

The Anatomy and Boundary of Adaptation under Temporal Tabular Shift

Tianyu Wang* Xi Vincent Wang† Lihui Wang‡ Mian Li§ Zhihao Liu¶

September 2026

Abstract

Prequential adaptation of frozen tabular foundation models under temporal drift, with each label revealed only after prediction, helps some deployments and harms others, yet current practice does not predict which. We study the sources and limits of these gains. A diagnostic anatomy attributes gains to four recurring mechanisms under a streaming protocol that removes three optimistic biases and quantifies a fourth. Within an agnostic total-variation drift class, the target conditional is only partially identified: its identified-set diameter, the *wall*, is irreducible from unlabeled data uniformly in sample size. A second, orthogonal L^2 projection wall quantifies what the frozen representation cannot express. Two canonical mechanism priors collapse the first wall. Under stated nuisance-rate conditions, the wall can be estimated from labeled historical windows at a $\sqrt{N}$ rate above the margin threshold $\gamma^* = d_0/(2\alpha_s)$. At $\gamma = 0$, the conditional lower-bound program depends on an open affinity estimate; the positive-margin lower branch also remains open. Semi-synthetic data illustrate the finite-sample mechanism with calibrated exponents. Stream-level proxies on eight industrial streams fall on the difficult side under a stated roughness bound, while the equality case $\gamma = \gamma^*$ remains unresolved.

1 Introduction

A frozen tabular foundation model (Hollmann et al., 2025; Qu et al., 2025) deployed on an industrial data stream can degrade as the world drifts. Its operator must repeatedly decide whether to adapt (correct the model online from whatever the stream reveals) or to freeze and wait for retraining. The empirical record for adaptation is mixed. In our case study on the eight industrial streams of the TabReD benchmark (Rubachev et al., 2025), streaming residual correction improves a frozen model on seven (by +0.59% to +3.94% RMSE, or up to +0.86 AUC points), while on the eighth it costs 2.28 AUC points. These outcomes motivate a diagnostic question that method rankings alone do not answer.

Three information states recur throughout, and each headline result is stated relative to one of them. Before a window’s labels arrive, the operator holds only the unlabeled stream and the

*Department of Production Engineering, KTH Royal Institute of Technology, Stockholm, Sweden. Email: tianyuwa@kth.se. Corresponding author.

†Department of Production Engineering, KTH Royal Institute of Technology, Stockholm, Sweden. Email: wangxi@kth.se.

‡Department of Production Engineering, KTH Royal Institute of Technology, Stockholm, Sweden. Email: lihuiw@kth.se.

§Global Institute of Future Technology, Shanghai Jiao Tong University, Shanghai, China. Email: mianli@sjtu.edu.cn.

¶Department of Production Engineering, KTH Royal Institute of Technology, Stockholm, Sweden. Email: zhihao1@kth.se.

frozen model; after each prediction, the revealed label joins a strictly-past prefix that supports prequential correction; and occasional fully labeled historical windows support learning where the boundary itself lies. Every measured gain in this paper belongs to the second state, the identification impossibility (the identified-set diameter, or Wall A, Section 6.1) to the first, and the margin-boundary theory of Part II to the third; the representation deficit (Wall B) binds in all three, with or without labels.

Method rankings presuppose that the shift is correctable and tend to treat failure as method immaturity. We ask a more basic question: *of the degradation a frozen model suffers under temporal tabular shift, how much is correctable at all, from what information, and at what price?* We answer this question in three parts.

The anatomy (Part I) attributes the evaluated system’s gains to four recurring gain mechanisms (global debias, freshness, local re-estimation in representation space, ranking recalibration) plus two canonical identification-collapse classes, covariate and label-marginal response. We do not assert these to be a unique additive decomposition of every adaptation algorithm; they are a mechanism-oriented audit of this deployment protocol. At matched sample size on one stream, the *oldest* labels degrade the corrector (-1.1%) while the newest improve it ($+4.9\%$): a staleness penalty in that case. On another stream, 93% of the corrector’s measured gain is reproduced by a running mean. Such examples show why we report gains with channel attribution and time-respecting, leak-free oracles. Part I therefore also quantifies four optimistic evaluation biases; two of them (the retrieval leak and the deduplication) reverse conclusions that the uncorrected protocol supports.

The boundary (Part II) uses partial identification to explain why the correctable residual signal measured in Part I can be small (Manski, 2003). Within one unlabeled window the drifted conditional is identified only up to a set whose diameter (the *wall*) reduces exactly to an averaged pointwise budget. We prove the wall irreducible: no unlabeled statistic shrinks it, uniformly in sample size. The wall is one of *two* obstructions, orthogonal in the projection geometry rather than statistically independent. The second is the information the frozen representation discarded, and the two prescribe different remedies (labels versus a new embedding). For a fixed number K of labeled historical windows holding N anchor labels in total, we establish a $\sqrt{N}$ -regular estimator for learning the wall from that history when the margin exponent γ exceeds $\gamma^* = d_0/(2\alpha_s)$ (intrinsic dimension d_0 , drift smoothness α_s), under stated nuisance-rate conditions. On the other side, the lower-bound program is conditional at $\gamma = 0$ on one open affinity estimate and remains open for $0 < \gamma < \gamma^*$. The equality case is not covered by the current log-rate argument. On eight TabReD streams, noisy stream-level proxies lie below the threshold under the rough-field bound $\alpha_s \leq 1$. For fifteen additional benchmarks without drift labels, we report only a sensitivity analysis under $(\alpha_s, \gamma) = (1, 1)$; these points are not data-set-specific phase measurements.

Beyond the boundary (Part III) develops a local-correction screening rule that needs no labels beyond those already consumed by the gated corrector. Retrospectively, the rule detects our harm case (HOMECREDIT-DEFAULT, Section 4.2). This part also gives an attribution procedure that splits residual error into Wall-A, Wall-B, and correctable shares, and it derives the rate at which anchors reduce the wall. The latter turns an unspecified need for more labels into a labeling requirement.

Our contributions follow the three parts of that question.

1. **What is correctable.** A diagnostic anatomy of prequential tabular adaptation: four measured gain mechanisms plus two identification-collapse classes, with per-stream attribution on eight industrial streams and a mechanistic harm case (Section 4). Its evidence layer is a measurement methodology that quantifies four optimistic biases in streaming evaluation.

The protocol removes future-neighbor oracle leakage (up to $4.5\times$, with a sign flip on one stream), temporal-twin noise floors, and duplicate-deflated intrinsic dimension. The fourth bias arises when a safety gate monitors one loss while harm accrues in another. We show that this mismatch is structural rather than tunable and measure its cost (Sections 5 and 7.1).

2. **What limits the remainder.** A partial-identification boundary: the exact reduction identity, the irreducibility theorem, the collapse classes, and the two-wall attribution (geometrically orthogonal components) with a deployable label-free wall estimand (Section 6); and a margin-indexed price for learning the wall from labeled history: a proved $\sqrt{N}$ sufficient branch above γ^* , a conditional $\gamma = 0$ lower-bound program below it, and an open positive-margin lower branch. Semi-synthetic data illustrate the sign-recovery mechanism with calibrated exponents; a proxy-and-sensitivity study covers twenty-three data sets (Sections 6.4–6.6).
3. **What to do about it.** Deployment diagnostics: the local-correction screening rule, evaluated on all eight streams with a one-sided error profile; the degradation-attribution procedure; and the price of labels (Section 7).

Section 2 situates this agenda in the five literatures it touches (empirical test-time adaptation, tabular foundation models, distribution-shift theory, representation-level limits, and drift monitoring) and in the two it borrows its formal tools from, partial identification and semiparametric efficiency theory. Section 3 then fixes the setting. The three parts occupy Sections 4–5 (Part I), Section 6 (Part II), and Section 7 (Part III).

Empirically, adaptation can help, harm, or reduce to a simple correction. The theory identifies which limitations follow from the stated drift and representation assumptions. For each claim, we state both the supporting evidence and the assumptions it requires.

2 Related Work

Test-Time Adaptation Practice. The modern test-time-adaptation (TTA) literature adapts a deployed model from the unlabeled test stream itself. That label-free constraint delimits the wall theory of Sections 6.1–6.3 rather than our measured system, which is prequential and consumes each label after its prediction (Section 3). Sun et al. (2020) update the model on each test input through an auxiliary self-supervised task before predicting; Wang et al. (2021) minimize prediction entropy, updating only normalization statistics and channel-wise affine parameters; Wang et al. (2022) extend the setting to continually changing targets, countering the error accumulation and forgetting that self-training induces with averaged pseudo-labels and stochastic weight restoration. These studies are mostly empirical and vision-centered, using corruption and style-shift benchmarks. On a large controlled benchmark, Zhao et al. (2023) show that reported gains are sensitive to hyperparameter and model selection, which is difficult without labels, and that well-configured methods still fail on some shift types.

Direct tabular TTA methods address the feature and shift structure that vision methods miss. TabLog learns adaptable logical rules (Ren et al., 2024), while AdapTable combines uncertainty calibration with target-label-distribution adjustment (Kim et al., 2024). With a small labeled target sample, Zeng et al. (2024) study representation choice under tabular $Y \mid X$ shift. These methods provide algorithmic baselines but do not ask which parts of conditional drift are determined by the time-respecting information available to a frozen deployment. Section 4.1 supplies the diagnostic taxonomy. Section 6 separates the limits of identification from those of the frozen representation. The attribution procedure in Section 7.2 then distinguishes method immaturity, channel exhaustion, and unidentifiability, which require different remedies.

Tabular Foundation Models. Frozen in-context tabular learners are the setting our theory takes as given. TabPFN (Hollmann et al., 2023) is a prior-fitted transformer that classifies small tables in a single forward pass, approximating Bayesian inference under a synthetic prior. Its foundation-model successor (Hollmann et al., 2025) outperforms tuned baselines on small-to-medium tables, and TabICL (Qu et al., 2025) scales tabular in-context learning to hundreds of thousands of rows. These models are distributed as frozen artifacts, and their context is the only adaptation interface. This is the fixed-embedding regime studied here (Section 3). Our measurements use the frozen backbone of a companion system built on this model family (Wang et al., 2026) (Section 3). For evaluation, TabReD (Rubachev et al., 2025) provides industrial streams with temporal splits on which method rankings reorder. Cai and Ye (2025) document the gap between i.i.d. and temporal evaluation for deep tabular methods. Drift-Resilient TabPFN (Helli et al., 2024) incorporates a drift prior into pretraining. This is complementary to our account, which takes whatever frozen model is deployed as given and asks what any post-hoc procedure can still recover from it.

Distribution-Shift Theory: Bounds versus Identification. Classical domain-adaptation theory controls target risk through computable divergences: Ben-David et al. (2010a) bound the target error by the source error plus a divergence between the two covariate laws that finite unlabeled samples estimate, plus the error of the best joint hypothesis; Mansour et al. (2009) generalize the divergence to arbitrary loss classes via the discrepancy distance. Covariate-shift correction operates under the complementary assumption that the conditional is invariant, so Shimodaira (2000) reweights the likelihood by the density ratio of the covariate laws, and Sugiyama et al. (2007) carry the same weighting into model selection. The taxonomy of data set shift (Quiñonero-Candela et al., 2009; Moreno-Torres et al., 2012) names the case both lines exclude: concept shift, where $p(y | x)$ itself moves. This machinery does not answer our question. Its label-free quantities are functionals of the covariate laws, whereas concept drift appears only in terms that unlabeled data do not determine: the joint-error term in the bounds or the invariance assumption behind reweighting. A bound conditions on that term. A deployment decision instead requires knowing whether the observable data determine the drifted conditional. Ben-David et al. (2010b) prove impossibility theorems for unsupervised domain adaptation, and Ben-David and Uner (2012) quantify the hardness without target labels. Those results do not measure the remaining ambiguity. Section 6 does: within the agnostic drift class, no unlabeled statistic shrinks the set of conditionals consistent with the data, uniformly in sample size (Theorem 6.2). We therefore treat identification as the primary question. The *wall* is the diameter of what the label-free observables leave undetermined, rather than another divergence to estimate.

Partial Identification and Label Shift. Partial identification in econometrics provides the relevant framework. Manski (2003) develops the program of reporting parameter values consistent with observable data and stated assumptions without forcing point identification; Molinari (2020) surveys its development. Our wall adapts a Manski-style identified set to frozen representations under temporal shift. The estimand is the set’s diameter, a function on covariate space (Section 6), and the observations arrive in streaming windows rather than in a single survey. Where Kong et al. (2022) impose a latent causal model whose changing components are themselves only partially identifiable, and use that partial identification to determine the target joint distribution, we retain the set of drifted conditionals and study its diameter. The label-shift literature is the collapse case. Lipton et al. (2018) recover the shifted label marginal from a black-box predictor’s confusion matrix, and Garg et al. (2020) unify the moment and likelihood variants of that estimation. The two drift classes studied here recover these identification mechanisms: covariate shift determines the conditional by assumption, while label shift identifies its finite-dimensional mixture weights. We use them as canonical collapse examples, not as an exhaustive characterization of every structural prior that could identify drift.

Representation-Level Limits. Our second obstruction is the information the frozen representation itself discarded, formalized as Wall B in Section 6.3. It instantiates the information-theoretic lower bounds of Zhao et al. (2019) (who show that enforcing invariant representations under shifted label marginals forces a floor on joint error) and the support/invertibility analysis of Johansson et al. (2019), as a per-window, label-free-deployable sufficiency deficit. Previous analyses of adaptation failure can conflate information lost by the representation with information that no representation could determine. Section 6.3 separates these quantities exactly (Theorem 6.4). The attribution procedure in Section 7.2 uses that separation to distinguish when labels are needed from when the embedding must change.

Semiparametric Efficiency and Non-Smooth Functionals. The margin-boundary analysis in Section 6.4 draws on two literatures. The proved regular branch ($\sqrt{N}$ rates for the wall as a functional of nuisance conditionals) builds on efficient-influence-function and debiased-machine-learning machinery (Chernozhukov et al., 2018; Kennedy, 2024), with the classical semiparametric backdrop of Bickel et al. (1993) and van der Vaart (1998). The conditional lower-bound program draws on the literature on non-smooth functionals. Lepski et al. (1999) study L_r norms of a regression function, and Cai and Low (2011) give composite-hypothesis lower bounds. The white-noise equivalences of Nussbaum (1996) and Brown and Low (1996) connect those results to our below-threshold analysis. The margin parameter that indexes our threshold is the Mammen–Tsybakov margin condition (Tsybakov, 2004; Audibert and Tsybakov, 2007) in a new role. There it governs classification rates; here it governs whether the *wall* (an absolute-value functional of drift) is smooth or non-smooth at the relevant scale. The margin-indexed threshold at $\gamma^* = d_0/(2\alpha_s)$ (the threshold of Section 6.4, combining the margin exponent with the drift smoothness α_s and the intrinsic dimension d_0 of the embedded data) and the conditional coherent-adversary construction for multi-window pooling appear to be new. The latter remains conditional on the affinity estimate identified in Appendix K of Online Appendix 1 (gap (2c), Section 6.4).¹

Drift Monitoring and Evaluation Practice. Our diagnostics build on the monitoring literature surveyed by Gama et al. (2014), whose detectors and adaptation strategies largely presume that labels arrive promptly enough to score the stream. Our local-correction screening rule (Section 7.1) is instead near-label-free by construction: it asks for no labels beyond those the corrector it gates already consumes, and the irreducibility result (Theorem 6.2) tells the operator which of its readings are measurements and which restate the assumed drift budget. Label-free risk monitors for test-time adaptation (Schirmer et al., 2025) are loss-denominated, as are the game-theoretic sequential tests they can be built on (Shafer, 2021; Ramdas et al., 2023). Our currency-mismatch result in Section 5.5 concerns safety gates scored in a loss the deployment does not rank by. It identifies what such monitors can detect and documents a failure in practice. Conformal prediction under covariate shift (Tibshirani et al., 2019) gives distribution-free coverage when only the covariate law moves and the likelihood ratio is estimable. That guarantee depends on conditional invariance, which our setting does not assume. Adaptive conformal inference (Gibbs and Candès, 2021) drops the invariance assumption and tracks arbitrary drift online, but its target is marginal coverage of prediction sets, not the attribution and identification questions asked here. The local-correction screening rule instead accounts for conditional drift through the drift budget. Our future-neighbor leak (Section 5.1) is the retrieval-oracle analogue of the look-ahead biases long managed in finance by purged cross-validation (López de Prado, 2018). Across the eight measured streams, 49.5–53.2% of the batch oracle’s retrieved neighbors come from the future, and removing them can reverse the sign of a signal ceiling. The four-bias protocol of Section 5 applies the same safeguards to noise

¹Online Appendix 1 is the supplementary document that accompanies this arXiv submission as the ancillary file `anc/supplementary.pdf`; it carries the proof appendices A–K under their original numbering, their full developments, and the status of every statement.

floors, intrinsic dimension, and safety gating. The prequential evaluation discipline itself is classical (Dawid, 1984). Online learning with expert advice (Cesa-Bianchi and Lugosi, 2006) operates on the same revealed-label stream and bounds regret against a comparator class without distributional assumptions; a regret bound says how well a corrector tracks its best comparator, not whether the information needed to correct exists at all. The latter is the identification question of Section 6.

3 Setting, Protocol, and Measurement Conventions

Model and Representation. Covariates live in a Polish space $\mathcal{X}$ and labels in $\mathcal{Y}$, binary $\{0, 1\}$ for classification (the primary case for our identification results) or real-valued for regression. A frozen pretrained tabular foundation model supplies a representation $\phi : \mathcal{X} \rightarrow \mathbb{R}^D$ (ambient width $D = 192$; the intrinsic dimension d of (A3), estimated as d_0 , is far smaller) and a readout h_0 , jointly defining the model’s source conditional $p_0(\cdot | x) = h_0(\phi(x))$ with mean $\eta_0(x)$. Neither ϕ nor h_0 is retrainable at deployment, so every deployed prediction is $\mathcal{F}_\phi$ -measurable, $\mathcal{F}_\phi = \sigma(\phi)$, a constraint with consequences (Theorem 6.5).

Stream and Prequential Protocol. Deployment data arrive as a temporally ordered stream $x_1, x_2, \dots$. The model predicts each x_t , the label y_t is revealed afterwards, and any corrector may use only the strictly-past labeled prefix. All adaptation gains in this paper are computed under this prequential discipline (Dawid, 1984). A *window* W is a contiguous segment of the stream with covariate law p_W and (unobserved) conditional $p_W(\cdot | x)$.

Observables and the Drift Budget. Within a window, the label-free information is: (O1) the unlabeled draws from p_W , hence any functional of the covariate law, including density ratios against the source; (O2) the frozen pair (ϕ, h_0) , hence $p_0(\cdot | x)$ pointwise; and optionally (O3) cross-window structure such as recurring regimes. The sole assumption linking the drifted conditional to the model is a total-variation *budget* $\beta : \mathcal{X} \rightarrow [0, 1]$, $p_W(\cdot | x) \in B_\beta(x) = \{q : D_{\text{TV}}(q, p_0(\cdot | x)) \leq \beta(x)\}$, with no coupling across covariates and no assumed drift mechanism (the *agnostic class*; Part II examines the consequences of this choice).

Data Sets. The empirical analysis uses the eight industrial temporal streams of TabReD (Rubachev et al., 2025) ($n_{\text{test}} = 4.6\text{k} - 60\text{k}$ points each; five regression, three classification, the latter with positive rates from 2.2% to 36%), extended in Section 6.5 by fifteen shift benchmarks across census, fraud, network-security, medical, credit, and bike-sharing domains. Deployed gains are taken from a companion system study by four of the present authors (Wang et al., 2026). That system is a frozen in-context tabular foundation model whose predictions are adjusted by an embedding-space k -nearest-neighbor residual corrector over a past-only buffer, with recency weighting, global debias, and per-region calibration. The present paper treats that system as a *measurement instrument*, not as a contribution of its own. Part I uses its recorded predictions for a retrospective channel audit. The measurements below are computed on eight per-stream slices of that deployment, one per stream, each holding the frozen predictions, embeddings, and residuals in temporal order; we call these the *probe streams*, and a corrector applied to them a *probe corrector*.

3.1 Standing Assumptions

The boundary theorems of Part II quote their hypotheses by name. We collect the names here, one line each, so that the semi-formal statements of Section 6 can be read without leaving the main text. In brief, the multi-window setting posits windows arriving in sequence, occasional windows carrying small anchor sets labeled missing completely at random (MCAR), a stationary drift response tying the budget to an observable novelty index, and regularity for the rest. The formal versions (with the measurability conventions, the failure modes each assumption excludes, and the

counterexamples that fix their scope) are given in the program contract (Online Appendix 1, “the formal program”) and in the standing-assumption blocks of Appendix B and Appendices A and C of Online Appendix 1. The single-window results (Sections 6.1–6.2) use the regularity conditions (S1)–(S7) together with the mechanism prior (B). The conditions require a Polish covariate space, Borel kernels and budget, i.i.d. unlabeled draws, and randomized Borel estimators; prior (B) states that the budget is the *only* constraint on the drifted conditional, with no cross-covariate coupling. The multi-window learning-the-wall results (Section 6.4) additionally assume, writing r_j for window j ’s scalar novelty (density-ratio) index, $\bar{b}_j = |\bar{\eta}_j - \eta_0|$, with $\bar{\eta}_j = \mathbb{E}[\eta_j \mid \mathcal{F}_\phi]$ (Section 6.3), for its ϕ -reachable drift magnitude, and $\mathcal{J}$ for the covered ratio interval (the subscript J in $\bar{\eta}_J$ below indexes the new window $K+1$, Section 6.4):

- (A0) *Windows and anchors.* K windows; window j carries m_j unlabeled i.i.d. draws from p_j and $k_j \geq k_{\min}$ labeled anchors drawn i.i.d. from p_j (MCAR anchor selection), $y \mid x \sim \text{Bern}(\eta_j(x))$, independent across windows and anchors.
- (A1) *Stationarity of the response curve.* A single $\Psi \in \mathcal{P}_{\text{mon}}$ with $\mathbb{E}[\bar{b}_j(x) \mid r_j(x) = r] = \Psi(r)$ for every window.
- (A2) *Support/relevance (rank condition).* The anchor-weighted pooled novelty law $\bar{\mu}$ has Lebesgue density $\geq \mu_{\min} > 0$ on $\mathcal{J}$.
- (A3) *Smoothness and intrinsic dimension.* Each η_j is (L, α_s) -Hölder in ϕ -coordinates on $\text{supp}(p_j)$, of doubling intrinsic dimension $\leq d$ (the model parameter that the deduplicated empirical estimate d_0 of Section 6.5 instantiates), with densities bounded above and below on the anchor region.
- (A4) *Shape.* $\Psi \in \mathcal{P}_{\text{mon}}^L$: nondecreasing, L_Ψ -Lipschitz, range ≤ 1 , and the estimator knows an upper bound on L_Ψ .
- (A5) *Ratio error.* Label-free ratio estimates with $\sup_R |\hat{r}_j - r_j| \leq \varepsilon_r$ on an event of probability $\geq 1 - \delta_r$.
- (A5') *Exogeneity of the ratio error.* The drift noise $\xi = \bar{b} - \Psi(r)$ satisfies $\mathbb{E}[\xi \mid r, \hat{r}] = 0$. This is not implied by (A5); Remark C.13 of Online Appendix 1 gives an errors-in-variables counterexample.
- (A6) *Novelty coverage for transfer.* $d\mu_{K+1}/d\bar{\mu} \leq \Gamma_{\text{cov}}$ on $\mathcal{J}$, where $\mu_{K+1} = \nu$ is the novelty law of the new window; the unprecedented-novelty mass $\pi_{\text{out}} = \mu_{K+1}(\mathcal{J}^c)$ is priced separately.
- (A7) *Density upper bounds.* μ_{K+1} has Lebesgue density $\leq f_{\max}$ on an ε_r -neighborhood of $\partial\mathcal{J}$, plus comparability constants for the stationarity test.
- (T) *Transversality.* $\bar{P}(\bar{\eta}_J(X) = \eta_0(X)) = 0$: the sign-ambiguous crossing set is null under the pooled anchor law.
- (Mgn $_\gamma$) *Margin.* $\bar{P}(0 < |\bar{\eta}_J - \eta_0| \leq t) \lesssim t^\gamma$ for some $\gamma > 0$, so drift mass thins polynomially near the crossing.

Of these, the MCAR selection in (A0) is the assumption most easily violated in practice; (A1) is anchor-falsifiable, unlike (B); and constant within-window novelty profiles reduce (A2) to identification at atoms only.

Currencies and Measurement Conventions. Regression is scored in RMSE, classification deployments in AUC. Loss-denominated quantities (squared error, log-loss) and ranking-denominated ones are *not* interchangeable on these streams, and we track the currency of every claim (Sections 4.1, 5.5). The boundary theory of Part II is stated in the proper-loss, conditional-mean currency throughout: identified sets, walls, and the margin-boundary rates price conditional-mean decisions. Ranking (AUC) sits outside that scope, and the paper treats it as a measured counterexample rather than a covered case. Channel C4’s gains (Section 4.1) are invisible to squared-error ceilings, HOMECREDIT-DEFAULT shows log-loss and AUC moving in opposite directions, and the local-correction screening rule’s ranking-currency false freeze on ECOM-OFFERS is priced to this

scope (Section 7.1). A ranking-currency analogue of the theory and of the screening rule is left open (Section 8). Unless stated otherwise, retrieval oracles and signal ceilings are *time-respecting*: they use either past-only data or strict out-of-time splits (Section 5.1). Noise floors are estimated from temporally decoupled pairs (Section 5.2), and intrinsic dimensions are estimated after exact-duplicate removal (Section 5.3). Every reported number was recomputed by a second, independently written implementation; the two preliminary readings that this recomputation overturned are noted where they arose (Sections 5.3 and 6.5).

Notation. The recurring symbols are collected below, with their defining locations.

ϕ, h_0	frozen embedding and readout (Section 3)	γ	margin exponent, (Mgn_γ)
$\eta_0(x)$	source conditional mean $p_0(1 x)$, $p_0 = h_0 \circ \phi$	γ^*	threshold $d_0/(2\alpha_s)$ (Section 6.4)
p_W, η_W	window covariate law (Section 3); $\mathbb{E}_{p_W}[y x]$	d_0	deduplicated intrinsic dimension of ϕ (Section 5.3)
$\beta(x)$	drift budget, a prior (Section 6.1)	α_s	drift-field Hölder smoothness, (A3)
$B_\beta(x)$	TV-ball of radius $\beta(x)$ around $p_0(\cdot x)$	a	rate exponent $\alpha_s/(2\alpha_s + d_0)$
$I(\beta)$	identified set: selections through B_β	K, k, N	windows, anchors per window, $N = Kk$ (Section 6.4)
$D_\Delta(\beta; R)$	its Δ -diameter: the wall (Section 6.1)	R	region: a measurable subset of $\mathcal{X}$
w_β	post-clip width $\min(1, \eta_0 + \beta) - \max(0, \eta_0 - \beta)$	$\Psi(r)$	drift response $\mathbb{E}[\bar{\eta}_J - \eta_0 R = r]$
$c(\beta; R)$	clip term, the subtractive label-free content	w	novelty representer $d\nu/d\bar{\mu}$
WallA	identification wall on ϕ -measurable acts (Section 6.3)		
WallB, WallB $_\rho$	$\mathbb{E}[\text{Var}(\eta_W \mathcal{F}_\phi)]$; deployed at stated resolution ρ (Section 6.3, Appendix J)	$\rho_\beta(x)$	price of freeze, $2 \max(0, \beta - \eta_0 - \frac{1}{2})$ (Section 6.1)

Three entries require additional qualification. The bound $\alpha_s \leq 1$ is an assumption about the roughness of the drift field, not a derived property; Section 6.5 and Appendix J discuss what happens when it fails. Per-data-set values of d_0 are tabulated in Section 6.5; the length of a test stream is written n_{test} , which is not the anchor count $N = Kk$. The letter R is overloaded: it denotes a region throughout, but in Section 6.4, Appendix H, and Appendices G and K of Online Appendix 1 it is also the scalar novelty index with law ν , in Section 6.4 the symbols $R_2, R_{\text{prod}}, \dots$ are remainder terms, and R^2 is always the coefficient of determination. The letters K, k, m , and D are also overloaded: K is the window count throughout (Appendix B writes the class count as $|\mathcal{Y}|$); k is the anchor count per window except as the k -NN neighbor count ($k = 20$) of Parts I and III; m denotes unlabeled draws, the target sample size, and the finest retrieval scale ($m = 5$) of Section 7.2; and D is the ambient width 192, the wall D_Δ , and the polynomial degree of Section 6.4. Wall A denotes the identification wall; its raw form is D_Δ and its deployable form WallA, which coincide on fiber-unions (Theorem 6.5). Each use is disambiguated where it occurs.

Reproducibility. All data sets are publicly documented research benchmarks; access-controlled sources are used under their applicable data-use terms. They comprise the TabReD streams (Rubachev et al., 2025) and the fifteen shift benchmarks of Section 6.5 (census, fraud, network-security, medical, credit, and bike-sharing sources). Appendix J specifies the full measurement protocol (probe and oracle configurations, noise-floor estimation, deduplication, and the threshold computation) at a level intended to permit independent re-implementation. The machine-verification programs behind the appendix proofs (exact rational enumeration and an independent cross-check written without shared code) are part of the reproduction deposit together with their outputs, and Section S4 of Online Appendix 1 indexes them. The measurement code and every archived input behind the primary empirical results are collected in a reproduction deposit that will be released in a public archive; its citable DOI will be recorded in a revised version of this manuscript. The deposit holds the eight per-stream probe slices (frozen predictions, embeddings, residuals), the stored embeddings behind the deduplication checks, the verification programs, and

a manifest with hashes. The historical system’s per-sample outputs (a tuned seven-prior ensemble, Section 5.6) belong to the companion study and are not redistributed; its values are retained only as a secondary comparison. Section 5.6’s single-prior reconstruction, fully covered by the deposit, is the primary reproducible result.

4 The Anatomy of Adaptation Gains

Prequential adaptation helps on most of the evaluated temporal tabular streams and harms on one, but existing evidence offers little guidance on when either outcome should occur. We ask which components of the observed gains can be attributed to recurring mechanisms. The gains do not generally amount to recovering the drifted conditional $p_t(y \mid x)$. Without labels, such recovery is impossible within the agnostic drift class (Part II); with revealed labels, it is rate-limited by the embedding’s intrinsic dimension (measured $d_0 \approx 3\text{--}18$; channel C3 below). We therefore treat the observed improvements as diagnostic channels, each with a theoretical status and a measured value on the evaluated streams.

Throughout, gains are measured on the eight TabReD streams under the prequential protocol of Section 3; all oracle and ceiling quantities follow the time-respecting measurement methodology of Section 5.

4.1 Four Gain Mechanisms and Two Collapse Classes

We organize the corrector’s observed improvement into four measured mechanisms (C1–C4) and one pair of identification-collapse classes (jointly labeled C5). For each mechanism, we specify why it is available, how its time-respecting ceiling is measured, its value on the eight streams (Table 1), and the conditions under which it is absent or harmful. The list is designed for mechanism attribution in the evaluated corrector, not as a mutually exclusive decomposition of every possible adaptation algorithm. Its levels differ deliberately: C1–C4 are measured mechanisms of the evaluated corrector, while C5 sits one level up, as a pair of identifiable drift classes (mechanism priors that collapse the identified set of Part II) with no measured value of its own. Theorem 6.2 (Part II) separately shows that the conditional law $p_t(y \mid x)$ is unidentified from unlabeled data beyond its ball constraint, while Theorem 6.3 gives two canonical mechanism classes under which that identified set collapses. *C1: Global and Slow-Timescale Debias.* The simplest channel tracks the slowly-moving mean of the residual $r_t = \hat{y}_t - y_t$ and subtracts it. The mechanism asks nothing of the representation. Whatever component of the drift is shared across the covariate space (a common shift in the target’s level) surfaces in the prequential residual stream as a slowly-moving mean, and a running average over strictly-past residuals estimates that mean with essentially no variance cost. The channel is trivially identifiable from prequential labels because it uses them only through their first moment: no retrieval, no locality, no model of *where* the drift acts.

Its contribution is measured by the counterfactual replay of Section 5.4: repeat the identical prequential protocol with the corrector reduced to its running-mean term alone, and compare the debias-only gain with the full system’s. On one stream, this simple channel accounts for most of the gain. On DELIVERY-ETA, the running-mean-only corrector reproduces 93% of the probe corrector’s +0.66% RMSE improvement, and the retrieval machinery contributes the small remainder. Refining the offset yields little more (on COOKING-TIME, region-resolved offsets add at most +0.07 percentage points over a single global offset).

