UNIFORM ENSEMBLES UNDER DEPENDENCE: CALIBRATION AND PROTECTED REPLACEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Temporal dependence does not by itself justify replacing uniform bagging. We separate the information available for learning, the intervention that changes an ensemble, and the evidence needed to approve that change. Uniform-bootstrap voting attains the minimax rate on a covariate-refresh family, while balanced spectral routing is exactly midpoint blocking on weighted paths, decreasing temporal kernels, and sufficiently large band graphs. For frozen replacements, independent-label certification has matching fixed-confidence cost $\Theta(\omega/\Delta^2)$ in disagreement ω and improvement Δ. A multiscale martingale certificate instead protects average conditional improvement along a dependent audit history. Across four chronological tasks, full residual stacking has the lowest mean later error on three, while graph projection improves on full stacking on none. Rolling-origin production evaluation records 23 approved replacements, including 2 forward-error reversals. A gradual within-audit drift experiment explains how a positive historical average can coexist with a harmful current predictor. Restarting at a predeclared recent window removes that outdated evidence but loses useful power under stable conditions. The resulting guarantee concerns the chosen audit window; future transfer still requires separate evidence or restrictions on drift.

1 INTRODUCTION

A dependent trajectory can contain much less information than its length suggests. Every ensemble trained on it shares that limitation. Adding bootstrap voters reduces their conditional randomization error, but cannot remove uncertainty in the common training sample. Whether uniform bagging wastes the available information is a different question. So is whether a proposed replacement improves prediction, and whether the audit budget can detect that improvement. These distinctions matter when a graph partition, a covariance reduction, or an impressive voter count is used to justify deployment. Figure 1 summarizes these distinctions.

We first establish what uniform voting can attain. On a reversible refresh chain with d observable feature cells, a finite cellwise bootstrap vote matches the minimax excess-risk order $\min\{1, \sqrt{dT/n}\}$, where T is the refresh time. For general finite-VC classes containing the Bayes classifier, bootstrap empirical-risk minimization inherits a blocked absolute-regularity rate. This broader result permits noisy labels, but does not assert sharp agnostic approximation error or that practical tree fitting is exact empirical-risk minimization. Dependence therefore creates an information limit without necessarily creating an additional uniform-resampling penalty.

Next, we ask whether the proposed intervention is different from a simpler one. Spectral seriation can recover an ordering from an affinity matrix (Atkins et al., 1998; Fogel et al., 2016). When chronological order is already known, the same mechanism can make a routing operation redundant. We prove exact equality with midpoint blocking on positive weighted paths and strictly decreasing distance kernels. For width-w band graphs, both combinatorial and normalized balanced Fiedler splits have this equality whenever the parent interval has at least $2w + 1$ vertices. Matched allocations and learner randomization then imply identical predictors, including nonlinear fits and hard votes. This equality gives a constructive deduplication rule: an ensemble search need not pay for a spectral candidate that is pathwise identical to its contiguous control.

The empirical intervention is simpler: fit residual weights on the uniform ensemble's existing heads, then compare the thresholded classifier with its baseline. Independent-label certification costs

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

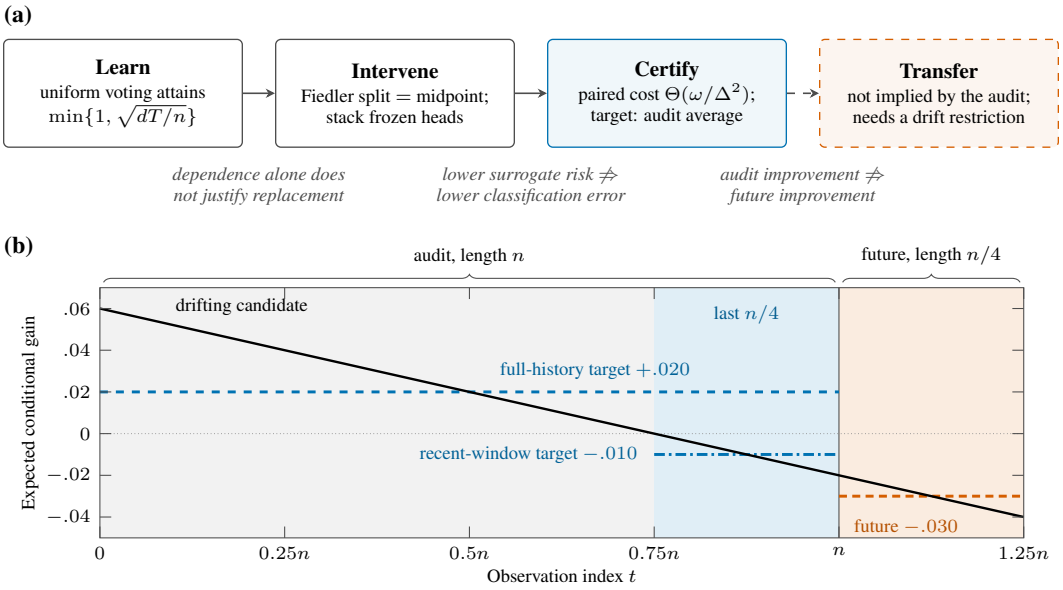


Figure 1: (a) Each stage establishes a different target, and a guarantee at one stage does not carry to the next. Uniform voting attains the dependence-limited rate (Section 3); temporal spectral routing reduces to midpoint blocking, and residual stacking reuses the uniform heads (Sections 4 and 5.1); certification protects a declared audit target (Sections 5.2 and 5.3); transfer to later data requires separate evidence (Appendix K). (b) Expected conditional gain $\mathbb{E}[\omega_t]\,\theta_{1,t}$ of the drifting candidate in the within-audit drift experiment, with $\mathbb{E}[\omega_t] = .20$. The full-history average remains positive after the current advantage has reversed: at $n = 16{,}000$, the full-history rule approves in 1,870 of 2,000 trajectories and every approval is future-harmful. A predeclared recent-window restart certifies the negative last-quarter average instead and makes no future-harmful deployment (Table 24).

