Adaptive Nuisance Selection for Weak-Identification-Robust Inference

Tamer Çetin*

March 2026

Abstract

Weak-identification-robust procedures are usually studied for a fixed score or nuisance pipeline. This paper studies adaptive selection of that pipeline. Under weak identification, data reuse in tuning can change the null law of otherwise robust statistics, and prediction-optimal tuning can attenuate residualized first-stage signal. For adaptive selection over curated nuisance libraries, with strength-selection theory developed in orthogonal PLIV/IV settings, the paper develops two primitives: leakage-free nested cross-fitting, which separates selection, estimation, and inference within each outer fold, and identification-aware cross-validation, which selects among near-prediction-optimal candidates using a feasible first-stage-strength proxy. The results show that leaky tuning can be first-order; screening can preserve orthogonal-learning rate requirements; prediction-only tuning can select pipelines with negligible residualized first-stage strength, whereas identification-aware tuning satisfies a near-oracle bound for the screened strength proxy and has no first-order effect within asymptotically equivalent screened sets; and cross-fitted orthogonal Anderson–Rubin inference remains uniformly valid, with ridge Anderson–Rubin and conditional QLR extensions.

Keywords: weak instruments; weak identification; machine learning; sample splitting; cross-fitting; identification-aware tuning; Anderson–Rubin; uniform inference.
JEL: C12, C13, C14, C36.

*UC Berkeley – tamercetin@berkeley.edu.

1 Introduction

Weak identification remains a central source of fragility in instrumental-variables and general moment-condition inference. Under weak instruments, Wald and ratio procedures can exhibit severe size distortions, which is why the modern identification-robust toolkit centers on Anderson–Rubin, pivotal LM, and conditional likelihood-ratio methods (Anderson and Rubin, 1949; Kleibergen, 2002; Moreira, 2003; Andrews, Moreira, and Stock, 2006). In classical low-dimensional settings these procedures deliver validity that is uniform in instrument strength (Staiger and Stock, 1997; Stock and Wright, 2000). The same literature includes K-type tests for GMM without assuming identification (Kleibergen, 2005), procedures with correct size under arbitrarily weak instruments (Moreira, 2009), and identification- and singularity-robust inference for general moment-condition models (Andrews and Guggenberger, 2019). Weak instruments remain empirically consequential even when conventional diagnostics appear reassuring (Bound et al., 1995; Lee et al., 2022), and Anderson–Rubin-type procedures have proved valuable well beyond textbook weak-IV environments (Keane and Neal, 2024).

Modern empirical work increasingly combines instruments with high-dimensional controls, flexible reduced forms, and machine-learned nuisance components, while orthogonal scores together with cross-fitting have become standard tools for preserving first-order properties under regular identification (Chernozhukov et al., 2018). This places the analysis within the broader orthogonal-score and locally robust literature (Belloni et al., 2015; Chernozhukov et al., 2022). Under weak identification, however, adaptive nuisance learning enters the inferential experiment itself. Hyperparameter choice, model search, and sample splitting shape not only the estimated nuisance functions but also the null law and identifying content of the score on which robust inference is based.

Two mechanisms are first-order. The first is leakage: when tuning reuses inference-fold observations, the selected nuisance pipeline need not be conditionally independent of the inference-fold score. Under weak identification, where relevant drifts operate on the same scale as the null fluctuation, that dependence can alter the limiting null law of otherwise robust procedures. The second is strength attenuation: prediction-optimal tuning need not preserve post-residualization identifying variation. Two candidate pipelines can be nearly indistinguishable in predictive risk and yet generate sharply different residualized first-stage signal, so that prediction-only tuning can attenuate local power and produce weakly informative robust confidence sets even when substantially stronger candidates are available within the same library. The distinction between predictive usefulness and identifying content is closely related to the broader emphasis on sharp and informative instruments in causal analysis (Kennedy et al., 2020). Weak-identification robustness is therefore a property of the adaptive pipeline that constructs the score.

The identification-aware strength-selection theory is developed for finite curated libraries in orthogonal PLIV/IV settings. The broader weakly identified moment-condition framework organizes the robust-inference results once the adaptive procedure has constructed a score. The analysis is organized around two primitives. Leakage-free nested cross-fitting (LF–NCF) imposes the filtration under which selection, estimation, and inference are separated within each outer fold. Identification-aware cross-validation (IACV) selects within a near-prediction-optimal screen by maximizing a feasible proxy for residualized first-stage strength. In the canonical homoskedastic Staiger–Stock PLIV benchmark, this criterion is monotone in the Anderson–Rubin noncentrality index. In the orthogonal PLIV environments studied here, it is used as an operational proxy for the residualized first-stage content that drives the informativeness of robust inference.

This focus links three literatures. First, it extends the weak-identification uniformity perspective of Andrews and Cheng (2012) and Andrews, Cheng, and Guggenberger (2020) to settings in which the moment process is itself the output of a data-adaptive learning rule. Second, it enters the orthogonal-score and locally robust literature by treating adaptive nuisance selection as part of the inferential object rather than as background regularization (Belloni et al., 2015; Chernozhukov et al., 2022). Third, it complements recent work on weak-identification-robust orthogonal inference with flexible nuisance adjustment in settings such as high-dimensional LATE, panel IV double machine learning, and efficient machine-learning instruments (Ma, 2025; Baiardi et al., 2026; Scheidegger et al., 2026), together with current software implementations of weak-IV-robust confidence sets in orthogonal-learning workflows (DoubleML Developers, 2026). Related adaptive-DML work by Çetin (2026a) studies leakage-safe nested cross-validation for the standard strongly identified PLIV DML estimator, and Çetin (2026b) studies Riesz-risk cross-validation for linear-functional DML. The present paper differs from both by focusing on weak-identification-robust inference, where adaptive tuning can alter the null law of robust statistics and can attenuate residualized first-stage strength. The contribution here is the adaptive selection layer itself: the paper characterizes the filtration, screening, and strength-preserving tuning rules under which data-dependent choice over a finite curated library preserves both validity and informativeness under weak identification.

The results are organized around that finite-library problem. A first set of results shows that leaky tuning can change the null law, including under full-sample search. A screening theorem then characterizes conditions under which adaptive selection preserves the nuisance-rate and product-rate restrictions required by orthogonal learning over growing curated libraries, with a transparent margin corollary and a worked sparse-linear verification. Propositions 3–5 show that prediction-only tuning can select asymptotically negligible proxy strength even when stronger candidates are available, whereas IACV is near-oracle for the screened strength proxy and asymptotically equivalent to the prediction-optimal selector under strong identification. On the inferential side, the paper establishes uniform size for cross-fitted orthogonal Anderson–Rubin inference over weak-identification classes that include Staiger–Stock local-to-zero sequences; develops a dimension-agnostic ridge Anderson–Rubin procedure under an effective-rank condition with Gaussian conditional bootstrap size control; and derives the corresponding conditional QLR construction in the sense of Andrews and Mikusheva (2016), with the homoskedastic Gaussian linear-IV specialization reducing to classical CLR (Moreira, 2003; Andrews, Moreira, and Stock, 2006).

The Monte Carlo section isolates the two mechanisms emphasized in the theory: null-law distortion from leaky tuning and losses of informative power when prediction tuning absorbs residualized identifying variation. The empirical section uses the quarter-of-birth benchmark to trace how weak-identification-robust reporting reshapes the inferential message of a canonical design, and the online appendix reports an enriched-control implementation of the adaptive tuning architecture in the same setting.

Section 2 introduces the weak-identification framework and the orthogonalized PLIV example. Section 3 develops LF–NCF and the leaky-tuning results. Section 4 defines the strength proxy and IACV and states the screening and strength-preservation results. Section 5 presents the main uniform-validity theorem package, including the ridge Anderson–Rubin and conditional QLR extensions. Section 6 reports Monte Carlo evidence mapped to the theorem mechanisms. Section 7 reports the empirical illustration. Appendix A contains the proofs for the main text. The online appendix records the full conditional QLR/CLR construction, learner-class verification, supplementary technical material, and additional Monte Carlo diagnostics.

2 Weakly Identified Moment Models and Orthogonal AR

2.1 General framework

Let $W \in \mathcal{W}$ denote observable data and P its distribution. An i.i.d. sample $W_1, \ldots, W_n$ is observed from P. Let $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$ be the target parameter and $\eta \in \mathcal{H}$ a nuisance parameter. Consider a moment function $g(W; \theta, \eta) \in \mathbb{R}^{d_g}$ satisfying

$$\mathbb{E}_P[g(W; \theta_0, \eta_0)] = 0. \tag{2.1}$$

Define the standardized sample moment

$$g_n(\theta, \eta) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g(W_i; \theta, \eta). \tag{2.2}$$

Let

$$G_P := \partial_\theta \mathbb{E}_P[g(W; \theta, \eta_0)]\Big|_{\theta=\theta_0} \in \mathbb{R}^{d_g \times d_\theta}.$$

Definition 1 (Strong and weak identification)**.** Strong identification at P holds if G_P has full column rank and $\sigma_{\min}(G_P) \geq c > 0$. A sequence $\{P_n\}$ exhibits weak identification if $\sigma_{\min}(G_{P_n}) \to 0$. A local-to-zero sequence satisfies $\sigma_{\min}(G_{P_n}) \asymp n^{-1/2}$.

Let

$$\Omega_P(\theta, \eta) := \mathbb{E}_P[g(W; \theta, \eta) g(W; \theta, \eta)^\top].$$

Under strong identification and local alternatives $\theta_n = \theta_0 + \Delta/\sqrt{n}$, regularity yields

$$g_n(\theta_0, \eta_0) \xrightarrow{d} N(-G_P \Delta, \Omega_P(\theta_0, \eta_0)), \tag{2.3}$$

so an oracle AR statistic has a noncentral χ^2 limit with index

$$\lambda(\Delta; P) = \Delta^\top G_P^\top \Omega_P(\theta_0, \eta_0)^{-1} G_P \Delta. \tag{2.4}$$

The weak-identification regime changes the relevant scaling. If $P = P_n$ varies with n so that $G_{P_n} = \widetilde{G}/\sqrt{n} + o(n^{-1/2})$, then the strong-ID local alternative $\theta_n = \theta_0 + \Delta/\sqrt{n}$ yields a vanishing drift and therefore trivial local power. Nontrivial weak-ID power instead arises for *fixed alternatives* $\theta = \theta_0 + \Delta$, in which case $\sqrt{n}\,\mathbb{E}_{P_n}[g(W;\theta_0,\eta_0)] \to -\widetilde{G}\Delta$ and the relevant noncentrality index becomes $\Delta^\top \widetilde{G}^\top \Omega^{-1} \widetilde{G}\Delta$.

2.2 Leading example: partially linear IV

Let $W = (Y, D, Z, X)$ with scalar Y and D, instruments $Z \in \mathbb{R}^{d_z}$, and controls X. Consider

$$Y = \theta_0 D + \gamma_0(X) + U, \tag{2.5}$$

$$\mathbb{E}[U \mid Z, X] = 0. \tag{2.6}$$

Define

$$Y^e := Y - \mu_0(X), \qquad D^e := D - m_0(X), \qquad Z^e := Z - r_0(X).$$

Then

$$E\Big[Z^e\Big(Y^e - \theta_0 D^e\Big)\Big] = 0. \tag{2.7}$$

Write the residualized reduced form as

$$D^e = Z^{e\top}\Pi_n + V,$$

where Staiger–Stock weak-IV sequences correspond to $\Pi_n = \pi/\sqrt{n}$ (Staiger and Stock, 1997).

Weak identification shrinks the first-stage signal, but it does not inherently make the nuisance regressions harder to learn. In PLIV, the local-to-zero term enters conditional means only at $n^{-1/2}$ scale. Throughout the PLIV setup, assume $\sup_n \mathbb{E}_{P_n}[\|Z\|^2] < \infty$.

Lemma 1 (Target drift under local-to-zero sequences)**.** *Suppose in PLIV that*

$$D = m_0^\star(X) + Z^\top\Pi_n + V, \qquad \mathbb{E}[V \mid Z, X] = 0, \qquad \Pi_n = \pi/\sqrt{n}.$$

Let

$$m_0^\star(x) := \mathbb{E}[D \mid X = x, \Pi \equiv 0],$$
$$m_{0,n}(x) := \mathbb{E}_{P_n}[D \mid X = x]$$
$$= m_0^\star(x) + \mathbb{E}_{P_n}[Z \mid X = x]^\top \Pi_n.$$

If

$$\sup_n \mathbb{E}_{P_n}\Big[\| \mathbb{E}_{P_n}[Z \mid X] \|^2\Big] < \infty,$$

then

$$\|m_{0,n} - m_0^\star\|_{L_2(P_n)} = O(n^{-1/2}).$$

Consequently, for any learner $\widehat{m}$,

$$\|\widehat{m} - m_{0,n}\|_{L_2(P_n)} \le \|\widehat{m} - m_0^\star\|_{L_2(P_n)} + O(n^{-1/2}),$$

and

$$\|\widehat{m} - m_0^\star\|_{L_2(P_n)} \le \|\widehat{m} - m_{0,n}\|_{L_2(P_n)} + O(n^{-1/2}).$$

In particular, an $O_{\mathbb{P}}(r_n)$ rate for either target transfers to the other up to $n^{-1/2}$, and if $r_n = o(n^{-1/4})$, both targets are learned at $o_{\mathbb{P}}(n^{-1/4})$ rate.

Proof. See Appendix A.1. □

2.3 Orthogonal score and cross-fitted AR

An orthogonal score $\psi(W;\theta,\eta) \in \mathbb{R}^{d_g}$ is used and satisfies

$$\mathbb{E}[\psi(W;\theta_0,\eta_0)] = 0, \qquad \partial_\eta \mathbb{E}[\psi(W;\theta_0,\eta)]\Big|_{\eta=\eta_0} = 0. \tag{2.8}$$

For PLIV, a standard orthogonal score is

$$\psi(W;\theta,\eta) = \Big(Y - \mu(X) - \theta(D - m(X))\Big)\Big(Z - r(X)\Big), \qquad \eta = (\mu, m, r). \tag{2.9}$$

Let $\widehat{\eta}_{-i}$ denote a nuisance estimate trained on data disjoint from observation i. Define

$$\widehat{\psi}_i(\theta) := \psi(W_i;\theta,\widehat{\eta}_{-i}), \qquad \widehat{g}_n(\theta) := \frac{1}{\sqrt{n}}\sum_{i=1}^n \widehat{\psi}_i(\theta), \qquad \widehat{\Omega}_n(\theta) := \frac{1}{n}\sum_{i=1}^n \widehat{\psi}_i(\theta)\widehat{\psi}_i(\theta)^\top. \tag{2.10}$$

The orthogonalized Anderson–Rubin statistic is

$$AR_n(\theta) := \widehat{g}_n(\theta)^\top \widehat{\Omega}_n(\theta)^{-1} \widehat{g}_n(\theta). \tag{2.11}$$

Inverting $AR_n(\theta)$ yields robust confidence sets that can be wide or unbounded when identification is weak. This is expected under uniformity: when the noncentrality index is near zero, tests cannot distinguish parameter values, and inverted sets must reflect that lack of information.

Remark 1 (Point estimation under weak identification)**.** Even with orthogonalization, point-estimation risk can be fundamentally discontinuous in weakly identified GMM problems; see Andrews and Mikusheva (2022a,b). This reinforces the emphasis on test inversion under weak identification, while strong-identification efficiency remains relevant when the Jacobian is uniformly nonsingular.

3 Leakage-Free Nested Cross-Fitting

3.1 A common bad pipeline

A common workflow is: use all observations to tune hyperparameters by cross-validation or model search, then cross-fit nuisances with the selected hyperparameters, then compute weak-ID robust inference on those same folds. Even if each observation is excluded from nuisance *training*, it may still influence the chosen hyperparameter $\widehat{\gamma}$ if $\widehat{\gamma}$ is selected using a criterion computed on all observations. Then $\widehat{\eta}_{-i}$ depends on W_i through $\widehat{\gamma}$, violating the no-own-observation measurability condition used to obtain foldwise conditional independence of inference-fold scores.

3.2 The LF–NCF algorithm

LF–NCF pays a finite-sample cost for this protection because three-way splitting leaves fewer observations for nuisance estimation than ordinary two-way cross-fitting. The simulation section is therefore designed to evaluate not only validity but also when that cost is offset by the gains from preserving weak-ID robustness and identifying strength.

For $i \in I_k^{\text{inf}}$, write $\widehat{\eta}_{-i} := \widehat{\eta}_k$. Then the leave-one-out notation $\widehat{\psi}_i(\theta) = \psi(W_i; \theta, \widehat{\eta}_{-i})$ and the foldwise notation $\widehat{\psi}_i(\theta) = \psi(W_i; \theta, \widehat{\eta}_k)$ coincide. Throughout the analysis, let

Algorithm 1 Leakage-Free Nested Cross-Fitting (LF–NCF)

Partition $\{1,\dots,n\}$ into K outer inference folds $I_1^{\mathrm{inf}},\dots,I_K^{\mathrm{inf}}$.
for $k=1,\dots,K$ **do**
  Set the inference fold equal to I_k^{inf}.
  Split the complement into disjoint selection and estimation sets:

$$(I_k^{\mathrm{inf}})^c = I_k^{\mathrm{sel}} \cup I_k^{\mathrm{est}}.$$

  For each $\gamma \in \Gamma_n$, train candidate nuisance fits $\widehat{\eta}_k(\gamma)$ using only I_k^{est}.
  Using only the selection fold I_k^{sel}, together with candidate fits trained on I_k^{est}, compute the selection criteria and choose $\widehat{\gamma}_k$.
  Set $\widehat{\eta}_k := \widehat{\eta}_k(\widehat{\gamma}_k)$, or equivalently refit the selected nuisance pipeline on I_k^{est}.
  Evaluate scores $\widehat{\psi}_i(\theta) = \psi(W_i;\theta,\widehat{\eta}_k)$ for $i \in I_k^{\mathrm{inf}}$.
end for
Aggregate the scores across folds to form $\widehat{g}_n(\theta)$, $\widehat{\Omega}_n(\theta)$, and $AR_n(\theta)$.

$n_{\mathrm{sel}} := |I_k^{\mathrm{sel}}|$ denote the size of the selection fold, and let $n_{\mathrm{inf}} := |I_k^{\mathrm{inf}}|$ denote the size of the inference fold.

Let

$$\mathcal{F}_k^{\mathrm{sel}} := \sigma(W_i : i \in I_k^{\mathrm{sel}}), \qquad \mathcal{F}_k^{\mathrm{est}} := \sigma(W_i : i \in I_k^{\mathrm{est}}), \qquad \mathcal{F}_k^{\mathrm{inf}} := \sigma(W_i : i \in I_k^{\mathrm{inf}}),$$

and define $\mathcal{F}_k^{\mathrm{train}} := \mathcal{F}_k^{\mathrm{sel}} \vee \mathcal{F}_k^{\mathrm{est}}$. By construction, $\mathcal{F}_k^{\mathrm{inf}} \perp \mathcal{F}_k^{\mathrm{train}}$.

Lemma 2 (Conditional i.i.d. structure)**.** *Under LF–NCF, for each outer fold k, conditional on $\mathcal{F}_k^{\mathrm{train}}$, the array $\{W_i : i \in I_k^{\mathrm{inf}}\}$ is i.i.d. and independent of $(\widehat{\gamma}_k, \widehat{\eta}_k)$.*

Proof of Lemma 2. Fix an outer fold index $k \in \{1,\dots,K\}$. Conditional on the realized fold assignment, the index sets I_k^{inf}, I_k^{sel}, and I_k^{est} are deterministic and disjoint, and $\{W_i\}_{i=1}^n$ remain i.i.d. under P.

Define the σ-fields

$$\mathcal{F}_k^{\mathrm{sel}} := \sigma(W_i : i \in I_k^{\mathrm{sel}}), \qquad \mathcal{F}_k^{\mathrm{est}} := \sigma(W_i : i \in I_k^{\mathrm{est}}), \qquad \mathcal{F}_k^{\mathrm{inf}} := \sigma(W_i : i \in I_k^{\mathrm{inf}}),$$

and let $\mathcal{F}_k^{\mathrm{train}} := \mathcal{F}_k^{\mathrm{sel}} \vee \mathcal{F}_k^{\mathrm{est}}$.

By construction of LF–NCF, the candidate nuisance fits are trained on I_k^{est}, and the selection criteria are evaluated on I_k^{sel}. Hence the selected hyperparameter $\widehat{\gamma}_k$ is measurable with respect to

$$\mathcal{F}_k^{\mathrm{train}} := \mathcal{F}_k^{\mathrm{sel}} \vee \mathcal{F}_k^{\mathrm{est}}.$$

The selected nuisance fit $\widehat{\eta}_k = \widehat{\eta}_k(\widehat{\gamma}_k)$ is also $\mathcal{F}_k^{\mathrm{train}}$-measurable.

Because the data are i.i.d., for any two disjoint index sets $A, B \subset \{1, \ldots, n\}$ the collections $\{W_i : i \in A\}$ and $\{W_i : i \in B\}$ are independent. Applying this with $A = I_k^{\mathrm{sel}} \cup I_k^{\mathrm{est}}$ and $B = I_k^{\mathrm{inf}}$ yields

$$\mathcal{F}_k^{\mathrm{train}} \perp \mathcal{F}_k^{\mathrm{inf}}.$$

Let $n_k := |I_k^{\mathrm{inf}}|$ and enumerate $I_k^{\mathrm{inf}} = \{i_1, \ldots, i_{n_k}\}$. For any bounded measurable functions $f_1, \ldots, f_{n_k}$,

$$\mathbb{E}\left[\prod_{\ell=1}^{n_k} f_\ell(W_{i_\ell}) \,\middle|\, \mathcal{F}_k^{\mathrm{train}}\right] = \mathbb{E}\left[\prod_{\ell=1}^{n_k} f_\ell(W_{i_\ell})\right] = \prod_{\ell=1}^{n_k} \mathbb{E}[f_\ell(W_1)],$$

where the first equality uses Step 2 and the second uses i.i.d. sampling. Hence the conditional law of $\{W_i : i \in I_k^{\mathrm{inf}}\}$ given $\mathcal{F}_k^{\mathrm{train}}$ is $P^{\otimes n_k}$.

By Step 1, $(\widehat{\gamma}_k, \widehat{\eta}_k)$ is $\mathcal{F}_k^{\mathrm{train}}$-measurable. Because $\mathcal{F}_k^{\mathrm{inf}} \perp \mathcal{F}_k^{\mathrm{train}}$, the inference-fold observations are independent of every $\mathcal{F}_k^{\mathrm{train}}$-measurable random element, and hence of $(\widehat{\gamma}_k, \widehat{\eta}_k)$. Equivalently, for any bounded measurable functions u and v,

$$\mathbb{E}\Big[u\Big((W_i)_{i \in I_k^{\mathrm{inf}}}\Big)\, v(\widehat{\gamma}_k, \widehat{\eta}_k) \,\Big|\, \mathcal{F}_k^{\mathrm{train}}\Big] = \mathbb{E}\Big[u\Big((W_i)_{i \in I_k^{\mathrm{inf}}}\Big) \,\Big|\, \mathcal{F}_k^{\mathrm{train}}\Big]\, v(\widehat{\gamma}_k, \widehat{\eta}_k),$$

which is exactly the claimed conditional independence. □

Remark 2 (Foldwise conditioning and the aggregated score)**.** Lemma 2 is a foldwise statement. It does not assert the existence of a single global conditioning σ-field under which the fully aggregated cross-fitted score is conditionally i.i.d. Rather, LF–NCF removes own-observation leakage: for every observation evaluated on an inference fold, the tuning and nuisance-fitting objects used to evaluate its score are measurable with respect to data outside that fold. The uniform AR theorem below then uses this foldwise measurability together with the selected-nuisance oracle-equivalence conditions in Assumption 5.

3.3 Leaky tuning changes the null law

The next proposition isolates the central mechanism: selection that uses inference-fold quadratic forms can change the null distribution of a weak-ID robust statistic at first order.

Proposition 1 (Leaky selection destroys pivotality)**.** *Fix $\theta = \theta_0$ and suppose two candidate nuisance pipelines $\widehat{\eta}^{(1)}$ and $\widehat{\eta}^{(2)}$ satisfy, under P,*

$$g_n^{(j)} := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(W_i; \theta_0, \widehat{\eta}^{(j)}) \xrightarrow{d} N(0, \Omega), \qquad j \in \{1, 2\},$$

with the same nonsingular limit Ω, and suppose $(g_n^{(1)}, g_n^{(2)})$ has a nondegenerate joint Gaussian limit. Consider the leaky *rule that selects $\widehat{\eta} = \widehat{\eta}^{(1)}$ if*

$$g_n^{(1)\top} \Omega^{-1} g_n^{(1)} \leq g_n^{(2)\top} \Omega^{-1} g_n^{(2)},$$

and selects $\widehat{\eta} = \widehat{\eta}^{(2)}$ otherwise, and then reports

$$AR_n(\theta_0) = g_n(\widehat{\eta})^\top \Omega^{-1} g_n(\widehat{\eta}).$$

Then $AR_n(\theta_0)$ converges to the distribution of $\min\{Q_1, Q_2\}$, where $Q_j = \left\|\Omega^{-1/2} G_j\right\|^2$ and (G_1, G_2) is the joint Gaussian limit. In particular, for any $\alpha \in (0, 1)$,

$$\mathbb{P}\Big(AR_n(\theta_0) > \chi^2_{d_g, 1-\alpha}\Big) \not\to \alpha,$$

so the usual critical value no longer delivers pivotal weak-ID inference.

Proof. See Appendix A.2. □

Corollary 1 (Global tuning that reuses inference observations is generally leaky)**.** *Let $\widehat{\gamma}$ be a tuning rule that is computed from the full sample $(W_1, \ldots, W_n)$. Suppose that for some outer fold k there exist two candidate values $\gamma_1, \gamma_2 \in \Gamma_n$ and an event $A \in \mathcal{F}_k^{\text{train}}$ with $\Pr(A) > 0$ such that on A,*

$$0 < \Pr\Big(\widehat{\gamma} = \gamma_1 \mid \mathcal{F}_k^{\text{train}}\Big) < 1.$$

Then $\widehat{\gamma}$ is not $\mathcal{F}_k^{\text{train}}$-measurable, so the conditional i.i.d. conclusion of Lemma 2 does not apply. In particular, weak-identification-robust pivotality is not guaranteed. Moreover, under the two-candidate selection design of Proposition 1, the resulting size distortion is of order $O(1)$.

Proof. See Appendix A.3. □

Proposition 2 (A stylized full-sample leaky search rule destroys weak-ID pivotality)**.** *Fix $d_g \geq 1$ and let $\Gamma = \{\gamma_1, \gamma_2\}$ index two candidate nuisance pipelines. For each $j \in \{1, 2\}$, let*

$$g_n^{(j)}(\theta_0) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi\Big(W_i; \theta_0, \hat{\eta}(\gamma_j)\Big) \in \mathbb{R}^{d_g}, \qquad AR_n^{(j)}(\theta_0) := \Big(g_n^{(j)}(\theta_0)\Big)^{\top} \Omega^{-1} g_n^{(j)}(\theta_0),$$

where $\Omega \in \mathbb{R}^{d_g \times d_g}$ is positive definite and does not depend on n.

Assume that under $H_0 : \theta = \theta_0$,

$$\Big(\Omega^{-1/2} g_n^{(1)}(\theta_0), \Omega^{-1/2} g_n^{(2)}(\theta_0)\Big) \Rightarrow (G_1, G_2), \tag{3.1}$$

where (G_1, G_2) is jointly Gaussian with nonsingular covariance matrix, $E[G_j] = 0$, and $\mathrm{Var}(G_j) = I_{d_g}$ for $j = 1, 2$. Consider the leaky global tuning rule

$$\widehat{\gamma} \in \arg\min_{\gamma \in \Gamma} AR_n^{(\gamma)}(\theta_0), \qquad AR_n^{(\widehat{\gamma})}(\theta_0) = \min\{AR_n^{(1)}(\theta_0), AR_n^{(2)}(\theta_0)\}.$$

Then

$$AR_n^{(\widehat{\gamma})}(\theta_0) \Rightarrow \min\{Q_1, Q_2\}, \qquad Q_j := \|G_j\|^2 \sim \chi^2_{d_g}.$$

Consequently, if $c_{1-\alpha}$ denotes the $(1-\alpha)$ quantile of $\chi^2_{d_g}$,

$$\Pr\Big(AR_n^{(\widehat{\gamma})}(\theta_0) > c_{1-\alpha}\Big) \to \Pr\Big(\min\{Q_1, Q_2\} > c_{1-\alpha}\Big) < \alpha,$$

so the nominal α AR test is not asymptotically correctly sized under this leaky tuning rule.

Proof of Proposition 2. By definition,

$$AR_n^{(j)}(\theta_0) = \Big\|\Omega^{-1/2} g_n^{(j)}(\theta_0)\Big\|^2, \qquad j = 1, 2.$$

The map $x \mapsto \|x\|^2$ is continuous on $\mathbb{R}^{d_g}$ and the map $(u, v) \mapsto \min\{\|u\|^2, \|v\|^2\}$ is continuous on $\mathbb{R}^{2d_g}$. Applying the continuous mapping theorem to (3.1) therefore yields

$$AR_n^{(\widehat{\gamma})}(\theta_0) = \min\left\{\Big\|\Omega^{-1/2} g_n^{(1)}(\theta_0)\Big\|^2,\ \Big\|\Omega^{-1/2} g_n^{(2)}(\theta_0)\Big\|^2\right\} \Rightarrow \min\{\|G_1\|^2, \|G_2\|^2\}.$$

This proves the weak convergence claim with $Q_j := \|G_j\|^2 \sim \chi^2_{d_g}$.

Now fix $\alpha \in (0,1)$ and let $c_{1-\alpha}$ be the $(1-\alpha)$ quantile of $\chi^2_{d_g}$ so that $\Pr(Q_1 > c_{1-\alpha}) = \alpha$. Because $\min\{Q_1, Q_2\} > c_{1-\alpha}$ implies $Q_1 > c_{1-\alpha}$,

$$\Pr(\min\{Q_1, Q_2\} > c_{1-\alpha}) \leq \Pr(Q_1 > c_{1-\alpha}) = \alpha.$$

To show strict inequality, note that (G_1, G_2) has a jointly Gaussian density that is strictly positive everywhere in $\mathbb{R}^{2d_g}$ because its covariance matrix is nonsingular. Hence the set

$$\mathcal{A} := \{(u, v) \in \mathbb{R}^{2d_g} : \ \|u\|^2 > c_{1-\alpha} \text{ and } \|v\|^2 < c_{1-\alpha}\}$$

is open and nonempty, and therefore has strictly positive probability under (G_1, G_2). On $\mathcal{A}$, $Q_1 > c_{1-\alpha}$ but $\min\{Q_1, Q_2\} \leq c_{1-\alpha}$, so

$$\Pr(\min\{Q_1, Q_2\} > c_{1-\alpha}) = \Pr(Q_1 > c_{1-\alpha}) - \Pr(Q_1 > c_{1-\alpha},\ Q_2 \leq c_{1-\alpha}) < \alpha.$$

Finally, continuity of the limiting cdf at $c_{1-\alpha}$ implies

$$\Pr\Big(AR_n^{(\widehat{\gamma})}(\theta_0) > c_{1-\alpha}\Big) \to \Pr\Big(\min\{Q_1, Q_2\} > c_{1-\alpha}\Big) < \alpha.$$

□

Remark 3 (Why LF–NCF matters)**.** Nested splitting is not new as a machine-learning device. The weak-identification issue is that reuse of inference-fold information inside tuning can be first-order for otherwise robust statistics. LF–NCF imposes a foldwise no-own-observation measurability restriction: the selected pipeline used on an inference fold is measurable with respect to data outside that fold. This rules out the inference-fold winner's-curse and adaptive selection-bias effects generated by score-evaluation reuse.

4 Identification-Aware Cross-Validation

4.1 Population power index and feasible strength proxy

In PLIV, the Jacobian of the orthogonal moment with respect to θ equals the negative residualized first-stage covariance:

$$G(\eta) := \partial_\theta \mathbb{E}[\psi(W;\theta,\eta)]\Big|_{\theta=\theta_0} = -\Pi(\eta),$$

where

$$\Pi(\eta) := \mathbb{E}[Z^e(\eta)D^e(\eta)] \in \mathbb{R}^{d_z}, \qquad \Sigma_Z(\eta) := \mathbb{E}[Z^e(\eta)Z^e(\eta)^\top], \tag{4.1}$$

with $Z^e(\eta) = Z - r(X)$ and $D^e(\eta) = D - m(X)$.

Define the effective concentration functional

$$S_n(\eta) := n\,\Pi(\eta)^\top \Sigma_Z(\eta)^{-1}\Pi(\eta). \tag{4.2}$$

In the canonical homoskedastic Staiger–Stock experiment, $S_n(\eta_0)$ is proportional to the concentration parameter and monotone in the AR noncentrality index. On a selection fold I^{sel} with $n_{\text{sel}} = |I^{\text{sel}}|$, define

$$\widehat{\Pi}(\gamma) := \frac{1}{n_{\text{sel}}} \sum_{i \in I^{\text{sel}}} \widehat{Z}_i^e(\gamma)\widehat{D}_i^e(\gamma), \tag{4.3}$$

$$\widehat{\Sigma}_Z(\gamma) := \frac{1}{n_{\text{sel}}} \sum_{i \in I^{\text{sel}}} \widehat{Z}_i^e(\gamma)\widehat{Z}_i^e(\gamma)^\top, \tag{4.4}$$

where $\widehat{Z}_i^e(\gamma) = Z_i - \widehat{r}_\gamma(X_i)$ and $\widehat{D}_i^e(\gamma) = D_i - \widehat{m}_\gamma(X_i)$. Introducing ridge stabilization $\kappa \geq 0$, define the feasible strength proxy

$$\widehat{S}_{n,\kappa}(\gamma) := n_{\text{sel}}\,\widehat{\Pi}(\gamma)^\top \Big(\widehat{\Sigma}_Z(\gamma) + \kappa I_{d_z}\Big)^{-1}\widehat{\Pi}(\gamma). \tag{4.5}$$

The exact local-power interpretation is sharpest in the canonical homoskedastic PLIV benchmark. Outside that benchmark, $\widehat{S}_{n,\kappa}$ is used as an operational proxy for residualized identifying strength, not as a universal exact power index for every weakly identified moment model.

