Assumption 4 is implied by simple boundedness and Lipschitz conditions.

Lemma 1 (Sufficient condition for Assumption 4)**.** *Suppose $\|\alpha(W)\|_2 \leq B$ and $\|\alpha_0(W)\|_2 \leq B$ a.s. for all α in the candidate class, and m satisfies the $L^2(P)$ continuity bound $\|m(\cdot, h)\|_{P,2} \leq C_m\|h\|_{P,2}$ for all admissible $h \in \mathcal{H}$. Then the Bernstein condition (4) holds with $\nu = (2B + 2C_m)^2$.*

Proof. See Appendix A.3. □

Assumption 5 (Refit stability for the representer learner)**.** For each $\kappa \in \Phi_n^\alpha$, let $\widehat{\alpha}_\kappa$ denote the estimator trained on the full sample used for inner CV, and let $\widehat{\alpha}_\kappa^{(-k)}$ denote the estimator trained on the sample with fold k removed (as in the definition of $\widehat{R}_{CV}(\kappa)$). Define the fold-average population risk $\overline{R}_K(\kappa) := \frac{1}{K}\sum_{k=1}^K R\Big(\widehat{\alpha}_\kappa^{(-k)}\Big)$.

Assume there exists a sequence $\delta_n = o(n^{-1/2})$ such that $\sup_{\kappa \in \Phi_n^\alpha} \left| R(\widehat{\alpha}_\kappa) - \overline{R}_K(\kappa) \right| \leq \delta_n$ in probability.

Remark 11 (On refit stability)**.** Assumption 5 is a genuine restriction: it is natural for stable procedures (e.g. strongly convex regularized ERM), but it can fail for highly unstable learners (e.g. certain tree ensembles, boosting, or unregularized neural networks). Importantly, the oracle inequality for the *fold-trained* objects in Theorem 1 does not rely on refitting. When refit stability is doubtful, one can omit the refit step and instead construct the final estimator using the same K folds as a cross-fitting scheme. After selecting $\widehat{\kappa}_\alpha$, keep the fold-trained representers $\{\widehat{\alpha}^{(-k)}_{\widehat{\kappa}_\alpha}\}_{k=1}^K$ and evaluate the score fold-wise: for each $i \in I_k$, plug in $\widehat{\alpha}^{(-k)}_{\widehat{\kappa}_\alpha}(W_i)$ (and analogously for g if desired). Equivalently, define $\widehat{\theta}_{\text{no-refit}}$ as the solution to $\frac{1}{n}\sum_{k=1}^K \sum_{i \in I_k} \psi\Big(W_i; \theta, \widehat{g}^{(-k)}_{\widehat{\varphi}_g}, \widehat{\alpha}^{(-k)}_{\widehat{\kappa}_\alpha}\Big) = 0$. No additional aggregation of the K objects is required; each fold-trained nuisance is used only on its own held-out fold.

Assumption 5 is implied, for example, by uniform stability of the learning rule for α (in the sense of algorithmic stability), which holds for many regularized ERM procedures under standard convexity and Lipschitz conditions.

4.2 Oracle inequality for the Riesz risk

Theorem 1 (Riesz oracle inequality)**.** *Suppose Assumptions 1–4 hold, and let* $\hat{\kappa}_\alpha \in \arg\min_{\kappa \in \Phi_n^\alpha} \widehat{R}_{CV}(\kappa)$*. Then, with probability tending to one,* $\overline{R}_K(\hat{\kappa}_\alpha) - R(\alpha_0) \leq (1 + o(1)) \inf_{\kappa \in \Phi_n^\alpha} \left\{\overline{R}_K(\kappa) - R(\alpha_0)\right\} + C\frac{\log|\Phi_n^\alpha|}{n}$*. Moreover, if Assumption 5 holds, then with probability tending to one,*

$$R(\widehat{\alpha}_{\hat{\kappa}_\alpha}) - R(\alpha_0) \leq (1 + o(1)) \inf_{\kappa \in \Phi_n^\alpha} \{R(\widehat{\alpha}_\kappa) - R(\alpha_0)\} + C\frac{\log|\Phi_n^\alpha|}{n} + 2\delta_n{}^5.$$

Proof. See Appendix A.4. □

Lemma 2 (Prediction-risk oracle inequality for the regression selector)**.** *Let* $\widehat{\mathrm{MSE}}_{CV}(\varphi)$ *be the K-fold inner-CV squared-loss criterion used to select g in Algorithm 1, and let* $\widehat{\varphi}_g \in \arg\min_{\varphi \in \Phi_n^g} \widehat{\mathrm{MSE}}_{CV}(\varphi)$*. Define the fold-average population prediction risk of the fold-trained estimators as:* $R_K^g(\varphi) := \frac{1}{K}\sum_{k=1}^K \mathbb{E}\Big[\|U - \widehat{g}^{(-k)}_\varphi(V)\|_2^2\Big]$.

[5]The $o(1)$ term depends only on n and $|\Phi_n^\alpha|$.

Under standard squared-loss analogues of Assumptions 1–4 (e.g. finite $q > 4$ moments and a Bernstein condition for the squared loss), the selected fold-trained *regression learner satisfies the oracle inequality:* $R_K^g(\widehat{\varphi}_g) \le (1 + o(1)) \inf_{\varphi \in \Phi_n^g} R_K^g(\varphi) + C\frac{\log|\Phi_n^g|}{n}$, *with probability tending to one. Since* $g_0(V) = \mathbb{E}[U \mid V]$, *we have* $R_K^g(\varphi) - \mathbb{E}\|U - g_0(V)\|_2^2 = \|\widehat{g}_\varphi - g_0\|_{P,2}^2$. *Moreover, if the regression learners satisfy the following* refit-stability *condition (analogous to Assumption 5): there exists* $\delta_{g,n} = o(n^{-1/2})$ *such that* $\sup_{\varphi \in \Phi_n^g} \left| \mathbb{E}\big[\|U - \widehat{g}_\varphi(V)\|_2^2\big] - R_K^g(\varphi) \right| \le \delta_{g,n}$ *in probability, then the same oracle inequality holds for the* refitted *estimator* $\widehat{g}_{\widehat{\varphi}_g}$ *(with an added* $2\delta_{g,n}$ *term):*

$$\mathbb{E}\Big[\big\|U - \widehat{g}_{\widehat{\varphi}_g}(V)\big\|_2^2\Big] - \mathbb{E}\big[\|U - g_0(V)\|_2^2\big] \le (1+o(1)) \inf_{\varphi \in \Phi_n^g} \Big\{ \|\widehat{g}_\varphi - g_0\|_{P,2}^2 \Big\} + C\,\frac{\log|\Phi_n^g|}{n} + 2\delta_{g,n}.$$

The lemma is applied conditionally within each outer fold of Algorithm 1; the outer-fold index is suppressed for notational simplicity.

Corollary 1 (Rates for the selected nuisances)**.** *Suppose Assumption 5 holds for the representer refit step and the regression refit-stability condition in Lemma 2 holds with* $\delta_{g,n} = o(n^{-1/2})$. *Suppose the libraries contain (possibly distinct) oracle elements with* $\|\hat{g}_{\varphi^*_{g,n}} - g_0\|_{P,2} = o_p(n^{-1/4})$ *and* $\|\hat{\alpha}_{\kappa^*_{\alpha,n}} - \alpha_0\|_{P,2} = o_p(n^{-1/4})$. *Then the selected (refitted) nuisances satisfy* $\|\hat{g}_{\hat{\varphi}_g} - g_0\|_{P,2} = o_p(n^{-1/4})$ *and* $\|\hat{\alpha}_{\hat{\kappa}_\alpha} - \alpha_0\|_{P,2} = o_p(n^{-1/4})$. *Consequently, the product-rate condition* $\|\hat{g}_{\hat{\varphi}_g} - g_0\|_{P,2}\,\|\hat{\alpha}_{\hat{\kappa}_\alpha} - \alpha_0\|_{P,2} = o_p(n^{-1/2})$ *required for Theorem 3 holds.*

4.3 Adaptive orthogonality under data-driven nuisance selection

Because the score (5) targets a linear functional, the population moment error admits an exact second–order representation. In particular, selecting $(\hat{g}, \hat{\alpha})$ from a finite library does not create an additional first–order drift.

Theorem 2 (Adaptive orthogonality under data-driven nuisance selection)**.** *Under Assumptions 1–4 and the sample–splitting scheme in Section 3, for each outer fold* ℓ *we have, conditional on the corresponding training sample,* $\mathbb{E}\big[\psi(W; \theta_0, \hat{g}_\ell, \hat{\alpha}_\ell)\big] = -\mathbb{E}\Big[\big(\hat{\alpha}_\ell(W) - \alpha_0(W)\big)^\top \big(\hat{g}_\ell(W) - g_0(W)\big)\Big]$, *and therefore* $\Big|\mathbb{E}\big[\psi(W; \theta_0, \hat{g}_\ell, \hat{\alpha}_\ell)\big]\Big| \le \|\hat{g}_\ell - g_0\|_{P,2}\,\|\hat{\alpha}_\ell - \alpha_0\|_{P,2}$.

Proof. See Appendix A.5. □

Remark 12 (Connection to structural decoupling)**.** The identity in Theorem 2 holds for any fold-specific nuisance pair $(\hat{g}_\ell, \hat{\alpha}_\ell)$ trained on an independent sample, including selections from the full product library $\Phi_n^g \times \Phi_n^\alpha$. In particular, no same-architecture restriction is required for adaptive orthogonality.

4.4 Asymptotic normality and efficiency

Because ψ is affine in θ with slope -1, cross–fitted estimator admits the closed form

$$\hat{\theta}_{\text{Adaptive}} = \frac{1}{n}\sum_{i=1}^{n}\Big[m\big(W_i, \hat{g}_{-i}\big) + \hat{\alpha}_{-i}(W_i)^\top\{U_i - \hat{g}_{-i}(W_i)\}\Big], \tag{13}$$

where $(\hat{g}_{-i}, \hat{\alpha}_{-i})$ denotes the nuisance pair trained on the fold that excludes observation i (including the inner selection and refitting step).

Theorem 3 (Asymptotic linearity, normality, and efficiency)**.** *Under Assumptions 1, 2, and 4 and the sample–splitting scheme, assume that the number of outer folds L is fixed. Suppose the fold–specific nuisance estimates satisfy the consistency condition*

$$\max_{1\le\ell\le L}\ \|\hat{g}_\ell - g_0\|_{P,2} = o_p(1) \quad \textit{and} \quad \max_{1\le\ell\le L}\ \|\hat{\alpha}_\ell - \alpha_0\|_{P,2} = o_p(1), \tag{14}$$

and the product–rate condition

$$\max_{1\le\ell\le L}\ \|\hat{g}_\ell - g_0\|_{P,2}\,\|\hat{\alpha}_\ell - \alpha_0\|_{P,2} = o_p(n^{-1/2}). \tag{15}$$

Let $\sigma^2_{\text{eff}} := \mathbb{E}[\psi(W;\theta_0,g_0,\alpha_0)^2] \in (0,\infty)$. Then $\sqrt{n}\big(\hat{\theta}_{\text{Adaptive}}-\theta_0\big) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \psi\big(W_i;\theta_0,g_0,\alpha_0\big)+ o_p(1)$, and hence $\sqrt{n}\big(\hat{\theta}_{\text{Adaptive}}-\theta_0\big) \xrightarrow{d} \mathcal{N}\big(0,\sigma^2_{\text{eff}}\big)$. Lemma 4 establishes that $\psi(\cdot;\theta_0,g_0,\alpha_0)$ is the unique efficient influence function for θ_0. Consequently, $\hat{\theta}_{\text{Adaptive}}$ attains the semiparametric efficiency bound σ^2_{eff} (Bickel et al., 1993; Severini and Tripathi, 2012).

Proof. See Appendix A.6. □

4.5 Variance estimation and inference

We employ the cross–fitted plug–in variance estimator:

$$\hat{\sigma}^2 := \frac{1}{n}\sum_{i=1}^{n}\big(\psi\Big(W_i;\hat{\theta}_{\text{Adaptive}},\hat{g}_{-i},\hat{\alpha}_{-i}\Big)\big)^2. \tag{16}$$

Lemma 3 (Consistency of variance estimator)**.** *Under the conditions of Theorem 3, the variance estimator is consistent:* $\hat{\sigma}^2 \xrightarrow{p} \sigma^2_{\text{eff}}$.

Proof. See Appendix A.7. □

Combining Theorem 3 with Lemma 3 yields the main inferential result.

Corollary 2 (Valid post-selection Wald coverage)**.** *The studentized statistic converges to a standard normal distribution:* $\frac{\sqrt{n}\left(\hat{\theta}_{\text{Adaptive}}-\theta_0\right)}{\hat{\sigma}} \xrightarrow{d} \mathcal{N}(0,1)$. *Consequently, the Wald interval* $CI_{1-\alpha} = \left[\hat{\theta}_{\text{Adaptive}} \pm z_{1-\alpha/2}\,\hat{\sigma}/\sqrt{n}\right]$ *has asymptotic coverage* $1-\alpha$*, preserving inference validity despite the data-driven selection of nuisance learners.*

Proof. Immediate from Theorem 3 and Lemma 3 by Slutsky's theorem. See Appendix A.7. □

4.6 Efficiency bounds for leading examples

When $\psi(\cdot;\theta_0,g_0,\alpha_0)$ is the efficient influence function, the asymptotic variance in Theorem 3 equals the semiparametric efficiency bound. We record closed–form expressions for three leading examples.

Example 1 (Average treatment effect)**.** For the ATE, the efficiency bound Hahn, 1998 is

$$\sigma^2_{\text{ATE}} = \mathbb{E}\left[\big(g_0(1,X)-g_0(0,X)-\theta_0\big)^2 + \frac{\text{Var}(Y \mid D=1,X)}{p_0(X)} + \frac{\text{Var}(Y \mid D=0,X)}{1-p_0(X)}\right].$$

Example 2 (Weighted average derivative)**.** For $\theta_0 = \mathbb{E}[w(X)\nabla_x g_0(X)]$, the efficiency bound Newey, 1994 is

$$\sigma^2_{\text{WAD}} = \mathbb{E}\Big[\big(w(X)\nabla_x g_0(X)-\theta_0\big)^2 + \text{Var}(Y \mid X)\,\alpha_0(X)^2\Big].$$

Example 3 (Difference-in-differences). Let $G \in \{0,1\}$ denote the treated group indicator and let ΔY be the two-period outcome change. Let $p := \Pr(G=1)$ and $p(X) := \Pr(G=1 \mid X)$, and define $\rho(X) := p(X)/(1-p(X))$. Writing $\mu_g(X) := \mathbb{E}[\Delta Y \mid G=g, X]$ and $\tau(X) := \mu_1(X) - \mu_0(X)$, the efficient score can be written in the form $\psi_0(W) = m(W,g_0) - \theta_0 + \alpha_0(W)(\Delta Y - g_0(W))$ with

$$m(W,g_0) = \frac{G}{p}\tau(X), \qquad \alpha_0(W) = \frac{G}{p} - \frac{1-G}{p}\rho(X), \qquad g_0(W) = \mu_G(X).$$

Therefore the semiparametric efficiency bound is

$$\sigma^2_{\mathrm{DiD}} = \mathbb{E}\left[\left(\tfrac{G}{p}\tau(X) - \theta_0\right)^2\right] + \mathbb{E}\big[\alpha_0(W)^2\ \mathrm{Var}(\Delta Y \mid G, X)\big]\,.$$

Under the conditions of Theorem 3 (in particular, the product-rate condition), the adaptive estimator is asymptotically efficient for ATE and DiD and therefore attains the corresponding semiparametric efficiency bounds. For WAD, the same conclusion holds when one works in a Sobolev-type Hilbert space in which the derivative functional is continuous (cf. Remark 3); all norms, Riesz representers, and product-rate conditions are then interpreted in that Hilbert geometry.

5 Finite–sample bounds

This section quantifies the accuracy of the Gaussian approximation in finite samples. Define $T_n := \sqrt{n}\big(\hat{\theta}_{\mathrm{Adaptive}} - \theta_0\big)$, $\quad S_n := \frac{1}{\sqrt{n}}\sum_{i=1}^n \psi\big(W_i; \theta_0, g_0, \alpha_0\big)$, $\quad R_n := T_n - S_n$. By Theorem 3, $R_n = o_p(1)$, but in finite samples the magnitude of R_n governs the error of the normal approximation.

Theorem 4 (Berry–Esseen bound for Adaptive Riesz-DML)**.** *Assume the conditions of Theorem 3 and suppose additionally that* $\mathbb{E}|\psi(W;\theta_0,g_0,\alpha_0)|^3 < \infty$ *and* $\sigma^2_{\mathrm{eff}} > 0$. *Then for any* $\varepsilon > 0$,

$$\sup_{x\in\mathbb{R}}\left|\Pr\left(\frac{T_n}{\sigma_{\mathrm{eff}}} \le x\right) - \Phi(x)\right| \le \frac{C_{\mathrm{BE}}}{\sqrt{n}}\,\frac{\mathbb{E}|\psi(W;\theta_0,g_0,\alpha_0)|^3}{\sigma^3_{\mathrm{eff}}} + \frac{\varepsilon}{\sqrt{2\pi}} + \Pr\left(\frac{|R_n|}{\sigma_{\mathrm{eff}}} > \varepsilon\right),$$

where C_{BE} *is the universal Berry–Esseen constant (e.g., van der Vaart, 1998).*

Proof. See Appendix A.8. □

Remark 13 (Controlling the remainder R_n)**.** The term $\Pr(|R_n|/\sigma_{\mathrm{eff}} > \varepsilon)$ collects the contribution of nuisance estimation. The remainder R_n admits a fold-wise decomposition over the L evaluation folds: $R_n = \sqrt{n}\cdot\frac{1}{L}\sum_{\ell=1}^L \Big\{(P_{n,\ell} - P)\big[\psi(W;\theta_0,\hat{g}_\ell,\hat{\alpha}_\ell) - \psi_0(W)\big] + P\big[\psi(W;\theta_0,\hat{g}_\ell,\hat{\alpha}_\ell) - \psi_0(W)\big]\Big\}$. The drift term (second term in braces) is bounded by $\|\hat{g}_\ell - g_0\|_{P,2}\|\hat{\alpha}_\ell - \alpha_0\|_{P,2}$ via the Riesz representation. Consequently, a sufficient condition for the remainder to vanish is the product-rate condition $\max_\ell \|\hat{g}_\ell - g_0\|_{P,2}\|\hat{\alpha}_\ell - \alpha_0\|_{P,2} = o_p(n^{-1/2})$, combined with standard empirical process control for the fluctuation term.

Corollary 3 (Coverage error for the reported studentized CI)**.** *Suppose the conditions of Theorem 4 hold and that the variance estimator $\hat{\sigma}^2$ is ratio-consistent for the asymptotic variance σ_{eff}^2, i.e. $\hat{\sigma}/\sigma_{\text{eff}} \to_p 1$. Then the usual studentized Wald interval*

$$\text{CI}_{0.95} := \left[\hat{\theta} \pm 1.96 \frac{\hat{\sigma}}{\sqrt{n}}\right]$$

satisfies

$$|\Pr(\theta_0 \in \text{CI}_{0.95}) - 0.95| \;\le\; C \cdot n^{-1/2} + o(1),$$

where C is the Berry–Esseen constant from Theorem 4.

Corollary 3 connects the Berry–Esseen bound directly to the confidence intervals reported in Sections 6 and 7. The last probability term isolates all finite-sample error induced by nuisance estimation and data-driven selection through the remainder R_n. In particular, if one can guarantee $\Pr(|R_n| > \sigma_{\text{eff}}\varepsilon_n) \to 0$ for some $\varepsilon_n \downarrow 0$, then the normal approximation error is $O(n^{-1/2}) + o(1)$, with the leading $n^{-1/2}$ constant governed by the third moment of the efficient influence function.

6 Simulations

We study the finite-sample performance of the Adaptive Riesz-DML estimator in controlled experiments. The target parameter is ATE $\theta_0 = \mathbb{E}\big[Y(1) - Y(0)\big]$, and we use the orthogonal score function $\psi(W;\theta,g,\alpha) \;=\; m(W,g) - \theta + \alpha(W)\{Y - g(D,X)\}$, with $m(W,g) = g(1,X) - g(0,X)$. The key object in our procedure is the Riesz representer α_0, which satisfies the Riesz identity $\mathbb{E}[m(W,h)] = \mathbb{E}[\alpha_0(W)h(W)]$ for all admissible directions h. In all designs below we set the true ATE to $\theta_0 = 1$ and evaluate bias, RMSE, and coverage of nominal 95% Wald confidence intervals. We partition the sample indices $\{1,\dots,n\}$ into L folds $(I_\ell)_{\ell=1}^L$. For each observation $i \in I_\ell$, let $\hat{g}_{-\ell}$ and $\hat{\alpha}_{-\ell}$ denote the nuisance functions estimated on the complementary sample I_ℓ^c. The estimator is constructed as $\hat{\theta} = n^{-1}\sum_{\ell=1}^L \sum_{i\in I_\ell} \hat{\phi}_i$, where $\hat{\phi}_i := m(W_i,\hat{g}_{-\ell}) + \hat{\alpha}_{-\ell}(W_i)\big(Y_i - \hat{g}_{-\ell}(W_i)\big)$. Standard errors[6] are computed using the centered score estimates $\hat{\psi}_i := \hat{\phi}_i - \hat{\theta}$: $\widehat{\text{se}}(\hat{\theta}) = \frac{1}{\sqrt{n}}\sqrt{\frac{1}{n-1}\sum_{i=1}^n \hat{\psi}_i^2}$, $\text{CI}_{0.95} = \hat{\theta} \pm 1.96\,\widehat{\text{se}}(\hat{\theta})$.

[6]For notational clarity, $\hat{\phi}_i$ denotes the uncentered score contribution, while $\hat{\psi}_i = \hat{\phi}_i - \hat{\theta}$ corresponds to the centered influence-function estimate used for variance estimation. These are algebraically equivalent representations of the same orthogonal score.