$\Theta(\omega/\Delta^2)$ at fixed confidence. A self-normalized temporal certificate instead protects average conditional improvement along the audit history, without a fitted mixing envelope. The four chronological tasks favor full stacking on three later periods and expose its failure on housing. Rolling-origin evaluation then distinguishes a single successful suffix from repeated transfer, and gradual drift within the audit tests whether old improvements can conceal a harmful current predictor. A fixed recent-window restart addresses this mismatch by changing the certified target, at the cost of fewer labels. Its window and error allocation are specified before outcomes are inspected.

2 Related Work

Learning from dependent observations uses mixing, blocking, and empirical-process bounds (Yu, 1994; Kuznetsov & Mohri, 2017; Paulin, 2012). Bagging is already an optimal i.i.d. realizable PAC learner (Larsen, 2023); robust-ERM bagging gives related VC-linear adversarial-learning guarantees (Montasser, 2026). These results address different sampling models: our VC bound allows noisy labels and absolute regularity under Bayes containment, whereas the refresh family identifies a matching dependence-specific classification benchmark. Stationary and circular block bootstraps are established dependence-aware competitors (Politis & Romano, 1991; 1994). Margin and PAC-Bayesian analyses control a learned vote through its moments or posterior (Germain et al., 2015; Masegosa et al., 2020; Bastian et al., 2026); our final comparisons condition on already fitted classifiers.

Spectral partitioning and seriation use Laplacian eigenvectors to recover groups or orderings (Fiedler, 1973; Von Luxburg et al., 2008; Atkins et al., 1998; Fogel et al., 2016). The ordering theory motivates our exact allocation comparison, including a sharp strict-ordering threshold for bands and a normalized-Laplacian extension. The difference-operator argument makes partition equality explicit for matched nonlinear learners, rather than only recovering an ordering. Residual weighting likewise builds on stacking, forecast combination, and covariance regularization (Wolpert, 1992; Breiman,

1996; Bates & Granger, 1969; Ledoit & Wolf, 2004). The empirical question is whether routing or graph compression improves the simpler uncompressed weighting procedure under matched budgets.

Blocked and hv-block validation address dependent performance estimation (Racine, 2000; Bergmeir & Benítez, 2012; Cerqueira et al., 2019). Our independent safeguard uses exact discordant-pair testing, standard multiplicity control, and matched-pair power planning (Mcnemar, 1947; Holm, 1979; Fagerland et al., 2013; Connor, 1987). Testing-based risk control supplies the broader framework (Angelopoulos et al., 2025). For arbitrary temporal dependence, our conditional-risk certificate specializes established self-normalized supermartingale bounds and confidence sequences (Howard et al., 2018; 2021). Recent empirical-Bernstein confidence sequences also target average conditional means (Chugg & Ramdas, 2025); we use a finite scale grid for a directly computable comparison on the same discordant counts. Its advantage is a justified audit target without a fitted mixing constant; its limitation is that this target does not automatically describe the subsequent deployment period.

3 Information Limits and Uniform Voting

Let $Z_t = (X_t, Y_t)$ be stationary with $Y_t \in \{-1, 1\}$, and let an independent test pair follow its stationary law π. Write $\eta(x) = \mathbb{E}[Y \mid X = x]$, $h^*(x) = \operatorname{sign} \eta(x)$, and resolve prediction ties toward $+1$. For a fitted classifier,

$$R(H) = \Pr\{H(X) \neq Y\}, \qquad \mathcal{E}(H) = R(H) - R^* = \mathbb{E}_X[|\eta(X)|\mathbf{1}\{H(X) \neq h^*(X)\}]. \quad (1)$$

Outer expectations also average over training and algorithmic randomization. A hard vote thresholds binary predictions; a soft vote thresholds real-valued scores. These are different learners.

3.1 A sharp covariate-refresh benchmark

Under π_θ, X is uniform on $[d]$ and $\Pr(Y = 1 \mid X = x) = 1/2 + \theta_x$, where $\theta \in [-1/4, 1/4]^d$. For $T \geq 1$, refresh the entire observed pair:

$$P_{\theta,T}(z, \cdot) = (1 - T^{-1})\delta_z + T^{-1}\pi_\theta, \qquad Z_1 \sim \pi_\theta. \quad (2)$$

This reversible chain has relaxation time T and worst-start total-variation mixing time comparable to T. The Bayes signs vary freely across cells.

Theorem 1 (Feature-dependent minimax benchmark)**.** *For universal constants $c, C > 0$ and every integer $n, d \geq 1$ and real $T \geq 1$,*

$$c \min\{1, \sqrt{dT/n}\} \leq \inf_{\widehat{H}} \sup_{\theta} \mathbb{E}_{\theta,T}\mathcal{E}(\widehat{H}) \leq C \min\{1, \sqrt{dT/n}\}. \quad (3)$$

The empirical cell sign $\operatorname{sign} \sum_t Y_t \mathbf{1}\{X_t = x\}$ attains the upper bound without knowing T.

Revealing refresh indicators gives $1 + \mathrm{Bin}(n-1, 1/T)$ independent covariate-label draws. Assouad's comparison across cell signs gives the lower bound. For the upper bound, $(d/n)\sum_t Y_t \mathbf{1}\{X_t = x\}$ is unbiased for $2\theta_x$ with variance at most $2dT/n$; a wrong sign costs at most its absolute estimation error. Appendix A includes empty cells and $n < dT$.