C1 is inexpensive and does not appear as a distinct architectural component, so its contribution is easy to misattribute to the full method. Any evaluation of a sophisticated adaptation method that does not report the debias-only baseline conflates this channel with genuine local structure

stream	task	reconstructed gain	dominant channel(s)	signal ceiling
SBERBANK-HOUSING	reg	+3.80% RMSE	C2 freshness	≤ 0
WEATHER	reg	+3.94% RMSE	C3 local	+0.025
COOKING-TIME	reg	+0.83% RMSE	C1 debias	≤ 0
DELIVERY-ETA	reg	+0.66% RMSE	C1 debias (93% on the probe)	≤ 0
MAPS-ROUTING	reg	+0.59% RMSE	C1/partial-scale C3	≤ 0
ECOM-OFFERS	cls	+0.86pp AUC	C4 ranking	≤ 0 (MSE)
HOMESITE-INSURANCE	cls	+0.74pp AUC	C4 ranking	+0.016 (MSE)
HOME CREDIT-DEFAULT	cls	−2.28pp AUC	harm case (Section 4.2)	≤ 0

Table 1: The anatomy on eight TabReD streams. The gain column reports the reproducible single-prior reconstruction over the frozen model; the dominant channel column records the counterfactual attribution of the historical seven-prior deployed system (the DELIVERY-ETA percentage is the probe-corrector replay of Section 5.4). The final column is the time-respecting (leak-free) squared-error signal ceiling. Ranking gains on classification streams flow through channel C4, which squared-error ceilings cannot register (Section 4.1). Ceilings are fractions of residual variance under the strict out-of-time protocol of Section 5.1. Table 3 gives bootstrap intervals and the non-redistributed seven-prior system’s figures as a secondary historical comparison.

(Section 5.4). Conversely, the channel is small where the drift is not a level shift. On SBERBANK-HOUSING and WEATHER the dominant channels are instead freshness and local structure (Table 1). *C2: Freshness (Drift Tracking)*. Under genuine distribution drift, *when* a labeled point was observed matters independently of how many are available. The residual field the corrector must estimate is itself moving, so an old label reports on a field that no longer holds. The channel is isolated by a size-matched design on SBERBANK-HOUSING: the same corrector is run with buffer size held fixed and only the buffer’s *composition* varied (the most recent labels, a random past-only subset, or the oldest labels of the stream). All three arms draw from the prequential past, so no future label (Section 5.1) can inflate any arm, and the spread across arms measures recency and nothing else. The arms deliver +4.9%, +4.1%, and −1.1% RMSE respectively. Stale supervision is actively *harmful*, not merely useless.

The matched-buffer comparison isolates freshness as a measurable channel: the freshest and stalest arms differ by about six points. A corrector whose buffer is never refreshed also inherits the stale-label penalty as that buffer ages. In this stream, tracking recency prevents that degradation; the experiment does not establish that online adaptation is necessary for every frozen predictor or deployment. Note also what C2 does *not* deliver: it consumes labels, the one currency the wall of Part II respects. Tracking recency from revealed labels is possible precisely because it never requires certifying, from unlabeled data, how far the world has moved (Theorem 6.2).

C3: Local Re-Estimation in Representation Space. Prediction errors of the frozen model are *spatially* organized in its own embedding. On WEATHER, the mean absolute residual difference between embedding nearest neighbors is 78% of that between random pairs (neighbors share errors), while the *temporal* autocorrelation of the residual stream is 0.025 at lag one. Errors live in space, not in time. The mechanism is a property of the frozen pair (ϕ, h_0) : points the embedding places together err together, so a corrector that indexes its memory by embedding position (a past-only k -nearest-neighbor regression of residuals over past labeled points) can re-estimate the frozen model’s error locally. Deployed, this delivers +3.9% RMSE on WEATHER.

This channel’s ceiling is the quantity most easily overstated, and its measurement is where the time-respecting methodology is essential. The natural oracle (batch leave-one-out k -NN regression of residuals on embeddings) draws 49.5–53.2% of its neighbors from the future of the query point

(Section 5.1). Every ceiling we report is instead time-respecting, computed with past-only growing buffers or strict out-of-time splits. On WEATHER the correction is first-order: the batch ceiling of 0.0995 of residual variance falls to 0.067 with the causally growing buffer and 0.0245 under the strict out-of-time split (both defined in Section 5.1), and the deployed corrector (RMSE 1.543, a +3.9% reduction against its own frozen arm) already sits near what even the leaking batch oracle reaches on the probe stream (1.513 against a frozen 1.615; Section 5.6). The channel is identifiable but rate-limited. Its ceiling is governed by the intrinsic dimension of the embedding (measured $d_0 \approx 3\text{--}18$ across streams, Table 4, robust to the deduplication check of Section 5.3), and Part II shows the resulting nonparametric rates are the binding constraint. The measured time-respecting ceilings of Section 5.6 are small *because* d_0 is not small.

The channel is absent in two distinct ways. On the classification streams its squared-error signal is zero or negative on two of the three and marginal on the third (Table 1; HOMESITE-INSURANCE’s +0.016 falls to 0.006 under the temporal-gap stress test of Section 5.1). The residual variance there is almost entirely Bernoulli label noise. Prediction-based floors run from 0.85 to 0.88 of residual variance across the three streams, and temporally decoupled variogram readings from 0.73 to 1.02, leaving no spatially organized residual for retrieval to recover, and what is correctable on those streams lives in the ranking currency of C4. And on HOMECREDIT-DEFAULT the channel turns actively harmful: at a 2.2% positive rate, local residual smoothing degenerates into noisy base-rate estimation (Section 4.2).

C4: Ranking and Recalibration (Classification). On classification streams a fourth channel appears that squared-error analysis cannot see. The mechanism is a currency phenomenon: the classification residual variance is almost entirely Bernoulli label noise (the noise-floor fractions already measured under C3), so in squared error there is nothing to correct. The time-respecting signal ceilings of Table 1 are zero, negative, or marginal on these streams. But the frozen model’s *scores* can still be systematically miscalibrated region by region, and offsets applied in the natural parameter move ranking and calibration in ways mean-squared-error accounting cannot register.

The measurement is guarded by an out-of-time design: region-resolved calibration offsets are fit strictly before a temporal cut and evaluated strictly after it, and the result is replicated across four temporal cut points (the first 40, 50, 60, and 70% of the stream) and six k -means codebook seeds, at the best of three codebook sizes (Appendix J). Measured this way, the offsets improve AUC by +1.3 to +1.9 points on HOMESITE-INSURANCE and +1.3 to +1.8 points on ECOM-OFFERS, while the deployed full systems on the same streams realize +0.7 and +0.9 points (Table 3, historical column; the reconstruction of Table 1 gives +0.74 and +0.86). These gains coexist with squared-error ceilings that are zero, negative, or marginal: a loss-denominated audit would declare the streams uncorrectable while ranking metrics improve.

The channel’s failure mode is the harm case of Section 4.2: on HOMECREDIT-DEFAULT, likelihood-optimal regional offsets *anti-align* with ranking in-sample, and no out-of-time offset configuration is positive at any granularity (the numbers are given there). The currency of measurement is therefore not an incidental detail. Section 5.5 shows the same mismatch defeats loss-based safety monitors.

C5: Identifiable Drift Classes (Covariate and Label-Marginal Response). Two canonical drift mechanisms are identifiable *even without target labels* (Theorem 6.3). Both are *collapse classes*: mechanism priors under which the identified set of Part II shrinks to a single conditional, so the wall is zero and no labels are required.

The first collapse is *prior-driven*. Under pure covariate shift (the window’s joint law is any covariate marginal paired with the unchanged source conditional) the identified set is the singleton $\{p_0(\cdot \mid \cdot)\}$, and the wall vanishes for every diagonal-vanishing discrepancy, every region, and every budget, already at $n = 0$. No unlabeled sample is even needed, because the collapse is the content

of the mechanism assumption itself (Proposition B.13). The frozen predictor remains correct where it matters, and the channel consists of leaving the conditional alone.

The second is *data-pinned*. Under label-marginal shift with known class-conditionals that are linearly independent as measures (and a window marginal consistent with the class), the target prior π_W is uniquely identified from unlabeled covariates by a moment system, $\pi_W = M^{-1}\mathbb{E}_{p_W}[T]$ for a bounded statistic T whose moment matrix M against the class-conditionals is nonsingular (Theorem B.16; this is the moment system of black-box shift estimation, Lipton et al., 2018; Garg et al., 2020). The identified set collapses to the single reweighted conditional, and π_W is estimated from unlabeled draws at the parametric rate (Proposition B.17). Unlike the covariate-shift collapse this one has genuine data content, but its hypothesis is not without cost: the known class-conditionals are label-derived side information obtained outside the window (Remark B.21).

A deployed corrector inherits both mechanisms automatically, and they delimit the “unlabeled-correctable” portion of any real drift. The channel’s failure mode is itself a theorem: *which* mechanism class a given window’s drift belongs to is unfalsifiable from unlabeled data. When the window marginal is consistent with the label-shift class, worlds drawn from the agnostic class, from covariate shift, and from label shift all induce the same unlabeled law at every sample size, so every label-free test between mechanism classes has power equal to size (Proposition B.20). Invoking C5 is a modeling commitment, not a measurement. The collapse is real when the mechanism prior is right, and whether it is right cannot be checked from unlabeled data.

4.2 The Eight-Stream Decomposition and the Harm Case

Table 1 summarizes the per-stream decomposition. The results show three recurring patterns.

Gain concentration. Every positive stream’s improvement (+0.6% to +4.6% RMSE and +0.7 to +0.9 AUC points on the historical deployed system, Table 3; +0.59% to +3.94% and +0.74 to +0.86 in the reconstruction of Table 1) is accounted for by C1–C4, with the mixture varying by stream (DELIVERY-ETA is almost purely C1, SBERBANK-HOUSING is dominated by C2, WEATHER by C3, and the classification streams by C4).

Observed harm. On HOMECREDIT-DEFAULT (default rate 2.2%), naive residual correction *costs* 2.46 AUC points on the deployed system (−2.28 in the reproducible reconstruction of Table 3; the probe corrector’s never-lock evaluation reads −2.36, Section 5.5), and the damage is mechanistic rather than accidental. Likelihood-optimal regional offsets anti-align with ranking on this stream (in-sample, offsets that improve log-loss by 2.1% simultaneously reduce AUC by 1.47 points), and no out-of-time offset configuration is positive at any granularity. Rare-positive streams turn local residual smoothing into noisy base-rate estimation. The harm case motivates the gate analysis in Section 5.5.

Proximity to the descriptive ceiling. The time-respecting (leak-free) signal ceilings of Section 5.6 are small (e.g. 2.5% of residual variance on WEATHER, less elsewhere), and the deployed systems capture the bulk of them. The measured room for local residual correction is small for this representation and protocol, consistent with the labeled-history boundary of Part II, whose nonparametric rates depend on d_0 (Section 6.4). We therefore compare adaptation methods against the time-respecting ceiling rather than only against each other.

5 Measurement Methodology for Streaming Adaptation

The preceding measurements depend on the evaluation protocol. Several common practices in streaming adaptation introduce optimistic bias; on our own data, two of them reverse a conclusion. This section documents five corrections. Each is illustrated with the measured size of the artifact

bias		measured artifact	corrected protocol
future-neighbor (Section 5.1)	leakage	49.5–53.2% future neighbors on 8/8 streams; ceilings inflated up to 4.5×; sign flip on SBERBANK-HOUSING	time-respecting oracles: past-only buffers or strict out-of-time splits
temporal-twin (Section 5.2)	noise floors	WEATHER nugget 0.0009 vs. decoupled floor 0.13–0.62 of residual variance	variogram on pairs $\geq 1,000$ steps apart
duplicate-deflated (Section 5.3)	d_0	UNSW_NB15 0.87 $\rightarrow$ 3.58; ACS_EMPLOYMENT 1.62 $\rightarrow$ 3.11; both leave the regular branch	deduplicate at four decimals before estimating d_0
currency-mismatched (Section 5.5)	gates	84% of AUC harm before the log-loss lock; −11.4 AUC points in the first 1%	monitor the deployment metric itself

Table 2: Four optimistic measurement biases for streaming adaptation, the measured size of the artifact each introduces (each magnitude is measured on the stream that exhibits it), and the corrected protocol. Every number is discussed in the corresponding subsection. The fifth correction, counterfactual channel attribution (Section 5.4), guards reporting rather than a number and is not tabulated.

it removes. Table 2 summarizes the four biases and their corrections, and Figure 1 plots the per-stream measurements behind three of them. The fifth item (Section 5.4) is a reporting correction rather than a numerical one.

5.1 Future-Neighbor Leakage Inflates Signal Ceilings

The natural oracle for “how much correctable signal exists” is a batch leave-one-out k -NN regression of residuals on embeddings. The mechanism of the bias is that batch retrieval is symmetric in time while deployment is not. On all eight TabReD streams, 49.5–53.2% of the oracle’s selected neighbors lie in the *future* of the query point. On a drifting stream this is not an incidental detail. The residual field itself moves, so a future neighbor that is close in embedding space carries the locally drifted residual the oracle is being asked to predict. The batch oracle is scored with information no deployed corrector can possess, and its “ceiling” overstates the correctable signal.

The leak is not benign, and its measured size changes conclusions (Figure 1a). Recomputing ceilings under a time-respecting protocol deflates the batch numbers by up to a factor of 4.5 (on HOMESITE-INSURANCE, R^2 falls from 0.0735 to 0.0164, and a temporal-gap stress test, which additionally withholds the 1,000 fitting rows nearest the cut, lowers even the corrected value to 0.006) and *reverses the sign* of the apparent signal on SBERBANK-HOUSING (batch +0.080, time-respecting −0.048), where the entire apparent squared-error signal was the leak.

The corrected procedure comes in two grades. The deployment-faithful grade restricts neighbors to the strictly-past prefix with a causally growing buffer, exactly as a deployed corrector would retrieve; the conservative grade is a strict out-of-time split: fit on the first half of the stream, evaluate on the second. On WEATHER, the batch value of 0.0995 becomes 0.067 with the causally growing buffer and 0.0245 under the strict split. Every ceiling in this paper is time-respecting, and we recommend the field adopt the same convention for any retrieval-based oracle on ordered data.

5.2 Temporal Twins Deflate Noise Floors

The irreducible-noise floor of a stream is naturally estimated from the variogram of residuals at vanishing embedding distance. Pairs that the embedding places arbitrarily close should disagree

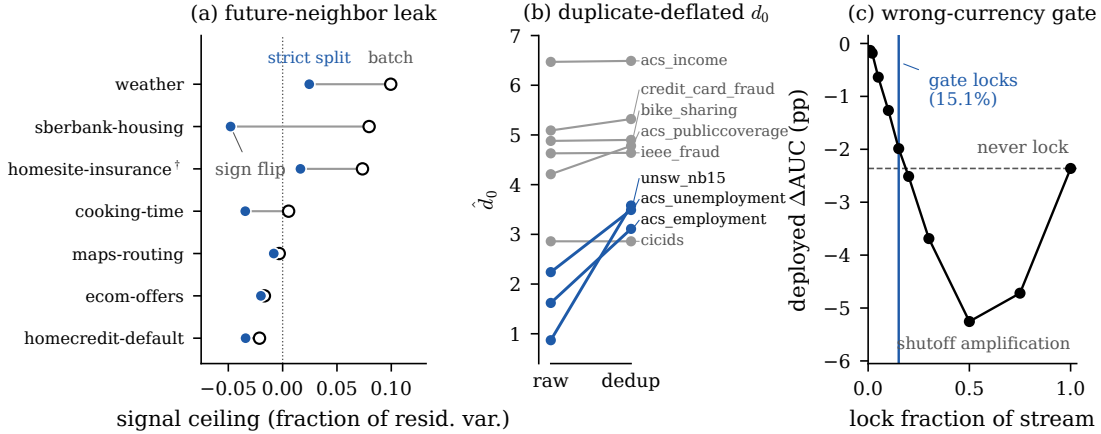


Figure 1: The measured size of three artifacts (per-stream intrinsic dimensions for panel (b) in Appendix J; the remaining values are part of the reproduction deposit). (a) Squared-error signal ceilings: the batch leave-one-out oracle (open) against the strict out-of-time split (filled), as fractions of residual variance. The leak inflates WEATHER by $4.1\times$ and HOMESITE-INSURANCE by $4.5\times$ and flips the sign on SBERBANK-HOUSING (the batch ceiling for DELIVERY-ETA was not recorded, so only its strict-split value is shown). The dagger flags HOMESITE-INSURANCE, whose strict-split value (filled) collapses further, to 0.006, under the temporal-gap stress test of Section 5.1. (b) Intrinsic dimension before and after four-decimal deduplication for the nine re-measured related benchmarks (accented: the three large movers, UNSW_NB15, ACS_EMPLOYMENT, and ACS_UNEMPLOYMENT; the TabReD streams move by at most 0.08 and are not shown; Section 5.3). (c) HOMECREDIT-DEFAULT: deployed AUC change of the probe corrector as a function of where the log-loss gate locks. The actual lock (15.1%, vertical line) arrives after 84% of the never-lock harm (dashed), and a mid-stream shutoff can more than double that harm (Section 5.5).

only through irreducible noise, so the variogram’s intercept (the *nugget*) reads off the floor. The reading is valid only if near pairs carry *independent* noise realizations. On temporally ordered tabular data they do not: the closest embedding pairs are *temporal twins*, near-duplicate records of the same physical event (on WEATHER, median stream separation 12 steps) that share their noise realization. A shared realization contributes nothing to pair disagreement, so the nugget reads duplication as correctability.

The size of the artifact is extreme. On WEATHER the naive nugget is 0.0009 (suggesting essentially all residual variance is correctable in principle), while the temporally decoupled estimate puts the floor at 0.13–0.62 of residual variance, over two orders of magnitude higher.

The corrected procedure restricts the variogram to pairs at least 1,000 steps apart in the stream. On WEATHER this leaves *no support* at small embedding distances (the twins were the only close pairs), so the decoupled floor is an extrapolation, and we report it as the range 0.13–0.62 rather than a point. Two rules follow: noise floors on ordered data must be computed from temporally decoupled pairs; and a near-zero nugget should be treated as a duplication symptom, not a promise of correctable signal. The artifact is the streaming sibling of the duplicate-deflated intrinsic dimension of Section 5.3, because both are driven by (near-)duplicate records. The measurement is scoped: on the three classification streams, where the residual field is a bounded Bernoulli residual rather than a smooth field, decoupling does not systematically raise the floor (the readings move from 0.98 to 0.73, from 1.05 to 1.02, and from 0.97 to 0.98; these readings are reported for contrast

only, since the decision-rule evaluation of Section 7.1 uses the Bernoulli floor $\mathbb{E}[p(1 - p)]$ on the classification streams). We claim the twin artifact for the regression streams we measured it on, not as a general law.

5.3 Duplicates Deflate Intrinsic Dimension

The channel-C3 ceiling and the boundary threshold of Part II both depend on the intrinsic dimension d_0 of the frozen embedding. The mechanism of the bias is elementary. Duplicate rows are endemic in tabular data (repeated transactions, network flows, census microdata), and they place many points at first-neighbor distance $r_1 \approx 0$, which breaks the standard estimators (TwoNN, Facco et al., 2017; the maximum-likelihood estimator, Levina and Bickel, 2004) and biases d_0 sharply *downward*. The direction of the bias is what makes it dangerous. A spuriously low d_0 manufactures exactly the “low-dimensional, inexpensively adaptable” reading that the threshold $\gamma^* = d_0/(2\alpha_s)$ of Part II rewards.

The bias is large enough to change conclusions. UNSW_NB15, a network-intrusion stream with 41% duplicate flows, reads $d_0 = 0.87$ raw and 3.58 after removing rows identical to four decimals. The raw value is not even reproducible across estimators (TwoNN gives 2.93 on the same raw data), an instability that is itself part of the diagnosis. ACS_EMPLOYMENT, a census stream with 16% of rows duplicated to four decimals, moves from 1.62 to 3.11 (Figure 1b). On the raw values, both data sets appear to fall into the regular ($\sqrt{N}$) branch of Part II. After deduplication, neither remains in that branch under the sensitivity inputs $(\alpha_s, \gamma) = (1, 1)$ (Appendix J); this correction does not constitute a measured phase call for the related benchmarks.

The corrected procedure removes rows identical to four decimal places, recomputes every d_0 estimate on the unique subset (for the essentially duplicate-free TabReD streams the raw and deduplicated readings agree to 0.08, Appendix J), and cross-checks estimators on borderline cases. The correction is inexpensive and, on clean data, inert (the eight TabReD streams are 99.8–100% duplicate-free and the six high- d_0 related benchmarks 99.97–100%; the d_0 estimates move by at most 0.08 and 0.10 respectively). It matters exactly on the data sets that look most favorable.

5.4 Architectural Attribution Miscredits Gains

Which channel produced a system’s gain cannot be read off its architecture or its learned weights. The artifact here is interpretive rather than numerical: an adaptation system’s gain is naturally credited to its most sophisticated component, because that component is what the paper is about. The DELIVERY-ETA decomposition of Section 4.1 is the cautionary measurement. On that stream, 93% of the probe corrector’s gain is reproduced by a running mean, and nothing in the system’s architecture gives any hint that its retrieval machinery is doing almost nothing.

The corrected procedure is *leave-one-channel-out counterfactual replay* on the same stream: repeat the identical prequential protocol with one channel disabled, and attribute to each channel the difference its removal makes. Disabling a channel means freezing the debias term at zero (C1 off), removing the buffer’s recency weighting (C2 off), or replacing the local corrector by the global one (C3 off). The C1–C3 entries of the dominant-channel column of Table 1 were produced this way (the C4 entries come from the out-of-time calibration design of Section 4.1), and the COOKING-TIME comparison of region-resolved against global offsets (+0.07 percentage points at most) is the same replay at finer granularity. We regard channel-resolved attribution tables as the minimum reporting standard for adaptation papers, for the same reason ablations are standard elsewhere: without them, trivial channels masquerade as method contributions.

5.5 Currency-Mismatched Gates Miss Deployment Harm

Because adaptation can harm (Section 4.2), deployed correctors need an online no-harm monitor. A safety gate is a measurement instrument too, and it can be biased in the same optimistic direction as an oracle. The natural gates are *loss*-denominated, because that is the currency in which anytime-valid sequential tests are available (Ramdas et al., 2023), while the deployment currency on classification streams is ranking. A statistically valid anytime monitor in the *wrong currency* can therefore miss deployment harm.

On HOMECREDIT-DEFAULT we ran an e-process (test-by-betting, Shafer, 2021) monitor on per-round *log-loss* regret of the corrector against the frozen model. The monitor is sound (its false-alarm rate under permutation nulls matches its nominal level), and it does eventually fire, locking the corrector out at 15.1% of the stream (Figure 1c). But 84% of the *AUC* harm is incurred before the lock: over the first 1% of the stream *AUC* falls by 11.4 points while the monitored log-loss barely moves, and from 2 to 5% of the stream the correction *improves* log-loss outright while the ranking harm keeps accruing. This is the anti-alignment of Section 4.2 seen from the monitor’s side. The gate analysis uses the probe corrector (configuration in Appendix J); its never-lock harm of -2.36 points reflects the same mechanism as the deployed system’s -2.46 . It is measured by the original probe-corrector code inside the gate experiment, whereas the -2.28 of Table 3 is the deposit’s reimplement of the same operator scored on the 55,901 points after the 100-point buffer; the two agree well within the bootstrap interval of Table 3. On this stream, offsets that help the likelihood hurt the ranking, so a gate watching the likelihood is reassured precisely while the deployment metric is being damaged.

The failure is structural in two senses. First, the information needed for an early stop does not exist in the monitored currency (log-loss is flat, then *improving*, while the worst of the harm accrues), so no threshold tuning, no faster alarm, no better betting scheme can repair the monitor. Second, the gate’s own action is not free in the deployment currency: mid-stream stopping introduces a score-scale discontinuity that can add further ranking harm. Locking at half-stream more than doubles the never-lock harm (-5.26 against -2.36 *AUC* points on the probe stream; Figure 1c). The evidence supports a simple operational rule: *monitor the metric you deploy on*. Constructing anytime-valid monitors for ranking currencies is, to our knowledge, open: the game-theoretic testing framework (Shafer, 2021; Ramdas et al., 2023) supplies the loss-denominated instruments used above, and we are not aware of a ranking-denominated counterpart.

5.6 Ceilings versus Realized Gains

With the corrected protocol in hand, the regression-stream audit finds small time-respecting signal ceilings (between 0 and 2.5% of residual variance in squared error); classification has a separate ranking-currency residual signal in channel C4. These quantities are descriptive point estimates, not confidence bounds or universal ceilings.

Reproducibility and Uncertainty of the Headline Gains. Table 3 reports a single-prior local corrector evaluated with a block bootstrap; this reconstruction is the primary empirical record. The corrector is a compact reimplement of the streaming k -NN probe corrector used in Sections 5.5 and 7.1 (configuration in Appendix J), applied to the probe streams; the reproduction deposit (Section 3) covers it in full. The secondary historical figures come from a tuned seven-prior ensemble whose per-sample outputs we do not redistribute. The reconstruction reproduces the deployed figure to its printed decimal on five of the eight streams, and on two more it agrees in sign and magnitude without matching the printed digit (COOKING-TIME $+0.83$ against $+0.9$; the harm case -2.28 against -2.5). Thus the qualitative seven-gain/one-harm pattern and its approximate magnitudes

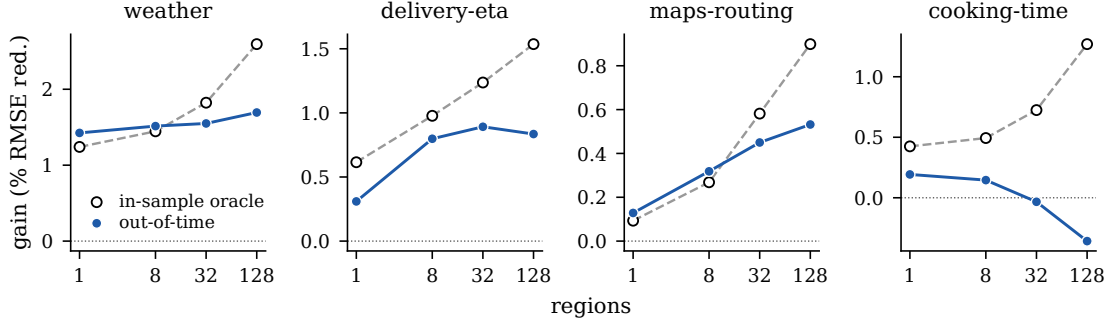


Figure 2: In-sample versus out-of-time gain as adaptation capacity grows: the regional-offset probe (k -means codebooks of 1, 8, 32, and 128 regions; protocol in Appendix J) on the four regression streams with adequate sample size for the split. Open, dashed: offsets fit with hindsight on the full stream. Filled: fit on the first half and evaluated on the second (a single out-of-time split, no error bars). The in-sample gain grows monotonically with codebook size on every stream while the out-of-time gain flattens, saturates, or reverses sign: added capacity yields apparent gain, not deployable gain. At small codebook sizes the arms can cross, as on WEATHER, since they score different halves of a drifting stream. SBERBANK-HOUSING (the shortest stream) is excluded because its evaluation half is too short to support the codebook sweep; the classification streams use the out-of-time protocol of Section 4.1.

are reproducible from the deposit; the exact historical figures still depend on access to the tuned system. On SBERBANK-HOUSING the reconstruction reads +3.8% against +4.6%: the historical system’s attribution assigns that stream’s gain to freshness (C2) rather than to the local channel this corrector implements, so a local-only corrector is not expected to recover it.

Every interval excludes zero, the harm case included. The last column measures the resolution of the evaluation itself. The half-width of the moving-block interval for the frozen model’s *own* metric exceeds the reconstructed gain on four of the eight streams, so a difference of a few tenths of a percent is resolvable only because the bootstrap is paired on the same resampled blocks. Unpaired comparisons at these effect sizes carry more evaluation noise than signal. This applies to two systems scored on different splits, and to any table of published numbers drawn from different evaluations.

On WEATHER, the tuned corrector achieves an RMSE of 1.543, while the batch oracle on the single-prior probe (scored with future neighbors it could not have had) reaches 1.513 against that probe’s frozen 1.615, little beyond the deployed value. The deployed gain itself is the +3.9% historical figure of Table 3, measured against the tuned system’s own frozen arm; the oracle figure bounds what the leak could have yielded on the probe, not what the deployment gained. Figure 2 shows the same fact in capacity-resolved form: growing the regional capacity of an offset corrector inflates the hindsight gain monotonically while the out-of-time gain flattens or reverses. Added fitting capacity converts drift into apparent signal, not into deployable gain. Part II explains why this ceiling is low within the stated agnostic drift class. It also separates the limitations addressed by additional labels from those addressed by a different representation.

stream	task	reconstruction	historical	95% CI	metric $\pm$
WEATHER	reg	+3.94%	+3.9%	[+3.60, +4.27]	1.43
SBERBANK-HOUSING	reg	+3.80%	+4.6%	[+2.54, +5.14]	7.80
COOKING-TIME	reg	+0.83%	+0.9%	[+0.57, +1.05]	0.86
DELIVERY-ETA	reg	+0.66%	+0.7%	[+0.45, +0.91]	0.97
MAPS-ROUTING	reg	+0.59%	+0.6%	[+0.43, +0.77]	0.96
ECOM-OFFERS	cls	+0.86pp	+0.9pp	[+0.29, +1.53]	0.69
HOMESITE-INSURANCE	cls	+0.74pp	+0.7pp	[+0.49, +0.98]	0.49
HOME CREDIT-DEFAULT	cls	−2.28pp	−2.5pp	[−3.41, −1.16]	1.38

Table 3: Reconstruction and historical comparison. The primary reconstruction is the single-prior streaming k -NN corrector ($k = 20$, unweighted neighbor mean, correction scale 0.7, block 64) run causally over the probe stream; the historical column reports the non-redistributed seven-prior deployment. Intervals are 95% moving-block bootstrap ($B = 1000$, block length $4\lceil n_{\text{test}}^{1/3} \rceil$, paired: each replicate scores both arms on the same resampled blocks); verdicts are unchanged at block lengths $2\lceil n_{\text{test}}^{1/3} \rceil$ and $8\lceil n_{\text{test}}^{1/3} \rceil$. The last column is the half-width of the same bootstrap applied to the frozen model’s own metric, on the gain scale. That half-width is the resolution of the evaluation itself.

6 The Boundary: What Unlabeled Data Can and Cannot Determine

Part I organized the observed gains into four measured mechanisms (plus two collapse classes) and reported descriptive signal ceilings. We use partial identification to study an informational limitation within the stated agnostic drift class (Manski, 2003). The label-free information state (labels not yet revealed) is nested inside the prequential protocol of Part I, not beside it: at each prediction time the arriving label has not yet been revealed, so every prequential step passes through this information state. The wall of Sections 6.1–6.2 is what a corrector faces before its labels arrive, Section 6.3 adds the wall that labels cannot remove, and Section 6.4 then shows what labeled history can recover and at what rate. Every statement is in the proper-loss, conditional-mean currency of Section 3; ranking is outside the theory’s scope, and we treat it as a measured counterexample. Main-text statements refer to the named assumptions of Section 3.1 and give the exact conclusions needed for interpretation. Each is followed by its core proof (routine steps cite appendix lemmas) and a pointer to the full counterpart in Appendix B, Appendix H, or Online Appendix 1, all of which retain the original letters A–K and statement numbers (A.14, G.1.4, etc.). There every claim is tagged *proven*, *partial* (a flagged step remains), *conditional* (on the named gap (2c)), *cited-standard*, *conjecture*, or *verified measurement*; the Status paragraph of Section 6.4 repeats the flags for the multi-window results, and the flags for Sections 6.1–6.3 are quoted where the statements occur.

6.1 The Identified Set and Its Diameter

Fix a deployment window W with observable covariate law p_W and unobserved conditional $p_W(\cdot | x)$. The frozen model supplies its source conditional $p_0(\cdot | x) = h_0(\phi(x))$. Absent labels, the only constraint linking the two is an assumed *drift budget* $\beta : \mathcal{X} \rightarrow [0, 1]$: we posit $p_W(\cdot | x) \in B_\beta(x) := \{q : D_{\text{TV}}(q, p_0(\cdot | x)) \leq \beta(x)\}$. The *identified set* $I(\beta)$ collects every conditional law consistent with

all label-free observables, and the *wall* is its diameter in prediction space,

$$D_{\Delta}(\beta; R) = \sup_{q, q' \in I(\beta)} \mathbb{E}_{x \sim p_W(\cdot | R)} [\Delta(f_q(x), f_{q'}(x))],$$

the worst disagreement between two worlds that no unlabeled quantity can distinguish. A natural first conjecture holds that D decomposes additively into an estimable covariate-geometry part and a budget part, $D = f(R) + g(\beta)$. That conjecture is *false*. Under the correct normalization the covariate geometry contributes no additive floor at all.