Because n_{sel} is common across candidates within a given outer fold, maximizing

$\widehat{S}_{n,\kappa}(\gamma)$ over Γ_n is equivalent to maximizing the normalized criterion

$$\overline{S}_{n,\kappa}(\gamma) := n_{\mathrm{sel}}^{-1}\widehat{S}_{n,\kappa}(\gamma) = \widehat{\Pi}(\gamma)^\top\Big(\widehat{\Sigma}_Z(\gamma)+\kappa I\Big)^{-1}\widehat{\Pi}(\gamma).$$

For a population nuisance value η, define the ridge-stabilized population strength target

$$S_{n,\kappa}(\eta) := n_{\mathrm{sel}}\,\Pi(\eta)^\top\Big(\Sigma_Z(\eta)+\kappa I_{d_z}\Big)^{-1}\Pi(\eta), \qquad \overline{S}_{n,\kappa}(\eta) := n_{\mathrm{sel}}^{-1}S_{n,\kappa}(\eta).$$

The normalized population analogue is $\overline{S}_{n,\kappa}(\eta)$ defined above. All argmax statements below may therefore be written in either normalized or unnormalized form.

Remark 4 (Heteroskedasticity and feasibility)**.** In heteroskedastic PLIV, exact weak-ID power for AR-type procedures depends on nuisance-dependent second moments. The feasible criterion $\widehat{S}_{n,\kappa}$ is used as an operational strength proxy: it targets residualized first-stage content while avoiding plug-in quantities that are not uniformly estimable under weak identification. In the canonical homoskedastic Staiger–Stock benchmark, Theorem 1 gives the exact local-power interpretation; under bounded conditional heteroskedasticity, the same proxy tracks the AR noncentrality index up to constants.

4.2 Canonical power theorem

Theorem 1 (Canonical Staiger–Stock experiment: AR power is monotone in effective concentration)**.** *Consider the triangular-array linear IV model*

$$Y = \theta D + U, \qquad D = Z^\top\Pi_n + V,$$

where $Z \in \mathbb{R}^{d_z}$ satisfies $\mathbb{E}[Z]=0$, $\mathbb{E}[ZZ^\top]=\Sigma_Z \succ 0$, $\mathbb{E}[\|Z\|^4]<\infty$, and Z is independent of (U,V). Assume $\mathbb{E}[U]=\mathbb{E}[V]=0$ and $\Pi_n=\pi/\sqrt{n}$ for a fixed $\pi\in\mathbb{R}^{d_z}$. Define

$$g_n(\vartheta) := \frac{1}{\sqrt{n}}\sum_{i=1}^n Z_i(Y_i-\vartheta D_i), \qquad \Omega_n(\vartheta) := \mathbb{E}[ZZ^\top(Y-\vartheta D)^2],$$

and the oracle AR statistic $AR_n(\vartheta) := g_n(\vartheta)^\top\Omega_n(\vartheta)^{-1}g_n(\vartheta)$. Then under the fixed *alternative* $\theta = \theta_0 + \Delta$ *with* $\Delta \neq 0$*,*

$$AR_n(\theta_0) \xrightarrow{d} \chi^2_{d_z}(\lambda_\Delta), \qquad \lambda_\Delta = (\Delta \Sigma_Z \pi)^\top \Omega_\Delta^{-1} (\Delta \Sigma_Z \pi),$$

where $\Omega_\Delta = \lim_{n\to\infty} \Omega_n(\theta_0) = \mathbb{E}[ZZ^\top (U + \Delta V)^2]$*. Under independence,*

$$\Omega_\Delta = \sigma^2_\Delta \Sigma_Z, \qquad \sigma^2_\Delta := \mathbb{E}[(U + \Delta V)^2],$$

so

$$\lambda_\Delta = \frac{\Delta^2}{\sigma^2_\Delta} \pi^\top \Sigma_Z \pi.$$

Consequently, the limiting rejection probability $\mathbb{P}(AR_n(\theta_0) > \chi^2_{d_z,1-\alpha})$ *is strictly increasing in* $\pi^\top \Sigma_Z \pi$*, equivalently in* $n\, \Pi_n^\top \Sigma_Z \Pi_n$*.*

Proof. See Appendix A.4. □

Remark 5 (Detection boundary under weak IV)**.** In the Staiger–Stock experiment, the strong-ID local alternative $\theta_n = \theta_0 + \Delta/\sqrt{n}$ yields a vanishing drift and hence asymptotic power equal to size. Theorem 1 therefore uses fixed alternatives $\theta = \theta_0 + \Delta$.

4.3 The IACV algorithm

Let Γ_n denote a finite library of nuisance pipelines. A generic element $\gamma \in \Gamma_n$ may index an entire tuple of learners and hyperparameters for (μ, m, r). On selection data, define a predictive-risk criterion

$$\widehat{R}_{\text{pred}}(\gamma) = \widehat{R}_Y(\gamma) + \widehat{R}_D(\gamma) + \widehat{R}_Z(\gamma),$$

where each term is an out-of-sample loss on the selection fold using nuisance fits trained on the estimation fold.

The screen is not, by itself, a certificate of DML product-rate conditions. Accordingly, Γ_n is treated as a curated library of candidate pipelines. To make this explicit, call a candidate γ *rate-valid* on fold k if the nuisance estimate $\widehat{\eta}_k(\gamma)$ trained on I_k^{est} satisfies

$$\|\widehat{\eta}_k(\gamma) - \eta_0\|_{L_2(P)} = o_P(n^{-1/4}), \qquad \sup_{\theta\in\Theta} \|\widehat{g}_n(\theta;\gamma) - g_{n,0}(\theta)\| = o_P(1), \tag{4.6}$$

together with the product-rate conditions implied by orthogonality.

Algorithm 2 Identification-Aware Cross-Validation (IACV) within LF–NCF

1: **Inputs:** candidate library Γ_n, predictive risk $\widehat{R}_{\mathrm{pred}}$, strength proxy $\widehat{S}_{n,\kappa}$, and tolerance rule ε_n or a 1-SE rule.

2: Compute $\widehat{R}_{\mathrm{pred}}(\gamma)$ and $\widehat{S}_{n,\kappa}(\gamma)$ for all $\gamma \in \Gamma_n$ on I^{sel} using nuisance fits trained on I^{est}.

3: Let $\widehat{R}_{\min} = \min_{\gamma \in \Gamma_n} \widehat{R}_{\mathrm{pred}}(\gamma)$ and define the prediction-quality screen

$$\widehat{\Gamma}_{\varepsilon_n} := \Big\{\gamma \in \Gamma_n : \ \widehat{R}_{\mathrm{pred}}(\gamma) \le \widehat{R}_{\min} + \varepsilon_n\Big\},$$

or replace ε_n by one estimated standard error.

4: Choose

$$\widehat{\gamma} \in \arg\max_{\gamma \in \widehat{\Gamma}_{\varepsilon_n}} \widehat{S}_{n,\kappa}(\gamma).$$

5: **Output:** $\widehat{\gamma}$ and the corresponding nuisance fits trained on I^{est}.

4.4 Screening preserves nuisance validity

For each $\gamma \in \Gamma_n$, let $R_{\mathrm{pred}}(\gamma)$ denote the population counterpart of $\hat{R}_{\mathrm{pred}}(\gamma)$, that is, the expectation of the same out-of-sample loss evaluated on an independent draw from P.

Theorem 2 (Rate-screening validity of IACV)**.** *For each outer fold k, suppose $\Gamma^{\mathrm{good}}_{n,k} \neq \varnothing$ and*

$$\Gamma_n = \Gamma^{\mathrm{good}}_{n,k} \cup \Gamma^{\mathrm{bad}}_{n,k}.$$

Assume the following hold uniformly over $P \in \mathcal{P}_n$:

(i) every $\gamma \in \Gamma^{\mathrm{good}}_{n,k}$ is rate-valid in the sense of (4.6)*;*

(ii) every $\gamma \in \Gamma^{\mathrm{bad}}_{n,k}$ violates (4.6) *with probability bounded away from zero;*

(iii) there exists a deterministic sequence $c_n > 0$ such that

$$\inf_{\gamma \in \Gamma^{\mathrm{bad}}_{n,k}} R_{\mathrm{pred}}(\gamma) - \inf_{\gamma \in \Gamma^{\mathrm{good}}_{n,k}} R_{\mathrm{pred}}(\gamma) \ge c_n;$$

(iv)

$$\sup_{\gamma \in \Gamma_n} \left|\hat{R}_{\mathrm{pred}}(\gamma) - R_{\mathrm{pred}}(\gamma)\right| = o_P(c_n), \qquad \varepsilon_n = o(c_n).$$

Then

$$\Pr\Big(\hat{\Gamma}_{\varepsilon_n,k} \subseteq \Gamma^{\text{good}}_{n,k},\ \hat{\Gamma}_{\varepsilon_n,k} \neq \varnothing\Big) \to 1$$

uniformly over $P \in \mathcal{P}_n$. Consequently, the IACV-selected nuisance $\hat{\eta}_k(\hat{\gamma}_k)$ is rate-valid with probability $1 - o(1)$.

Proof. See Appendix A.5. □

Corollary 2 (Simple margin-screening sufficient condition)**.** *Suppose there exists a rate-valid candidate $\gamma^\dagger \in \Gamma_n$ such that $\widehat{\eta}_k(\gamma^\dagger)$ satisfies* (4.6) *on each outer fold. Suppose further that there exists a deterministic constant $\delta > 0$ such that, uniformly over $P \in \mathcal{P}_n$, with probability $1 - o(1)$,*

$$\widehat{R}_{\text{pred}}(\gamma) \geq \widehat{R}_{\text{pred}}(\gamma^\dagger) + \delta \qquad \textit{for every candidate } \gamma \textit{ that is not rate-valid.}$$

If the screen tolerance satisfies $\varepsilon_n = o_P(1)$, then

$$\Pr\Big(\widehat{\Gamma}_{\varepsilon_n,k} \subseteq \Gamma^{\text{good}}_{n,k}\Big) \to 1$$

uniformly over $P \in \mathcal{P}_n$. Consequently, the IACV-selected nuisance is rate-valid with probability $1 - o(1)$.

Remark 6 (Interpretation of the screening result)**.** Theorem 2 and Corollary 2 are sufficient-condition results for curated libraries. They do not claim that predictive risk generically separates rate-valid from rate-invalid pipelines in unrestricted adaptive libraries. Their role is narrower: if the candidate menu is curated so that prediction quality is informative enough to screen out rate-invalid pipelines, then adaptive selection can preserve the nuisance-rate conditions needed for weak-identification-robust orthogonal inference.

Proof of Corollary 2. Again fix an outer fold k and suppress the fold index. Let

$$\Gamma^{\text{good}}_n := \{\gamma \in \Gamma_n : \widehat{\eta}(\gamma) \text{ is rate-valid}\}, \qquad \Gamma^{\text{bad}}_n := \Gamma_n \setminus \Gamma^{\text{good}}_n.$$

By assumption, $\gamma^\dagger \in \Gamma^{\text{good}}_n$. Define the event

$$\mathcal{M}_n := \Big\{\widehat{R}_{\text{pred}}(\gamma) \geq \widehat{R}_{\text{pred}}(\gamma^\dagger) + \delta \text{ for every } \gamma \in \Gamma^{\text{bad}}_n\Big\}.$$

The corollary assumes

$$\sup_{P\in\mathcal{P}_n} \Pr_P(\mathcal{M}_n^c) \to 0.$$

Let

$$\widehat{R}_{\min} := \min_{\gamma\in\Gamma_n} \widehat{R}_{\text{pred}}(\gamma), \qquad \widehat{\Gamma}_{\varepsilon_n} := \{\gamma \in \Gamma_n : \widehat{R}_{\text{pred}}(\gamma) \le \widehat{R}_{\min} + \varepsilon_n\}.$$

Also define

$$\mathcal{B}_n := \{\varepsilon_n \le \delta/2\}.$$

If ε_n is deterministic and $\varepsilon_n = o(1)$, then $\mathcal{B}_n$ holds for all sufficiently large n. If $\varepsilon_n = o_P(1)$, then

$$\sup_{P\in\mathcal{P}_n} \Pr_P(\mathcal{B}_n^c) \to 0.$$

Thus in either case,

$$\sup_{P\in\mathcal{P}_n} \Pr_P(\mathcal{B}_n^c) \to 0. \tag{4.7}$$

Now work on the event $\mathcal{M}_n \cap \mathcal{B}_n$. For any $\gamma \in \Gamma_n^{\text{bad}}$,

$$\widehat{R}_{\text{pred}}(\gamma) \ge \widehat{R}_{\text{pred}}(\gamma^\dagger) + \delta.$$

Since $\widehat{R}_{\min} \le \widehat{R}_{\text{pred}}(\gamma^\dagger)$, it follows that

$$\widehat{R}_{\text{pred}}(\gamma) \ge \widehat{R}_{\min} + \delta > \widehat{R}_{\min} + \varepsilon_n,$$

where the strict inequality uses $\varepsilon_n \le \delta/2$. Hence

$$\gamma \notin \widehat{\Gamma}_{\varepsilon_n} \qquad \text{for every } \gamma \in \Gamma_n^{\text{bad}}.$$

Therefore

$$\widehat{\Gamma}_{\varepsilon_n} \subseteq \Gamma_n^{\text{good}} \qquad \text{on } \mathcal{M}_n \cap \mathcal{B}_n. \tag{4.8}$$

To prove nonemptiness, let

$$\widehat{\gamma}_n^{\min} \in \arg\min_{\gamma\in\Gamma_n} \widehat{R}_{\text{pred}}(\gamma).$$

If $\widehat{\gamma}_n^{\min} \in \Gamma_n^{\mathrm{bad}}$, then on $\mathcal{M}_n$,

$$\widehat{R}_{\min} = \widehat{R}_{\mathrm{pred}}(\widehat{\gamma}_n^{\min}) \geq \widehat{R}_{\mathrm{pred}}(\gamma^\dagger) + \delta > \widehat{R}_{\mathrm{pred}}(\gamma^\dagger),$$

a contradiction to the definition of $\widehat{R}_{\min}$. Thus

$$\widehat{\gamma}_n^{\min} \in \Gamma_n^{\mathrm{good}} \qquad \text{on } \mathcal{M}_n.$$

Since $\widehat{R}_{\mathrm{pred}}(\widehat{\gamma}_n^{\min}) = \widehat{R}_{\min}$,

$$\widehat{\gamma}_n^{\min} \in \widehat{\Gamma}_{\varepsilon_n}.$$

Hence

$$\widehat{\Gamma}_{\varepsilon_n} \neq \varnothing \qquad \text{on } \mathcal{M}_n. \tag{4.9}$$

Combining (4.8) and (4.9),

$$\mathcal{M}_n \cap \mathcal{B}_n \subseteq \Big\{ \widehat{\Gamma}_{\varepsilon_n} \subseteq \Gamma_n^{\mathrm{good}},\ \widehat{\Gamma}_{\varepsilon_n} \neq \varnothing \Big\}.$$

Therefore, by the union bound and (4.7),

$$\sup_{P \in \mathcal{P}_n} \Pr_P\Big(\widehat{\Gamma}_{\varepsilon_n} \subseteq \Gamma_n^{\mathrm{good}},\ \widehat{\Gamma}_{\varepsilon_n} \neq \varnothing \Big) \to 1.$$

Consequently, any selector taking values in $\widehat{\Gamma}_{\varepsilon_n}$ is rate-valid with probability $1 - o(1)$, uniformly over $P \in \mathcal{P}_n$. This proves the corollary. $\square$

4.5 Prediction-only tuning can kill attainable strength

Proposition 3 (Prediction-only tuning can select asymptotically negligible proxy strength)**.** *Consider the canonical homoskedastic Gaussian PLIV experiment of Theorem 1. Suppose $\Gamma_n = \{\gamma_1, \gamma_2\}$ contains two candidate nuisance pipelines whose candidate-specific AR statistics are correctly calibrated under $H_0 : \theta = \theta_0$. Under the fixed alternative $\theta = \theta_0 + \Delta$, let $\lambda_{j,n}$ denote the AR noncentrality index generated by candidate γ_j, and suppose*

$$\lambda_{1,n} \to \lambda_1 > 0, \qquad \lambda_{2,n} \to 0.$$

If

$$\widehat{R}_{\text{pred}}(\gamma_1) - \widehat{R}_{\text{pred}}(\gamma_2) = o_{\mathbb{P}}(1), \qquad \liminf_{n\to\infty} \mathbb{P}\Big(\widehat{R}_{\text{pred}}(\gamma_2) \le \widehat{R}_{\text{pred}}(\gamma_1)\Big) > 0,$$

then prediction-only selection chooses γ_2 with nonvanishing probability. Under any fixed alternative $\theta = \theta_0 + \Delta$, this places nonvanishing probability on a pipeline whose candidate-specific AR power converges to size. Along any subsequence for which

$$\mathbb{P}(\widehat{\gamma}_n^{\text{pred}} = \gamma_2) \to 1,$$

the rejection probability of the prediction-only AR statistic converges to the nominal size α.

This proposition is a power/informativeness example, not a selected-nuisance validity theorem. The candidates need not both satisfy the global rate-validity condition (4.6); indeed, if both converged to the same oracle nuisance at the rate in (4.6), their residualized strength would be asymptotically equivalent. If, along a subsequence,

$$\mathbb{P}(\widehat{\gamma}_n^{\text{pred}} = \gamma_2) \to p \in (0, 1]$$

and the selected-statistic rejection probability admits the corresponding mixture limit, then the limiting rejection probability is

$$(1 - p) \Pr(\chi^2_{d_z}(\lambda_1) > c_{1-\alpha}) + p\alpha,$$

which is strictly below the fixed-γ_1 power whenever $p > 0$.

Proof. See Appendix A.6. □

4.6 Near-oracle strength and no first-order downside

The next proposition states that, within the screened set, IACV is asymptotically near-oracle for the strength target.

Proposition 4 (Near-oracle strength within the screen)**.** *Let*

$$\hat{\gamma}_k \in \arg\max_{\gamma \in \widehat{\Gamma}_{\varepsilon_n,k}} \widehat{S}_{n,\kappa}(\gamma),$$

and suppose $\widehat{\Gamma}_{\varepsilon_n,k} \neq \varnothing$ *with probability* $1 - o(1)$*. Define*

$$\overline{S}_{n,\kappa}(\gamma) := n_{\mathrm{sel}}^{-1}\widehat{S}_{n,\kappa}(\gamma), \qquad \overline{S}_{n,\kappa}(\eta(\gamma)) := n_{\mathrm{sel}}^{-1}S_{n,\kappa}(\eta(\gamma)),$$

and let

$$\Delta_n := \sup_{\gamma\in\widehat{\Gamma}_{\varepsilon_n,k}} \left|\overline{S}_{n,\kappa}(\gamma) - \overline{S}_{n,\kappa}(\eta(\gamma))\right|.$$

Then, on the event $\widehat{\Gamma}_{\varepsilon_n,k} \neq \varnothing$*,*

$$\overline{S}_{n,\kappa}(\eta(\hat{\gamma}_k)) \geq \sup_{\gamma\in\widehat{\Gamma}_{\varepsilon_n,k}} \overline{S}_{n,\kappa}(\eta(\gamma)) - 2\Delta_n.$$

Equivalently,

$$S_{n,\kappa}(\eta(\hat{\gamma}_k)) \geq \sup_{\gamma\in\widehat{\Gamma}_{\varepsilon_n,k}} S_{n,\kappa}(\eta(\gamma)) - 2n_{\mathrm{sel}}\Delta_n.$$

Proof. See Appendix A.7. □

The proposition is a regret statement for the screened strength proxy. In exact weak-identification regimes, the feasible proxy can be noisy on the same scale as the local first-stage signal. The result should therefore not be read as selection consistency for a population power-maximizing pipeline; its role is to show that, conditional on the screened menu and observed proxy, IACV does not leave estimated residualized strength on the table.

Proposition 5 (No first-order downside under regular identification)**.** *Let* $\widehat{\gamma}_k^{\mathrm{pred}}$ *denote the prediction-only selector and* $\widehat{\gamma}_k^{IACV}$ *the IACV selector computed from the same outer fold. Suppose the prediction-quality screen uses a sequence* $\varepsilon_n \downarrow 0$ *such that, with probability* $1 - o(1)$*, either*

$$\widehat{\Gamma}_{\varepsilon_n,k} = \{\widehat{\gamma}_k^{\mathrm{pred}}\},$$

or all candidates in $\widehat{\Gamma}_{\varepsilon_n,k}$ *are rate-valid in the sense of* (4.6) *and generate nuisance fits that are* $o_{\mathbb{P}}(n^{-1/4})$*-equivalent. Then*

$$\sup_{\theta\in\Theta} \left\|\widehat{g}_n^{IACV}(\theta) - \widehat{g}_n^{\mathrm{pred}}(\theta)\right\| = o_{\mathbb{P}}(1).$$

In the singleton-screen case,

$$\mathbb{P}\Big(\widehat{\gamma}_k^{IACV} = \widehat{\gamma}_k^{\mathrm{pred}}\Big) \to 1.$$

If, in addition, the strong-identification regularity conditions of Theorem 5 hold for the selected moment problem, then the corresponding GMM estimators are asymptotically equivalent.

In the canonical PLIV benchmark, a sufficient primitive for regular identification is that the normalized population strength

$$\overline{S}_{n,\kappa}(\eta(\gamma)) = \Pi(\eta(\gamma))^\top \Big(\Sigma_Z(\eta(\gamma)) + \kappa I\Big)^{-1} \Pi(\eta(\gamma))$$

is bounded away from zero over the selected screen with probability approaching one.

Proof. See Appendix A.8. □

Remark 7 (Practical defaults)**.** A stable default is a one-standard-error rule for the prediction-quality screen. For κ, a conservative default is

$$\kappa = \kappa_0 \cdot \frac{\operatorname{tr}(\widehat{\Sigma}_Z)}{d_z}, \qquad \kappa_0 \in [10^{-6}, 10^{-2}],$$

with sensitivity analysis over a short grid. If d_z is small and $\widehat{\Sigma}_Z$ is well-conditioned, one may set $\kappa = 0$.

5 Main Results

5.1 Assumptions

Assumption 1 (Sampling and triangular arrays)**.** $\{W_i\}_{i=1}^n$ are i.i.d. from $P \in \mathcal{P}_n$, where $\mathcal{P}_n$ may include weak-identification triangular arrays.

Assumption 2 (Moments and covariance regularity)**.** There exists $q > 4$ such that

$$\sup_{P\in\mathcal{P}_n} \sup_{\theta\in\Theta} \mathbb{E}_P[\|\psi(W;\theta,\eta_0)\|^q] < \infty.$$

For θ in a neighborhood of θ_0, the covariance matrix $\Omega_P(\theta) = \mathbb{E}_P[\psi\psi^\top]$ has eigenvalues bounded away from 0 and ∞ uniformly over $P \in \mathcal{P}_n$.

Assumption 3 (Orthogonality)**.** The score $\psi(W;\theta,\eta)$ is Neyman-orthogonal at (θ_0, η_0) as in (2.8).

Assumption 4 (Leakage-free adaptivity)**.** LF–NCF is used. For each fold k, $(\widehat{\gamma}_k, \widehat{\eta}_k)$ is measurable with respect to $\mathcal{F}_k^{\text{train}}$ and independent of $\mathcal{F}_k^{\text{inf}}$.

Assumption 5 (Selected nuisance rates and product conditions)**.** Uniformly over $P \in \mathcal{P}_n$ and folds k,

$$\|\widehat{\eta}_k - \eta_0\|_{L_2(P)} = o_P(n^{-1/4}),$$

and the product-rate conditions implied by orthogonality yield

$$\sup_{\theta \in \Theta} \|\widehat{g}_n(\theta) - g_{n,0}(\theta)\| = o_P(1),$$

where

$$g_{n,0}(\theta) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(W_i; \theta, \eta_0).$$

In addition,

$$\frac{1}{n} \sum_{i=1}^{n} \Big\| \psi(W_i; \theta_0, \widehat{\eta}_{-i}) - \psi(W_i; \theta_0, \eta_0) \Big\|^2 = o_P(1).$$

Assumption 6 (Library size / complexity)**.** The candidate library Γ_n is finite with $\log |\Gamma_n| = o(n_{\text{sel}})$, and the concentration bounds used to control $\widehat{R}_{\text{pred}}$ and $\widehat{S}_{n,\kappa}$ hold uniformly over $\Gamma_n \times \mathcal{P}_n$.

Assumption 7 (Weak-identification uniformity class)**.** $\mathcal{P}_n$ includes both strong-ID and weak-ID sequences, including Staiger–Stock local-to-zero sequences in linear IV. The baseline AR theorem below is stated for fixed d_g.

Assumption 8 (Effective-rank ridge regime)**.** Uniformly over $P \in \mathcal{P}_n$:

(i) there exists $q > 8$ such that

$$E_P\Big[\|\psi(W; \theta_0, \eta_0)\|^q\Big] \le C_q;$$

(ii) the ridge level $\rho_n \ge 0$ satisfies

$$\lambda_{\min}\Big(\Omega_P(\theta_0) + \rho_n I_{d_{g,n}}\Big) \ge c > 0;$$

(iii) the effective rank

$$r_{\text{eff}}(\rho_n) := \operatorname{tr}\Big(\Omega_P(\theta_0)\Big(\Omega_P(\theta_0) + \rho_n I_{d_{g,n}}\Big)^{-1}\Big)$$

satisfies $r_{\text{eff}}(\rho_n)^4/n \to 0$;

(iv) with

$$\Omega_{P,\rho_n} := \Omega_P(\theta_0) + \rho_n I_{d_{g,n}},$$

one has

$$\|\Omega_{P,\rho_n}^{-1/2}\Big(\widehat{g}_n(\theta_0) - g_{n,0}(\theta_0)\Big)\| = o_P\Big(r_{\text{eff}}(\rho_n)^{-1/2}\Big),$$
$$\Big\|\Omega_{P,\rho_n}^{-1/2}\Big(\widehat{\Omega}_n(\theta_0) - \Omega_P(\theta_0)\Big)\Omega_{P,\rho_n}^{-1/2}\Big\|_{\text{op}} = o_P\Big(r_{\text{eff}}(\rho_n)^{-1}\Big),$$

and, for the Gaussian multiplier bootstrap analogue,

$$\|\Omega_{P,\rho_n}^{-1/2}\Big(\widehat{g}_n^*(\theta_0) - g_{n,0}^*(\theta_0)\Big)\| = o_{P^*}\Big(r_{\text{eff}}(\rho_n)^{-1/2}\Big) \qquad \text{in } P\text{-probability};$$

(v) the multipliers are i.i.d. standard normal and independent of the data;

(vi) letting

$$\Omega_{n,0} := \frac{1}{n}\sum_{i=1}^{n} \psi(W_i;\theta_0,\eta_0)\psi(W_i;\theta_0,\eta_0)^\top, \qquad B_{0n} := \Omega_{P,\rho_n}^{-1/2}\Omega_{n,0}\Omega_{P,\rho_n}^{-1/2},$$

and $\mu_{1,n} := \lambda_{\max}(B_{0n})$, there exists $\underline{\lambda} > 0$ such that

$$\inf_{P\in\mathcal{P}_n} P_P(\mu_{1,n} \geq \underline{\lambda}) \to 1;$$

(vii) letting

$$\widehat{A}_n := \widehat{\Omega}_n(\theta_0) + \rho_n I_{d_{g,n}}, \qquad \widehat{B}_n := \widehat{A}_n^{-1/2}\widehat{\Omega}_n(\theta_0)\widehat{A}_n^{-1/2},$$

and $\widehat{\mu}_{1,n} := \lambda_{\max}(\widehat{B}_n)$, there exists $\underline{\lambda}^* > 0$ such that

$$\inf_{P\in\mathcal{P}_n} P_P(\widehat{\mu}_{1,n} \geq \underline{\lambda}^*) \to 1.$$

5.2 Uniform size of orthogonal AR after adaptive learning

The next theorem is a uniform inference result conditional on the selected nuisance satisfying the oracle-equivalence and product-rate requirements in Assumption 5. The screening results in Section 4 provide sufficient conditions under which adaptive selection over a curated library delivers those requirements; the AR theorem itself is stated at the level of the selected score.

Theorem 3 (Uniform size of orthogonal AR with LF–NCF)**.** *Under Assumptions 1–7,*

$$\sup_{P\in\mathcal{P}_n}\left|\mathbb{P}_P\Big(AR_n(\theta_0)>\chi^2_{d_g,1-\alpha}\Big)-\alpha\right|\to 0.$$

Equivalently, the inverted set

$$\mathcal{C}^{AR}_{1-\alpha}:=\{\theta\in\Theta: AR_n(\theta)\le\chi^2_{d_g,1-\alpha}\}$$

has asymptotic coverage $1-\alpha$ *uniformly over* $P\in\mathcal{P}_n$*.*

Proof. See Appendix A.9. □

The ridge Anderson–Rubin and conditional QLR results below are extensions of the main finite-library selection theory. They show that the LF–NCF filtration can be combined with other weak-identification-robust procedures once the adaptive score has been constructed. The core selection results remain the leakage-free filtration, prediction-quality screening, and identification-aware tuning rules developed above.

5.3 Dimension-agnostic ridge AR with Gaussian conditional bootstrap

For $\rho\ge 0$, define the ridge-regularized covariance estimator

$$\widehat{\Omega}_{n,\rho}(\theta_0):=\widehat{\Omega}_n(\theta_0)+\rho I_{d_{g,n}}.$$

For the sequence ρ_n, define

$$AR_{n,\rho_n}(\theta_0):=\widehat{g}_n(\theta_0)^\top\widehat{\Omega}_{n,\rho_n}(\theta_0)^{-1}\widehat{g}_n(\theta_0).$$

For Gaussian conditional bootstrap calibration in Theorem 4, let $\{\xi_i\}_{i=1}^n$ be i.i.d. standard normal multipliers independent of the data and define

$$\widehat{g}^*_n(\theta_0):=\frac{1}{\sqrt{n}}\sum_{i=1}^n\xi_i\widehat{\psi}_i(\theta_0),$$

and

$$AR^*_{n,\rho_n}(\theta_0):=\widehat{g}^*_n(\theta_0)^\top\widehat{\Omega}_{n,\rho_n}(\theta_0)^{-1}\widehat{g}^*_n(\theta_0).$$

The oracle distribution-function approximation in Online Appendix Section OA.1 continues to hold for more general mean-zero, variance-one multipliers with finite qth moments; compare Pouzo (2015) for related high-dimensional quadratic-form bootstrap arguments. The random-critical-value size-control theorem below is stated for Gaussian multipliers.

Theorem 4 (Dimension-agnostic uniform size of ridge AR with Gaussian conditional bootstrap)**.** *Assume LF–NCF, Assumptions 1–6, and Assumption 8. Let $c^*_{1-\alpha}$ denote the conditional $(1-\alpha)$ quantile of $AR^*_{n,\rho_n}(\theta_0)$ given the realized scores and the LF–NCF training outcomes. Then*

$$\limsup_{n\to\infty} \sup_{P\in\mathcal{P}_n} \mathbb{P}_P\Big(AR_{n,\rho_n}(\theta_0) > c^*_{1-\alpha}\Big) \le \alpha.$$

The result is dimension-agnostic: it does not require the researcher to impose a fixed-d_g, many-d_g, or $d_g > n$ asymptotic regime ex ante.

Proof. See Appendix A.10. □

Remark 8 (Practical choice of the ridge level)**.** A leakage-free data-dependent choice of ρ_n can be made on non-inference data only. One simple rule is to choose the smallest value in a prespecified grid $\mathcal{R}_n \subset (0,\infty)$ such that the estimated effective rank satisfies

$$\widehat{r}_{\mathrm{eff}}(\rho)^4/n_{\mathrm{inf}} \le \tau_n, \qquad \tau_n \downarrow 0,$$

where

$$\widehat{r}_{\mathrm{eff}}(\rho) := \mathrm{tr}\Big(\widehat{\Omega}_n(\theta_0)\Big(\widehat{\Omega}_n(\theta_0) + \rho I\Big)^{-1}\Big).$$

Because this rule uses only training or selection observations, it preserves the measurability structure required by LF–NCF.

5.4 Strong-identification efficiency

When the Jacobian is bounded away from singularity, the weak-identification concerns motivating robust test inversion become asymptotically irrelevant for point estimation. Let

$$\widehat{m}_n(\theta) := \frac{1}{n}\sum_{i=1}^n \widehat{\psi}_i(\theta), \qquad \widehat{W}_n := \widehat{\Omega}_n(\widetilde{\theta})^{-1},$$

where $\widetilde{\theta}$ is a preliminary consistent estimator. Define the cross-fitted GMM estimator

$$\widehat{\theta} \in \arg\min_{\theta \in \Theta} \ \widehat{m}_n(\theta)^\top \widehat{W}_n \widehat{m}_n(\theta).$$

Theorem 5 (Efficiency under strong identification)**.** *Suppose $\sigma_{\min}(G_P) \geq c > 0$ uniformly over $P \in \mathcal{P}_n$, the score is continuously differentiable in θ on a neighborhood of θ_0, a uniform law of large numbers holds for its Jacobian (van der Vaart and Wellner, 1996), and the preliminary estimator $\widetilde{\theta}$ is consistent. Suppose further that there exists a neighborhood $\mathcal{N}_0$ of θ_0 such that*

$$\inf_{\theta \in \Theta \setminus \mathcal{N}_0} Q_P(\theta) > Q_P(\theta_0) = 0,$$

and that

$$\sup_{\theta \in \Theta \setminus \mathcal{N}_0} |Q_n(\theta) - Q_P(\theta)| = o_{\mathbb{P}}(1).$$

Then the global LF–NCF GMM estimator $\widehat{\theta}$ based on the orthogonal score is consistent and root-n asymptotically normal. If the score is chosen to coincide with the semiparametrically efficient influence function, then $\widehat{\theta}$ attains the semiparametric efficiency bound.

Proof. See Appendix A.11. □

Remark 9. A primitive sufficient condition for the displayed global-separation assumption is that $\Theta \setminus \mathcal{N}_0$ is compact, Q_P is continuous on Θ, and $Q_P(\theta) > 0$ for every $\theta \neq \theta_0$. The theorem is stated directly in terms of the separation condition because compactness is not otherwise required.