We use the finite-sample Bessel correction $(n-1)^{-1}\sum_{i=1}^{n}\hat{\psi}_i^2$ when reporting standard errors; this is asymptotically equivalent to $n^{-1}\sum_{i=1}^{n}\hat{\psi}_i^2$ analyzed in Lemma 3.

6.1 Library of learners

We use a rich finite library of candidate learners for the regression nuisance $g_0(d,x) = \mathbb{E}[Y \mid D=d, X=x]$ and for the Riesz representer α_0. We estimate g_0 using supervised learning on features (D,X) and select the algorithm (and, when applicable, its regularization parameter) by K-fold cross-validation minimizing MSE. The candidate set includes linear learners (Lasso, Ridge, Elastic Net, Linear SVR) and nonlinear learners (Decision Trees, Random Forests, GBM, RBF-SVR, and a feedforward neural network; XGBoost is included when available in the computing environment). Nonlinear learners are run with stable default hyperparameters (e.g. bounded depth trees) to keep computation tractable.

We construct a hybrid library consisting of plugin and direct riesz estimators. For each probabilistic classifier $\widehat{\pi}(X)$ for the propensity score $p_0(X) = \mathbb{P}(D=1 \mid X)$ in the propensity library, we form the plugin ATE representer $\hat{\alpha}_{\text{plugin}}(W) = \frac{D}{\widehat{\pi}(X)} - \frac{1-D}{1-\widehat{\pi}(X)}$, with $\widehat{\pi}(X)$ clipped away from $\{0,1\}$ to ensure finite weights.The propensity library mirrors the flexibility of the g-library. We implement *Riesz ridge regression*, which directly minimizes the empirical Riesz risk over a polynomial basis in X. In the ATE setting, each representer candidate is a function $\alpha(d,x)$ on $\{0,1\}\times\mathcal{X}$ with *two* treatment-state components, and we parameterize both $\alpha(1,\cdot)$ and $\alpha(0,\cdot)$ using the same polynomial basis in X (with separate ridge-penalized coefficient vectors), over a small grid of degrees and ridge penalties. For each representer candidate α, we write its observed evaluation as $\alpha(W) = \alpha(D,X)$. In the ATE design, the linear functional acts on any such function $h(d,x)$ via $m(W,h) = h(1,X) - h(0,X)$, so the corresponding Riesz risk specializes to $R(\alpha) = \mathbb{E}[\alpha(D,X)^2 - 2\{\alpha(1,X) - \alpha(0,X)\}]$. We estimate $R(\alpha)$ by cross-validation and select $\hat{\alpha}$ as the minimizer of the cross-validated Riesz risk over the union of plugin and direct-Riesz candidates. The final estimator uses outer-fold cross-fitting: nuisance learning (including model selection) is performed on the training folds only, and the score is evaluated on held-out folds to preserve independence.

For plugin ATE representers of the form $\hat{\alpha}(d,X) = \frac{d}{\widehat{p}(X)} - \frac{1-d}{1-\widehat{p}(X)}$, the functional term in the risk is evaluated by treating $\hat{\alpha}(\cdot,X)$ as a function of the counterfactual

treatment state and computing $m(W_i, \hat{\alpha}) = \hat{\alpha}(1, X_i) - \hat{\alpha}(0, X_i) = \frac{1}{\widehat{p}(X_i)} + \frac{1}{1-\widehat{p}(X_i)}$. Although only $D_i \in \{0, 1\}$ is observed, evaluating $m(W_i, \hat{\alpha})$ requires both treatment-state evaluations; this is immediate here because $\hat{\alpha}(d, X)$ is available in closed form once $\widehat{p}$ is learned (with $\widehat{p}$ clipped away from $\{0, 1\}$ to ensure finite weights).

6.2 Data-generating processes

We report results for two benchmark DGPs and a phase-transition design. All experiments use $n = 1000$ and dimension $d = 200$.

DGP 1 (Sparse Linear). We draw $X \in \mathbb{R}^d$ with $X \sim \mathcal{N}(0, I_d)$, choose a sparsity level $s = 5$, and draw sparse coefficient vectors $\theta, \beta \in [-1, 1]^s$ independently each replication. Define the propensity index $\eta_p(X) = X_{1:s}^\top(\beta/\sqrt{s})$, $p_0(X) = \text{expit}(\eta_p(X))$, and clip $p_0(X)$ to $[0.05, 0.95]$ to ensure overlap. Then draw $D \sim \text{Bernoulli}(p_0(X))$. The outcome satisfies $Y = 1 \cdot D + X_{1:s}^\top\theta + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, 1)$. This is a sparse, approximately linear setting in which linear learners for the regression nuisance g_0 (and for the propensity index) are expected to perform well. However, even in this regime the ATE Riesz representer $\alpha_0(W)$ is generally nonlinear in X because of its inverse-propensity form (e.g. $\alpha_0(W) = D/e_0(X) - (1-D)/(1-e_0(X))$ for the ATE); hence a fully linear *symmetric* pipeline that also constrains α to be linear in W (such as Fixed Lasso/Ridge) need not be optimal even at the sparse-linear endpoint.

DGP 2 (Highly Nonlinear). We draw $X \sim \text{Uniform}[-1, 1]^d$ with $d = 200$. Define the nonlinear propensity index $\eta_p(X) = \sin(\pi X_1 X_2) + X_3^2 - X_4 + 0.5X_5$, $p_0(X) = \text{expit}(\eta_p(X))$. We clip $p_0(X)$ to $[0.05, 0.95]$ to ensure overlap and draw $D \sim \text{Bernoulli}(p_0(X))$. The outcome is generated as $Y = \theta_0 D + \sin(X_1) + \cos(X_2) + (X_3 + X_4)^2 + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, 1)$, with $\theta_0 = 1$. This design features complex nonlinearities in both the treatment assignment and the outcome mechanism, challenging linear approximations.

6.3 Simulation Results

6.3.1 Main Results: Efficiency, Adaptivity, and Inference

This section evaluates the finite-sample behavior of the Adaptive Riesz-DML estimator under nuisance ambiguity, focusing on efficiency, adaptivity, and inferential

validity. Table 1 reports outer-fold selection frequencies for the outcome regression nuisance $\hat{g}$, which is chosen by minimizing inner-fold cross-validated mean squared error. In the sparse linear design (DGP 1), selection concentrates entirely on regularized linear learners, with Lasso selected in 89.6% of outer folds and Elastic Net in the remaining 10.4%. This pattern is consistent with the underlying sparse linear structure and reflects effective exclusion of high-variance nonlinear learners.

Table 1: Nuisance Learner Selection Frequencies $\hat{g}$ (Percentage of Outer Folds)

Candidate Learner	**DGP 1: Sparse Linear**	**DGP 2: Highly Nonlinear**
Linear Class (Total: 100.0% / 40.4%)		
Lasso	89.6%	36.1%
Elastic Net	10.4%	4.3%
Ridge	0.0%	0.0%
Linear SVR	0.0%	0.0%
Nonlinear Class (Total: 0.0% / 59.6%)		
XGBoost	0.0%	31.9%
GBM	0.0%	26.7%
Random Forest	0.0%	1.0%
Neural Network	0.0%	0.0%
Decision Tree	0.0%	0.0%
Kernel SVR (RBF)	0.0%	0.0%

Note: Entries report the selection frequency of the outcome regression learner $\hat{g}$. In DGP 2, the procedure selects nonlinear methods in the majority of cases ($\approx 60\%$), but retains linear estimators in cases where finite-sample variance reduction is prioritized over bias reduction. Selection frequencies are reported for the *outcome regression* learner g under the K-fold MSE criterion described above. This criterion is aligned with prediction of U but is not, in finite samples, identical to minimizing the RMSE of the target θ, which depends jointly on (g, α) and also reflects the additional nested sample-splitting used for selection. Consequently, in the correctly specified sparse-linear DGP1, a well-tuned fixed linear pipeline (notably Fixed LinearSVR in Table 2) can outperform the adaptive procedure in RMSE even if LinearSVR is rarely selected for g by MSE. The main advantage of the adaptive procedure is robustness when nonlinear structure is present, as seen in DGP2.

In the highly nonlinear design (DGP 2), selection shifts toward flexible learners: XGBoost (31.9%) and GBM (26.7%) together account for nearly 60% of selections. Nevertheless, linear specifications remain selected in 40.4% of folds, indicating a finite-sample bias–variance trade-off in which linear approximations are retained when they reduce estimation variance. The representer $\hat{\alpha}$ is selected independently using inner-fold cross-validated Riesz risk over plugin and direct-Riesz candidates, allowing the weighting architecture to adapt separately from the outcome regression.

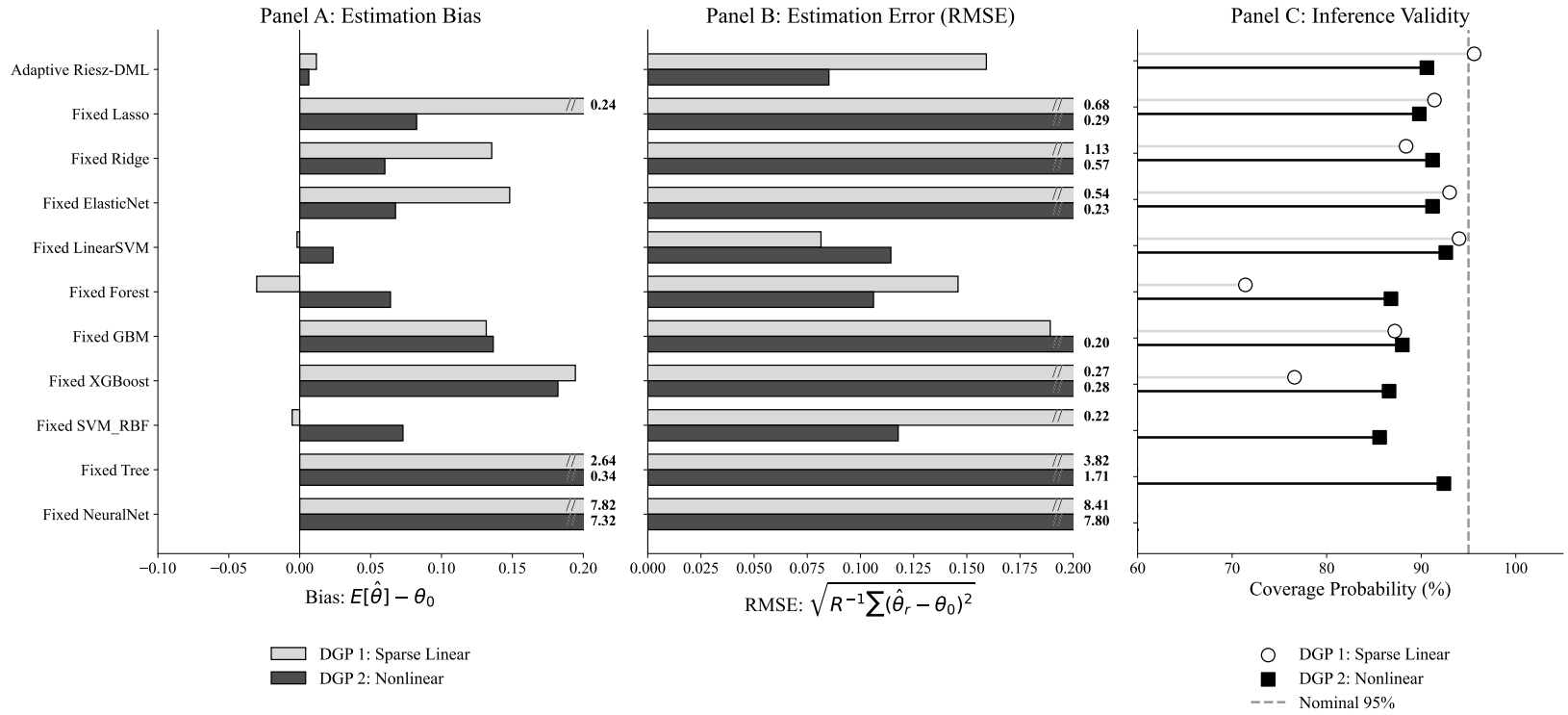

Figure 1: **Finite-Sample Performance: Efficiency and Inference.** Performance of the Adaptive Riesz-DML estimator compared to fixed architectures across $R = 500$ replications with $N = 1000$. Panel A and Panel B report Bias and RMSE (relative to $\theta_0 = 1$). Bars are truncated at 0.20 for visibility; large outliers (e.g., Fixed NeuralNet (default)) are annotated with their true values. In the sparse-linear design (DGP1), the adaptive estimator is competitive but can be outperformed by a well-specified fixed linear pipeline (notably Fixed LinearSVR). Under nonlinear structure (DGP2), the adaptive procedure substantially improves over fixed baselines. Panel C displays coverage probabilities for nominal 95% confidence intervals (dashed line). The Adaptive estimator delivers near-nominal coverage in the sparse linear design and substantially improved (though still below nominal) coverage in the nonlinear design at $n = 1000$. This gap is consistent with finite-sample effects emphasized by the Berry–Esseen bound in Section 5.

Figure 1 summarizes the efficiency and inferential consequences of this adaptivity. In the fully linear DGP, some fixed linear baselines attain lower RMSE than the adaptive procedure, as expected when the model class is correctly specified[7]. The purpose of Adaptive Riesz-DML is robustness across regimes: it preserves (near) nominal coverage and avoids large performance deterioration when the DGP deviates from linearity, where several fixed learners exhibit substantial undercoverage and/or higher RMSE. For example, Fixed Forest attains RMSE 0.146 but covers only 71.4% of nominal 95% confidence intervals. The Adaptive Riesz-DML estimator achieves RMSE 0.159 while maintaining near-nominal coverage (95.6%), providing a robust compromise between efficiency and inference. In the highly nonlinear design, Adaptive Riesz-DML achieves the lowest RMSE among all considered methods (0.085), improving upon the best fixed alternative (Fixed Forest, RMSE 0.106). Coverage in this regime is 90.6%, which is below nominal but substantially higher than the severe under-coverage observed for highly unstable fixed learners such as neural networks (coverage below 10%). Overall, the results demonstrate that Adaptive Riesz-DML effectively safeguards inference while adapting to structural complexity without requiring ex ante architectural commitments.

6.3.2 Phase Transition: Performance Along a Smooth Complexity Path

To visualize how performance evolves as the propensity and outcome indices move continuously from sparse-linear to highly nonlinear, we consider a phase-transition design indexed by $\gamma \in [0, 1]$. At $\gamma = 0$ the indices are sparse-linear and at $\gamma = 1$ they follow the nonlinear specification; intermediate values convexly mix the indices. Note that—even at $\gamma = 0$—the ATE representer is an inverse-propensity weight and is therefore typically nonlinear in X, so a symmetric linear pipeline need not be optimal at the 'linear-index' endpoint. Figure 2 reports RMSE and empirical coverage along this path for Adaptive Riesz–DML and two fixed symmetric pipelines

[7]In DGP1, the Fixed LinearSVR pipeline attains the smallest RMSE for $\hat{\theta}$ even though the inner-CV selector for the regression nuisance g almost never chooses LinearSVR. This is not contradictory: the g selector is based on prediction risk (MSE), whereas the finite-sample DML error for $\hat{\theta}$ is governed by the interaction of the two nuisance errors through the orthogonal score, in particular the product $\|\hat{g} - g_0\|_{P,2} \, \|\hat{\alpha} - \alpha_0\|_{P,2}$. In DGP1, several linear outcome learners deliver very similar MSE, and our stable selection rule favors sparse linear models when their MSE is statistically indistinguishable from more complex alternatives. A symmetric LinearSVR pipeline can nevertheless yield a smaller $\hat{\theta}$ RMSE if its representer estimate $\hat{\alpha}$ is especially accurate (or yields a favorable bias–variance tradeoff) in that design.

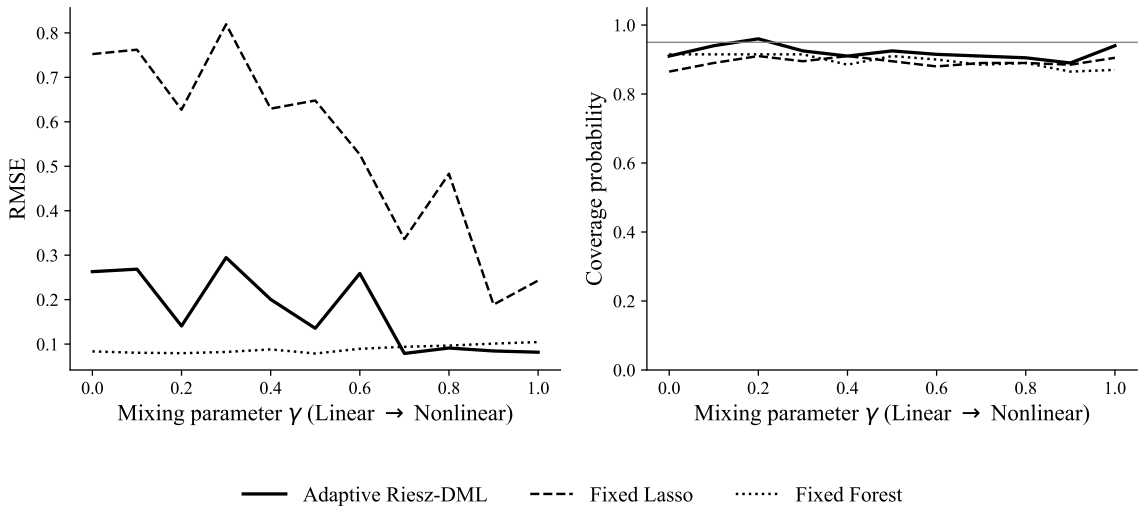

Figure 2: **Phase-transition experiment.** Left panel: RMSE of $\hat{\theta}$ along the mixing parameter $\gamma \in [0, 1]$. Right panel: empirical coverage of nominal 95% confidence intervals. The solid curve is Adaptive Riesz–DML, the dashed curve is Fixed Lasso, and the dotted curve is Fixed Forest. Although $\gamma = 0$ corresponds to sparse indices, the ATE representer is still an inverse-propensity weight and thus nonlinear in X.

(Fixed Lasso and Fixed Forest). In this design, Fixed Forest attains the lowest RMSE throughout the path, whereas Fixed Lasso[8] exhibits substantially larger errors near the sparse-linear endpoint. Adaptive Riesz–DML markedly improves over the fixed linear pipeline and delivers stable coverage close to the nominal level across γ. These results illustrate that (i) the identity of the best fixed architecture can be design dependent, and (ii) data-driven selection stabilizes performance and inference relative to committing to a single pipeline ex ante.[9]

Panel B reveals a critical insight regarding inference. While the Fixed Random Forest achieves reasonable prediction error in the linear regime, its coverage probability drops significantly below the nominal 95% level. This under-coverage likely

[8]Throughout, "Fixed Lasso/Ridge" refers to a *symmetric* pipeline that fits the Riesz representer α *within the same (sparse) linear learner class* by minimizing the Riesz loss over $W = (D, X)$, rather than a plug-in ATE-IPW construction that sets $\hat{\alpha}(W) = D/\hat{e}(X) - (1 - D)/(1 - \hat{e}(X))$ after estimating $\hat{e}$ by (say) lasso/logit. Even in DGP 1, $\alpha_0(W)$ is nonlinear in X because of the inverse-propensity form, so restricting α to be linear can be a substantial approximation error in finite samples and can inflate both dispersion and bias of the resulting DML estimator.

[9]This phase-transition experiment uses Uniform covariates $X \sim \text{Unif}([-1, 1]^p)$, whereas the benchmark results in Table 2 use Gaussian covariates $X \sim N(0, I_d)$. Hence the phase-transition plot is not intended to match the endpoint levels in Table 2 exactly; it is included to illustrate how RMSE and coverage evolve smoothly as the nuisance structure becomes more nonlinear.

reflects finite-sample deviations from the influence-function normal approximation for the fixed forest pipeline—for example, a non-negligible second-order remainder (drift) and/or heavier-tailed influence-function realizations and noisier variance estimation when $\hat{g}$ and $\hat{\alpha}$ are obtained from highly adaptive learners. While centering bias can contribute, in these simulations the forest pipeline's mean bias is small relative to its Monte Carlo dispersion, suggesting that variance underestimation and/or non-Gaussianity also play an important role. In contrast, the adaptive procedure tends to select more stable nuisance fits in this regime and delivers coverage closer to nominal.

Table 2: Main Simulation Results: Estimation Error and Coverage

	DGP 1: Sparse Linear			**DGP 2: Highly Nonlinear**		
Estimator	**Bias**	**RMSE**	**Cov 95%**	**Bias**	**RMSE**	**Cov 95%**
Adaptive Strategy						
Adaptive Riesz-DML	0.012	0.159	0.956	0.006	0.085	0.906
Fixed Benchmarks						
Fixed Boosting	0.132	0.189	0.872	0.136	0.201	0.880
Fixed ElasticNet	0.148	0.539	0.930	0.068	0.233	0.912
Fixed Forest	-0.030	0.146	0.714	0.064	0.106	0.868
Fixed Lasso	0.235	0.680	0.914	0.082	0.289	0.898
Fixed LinearSVR	-0.002	0.081	0.940	0.024	0.114	0.926
Fixed NeuralNet (default)	7.823	8.407	0.106	7.322	7.803	0.098
Fixed RBF SVR	-0.005	0.217	0.576	0.073	0.118	0.856
Fixed Ridge	0.135	1.130	0.884	0.060	0.566	0.912
Fixed Tree	2.641	3.824	0.550	0.344	1.715	0.924
Fixed XGBoost	0.194	0.268	0.766	0.182	0.279	0.866

Note: Bias, RMSE, and empirical coverage of nominal 95% confidence intervals across 500 replications. The Fixed benchmarks correspond to standard DML practice: outcome learners are tuned by MSE and propensity learners (where applicable) are tuned by log-loss. Adaptive Riesz-DML automatically adapts to the regime, maintaining valid coverage in the linear setting (where Forests undercover) and achieving the lowest RMSE in the nonlinear setting (where Linear models fail). Note the extreme failure of the *default/untuned* NeuralNet baseline across both regimes, illustrating that off-the-shelf high-complexity architectures can induce heavy-tailed sampling behavior when used without inference-targeted tuning.