Theorem 6.1 (Exact reduction). *Assume the single-window regularity conditions (S1)–(S4) of Appendix A of Online Appendix 1 and a bounded discrepancy kernel G of Carathéodory class (C) or x -free finite-range class (U') (the kernel conditions (S5) of that appendix, not the mechanism prior (S5) of Appendix B). The mechanism prior (B) is encoded in the definition of $I(\beta)$ as all measurable pointwise selections. Then the pointwise diameter $\delta_{\beta}^{\Delta}(x) = \sup_{a, b \in B_{\beta}(x)} G(x, a, b)$ is measurable with its supremum attained, and*

$$D_{\Delta}(\beta; R) = \sup_{q, q' \in I(\beta)} \mathbb{E}_{x \sim p_W(\cdot | R)} [G(x, q(\cdot | x), q'(\cdot | x))] = \mathbb{E}_{x \sim p_W(\cdot | R)} [\delta_{\beta}^{\Delta}(x)],$$

the outer supremum attained by a measurable (Kuratowski–Ryll–Nardzewski-selected) pair $q^*, q'^* \in I(\beta)$, so covariate geometry enters only as the averaging measure. In the binary case with the mean-gap discrepancy (Corollary A.16), with $\mu = p_W(\cdot | R)$,

$$\begin{aligned} D_{\text{mean}}(\beta; R) &= \mathbb{E}_{\mu}[w_{\beta}] = 2 \mathbb{E}_{\mu}[\beta] - c(\beta; R), \\ c(\beta; R) &= \mathbb{E}_{\mu}[(\eta_0 + \beta - 1)^+ + (\beta - \eta_0)^+] \in [0, \mathbb{E}_{\mu}\beta], \end{aligned}$$

so $\mathbb{E}_{\mu}[\beta] \leq D_{\text{mean}} \leq 2 \mathbb{E}_{\mu}[\beta]$, attained by the explicit endpoint pair $\Pi_{[0,1]}(\eta_0 \pm \beta)$. The decision-flip diameter is $D_{01}(\beta; R) = \mu(|\eta_0 - \frac{1}{2}| \leq \beta)$ (tie-inclusive convention; the tie-broken convention differs only on $\{\eta_0 - \beta = \frac{1}{2}\}$).

The clip term $c \geq 0$ is the only label-free-measurable content of the identity, and it is purely subtractive: confident source predictions pull the wall down, but nothing in the unlabeled data pushes it up.

Proof idea (Appendix A of Online Appendix 1). The budget ball $B_{\beta}(x)$ is nonempty, compact, convex, with an exact ball-distance formula that makes the correspondence $\Gamma_{\beta} : x \rightrightarrows B_{\beta}(x)$ weakly measurable. The Kuratowski–Ryll–Nardzewski selection theorem then supplies measurable selections, and a Castaing family of selections dense in every ball. The measurable maximum theorem (re-proved in full in Appendix A of Online Appendix 1, since the argument rests on it) delivers measurability of δ_{β}^{Δ} and a measurable argmax pair. The prior (B) enters *exactly once*: because the budget constrains each x separately, $I(\beta)$ is *decomposable* (splicing two members along any measurable set stays feasible), and an interchange-of-sup-and-integral lemma in the mechanism of Rockafellar’s interchange theorem (splicing directedness plus a countable envelope) upgrades the trivial “ $\leq$ ” to equality. Any cross- x prior (O3 smoothness, recurrence) breaks decomposability, and only “ $\leq$ ” remains valid (Remark A.12). The binary corollary is then clip algebra on the explicit endpoint selections. The normalization anchor $D_{\Delta}(0; R) = 0$ (hence the refutation $f \equiv 0$ of the additive conjecture) holds for diagonal-vanishing kernels (Remark A.17).

Proof of the binary formulas. Write $\underline{\eta} = \max(0, \eta_0 - \beta)$, $\bar{\eta} = \min(1, \eta_0 + \beta)$, so $B_{\beta}(x) = [\underline{\eta}, \bar{\eta}]$ and the pointwise diameter under the mean-gap kernel $G(x, a, b) = |\eta_a - \eta_b|$ (class (C), diagonal-vanishing)

is $\delta_\beta^{\text{mean}} = w_\beta = \bar{\eta} - \underline{\eta}$, attained by the *explicit* endpoint pair $q^*(1 \mid \cdot) = \Pi_{[0,1]}(\eta_0 + \beta) = \bar{\eta}$, $q'^*(1 \mid \cdot) = \Pi_{[0,1]}(\eta_0 - \beta) = \underline{\eta}$. Both are measurable everywhere-selections needing no selection theory. The reduction identity (the general part of the theorem) then gives $D_{\text{mean}}(\beta; R) = \mathbb{E}_\mu[w_\beta]$. The clip identity is algebra on the endpoints: substituting $\min(1, u) = u - (u - 1)^+$ at $u = \eta_0 + \beta$ and $\max(0, \eta_0 - \beta) = (\eta_0 - \beta) + (\beta - \eta_0)^+$,

$$w_\beta = 2\beta - (\eta_0 + \beta - 1)^+ - (\beta - \eta_0)^+, \quad \text{whence} \quad D_{\text{mean}}(\beta; R) = 2\mathbb{E}_\mu[\beta] - c(\beta; R)$$

with $c(\beta; R) = \mathbb{E}_\mu[(\eta_0 + \beta - 1)^+ + (\beta - \eta_0)^+] \geq 0$. The sandwich follows from the pointwise bound $0 \leq (\eta_0 + \beta - 1)^+ + (\beta - \eta_0)^+ \leq \beta$ (a three-case check using $\beta \leq 1$, $\eta_0 \in [0, 1]$), so $c \in [0, \mathbb{E}_\mu\beta]$ and $\mathbb{E}_\mu[\beta] \leq D_{\text{mean}} \leq 2\mathbb{E}_\mu[\beta]$, with $c = 0$ iff $\beta \leq \eta_0 \leq 1 - \beta$ μ -a.e. (the interior regime). For decision flips, $B_\beta(x)$ meets the cell $\{a : \eta_a \geq \frac{1}{2}\}$ iff $\eta_0 + \beta \geq \frac{1}{2}$ and the cell $\{a : \eta_a \leq \frac{1}{2}\}$ iff $\eta_0 - \beta \leq \frac{1}{2}$, so the pointwise flip indicator is $\mathbf{1}[|\eta_0 - \frac{1}{2}| \leq \beta]$ (tie-inclusive) resp. $\mathbf{1}[\eta_0 - \beta < \frac{1}{2} \leq \eta_0 + \beta]$ (tie-broken), and averaging gives $D_{01}(\beta; R) = \mu(|\eta_0 - \frac{1}{2}| \leq \beta)$, the two conventions differing only when $\mu(\eta_0 - \beta = \frac{1}{2}) > 0$. $\square$

Formal version and the measurability machinery: Theorem A.14, Appendix A of Online Appendix 1; the binary clip algebra above is Corollary A.16.

An immediate reading is that in the calibrated interior regime the wall $2\mathbb{E}[\beta]$ is a restatement of the assumed budget with zero data content. The operational quantities are instead the *decision-flip fraction* $D_{01}(\beta; R) = \Pr_\mu[|\eta_0(x) - \frac{1}{2}| \leq \beta(x)]$ (the diameter under the decision discrepancy Δ_{01} ; how often the drift could change the decision) and the *price of freeze* $\rho_\beta(x) = 2 \max(0, \beta(x) - |\eta_0(x) - \frac{1}{2}|)$ (worst-case regret of trusting the frozen model). Both are computable from (η_0, β) and both are used by the local-correction screening rule of Part III.

6.2 Irreducibility: The Wall Is Assumed, Not Measured

Theorem 6.2 (Label-free irreducibility). *Assume the standing conditions (S1)–(S7) of Appendix B with binary $\mathcal{Y}$: Polish $\mathcal{X}$, Borel kernel p_0 and budget β , i.i.d. unlabeled draws $x_{1:n} \sim p_W$ with y never observed, the mechanism prior (B) , and randomized Borel estimators. Then for every n , including $n = \infty$, over the worlds $\mathcal{W} = \{p_W \otimes q : q \in I(\beta)\}$ and the region-mean functional $\theta(P) = \mathbb{E}_\mu[\eta_q]$,*

$$\inf_{\hat{\theta}_n \text{ randomized}} \sup_{P \in \mathcal{W}} \mathbb{E}|\hat{\theta}_n - \theta(P)| = \frac{1}{2} D_{\text{mean}}(\beta; R) = \frac{1}{2} \mathbb{E}_{p_W(\cdot|R)}[w_\beta],$$

constant in n . The lower bound is attained already at the two-point subfamily $\{P_-, P_+\}$ built from $\eta_\pm = \Pi_{[0,1]}(\eta_0 \pm \beta)$, and the upper bound by the zero-data midpoint estimator. The risk is strictly positive whenever $p_W(R \cap \{\beta > 0\}) > 0$, and the wall height is a functional of the prior (β, p_0, p_W) alone. Moreover, any test of one drift explanation q against another q' satisfies $\mathbb{E}_{P'}[\psi_n] = \mathbb{E}_P[\psi_n]$ (power equals size, Corollary B.9), and any confidence procedure with uniform coverage $1 - \alpha$ has expected θ -diameter at least $(1 - 2\alpha) D_{\text{mean}}(\beta; R)$ for every n (Corollary B.10, outer-expectation form, under its measurability convention).

Proof. A two-point argument in the tradition of Le Cam (Le Cam, 1986; Tsybakov, 2009), resting on an exact ancillarity. Throughout, $\mu = p_W(\cdot \mid R)$ and $\underline{\eta} = \max(0, \eta_0 - \beta)$, $\bar{\eta} = \min(1, \eta_0 + \beta)$, so $w_\beta = \bar{\eta} - \underline{\eta}$.

Worlds and ancillarity. Each world is the joint law $P = p_W \otimes q$ on $\mathcal{X} \times \mathcal{Y}$, defined by $P(E) = \sum_y \int_{E_y} q(y \mid x) p_W(dx)$, a probability measure with x -marginal p_W and p_W -a.e.-unique disintegration q (Lemma B.1). Push the n -fold product through the observation map $\Pi_A : (x, y) \mapsto x$.

The product–pushforward identity $(\Pi_A^{(n)})_{\#} P^{\otimes n} = ((\Pi_A)_{\#} P)^{\otimes n}$ (Lemma B.2, via agreement on measurable rectangles and Dynkin’s π – λ theorem) together with $(\Pi_A)_{\#} P = p_W$ (which uses only $q(\mathcal{Y} \mid x) = 1$) gives

$$Q_P^{(n)} = p_W^{\otimes n} \quad \text{for every } q \in I(\beta) \text{ and every } n,$$

including $n = \infty$ (infinite product measure plus the cylinder π -system; Proposition B.3). Marginalizing the never-observed label erases q from the model before any data are drawn. The unlabeled sample is exactly ancillary for the conditional, and $D_{\text{TV}}(Q_P^{(n)}, Q_{P'}^{(n)}) = 0$ for any two worlds.

Lower bound. The clipped functions $\eta_{\pm} = \Pi_{[0,1]}(\eta_0 \pm \beta)$ define Borel Markov kernels $q_{\pm}$ with $q_{\pm}(\cdot \mid x) \in B_{\beta}(x)$ for every x (by the binary TV identity, $|\eta_+ - \eta_0| = \min(1 - \eta_0, \beta) \leq \beta$ and symmetrically), so $q_{\pm} \in I(\beta)$, and since $\eta_{\pm} = \bar{\eta}$ resp. $\underline{\eta}$, $\theta(P_+) - \theta(P_-) = \mathbb{E}_{\mu}[w_{\beta}]$ (Lemma B.6). This is the full mean-gap diameter $D_{\text{mean}}(\beta; R)$, because every $q \in I(\beta)$ has $\eta_q \in [\underline{\eta}, \bar{\eta}]$ μ -a.e. (Lemma B.7). The risk of any randomized estimator factorizes through $(Q_P^{(n)}, \theta(P))$ by change of variables for pushforwards, so Le Cam’s two-point inequality (Lemma B.5, proved there in a form covering arbitrary Markov-kernel randomized rules) with $D_{\text{TV}} = 0$ gives, already over the subfamily $\{P_-, P_+\}$,

$$\sup_{P \in \{P_-, P_+\}} \mathbb{E}|\hat{\theta}_n - \theta(P)| \geq \frac{1}{2} |\theta(P_+) - \theta(P_-)| (1 - 0) = \frac{1}{2} \mathbb{E}_{\mu}[w_{\beta}].$$

Enlarging the supremum to all of $\mathcal{W}$ preserves the bound.

Upper bound. No data are needed. For every $q \in I(\beta)$, $\theta(P_q) \in [\mathbb{E}_{\mu}\underline{\eta}, \mathbb{E}_{\mu}\bar{\eta}]$, an interval of length $\mathbb{E}_{\mu}[w_{\beta}]$, so the zero-data midpoint estimator $\hat{\theta}^* = \mathbb{E}_{\mu}[(\underline{\eta} + \bar{\eta})/2]$ has worst-case error $\frac{1}{2} \mathbb{E}_{\mu}[w_{\beta}]$. The two bounds meet, constant in n . Strict positivity when $p_W(R \cap \{\beta > 0\}) > 0$ follows from $w_{\beta} \geq \beta$ (valid also at the both-clips boundary via $\beta \leq 1$), and the value is a functional of the prior (β, p_0, p_W) alone.

Corollaries. Power = size is immediate. Any test ψ_n of q against q' has $\mathbb{E}_{P'}[\psi_n] = \mathbb{E}_P[\psi_n] = \mathbb{E}_{p_W^{\otimes n}}[\psi_n]$, since all observation laws coincide (Corollary B.9). For confidence sets: uniform coverage $1 - \alpha$ puts each of $q_{\pm}$ in C_n with probability $\geq 1 - \alpha$ under the common observation law $p_W^{\otimes n}$, hence both with probability $\geq 1 - 2\alpha$ (union bound). On that event $\text{diam}_{\theta} C_n \geq D_{\text{mean}}(\beta; R)$, and taking (outer) expectations, since the diameter need not be measurable, gives the $(1 - 2\alpha) D_{\text{mean}}(\beta; R)$ bound for every n (Corollary B.10, outer-expectation form). $\square$

Formal version: Theorem B.8 with Corollaries B.9–B.10, Appendix B. The constant is unaffected even by an oracle revealing p_W itself, because within (B) the observation functional factors through p_W alone (Remark B.11).

The height of the wall is an assumption, not a measurement. An unlabeled monitor that reports small drift is therefore reporting the chosen prior.

The wall shrinks only when the model includes structure that the label-free observation map cannot see. The program contract (Online Appendix 1, “the formal program”) enumerates these cases. (a) *Labels:* writing $\bar{\beta}_R = \mathbb{E}_{\mu}[\beta]$, k anchor labels drawn i.i.d. from the region pin its mean to $\min(2\bar{\beta}_R, \tilde{O}(k^{-1/2}))$ (the pure $k^{-1/2}$ collapse operating only once $k \gtrsim \bar{\beta}_R^{-2}$, below which the budget still binds), while *pointwise* recovery needs anchors plus Lipschitz-in- ϕ smoothness at the nonparametric rate $\tilde{O}(k^{-\alpha(R)})$, $\alpha(R) \asymp 1/(2 + d_R)$ with d_R (the analogue of d in (A3)) the intrinsic dimension of the region’s support in ϕ -space. In this branch, covariate geometry governs the rate. (b) *A drift-transport prior* $\beta(x) = \Psi(r_W(x))$, with r_W the window’s density ratio against the source, ties the budget to the estimable density ratio, letting unlabeled data constrain the effective wall (the coupling that Section 6.4 makes learnable from history). (c) *(O3) recurrence* plus a stationarity prior amounts to transferred anchors. Part III quantifies these label requirements.

Two canonical mechanism classes illustrate such a collapse of the wall.

Theorem 6.3 (Canonical collapse classes). (i) Covariate shift, prior-driven. *For the class $\mathcal{C}_{\text{cov}} = \{p \otimes p_0\}$, every observed p_W pins $I_{\mathcal{C}_{\text{cov}}}(p_W) = \{p_0(\cdot \mid \cdot)\}$, so $D_{\Delta}^{\mathcal{C}_{\text{cov}}}(R) = 0$ for every diagonal-vanishing discrepancy, every R , every β , already at $n = 0$. The frozen conditional remains valid and the collapse class C5 applies.* (ii) Label shift, data-pinned. *Fix known class-conditionals $\{\mu_y\}$, linearly independent in $M(\mathcal{X})$ (equivalently, the mixture map $\pi \mapsto \sum_y \pi(y)\mu_y$ is injective). Then for p_W in the class the target prior is identified from unlabeled covariates alone by the linear moment system $\pi_W = M^{-1}\mathbb{E}_{p_W}[T]$, for a bounded Borel moment vector T with nonsingular $M_{jy} = \mathbb{E}_{\mu_y}[T_j]$ (the moment system of [Lipton et al., 2018](#)). The Bayes formula pins $I_{\mathcal{C}_{\text{ls}}}(p_W) = \{q_{\pi_W}\}$ and again $D_{\Delta}^{\mathcal{C}_{\text{ls}}}(R) = 0$, with π_W estimated from n unlabeled draws at rate $\mathbb{E}\|\hat{\pi} - \pi_W\|_2 \leq \|M^{-1}\|_{\text{op}}\sqrt{|\mathcal{Y}|/(4n)}$ (Proposition B.17).*

Proof. (i) is pointwise pinning. By [Lemma B.1](#) the disintegration of a joint law over its covariate marginal is p_W -a.e. unique; since every member of $\mathcal{C}_{\text{cov}}$ has conditional p_0 , whatever p_W is observed the identified set is the singleton $\{p_0(\cdot \mid \cdot)\}$, on which every diagonal-vanishing discrepancy vanishes: $D_{\Delta}^{\mathcal{C}_{\text{cov}}}(R) = 0$ for every R and β , with no data required.

(ii) is a moment system. First, linear independence of $\{\mu_y\}$ is equivalent to injectivity of the mixture map: any relation $\sum_y c_y \mu_y = 0$ among probability measures has $\sum_y c_y = (\sum_y c_y \mu_y)(\mathcal{X}) = 0$ automatically, so affine and linear independence coincide, and a nonzero relation would produce two distinct priors with equal mixtures ([Lemma B.14](#)). Next, the moment vector exists. Evaluation functionals $\nu \mapsto \nu(A)$ separate finite measures (Hahn–Jordan), so a point-separating subset of the dual of the $|\mathcal{Y}|$ -dimensional span $V = \text{span}\{\mu_y\}$ spans V^* , and one selects sets $A_1, \dots, A_{|\mathcal{Y}|}$ whose evaluations form a basis. The indicator moments $T_j = \mathbf{1}_{A_j}$ are bounded Borel with $M_{jy} = \mathbb{E}_{\mu_y}[T_j]$ nonsingular ($Mc = 0$ forces $\sum_y c_y \mu_y$ to be annihilated by a basis of V^* , hence $c = 0$). Moment linearity gives $\mathbb{E}_{p_W}[T] = M\pi$ for every mixture, which inverts to $\pi_W = M^{-1}\mathbb{E}_{p_W}[T]$. Bayes pinning ([Lemma B.15](#), with everywhere-nonnegative Radon–Nikodym versions g_y so the formula defines a Markov kernel at every x) exhibits $q_{\pi_W}(y \mid x) = \pi_W(y)g_y(x)/m_{\pi_W}(x)$ as the p_W -a.e.-unique disintegration, so $I_{\mathcal{C}_{\text{ls}}}(p_W) = \{q_{\pi_W}\}$ and $D_{\Delta}^{\mathcal{C}_{\text{ls}}}(R) = 0$ ([Theorem B.16](#)). The estimation rate is [Proposition B.17](#), where $\hat{\pi} = M^{-1}\hat{b}$ with $\hat{b}_j$ the empirical moment has coordinate variances $\leq 1/(4n)$, and Jensen plus the operator norm give the stated bound. The two collapses thus have different currencies: (i) is prior-driven at $n = 0$; (ii) uses unlabeled data, at the $n^{-1/2}$ rate. $\square$

Formal version: [Proposition B.13](#) and [Theorem B.16](#) (with Lemmas B.14–B.15 and Proposition B.17), Appendix B; the unfalsifiability claim below is [Proposition B.20](#).

The agnostic class itself cannot be falsified from unlabeled data. If p_W is consistent with label shift, the worlds $p_W \otimes q$ under (B), the covariate-shift world, and the label-shift world all induce the same $p_W^{\otimes n}$ for every n , so every label-free test between mechanism classes has power equal to size. The defensible posture is therefore to read D as the *maximal honest wall*: an upper envelope over mechanism priors that respect the budget (any class with $I_{\mathcal{C}}(p_W) \subseteq I(\beta)$ satisfies $D^{\mathcal{C}} \leq D$; [Theorem B.18](#)), which the two collapse classes lower to zero. The scoping matters. A *budget-free* prior can sit above D ([Remark B.19](#)), so “maximal” is claimed only among budget-respecting priors.

6.3 The Second Wall: What the Representation Discarded

The wall of Sections 6.1–6.2 exists for any representation. A second, distinct obstruction is contributed by the frozen representation itself.

Theorem 6.4 (Wall B and orthogonality). *Let $\eta_W(x) = \mathbb{E}_{p_W}[y | x]$, $\bar{\eta}_W = \mathbb{E}[\eta_W | \mathcal{F}_\phi]$ (the $L^2(p_W)$ projection onto $\mathcal{F}_\phi = \sigma(\phi)$, the ϕ -generated σ -algebra), and define $\text{WallB} = \mathbb{E}[\text{Var}(\eta_W | \mathcal{F}_\phi)] = \|\eta_W - \bar{\eta}_W\|_{L^2(p_W)}^2$, equivalently in log-loss the discarded conditional information $\text{WallB}_{\log} = I(x; y | \phi(x)) = I(x; y) - I(\phi(x); y) \geq 0$: the part of the target that no readout of the frozen embedding can express, labels or not (a functional of (p_W, ϕ) alone, in which β and p_0 never appear). Then for any deployed predictor $\hat{p} = g \circ \phi$, with $L_x^* = \mathbb{E}_{p_W}[\eta_W(1 - \eta_W)]$ the full-covariate Bayes squared-error risk,*

$$\mathbb{E}_{p_W}(y - \hat{p})^2 - L_x^* = \underbrace{\|\eta_W - \bar{\eta}_W\|^2}_{\text{WallB} \in \mathcal{F}_\phi^\perp} + \underbrace{\|\bar{\eta}_W - g \circ \phi\|^2}_{\text{reachable error} \in \mathcal{F}_\phi},$$

and the deployable Wall A (a supremum over $I(\beta)$ of $\mathcal{F}_\phi$ -measurable acts) lives entirely in $\mathcal{F}_\phi$, so the two walls are L^2 -orthogonal subspace components: geometric orthogonality, not statistical independence. They are certifiably distinct: there is a world with ϕ invertible ($\text{WallB} = 0$) and $\beta > 0$ interior, where $\text{WallA} = \mathbb{E}[\text{diam } B_\beta] > 0$; and a one-fiber world whose ϕ -level conditional is identified by anchors, where $\text{WallA} = 0$ and $\text{WallB} = \text{Var}(\eta_W | \phi) > 0$. They are coupled only through $\text{WallB} \leq \mathbb{E}[\beta^2]$ (the drift ball is centered at the fiber-constant p_0), so $\beta \rightarrow 0$ caps both walls. For injective ϕ (generic at $D = 192$) the exact WallB vanishes identically; the deployable object is the resolution- ρ deficit WallB_ρ defined below and used in Section 7.2, tied to the readout's operating scale, and every empirical Wall-B statement in this paper is at a stated ρ .

Proof. Pure $L^2(p_W)$ projection geometry. Since $\mathcal{F}_\phi \subseteq \mathcal{F}_x$, the square-integrable ϕ -measurable functions form a closed subspace of $L^2(p_W)$, and $\bar{\eta}_W = \mathbb{E}[\eta_W | \mathcal{F}_\phi]$ is the orthogonal projection of η_W onto it. The residual $\eta_W - \bar{\eta}_W$ is orthogonal to every $g \circ \phi$, since $\mathbb{E}[(\eta_W - \bar{\eta}_W)g(\phi)] = 0$ by the tower property. Any deployed predictor $\hat{p} = g \circ \phi$ lies in the subspace, so

$$\|\eta_W - g \circ \phi\|^2 = \|\eta_W - \bar{\eta}_W\|^2 + \|\bar{\eta}_W - g \circ \phi\|^2 \quad (\text{Pythagoras}),$$

and adding the pointwise decomposition $\mathbb{E}_{p_W}(y - \hat{p})^2 = L_x^* + \|\eta_W - \hat{p}\|^2$ gives the displayed identity. The deployable Wall A is a supremum over $I(\beta)$ of differences of $\mathcal{F}_\phi$ -measurable acts, hence lives entirely in the subspace. The log-loss form is the data-processing deficit $I(x; y) - I(\phi(x); y) = I(x; y | \phi(x)) \geq 0$.

Distinctness. World (i): $\mathcal{X} = \{a, b\}$, $\phi = \text{id}$ (invertible), $p_0 \equiv 0.5$, $\beta = 0.3$ interior. Fibers are singletons, so $\text{WallB} = 0$, while $\text{WallA} = \mathbb{E}[\text{diam } B_\beta] = 0.6 > 0$. World (ii): a single fiber (ϕ constant), $\eta_W(a) = 0.9$, $\eta_W(b) = 0.1$ with equal weights, so $\bar{\eta}_W \equiv 0.5$ and $\text{WallB} = \text{Var}(\eta_W | \phi) = 0.16$ ($\text{WallB}_{\log} = 0.368$ nats). Anchors identifying the ϕ -level conditional pin the one deployable number, so $\text{WallA} = 0$. This world must be realized by anchors, *not* by $\beta = 0$, because of the coupling next.

Coupling. The drift ball is centered at p_0 , which is fiber-constant ($\eta_0 = h_0 \circ \phi$ is $\mathcal{F}_\phi$ -measurable), so $|\eta_W - \eta_0| \leq \beta$ pointwise, and since the conditional variance minimizes conditional squared deviation over $\mathcal{F}_\phi$ -measurable centers, $\text{Var}(\eta_W | \mathcal{F}_\phi) \leq \mathbb{E}[(\eta_W - \eta_0)^2 | \mathcal{F}_\phi] \leq \mathbb{E}[\beta^2 | \mathcal{F}_\phi]$. Taking expectations, $\text{WallB} \leq \mathbb{E}[\beta^2]$, so $\beta \rightarrow 0$ caps both walls and the orthogonality is geometric (subspace), not statistical independence. $\square$

Two qualifications apply to the split: it is exact *pointwise-in-world* (the worst-case certificate obeys only “ $\leq$ ”), and ϕ -resolution is a *shared* lever, since refining ϕ weakly decreases WallB (tower property) yet can *increase* deployable Wall A on a sub-fiber region (Theorem 6.5); the reading of the two walls as mutually inert interventions holds only for fiber-union regions.

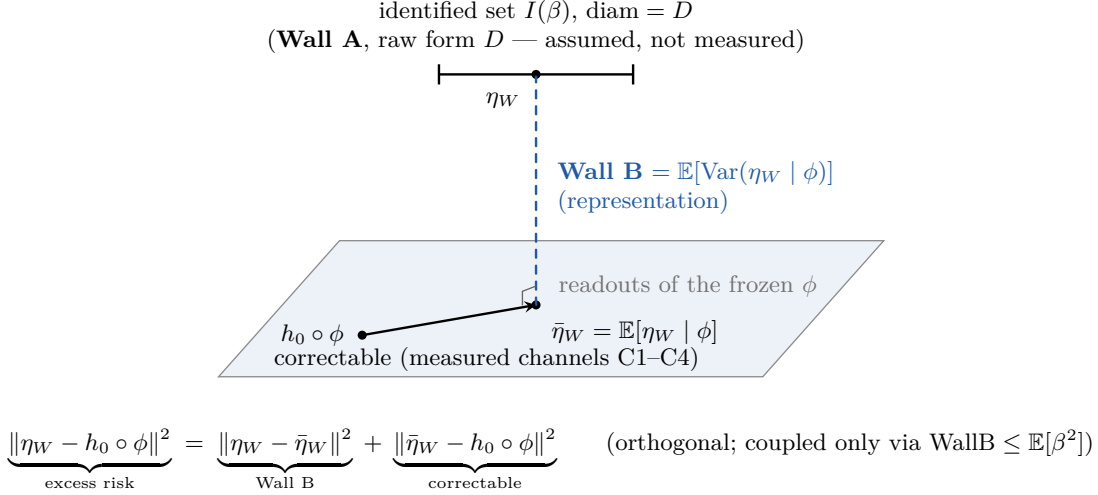


Figure 3: The two walls are orthogonal in the $L^2(p_W)$ projection geometry (geometric orthogonality, not statistical independence), coupled through $\text{WallB} \leq \mathbb{E}[\beta^2]$. The excess risk of a deployed predictor splits Pythagorean-wise into an out-of-plane Wall-B component (what no readout of the frozen ϕ can express, labels or not) and an in-plane component reachable by the channels of Part I. Wall A is not a distance but a *diameter*: the identified set $I(\beta)$ of worlds that no unlabeled quantity distinguishes, sitting at η_W regardless of the representation. Formal statements: the program contract and Appendix F of Online Appendix 1.

Formal version: the two-walls orthogonality block of the program contract (Online Appendix 1, “the formal program”). The deployable fiber-wise form is developed in Appendix F of Online Appendix 1 (theorem cluster F.4: the reduction and dichotomy theorems).

One measurement qualification comes from the program contract: the resolution- ρ deficit of the theorem’s last sentence is $\text{WallB}_\rho = \mathbb{E}[\text{Var}(\eta_W \mid \phi \text{ coarsened to } \rho\text{-balls})]$, monotone in ρ and converging to WallB as $\rho \rightarrow 0$.

The practical content is an *attribution* (Figure 3): residual error decomposes into a Wall-A share (acquire labels), a Wall-B share (change or fine-tune the representation), and a correctable share (run the channels of Part I). Because the two walls respond to different interventions, conflating them (as “adaptation failed” reports implicitly do) prescribes the wrong remedy.

One further result makes Wall A deployable on the frozen readout, whose predictions are constant on the fibers of ϕ :

Theorem 6.5 (Borrowed certainty). *Assume the standing conditions (P1)–(P3) of Appendix F of Online Appendix 1 (standard Borel $\mathcal{X}$, binary $\mathcal{Y}$, Borel ϕ, h_0, β with β not required ϕ -measurable, and a proper regular conditional distribution ρ_z on the fibers of $\nu = \phi_\# p_W$, not to be confused with the resolution ρ) together with the modeling choice (D-proj): the deployable act conditions on the whole window, $\bar{\eta}_q = \mathbb{E}_{p_W}[\eta_q \mid \mathcal{F}_\phi]$ (the frozen readout is one ϕ -function of the deployment window, ignorant of R). Then, writing $\bar{w}_\beta(z) = \mathbb{E}_{p_W}[w_\beta \mid \phi = z]$, $\mu_R = p_W(\cdot \mid R)$, and $\nu_R = \phi_\# \mu_R$: (i) for any region R ,*

$$\text{WallA}(\beta; R) = \mathbb{E}_{z \sim \nu_R} [\mathbb{E}_{p_W}[w_\beta \mid \phi = z]],$$

the inner conditional expectation running over the full fiber population (fiber points outside R included), attained by the explicit pair (q_+, q_-) ; (ii) on fiber-unions $R = \phi^{-1}(B_R)$ the deployable

and raw walls coincide, $\text{WallA}(\beta; R) = D_{\text{mean}}(\beta; R)$; (iii) on sub-fiber regions the gap

$$D_{\text{mean}}(\beta; R) - \text{WallA}(\beta; R) = \mathbb{E}_{z \sim \nu_R} [\mathbb{E}_{\mu_R}[w_\beta \mid \phi = z] - \mathbb{E}_{p_W}[w_\beta \mid \phi = z]]$$

is signed: *strictly positive (deployable wall smaller) when R over-selects high- w_β points whose fiber-mates are pinned (the drifting cell borrows certainty), negative when R under-selects. The resulting quantity is a label-free plug-in estimand given $(\phi, \eta_0, \beta, p_W)$ (Corollary F.5; its estimand identity is proven, its coarse- ϕ estimator rate is graded partial).*

Proof. The engine is the conditional Aumann-interval description of the reachable acts (Lemma F.1): with $\eta_\pm = \Pi_{[0,1]}(\eta_0 \pm \beta)$ and $\bar{\eta}_\pm(z) = \int \eta_\pm d\rho_z$ (ρ_z the proper regular conditional distribution on the fiber over z), for ν -a.e. z , as q ranges over $I(\beta)$ the fiber-average $\bar{\eta}_q(z)$ sweeps *exactly* the interval $[\bar{\eta}_-(z), \bar{\eta}_+(z)]$ of width $\bar{w}_\beta(z)$, both endpoints attained by the explicit everywhere-selections $q_\pm$. (Feasibility off a p_W -null set transfers to ν -a.e. fibers since $\int \rho_z(N_q) d\nu = p_W(N_q) = 0$; surjectivity comes from the convex combinations $\lambda\eta_+ + (1-\lambda)\eta_-$, measurable selections of the convex-valued Γ_β .)