5.5 Conditional QLR / CLR under leakage-free adaptivity

For each $\theta \in \Theta$, define the cross-fitted score contribution

$$\widehat{\psi}_i(\theta) := \psi(W_i; \theta, \widehat{\eta}_{-i}),$$

the cross-fitted moment process

$$\widehat{g}_n(\theta) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \widehat{\psi}_i(\theta),$$

and the covariance kernel estimator

$$\widehat{\Sigma}_n(\theta, \vartheta) := \frac{1}{n}\sum_{i=1}^{n} \widehat{\psi}_i(\theta)\widehat{\psi}_i(\vartheta)^\top, \qquad \theta, \vartheta \in \Theta.$$

Define the conditional quasi-likelihood ratio statistic

$$QLR_n(\theta_0) := \widehat{g}_n(\theta_0)^\top \widehat{\Sigma}_n(\theta_0, \theta_0)^{-1}\widehat{g}_n(\theta_0) - \inf_{\theta\in\Theta} \widehat{g}_n(\theta)^\top \widehat{\Sigma}_n(\theta, \theta)^{-1}\widehat{g}_n(\theta).$$

Following Andrews and Mikusheva (2016), define the sufficient-statistic transform

$$\widehat{h}_n(\theta) := \widehat{g}_n(\theta) - \widehat{\Sigma}_n(\theta, \theta_0)\widehat{\Sigma}_n(\theta_0, \theta_0)^{-1}\widehat{g}_n(\theta_0), \qquad \theta \in \Theta,$$

and let $\widehat{c}^{\,\mathrm{CQLR}}_{1-\alpha}(\theta_0)$ denote the exact conditional critical value, or the simulation critical value of Algorithm 1 with $B \to \infty$, constructed in Online Appendix OA.2; the full algorithm and the linear-IV CLR specialization are recorded in Online Appendix OA.2.

Theorem 6 (Uniform size of cross-fitted conditional QLR under LF–NCF)**.** *Assume LF–NCF and suppose the cross-fitted moment process $\{\widehat{g}_n(\theta) : \theta \in \Theta\}$ and covariance kernel $\widehat{\Sigma}_n(\theta, \vartheta)$ satisfy Assumption OA.1 in Online Appendix OA.2. For a fixed null value θ_0, let*

$$\mathcal{P}_n(\theta_0) := \{P \in \mathcal{P}_n : \mathbb{E}_P[\psi(W; \theta_0, \eta_0)] = 0\}.$$

Then, for any fixed $\varepsilon > 0$,

$$\limsup_{n\to\infty} \sup_{P\in\mathcal{P}_n(\theta_0)} \Pr_P\Big(QLR_n(\theta_0) > \widehat{c}^{\,\mathrm{CQLR}}_{1-\alpha}(\theta_0) + \varepsilon\Big) \le \alpha.$$

If the Gaussian benchmark difference

$$T^{\mathrm{G}}_{n,P} := QLR^{\mathrm{G}}_{n,P}(\theta_0) - c^{\mathrm{CQLR}}_{1-\alpha,n,P}(\theta_0),$$

defined in Online Appendix OA.2, has a distribution that is continuous at zero uniformly over $P \in \mathcal{P}_n(\theta_0)$, the slack ε may be omitted.

Proof. See Online Appendix OA.2. □

Remark 10. The result shows that the LF–NCF filtration is compatible with the Andrews–Mikusheva conditional-inference architecture, so the adaptive-pipeline theory extends beyond Anderson–Rubin. The online appendix records the full conditional construction and proves that, in the canonical homoskedastic Gaussian linear-IV model, the resulting conditional QLR procedure reduces to the classical CLR procedure.

6 Monte Carlo Evidence

The theorem package isolates two failures that are specific to adaptive learning under weak identification: leaky tuning can change the null law, and prediction-only tuning can attenuate attainable identification strength. The simulation evidence is therefore targeted rather than omnibus. Each design is chosen to isolate one mechanism as cleanly as possible.

The first design is a stylized Gaussian experiment that directly implements Propositions 1 and 2. Two candidate score processes are individually pivotal under the null, but the leaky rule reports the smaller of the two inference-sample quadratic forms. This design isolates null-law distortion without nuisance-estimation noise. The second design is a homoskedastic partially linear IV design calibrated to the Staiger–Stock local-to-zero regime (Staiger and Stock, 1997): $p = 150$, $X \sim N(0, \Sigma_X)$ with $(\Sigma_X)_{jk} = 0.5^{|j-k|}$, $Z \sim N(0, 1)$ independent of X, (U, V) jointly Gaussian with mean zero, unit marginal variances, and correlation 0.6, $m_0(X) = \sum_{j=1}^{5} X_j, \qquad \gamma_0(X) = \sum_{j=1}^{5} X_j$, $D = m_0(X) + (\pi/\sqrt{n})Z + V, \qquad Y = \theta_0 D + \gamma_0(X) + U$, with $n \in \{500, 2000\}$ and $\pi \in \{0, 1, 3, 10, 30\}$. The third design is a proxy-rich modification of the same PLIV experiment in which instruments are partially predictable from covariates, so that flexible nuisance learners have scope to absorb residualized first-stage signal and thereby generate strength killing. Supplementary null-size and heteroskedastic diagnostics are reported in the online appendix.

Three procedures are compared: a leaky global-tuning benchmark, LF–NCF with prediction-only tuning, and LF–NCF with IACV. The reported objects are null rejection frequencies for AR at θ_0, rejection probabilities under fixed alternatives $\theta = \theta_0 + \Delta$, the average selected strength proxy $\widehat{S}_{n,\kappa}(\widehat{\gamma})$, and the shape of inverted AR confidence sets on a finite parameter grid, summarized by grid-boundary frequency and average bounded length.

Table 1 reports the stylized Gaussian experiment. Each fixed-candidate AR statistic is correctly calibrated under the null, but the leaky rule that reports the smaller inference-sample quadratic form is not. When the correlation between the two candidate statistics is low, the null rejection probability of the leaky procedure falls far below the nominal 5 percent level; when the correlation rises, the distortion remains first-order. By contrast, LF–NCF, which selects on an independent sample and evaluates on the inference sample, remains correctly calibrated throughout. The table is therefore the finite-sample analogue of Proposition 2.

Table 1: Toy leakage experiment: null-law distortion from inference-sample selection

ρ	Leaky size	(s.e.)	LF–NCF size	(s.e.)
0.0	0.002	0.000	0.049	0.001
0.3	0.005	0.000	0.051	0.001
0.6	0.013	0.001	0.050	0.001
0.9	0.030	0.001	0.049	0.001

Notes. Nominal 5 percent test with $d_g = 1$ and 50,000 Monte Carlo replications. The leaky procedure reports the smaller of two inference-sample quadratic forms. LF–NCF selects using an independent sample and evaluates on the inference sample.

The proxy-rich stress design makes the strength-preservation channel visible. Table 2 reports power and average selected strength at a representative weak-identification cell, $(n, \pi) = (500, 3)$. Relative to prediction-only tuning, LF–NCF+IACV raises the average strength proxy from 15.532 to 17.009 and increases rejection probability from 0.095 to 0.124. Confidence-set geometry is reported separately in Table 3. The finite-sample gains are moderate rather than dramatic, which is the natural pattern when identification is weak but not absent, yet the direction of change matches the theoretical strength-preservation results exactly.

Table 2: Proxy-rich stress design: power and selected strength at a representative weak-IV cell

Method	Power	(s.e.)	Avg. $\widehat{S}_{n,\kappa}$
LF–NCF + IACV	0.124	0.007	17.009
LF–NCF (pred-only)	0.095	0.007	15.532

Notes. The representative cell uses $(n, \pi) = (500, 3)$ in the proxy-rich design. Both rows are leakage-free; differences arise from the tuning objective rather than from sample reuse. Confidence-set geometry is reported separately in Table 3.

Table 3 reports confidence-set diagnostics. Because boundary-touch frequencies are very high in this weak-identification design, average bounded length is computed on a small selected subset of replications and should be read as a secondary diagnostic.

The more stable feature is the bounded-set frequency itself: at $n = 500$, the boundary-touch frequency falls from 0.999 under prediction-only tuning to 0.995 under IACV, so bounded sets occur about five times as often under IACV. At $n = 2000$, boundary-touch frequencies are identical at three decimal places, while bounded lengths remain slightly shorter under IACV.

Table 3: Proxy-rich stress design: confidence-set diagnostics

Method	n	Grid-boundary freq.	(s.e.)	Avg. bounded length
LF–NCF + IACV	500	0.995	0.002	1.789
LF–NCF (pred-only)	500	0.999	0.001	1.836
LF–NCF + IACV	2000	0.993	0.002	1.786
LF–NCF (pred-only)	2000	0.993	0.002	1.802

Notes. Confidence sets are obtained by grid inversion of the orthogonal AR statistic. A set is classified as reaching the grid boundary when the accepted region touches the boundary of the numerical search interval.

The direction of the finite-sample distortion depends on the selector map. The stylized Gaussian experiment uses a selector that reports the smaller of two inference-sample quadratic forms, so underrejection is the natural manifestation of the null-law failure. By contrast, more realistic leaky predictive-tuning rules can distort in the opposite direction once the tuning map interacts with nuisance-estimation error and covariance estimation. The theory only requires that the null law can change at first order; the simulations show that the sign of the distortion is a property of the leaky rule, not a universal feature of weak-identification leakage.

The homoskedastic baseline design shows little separation between LF–NCF prediction-only tuning and LF–NCF+IACV, which is the finite-sample counterpart of Proposition 5. The online appendix reports null rejection rates for the full baseline PLIV design: at $n = 500$, the leaky benchmark rejects between 0.156 and 0.179 across the π grid, whereas the two leakage-free procedures lie between 0.075 and 0.077. These baseline null-size results should be read as finite-sample diagnostics rather than as sharp finite-sample guarantees. At $n = 500$, the leakage-free procedures remain mildly oversized relative to the nominal five percent level, reflecting the combined finite-sample costs of three-way splitting, nuisance estimation, and covariance estimation. The asymptotic theory identifies the filtration and orthogonal-score conditions under which size is restored; the simulations show that these protections can still involve non-negligible finite-sample costs in moderate samples. The online appendix also reports a heteroskedastic baseline design in which the two leakage-free procedures are again nearly indistinguishable at a representative weak-IV cell, with power 0.102 under IACV and 0.101 under prediction-only tuning. The finite-sample message is therefore the same as the theoretical one: when the candidate library leaves little room for systematic strength killing, IACV is approximately neutral; when such room exists, IACV improves weak-ID informativeness.

The simulation evidence supports two central claims in the paper. The stylized Gaussian design shows that leakage can alter the null law of weak-identification-robust statistics at first order. The proxy-rich PLIV design shows that prediction-only tuning can attenuate attainable strength and that IACV mitigates that attenuation in the directions predicted by the theory. The supplementary baseline and heteroskedastic designs show that these gains are not mechanical: they emerge when the library allows strength killing and disappear when it does not.

7 Empirical Illustration: Quarter-of-Birth Benchmark

The AK application is an illustration of weak-identification-robust reporting in a canonical design. The benchmark block reports the fixed-control AK91 specifications using the official data archive and the cohort-specific designs underlying Tables IV, V, and VI of Angrist and Krueger (1991). It shows how weak-identification-robust reporting shifts the inferential message even in a canonical and heavily studied application. The enriched-control exercise in the online appendix shows how the adaptive tuning architecture behaves in the same setting when the nuisance dictionary is enlarged.

Table 4 reports conventional 2SLS coefficients together with Anderson–Rubin and conditional likelihood ratio p-values for the cohort-specific instrumental-variables specifications corresponding to columns (2), (4), (6), and (8) of the AK91 tables. In the canonical quarter-of-birth design, economically modest specification changes already move the weak-identification-robust inferential message. For the 1930–39 cohort, the basic year-of-birth specification yields strong Wald, AR, and CLR rejection, whereas adding age and age squared drives the AR and CLR p-values above conventional significance levels. The 1920–29 cohort displays the same qualitative fragility.

Table 4: Angrist–Krueger quarter-of-birth benchmark: 2SLS coefficients and weak-identification-robust p-values

Cohort	Spec.	2SLS	Wald p	AR p	CLR p	J p
IV	(2)	0.0769 (0.0150)	3.2×10^{-7}	0.0085	5.2×10^{-4}	0.17
IV	(4)	0.1352 (0.0337)	6.0×10^{-5}	0.095	0.11	0.80
IV	(6)	0.0669 (0.0151)	9.4×10^{-6}	0.028	0.0020	0.23
IV	(8)	0.1039 (0.0341)	0.0023	0.20	0.27	0.69
V	(2)	0.0891 (0.0161)	3.2×10^{-8}	0.013	1.2×10^{-5}	0.66
V	(4)	0.0655 (0.0280)	0.019	0.64	0.38	0.71
V	(6)	0.0806 (0.0164)	8.8×10^{-7}	0.064	7.4×10^{-5}	0.80
V	(8)	0.0509 (0.0279)	0.069	0.85	0.51	0.87
VI	(2)	0.0553 (0.0138)	5.8×10^{-5}	5.6×10^{-11}	0.0089	5.4×10^{-10}
VI	(4)	0.1293 (0.0191)	1.4×10^{-11}	1.6×10^{-9}	1.9×10^{-10}	0.0032
VI	(6)	0.0393 (0.0145)	0.0067	1.0×10^{-8}	0.19	1.1×10^{-8}
VI	(8)	0.1138 (0.0200)	1.3×10^{-8}	2.5×10^{-7}	1.3×10^{-7}	0.0025

Notes. Cohorts IV, V, and VI correspond to men born 1920–29, 1930–39, and 1940–49, respectively. Specifications (2), (4), (6), and (8) mirror the cohort-specific instrumental-variables columns in the public AK91 replication files: year-of-birth controls only; year-of-birth controls plus age and age squared; year-of-birth plus region and demographic controls; and the full specification with both region/demographic and age controls. Standard errors are shown in parentheses. The reported Wald, AR, CLR, and J-test p-values follow the modern weak-IV benchmark computed from the official MIT archive. For columns (4) and (8), the public modern replication used here includes age terms in both stages, so the benchmark is not an exact transcription of the printed AK91 columns (Angrist and Krueger, 1991; Angrist, n.d.; ivmodels Developers, n.d.).

The 1940–49 cohort often continues to reject, but the overidentification J-test rejects in several specifications, suggesting specification instability rather than cleanly informative identification. Specification VI(6) is especially diagnostic: the AR p-value is 1.0×10^{-8}, while the CLR p-value is 0.19, and the J-test also rejects at 1.1×10^{-8}. That stark AR/CLR divergence is consistent with the same misspecification flagged by the overidentification test. The empirical lesson is the same as in the theory: robustness is not a property of the coefficient estimate alone, but of the full inferential pipeline used to construct and evaluate the score. Online Appendix Table 3 reports an enriched-control implementation of the cohort-V quarter-of-birth design. The exercise keeps the excluded instruments fixed and enlarges only the nuisance side, so it directly evaluates how the paper's tuning architecture moves the selected strength proxy and the associated weak-identification-robust inferential objects in the directions predicted by the theory.

8 Conclusion

Weak identification turns tuning and selection into first-order inferential objects. The paper isolates two pipeline failure modes—leakage and strength killing—and develops two primitives to address them. LF–NCF enforces the foldwise no-own-observation measurability condition that rules out leaky adaptivity and supports the oracle-equivalence argument used for uniform weak-identification validity. IACV treats identification strength as a tuning objective by targeting a feasible proxy for residualized first-stage strength, with an exact weak-ID Anderson–Rubin power interpretation in the canonical homoskedastic Staiger–Stock PLIV benchmark. The Monte Carlo evidence confirms that both mechanisms are visible in finite samples: leaky tuning distorts null rejection frequencies, and strength-preserving tuning improves power and confidence-set informativeness precisely in the designs where nuisance learners can absorb identifying variation.

The contribution is a theory of adaptive learning under weak identification for selection over a finite curated library of nuisance pipelines. The results show that leaky tuning can change the null law, screening can preserve nuisance validity, prediction-only tuning can destroy attainable strength, and identification-aware tuning can improve weak-ID informativeness without a first-order cost under strong identification. The same filtration also supports a dimension-agnostic ridge Anderson–Rubin extension and conditional QLR / CLR inference. Taken together, these results imply that under weak identification, robustness is a property of the adaptive pipeline that produces the score, not merely of the statistic finally reported.

A Proofs of the Main Results

This appendix contains the proofs of the lemmas, propositions, corollaries, and theorems stated in the main text. The online appendix records the oracle ridge-bootstrap support lemmas, the full conditional QLR / CLR construction, a worked learner-class verification for sparse linear nuisance estimators, supplementary technical notes, and additional Monte Carlo diagnostics.

A.1 Proof of Lemma 1

By the definition of $m_{0,n}$,

$$m_{0,n}(X) - m_0^\star(X) = \mathbb{E}_{P_n}[Z \mid X]^\top \Pi_n.$$

Therefore, by Cauchy–Schwarz,

$$\|m_{0,n} - m_0^\star\|_{L_2(P_n)} \le \|\mathbb{E}_{P_n}[Z \mid X]\|_{L_2(P_n)} \|\Pi_n\|.$$

The assumed uniform second-moment bound gives

$$\|\mathbb{E}_{P_n}[Z \mid X]\|_{L_2(P_n)} = O(1),$$

and $\|\Pi_n\| = O(n^{-1/2})$, so

$$\|m_{0,n} - m_0^\star\|_{L_2(P_n)} = O(n^{-1/2}).$$

The two displayed inequalities follow from the triangle inequality:

$$\|\widehat{m} - m_{0,n}\|_{L_2(P_n)} \le \|\widehat{m} - m_0^\star\|_{L_2(P_n)} + \|m_{0,n} - m_0^\star\|_{L_2(P_n)}$$

and

$$\|\widehat{m} - m_0^\star\|_{L_2(P_n)} \le \|\widehat{m} - m_{0,n}\|_{L_2(P_n)} + \|m_{0,n} - m_0^\star\|_{L_2(P_n)}.$$

Substituting the $O(n^{-1/2})$ drift bound gives the two displayed inequalities. If either of the two distances is $O_{\mathbb{P}}(r_n)$, the corresponding opposite distance is $O_{\mathbb{P}}(r_n + n^{-1/2})$. If $r_n = o(n^{-1/4})$, then $r_n + n^{-1/2} = o(n^{-1/4})$, which proves the final claim.

A.2 Proof of Proposition 1

For $j \in \{1, 2\}$, define the candidate-specific quadratic forms

$$Q_n^{(j)} := g_n^{(j)\top} \Omega^{-1} g_n^{(j)} = \left\| \Omega^{-1/2} g_n^{(j)} \right\|^2.$$

Let $\widehat{J}_n \in \{1,2\}$ denote the index selected by the leaky rule, that is,

$$\widehat{J}_n = \begin{cases} 1, & \text{if } Q_n^{(1)} \leq Q_n^{(2)}, \\ 2, & \text{if } Q_n^{(2)} < Q_n^{(1)}. \end{cases}$$

Since the reported statistic is computed from the selected candidate,

$$AR_n(\theta_0) = g_n(\widehat{\eta})^\top \Omega^{-1} g_n(\widehat{\eta}) = Q_n^{(\widehat{J}_n)} = \min\{Q_n^{(1)}, Q_n^{(2)}\}.$$

Thus the proposition reduces to identifying the limit law of the minimum of the two candidate quadratic forms and comparing that limit with the usual $\chi^2_{d_g}$ law.

Step 1: joint weak limit of the candidate quadratic forms. By assumption,

$$\left(g_n^{(1)}, g_n^{(2)}\right) \xrightarrow{d} (G_1, G_2),$$

where (G_1, G_2) is jointly Gaussian in $\mathbb{R}^{2d_g}$, each marginal has mean zero and covariance matrix Ω, and the joint Gaussian limit is nondegenerate. Because Ω is nonsingular, the linear map

$$T : \mathbb{R}^{2d_g} \to \mathbb{R}^{2d_g}, \qquad T(x,y) := \left(\Omega^{-1/2}x, \Omega^{-1/2}y\right),$$

is continuous. Hence the continuous mapping theorem gives

$$\left(\Omega^{-1/2}g_n^{(1)}, \Omega^{-1/2}g_n^{(2)}\right) \xrightarrow{d} (H_1, H_2),$$

where

$$H_j := \Omega^{-1/2}G_j, \qquad j \in \{1,2\}.$$

Since $G_j \sim N(0, \Omega)$, it follows that

$$H_j \sim N(0, I_{d_g}), \qquad j \in \{1,2\}.$$

Define

$$Q_j := \|H_j\|^2, \qquad j \in \{1,2\}.$$

Then each Q_j has the $\chi^2_{d_g}$ distribution.

Now consider the map

$$m : \mathbb{R}^{2d_g} \to \mathbb{R}, \qquad m(u, v) := \min\{\|u\|^2, \|v\|^2\}.$$

The map m is continuous on $\mathbb{R}^{2d_g}$. Applying the continuous mapping theorem once more yields

$$AR_n(\theta_0) = \min\left\{\left\|\Omega^{-1/2} g_n^{(1)}\right\|^2, \left\|\Omega^{-1/2} g_n^{(2)}\right\|^2\right\} \xrightarrow{d} \min\{Q_1, Q_2\}.$$

This proves the first claim of the proposition.

Step 2: the usual $\chi^2_{d_g}$ critical value is no longer correct. Fix $\alpha \in (0, 1)$ and let

$$c_{1-\alpha} := \chi^2_{d_g, 1-\alpha}$$

denote the usual $(1 - \alpha)$ quantile of the $\chi^2_{d_g}$ law. Since $Q_1 \sim \chi^2_{d_g}$,

$$\Pr(Q_1 > c_{1-\alpha}) = \alpha.$$

Moreover,

$$\{\min\{Q_1, Q_2\} > c_{1-\alpha}\} = \{Q_1 > c_{1-\alpha},\, Q_2 > c_{1-\alpha}\},$$

so

$$\Pr(\min\{Q_1, Q_2\} > c_{1-\alpha}) = \Pr(Q_1 > c_{1-\alpha},\, Q_2 > c_{1-\alpha}) \leq \Pr(Q_1 > c_{1-\alpha}) = \alpha.$$

To prove strict inequality, it suffices to show

$$\Pr(Q_1 > c_{1-\alpha},\, Q_2 \leq c_{1-\alpha}) > 0.$$

Because the joint Gaussian law of (G_1, G_2) is nondegenerate, its covariance matrix on $\mathbb{R}^{2d_g}$ is nonsingular. The same is therefore true for the transformed vector (H_1, H_2). Consequently, (H_1, H_2) admits a Gaussian density on $\mathbb{R}^{2d_g}$ that is strictly positive everywhere. Consider the set

$$\mathcal{A}_{1-\alpha} := \left\{(u, v) \in \mathbb{R}^{2d_g} :\ \|u\|^2 > c_{1-\alpha} \text{ and } \|v\|^2 < c_{1-\alpha}\right\}.$$

This set is open and nonempty. Since the density of (H_1, H_2) is strictly positive everywhere,

$$\Pr\Big((H_1, H_2) \in \mathcal{A}_{1-\alpha}\Big) > 0.$$

But on $\mathcal{A}_{1-\alpha}$,

$$Q_1 = \|H_1\|^2 > c_{1-\alpha}, \qquad Q_2 = \|H_2\|^2 < c_{1-\alpha},$$

hence

$$\Pr(Q_1 > c_{1-\alpha},\, Q_2 \le c_{1-\alpha}) > 0.$$

Therefore

$$\Pr(\min\{Q_1, Q_2\} > c_{1-\alpha}) = \Pr(Q_1 > c_{1-\alpha}) - \Pr(Q_1 > c_{1-\alpha},\, Q_2 \le c_{1-\alpha}) < \alpha.$$

Step 3: transfer the strict size distortion back to the finite-sample statistic. Since Q_1 and Q_2 have continuous distributions, the random variable $\min\{Q_1, Q_2\}$ is also continuously distributed. Indeed, for any $t \in \mathbb{R}$,

$$\Pr\Big(\min\{Q_1, Q_2\} = t\Big) \le \Pr(Q_1 = t) + \Pr(Q_2 = t) = 0.$$

Hence $c_{1-\alpha}$ is a continuity point of the limiting cdf of $\min\{Q_1, Q_2\}$. By the Portmanteau theorem,

$$\Pr\Big(AR_n(\theta_0) > c_{1-\alpha}\Big) \to \Pr\Big(\min\{Q_1, Q_2\} > c_{1-\alpha}\Big) < \alpha.$$

In particular,

$$\Pr\Big(AR_n(\theta_0) > \chi^2_{d_g, 1-\alpha}\Big) \not\to \alpha.$$

Thus the usual $\chi^2_{d_g}$ critical value no longer yields pivotal weak-identification inference under this leaky selection rule.

A.3 Proof of Corollary 1

Fix an outer fold k, and write

$$\mathcal{F}_k^{\text{train}} := \mathcal{F}_k^{\text{sel}} \vee \mathcal{F}_k^{\text{est}}$$

for the LF–NCF training σ-field attached to that fold. The claim has two parts. First, a full-sample tuning rule that uses inference-fold observations in a nondegenerate way cannot be $\mathcal{F}_k^{\text{train}}$-measurable. Second, once that measurability fails, the conditional i.i.d. structure behind Lemma 2 is no longer available, and the leakage can produce first-order null-law distortion. The second claim is established by reduction to Proposition 1.

Step 1: nondegenerate full-sample tuning is not training-measurable. Suppose, to the contrary, that $\widehat{\gamma}$ is $\mathcal{F}_k^{\text{train}}$-measurable. Then for every candidate value $\gamma \in \Gamma_n$, the event

$$\{\widehat{\gamma} = \gamma\} \in \mathcal{F}_k^{\text{train}}.$$

In particular,

$$\mathbf{1}\{\widehat{\gamma} = \gamma_1\}$$

is $\mathcal{F}_k^{\text{train}}$-measurable. Therefore, by the defining property of conditional expectation,

$$\mathbb{E}\Big[\mathbf{1}\{\widehat{\gamma} = \gamma_1\} \mid \mathcal{F}_k^{\text{train}}\Big] = \mathbf{1}\{\widehat{\gamma} = \gamma_1\} \qquad \text{a.s.}$$

But

$$\mathbb{E}\Big[\mathbf{1}\{\widehat{\gamma} = \gamma_1\} \mid \mathcal{F}_k^{\text{train}}\Big] = \Pr(\widehat{\gamma} = \gamma_1 \mid \mathcal{F}_k^{\text{train}}),$$

so

$$\Pr(\widehat{\gamma} = \gamma_1 \mid \mathcal{F}_k^{\text{train}}) = \mathbf{1}\{\widehat{\gamma} = \gamma_1\} \qquad \text{a.s.} \tag{A.1}$$

The right-hand side takes values only in $\{0, 1\}$. Hence

$$\Pr(\widehat{\gamma} = \gamma_1 \mid \mathcal{F}_k^{\text{train}}) \in \{0, 1\} \qquad \text{a.s.}$$

This contradicts the stated hypothesis that there exists an event $A \in \mathcal{F}_k^{\text{train}}$ with $\Pr(A) > 0$ on which

$$0 < \Pr(\widehat{\gamma} = \gamma_1 \mid \mathcal{F}_k^{\text{train}}) < 1.$$

Therefore $\widehat{\gamma}$ cannot be $\mathcal{F}_k^{\text{train}}$-measurable.

Step 2: the conditional i.i.d. argument of Lemma 2 no longer applies. Lemma 2 requires that the selected nuisance objects used on fold k be measurable with respect to the training σ-field $\mathcal{F}_k^{\text{train}}$. More precisely, the proof of Lemma 2 uses the fact that

$$(\widehat{\gamma}_k, \widehat{\eta}_k) \quad \text{is} \quad \mathcal{F}_k^{\text{train}}\text{-measurable},$$

together with the independence of the disjoint inference block $\mathcal{F}_k^{\text{inf}}$ from $\mathcal{F}_k^{\text{train}}$, to conclude that conditional on $\mathcal{F}_k^{\text{train}}$, the inference-fold observations remain i.i.d. and independent of the tuning/fitting choices.

Once $\widehat{\gamma}$ is not $\mathcal{F}_k^{\text{train}}$-measurable, that argument fails at its first step: the selected nuisance pipeline is no longer fixed when conditioning on the training data alone. Consequently, the reported cross-fitted score

$$\widehat{g}_n(\theta_0) = \frac{1}{\sqrt{n}} \sum_{\ell=1}^{K} \sum_{i \in I_\ell^{\text{inf}}} \psi\Big(W_i; \theta_0, \widehat{\eta}_\ell(\widehat{\gamma})\Big) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi\Big(W_i; \theta_0, \widehat{\eta}_{-i}(\widehat{\gamma})\Big)$$

need not be representable, on fold k, as an average of summands that are conditionally i.i.d. given $\mathcal{F}_k^{\text{train}}$. Thus the pivotality argument for weak-identification-robust statistics based on conditional i.i.d. inference-fold scores is no longer available.

This proves the first assertion of the corollary: a genuinely full-sample tuning rule that reuses inference-fold observations in a nondegenerate way is outside the LF–NCF measurability structure, so weak-identification-robust pivotality is not guaranteed.

Step 3: first-order failure can occur. To see that the resulting failure can be first-order rather than merely technical, specialize to the two-candidate leaky selection design in Proposition 1. There, the selected candidate is determined by a criterion computed from the same observations used to form the reported quadratic form, so the selector is a full-sample rule. Let

$$Q_n^{(j)} := g_n^{(j)\top} \Omega^{-1} g_n^{(j)}, \qquad j \in \{1, 2\},$$

and define the selected index by

$$\widehat{J}_n = \begin{cases} 1, & \text{if } Q_n^{(1)} \le Q_n^{(2)}, \\ 2, & \text{if } Q_n^{(2)} < Q_n^{(1)}. \end{cases}$$

Then

$$AR_n(\theta_0) = Q_n^{(\widehat{J}_n)} = \min\{Q_n^{(1)}, Q_n^{(2)}\}.$$

The event $\{\widehat{J}_n = 1\} = \{Q_n^{(1)} \le Q_n^{(2)}\}$ depends on the same realized quadratic forms that enter the reported statistic. Under the joint weak-convergence assumptions of Proposition 1,

$$AR_n(\theta_0) \xrightarrow{d} \min\{Q_1, Q_2\},$$

where $Q_1, Q_2 \sim \chi^2_{d_g}$ marginally, but

$$\Pr\Big(\min\{Q_1, Q_2\} > \chi^2_{d_g,1-\alpha}\Big) < \alpha.$$

Hence

$$\Pr\Big(AR_n(\theta_0) > \chi^2_{d_g,1-\alpha}\Big) \to \Pr\Big(\min\{Q_1, Q_2\} > \chi^2_{d_g,1-\alpha}\Big) < \alpha.$$

Thus the size distortion is of order $O(1)$. Combining Steps 1–3 proves the corollary.

A.4 Proof of Theorem 1

Work throughout under the fixed alternative

$$H_1 : \theta = \theta_0 + \Delta, \qquad \Delta \neq 0,$$

with Δ held fixed as $n \to \infty$. Recall that

$$Y = \theta D + U, \qquad D = Z^\top \Pi_n + V, \qquad \Pi_n = \pi/\sqrt{n},$$

where $Z \in \mathbb{R}^{d_z}$ satisfies

$$\mathbb{E}[Z] = 0, \qquad \mathbb{E}[ZZ^\top] = \Sigma_Z \succ 0,$$

and Z is independent of (U, V). Also,

$$\mathbb{E}[U] = 0, \qquad \mathbb{E}[V] = 0, \qquad \sigma_\Delta^2 := \mathbb{E}[(U + \Delta V)^2] \in (0, \infty).$$

The proof has four parts. First, the mean and covariance limit of $g_n(\theta_0)$ are derived. Second, the limit of $\Omega_n(\theta_0)$ is established. Third, the AR statistic is shown to converge to a noncentral χ^2 law. Fourth, the limiting rejection probability is shown to be strictly increasing in the effective concentration index.

Step 1: expansion of $g_n(\theta_0)$ under the fixed alternative. By definition,

$$g_n(\theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i \Big(Y_i - \theta_0 D_i \Big).$$

Under $H_1 : \theta = \theta_0 + \Delta$,

$$Y_i - \theta_0 D_i = (\theta_0 + \Delta) D_i + U_i - \theta_0 D_i = \Delta D_i + U_i.$$

Using the reduced form $D_i = Z_i^\top \Pi_n + V_i$ and $\Pi_n = \pi/\sqrt{n}$,

$$Y_i - \theta_0 D_i = U_i + \Delta V_i + \Delta Z_i^\top \pi/\sqrt{n}.$$

Substituting into $g_n(\theta_0)$,

$$\begin{aligned} g_n(\theta_0) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i \Big(U_i + \Delta V_i + \Delta Z_i^\top \pi/\sqrt{n} \Big) \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i (U_i + \Delta V_i) + \frac{\Delta}{n} \sum_{i=1}^{n} Z_i Z_i^\top \pi. \end{aligned} \tag{A.2}$$

Define

$$S_n := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i (U_i + \Delta V_i), \qquad B_n := \Delta \Big(\frac{1}{n} \sum_{i=1}^{n} Z_i Z_i^\top \Big) \pi.$$

Then $g_n(\theta_0) = S_n + B_n$.

Mean and covariance of the stochastic term. Because Z is independent of (U, V) and $\mathbb{E}[U + \Delta V] = 0$,

$$\mathbb{E}[Z(U + \Delta V)] = \mathbb{E}[Z]\mathbb{E}[U + \Delta V] = 0.$$

Hence S_n is a centered sum. Its covariance matrix is

$$\mathrm{Var}(S_n) = \mathrm{Var}\Big(Z(U + \Delta V) \Big) = \mathbb{E}\Big[Z Z^\top (U + \Delta V)^2 \Big] =: \Omega_\Delta. \tag{A.3}$$

Since d_z is fixed and $\sigma_\Delta^2 < \infty$, the vector $Z(U + \Delta V)$ has finite second moments. Therefore the multivariate central limit theorem yields

$$S_n \Rightarrow N(0, \Omega_\Delta). \tag{A.4}$$

Deterministic limit of the drift term. By the law of large numbers,

$$\frac{1}{n}\sum_{i=1}^{n} Z_i Z_i^\top \to \Sigma_Z \qquad \text{in probability.}$$

Hence

$$B_n = \Delta\Big(\frac{1}{n}\sum_{i=1}^{n} Z_i Z_i^\top\Big)\pi \to \Delta\Sigma_Z\pi \qquad \text{in probability.} \tag{A.5}$$

Combining (A.4), (A.5), and (A.2), Slutsky's theorem gives

$$g_n(\theta_0) \Rightarrow N(\Delta\Sigma_Z\pi, \Omega_\Delta). \tag{A.6}$$

Step 2: limit of $\Omega_n(\theta_0)$. By definition,

$$\Omega_n(\theta_0) = \mathbb{E}_{P_n}\Big[ZZ^\top(Y-\theta_0 D)^2\Big].$$

Under H_1,

$$Y - \theta_0 D = U + \Delta V + \Delta Z^\top\pi/\sqrt{n}.$$

Therefore

$$\begin{aligned}
\Omega_n(\theta_0) &= \mathbb{E}\Big[ZZ^\top\Big(U+\Delta V+\Delta Z^\top\pi/\sqrt{n}\Big)^2\Big] \\
&= \mathbb{E}\Big[ZZ^\top(U+\Delta V)^2\Big] + \frac{2\Delta}{\sqrt{n}}\mathbb{E}\Big[ZZ^\top(U+\Delta V)(Z^\top\pi)\Big] \\
&\quad + \frac{\Delta^2}{n}\mathbb{E}\Big[ZZ^\top(Z^\top\pi)^2\Big].
\end{aligned} \tag{A.7}$$

The first term is Ω_Δ by definition (A.3). The middle term vanishes exactly, because Z is independent of (U, V) and $\mathbb{E}[U + \Delta V] = 0$:

$$\mathbb{E}\Big[ZZ^\top(U+\Delta V)(Z^\top\pi)\Big] = \mathbb{E}[U+\Delta V]\,\mathbb{E}\Big[ZZ^\top(Z^\top\pi)\Big] = 0.$$

The last term is $O(n^{-1})$ entrywise. Indeed, $\mathbb{E}[\|Z\|^4] < \infty$ implies that the matrix $\mathbb{E}[ZZ^\top(Z^\top\pi)^2]$ has finite entries. Hence

$$\frac{\Delta^2}{n}\mathbb{E}\Big[ZZ^\top(Z^\top\pi)^2\Big] \to 0.$$

Thus (A.7) implies

$$\Omega_n(\theta_0) \to \Omega_\Delta. \tag{A.8}$$

Under the maintained independence of Z and (U, V), the matrix Ω_Δ simplifies to

$$\begin{aligned} \Omega_\Delta &= \mathbb{E}\Big[ZZ^\top (U + \Delta V)^2\Big] = \mathbb{E}[ZZ^\top]\,\mathbb{E}[(U + \Delta V)^2] \\ &= \Sigma_Z\,\sigma_\Delta^2. \end{aligned} \tag{A.9}$$

Because $\Sigma_Z \succ 0$ and $\sigma_\Delta^2 > 0$, it follows that $\Omega_\Delta \succ 0$.