6.3.3 Mechanism Stress Test: Off-Diagonal Optimality

A central implication of Definition 1 is that *symmetric* (diagonal) restrictions can be first-order suboptimal when the effective complexity of the regression nuisance g_0 differs from that of the Riesz representer α_0. To make this mechanism empirically un-

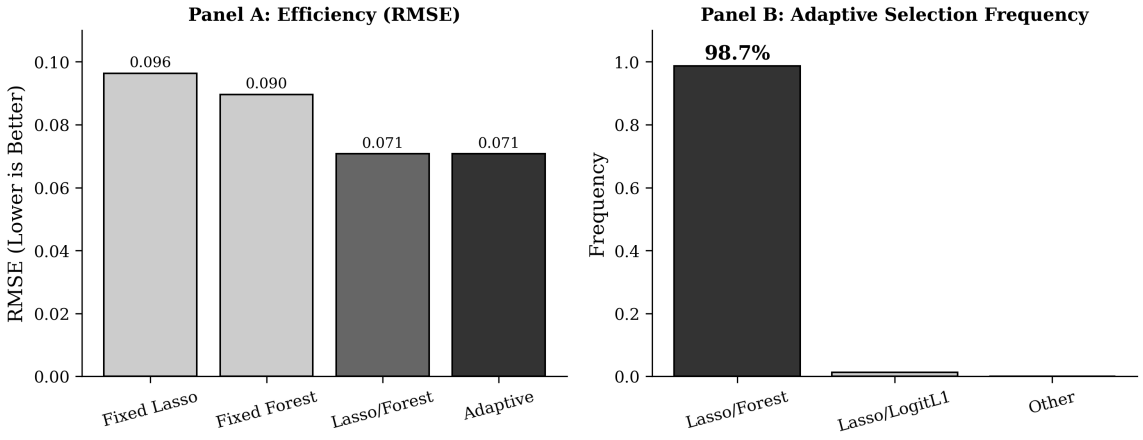

Figure 3: **Mechanism Stress Test Results.** Panel A: RMSE comparison ($N = 1000$). Symmetric architectures (Fixed Lasso, Fixed Forest) are outperformed by the Decoupled Oracle (Lasso/Forest). The Adaptive estimator matches the Oracle. Panel B: Selection frequency showing near-perfect convergence to the decoupled pair.

mistakable, we construct a design with Easy g_0 / Hard α_0[10] Figure 3 summarizes the results ($N = 1000$, 500 replications). Panel A confirms that symmetric architectures fail in this regime. The Fixed Lasso (symmetric linear) yields the highest error (RMSE ≈ 0.096) because the linear propensity fails to approximate the nonlinear Riesz representer. The Fixed Forest (symmetric nonlinear) reduces bias but remains inefficient (RMSE ≈ 0.090) due to overfitting the simple outcome. In contrast, Adaptive Riesz-DML effectively decouples the estimation problem, matching the performance of the Oracle Pair (Lasso/Forest) with an RMSE of 0.071. Panel B reveals the mechanism: the adaptive procedure correctly identifies the off-diagonal architecture (Lasso for g, Forest for p) in 98.7% of outer-fold selection decisions, thereby recovering the oracle efficiency without ex-ante knowledge of the structural asymmetry.

6.3.4 Distributional Properties and Finite-Sample Stability

Figure 4 presents the Empirical Cumulative Distribution Function (ECDF) of the absolute estimation error $|\hat{\theta} - \theta_0|$ for the full library of estimators. The split-axis visualization highlights two distinct classes of failure modes that fixed-architecture DML can suffer from, both of which are mitigated by the adaptive procedure. First,

[10]Let $X \in \mathbb{R}^d$ with $X \sim \mathcal{N}(0, I_d)$ and $d = 200$. Let $D \sim \text{Bernoulli}(p_0(X))$ with $p_0(X) = \Lambda(2\sin(3X_1X_2) + 1.5\exp(-X_3^2) + 0.5X_4 - 0.5X_5)$, where $p_0(X)$ is clipped to $[\underline{p}, \bar{p}]$ to enforce overlap and $\Lambda(\cdot)$ denotes the logistic link. The outcome regression is sparse linear: $Y = \theta_0 D + X^\top \beta_0 + \varepsilon$.

in the competitive regime (errors < 1.0, left panels), we observe the cost of systematic bias. In the nonlinear design (DGP 2), fixed linear estimators like Lasso (blue dashed line) flatten out early, indicating a high probability of moderate bias due to model misspecification. In contrast, the Adaptive Riesz-DML tracks the oracle-like performance of the best nonlinear learners, lying closest to the y-axis (zero error).

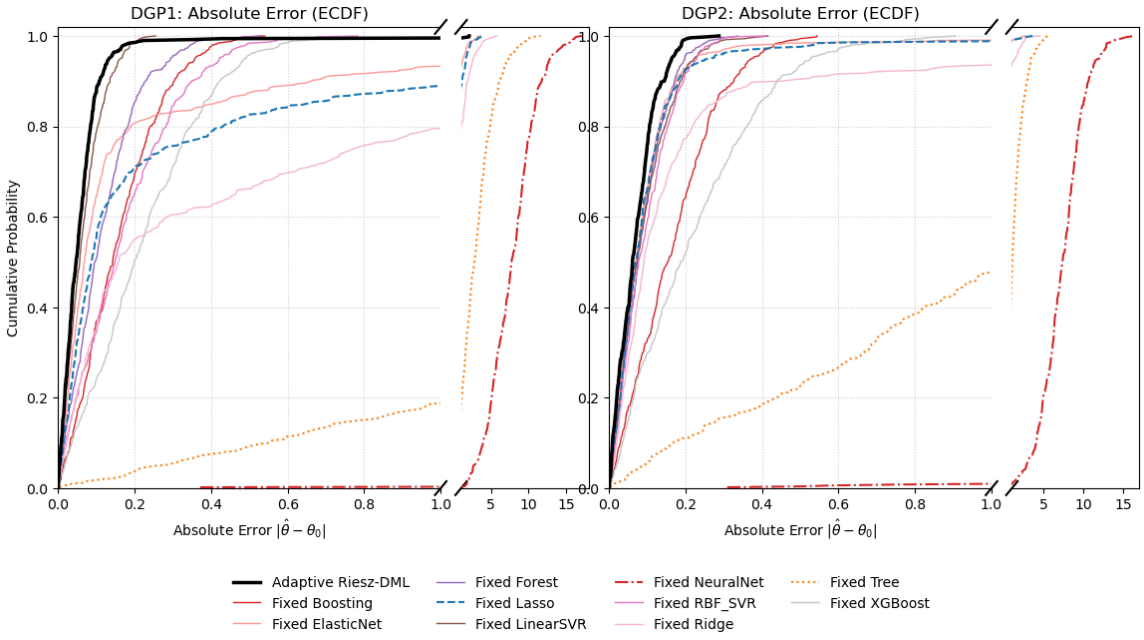

Figure 4: **Empirical CDF of absolute estimation errors.** The bold black curve plots the distribution of $|\hat{\theta} - \theta_0|$ for Adaptive Riesz–DML; dashed curves correspond to fixed symmetric pipelines. In the nonlinear design (DGP 2), the adaptive procedure is close to the best-performing fixed pipeline throughout the distribution and dominates fixed linear baselines. In the sparse-linear design (DGP 1), the best fixed linear pipeline can dominate in the extreme left tail (very small errors), but the adaptive estimator remains competitive and, importantly, avoids the heavy-tail behavior exhibited by overly complex fixed learners.

Second, the outlier regime (errors > 1.0, right panels) reveals the risk of catastrophic variance. Complex fixed learners, particularly the Fixed Neural Network (red dash-dot) and Decision Tree (orange dotted), exhibit heavy-tailed error distributions extending beyond 15.0 standard deviations. This confirms that blindly relying on high-complexity architectures without Riesz-aware selection is dangerous in finite samples. Crucially, the Adaptive estimator avoids this tail risk entirely. By selecting architectures based on the Riesz risk—which balances approximation bias against

variance—it effectively filters out these unstable candidates, delivering inference that is both accurate (low bias) and robust (low variance).

7 Empirical Application: 401(k) Eligibility and Wealth

To illustrate the empirical role of adaptive architecture selection, we revisit the canonical DML analysis of the effect of 401(k) eligibility on net financial assets (Chernozhukov et al., 2018). The application is well-suited to our framework because plausible fixed-learner specifications can yield materially different conclusions, a manifestation of the Rashomon effect (Breiman, 2001). We use the 1991 Survey of Income and Program Participation (SIPP) sample with $N = 9{,}915$ and target the Average Treatment Effect (ATE). We benchmark our results against the fixed-specification interactive regression estimates reported in Chernozhukov et al. (2018).

7.1 Implementation and Riesz-Consistent Selection

We implement Adaptive Riesz-DML with $L = 5$ outer folds for cross-fitting and nested $K = 5$ cross-validation within each training fold to select nuisance learners from a finite library including regularized linear learners and flexible nonlinear learners, together with parametric baselines. For the ATE, the efficient orthogonal score depends on the propensity score $p_0(X)$ through the representer $\alpha_0(W) = \frac{D}{p_0(X)} - \frac{1-D}{1-p_0(X)}$, so α_0 is a known functional of p_0. Accordingly, within each propensity architecture we tune hyperparameters by log-loss (a proper scoring rule), but we *select among architectures* using the cross-validated Riesz criterion evaluated on the implied weights $\hat{\alpha}(W) = D/\widehat{p}(X) - (1-D)/(1-\widehat{p}(X))$. In terms of Algorithm 1, each propensity *architecture* together with its internal log-loss tuning rule is treated as a single candidate procedure $\kappa \in \Phi_n^\alpha$: it outputs a tuned $\widehat{p}_\kappa$ on each training fold and the implied weights $\widehat{\alpha}_\kappa(W) = D/\widehat{p}_\kappa(X) - (1-D)/(1-\widehat{p}_\kappa(X))$. The inner-CV Riesz criterion is then applied *across* these tuned candidates. We do not additionally retune propensity hyperparameters by the Riesz criterion to avoid a computationally costly nested search, and because the log-loss tuned grids are already narrow in our implementation. Recall that for ATE the Riesz criterion uses $m(W, \alpha) = \alpha(1, X) - \alpha(0, X)$, so the empirical criterion is computed by evaluating the fitted representer at both treatment states for each X_i. Outcome regression learners are selected by held-out

MSE. All reported standard errors use the usual cross-fitted orthogonal-score variance estimate[11].

7.2 Results: Ambiguity under Fixed Learners, and Decoupled Adaptivity

Table 3 summarizes the results. Panel A reports the published benchmarks from Chernozhukov et al. (2018), where the Lasso specification relies on extensive feature engineering (polynomials and interactions). Panel B reports fixed-learner benchmarks using raw features. These fixed specifications span a wide range, from an imprecise estimate under Double Lasso to substantially larger effects under tree-based learners. This dispersion reflects sensitivity to *ex ante* architectural commitments when the two nuisance components exhibit different effective complexity levels.

Panel C reports the proposed Adaptive Riesz-DML estimate. The adaptive procedure delivers a single estimate that is close to the robust nonlinear fixed benchmarks, but is obtained without manual trial-and-error over learner classes. Importantly, adaptivity does not come with a large precision cost: the adaptive standard error is comparable to the most precise fixed nonlinear specifications in Panel B.

Panel D provides a transparent mechanism. In multiple folds, the selector chooses a structured (regularized linear) learner for the outcome regression $\hat{g}$ while selecting a flexible nonlinear learner for the propensity score $\widehat{p}$. This fold-wise pattern is consistent with *asymmetric nuisance complexity* in the 401(k) application and directly illustrates the value of structural decoupling in the sense of Definition 1. It also aligns with Proposition 2, which formalizes that symmetric (diagonal) restrictions such as Double Lasso or Double Forest can be first-order suboptimal when the best approximation regimes for g_0 and α_0 differ.[12]

Figure 5 unpacks the mechanism driving these efficiency gains by tracking the evolution of learner selection probabilities as the sample size N grows. The stacked areas

[11] We implement the stable ε-minimizer variant described in Section 3.6.

[12] A natural concern is whether the selection of nonlinear propensity models (e.g., Boosting) in Panel D is driven by the true data generating process or simply by a deficiency in the linear candidate set. Comparison with Panel A provides robustness. The Double Lasso benchmark from Chernozhukov et al. (2018) (Panel A) utilized extensive manual feature engineering (polynomials and interactions) to achieve an estimate of $7,170$. Our Adaptive procedure, utilizing only raw features, recovers a similar positive effect ($7,565$) by selecting tree-based methods. This suggests that the nonlinearity discovered by the Riesz-Risk selector serves as an automated substitute for manual feature engineering.

Table 3: **Empirical Results: Conventional vs. Adaptive DML**

Estimator Specification	**ATE**	**S.E.**	**Structural Assumption/Implication**
Panel A: Literature Benchmarks (Chernozhukov et al., 2018)			
Double Lasso (w/ interactions)	7,170	1,201	Sparsity in *engineered* features
Double Boosting (GBM / GBM)	7,713	1,155	Homogeneous Boosting
Double Random Forest (RF / RF)	8,105	1,242	Homogeneous Non-Linearity
Panel B: Fixed DML Benchmarks (This Paper, Raw Features)			
Double Lasso (No interactions)	2,723	3,827	*Fails* (Bias due to p_0 nonlinearity)
Double Boosting (GBM / GBM)	8,559	1,123	Robust Nonlinear Baseline
Double Random Forest (RF / RF)	7,969	1,115	Robust Nonlinear Baseline
Panel C: Adaptive Riesz-DML (Proposed Method)			
Adaptive Selector	**7,565**	**1,160**	**Decoupled / Endogenous Selection**
95% CI: [5,291, 9,839]			

Panel D: Anatomy of Adaptive Selection (Mechanism)			
Fold	**Outcome ($\hat{g}$)**	**Propensity ($\hat{p}$)**	**Selected Architecture Pattern**
1	Forest	Boosting	Nonlinear / Nonlinear
2	**Lasso**	**Boosting**	**Decoupled (Linear / Nonlinear)**
3	Forest	Forest	Nonlinear / Nonlinear
4	Forest	Forest	Nonlinear / Nonlinear
5	**Ridge**	**Forest**	**Decoupled (Linear / Nonlinear)**

Note: $N = 9,915$. Panel A reports the published benchmarks from Chernozhukov et al. (2018). Double Lasso is Linear DML (Lasso / Logit). Panel B reports our replication on raw features; the statistically insignificant estimate for Double Lasso (2,723) confirms that linear models fail without manual feature engineering. Panel C shows that Adaptive Riesz-DML recovers the robust positive effect (7,565) automatically. Panel D confirms the mechanism: the selector discovers *asymmetric complexity* in Folds 2 and 5.

represent the probability mass assigned to each learner class, visualizing the adaptive procedure's trajectory toward the oracle selector. A stark asymmetry emerges between the two nuisance functions. Panel B reveals a rapid *regime shift* in the propensity score model: while simple parametric benchmarks (Linear/Logit) have non-negligible weight at $N = 1,000$, the probability mass for nonparametric learners converges toward unity as N approaches the full sample size. This suggests that the treatment assignment mechanism—the decision to participate in a 401(k)—is highly nonlinear and cannot be adequately captured by sparse linear approximations.

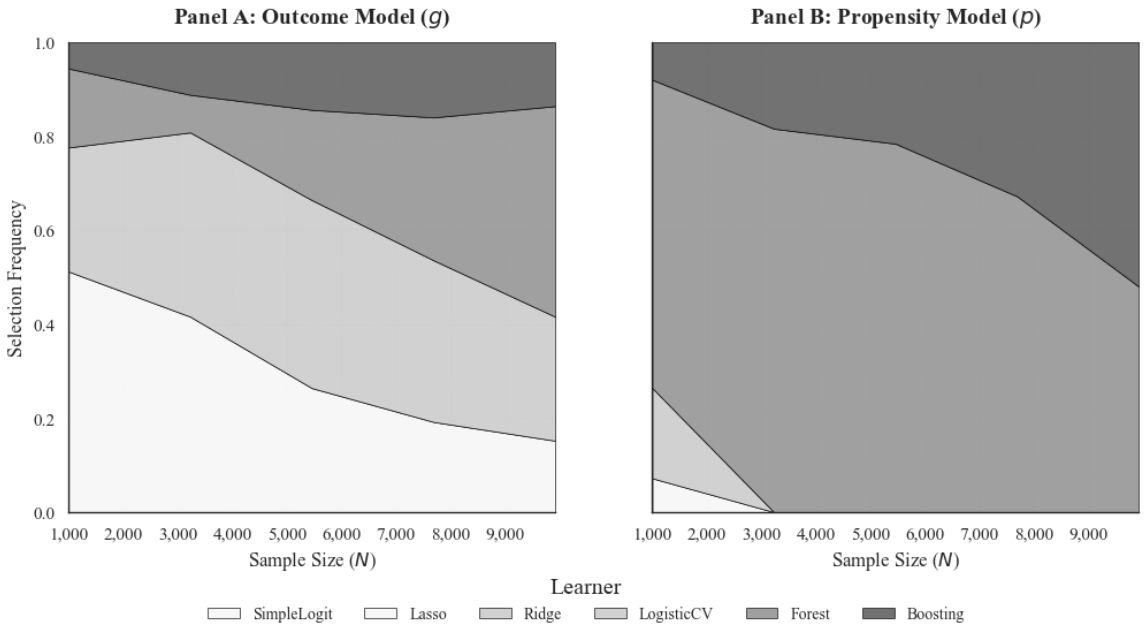

Figure 5: **Learner selection frequencies across sample size.** For each subsample size N, the adaptive procedure is run repeatedly; the figure reports the fold-level frequency with which each learner is selected for the outcome regression g (Panel A) and the propensity score p (Panel B). The stacked areas represent the cumulative probability mass assigned to each learner class, summing to 1.0. Panel A illustrates the persistence of sparsity, where regularized linear methods (lighter regions) retain a substantial selection share, illustrating the asymmetric complexity of the nuisance functions. Panel B illustrates a transition to nonlinearity, where the share of tree-based methods (darker regions) dominates as N grows.

Panel A exhibits structural stability. The outcome regression continues to allocate the majority of probability mass to regularized linear learners even at moderate to large sample sizes. This persistence implies that the conditional expectation of

net financial assets is well-approximated by a sparse linear combination of covariates. The ability of the adaptive estimator to decouple these choices—employing high-complexity models for p_0 while retaining low-variance linear models for g_0—prevents the regularization bias that would arise from a fixed linear architecture and the variance inflation that would arise from a fixed nonlinear architecture.

Figure 6 reports two applied diagnostics. Panel A shows a learning curve: dispersion shrinks quickly with N and the mean stabilizes near the full-sample value. Panel B shows that the full-sample Adaptive Riesz-DML estimate is stable across $K \in \{2, 5, 10\}$, suggesting limited sensitivity to the details of the nested selection stage. Additional inner-CV diagnostic plots that summarize the selection metrics used for $\hat{g}$ and $\widehat{p}$ are reported in Appendix D.

Panel A: Learning curve **Panel B: Sensitivity to inner folds**

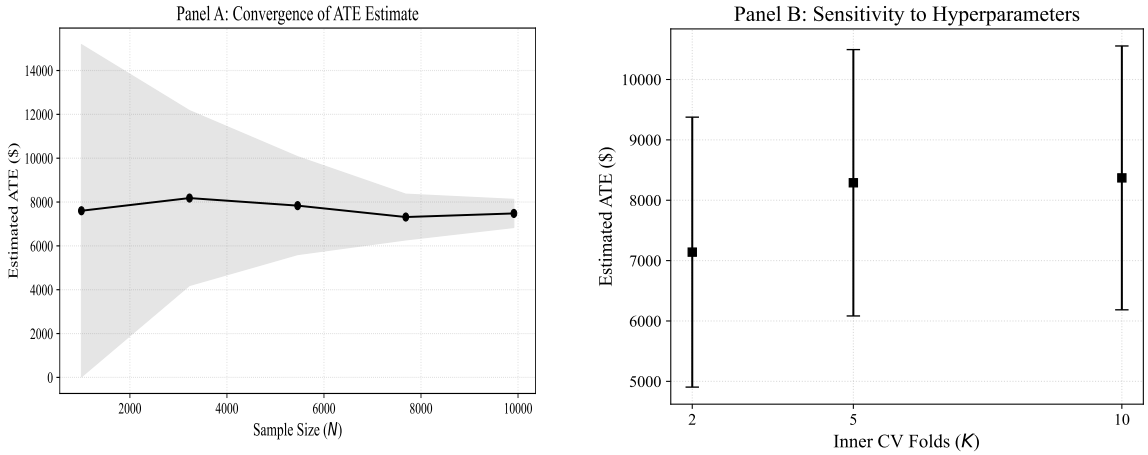

Figure 6: **Diagnostics for the Adaptive Riesz-DML ATE in the 401(k) application.** Panel A plots the mean ATE estimate across repeated subsamples, with the shaded band showing the 10th–90th percentiles (dispersion shrinks rapidly with N). Panel B shows that the full-sample estimate is stable across inner-CV fold choices $K \in \{2, 5, 10\}$.

7.3 Diagnosing the Instability: Riesz Representer Weights

For the ATE, the Neyman–orthogonal DML score can be written in terms of the inverse-propensity Riesz representer weight $\hat{\alpha}(W) = \frac{D}{\hat{p}(X)} - \frac{1-D}{1-\hat{p}(X)}$. Because $\hat{\alpha}(W)$ diverges as $\hat{p}(X) \to 0$ or $\hat{p}(X) \to 1$, the *tail behavior* of the implied representer weights provides a direct diagnostic for whether an estimator is implicitly relying on a small

set of high-leverage observations, and for whether a particular nuisance architecture (especially for $\hat{p}$) is effectively imposing strong regularization or implicit trimming. This diagnostic is tightly linked to our theory: the inferential object depends on the Riesz representer, so nuisance choices that look adequate for prediction can still be poorly aligned with the score geometry that governs inference.