(i), *general R .* For $q, q' \in I(\beta)$ the lemma gives $|\bar{\eta}_q(z) - \bar{\eta}_{q'}(z)| \leq \bar{w}_\beta(z)$ ν -a.e. Since $\nu_R = \phi_\# \mu_R \ll \nu$, integrating against ν_R bounds the deployable wall by $\int \bar{w}_\beta d\nu_R$, and the pair (q_+, q_-) attains the bound because $\bar{\eta}_{q_+} - \bar{\eta}_{q_-} = \bar{w}_\beta \circ \phi \geq 0$ pointwise. Neither step uses $R \in \mathcal{F}_\phi$, so $\text{WallA}(\beta; R) = \mathbb{E}_{z \sim \nu_R} [\mathbb{E}_{p_W}[w_\beta \mid \phi = z]]$, the inner expectation over the full fiber population.

(ii), *fiber-unions.* When $R = \phi^{-1}(B_R)$, conditioning on the $\mathcal{F}_\phi$ -event R leaves each within-fiber law ρ_z ($z \in B_R$) unchanged, and a tower computation collapses (i) to the raw wall: $\int_{\phi^{-1}(B)} w_\beta d\mu_R = \int_{\phi^{-1}(B)} \bar{w}_\beta \circ \phi d\mu_R$ for every Borel B , whence $\mathbb{E}_{\mu_R}[\bar{w}_\beta \circ \phi] = \mathbb{E}_{\mu_R}[w_\beta] = D_{\text{mean}}(\beta; R)$.

(iii), *sub-fiber regions.* Exactly that tower step fails: $\bar{w}_\beta \circ \phi = \mathbb{E}_{p_W}[w_\beta \mid \mathcal{F}_\phi]$ is *not* a version of $\mathbb{E}_{\mu_R}[w_\beta \mid \mathcal{F}_\phi]$ when R is sub-fiber, because conditioning on R tilts the within-fiber law. Disintegrating both walls under ν_R leaves precisely the stated signed gap (the discrepancy between the R -tilted and full-population fiber averages, with the sign of the reweighting).

Check on the worked example (Corollary F.3, Figure 4): a two-point fiber with $\eta_0 \equiv 0.5$, $\beta = (0.3, 0)$, equal weights, R the drifting cell $\{a\}$. Pointwise the cell admits any $\eta_q(a) \in [0.2, 0.8]$ (raw wall 0.60), but the pinned mate forces $\bar{\eta}_q(z_0) = \frac{1}{2}\eta_q(a) + 0.25 \in [0.35, 0.65]$, so $\text{WallA} = 0.30$, half the raw wall, with gap 0.30 matching (iii). $\square$

The multi-class structure stays in Appendix F of Online Appendix 1 (cluster F.4). The whole binary picture carries over verbatim to TV/mean-gap discrepancies (concentric within-fiber balls share a common worst direction), while non-monotone discrepancies such as ℓ_2 incur an additional strict directional-misalignment suppression.

Formal version: the proof above transcribes Theorems F.2–F.3, Appendix F of Online Appendix 1. Remaining appendix-grade pieces are Corollary F.5 (the label-free estimand form) and Proposition F.4 with cluster F.4 (multi-class structure). Figure 4 draws the theorem's worked example (Corollary F.3) to scale.

One hypothesis deserves separate mention because the phenomenon rests on it. (D-proj) is semantic, not a theorem. Under the alternative act model $\bar{\eta}_q = \mathbb{E}_{\mu_R}[\eta_q \mid \mathcal{F}_\phi]$ the inner and outer measures coincide and borrowing vanishes for every R . But it is forced by the definition of the deployable act itself (a frozen readout emits one ϕ -function of the whole deployment window and does not know the analyst's region), so the appendix's classification of (D-proj) as a modeling choice is conservative rather than a weakness (Remark F.3). The same remark records the shared-lever effect quoted under Theorem 6.4. As ϕ refines, $\text{WallA} \rightarrow D_{\text{mean}}(\beta; R)$, while coarsening merges drifting cells with pinned mates and lowers the deployable wall.

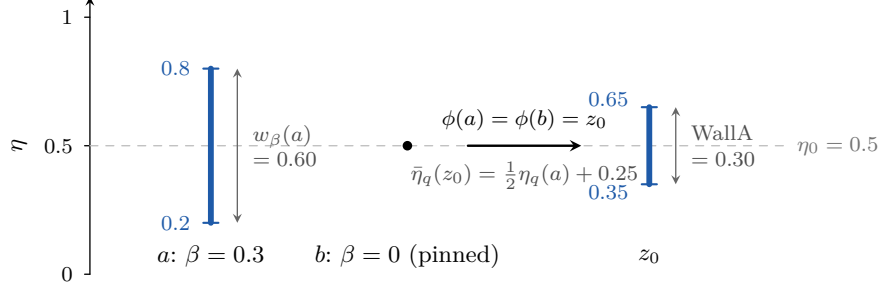


Figure 4: Borrowed certainty, drawn to the numbers of Corollary F.3. Left: pointwise, the drifting cell a admits any $\eta \in [0.2, 0.8]$ (raw wall 0.60) while its fiber-mate b is pinned at 0.5. Right: the frozen readout emits one number per fiber, so the deployable identified interval at z_0 is $[0.35, 0.65]$, which gives WallA = 0.30, half the raw wall. On fiber-unions deployable and pointwise walls coincide (Theorem F.2); on sub-fiber regions the drifting cell borrows certainty from its pinned mates (Theorem F.3).

6.4 Learning the Wall from History

The wall is a prior within one window, but deployments see many windows, and occasionally labels. The question is whether the wall's height can be *learned* from labeled history and transferred forward, converting the assumption into a measurement. A threshold calculation identifies one sufficient regular branch and a corresponding, still incomplete lower-bound program. Both are governed by one geometric quantity: how much drift mass sits near the decision boundary.

Let d_0 be the (deduplicated) intrinsic dimension of the embedding (instantiating the model parameter d of assumption (A3)), α_s the smoothness of the drift-response field, and γ the margin exponent of (Mgn_γ) ($\Pr[0 < |\bar{\eta} - \eta_0| \leq t] \lesssim t^\gamma$). Define

$$a = \frac{\alpha_s}{2\alpha_s + d_0} \in \left(0, \frac{1}{2}\right], \quad \gamma^* = \frac{1}{2a} - 1 = \frac{d_0}{2\alpha_s}, \quad \text{so that} \quad a(1 + \gamma^*) = \frac{1}{2} \text{ identically.}$$

The estimand is the covered-part wall $D_{\text{cov}} = 2 \int_{\mathcal{J}} \Psi d\nu$, $\Psi(r) = \mathbb{E}[|\bar{\eta}_J - \eta_0| \mid R = r]$ (here R is the novelty index of Section 3.1, not a region; $\mathcal{J}$ is the covered ratio interval of (A2) and the subscript J indexes the new window), of a new anchor-free window with novelty law ν and representer $w = d\nu/d\bar{\mu}$.

Theorem 6.6 (Regular branch of the margin boundary). *Assume (A0)–(A7), (A5'), the interior regime, (T), and (Mgn $_\gamma$), with K historical windows of k MCAR anchor labels each ($N = Kk$), the fold-balanced anchor-level cross-fitting of Definition G.1.2 (including held-out Stage-1 evaluation), the Stage-1 sup-norm rate $\varepsilon_1 \asymp (\log k/k)^a$, and the DML side conditions (H1)–(H4) and (H6) of Theorem G.1.4: overlap, cross-fitting, an $o_P(N^{-1/2})$ product rate for the two nuisances, proportional target size, and a sup-norm density-ratio rate (H6). If $\gamma > \gamma^*$ (drift stays clear of the boundary), the wall is estimable at the parametric rate: the sign-corrected, cross-fitted one-step estimator built on the efficient influence function $\varphi = 2[w(b - \Psi) + (\Psi - \theta)]$, where $b := |\bar{\eta}_J - \eta_0|$ is the drift magnitude and $\theta := D_{\text{cov}}/2$, satisfies*

$$\sqrt{N}(\hat{D}_{\text{cov}} - D_{\text{cov}}) \Rightarrow \mathcal{N}(0, V), \quad V = 4 \mathbb{E}_{\bar{\mu}}[w^2 \sigma_b^2] + 4\tau_{NT} \text{Var}_\nu(\Psi),$$

where $\sigma_b^2(r) := \text{Var}(b \mid R=r)$ is the conditional variance of the drift magnitude and $\tau_{NT} := \lim N/m$ is the anchor-to-target size ratio of (H4), and no regular estimator improves on the rate. By the

Hájek–Le Cam convolution and local-asymptotic-minimax theorems (Hájek, 1970; Le Cam, 1986), $\liminf_N N \text{Var}(T_N) \geq V_{A1}^* > 0$ for every regular T_N , where $V_{A1}^* \leq V^*$ is the efficient constant on the stationarity-constrained tangent (attainment by inverse-variance pooling is graded partial in the appendix; $= V^*$ only under disjoint novelty coverage). For fixed K , (A6) overlap, (T), and $\gamma > \gamma^*$ are sufficient for this conclusion modulo the Appendix-G regime-(I) qualifications (Theorem I.3.2). At equality the exponents meet, but the displayed Stage-1 rate retains a logarithmic factor and the present argument does not establish $\sqrt{N}$ -regularity.

Conditional lower-bound program (not part of the proved regular theorem). One open analytic input conditions the program’s $\gamma = 0$ results. Gap (2c) is a quantitative resummation estimate for the permutation-mixture affinity of the derandomized lower-bound construction of Appendix K of Online Appendix 1: its band structure and first band are proven, and exact-rational enumeration supports the general estimate, which remains open. The extension to $0 < \gamma < \gamma^*$ additionally requires a margin adaptation that Remark H.8(iv) leaves open. We therefore state only the $\gamma = 0$ consequences (no margin: (Mgn₀) is vacuous), explicitly conditional on gap (2c).

Proposition 6.7 (Fixed- K lower bound at $\gamma = 0$; conditional on gap (2c)). *Under the hypotheses of Theorem 6.6 with $\gamma = 0$ and each fixed $K \geq 1$, and conditional on the Appendix-K affinity estimate (2c) together with the cited transfer steps in Theorem E.3,*

$$\inf_{\hat{\theta}} \sup \mathbb{E}|\hat{\theta} - \theta| \gtrsim (N/K)^{-a}/\sqrt{K} \quad (\text{up to logarithmic factors}).$$

Because K is fixed and $a < 1/2$, this conditional floor is larger than $N^{-1/2}$ and hence implies non-regularity. At $K = 1$ it meets the available Cai–Low upper bound up to logarithmic factors.

Proposition 6.8 (Growing- K rate at $\gamma = 0$; conditional on gap (2c)). *Under the hypotheses of Theorem 6.6 with $\gamma = 0$ and $k^\delta \leq K \leq \text{poly}(k)$ for some fixed $\delta > 0$, and conditional on the graded affinity estimate (2c) together with the cited transfer steps in Theorem E.3, the minimax rate is*

$$\inf_{\hat{\theta}} \sup \mathbb{E}|\hat{\theta} - \theta| \asymp (N/K)^{-a} \quad (\text{up to polylog}), \quad \text{not } k^{-a}/\sqrt{K}.$$

Conditionally on gap (2c), accumulating more windows does not improve the worst-case rate: *independent-window averaging reduces the stochastic error by $\sqrt{K}$ unconditionally, but an adversary that drifts coherently across windows keeps the non-smoothness bias at the single-window floor.*

Proof of Theorem 6.6, given the Stage-1 inputs. The estimand is linear in Ψ : $\theta = \langle \Psi, w \rangle_{L^2(\bar{\mu})}$ with $D_{\text{cov}} = 2\theta$, so it is pathwise differentiable with the two-sample efficient influence function φ above (source gradient $w(b - \Psi)$, target gradient $\Psi - \theta$; Proposition G.1.1). The estimator is the cross-fitted one-step (DML2) of Definition G.1.2: split every window’s anchors across all folds with equal per-window quotas, train $(\hat{\Psi}^{(-\ell)}, \hat{w}^{(-\ell)})$ and each Stage-1 $\hat{\eta}_j^{(-\ell)}$ off the evaluation anchors in fold ℓ , and read the held-out drift through the recovered sign, $\hat{b}^{\text{obs}} = \hat{s}(y - \eta_0)$, where $g := \bar{\eta}_J - \eta_0$, $s := \text{sign } g$, and $\hat{s} := \text{sign}(\hat{\eta}_J - \eta_0)$.

Step 1: sign recovery; the threshold computation. On the Stage-1 sup-norm event $\|\hat{\eta}_J - \bar{\eta}_J\|_\infty \leq \varepsilon_1$ (probability $\geq 1 - \sum_j k_j^{-2}$, Lemma C.11’), the *straddle* holds pointwise: a sign flip forces $0 < b \leq \varepsilon_1$ (Lemma H.1). Hence under (Mgn _{γ}) the weighted sign error obeys $\mathbb{E}[b \mathbf{1}(\hat{s} \neq s)] \leq \varepsilon_1 \bar{P}(0 < b \leq \varepsilon_1) \leq C\varepsilon_1^{1+\gamma}$, and, via the pointwise identity $(\hat{s} - s)g = -2b \mathbf{1}(\hat{s} \neq s)$, the *entire* outcome bias of $\hat{b}^{\text{obs}}$ is $\|\beta_1^{\text{sign}}\|_{L^1(\bar{\mu})} \leq 2C\varepsilon_1^{1+\gamma}$. Wherever the sign is right, the raw residual is conditionally

unbiased, so no Jensen/smoothing bias is ever incurred. With $\varepsilon_1 \asymp k^{-a}$ (logs absorbed by the strict inequality) and $k = N/K$, K fixed,

$$\varepsilon_1^{1+\gamma} = o(N^{-1/2}) \iff (1+\gamma)a > \frac{1}{2} \iff \gamma > \gamma^* = \frac{1}{2a} - 1 = \frac{d_0}{2\alpha_s}.$$

This computation *is* the threshold.

Step 2: one-step decomposition. Write $\hat{D}_{\text{cov}} - D_{\text{cov}} = 2[(\nu_m - \nu)\hat{\Psi} + (\bar{P}_N - \bar{P})(\hat{w}(\hat{b}^{\text{obs}} - \hat{\Psi}))] + 2R_2$ with $R_2 = R_{\text{prod}} + R_{\text{ratio}} + R_{\text{stage1}}^{\text{sign}}$. The Neyman-orthogonal *product* remainder is (Lemma G.1.3)

$$\frac{1}{2} \mathbb{E}[\hat{D}_{\text{cov}} \mid \hat{\Psi}, \hat{w}] - \theta = \int (\hat{\Psi} - \Psi)(w - \hat{w}) d\bar{\mu}, \quad |R_{\text{prod}}| \leq \|\hat{\Psi} - \Psi\|_{L^2(\bar{\mu})} \|\hat{w} - w\|_{L^2(\bar{\mu})}$$

(double robustness in (Ψ, w)), so the cube-root shape rate of the isotonic $\hat{\Psi}$ enters only *multiplied* by the fast, one-dimensional representer error, and (H3) gives $R_{\text{prod}} = o_P(N^{-1/2})$; (H6) handles R_{ratio} ; Step 1 gives $|R_{\text{stage1}}^{\text{sign}}| \leq \Gamma_{\text{cov}} \|\beta_1^{\text{sign}}\|_{L^1} = o(N^{-1/2})$ above threshold.

Step 3: CLT. Cross-fitting replaces any Donsker condition. [Lemma G.14'](#) proves the fold-conditional empirical-to-population step under the window-stratified (non-i.i.d.) fold-balanced design with equal quotas $k_j \equiv k$. The published i.i.d. DML theorem ([Chernozhukov et al., 2018](#), Thm. 3.1) does not apply as stated, and only its Step-2 device is imported. What remains is the empirical process of the fixed φ on two independent bounded mean-zero averages. Lindeberg–Feller ([van der Vaart, 1998](#), Prop. 2.27) plus Slutsky give $\sqrt{N}(\hat{D}_{\text{cov}} - D_{\text{cov}}) \Rightarrow \mathcal{N}(0, V)$, where the variance identification $V = 4\mathbb{E}_{\bar{\mu}}[w^2\sigma_b^2] + 4\tau_{NT} \text{Var}_{\nu}(\Psi)$ *requires* stationarity (A1): with nonvanishing per-window score means the stratified variance is strictly below the mixture V (the CI constant is conservative; the estimator stays centered on the pooled θ).

Step 4: no regular estimator improves. The stratified anchor experiment is LAN in the interior regime, so the Hájek–Le Cam convolution and LAM theorems applied on the stationarity-constrained tangent $\mathcal{T}_{A1}$ give $\liminf_N N \text{Var}(T_N) \geq V_{A1}^* > 0$ for every regular T_N (Theorem G.3.2; for the non-i.i.d. experiment via the convolution theorem of [McNeney and Wellner, 2000](#)).

Remainders left to the appendix, as noted there: the efficient constant V_{A1}^* under overlapping novelty coverage (and the GMM/inverse-variance pooling attaining it) is partial; and the ratio channel keeps (H6) at $\varepsilon_r = o(N^{-1/2})$ unconditionally, ε_r^2 only under (A5') plus the conjectural DML-in- r orthogonalization (Conjecture G.2.3). $\square$

The conditional lower-bound program (Appendix H; Appendices I and K of Online Appendix 1). The candidate non-regularity mechanism lives entirely in the labels $\rightarrow |\cdot|$ map. The constructions behind Propositions 6.7 and 6.8 have three stages; their dependence on gap (2c) is carried throughout.

Single window (Theorems E.3 and H.6). By LP duality of the moment problem, two symmetric priors ν_0, ν_1 on $[-1, 1]$ matching moments to degree D can differ in $\mathbb{E}|v|$ by $2\delta_D$, $\delta_D = \beta_*/D(1+o(1))$ with Bernstein’s constant $\beta_* \approx 0.2802$, the extremal pair sitting on the Chebyshev alternation points of the best degree- D approximant of $|x|$ (Proposition H.5, following [Cai and Low, 2011](#); see also [Lepski et al., 1999](#)). Partition the d_0 -dimensional ϕ -support into $m_h = h^{-d_0}$ cells of side $h = k^{-1/(2\alpha_s+d_0)}$. On each cell set the drift $g = Av_c$ at Hölder-maximal amplitude $A = h^{\alpha_s} = k^{-a}$, with i.i.d. cell heights $v_c \sim \nu_\iota$, $\iota \in \{0, 1\}$. With $A^2 \cdot$ (anchors per cell) $= \Theta(1)$ and $D \sim \log k / \log \log k$ so that $(D+1)! \gtrsim m_h$, the mixture χ^2 stays $O(1)$ while the functional gap is $A \cdot 2\delta_D \asymp k^{-a}(\log \log k / \log k)$. Conditional on the Appendix-K affinity estimate, Le Cam’s two-point method gives the $K = 1$ floor (the sharp constant and the uniform-over-cells Bernoulli $\leftrightarrow$ Gaussian transfer are cited results, [Cai and Low, 2011](#); [Nussbaum, 1996](#); Theorem E.3 carries this external dependence).

Fixed K (Theorem H.7). Place the K windows on *disjoint* novelty cells (Stage-1 cannot pool), each of ν -mass $1/K$, and let each window be active with probability p under Λ_0 and p' under Λ_1 , $|p' - p| = c/\sqrt{K}$, the active state being the fuzzy prior above, coherently oriented. The affinities tensorize, $\chi^2(\Lambda_1\|\Lambda_0) = \prod_j (1 + \chi_j^2) - 1 \leq e^{\sum_j \chi_j^2} - 1$ with $\sum_j \chi_j^2 \asymp K \cdot (c^2/K) \cdot O(1) = O(1)$, so $\text{TV} \leq \frac{1}{2}$, while the functional separates by $(p' - p) \cdot 2\delta_D A \asymp (N/K)^{-a}/\sqrt{K}$ up to the logarithmic factor $\log \log k / \log k$ of the single-window case. Conditional on gap (2c), Le Cam gives the fixed- K floor, which is enough for non-regularity, since $(k^{-a}/\sqrt{K})/N^{-1/2} = k^{1/2-a} \rightarrow \infty$ and no logarithmic factor changes that. Assouad is inapplicable here, because the functional sees the sign pattern $\tau \in \{0, 1\}^K$ only through the scalar $\sum_j \tau_j$, a Hamming *upper* bound where Assouad needs a lower one (Remark H.8), which is exactly why this bound carries the $\sqrt{K}$ deficit.

Growing K : degree-tuning (Theorem I.2.1 and Proposition I.2.2). The amplitude-tuned bound loses $\sqrt{K}$ in the gap. The general- K bound keeps the amplitude and tunes the *degree*. At the Cai–Low bandwidth the per-window affinity $\chi_j^2 \asymp m_h/(D+1)!$ falls *super-exponentially* in the matched degree while the gap $\asymp A\delta_D \asymp A/D$ falls only *polynomially*. Choosing $(D+1)! \asymp m_h K/c_0$ with $c_0 \leq \ln 2$ drives $\chi_j^2 \asymp c_0/K$, so that on independent windows $(1 + \chi_j^2)^K - 1 \leq e^{c_0} - 1 \leq 1$ and $\text{TV} \leq \frac{1}{2}$ (Ingster–Suslina tensorization; Ingster and Suslina, 2003), while raising the degree from D_1 ($(D_1+1)! \asymp m_h$) costs the gap only the factor $1 + O(\log K / \log k)$. Conditional on the permutation-mixture affinity estimate, placing the *same* moment-matched prior coherently on all K windows therefore retains a gap of order $k^{-a}/\text{polylog}$ at $\text{TV} \leq \frac{1}{2}$, and the functional concentrates (variance $\asymp k^{-2a}/(Km_h) \ll$ the squared gap), so the two-fuzzy-hypotheses Le Cam bound (Tsybakov, 2009, Thm. 2.15) yields the floor $r_k/\text{polylog}(k)$, $r_k := (k/\log k)^{-a} = (N/K)^{-a}$ up to the $(\log k)^a$ factor, for $k^\delta \leq K \leq \text{poly}(k)$, which would refute the $k^{-a}/\sqrt{K}$ averaging gain. The competing Cauchy–Schwarz ceiling $\chi_j^2 = (\delta_j/r_k)^2$, δ_j the per-window functional gap, which would force the $r_k/\sqrt{K}$ cap, holds only along the amplitude curve and is not a valid universal lower-bound cap (Proposition I.2.2).

Admissibility of the coherent construction. The construction’s admissibility is established in Appendix K of Online Appendix 1, at every K . A naive i.i.d.-heights construction is *inadmissible* at every $K \geq 1$ (its realized response curve is a.s. non-monotone); the admissible version is a derandomized, phase-complementary construction with *exactly constant* Ψ . The derandomization trades the i.i.d. tensorization for a permutation-mixture affinity bound, proven modulo the single graded gap (2c). Because the derandomization is what makes the construction admissible already at $K = 1$, Propositions 6.7 and 6.8 carry the same conditioning. If (2c) failed entirely, both conditional floors would lapse; the bracket $[r_k/\sqrt{K}, r_k]$ of Remark H.8(ii) would then remain only as the candidate range for the growing- K rate, with neither end established.

Reconciliation and refinements. The certified-sign regime of Appendix G of Online Appendix 1, where the signed drift is observed and the Stage-1 bias vanishes for every γ , is the *perfect-margin endpoint* of the regular-side calculation. Its side condition $\varepsilon_1^{1+\gamma} = o(N^{-1/2})$ is algebraically identical to $\gamma > \gamma^*$, and its “outcome-bias hypothesis fails for all $K \geq 1$ ” is exactly the no-margin instance $\gamma = 0$ (Theorem H.11). Below threshold the candidate rate-relevant achiever is the polynomial-debiased U -statistic of Proposition H.10, which confines poly-debiasing to the near-crossing layer (mass $\lesssim \varepsilon_1^\gamma$) and runs the linear one-step on the bulk. It is *not* the naive plug-in $|\hat{\eta}_J - \eta_0|$, whose one-signed Jensen bias is coherent on every window and therefore never averages. Two refinements (Corollary H.4(b); Theorem I.3.2 of Online Appendix 1): for *growing* $K = K_N$ with $\log K / \log N \rightarrow \lambda \in (0, \frac{1}{2})$, the parametric branch is established only for $\gamma > \gamma^*(\lambda) = \frac{1}{2a(1-\lambda)} - 1 > \gamma^*$, leaving the interval $(\gamma^*, \gamma^*(\lambda))$ with a proven $N^{-1/2}$ lower bound but no proven attaining estimator; and rates of non-smooth functionals routinely jump at phase boundaries, so continuity across γ^* is

not a valid selection principle. The below-threshold rate is fixed by the two-sided bracket, not by matching the parametric branch at the threshold. Finally, the deployment reading of the $\sqrt{K}$ (Remark I.3.3): the gain is real but sits on the wrong error component. It averages the stochastic error unconditionally and helps on *bias-incoherent* recurrences, while a worst-case adversary drifts coherently and pays no averaging at all. Regime recurrence therefore supplies a $\sqrt{K}$ lever *under an incoherence prior on the recurrences*, degrading to the bare per-window price under coherent drift. Section 6.5 examines how the resulting threshold behaves under stream-level proxies and an explicit sensitivity scenario; it does not identify a worst-case price for every data set.

Formal statements and proofs: the regular branch is Theorem G.1.4 (DML2 CLT) with the matching regularity bound Theorem G.3.2 and its regime-(II) extension Theorem H.3. The conditional $\gamma = 0$ lower-bound statements are Theorem H.7 (fixed K) and Theorem I.2.1, Proposition I.2.2, and Theorem I.3.1 (general K). The positive-margin adaptation below the threshold remains open. Appendix H and Appendices G, I, and K of Online Appendix 1 give the full status record.

Status. The upper rate, the threshold identity, and the sign-recovery reduction behind the magnitude case are theorems under their stated hypotheses. One hypothesis deserves separate mention: Theorem 6.6 *assumes* the sup-norm density-ratio rate (H6), and the program that would establish (H6) from primitives is itself graded partial in Online Appendix 1, so the theorem certifies the implication, not the availability of its premise. The efficient *constant* under overlapping windows and the second-order behavior of the density-ratio channel are graded partial in the appendix. The growing- K admissibility step of the coherent construction is settled exactly (Appendix K of Online Appendix 1, via the derandomized exactly-constant construction). What remains there is the single graded affinity estimate (gap (2c)); it conditions the $\gamma = 0$ lower bounds at every K , fixed and growing alike (Propositions 6.7 and 6.8), and the margin- γ adaptation of the fixed- K bound remains open (Remark H.8).

Estimating the wall requires estimating $|\bar{\eta} - \eta_0|$, and the absolute value is non-smooth where drift crosses the boundary. The margin condition keeps this kink rare enough for sign recovery to restore a smooth problem. For deployment, the wall is learnable from history at low label cost when drift stays away from decisions. The conditional lower-bound program explains why crossings are the difficult case, but it does not yet prove that conclusion throughout the positive-margin sub-threshold region.

6.5 Proxy Estimates and Sensitivity Analysis on Twenty-Three Data Sets

The threshold depends on three quantities that are difficult to estimate jointly: intrinsic dimension d_0 , drift smoothness α_s , and margin exponent γ . We therefore separate two evidence levels. On eight TabReD streams, labels permit stream-level proxy estimates of all three quantities, although the smoothness and margin estimates are noisy. On fifteen further shift benchmarks, labels for the relevant temporal drift field are unavailable; those data sets contribute only d_0 and are shown under the favorable rough-field scenario $(\alpha_s, \gamma) = (1, 1)$. Figure 5 visualizes both levels and Table 4 records the inputs. Appendix J gives the protocol and its limitations.

For the TabReD streams, the filled-circle locations should be read as estimated inputs, not exact coordinates. Six of the eight $\hat{\gamma}$ values lie in $[0.95, 1.06]$, while the two-range slope check of Appendix J disagrees on five streams. Nevertheless, if the rough-field bound $\alpha_s \leq 1$ is accepted and the measured margins are representative, then $2\alpha_s\hat{\gamma} < d_0$ for all eight streams; the closest is SBERBANK-HOUSING, with a favorable left side of 2.12 (at $\alpha_s = 1$; 1.63 at the measured $\hat{\alpha}_s = 0.77$) against $d_0 = 3.01$. This is conditional evidence for the difficult branch, not an uncertainty-quantified phase assignment. The known biases run in the favorable direction for this reading: the α_s proxy over-states smoothness (Appendix J), inflating the left side, so the below-threshold comparison is

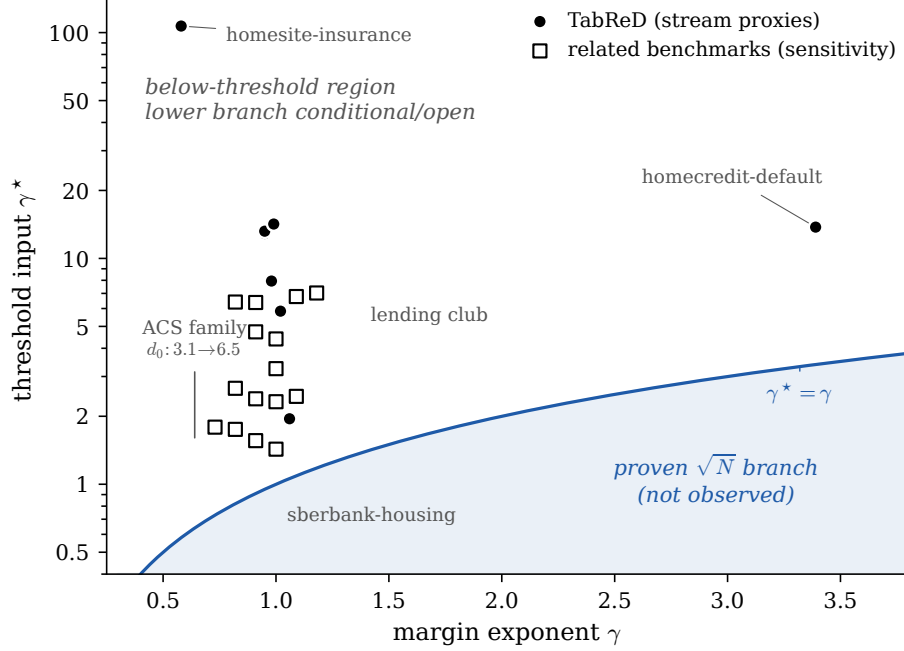


Figure 5: Proxy estimates and sensitivity scenarios relative to the margin boundary (log vertical scale). Filled circles use the stream-level $\hat{\gamma}$ and $\hat{\alpha}_s$ available for eight TabReD streams. Open squares are not measured phase locations: they fix $(\alpha_s, \gamma) = (1, 1)$ and plot $\gamma^* = d_0/2$ for fifteen related benchmarks. Horizontal dodging of the squares (at most 0.27) is purely for legibility. The shaded region $\gamma > \gamma^*$ is the strict sufficient branch for $\sqrt{N}$ regularity; equality is not classified by the present theory. All proxy points and all sensitivity scenarios fall on the below-threshold side, conditional on their stated inputs; the lower branch there is conditional or open, depending on γ .

conservative.