Step 3: asymptotic law of the AR statistic. The oracle AR statistic is

$$AR_n(\theta_0) = g_n(\theta_0)^\top \Omega_n(\theta_0)^{-1} g_n(\theta_0).$$

By (A.6) and (A.8),

$$\Big(g_n(\theta_0), \Omega_n(\theta_0)\Big) \Rightarrow \Big(G, \Omega_\Delta\Big), \qquad G \sim N(\Delta\Sigma_Z\pi, \Omega_\Delta).$$

Since Ω_Δ is positive definite, the map

$$(x, A) \mapsto x^\top A^{-1} x$$

is continuous on $\mathbb{R}^{d_z} \times \mathbb{S}_{++}^{d_z}$. Hence the continuous mapping theorem yields

$$AR_n(\theta_0) \Rightarrow G^\top \Omega_\Delta^{-1} G. \tag{A.10}$$

Write

$$H := \Omega_\Delta^{-1/2} G.$$

Then

$$H \sim N\Big(\Omega_\Delta^{-1/2}\Delta\Sigma_Z\pi,\ I_{d_z}\Big),$$

and

$$G^\top \Omega_\Delta^{-1} G = \|H\|^2.$$

Therefore the limit in (A.10) is a noncentral chi-square random variable with d_z degrees of freedom and noncentrality parameter

$$\lambda_\Delta = (\Delta\Sigma_Z\pi)^\top \Omega_\Delta^{-1}(\Delta\Sigma_Z\pi).$$

That is,

$$AR_n(\theta_0) \Rightarrow \chi^2_{d_z}(\lambda_\Delta).$$

Using (A.9),

$$\begin{aligned}\lambda_\Delta &= (\Delta\Sigma_Z\pi)^\top(\sigma^2_\Delta\Sigma_Z)^{-1}(\Delta\Sigma_Z\pi)\\ &= \frac{\Delta^2}{\sigma^2_\Delta}\,\pi^\top\Sigma_Z^\top\Sigma_Z^{-1}\Sigma_Z\pi = \frac{\Delta^2}{\sigma^2_\Delta}\,\pi^\top\Sigma_Z\pi,\end{aligned}$$

since Σ_Z is symmetric. This proves the displayed formula for λ_Δ.

Step 4: strict monotonicity of local power in effective concentration. Fix $\alpha \in (0,1)$ and let

$$c_{1-\alpha} := \chi^2_{d_z,1-\alpha}.$$

By Step 3,

$$\Pr\Big(AR_n(\theta_0) > c_{1-\alpha}\Big) \to \Pr\Big(\chi^2_{d_z}(\lambda_\Delta) > c_{1-\alpha}\Big).$$

It remains to show that the right-hand side is strictly increasing in λ_Δ, and hence in $\pi^\top\Sigma_Z\pi$.

Let

$$X_\lambda \sim \chi^2_{d_z}(\lambda).$$

A standard representation is

$$X_\lambda \overset{d}{=} (Z_1 + \sqrt{\lambda})^2 + W,$$

where $Z_1 \sim N(0,1)$, $W \sim \chi^2_{d_z-1}$, and Z_1 and W are independent; when $d_z = 1$, $W \equiv 0$. Fix $c > 0$. Conditional on $W = w < c$,

$$\Pr(X_\lambda > c \mid W = w) = \Pr\left((Z_1 + \sqrt{\lambda})^2 > c - w\right).$$

Let $a := \sqrt{\lambda} \geq 0$ and $t_w := \sqrt{c - w} > 0$. Then

$$\begin{aligned} h_w(a) &:= \Pr\Big((Z_1 + a)^2 > t_w^2\Big) \\ &= \Pr(Z_1 > t_w - a) + \Pr(Z_1 < -t_w - a) \\ &= 1 - \Phi(t_w - a) + \Phi(-t_w - a). \end{aligned}$$

Differentiating with respect to a,

$$h_w'(a) = \phi(t_w - a) - \phi(t_w + a),$$

where ϕ is the standard normal density. For every $a > 0$ and $t_w > 0$,

$$|t_w - a| < t_w + a,$$

hence, since $\phi(x)$ is strictly decreasing in $|x|$,

$$\phi(t_w - a) > \phi(t_w + a).$$

Therefore

$$h_w'(a) > 0 \qquad \text{for every } a > 0 \text{ and every } w < c.$$

So $h_w(a)$ is strictly increasing in $a = \sqrt{\lambda}$, and therefore in λ, whenever $w < c$. If $w \geq c$, then $\Pr(X_\lambda > c \mid W = w) = 1$ for all λ, so there is no decrease there either.

Now integrate over W. Since $c = c_{1-\alpha} > 0$ and $W \sim \chi^2_{d_z - 1}$,

$$\Pr(W < c_{1-\alpha}) > 0.$$

Hence there is a set of w-values with positive probability on which the conditional tail probability is strictly increasing in λ. It follows that

$$\lambda \mapsto \Pr\Big(\chi^2_{d_z}(\lambda) > c_{1-\alpha}\Big)$$

is strictly increasing on $[0, \infty)$.

Finally, under (A.9),

$$\lambda_\Delta = \frac{\Delta^2}{\sigma_\Delta^2}\, \pi^\top \Sigma_Z \pi.$$

For fixed $\Delta \neq 0$ and fixed $\sigma_\Delta^2 \in (0, \infty)$, this is a strictly increasing affine transformation of $\pi^\top \Sigma_Z \pi$. Since

$$n\, \Pi_n^\top \Sigma_Z \Pi_n = n\Big(\frac{\pi}{\sqrt{n}}\Big)^\top \Sigma_Z \Big(\frac{\pi}{\sqrt{n}}\Big) = \pi^\top \Sigma_Z \pi,$$

the same monotonicity statement holds in the effective concentration index $n\, \Pi_n^\top \Sigma_Z \Pi_n$.

This proves the theorem.

A.5 Proof of Theorem 2

Fix an outer fold k. To simplify notation, suppress the fold index k throughout this proof and write

$$\Gamma_n^{\text{good}} := \Gamma_{n,k}^{\text{good}}, \qquad \Gamma_n^{\text{bad}} := \Gamma_{n,k}^{\text{bad}}, \qquad \widehat{\Gamma}_{\varepsilon_n} := \widehat{\Gamma}_{\varepsilon_n,k}.$$

Also write

$$\widehat{R}(\gamma) := \widehat{R}_{\text{pred}}(\gamma), \qquad R(\gamma) := R_{\text{pred}}(\gamma).$$

By assumption, $\Gamma_n^{\text{good}} \neq \varnothing$ and

$$\Gamma_n = \Gamma_n^{\text{good}} \cup \Gamma_n^{\text{bad}}.$$

Since the library Γ_n is finite, the infima below are attained. Define

$$\underline{R}_n^{\text{good}} := \min_{\gamma \in \Gamma_n^{\text{good}}} R(\gamma), \qquad \underline{R}_n^{\text{bad}} := \min_{\gamma \in \Gamma_n^{\text{bad}}} R(\gamma).$$

Assumption (iii) of Theorem 2 states that

$$\underline{R}_n^{\text{bad}} - \underline{R}_n^{\text{good}} \geq c_n. \tag{A.11}$$

Pick any

$$\gamma_n^\star \in \arg\min_{\gamma \in \Gamma_n^{\text{good}}} R(\gamma),$$

so that

$$R(\gamma_n^\star) = \underline{R}_n^{\text{good}}.$$

The proof proceeds by working on a high-probability event on which empirical risks are uniformly close to their population counterparts and the screening tolerance is smaller than the risk-separation scale.

Step 1: define the high-probability screening event. Let

$$\mathcal{E}_n := \left\{ \sup_{\gamma \in \Gamma_n} \left| \widehat{R}(\gamma) - R(\gamma) \right| \le \frac{c_n}{4} \right\} \cap \left\{ \varepsilon_n \le \frac{c_n}{4} \right\}.$$

Because

$$\sup_{\gamma \in \Gamma_n} \left| \widehat{R}(\gamma) - R(\gamma) \right| = o_P(c_n)$$

uniformly over $P \in \mathcal{P}_n$, it follows that

$$\sup_{P \in \mathcal{P}_n} \Pr_P \left(\sup_{\gamma \in \Gamma_n} \left| \widehat{R}(\gamma) - R(\gamma) \right| > \frac{c_n}{4} \right) \to 0.$$

Since $\varepsilon_n = o(c_n)$, the second event in the definition of $\mathcal{E}_n$ holds deterministically for all sufficiently large n. Hence

$$\sup_{P \in \mathcal{P}_n} \Pr_P(\mathcal{E}_n^c) \to 0. \tag{A.12}$$

Step 2: on $\mathcal{E}_n$, every bad candidate is empirically separated from the good class. Fix any $\gamma \in \Gamma_n^{\mathrm{bad}}$. On $\mathcal{E}_n$, the uniform concentration event gives $\widehat{R}(\gamma) \ge R(\gamma) - c_n/4$. Because $\gamma \in \Gamma_n^{\mathrm{bad}}$, one has $R(\gamma) \ge \underline{R}_n^{\mathrm{bad}}$. The population separation condition (A.11) then yields $\underline{R}_n^{\mathrm{bad}} \ge \underline{R}_n^{\mathrm{good}} + c_n$, and by definition of $\gamma_n^\star$ the latter equals $R(\gamma_n^\star) + c_n$. Combining these bounds and then using once more the concentration event, $\widehat{R}(\gamma_n^\star) \le R(\gamma_n^\star) + c_n/4$, gives

$$\begin{aligned}
\widehat{R}(\gamma) &\ge R(\gamma) - \frac{c_n}{4} && \text{(A.13)}\\
&\ge \underline{R}_n^{\mathrm{bad}} - \frac{c_n}{4} && \text{(A.14)}\\
&\ge \underline{R}_n^{\mathrm{good}} + \frac{3c_n}{4} && \text{(A.15)}\\
&= R(\gamma_n^\star) + \frac{3c_n}{4} && \text{(A.16)}\\
&\ge \widehat{R}(\gamma_n^\star) + \frac{c_n}{2}. && \text{(A.17)}
\end{aligned}$$

Thus, on $\mathcal{E}_n$,

$$\widehat{R}(\gamma) \geq \widehat{R}(\gamma_n^\star) + \frac{c_n}{2} \qquad \text{for every } \gamma \in \Gamma_n^{\text{bad}}. \tag{A.18}$$

Step 3: on $\mathcal{E}_n$, an empirical minimizer must be good. Let

$$\widehat{\gamma}_n^{\min} \in \arg\min_{\gamma \in \Gamma_n} \widehat{R}(\gamma), \qquad \widehat{R}_{\min} := \widehat{R}(\widehat{\gamma}_n^{\min}).$$

Suppose, for contradiction, that $\widehat{\gamma}_n^{\min} \in \Gamma_n^{\text{bad}}$. Then (A.18) implies

$$\widehat{R}_{\min} = \widehat{R}(\widehat{\gamma}_n^{\min}) \geq \widehat{R}(\gamma_n^\star) + \frac{c_n}{2} > \widehat{R}(\gamma_n^\star),$$

which contradicts the definition of $\widehat{\gamma}_n^{\min}$ as an empirical minimizer. Therefore,

$$\widehat{\gamma}_n^{\min} \in \Gamma_n^{\text{good}} \qquad \text{on } \mathcal{E}_n. \tag{A.19}$$

Step 4: on $\mathcal{E}_n$, the screen is nonempty. By definition,

$$\widehat{R}(\widehat{\gamma}_n^{\min}) = \widehat{R}_{\min}.$$

Therefore

$$\widehat{\gamma}_n^{\min} \in \widehat{\Gamma}_{\varepsilon_n}$$

deterministically, because $\varepsilon_n \geq 0$. Hence

$$\widehat{\Gamma}_{\varepsilon_n} \neq \varnothing.$$

Combining this deterministic inclusion with (A.19) yields the stronger property

$$\widehat{\Gamma}_{\varepsilon_n} \cap \Gamma_n^{\text{good}} \neq \varnothing \qquad \text{on } \mathcal{E}_n. \tag{A.20}$$

Step 5: on $\mathcal{E}_n$, the screen excludes all bad candidates. Fix $\gamma \in \Gamma_n^{\text{bad}}$. By (A.18),

$$\widehat{R}(\gamma) \geq \widehat{R}(\gamma_n^\star) + \frac{c_n}{2}.$$

Since $\widehat{R}_{\min} \leq \widehat{R}(\gamma_n^\star)$, this implies

$$\widehat{R}(\gamma) \geq \widehat{R}_{\min} + \frac{c_n}{2}.$$

On $\mathcal{E}_n$, $\varepsilon_n \le c_n/4$, so

$$\widehat{R}(\gamma) > \widehat{R}_{\min} + \varepsilon_n.$$

Thus $\gamma \notin \widehat{\Gamma}_{\varepsilon_n}$. Because $\gamma \in \Gamma_n^{\mathrm{bad}}$ was arbitrary,

$$\widehat{\Gamma}_{\varepsilon_n} \subseteq \Gamma_n^{\mathrm{good}} \qquad \text{on } \mathcal{E}_n. \tag{A.21}$$

Step 6: conclude the theorem. From (A.20) and (A.21),

$$\mathcal{E}_n \subseteq \Big\{\widehat{\Gamma}_{\varepsilon_n} \subseteq \Gamma_n^{\mathrm{good}},\ \widehat{\Gamma}_{\varepsilon_n} \neq \varnothing\Big\}.$$

Therefore, by (A.12),

$$\sup_{P\in\mathcal{P}_n} \Pr_P\Big(\widehat{\Gamma}_{\varepsilon_n} \subseteq \Gamma_n^{\mathrm{good}},\ \widehat{\Gamma}_{\varepsilon_n} \neq \varnothing\Big) \to 1.$$

This proves the first claim of Theorem 2.

Finally, let $\widehat{\gamma}_k$ be any selector taking values in $\widehat{\Gamma}_{\varepsilon_n,k}$. On the event

$$\Big\{\widehat{\Gamma}_{\varepsilon_n,k} \subseteq \Gamma_{n,k}^{\mathrm{good}},\ \widehat{\Gamma}_{\varepsilon_n,k} \neq \varnothing\Big\},$$

one has $\widehat{\gamma}_k \in \Gamma_{n,k}^{\mathrm{good}}$. By assumption (i), every $\gamma \in \Gamma_{n,k}^{\mathrm{good}}$ is rate-valid in the sense of (4.6). Hence the selected nuisance $\widehat{\eta}_k(\widehat{\gamma}_k)$ is rate-valid with probability $1 - o(1)$, uniformly over $P \in \mathcal{P}_n$. This proves the final claim.

A.6 Proof of Proposition 3

Let

$$A_n := \{\widehat{\gamma}_n^{\mathrm{pred}} = \gamma_2\}.$$

Under the deterministic tie-breaking rule used by prediction-only selection,

$$A_n = \{\widehat{R}_{\mathrm{pred}}(\gamma_2) \le \widehat{R}_{\mathrm{pred}}(\gamma_1)\}.$$

Therefore the assumption

$$\liminf_{n\to\infty} \mathbb{P}\Big(\widehat{R}_{\mathrm{pred}}(\gamma_2) \le \widehat{R}_{\mathrm{pred}}(\gamma_1)\Big) > 0$$

implies

$$\liminf_{n\to\infty} \mathbb{P}(A_n) > 0.$$

Thus prediction-only tuning selects the low-strength candidate with nonvanishing probability.

For $j \in \{1, 2\}$, define the fixed-candidate statistic

$$AR_n^{(j)}(\theta_0) := AR_n(\theta_0; \gamma_j),$$

and let

$$c_{1-\alpha} := \chi^2_{d_z,1-\alpha}.$$

By the candidate-specific asymptotic experiment assumed in the proposition, under the fixed alternative $\theta = \theta_0 + \Delta$,

$$AR_n^{(1)}(\theta_0) \Rightarrow \chi^2_{d_z}(\lambda_1), \qquad AR_n^{(2)}(\theta_0) \Rightarrow \chi^2_{d_z}(0),$$

where $\lambda_{1,n} \to \lambda_1 > 0$ and $\lambda_{2,n} \to 0$. Equivalently, the second candidate has limiting noncentrality zero. Hence, by continuity of the rejection probability at the fixed critical value $c_{1-\alpha}$,

$$\mathbb{P}\Big(AR_n^{(2)}(\theta_0) > c_{1-\alpha}\Big) \to \Pr\Big(\chi^2_{d_z}(0) > c_{1-\alpha}\Big) = \alpha.$$

Thus γ_2 is a candidate pipeline whose fixed-candidate AR power converges to size.

The selected prediction-only statistic is

$$AR_n^{\mathrm{pred}}(\theta_0) := AR_n(\theta_0; \widehat{\gamma}_n^{\mathrm{pred}}).$$

On the event A_n, one has

$$AR_n^{\mathrm{pred}}(\theta_0) = AR_n^{(2)}(\theta_0).$$

Therefore prediction-only tuning places nonvanishing probability on a pipeline whose fixed-candidate AR power converges to the nominal size.

Now take any subsequence along which

$$\mathbb{P}(\widehat{\gamma}_n^{\mathrm{pred}} = \gamma_2) \to 1.$$

Let

$$B_n := \{AR_n^{\mathrm{pred}}(\theta_0) > c_{1-\alpha}\}, \qquad C_n := \{AR_n^{(2)}(\theta_0) > c_{1-\alpha}\}.$$

On A_n, the events B_n and C_n coincide. Hence their symmetric difference is contained in A_n^c:

$$B_n \triangle C_n \subseteq A_n^c.$$

It follows that

$$|\mathbb{P}(B_n) - \mathbb{P}(C_n)| \leq \mathbb{P}(B_n \triangle C_n) \leq \mathbb{P}(A_n^c) = \mathbb{P}(\widehat{\gamma}_n^{\mathrm{pred}} \neq \gamma_2) \to 0.$$

Since

$$\mathbb{P}(C_n) = \mathbb{P}\Big(AR_n^{(2)}(\theta_0) > c_{1-\alpha}\Big) \to \alpha,$$

we conclude that

$$\mathbb{P}\Big(AR_n^{\mathrm{pred}}(\theta_0) > c_{1-\alpha}\Big) \to \alpha$$

along that subsequence. This proves the proposition.

For the optional mixture statement, assume the stated mixture limit holds along a subsequence with

$$\mathbb{P}(A_n) \to p \in (0, 1].$$

The fixed-candidate limits imply

$$\mathbb{P}\Big(AR_n^{(1)}(\theta_0) > c_{1-\alpha}\Big) \to \Pr\Big(\chi^2_{d_z}(\lambda_1) > c_{1-\alpha}\Big)$$

and

$$\mathbb{P}\Big(AR_n^{(2)}(\theta_0) > c_{1-\alpha}\Big) \to \alpha.$$

Under the mixture-limit condition,

$$\mathbb{P}\Big(AR_n^{\mathrm{pred}}(\theta_0) > c_{1-\alpha}\Big) \to (1-p)\Pr\Big(\chi^2_{d_z}(\lambda_1) > c_{1-\alpha}\Big) + p\alpha.$$

Since $\lambda_1 > 0$, the noncentral χ^2 rejection probability at the central $(1-\alpha)$-critical value is strictly larger than α. Therefore, when $p > 0$,

$$(1-p)\Pr\Big(\chi^2_{d_z}(\lambda_1) > c_{1-\alpha}\Big) + p\alpha < \Pr\Big(\chi^2_{d_z}(\lambda_1) > c_{1-\alpha}\Big).$$

This gives the asserted strict power gap.

A.7 Proof of Proposition 4

Let

$$E_k := \left\{\widehat{\Gamma}_{\varepsilon_n,k} \neq \varnothing\right\}.$$

The argument below is deterministic on the event E_k; the assumption $\Pr(E_k) \to 1$ ensures that the proposition is asymptotically non-vacuous.

Because $n_{\text{sel}} > 0$ is common across all candidates in $\widehat{\Gamma}_{\varepsilon_n,k}$, maximizing the unnormalized criterion $\widehat{S}_{n,\kappa}(\gamma)$ is equivalent on E_k to maximizing the normalized criterion $\overline{S}_{n,\kappa}(\gamma)$. Hence, by the definition of $\hat{\gamma}_k$,

$$\overline{S}_{n,\kappa}(\hat{\gamma}_k) \geq \overline{S}_{n,\kappa}(\gamma) \qquad \text{for every } \gamma \in \widehat{\Gamma}_{\varepsilon_n,k}. \tag{A.22}$$

Next, by the definition of Δ_n, on the event E_k,

$$\left|\overline{S}_{n,\kappa}(\gamma) - \overline{S}_{n,\kappa}(\eta(\gamma))\right| \leq \Delta_n \qquad \text{for every } \gamma \in \widehat{\Gamma}_{\varepsilon_n,k}. \tag{A.23}$$

Applying (A.23) with $\gamma = \hat{\gamma}_k$ yields

$$\overline{S}_{n,\kappa}(\eta(\hat{\gamma}_k)) \geq \overline{S}_{n,\kappa}(\hat{\gamma}_k) - \Delta_n. \tag{A.24}$$

Now fix an arbitrary $\gamma \in \widehat{\Gamma}_{\varepsilon_n,k}$. Combining (A.24) with the argmax property (A.22),

$$\overline{S}_{n,\kappa}(\eta(\hat{\gamma}_k)) \geq \overline{S}_{n,\kappa}(\hat{\gamma}_k) - \Delta_n \geq \overline{S}_{n,\kappa}(\gamma) - \Delta_n.$$

Applying (A.23) again, this time to the arbitrary candidate γ, gives

$$\overline{S}_{n,\kappa}(\eta(\hat{\gamma}_k)) \geq \overline{S}_{n,\kappa}(\eta(\gamma)) - 2\Delta_n.$$

Because the choice of $\gamma \in \widehat{\Gamma}_{\varepsilon_n,k}$ was arbitrary, taking the supremum over the screened set yields

$$\overline{S}_{n,\kappa}(\eta(\hat{\gamma}_k)) \geq \sup_{\gamma \in \widehat{\Gamma}_{\varepsilon_n,k}} \overline{S}_{n,\kappa}(\eta(\gamma)) - 2\Delta_n, \tag{A.25}$$

which is the first displayed inequality in the proposition.

Finally, multiplying both sides by the common positive factor n_{sel} gives

$$S_{n,\kappa}(\eta(\hat{\gamma}_k)) \geq \sup_{\gamma \in \widehat{\Gamma}_{\varepsilon_n,k}} S_{n,\kappa}(\eta(\gamma)) - 2n_{\text{sel}}\Delta_n,$$

which is the equivalent unnormalized form.

A.8 Proof of Proposition 5

For each outer fold k, let

$$\widehat{\gamma}_k^{\text{pred}}$$

denote the prediction-only selector and

$$\widehat{\gamma}_k^{IACV}$$

the IACV selector computed from the same fold partition. Let the corresponding selected nuisance fits be

$$\widehat{\eta}_k^{\text{pred}} := \widehat{\eta}_k(\widehat{\gamma}_k^{\text{pred}}), \qquad \widehat{\eta}_k^{IACV} := \widehat{\eta}_k(\widehat{\gamma}_k^{IACV}).$$

Define the associated cross-fitted moment processes by

$$\widehat{g}_n^{\text{pred}}(\theta) := \frac{1}{\sqrt{n}} \sum_{k=1}^{K} \sum_{i \in I_k^{\text{inf}}} \psi(W_i; \theta, \widehat{\eta}_k^{\text{pred}}), \qquad \widehat{g}_n^{IACV}(\theta) := \frac{1}{\sqrt{n}} \sum_{k=1}^{K} \sum_{i \in I_k^{\text{inf}}} \psi(W_i; \theta, \widehat{\eta}_k^{IACV}),$$

and let

$$g_{n,0}(\theta) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(W_i; \theta, \eta_0)$$

denote the oracle moment process.

The proof has three parts. First, the singleton-screen branch gives exact equality of the selectors. Second, on the non-singleton branch the two selected pipelines are shown to produce asymptotically equivalent moment processes. Third, under the strong-identification conditions of Theorem 5, the corresponding GMM estimators are shown to have the same first-order asymptotic expansion and are therefore asymptotically equivalent.

Step 1: exact equality on the singleton-screen event. For each fold k, define

$$A_{n,k} := \left\{ \widehat{\Gamma}_{\varepsilon_n,k} = \{\widehat{\gamma}_k^{\text{pred}}\} \right\},$$

and

$$B_{n,k} := \begin{Bmatrix} \text{all candidates in } \widehat{\Gamma}_{\varepsilon_n,k} \text{ are rate-valid and generate} \\ \text{nuisance fits that are } o_P(n^{-1/4})\text{-equivalent} \end{Bmatrix}.$$

By the screening-equivalence hypothesis in Proposition 5, for each fold k,

$$\Pr(A_{n,k} \cup B_{n,k}) \to 1.$$

On the event $A_{n,k}$, the screened set contains only the prediction-only selector. Since IACV maximizes the strength proxy over the screened set, it must choose the same candidate:

$$\widehat{\gamma}_k^{IACV} = \widehat{\gamma}_k^{\text{pred}} \qquad \text{on } A_{n,k}.$$

Therefore, if the singleton-screen branch itself occurs with probability $1 - o(1)$, then

$$\Pr\Big(\widehat{\gamma}_k^{IACV} = \widehat{\gamma}_k^{\text{pred}}\Big) \to 1.$$

This proves the last assertion of the proposition.

Step 2: the two selected moment processes are asymptotically equivalent. Now consider the non-singleton branch $B_{n,k}$. By assumption, on $B_{n,k}$ all candidates in $\widehat{\Gamma}_{\varepsilon_n,k}$ are rate-valid and pairwise $o_P(n^{-1/4})$-equivalent in the nuisance norm. In particular, both selected candidates, $\widehat{\gamma}_k^{\text{pred}}$ and $\widehat{\gamma}_k^{IACV}$, belong to the screened set on $B_{n,k}$. Hence their selected nuisances are rate-valid. Therefore the selected-nuisance oracle-equivalence bound used in Theorem 3 applies to each of the two selectors:

$$\sup_{\theta \in \Theta} \Big\| \widehat{g}_n^{\text{pred}}(\theta) - g_{n,0}(\theta) \Big\| = o_P(1), \tag{A.26}$$

and

$$\sup_{\theta \in \Theta} \Big\| \widehat{g}_n^{IACV}(\theta) - g_{n,0}(\theta) \Big\| = o_P(1). \tag{A.27}$$

Indeed, (A.26) and (A.27) are just the conclusion of Assumption 5 applied candidate-wise to the two selected pipelines.

By the triangle inequality,

$$\sup_{\theta\in\Theta}\left\|\widehat{g}_n^{IACV}(\theta)-\widehat{g}_n^{\mathrm{pred}}(\theta)\right\| \le \sup_{\theta\in\Theta}\left\|\widehat{g}_n^{IACV}(\theta)-g_{n,0}(\theta)\right\|+\sup_{\theta\in\Theta}\left\|\widehat{g}_n^{\mathrm{pred}}(\theta)-g_{n,0}(\theta)\right\| \\ = o_P(1). \tag{A.28}$$

This proves the first displayed conclusion of the proposition on the event $A_{n,k}\cup B_{n,k}$, and hence unconditionally because that event has probability $1-o(1)$.

For (A.28), candidatewise rate-validity is enough: each selected score is $o_{\mathbb{P}}(1)$-equivalent to the same oracle score $g_{n,0}$. Candidatewise rate-validity also implies pairwise $o_{\mathbb{P}}(n^{-1/4})$-equivalence of the nuisance fits by the triangle inequality. Thus the pairwise clause in $B_{n,k}$ is redundant mathematically, but it makes explicit the economic content of the non-singleton branch: the surviving candidates are already indistinguishable at the nuisance-rate scale relevant for first-order orthogonal inference.

Step 3: asymptotic equivalence of the strong-identification GMM estimators. Define

$$\widehat{m}_n^{\mathrm{pred}}(\theta):=n^{-1/2}\widehat{g}_n^{\mathrm{pred}}(\theta),\qquad \widehat{m}_n^{IACV}(\theta):=n^{-1/2}\widehat{g}_n^{IACV}(\theta).$$

Then (A.28) implies

$$\sup_{\theta\in\Theta}\left\|\widehat{m}_n^{IACV}(\theta)-\widehat{m}_n^{\mathrm{pred}}(\theta)\right\|=o_P(n^{-1/2}). \tag{A.29}$$

Now let $\widehat{\theta}^{\mathrm{pred}}$ and $\widehat{\theta}^{IACV}$ denote the corresponding strong-identification GMM estimators, that is,

$$\widehat{\theta}^{a}\in\arg\min_{\theta\in\Theta}\widehat{m}_n^{a}(\theta)^{\top}\widehat{W}_n^{a}\widehat{m}_n^{a}(\theta),\qquad a\in\{\mathrm{pred},IACV\},$$

with $\widehat{W}_n^a$ the corresponding consistent weight matrix estimator.

The final sentence of Proposition 5 imposes the strong-identification regularity conditions of Theorem 5 for the selected moment problem. These conditions, not divergence of the strength proxy, deliver the local quadratic curvature required for regular GMM:

$$\lambda_{\min}\Big(G_P^{\top}\Omega_P(\theta_0)^{-1}G_P\Big)\ge c>0.$$

The strength proxy is used to define the identification-aware selection objective; it is not used here to infer strong identification.

Under the regularity conditions of Theorem 5, both selected empirical moment functions admit the same first-order expansion because

$$\sup_{\theta\in\Theta}\left\|\widehat{m}_n^{IACV}(\theta)-\widehat{m}_n^{\mathrm{pred}}(\theta)\right\|=o_P(n^{-1/2})$$

and both are oracle-equivalent to the same population moment. Applying the expansion in the proof of Theorem 5 to each selected GMM estimator gives

$$\sqrt{n}(\widehat{\theta}^{\mathrm{pred}}-\theta_0)=-\left(G_P^\top\Omega_P(\theta_0)^{-1}G_P\right)^{-1}G_P^\top\Omega_P(\theta_0)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n\psi(W_i;\theta_0,\eta_0)+o_P(1), \tag{A.30}$$

$$\sqrt{n}(\widehat{\theta}^{IACV}-\theta_0)=-\left(G_P^\top\Omega_P(\theta_0)^{-1}G_P\right)^{-1}G_P^\top\Omega_P(\theta_0)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n\psi(W_i;\theta_0,\eta_0)+o_P(1). \tag{A.31}$$

Subtracting (A.30) from (A.31) gives

$$\sqrt{n}\left(\widehat{\theta}^{IACV}-\widehat{\theta}^{\mathrm{pred}}\right)=o_P(1).$$

Hence the two strong-identification GMM estimators are asymptotically equivalent.

This completes the proof.

A.9 Proof of Theorem 3

For brevity, write

$$\psi_{0,i}:=\psi(W_i;\theta_0,\eta_0),\qquad \widehat{\psi}_i:=\psi(W_i;\theta_0,\widehat{\eta}_{-i}),$$

where, for $i\in I_k^{\mathrm{inf}}$, the notation $\widehat{\eta}_{-i}$ means the fold-k training output $\widehat{\eta}_k$. Also write

$$g_{n,0}:=g_{n,0}(\theta_0)=\frac{1}{\sqrt{n}}\sum_{i=1}^n\psi_{0,i},\qquad \widehat{g}_n:=\widehat{g}_n(\theta_0)=\frac{1}{\sqrt{n}}\sum_{i=1}^n\widehat{\psi}_i,$$

and define the oracle and feasible covariance matrices

$$\Omega_P:=\Omega_P(\theta_0)=E_P[\psi(W;\theta_0,\eta_0)\psi(W;\theta_0,\eta_0)^\top],$$

$$\Omega_{n,0} := \frac{1}{n}\sum_{i=1}^{n} \psi_{0,i}\psi_{0,i}^{\top}, \qquad \widehat{\Omega}_n := \frac{1}{n}\sum_{i=1}^{n} \widehat{\psi}_i\widehat{\psi}_i^{\top}.$$

The oracle Anderson–Rubin statistic is

$$AR_{n,0} := g_{n,0}^{\top}\Omega_P^{-1}g_{n,0},$$

and the feasible statistic is

$$AR_n(\theta_0) = \widehat{g}_n^{\top}\widehat{\Omega}_n^{-1}\widehat{g}_n.$$

The proof has four steps. First, the oracle score satisfies a uniform central limit theorem. Second, the feasible score and covariance matrix are shown to be asymptotically equivalent to their oracle counterparts. Third, the feasible statistic is shown to differ from the oracle statistic by $o_P(1)$. Fourth, the size statement follows by Slutsky and continuity of the χ^2 law.