Figure 7 visualizes the distribution of $\hat{\alpha}(W)$ under the baseline overlap stabilization used throughout ($\tau = 0.01$).[13] Panel A compares a symmetric linear benchmark (Fixed Linear: Lasso outcome regression with a simple logistic propensity; dashed gray) against Adaptive Riesz-DML (solid black), which selects the propensity learner by held-out Riesz risk. The Fixed Linear benchmark exhibits a noticeably more concentrated right tail on the displayed scale, consistent with a propensity architecture that does not generate many low-$\hat{p}(X)$ treated observations and therefore assigns less leverage to those units in the score. In the same baseline specification, the benchmark yields a much smaller and statistically insignificant estimate: 2,723 with SE 3,827 (Table 4, Panel A, $\tau = 0.01$). In contrast, Adaptive Riesz-DML delivers a precisely estimated positive effect of 7,565 with SE 1,160 under the same stabilization.

Table 4 reports a formal sensitivity analysis with respect to overlap stabilization, using two standard diagnostics. *Clipping* replaces $\hat{p}(X)$ by $\tilde{p}(X) \in [\tau, 1-\tau]$ and recomputes the orthogonal score using all observations. *Dropping* discards observations with $\hat{p}(X) \notin [\tau, 1-\tau]$ and recomputes the score on the retained subsample. These diagnostics are not proposed as preferred estimands; rather, they assess whether conclusions are sensitive to the most extreme implied weights. When dropping is applied, the resulting estimate can be interpreted as inference for a trimmed version of the target parameter on the retained (overlap) subpopulation; when clipping is applied, the procedure corresponds to a stabilized weighting scheme that limits the influence of extreme implied weights.

Two empirical patterns emerge. First, the Fixed Linear benchmark is highly sensitive to stabilization. Under clipping, its estimate rises from 2,723 at $\tau = 0.01$ to 4,878 at $\tau = 0.10$, and under dropping it rises further to 6,121 at $\tau = 0.10$, with corresponding changes in precision. Second, the estimates produced by standard DML and Adaptive Riesz-DML are comparatively stable under clipping: both remain

[13] In this application the fitted propensities lie within $[0.01, 0.99]$ for all three specifications, so the baseline bound is effectively non-binding; Table 4 considers more aggressive stabilization to stress the implied weights.

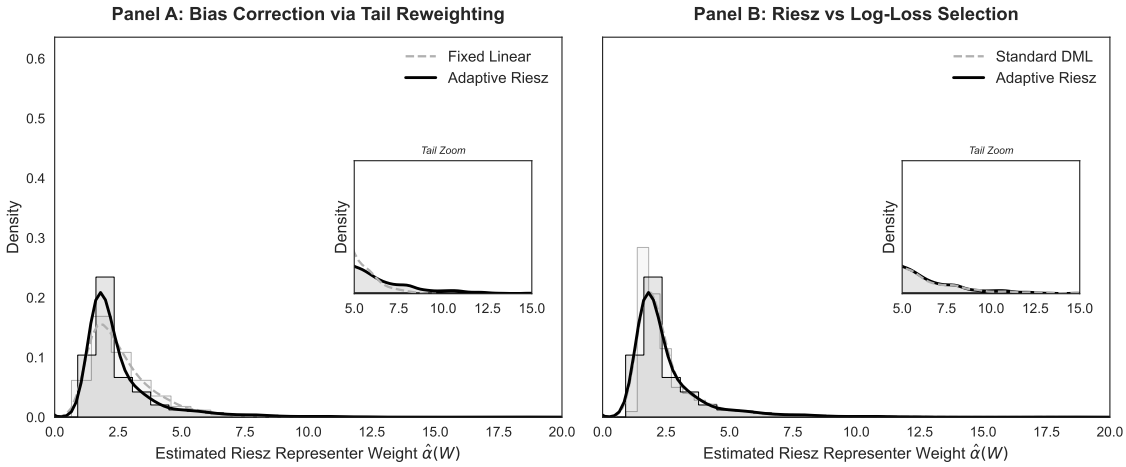

Figure 7: **Distribution of implied Riesz representer weights in the 401(k) application.** The figure plots the empirical distribution of the inverse-propensity representer weights $\hat{\alpha}(W) = D/\tilde{p}(X) - (1-D)/(1-\tilde{p}(X))$ constructed from cross-fitted propensity predictions, with baseline stabilization $\tilde{p}(X) = \min\{\max\{\hat{p}(X), 0.01\}, 0.99\}$. The horizontal axis is truncated to focus on the right tail (treated weights); the inset zooms in on $\hat{\alpha} \in [5, 15]$. Panel A compares a symmetric linear benchmark (Fixed Linear: dashed gray) with Adaptive Riesz-DML (solid black); the adaptive procedure yields a thicker right tail on the displayed scale. Panel B compares propensity selection by log-loss (Standard DML: dashed gray) with selection by Riesz risk (Adaptive Riesz-DML: solid black); the implied weight distributions are very similar in this application.

essentially unchanged across $\tau \in \{0.01, 0.05, 0.10\}$ with SEs around 1.16k.

Under dropping, both increase modestly at $\tau = 0.10$ (to 7,935 for Standard DML and 7,991 for Adaptive Riesz), reflecting the fact that removing the most extreme $\hat{p}(X)$ observations mechanically changes the score contributions of those regions. Importantly, the effective sample size (ESS) remains large for both methods throughout—for Adaptive Riesz, ESS is at least about 6,900 even when dropping at $\tau = 0.05$ and increases to about 7,200 at $\tau = 0.10$—indicating that inference is not driven by a vanishing set of extreme-weight observations.

Finally, Panel B of Figure 7 contrasts propensity selection by prediction (log-loss; dashed gray) with inference-targeted Riesz risk (solid black). The weight distributions are very similar in this application, which is consistent with the proximity of the corresponding ATEs in Table 4 at $\tau = 0.01$ (Standard DML: 7,501; Adaptive Riesz: 7,565). This alignment is not generic: the mechanism stress test in Section 6.3.3

Table 4: **Sensitivity of 401(k) ATE estimates to propensity-score stabilization.**

Method	Trim (τ)	ATE	SE	ESS	N
Panel A: Clipping					
Double Lasso	0.01	2,723	3,827	7,381	9,915
Double Lasso	0.05	3,853	2,872	7,466	9,915
Double Lasso	0.10	4,878	2,040	7,618	9,915
Standard DML	0.01	7,501	1,169	7,258	9,915
Standard DML	0.05	7,501	1,169	7,259	9,915
Standard DML	0.10	7,505	1,168	7,493	9,915
Adaptive Riesz	**0.01**	**7,565**	**1,160**	**6,914**	**9,915**
Adaptive Riesz	0.05	7,568	1,160	6,949	9,915
Adaptive Riesz	0.10	7,589	1,159	7,372	9,915
Panel B: Dropping					
Double Lasso	0.01	2,723	3,827	7,381	9,915
Double Lasso	0.05	5,643	1,871	7,547	9,907
Double Lasso	0.10	6,121	1,284	7,704	9,864
Standard DML	0.01	7,501	1,169	7,258	9,915
Standard DML	0.05	7,525	1,175	7,227	9,865
Standard DML	0.10	7,935	1,246	7,328	9,287
Adaptive Riesz	**0.01**	**7,565**	**1,160**	**6,914**	**9,915**
Adaptive Riesz	0.05	7,604	1,172	6,926	9,816
Adaptive Riesz	0.10	7,991	1,250	7,199	9,168

Notes: The table reports cross-fitted ATE estimates based on the orthogonal DML score. *Clipping* replaces the cross-fitted propensity by $\tilde{p}(X) = \min\{\max\{\hat{p}(X), \tau\}, 1-\tau\}$ and recomputes the score using all observations. *Dropping* discards observations with $\hat{p}(X) \notin [\tau, 1-\tau]$ and recomputes the score on the retained subsample (using the same cross-fitted nuisance predictions). Standard errors are computed as the sample standard deviation of the recomputed score divided by $\sqrt{N}$, where N is the number of retained observations. ESS is computed from the *positive* inverse-propensity weights $w_i = D_i/\tilde{p}(X_i) + (1-D_i)/(1-\tilde{p}(X_i))$ via $\text{ESS} = (\sum_i w_i)^2 / \sum_i w_i^2$ (using the retained sample under dropping). Double Lasso is Linear DML (Lasso / Logit). Because the stable ε-minimizer (one-standard-error rule) used for the reported fold-level learners may differ from the strict inner-CV argmin reported in Appendix D for diagnostic purposes, the diagnostic argmin in Appendix D need not coincide with the learner recorded for each outer fold.

shows that predictive selection can fail precisely when the representer is harder than the regression nuisance. The role of Riesz-risk selection is therefore not to guarantee a distinct choice in every application, but to provide a principled safeguard when prediction and inference objectives diverge.

8 Conclusion

A central message of semiparametric theory is that valid inference is driven by the geometry of the efficient influence function, not by predictive fit of auxiliary regressions. In DML implementations, however, the researcher typically commits *ex ante* to a particular double learning pipeline and tunes it using prediction-oriented criteria. For linear-functional targets, this practice is conceptually incomplete: the efficient score depends on a Riesz representer that is defined by a Hilbert-space duality (the Riesz identity) and is estimable via a functional-specific quadratic criterion, the Riesz risk. As a consequence, pre-determined predictive model selection can be weakly informative about the quality of the representer and, more importantly, symmetric architecture restrictions can be first-order by excluding the nuisance pair whose error product vanishes at the required $o_p(n^{-1/2})$ rate.

This paper proposes *Adaptive Riesz-DML*, a fully data-driven estimator that makes architecture choice an explicit part of the inferential problem. The method is structurally *decoupled*: it selects the regression nuisance g by cross-validated prediction loss, while selecting the representer α by cross-validated Riesz risk computed on held-out folds. This separation aligns the selection objective with each nuisance component's role in the score and permits off-diagonal pairs in the product library, thereby operationalizing the possibility that g_0 and α_0 inhabit different effective complexity regimes.

On the theory side, the estimator provides a fast-rate oracle inequality for Riesz-risk cross-validation over a growing finite library and shows that nested sample splitting preserves the orthogonal-score structure under non-parametric data-driven selection. For linear-functional scores, the population moment error satisfies the exact second-order identity $P\,\psi(W;\theta_0,g,\alpha) = \langle g-g_0, \alpha_0-\alpha\rangle_P$. Theorem 2 shows that this same second-order representation continues to hold fold-wise conditional on the training sample when $(\hat{g},\hat{\alpha})$ are chosen data-dependently from libraries via nested sample splitting; we refer to this post-selection second-order property as *adaptive orthogonal-*

ity. Under standard moment and product-rate conditions, the resulting estimator is asymptotically linear with influence function $\psi(\cdot; \theta_0, g_0, \alpha_0)$ and therefore attains the semiparametric efficiency bound.

Empirically, the adaptive DML estimator stabilizes inference across regimes where fixed architectures in conventional DML exhibit either misspecification-driven bias or high-variance failures, and the 401(k) application illustrates how the method can endogenize architectural choices that are otherwise a source of researcher degrees of freedom. In particular, the fold-level selections transparently reveal *asymmetric complexity* (e.g., nonlinear propensity/weight architectures paired with comparatively structured outcome regressions), providing a reproducible diagnostic for when symmetric DML pipelines are likely to be biased.

Several directions remain open. A natural extension is to sharpen finite-sample guarantees for large libraries (including computationally efficient search procedures) while retaining inference validity. Another is to broaden the analysis to settings with stronger ill-posedness and richer moment restrictions, where source conditions and regularization interact more tightly with representer estimation. Finally, further work on principled stabilization under limited overlap—viewed explicitly as a change in estimand or as a sensitivity analysis—would complement the representer-oriented diagnostics emphasized here.

References

Abadie, A. (2003): "Semiparametric Instrumental Variable Estimation of Treatment Response Models," *Journal of Econometrics*, 113, 231–263.

Argañaraz, F. (2025): "Automatic Debiased Machine Learning of Structural Parameters with General Conditional Moments," arXiv preprint arXiv:2512.08423.

Athey, S., J. Tibshirani, and S. Wager (2019): "Generalized Random Forests," *The Annals of Statistics*, 47, 1148–1178.

Athey, S., G. W. Imbens, and S. Wager (2018): "Approximate Residual Balancing: Debiased Inference of Average Treatment Effects in High Dimensions," *Journal of the Royal Statistical Society: Series B*, 80, 597–623.

Bach, P., O. Schacht, V. Chernozhukov, S. Klaassen, and M. Spindler (2024): "Hyperparameter Tuning for Causal Inference with Double Machine Learning: A Simulation Study," *Proceedings of Machine Learning Research*, 236, 1065–1117.

Belloni, A., and V. Chernozhukov (2011): "ℓ_1-Penalized Quantile Regression in High-Dimensional Sparse Models," *The Annals of Statistics*, 39, 82–130.

Bennett, A., N. Kallus, X. Mao, W. K. Newey, V. Syrgkanis, and M. Uehara (2023): "Source Condition Double Robust Inference on Functionals of Inverse Problems," arXiv preprint arXiv:2307.13793.

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

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009): "Simultaneous Analysis of Lasso and Dantzig Selector," *The Annals of Statistics*, 37, 1705–1732.

Breiman, L. (2001): "Statistical Modeling: The Two Cultures," *Statistical Science*, 16(3), 199–231.

Bruns-Smith, D., O. Dukes, A. Feller, and E. L. Ogburn (2025): "Augmented Balancing Weights as Linear Regression," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, forthcoming.

Chen, X., and D. Pouzo (2012): "Estimation of Nonparametric Conditional Moment Models with Possibly Nonsmooth Generalized Residuals," *Econometrica*, 80, 277–321.

Chen, X., and H. White (1999): "Improved Rates and Asymptotic Normality for Nonparametric Neural Network Estimators," *IEEE Transactions on Information Theory*, 45, 682–691.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018): "Double/Debiased Machine Learning for Treatment and Structural Parameters," *The Econometrics Journal*, 21, C1–C68.

Chernozhukov, V., W. K. Newey, and R. Singh (2022a): "Automatic Debiased Machine Learning of Causal and Structural Effects," *Econometrica*, 90(3), 967–1027.

Chernozhukov, V., W. K. Newey, and R. Singh (2022b): "Debiased Machine Learning of Global and Local Parameters Using Regularized Riesz Representers," *The Econometrics Journal*, 25(3), 576–601.

Chernozhukov, V., W. K. Newey, V. Quintas-Martínez, and V. Syrgkanis (2024): "RieszNet and ForestRiesz: Automatic Debiased Machine Learning with Neural Nets and Random Forests," arXiv preprint arXiv:2104.14737v3.

Chernozhukov, V., W. K. Newey, V. Quintas-Martínez, and V. Syrgkanis (2024): "Automatic Debiased Machine Learning via Riesz Regression," arXiv preprint arXiv:arXiv:2104.14737.

Cory-Wright, R., and A. Gómez (2025): "Stability Regularized Cross-Validation," arXiv preprint arXiv:2505.06927.

Hahn, J. (1998): "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 66, 315–331.

Hirshberg, D. A., and S. Wager (2021): "Augmented Minimax Linear Estimation," *The Annals of Statistics*, 49, 3406–3436.

Kallus, N., X. Mao, and M. Uehara (2024): "Localized Debiased Machine Learning: Efficient Inference on Quantile Treatment Effects and Beyond," *Journal of Machine Learning Research*, 25, 1–59.

Kato, M. (2025): "ScoreMatchingRiesz: Auto-DML with Infinitesimal Classification," arXiv preprint arXiv:2512.20523.

Luo, Y., M. Spindler, and J. Kueck (2025): "High-Dimensional L_2 Boosting: Rate of Convergence," *Journal of Machine Learning Research*, 26, 1–54.

Mullainathan, S., and J. Spiess (2017): "Machine Learning: An Applied Econometric Approach," *Journal of Economic Perspectives*, 31(2), 87–106.

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

Newey, W. K., and J. L. Powell (2003): "Instrumental Variable Estimation of Nonparametric Models," *Econometrica*, 71, 1565–1578.

Quintas-Martínez, V. M. (2025): "Machine Learning for Causal Estimation," Ph.D. Thesis, Massachusetts Institute of Technology.

Santos, A. (2011): "Instrumental Variable Methods for Recovering Continuous Linear Functionals," *Journal of Econometrics*, 161, 129–146.

Santos, A. (2012): "Inference in Nonparametric Instrumental Variables with Partial Identification," *Econometrica*, 80, 213–275.

Schmidt-Hieber, J. (2020): "Nonparametric Regression Using Deep Neural Networks with ReLU Activation Function," *The Annals of Statistics*, 48, 1875–1897.

Severini, T. A., and G. Tripathi (2012): "Efficiency Bounds for Estimating Linear Functionals of Nonparametric Regression Models with Endogenous Regressors," *Journal of Econometrics*, 170, 491–498.

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

van der Laan, M. J., E. C. Polley, and A. E. Hubbard (2007): "Super Learner," *Statistical Applications in Genetics and Molecular Biology*, 6(1).

Wager, S., and G. Walther (2016): "Adaptive Concentration of Regression Trees, with Application to Random Forests," arXiv preprint arXiv:1503.06388.

Zubizarreta, J. R. (2015): "Stable Weights that Balance Covariates for Estimation with Incomplete Outcome Data," *Journal of the American Statistical Association*, 110, 910–922.

A Proofs

A.1 Proof of Proposition 1

Let $\mathcal{H}$ denote the relevant Hilbert space. In the scalar case, $\mathcal{H} = L^2(P)$ with inner product $\langle f, h\rangle_P := \mathbb{E}[f(W)h(W)]$ and norm $\|f\|_P^2 := \mathbb{E}[f(W)^2]$. (If g is J-vector-valued, take $\mathcal{H} = L^2(P)^J$ with $\langle f, h\rangle_P := \mathbb{E}[f(W)^\top h(W)]$ and $\|f\|_P^2 := \mathbb{E}[\|f(W)\|_2^2]$.)

Define the continuous linear functional $T : \mathcal{H} \to \mathbb{R}$ by $T(g) := \mathbb{E}[m(W, g)]$. By the Riesz–Fréchet theorem, there exists a unique $\alpha_0 \in \mathcal{H}$ such that $T(g) = \langle \alpha_0, g\rangle_P \quad$ for all $g \in \mathcal{H}$.

For any $\alpha \in \mathcal{H}$, define the Riesz risk

$$\begin{aligned} R(\alpha) &:= \mathbb{E}\big[\|\alpha(W)\|_2^2 - 2\, m(W, \alpha)\big] \\ &= \|\alpha\|_P^2 - 2T(\alpha) \\ &= \|\alpha\|_P^2 - 2\langle \alpha_0, \alpha\rangle_P. \end{aligned} \tag{17}$$

Hence, $R(\alpha) - R(\alpha_0) = \big(\|\alpha\|_P^2 - 2\langle \alpha_0, \alpha\rangle_P\big) - \big(\|\alpha_0\|_P^2 - 2\langle \alpha_0, \alpha_0\rangle_P\big) = \|\alpha\|_P^2 + \|\alpha_0\|_P^2 - 2\langle \alpha_0, \alpha\rangle_P = \|\alpha - \alpha_0\|_P^2 \geq 0$. Moreover, $\|\alpha - \alpha_0\|_P^2 = 0$ if and only if $\alpha = \alpha_0$ P-a.s. Therefore, α_0 is the unique minimizer of $R(\alpha)$ over $\mathcal{H}$.

A.2 Proposition 2

Proposition 2 is an informal summary of the following rate-based statement.

Proposition 4 (Formal rate-based version of Proposition 2)**.** *Let Φ_n^g and Φ_n^α be index sets generating nuisance libraries $\{\hat{g}_{\varphi,n} : \varphi \in \Phi_n^g\}$ and $\{\hat{\alpha}_{\kappa,n} : \kappa \in \Phi_n^\alpha\}$. Fix indices $\varphi_1, \varphi_2 \in \Phi_n^g$ and $\kappa_1, \kappa_2 \in \Phi_n^\alpha$, and write $\hat{g}_{j,n} := \hat{g}_{\varphi_j,n}$ and $\hat{\alpha}_{j,n} := \hat{\alpha}_{\kappa_j,n}$ for $j = 1, 2$. Suppose there exist constants $0 < c \leq C < \infty$ and exponents $a_g, a_\alpha, b_g, b_\alpha \in [0, 1/2)$ such that, with probability tending to one,*

$$c\, n^{-a_g} \leq \|\hat{g}_{1,n} - g_0\|_{P,2} \leq C\, n^{-a_g}, \qquad c\, n^{-b_g} \leq \|\hat{g}_{2,n} - g_0\|_{P,2} \leq C\, n^{-b_g}, \tag{18}$$

$$c\, n^{-b_\alpha} \leq \|\hat{\alpha}_{1,n} - \alpha_0\|_{P,2} \leq C\, n^{-b_\alpha}, \qquad c\, n^{-a_\alpha} \leq \|\hat{\alpha}_{2,n} - \alpha_0\|_{P,2} \leq C\, n^{-a_\alpha}. \tag{19}$$

Assume the exponents satisfy

$$a_g + a_\alpha > \frac{1}{2}, \qquad a_g + b_\alpha < \frac{1}{2}, \qquad b_g + a_\alpha < \frac{1}{2}. \tag{20}$$

Consider the diagonal (symmetric) restriction $\mathcal{S}_n^{\mathrm{diag}} := \{(\varphi_1, \kappa_1), (\varphi_2, \kappa_2)\} \subseteq \Phi_n^g \times \Phi_n^\alpha$, *which corresponds to allowing only the coupled estimator pairs* $\{(\hat{g}_{1,n}, \hat{\alpha}_{1,n}), (\hat{g}_{2,n}, \hat{\alpha}_{2,n})\}$. *In contrast, the decoupled pair* $(\hat{g}_{1,n}, \hat{\alpha}_{2,n})$ *(corresponding to the off-diagonal index pair* (φ_1, κ_2)*) satisfies the product-rate condition.*

Proof. Work on the event where (18)–(19) hold.