For the fifteen related benchmarks, neither α_s nor γ is observed. Their open squares answer a narrower question, namely whether, under the generic transversal-margin hypothesis $\gamma = 1$ and the favorable roughness value $\alpha_s = 1$, the estimated dimension alone would place the scenario below threshold. The answer is yes for every displayed $d_0 > 2$. Changing either assumption can change that answer, so these rows cannot support the claim that all fifteen data sets have measured non-regular walls. All fifteen were re-estimated after deduplication: the low-uniqueness members of the nine re-measured data sets moved materially (two left the low-dimensional reading entirely), while the six high- d_0 rows are duplicate-free to four decimals (99.97–100% unique) and move by at most 0.10, comparable to the TabReD-8 robustness check (at most 0.08).

The study therefore supplies a consistent empirical warning, not a population law: the observed TabReD proxies and a favorable sensitivity scenario do not populate the regular branch. Establishing an empirical phase map would require repeated labeled windows, uncertainty intervals for (d_0, α_s, γ) , and sensitivity to the dimension and variogram estimators. The semi-synthetic experiment of the next subsection supplies what the stream proxies cannot: controlled inputs whose effective exponents can be calibrated and checked numerically. The ACS ordering (employment, unemployment, public coverage, income) remains a useful descriptive pattern in d_0 , but is not by itself a validation of the theoretical threshold.

data set	domain	d_0	α_s input	γ input	γ^*	status
ECOM-OFFERS	TabReD (cls)	3.61	0.14	0.95	12.89	proxy below
SBERBANK-HOUSING	TabReD (reg)	3.01	0.77	1.06	1.95	proxy below
WEATHER	TabReD (reg)	11.10	0.95	1.02	5.84	proxy below
COOKING-TIME	TabReD (reg)	11.12	0.70	0.98	7.94	proxy below
HOMESITE-INSURANCE	TabReD (cls)	10.69	0.05	0.58	106.90	proxy below
HOME CREDIT-DEFAULT	TabReD (cls)	13.76	0.50	3.39	13.76	proxy below
DELIVERY-ETA	TabReD (reg)	16.63	0.63	0.95	13.20	proxy below
MAPS-ROUTING	TabReD (reg)	17.61	0.62	0.99	14.20	proxy below
CICIDS	security	2.86	[1]	[1]	1.43	sensitivity below
ACS_EMPLOYMENT	census	3.11	[1]	[1]	1.56	sensitivity below
ACS_UNEMPLOYMENT	census	3.49	[1]	[1]	1.75	sensitivity below
UNSW_NB15	security	3.58	[1]	[1]	1.79	sensitivity below
IEEE_FRAUD	fraud	4.64	[1]	[1]	2.32	sensitivity below
ACS_PUBLIC COVERAGE	census	4.78	[1]	[1]	2.39	sensitivity below
BIKE_SHARING	reg	4.90	[1]	[1]	2.45	sensitivity below
CREDIT_CARD_FRAUD	fraud	5.32	[1]	[1]	2.66	sensitivity below
ACS_INCOME	census	6.49	[1]	[1]	3.25	sensitivity below
DIABETES_READMISSION	medical	8.79	[1]	[1]	4.40	sensitivity below
BRFSS_DIABETES	medical	9.46	[1]	[1]	4.73	sensitivity below
SEPSIS	medical	12.75	[1]	[1]	6.38	sensitivity below
BAF	fraud	12.81	[1]	[1]	6.41	sensitivity below
LENDING_CLUB	credit	13.55	[1]	[1]	6.78	sensitivity below
MIMIC_IV_MORTALITY	medical	14.05	[1]	[1]	7.03	sensitivity below

Table 4: Inputs to Figure 5. In the top block, $\hat{\alpha}_s$ and $\hat{\gamma}$ are noisy stream-level proxies and $\gamma^* = d_0/(2\hat{\alpha}_s)$ is recomputed from the displayed values. In the lower blocks, bracketed values are assumptions for a sensitivity analysis, so $\gamma^* = d_0/2$ by construction. All fifteen related benchmarks use deduplicated d_0 ; the six high- d_0 rows are 99.97–100% unique, so their correction is bounded by subsample noise (Appendix J). “Below” means $\gamma < \gamma^*$ for the stated inputs, not a population-level phase classification.

6.6 Semi-Synthetic Illustration of the Threshold Mechanism

The stream proxies of Section 6.5 cannot manipulate (d_0, α_s, γ) ; a generated control can vary them separately. Covariates are $Z \sim \text{Unif}[0, 1]^{d_0}$ with ϕ the identity, so the doubling dimension is d_0 . The signed drift is $g(z) = \text{sign}(W(z_1)) |W(z_1)|^{1/\gamma}$, a function of the first coordinate z_1 only, where W is a normalized Weierstrass field of exponent $\alpha_0 = \gamma\alpha_s \leq 1$. The construction targets a Hölder- α_s drift magnitude and margin exponent γ ; finite-grid self-tests report the corresponding effective slopes rather than certify the asymptotic exponents. Labels are Bernoulli with $\eta_j = \frac{1}{2} + 0.45g$ over four windows. The novelty representer is $w \equiv 1$ and known, which removes the representer/ratio channel entirely; the simulation therefore isolates the sign-recovery mechanism that creates the threshold, not the DML nuisance theory. The estimator is the specified two-stage procedure: per-window anchor split, Stage-1 κ -NN with the non-adaptive tuning $\kappa \asymp k^{2\alpha_s/(2\alpha_s+d_0)}$ (up to the logarithmic factor) of Lemma C.11', cross-fitted sign recovery, sign-corrected residual mean. Four arms vary one parameter each across $\gamma^* = d_0/(2\alpha_s)$. The design, prediction, and falsification criterion were fixed before the sweep.

arm	d_0	α_s	γ	γ^*	predicted	measured
A (above)	1	0.75	1	0.67	−0.500	−0.509 ± 0.024
B (below)	8	0.75	1	5.33	−0.158	−0.138 ± 0.002
C (below)	1	0.375	0.75	1.33	−0.375	−0.305 ± 0.008
D (above)	1	0.375	2	1.33	−0.500	−0.515 ± 0.023

Slopes are weighted log-log fits of RMSE against N over the largest five sample sizes (N up to 131,072; 100–400 seeds per point). The above-threshold arms are variance-dominated with slopes at the regular $-\frac{1}{2}$; the below-threshold arms are bias-dominated and sit 181 and 24 standard errors away from $-\frac{1}{2}$, with point values on the same side of $-\frac{1}{2}$ as, though not at, the candidate bias-rate prediction $-a(1+\gamma)$. The experiment illustrates the sign-recovery mechanism; it does not establish the conditional lower bound. Arms A/B differ only in d_0 , arms C/D only in γ : each parameter alone moves the scaling across the boundary. The generating code, its outputs, and the pre-registered design are part of the reproduction deposit (Section 3).

7 Beyond the Boundary: Diagnostics and Decision Rules

This section develops three procedures from quantities that remain computable under the boundary results.

7.1 The Local-Correction Screening Rule

The local-correction screening rule is a per-region, per-window computation, and it asks for no labels beyond those the corrector it gates already consumes. *Inputs requiring no labels of their own.* The frozen readout η_0 comes with the model. The drift-budget profile β is a prior, by Theorem 6.2, and not a measurement, unless it is upgraded to the history-transferred estimate $\hat{D}_{\text{cov}}$ of Section 6.4 under the remaining assumptions of Section 3.1 that Theorem 6.6 invokes ((A1) stationarity, (A4) shape, (A5') exogeneity, (A6) coverage, (A0) MCAR anchors, with the uncovered mass $\hat{\pi}_{\text{out}}$ reported alongside). The remaining inputs are the embedding ϕ and the local geometry: neighborhood density and freshness of the labeled buffer, and the deduplicated local intrinsic dimension of Section 5.3. *The noise floor.* For classification it is the Bernoulli quantity $\mathbb{E}[p(1-p)]$ over the frozen predictions, which needs no labels at all. For regression it is the short-range decoupled semivariance over the stream's residuals, read at the smallest available distances among pairs at least 1,000 steps apart and without extrapolation to zero distance (Section 5.2 estimates the extrapolated nugget instead, and reports it as a range). Our retrospective evaluation computes that floor over the whole stream; a deployment would compute it over the revealed prefix. From these, three quantities are computed per region.

1. *Decision-flip fraction* $\Pr[|\eta_0 - \frac{1}{2}| \leq \beta]$: the mass on which drift within budget could change the decision (the identified-set diameter in the decision discrepancy Δ_{01}). Where it is near zero, even the full assumed budget cannot move decisions, and freezing is safe in the decision currency whatever the wall's height in the mean currency.
2. *Price of freeze* $\rho_\beta(x) = 2 \max(0, \beta(x) - |\eta_0(x) - \frac{1}{2}|)$, the worst-case regret of trusting the frozen model at x . Its region average prices the freeze option, so adaptation is worth its risk only where that price is material.
3. *Deployable Wall A*, via the fiber-average formula of Theorem 6.5: $\text{WallA}(\beta; R) = \mathbb{E}_{z \sim \nu_R}[\mathbb{E}_{p_W}[w_\beta \mid \phi = z]]$, where $w_\beta = \min(1, \eta_0 + \beta) - \max(0, \eta_0 - \beta)$ is the post-clip interval width and the inner average runs over the full fiber population. For injective ϕ (generic at embedding dimension 192) fibers are singletons and this is the plain plug-in average of w_β over the region's unlabeled draws, accurate to $O_P(m^{-1/2})$ given the budget profile. For a coarse or resolution-limited ϕ the inner average is estimated by k -NN in ϕ over the *full* window (the operational form of the borrowing), at a rate graded partial in Appendix F of Online Appendix 1 (Corollary F.5). The mandatory order of operations is to average the post-clip

width across the fiber, never to average β first and clip after. The post-clip width is concave in β , so the latter overstates the wall.

These quantities define the *local-correction screening rule*. The rule adapts where the local estimate is well determined against the calibrated floor and the flip fraction is material, and freezes where the identified set is wide (the wall output is then a restatement of the prior, and acting on it would be acting on an assumption) or the buffer sparse or stale. When the budget is transferred rather than assumed, its semantics must be reported with it. The mean-response forecast is a calibrated forecast of typical drift that upper-bounds the mean realized width, not a worst-case certificate.

The two gates that decide every verdict in the evaluation below (the calibrated noise floor and the local-determinacy statistic) are empirical instruments, not consequences of the Part II theorems. What the theory contributes to the rule is interpretation: Theorem 6.2 makes the budget β a prior, the three quantities above inherit that status, and the theory thereby tells the operator which of the rule’s readings are measurements and which restate an assumption. On the evaluated streams the prior-dependent quantities corroborate verdicts the empirical gates have already forced (the zero flip fraction on HOMECREDIT-DEFAULT confirms a freeze the determinacy gate decided on its own); they decide none.

The rule on all eight streams. We call this a screening rule rather than a certificate because it has no finite-sample false-adapt guarantee. The evaluation below is retrospective; Section 8 collects the corresponding soundness limitations. We apply the rule to all eight streams, not only to the harm case, so that its false-freeze behavior is measured alongside its false-adapt behavior. Table 5 instantiates the two label-free gates against the realized outcomes of Table 1. The first gate is room above the calibrated floor ($1 - \text{floor}$); the second is a local estimate determined out-of-sample under the deployed strictly-past retrieval (past-only buffer, $k = 20$, unweighted neighbor mean). The rule is fixed before the outcomes are consulted, and the eight verdicts are unchanged over the whole box $\tau_{\text{room}} \in [0, 0.108)$, $\tau_{\text{sig}} \in [0, 0.0616)$ of the two thresholds. The operating point is $\tau_{\text{room}} = 0.10$ with $\tau_{\text{sig}} = 0$, that is, a strict positivity test on the local statistic. WEATHER is the stream nearest a boundary of that box: its room is 0.108, so a room threshold of 0.2 would freeze it.

In this retrospective evaluation, *the rule makes no false adapt on any stream*: every stream it clears improved under adaptation (+3.94% and +3.80% RMSE, +0.74 AUC points), and the one stream that is harmed is frozen. On HOMECREDIT-DEFAULT the freeze is decided by the local-determinacy gate alone ($R^2_{\text{causal}} = -0.021$): the room gate passes, because a Bernoulli floor of 0.85 still leaves 0.15 of nominal room. The decision-currency quantities corroborate the same verdict independently. The flip fraction is identically zero for every budget up to $\beta = 0.25$ (the predictions concentrate around a mean of 0.019 with no mass in $[0.25, 0.75]$), so within-budget drift cannot move a single decision, and the price of freeze is zero. The calibrated floor changes how the stream *reads* rather than what is decided here: a variogram floor of the kind the regression streams use reads 0.73 on this stream against the Bernoulli calibration’s 0.85, that is 0.27 of nominal room against 0.15, which is the difference between a stream that looks worth adapting and one that barely does. Even so, here the local-determinacy gate refuses either way.

One of the three clearances warrants a caution the other two do not. On HOMESITE-INSURANCE the clearing statistic is $R^2_{\text{causal}} = +0.070$ on the growing buffer, but the stream’s strict-split ceiling is 0.016 and falls to 0.006 under the temporal-gap stress test of Section 5.1, so the local squared-error signal that clears the gate may be partly residual temporal autocorrelation rather than durable spatial structure. The realized +0.74pp gain traveled through the ranking channel C4, which the gate does not observe: the verdict is right, but for reasons the rule cannot claim credit for. A stress-tested variant of the determinacy gate (scoring the local R^2 under a temporal gap) would

stream	task	floor	R^2_{causal}	verdict	realized	outcome
WEATHER	reg	0.89	+0.052	adapt	+3.94%	gain kept
SBERBANK-HOUSING	reg	0.66	+0.021	adapt	+3.80%	gain kept
HOMESITE-INSURANCE	cls	0.88	+0.070	adapt	+0.74pp	gain kept
HOME CREDIT-DEFAULT	cls	0.85	−0.021	freeze	−2.28pp	harm avoided
COOKING-TIME	reg	0.92	−0.009	freeze	+0.83%	forgone (C1)
DELIVERY-ETA	reg	0.97	−0.018	freeze	+0.66%	forgone (C1, 93% probe)
MAPS-ROUTING	reg	0.88	−0.006	freeze	+0.59%	forgone (C1/C3 mix)
ECOM-OFFERS	cls	0.87	−0.022	freeze	+0.86pp	forgone (C4 ranking)

Table 5: The local-correction screening rule on all eight streams, with the thresholds fixed before the outcomes were consulted, under the rule stated above: adapt only if $1 - \text{floor} \geq \tau_{\text{room}}$ and $R^2_{\text{causal}} > \tau_{\text{sig}}(1 - \text{floor})$, at the operating point $\tau_{\text{room}} = 0.10$, $\tau_{\text{sig}} = 0$; verdicts are constant over $\tau_{\text{room}} \in [0, 0.108)$, $\tau_{\text{sig}} \in [0, 0.0616)$. Floors are the calibrated estimators of Section 7.1, as fractions of residual variance; R^2_{causal} is the local-determinacy statistic of Appendix J, the strictly-past k -NN R^2 under the deployed retrieval; it is the out-of-sample R^2 of the streaming corrector itself (block 64, first 100 points reserved; Appendix J) and not the growing-buffer oracle ceiling of Section 5.1, which on WEATHER reads 0.067 against the +0.052 printed here. Gate inputs and the realized column are computed on the single-prior probe streams (Appendix J). The rule speaks only to the local channel C3.

demote this stream to a marginal call and is the natural tightening; we flag rather than adopt it, since changing the gate after seeing the outcomes would forfeit the fixed-in-advance status of Table 5.

The price of that safety is conservatism, and it has two distinct causes worth separating. Three regression streams are frozen although adaptation helped them (+0.83%, +0.66%, +0.59% RMSE forgone). Two of the three are attributed by Table 1 to the *global* debias channel C1 (on DELIVERY-ETA, 93% of the probe corrector’s gain is a running mean), and the third, MAPS-ROUTING, to a C1/partial-scale-C3 mixture; in all three the rule reports that the *local* field carries no out-of-sample signal, which for the first two is the full explanation and for the third is a partial one. A gate for the inexpensive global channels is a different and easier instrument, which we do not build here. The fourth false freeze is the one our own theory predicts: ECOM-OFFERS gains +0.86 AUC points through ranking recalibration (C4) while its squared-error local field is empty, and a loss-denominated gate cannot see a ranking gain. This is the currency limitation of Section 5.5 recurring inside the rule itself, now with a price attached. For ranking deployments the loss-denominated components are necessary but not sufficient, and a practitioner who freezes on them forgoes gains of this size.

Across the eight retrospective streams, each of the four forgone gains is of the same order as the single avoided harm. The rule’s case rests on loss asymmetry rather than on expected value; it is built for deployments where one undetected degradation costs more than several missed sub-point gains.

Formal statements: the decision-flip and price-of-freeze functionals are the canonical discrepancy instances of the program contract (Online Appendix 1, “the formal program”); the deployable estimand, its plug-in, and the clip/average non-commutation are Theorem F.3, Corollary F.5, and Remark F.5, Appendix F of Online Appendix 1; the budget-transfer option, its error budget, and its forecast-not-certificate semantics are Theorem C.15 and Remark C.16, Appendix C.

7.2 The Degradation-Attribution Procedure

Given a labeled audit sample (periodic or delayed labels suffice), the orthogonal split of Theorem 6.4 operationalizes as a three-way attribution of a window’s excess error: the *Wall-B share* $\hat{\mathbb{E}}[\text{Var}(\eta_W | \phi)]$ (at a stated fiber resolution), the *correctable share* (the time-respecting-oracle ceiling of the combined channels; the worked example below measures its local, C3 slice, folding the temporal and ranking channels into the remainder), and the *Wall-A remainder*. Two estimates are required; the third share follows by subtraction.

Step 1 (Wall-B share, needs the audit labels). Because an injective embedding has singleton fibers and hence raw $\text{WallB} \equiv 0$, the deployable object is the resolution- ρ deficit WallB_ρ : coarsen ϕ to ρ -balls, estimate the conditional mean of the audit labels at the finest resolution the sample supports, and report the average variance of that mean within the ρ -cells. The resolution ρ must be stated with the estimate (WallB_ρ is monotone in ρ and recovers WallB as $\rho \rightarrow 0$), so the share is meaningful only relative to the fiber scale at which the readout is actually operated; Appendix J gives the selection protocol that ties ρ to the deployed retrieval scale. A nonzero Wall-B share does not disable the history-transfer machinery of Section 6.4. That theory applies verbatim to the ϕ -reachable drift, with the within-fiber discrepancy priced by $\text{WallB}^{1/2}$ and reported separately (the (A3) disclosure of Appendix C, Online Appendix 1).

Step 2 (Wall-A envelope, label-free). The borrowed-certainty plug-in of Section 7.1, computed from (ϕ, η_0, β) alone, prices the Wall-A remainder *ex ante*. It is the diameter of the identified set under the assumed budget (an upper envelope on what unidentified drift could contribute, not an estimate of the realized share). The realized Wall-A share is then the remainder of the audit window’s excess error after subtracting the Wall-B share and the correctable share, the latter already measured as the time-respecting (leak-free) ceiling of the combined channels under the strict out-of-time protocol of Section 5.1.

Each share suggests a different intervention: representation changes for Wall B, engineering changes for the correctable share, and additional labels for the Wall-A remainder. Table 6 works the attribution on all eight streams, in the one currency common to all three shares (squared error as a fraction of residual variance). The measured Wall-B share is indistinguishable from zero on every stream: the fine-scale arm is strictly worse out of sample everywhere, so no conditional-mean structure below the deployed retrieval scale is detectable at these sample sizes, and representation work is not the binding constraint. The main cross-stream differences lie in the split between the local correctable share and the remainder: WEATHER and HOMESITE-INSURANCE are the only streams where the local channel has anything to recover (0.025 of 0.108 and 0.016 of 0.125), SBERBANK-HOUSING has the largest room (0.34) with none of it locally correctable (its realized gain traveled through freshness, C2, which a spatial split cannot see), and on HOMECREDIT-DEFAULT the entire 0.15 is remainder and naive correction harms. The remainder necessarily bundles the realized Wall-A share with the temporal channels (C1, C2) and ranking-currency gains (C4) that a strict spatial split cannot register. The measurements therefore do not support a uniform remedy across streams.

Formal statements: the Pythagorean split, the WallB_ρ object, and the orthogonality qualifications are the two-walls block of the program contract (Online Appendix 1, “the formal program”); the ϕ -reachable estimand and its Wall-B disclosure are assumption (A3) of Appendix C of Online Appendix 1; the deployable Wall-A envelope is Corollary F.5, Appendix F.

stream	room	correctable (local)	WallB $_{\rho}$	remainder
WEATHER	0.108	0.025	0.000	0.084
SBERBANK-HOUSING	0.337	0.000	0.000	0.337
COOKING-TIME	0.075	0.000	0.000	0.075
DELIVERY-ETA	0.032	0.000	0.000	0.032
MAPS-ROUTING	0.117	0.000	0.000	0.117
ECOM-OFFERS	0.134	0.000	0.000	0.134
HOMESITE-INSURANCE	0.125	0.016	0.000	0.108
HOME CREDIT-DEFAULT	0.146	0.000	0.000	0.146

Table 6: Worked three-way attribution on the eight probe streams, as fractions of residual variance on the evaluation half. Room is 1 minus the calibrated floor (Table 5); the correctable column is the strict out-of-time $k=20$ ceiling of Table 1 (its local, C3 slice); WallB $_{\rho}$ is the additional variance a finest-supported ($m=5$) readout explains over the deployed scale, out of sample, at the protocol resolution ρ (median 20th-neighbor distance, 1.9–4.0 across streams; Appendix J); the remainder bundles the realized Wall-A share with the temporal and ranking channels a spatial split cannot see. Room is measured on the full rule-evaluation window and treated as a common fraction under approximate stationarity of the residual variance across halves. Rows may differ in the last digit from column arithmetic; shares are rounded independently; the computation is part of the reproduction deposit (Appendix J).

7.3 The Price of Labels

When the Wall-A share dominates, additional labels are required. The boundary theory gives two rates.

Region average, parametric. k anchors drawn i.i.d. from the region pin the region-mean wall to $\min(2\bar{\beta}_R, \tilde{O}(k^{-1/2}))$. Averaging a bounded quantity over a fixed region is a parametric problem, so no geometry enters the rate. The pure $k^{-1/2}$ collapse is operative only once $k \gtrsim \bar{\beta}_R^{-2}$. Below that threshold the assumed budget still binds, so the pin sits at the prior level $2\bar{\beta}_R$, not at the anchor rate.

Pointwise, nonparametric. Recovering the wall’s height point by point requires anchors *plus* Lipschitz-in- ϕ smoothness, at the local nonparametric rate $\tilde{O}(k^{-1/(2+d_0)})$. Each point can only be estimated from anchors nearby in the embedding, and the intrinsic dimension d_0 of the region’s support in ϕ -space governs how many anchors are ever local. This is the same deduplicated d_0 that governs channel C3 and the margin threshold. Higher intrinsic dimension both weakens unlabeled correction and raises the label requirement.

Up to suppressed constants and logarithmic factors, the rates imply the following costs at two points of the proxy map (Table 4): past the $\bar{\beta}_R^{-2}$ threshold, halving region-averaged uncertainty costs $4\times$ the labels, whatever the geometry. At $d_0 = 3$ (the floor among the measured streams, where SBERBANK-HOUSING sits), halving *pointwise* uncertainty costs $2^{2+3} = 32\times$. At $d_0 = 13$ (the HOME CREDIT-DEFAULT / LENDING CLUB neighborhood, $d_0 = 13.8$ and 13.6 ; the map’s top, MAPS-ROUTING, sits at 17.6), it costs $2^{2+13} \approx 3 \times 10^4\times$. Equivalently, $N = 10^4$ labels yield region-averaged resolution of order $N^{-1/2} = 10^{-2}$ on either stream, but pointwise resolution of order $N^{-1/5} \approx 0.16$ at $d_0 = 3$ and $N^{-1/15} \approx 0.54$ at $d_0 = 13$. On the high-dimensional streams a pointwise drift map is out of reach at any realistic labeling budget, while the regional audit is unaffected. Anchors yield the most where the pointwise diameter is large and the region straddles the decision boundary. Under the measured $d_0 \approx 3\text{--}18$, pointwise certification is costly, whereas region-averaged certification is comparatively inexpensive. The labeling budget should therefore

target regional wall audits and decision-flip monitoring rather than pointwise drift maps.

Formal statements: both rates are Branch 2(a) of the decisive lemma in the program contract (Online Appendix 1, “the formal program”), with the generic intrinsic dimension there instantiated by the deduplicated d_0 of Table 4 (Appendix J); the multi-window refinement of the pointwise line (the anchor rate $k^{-\alpha_s/(2\alpha_s+d_0)}$ and its margin-graded version) is given by the margin-boundary results of Section 6.4, Appendix H, and Appendices C, G, and I of Online Appendix 1.

7.4 Deployment Recommendations

Together, the three procedures suggest the following deployment practice. Deploy the combined channels where an out-of-time validation shows a positive gain. Screen the local channel with the rule of Section 7.1: on the eight retrospective streams it made no false adapt, at the measured price of freezing four streams whose gains traveled wholly or partly through channels it does not see; gates for the temporal and ranking channels are open instruments, so in ranking deployments the deployment currency itself must be monitored (Section 5.5). Attribute residual degradation with the three-way split, and allocate labels according to the price of labels when the attribution puts the mass on the identifiability wall.

8 Conclusion

We studied two opposing claims about adaptation under drift: that it recovers drift, and that unlabeled adaptation cannot help. The empirical results support neither claim as a general rule. Across eight temporal streams, the observed improvements and one harm case can be attributed to a small set of mechanisms under a leakage-corrected protocol. Under the stated agnostic drift class, the theory then separates uncertainty caused by non-identifiability from error caused by the frozen representation. The margin-indexed analysis proves the regular branch and states the remaining condition in the lower-bound program explicitly. TabReD proxies and a sensitivity analysis on fifteen additional benchmarks place the examined settings on the difficult side under their stated assumptions, but do not establish a universal empirical phase law. These distinctions determine whether a deployment should improve the correction method, change the representation, or acquire labels.

Several limitations bound these claims. The equality case $\gamma = \gamma^*$ is not covered by the current log-rate argument. Above the threshold, the efficient constant under overlapping windows is graded partial, and the rate statement itself holds only under the stated nuisance-rate hypotheses, (H6) included, whose derivation from primitives is graded partial. Below the threshold, the lower bound at $\gamma = 0$ is conditional on one graded affinity estimate (gap (2c)), for fixed and growing K alike; the margin adaptation for $0 < \gamma < \gamma^*$ remains open (Remark H.8); Online Appendix 1 records the status of every statement. On the empirical side, the proxy map estimates drift smoothness and margin from single streams, so the point estimates are noisy even where the verdict, resting only on an inequality under the rough-field bound, is insensitive to that noise. All measurements use a single backbone family, so the map’s d_0 floor should be re-established per representation. Schema and semantic drift, where the covariate space itself changes, lie outside our formalism altogether, and are in our view the deepest open formal gap in the area.

The screening rule carries qualifications of its own. Its soundness can be checked only with labels the deployment lacks, so it must be validated on semi-synthetic controls with matched covariate laws; Section 6.6 gives the generator we would start from, and a community-scale study is future work. One-sidedness on all eight streams in retrospective evaluation (Table 5) is an observation on

one deployment, not a bound: soundness on unseen streams is not established. Finally, the loss-denominated gates forgo ranking gains of measured size, so a sound ranking-currency gate with the same one-sided profile remains open, as does a ranking-currency analogue of the wall theory itself.

Author Contributions

Contributor roles follow the CRediT taxonomy. **Tianyu Wang:** Conceptualization; Methodology; Formal analysis; Software; Data curation; Investigation; Visualization; Writing – original draft. **Xi Vincent Wang:** Conceptualization; Funding acquisition; Supervision; Writing – review & editing. **Lihui Wang:** Funding acquisition; Supervision; Writing – review & editing. **Mian Li:** Conceptualization; Methodology; Funding acquisition; Writing – review & editing. **Zhihao Liu:** Software; Writing – review & editing.

Acknowledgments and Disclosure of Funding

The work presented here was supported by the Swedish research centre of eXcellence in PRoduction RESearch (XPRES) and the National Natural Science Foundation of China (Grant No. 52275263). The authors declare no competing interests.

References

- Jean-Yves Audibert and Alexandre B. Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608–633, 2007.
- Shai Ben-David and Ruth Urner. On the hardness of domain adaptation and the utility of unlabeled target samples. In *Proceedings of the 23rd International Conference on Algorithmic Learning Theory*, volume 7568 of *Lecture Notes in Computer Science*, pages 139–153. Springer, 2012. doi: 10.1007/978-3-642-34106-9_14.
- Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1–2):151–175, 2010a.
- Shai Ben-David, Tyler Lu, Teresa Luu, and Dávid Pál. Impossibility theorems for domain adaptation. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 129–136, 2010b.
- Peter J. Bickel, Chris A. J. Klaassen, Ya’acov Ritov, and Jon A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press, Baltimore, 1993. Reprinted by Springer, New York, 1998.
- Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. Introduction to statistical learning theory. In *Advanced Lectures on Machine Learning*, volume 3176 of *Lecture Notes in Computer Science*, pages 169–207. Springer, Berlin, 2004.
- Lawrence D. Brown and Mark G. Low. Asymptotic equivalence of nonparametric regression and white noise. *The Annals of Statistics*, 24(6):2384–2398, 1996.

- Haorun Cai and Han-Jia Ye. Understanding the limits of deep tabular methods with temporal shift. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 6366–6386, 2025. arXiv:2502.20260.
- T. Tony Cai and Mark G. Low. Testing composite hypotheses, Hermite polynomials and optimal estimation of a nonsmooth functional. *The Annals of Statistics*, 39(2):1012–1041, 2011.
- Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, Cambridge, 2006. doi: 10.1017/CBO9780511546921.
- Kamalika Chaudhuri and Sanjoy Dasgupta. Rates of convergence for the cluster tree. In *Advances in Neural Information Processing Systems*, volume 23, 2010.
- Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.
- A. Philip Dawid. Present position and potential developments: Some personal views: Statistical theory: The prequential approach. *Journal of the Royal Statistical Society. Series A (General)*, 147(2):278–292, 1984. doi: 10.2307/2981683.
- Elena Facco, Maria d’Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*, 7:12140, 2017.
- João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4):44:1–44:37, 2014.
- Saurabh Garg, Yifan Wu, Sivaraman Balakrishnan, and Zachary C. Lipton. A unified view of label shift estimation. In *Advances in Neural Information Processing Systems*, volume 33, 2020.
- Isaac Gibbs and Emmanuel J. Candès. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems*, volume 34, pages 1660–1672, 2021.
- Jaroslav Hájek. A characterization of limiting distributions of regular estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 14(4):323–330, 1970.
- Kai Helli, David Schnurr, Noah Hollmann, Samuel Müller, and Frank Hutter. Drift-resilient TabPFN: In-context learning temporal distribution shifts on tabular data. In *Advances in Neural Information Processing Systems*, volume 37, 2024. arXiv:2411.10634.
- Noah Hollmann, Samuel Müller, Katharina Eggersperger, and Frank Hutter. TabPFN: A transformer that solves small tabular classification problems in a second. In *International Conference on Learning Representations*, 2023.
- Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeyer, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025.
- Yuri I. Ingster and Irina A. Suslina. *Nonparametric Goodness-of-Fit Testing Under Gaussian Models*, volume 169 of *Lecture Notes in Statistics*. Springer, New York, 2003.
- Heinrich Jiang. Non-asymptotic uniform rates of consistency for k-NN regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3999–4006, 2019.