Step 1: uniform CLT for the oracle score. Under the null,

$$E_P[\psi(W;\theta_0,\eta_0)] = 0 \qquad \text{for every } P \in \mathcal{P}_n$$

by the defining moment condition. By Assumption 2, there exists $q > 4$ such that

$$\sup_{P\in\mathcal{P}_n} E_P[\|\psi(W;\theta_0,\eta_0)\|^q] < \infty.$$

Let $a \in \mathbb{R}^{d_g}$ be any fixed nonzero vector and define

$$X_{i,P}(a) := a^{\top}\psi_{0,i}.$$

Then

$$E_P[X_{i,P}(a)] = 0, \qquad \mathrm{Var}_P(X_{i,P}(a)) = a^{\top}\Omega_P a.$$

By Assumption 2, the eigenvalues of Ω_P are bounded away from zero and infinity uniformly over $P \in \mathcal{P}_n$. Hence there exist constants $0 < c_\Omega \le C_\Omega < \infty$ such that

$$c_\Omega\|a\|^2 \le a^{\top}\Omega_P a \le C_\Omega\|a\|^2 \qquad \text{for all } P \in \mathcal{P}_n.$$

Moreover, Hölder's inequality gives

$$E_P|X_{i,P}(a)|^q \le \|a\|^q E_P[\|\psi_{0,i}\|^q] \le C_q\|a\|^q$$

uniformly over $P \in \mathcal{P}_n$. Therefore the Lyapunov ratio with exponent $\delta := q-2 > 2$ satisfies

$$\begin{aligned}
\frac{\sum_{i=1}^n E_P|X_{i,P}(a)|^{2+\delta}}{\Big(\sum_{i=1}^n \mathrm{Var}_P(X_{i,P}(a))\Big)^{1+\delta/2}} &= \frac{n\,E_P|X_{i,P}(a)|^q}{\Big(n\,a^\top\Omega_P a\Big)^{q/2}} \\
&\le \frac{n\,C_q\|a\|^q}{(n\,c_\Omega\|a\|^2)^{q/2}} = \frac{C_q}{c_\Omega^{q/2}}\,n^{1-q/2} \to 0
\end{aligned}$$

uniformly over $P \in \mathcal{P}_n$. By the Lyapunov central limit theorem,

$$a^\top g_{n,0} = \frac{1}{\sqrt{n}}\sum_{i=1}^n X_{i,P}(a) \Rightarrow N(0, a^\top\Omega_P a)$$

uniformly over $P \in \mathcal{P}_n$. Since d_g is fixed, the Cramér–Wold device yields

$$g_{n,0} \Rightarrow N(0,\Omega_P) \qquad \text{uniformly over } P \in \mathcal{P}_n. \tag{A.32}$$

Now premultiply by $\Omega_P^{-1/2}$. Because Ω_P is uniformly nonsingular, the map $x \mapsto \Omega_P^{-1/2}x$ is continuous uniformly over P, so

$$\Omega_P^{-1/2}g_{n,0} \Rightarrow N(0, I_{d_g}) \qquad \text{uniformly over } P \in \mathcal{P}_n.$$

Hence, by another application of the continuous mapping theorem,

$$AR_{n,0} = g_{n,0}^\top\Omega_P^{-1}g_{n,0} = \|\Omega_P^{-1/2}g_{n,0}\|^2 \Rightarrow \chi^2_{d_g} \qquad \text{uniformly over } P \in \mathcal{P}_n. \tag{A.33}$$

Step 2: feasible score and covariance are asymptotically equivalent to oracle objects. Assumption 5 gives

$$\widehat{g}_n - g_{n,0} = o_P(1) \qquad \text{uniformly over } P \in \mathcal{P}_n. \tag{A.34}$$

To control the covariance matrix, first note that

$$\widehat{\Omega}_n - \Omega_P = (\widehat{\Omega}_n - \Omega_{n,0}) + (\Omega_{n,0} - \Omega_P).$$

The second term is an ordinary sample-covariance fluctuation. Since d_g is fixed and Assumption 2 gives a uniform $q > 4$ moment bound,

$$\Omega_{n,0} - \Omega_P = o_P(1) \qquad \text{uniformly over } P \in \mathcal{P}_n. \tag{A.35}$$

Now consider the first term. Define

$$\Delta_{\psi,i} := \widehat{\psi}_i - \psi_{0,i}.$$

Then

$$\begin{aligned} \widehat{\psi}_i \widehat{\psi}_i^\top - \psi_{0,i}\psi_{0,i}^\top &= (\widehat{\psi}_i - \psi_{0,i})\widehat{\psi}_i^\top + \psi_{0,i}(\widehat{\psi}_i - \psi_{0,i})^\top \\ &= \Delta_{\psi,i}\widehat{\psi}_i^\top + \psi_{0,i}\Delta_{\psi,i}^\top. \end{aligned} \tag{A.36}$$

Since d_g is fixed, the operator norm is dominated by the Frobenius norm, so

$$\begin{aligned} \|\widehat{\Omega}_n - \Omega_{n,0}\|_{\mathrm{op}} &\leq \frac{1}{n}\sum_{i=1}^n \left\| \Delta_{\psi,i}\widehat{\psi}_i^\top + \psi_{0,i}\Delta_{\psi,i}^\top \right\|_{\mathrm{op}} \\ &\leq \frac{1}{n}\sum_{i=1}^n \|\Delta_{\psi,i}\| \, \|\widehat{\psi}_i\| + \frac{1}{n}\sum_{i=1}^n \|\psi_{0,i}\| \, \|\Delta_{\psi,i}\|. \end{aligned} \tag{A.37}$$

By Cauchy–Schwarz,

$$\frac{1}{n}\sum_{i=1}^n \|\Delta_{\psi,i}\| \, \|\widehat{\psi}_i\| \leq \left(\frac{1}{n}\sum_{i=1}^n \|\Delta_{\psi,i}\|^2\right)^{1/2} \left(\frac{1}{n}\sum_{i=1}^n \|\widehat{\psi}_i\|^2\right)^{1/2}, \tag{A.38}$$

$$\frac{1}{n}\sum_{i=1}^n \|\psi_{0,i}\| \, \|\Delta_{\psi,i}\| \leq \left(\frac{1}{n}\sum_{i=1}^n \|\psi_{0,i}\|^2\right)^{1/2} \left(\frac{1}{n}\sum_{i=1}^n \|\Delta_{\psi,i}\|^2\right)^{1/2}. \tag{A.39}$$

By the same orthogonality expansion and product-rate argument used to verify Assumption 5, the feasible score perturbation satisfies

$$\frac{1}{n}\sum_{i=1}^n \|\Delta_{\psi,i}\|^2 = o_P(1) \qquad \text{uniformly over } P \in \mathcal{P}_n. \tag{A.40}$$

Moreover,

$$\frac{1}{n}\sum_{i=1}^n \|\psi_{0,i}\|^2 = O_P(1) \qquad \text{and} \qquad \frac{1}{n}\sum_{i=1}^n \|\widehat{\psi}_i\|^2 = O_P(1)$$

uniformly over $P \in \mathcal{P}_n$: the first bound follows from the uniform moment assumption, and the second follows from

$$\|\widehat{\psi}_i\|^2 \le 2\|\psi_{0,i}\|^2 + 2\|\Delta_{\psi,i}\|^2$$

together with (A.40). Therefore

$$\widehat{\Omega}_n - \Omega_{n,0} = o_P(1) \qquad \text{uniformly over } P \in \mathcal{P}_n. \tag{A.41}$$

Combining (A.35) and (A.41),

$$\widehat{\Omega}_n - \Omega_P = o_P(1) \qquad \text{uniformly over } P \in \mathcal{P}_n. \tag{A.42}$$

Step 3: the feasible and oracle AR statistics differ by $o_P(1)$. Since the eigenvalues of Ω_P are uniformly bounded away from zero and $\widehat{\Omega}_n - \Omega_P = o_P(1)$, Weyl's inequality implies that $\widehat{\Omega}_n$ is invertible with probability tending to one uniformly over $P \in \mathcal{P}_n$, and

$$\|\widehat{\Omega}_n^{-1}\|_{\text{op}} = O_P(1) \qquad \text{uniformly over } P \in \mathcal{P}_n. \tag{A.43}$$

Moreover, using the resolvent identity,

$$\widehat{\Omega}_n^{-1} - \Omega_P^{-1} = -\Omega_P^{-1}(\widehat{\Omega}_n - \Omega_P)\widehat{\Omega}_n^{-1}, \tag{A.44}$$

so (A.42) and (A.43) imply

$$\widehat{\Omega}_n^{-1} - \Omega_P^{-1} = o_P(1) \qquad \text{uniformly over } P \in \mathcal{P}_n. \tag{A.45}$$

Now write

$$\Delta_{g,n} := \widehat{g}_n - g_{n,0}.$$

Then

$$\begin{aligned} AR_n(\theta_0) - AR_{n,0} &= \widehat{g}_n^\top \widehat{\Omega}_n^{-1} \widehat{g}_n - g_{n,0}^\top \Omega_P^{-1} g_{n,0} \\ &= (g_{n,0} + \Delta_{g,n})^\top \widehat{\Omega}_n^{-1} (g_{n,0} + \Delta_{g,n}) - g_{n,0}^\top \Omega_P^{-1} g_{n,0} \\ &= \Delta_{g,n}^\top \widehat{\Omega}_n^{-1} \Delta_{g,n} + 2 g_{n,0}^\top \widehat{\Omega}_n^{-1} \Delta_{g,n} + g_{n,0}^\top (\widehat{\Omega}_n^{-1} - \Omega_P^{-1}) g_{n,0}. \end{aligned} \tag{A.46}$$

By (A.34), $\Delta_{g,n} = o_P(1)$ uniformly, and by (A.43),

$$\Delta_{g,n}^\top \widehat{\Omega}_n^{-1} \Delta_{g,n} = o_P(1).$$

Also, (A.33) implies $AR_{n,0} = O_P(1)$ uniformly. Because the maximum eigenvalue of Ω_P is bounded above uniformly over $P \in \mathcal{P}_n$, there exists $c_\Omega > 0$ such that

$$\Omega_P^{-1} \succeq c_\Omega I$$

for all $P \in \mathcal{P}_n$. Consequently,

$$AR_{n,0} = g_{n,0}^\top \Omega_P^{-1} g_{n,0} \geq c_\Omega \|g_{n,0}\|^2.$$

Hence $\|g_{n,0}\|^2 \leq c_\Omega^{-1} AR_{n,0} = O_P(1)$ uniformly, and therefore

$$\|g_{n,0}\| = O_P(1) \qquad \text{uniformly over } P \in \mathcal{P}_n. \tag{A.47}$$

Using (A.47), (A.43), and $\Delta_{g,n} = o_P(1)$, we obtain

$$|2 g_{n,0}^\top \widehat{\Omega}_n^{-1} \Delta_{g,n}| \leq 2 \|g_{n,0}\| \, \|\widehat{\Omega}_n^{-1}\|_{\mathrm{op}} \, \|\Delta_{g,n}\| = o_P(1),$$

and by (A.45),

$$g_{n,0}^\top (\widehat{\Omega}_n^{-1} - \Omega_P^{-1}) g_{n,0} = o_P(1).$$

Returning to (A.46), we conclude

$$AR_n(\theta_0) - AR_{n,0} = o_P(1) \qquad \text{uniformly over } P \in \mathcal{P}_n. \tag{A.48}$$

Step 4: conclude uniform size. From (A.33) and (A.48), Slutsky's theorem gives

$$AR_n(\theta_0) \Rightarrow \chi^2_{d_g} \qquad \text{uniformly over } P \in \mathcal{P}_n.$$

The $\chi^2_{d_g}$ distribution is continuous, so $c_{1-\alpha} := \chi^2_{d_g, 1-\alpha}$ is a continuity point. Hence

$$\sup_{P \in \mathcal{P}_n} \Big| P_P\Big(AR_n(\theta_0) > c_{1-\alpha}\Big) - \alpha \Big| \to 0.$$

This proves the size statement.

The coverage statement follows by inversion. Define

$$\mathcal{C}_{1-\alpha}^{AR} := \{\theta \in \Theta : AR_n(\theta) \leq \chi^2_{d_g,1-\alpha}\}.$$

Since $\theta_0 \in \mathcal{C}_{1-\alpha}^{AR}$ if and only if $AR_n(\theta_0) \leq \chi^2_{d_g,1-\alpha}$,

$$P_P\Big(\theta_0 \in \mathcal{C}_{1-\alpha}^{AR}\Big) = 1 - P_P\Big(AR_n(\theta_0) > \chi^2_{d_g,1-\alpha}\Big).$$

Therefore

$$\sup_{P \in \mathcal{P}_n} \Big| P_P\Big(\theta_0 \in \mathcal{C}_{1-\alpha}^{AR}\Big) - (1-\alpha) \Big| \to 0.$$

This proves the theorem.

A.10 Proof of Theorem 4

Write

$$A_n := \Omega_P(\theta_0) + \rho_n I_{d_{g,n}}, \qquad \widehat{A}_n := \widehat{\Omega}_n(\theta_0) + \rho_n I_{d_{g,n}}.$$

Define the oracle and feasible ridge statistics

$$Q_{0n} := g_{n,0}(\theta_0)^\top A_n^{-1} g_{n,0}(\theta_0), \qquad Q_n := \widehat{g}_n(\theta_0)^\top \widehat{A}_n^{-1} \widehat{g}_n(\theta_0) = AR_{n,\rho_n}(\theta_0),$$

and the corresponding multiplier-bootstrap analogues

$$Q^*_{0n} := g^*_{n,0}(\theta_0)^\top A_n^{-1} g^*_{n,0}(\theta_0), \qquad Q^*_n := \widehat{g}^*_n(\theta_0)^\top \widehat{A}_n^{-1} \widehat{g}^*_n(\theta_0) = AR^*_{n,\rho_n}(\theta_0).$$

Let

$$F_n(t) := P_P(Q_n \leq t), \qquad F_{0n}(t) := P_P(Q_{0n} \leq t),$$

and

$$G_n(t) := P^*_P(Q^*_n \leq t \mid \mathcal{G}_n), \qquad G_{0n}(t) := P^*_P(Q^*_{0n} \leq t \mid \mathcal{G}_n).$$

Let

$$c^*_{1-\alpha} := \inf\{t \in \mathbb{R} : G_n(t) \geq 1 - \alpha\}$$

denote the conditional $(1-\alpha)$ quantile of the feasible bootstrap statistic.

The proof has three steps. First, the feasible and oracle ridge statistics are shown to be asymptotically equivalent, both for the statistic and for its bootstrap analogue.

Second, the distribution function of the feasible statistic is shown to be uniformly close to the conditional distribution function of the feasible bootstrap statistic. Third, anti-concentration is used to transfer this cdf approximation into one-sided uniform size control for the random critical value.

Step 1: feasible-to-oracle reduction. To simplify notation, write

$$g_n := g_{n,0}(\theta_0), \qquad \widehat{g}_n := \widehat{g}_n(\theta_0), \qquad \delta_n := \widehat{g}_n - g_n,$$

and

$$g_n^* := g_{n,0}^*(\theta_0), \qquad \widehat{g}_n^* := \widehat{g}_n^*(\theta_0), \qquad \delta_n^* := \widehat{g}_n^* - g_n^*.$$

Also define the rescaled covariance perturbation

$$E_n := A_n^{-1/2}\Big(\widehat{\Omega}_n(\theta_0) - \Omega_P(\theta_0)\Big)A_n^{-1/2}.$$

Then

$$\widehat{A}_n = A_n^{1/2}(I + E_n)A_n^{1/2}.$$

By Assumption 8(ii), A_n is uniformly positive definite. By Assumption 8(iv),

$$\|E_n\|_{\mathrm{op}} = o_P(1).$$

Hence on an event of probability tending to one,

$$\widehat{A}_n^{-1} = A_n^{-1/2}(I + E_n)^{-1}A_n^{-1/2},$$

and moreover

$$\|(I + E_n)^{-1}\|_{\mathrm{op}} \le 2, \qquad \|(I + E_n)^{-1} - I\|_{\mathrm{op}} \le 2\|E_n\|_{\mathrm{op}}. \tag{A.49}$$

Now expand the statistic:

$$\begin{aligned} Q_n - Q_{0n} &= (\widehat{g}_n - g_n)^\top \widehat{A}_n^{-1}(\widehat{g}_n - g_n) + 2g_n^\top \widehat{A}_n^{-1}(\widehat{g}_n - g_n) + g_n^\top(\widehat{A}_n^{-1} - A_n^{-1})g_n \\ &= \delta_n^\top \widehat{A}_n^{-1}\delta_n + 2g_n^\top \widehat{A}_n^{-1}\delta_n + g_n^\top(\widehat{A}_n^{-1} - A_n^{-1})g_n. \end{aligned} \tag{A.50}$$

For the first term,

$$\delta_n^\top \widehat{A}_n^{-1}\delta_n = (A_n^{-1/2}\delta_n)^\top (I+E_n)^{-1}(A_n^{-1/2}\delta_n),$$

so by (A.49),

$$\delta_n^\top \widehat{A}_n^{-1}\delta_n \le 2\|A_n^{-1/2}\delta_n\|^2.$$

Assumption 8(iv) gives

$$\|A_n^{-1/2}\delta_n\| = o_P\Big(r_{\mathrm{eff}}(\rho_n)^{-1/2}\Big),$$

hence in particular

$$\delta_n^\top \widehat{A}_n^{-1}\delta_n = o_P(1). \tag{A.51}$$

For the second term,

$$2g_n^\top \widehat{A}_n^{-1}\delta_n = 2(A_n^{-1/2}g_n)^\top (I+E_n)^{-1}(A_n^{-1/2}\delta_n).$$

Therefore

$$|2g_n^\top \widehat{A}_n^{-1}\delta_n| \le 2\|(I+E_n)^{-1}\|_{\mathrm{op}}\,\|A_n^{-1/2}g_n\|\,\|A_n^{-1/2}\delta_n\|.$$

By (A.49),

$$|2g_n^\top \widehat{A}_n^{-1}\delta_n| \le 4\|A_n^{-1/2}g_n\|\,\|A_n^{-1/2}\delta_n\|.$$

Now

$$\|A_n^{-1/2}g_n\|^2 = Q_{0n},$$

and

$$E_P[Q_{0n}] = \mathrm{tr}\Big(\Omega_P(\theta_0)A_n^{-1}\Big) = r_{\mathrm{eff}}(\rho_n).$$

Hence

$$Q_{0n} = O_P(r_{\mathrm{eff}}(\rho_n)), \qquad \|A_n^{-1/2}g_n\| = O_P\Big(r_{\mathrm{eff}}(\rho_n)^{1/2}\Big).$$

Combined with Assumption 8(iv),

$$\|A_n^{-1/2}\delta_n\| = o_P\Big(r_{\mathrm{eff}}(\rho_n)^{-1/2}\Big),$$

this yields

$$2g_n^\top \widehat{A}_n^{-1}\delta_n = o_P(1). \tag{A.52}$$

For the third term, use

$$\widehat{A}_n^{-1} - A_n^{-1} = A_n^{-1/2}\Big((I+E_n)^{-1} - I\Big)A_n^{-1/2}.$$

Thus

$$\begin{aligned} |g_n^\top(\widehat{A}_n^{-1} - A_n^{-1})g_n| &= \Big|(A_n^{-1/2}g_n)^\top\Big((I+E_n)^{-1} - I\Big)(A_n^{-1/2}g_n)\Big| \\ &\le \|A_n^{-1/2}g_n\|^2\,\|(I+E_n)^{-1} - I\|_{\mathrm{op}} \\ &\le 2\|A_n^{-1/2}g_n\|^2\,\|E_n\|_{\mathrm{op}}. \end{aligned}$$

Since $\|A_n^{-1/2}g_n\|^2 = O_P(r_{\mathrm{eff}}(\rho_n))$ and Assumption 8(iv) gives

$$\|E_n\|_{\mathrm{op}} = o_P\Big(r_{\mathrm{eff}}(\rho_n)^{-1}\Big),$$

it follows that

$$g_n^\top(\widehat{A}_n^{-1} - A_n^{-1})g_n = o_P(1). \tag{A.53}$$

Combining (A.51), (A.52), and (A.53) in (A.50), we obtain

$$Q_n - Q_{0n} = o_P(1). \tag{A.54}$$

Exactly the same argument applies to the bootstrap analogue:

$$Q_n^* - Q_{0n}^* = \delta_n^{*\top}\widehat{A}_n^{-1}\delta_n^* + 2g_n^{*\top}\widehat{A}_n^{-1}\delta_n^* + g_n^{*\top}(\widehat{A}_n^{-1} - A_n^{-1})g_n^*. \tag{A.55}$$

Conditional on $\mathcal{G}_n$, one has

$$E_P^*\Big[\|A_n^{-1/2}g_n^*\|^2 \mid \mathcal{G}_n\Big] = \mathrm{tr}(\Omega_{n,0}A_n^{-1}) = r_{\mathrm{eff}}(\rho_n) + o_P(1),$$

so

$$\|A_n^{-1/2}g_n^*\| = O_{P^*}\Big(r_{\mathrm{eff}}(\rho_n)^{1/2}\Big) \qquad \text{in } P\text{-probability.}$$

By the bootstrap analogue added to Assumption 8(iv),

$$\|A_n^{-1/2}\delta_n^*\| = o_{P^*}\Big(r_{\mathrm{eff}}(\rho_n)^{-1/2}\Big) \qquad \text{in } P\text{-probability.}$$

Therefore
$$Q_n^* - Q_{0n}^* = o_{P^*}(1) \qquad \text{in } P\text{-probability}. \tag{A.56}$$

Step 2: cdf approximation for the feasible statistic and feasible bootstrap. Choose a deterministic sequence $\eta_n \downarrow 0$ such that

$$P_P(|Q_n - Q_{0n}| > \eta_n) \to 0, \qquad P_P^*(|Q_n^* - Q_{0n}^*| > \eta_n \mid \mathcal{G}_n) \to_P 0. \tag{A.57}$$

Such a sequence exists by a standard diagonal argument applied to (A.54) and (A.56).

By Lemma OA.1,

$$d_{0n} := \sup_{t\in\mathbb{R}} |F_{0n}(t) - G_{0n}(t)| \to_P 0 \qquad \text{uniformly over } P \in \mathcal{P}_n. \tag{A.58}$$

By Lemma OA.2, for the deterministic sequence $\eta_n \downarrow 0$,

$$\omega_n := \sup_{t\in\mathbb{R}} \Big(G_{0n}(t+\eta_n) - G_{0n}(t-\eta_n)\Big) \to_P 0 \qquad \text{uniformly over } P \in \mathcal{P}_n. \tag{A.59}$$

Now fix $t \in \mathbb{R}$. Since

$$\{Q_n \le t\} \subseteq \{Q_{0n} \le t + \eta_n\} \cup \{|Q_n - Q_{0n}| > \eta_n\},$$

one has
$$F_n(t) \le F_{0n}(t+\eta_n) + P_P(|Q_n - Q_{0n}| > \eta_n).$$

Using (A.58),
$$F_{0n}(t+\eta_n) \le G_{0n}(t+\eta_n) + d_{0n},$$

and then
$$G_{0n}(t+\eta_n) \le G_{0n}(t) + \omega_n.$$

Thus
$$F_n(t) \le G_{0n}(t) + \omega_n + d_{0n} + P_P(|Q_n - Q_{0n}| > \eta_n).$$

Similarly,
$$\{Q_{0n} \le t - \eta_n\} \subseteq \{Q_n \le t\} \cup \{|Q_n - Q_{0n}| > \eta_n\},$$

so

$$F_n(t) \geq F_{0n}(t-\eta_n) - P_P(|Q_n - Q_{0n}| > \eta_n) \geq G_{0n}(t) - \omega_n - d_{0n} - P_P(|Q_n - Q_{0n}| > \eta_n).$$

Therefore

$$\sup_{t \in \mathbb{R}} |F_n(t) - G_{0n}(t)| \leq \omega_n + d_{0n} + P_P(|Q_n - Q_{0n}| > \eta_n). \tag{A.60}$$

Apply the same argument conditionally to the bootstrap statistics. Since

$$\{Q_n^* \leq t\} \subseteq \{Q_{0n}^* \leq t + \eta_n\} \cup \{|Q_n^* - Q_{0n}^*| > \eta_n\},$$

and similarly for the lower bound, we obtain

$$\sup_{t \in \mathbb{R}} |G_n(t) - G_{0n}(t)| \leq \omega_n + P_P^*(|Q_n^* - Q_{0n}^*| > \eta_n \mid \mathcal{G}_n) \to_P 0. \tag{A.61}$$

Combining (A.60), (A.61), and (A.57),

$$\sup_{t \in \mathbb{R}} |F_n(t) - G_n(t)| \leq \sup_t |F_n(t) - G_{0n}(t)| + \sup_t |G_n(t) - G_{0n}(t)| \to_P 0.$$

Hence

$$d_n := \sup_{t \in \mathbb{R}} |F_n(t) - G_n(t)| \to_P 0 \qquad \text{uniformly over } P \in \mathcal{P}_n. \tag{A.62}$$

Step 3: size from the random critical value. Let

$$F_n(t) := \mathbb{P}_P(Q_n \leq t), \qquad G_n(t) := \mathbb{P}_P^*(Q_n^* \leq t \mid \mathcal{G}_n),$$

and let

$$c_{n,1-\alpha} := \inf\{t \in \mathbb{R} : F_n(t) \geq 1 - \alpha\}.$$

By (A.62),

$$d_n := \sup_{t \in \mathbb{R}} |G_n(t) - F_n(t)| = o_{\mathbb{P}}(1)$$

uniformly over $P \in \mathcal{P}_n$.

Choose a deterministic sequence $a_n \downarrow 0$ such that the quantile-separation quantity

$$\tau_n := 1 - \alpha - F_n(c_{n,1-\alpha} - a_n)$$

satisfies

$$\tau_n > 0, \qquad d_n/\tau_n = o_{\mathbb{P}}(1)$$

uniformly over $P \in \mathcal{P}_n$, and such that the local oscillation modulus

$$\omega_n(a_n) := F_n(c_{n,1-\alpha}) - F_n(c_{n,1-\alpha} - a_n)$$

satisfies

$$\sup_{P \in \mathcal{P}_n} \omega_n(a_n) \to 0.$$

The existence of such a sequence follows from the anti-concentration and continuity bounds for the feasible Gaussian quadratic-form law established in the preceding steps and Online Appendix Lemmas OA.1–OA.2.

We first show that the random bootstrap critical value is not too small. Let

$$E_n := \{d_n < \tau_n\}.$$

On E_n,

$$G_n(c_{n,1-\alpha} - a_n) \leq F_n(c_{n,1-\alpha} - a_n) + d_n = 1 - \alpha - \tau_n + d_n < 1 - \alpha.$$

Because

$$c^*_{1-\alpha} := \inf\{t : G_n(t) \geq 1 - \alpha\},$$

the last display implies

$$c^*_{1-\alpha} \geq c_{n,1-\alpha} - a_n$$

on E_n. Since $d_n/\tau_n = o_{\mathbb{P}}(1)$, we have

$$\mathbb{P}_P(E_n) \to 1$$

uniformly over $P \in \mathcal{P}_n$. Therefore

$$\mathbb{P}_P\Big(c^*_{1-\alpha} \geq c_{n,1-\alpha} - a_n\Big) \to 1$$

uniformly over $P \in \mathcal{P}_n$.

Now use this deterministic lower bound for the random critical value. Since

$$\{Q_n > c^*_{1-\alpha}\} \subseteq \{Q_n > c_{n,1-\alpha} - a_n\} \cup E_n^c,$$

we obtain

$$\mathbb{P}_P(Q_n > c^*_{1-\alpha}) \leq \mathbb{P}_P(Q_n > c_{n,1-\alpha} - a_n) + \mathbb{P}_P(E_n^c).$$

The second term is $o(1)$ uniformly. For the first term,

$$\mathbb{P}_P(Q_n > c_{n,1-\alpha} - a_n) = 1 - F_n(c_{n,1-\alpha} - a_n).$$

Because $F_n(c_{n,1-\alpha}) \geq 1 - \alpha$,

$$1 - F_n(c_{n,1-\alpha} - a_n) \leq \alpha + \Big[F_n(c_{n,1-\alpha}) - F_n(c_{n,1-\alpha} - a_n)\Big] \leq \alpha + \omega_n(a_n).$$

By construction, $\omega_n(a_n) = o(1)$ uniformly. Hence

$$\limsup_{n\to\infty} \sup_{P\in\mathcal{P}_n} \mathbb{P}_P(Q_n > c^*_{1-\alpha}) \leq \alpha.$$

Since $Q_n = AR_{n,\rho_n}(\theta_0)$, this proves the stated one-sided uniform size control.

A.11 Proof of Theorem 5

Define the population moment

$$m_P(\theta) := E_P[\psi(W;\theta,\eta_0)],$$

and let

$$G_P(\theta) := \partial_\theta m_P(\theta), \qquad G_P := G_P(\theta_0).$$

Write

$$\widehat{m}_n(\theta) := \frac{1}{n}\sum_{i=1}^n \widehat{\psi}_i(\theta), \qquad m_{n,0}(\theta) := \frac{1}{n}\sum_{i=1}^n \psi(W_i;\theta,\eta_0),$$

so that

$$\widehat{g}_n(\theta) = \sqrt{n}\,\widehat{m}_n(\theta), \qquad g_{n,0}(\theta) = \sqrt{n}\,m_{n,0}(\theta).$$

Let

$$\Omega_P := \Omega_P(\theta_0), \qquad W_P := \Omega_P^{-1},$$

and define the feasible criterion

$$Q_n(\theta) := \widehat{m}_n(\theta)^\top \widehat{W}_n \, \widehat{m}_n(\theta), \qquad \widehat{W}_n := \widehat{\Omega}_n(\tilde{\theta})^{-1}.$$

The corresponding population criterion is

$$Q_P(\theta) := m_P(\theta)^\top W_P \, m_P(\theta).$$

The proof has four steps. First, the population criterion is shown to admit a locally quadratic expansion with unique minimizer θ_0. Second, the feasible criterion is shown to converge uniformly to the population criterion on a neighborhood of θ_0, which yields consistency of $\widehat{\theta}$. Third, the first-order condition is expanded to obtain an asymptotic linear representation. Fourth, asymptotic normality and efficiency follow.

Step 1: local quadratic identification of the population criterion. By the assumptions of Theorem 5, $m_P(\theta_0) = 0$, θ_0 is an interior point of Θ, and m_P is continuously differentiable on a neighborhood $\mathcal{N}$ of θ_0. Thus, for $\theta \in \mathcal{N}$,

$$m_P(\theta) = G_P(\theta - \theta_0) + r_P(\theta), \qquad \frac{\|r_P(\theta)\|}{\|\theta - \theta_0\|} \to 0 \quad \text{as } \theta \to \theta_0. \tag{A.63}$$

Since G_P has full column rank and $W_P = \Omega_P^{-1}$ is positive definite,

$$M_P := G_P^\top W_P G_P$$

is positive definite. Let $\lambda_{\min}(M_P) > 0$ denote its smallest eigenvalue. Using (A.63),

$$\begin{aligned} Q_P(\theta) &= \Big(G_P(\theta - \theta_0) + r_P(\theta)\Big)^\top W_P \Big(G_P(\theta - \theta_0) + r_P(\theta)\Big) \\ &= (\theta - \theta_0)^\top M_P(\theta - \theta_0) + 2(\theta - \theta_0)^\top G_P^\top W_P r_P(\theta) + r_P(\theta)^\top W_P r_P(\theta). \end{aligned} \tag{A.64}$$

Because W_P is bounded and $r_P(\theta) = o(\|\theta - \theta_0\|)$, the last two terms in (A.64) are $o(\|\theta - \theta_0\|^2)$. Hence

$$Q_P(\theta) = (\theta - \theta_0)^\top M_P(\theta - \theta_0) + o(\|\theta - \theta_0\|^2) \qquad \text{as } \theta \to \theta_0. \tag{A.65}$$

Since M_P is positive definite, there exists a neighborhood $\mathcal{N}_0 \subseteq \mathcal{N}$ and a constant

$c_0 > 0$ such that

$$Q_P(\theta) \geq c_0 \|\theta - \theta_0\|^2 \qquad \text{for every } \theta \in \mathcal{N}_0. \tag{A.66}$$

Therefore Q_P has a unique minimizer at θ_0 on $\mathcal{N}_0$.

Step 2: uniform convergence of the feasible criterion and consistency of $\widehat{\theta}$. The theorem assumptions imply $\tilde{\theta} \to_P \theta_0$ and

$$\widehat{\Omega}_n(\tilde{\theta}) \to_P \Omega_P(\theta_0).$$

Since $\Omega_P(\theta_0)$ is nonsingular, continuity of matrix inversion gives

$$\widehat{W}_n \to_P W_P. \tag{A.67}$$

Next, by Assumption 5,

$$\sup_{\theta \in \mathcal{N}_0} \sqrt{n}\, \|\widehat{m}_n(\theta) - m_{n,0}(\theta)\| = o_P(1).$$

Since $m_{n,0}(\theta)$ is an empirical average indexed by a fixed-dimensional parameter θ over a neighborhood $\mathcal{N}_0$, and the theorem assumptions impose local continuity and finite moments for the oracle score, the corresponding class is Glivenko–Cantelli on $\mathcal{N}_0$. Therefore

$$\sup_{\theta \in \mathcal{N}_0} \|m_{n,0}(\theta) - m_P(\theta)\| = o_P(1).$$

Therefore

$$\sup_{\theta \in \mathcal{N}_0} \|\widehat{m}_n(\theta) - m_P(\theta)\| = o_P(1). \tag{A.68}$$

Now decompose

$$\begin{aligned} Q_n(\theta) - Q_P(\theta) = {} & \widehat{m}_n(\theta)^\top (\widehat{W}_n - W_P) \widehat{m}_n(\theta) \\ & + \Big(\widehat{m}_n(\theta) - m_P(\theta)\Big)^\top W_P \widehat{m}_n(\theta) + m_P(\theta)^\top W_P \Big(\widehat{m}_n(\theta) - m_P(\theta)\Big). \end{aligned} \tag{A.69}$$

Because m_P is continuous on the fixed neighborhood $\mathcal{N}_0$, one has

$$\sup_{\theta \in \mathcal{N}_0} \|m_P(\theta)\| < \infty.$$

Together with (A.68), this implies

$$\sup_{\theta \in \mathcal{N}_0} \|\widehat{m}_n(\theta)\| = O_P(1).$$

Hence (A.67), (A.68), and (A.69) yield

$$\sup_{\theta \in \mathcal{N}_0} |Q_n(\theta) - Q_P(\theta)| = o_P(1). \tag{A.70}$$

To prove consistency, fix $\varepsilon > 0$ small enough that the closed annulus

$$B_\varepsilon := \{\theta \in \mathcal{N}_0 : \|\theta - \theta_0\| \geq \varepsilon\}$$

is nonempty. By (A.66),

$$\inf_{\theta \in B_\varepsilon} Q_P(\theta) \geq c_0 \varepsilon^2.$$

Also $Q_P(\theta_0) = 0$, and by (A.70),

$$Q_n(\theta_0) = o_P(1), \qquad \sup_{\theta \in \mathcal{N}_0} |Q_n(\theta) - Q_P(\theta)| = o_P(1).$$

By the added global separation condition and the uniform convergence

$$\sup_{\theta \in \Theta \setminus \mathcal{N}_0} |Q_n(\theta) - Q_P(\theta)| = o_P(1),$$

one also has, with probability tending to one,

$$\inf_{\theta \in \Theta \setminus \mathcal{N}_0} Q_n(\theta) > Q_n(\theta_0).$$

Hence any global minimizer $\widehat{\theta}$ must lie in $\mathcal{N}_0 \setminus B_\varepsilon$ with probability tending to one, which proves $\widehat{\theta} \to_P \theta_0$.