Theorem 2 (Finite cellwise uniform voting)**.** *Each of m conditionally independent voters takes the majority of an odd number b of labels drawn uniformly with replacement from the query cell. Empty cells predict $+1$. Their hard vote satisfies*

$$\sup_{\theta} \mathbb{E}\mathcal{E}(H_{m,b}) \leq \min\left\{\frac{1}{2}, \sqrt{\frac{2dT}{n}} + \frac{1}{\sqrt{m}}\right\}. \quad (4)$$

Thus $m \geq n/(dT)$ suffices for the minimax order when $n \geq dT$.

An odd bootstrap majority preserves the empirical cell sign and amplifies its conditional mean. Weighting its finite-voter fluctuation by empirical cell occupancy gives the $m^{-1/2}$ term without dividing by a small random count (Appendix B). This is statistical attainability; the empirical cell-sign learner is cheaper.

The observable-cell construction is not the only attainable case. A uniform-bootstrap ERM vote over a finite-VC class containing the Bayes rule also attains a blocked absolute-regularity rate, with noisy labels allowed (Theorem 7). Appendix C gives the statement and proof; it retains the approximation-error factor under misspecification rather than claiming an agnostic optimal rate.

4 When Spectral Routing Is Only Blocking

Routing changes the training allocation. The next result identifies graph families for which that allocation is already determined by chronological order.

Theorem 3 (Temporal Fiedler routing equals midpoint blocking)**.** *Consider a nearest-neighbor path with arbitrary positive edge weights, or a complete graph with $W_{ij} = \psi(|i-j|)$ for $i \neq j$, where $\psi(1) > \cdots > \psi(n-1) > 0$. For the combinatorial Laplacian, every induced interval of size $s \geq 2$ has a simple Fiedler eigenvalue and a strictly monotone Fiedler vector up to sign. Orient it positive at the left endpoint and assign its largest $\lceil s/2 \rceil$ coordinates to the left child. Recursive balanced bisection gives midpoint intervals. The same partition conclusion holds for the symmetric normalized Laplacian of an unweighted path. Identical head allocations, resampling seeds, fitting, and aggregation imply identical spectral and contiguous predictors for every training realization.*

The first-difference operator satisfies $D_\nabla L = M D_\nabla$, with M irreducible and nonpositive off-diagonal. Perron–Frobenius makes the Fiedler differences strictly one-signed. This is the ordering mechanism behind spectral seriation, specialized to a pathwise ensemble comparison (Atkins et al., 1998; Fogel et al., 2016). Appendix D.1 gives the proof and the normalized-path argument.

Corollary 4 (Band graphs and the strict-ordering threshold)**.** *Let $W_{ij} = \mathbf{1}\{0 < |i-j| \leq w\}$ on an interval of size s, with integer $w \geq 1$. If $s \geq 2w+1$, the combinatorial and normalized Fiedler eigenvalues are simple. Their combinatorial and generalized random-walk eigenvectors are strictly monotone up to sign. The symmetric normalized Fiedler vector has a positive left half and negative right half, with a central zero when s is odd. Both balanced median splits are midpoint splits. Recursive equality with contiguous routing holds whenever every split parent satisfies this size condition. For $3 \leq s < 2w+1$, strict ordering can fail through a repeated eigenvalue or equal central coordinates.*

The band difference operator has backward links when $i \geq w+1$ and forward links when $i \leq s-w-1$. Their digraph is strongly connected at $s \geq 2w+1$. Degree normalization preserves the required off-diagonal signs; reflection symmetry supplies the midpoint sign split even when normalized coordinates are not globally monotone (Appendix D.2). The size threshold concerns parents being split, not the final leaf sizes. This explains the small-interval exceptions in the recursive audit rather than excluding band graphs altogether.

These results are not a no-gain theorem for unrestricted graphs. On four vertices, edges $(1,3)$ and $(2,4)$ of weight one, with all other edges of weight $\varepsilon \in (0,1)$, yield the simple noncontiguous Fiedler split $\{1,3\} \mid \{2,4\}$ under both Laplacians. Conversely, an index graph cannot identify a transition-mixing parameter, and balanced linear regrouping preserves the full-sample mean. Appendix D separates these obstructions from genuine weighting opportunities caused by heterogeneous residuals.

5 Calibrating and Certifying a Replacement

5.1 Residual weighting on frozen uniform heads

Training fixes scores $f : \mathcal{X} \to [-1,1]^m$. A disjoint calibration segment estimates $Q = \mathbb{E}[(f(X) - Y\mathbf{1})(f(X) - Y\mathbf{1})^\top]$ by $\widehat{Q}$. For simplex weights,

$$w^\top Q w = \mathbb{E}[(w^\top f(X) - \eta(X))^2] + \nu, \qquad \nu = \mathbb{E}[1 - \eta(X)^2]. \tag{5}$$

Full stacking minimizes $w^\top \widehat{Q} w$ subject to $w \geq 0$, $\mathbf{1}^\top w = 1$, and $\|w\|_2^2 \leq 2/m$. Graph stacking replaces $\widehat{Q}$ by $K = P\widehat{Q}P$ and minimizes $w^\top K w + d\|w\|_2^2$, where the training-fixed projector P has discrepancy $d = \|K - \widehat{Q}\|_{\mathrm{op}}$. Both reuse identical heads and calibration labels. The residual oracle in Proposition 10 is a surrogate statement, not evidence that projection helps or that classification improves. Every resulting thresholded classifier is frozen before the final comparison.