- Fredrik D. Johansson, David Sontag, and Rajesh Ranganath. Support and invertibility in domain-invariant representations. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 527–536, 2019.
- Edward H. Kennedy. Semiparametric doubly robust targeted double machine learning: A review. In *Handbook of Statistical Methods for Precision Medicine*, pages 207–236. Chapman and Hall/CRC, Boca Raton, FL, 2024. doi: 10.1201/9781003216223-10.
- Changhun Kim, Taewon Kim, Seungyeon Woo, June Yong Yang, and Eunho Yang. AdapTable: Test-time adaptation for tabular data via shift-aware uncertainty calibrator and label distribution handler. In *Table Representation Learning Workshop at NeurIPS*, 2024.
- Lingjing Kong, Shaoan Xie, Weiran Yao, Yujia Zheng, Guangyi Chen, Petar Stojanov, Victor Akinwande, and Kun Zhang. Partial disentanglement for domain adaptation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 11455–11472, 2022.
- Lucien Le Cam. *Asymptotic Methods in Statistical Decision Theory*. Springer Series in Statistics. Springer, New York, 1986.
- Oleg Lepski, Arkadi Nemirovski, and Vladimir Spokoiny. On estimation of the L_r norm of a regression function. *Probability Theory and Related Fields*, 113(2):221–253, 1999.
- Elizaveta Levina and Peter J. Bickel. Maximum likelihood estimation of intrinsic dimension. In *Advances in Neural Information Processing Systems*, volume 17, 2004.
- Zachary C. Lipton, Yu-Xiang Wang, and Alexander J. Smola. Detecting and correcting for label shift with black box predictors. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3122–3130, 2018.
- Marcos López de Prado. *Advances in Financial Machine Learning*. Wiley, Hoboken, NJ, 2018.
- Charles F. Manski. *Partial Identification of Probability Distributions*. Springer Series in Statistics. Springer, New York, 2003.
- Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009.
- Brad McNeney and Jon A. Wellner. Application of convolution theorems in semiparametric models with non-i.i.d. data. *Journal of Statistical Planning and Inference*, 91(2):441–480, 2000.
- Francesca Molinari. Microeconometrics with partial identification. In *Handbook of Econometrics, Volume 7A*, Handbooks in Economics, pages 355–486. Elsevier, Amsterdam, 2020.
- Jose G. Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V. Chawla, and Francisco Herrera. A unifying view on dataset shift in classification. *Pattern Recognition*, 45(1):521–530, 2012.
- Michael Nussbaum. Asymptotic equivalence of density estimation and Gaussian white noise. *The Annals of Statistics*, 24(6):2399–2430, 1996.

- Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICL: A tabular foundation model for in-context learning on large data. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 50817–50847, 2025. arXiv:2502.05564.
- Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence, editors. *Dataset Shift in Machine Learning*. MIT Press, Cambridge, MA, 2009.
- Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023. doi: 10.1214/23-STS894.
- Weijieying Ren, Xiaoting Li, Huiyuan Chen, Vineeth Rakesh, Zhuoyi Wang, Mahashweta Das, and Vasant G. Honavar. TabLog: Test-time adaptation for tabular data using logic rules. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 42417–42427, 2024.
- Ivan Rubachev, Nikolay Kartashev, Yury Gorishniy, and Artem Babenko. TabReD: Analyzing pitfalls and filling the gaps in tabular deep learning benchmarks. In *International Conference on Learning Representations*, 2025. arXiv:2406.19380.
- Mona Schirmer, Metod Jazbec, Christian A. Naesseth, and Eric Nalisnick. Monitoring risks in test-time adaptation. In *Advances in Neural Information Processing Systems*, volume 38, 2025. arXiv:2507.08721.
- Glenn Shafer. Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 184(2):407–431, 2021. doi: 10.1111/rssa.12647.
- Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.
- Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8:985–1005, 2007.
- Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei A. Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9229–9248, 2020.
- Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems*, volume 32, pages 2526–2536, 2019.
- Alexandre B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.
- Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, New York, 2009.
- Aad W. van der Vaart. *Asymptotic Statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998.

- Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer, New York, 1996.
- Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021.
- Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7201–7211, 2022.
- Tianyu Wang, Mian Li, Xi Vincent Wang, and Lihui Wang. SNR-adaptive local posterior correction for frozen tabular predictors. Companion manuscript by four of the present authors; preprint forthcoming, 2026.
- Yibo Zeng, Jiashuo Liu, Henry Lam, and Hongseok Namkoong. LLM embeddings improve test-time adaptation to tabular $Y \mid X$ -shifts. In *Table Representation Learning Workshop at NeurIPS*, 2024. arXiv:2410.07395.
- Han Zhao, Rémi Tachet des Combes, Kun Zhang, and Geoffrey J. Gordon. On learning invariant representations for domain adaptation. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7523–7532, 2019.
- Hao Zhao, Yuejiang Liu, Alexandre Alahi, and Tao Lin. On pitfalls of test-time adaptation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 42058–42080, 2023.

Reader’s guide to the appendices

Where each appendix appears. This document keeps the appendices a reader needs to check the headline results: **B** (the label-free irreducibility core behind Theorem 6.2), **H** (the fixed- K margin boundary behind Theorem 6.6), the auxiliary Lemmas C.11’ and G.14’ that Theorem 6.6 quotes, and **J** (the complete measurement protocol). Appendices **A**, **C**, **D–F**, **G**, **I**, **K** and the program contract appear only in Online Appendix 1, the supplementary document that accompanies this arXiv submission as the ancillary file `anc/supplementary.pdf`. Its first half reproduces all of Appendices A–K (B, H, and J included), the program contract, and the two auxiliary lemmas, with their original letters and statement numbers unchanged, so a statement cited by a letter absent from this document (for example Theorem G.1.4 or Corollary F.5) appears there under the same number. The lettering in this document therefore runs B, H, J by design.

The appendices can be read at three depths. **(i) Contract level:** read the statement-status conventions (Online Appendix 1, Appendices A–K), the status note closing Appendix H, and the per-section status tables of Online Appendix 1, §S1; this suffices to verify the grade (*proven*, *partial*, *conditional*, *cited-standard*, or *conjecture*) of every claim cited from the main text. **(ii) Pillar level:** additionally read the six pillar proofs, namely Thm. A.14 (exact reduction identity), Thm. B.8 (label-free irreducibility), Thm. F.2–Cor. F.5 (deployable Wall A reduction), Thm. G.1.4 ($\sqrt{N}$ CLT), Thm. H.9 (margin-indexed regular branch; graded partial, see the status note of Appendix H), and Thm. I.3.2 (the strict branches above the threshold and the growing- K remainder). **(iii) Full:** everything, including Appendix K of Online Appendix 1 (exact admissibility and the permutation-mixture affinity bound) and the auxiliary-lemmas section here (Lemmas C.11’ and G.14’); at this depth most of the reading is in Online Appendix 1: its first half carries the verified statements of all of A–K (proof outlines and consumed outputs kept, rendered numbering unchanged), and the full developments live in §S7 (the selection and measurability machinery Lemmas A.1–A.13 of Appendix A, with the auxiliary OP1 remark C.4 of Appendix C), §S5 (the A5’ orthogonalization program of Appendix D), §S6 (the C.20 program of Appendix E), and §S3 (the complete staged proof of Lemma B.5’, Appendix K’s permutation-mixture affinity bound, numbered Lemma K.8 there and unrelated to Lemma B.5 of Appendix B, together with the proofs of K.1, K.3, and K.4). Proofs that were promoted into the main text (Section 6) are replaced in Appendices B, F, and the contract by one-line pointers; the statements and status tags are unchanged.

Online Appendix 1 also records, per appendix, the status of every statement, the ranked open points, and a status summary (§S1); the comparison with the published theorems behind Lemmas C.11’ and G.14’ (§S2); the complete staged proof of Lemma B.5’ with its machine-checked cases (§S3); a one-page index of the verification programs (§S4); and the full developments of the abridged Appendices D and E (§S5, §S6). The verification code and its outputs are part of the reproduction deposit (Section 3); they are not part of this arXiv source package.

B Ancillarity, the Le Cam two-point wall, and its (B)-contingent scope

B.0 Standing assumptions

(S1) $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$ Polish with Borel σ -algebra; $\mathcal{Y}$ finite with $2^{\mathcal{Y}}$; primary case $\{0, 1\}$. **(S2)** $p_0(\cdot \mid x) = h_0(\phi(x))$ a Markov kernel (ϕ, h_0 Borel); $\eta_0(x) = p_0(1 \mid x)$. **(S3)** $\beta : \mathcal{X} \rightarrow [0, 1]$ Borel (used by the extremal pair of Lemma B.6). **(S4)** (O1) i.i.d. $x_1, \dots, x_n \sim p_W$; observation map $\Pi_A : (x, y) \mapsto x$; y never observed; the observation law of a world P is $Q_P^{(n)} := (\Pi_A^{(n)})_{\#} P^{\otimes n}$. **(S5)**

(B) $I(\beta)$ as in Appendix A of Online Appendix 1 (and Section 6.1 of the main paper). **(S6)** $R \in \mathcal{B}_{\mathcal{X}}$, $p_W(R) > 0$; $\mu := p_W(\cdot \mid R)$. **(S7)** An estimator is a Borel map $\hat{\theta}_n : \mathcal{X}^n \rightarrow \mathbb{R}$ (for the infinite-sequence claims: on $\mathcal{X}^{\mathbb{N}}$); a randomized estimator is Borel on $\mathcal{X}^n \times [0, 1]$ (resp. $\mathcal{X}^{\mathbb{N}} \times [0, 1]$) with independent $U \sim \text{Unif}[0, 1]$; general randomized procedures given by Markov kernels are covered by the kernel-rule case of Lemma B.5. Binary TV identity: $D_{\text{TV}}(a, b) = |\eta_a - \eta_b|$; $B_{\beta}(x) = [\underline{\eta}, \bar{\eta}]$ with $\underline{\eta} = \max(0, \eta_0 - \beta)$, $\bar{\eta} = \min(1, \eta_0 + \beta)$, with $w_{\beta} = \bar{\eta} - \underline{\eta} \in [\beta, 2\beta]$.

B.1 Construction and ancillarity

Lemma B.1 (world construction). [proven] *For $p_W \in \mathcal{P}(\mathcal{X})$ and any Markov kernel q , $P = p_W \otimes q$ defined by $P(E) = \sum_y \int_{E_y} q(y \mid x) p_W(dx)$ is a probability measure on $(\mathcal{X} \times \mathcal{Y}, \mathcal{B}_{\mathcal{X}} \otimes 2^{\mathcal{Y}})$ with X -marginal p_W and p_W -a.e.-unique disintegration q .*

Proof. Since $\mathcal{Y}$ is finite, every product-measurable E decomposes as $\bigcup_y E_y \times \{y\}$ with Borel sections; countable additivity holds per y by monotone convergence, and the total mass is 1. For uniqueness, note that for each y both versions are Radon–Nikodym derivatives of $A \mapsto P(A \times \{y\}) \leq p_W(A)$ with respect to p_W . $\square$

Lemma B.2 (product pushforward). [proven] $(\Pi_A^{(n)})_{\#} P^{\otimes n} = ((\Pi_A)_{\#} P)^{\otimes n}$.

Proof. The two sides agree on measurable rectangles, which form a generating π -system, and both are probability measures, so they coincide by Dynkin’s π – λ theorem. $\square$

Proposition B.3 (ancillarity, $n \leq \infty$). [proven] *For every Markov kernel q and every n : $Q_{p_W \otimes q}^{(n)} = p_W^{\otimes n}$; hence $D_{\text{TV}}(Q_P^{(n)}, Q_{P'}^{(n)}) = 0$ for all $q, q' \in I(\beta)$. The same holds for the infinite-sequence law on $(\mathcal{X}^{\mathbb{N}}, \mathcal{B}_{\mathcal{X}}^{\otimes \mathbb{N}})$.*

Proof. The identity $(\Pi_A)_{\#} P = p_W$ uses only $q(\mathcal{Y} \mid x) = 1$, and Lemma B.2 gives the finite case. For $n = \infty$, the i.i.d. infinite product $P^{\otimes \mathbb{N}}$ exists on an arbitrary probability space by the infinite-product-measure theorem (equivalently Ionescu–Tulcea with constant kernels; Kolmogorov extension is also available under (S1)); the coordinatewise map $\Pi_A^{(\mathbb{N})}$ is product-measurable; cylinder sets form a generating π -system on which the two observation laws agree by the finite case, and Dynkin’s theorem concludes. $\square$

Remark B.4 (terminology). With p_W fixed and $q \in I(\beta)$ the parameter, the full data $x_{1:n}$ is **ancillary** for q (exact, uniform in n): marginalizing the never-observed y erases q from the model before any data are drawn.

B.2 Le Cam two-point and the irreducibility theorem

The functional. $\theta(P) := \mathbb{E}_{\mu}[\eta_P]$, $\eta_P := dP(\cdot \times \{1\})/dp_W$ (a Radon–Nikodym derivative, which exists because $P(A \times \{1\}) \leq p_W(A)$ with both measures finite, and satisfies $0 \leq \eta_P \leq 1$ a.e.). By Lemma B.1’s uniqueness and $\mu \ll p_W$, $\theta(p_W \otimes q) = \mathbb{E}_{\mu}[\eta_q]$ is a well-defined functional of the world.

Lemma B.5 (two-point inequality; randomized and kernel rules). [proven] *For probability measures Q_0, Q_1 on $(\Omega, \mathcal{F})$ and $\theta_0, \theta_1 \in \mathbb{R}$: every randomized estimator satisfies $\max_i \mathbb{E}|\hat{\theta} - \theta_i| \geq \frac{1}{2}|\theta_0 - \theta_1|(1 - D_{\text{TV}}(Q_0, Q_1))$.*

Proof. Let $\nu = Q_0 + Q_1$, $g_i = dQ_i/d\nu$, and $d\mu_\wedge = \min(g_0, g_1)d\nu$, so that $\mu_\wedge(\Omega) = 1 - D_{\text{TV}}(Q_0, Q_1)$. Pointwise, $|\hat{\theta} - \theta_0| + |\hat{\theta} - \theta_1| \geq |\theta_0 - \theta_1|$, and $\mu_\wedge \leq Q_i$; integrating against $\mu_\wedge$, summing, and halving gives the bound. For randomization, replace Q_i by $Q_i \otimes \lambda$, which has the same TV distance. **Kernel rules:** for a Markov kernel rule $K(\omega, dt)$, $\int (|t - \theta_0| + |t - \theta_1|)K(\omega, dt) \geq |\theta_0 - \theta_1|$ pointwise; integrate against $\mu_\wedge$ to obtain the same bound. So “no label-free procedure improves” covers all randomized procedures, not only $\text{Unif}[0, 1]$ -randomized maps. $\square$

Lemma B.6 (extremal pair). [proven] $\eta_\pm = \Pi_{[0,1]}(\eta_0 \pm \beta)$ define Borel Markov kernels $q_\pm$ with $q_\pm(\cdot | x) \in B_\beta(x)$ for *every* x (binary TV identity: $|\eta_+ - \eta_0| = \min(1 - \eta_0, \beta) \leq \beta$, symmetrically), so $q_\pm \in I(\beta)$; and $\theta(P_+) - \theta(P_-) = \mathbb{E}_\mu[w_\beta]$. $\square$

Lemma B.7 (binary mean diameter: direct). [proven] $D_{\text{mean}}(\beta; R) = \mathbb{E}_\mu[w_\beta]$, attained by (q_-, q_+) .

Proof. For the upper bound, $\eta_q, \eta_{q'} \in [\underline{\eta}, \bar{\eta}]$ p_W -a.e., hence μ -a.e., and integrating gives the bound; attainment is Lemma B.6. $\square$

(This instantiates Theorem A.14 with $\delta_\beta^{\text{mean}} = w_\beta$; the argument does not use Theorem A.14.)

Theorem B.8 (label-free irreducibility within (B)). [proven] Assume (S1)–(S7), binary $\mathcal{Y}$. For every n (including $n = \infty$), over $\mathcal{W} = \{p_W \otimes q : q \in I(\beta)\}$:

$$\inf_{\hat{\theta}_n \text{ randomized}} \sup_{P \in \mathcal{W}} E|\hat{\theta}_n - \theta(P)| = \frac{1}{2} D_{\text{mean}}(\beta; R) = \frac{1}{2} E_{p_W(\cdot|R)}[w_\beta],$$

constant in n , attained at the two-point subfamily $\{P_-, P_+\}$ (lower bound) and by the zero-data midpoint estimator $\hat{\theta}^* = \mathbb{E}_\mu[(\underline{\eta} + \bar{\eta})/2]$ (upper bound). If $p_W(R \cap \{\beta > 0\}) > 0$ the bound is strictly positive ($w_\beta \geq \beta$, valid also at the both-clips boundary via $\beta \leq 1$). The wall height is a functional of the prior (β, p_0, p_W) alone.

Proof. The proof has been promoted to the main text: the proof of Theorem 6.2 (Section 6.2) transcribes it in full, and the inputs it consumes (Lemmas B.1–B.7 above) are unchanged. $\square$

Corollary B.9 (power = size). [proven] Any test $\psi_n : \mathcal{X}^n \rightarrow [0, 1]$ of q vs q' satisfies $E_{P'}[\psi_n] = \mathbb{E}_P[\psi_n] = E_{p_W^{\otimes n}}[\psi_n]$. $\square$

Corollary B.10 (confidence sets cannot shrink; outer-expectation form). [proven] Let $C_n : \mathcal{X}^n \rightarrow 2^{I(\beta)}$ satisfy the measurability convention ($\{x_{1:n} : q' \in C_n\}$ measurable for each fixed q') and uniform coverage $P_q^{(n)}(q \in C_n) \geq 1 - \alpha$ for all $q \in I(\beta)$. Then, with $\text{diam}_\theta C = \sup_{q, q' \in C} |\theta(P_q) - \theta(P_{q'})|$, the pointwise minorant bound

$$\text{diam}_\theta C_n \geq D_{\text{mean}}(\beta; R) \cdot \mathbf{1}_{\{q_+ \in C_n\} \cap \{q_- \in C_n\}}$$

holds, the minorant is measurable by the convention, and taking (inner, hence also outer) expectations under $\mu_n := p_W^{\otimes n}$ with the union bound $\mu_n(q_+ \in C_n, q_- \in C_n) \geq 1 - 2\alpha$ (each marginal event has Q -probability $\geq 1 - \alpha$ by Prop. B.3, since all observation laws coincide) gives $E_*[\text{diam}_\theta C_n] \geq (1 - 2\alpha)D_{\text{mean}}(\beta; R)$ for every n . As a supremum over an uncountable data-dependent family, $\text{diam}_\theta C_n$ itself need not be measurable, which is why the statement takes inner/outer-expectation form. $\square$

Remark B.11 (uniformity; oracle p_W ; other functionals). The constant is the same for every n , including $n = \infty$, and is unaffected by an oracle revealing p_W itself: within (B) the observation functional $P \mapsto Q_P^{(n)}$ factors through p_W alone. For any real functional θ on $\mathcal{W}$, the same proof gives minimax risk $\geq \frac{1}{2} \sup_{q,q'} |\theta(P_q) - \theta(P_{q'})|$; the extremal pairs for Δ_{01} and ρ_β are deferred (Online Appendix 1, §S1, open point 6).

B.3 Scope delimitation: the scope theorems

Definition B.12 (class-restricted identified set). [proven-well-posed as scoped] For a set $\mathcal{C}$ of joint laws, $I_{\mathcal{C}}(p_W) := \{q : q \text{ is a disintegration of some } P \in \mathcal{C} \text{ with } (\Pi_A)_\# P = p_W\}$ (modulo p_W -a.e. equality), and $D_{\Delta}^{\mathcal{C}}(R) := \sup_{q,q' \in I_{\mathcal{C}}(p_W)} \mathbb{E}_\mu[\Delta(f_q, f_{q'})]$ with $\sup \emptyset := 0$. *Scope note:* for non-singleton classes and general Δ this presupposes the measurability of the integrand, so carry the kernel hypotheses (S5) of Appendix A of Online Appendix 1 (not the mechanism prior (S5) above); in this section only singleton/empty cases are used, where the integrand is $\equiv 0$ under (D0), which we assume of any discrepancy here.

Proposition B.13 (covariate-shift collapse: prior-driven). [proven] $\mathcal{C}_{\text{cov}} = \{p \otimes p_0\}$: for every observed p_W , $I_{\mathcal{C}_{\text{cov}}}(p_W) = \{p_0(\cdot | \cdot)\}$ and $D_{\Delta}^{\mathcal{C}_{\text{cov}}}(R) = 0$ at $n = 0$, for every (D0)-discrepancy, every R , every β .

Proof. The proof has been promoted to the main text (proof of Theorem 6.3(i), Section 6.2); it rests on the uniqueness clause of Lemma B.1. $\square$

Lemma B.14 (moment identification of π_W). [proven] Fix known $\mu_y \in \mathcal{P}(\mathcal{X})$, $y \in \mathcal{Y} = \{1, \dots, |\mathcal{Y}|\}$. The following are equivalent: (1) $\pi \mapsto \sum_y \pi(y) \mu_y$ is injective on $\Delta(\mathcal{Y})$; (2) $\{\mu_y\}$ are linearly independent in $M(\mathcal{X})$; (3) there is a bounded Borel $T : \mathcal{X} \rightarrow \mathbb{R}^{|\mathcal{Y}|}$ with nonsingular moment matrix $M_{jy} = \mathbb{E}_{\mu_y}[T_j]$ and $\mathbb{E}_{p_W}[T] = M\pi$ for every $p_W = \sum_y \pi(y) \mu_y$. In particular $\pi_W = M^{-1} \mathbb{E}_{p_W}[T]$.

Proof. The proof has been promoted to the main text and is reproduced inside the proof of Theorem 6.3(ii) (Section 6.2). That proof assembles the total-mass observation (affine $\iff$ linear independence for probability measures), the two-priors converse, the Hahn–Jordan/annihilator selection of the indicator moments $T_j = \mathbf{1}_{A_j}$ with nonsingular M , and moment linearity; (3) $\implies$ (1) follows from nonsingularity. $\square$

Lemma B.15 (Bayes pinning; nonnegative RN versions). [proven] $\lambda = \sum_y \mu_y$, $g_y = d\mu_y/d\lambda$ **chosen nonnegative everywhere** (replace any version by $\max(g_y, 0)$, another version, and one needed so that $q_\pi(\cdot | x) \in \Delta(\mathcal{Y})$ at every x , as Lemma B.1’s kernel framework requires); $m_\pi = \sum_y \pi(y) g_y$; $q_\pi(y | x) = \pi(y) g_y(x) / m_\pi(x)$ on $\{m_\pi > 0\}$, $:= \pi$ elsewhere. Then q_π is a Markov kernel and the p_π -a.e.-unique disintegration of $P_\pi(A \times \{y\}) = \pi(y) \mu_y(A)$.

Proof. One has $p_\pi(\{m_\pi = 0\}) = 0$; that q_π is a Markov kernel and a disintegration of P_π is a direct verification, and uniqueness follows from Lemma B.1. $\square$

Theorem B.16 (label-shift collapse: data-pinned). [proven] Under Lemma B.14’s condition (2) and $p_W \in \mathcal{M}_{\text{ls}}$, the set of X -marginals of the label-shift family $\mathcal{C}_{\text{ls}} := \{P_\pi : \pi \in \Delta(\mathcal{Y})\}$ of Lemma B.15 (else the class is falsified and $D^{\mathcal{C}} := 0$ vacuously): π_W is unique, recovered by $\pi_W = M^{-1}\mathbb{E}_{p_W}[T]$, $I_{\mathcal{C}_{\text{ls}}}(p_W) = \{q_{\pi_W}\}$, and $D_{\Delta}^{\mathcal{C}_{\text{ls}}}(R) = 0$ for all R and $(D0)$ - Δ . Binary: $q_W(1 | x) = \pi_W(1)g_1(x)/(\pi_W(1)g_1(x) + \pi_W(0)g_0(x))$. $\square$

Proposition B.17 ($\sqrt{n}$ -estimability of π_W). [proven] $\hat{\pi} = M^{-1}\hat{b}$, $\hat{b}_j = n^{-1} \sum_i T_j(x_i)$: $\mathbb{E}\|\hat{\pi} - \pi_W\|_2 \leq \|M^{-1}\|_{\text{op}} \sqrt{|\mathcal{Y}|/(4n)}$ (the variance is $\leq 1/(4n)$ per coordinate since each T_j is $[0, 1]$ -valued; Jensen’s inequality and the operator-norm bound then give the result). $\square$

Rate transfer to the θ -functional requires a lower bound on the mixture density m_{π_W} on the relevant region. That lower bound is [conjecture], with the failure mode (boundary π_W , barely-overlapping class-conditional supports) explicitly delimiting it.

Theorem B.18 ((B)-contingency of irreducibility; scoped maximality). [proven] Fix the observation model (S_4) . Over the identical channel:

1. (B): wall $D_{\text{mean}}(\beta; R) = \mathbb{E}_\mu[w_\beta] > 0$ whenever $p_W(R \cap \{\beta > 0\}) > 0$, invariant to n (Theorem B.8);
2. $\mathcal{C}_{\text{cov}}$: $D^{\mathcal{C}} = 0$ identically, at $n = 0$ (Prop. B.13);
3. $\mathcal{C}_{\text{ls}}$ (known, linearly independent class-conditionals): $D^{\mathcal{C}} = 0$ at population level, π_W recovered at rate $n^{-1/2}$ (Thm. B.16, Prop. B.17);
4. (maximality, **scoped**): any mechanism class $\mathcal{C}$ over the same channel whose compatible conditionals embed in the budget, $I_{\mathcal{C}}(p_W) \subseteq I(\beta)$, satisfies $D_{\Delta}^{\mathcal{C}}(R) \leq D_{\Delta}(\beta; R)$ (monotonicity of the pairwise supremum). Thus (B) is maximal **among budget- β -respecting mechanism priors**. Note $\mathcal{C}_{\text{ls}}$ does not embed in $I(\beta)$ in general. Its collapse is by the direct computation (3), not by (4).

Hence label-free irreducibility is a property of the mechanism prior (B), not of label-free observation per se. $\square$

Remark B.19 (the unrestricted “maximal wall” reading does not hold: heuristic only).

The budget-free class $\mathcal{C}_{\text{all}} = \{p_W \otimes q : q \text{ any kernel}\}$ is a *less* committal prior over the same channel; by Lemma B.1 every kernel is compatible with every marginal, so $I_{\mathcal{C}_{\text{all}}}(p_W)$ is all kernels and $D_{\text{mean}}^{\mathcal{C}_{\text{all}}}(R) = 1$ (take $q \equiv \delta_1$, $q' \equiv \delta_0$), exceeding $\mathbb{E}_\mu[w_\beta]$ whenever the latter is < 1 . So “any defensible coupling prior can only lower the wall” holds only under clause (4)’s embedding restriction; the “upper envelope” reading of Section 6.2 should be read as scoped to budget-respecting priors, and a precise envelope formulation over non-embedding priors (re-centered or enlarged balls) remains **unformulated** (Online Appendix 1, §S1, open point 8).

Proposition B.20 (the mechanism class is itself unfalsifiable from unlabeled data). [proven] If $p_W \in \mathcal{M}_{\text{ls}}$, the worlds $p_W \otimes q$ ($q \in I(\beta)$, e.g. $q = p_0$), $p_W \otimes p_0 \in \mathcal{C}_{\text{cov}}$, and $P_{\pi_W} \in \mathcal{C}_{\text{ls}}$ all induce $p_W^{\otimes n}$ for every n ; every label-free test between mechanism classes has power = size. The choice among (B), covariate shift, label shift is a prior, not a measurable fact of the window. $\square$

Remark B.21 (two flavors of collapse; the currency). $\mathcal{C}_{\text{cov}}$ collapses **prior-driven** ($n = 0$); $\mathcal{C}_{\text{ls}}$ collapses **data-pinned** (the class reduces the unknown to $\pi \in \Delta(\mathcal{Y})$, identified and $n^{-1/2}$ -estimated from unlabeled data). Under (B) neither route exists: the observation law is constant

in the unknown. The currency of class (3) of Theorem B.18 is “known $\{\mu_y\}$ ”, label-derived side information obtained outside the window: exactly the currency to which the program’s deliverable is directed.

H The magnitude-wall margin boundary (Conjecture E.6 of Online Appendix 1, regime (II))

Inherited results. Section G resolved the covered-part covariate wall $D_{\text{cov}} = 2 \int_{\mathcal{J}} \Psi d\nu$, $\Psi(r) = \mathbb{E}[b \mid R=r]$, at the parametric rate in the *direct/sign-coherent* regime (I), and left *one* decisive obstruction: the *raw magnitude* wall $b = |\bar{\eta}_J - \eta_0|$ of regime (II), where the non-smooth $|\cdot|$ Stage-1 map makes the outcome-bias hypothesis (H5) fail “for all $K \geq 1$ ” (Rmk. G.1.5 of Appendix G, Online Appendix 1). Section H establishes the regular branch and records the conditional/open status of the lower-bound program. It assembles two derivations (Route A constructive, Route B lower-bound) and their synthesis, and records two qualifications: (i) no σ^2/b heavy tail, and hence no ε_1^2 bias-cap / “ $\alpha_s > d/2$ ” clause, arises in the synthesis for Conjecture E.6 (remark following Lem. H.1), and (ii) a general- K Assouad lower bound at the exact rate $(N/K)^{-a}$ does not apply (the functional is a scalar projection $\sum_j \tau_j$; the fuzzy Le Cam bound of Thm. H.7 loses a $\sqrt{K}$ factor). All notation is inherited from Sections C and G.

Inherited notation and standing conditions. $g(x) := \bar{\eta}_J(x) - \eta_0(x)$, $b := |g|$, $s := \text{sign } g$, plug-in sign $\hat{s} := \text{sign}(\hat{\eta}_J - \eta_0)$ from the per-window within-split Stage-1 fit $\hat{\eta}_J$; $\mathbb{E}[y - \eta_0 \mid \mathcal{F}_\phi] = g$ (observation condition O2 of Appendix C, Online Appendix 1, plus Stage-1 in ϕ -coordinates); representer $w = d\nu/d\bar{\mu} \leq \Gamma_{\text{cov}}$ (A6); efficient influence function (EIF) $\phi = 2[w(b - \Psi) + (\Psi - \theta)]$ (Prop. G.1.1). Set

$$a := \frac{\alpha_s}{2\alpha_s + d} \in \left(0, \frac{1}{2}\right) \quad (\text{since } d \geq 1), \quad \varepsilon_1 \asymp \left(\frac{\log k}{k}\right)^a \quad (\text{Lemma C.11'}), \quad k = N/K, \quad N = Kk.$$

Standing: **(T)** $\bar{P}(g = 0) = 0$ (crossing set is $\bar{\mu}$ -null); **(Mgn $_\gamma$)** $\bar{P}(0 < b \leq t) \leq Ct^\gamma$. Throughout K is fixed and $k \rightarrow \infty$ unless a growing- K regime is named. Define the *margin threshold*

$$\boxed{\gamma^* := \frac{1}{2a} - 1 = \frac{d}{2\alpha_s}} \quad \Longleftrightarrow \quad a(1 + \gamma^*) = \frac{1}{2}.$$

H.1 Route A: $\sqrt{N}$ above the margin threshold

The construction estimates b by *recovering its sign* and reading the *raw* Bernoulli residual, so no Jensen/smoothing bias is ever incurred; the only error is misclassifying the sign, which the margin makes second-order.

Lemma H.1 (weighted sign error; the exact outcome-bias identity). *[proven] On the Stage-1 sup-norm event $E_1 = \{\|\hat{\eta}_J - \bar{\eta}_J\|_\infty \leq \varepsilon_1\}$ (prob. $\geq 1 - \sum_j k_j^{-2}$, Lemma C.11') the straddle holds pointwise,*

$$\hat{s}(x) \neq s(x) \implies 0 < b(x) \leq \varepsilon_1, \quad (\star)$$

because $\hat{g} - g = \hat{\eta}_J - \bar{\eta}_J$ gives $|\hat{g} - g| \leq \varepsilon_1$, so a sign flip forces $0 < b = |g| \leq |g - \hat{g}| \leq \varepsilon_1$. Hence, under (Mgn_γ) ,

$$\bar{P}(\hat{s} \neq s) \leq \bar{P}(0 < b \leq \varepsilon_1) \leq C\varepsilon_1^\gamma, \quad \boxed{\mathbb{E}_{\bar{P}}[b \mathbf{1}(\hat{s} \neq s)] \leq \varepsilon_1 \bar{P}(0 < b \leq \varepsilon_1) \leq C\varepsilon_1^{1+\gamma}.}$$

The sign-recovered signed observation $\hat{b}^{\text{obs}} := \hat{s}(y - \eta_0)$ has, via the pointwise identity $(\hat{s} - s)g = -2b \mathbf{1}(\hat{s} \neq s)$,

$$\begin{aligned} \beta_1^{\text{sign}}(R) &:= \mathbb{E}[\hat{b}^{\text{obs}} - b \mid R] = \mathbb{E}[(\hat{s} - s)g \mid R] \\ &= -2 \mathbb{E}[b \mathbf{1}(\hat{s} \neq s) \mid R], \quad \|\beta_1^{\text{sign}}\|_{L^1(\bar{\mu})} \leq 2C\varepsilon_1^{1+\gamma}. \end{aligned}$$

Off E_1 the contribution is $\leq O(1) \sum_j k_j^{-2} \asymp K^3 N^{-2} = o(N^{-1/2})$ (K fixed). A dyadic-shell refinement of $(\star)$ gives the same order with a γ -independent constant.