Step 3: asymptotic linear representation. Let

$$\widehat{G}_n(\theta) := \partial_\theta \widehat{m}_n(\theta).$$

Because $\widehat{W}_n$ does not depend on θ, the first-order condition for the GMM minimizer $\widehat{\theta}$ is

$$0 = \widehat{G}_n(\widehat{\theta})^\top \widehat{W}_n \widehat{m}_n(\widehat{\theta}). \tag{A.71}$$

By the integral form of the mean-value expansion,

$$\widehat{m}_n(\widehat{\theta}) = \widehat{m}_n(\theta_0) + \overline{G}_n(\widehat{\theta} - \theta_0),$$

where

$$\overline{G}_n := \int_0^1 \widehat{G}_n\Big(\theta_0 + t(\widehat{\theta} - \theta_0)\Big)\, dt.$$

Substituting into (A.71),

$$0 = \widehat{G}_n(\widehat{\theta})^\top \widehat{W}_n \widehat{m}_n(\theta_0) + \widehat{G}_n(\widehat{\theta})^\top \widehat{W}_n \overline{G}_n(\widehat{\theta} - \theta_0).$$

Thus

$$\sqrt{n}(\widehat{\theta} - \theta_0) = -\Big(\widehat{G}_n(\widehat{\theta})^\top \widehat{W}_n \overline{G}_n\Big)^{-1} \widehat{G}_n(\widehat{\theta})^\top \widehat{W}_n \sqrt{n}\, \widehat{m}_n(\theta_0), \tag{A.72}$$

provided the displayed matrix is invertible.

Now use the theorem's differentiability and stochastic equicontinuity assumptions. Since $\widehat{\theta} \to_P \theta_0$,

$$\widehat{G}_n(\widehat{\theta}) - G_P = o_P(1), \qquad \overline{G}_n - G_P = o_P(1).$$

Together with (A.67),

$$\widehat{G}_n(\widehat{\theta})^\top \widehat{W}_n \overline{G}_n = G_P^\top W_P G_P + o_P(1) = M_P + o_P(1).$$

Since M_P is positive definite, it is invertible, and therefore

$$\Big(\widehat{G}_n(\widehat{\theta})^\top \widehat{W}_n \overline{G}_n\Big)^{-1} = M_P^{-1} + o_P(1). \tag{A.73}$$

Next,

$$\sqrt{n}\, \widehat{m}_n(\theta_0) = \widehat{g}_n(\theta_0) = g_{n,0}(\theta_0) + o_P(1) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi(W_i; \theta_0, \eta_0) + o_P(1),$$

where the first equality is the definition of $\widehat{g}_n$, the second is Assumption 5, and the third is the definition of $g_{n,0}$. Substituting this and (A.73) into (A.72) yields

$$\sqrt{n}(\widehat{\theta}-\theta_0) = -M_P^{-1}G_P^\top W_P \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi(W_i;\theta_0,\eta_0) + o_P(1). \tag{A.74}$$

Step 4: asymptotic normality and efficiency. By Assumption 2, the oracle score has a uniform $q > 4$ moment bound, so the multivariate central limit theorem implies

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi(W_i;\theta_0,\eta_0) \Rightarrow N(0,\Omega_P).$$

Applying (A.74) and Slutsky's theorem gives

$$\sqrt{n}(\widehat{\theta}-\theta_0) \Rightarrow N\Big(0,\ M_P^{-1}G_P^\top W_P\Omega_P W_P G_P M_P^{-1}\Big).$$

Since $W_P = \Omega_P^{-1}$,

$$M_P^{-1}G_P^\top W_P\Omega_P W_P G_P M_P^{-1} = M_P^{-1}G_P^\top\Omega_P^{-1}G_P M_P^{-1} = M_P^{-1}.$$

Therefore

$$\sqrt{n}(\widehat{\theta}-\theta_0) \Rightarrow N\Big(0,\ (G_P^\top\Omega_P^{-1}G_P)^{-1}\Big).$$

This proves the asymptotic normality claim.

Finally, if $\psi(W;\theta,\eta_0)$ is chosen to equal the semiparametrically efficient influence function for θ, then the variance matrix

$$(G_P^\top\Omega_P^{-1}G_P)^{-1}$$

is the semiparametric efficiency bound. Hence $\widehat{\theta}$ attains that bound.

This completes the proof.

Declaration of generative AI and AI-assisted technologies in the writing process

The author used AI-assisted tools for language editing, structure of the paper, LaTeX revision, and Python coding support during the preparation of this paper. The author reviewed and edited all content and takes full responsibility for the manuscript.

References

Anderson, T. W. and H. Rubin (1949). Estimation of the parameters of a single equation in a complete system of stochastic equations. *Annals of Mathematical Statistics* 20(1), 46–63.

Andrews, D. W. K., M. J. Moreira, and J. H. Stock (2006). Optimal two-sided invariant similar tests for instrumental variables regression. *Econometrica* 74(3), 715–752.

Andrews, D. W. K. and X. Cheng (2012). Estimation and Inference with Weak, Semi-strong, and Strong Identification. *Econometrica* 80(5), 2153–2211.

Andrews, I. and A. Mikusheva (2016). Conditional inference with a functional nuisance parameter. *Econometrica* 84(4), 1571–1612.

Andrews, D. W. K., X. Cheng, and P. Guggenberger (2020). Generic Results for Establishing the Asymptotic Size of Confidence Sets and Tests. *Journal of Econometrics* 218(2), 496–531.

Andrews, I. and A. Mikusheva (2022a). Optimal Decision Rules for Weak GMM. *Econometrica* 90(2), 715–748.

Andrews, I. and A. Mikusheva (2022b). GMM is Inadmissible Under Weak Identification. arXiv preprint arXiv:2204.12462.

Andrews, D. W. and P. Guggenberger (2019). Identification- and singularity-robust inference for moment condition models. *Quantitative Economics* 10, 1703–1746.

Andrews, I., J. H. Stock, and L. Sun (2019). Weak instruments in instrumental variables regression: Theory and practice. *Annual Review of Economics* 11, 727–753.

Angrist, J. D. and A. B. Krueger (1991). Does Compulsory School Attendance Affect Schooling and Earnings? *Quarterly Journal of Economics* 106(4), 979–1014.

Angrist, J. D. (n.d.). *Angrist Data Archive*. MIT Economics. `https://economics.mit.edu/people/faculty/josh-angrist/angrist-data-archive`. Accessed March 2026.

Belloni, A., V. Chernozhukov, and K. Kato (2015). Uniform post-selection inference for least absolute deviation regression and other Z-estimation problems. *Biometrika* 102, 77–94.

Belloni, A., V. Chernozhukov, I. Fernandez-Val, and C. Hansen (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica* 85, 233–298.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica* 80(6), 2369–2429.

Belloni, A., V. Chernozhukov, and C. Hansen (2014). Inference on Treatment Effects after Selection among High-Dimensional Controls. *Review of Economic Studies* 81(2), 608–650.

Bia, M., M. Huber, and L. Lafférs (2024). Double machine learning for sample selection models. *Journal of Business & Economic Statistics* 42(3), 958–969.

Baiardi, A., P. S. Clarke, A. A. Naghi, and A. Polselli (2026). Double machine learning for static panel data with instrumental variables: new method and applications. *arXiv* preprint arXiv:2603.20464.

Bühlmann, P. and S. van de Geer (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer.

Çetin, T. (2026a). Nested cross-validated double machine learning for the partially linear IV model. Working paper.

Çetin, T. (2026b). Adaptive double machine learning via Riesz-risk cross-validation. Working paper.

Chao, J. C. and N. R. Swanson (2005). Consistent estimation with a large number of weak instruments. *Econometrica* 73(5), 1673–1702.

Chao, J. C., N. R. Swanson, J. A. Hausman, W. K. Newey, and T. Woutersen (2012). Asymptotic distribution of the jackknife instrumental variables estimator in a heteroskedastic instrumental variables regression with many instruments. *Econometric Theory* 28(1), 42–86.

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(1), C1–C68.

Bound, J., D. A. Jaeger, and R. M. Baker (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association* 90, 443–450.

Chernozhukov, V., J. C. Escanciano, H. Ichimura, W. K. Newey, and J. M. Robins (2022). Locally robust semiparametric estimation. *Econometrica* 90, 1501–1535.

Chen, J.-E., C.-H. Huang, and J.-J. Tien (2021). Debiased / double machine learning for instrumental variable quantile regressions. *Econometrics* 9(2), 15.

DoubleML Developers (2026). *Python: Confidence Intervals for Instrumental Variables Models That Are Robust to Weak Instruments*. DoubleML documentation example page. `https://docs.doubleml.org/stable/examples/py_double_ml_robust_iv.html`. Accessed March 2026.

Dufour, J.-M. (2003). Identification, weak instruments, and statistical inference in econometrics. *Canadian Journal of Economics* 36, 767–808.

ivmodels Developers (n.d.). *Does Compulsory School Attendance Affect Schooling and Earnings?* ivmodels documentation example page. Replication note based on the official MIT AK91 archive. `https://ivmodels.readthedocs.io/en/latest/examples/angrist1991does.html`. Accessed March 2026.

Keane, M. P. and T. Neal (2024). A practical guide to weak instruments. *Annual Review of Economics* 16, 185–212.

Kennedy, E. H., S. Balakrishnan, and M. G'Sell (2020). Sharp instruments for classifying compliers and generalizing causal effects. *The Annals of Statistics* 48, 2008–2030.

Kleibergen, F. (2005). Testing parameters in GMM without assuming that they are identified. *Econometrica* 73, 1103–1123.

Kleibergen, F. (2002). Pivotal statistics for testing structural parameters in instrumental variables regression. *Econometrica* 70(5), 1781–1803.

Lee, D. S., J. McCrary, M. J. Moreira, and J. Porter (2022). Valid t-ratio inference for IV. *American Economic Review* 112, 3260–3290.

Ma, Y. (2025). Identification-robust inference for the LATE with high-dimensional covariates. arXiv preprint arXiv:2302.09756, revised November 2025.

Matsushita, Y. and T. Otsu (2024). A jackknife Lagrange multiplier test with many weak instruments. *Econometric Theory* 40(2), 447–470.

Mikusheva, A. and L. Sun (2022). Inference with Many Weak Instruments. *The Review of Economic Studies* 89(5), 2663–2686.

Mikusheva, A. and L. Sun (2024). Weak identification with many instruments. *The Econometrics Journal* 27(2), C1–C28.

Montiel Olea, J. L. and C. Pflueger (2013). A Robust Test for Weak Instruments. *Journal of Business & Economic Statistics* 31(3), 358–369.

Moreira, M. J. (2003). A conditional likelihood ratio test for structural models. *Econometrica* 71(4), 1027–1048.

Moreira, M. J. (2009). Tests with correct size when instruments can be arbitrarily weak. *Journal of Econometrics* 152, 131–140.

Pouzo, D. (2015). Bootstrap Consistency for Quadratic Forms of Sample Averages with Increasing Dimension. *Electronic Journal of Statistics* 9, 1273–1307.

Scheidegger, C., Z. Guo, and P. Bühlmann (2026). Inference for heterogeneous treatment effects with efficient instruments and machine learning. *Electronic Journal of Statistics* 20(1), 718–770.

Smucler, E., L. Lanni, and D. Masip (2025). A note on the properties of the confidence set for the local average treatment effect obtained by inverting the score test. *arXiv* preprint arXiv:2506.10449.

Staiger, D. and J. H. Stock (1997). Instrumental variables regression with weak instruments. *Econometrica* 65(3), 557–586.

Stock, J. H. and J. H. Wright (2000). GMM with weak identification. *Econometrica* 68(5), 1055–1096.

Stock, J. H. and M. Yogo (2005). Testing for weak instruments in linear IV regression. In *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg* (D. W. K. Andrews and J. H. Stock, eds.), 80–108. Cambridge University Press.

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

Sun, B. and Z. Tan (2022). High-dimensional model-assisted inference for local average treatment effects with instrumental variables. *Journal of Business & Economic Statistics* 40, 1732–1744.

van der Vaart, A. W. and J. A. Wellner (1996). *Weak Convergence and Empirical Processes*. Springer.

Online Appendix for
Adaptive Nuisance Selection for
Weak-Identification-Robust Inference

Tamer Çetin

March 2026

This online appendix records six auxiliary blocks that support the main text. Section OA.1 proves the oracle ridge quadratic-form bootstrap lemma used in Appendix A.10 of the main text. Section OA.2 gives the full conditional QLR / CLR construction compatible with LF–NCF, including the simulation algorithm, the proof of Theorem 6 in the main text, and the homoskedastic Gaussian linear-IV specialization to CLR. Section OA.3 gives a worked learner-class verification for sparse linear nuisance estimators based on lasso over a finite library. Section OA.4 records short supplementary technical notes. Section OA.5 reports supplementary Monte Carlo diagnostics for the baseline homoskedastic and heteroskedastic PLIV designs. Section OA.6 records empirical implementation details and an enriched-control Angrist–Krueger exercise. Theorem 4 in the main text additionally relies on the oracle ridge-bootstrap lemma proved in Section OA.1; the remaining sections sharpen implementation, non-vacuity, and finite-sample interpretation.

OA.1 Oracle Ridge Bootstrap under Effective-Rank Control

Throughout this section, write

$$A_n := \Omega_P(\theta_0) + \rho_n I_{d_{g,n}}, \qquad \widehat{A}_n := \widehat{\Omega}_n(\theta_0) + \rho_n I_{d_{g,n}}, \qquad \Sigma_n := A_n^{-1/2}\Omega_P(\theta_0)A_n^{-1/2}.$$

Also define the oracle and feasible score-covariance matrices

$$\Omega_{n,0} := \frac{1}{n}\sum_{i=1}^{n}\psi(W_i;\theta_0,\eta_0)\psi(W_i;\theta_0,\eta_0)^\top, \qquad \widehat{\Omega}_n(\theta_0) := \frac{1}{n}\sum_{i=1}^{n}\widehat{\psi}_i(\theta_0)\widehat{\psi}_i(\theta_0)^\top.$$

Define

$$B_{0n} := A_n^{-1/2}\Omega_{n,0}A_n^{-1/2}, \qquad \widehat{B}_n := \widehat{A}_n^{-1/2}\widehat{\Omega}_n(\theta_0)\widehat{A}_n^{-1/2}.$$

Let

$$X_{i,n} := A_n^{-1/2}\psi(W_i;\theta_0,\eta_0) \in \mathbb{R}^{d_{g,n}},$$

so that

$$Q_{0n} = g_{n,0}(\theta_0)^\top A_n^{-1} g_{n,0}(\theta_0) = \left\| \frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_{i,n} \right\|^2.$$

With Gaussian multiplier weights $\{\xi_i\}_{i=1}^n$, define the oracle bootstrap analogue

$$Q^*_{0n} = g^*_{n,0}(\theta_0)^\top A_n^{-1} g^*_{n,0}(\theta_0) = \left\| \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \xi_i X_{i,n} \right\|^2.$$

Here $\mathcal{G}_n$ denotes the σ-field generated by the realized data and all LF–NCF training outcomes, as in the main paper.

Lemma OA.1 (Oracle ridge quadratic-form bootstrap)**.** *Under Assumption 8(i)–(v),*

$$\sup_{t\in\mathbb{R}} |P_P(Q_{0n} \le t) - P^*_P(Q^*_{0n} \le t \mid \mathcal{G}_n)| \to_P 0$$

uniformly over $P \in \mathcal{P}_n$.

Proof. Fix $P \in \mathcal{P}_n$. The transformed triangular array $\{X_{i,n}\}_{i=1}^n$ falls within the scope of the weighted-bootstrap consistency result for quadratic forms of sample averages in Pouzo (2015), applied row by row and then uniformly over $P \in \mathcal{P}_n$.

To make the mapping into that theorem completely explicit, note first that Assumption 8(ii) gives

$$\lambda_{\min}(A_n) = \lambda_{\min}\Big(\Omega_P(\theta_0) + \rho_n I_{d_{g,n}}\Big) \ge c > 0.$$

Hence A_n is symmetric positive definite. Hence a spectral decomposition

$$A_n = U_n \operatorname{diag}(\lambda_{1,n}(A_n), \ldots, \lambda_{d_{g,n},n}(A_n))U_n^\top$$

exists with U_n orthogonal and each eigenvalue strictly positive. Consequently,

$$A_n^{-1/2} = U_n \operatorname{diag}(\lambda_{1,n}(A_n)^{-1/2}, \ldots, \lambda_{d_{g,n},n}(A_n)^{-1/2}) U_n^\top$$

is deterministic for fixed P and is the unique symmetric positive definite square root of A_n^{-1}. This deterministic linear map is the only transformation applied to the oracle score vector $\psi(W_i; \theta_0, \eta_0)$.

Recall that

$$X_{i,n} = A_n^{-1/2} \psi(W_i; \theta_0, \eta_0), \qquad A_n = \Omega_P(\theta_0) + \rho_n I_{d_{g,n}},$$

so the matrix A_n, and hence $A_n^{-1/2}$, is deterministic once P is fixed. Because $W_1, \ldots, W_n$ are i.i.d. under P, it follows immediately that for each row n the vectors $X_{1,n}, \ldots, X_{n,n}$ are i.i.d. as well. Moreover,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n X_{i,n} = A_n^{-1/2} \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi(W_i; \theta_0, \eta_0) = A_n^{-1/2} g_{n,0}(\theta_0),$$

so that

$$\left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n X_{i,n} \right\|^2 = g_{n,0}(\theta_0)^\top A_n^{-1} g_{n,0}(\theta_0) = Q_{0n}.$$

The last identity can be written out directly as

$$\begin{aligned}
\left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n X_{i,n} \right\|^2 &= \left(A_n^{-1/2} g_{n,0}(\theta_0) \right)^\top \left(A_n^{-1/2} g_{n,0}(\theta_0) \right) \\
&= g_{n,0}(\theta_0)^\top A_n^{-1/2} A_n^{-1/2} g_{n,0}(\theta_0) \\
&= g_{n,0}(\theta_0)^\top A_n^{-1} g_{n,0}(\theta_0),
\end{aligned}$$

where symmetry of $A_n^{-1/2}$ is used in the second line. Likewise, conditional on the data,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i X_{i,n} = A_n^{-1/2} \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i \psi(W_i; \theta_0, \eta_0) = A_n^{-1/2} g_{n,0}^*(\theta_0),$$

whence

$$\left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i X_{i,n} \right\|^2 = g_{n,0}^*(\theta_0)^\top A_n^{-1} g_{n,0}^*(\theta_0) = Q_{0n}^*.$$

Similarly,

$$\left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i X_{i,n} \right\|^2 = \left(A_n^{-1/2} g_{n,0}^*(\theta_0) \right)^\top \left(A_n^{-1/2} g_{n,0}^*(\theta_0) \right)$$
$$= g_{n,0}^*(\theta_0)^\top A_n^{-1} g_{n,0}^*(\theta_0).$$

Thus the statistic and its multiplier analogue are exactly the quadratic forms to which the high-dimensional weighted-bootstrap theorem is to be applied after the deterministic linear transformation induced by $A_n^{-1/2}$.

Step 1: centering and row-wise i.i.d. structure. By the null moment restriction in the main paper,

$$E_P[\psi(W; \theta_0, \eta_0)] = 0.$$

Therefore,

$$E_P[X_{i,n}] = A_n^{-1/2} E_P[\psi(W_i; \theta_0, \eta_0)] = 0.$$

The linearity step is legitimate because $A_n^{-1/2}$ is deterministic under fixed P and the moment condition in Assumption 8(i) guarantees that $E_P \|\psi(W_i; \theta_0, \eta_0)\| < \infty$. For independence, if $B_1, \ldots, B_n \subseteq \mathbb{R}^{d_{g,n}}$ are Borel sets, then

$$P_P(X_{1,n} \in B_1, \ldots, X_{n,n} \in B_n) = P_P\Big(\psi(W_1; \theta_0, \eta_0) \in A_n^{1/2} B_1, \ldots, \psi(W_n; \theta_0, \eta_0) \in A_n^{1/2} B_n\Big)$$
$$= \prod_{i=1}^{n} P_P\Big(\psi(W_i; \theta_0, \eta_0) \in A_n^{1/2} B_i\Big)$$
$$= \prod_{i=1}^{n} P_P(X_{i,n} \in B_i),$$

so the deterministic transformation preserves the row-wise i.i.d. structure. Since the map $w \mapsto A_n^{-1/2} \psi(w; \theta_0, \eta_0)$ is deterministic for fixed P, the array $\{X_{i,n}\}_{i=1}^n$ is mean zero and i.i.d. within each row n. This is the basic sampling structure required in Pouzo (2015).

Step 2: covariance matrix and effective-rank characterization. The covariance matrix of the transformed summands is

$$E_P[X_{i,n} X_{i,n}^\top] = A_n^{-1/2} E_P\Big[\psi(W_i; \theta_0, \eta_0) \psi(W_i; \theta_0, \eta_0)^\top\Big] A_n^{-1/2} = A_n^{-1/2} \Omega_P(\theta_0) A_n^{-1/2} =: \Sigma_n.$$

Hence the covariance of the normalized sample average $n^{-1/2}\sum_{i=1}^{n} X_{i,n}$ is precisely Σ_n.

Indeed, because the row is i.i.d. and mean zero,

$$\begin{aligned}\mathrm{Var}_P\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_{i,n}\right) &= \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}_P(X_{i,n}) + \frac{1}{n}\sum_{i\neq j}\mathrm{Cov}_P(X_{i,n}, X_{j,n}) \\ &= \frac{1}{n}\sum_{i=1}^{n} E_P[X_{i,n}X_{i,n}^\top] + \frac{1}{n}\sum_{i\neq j} 0 \\ &= E_P[X_{1,n}X_{1,n}^\top] = \Sigma_n.\end{aligned}$$

The cross-covariances vanish because independence together with mean zero gives $E_P[X_{i,n}X_{j,n}^\top] = E_P[X_{i,n}]E_P[X_{j,n}]^\top = 0$ for $i \neq j$.

Next, connect Σ_n to the effective-rank quantity in Assumption 8. By cyclicity of the trace,

$$\mathrm{tr}(\Sigma_n) = \mathrm{tr}\Big(A_n^{-1/2}\Omega_P(\theta_0)A_n^{-1/2}\Big) = \mathrm{tr}\Big(\Omega_P(\theta_0)A_n^{-1}\Big) = r_{\mathrm{eff}}(\rho_n).$$

Thus the trace parameter appearing in the theorem of Pouzo (2015) is exactly the ridge effective rank used in the main paper.

This identity is useful because it shows that the relevant "dimension" for the quadratic form is not the ambient score dimension $d_{g,n}$ itself, but rather the effective-rank quantity induced by ridge regularization. In particular, directions in score space associated with eigenvalues of $\Omega_P(\theta_0)$ that are small relative to ρ_n contribute little to $\mathrm{tr}(\Sigma_n)$, since the eigenvalues of Σ_n equal $\lambda_j(\Omega_P(\theta_0))/(\lambda_j(\Omega_P(\theta_0)) + \rho_n) \in [0, 1]$.

Also record the uniform operator-norm bound. Since

$$\Omega_P(\theta_0) \preceq \Omega_P(\theta_0) + \rho_n I_{d_{g,n}} = A_n,$$

pre- and post-multiplication by the positive semidefinite matrix $A_n^{-1/2}$ yields

$$A_n^{-1/2}\Omega_P(\theta_0)A_n^{-1/2} \preceq A_n^{-1/2}A_nA_n^{-1/2} = I_{d_{g,n}}.$$

Equivalently,

$$0 \preceq \Sigma_n \preceq I_{d_{g,n}}, \qquad \|\Sigma_n\|_{\mathrm{op}} \leq 1. \tag{OA.1.1}$$

To see the operator-norm statement explicitly, let $v \in \mathbb{R}^{d_{g,n}}$ satisfy $\|v\| = 1$. Then

(OA.1.1) implies

$$0 \leq v^\top \Sigma_n v \leq v^\top I_{d_{g,n}} v = 1.$$

Taking the supremum over unit vectors yields

$$\|\Sigma_n\|_{\mathrm{op}} = \sup_{\|v\|=1} v^\top \Sigma_n v \leq 1,$$

because Σ_n is symmetric nonnegative semidefinite. In particular, the covariance matrices of the transformed array are uniformly bounded in operator norm over $P \in \mathcal{P}_n$.

Step 3: uniform high-order moments of the transformed array. Assumption 8(ii) gives the uniform lower eigenvalue bound

$$\lambda_{\min}(A_n) = \lambda_{\min}\Big(\Omega_P(\theta_0) + \rho_n I_{d_{g,n}}\Big) \geq c > 0.$$

Hence

$$\|A_n^{-1/2}\|_{\mathrm{op}} = \lambda_{\min}(A_n)^{-1/2} \leq c^{-1/2}.$$

The equality follows from the spectral decomposition displayed above: the eigenvalues of $A_n^{-1/2}$ are exactly $\lambda_j(A_n)^{-1/2}$, so the operator norm is the largest of these numbers, namely $\lambda_{\min}(A_n)^{-1/2}$. Using the submultiplicativity of the Euclidean norm under linear maps,

$$\|X_{i,n}\| = \|A_n^{-1/2}\psi(W_i;\theta_0,\eta_0)\| \leq \|A_n^{-1/2}\|_{\mathrm{op}}\, \|\psi(W_i;\theta_0,\eta_0)\|,$$

and therefore, for the q in Assumption 8(i),

$$\begin{aligned} E_P[\|X_{i,n}\|^q] &= E_P\Big[\|A_n^{-1/2}\psi(W_i;\theta_0,\eta_0)\|^q\Big] \\ &\leq \|A_n^{-1/2}\|_{\mathrm{op}}^q E_P[\|\psi(W_i;\theta_0,\eta_0)\|^q] \\ &\leq c^{-q/2} C_q. \end{aligned} \tag{OA.1.2}$$

More explicitly, the first inequality uses

$$\|A_n^{-1/2}\psi(W_i;\theta_0,\eta_0)\|^q \leq \|A_n^{-1/2}\|_{\mathrm{op}}^q\, \|\psi(W_i;\theta_0,\eta_0)\|^q$$

pointwise, and then expectation is taken on both sides. The final inequality substitutes the uniform eigenvalue bound and the moment bound from Assumption 8(i). This verifies a row-uniform qth moment bound for the transformed triangular array. Because Assumption 8(i) imposes $q > 8$, the moment requirement in the quadratic-form bootstrap theorem is satisfied with slack.

Step 4: growth restriction in terms of effective rank. The high-dimensional theorem used below controls the bootstrap approximation through the trace of the row covariance matrix. Here that trace is exactly $r_{\text{eff}}(\rho_n)$. Accordingly, Assumption 8(iii) becomes

$$\frac{\operatorname{tr}(\Sigma_n)^4}{n} = \frac{r_{\text{eff}}(\rho_n)^4}{n} \to 0$$

uniformly over $P \in \mathcal{P}_n$, which is precisely the required dimension-growth condition after the ridge normalization.

Because the equality is exact, no additional comparison argument is needed: the growth regime assumed for the ridge effective rank transfers verbatim to the trace growth condition in Pouzo (2015). This is the point at which the proof imports the main paper's effective-rank assumption into the bootstrap theorem.

Step 5: multiplier array. Assumption 8(v) gives i.i.d. standard normal multipliers independent of the data. Hence the multiplier conditions in Pouzo (2015) are immediate: the multipliers are centered, have unit variance, and possess moments of all orders.

In particular,

$$E[\xi_i] = 0, \qquad E[\xi_i^2] = 1, \qquad E[|\xi_i|^m] < \infty \text{ for every } m \geq 1,$$

and the collection $\{\xi_i\}_{i=1}^n$ is independent of $\{W_i\}_{i=1}^n$. Conditional on the realized data, the only randomness in Q_{0n}^* therefore comes from a multiplier array satisfying the exact moment and independence conditions required for the weighted bootstrap.

Step 6: application of the weighted-bootstrap theorem. The triangular array $\{X_{i,n}\}_{i=1}^n$ has now been verified, uniformly over $P \in \mathcal{P}_n$, to satisfy all ingredients required by the theorem of Pouzo (2015) for quadratic forms of sample averages:

(a) row-wise i.i.d. sampling and mean zero;

(b) covariance matrix Σ_n with $\|\Sigma_n\|_{\mathrm{op}} \leq 1$;

(c) a uniform qth moment bound (OA.1.2) for some $q > 8$;

(d) the growth restriction $\mathrm{tr}(\Sigma_n)^4/n \to 0$;

(e) i.i.d. mean-zero, variance-one multipliers with sufficiently many moments.

Each item matches a specific ingredient of the theorem in Pouzo (2015): part (a) above verifies the triangular-array sampling structure, parts (b)–(d) provide the covariance and moment controls, and part (e) supplies the admissible multiplier scheme. Accordingly, that result can be invoked with row covariance matrix Σ_n and statistic equal to the squared Euclidean norm of the normalized sample average. Applying that theorem to the statistic

$$\left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_{i,n} \right\|^2$$

and its multiplier version

$$\left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i X_{i,n} \right\|^2,$$

one obtains

$$\sup_{t \in \mathbb{R}} \left| P_P\left(\left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_{i,n} \right\|^2 \leq t \right) - P_P^*\left(\left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i X_{i,n} \right\|^2 \leq t \mid \mathcal{G}_n \right) \right| \to_P 0$$

uniformly over $P \in \mathcal{P}_n$. Substituting the identities already established for Q_{0n} and Q_{0n}^* yields

$$\sup_{t \in \mathbb{R}} |P_P(Q_{0n} \leq t) - P_P^*(Q_{0n}^* \leq t \mid \mathcal{G}_n)| \to_P 0$$

uniformly over $P \in \mathcal{P}_n$, which is the claim. □

Remark OA.1. Lemma OA.1 is stated under the moment conditions of the main paper's effective-rank regime, so the oracle cdf approximation remains valid for more general mean-zero, variance-one multipliers with finite qth moments. The random-critical-value size-control theorem in the main paper is nevertheless stated for Gaussian multipliers, because continuity of the feasible bootstrap law is then immediate from Lemma OA.3.

Lemma OA.2 (Anti-concentration of the oracle Gaussian ridge-bootstrap law)**.** *Let*

$$B_{0n} := A_n^{-1/2}\Omega_{n,0}A_n^{-1/2}, \qquad \Omega_{n,0} := \frac{1}{n}\sum_{i=1}^{n}\psi(W_i;\theta_0,\eta_0)\psi(W_i;\theta_0,\eta_0)^\top,$$

and let

$$\mu_{1,n} \geq \mu_{2,n} \geq \cdots \geq \mu_{d_{g,n},n} \geq 0$$

denote the eigenvalues of B_{0n}. Let

$$G_{0n}(t) := P_P^*(Q_{0n}^* \leq t \mid \mathcal{G}_n)$$

be the conditional cdf of the oracle Gaussian ridge-bootstrap statistic. Suppose there exists a constant $\underline{\lambda} > 0$ such that

$$\inf_{P\in\mathcal{P}_n} P_P(\mu_{1,n} \geq \underline{\lambda}) \to 1.$$

Then there exists an absolute constant $C < \infty$ such that for every deterministic sequence $\eta_n \downarrow 0$,

$$\sup_{P\in\mathcal{P}_n} P_P\left(\sup_{t\in\mathbb{R}}\Big(G_{0n}(t+\eta_n) - G_{0n}(t-\eta_n)\Big) > C\sqrt{\eta_n/\underline{\lambda}}\right) \to 0.$$

In particular,

$$\sup_{t\in\mathbb{R}}\Big(G_{0n}(t+\eta_n) - G_{0n}(t-\eta_n)\Big) = o_P(1)$$

uniformly over $P \in \mathcal{P}_n$.

Proof. Fix $P \in \mathcal{P}_n$, and condition on $\mathcal{G}_n$. Define

$$X_{i,n} := A_n^{-1/2}\psi(W_i;\theta_0,\eta_0), \qquad B_{0n} := A_n^{-1/2}\Omega_{n,0}A_n^{-1/2} = \frac{1}{n}\sum_{i=1}^{n} X_{i,n}X_{i,n}^\top.$$

Then $\{X_{i,n}\}_{i=1}^n$ and B_{0n} are fixed.

The objective is to bound the oscillation of the conditional distribution function of Q_{0n}^* over intervals of radius η_n. Because conditioning on $\mathcal{G}_n$ freezes the array $\{X_{i,n}\}_{i=1}^n$, the problem becomes finite-dimensional and deterministic conditional on the data: the only remaining randomness comes from the Gaussian multipliers. The proof therefore proceeds by writing the bootstrap quadratic form as a weighted sum of independent

χ_1^2 variables and then showing that the largest non-negligible weight already forces an anti-concentration bound.

Step 1: conditional Gaussian representation of the oracle bootstrap statistic. By construction,

$$Q_{0n}^* = \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i X_{i,n} \right\|^2, \qquad \xi_i \overset{iid}{\sim} N(0,1).$$

Conditional on $\mathcal{G}_n$, the vector

$$S_n^* := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i X_{i,n}$$

is Gaussian with mean zero. Its conditional covariance matrix is

$$\begin{aligned} E_P^*\Big[S_n^*(S_n^*)^\top \mid \mathcal{G}_n\Big] &= \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} E_P^*[\xi_i \xi_j] X_{i,n} X_{j,n}^\top \\ &= \frac{1}{n} \sum_{i=1}^{n} X_{i,n} X_{i,n}^\top \\ &= A_n^{-1/2} \Omega_{n,0} A_n^{-1/2} = B_{0n}, \end{aligned}$$

where the second equality uses $E_P^*[\xi_i \xi_j] = \mathbb{1}\{i = j\}$. Hence, conditional on $\mathcal{G}_n$,

$$S_n^* \overset{d}{=} N(0, B_{0n}), \qquad Q_{0n}^* = \|S_n^*\|^2.$$

If $Z_n \sim N(0, I_{d_{g,n}})$, then the standard covariance-factorization identity gives

$$S_n^* \overset{d}{=} B_{0n}^{1/2} Z_n,$$

and therefore

$$Q_{0n}^* \overset{d}{=} Z_n^\top B_{0n} Z_n.$$

The covariance calculation can be unpacked slightly further. Since

$$S_n^*(S_n^*)^\top = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \xi_i \xi_j X_{i,n} X_{j,n}^\top,$$

conditional expectation with respect to the multipliers acts only on the scalar products $\xi_i\xi_j$. Independence and standard normality imply

$$E_P^*[\xi_i\xi_j] = 0 \quad (i \neq j), \qquad E_P^*[\xi_i^2] = 1,$$

so all cross terms vanish and only the diagonal terms remain. This is exactly why the bootstrap covariance equals the empirical covariance matrix B_{0n}.