5.2 Independent and mixing-adjusted certification

Let H_0 be uniform bagging and $H_1, \ldots, H_J$ its frozen alternatives, with $J \geq 1$. For this subsection, audit data have a common target law and are independent of candidate construction; a separated

continuation requires the additional gap term below. On the same k audit observations, let B_r count baseline errors corrected by H_r, C_r count new errors, and $D_r = B_r + C_r$. The exact paired tail is

$$p_r = \Pr\{\mathrm{Bin}(D_r, 1/2) \geq B_r\}, \qquad p_r = 1 \text{ if } D_r = 0. \tag{6}$$

Approve $p_r \leq \delta/J$, select the approved candidate with largest empirical gain, and otherwise use H_0. Ties follow a fixed index order. Independent audit observations give coupling cost $\gamma = 0$. A stationary absolutely regular trajectory, independent of construction and thinned at lag a, gives $\gamma = (k-1)\beta(a)$.

Theorem 5 (Paired protection and tight fixed-confidence cost)**.** *The probability that any approved candidate has risk at least $R(H_0)$ is at most $\delta + \gamma$. A candidate with $\Delta = R(H_0) - R(H_r) > 0$ and $\omega = \Pr(H_r(X) \neq H_0(X))$ qualifies with probability at least $1 - \epsilon - \gamma$ if*

$$k \geq \frac{32\omega}{\Delta^2} \max\left\{\log\frac{J}{\delta}, \log\frac{2}{\epsilon}\right\}. \tag{7}$$

Conversely, for $0 < \Delta \leq \omega/2$ and $\delta_0 + \epsilon_0 < 1$, beneficial and harmful laws with the same classifiers and disagreement require any rule having harmful-deployment probability at most δ_0 and useful-deployment probability at least $1 - \epsilon_0$ to use

$$k \geq \frac{\omega}{4\Delta^2} \max\{\mathrm{kl}(1-\epsilon_0, \delta_0), \mathrm{kl}(1-\delta_0, \epsilon_0)\} \tag{8}$$

independent labels. Here $\delta, \epsilon, \delta_0, \epsilon_0 \in (0,1)$ and kl is binary relative entropy.

Only disagreements reveal the risk difference. The paired loss difference has mean Δ and second moment ω; reversing its conditional advantage gives single-observation KL at most $4\Delta^2/\omega$. Exact testing, concentration, and data processing prove the result (Appendix F). The fixed-family, fixed-confidence cost is therefore $\Theta(\omega/\Delta^2)$, without a matching growing-J claim. For construction and audit on one sequence, add the construction-to-audit coupling $\beta(g)$ and require $\delta + \gamma \leq \alpha$. An unknown mixing envelope cannot be replaced by a nominal p-value.

5.3 A TEMPORAL CERTIFICATE WITHOUT A MIXING ENVELOPE

When independence cannot be justified, a different target remains accessible. Let $\mathcal{F}_0$ contain all construction and calibration information, and let $\mathcal{F}_t$ additionally contain the first t revealed audit examples. Define

$$U_{r,t} = \mathbf{1}\{H_0(X_t) \neq Y_t\} - \mathbf{1}\{H_r(X_t) \neq Y_t\}, \qquad \overline{\Delta}_{r,t} = \frac{1}{t}\sum_{i=1}^{t} \mathbb{E}[U_{r,i} \mid \mathcal{F}_{i-1}]. \tag{9}$$

This is average conditional improvement over the audited history. It equals the population risk difference for an independent common target law, but need not equal a stationary or future-period risk difference under temporal dependence.

Theorem 6 (Simultaneous conditional-risk certificate)**.** *Let all candidates be $\mathcal{F}_0$-measurable and let the audit be adapted, with arbitrary dependence. Fix a nonempty finite $\Lambda \subset (0,1)$, $G = |\Lambda|$, $\alpha \in (0,1)$, and $\psi(\lambda) = -\log(1-\lambda) - \lambda$. With probability at least $1 - \alpha$, simultaneously for every $r \in [J]$ and $t \geq 1$,*

$$\overline{\Delta}_{r,t} \geq \underline{\Delta}_{r,t} := \frac{1}{t}\max_{\lambda \in \Lambda}\left\{S_{r,t} - \frac{\psi(\lambda)}{\lambda} D_{r,t} - \frac{\log(JG/\alpha)}{\lambda}\right\}, \tag{10}$$

where $S_{r,t} = \sum_{i \leq t} U_{r,i}$ and $D_{r,t} = \sum_{i \leq t} U_{r,i}^2$. Approving only positive lower bounds protects every approved candidate for this conditional-risk target.

The inequality $\exp\{\lambda u - \psi(\lambda)u^2\} \leq 1 + \lambda u$ for $u \in \{-1, 0, 1\}$ gives the requisite supermartingale. Ville's inequality and a union bound over candidates and scales prove Equation (10), a specialization of self-normalized confidence-sequence methods (Howard et al., 2018; 2021). We use $\Lambda_9 = \{2^{-j} : j = 1, \ldots, 9\}$ and decide at the fixed audit endpoint. This includes $1/512$ and guarantees the sufficient scale coverage when $\Delta/\omega \geq 1/128$ (Corollary 12). The correction is $\log(9J/\alpha)$; extra scales require no additional labels, but do incur a multiplicity cost. Later-period transfer is a separate target, evaluated in Appendix K.

For a predeclared lookback h at audit endpoint n, the same supermartingale can restart at $n - h$ and use only the last h observations. This recent-window rule certifies their average conditional improvement, not the full historical average. A single fixed restart uses the same confidence level; several starts require an explicit allocation across restarts (Proposition 14). We evaluate the full and recent rules separately, with identical frozen candidates and the same nine scales.