(i) Symmetric pair $(1,1)$. By the lower bounds, $\|\hat{g}_{1,n} - g_0\|_{P,2} \, \|\hat{\alpha}_{1,n} - \alpha_0\|_{P,2} \geq c^2 \, n^{-(a_g + b_\alpha)}$. Since $a_g + b_\alpha < 1/2$, it follows that $\frac{\|\hat{g}_{1,n} - g_0\|_{P,2} \, \|\hat{\alpha}_{1,n} - \alpha_0\|_{P,2}}{n^{-1/2}} \geq c^2 \, n^{\frac{1}{2} - (a_g + b_\alpha)} \to \infty$, so the product cannot be $o(n^{-1/2})$.

(ii) Symmetric pair $(2,2)$. Similarly, by the lower bounds, $\|\hat{g}_{2,n} - g_0\|_{P,2} \, \|\hat{\alpha}_{2,n} - \alpha_0\|_{P,2} \geq c^2 \, n^{-(b_g + a_\alpha)}$. Since $b_g + a_\alpha < 1/2$, this product is not $o(n^{-1/2})$.

(iii) Decoupled pair $(1,2)$. By the upper bounds, $\|\hat{g}_{1,n} - g_0\|_{P,2} \, \|\hat{\alpha}_{2,n} - \alpha_0\|_{P,2} \leq C^2 \, n^{-(a_g + a_\alpha)}$. Since $a_g + a_\alpha > 1/2$, we have $n^{-(a_g + a_\alpha)} = o(n^{-1/2})$, so the product-rate condition holds. □

A.3 Proof of Lemma 1

Fix any (possibly random) $\alpha \in \mathcal{H}$ satisfying the boundedness condition, and define $\Delta(W) := \alpha(W) - \alpha_0(W)$. Recall $\ell(W, \alpha) = \|\alpha(W)\|_2^2 - 2m(W, \alpha)$. By linearity of m, $\ell(W, \alpha) - \ell(W, \alpha_0) = \|\alpha(W)\|_2^2 - \|\alpha_0(W)\|_2^2 - 2m(W, \alpha - \alpha_0) = (\alpha(W) + \alpha_0(W))^\top \Delta(W) - 2m(W, \Delta)$. Conditional on α, the law of W is P, so $\mathbb{E}[(\ell(W, \alpha) - \ell(W, \alpha_0))^2 \mid \alpha] = \|\ell(\cdot, \alpha) - \ell(\cdot, \alpha_0)\|_{P,2}^2$. Using the triangle inequality, Cauchy–Schwarz, and the bounds $\|\alpha\|_\infty, \|\alpha_0\|_\infty \leq B$, $\|\ell(\cdot, \alpha) - \ell(\cdot, \alpha_0)\|_{P,2} \leq \|(\alpha + \alpha_0)^\top \Delta\|_{P,2} + 2\|m(\cdot, \Delta)\|_{P,2} \leq (2B + 2C_m)\|\Delta\|_{P,2}$. Squaring this yields $\mathbb{E}[(\ell(W, \alpha) - \ell(W, \alpha_0))^2 \mid \alpha] \leq (2B + 2C_m)^2 \, \|\alpha - \alpha_0\|_{P,2}^2$. Since $R(\alpha) - R(\alpha_0) = \|\alpha - \alpha_0\|_{P,2}^2$, the Bernstein condition holds with $\nu = (2B + 2C_m)^2$. □

A.4 Proof of Theorem 1

Throughout the proof we invoke Assumption 4. Define the (pointwise) Riesz loss $\ell(W, \alpha) := \|\alpha(W)\|_2^2 - 2\, m(W, \alpha), \text{so that} R(\alpha) = \mathbb{E}\big[\ell(W, \alpha)\big]$.

Step 0: Setup, folds, measurability, and conditional independence. The sample is partitioned into K folds $\{I_k\}_{k=1}^K$. For each $\kappa \in \Phi_n^\alpha$, the fold-k estimator $\hat{\alpha}_\kappa^{(-k)}$ is trained using only observations indexed by $I_k^c := \{1, \ldots, n\} \setminus I_k$. Let $\mathcal{F}_k :=$

$\sigma\big(\{W_i : i \in I_k^c\}\big)$ denote the training σ–field in fold k. By construction, $\hat{\alpha}_\kappa^{(-k)}$ is $\mathcal{F}_k$–measurable. Since the data are i.i.d. and $I_k \cap I_k^c = \emptyset$, conditional on $\mathcal{F}_k$ the validation observations $\{W_i : i \in I_k\}$ are i.i.d. draws from P and are independent of $\hat{\alpha}_\kappa^{(-k)}$. We assume folds are approximately balanced: there exist constants $0 < c_- \le c_+ < \infty$ such that

$$c_- \frac{n}{K} \;\le\; |I_k| \;\le\; c_+ \frac{n}{K} \qquad \text{for all } k \in \{1, \dots, K\}. \tag{21}$$

Step 1: Work with *excess* CV risk. The K–fold cross-validated risk is $\widehat{R}_{CV}(\kappa) := \frac{1}{K}\sum_{k=1}^K \widehat{R}_k(\kappa), \widehat{R}_k(\kappa) := \frac{1}{|I_k|}\sum_{i\in I_k} \ell\Big(W_i, \hat{\alpha}_\kappa^{(-k)}\Big)$. Introduce the *excess* (oracle-centered) CV criterion $\widehat{\mathcal{E}}_{CV}(\kappa) := \widehat{R}_{CV}(\kappa) - \widehat{R}_{CV}(\alpha_0), \widehat{R}_{CV}(\alpha_0) := \frac{1}{K}\sum_{k=1}^K \frac{1}{|I_k|}\sum_{i\in I_k} \ell(W_i, \alpha_0)$. Since $\widehat{R}_{CV}(\alpha_0)$ does not depend on κ, any minimizer of $\widehat{R}_{CV}$ also minimizes $\widehat{\mathcal{E}}_{CV}$. Let $\hat{\kappa}_\alpha \in \arg\min_{\kappa\in\Phi_n^\alpha} \widehat{R}_{CV}(\kappa) = \arg\min_{\kappa\in\Phi_n^\alpha} \widehat{\mathcal{E}}_{CV}(\kappa)$.

Define the fold-average *population* excess risk $\bar{\mathcal{E}}_K(\kappa) := \frac{1}{K}\sum_{k=1}^K \mathcal{E}_k(\kappa), \qquad \mathcal{E}_k(\kappa) := R(\hat{\alpha}_\kappa^{(-k)}) - R(\alpha_0)$. Then, for any $\kappa \in \Phi_n^\alpha$,

$$\begin{aligned}
\bar{\mathcal{E}}_K(\hat{\kappa}_\alpha) &\le \widehat{\mathcal{E}}_{CV}(\hat{\kappa}_\alpha) + \big|\widehat{\mathcal{E}}_{CV}(\hat{\kappa}_\alpha) - \bar{\mathcal{E}}_K(\hat{\kappa}_\alpha)\big| \\
&\le \widehat{\mathcal{E}}_{CV}(\kappa) + \big|\widehat{\mathcal{E}}_{CV}(\hat{\kappa}_\alpha) - \bar{\mathcal{E}}_K(\hat{\kappa}_\alpha)\big| \\
&\le \bar{\mathcal{E}}_K(\kappa) + \big|\widehat{\mathcal{E}}_{CV}(\kappa) - \bar{\mathcal{E}}_K(\kappa)\big| + \big|\widehat{\mathcal{E}}_{CV}(\hat{\kappa}_\alpha) - \bar{\mathcal{E}}_K(\hat{\kappa}_\alpha)\big|.
\end{aligned} \tag{22}$$

In particular, letting $\kappa^* \in \arg\min_{\kappa\in\Phi_n^\alpha} \bar{\mathcal{E}}_K(\kappa)$,

$$\bar{\mathcal{E}}_K(\hat{\kappa}_\alpha) \le \bar{\mathcal{E}}_K(\kappa^*) + 2 \sup_{\kappa\in\Phi_n^\alpha} \big|\widehat{\mathcal{E}}_{CV}(\kappa) - \bar{\mathcal{E}}_K(\kappa)\big|. \tag{23}$$

Thus it remains to control the uniform deviation $\sup_\kappa |\widehat{\mathcal{E}}_{CV}(\kappa) - \bar{\mathcal{E}}_K(\kappa)|$.

Step 2: Centering and boundedness of the excess loss. Fix κ and fold k. Define the fold-k empirical excess loss average $\widehat{\mathcal{E}}_k(\kappa) := \frac{1}{|I_k|}\sum_{i\in I_k} \Big\{\ell\Big(W_i, \hat{\alpha}_\kappa^{(-k)}\Big) - \ell(W_i, \alpha_0)\Big\}$, so that $\widehat{\mathcal{E}}_{CV}(\kappa) = \frac{1}{K}\sum_{k=1}^K \widehat{\mathcal{E}}_k(\kappa)$. Conditional on $\mathcal{F}_k$, $\hat{\alpha}_\kappa^{(-k)}$ is fixed and $\{W_i : i \in I_k\}$ are i.i.d. from P, so

$$\mathbb{E}\Big[\widehat{\mathcal{E}}_k(\kappa) \mid \mathcal{F}_k\Big] = \mathbb{E}_W\Big[\ell\big(W, \hat{\alpha}_\kappa^{(-k)}\big) - \ell(W, \alpha_0)\Big] = R(\hat{\alpha}_\kappa^{(-k)}) - R(\alpha_0) = \mathcal{E}_k(\kappa). \tag{24}$$

Hence each $\widehat{\mathcal{E}}_k(\kappa) - \mathcal{E}_k(\kappa)$ is conditionally mean-zero. Next, by Assumption 1 there exists $B < \infty$ such that a.s. uniformly over (κ, k), $\big|\hat{\alpha}_\kappa^{(-k)}(W)\big| \le B$ and $\big|m(W, \hat{\alpha}_\kappa^{(-k)})\big| \le$

B and also $|\alpha_0(W)| \le B, \quad |m(W,\alpha_0)| \le B$. Therefore, for any such α, $|\ell(W,\alpha)| = \big|\|\alpha(W)\|_2^2 - 2m(W,\alpha)\big| \le B^2 + 2B =: M_0$, and hence

$$\big|\ell(W,\alpha)-\ell(W,\alpha_0)\big| \le 2M_0 \Rightarrow \Big|\big(\ell(W,\alpha)-\ell(W,\alpha_0)\big)-\mathbb{E}[\ell(W,\alpha)-\ell(W,\alpha_0)]\Big| \le 4M_0 =: b. \tag{25}$$

Step 3: Conditional variance control via the Bernstein condition. Fix κ and k. Define, for $i \in I_k$, $Z_{i,k}(\kappa) := \Big\{\ell\Big(W_i, \hat\alpha_\kappa^{(-k)}\Big) - \ell(W_i,\alpha_0)\Big\} - \mathcal{E}_k(\kappa)$. By (24), $\mathbb{E}[Z_{i,k}(\kappa) \mid \mathcal{F}_k] = 0$. Moreover, by (25), $|Z_{i,k}(\kappa)| \le b$ almost surely. Finally, by Assumption 4 (applied conditionally on $\mathcal{F}_k$, since $\hat\alpha_\kappa^{(-k)}$ is $\mathcal{F}_k$–measurable),

$$\begin{aligned}\mathrm{Var}\big(Z_{i,k}(\kappa) \mid \mathcal{F}_k\big) &\le \mathbb{E}\big[Z_{i,k}(\kappa)^2 \mid \mathcal{F}_k\big] \le \mathbb{E}\Big[\big(\ell(W,\hat\alpha_\kappa^{(-k)}) - \ell(W,\alpha_0)\big)^2 \mid \mathcal{F}_k\Big] \\ &\le \nu\Big\{R\big(\hat\alpha_\kappa^{(-k)}\big) - R(\alpha_0)\Big\} = \nu\mathcal{E}_k(\kappa).\end{aligned} \tag{26}$$

Step 4: Bernstein inequality + union bound (fast-rate uniform deviation). Conditional on $\mathcal{F}_k$, the variables $\{Z_{i,k}(\kappa) : i \in I_k\}$ are i.i.d. mean-zero, bounded by b, and have conditional variance at most $\nu\mathcal{E}_k(\kappa)$ by (26). Bernstein's inequality therefore yields, for any $t > 0$,

$$\mathbb{P}\left(\Big|\widehat{\mathcal{E}}_k(\kappa) - \mathcal{E}_k(\kappa)\Big| > \sqrt{\frac{2\nu\,\mathcal{E}_k(\kappa)\,t}{|I_k|}} + \frac{b\,t}{3|I_k|} \;\Big|\; \mathcal{F}_k\right) \le 2e^{-t}. \tag{27}$$

Take a union bound over $\kappa \in \Phi_n^\alpha$ and $k \in \{1,\dots,K\}$. Let $\delta \in (0,1)$ and set $t := \log\Big(\frac{2K|\Phi_n^\alpha|}{\delta}\Big)$. Then with probability at least $1-\delta$, simultaneously for all (κ,k),

$$\Big|\widehat{\mathcal{E}}_k(\kappa) - \mathcal{E}_k(\kappa)\Big| \le \sqrt{\frac{2\nu\,\mathcal{E}_k(\kappa)\,t}{|I_k|}} + \frac{b\,t}{3|I_k|}. \tag{28}$$

Average (28) over k and use (21). By Jensen's inequality (concavity of $\sqrt{\cdot}$), $\frac{1}{K}\sum_{k=1}^K \sqrt{\mathcal{E}_k(\kappa)} \le \sqrt{\frac{1}{K}\sum_{k=1}^K \mathcal{E}_k(\kappa)} = \sqrt{\bar{\mathcal{E}}_K(\kappa)}$. Also $\frac{1}{|I_k|} \lesssim \frac{K}{n}$ and $\frac{1}{\sqrt{|I_k|}} \lesssim \sqrt{\frac{K}{n}}$. Hence, on the same event, simultaneously for all $\kappa \in \Phi_n^\alpha$,

$$\Big|\widehat{\mathcal{E}}_{CV}(\kappa) - \bar{\mathcal{E}}_K(\kappa)\Big| \le C_1\sqrt{\bar{\mathcal{E}}_K(\kappa)\,\frac{t}{n}} + C_2\,\frac{t}{n}, \qquad t = \log\left(\frac{2K|\Phi_n^\alpha|}{\delta}\right), \tag{29}$$

for constants C_1, C_2 depending only on (ν, b, K, c_-, c_+).

Step 5: Solve the self-bounding inequality and obtain the oracle bound. Convert the empirical minimizer into a population oracle inequality. Let $\hat{\kappa}_\alpha \in \arg\min_{\kappa \in \Phi_n^\alpha} \widehat{\mathcal{E}}_K(\kappa)$ and let $\kappa^* \in \arg\min_{\kappa \in \Phi_n^\alpha} \overline{\mathcal{E}}_K(\kappa)$. Then, we have

$$\begin{aligned}
\overline{\mathcal{E}}_K(\hat{\kappa}_\alpha) &\le \widehat{\mathcal{E}}_K(\hat{\kappa}_\alpha) + \left|\overline{\mathcal{E}}_K(\hat{\kappa}_\alpha) - \widehat{\mathcal{E}}_K(\hat{\kappa}_\alpha)\right| \\
&\le \widehat{\mathcal{E}}_K(\kappa^*) + \left|\overline{\mathcal{E}}_K(\hat{\kappa}_\alpha) - \widehat{\mathcal{E}}_K(\hat{\kappa}_\alpha)\right| \\
&\le \overline{\mathcal{E}}_K(\kappa^*) + \left|\widehat{\mathcal{E}}_K(\kappa^*) - \overline{\mathcal{E}}_K(\kappa^*)\right| + \left|\overline{\mathcal{E}}_K(\hat{\kappa}_\alpha) - \widehat{\mathcal{E}}_K(\hat{\kappa}_\alpha)\right| \\
&\le \overline{\mathcal{E}}_K(\kappa^*) + C_1\sqrt{\overline{\mathcal{E}}_K(\kappa^*)\,\frac{t}{n}} + C_2\frac{t}{n} + C_1\sqrt{\overline{\mathcal{E}}_K(\hat{\kappa}_\alpha)\,\frac{t}{n}} + C_2\frac{t}{n}.
\end{aligned}$$

To remove the square-root term involving $\overline{\mathcal{E}}_K(\hat{\kappa}_\alpha)$ on the right-hand side, apply $2ab \le \eta a^2 + \eta^{-1}b^2$ with $a = \sqrt{\overline{\mathcal{E}}_K(\hat{\kappa}_\alpha)}$ and $b = C_1\sqrt{t/n}$ (e.g. $\eta = 1/2$), and then rearrange.

Step 6: Transfer to the refitted estimator. By Assumption 5, uniformly over $\kappa \in \Phi_n^\alpha$ we have $\left|R(\widehat{\alpha}_\kappa) - \overline{R}_K(\kappa)\right| \le \delta_n$ in probability. Therefore, $R(\widehat{\alpha}_{\widehat{\kappa}_\alpha}) \le \overline{R}_K(\widehat{\kappa}_\alpha) + \delta_n$, $\inf_{\kappa \in \Phi_n^\alpha} \overline{R}_K(\kappa) \le \inf_{\kappa \in \Phi_n^\alpha} R(\widehat{\alpha}_\kappa) + \delta_n$. Combining these inequalities with the oracle bound for $\overline{R}_K(\widehat{\kappa}_\alpha)$ yields the refit inequality with an additional $2\delta_n$ term.

Finally, choose $\delta := n^{-2}$ so that $t = \log(2K|\Phi_n^\alpha|n^2) \asymp \log|\Phi_n^\alpha| + \log n$ and $1-\delta \to 1$. Under Assumption 2 ($\log|\Phi_n^\alpha| = O(\log n)$), the remainder term is of order $\log|\Phi_n^\alpha|/n$. Combining $\bar{\mathcal{E}}_K(\hat{\kappa}_\alpha) \le \frac{1+\eta}{1-\eta}\,\bar{\mathcal{E}}_K(\kappa^*) + \frac{1}{1-\eta}\left(\frac{2C_1^2}{\eta} + 4C_2\right)\frac{t}{n}$ with the refit transfer argument yields, with probability tending to one, $R(\hat{\alpha}_{\hat{\kappa}_\alpha}) - R(\alpha_0) \le (1+o(1))\inf_{\kappa \in \Phi_n^\alpha}\{R(\hat{\alpha}_\kappa) - R(\alpha_0)\} + C\,\frac{\log|\Phi_n^\alpha|}{n} + 2\delta_n$. Using the variational identity $R(\alpha) - R(\alpha_0) = \|\alpha - \alpha_0\|_{P,2}^2$ completes the proof.

A.5 Proof of Theorem 2

Note on notation. We use the main-text notation throughout. Let $U \in \mathbb{R}^J$ be the signal and $g_0(v) := \mathbb{E}[U \mid V = v]$ the regression nuisance (so $\eta_0 = (g_0, \alpha_0)$). Recall $\mathcal{V} = \sigma(V)$. For readability we present the argument for the scalar case $J = 1$; the extension to general J follows by replacing products with inner products, e.g. $\alpha(W)\{U - g(W)\}$ becomes $\alpha(W)^\top\{U - g(W)\}$ and squares with squared Euclidean norms (e.g. $\alpha(W)^2$ becomes $\|\alpha(W)\|_2^2$). Throughout Appendix A.5 and Appendix A.6, we write $Y := U$ for the scalar ($J = 1$) signal. The data-driven (cross-validated) choice of hyperparameters does not destroy Neyman–orthogonality of the score. The

key fact for the linear-functional score in (5) is that the population moment error admits an *exact* second-order representation in the nuisance errors. This allows us to treat the selected hyperparameters as a random index without taking derivatives with respect to them.

Step 0: Outer-fold conditioning, measurability, and independence. Fix an outer fold index $l \in \{1, \dots, L\}$ and write I_l for the evaluation indices and $I_l^c = \{1, \dots, n\} \setminus I_l$ for the corresponding training indices. Define the training σ–field $\mathcal{F}_l := \sigma\big(\{W_i : i \in I_l^c\}\big)$. By construction of Algorithm 1, the selected hyperparameters $(\hat{\varphi}_{g,l}, \hat{\kappa}_{\alpha,l})$ are computed using only the training sample I_l^c, and the nuisance estimators are refit on I_l^c, producing $\hat{g}_l := \hat{g}_{\hat{\varphi}_{g,l}}$, $\hat{\alpha}_l := \hat{\alpha}_{\hat{\kappa}_{\alpha,l}}$. Since Φ_n^g and Φ_n^α are finite (Assumption 2), the map $I_l^c \mapsto (\hat{\varphi}_{g,l}, \hat{\kappa}_{\alpha,l})$ is measurable and therefore $\hat{g}_l$ and $\hat{\alpha}_l$ are $\mathcal{F}_l$–measurable random functions. Because the observations are i.i.d. and I_l is disjoint from I_l^c, conditional on $\mathcal{F}_l$ the evaluation observations $\{W_i : i \in I_l\}$ are i.i.d. draws from P and are independent of $(\hat{g}_l, \hat{\alpha}_l)$. Equivalently, conditional on $\mathcal{F}_l$, we may treat $(\hat{g}_l, \hat{\alpha}_l)$ as fixed functions and $W \sim P$ as a fresh draw independent of $\mathcal{F}_l$. Let $\mathcal{V} := \sigma(V)$ denote the σ-field generated by the conditioning variables V (as in Section 2.1). We assume that the nuisance classes for g and α are contained in $L^2(P; \mathcal{V})$, i.e., their realizations are $\mathcal{V}$-measurable and square-integrable. Thus, $\hat{g}_l$ and $\hat{\alpha}_l$ are $\mathcal{V}$–measurable functions for each realization of the training data.