Remark (plug-in smoothing bias versus sign-error bias). The naive plug-in $|\hat{\eta}_J - \eta_0|$ carries the deterministic Stage-1 *smoothing* bias $\asymp \varepsilon_1$ for all γ (this is exactly Rmk. G.1.5(a)) and is **not** the estimator here. Route A uses the *raw* label, for which $\mathbb{E}[y - \eta_0 \mid \mathcal{F}_\phi] = g$ is unbiased and the *only* bias is the weighted sign error above; consequently $\|\beta_1^{\text{sign}}\|_{L^1} \lesssim \varepsilon_1^{1+\gamma}$ with no cap at ε_1^2 . (No heavy tail σ^2/b arises: for sub-Gaussian noise the pointwise $|\cdot|$ -bias decays super-exponentially for $b \gg \sigma$, so the integrated Jensen bias converges to $\varepsilon_1^{1+\gamma}$ for every $\gamma > 0$; no $\alpha_s > d/2$ clause is needed.)

Proposition H.2 (linearization: regime (II) $\rightarrow$ regime (I)). *[proven] With the sign recovered, $b = sg$ is linear in $\bar{\eta}_J$ on the correctly-signed region, so $\hat{b}^{\text{obs}} = \hat{s}(y - \eta_0)$ is the regime-(I) direct/sign-coherent object of Section G: conditionally on the frozen $\hat{s}$ the label fluctuation $y - \bar{\eta}_J$ is mean-zero given R (variance, not bias), and the entire outcome bias is β_1^{sign} of Lem. H.1. The fold-balanced anchor-level split of Def. G.1.2 freezes $\hat{s}$ off the evaluation fold, so $\hat{b}_{ji}^{\text{obs}}$ has conditionally-independent, mean-zero fluctuation given R (this neutralizes the second, in-sample-fluctuation obstruction of Rmk. G.1.5(b)). The EIF $\phi = 2[w(b - \Psi) + (\Psi - \theta)]$, the DML2 one-step $\hat{D}_{\text{cov}}$ (Def. G.1.2 with $\hat{b}^{\text{obs}}$), and the double-robust product remainder $R_{\text{prod}} = \int (\hat{\Psi} - \Psi)(w - \hat{w}) d\bar{\mu}$ (Lem. G.1.3) apply verbatim.*

Remark (no orthogonality in the sign nuisance). We do not invoke “Neyman-orthogonality in the sign nuisance”: $s \in \{\pm 1\}$ admits no score/tangent, so the Gateaux-derivative-in- s language is not defined. The $\varepsilon_1^{1+\gamma}$ bias is established self-containedly by the exact identity and margin bound of Lem. H.1; nothing downstream uses orthogonality-in- s .

Theorem H.3 (Route A $\sqrt{N}$ CLT above threshold). *[partial: new argument proven; CLT inherits the Section-G regime-(I) foundation] Assume (A0)–(A7), (A5'), interior regime, (T), (Mgn $_\gamma$) with $\gamma > \gamma^*$, K fixed, $k \rightarrow \infty$, the within-window split (Def. G.1.2), the sup-norm rate $\varepsilon_1 \asymp (\log k/k)^a$ (Lemma C.11'), and (H1)–(H4), (H6). Then the sign-corrected DML2 one-step*

$$\hat{D}_{\text{cov}} = \frac{2}{m} \sum_l \hat{\Psi}(\hat{r}_{K+1}(\tilde{x}_l)) + \frac{2}{N} \sum_\ell \sum_{(j,i) \in I_\ell} \hat{w}^{(-\ell)}(\hat{r}_j) (\hat{b}_{ji}^{\text{obs}} - \hat{\Psi}^{(-\ell)}(\hat{r}_j))$$

satisfies $\sqrt{N}(\hat{D}_{\text{cov}} - D_{\text{cov}}) \Rightarrow \mathcal{N}(0, V)$, $V = 4\mathbb{E}_{\bar{\mu}}[w^2 \sigma_b^2] + 4\tau_{NT} \text{Var}_\nu(\Psi)$ with $\tau_{NT} := \lim N/m$ (Prop. G.1.1; efficient constant V_{A1}^* under the generalized-method-of-moments pooling of Lem. G.3.1), for the magnitude wall $D_{\text{cov}} = 2 \int_{\mathcal{J}} \Psi d\nu$, $\Psi(r) = \mathbb{E}[|\bar{\eta}_J - \eta_0| \mid R=r]$. The Hájek–Le Cam local-asymptotic-minimax (LAM) bound (Thm. G.3.2(a)) supplies the matching $N^{-1/2}$ floor at fixed super-threshold γ , so the rate is two-sided there.

Key steps and the threshold computation. Decompose as in Thm. G.1.4: $\hat{D}_{\text{cov}} - D_{\text{cov}} = 2[(\nu_m - \nu)\hat{\Psi} + (\bar{P}_N - \bar{P})(\hat{w}(\hat{b}^{\text{obs}} - \hat{\Psi}))] + 2R_2$, $R_2 = R_{\text{prod}} + R_{\text{ratio}} + R_{\text{stage1}}^{\text{sign}}$, with $R_{\text{stage1}}^{\text{sign}} = \int \hat{w} \beta_1^{\text{sign}} d\bar{\mu}$. Then $R_{\text{prod}} = o_P(N^{-1/2})$ by (H3) (Lem. G.1.3), $R_{\text{ratio}} = o_P(N^{-1/2})$ by (H6), and $|R_{\text{stage1}}^{\text{sign}}| \leq \Gamma_{\text{cov}} \|\beta_1^{\text{sign}}\|_{L^1} \leq 2CT_{\text{cov}}\varepsilon_1^{1+\gamma}$. With $\varepsilon_1 \asymp k^{-a}$ (logs absorbed by the strict inequality) and $k = N/K$,

$$\varepsilon_1^{1+\gamma} \asymp (N/K)^{-(1+\gamma)a} \stackrel{K \text{ fixed}}{\asymp} N^{-(1+\gamma)a},$$

$$N^{-(1+\gamma)a} = o(N^{-1/2}) \iff (1+\gamma)a > \frac{1}{2} \iff \boxed{\gamma > \gamma^* = \frac{1}{2a} - 1 = \frac{d}{2\alpha_s}}.$$

Cross-fitting replaces any Donsker condition; Lindeberg–Feller on the two independent bounded mean-zero averages plus Slutsky give the CLT (Lemma G.14' with the Lindeberg–Feller step, [auxiliary-lemmas section](#)). $\square$

Qualifications. Theorem H.3 is a *reduction*, not an unconditional closure. Its *new* content is *proven*, namely the weighted sign error (Lem. H.1), the bias identity, and the threshold γ^* . Its CLT conclusion inherits the two Section-G regime-(I) qualifications graded *partial*, (i) and (ii), and one input, (iii): (i) the efficient constant is V_{A1}^* (GMM/inverse-variance pooling), $= V^*$ only under disjoint novelty coverage; (ii) the ratio channel (A5') is *not* removed, so (H6) stands as $\varepsilon_r = o(N^{-1/2})$ (Conj. G.2.3 open); (iii) the straddle ($\star$) uses the *sup-norm* Stage-1 rate, supplied by Lemma C.11' ([auxiliary-lemmas section](#)) under the minimal-mass reading of (A3); an L^2 -only Stage-1 rate would give a different Audibert–Tsybakov exponent. Thus the correct reading is “*reduces regime (II) to regime (I) above the margin threshold, modulo the Section-G regime-(I) qualifications.*”

Corollary H.4 (reach limitation and the growing- K threshold). [*proven*] (a) Generic-margin reach. Under (T) with a genuine crossing (bounded positive density of g at 0, which is the generic interior case and the only one where the sign is truly ambiguous), the margin exponent is $\gamma = 1$ (the tube $\{0 < b \leq t\}$ has mass $\asymp t$). Then $\gamma > \gamma^*$ requires $\gamma^* < 1$, i.e. $\boxed{d < 2\alpha_s}$. So at the generic transversal margin Route A's $\sqrt{N}$ is non-vacuous only in low intrinsic dimension / high smoothness; for large- d representations with $d \geq 2\alpha_s$ the generic crossing lies at or below threshold: below it falls under Route B, and at equality no claim is made (Thm. H.9(c)). This is a scope statement, not a defect: $\gamma^* > 0$ always ($a < \frac{1}{2}$), and the naive magnitude reconstruction is the $\gamma = 0$ case, which needs $a > \frac{1}{2}$ (impossible for $K \geq 1$). That is the exact failure of (H5) that Route A removes. (b) Growing K . If $K = K_N$ with $\log K / \log N \rightarrow \lambda \in [0, \frac{1}{2})$ (so $k \asymp N^{1-\lambda}$), solving $(N/K)^{-(1+\gamma)a} = o(N^{-1/2})$ gives $\gamma^*(\alpha_s, d, \lambda) = \frac{1}{2a(1-\lambda)} - 1$. The restriction $\lambda < \frac{1}{2}$ is necessary: the off- E_1 failure budget is $\sum_j k_j^{-2} = Kk^{-2} \asymp N^{3\lambda-2}$, which is $o(N^{-1/2})$ iff $\lambda < \frac{1}{2}$; for $\lambda \geq \frac{1}{2}$ the centering is uncontrolled unless the Stage-1 tail is strengthened from k^{-2} to k^{-c} with $Kk^{-c} = o(N^{-1/2})$ (supplied by the tail generalization of Lemma C.11', [auxiliary-lemmas section](#)). Along growing K the CLT additionally requires a variance-stabilization hypothesis; see the hypothesis check following Lemma G.14' ([auxiliary-lemmas section](#)).

H.2 Route B: the Cai–Low floor below the margin threshold

Below threshold the wall is a genuine non-smooth-functional object. We state the mechanism, the conditional $\gamma = 0$ floor, and the remaining gaps: the affinity estimate needed by the lower-bound tool, gap (2c) (the quantitative resummation estimate for the permutation-mixture affinity, Appendix K of Online Appendix 1), is open, and the positive-margin adaptation is unwritten.

Proposition H.5 (polynomial-approximation moment duality). *[proven] With $\delta_D := \inf_{\deg p \leq D} \sup_{|v| \leq 1} ||v| - p(v)| = \beta_*/D(1 + o(1))$ (Bernstein constant $\beta_* \approx 0.2802$),*

$$\sup \left\{ \mathbb{E}_{\nu_1} |v| - \mathbb{E}_{\nu_0} |v| : \nu_0, \nu_1 \in \mathcal{P}([-1, 1]), \int v^\ell d\nu_0 = \int v^\ell d\nu_1, \ell = 0, \dots, D \right\} = 2\delta_D$$

by LP duality of the moment problem; the extremal pair sits on the Chebyshev alternation points of the best degree- D approximant. Two drift-height priors matching D moments are near-indistinguishable from Bernoulli labels, yet the $|\cdot|$ -functional they induce differs by $\asymp 1/D$. This non-regularity lives entirely in the labels $\rightarrow b$ map, not in the linear $\int(\cdot) d\nu$ (Prop. G.1.5).

Theorem H.6 (single-window floor, $K = 1$). *[partial: cited-standard (Thm. E.3); shape-class admissibility is supplied by the Appendix-K derandomization already at $K = 1$; its affinity bound is modulo gap (2c) (Thm. K.6)]* For $K = 1$, k anchors, over the Hölder(α_s) interior class in the zero-crossing (no-margin) regime, $\inf_{\hat{\theta}} \sup \mathbb{E} |\hat{\theta} - \theta| \gtrsim k^{-a}/\text{polylog}(k)$, i.e. k^{-a} up to logarithmic factors. Partition the d -dimensional ϕ -support into $m = h^{-d}$ cells of side $h = k^{-1/(2\alpha_s+d)}$; on each cell set $g = Av_c$ with $A = h^{\alpha_s} = k^{-a}$ (Hölder-maximal); draw $v_c \stackrel{iid}{\sim} \nu_c$ (the moment-matched pair of Prop. H.5). With $A^2 n_c = kh^{2\alpha_s+d} = \Theta(1)$ and $D \sim \log k / \log \log k$ so $(D+1)! \gtrsim m$, the mixture $\chi^2 = O(1)$ while the functional gap is $A \cdot 2\delta_D \asymp k^{-a}(\log \log k / \log k)$; Le Cam's two-point method then gives the bound. This is the Cai–Low/Lepski–Nemirovski–Spokoiny $\int |f|$ rate = the pointwise regression rate (no averaging gain, never $\sqrt{N}$ -parametric); the sketch gives $k^{-a} \log \log k / \log k$, and the exact logarithmic power together with the uniform-over-cells Bernoulli $\leftrightarrow$ Gaussian transfer are cited from Cai and Low (2011) and Nussbaum (1996) (the cited-standard step Thm. E.3 already carries).

Theorem H.7 (fixed- K lower bound at $\gamma = 0$). *[conditional: the admissible-class version routes through the Appendix-K derandomization and requires the open affinity gap (2c) (Thm. K.6), in addition to the cited-standard transfer steps in Thm. E.3]* Place the K windows on disjoint novelty cells $\mathcal{J}_1, \dots, \mathcal{J}_K$ (Stage-1 cannot pool, Rmk. C.14), each of ν -mass $1/K$, $w \equiv 1$. Let each window be independently “active” with probability p under Λ_0 and p' under Λ_1 , with $|p' - p| = c/\sqrt{K}$, active state = the fuzzy prior ν_1 of Prop. H.5 at budget k , coherently oriented. Then the total data-affinity is bounded,

$$\chi^2(\Lambda_1 \| \Lambda_0) = \prod_{j=1}^K (1 + \chi_j^2) - 1 \leq e^{\sum_j \chi_j^2} - 1, \quad \sum_{j=1}^K \chi_j^2 \asymp K \cdot \frac{c^2}{K} \cdot O(1) = O(1),$$

so $\text{TV}(\Lambda_1, \Lambda_0) \leq \frac{1}{2}$ for c sufficiently small, while the functional separates by $\mathbb{E}_{\Lambda_1} \theta - \mathbb{E}_{\Lambda_0} \theta = (p' - p) \cdot 2\delta_D A \asymp (N/K)^{-a} \frac{\log \log k}{\log k} / \sqrt{K}$, the same logarithmic factor as in Thm. H.6 (it comes from $\delta_D \asymp 1/D$ at the degree $D \sim \log k / \log \log k$ needed for $\chi^2 = O(1)$). Le Cam gives, conditional on gap (2c), in the no-margin case $\gamma = 0$,

$$\inf_{\hat{\theta}} \sup \mathbb{E} |\hat{\theta} - \theta| \geq c' \frac{(N/K)^{-a}}{\sqrt{K} \text{polylog}(k)} = c' \frac{k^{-a}}{\sqrt{K} \text{polylog}(k)},$$

with the same logarithmic power as in Thm. H.6. For fixed K this is k^{-a} up to logarithmic factors, and $k^{-a}/(\sqrt{K} \text{polylog}(k))/N^{-1/2} = k^{1/2-a}/\text{polylog}(k) \rightarrow \infty$ (since $a < \frac{1}{2}$): $\sqrt{N}$ -estimability is impossible and θ is non-regular, for every fixed $K \geq 1$.

Remark H.8 (why Assouad does not apply, and the $\sqrt{K}$ gap). A general- K derivation might invoke Assouad over $\tau \in \{0, 1\}^K$ with $\theta(P_\tau) = \theta_{\text{base}} + (A\delta_D/K) \sum_j \tau_j$. But this functional depends on

τ only through the scalar $\sum_j \tau_j$: the induced separation is $|\theta_\tau - \theta_{\tau'}| = (A\delta_D/K)|\sum(\tau - \tau')| \leq (A\delta_D/K)\rho_H(\tau, \tau')$, an *upper* bound by Hamming distance, whereas Assouad's lemma as stated in Tsybakov (2009, Thm. 2.12) requires a *lower* bound $\geq 2s\rho_H$; per-coordinate bit decoding is impossible for a projection functional. The correct tool is the fuzzy Le Cam bound of Thm. H.7, which costs a *genuine* $\sqrt{K}$ relative to the rate $(N/K)^{-a}$. Consequences:

- (i) **Fixed K :** non-regularity/ $\sqrt{N}$ -impossibility is Thm. H.7's conclusion, conditional on gap (2c) and the Thm. E.3 transfer steps; the rate is k^{-a} up to logarithmic factors and the constant $\sqrt{K}$.
- (ii) **Growing K :** the $\sqrt{K}$ gap diverges. Only the sandwich $[(N/K)^{-a}/(\sqrt{K} \text{polylog}), (N/K)^{-a} \cdot \text{polylog}]$ is available: its lower end is Thm. H.7's floor, conditional on gap (2c), and its upper end is the minimax rate conditional on the same gap (Thm. I.2.1 of Appendix I, Online Appendix 1; Proposition 6.8). For the variance-averaged estimator the coherent linear functional does average down by $\sqrt{K}$.
- (iii) **Admissibility:** the $|\cdot|$ -nonsmoothness caps $\Psi \lesssim A = k^{-a}$ per cell, so keeping all 2^K sign patterns monotone-Lipschitz (A4) forces $K \cdot 2\delta_D A \lesssim A$, i.e. $K \lesssim \log k / \log \log k$. Beyond that the cube is inadmissible in $\mathcal{P}_{\text{mon}}^L$. This cap applies to the i.i.d.-heights cube only; the derandomized, phase-complementary construction of Appendix K (Online Appendix 1) is admissible at every K ; only its permutation-mixture affinity bound remains modulo gap (2c).
- (iv) **Margin realizability:** the zero-crossing-with-density drift gives $\gamma = 1$, which is sub-threshold ($\gamma < \gamma^*$) iff $\alpha_s < d/2$; for $\alpha_s \geq d/2$ a higher-order-vanishing ($\gamma < 1$) crossing must be built and its Hölder membership checked (not carried out here); the margin adaptation of Thm. H.7 to $0 < \gamma < \gamma^*$ is unwritten.
- (v) **Upper bound:** the crude per-window plug-in has coherent, non-cancelling $|\cdot|$ -Jensen bias $\asymp k^{-a}$ but is *not* a certificate of the floor; the matching achiever is the poly-debiased U-statistic of Prop. H.10 (rate-sharp only at $K = 1$).

H.3 The regular branch, boundary estimator, and lower-bound status

Theorem H.9 (regular branch and lower-bound status for the magnitude wall). *[partial] Let $\gamma^* = d/(2\alpha_s)$, $a = \alpha_s/(2\alpha_s + d) \in (0, \frac{1}{2})$. Under (A0)–(A7), (A5'), (T), (Mgn $_\gamma$), K fixed, $k \rightarrow \infty$, and for the within-window-split one-step, the following hold:*

- (a) Above threshold $\gamma > \gamma^*$ (Route A): D_{cov} is regular, rate $N^{-1/2}$, EIF $\phi = 2[w(b - \Psi) + (\Psi - \theta)]$, efficient constant V_{A1}^* (Thm. H.3), modulo the Section-G regime-(I) qualifications.
- (b) Below threshold $\gamma < \gamma^*$ (Route B): at $\gamma = 0$, Thm. H.7 gives a conditional fixed- K lower bound $(N/K)^{-a}/\sqrt{K}$ up to logarithmic factors, implying non-regularity if gap (2c) and the cited transfer steps hold. For $0 < \gamma < \gamma^*$, both the proposed margin-adaptive upper bound (Prop. H.10) and the matching lower-bound adaptation are open.
- (c) Boundary point: $a(1 + \gamma^*) = \frac{1}{2}$ identically, so the two polynomial exponents meet at $\gamma = \gamma^*$. This identity does not settle the boundary case: the Stage-1 rate contains a logarithmic factor, and Route A requires the strict little-o condition $\varepsilon_1^{1+\gamma} = o(N^{-1/2})$.

Consequently, for fixed K (the first line theorem-grade modulo the regime-(I) qualifications, in particular hypothesis (H6); the second conditional on gap (2c)),

$$\inf_{\hat{D}} \sup \mathbb{E}|\hat{D} - D_{\text{cov}}| \asymp N^{-1/2} \quad (\gamma > \gamma^*),$$

$$\inf_{\hat{D}} \sup \mathbb{E}|\hat{D} - D_{\text{cov}}| \gtrsim (N/K)^{-a}/(\sqrt{K} \text{polylog}(k)) \quad (\gamma = 0; \text{ conditional}).$$

Scope. The $\gamma = 0$ lower side is conditional on gap (2c) even at $K = 1$, because the admissible construction depends on it.

Proposition H.10 (the boundary estimator: Cai–Low/LNS polynomial debiasing). *[partial ($K=1$); conjecture ($K>1$)] Per window and per width- h cell ($n_c = kh^d$ anchors, drift $\mu_C = \bar{\eta}_j - \eta_0$), the estimator proceeds in four steps: (1) take the best minimax polynomial P_{2D} of $|x|$ on $[-M, M]$, $M \asymp k^{-a}$, with uniform error $\asymp M/D$; (2) form the unbiased U -statistic $\hat{U}_\ell = \binom{n_c}{\ell}^{-1} \sum_{\text{distinct}} \prod_p (y_{i_p} - \eta_0)$ for μ_C^ℓ (η_0 known); (3) set $|\widehat{\mu_C}| = \sum_\ell c_\ell \hat{U}_\ell$, which has bias $\leq M/D$ and variance $\asymp M^2 2^{O(D)}/n_c$; (4) choose the degree $D^* \asymp \log n_c$ so that bias² $\asymp$ var. Aggregating $\frac{1}{N} \sum_{\text{cells}} |\widehat{\mu_C}| w$ attains $k^{-a} = (N/K)^{-a}$ up to logarithmic factors; with Thm. H.6’s lower bound this is rate-sharp up to logarithmic factors at $(\gamma=0, K=1)$, conditional on gap (2c). This is the estimator that realizes the clean $\varepsilon_1^{1+\gamma}$ rate (raw-label U -statistics carry no $\hat{\eta}$ -smoothing bias), together with Route A’s sign-recovery step, and not the plug-in $|\hat{\eta}_J - \eta_0|$. For $0 < \gamma < \gamma^*$ the margin-adaptive version applies poly-debiasing only in the near-crossing layer (mass $\lesssim \varepsilon_1^\gamma$) and the linear-bulk one-step elsewhere, targeting $(N/K)^{-a(1+\gamma)}$ (the aggregation of the $2^{O(D)}/n_c$ cell variances is not carried out here, so this achievability is partial). The general- K aggregated lower bound is the $\sqrt{K}$ gap of Rmk. H.8.*

Theorem H.11 (regular branch of Conjecture E.6; Section G as the perfect-margin endpoint). *[partial] For fixed K , the covered-part covariate wall $D_{\text{cov}} = 2 \int_{\mathcal{J}} \Psi \, d\nu$ is $\sqrt{N}$ -regular (rate $N^{-1/2}$, EIF ϕ , efficient constant V_{A1}^*) under*

$$\underbrace{w \leq \Gamma_{\text{cov}}}_{(A6) \text{ overlap}} \wedge \underbrace{\bar{P}(\bar{\eta}_J = \eta_0) = 0}_{(T)} \wedge \underbrace{\gamma > \gamma^* = \frac{d}{2\alpha_s}}_{(Mgn_\gamma)},$$

for all α_s (no $\alpha_s > d/2$ clause; the near-crossing part has mass $\lesssim \varepsilon_1^\gamma$ and the bulk $\int (\bar{\eta}_J - \eta_0) s w \, d\nu$ is linear-in- Ψ , hence $\sqrt{N}$ -regular at any smoothness by Prop. G.1.5, so that no higher-order influence function is needed). At $\gamma = 0$, the conditional lower bound of Theorem H.7 implies non-regularity if gap (2c) holds. For $0 < \gamma < \gamma^*$, the lower-bound adaptation remains open; no claim is made here at $\gamma = \gamma^*$.

Reconciliation. Section G’s regime (I) is the certified-sign / $\beta_1 \equiv 0$ branch: the signed drift is observed, so $\beta_1 \equiv 0$ for every γ . It is the perfect-margin endpoint of the magnitude problem in the precise sense that as the sign-recovery penalty $\varepsilon_1^{1+\gamma} \rightarrow 0$ (margin $\gamma \rightarrow \infty$, or the sign supplied) Route A’s object converges to regime (I)’s directly-observed object. Section G’s own side condition $\varepsilon_1^{1+\gamma} = o(N^{-1/2})$ (Thm. G.1.4) is algebraically identical to $\gamma > \gamma^*$; and its “(H5) fails for all $K \geq 1$ ” is exactly the $\gamma = 0$ (no-margin) instance, since then $\|\beta_1\|_{L^1} \asymp \varepsilon_1$ needs a $> \frac{1}{2}$. Thus Section G sits on the regular side of the frontier: (H5) holds above γ^* ; equality remains unresolved, and its failure below the threshold is conditional/open as described above. With bounded overlap and transversality, the magnitude wall recovers $\sqrt{N}$ when drift stays sufficiently clear of the decision boundary $\bar{\eta}_J = \eta_0$. The conditional lower-bound program points to a non-parametric Cai–Low price for near-boundary drift but does not yet establish it throughout the sub-threshold region.

H.4 Summary of status (Section H)

The per-claim status record for this section (claims, assumptions and tools used, weakest steps) is in §S1 of Online Appendix 1. In summary, Conjecture E.6 (regime II, the magnitude wall) has an

established regular branch above $\gamma^* = d/(2\alpha_s)$, where the wall is $\sqrt{N}$ -regular (Route A, Thm. H.3, modulo the Section-G regime-(I) qualifications). At $\gamma = 0$, the fixed- K lower bound is conditional on the Appendix-K gap (2c); the positive-margin adaptation below the threshold is open. Equality is unresolved under the logarithmic Stage-1 rate. The general/growing- K conditional program is taken up in Appendix I of Online Appendix 1.

Auxiliary lemmas C.11' and G.14'

Scope. This section proves two steps that Appendices C and G of Online Appendix 1 and Appendix H use: the Stage-1 sup-norm k -NN rate behind Lem. C.11 (Lemma C.11' below; it is non-adaptive, tuning κ with known (α_s, d) , which suffices downstream), and the empirical-to-population and CLT steps of Thm. G.1.4, used again in Thm. H.3 (Lemma G.14' and the Lindeberg–Feller step below). The comparison with the published theorems is in §S2 of Online Appendix 1. Step (3) of Thm. C.12 is not covered here (closing paragraph).

Item 1: Stage-1 sup-norm k -NN (Lem. C.11). The published uniform k -NN rates do not cover the sup-norm rate in the form Lem. C.11 needs (§S2 of Online Appendix 1 compares them); the lemma below proves it directly, following the scheme of Jiang (2019) under (A3)-compatible hypotheses with two standard concentration inequalities.

Lemma C.11' (sup-norm variable-radius k -NN under minimal mass). [proven] *Let $x_1, \dots, x_n$ be i.i.d. q on a region $R \subset \mathbb{R}^D$ (ϕ -coordinates, $D = \text{ambient representation dimension}$), $y_i \in [0, 1]$, $\eta(x) = \mathbb{E}[y \mid x]$ (L, α_s)-Hölder on R . Assume the minimal-mass condition*

$$(M) \quad q(B(x, r)) \geq c_0 r^d \quad \forall x \in R, 0 < r \leq r_0.$$

Scope of (M). When the lemma is invoked per window $j = 1, \dots, K$ (as in Thm. H.3's straddle step), (M) is required with a single pair of constants (c_0, r_0) uniform over j , and must hold on the full anchor-region support of p_j (every $x \in R_j$, not merely the realized anchor sites), because H.1's straddle reads the sign at population points ($\beta_1^{\text{sign}}(R) = \mathbb{E}[\cdot \mid R]$ integrates over $X \sim \bar{P}$ on that support). Let $\hat{\eta}$ be the κ -NN average (the convention of Jiang (2019): average of y_i over $N_\kappa(x) = B(x, r_\kappa(x)) \cap \{x_i\}$, $r_\kappa(x) = \kappa$ -NN radius; ties enlarge N_κ , which only helps) with $\kappa = \lceil n^{2\alpha_s/(2\alpha_s+d)} (\log n)^{d/(2\alpha_s+d)} \rceil$. (Log convention: throughout this lemma and its proof, $\log$ denotes the natural logarithm, $\log = \ln$.) Then there is $n_0 = n_0(D, d, \alpha_s, c_0, r_0)$ such that for $n \geq n_0$, with probability $\geq 1 - n^{-2}$,

$$\sup_{x \in R} |\hat{\eta}(x) - \eta(x)| \leq \left[L(4/c_0)^{\alpha_s/d} + 2\sqrt{D+4} \right] \left(\frac{\log n}{n} \right)^{\alpha_s/(2\alpha_s+d)}.$$

Proof. Step 0 ((A3) $\Rightarrow$ (M)). If q is measure-doubling with exponent d on R ($q(B(x, 2r)) \leq 2^d q(B(x, r))$), the operative meaning of “doubling intrinsic dimension $\leq d$ ” and $\Delta = \text{diam } R$, then iterating $\ell = \lceil \log_2(\Delta/r) \rceil$ times gives $q(B(x, r)) \geq 2^{-d\ell} q(B(x, \Delta)) \geq (r/2\Delta)^d q(R)$: (M) holds with $c_0 = q(R)(2\Delta)^{-d}$, $r_0 = \Delta$. Alternatively, density bounds w.r.t. an Ahlfors- d -regular reference measure give (M) directly. Assumption (A3) is read here in its minimal-mass form (M); this argument records when (M) follows from the stronger measure-doubling reading printed in Section 3.1.

Step 1 (uniform radius; bias). Balls in $\mathbb{R}^D$ have VC dimension $D + 1$. By the VC relative-deviation bound (Bousquet et al., 2004, Thm. 5.1, as paraphrased in Chaudhuri and Dasgupta,

2010, Thm. 15): w.p. $\geq 1 - \delta/2$, every ball B satisfies $q_n(B) \geq q(B) - \beta_n \sqrt{q(B)}$ with $\beta_n = \sqrt{(4/n)((D+1) \ln 2n + \ln(16/\delta))}$. Set $h := (2\kappa/(c_0 n))^{1/d}$ (require $h \leq r_0$ and $\kappa \geq 8((D+1) \ln 2n + \ln(16/\delta))$), i.e. $\beta_n^2 \leq \kappa/2n$; both hold for $n \geq n_0$). For any $x \in R$, $q(B(x, h)) \geq c_0 h^d = 2\kappa/n$, and $t - \beta_n \sqrt{t} \geq \kappa/n$ for all $t \geq 2\kappa/n$ under $\beta_n^2 \leq \kappa/2n$; hence $q_n(B(x, h)) \geq \kappa/n$, i.e. $\sup_{x \in R} r_\kappa(x) \leq h$. Since every point of $N_\kappa(x)$ lies within $r_\kappa(x) \leq h$: $||N_\kappa(x)|^{-1} \sum_{i \in N_\kappa(x)} \eta(x_i) - \eta(x)| \leq Lh^{\alpha_s}$, uniformly. (No cover argument or per-cover-point stability of the NN sets is needed.)

Step 2 (design-conditional noise; realizable-NN-set counting). Condition on the design $(x_1, \dots, x_n)$: $\xi_i := y_i - \eta(x_i)$ are independent, mean-zero, $|\xi_i| \leq 1$. Every realizable $N_\kappa(x)$ is a subset of the design cut out by a closed Euclidean ball, of size $\geq \kappa$; by Sauer's lemma (VC dim $D+1$) the number of distinct ball-cut subsets is $M \leq (n+1)^{D+1}$. (We use Sauer's bound instead of the arrangement count in Jiang (2019, Lem. 3); it is conservative.) Hoeffding's inequality for a fixed set A with $|A| = s \geq \kappa$ gives $\Pr(|s^{-1} \sum_{i \in A} \xi_i| > t) \leq 2e^{-st^2/2} \leq 2e^{-\kappa t^2/2}$. A union bound with $t = \sqrt{2((D+1) \ln(n+1) + \ln(4/\delta))/\kappa}$ gives, w.p. $\geq 1 - \delta/2$ conditionally on *any* design (hence unconditionally), $\sup_x ||N_\kappa(x)|^{-1} \sum_{i \in N_\kappa(x)} \xi_i| \leq t$.

Step 3 (combine). Take $\delta = n^{-2}$. Then $\ln(4/\delta) \leq 2 \ln n + \ln 4$, and with the stated κ the bias is $\leq Lh^{\alpha_s} = L(2\kappa/(c_0 n))^{\alpha_s/d} \leq L(4/c_0)^{\alpha_s/d} (\log n/n)^{\alpha_s/(2\alpha_s+d)}$ (without the ceiling $(\kappa/n)^{\alpha_s/d} = (\log n/n)^{\alpha_s/(2\alpha_s+d)}$ exactly; the ceiling on κ costs at most a factor 2, absorbed into $4/c_0$), while the noise is $\leq 2\sqrt{D+4} (\log n/n)^{\alpha_s/(2\alpha_s+d)}$ for $n \geq n_0$. The failure probability is $\leq n^{-2}$. $\square$

Tail generalization. Repeating the proof with $\delta = n^{-c}$ changes only constants (κ 's side condition and the noise constant scale with c): Lemma C.11' holds with probability $\geq 1 - n^{-c}$ for any fixed $c > 0$, which supplies the k^{-c} tail used in Cor. H.4(b).