6 EXPERIMENTS

6.1 DISAGREEMENT, SCALE COVERAGE, AND CERTIFICATION POWER

The known-risk comparison fixes thirteen classifiers, with baseline risk .30, one gain .01, five zero gains, and six gains $-.01$. All rules use identical observations in each of $20,000$ independent repetitions. At $k = 8,000$, exact paired useful-deployment probability decreases from .9078 at disagreement .05 to .1046 at disagreement .40. The nine-scale temporal probabilities are .3287 and .0046; its all-null approval count is 7/20,000, and every gain-.08 positive control passes. Complete counts distinguish useful, harmful, equal-risk, and fallback decisions (Tables 5 and 21). Exact paired 90% qualification budgets have normalized cost $k\Delta^2/\omega$ between 15.39 and 15.61 across the gain/disagreement sweep, consistent with the matching bound.

A wider scale grid does not uniformly increase finite-budget power. At $\Delta = .01$, $\omega = .40$, the five-scale grid misses the sufficient interval $[\Delta/(8\omega), \Delta/(4\omega)]$, while Λ_9 includes $1/256$. Exact marginal power rises from 0.8516 to 0.9554 at $k = 128,000$, and the attaining 90% budget decreases from 149,304 to 111,337. At $k = 8,000$, however, power falls from 0.0074 to 0.0046: the additional small scales do not determine the boundary there, but the $\log(9/5)$ correction still applies. Table 22 reports this tradeoff without attributing the entire low-budget failure to grid truncation.

A predictable-regime stress test varies disagreement between .04 and .36. Across $2,000$ length-$4,000$ trajectories per setting, temporal useful-deployment frequencies are 0.9360, 0.9195, and 0.8775 at regime refresh times 1, 50, and 200, with no harmful endpoint deployment. The largest anytime coverage-failure frequency is 0.0025(Table 23). Separately, nine covariate-refresh settings give normalized finite-vote excess risk between .1564 and .1791 (Table 8). The first study tests certification under changing conditional means; the latter tests the training information rate.

6.2 TEMPORAL ROUTING AND THE LIMITS OF GRAPH PROJECTION

The band threshold explains the recursive spectral audit. All 51 noncontiguous outcomes occur on intervals of sizes 7, 8, or 15 below $2w + 1$, with a tied median or repeated Fiedler eigenvalue. All 720 exponential-kernel splits are contiguous (Table 9). A separate width-1 through width-16 audit checks 905 eligible band intervals per Laplacian: every combinatorial and normalized split is the midpoint split, and every generalized Fiedler vector is strictly monotone. The maximum eigenpair residuals are 2.18×10^{-14} and 1.57×10^{-15}, respectively (Appendix H.4).

A learned graph has a small, controlled weighting benefit, but not a demonstrated real-data role. In the heterogeneous-score model, a rank-two graph learned from 256 additional unlabeled score vectors gives excess risk .00731 at $k_c = 128$, compared with .00830 for full stacking and .00851 for a matched-penalty control. The paired learned-minus-full interval is $[-.001208, -.000773]$ over 100 pilot/calibration repetitions on a fixed test-feature pool (Table 10). The model is constructed so score similarity exposes residual groups. This result neither removes the extra pilot cost nor establishes a benefit on naturally occurring series.

6.3 STACKING, CERTIFICATION, AND CHRONOLOGICAL TRANSFER

We evaluate next-step increases in weekly Mauna Loa CO_2, hourly Victoria electricity demand, monthly U.S. meat production, and monthly housing sales across Texas MLS areas (Keeling & Whorf, 2004; statsmodels Developers, 2026a; Plotnine Contributors, 2026; ggplot2 Contributors, 2026). The production and housing panels supply 2,268 and 2,615 actual audit labels, rather than repeated draws from a small pool. Features use current levels, five backward differences, calendar phases, and panel entity indicators. Every routing or resampling family fits 64 trees with matched row-presentation budgets; full and graph stacking reuse the uniform heads. Stationary and circular

Table 1: Later-period classification error, mean ± standard deviation over ten learner seeds on each fixed chronological split. Every fitted family has 64 trees and matched row-presentation budgets within its task; both stacking candidates reuse the uniform heads. These deviations measure learner randomization, not uncertainty over independent environments.

Method	Climate	Electricity	Production	Housing
Uniform bagging	$.2800 \pm .0055$	$.3386 \pm .0286$	$.3239 \pm .0081$	$.2323 \pm .0118$
Contiguous / path, $P = 4$	$.2930 \pm .0095$	$.3473 \pm .0282$	$.3391 \pm .0117$	$.2266 \pm .0048$
Random balanced, $P = 4$	$.2812 \pm .0062$	$.3391 \pm .0272$	$.3246 \pm .0088$	$.2365 \pm .0071$
Contiguous / path, $P = 8$	$.2957 \pm .0132$	$.3463 \pm .0297$	$.3405 \pm .0206$	$.2550 \pm .0147$
Random balanced, $P = 8$	$.2858 \pm .0092$	$.3238 \pm .0168$	$.3242 \pm .0159$	$.2389 \pm .0122$
Stationary bootstrap	$.2827 \pm .0062$	$.3480 \pm .0220$	$.3305 \pm .0120$	$.2383 \pm .0121$
Circular block bootstrap	$.2853 \pm .0130$	$.3407 \pm .0299$	$.3237 \pm .0069$	$.2380 \pm .0208$
Full residual stacking	$.2742 \pm .0062$	$.3133 \pm .0180$	$.3062 \pm .0127$	$.2519 \pm .0118$
Graph residual stacking	$.2793 \pm .0061$	$.3311 \pm .0205$	$.3200 \pm .0127$	$.2555 \pm .0131$

Table 2: Actual-budget replacement decisions. Ref. is the exact-binomial independent-observation reference, not a temporal certificate. Temporal is Equation (10) at $\alpha = .05$ for average conditional audit improvement. Both counts are seeds with any approved replacement out of ten; deployed error includes uniform fallback. Later errors are mean ± standard deviation over learner seeds on fixed suffixes.