Step 1: Define the population moment map. Recall the orthogonal score $\psi(W; \theta, g, \alpha) = m(W, g) - \theta + \alpha(W)\{U - g(W)\}$. Define the population moment (at the true target θ_0) as the map $\Psi(g, \alpha) := \mathbb{E}\big[\psi(W; \theta_0, g, \alpha)\big]$. Using $\theta_0 = \mathbb{E}[m(W, g_0)]$, we may rewrite $\Psi(g, \alpha)$ as

$$\Psi(g, \alpha) = \mathbb{E}\big[m(W, g) - m(W, g_0)\big] + \mathbb{E}\big[\alpha(W)\{U - g(W)\}\big]. \tag{30}$$

Step 2: Linearity of m and the Riesz representation. Because $m(W, \cdot)$ is linear in its second argument, $m(W, g) - m(W, g_0) = m(W, g - g_0)$, and therefore

$$\mathbb{E}\big[m(W, g) - m(W, g_0)\big] = \mathbb{E}\big[m(W, g - g_0)\big]. \tag{31}$$

By the defining property of the Riesz representer α_0, for every admissible direction h we have $\mathbb{E}\big[m(W,h)\big] \;=\; \mathbb{E}\big[\alpha_0(W)\,h(W)\big]$. Applying this with $h = g - g_0$ yields

$$\mathbb{E}\big[m(W, g - g_0)\big] = \mathbb{E}\big[\alpha_0(W)\{g(W) - g_0(W)\}\big]. \tag{32}$$

Step 3: Conditional mean property and elimination of the first-order term in α. Assume that the regression nuisance satisfies the conditional mean restriction

$$g_0(W) = \mathbb{E}[U \mid \mathcal{V}] \qquad \Longleftrightarrow \qquad \mathbb{E}[U - g_0(W) \mid \mathcal{V}] = 0. \tag{33}$$

Let α be any $\mathcal{V}$-measurable function in $L^2(P)$. Then, by iterated expectation,

$$\begin{aligned}\mathbb{E}[\alpha(W)\{U - g_0(W)\}] &= \mathbb{E}[\mathbb{E}[\alpha(W)\{U - g_0(W)\} \mid \mathcal{V}]] \\ &= \mathbb{E}[\alpha(W)\,\mathbb{E}[U - g_0(W) \mid \mathcal{V}]] = 0.\end{aligned} \tag{34}$$

Now decompose the residual as

$$U - g(W) = \{U - g_0(W)\} - \{g(W) - g_0(W)\}.$$

Multiplying by $\alpha(W)$ and taking expectations gives

$$\begin{aligned}\mathbb{E}[\alpha(W)\{U - g(W)\}] &= \mathbb{E}[\alpha(W)\{U - g_0(W)\}] - \mathbb{E}[\alpha(W)\{g(W) - g_0(W)\}] \\ &= -\mathbb{E}[\alpha(W)\{g(W) - g_0(W)\}]\,,\end{aligned} \tag{35}$$

where the last equality uses (34). Therefore, Term (II) simplifies to

$$\text{(II)} = -\mathbb{E}[\alpha(W)\{g(W) - g_0(W)\}]\,.$$

Step 4: Exact second-order representation of the population moment error. We now substitute the identities derived in Steps 2 and 3 into the decomposition of $\Psi(g, \alpha)$ established in (30). Recall the decomposition:

$$\Psi(g,\alpha) = \underbrace{\mathbb{E}\big[m(W,g) - m(W,g_0)\big]}_{\text{Term (I)}} + \underbrace{\mathbb{E}\big[\alpha(W)\{U - g(W)\}\big]}_{\text{Term (II)}}.$$

By the linearity of m and the Riesz representation derived in (31)–(32), Term (I) becomes: (I) $= \mathbb{E}\big[\alpha_0(W)\{g(W) - g_0(W)\}\big]$.

By the residual decomposition and the conditional mean restriction (Step 3), Term (II) becomes: (II) $= -\mathbb{E}\big[\alpha(W)\{g(W) - g_0(W)\}\big]$. Adding these two terms and factoring the common difference $g(W) - g_0(W)$ yields the exact product representation:

$$\begin{aligned}\Psi(g,\alpha) &= \mathbb{E}\big[\alpha_0(W)\{g(W) - g_0(W)\}\big] - \mathbb{E}\big[\alpha(W)\{g(W) - g_0(W)\}\big] \\ &= \mathbb{E}\Big[\big(\alpha_0(W) - \alpha(W)\big)\big(g(W) - g_0(W)\big)\Big] \\ &= -\mathbb{E}\Big[\big(\alpha(W) - \alpha_0(W)\big)\big(g(W) - g_0(W)\big)\Big]. \end{aligned} \tag{36}$$

Step 5: Application to adaptively selected nuisances (conditional identity). We now apply the identity (36) to the fold-specific, adaptively selected nuisances. Let $\mathcal{F}_l := \sigma(\{W_i : i \in I_l^c\})$ denote the σ–field generated by the training fold. By Step 0 (measurability of the selection/refitting map over a finite library), the estimators $\hat{g}_l$ and $\hat{\alpha}_l$ are $\mathcal{F}_l$–measurable. Since the observations are i.i.d. and I_l is disjoint from I_l^c, conditional on $\mathcal{F}_l$ the evaluation observations $\{W_i : i \in I_l\}$ are i.i.d. draws from P and are independent of $(\hat{g}_l, \hat{\alpha}_l)$. Equivalently, conditional on $\mathcal{F}_l$ we may treat $(\hat{g}_l, \hat{\alpha}_l)$ as fixed functions and let $W \sim P$ be a fresh draw independent of $\mathcal{F}_l$. Therefore, applying (36) conditionally yields

$$\mathbb{E}\big[\psi(W;\theta_0,\hat{g}_l,\hat{\alpha}_l) \mid \mathcal{F}_l\big] = -\mathbb{E}\Big[\big(\hat{\alpha}_l(W) - \alpha_0(W)\big)\big(\hat{g}_l(W) - g_0(W)\big) \mid \mathcal{F}_l\Big]. \tag{37}$$

Step 6: Conditional Cauchy–Schwarz bound. Define the fold–specific nuisance errors $U_l(W) := \hat{\alpha}_l(W) - \alpha_0(W)$, $V_l(W) := \hat{g}_l(W) - g_0(W)$. Then (37) can be written as $\mathbb{E}\big[\psi(W;\theta_0,\hat{g}_l,\hat{\alpha}_l) \mid \mathcal{F}_l\big] = -\mathbb{E}\big[U_l(W)V_l(W) \mid \mathcal{F}_l\big]$. Applying the conditional Cauchy–Schwarz inequality yields

$$\begin{aligned}\Big|\mathbb{E}\big[\psi(W;\theta_0,\hat{g}_l,\hat{\alpha}_l) \mid \mathcal{F}_l\big]\Big| &= \Big|\mathbb{E}\big[U_l(W)V_l(W) \mid \mathcal{F}_l\big]\Big| \\ &\le \Big(\mathbb{E}\big[U_l(W)^2 \mid \mathcal{F}_l\big]\Big)^{1/2}\Big(\mathbb{E}\big[V_l(W)^2 \mid \mathcal{F}_l\big]\Big)^{1/2}. \end{aligned} \tag{38}$$

By construction of cross–fitting, $(\hat{g}_l, \hat{\alpha}_l)$ are $\mathcal{F}_l$–measurable, and conditional on $\mathcal{F}_l$ the evaluation draw $W \sim P$ is independent of $\mathcal{F}_l$. For any $\mathcal{F}_l$–measurable $f \in L^2(P)$, $\mathbb{E}\big[f(W)^2 \mid \mathcal{F}_l\big] = \int f(w)^2\, dP(w) = \|f\|_{P,2}^2$ a.s. Applying this identity with $f = U_l$ and $f = V_l$ gives $\mathbb{E}\big[U_l(W)^2 \mid \mathcal{F}_l\big] = \|\hat{\alpha}_l - \alpha_0\|_{P,2}^2$, $\mathbb{E}\big[V_l(W)^2 \mid \mathcal{F}_l\big] = \|\hat{g}_l - g_0\|_{P,2}^2$.

Substituting back into (38) yields the pathwise (almost sure) bound

$$\left|\mathbb{E}\big[\psi(W;\theta_0,\hat{g}_l,\hat{\alpha}_l)\mid\mathcal{F}_l\big]\right| \le \|\hat{g}_l - g_0\|_{P,2}\,\|\hat{\alpha}_l - \alpha_0\|_{P,2}. \tag{39}$$

Crucially, (39) holds *pathwise* for each realization of the training sample (equivalently, conditional on $\mathcal{F}_l$), and therefore is uniform in the realized hyperparameter selections $(\hat{\varphi}_{g,l}, \hat{\kappa}_{\alpha,l})$; no continuity or differentiability of the selection map is required. Steps 1–6 establish the identity and Cauchy–Schwarz bound exactly (conditional on $\mathcal{D}^{\text{train}}$).

Step 7 (Asymptotic implication used in Theorem 3). Fix a fold $l \in \{1,\dots,L\}$. By the pathwise bound (39), we have $\mathcal{F}_l$–a.s. (i.e. for P–almost every realization of the training sample) that

$$\left|\mathbb{E}\big[\psi(W;\theta_0,\hat{g}_l,\hat{\alpha}_l)\mid\mathcal{F}_l\big]\right| \le \|\hat{g}_l - g_0\|_{P,2}\,\|\hat{\alpha}_l - \alpha_0\|_{P,2}. \tag{40}$$

To convert (40) into an $o_p(n^{-1/2})$ bound, let $\varepsilon > 0$ be arbitrary. Then, using (40) and the monotonicity of probability,

$$\mathbb{P}\Big(\sqrt{n}\left|\mathbb{E}\big[\psi(W;\theta_0,\hat{g}_l,\hat{\alpha}_l)\mid\mathcal{F}_l\big]\right| > \varepsilon\Big) \le \mathbb{P}\big(\sqrt{n}\,\|\hat{g}_l - g_0\|_{P,2}\,\|\hat{\alpha}_l - \alpha_0\|_{P,2} > \varepsilon\big). \tag{41}$$

By the product-rate condition for DML inference, $\|\hat{g}_l - g_0\|_{P,2}\,\|\hat{\alpha}_l - \alpha_0\|_{P,2} = o_p(n^{-1/2})$, the right-hand side of (41) converges to 0 as $n \to \infty$. Hence

$$\mathbb{E}\big[\psi(W;\theta_0,\hat{g}_l,\hat{\alpha}_l)\mid\mathcal{F}_l\big] = o_p(n^{-1/2}). \tag{42}$$

Since the number of outer folds L is fixed, the same rate holds uniformly over ℓ: indeed, for any $\varepsilon > 0$,

$$\mathbb{P}\Big(\max_{1\le l\le L}\sqrt{n}\left|\mathbb{E}\big[\psi(W;\theta_0,\hat{g}_l,\hat{\alpha}_l)\mid\mathcal{F}_l\big]\right| > \varepsilon\Big) \le \sum_{l=1}^{L}\mathbb{P}\Big(\sqrt{n}\left|\mathbb{E}\big[\psi(W;\theta_0,\hat{g}_l,\hat{\alpha}_l)\mid\mathcal{F}_l\big]\right| > \varepsilon\Big) \;\to\; 0, \tag{43}$$

where the inequality is the union bound and the convergence uses (42) for each fixed ℓ. Therefore, $\max_{1\le l\le L}\left|\mathbb{E}\big[\psi(W;\theta_0,\hat{g}_l,\hat{\alpha}_l)\mid\mathcal{F}_l\big]\right| = o_p(n^{-1/2})$. In particular, any cross-fitted average over folds inherits the same order because it is a convex combination of the fold-specific terms and ℓ is fixed. This establishes the usual $o_p(n^{-1/2})$ implication

of the non-asymptotic bound under a product-rate condition, which is the only part of this step used later in Theorem 3. □

Remark. The argument is non-asymptotic up to the final product-rate step: the exact identity (36) and the conditional bound (39) hold for any pair of $\mathcal{V}$–measurable functions (g, α) trained on an independent sample, including data-dependent selections from a finite library. In particular, no derivatives with respect to hyperparameters are required, which justifies treating hyperparameters as a tertiary nuisance parameter indexed over Φ_n.

A.6 Proof of Theorem 3

We establish the asymptotic normality of the adaptive cross–fitted estimator $\hat{\theta}_{\text{Adaptive}}$ solving

$$\frac{1}{n}\sum_{i=1}^{n}\psi\Big(W_i;\hat{\theta}_{\text{Adaptive}},\hat{g}_{-i},\hat{\alpha}_{-i}\Big)=0,\qquad \hat{\eta}_{-i}:=(\hat{g}_{-i},\hat{\alpha}_{-i}).$$

Throughout, write $Pf=\mathbb{E}[f(W)]$ and $P_nf=n^{-1}\sum_{i=1}^{n}f(W_i)$.

Step 0: Cross–fitting notation and conditioning. Let $(I_l)_{l=1}^{L}$ be the outer folds. For each fold ℓ, let $\hat{\eta}_l=(\hat{g}_l,\hat{\alpha}_l)$ denote the nuisance estimates trained on I_l^c (including the inner CV selection and refitting), and define the cross–fitted nuisance for observation i by $\hat{\eta}_{-i}:=\hat{\eta}_l$ whenever $i\in I_l$. Let $\mathcal{F}_l:=\sigma\big(\{W_j:j\in I_l^c\}\big)$ be the training σ–field for fold ℓ. Then $\hat{\eta}_l$ is $\mathcal{F}_l$–measurable and, conditional on $\mathcal{F}_l$, the evaluation observations $\{W_i:i\in I_l\}$ are i.i.d. draws from P and independent of $\hat{\eta}_l$.

Step 1: Linearization in the target parameter. Recall the score[14]

$$\psi(W;\theta,g,\alpha)=m(W,g)-\theta+\alpha(W)^{\top}\{U-g(W)\}. \tag{44}$$

This score is affine in θ with slope -1. Hence, for any nuisance $\eta=(g,\alpha)$, $\psi(W;\hat{\theta},\eta)-\psi(W;\theta_0,\eta)=-(\hat{\theta}-\theta_0)$, and therefore the sample moment equation implies the *exact*

[14]If $J=1$, interpret U and $g(W)$ as scalars and drop the transpose. The argument below is written in this scalar notation for readability; the $J>1$ case follows by replacing products with inner products.

identity

$$0 = P_n\psi\Big(W; \hat{\theta}_{\text{Adaptive}}, \hat{\eta}_-\Big) = P_n\psi(W; \theta_0, \hat{\eta}_-) - \big(\hat{\theta}_{\text{Adaptive}} - \theta_0\big), \tag{45}$$

where $\hat{\eta}_-$ denotes the observation-specific cross–fitted nuisance map $i \mapsto \hat{\eta}_{-i}$. Rearranging (45) yields

$$\hat{\theta}_{\text{Adaptive}} - \theta_0 = P_n\psi(W; \theta_0, \hat{\eta}_-)\,. \tag{46}$$

Consequently,

$$\sqrt{n}\big(\hat{\theta}_{\text{Adaptive}} - \theta_0\big) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi(W_i; \theta_0, \hat{\eta}_{-i})\,. \tag{47}$$

(Remark: In settings where ψ is not affine in θ, one can replace this step by a mean-value expansion with Jacobian $\Gamma_0 = \partial_\theta P\psi(W; \theta_0, \eta_0)$; the rest of the argument below is unchanged, up to the usual premultiplication by Γ_0^{-1}.

Step 2: Decomposition into oracle score and empirical remainder. Let $\eta_0 = (g_0, \alpha_0)$ denote the true nuisance parameters. We decompose the estimator's error by adding and subtracting the unobserved oracle score $\psi(W_i; \theta_0, \eta_0)$ inside the summation in (47):

$$\begin{aligned}
\sqrt{n}\big(\hat{\theta}_{\text{Adaptive}} - \theta_0\big) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi(W_i; \theta_0, \hat{\eta}_{-i}) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big[\psi(W_i; \theta_0, \eta_0) \;+\; \Big(\psi(W_i; \theta_0, \hat{\eta}_{-i}) - \psi(W_i; \theta_0, \eta_0)\Big)\Big] \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi(W_i; \theta_0, \eta_0) \;+\; \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big\{\psi(W_i; \theta_0, \hat{\eta}_{-i}) - \psi(W_i; \theta_0, \eta_0)\Big\}.
\end{aligned}$$

Defining the first term as the oracle process S_n and the second term as the nuisance estimation remainder T_n, we obtain the representation:

$$\sqrt{n}\big(\hat{\theta}_{\text{Adaptive}} - \theta_0\big) \;=\; \underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi(W_i; \theta_0, \eta_0)}_{=:S_n} \;+\; \underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big\{\psi(W_i; \theta_0, \hat{\eta}_{-i}) - \psi(W_i; \theta_0, \eta_0)\Big\}}_{=:T_n}. \tag{48}$$

The term S_n is a normalized sum of i.i.d. mean-zero random variables (the efficient influence functions). The term T_n captures the stochastic error arising from estimating

η_0 with $\hat{\eta}_{-i}$. We proceed to show that $T_n = o_p(1)$, which implies asymptotic normality via the Central Limit Theorem applied to S_n.

Step 3: $L^2(P)$ Lipschitz bound for the score difference. Recall $\psi(W;\theta,g,\alpha) = m(W,g) - \theta + \alpha(W)^\top\{U - g(W)\}$. For any (g,α), consider the difference relative to the oracle nuisances $\eta_0 = (g_0,\alpha_0)$:

$$\psi(W;\theta_0,g,\alpha) - \psi(W;\theta_0,g_0,\alpha_0) = m(W,g-g_0) \\ - \alpha(W)^\top\{g(W) - g_0(W)\} + \{\alpha(W) - \alpha_0(W)\}^\top\{U - g_0(W)\}.$$

Applying the $L^2(P)$ norm and the triangle inequality yields

$$\|\psi(\cdot;\theta_0,g,\alpha) - \psi(\cdot;\theta_0,g_0,\alpha_0)\|_{P,2} \le \|m(\cdot,g-g_0)\|_{P,2} + \|\alpha^\top(g-g_0)\|_{P,2} + \|(\alpha-\alpha_0)^\top(U-g_0)\|_{P,2}. \tag{49}$$

By Assumption 1, $\|m(\cdot,h)\|_{P,2} \le C_m\|h\|_{P,2}$. Moreover, by the interpolation inequality $\|f\|_{P,4} \le \|f\|_{P,\infty}^{1/2}\|f\|_{P,2}^{1/2}$ and the clipping bound $\|\alpha-\alpha_0\|_{P,\infty} \le \|\alpha\|_{P,\infty} + \|\alpha_0\|_{P,\infty} \le 2B$, we have $\|\alpha-\alpha_0\|_{P,4} \le (2B)^{1/2}\|\alpha-\alpha_0\|_{P,2}^{1/2}$. Therefore, by Hölder inequality, $\left\|(\alpha-\alpha_0)^\top(U-g_0)\right\|_{P,2} \le \|\alpha-\alpha_0\|_{P,4}\|U-g_0\|_{P,4} \le (2B)^{1/2}\|U-g_0\|_{P,4}\|\alpha-\alpha_0\|_{P,2}^{1/2}$.

Combining yields

$$\|\psi(\cdot;\theta_0,g,\alpha) - \psi(\cdot;\theta_0,g_0,\alpha_0)\|_{P,2} \le (C_m+B)\|g-g_0\|_{P,2} + C_{U,4}\|\alpha-\alpha_0\|_{P,2}^{1/2}, \tag{50}$$

where $C_{U,4} := (2B)^{1/2}\|U-g_0\|_{P,4} < \infty$.

Step 4: Fold-wise decomposition and control of the bias term. We analyze the remainder T_n by partitioning the sum over the ℓ folds. Let $n_l = |I_l|$ denote the number of observations in fold ℓ. We can rewrite T_n as:

$$T_n \;=\; \sum_{l=1}^{L} \frac{n_l}{\sqrt{n}} P_{n,l}\Delta_l, \tag{51}$$

where $P_{n,l}f := n_l^{-1}\sum_{i\in I_l} f(W_i)$ denotes the empirical measure on fold ℓ. Conditional on the training σ-field $\mathcal{F}_l$, the function Δ_l is fixed, and the data $\{W_i\}_{i\in I_l}$ are i.i.d. draws from P. We decompose the empirical mean $P_{n,l}\Delta_l$ into its conditional

expectation (bias) and a centered empirical process (fluctuation):

$$P_{n,l}\Delta_l \;=\; \underbrace{P\Delta_l}_{\text{Bias}} \;+\; \underbrace{(P_{n,l}-P)\Delta_l}_{\text{Fluctuation}}, \tag{52}$$

where $P\Delta_l := \mathbb{E}[\Delta_l(W) \mid \mathcal{F}_l]$.

We first control the bias term $P\Delta_l$. By the definition of $\Delta_l(W) = \psi(W;\theta_0,\hat{\eta}_l) - \psi(W;\theta_0,\eta_0)$, and using the fact that the oracle score has mean zero ($\mathbb{E}[\psi(W;\theta_0,\eta_0)] = 0$), the conditional expectation simplifies to: $P\Delta_l \;=\; \mathbb{E}\big[\psi(W;\theta_0,\hat{\eta}_l) \mid \mathcal{F}_l\big] - 0$.

We now invoke Theorem 2 (Adaptive Orthogonality). Conditional on $\mathcal{F}_\ell$,

$$\big|P\Delta_\ell\big| = \Big|\mathbb{E}\big[\psi(W;\theta_0,\hat{g}_\ell,\hat{\alpha}_\ell) \mid \mathcal{F}_\ell\big]\Big| \le \|\hat{g}_\ell - g_0\|_{P,2}\,\|\hat{\alpha}_\ell - \alpha_0\|_{P,2}. \tag{53}$$

Under the product-rate condition (Assumption 2 and Theorem 1), $\|\hat{g}_\ell - g_0\|_{P,2}\,\|\hat{\alpha}_\ell - \alpha_0\|_{P,2} = o_p(n^{-1/2})$, and therefore, by (53), $|P\Delta_\ell| = o_p(n^{-1/2})$.