Because any centered Gaussian vector is completely characterized by its covariance matrix, the law of $S_n^* \mid \mathcal{G}_n$ is the same as that of $B_{0n}^{1/2} Z_n$, where $Z_n \sim N(0, I_{d_{g,n}})$. Squaring the Euclidean norm then converts the bootstrap statistic into a Gaussian quadratic form with deterministic matrix B_{0n}.

Step 2: spectral decomposition and reduction to independent χ_1^2 coordinates. Because B_{0n} is symmetric and positive semidefinite, the spectral theorem yields an orthogonal matrix U_n such that

$$U_n^\top B_{0n} U_n = \operatorname{diag}(\mu_{1,n}, \mu_{2,n}, \ldots, \mu_{d_{g,n},n}), \qquad \mu_{1,n} \geq \cdots \geq \mu_{d_{g,n},n} \geq 0.$$

Since orthogonal transformations preserve the standard Gaussian law, $U_n^\top Z_n \sim N(0, I_{d_{g,n}})$. Writing $(Z_1, \ldots, Z_{d_{g,n}})^\top := U_n^\top Z_n$, one obtains the conditional representation

$$Q_{0n}^* \overset{d}{=} \sum_{j=1}^{d_{g,n}} \mu_{j,n} Z_j^2, \tag{OA.1.3}$$

where $Z_1, \ldots, Z_{d_{g,n}}$ are i.i.d. standard normal.

This representation is immediate from the orthogonal invariance of Z_n: indeed,

$$\begin{aligned} Z_n^\top B_{0n} Z_n &= Z_n^\top U_n \operatorname{diag}(\mu_{1,n}, \ldots, \mu_{d_{g,n},n}) U_n^\top Z_n \\ &= (U_n^\top Z_n)^\top \operatorname{diag}(\mu_{1,n}, \ldots, \mu_{d_{g,n},n}) (U_n^\top Z_n) \\ &\overset{d}{=} \sum_{j=1}^{d_{g,n}} \mu_{j,n} Z_j^2. \end{aligned}$$

Thus the conditional distribution of the full quadratic form is completely determined by the eigenvalues of B_{0n}.

A key observation is that at least one coordinate receives non-negligible weight on the event $\{\mu_{1,n} \geq \underline{\lambda}\}$. To exploit that coordinate, define

$$V_n := \sum_{j=2}^{d_{g,n}} \mu_{j,n} Z_j^2.$$

Then V_n is independent of Z_1 and

$$Q_{0n}^* \overset{d}{=} \mu_{1,n} Z_1^2 + V_n.$$

The independence follows because V_n is measurable with respect to $(Z_2, \ldots, Z_{d_{g,n}})$, while Z_1 is independent of that vector. Hence conditional on V_n, the randomness in Q_{0n}^* is driven entirely by the one-dimensional term $\mu_{1,n} Z_1^2$.

Step 3: reduction of anti-concentration to a one-dimensional problem. Fix $\eta > 0$ and $t \in \mathbb{R}$. Conditional on $(V_n, \mathcal{G}_n)$,

$$\begin{aligned} &P_P^*\Big(t - \eta < Q_{0n}^* \leq t + \eta \mid V_n, \mathcal{G}_n\Big) \\ &\qquad = P\Big(t - \eta - V_n < \mu_{1,n} Z_1^2 \leq t + \eta - V_n\Big). \end{aligned}$$

Taking expectations with respect to V_n conditional on $\mathcal{G}_n$ gives

$$\begin{aligned} G_{0n}(t+\eta) - G_{0n}(t-\eta) &= P_P^*\Big(t - \eta < Q_{0n}^* \leq t + \eta \mid \mathcal{G}_n\Big) \\ &= E_P^*\Big[P\Big(t - \eta - V_n < \mu_{1,n} Z_1^2 \leq t + \eta - V_n\Big) \Big| \mathcal{G}_n\Big] \\ &\leq \sup_{s \in \mathbb{R}} P\Big(s < \mu_{1,n} Z_1^2 \leq s + 2\eta\Big). \end{aligned} \tag{OA.1.4}$$

Thus the conditional anti-concentration problem for the full quadratic form is reduced to an anti-concentration bound for one scaled χ_1^2 variable.

The last inequality is simply a re-centering argument. For each realization of V_n, set

$$s = t - \eta - V_n.$$

Then the inner probability equals

$$P\Big(s < \mu_{1,n} Z_1^2 \leq s + 2\eta\Big),$$

and this is bounded above by the supremum over all real s. Averaging with respect to the law of $V_n \mid \mathcal{G}_n$ cannot increase that supremum. The many-dimensional problem is therefore reduced exactly, not approximately, to a one-dimensional anti-concentration bound.

Step 4: one-dimensional anti-concentration bound. There exists an absolute constant $C_0 < \infty$ such that for every $\mu > 0$, every $h > 0$, and every $s \in \mathbb{R}$,

$$P\Big(s < \mu Z^2 \le s + h\Big) \le C_0 \sqrt{h/\mu}, \qquad Z \sim N(0,1). \tag{OA.1.5}$$

The proof proceeds by splitting according to the location of the interval $(s, s+h]$.

Case 1: $s + h \le 0$. Since $\mu Z^2 \ge 0$ almost surely,

$$P\Big(s < \mu Z^2 \le s + h\Big) = 0,$$

so (OA.1.5) is trivial.

Case 2: $s < h$. Then the interval $(s, s+h]$ intersects $[0, \infty)$ in a subset of $[0, 2h]$, and thus

$$P\Big(s < \mu Z^2 \le s + h\Big) \le P(0 \le \mu Z^2 \le 2h) = P\Big(|Z| \le \sqrt{2h/\mu}\Big).$$

Using the elementary Gaussian bound

$$P(|Z| \le x) = \int_{-x}^{x} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz \le \frac{2x}{\sqrt{2\pi}} = \sqrt{\frac{2}{\pi}}\, x, \qquad x \ge 0,$$

one obtains

$$P\Big(s < \mu Z^2 \le s + h\Big) \le \sqrt{\frac{2}{\pi}} \sqrt{\frac{2h}{\mu}} = \frac{2}{\sqrt{\pi}} \sqrt{h/\mu}.$$

The set inclusion used here is worth recording explicitly. If $s < h$ and $y \in (s, s+h] \cap [0, \infty)$, then $0 \le y \le s + h < 2h$. Hence

$$(s, s+h] \cap [0, \infty) \subseteq [0, 2h].$$

Since μZ^2 is nonnegative, only this intersection matters.

Case 3: $s \ge h$. Let $Y_\mu := \mu Z^2$. Since Z^2 has density $y \mapsto (2\pi y)^{-1/2} e^{-y/2}$ on $(0, \infty)$, the change of variables $y \mapsto y/\mu$ gives the density of Y_μ:

$$f_\mu(y) = \frac{1}{\sqrt{2\pi\mu y}} \exp\Big(-\frac{y}{2\mu}\Big), \qquad y > 0.$$

Moreover,

$$\frac{d}{dy} \log f_\mu(y) = -\frac{1}{2y} - \frac{1}{2\mu} < 0, \qquad y > 0,$$

so f_μ is strictly decreasing on $(0, \infty)$. Therefore,

$$P\Big(s < Y_\mu \le s + h\Big) = \int_s^{s+h} f_\mu(y)\, dy \le h \sup_{y \in [s, s+h]} f_\mu(y) = h f_\mu(s) \le \frac{h}{\sqrt{2\pi\mu s}}.$$

Since $s \ge h$, it follows that $h/s \le 1$, and hence

$$\frac{h}{\sqrt{2\pi\mu s}} = \frac{1}{\sqrt{2\pi}} \sqrt{\frac{h}{\mu}} \sqrt{\frac{h}{s}} \le \frac{1}{\sqrt{2\pi}} \sqrt{\frac{h}{\mu}}.$$

Thus (OA.1.5) holds in Case 3 as well.

For completeness, the change-of-variables formula is as follows. Since $Y_\mu = \mu Z^2$, for $y > 0$,

$$P(Y_\mu \le y) = P\left(Z^2 \le \frac{y}{\mu}\right),$$

and differentiating with respect to y gives

$$f_\mu(y) = \frac{1}{\mu} f_{Z^2}(y/\mu) = \frac{1}{\mu} \cdot \frac{1}{\sqrt{2\pi(y/\mu)}} e^{-y/(2\mu)} = \frac{1}{\sqrt{2\pi\mu y}} e^{-y/(2\mu)}.$$

Monotonicity of the density then implies that among all intervals of a given length lying in $[h, \infty)$, the largest probability mass is attained by pushing the interval as far left as allowed; this heuristic is exactly what the bound by $h f_\mu(s)$ formalizes.

Combining the three cases, (OA.1.5) is valid with, for example,

$$C_0 := \frac{2}{\sqrt{\pi}}.$$

Step 5: conclusion on the event of a nondegenerate leading eigenvalue. Returning to (OA.1.4) and applying (OA.1.5) with $h = 2\eta$ and $\mu = \mu_{1,n}$, one obtains for every $t \in \mathbb{R}$,

$$G_{0n}(t+\eta) - G_{0n}(t-\eta) \le C_0 \sqrt{2\eta/\mu_{1,n}}.$$

Hence, on the event

$$E_n := \{\mu_{1,n} \ge \underline{\lambda}\},$$

the deterministic bound holds

$$\sup_{t\in\mathbb{R}} \Big(G_{0n}(t+\eta) - G_{0n}(t-\eta) \Big) \le C_0 \sqrt{2\eta/\underline{\lambda}}.$$

Define

$$C := C_0\sqrt{2} = \frac{2\sqrt{2}}{\sqrt{\pi}}.$$

Then on E_n,

$$\sup_{t\in\mathbb{R}} \Big(G_{0n}(t+\eta) - G_{0n}(t-\eta) \Big) \le C \sqrt{\eta/\underline{\lambda}}.$$

This is the key deterministic inequality: once the largest eigenvalue of B_{0n} is bounded away from zero, the entire conditional cdf admits a modulus of continuity of order $\sqrt{\eta}$. The bound is uniform in t because the one-dimensional supremum in (OA.1.4) was already uniform in the shift parameter.

Step 6: uniform probability statement over $\mathcal{P}_n$. Now set $\eta = \eta_n \downarrow 0$. By the hypothesis of the lemma,

$$\inf_{P\in\mathcal{P}_n} P_P(E_n) \to 1.$$

Therefore,

$$\begin{aligned}\sup_{P\in\mathcal{P}_n} P_P\bigg(\sup_{t\in\mathbb{R}} \Big(G_{0n}(t+\eta_n) - G_{0n}(t-\eta_n) \Big) > C\sqrt{\eta_n/\underline{\lambda}} \bigg) \\ \le \sup_{P\in\mathcal{P}_n} P_P(E_n^c) \to 0.\end{aligned}$$

This proves the first displayed assertion.

Finally, because $\eta_n \downarrow 0$, the deterministic envelope $C\sqrt{\eta_n/\underline{\lambda}}$ converges to zero. Hence

$$\sup_{t\in\mathbb{R}} \Big(G_{0n}(t+\eta_n) - G_{0n}(t-\eta_n) \Big) = o_P(1)$$

uniformly over $P \in \mathcal{P}_n$. This proves the lemma. $\square$

Lemma OA.3 (Continuity of the feasible Gaussian ridge-bootstrap law)**.** *Assume the multipliers are i.i.d. standard normal. Conditional on $\mathcal{G}_n$, if $\widehat{\mu}_{1,n} > 0$, then the bootstrap statistic Q_n^* has a continuous distribution function on $\mathbb{R}$.*

On the event $\lambda_{\min}(\widehat{A}_n) > 0$, which has probability tending to one under the feasible covariance approximation and Assumption 8(ii), the following identities are well defined. More precisely, if

$$\widehat{\lambda}_{1,n} \geq \widehat{\lambda}_{2,n} \geq \cdots \geq \widehat{\lambda}_{d_g,n,n} \geq 0$$

denote the eigenvalues of $\widehat{B}_n$, and $r_n := \#\{j : \widehat{\lambda}_{j,n} > 0\}$, then conditional on $\mathcal{G}_n$,

$$Q_n^* \stackrel{d}{=} \sum_{j=1}^{r_n} \widehat{\lambda}_{j,n} Z_j^2,$$

where $Z_1, \ldots, Z_{r_n}$ are i.i.d. standard normal. In particular,

$$P_P^*(Q_n^* = t \mid \mathcal{G}_n) = 0 \qquad \textit{for every } t \in \mathbb{R}.$$

Under Assumption 8(vii), the event $\{\widehat{\mu}_{1,n} > 0\}$ occurs with probability approaching one uniformly over $P \in \mathcal{P}_n$.

Proof. Conditional on $\mathcal{G}_n$, the vectors $\widehat{\psi}_i(\theta_0)$ are fixed and the multipliers $\xi_1, \ldots, \xi_n$ are i.i.d. $N(0, 1)$. Therefore

$$\widehat{g}_n^*(\theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i \widehat{\psi}_i(\theta_0)$$

is a conditionally Gaussian random vector, because it is a deterministic linear combination of jointly Gaussian multipliers. Its conditional mean and covariance are computed explicitly. First,

$$E_P^*\Big[\widehat{g}_n^*(\theta_0) \mid \mathcal{G}_n\Big] = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} E_P^*[\xi_i] \widehat{\psi}_i(\theta_0) = 0,$$

since each multiplier has mean zero. Second, using independence of the multipliers and $E_P^*[\xi_i\xi_j] = \mathbb{1}\{i = j\}$,

$$\begin{aligned}
E_P^*\Big[\widehat{g}_n^*(\theta_0)\widehat{g}_n^*(\theta_0)^\top \mid \mathcal{G}_n\Big]. &= E_P^*\left[\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i\widehat{\psi}_i(\theta_0)\right)\left(\frac{1}{\sqrt{n}}\sum_{j=1}^n \xi_j\widehat{\psi}_j(\theta_0)\right)^\top \middle| \mathcal{G}_n\right] \\
&= \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n E_P^*[\xi_i\xi_j]\,\widehat{\psi}_i(\theta_0)\widehat{\psi}_j(\theta_0)^\top \\
&= \frac{1}{n}\sum_{i=1}^n \widehat{\psi}_i(\theta_0)\widehat{\psi}_i(\theta_0)^\top \\
&= \widehat{\Omega}_n(\theta_0).
\end{aligned}$$

Hence, conditional on $\mathcal{G}_n$,

$$\widehat{g}_n^*(\theta_0) \overset{d}{=} \widehat{\Omega}_n(\theta_0)^{1/2}Z, \qquad Z \sim N(0, I_{d_{g,n}}).$$

This is simply the standard representation of a centered Gaussian vector with covariance matrix $\widehat{\Omega}_n(\theta_0)$.

Substituting this representation into the bootstrap quadratic form gives

$$\begin{aligned}
Q_n^* &= \widehat{g}_n^*(\theta_0)^\top\widehat{A}_n^{-1}\widehat{g}_n^*(\theta_0) \\
&\overset{d}{=} \left(\widehat{\Omega}_n(\theta_0)^{1/2}Z\right)^\top\widehat{A}_n^{-1}\left(\widehat{\Omega}_n(\theta_0)^{1/2}Z\right) \\
&= Z^\top\widehat{\Omega}_n(\theta_0)^{1/2}\widehat{A}_n^{-1}\widehat{\Omega}_n(\theta_0)^{1/2}Z.
\end{aligned}$$

On the event $\lambda_{\min}(\widehat{A}_n) > 0$, which has probability tending to one under the feasible covariance approximation and Assumption 8(ii), the following identities are well defined. Let

$$M_n := \widehat{\Omega}_n(\theta_0)^{1/2}\widehat{A}_n^{-1}\widehat{\Omega}_n(\theta_0)^{1/2}.$$

Since

$$\widehat{A}_n = \widehat{\Omega}_n(\theta_0) + \rho_n I,$$

the matrices $\widehat{A}_n$ and $\widehat{\Omega}_n(\theta_0)$ commute, and so do their spectral powers. Hence

$$M_n = \widehat{A}_n^{-1/2}\widehat{\Omega}_n(\theta_0)\widehat{A}_n^{-1/2} = \widehat{B}_n$$

as an exact algebraic identity.

By the spectral theorem there exists an orthogonal matrix U_n such that

$$U_n^\top \widehat{B}_n U_n = \operatorname{diag}(\widehat{\lambda}_{1,n}, \dots, \widehat{\lambda}_{r_n,n}, 0, \dots, 0),$$

with $r_n = \#\{j : \widehat{\lambda}_{j,n} > 0\}$. Because a standard Gaussian vector is invariant under orthogonal transformations, $U_n^\top Z \sim N(0, I_{d_{g,n}})$. Writing

$$U_n^\top Z = (Z_1, \dots, Z_{d_{g,n}})^\top$$

with $Z_1, \dots, Z_{d_{g,n}}$ i.i.d. standard normal, one obtains

$$\begin{aligned}
Q_n^* &\overset{d}{=} Z^\top M_n Z = Z^\top \widehat{B}_n Z \\
&= Z^\top U_n \operatorname{diag}(\widehat{\lambda}_{1,n}, \dots, \widehat{\lambda}_{r_n,n}, 0, \dots, 0)\, U_n^\top Z \\
&\overset{d}{=} \sum_{j=1}^{r_n} \widehat{\lambda}_{j,n} Z_j^2.
\end{aligned}$$

This proves the claimed conditional representation.

Suppose now that $\widehat{\mu}_{1,n} > 0$. Since $\widehat{\mu}_{1,n}$ is the largest eigenvalue of $\widehat{B}_n$, this implies $\widehat{\lambda}_{1,n} > 0$, and therefore $r_n \geq 1$. Write

$$Y_n := \widehat{\lambda}_{1,n} Z_1^2, \qquad V_n := \sum_{j=2}^{r_n} \widehat{\lambda}_{j,n} Z_j^2,$$

so that $Q_n^* = Y_n + V_n$. Because Y_n is measurable with respect to Z_1 alone while V_n is measurable with respect to $(Z_2, \dots, Z_{r_n})$, and the coordinates of Z are independent, Y_n and V_n are independent conditional on $\mathcal{G}_n$.

Continuity of the law of Y_n is verified next. Since Z_1^2 has density

$$f_{Z_1^2}(u) = \frac{1}{\sqrt{2\pi u}} e^{-u/2}, \qquad u > 0,$$

a change of variables $y = \widehat{\lambda}_{1,n} u$ yields, for $y > 0$,

$$f_{Y_n}(y) = \frac{1}{\widehat{\lambda}_{1,n}} f_{Z_1^2}\left(\frac{y}{\widehat{\lambda}_{1,n}}\right) = \frac{1}{\sqrt{2\pi \widehat{\lambda}_{1,n} y}} \exp\Big(-\frac{y}{2\widehat{\lambda}_{1,n}}\Big).$$

Thus Y_n has a density on $(0, \infty)$, so its distribution function F_{Y_n} is continuous on $\mathbb{R}$.

Now condition again on $\mathcal{G}_n$ and on V_n. By independence of Y_n and V_n, for any $t \in \mathbb{R}$,

$$\begin{aligned} P_P^*(Q_n^* \leq t \mid \mathcal{G}_n) &= P_P^*(Y_n + V_n \leq t \mid \mathcal{G}_n) \\ &= E_P^*\Big[P_P^*(Y_n \leq t - V_n \mid V_n, \mathcal{G}_n) \mid \mathcal{G}_n\Big] \\ &= E_P^*\Big[F_{Y_n}(t - V_n) \mid \mathcal{G}_n\Big]. \end{aligned}$$

This is the convolution formula for the cdf of a sum of two independent random variables. To show continuity in t, let $t_m \to t$. Since F_{Y_n} is continuous,

$$F_{Y_n}(t_m - V_n) \to F_{Y_n}(t - V_n) \qquad \text{almost surely under } P_P^*(\cdot \mid \mathcal{G}_n).$$

Also, $0 \leq F_{Y_n}(\cdot) \leq 1$. Therefore dominated convergence gives

$$E_P^*\Big[F_{Y_n}(t_m - V_n) \mid \mathcal{G}_n\Big] \to E_P^*\Big[F_{Y_n}(t - V_n) \mid \mathcal{G}_n\Big].$$

Hence the map

$$t \mapsto P_P^*(Q_n^* \leq t \mid \mathcal{G}_n)$$

is continuous on $\mathbb{R}$. A distribution function is continuous at t if and only if there is no point mass at t, so it follows that

$$P_P^*(Q_n^* = t \mid \mathcal{G}_n) = 0 \qquad \text{for every } t \in \mathbb{R}.$$

Finally, Assumption 8(vii) states that

$$\inf_{P \in \mathcal{P}_n} P_P(\widehat{\mu}_{1,n} \geq \underline{\lambda}^*) \to 1$$

for some $\underline{\lambda}^* > 0$. Since the event $\{\widehat{\mu}_{1,n} \geq \underline{\lambda}^*\}$ is contained in $\{\widehat{\mu}_{1,n} > 0\}$, for every $P \in \mathcal{P}_n$,

$$P_P(\widehat{\mu}_{1,n} > 0) \geq P_P(\widehat{\mu}_{1,n} \geq \underline{\lambda}^*).$$

Taking the infimum over $P \in \mathcal{P}_n$ on both sides yields

$$\inf_{P \in \mathcal{P}_n} P_P(\widehat{\mu}_{1,n} > 0) \geq \inf_{P \in \mathcal{P}_n} P_P(\widehat{\mu}_{1,n} \geq \underline{\lambda}^*) \to 1.$$

Thus the event $\{\widehat{\mu}_{1,n} > 0\}$ occurs with probability approaching one, uniformly over

$P \in \mathcal{P}_n$. This proves the final assertion. $\square$

OA.2 Conditional QLR / CLR with Cross-Fitted Orthogonal Moments

This section records the full conditional-inference construction that underlies Theorem 6 in the main text. The construction follows Andrews and Mikusheva (2016) and adapts it to the leakage-free filtration induced by LF–NCF.

OA.2.1 Setup

For each $\theta \in \Theta$, define

$$\widehat{\psi}_i(\theta) := \psi(W_i; \theta, \widehat{\eta}_{-i}), \qquad \psi_{0,i}(\theta) := \psi(W_i; \theta, \eta_0),$$

$$\widehat{g}_n(\theta) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \widehat{\psi}_i(\theta), \qquad g_{n,0}(\theta) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_{0,i}(\theta),$$

and

$$\widehat{\Sigma}_n(\theta, \vartheta) := \frac{1}{n} \sum_{i=1}^{n} \widehat{\psi}_i(\theta) \widehat{\psi}_i(\vartheta)^\top, \qquad \Sigma_{n,0}(\theta, \vartheta) := \frac{1}{n} \sum_{i=1}^{n} \psi_{0,i}(\theta) \psi_{0,i}(\vartheta)^\top.$$

Let

$$\Sigma_P(\theta, \vartheta) := \mathbb{E}_P\left[\psi(W; \theta, \eta_0) \psi(W; \vartheta, \eta_0)^\top\right].$$

Assumption OA.1 (Process regularity for conditional inference)**.** Uniformly over $P \in \mathcal{P}_n$:

(i) there exists $q > 4$ such that

$$\sup_{P \in \mathcal{P}_n(\theta_0)} \sup_{\theta \in \Theta} \mathbb{E}_P[\|\psi(W; \theta, \eta_0)\|^q] < \infty;$$

(ii) the class

$$\{\psi(\cdot; \theta, \eta_0) : \theta \in \Theta\}$$

is P-Donsker, uniformly over $\mathcal{P}_n$, and the associated Gaussian approximation admits a version with almost surely uniformly continuous sample paths on Θ;

(iii) the product class

$$\Big\{\psi(\cdot;\theta,\eta_0)\psi(\cdot;\vartheta,\eta_0)^\top : (\theta,\vartheta)\in\Theta^2\Big\}$$

is uniformly Glivenko–Cantelli, so that

$$\sup_{\theta,\vartheta\in\Theta}\|\Sigma_{n,0}(\theta,\vartheta)-\Sigma_P(\theta,\vartheta)\|_{\mathrm{op}}=o_P(1)$$

uniformly over $P\in\mathcal{P}_n$;

(iv) there exist constants $0<\underline{c}<\overline{C}<\infty$ such that

$$\inf_{P\in\mathcal{P}_n}\inf_{\theta\in\Theta}\lambda_{\min}(\Sigma_P(\theta,\theta))\ge\underline{c},\qquad \sup_{P\in\mathcal{P}_n(\theta_0)}\sup_{\theta\in\Theta}\|\Sigma_P(\theta,\theta)\|_{\mathrm{op}}\le\overline{C};$$

(v) under LF–NCF and the nuisance-rate assumptions of the main paper,

$$\sup_{\theta\in\Theta}\|\widehat{g}_n(\theta)-g_{n,0}(\theta)\|=o_P(1),$$

and

$$\sup_{\theta,\vartheta\in\Theta}\|\widehat{\Sigma}_n(\theta,\vartheta)-\Sigma_{n,0}(\theta,\vartheta)\|_{\mathrm{op}}=o_P(1).$$

Lemma OA.4 (Oracle Gaussian approximation)**.** *Under Assumption OA.1(i)–(iii), there exists a Gaussian process $\mathbb{G}_{n,P}(\cdot)$ in $\ell^\infty(\Theta)$ with mean function*

$$\mu_{n,P}(\theta):=\mathbb{E}_P[g_{n,0}(\theta)]$$

and covariance kernel

$$\mathrm{Cov}\Big(\mathbb{G}_{n,P}(\theta),\mathbb{G}_{n,P}(\vartheta)\Big)=\Sigma_P(\theta,\vartheta)-n^{-1}\mu_{n,P}(\theta)\mu_{n,P}(\vartheta)^\top$$

such that

$$g_{n,0}(\cdot)\Rightarrow\mathbb{G}_{n,P}(\cdot)\qquad \textit{in } \ell^\infty(\Theta),$$

uniformly over $P\in\mathcal{P}_n$.

Proof. For each θ, $g_{n,0}(\theta)$ is a rescaled empirical average with finite-sample mean $\mu_{n,P}(\theta)=\sqrt{n}\,\mathbb{E}_P[\psi(W;\theta,\eta_0)]$. Its covariance kernel is

$$\mathrm{Cov}\Big(g_{n,0}(\theta),g_{n,0}(\vartheta)\Big)=\Sigma_P(\theta,\vartheta)-n^{-1}\mu_{n,P}(\theta)\mu_{n,P}(\vartheta)^\top.$$

Under Assumption OA.1(i)–(ii), the class $\{\psi(\cdot;\theta,\eta_0):\theta\in\Theta\}$ is uniformly Donsker with uniformly bounded q-moments, so the standard empirical-process Gaussian approximation applies uniformly over $P\in\mathcal{P}_n$; see van der Vaart and Wellner (1996). Assumption OA.1(iii) supplies the required covariance regularity. This yields the stated Gaussian approximation. □

Lemma OA.5 (Feasible Gaussian approximation under LF–NCF)**.** *Under LF–NCF, Assumption OA.1, and the orthogonality and nuisance-rate conditions in the main paper,*

$$\widehat{g}_n(\cdot)\Rightarrow\mathbb{G}_{n,P}(\cdot)\qquad in\ \ell^\infty(\Theta),$$

uniformly over $P\in\mathcal{P}_n$, with the same Gaussian approximation as in Lemma OA.4. Moreover,

$$\sup_{\theta,\vartheta\in\Theta}\|\widehat{\Sigma}_n(\theta,\vartheta)-\Sigma_P(\theta,\vartheta)\|_{\mathrm{op}}=o_P(1)$$

uniformly over $P\in\mathcal{P}_n$.

Proof. Assumption OA.1(v) gives feasible-oracle equivalence uniformly over Θ and Θ^2:

$$\sup_{\theta\in\Theta}\|\widehat{g}_n(\theta)-g_{n,0}(\theta)\|=o_P(1),\qquad\sup_{\theta,\vartheta\in\Theta}\|\widehat{\Sigma}_n(\theta,\vartheta)-\Sigma_{n,0}(\theta,\vartheta)\|_{\mathrm{op}}=o_P(1).$$

The uniform covariance-kernel law of large numbers in Assumption OA.1(iii) gives

$$\sup_{\theta,\vartheta\in\Theta}\|\Sigma_{n,0}(\theta,\vartheta)-\Sigma_P(\theta,\vartheta)\|_{\mathrm{op}}=o_P(1).$$

Lemma OA.4 yields

$$g_{n,0}(\cdot)\Rightarrow\mathbb{G}_{n,P}(\cdot)\qquad\text{in }\ell^\infty(\Theta),$$

and Assumption OA.1(v) gives

$$\sup_{\theta\in\Theta}\|\widehat{g}_n(\theta)-g_{n,0}(\theta)\|=o_P(1).$$

Therefore

$$\widehat{g}_n(\cdot)\Rightarrow\mathbb{G}_{n,P}(\cdot)\qquad\text{in }\ell^\infty(\Theta),$$

by the standard perturbation argument for weak convergence in the supremum norm.

$\square$

OA.2.2 Conditional QLR statistic and simulated critical value

Define the feasible nonsingular domain

$$\Theta_n^{\mathrm{ns}} := \Big\{\theta \in \Theta : \widehat{\Sigma}_n(\theta,\theta) \text{ is nonsingular}\Big\}.$$

For each null value $\theta_0 \in \Theta_n^{\mathrm{ns}}$, define

$$QLR_n(\theta_0) := \widehat{g}_n(\theta_0)^\top \widehat{\Sigma}_n(\theta_0,\theta_0)^{-1}\widehat{g}_n(\theta_0) - \inf_{\theta\in\Theta_n^{\mathrm{ns}}} \widehat{g}_n(\theta)^\top \widehat{\Sigma}_n(\theta,\theta)^{-1}\widehat{g}_n(\theta).$$

Under Assumption OA.1(iii)–(v), $\Theta_n^{\mathrm{ns}} = \Theta$ with probability tending to one; the restriction to Θ_n^{ns} only makes the finite-sample definition explicit.

Following Andrews and Mikusheva (2016), define the feasible sufficient-statistic transform

$$\widehat{h}_n(\theta) := \widehat{g}_n(\theta) - \widehat{\Sigma}_n(\theta,\theta_0)\widehat{\Sigma}_n(\theta_0,\theta_0)^{-1}\widehat{g}_n(\theta_0), \qquad \theta \in \Theta.$$

Algorithm 1 Conditional QLR test / confidence set under LF–NCF

Input: a grid $\Theta_{\text{grid}} \subset \Theta_n^{\text{ns}}$ (or an optimizer), simulation size B, and level α.

Compute $\widehat{g}_n(\theta)$ and $\widehat{\Sigma}_n(\theta, \vartheta)$ for $\theta, \vartheta \in \Theta_{\text{grid}}$.

for each $\theta_0 \in \Theta_{\text{grid}}$ **do**

  Compute $\widehat{h}_n(\theta)$ for $\theta \in \Theta_{\text{grid}}$.

  Simulate i.i.d. draws

$$\xi^{(b)} \sim N\Big(0, \widehat{\Sigma}_n(\theta_0, \theta_0)\Big), \qquad b = 1, \dots, B.$$

  Form

$$\widehat{g}_n^{*(b)}(\theta) := \widehat{h}_n(\theta) + \widehat{\Sigma}_n(\theta, \theta_0)\widehat{\Sigma}_n(\theta_0, \theta_0)^{-1}\xi^{(b)}.$$

  Compute

$$QLR_n^{*(b)}(\theta_0) := \widehat{g}_n^{*(b)}(\theta_0)^\top \widehat{\Sigma}_n(\theta_0, \theta_0)^{-1}\widehat{g}_n^{*(b)}(\theta_0) - \inf_{\theta \in \Theta_{\text{grid}}} \widehat{g}_n^{*(b)}(\theta)^\top \widehat{\Sigma}_n(\theta, \theta)^{-1}\widehat{g}_n^{*(b)}(\theta).$$

  Let $\widehat{c}_{1-\alpha}^{\text{CQLR}}(\theta_0)$ be the empirical $(1-\alpha)$-quantile of $\{QLR_n^{*(b)}(\theta_0)\}_{b=1}^B$.

end for

Output the inverted confidence set

$$C_{1-\alpha}^{\text{CQLR}} := \Big\{\theta_0 \in \Theta_{\text{grid}} : QLR_n(\theta_0) \le \widehat{c}_{1-\alpha}^{\text{CQLR}}(\theta_0)\Big\}.$$

The theorem below is stated for the exact infimum over Θ. A grid implementation is covered whenever the discretization error in the displayed infima is $o_P(1)$, uniformly over $P \in \mathcal{P}_n$. In the algorithm, this corresponds to taking Θ_{grid} sufficiently fine or using a numerical optimizer over Θ.

For a fixed null value θ_0, write

$$\mathcal{P}_n(\theta_0) := \Big\{P \in \mathcal{P}_n : \mathbb{E}_P[\psi(W; \theta_0, \eta_0)] = 0\Big\}.$$

Theorem OA.1 (Uniform size of cross-fitted conditional QLR under LF–NCF)**.** *Assume LF–NCF and Assumption OA.1. Fix a null value $\theta_0 \in \Theta$ and consider $P \in \mathcal{P}_n$ satisfying*

$$\mathbb{E}_P[\psi(W; \theta_0, \eta_0)] = 0.$$

Suppose $QLR_n(\theta_0)$ is computed as above and the conditional critical value $\widehat{c}^{\mathrm{CQLR}}_{1-\alpha}(\theta_0)$ is computed by Algorithm 1 with $B \to \infty$. Then, for any fixed $\varepsilon > 0$,

$$\limsup_{n\to\infty} \sup_{P\in\mathcal{P}_n(\theta_0)} \Pr_P\Big(QLR_n(\theta_0) > \widehat{c}^{\mathrm{CQLR}}_{1-\alpha}(\theta_0) + \varepsilon\Big) \leq \alpha.$$

If the Gaussian benchmark difference

$$T^{\mathrm{G}}_{n,P} := QLR^{\mathrm{G}}_{n,P}(\theta_0) - c^{\mathrm{CQLR}}_{1-\alpha,n,P}(\theta_0)$$

has a distribution that is continuous at zero uniformly over $P \in \mathcal{P}_n(\theta_0)$, the slack ε may be omitted.

Proof. The argument is organized in four steps.