Task	Audit labels	Ref.	Temporal	Temporal later error
Climate	309	$0/10$	$0/10$	$.2800 \pm .0055$
Electricity	533	$5/10$	$0/10$	$.3386 \pm .0286$
Production	$2,268$	$9/10$	$7/10$	$.3090 \pm .0111$
Housing	$2,615$	$0/10$	$0/10$	$.2323 \pm .0118$

bootstraps preserve monthly cross-sections in the panels. Training, calibration, audit, and later test periods are chronological and separated by six-observation or six-month embargoes. No later labels determine a candidate, split, or decision. Appendix I gives the panel construction; the two smaller annual/monthly series appear only in Appendix H.7.

Full residual stacking has the lowest mean later error on climate, electricity, and production, but worsens housing error in all ten learner seeds (Table 1). Graph projection is worse than full stacking on every task. Contiguous routing hurts the first three tasks; its four-group variant has the best housing mean but no qualifying audit advantage. Block bootstraps do not consistently improve uniform bagging. These comparisons support full stacking as a useful candidate, not an unconditional replacement or a role for graph projection.

The temporal certificate yields a positive, target-specific result on production (Table 2). Full stacking qualifies in 6/10 seeds. The final selector deploys full stacking in 6 seeds, stationary bootstrap in 1, and uniform in the other 3. All 7 actual replacements have lower error than their matched baseline on the untouched later period. Mean deployed error is .3090, versus .3239 for uniform; raw full stacking reaches .3062. The certification concerns average conditional audit improvement, while this later comparison establishes observed transfer on one fixed suffix. The ten seeds measure learner randomization, not ten independent environments.

Housing supplies the necessary failure case. Full stacking increases later error from .2323 to .2519, while every temporal decision retains uniform. Climate and electricity also have no positive temporal lower bound. Electricity's $5/10$ exact-binomial approvals remain independent-observation reference decisions, not mixing-adjusted certificates. On climate, a 309-label audit needs gain $.0704$ for 90% exact paired power at disagreement $.10$ with $.025$ testing budget and eight alternatives; the largest observed gain is $.0129$ (Table 14). Larger nominal resampling counts cannot close this information gap. The finite-pool comparison is therefore isolated in Appendix H.10.

Within-audit drift makes 1,870/2,000 full-history approvals future-harmful at $n = 16,000$; a predeclared last-quarter restart has no future-harmful deployment, but stable-control useful deployment falls from 1.0000 to 0.9210 (Table 24). On seven fixed production origins, expanding-history certification approves 23/70 seed-origin decisions with 2 forward-error reversals, whereas the recent-window

rule with a simultaneous restart correction retains uniform throughout, exposing the corresponding recency–power tradeoff on real data (Table 25).

7 Discussion and Conclusion

Uniform voting can attain a dependence-limited learning rate without eliminating shared training uncertainty, and temporal spectral routing can be exactly redundant with blocking. In the four chronological tasks, full residual stacking improves three later-period means but fails on housing; graph projection does not improve on it. A safeguard must therefore assess the actual thresholded replacement rather than its surrogate objective or geometric description. Paired certification has tight independent-label cost in ω/Δ^2, while the temporal extension protects an explicitly history-conditional target. The transfer comparisons identify a limitation within that target, not only after an independent change of regime: early gains can dominate the audit average after the current advantage has reversed. A fixed recent-window restart removes those early gains but spends fewer labels, and our stable controls and rolling origins show the resulting loss of power. These are separate rules, not an audit-adaptive search over favorable windows. The real studies use fixed environments and learner seeds, so their forward-error counts do not estimate a universal population transfer failure rate. A useful replacement requires evidence for the intended time window and an additional justification for any claim about future risk.

Ethics Statement

This work complies with the ICLR Code of Ethics. It involves no human subjects, crowdsourcing, or newly collected data. All empirical analyses use publicly distributed datasets: the Mauna Loa CO_2 record, Victoria electricity demand, NOAA Niño 1+2 sea-surface temperatures, and annual sunspot counts distributed with statsmodels, and U.S. meat production and Texas housing sales distributed with plotnine. The housing panel contains monthly sales counts aggregated to multiple-listing-service areas. It includes no individual transactions, prices, or personally identifiable information.

The main risk we foresee is misinterpretation of a certificate. An approved replacement is protected for average conditional improvement over the declared audit window, not for future deployment. Our within-audit drift experiment and the two rolling-origin reversals show that treating the audit guarantee as a forecast can deploy a harmful predictor. Anyone applying these procedures in consequential settings, such as energy or housing-market forecasting, should state the certified window explicitly and obtain separate evidence, or restrictions on drift, before claiming future benefit. The certificates also concern aggregate binary loss. They do not assess whether a replacement shifts errors across subpopulations, so decisions affecting people would require a separate group-level evaluation. We report failure cases, including the housing panel, alongside successes to support this reading.

Reproducibility Statement

The appendices give complete proofs, chronological splits, input identities, candidate budgets, optimization diagnostics, and per-seed paired counts. Embedded programs and numerical records reproduce the controlled comparisons, panel analyses, multiscale certification, within-audit drift, and production rolling origins. The latter analyses save the fixed window definitions, restart budgets, decisions, and separate audit and future targets. The two panel datasets were recomputed using the stated model specifications before the rolling-origin analysis.