We now evaluate the contribution of this bias term to the total remainder T_n. Recall from (51) that the bias component is the weighted sum of these conditional expectations. Substituting the rate obtained above: $\sum_{l=1}^{L} \frac{n_l}{\sqrt{n}} P\Delta_l = \sum_{l=1}^{L} \frac{n_l}{n}\Big(\sqrt{n}\cdot P\Delta_l\Big) = \sum_{l=1}^{L} \underbrace{\frac{n_l}{n}}_{\le 1}\Big(\sqrt{n}\cdot o_p(n^{-1/2})\Big) = \sum_{l=1}^{L} \frac{n_l}{n}\cdot o_p(1)$. Since the number of folds ℓ is fixed and $\sum_{l=1}^{L} \frac{n_l}{n} = 1$, the linear combination of $o_p(1)$ terms remains $o_p(1)$. Thus, $\sum_{l=1}^{L} \frac{n_l}{\sqrt{n}} P\Delta_l \;=\; o_p(1)$, proving that the systematic bias component of the remainder T_n is asymptotically negligible.

Step 5: Control of the stochastic fluctuation term. It remains to control the fluctuation component of the remainder T_n. Recall from (52) that $P_{n,l}\Delta_l = P\Delta_l + (P_{n,l}-P)\Delta_l$. Having established in Step 4 that the bias component sums to $o_p(1)$, we focus on the weighted sum of the centered empirical processes:

$$T_{n,\text{fluc}} \;=\; \sum_{l=1}^{L} \frac{n_l}{\sqrt{n}}(P_{n,l}-P)\Delta_l \;=\; \sum_{l=1}^{L} \sqrt{\frac{n_l}{n}}\,\underbrace{\sqrt{n_l}(P_{n,l}-P)\Delta_l}_{=:\mathbb{G}_{n,l}(\Delta_l)}.$$

We analyze the term $\mathbb{G}_{n,l}(\Delta_l)$ conditional on the training data $\mathcal{F}_l$. Conditional on $\mathcal{F}_l$, the function Δ_l is fixed, and the observations in fold ℓ are i.i.d. draws from P. The conditional expectation is zero by construction: $\mathbb{E}\big[(P_{n,l}-P)\Delta_l \mid \mathcal{F}_l\big] = 0$.

The conditional variance is determined by the second moment of the score difference: $\mathrm{Var}\big(\mathbb{G}_{n,l}(\Delta_l) \mid \mathcal{F}_l\big) = n_l \,\mathrm{Var}\big(P_{n,l}\Delta_l \mid \mathcal{F}_l\big) = n_l \cdot \frac{1}{n_l}\,\mathrm{Var}\big(\Delta_l(W) \mid \mathcal{F}_l\big) \le \mathbb{E}\big[\Delta_l(W)^2 \mid \mathcal{F}_l\big] = \|\Delta_l\|_{P,2}^2$. Fix any $t > 0$. Conditional on $\mathcal{F}_\ell$, we have

$$\Pr\left(\frac{\big|\mathbb{G}_{n,\ell}(\Delta_\ell)\big|}{\|\Delta_\ell\|_{P,2}} > t \;\middle|\; \mathcal{F}_\ell \right) \le \frac{\mathrm{Var}(\mathbb{G}_{n,\ell}(\Delta_\ell) \mid \mathcal{F}_\ell)}{t^2 \|\Delta_\ell\|_{P,2}^2} \le t^{-2},$$

where the first inequality is Chebyshev's inequality and the second uses $\mathrm{Var}(\mathbb{G}_{n,\ell}(\Delta_\ell) \mid \mathcal{F}_\ell) \le \|\Delta_\ell\|_{P,2}^2$. Hence,

$$\frac{\mathbb{G}_{n,\ell}(\Delta_\ell)}{\|\Delta_\ell\|_{P,2}} = O_p(1) \quad \text{(conditionally on } \mathcal{F}_\ell).$$

Combining this with Step 3, which gives $\|\Delta_\ell\|_{P,2} = o_p(1)$, yields $\mathbb{G}_{n,\ell}(\Delta_\ell) = o_p(1)$. Since ℓ is fixed and $\sqrt{n_\ell/n} \le 1$,

$$T_{n,\mathrm{fluc}} = \sum_{\ell=1}^{L} \sqrt{\frac{n_\ell}{n}}\, \mathbb{G}_{n,\ell}(\Delta_\ell) = o_p(1). \tag{54}$$

Step 6: Asymptotic normality of the oracle score process. Let $Z_i := \psi(W_i; \theta_0, \eta_0)$ with $\eta_0 = (g_0, \alpha_0)$. Then $\{Z_i\}_{i=1}^n$ are i.i.d. We first verify that $\mathbb{E}[Z_i] = 0$:

$$\begin{aligned}
\mathbb{E}[Z_i] &= \mathbb{E}[m(W, g_0) - \theta_0] + \mathbb{E}\big[\alpha_0(W)^\top \{U - g_0(W)\}\big] \\
&= 0 + \mathbb{E}\big[\mathbb{E}\big[\alpha_0(W)^\top \{U - g_0(W)\} \mid \mathcal{X}\big]\big] \\
&= \mathbb{E}\big[\alpha_0(W)^\top \mathbb{E}[U - g_0(W) \mid \mathcal{X}]\big] \\
&= 0,
\end{aligned}$$

where $\theta_0 = \mathbb{E}[m(W, g_0)]$, α_0 is $\mathcal{X}$-measurable, and $g_0(W) = \mathbb{E}[U \mid \mathcal{X}]$.

Next, under Assumption 1, Z_i has finite second moment, so the variance $\sigma_{\mathrm{eff}}^2 := \mathrm{Var}(Z_i) = \mathbb{E}[\psi(W; \theta_0, \eta_0)^2] < \infty$ is well-defined. Hence, by the Lindeberg–Lévy CLT,

$$S_n := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i \;\Rightarrow\; \mathcal{N}(0, \sigma_{\mathrm{eff}}^2). \tag{55}$$

Step 7: Asymptotic normality via Slutsky's Theorem. We combine the results from the previous steps to establish the final limiting distribution. Recall the

decomposition from (48): $\sqrt{n}\big(\hat{\theta}_{\text{Adaptive}} - \theta_0\big) = S_n + T_n$. From Step 6 (Equation 55), we established that the oracle score component converges in distribution: $S_n \xrightarrow{d} \mathcal{N}\big(0, \sigma_{\text{eff}}^2\big)$. From Steps 4 and 5 (Equation (54)), we established that the nuisance estimation remainder converges in probability to zero: $T_n \xrightarrow{p} 0$. Slutsky's Theorem states that if a sequence of random variables X_n converges in distribution to X and another sequence Y_n converges in probability to a constant c, then their sum $X_n + Y_n$ converges in distribution to $X + c$. Applying this with $X_n = S_n$ and $Y_n = T_n$, we obtain: $\sqrt{n}\big(\hat{\theta}_{\text{Adaptive}} - \theta_0\big) = S_n + T_n \xrightarrow{d} \mathcal{N}\big(0, \sigma_{\text{eff}}^2\big) + 0 = \mathcal{N}\big(0, \sigma_{\text{eff}}^2\big)$. This completes the proof of Theorem 3. □

Remark 14 (Generalization to non-affine scores)**.** If ψ is nonlinear in θ, Step 1 is replaced by a mean-value expansion of the estimating equation $P_n\psi(W; \hat{\theta}, \hat{\eta}_-) = 0$ around θ_0: $0 = P_n\psi(W; \theta_0, \hat{\eta}_-) + \hat{\Gamma}_n(\hat{\theta} - \theta_0)$, where $\hat{\Gamma}_n$ is the empirical Jacobian and $\hat{\Gamma}_n \xrightarrow{p} \Gamma_0 := \mathbb{E}[\partial_\theta \psi(W; \theta_0, \eta_0)]$. Provided $\Gamma_0 \neq 0$ and the corresponding remainder term $R_n = o_p(1)$ (with $R_n = T_n$ in the affine case treated above), Slutsky's theorem yields $\sqrt{n}(\hat{\theta} - \theta_0) \Rightarrow \mathcal{N}(0, \Gamma_0^{-2}\sigma_{\text{eff}}^2)$.

A.7 Proof of Lemma 3 and Corollary 2

We establish the consistency of the plug-in variance estimator $\hat{\sigma}^2$ (Lemma 3) and the resulting validity of the confidence intervals (Corollary 2).

Recall the definition $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \Big(\psi\big(W_i\,;\, \hat{\theta}_{\text{Adaptive}},\, \hat{g}_{-i},\, \hat{\alpha}_{-i}\big)\Big)^2$, where $\hat{g}_{-i}$ and $\hat{\alpha}_{-i}$ denote the nuisance estimators trained on the fold I_k^c such that $i \in I_k$.

Step 1: Decomposition of the variance estimator. The score function $\psi(W; \theta, g, \alpha) = m(W, g) - \theta + \alpha(W)\{U - g(W)\}$ is linear in θ. Specifically, $\psi(W; \theta, g, \alpha) = \psi(W; \theta_0, g, \alpha) - (\theta - \theta_0)$. Substituting this into the definition of $\hat{\sigma}^2$, we expand the square:

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^n \Big(\psi\big(W_i; \theta_0, \hat{g}_{-i}, \hat{\alpha}_{-i}\big) - (\hat{\theta}_{\text{Adaptive}} - \theta_0)\Big)^2 \\ &= \underbrace{\frac{1}{n}\sum_{i=1}^n \psi\big(W_i; \theta_0, \hat{g}_{-i}, \hat{\alpha}_{-i}\big)^2}_{=:\hat{\Sigma}_n} - 2(\hat{\theta}_{\text{Adaptive}} - \theta_0)\underbrace{\frac{1}{n}\sum_{i=1}^n \psi\big(W_i; \theta_0, \hat{g}_{-i}, \hat{\alpha}_{-i}\big)}_{=:\bar{\Psi}_n} + (\hat{\theta}_{\text{Adaptive}} - \theta_0)^2.\end{aligned}$$

By the definition of the estimator $\hat{\theta}_{\text{Adaptive}}$, it solves $\frac{1}{n}\sum_{i=1}^{n}\psi(W_i;\hat{\theta}_{\text{Adaptive}},\hat{g}_{-i},\hat{\alpha}_{-i}) = 0$. Using the linearity in θ, this implies $\bar{\Psi}_n - (\hat{\theta}_{\text{Adaptive}} - \theta_0) = 0 \implies \bar{\Psi}_n = \hat{\theta}_{\text{Adaptive}} - \theta_0$. Substituting this back into the expansion yields $\hat{\sigma}^2 = \hat{\Sigma}_n - 2(\hat{\theta}_{\text{Adaptive}} - \theta_0)^2 + (\hat{\theta}_{\text{Adaptive}} - \theta_0)^2 = \hat{\Sigma}_n - (\hat{\theta}_{\text{Adaptive}} - \theta_0)^2$. From Theorem 3, we know that $\hat{\theta}_{\text{Adaptive}} - \theta_0 = O_p(n^{-1/2})$, so $(\hat{\theta}_{\text{Adaptive}} - \theta_0)^2 = O_p(n^{-1})$. Thus,

$$\hat{\sigma}^2 = \hat{\Sigma}_n + o_p(1). \tag{56}$$

It remains to show that $\hat{\Sigma}_n \xrightarrow{p} \sigma^2_{\text{eff}}$.

Step 2: Consistency of the nuisance-plug-in second moment. We decompose $\hat{\Sigma}_n$ by adding and subtracting the oracle score $\psi_0(W_i) := \psi(W_i;\theta_0,g_0,\alpha_0)$. Let $\hat{\psi}_i := \psi(W_i;\theta_0,\hat{g}_{-i},\hat{\alpha}_{-i})$. Then

$$\hat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^{n}\hat{\psi}_i^2 = \frac{1}{n}\sum_{i=1}^{n}\psi_0(W_i)^2 + \frac{1}{n}\sum_{i=1}^{n}\big(\hat{\psi}_i^2 - \psi_0(W_i)^2\big). \tag{57}$$

Note that the cross-fitted nuisances differ across folds, so we apply the conditional law of large numbers *within each fold* conditional on its training sample $\mathcal{F}_\ell$, treating $\Delta_\ell(\cdot)$ as fixed on the evaluation fold I_ℓ. This yields $|I_\ell|^{-1}\sum_{i\in I_\ell}\Delta_i^2 = \mathbb{E}[\Delta_\ell(W)^2 \mid \mathcal{F}_\ell] + o_p(1) = \|\Delta_\ell\|_{P,2}^2 + o_p(1)$, and averaging over ℓ gives $n^{-1}\sum_{i=1}^{n}\Delta_i^2 = o_p(1)$.

The first term converges to $\sigma^2_{\text{eff}} = \mathbb{E}[\psi_0(W)^2]$ by the Law of Large Numbers, since observations are i.i.d. and moments exist by Assumption 1. For the second term, we use the algebraic identity $|a^2 - b^2| \le |a-b|^2 + 2|b||a-b|$:

$$\left|\frac{1}{n}\sum_{i=1}^{n}(\hat{\psi}_i^2 - \psi_0(W_i)^2)\right| \le \frac{1}{n}\sum_{i=1}^{n}(\hat{\psi}_i - \psi_0(W_i))^2 + 2\sqrt{\frac{1}{n}\sum_{i=1}^{n}\psi_0(W_i)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\psi}_i - \psi_0(W_i))^2}.$$

Write the cross-fitted average as a sum over outer folds:

$$\frac{1}{n}\sum_{i=1}^{n}\Delta_i^2 = \sum_{\ell=1}^{L}\frac{|I_\ell|}{n}\left(\frac{1}{|I_\ell|}\sum_{i\in I_\ell}\Delta_i^2\right).$$

Within each fold ℓ, $(\hat{g}_{-i},\hat{\alpha}_{-i}) = (\hat{g}_\ell,\hat{\alpha}_\ell)$ is fixed conditional on $\mathcal{F}_\ell$, so the conditional LLN applies to $|I_\ell|^{-1}\sum_{i\in I_\ell}\Delta_i^2$. Recall $\Delta_i := \psi(W_i;\theta_0,\hat{g}_{-i},\hat{\alpha}_{-i}) - \psi_0(W_i)$. By the score definition, $\Delta_i := m(W_i,\widehat{g}_{-i} - g_0) + (\widehat{\alpha}_{-i}(W_i) - \alpha_0(W_i))^\top(U_i - g_0(W_i)) -$

$\widehat{\alpha}_{-i}(W_i)^\top(\widehat{g}_{-i}(W_i) - g_0(W_i))$. Thus, conditional on $\mathcal{F}_\ell$, using $(a+b+c)^2 \le 3(a^2 + b^2 + c^2)$,

$$\mathbb{E}[\Delta_i^2 \mid \mathcal{F}_\ell] \le 3\,\mathbb{E}\big[m(W_i, \widehat{g}_{-i} - g_0)^2 \mid \mathcal{F}_\ell\big] + 3\,\mathbb{E}\Big[\Big((\widehat{\alpha}_{-i}(W_i) - \alpha_0(W_i))^\top(U_i - g_0(W_i))\Big)^2 \;\Big|\; \mathcal{F}_\ell\Big] + 3\,\mathbb{E}\Big[\big(\widehat{\alpha}_{-i}(W_i)^\top(\widehat{g}_{-i}(W_i) - g_0(W_i))\big)^2 \;\Big|\; \mathcal{F}_\ell\Big].$$

By the L^2 continuity of m, the first term is bounded by $3C_m^2\|\widehat{g}_{-i} - g_0\|_{P,2}^2$. For the second term, we do not require an essential-supremum bound on the conditional variance. Instead, by Cauchy–Schwarz and Hölder's inequality with conjugate exponents $p = \frac{q}{q-2}$ and $p' = \frac{q}{2}$ (valid since $q > 4$), we have

$$\begin{aligned}\mathbb{E}\Big[\big\{(\widehat{\alpha}_{-i}(W) - \alpha_0(W))^\top(U - g_0(W))\big\}^2 \mid \mathcal{F}_\ell\Big] &\le \mathbb{E}\big[\|\widehat{\alpha}_{-i}(W) - \alpha_0(W)\|_2^2\,\|U - g_0(W)\|_2^2 \mid \mathcal{F}_\ell\big] \\ &\le \|\widehat{\alpha}_{-i} - \alpha_0\|_{P,\frac{2q}{q-2}}^2\,\|U - g_0\|_{P,q}^2. \qquad (58)\end{aligned}$$

Using the clipping bound $\|\widehat{\alpha}_{-i} - \alpha_0\|_\infty \le 2B$ (Assumption 1) and the interpolation inequality $\|h\|_{P,r} \le \|h\|_\infty^{1-2/r}\|h\|_{P,2}^{2/r}$ with $r = \frac{2q}{q-2}$, we obtain

$$\|\widehat{\alpha}_{-i} - \alpha_0\|_{P,\frac{2q}{q-2}}^2 \le (2B)^{4/q}\,\|\widehat{\alpha}_{-i} - \alpha_0\|_{P,2}^{2-4/q}.$$

Therefore,

$$\mathbb{E}\Big[\big\{(\widehat{\alpha}_{-i}(W) - \alpha_0(W))^\top(U - g_0(W))\big\}^2 \mid \mathcal{F}_\ell\Big] \le C_{U,q}^2\,\|\widehat{\alpha}_{-i} - \alpha_0\|_{P,2}^{2-4/q}, \qquad (59)$$

where $C_{U,q}^2 := (2B)^{4/q}\|U - g_0\|_{P,q}^2 < \infty$ by Assumption 1. For the third term, using the uniform bound $\|\widehat{\alpha}_{-i}(W)\|_2 \le B$ (from the clipping/boundedness condition), $\mathbb{E}\Big[\big(\widehat{\alpha}_{-i}(W_i)^\top(\widehat{g}_{-i}(W_i) - g_0(W_i))\big)^2 \;\Big|\; \mathcal{F}_\ell\Big] \le B^2\|\widehat{g}_{-i} - g_0\|_{P,2}^2$. Hence, conditional on $\mathcal{F}_\ell$,

$$\mathbb{E}[\Delta_i^2 \mid \mathcal{F}_\ell] \le 3C_m^2\|\widehat{g}_{-i} - g_0\|_{P,2}^2 + 3B^2\|\widehat{g}_{-i} - g_0\|_{P,2}^2 + 3C_{U,q}^2\|\widehat{\alpha}_{-i} - \alpha_0\|_{P,2}^{2-4/q}.$$

Since $q > 4$, the exponent $2 - 4/q > 1$, and consistency of the selected nuisances implies $\|\widehat{g}_{-i} - g_0\|_{P,2} \to_p 0$ and $\|\widehat{\alpha}_{-i} - \alpha_0\|_{P,2} \to_p 0$. Therefore $\mathbb{E}[\Delta_i^2 \mid \mathcal{F}_\ell] \to_p 0$ uniformly over folds.

Averaging over i and using a foldwise conditional law of large numbers yields

$n^{-1}\sum_{i=1}^{n}\Delta_i^2 = o_p(1)$ by Corollary 1. Substituting this back into (57), both error terms vanish, proving $\hat{\Sigma}_n \xrightarrow{p} \sigma_{\text{eff}}^2$. Combining with (56), we have $\hat{\sigma}^2 \xrightarrow{p} \sigma_{\text{eff}}^2$.

Step 3: Slutsky's theorem and confidence interval validity. By Theorem 3,

$$\sqrt{n}(\hat{\theta}_{\text{Adaptive}} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \sigma_{\text{eff}}^2).$$

Since $\hat{\sigma} \xrightarrow{p} \sigma_{\text{eff}}$ (by continuous mapping of $\hat{\sigma}^2$), Slutsky's theorem implies

$$\frac{\sqrt{n}(\hat{\theta}_{\text{Adaptive}} - \theta_0)}{\hat{\sigma}} = \frac{\sqrt{n}(\hat{\theta}_{\text{Adaptive}} - \theta_0)}{\sigma_{\text{eff}}} \cdot \frac{\sigma_{\text{eff}}}{\hat{\sigma}} \xrightarrow{d} \mathcal{N}(0,1) \cdot 1 = \mathcal{N}(0,1).$$

The validity of the confidence interval follows immediately from the definition of convergence in distribution:

$$\begin{aligned}\mathbb{P}(\theta_0 \in CI_{1-\alpha}) &= \mathbb{P}\left(-z_{1-\alpha/2} \leq \sqrt{n}\frac{\hat{\theta}_{\text{Adaptive}} - \theta_0}{\hat{\sigma}} \leq z_{1-\alpha/2}\right) \\ &\to \Phi(z_{1-\alpha/2}) - \Phi(-z_{1-\alpha/2}) = (1-\alpha/2) - (\alpha/2) = 1-\alpha.\end{aligned}$$

A.8 Proof of Theorem 4

Let $T_n := \sqrt{n}(\hat{\theta}_{\text{Adaptive}} - \theta_0)$, $S_n := n^{-1/2}\sum_{i=1}^{n}\psi(W_i; \theta_0, g_0, \alpha_0)$, and $R_n := T_n - S_n$. Fix $\varepsilon > 0$ and write $\sigma := \sigma_{\text{eff}}$.