Item 2: empirical-to-population / CLT steps of Thm. G.1.4 (echoed in Thm. H.3). Chernozhukov et al. (2018) (henceforth CCDDHNR), Theorem 3.1, assumes an i.i.d. sample, while the paper's anchor design is window-stratified (independent, not identically distributed; inid), so it does not apply as published (§S2 of Online Appendix 1). Its fold-conditional proof device and its model-free Lemma 6.1 carry over verbatim; the short inid extension is written out below, and the CLT is then Lindeberg–Feller (van der Vaart, 1998, Prop. 2.27).

Lemma G.14' (fold-conditional empirical-to-population under window-stratified sampling). [proven] Fix fold ℓ with evaluation windows I_ℓ , each contributing k i.i.d. anchors (all windows independent), $N_\ell = |I_\ell|k$, $\bar{P}_\ell := |I_\ell|^{-1} \sum_{j \in I_\ell} P_j$. Let $\hat{f}_\ell$ be measurable w.r.t. the training data (independent of the evaluation anchors) with $\|\hat{f}_\ell\|_{L^2(\bar{P}_\ell)} < \infty$. Then

$$\mathbb{E} \left[((\bar{P}_{N_\ell} - \bar{P}_\ell) \hat{f}_\ell)^2 \mid \text{train} \right] = \frac{1}{N_\ell^2} \sum_{j \in I_\ell} k \text{Var}_{P_j}(\hat{f}_\ell) \leq \frac{\|\hat{f}_\ell\|_{L^2(\bar{P}_\ell)}^2}{N_\ell},$$

so $(\bar{P}_{N_\ell} - \bar{P}_\ell) \hat{f}_\ell = O_P(N_\ell^{-1/2} \|\hat{f}_\ell\|_{L^2(\bar{P}_\ell)})$ unconditionally by CCDDHNR Lemma 6.1(b); in particular $= o_P(N^{-1/2})$ whenever $\|\hat{f}_\ell\|_{L^2(\bar{P}_\ell)} = o_P(1)$ and $N_\ell \asymp N$.

Proof. Conditionally on the training data $\hat{f}_\ell$ is a fixed function; the evaluation anchors are independent with $\mathbb{E}[\bar{P}_{N_\ell} \hat{f}_\ell \mid \text{train}] = |I_\ell|^{-1} \sum_j \mathbb{E}_{P_j} \hat{f}_\ell = \bar{P}_\ell \hat{f}_\ell$, an identity that requires equal quotas $k_j \equiv k$ matching $\bar{P}$'s uniform mixture weights (A0); for unequal k_j , $\bar{P}$ must be re-weighted to $\sum_j (k_j/N) P_j$, which is precisely Appendix C's Convention C.0 (anchor-weighted pool $\bar{\mu} = N^{-1} \sum_j k_j \mu_j$): unequal quotas are handled by the already-weighted mixture, provided $\bar{P}$ (and $\bar{\mu}$) are read consistently in

that convention; Appendix G’s uniform mixture is the equal- k special case. The variance splits over independent terms, $\text{Var}_{P_j} \leq \mathbb{E}_{P_j}[\hat{f}^2]$, and averaging over $j \in I_\ell$ gives the displayed bound. Chebyshev’s inequality, followed by Lemma 6.1(b), gives the unconditional rate. $\square$

Application closing step (2). Take $\hat{f}_\ell = \hat{w}^{(-\ell)}(b - \hat{\Psi}^{(-\ell)}) - w(b - \Psi)$; since $b, \hat{\Psi}, \Psi \in [0, 1]$ and $w \leq \Gamma_{\text{cov}}$, $\|\hat{f}_\ell\|_{L^2(\bar{P}_\ell)} \leq \|\hat{w} - w\|_{L^2} + \Gamma_{\text{cov}}\|\hat{\Psi} - \Psi\|_{L^2} = o_P(1)$ by (H3). For the target term, $\hat{\Psi}$ is independent of the target sample and ν_m is i.i.d., so $\text{Var} \leq m^{-1}\|\hat{\Psi} - \Psi\|_{L^2(\nu)}^2 \leq m^{-1}\Gamma_{\text{cov}}\|\hat{\Psi} - \Psi\|_{L^2(\bar{\mu})}^2$, and (H4), $m \geq cN$, gives $o_P(N^{-1/2})$. Both displays of step (2) are proven for fold-balanced designs, in which every window contributes equally to every fold (the anchor-level design of Def. G.1.2), so that $\bar{P}_\ell = \bar{P}$ for each fold. The theorem uses this design. Window-level folding is not covered: there a fold-centering term $\sum_\ell (N_\ell/N)(\bar{P}_\ell - \bar{P})\hat{f}_\ell$, of first order in the nuisance error, appears and is not controlled by the three proof steps. Unequal quotas are handled by Convention C.0’s anchor-weighted mixture, as noted in the proof.

Step (3), the CLT: precise citation and hypothesis check. The remaining terms are $2N^{-1/2} \sum_{j,i} \phi_{\text{src}}(O_{ji}) + 2\sqrt{N/m} \cdot m^{-1/2} \sum_l \phi_{\text{tgt}}(\tilde{O}_l)$, $\phi_{\text{src}} = w(b - \Psi)$, $\phi_{\text{tgt}} = \Psi - \theta$: a row-independent triangular array of *bounded* summands ($|\phi_{\text{src}}| \leq \Gamma_{\text{cov}}$, $|\phi_{\text{tgt}}| \leq 1$). The CLT used is the Lindeberg–Feller CLT for independent non-identically distributed triangular arrays, namely [van der Vaart \(1998, Prop. 2.27, §2.8\)](#) (independent mean-zero rows, $\sum_i \text{Var} \rightarrow \Sigma$ plus the Lindeberg condition imply $\sum_i Y_{ni} \Rightarrow \mathcal{N}(0, \Sigma)$), plus Cramér–Wold and Slutsky; CCDDHNR’s own proof uses exactly this pair. *Hypothesis check:* (i) Independence across windows, anchors, and the target sample holds by (A0) and two-sample independence. (ii) The Lindeberg condition holds automatically: summands are bounded by $C/\sqrt{N}$ after scaling, so the Lindeberg sum vanishes for $N > (C/(\epsilon\sqrt{V}))^2$, given $V > 0$. (iii) Variance identification is the genuine content: under the fixed-quota design the CLT variance is $N^{-1} \sum_j k \text{Var}_{P_j}(\phi_{\text{src}})$, which equals the mixture variance $\text{Var}_{\bar{P}}(\phi_{\text{src}})$ used in Prop. G.1.1’s *V if and only if the per-window means vanish*, $\mathbb{E}_{P_j}[\phi_{\text{src}}] = 0 \forall j$. That is exactly what stationarity (A1) delivers ($\mathbb{E}[b \mid R=r, J=j] = \Psi(r)$ for every j). Without (A1) the stratified variance is $V - K^{-1} \sum_j (\mathbb{E}_{P_j} \phi_{\text{src}})^2$ (strictly $< V$ unless every per-window mean vanishes), so the CI constant V is conservative; the one-step estimator nevertheless *remains centered on the pooled* θ , and bias arises only relative to the *transfer estimand* $D_{\text{cov}}^{(K+1)}$, whose very definition requires (A1) extended to window $K+1$. (A1) is essential for the CLT constant. (iv) For growing $K = K_N$ (Cor. [H.4\(b\)](#)), convergence of $|I_\ell|^{-1} \sum_j \text{Var}_{P_j}$ is a *new* variance-stabilization hypothesis.

What is not covered. Step (3) of the proof of Theorem C.12 (Online Appendix 1), the inid extension of the *ratio-type* bracketing argument for the shape-constrained least-squares estimator ([van der Vaart and Wellner, 1996, Thm. 2.7.5](#)), is a separate open point and is not settled here.

J Measurement protocol and the proxy/sensitivity study

This appendix specifies the measurement protocol behind the empirical sections of the paper, namely the probe, oracle, and gate configurations of Sections 4–5, the bootstrap reconstruction of Section 5.6, and the decision-rule and attribution evaluations of Sections 7.1–7.2, and it documents the two evidence levels used in Section 6.5. Eight TabReD streams provide noisy, single-stream proxies for all threshold ingredients. Fifteen related benchmarks lack the labeled temporal drift field needed to estimate α_s and γ ; they therefore enter only a sensitivity analysis at $(\alpha_s, \gamma) = (1, 1)$. The appendix also records the deduplication correction that invalidated two raw low-dimensional readings (Section J.3). These analyses motivate, but do not establish, a population-level claim

about which phase real data sets occupy. Printed values are transcribed from the measurement record of the reproduction deposit (the per-stream record of each estimate and its provenance).

J.1 Measurement protocol

Data sets. Twenty-three real tabular data sets: the eight industrial TabReD streams (five regression, three classification; $n_{\text{test}} = 4,647\text{--}59,951$ test points), and fifteen shift benchmarks with precomputed test embeddings of the frozen tabular foundation model, namely the four ACS census tasks (the canonical shift benchmark), fraud (BAF, IEEE_FRAUD, CREDIT_CARD_FRAUD), network security (UNSW_NB15, CICIDS), medical (BRFSS_DIABETES, DIABETES_READMISSION, SEPSIS, MIMIC_IV_MORTALITY), credit (LENDING_CLUB), and BIKE_SHARING ($n_{\text{test}} = 6,101\text{--}1,162,213$).

Ingredients of the threshold. A data-set-specific phase assessment requires three quantities and a sampling protocol that supports their joint interpretation.

(i) *Deduplication first.* Rows identical to four decimal places are removed and every intrinsic-dimension estimate is recomputed on the unique subset. The step matters because duplicate rows produce first-neighbor distances $r_1 \approx 0$ that break both estimators below and bias d_0 sharply downward (Section J.3).

(ii) *Intrinsic dimension d_0 .* For the TabReD streams, the eight d_0 values printed in Table 4 are TwoNN (Facco et al., 2017) estimates on a 12,000-row subsample of the raw streams. The streams are 99.8–100% unique at four decimals and deduplication moves the estimate by at most 0.08 (Table 8, where the check is run with the MLE estimator), so the raw and deduplicated readings are interchangeable at the printed precision. For the fifteen related benchmarks, d_0 is the Levina–Bickel MLE (Levina and Bickel, 2004) on a 12,000-row subsample, with TwoNN and PCA cross-checks on borderline cases. The two estimators do not agree pointwise on low-uniqueness data (an instability documented in Section J.3); the deduplicated MLE values are the corrected record.

(iii) *Drift smoothness α_s .* On the TabReD streams (where labels exist), α_s is the Hölder exponent of the drift/residual field, read as half the variogram log-log slope. The slope saturates at 2 in the limit (finite-sample slopes can slightly exceed it), so the variogram cannot certify $\alpha_s > 1$; we therefore work under the rough-field bound $\alpha_s \leq 1$. Three qualifications are recorded in the measurement record: the estimate is taken on the k -NN-smoothed drift field and is therefore biased high (the raw-residual Hölder on WEATHER is ~ 0.5); the point estimates are unstable: the slope is refit on the lower half of the variogram bins alone, and on 5/8 streams the two estimates differ by more than 0.2 (SBERBANK-HOUSING, DELIVERY-ETA, HOMECREDIT-DEFAULT, HOMESITE-INSURANCE, ECOM-OFFERS; differences 0.22–0.76, against at most 0.11 on the other three), with four of the sixteen values sitting on the clip bounds $[0.05, 1]$; and the field is proxied from residuals on a *single* stream, not the idealized multi-window drift response Ψ of the theory. We therefore report them as stream-level proxies and separately inspect the rough-field bound $\alpha_s \leq 1$. On the fifteen label-free benchmarks we also measured the readout-field smoothness $\alpha(\eta_0)$, the variogram exponent of $\eta_0(x) = \sigma(z_0(x))$, where z_0 is the logit of the frozen readout $h_0 \circ \phi$ and σ the logistic function; this is a label-free *proxy* for α_s . Because no result here justifies substituting it for drift smoothness, it is not used in the main table’s threshold calculation.

(iv) *Margin exponent γ .* On the TabReD streams γ is fit from the same residual field (the same single-stream proxy qualification applies). The measured values sit near the generic transversal value $\gamma \approx 1$, ranging from 0.58 (HOMESITE-INSURANCE) to 3.39 (HOMECREDIT-DEFAULT). At that value the drift crosses the decision boundary with a density. On the related benchmarks γ is not measurable; $\gamma = 1$ is an explicit sensitivity assumption, not an estimate.

Probe and oracle configurations. All measurements of Sections 4–5 run on precomputed predictions and 192-dimensional embeddings of the frozen backbone. The regional-offset probe fits a k -means codebook of 1, 8, 32, or 128 regions and per-region offsets (mean residual for regression; a one-dimensional Newton logit offset for classification), regions with fewer than 30 fitting points shrinking fully to the global offset; the out-of-time arm fits codebook and offsets on the first half of the stream and evaluates on the second. The C4 replication (Section 4.1) repeats the out-of-time arm at cut fractions 0.4, 0.5, 0.6, and 0.7, with k -means seeds 0–5 (four initialization restarts per fit) and codebook sizes 8, 32, and 128 regions; the quoted ranges are the per-cut seed means at the best codebook size, selected on the same out-of-time evaluation (8 regions on HOMESITE-INSURANCE: +1.33, +1.53, +1.93, +1.76; 128 regions on ECOM-OFFERS: +1.30, +1.44, +1.79, +1.80; seed standard errors at most 0.007 on HOMESITE-INSURANCE and 0.12 on ECOM-OFFERS). At the other codebook sizes every cut on both streams remains positive (ECOM-OFFERS with 8 regions: +0.56 to +0.79). Signal ceilings use leave-one-out k -NN regression of residuals on embeddings ($k = 20$) in three grades: batch (time-symmetric), causally growing buffer, and strict out-of-time split. The temporal-gap stress test on HOMESITE-INSURANCE truncates the strict split’s fitting prefix G rows before the cut, so no retrieved neighbor lies within G stream positions of the evaluation half; $G = 1,000$ (the decoupling distance also used for the variogram) gives $R^2 = 0.006$, and the archived sweep reads 0.016, 0.013, 0.011, 0.006, 0.008 at $G = 0, 250, 500, 1,000, 2,000$. Noise floors use the temporally decoupled variogram (pairs $\geq 1,000$ steps apart). The gate probe couples a streaming k -NN corrector ($k = 20$, correction scale 0.7, block size 64) to a test-by-betting e-process on per-round log-loss regret (burn-in $n_{\text{burn}} \geq 20$, level 0.05). The corrector’s remaining constants (a minimum past buffer of 100 points; for classification, clipping of the corrected probability to $[10^{-6}, 1 - 10^{-6}]$) are listed with the reconstruction below and are the values in the deposited code; the e-process monitor itself is not part of the deposit. The drift-smoothness proxy α_s is half the log-log slope of the variogram of the 20-NN-smoothed residual field against embedding distance, over 15 log-spaced bins spanning the 2nd to 60th percentile of 20-NN distances on a 12,000-row subsample (bins with fewer than 50 pairs dropped; the resulting exponent clipped to $[0.05, 1]$); the two-range check refits the slope on the lower half of those bins. The margin exponent γ is the log-log slope of the empirical CDF of the smoothed residual magnitude between its 5th and 40th percentiles.

Resolution selection for WallB_ρ . The resolution is not a free parameter: it must match the fiber scale at which the deployed readout pools. The protocol sets ρ to the median distance to the k -th retrieved neighbor under the deployed retrieval ($k = 20$) over the window’s unlabeled draws, the operating scale of the corrector’s local averaging, and reports WallB_ρ together with the half- and double-resolution values $\text{WallB}_{\rho/2}$ and $\text{WallB}_{2\rho}$ as a monotone sensitivity band (within-cell estimation as in Section 7.2, Step 1). Choosing ρ below the retrieval scale reports capacity the readout cannot use; above it, reachable structure is misattributed to the wall. On the eight probe streams the resulting ρ (median Euclidean distance in the 192-dimensional embedding to the 20th retrieved neighbor over the evaluation half, as recorded in the reproduction deposit) ranges from 1.95 (HOMECREDIT-DEFAULT) to 3.97 (SBERBANK-HOUSING): WEATHER 2.28, ECOM-OFFERS 2.20, DELIVERY-ETA 2.53, MAPS-ROUTING 2.53, COOKING-TIME 2.79, HOMESITE-INSURANCE 2.94.

Worked attribution (Table 6). The three-way attribution instantiates Section 7.2 with both scales out of sample: fit on the first half of the stream, evaluate on the second, k -NN residual means at the deployed scale ($k = 20$) and at the finest supported scale ($m = 5$). $\text{WallB}_\rho = \max(0, R_{m=5}^2 - R_{k=20}^2)$; the fine arm’s variance penalty biases the difference downward, so a positive reading is unlikely to be noise, at the price that a true sub- ρ signal smaller than that penalty reads

as zero (the estimate is a detection, not a bound). The worked example computes the strict-split analogue of the protocol resolution (ρ from the fixed fit-half retrieval rather than the growing deployed buffer) and reports the single operating scale; the $\rho/2$, 2ρ band of the selection protocol applies to the within-cell estimator, not to this detection form. The $k = 20$ column is checked against the archived strict-split ceilings behind Table 1 (numeric where printed, sign elsewhere) before the shares are computed.

Screening-rule evaluation (Table 5). The eight-stream evaluation instantiates the two label-free gates of Section 7.1 with the rule fixed in advance of the outcomes. The floor is the calibrated estimator of that section: Bernoulli $\mathbb{E}[p(1 - p)]$ over the frozen predictions for classification, and for regression the short-range decoupled semivariance: the mean of $\frac{1}{2}(r_i - r_j)^2$ over the smallest 1% (and at least 50) of the *available* embedding distances among 400,000 uniformly sampled index pairs at least 1,000 stream steps apart. This is not the extrapolated nugget of Section 5.2, which reads the variogram down to zero distance where WEATHER has no support. The rule uses the non-extrapolated short-range value deliberately. On WEATHER, the one stream for which the extrapolated range is reported, the short-range value is 0.89 against 0.13–0.62, so it is the conservative reading for a gate that has to justify adapting. Both floors are divided by the residual variance of the stream. The local-determinacy statistic is the out-of-sample R^2 of the strictly-past k -NN estimate of the residual field under the deployed retrieval (growing buffer, $k = 20$, unweighted neighbor mean, block 64, first 100 points reserved as buffer); every stream is scored in full, so the evaluated count is $n_{\text{test}} - 100$ throughout (the longest stream, MAPS-ROUTING, has $n_{\text{test}} = 59,951$ and evaluates 59,851 points). The rule fires only if $1 - \text{floor} \geq \tau_{\text{room}}$ and $R^2_{\text{causal}} > \tau_{\text{sig}}(1 - \text{floor})$, at the operating point $\tau_{\text{room}} = 0.10$, $\tau_{\text{sig}} = 0$. The reported verdicts hold for every $(\tau_{\text{room}}, \tau_{\text{sig}})$ in $[0, 0.108) \times [0, 0.0616)$, the binding streams being WEATHER (room 0.108) and SBERBANK-HOUSING ($R^2/\text{room} = 0.0616$) respectively; both thresholds are swept on a coarse grid in the archived output, and the exact edges are computed there from the archived R^2 and room columns. Decision-flip fractions, freeze prices, and post-clip wall widths are reported over an assumed-budget grid $\beta \in \{0.01, 0.05, 0.10, 0.25\}$, since β is a prior rather than a measurement. The archived output also records, per classification stream, the variogram floor used for the calibration contrast and the prediction mean and maximum behind the flip-fraction statement.

Reconstruction and bootstrap (Table 3). The reconstruction reimplements the archived streaming k -NN operator (the corrector re-implemented in the reproduction deposit; past-only buffer, $k = 20$ Euclidean neighbors, *unweighted* residual mean, correction $p - 0.7\hat{r}$ applied in the prediction space and clipped to $[10^{-6}, 1 - 10^{-6}]$ for classification, block 64, first 100 points reserved), applies it to the single precomputed prior of each probe stream, and scores % RMSE reduction (regression) or AUC points (classification) against the frozen arm. The streaming pass is run once on the true stream order; the bootstrap resamples evaluation positions in moving blocks of length $4\lceil n_{\text{test}}^{1/3} \rceil$ ($B = 1000$), scoring both arms on the same resample so the interval is for the paired difference, and the reported verdicts are unchanged at block multipliers 2 and 8. The metric-resolution column applies the identical resampling to the frozen arm alone. Deployed gains are *not* recomputed here: the tuned system is a seven-prior ensemble whose per-sample outputs are not redistributed, and the reconstruction is offered as an auditable point estimate of what the local channel alone delivers.

Threshold check. The proven sufficient branch for $\sqrt{N}$ regularity requires $\gamma > \gamma^* = d_0/(2\alpha_s)$, equivalently

$$2\alpha_s\gamma > d_0.$$

The strict reverse inequality is the lower branch, conditional on gap (2c) at $\gamma = 0$ and open for $0 < \gamma < \gamma^*$ (Section 6.4); the equality case is unresolved by the current logarithmic-rate proof. Under the sensitivity inputs $(\alpha_s, \gamma) = (1, 1)$, the check reduces to $d_0 > 2$. This is useful for stress testing but cannot turn d_0 alone into a data-set-specific phase measurement. Below threshold, the candidate upper-rate exponent is $a(1 + \gamma)$ with $a = \alpha_s/(2\alpha_s + d_0)$; its matching scope is stated precisely in Appendix H and Appendix I of Online Appendix 1.

J.2 The proxy and sensitivity display for the twenty-three data sets

The full display is Figure 5 in the main text, with its inputs in Table 4 (Section 6.5). Its filled points are stream-level proxies; its open points are favorable sensitivity scenarios. Both sets lie below the threshold for their stated inputs, but only the former use observed drift labels, and neither set carries uncertainty intervals.

How far below. We record, for each TabReD stream, the smoothness α_s it would need for the $\sqrt{N}$ regime against the measured upper bound: SBERBANK-HOUSING needs $\alpha_s > 1.42$ (measured ~ 0.77), ECOM-OFFERS > 1.90 (~ 0.14), HOMECREDIT-DEFAULT > 2.03 (~ 0.50), and the remaining five streams need > 5.44 – 9.22 against measured values of 0.05 – 0.95 . Even the two lowest- d_0 streams, which are exactly the data sets closest to the boundary, fall short; this is consistent with SBERBANK-HOUSING being the leading empirical example of the earlier probes (largest historical gain, $+4.6\%$; Table 3, historical column; $+3.80\%$ in the reconstruction). The plug-in upper-rate exponent $a(1 + \gamma)$ is small throughout: 0.007 – 0.35 , mostly below 0.15 , with extremes 0.007 (HOMESITE-INSURANCE) and 0.347 (SBERBANK-HOUSING; from unrounded inputs, the printed inputs $d_0 = 3.01$, $\alpha_s = 0.77$, $\gamma = 1.06$ give 0.349). These values quantify the implication of the proxy inputs, not a minimax rate measured from the streams. Appendix I of Online Appendix 1 proves the growing- K negative only conditionally at $\gamma = 0$; it does not justify a blanket claim that additional windows never help these data sets.

If the rough-field bound fails. The bound $\alpha_s \leq 1$ is an assumption (Section 3), and the verdicts above depend on it. For $\alpha_s > 1$ the threshold $\gamma^* = d_0/(2\alpha_s)$ falls, and the required values just listed are exactly the crossing points under the printed d_0 and γ inputs: SBERBANK-HOUSING would enter the sufficient branch at $\alpha_s > 1.42$, ECOM-OFFERS at > 1.90 , HOMECREDIT-DEFAULT at > 2.03 , and the remaining five streams only at > 5.44 – 9.22 . Because the variogram estimator saturates at slope 2, it cannot detect $\alpha_s > 1$, so such a crossing would be invisible to the proxy of item (iii): the map can place a stream below threshold under the bound, but it cannot rule out a smoother drift field that places the two lowest- d_0 streams above it.

J.3 The deduplication correction

Mechanism. Duplicate rows place many points at first-neighbor distance $r_1 \approx 0$ and bias both intrinsic-dimension estimators downward (the mechanism is described in Section 5.3); the artifact is a sibling of the temporal-twin variogram-nugget artifact of Section 5.2, both being duplicate-driven. The consequence for protocol is stated in Section 5.3 and enforced here: intrinsic dimension on FM embeddings must be measured after deduplication. Numerically, UNSW_NB15 moves from $d_0 = 0.87$ to 3.58 and ACS_EMPLOYMENT from 1.62 to 3.11 (Table 7).

Table 7: The deduplication correction: unique fraction at four-decimal precision and raw versus deduplicated MLE d_0 . Upper block: the nine re-measured related benchmarks; two (IEEE_FRAUD, ACS_INCOME) are already duplicate-free at this precision. Lower block: the six high- d_0 benchmarks, essentially duplicate-free, where the shift is a noise bound rather than a correction. The final column computes $d_0/2$, the threshold under the sensitivity value $\alpha_s = 1$; it is not a measured drift threshold.

data set	unique%	d_0 raw (MLE)	d_0 dedup	$d_0/2$ scenario
UNSW_NB15	58.8%	0.87	3.58	1.79
ACS_EMPLOYMENT	84.3%	1.62	3.11	1.56
ACS_UNEMPLOYMENT	93.1%	2.24	3.49	1.75
CICIDS	96.0%	2.86	2.86	1.43
ACS_PUBLICCOVERAGE	98.2%	4.21	4.78	2.39
IEEE_FRAUD	100%	4.63	4.64	2.32
BIKE_SHARING	99.9%	4.88	4.90	2.45
CREDIT_CARD_FRAUD	97.8%	5.09	5.32	2.66
ACS_INCOME	100%	6.47	6.49	3.25
DIABETES_READMISSION	100%	8.77	8.79	4.40
BRFSS_DIABETES	99.97%	9.46	9.46	4.73
SEPSIS	100%	12.75	12.75	6.38
BAF	100%	12.77	12.81	6.41
LENDING_CLUB	100%	13.57	13.55	6.78
MIMIC_IV_MORTALITY	100%	14.02	14.05	7.03

The two artifact cases. The raw readings placed two data sets in or near the $\sqrt{N}$ -learnable regime. Both readings were duplicate inflation.

UNSW_NB15 is 58.8% unique (41% duplicate flows). The raw MLE gives $d_0 = 0.87$; after deduplication $d_0 = 3.58$, giving the favorable-scenario threshold $d_0/2 = 1.79$. The raw value is itself estimator-unstable: a TwoNN pass on the same data gives 2.93. On low-uniqueness data the raw numbers are not reproducible across estimators, which is part of the diagnosis.

ACS_EMPLOYMENT is 84.3% unique. On the raw data, TwoNN gives 0.30 (unstable, with 4.6% zero-1NN pairs), the MLE gives 1.62, the PCA participation ratio is 6.2 (11 PCs for 90% variance), and the readout field is smooth ($\alpha(\eta_0) \approx 1.12$, a finite-sample variogram slope ~ 2.2 , above the limiting value 2; 89% of η_0 in the interior $[0.02, 0.98]$). On that evidence alone the data set could be above threshold: at $d_0 \approx 1.6$ and $\gamma \approx 1$ the requirement is $\alpha_s > 0.81$, plausibly met if drift smoothness tracks readout smoothness. The definitive α_s (which needs the true labels) is not measured, since the effort would be disproportionate for a boundary estimate. Deduplication closes the question from the other side under the same sensitivity inputs: $d_0 = 1.62 \rightarrow 3.11$, so $d_0/2 = 1.56 > 1$. The apparent explanation of the raw reading (a low-complexity binary target concentrating the embedding on a near-1-D discriminative direction) rests on the inflated estimate and does not remain valid after the correction.

Table 7 gives the full re-measured block, including ACS_UNEMPLOYMENT (93.1% unique, $2.24 \rightarrow 3.49$, raw reading borderline) and six higher-uniqueness data sets whose values move little. The lower block extends the identical protocol to the six high- d_0 benchmarks, on which the correction is inert: all six are 99.97–100% unique at four decimals (five carry literally no duplicate rows), so for them the raw-versus-deduplicated difference is re-subsampling noise, which bounds the deduplication effect itself by 0.10 (against the same-run raw re-estimate archived with the verification code; against the raw values of the measurement record the shift is at most 0.04).

Table 8: Deduplication robustness check on the eight TabReD streams (from the measurement record). All streams are essentially duplicate-free, so their intrinsic-dimension inputs are insensitive to this correction. Columns are rounded independently, so a printed raw/dedup pair can differ from the printed shift in the last digit.

data set	unique%	d_0 raw	d_0 dedup	shift
ECOM-OFFERS	99.9%	2.95	2.94	0.00
SBERBANK-HOUSING	99.9%	3.06	3.08	+0.02
WEATHER	99.8%	11.06	11.13	+0.07
COOKING-TIME	100%	10.64	10.57	−0.07
HOMESITE-INSURANCE	100%	10.40	10.34	−0.06
HOME CREDIT-DEFAULT	100%	12.23	12.31	+0.08
DELIVERY-ETA	100%	14.82	14.85	+0.03
MAPS-ROUTING	100%	15.62	15.58	−0.04

The TabReD-8 streams pass the same check. Applying the identical protocol (unique rows at four decimals, recompute MLE d_0 on the unique subset, 12,000-row subsample) to the eight TabReD streams gives Table 8: all are 99.8–100% unique, every shift is at most 0.08 in absolute value, and the minimum deduplicated d_0 is 2.94. Thus the dimension input for the TabReD sensitivity check is duplicate-robust; only the low-uniqueness related data sets (network flows, census microdata) were materially affected. Note that this check uses the MLE estimator, whose raw point values differ from the TwoNN values of Table 4 (for example DELIVERY-ETA 14.82 MLE against 16.63 TwoNN). That is an estimator difference, not a data difference; the quantity the check certifies is the *deduplication shift*, which is at most 0.08 in absolute value everywhere.

Corrected conclusion. Deduplication removes the apparent low-dimensional examples: the lowest corrected estimates in the re-measured block are CICIDS at 2.86 and ACS_EMPLOYMENT at 3.11, while the minimum TabReD MLE is 2.94. Consequently none of these rows enters the regular branch under the specific sensitivity values $(\alpha_s, \gamma) = (1, 1)$. This does not show that their true drift fields are below threshold, because neither input is measured for the related benchmarks. It does rule out reading ACS_EMPLOYMENT or UNSW_NB15 as an observed $\sqrt{N}$ -learnable demonstration data set.

J.4 Reading the display

Raw versus deduplicated readings. The raw and deduplicated columns of Table 7 differ materially on three data sets only (UNSW_NB15, ACS_EMPLOYMENT, ACS_UNEMPLOYMENT), all of low uniqueness; on the twelve others the two readings agree to within subsample noise. The raw readings placed the first two in or near the regular branch under the sensitivity inputs and the third at its border; the deduplicated readings place all three below it. Only the deduplicated column enters Table 4 and Figure 5.

The dimension input drives the threshold spread. Under the common sensitivity value $\alpha_s = 1$, the threshold input $d_0/2$ ranges from 1.43 (CICIDS) to 7.03 (MIMIC_IV_MORTALITY). This spread is driven by intrinsic dimension. Interpreting it as wall height would additionally require measuring the drift smoothness, so the display supports comparative stress testing rather than a ranked empirical ordering of walls.

The ACS family gradient. The cleanest cross-data-set pattern is the controlled sweep within the ACS census family (same data, same representation, same shift, varying only the prediction target). The deduplicated d_0 rises with target complexity: employment 3.11, unemployment 3.49, public coverage 4.78, income 6.49. The ordering (income > coverage > unemployment > employment) already held in the raw (pre-deduplication) values and persists after the correction; what the correction removed is only the reading that the low end of the sweep crosses the boundary. Under a common (α_s, γ) scenario this ordering also orders the threshold inputs. It is a descriptive pattern, not an independent test of the threshold of Theorem 6.6.

Reliability of the inputs. The deduplicated d_0 values and their cross-data-set ordering are the strongest part of the record, although estimator sensitivity remains. The TabReD α_s and γ values are noisy proxies from a single residual field, with two-range slope disagreement on five of eight streams and upward bias from k -NN smoothing. For the related benchmarks, both drift quantities are unobserved. A phase-validation study would need repeated labeled windows, uncertainty intervals for all three inputs, and a pre-specified rule for propagating them through $\gamma^* = d_0/(2\alpha_s)$.