Step 1: joint weak convergence of the objects entering the CQLR construction. By Lemma OA.5,

$$\widehat{g}_n(\cdot) \Rightarrow \mathbb{G}_{n,P}(\cdot) \qquad \text{in } \ell^\infty(\Theta),$$

uniformly over $P \in \mathcal{P}_n$, and

$$\sup_{\theta,\vartheta\in\Theta} \|\widehat{\Sigma}_n(\theta,\vartheta) - \Sigma_P(\theta,\vartheta)\|_{\mathrm{op}} = o_P(1).$$

Define the feasible sufficient-statistic transform

$$\widehat{h}_n(\theta) := \widehat{g}_n(\theta) - \widehat{\Sigma}_n(\theta,\theta_0)\widehat{\Sigma}_n(\theta_0,\theta_0)^{-1}\widehat{g}_n(\theta_0), \qquad \theta\in\Theta,$$

and its Gaussian-population analogue

$$h_{n,P}(\theta) := \mathbb{G}_{n,P}(\theta) - \Sigma_P(\theta,\theta_0)\Sigma_P(\theta_0,\theta_0)^{-1}\mathbb{G}_{n,P}(\theta_0).$$

Because evaluation, matrix multiplication, and inversion are continuous on the set of nonsingular matrices, and Assumption OA.1(iv) guarantees nonsingularity of $\Sigma_P(\theta_0,\theta_0)$, the continuous mapping theorem gives

$$\Big(\widehat{g}_n(\theta_0), \widehat{h}_n(\cdot), \widehat{\Sigma}_n(\cdot,\cdot)\Big) \Rightarrow \Big(\mathbb{G}_{n,P}(\theta_0), h_{n,P}(\cdot), \Sigma_P(\cdot,\cdot)\Big)$$

uniformly over $P \in \mathcal{P}_n$.

Step 2: finite-sample Gaussian experiment and conditional nuisance elimination. Write

$$\mathbb{G}_{n,P}(\cdot) = \mu_{n,P}(\cdot) + \mathbb{Z}_{n,P}(\cdot),$$

where $\mu_{n,P}(\theta) := \mathbb{E}_P[g_{n,0}(\theta)]$ and $\mathbb{Z}_{n,P}(\cdot)$ is a mean-zero Gaussian process with covariance kernel

$$\operatorname{Cov}\Big(\mathbb{Z}_{n,P}(\theta), \mathbb{Z}_{n,P}(\vartheta)\Big) = \Sigma_P(\theta, \vartheta) - n^{-1}\mu_{n,P}(\theta)\mu_{n,P}(\vartheta)^\top.$$

Then

$$\begin{aligned} h_{n,P}(\theta) = \mu_{n,P}(\theta) - \Sigma_P(\theta, \theta_0)\Sigma_P(\theta_0, \theta_0)^{-1}\mu_{n,P}(\theta_0) \\ + \mathbb{Z}_{n,P}(\theta) - \Sigma_P(\theta, \theta_0)\Sigma_P(\theta_0, \theta_0)^{-1}\mathbb{Z}_{n,P}(\theta_0). \end{aligned}$$

Under the null, $\mu_{n,P}(\theta_0) = 0$. Hence

$$\operatorname{Cov}\Big(\mathbb{G}_{n,P}(\theta), \mathbb{G}_{n,P}(\theta_0)\Big) = \Sigma_P(\theta, \theta_0), \qquad \operatorname{Cov}\Big(\mathbb{G}_{n,P}(\theta_0), \mathbb{G}_{n,P}(\theta_0)\Big) = \Sigma_P(\theta_0, \theta_0).$$

Therefore, for every $\theta \in \Theta$,

$$\begin{aligned} \operatorname{Cov}\Big(h_{n,P}(\theta), \mathbb{G}_{n,P}(\theta_0)\Big) &= \operatorname{Cov}\Big(\mathbb{G}_{n,P}(\theta), \mathbb{G}_{n,P}(\theta_0)\Big) \\ &\quad - \Sigma_P(\theta, \theta_0)\Sigma_P(\theta_0, \theta_0)^{-1} \operatorname{Cov}\Big(\mathbb{G}_{n,P}(\theta_0), \mathbb{G}_{n,P}(\theta_0)\Big) \\ &= \Sigma_P(\theta, \theta_0) - \Sigma_P(\theta, \theta_0)\Sigma_P(\theta_0, \theta_0)^{-1}\Sigma_P(\theta_0, \theta_0) \\ &= 0. \end{aligned}$$

Hence every finite-dimensional restriction of $h_{n,P}(\cdot)$ is uncorrelated with $\mathbb{G}_{n,P}(\theta_0)$. Because the joint law is Gaussian, uncorrelatedness implies independence. This is the exact sufficient-statistic reduction used in the conditional-inference framework of Andrews and Mikusheva (2016).

Step 3: joint weak convergence of the feasible statistic and critical value. Define

$$\mathcal{Q}(g, \Sigma) := g(\theta_0)^\top \Sigma(\theta_0, \theta_0)^{-1} g(\theta_0) - \inf_{\theta \in \Theta} g(\theta)^\top \Sigma(\theta, \theta)^{-1} g(\theta),$$

and let $\mathcal{C}_{1-\alpha}(h, \Sigma)$ denote the $(1-\alpha)$-quantile of the conditional Gaussian QLR law generated after conditioning on the residual/sufficient process h and using covariance kernel Σ, where the randomness is over the Gaussian draw used in the conditional reconstruction. Then

$$QLR_n(\theta_0) = \mathcal{Q}(\widehat{g}_n, \widehat{\Sigma}_n), \qquad \widehat{c}_{1-\alpha}^{\mathrm{CQLR}}(\theta_0) = \mathcal{C}_{1-\alpha}\Big(\widehat{h}_n(\cdot), \widehat{\Sigma}_n(\cdot,\cdot)\Big).$$

The Gaussian benchmark objects are

$$QLR_{n,P}^{\mathrm{G}}(\theta_0) = \mathcal{Q}(\mathbb{G}_{n,P}, \Sigma_P), \qquad c_{1-\alpha,n,P}^{\mathrm{CQLR}}(\theta_0) = \mathcal{C}_{1-\alpha}\Big(h_{n,P}(\cdot), \Sigma_P(\cdot,\cdot)\Big).$$

Under the continuity condition in Andrews and Mikusheva (2016), both maps are almost surely continuous at the Gaussian limit object. Hence the continuous mapping theorem applied to

$$\Big(\widehat{g}_n(\cdot), \widehat{h}_n(\cdot), \widehat{\Sigma}_n(\cdot,\cdot)\Big) \Rightarrow \Big(\mathbb{G}_{n,P}(\cdot), h_{n,P}(\cdot), \Sigma_P(\cdot,\cdot)\Big)$$

yields

$$\Big(QLR_n(\theta_0), \widehat{c}_{1-\alpha}^{\mathrm{CQLR}}(\theta_0)\Big) \Rightarrow \Big(QLR_{n,P}^{\mathrm{G}}(\theta_0), c_{1-\alpha,n,P}^{\mathrm{CQLR}}(\theta_0)\Big)$$

uniformly over $P \in \mathcal{P}_n$. Therefore, with

$$T_n := QLR_n(\theta_0) - \widehat{c}_{1-\alpha}^{\mathrm{CQLR}}(\theta_0), \qquad T_{n,P}^{\mathrm{G}} := QLR_{n,P}^{\mathrm{G}}(\theta_0) - c_{1-\alpha,n,P}^{\mathrm{CQLR}}(\theta_0),$$

another application of the continuous mapping theorem gives

$$T_n \Rightarrow T_{n,P}^{\mathrm{G}}$$

uniformly over $P \in \mathcal{P}_n$.

Step 4: transfer the Gaussian size bound. The conditional-inference result of Andrews and Mikusheva (2016) implies

$$\Pr\Big(T_{n,P}^{\mathrm{G}} > 0\Big) \leq \alpha$$

uniformly over $P \in \mathcal{P}_n(\theta_0)$. Let $A_\varepsilon := [\varepsilon, \infty)$, which is closed. Since $\varepsilon > 0$, $A_\varepsilon \subset (0, \infty)$, and hence

$$\Pr\Big(T^{\mathrm{G}}_{n,P} \in A_\varepsilon\Big) \le \Pr\Big(T^{\mathrm{G}}_{n,P} > 0\Big) \le \alpha$$

uniformly over $P \in \mathcal{P}_n(\theta_0)$. Since

$$\{T_n > \varepsilon\} \subseteq \{T_n \in A_\varepsilon\},$$

the portmanteau theorem yields

$$\limsup_{n\to\infty} \sup_{P\in\mathcal{P}_n(\theta_0)} \Pr_P(T_n > \varepsilon) \le \limsup_{n\to\infty} \sup_{P\in\mathcal{P}_n(\theta_0)} \Pr_P(T_n \in A_\varepsilon) \le \sup_{P\in\mathcal{P}_n(\theta_0)} \Pr\Big(T^{\mathrm{G}}_{n,P} \in A_\varepsilon\Big) \le \alpha.$$

Equivalently,

$$\limsup_{n\to\infty} \sup_{P\in\mathcal{P}_n(\theta_0)} \Pr_P\Big(QLR_n(\theta_0) > \widehat{c}^{\,\mathrm{CQLR}}_{1-\alpha}(\theta_0) + \varepsilon\Big) \le \alpha.$$

If the law of $T^{\mathrm{G}}_{n,P}$ is continuous at zero uniformly over $P \in \mathcal{P}_n(\theta_0)$, apply the same closed-set argument with $A_0 := [0, \infty)$. Uniform continuity at zero gives

$$\Pr\Big(T^{\mathrm{G}}_{n,P} \in A_0\Big) = \Pr\Big(T^{\mathrm{G}}_{n,P} \ge 0\Big) = \Pr\Big(T^{\mathrm{G}}_{n,P} > 0\Big) \le \alpha$$

uniformly over $P \in \mathcal{P}_n(\theta_0)$. Since

$$\{T_n > 0\} \subseteq \{T_n \in A_0\},$$

the portmanteau theorem yields

$$\limsup_{n\to\infty} \sup_{P\in\mathcal{P}_n(\theta_0)} \Pr_P(T_n > 0) \le \alpha.$$

$\square$

Proposition OA.1 (CQLR reduces to CLR in homoskedastic Gaussian linear IV)**.** *In the homoskedastic Gaussian linear-IV model with one endogenous regressor and the standard linear-IV moments, the conditional QLR test based on the statistic above and Algorithm 1 coincides with Moreira's (2003) conditional likelihood ratio test, equivalently with the implementation in Andrews, Moreira, and Stock (2006).*

Proof. Let

$$Z_n := (Z_1, \ldots, Z_n)^\top \in \mathbb{R}^{n\times d_z}, \qquad Y_n := (Y_1, \ldots, Y_n)^\top, \qquad D_n := (D_1, \ldots, D_n)^\top,$$

and similarly $U_n := (U_1, \ldots, U_n)^\top$, $V_n := (V_1, \ldots, V_n)^\top$. Consider the homoskedastic Gaussian linear-IV model with one endogenous regressor,

$$Y_i = D_i\theta_0 + U_i, \qquad D_i = Z_i^\top \pi + V_i,$$

where, conditional on Z_n, the reduced-form errors (U_n, V_n) are jointly Gaussian, homoskedastic, and satisfy $E[U_n \mid Z_n] = E[V_n \mid Z_n] = 0$. The standard linear-IV moment condition is based on

$$\psi(W_i; \theta) = Z_i(Y_i - D_i\theta).$$

Conditional on Z_n,

$$g_n(\theta) = \frac{1}{\sqrt{n}} Z_n^\top (Y_n - D_n\theta) = \frac{1}{\sqrt{n}} Z_n^\top U_n + (\theta_0 - \theta)\frac{1}{\sqrt{n}} Z_n^\top (Z_n\pi + V_n).$$

Hence, conditional on Z_n, $g_n(\theta)$ is Gaussian for every fixed θ, and the process $\theta \mapsto g_n(\theta)$ is affine Gaussian conditional on Z_n. This is the exact finite-sample conditional Gaussian experiment underlying the classical CLR construction as in Moreira (2003) and Andrews, Moreira, and Stock (2006).

That affine representation can be written directly from the structural equations:

$$\begin{aligned}
g_n(\theta) &= \frac{1}{\sqrt{n}} Z^\top \Big(Y - D\theta\Big) \\
&= \frac{1}{\sqrt{n}} Z^\top \Big(D\theta_0 + U - D\theta\Big) \\
&= \frac{1}{\sqrt{n}} Z^\top U + (\theta_0 - \theta)\frac{1}{\sqrt{n}} Z^\top D \\
&= \frac{1}{\sqrt{n}} Z^\top U + (\theta_0 - \theta)\frac{1}{\sqrt{n}} Z^\top (Z\pi + V).
\end{aligned}$$

Hence

$$g_n(\theta) = a_n - b_n\theta, \qquad a_n := \frac{1}{\sqrt{n}} Z^\top Y, \quad b_n := \frac{1}{\sqrt{n}} Z^\top D,$$

so $\theta \mapsto g_n(\theta)$ is affine. Conditional on Z_n, every finite collection $(g_n(\theta_1), \ldots, g_n(\theta_m))$

is jointly Gaussian because it is a linear transformation of the conditional Gaussian reduced-form errors (U_n, V_n).

In this specialization, orthogonal-score estimation and oracle-score estimation coincide, so the feasible objects in Algorithm 1 reduce to their population Gaussian counterparts. In particular, the covariance kernel takes the form

$$\Sigma_n(\theta,\tilde{\theta}) := \operatorname{Var}\Big(g_n(\theta), g_n(\tilde{\theta})\Big) = \frac{1}{n}\operatorname{Var}\Big(Z^\top(Y-D\theta), Z^\top(Y-D\tilde{\theta})\Big),$$

and homoskedasticity implies that the dependence on θ and $\tilde{\theta}$ enters only through the scalar residual combinations $Y-D\theta$ and $Y-D\tilde{\theta}$, not through any additional nonparametric nuisance.

The conditional QLR construction in Algorithm 1 uses the transformed process

$$\widehat{h}_n(\theta) = \widehat{g}_n(\theta) - \widehat{\Sigma}_n(\theta,\theta_0)\widehat{\Sigma}_n(\theta_0,\theta_0)^{-1}\widehat{g}_n(\theta_0).$$

In the present model, replacing hats by their Gaussian linear-IV analogues gives

$$h_n(\theta) = g_n(\theta) - \Sigma_n(\theta,\theta_0)\Sigma_n(\theta_0,\theta_0)^{-1}g_n(\theta_0).$$

Because the pair $(g_n(\theta), g_n(\theta_0))$ is jointly Gaussian, this is exactly the linear projection residual from regressing $g_n(\theta)$ on $g_n(\theta_0)$. Therefore,

$$\operatorname{Cov}\Big(h_n(\theta), g_n(\theta_0)\Big) = 0.$$

For jointly Gaussian vectors, zero covariance implies independence, so $h_n(\theta)$ is independent of $g_n(\theta_0)$ for each fixed θ. More generally, the entire residual process $\{h_n(\theta) : \theta \in \Theta\}$ is independent of the sufficient statistic generated by $g_n(\theta_0)$ after conditioning on the classical reduced-form Gaussian experiment. This is the same orthogonal decomposition that underlies Moreira's conditioning argument.

To see the covariance cancellation directly, write

$$\begin{aligned}
\operatorname{Cov}\Big(h_n(\theta), g_n(\theta_0)\Big) &= \operatorname{Cov}\Big(g_n(\theta), g_n(\theta_0)\Big) - \Sigma_n(\theta,\theta_0)\Sigma_n(\theta_0,\theta_0)^{-1}\operatorname{Cov}\Big(g_n(\theta_0), g_n(\theta_0)\Big) \\
&= \Sigma_n(\theta,\theta_0) - \Sigma_n(\theta,\theta_0)\Sigma_n(\theta_0,\theta_0)^{-1}\Sigma_n(\theta_0,\theta_0) \\
&= 0.
\end{aligned}$$

Thus the conditional law used in Algorithm 1 is exactly the Gaussian law obtained after partialling out the sufficient statistic for the nuisance strength parameter. In the present Gaussian linear-IV model, this transformation is a one-to-one linear transform of the classical sufficient statistic for the nuisance reduced-form parameter governing first-stage strength. This is exactly the sufficient-statistic reduction underlying the CLR test of Moreira (2003) and the invariant formulation in Andrews, Moreira, and Stock (2006).

Next consider the statistic itself. The conditional QLR statistic is

$$QLR_n(\theta_0) = \widehat{g}_n(\theta_0)^\top \widehat{\Sigma}_n(\theta_0, \theta_0)^{-1} \widehat{g}_n(\theta_0) - \inf_{\theta \in \Theta} \widehat{g}_n(\theta)^\top \widehat{\Sigma}_n(\theta, \theta)^{-1} \widehat{g}_n(\theta).$$

In the Gaussian benchmark, define

$$Q_n^{\mathrm{G}}(\theta) := g_n(\theta)^\top \Sigma_n(\theta, \theta)^{-1} g_n(\theta), \qquad QLR_n^{\mathrm{G}}(\theta_0) := Q_n^{\mathrm{G}}(\theta_0) - \inf_{\theta \in \Theta} Q_n^{\mathrm{G}}(\theta).$$

Because $g_n(\theta) = a_n - b_n\theta$, the criterion $Q_n^{\mathrm{G}}(\theta)$ is the Gaussian quadratic form obtained by evaluating the log-likelihood ratio at θ after concentrating out the nuisance reduced-form parameter. With one endogenous regressor and any finite number of instruments, the Gaussian likelihood is a quadratic function of the reduced-form coefficient vector, so profiling over θ yields the likelihood-ratio statistic used in Moreira's CLR construction.

More concretely, under homoskedastic Gaussian linear IV the pair of reduced-form coefficient estimators can be rotated into the canonical statistics customarily denoted (S, T) in the CLR literature. The statistic S is sufficient for the nuisance parameter proportional to π, while T contains the information about θ_0 after conditioning on S; see Moreira (2003); Andrews, Moreira, and Stock (2006). The conditional QLR statistic above is invariant to this nonsingular linear reparameterization of the Gaussian experiment. After the rotation, Algorithm 1 computes exactly the same conditional quadratic-form comparison based on (S, T) that defines CLR.

Equivalently, the infimum over θ in

$$QLR_n^{\mathrm{G}}(\theta_0) = Q_n^{\mathrm{G}}(\theta_0) - \inf_{\theta \in \Theta} Q_n^{\mathrm{G}}(\theta)$$

plays the same role as profiling the Gaussian likelihood over the structural parameter under the null and under the unrestricted alternative. Since the Gaussian model is fully parametric and the orthogonal-moment criterion coincides with the Gaussian score-based quadratic form, no gap remains between the conditional QLR objective and the classical likelihood-ratio objective. Under homoskedastic Gaussian linear IV with one endogenous regressor, the covariance kernel does not introduce any nonlinear nuisance structure beyond the classical concentration/strength parameter, and the Gaussian quasi-likelihood ratio above is exactly the Gaussian likelihood-ratio statistic used by the CLR procedure. This identification is the linear-IV specialization established in the conditional Gaussian framework of Andrews and Mikusheva (2016).

Therefore, both ingredients of the procedure coincide with CLR:

(i) the conditioning statistic $\widehat{h}_n$ is equivalent to the classical sufficient statistic for first-stage strength;

(ii) the objective function defining QLR_n is the CLR likelihood-ratio statistic.

Consequently, the simulated conditional critical value produced by Algorithm 1 coincides with the CLR critical value, and the resulting conditional QLR test coincides with Moreira's (2003) CLR test, equivalently with the implementation in Andrews, Moreira, and Stock (2006). □

OA.3 Worked Verification for Sparse Linear Nuisances

This section verifies the nuisance-rate and library-complexity conditions for a standard sparse-linear learner class. The verification makes the main paper's curated-library conditions operational in a standard sparse-linear setting.

Proposition OA.2 (Verification of nuisance-rate and screening conditions under lasso)**.** *Let $X \in \mathbb{R}^{p_n}$. For each nuisance component $q \in \{\mu, m, r_1, \ldots, r_{d_z}\}$, suppose*

$$q_0(x) = x^\top \beta_{0,q}, \qquad \|\beta_{0,q}\|_0 \le s_n,$$

and suppose the regression errors are uniformly sub-Gaussian and the design satisfies a uniform restricted-eigenvalue condition. Let the candidate library be a finite grid of lasso penalties Λ_n with $\log|\Lambda_n| = O(\log n)$. Suppose there exists a subset $\Lambda_n^{\text{good}} \subseteq \Lambda_n$

of theoretically valid penalty levels such that for every $\lambda \in \Lambda_n^{\text{good}}$,

$$\|\widehat{q}_k(\lambda) - q_0\|_{L_2(P)} = O_P\left(\sqrt{\frac{s_n \log p_n}{n}}\right)$$

uniformly over folds k. If

$$s_n \log p_n = o(n^{1/2}),$$

then

$$\sup_{k,\lambda \in \Lambda_n^{\text{good}}} \|\widehat{q}_k(\lambda) - q_0\|_{L_2(P)} = o_P(n^{-1/4}).$$

If, in addition, the prediction losses and the component products entering $\widehat{\Pi}(\gamma)$ and $\widehat{\Sigma}_Z(\gamma)$ have uniformly sub-exponential tails, then

$$\sup_{\gamma \in \Gamma_n} \left|\widehat{R}_{\text{pred}}(\gamma) - R_{\text{pred}}(\gamma)\right| = O_P\left(\sqrt{\frac{\log |\Gamma_n|}{n_{\text{sel}}}}\right),$$

and

$$\sup_{\gamma \in \Gamma_n} \left|\overline{S}_{n,\kappa}(\gamma) - \overline{S}_{n,\kappa}(\eta(\gamma))\right| = O_P\left(\sqrt{\frac{\log |\Gamma_n|}{n_{\text{sel}}}}\right),$$

for the finite curated library Γ_n built from Λ_n^{good}.

Proof. The first claim is the standard lasso prediction-rate bound under uniform restricted-eigenvalue conditions and sub-Gaussian errors; see Bühlmann and van de Geer (2011) and Belloni, Chen, Chernozhukov, and Hansen (2012). Uniformity over the finite penalty subset Λ_n^{good} follows from a union bound. Under $s_n \log p_n = o(n^{1/2})$, the resulting rate is $o_P(n^{-1/4})$.

For the prediction criterion, Bernstein inequalities applied to the selection-fold loss averages and a union bound over the finite library give the displayed uniform concentration of $\widehat{R}_{\text{pred}}$. For the strength criterion, the Bernstein step is applied to the component averages

$$\widehat{\Pi}(\gamma) = n_{\text{sel}}^{-1} \sum_{i \in I^{\text{sel}}} \widehat{Z}_i^e(\gamma) \widehat{D}_i^e(\gamma), \qquad \widehat{\Sigma}_Z(\gamma) = n_{\text{sel}}^{-1} \sum_{i \in I^{\text{sel}}} \widehat{Z}_i^e(\gamma) \widehat{Z}_i^e(\gamma)^\top.$$

Under the stated tail conditions, uniformly over $\gamma \in \Gamma_n$,

$$\|\widehat{\Pi}(\gamma) - \Pi(\eta(\gamma))\| + \|\widehat{\Sigma}_Z(\gamma) - \Sigma_Z(\eta(\gamma))\|_{\mathrm{op}} = O_{\mathbb{P}}\left(\sqrt{\frac{\log|\Gamma_n|}{n_{\mathrm{sel}}}}\right).$$

On the event, whose probability tends to one under the stated eigenvalue condition, that $\lambda_{\min}(\Sigma_Z(\eta(\gamma)) + \kappa I)$ is uniformly bounded away from zero, the map

$$(v, M) \mapsto v^\top (M + \kappa I)^{-1} v$$

is locally Lipschitz on the relevant compact neighborhood. The uniform component concentration therefore propagates to

$$\sup_{\gamma \in \Gamma_n} \left|\overline{S}_{n,\kappa}(\gamma) - \overline{S}_{n,\kappa}(\eta(\gamma))\right| = O_{\mathbb{P}}\left(\sqrt{\frac{\log|\Gamma_n|}{n_{\mathrm{sel}}}}\right).$$

□

OA.4 Supplementary Technical Notes

Remark OA.2 (Heteroskedasticity-weighted strength as an optional extension)**.** The main paper uses the feasible effective-concentration proxy $\widehat{S}_{n,\kappa}$ because it remains operational under weak identification without requiring a uniformly consistent estimator of θ_0. If a leakage-free anchor value θ_{ref} is available from the training sample, a heteroskedasticity-weighted refinement can be built by replacing the unweighted residualized first-stage covariance with a covariance weighted by the orthogonalized structural residual $U(\theta_{\mathrm{ref}}, \eta) = Y - \mu(X) - \theta_{\mathrm{ref}}(D - m(X))$. This refinement is natural but not needed for the main theorem package.

Remark OA.3 (Alternative sufficient conditions for ridge AR)**.** The effective-rank condition in the main paper is a transparent sufficient condition rather than a sharp frontier. The quadratic-form bootstrap argument may also be verified under alternative trace-ratio conditions used in the high-dimensional quadratic-form literature. The main text therefore emphasizes the effective-rank regime because it is dimension-agnostic and easy to interpret.

OA.5 Additional Monte Carlo Diagnostics

This section records two supplementary diagnostics for the Monte Carlo designs used in the main paper. The first table reports empirical null rejection rates for the full baseline homoskedastic PLIV design. The second table reports a representative weak-identification cell from the heteroskedastic baseline design. These results are supplementary: the stylized toy experiment in the main paper remains the cleanest finite-sample illustration of the leakage proposition, and the proxy-rich stress design remains the cleanest illustration of strength killing.

Finite-sample null-size calibration in the full baseline PLIV design is reported in Table 1. At $n = 500$, the leaky benchmark is clearly distorted, with rejection rates between 0.156 and 0.179 across the π grid, whereas the two leakage-free procedures lie between 0.075 and 0.077. At $n = 2000$, the leaky global procedure is numerically closer to nominal than the two leakage-free procedures in this benign baseline design. The finite-sample message is therefore not that the leaky rule is uniformly worse at every sample size, but that leakage can generate first-order distortions and that LF–NCF trades some finite-sample splitting cost for guaranteed no-own-observation reuse.

Table 1: Baseline homoskedastic PLIV design: empirical null rejection rates

Panel A: $n = 500$

π	LF–NCF (pred-only)	LF–NCF + IACV	Leaky global
0	0.075	0.076	0.179
1	0.075	0.077	0.172
3	0.075	0.077	0.156
10	0.075	0.076	0.156
30	0.075	0.076	0.158

Panel B: $n = 2000$

π	LF–NCF (pred-only)	LF–NCF + IACV	Leaky global
0	0.077	0.083	0.071
1	0.077	0.084	0.071
3	0.077	0.083	0.070
10	0.077	0.083	0.070
30	0.077	0.083	0.071

Notes: Null rejection rates are evaluated at the nominal 5 percent level in the full baseline homoskedastic PLIV design. The leaky benchmark is included only as a finite-sample calibration diagnostic; the stylized Gaussian experiment in the main text remains the cleanest mechanism-isolating illustration of leakage.

Table 2 reports the heteroskedastic baseline design. The qualitative message is unchanged. When the design leaves little room for systematic strength killing, LF–NCF+IACV and LF–NCF with prediction-only tuning are nearly indistinguishable in power, strength, and confidence-set geometry. This mirrors the benign homoskedastic baseline and confirms that the gains from IACV are not mechanical.

Table 2: Heteroskedastic baseline design: representative weak-IV cell

Method	Power	(s.e.)	Avg. $\widehat{S}_{n,\kappa}$	Grid-boundary freq.	Avg. CI length
LF–NCF + IACV	0.102	0.007	14.390	0.931	1.629
LF–NCF (pred-only)	0.101	0.007	14.350	0.933	1.630

Notes: Representative cell in the heteroskedastic baseline design. The two rows are both leakage-free; the absence of separation confirms that IACV is approximately neutral when the candidate library leaves little scope for strength killing.

OA.6 Empirical Implementation Details and Enriched-Control AK Evidence

Table 3 reports an enriched-control implementation of the AK91 cohort-V design using the same three procedures studied in the paper: leaky full-sample predictive tuning, LF–NCF with prediction-only tuning, and LF–NCF + IACV. The table shows how the tuning architecture shifts the selected strength proxy and the associated weak-identification-robust p-values in a demanding empirical specification.

The sample is cohort V from the official AK91 archive, that is, men born 1930–39 in the 1980 Census extract (Angrist and Krueger, 1991; Angrist, n.d.). The outcome is log weekly wage and the endogenous regressor is years of schooling. The excluded instruments are the 30 year-by-quarter dummies used in the classical AK91 benchmark. The baseline exogenous controls are year-of-birth dummies, age, age squared, race, marital status, SMSA status, and the eight region indicators.

Let X^{base} denote the baseline exogenous controls. The enriched nuisance dictionary X^{rich} consists of X^{base} together with the following predetermined transformations: year-of-birth by region interactions, age by year-of-birth interactions, age by region interactions, age-squared by year-of-birth interactions, race by region interactions, SMSA by region interactions, and marital-status by region interactions. No excluded instrument or transformation involving an excluded instrument is included in X^{rich}.

The nuisance functions are estimated by sparse linear learners. For each nuisance component, the candidate library is the lasso-penalty grid

$$\Lambda_n = \{0.25,\, 0.5,\, 1,\, 2,\, 4\} \times \lambda_{\mathrm{plugin}},$$

and the same penalty scale grid is used for the outcome, treatment, and instrument nuisance regressions. Instrument residualization is implemented componentwise. The empirical pipeline therefore matches the sparse-linear learner-class verification in Section OA.3. The implementation mirrors the paper's theoretical architecture: leakage-free nested cross-fitting separates selection, estimation, and inference samples; the prediction-quality screen uses a one-standard-error rule; and the strength proxy is

$$\widehat{S}_{n,\kappa}(\gamma) = n_{\mathrm{sel}} \widehat{\Pi}(\gamma)^{\top} \left(\widehat{\Sigma}_Z(\gamma) + \kappa I\right)^{-1} \widehat{\Pi}(\gamma),$$

with

$$\kappa = \kappa_0 \frac{\mathrm{tr}(\widehat{\Sigma}_Z)}{d_z}, \qquad \kappa_0 = 10^{-4}.$$

For the enriched-control implementation reported in Table 3, the implementation uses $K_{\mathrm{outer}} = 2$ outer folds, $K_{\mathrm{inner}} = 2$ inner folds, $n_{\mathrm{reps}} = 10$ repeated random outer partitions, $B = 2000$ conditional QLR simulations, and an adaptive AR inversion grid initialized on $[-0.5, 0.5]$ and allowed to expand up to $[-5, 5]$. For computational feasibility in this implementation, the working sample is capped at $n = 2000$. The reported quantities are the median orthogonal PLIV point estimate, the median AR and CQLR p-values, the median selected strength proxy, the boundary-touch share of AR confidence sets under adaptive grid expansion, and the median length of bounded AR sets when such sets occur.

OA.6.1 Enriched-control AK evidence

Table 3 reports an enriched-control implementation of the AK91 cohort-V design using the same three procedures studied in the paper: The leaky full-sample procedure tunes and evaluates on overlapping observations and is reported as the empirical counterpart of the score-evaluation reuse mechanism analyzed in the main text. The evidence is directionally consistent with the paper's strength-preservation mechanism. Relative to LF–NCF with prediction-only tuning, LF–NCF + IACV raises the median selected strength proxy from 366.7 to 425.3, increases the median point estimate from 0.0609 to 0.0752, lowers the median AR p-value from 0.5324 to 0.3393, and lowers the median CQLR p-value from 0.6699 to 0.3741. Those directional changes match the paper's identification-aware tuning mechanism.

At the same time, the AR confidence set reaches the numerical search boundary in every replication for all three procedures, even after adaptive grid expansion. This severe weak-identification geometry should be interpreted in light of the computationally restricted working sample, which is capped at $n = 2000$, combined with the many-quarter instrument set. The original AK design relies on much larger samples to accumulate first-stage information. Within this deliberately small working sample, the tuning architecture nevertheless shifts the selected strength proxy and the associated robust p-values in the direction predicted by the theory.

Table 3: AK91 cohort V: enriched-control implementation

Procedure	med. $\widehat{\theta}$	med. AR p	med. CQLR p	med. $\widehat{S}_{n,\kappa}$	boundary-touch share	med. bounded length
Leaky full-sample tuning	0.0566	0.7483	0.7561	334.5182	1.0000	–
LF–NCF (pred-only)	0.0609	0.5324	0.6699	366.6709	1.0000	–
LF–NCF + IACV	0.0752	0.3393	0.3741	425.3318	1.0000	–

Notes: Sample is the AK91 cohort-V extract (men born 1930–39 in the 1980 Census extract), with the 30 year-by-quarter dummies as excluded instruments. The nuisance dictionary augments the baseline AK controls with the enriched interactions described in the main paper. Entries are medians across repeated random outer partitions. This implementation run uses $n_{\text{reps}} = 10$, $K_{\text{outer}} = 2$, $K_{\text{inner}} = 2$, $B = 2000$ conditional QLR simulations, and an adaptive AR inversion grid initialized on $[-0.5, 0.5]$ and allowed to expand up to $[-5, 5]$, with the working sample capped at $n = 2000$. "Boundary-touch share" denotes the fraction of replications for which the AR confidence set reached the numerical search boundary; because this occurred in every replication, no bounded-set length is reported. The leaky full-sample procedure is included only as a comparison benchmark and is not covered by the paper's validity theory.

References

Andrews, D. W. K., M. J. Moreira, and J. H. Stock (2006). Optimal two-sided invariant similar tests for instrumental variables regression. *Econometrica* 74(3), 715–752.

Andrews, I. and A. Mikusheva (2016). Conditional inference with a functional nuisance parameter. *Econometrica* 84(4), 1571–1612.

Angrist, J. D. and A. B. Krueger (1991). Does Compulsory School Attendance Affect Schooling and Earnings? *Quarterly Journal of Economics* 106(4), 979–1014.

Angrist, J. D. (n.d.). Angrist Data Archive. `https://economics.mit.edu/people/faculty/josh-angrist/angrist-data-archive`. MIT Economics. Accessed March 2026.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica* **80**(6), 2369–2429.

Bühlmann, P. and S. van de Geer (2011): *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer.

Moreira, M. J. (2003). A conditional likelihood ratio test for structural models. *Econometrica* 71(4), 1027–1048.

Pouzo, D. (2015): "Bootstrap Consistency for Quadratic Forms of Sample Averages with Increasing Dimension," *Electronic Journal of Statistics* 9, 1273–1307.

van der Vaart, A. W. and J. A. Wellner (1996): *Weak Convergence and Empirical Processes*. Springer.