AI Use Statement

An AI assistant contributed to literature retrieval, mathematical development and checking, experimental design, code development and execution, numerical analysis, and manuscript writing. This assistance included the band-graph extension, temporal certification analysis, panel computations, multiscale power calculations, within-audit drift experiment, and rolling-origin evaluation. The authors retain responsibility for the scientific claims, source attribution, and submitted material.

REFERENCES

Anastasios N Angelopoulos, Stephen Bates, Emmanuel J Candès, Michael I Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025.

Jonathan E Atkins, Erik G Boman, and Bruce Hendrickson. A spectral algorithm for seriation and the consecutive ones problem. *SIAM Journal on Computing*, 28(1):297–310, 1998.

Julien Bastian, Benjamin Leblanc, Pascal Germain, Amaury Habrard, Guillaume Metzler, Emilie Morvant, and Paul Viallard. On the disintegration of the stochastic majority vote: From pac-bayesian bounds to a self-bounding algorithm, 2026. URL https://arxiv.org/abs/2609.16803.

John M Bates and Clive WJ Granger. The combination of forecasts. *Journal of the operational research society*, 20(4):451–468, 1969.

Christoph Bergmeir and José M. Benítez. On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191:192–213, 2012. ISSN 0020-0255. doi: https://doi.org/10.1016/j.ins.2011.12.028. Data Mining for Software Trustworthiness.

Leo Breiman. Stacked regressions. *Machine learning*, 24(1):49–64, 1996.

Vítor Cerqueira, Luís Torgo, and Igor Mozetič. Evaluating time series forecasting models: an empirical study on performance estimation methods. *Machine Learning*, 109:1997 – 2028, 2019.

Ben Chugg and Aaditya Ramdas. Closed-form empirical bernstein confidence sequences for scalars and matrices. *arXiv preprint arXiv:2512.21300*, 2025.

Robert J. Connor. Sample size for testing differences in proportions for the paired-sample design. *Biometrics*, 43 1:207–11, 1987.

Morten Wang Fagerland, Stian Lydersen, and Petter Laake. The mcnemar test for binary matched-pairs data: mid-p and asymptotic are better than exact conditional. *BMC Medical Research Methodology*, 13:91 – 91, 2013. doi: 10.1186/1471-2288-13-91.

Miroslav Fiedler. Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, 23:298–305, 1973.

Fajwel Fogel, Alexandre d'Aspremont, and Milan Vojnovic. Spectral ranking using seriation. *Journal of Machine Learning Research*, 17(88):1–45, 2016.

Pascal Germain, Alexandre Lacasse, Francois Laviolette, Mario March, and Jean-Francis Roy. Risk bounds for the majority vote: From a pac-bayesian analysis to a learning algorithm. *Journal of Machine Learning Research*, 16(26):787–860, 2015.

ggplot2 Contributors. Housing sales in TX. Dataset documentation; source: Texas A&M University Real Estate Center, 2026. URL https://ggplot2.tidyverse.org/reference/txhousing.html. Numerical study uses txhousing.csv distributed with plotnine 0.15.3. Accessed September 23, 2026.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pp. 65–70, 1979.

Steven R. Howard, Aaditya Ramdas, Jon D. McAuliffe, and Jasjeet S. Sekhon. Time-uniform chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 2018.

Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021.

Charles D. Keeling and Timothy P. Whorf. Atmospheric CO2 concentrations derived from flask air samples at sites in the SIO network, 2004. In *Trends: A Compendium of Data on Global Change*. Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, U.S. Department of Energy, Oak Ridge, Tennessee, USA.

Vitaly Kuznetsov and Mehryar Mohri. Generalization bounds for non-stationary mixing processes. *Machine Learning*, 106:93–117, 2017. doi: 10.1007/s10994-016-5588-2.

Kasper Green Larsen. Bagging is an optimal pac learner. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 450–468. PMLR, 2023.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411, 2004.

Andrés Masegosa, Stephan Lorenzen, Christian Igel, and Yevgeny Seldin. Second order pac-bayesian bounds for the weighted majority vote. *Advances in Neural Information Processing Systems*, 33: 5263–5273, 2020.

Quinn Mcnemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12:153–157, 1947. doi: 10.1007/BF02295996.

Omar Montasser. Bagging robustly learns vc classes with linear sample complexity. *arXiv preprint arXiv:2608.13514*, 2026.

D. Paulin. Concentration inequalities for markov chains by marton couplings and spectral methods. *Electronic Journal of Probability*, 20:1–32, 2012.

Plotnine Contributors. US meat production. Dataset documentation; source: USDA Economic Research Service, Livestock and Meat Domestic Data, 2026. URL `https://plotnine.org/reference/meat.html`. Numerical study uses meat.csv distributed with plotnine 0.15.3. Accessed September 23, 2026.

Dimitris N Politis and Joseph P Romano. *A circular block-resampling procedure for stationary data*. Purdue University. Department of Statistics, 1991.

Dimitris N Politis and Joseph P Romano. The stationary bootstrap. *Journal of the American Statistical association*, 89(428):1303–1313, 1994.

Jeffrey S. Racine. Consistent cross-validatory model-selection for dependent data: hv-block cross-validation. *Journal of Econometrics*, 99:39–61, 2000.

statsmodels Developers. MSTL decomposition: Victoria electricity demand example and test data. statsmodels documentation; data fixture distributed with version 0.14.6, 2026a. URL `https://www.statsmodels.org/stable/examples/notebooks/generated/mstl_decomposition.html`. Data file: statsmodels/tsa/stl/tests/results/mstl_elec_vic.csv, version 0.14.6. Accessed September 23, 2026.

statsmodels Developers. El niño: Sea surface temperatures. statsmodels 0.14.6 dataset; NOAA ERSST.V3B source, 2026b. URL `https://www.statsmodels.org/stable/datasets/generated/elnino.html`. Monthly Niño 1+2 temperatures, 1950–2010. Accessed September 23, 2026.

statsmodels Developers. Yearly sunspots data 1700–2008. statsmodels 0.14.6 dataset; National Geophysical Data Center source, 2026c. URL `https://www.statsmodels.org/stable/datasets/generated/sunspots.html`. Historical packaged series. Accessed September 23, 2026.