Step 1: Reduction to the oracle sum plus the remainder. For any $x \in \mathbb{R}$, $\{T_n/\sigma \leq x\} \subseteq \{S_n/\sigma \leq x + \varepsilon\} \cup \{|R_n|/\sigma > \varepsilon\}$, hence $\Pr(T_n/\sigma \leq x) \leq \Pr(S_n/\sigma \leq x+\varepsilon) + \Pr(|R_n|/\sigma > \varepsilon)$. Similarly, $\{S_n/\sigma \leq x - \varepsilon\} \subseteq \{T_n/\sigma \leq x\} \cup \{|R_n|/\sigma > \varepsilon\}$, so

$$\Pr(T_n/\sigma \leq x) \geq \Pr(S_n/\sigma \leq x - \varepsilon) - \Pr(|R_n|/\sigma > \varepsilon).$$

Step 2: Convert the shift into a Φ-error. Using the upper bound,

$$\begin{aligned}\Pr(T_n/\sigma \leq x) - \Phi(x) &\leq \Big(\Pr(S_n/\sigma \leq x+\varepsilon) - \Phi(x+\varepsilon)\Big) + \Big(\Phi(x+\varepsilon) - \Phi(x)\Big) + \Pr(|R_n|/\sigma > \varepsilon) \\ &\leq \sup_{u\in\mathbb{R}}\Big|\Pr(S_n/\sigma \leq u) - \Phi(u)\Big| + \frac{\varepsilon}{\sqrt{2\pi}} + \Pr(|R_n|/\sigma > \varepsilon),\end{aligned}$$

where we used $\sup_x \phi(x) = 1/\sqrt{2\pi}$ for the standard normal density ϕ. The corresponding lower bound yields the same inequality for $\Phi(x) - \Pr(T_n/\sigma \le x)$. Therefore,

$$\sup_{x\in\mathbb{R}} |\Pr(T_n/\sigma \le x) - \Phi(x)| \le \sup_{u\in\mathbb{R}} |\Pr(S_n/\sigma \le u) - \Phi(u)| + \frac{\varepsilon}{\sqrt{2\pi}} + \Pr(|R_n|/\sigma > \varepsilon).$$

Step 3: Apply the classical Berry–Esseen inequality to S_n. The summands $\psi(W_i; \theta_0, g_0, \alpha_0)$ are i.i.d., mean zero, and have variance σ^2. Under $\mathbb{E}|\psi(W; \theta_0, g_0, \alpha_0)|^3 < \infty$, the Berry–Esseen theorem implies

$$\sup_{u\in\mathbb{R}} |\Pr(S_n/\sigma \le u) - \Phi(u)| \le \frac{C_{\mathrm{BE}}}{\sqrt{n}} \frac{\mathbb{E}|\psi(W; \theta_0, g_0, \alpha_0)|^3}{\sigma^3}.$$

Combining Steps 2–3 yields the desired bound. □

B Efficient influence function for the Riesz score

Lemma 4 (Efficient influence function of the Riesz score)**.** *Let $\mathcal{P}$ be a dominated nonparametric model for the law of W with regular parametric submodels. Let $U = u(W) \in \mathbb{R}^J$ and let $\mathcal{X}$ be a σ–field. For $P \in \mathcal{P}$, define the regression nuisance $g_P := \mathbb{E}_P[U \mid \mathcal{X}]$ and the target functional $\theta(P) := \mathbb{E}_P[m(W, g_P)]$. Let $P_0 \in \mathcal{P}$ be the true law and write (θ_0, g_0) for the induced objects. Assume that the map $h \mapsto \mathbb{E}_{P_0}[m(W, h)]$ is linear and continuous on $L^2(P_0)^J$ restricted to $\mathcal{X}$–measurable h, so that there exists a unique $\mathcal{X}$–measurable $\alpha_0 \in L^2(P_0)^J$ satisfying the Riesz identity $\mathbb{E}_{P_0}[m(W, h)] = \mathbb{E}_{P_0}\big[\alpha_0(W)^\top h(W)\big]$ for all $\mathcal{X}$–measurable $h \in L^2(P_0)^J$. Then θ is pathwise differentiable at P_0 with efficient influence function $\psi_0(W) = m(W, g_0) - \theta_0 + \alpha_0(W)^\top\{U - g_0(W)\}$.*

Consequently, the semiparametric efficiency bound equals $\sigma^2_{\mathrm{eff}} := \mathrm{Var}_{P_0}(\psi_0(W)) = \mathbb{E}_{P_0}\big[\psi_0(W)^2\big]$, where the last equality uses $\mathbb{E}_{P_0}[\psi_0(W)] = 0$.

Proof. Let $\{P_t : t \in (-\varepsilon, \varepsilon)\}$ be an arbitrary regular parametric submodel through P_0 with density p_t (w.r.t. a common dominating measure) and score $s(W) := \partial_t \log p_t(W)|_{t=0} \in L^2_0(P_0)$. Write $\mathbb{E}_t[\cdot]$ for expectation under P_t and denote $g_t := g_{P_t}$.

Step 1 (Pathwise derivative of $\theta(P_t)$). Define $f_t(W) := m(W, g_t)$. For regular

submodels,

$$\left.\frac{d}{dt}\right|_{t=0} \mathbb{E}_t[f_t(W)] = \mathbb{E}_0[f_0(W)\, s(W)] + \mathbb{E}_0\left[\left.\frac{d}{dt}\right|_{t=0} f_t(W)\right].$$

Since $f_0(W) = m(W, g_0)$, the first term equals $\mathbb{E}_0[m(W, g_0)s(W)]$.

Step 2 (Linearity + Riesz representation for the g_t-term). By linearity and continuity of $m(W, \cdot)$,

$$\left.\frac{d}{dt}\right|_{t=0} f_t(W) = m(W, \dot{g}), \qquad \dot{g} := \left.\frac{d}{dt}\right|_{t=0} g_t.$$

By the Riesz representation property at P_0, $\mathbb{E}_0[m(W, \dot{g})] = \mathbb{E}_0\left[\alpha_0(W)^\top \dot{g}(W)\right]$.

Step 3 (Derivative of the conditional mean). A standard calculation for regular submodels yields

$$\dot{g}(W) = \mathbb{E}_0\big[(U - g_0(W))\, s(W) \mid \mathcal{X}\big].$$

Because α_0 is $\mathcal{X}$–measurable, iterated expectations give

$$\mathbb{E}_0\left[\alpha_0(W)^\top \dot{g}(W)\right] = \mathbb{E}_0\big[\alpha_0(W)^\top (U - g_0(W))\, s(W)\big].$$

Step 4 (Identify the gradient and conclude efficiency). Recall $\psi_0(W) = m(W, g_0) - \theta_0 + \alpha_0(W)^\top\{U - g_0(W)\}$. Combining Steps 1–3 yields

$$\begin{aligned}
\left.\frac{d}{dt}\right|_{t=0} \theta(P_t) &= \mathbb{E}_0[m(W, g_0)\, s(W)] + \mathbb{E}_0[m(W, \dot{g})] \\
&= \mathbb{E}_0\left[\{m(W, g_0) + \alpha_0(W)^\top (U - g_0(W))\}\, s(W)\right] \\
&= \mathbb{E}_0\left[\{m(W, g_0) - \theta_0 + \alpha_0(W)^\top (U - g_0(W))\}\, s(W)\right] \\
&= \mathbb{E}_0[\psi_0(W)\, s(W)]\,,
\end{aligned}$$

where the third line uses $\mathbb{E}_0[s(W)] = 0$.

In the dominated nonparametric model, the tangent space equals $L_0^2(P_0)$; hence the unique $\psi_0 \in L_0^2(P_0)$ representing the derivative is the efficient influence function, and the efficiency bound is $\mathbb{E}_0[\psi_0(W)^2]$. □

C Additional Simulation Results

C.1 Robustness to Dimensionality

To ensure that the adaptivity results are not specific to the high-dimensional setting ($d = 200$) presented in the main text, we repeated the phase-transition experiment with $p = 20$. Figure 8 reports the results. The findings mirror the high-dimensional case: the Adaptive Riesz-DML estimator successfully tracks the best-performing model in terms of RMSE and maintains valid coverage where fixed estimators fail. Notably, the under-coverage of the Random Forest in the linear regime is slightly less severe in low dimensions but remains present, reinforcing the necessity of adaptive selection even in simpler settings.

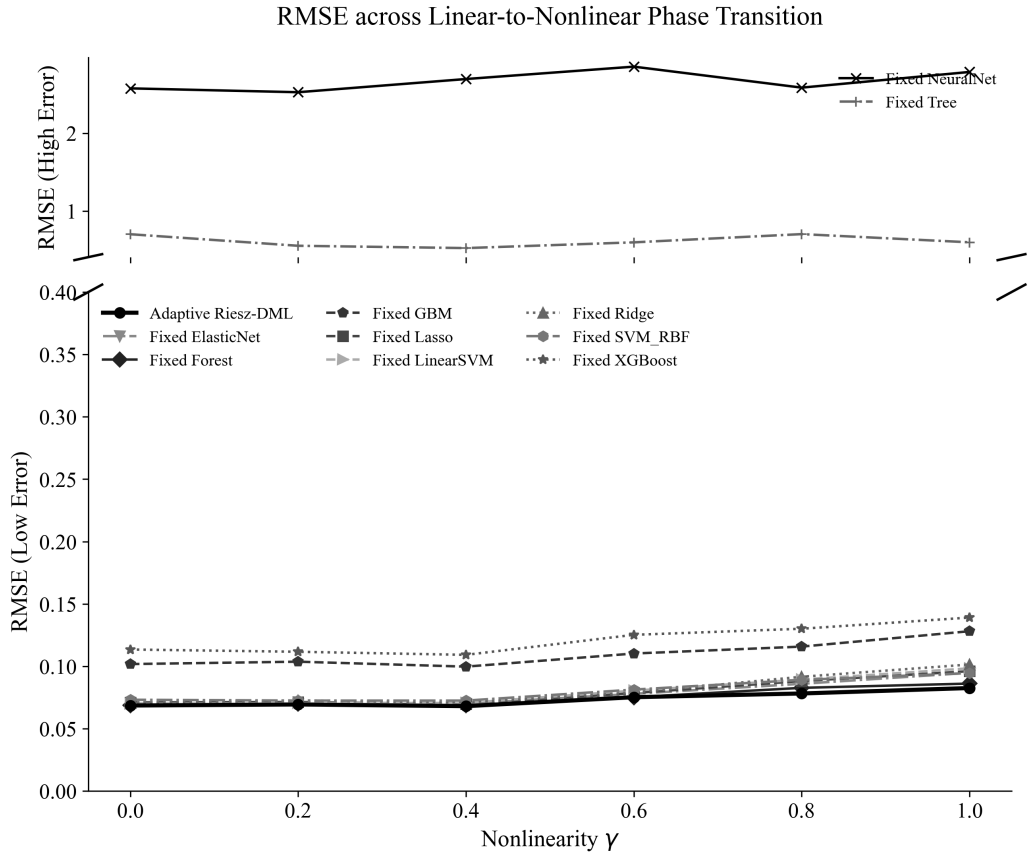

Figure 8: **Low-dimensional Phase Transition ($d = 20$).** Replication of the phase transition experiment in a low-dimensional setting. The qualitative dominance of the Adaptive estimator (solid black line) is preserved.

C.2 Finite-Sample Distributional Diagnostics

To verify the theoretical prediction that adaptive selection preserves inference while minimizing risk, we examine the full sampling distribution of the estimators beyond the first two moments. Figure 9 summarizes the distribution of estimation errors $(\hat{\theta} - \theta_0)$ for the full library of learners. The results highlight two distinct failure modes of fixed strategies that the adaptive procedure successfully avoids. First, in the sparse linear regime (DGP 1, left panel), high-complexity learners such as the Fixed Neural Network exhibit extreme variance, with error whiskers extending significantly beyond the scale of the linear benchmarks. This confirms that unconditionally using flexible architectures can induce heavy-tailed sampling distributions when the underlying signal is simple. Second, in the nonlinear regime (DGP 2, right panel), fixed linear estimators exhibit systematic bias, with median errors visibly shifted away from zero. In contrast, the Adaptive Riesz-DML estimator (dark grey) consistently recovers a centered, low-variance distribution. In DGP 1, Adaptive Riesz–DML yields a concentrated error distribution and near-nominal coverage, though the smallest RMSE is attained by the best-performing fixed linear benchmark in this design. In DGP 2, it remains centered at zero, effectively filtering out the biased linear candidates. This confirms the adaptive orthogonality result: the estimator automatically navigates the bias-variance trade-off to preserve valid inference.

Finally, we examine the price of adaptivity. A central concern in semiparametric theory is whether searching over a complex library inflates variance when a simple model would suffice. Figure 10 provides a granular test of this trade-off by plotting the absolute estimation error of the Adaptive estimator against the best available fixed benchmark for each replication. In the sparse linear regime (DGP 1, left panel), the replications cluster tightly along the 45° line. This provides direct visual confirmation of the oracle property: the price of adaptivity—the efficiency loss incurred by not knowing the true linear structure ex ante—is statistically negligible. The adaptive procedure effectively collapses to the optimal linear benchmark when the data supports it. In contrast, the nonlinear regime (DGP 2, right panel) reveals the asymmetric payoff of this strategy. The cloud of points disperses into the upper-left region, indicating that when the estimators diverge, the fixed benchmark typically incurs significantly larger errors than the adaptive estimator. This confirms that the procedure offers substantial robustness against misspecification bias in complex settings while incurring virtually no penalty in simple ones.

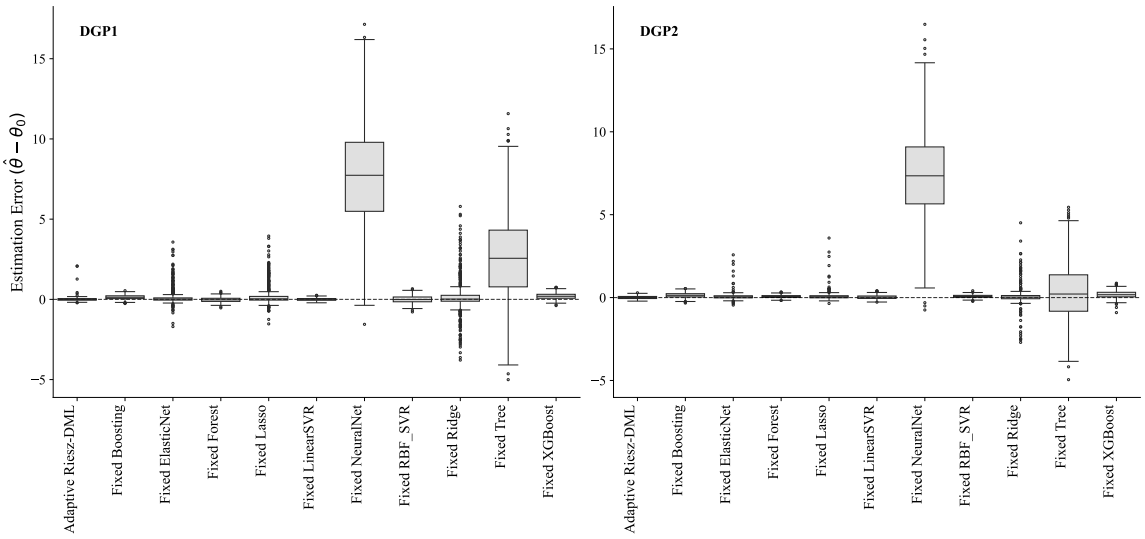

Figure 9: **Distribution of estimation errors across regimes.** Boxplots display the interquartile range (box), median (line), and 1.5×IQR whiskers of the estimation error $(\hat{\theta} - \theta_0)$ for all candidate learners. The Adaptive Riesz-DML estimator (dark grey) consistently achieves a centered, low-variance distribution. Note that fixed learners exhibit regime-specific failures: linear learners suffer bias in DGP 2 (shifted medians), while complex learners like Neural Networks suffer extreme variance in both regimes (wide whiskers). The comparatively larger RMSEs for some fixed linear benchmarks in DGP 1 are driven by rare but large outliers, which appear only as scattered points in the boxplots and can be visually muted on a vertical scale dominated by the catastrophic Tree/NeuralNet errors.

D Additional Empirical Diagnostics for the 401(k) Application

This appendix reports additional diagnostics from the nested cross-validation stage that inform nuisance learning and the selection decisions summarized in Table 3. For readability, we present high-contrast diagnostic summaries; these figures complement the main text by showing that (i) flexible nonlinear propensity learners tend to achieve better classification fit and better representer-oriented diagnostics, while (ii) differences across outcome learners are comparatively modest, consistent with asymmetric nuisance complexity.

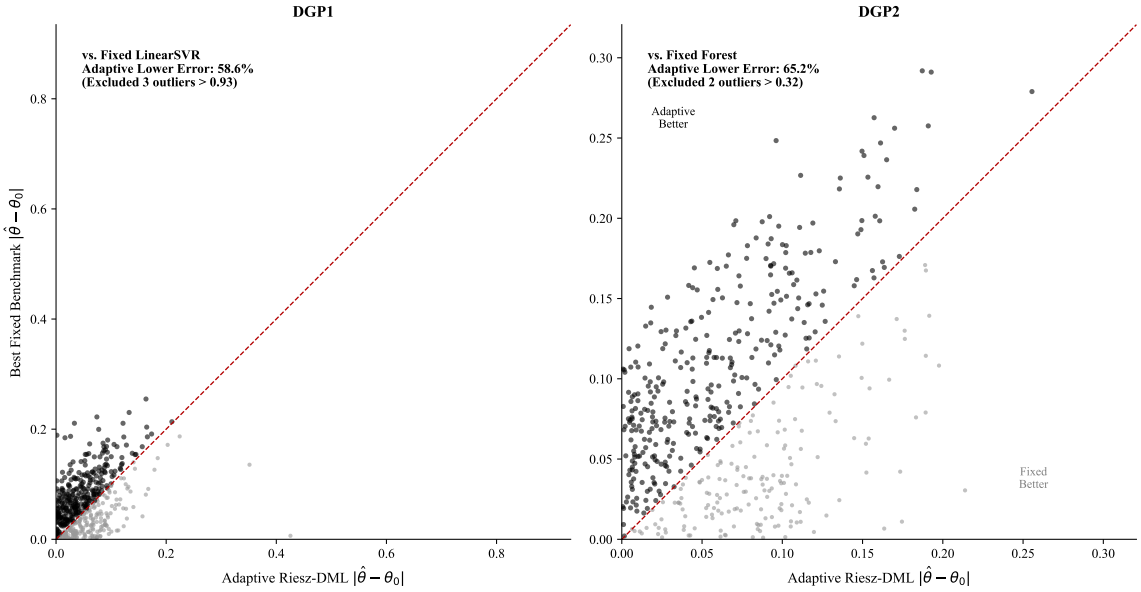

Figure 10: **Paired comparison of absolute estimation errors $|\hat{\theta} - \theta_0|$.** The x-axis represents the Adaptive Riesz-DML estimator, and the y-axis represents the best-performing fixed benchmark (LinearSVR for DGP 1, Random Forest for DGP 2). Points above the red 45° dashed line indicate replications where the Adaptive estimator achieved lower error. In the linear regime (DGP 1), the Adaptive estimator tracks the oracle benchmark, with the bulk of the distribution lying on the diagonal. In the nonlinear regime (DGP 2), the Adaptive estimator frequently outperforms the benchmark (points in the upper-left region). To visualize the density of the primary distribution, axes are zoomed to the 99.5% quantile. As noted in the figure text, a small number of extreme outliers (e.g., $< 1\%$ of draws) with errors exceeding the axis limits are excluded from the visual but included in all tabular RMSE calculations.

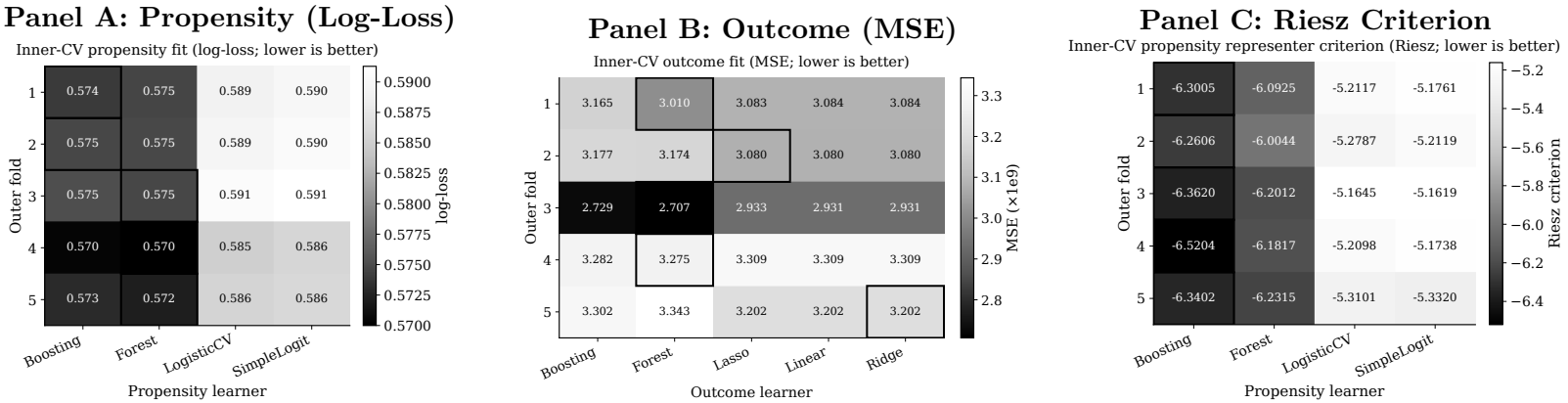

Figure 11: **Inner-CV diagnostics for nuisance selection in the 401(k) application.** Panel A reports inner-CV log-loss for propensity learners (used for hyperparameter tuning within architectures). Panel B reports inner-CV MSE for outcome learners. Panel C reports the representer-oriented diagnostic used to compare propensity architectures via the implied weights $\hat{\alpha}$. Highlighted winners correspond to strict argmins of mean inner-CV criteria. The diagnostics display the strict inner-CV argmins of the mean criteria. However, the main application tables implement the stable ε-minimizer rule from Section 3.6, so the selected learner in a fold need not coincide with the strict argmin when several candidates are statistically indistinguishable. In folds where Forest is selected in Table 3 despite Boosting being the strict argmin in Panel C, the Forest and Boosting criteria lie within the estimated one-standard-error band, and the ε-minimizer rule selects the simplest/stablest architecture among near-minimal candidates.