Ulrike Von Luxburg, Mikhail Belkin, and Olivier Bousquet. Consistency of spectral clustering. *The Annals of Statistics*, pp. 555–586, 2008.

David H Wolpert. Stacked generalization. *Neural networks*, 5(2):241–259, 1992.

Bin Yu. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, pp. 94–116, 1994.

A Classification Limits with Observable Covariates

A.1 Stationarity, relaxation, and mixing

Write $\alpha = 1/T$ and $\rho = 1 - \alpha$. For distinct states z, z', detailed balance is $\pi_\theta(z)P(z,z') = \alpha\pi_\theta(z)\pi_\theta(z')$. Thus the kernel is reversible. Every mean-zero function in $L^2(\pi_\theta)$ is an eigenfunction with eigenvalue ρ, and

$$P^t(z,\cdot) = \rho^t\delta_z + (1-\rho^t)\pi_\theta. \tag{11}$$

The largest stationary point mass is at most $3/(4d)$, and the smallest is at most $1/(2d)$. Consequently $\sup_z \|\delta_z - \pi_\theta\|_{\mathrm{TV}}$ is between $1/2$ and 1. Equation (11) gives $T_{\mathrm{mix}} \asymp T$, with universal constants, including $T = 1$. More generally, relaxation and worst-start total-variation mixing need not be comparable. In particular, an unbounded Gaussian AR(1) chain does not have finite worst-start total-variation mixing time; its persistence parameter must not be substituted into this finite-state conclusion.

For any square-integrable g, the refresh kernel gives

$$\begin{aligned}\mathrm{Cov}(g(Z_t), g(Z_{t+h})) &= \rho^h \,\mathrm{Var}_\pi(g),\\ \mathrm{Var}\left(\frac{1}{n}\sum_{t=1}^n g(Z_t)\right) &= \frac{\mathrm{Var}_\pi(g)}{n}\left[1 + 2\sum_{h=1}^{n-1}\left(1 - \frac{h}{n}\right)\rho^h\right].\end{aligned} \tag{12}$$

The bracket is at most $1 + 2\rho/(1-\rho) = 2T - 1 \le 2T$.

A.2 Proof of Theorem 1

Let $\mu_x = 2\theta_x$ and $S_x = \sum_t Y_t \mathbf{1}\{X_t = x\}$. Since X is uniform, $\mathbb{E}[(d/n)S_x] = \mu_x$. Apply Equation (12) to $g_x(Z) = Y\mathbf{1}\{X = x\}$, whose variance is at most $1/d$. It follows that

$$\widehat{u}_x = \frac{d}{n}S_x, \qquad \mathrm{Var}(\widehat{u}_x) \le \frac{2dT}{n}. \tag{13}$$

Whenever $\operatorname{sign}\widehat{u}_x \neq \operatorname{sign}\mu_x$, the inequality $|\mu_x| \le |\widehat{u}_x - \mu_x|$ holds, including empty cells under the stated tie rule. Hence

$$\mathbb{E}\mathcal{E}(\widehat{H}) \le \frac{1}{d}\sum_{x=1}^d \mathbb{E}|\widehat{u}_x - \mu_x| \le \sqrt{\frac{2dT}{n}}. \tag{14}$$

Excess risk is also at most $1/2$ throughout this family.

For the lower bound, index a subfamily by $\sigma \in \{-1,1\}^d$ with $\mu_x = a\sigma_x$, where $0 < a \le 1/2$. Augment the trajectory with the initial state, independent refresh indicators, and the new state drawn at each refresh. Conditional on the indicators, these new states are independent draws from π_σ. The original trajectory is a deterministic function of this augmented experiment. Let

$$L = 1 + (n-1)/T. \tag{15}$$

If σ and $\sigma^{(x)}$ differ only at cell x, the KL chain rule and data processing give

$$\begin{aligned}\mathrm{KL}(\mathbb{P}_\sigma^n \| \mathbb{P}_{\sigma^{(x)}}^n) &\le L\,\mathrm{KL}(\pi_\sigma \| \pi_{\sigma^{(x)}})\\ &= \frac{La}{d}\log\frac{1+a}{1-a} \le \frac{4La^2}{d}.\end{aligned} \tag{16}$$

The final inequality follows from $\log((1+a)/(1-a)) \le 4a$ for $a \le 1/2$. Choose $a = \frac{1}{8}\min\{1, \sqrt{d/L}\}$. Every neighboring KL is at most $1/16$, so Pinsker's inequality bounds the neighboring total variation by $1/\sqrt{32} < 1/4$.

For any possibly randomized learner, its prediction at cell x is a test of σ_x. Pairing every sign vector with its neighbor $\sigma^{(x)}$, the sum of the two testing errors is at least $1 - \mathrm{TV} > 3/4$. Averaging over the hypercube yields average error at least $3/8$ at each cell. Therefore

$$2^{-d}\sum_\sigma \mathbb{E}_\sigma\mathcal{E}(\widehat{H}) = \frac{a}{d}\sum_{x=1}^d 2^{-d}\sum_\sigma \Pr_\sigma\{\widehat{H}(x) \neq \sigma_x\} \ge \frac{3a}{8}. \tag{17}$$

A supremum is at least this average. Since $L \le 1 + n/T$ and $d \ge 1$, the quantity $\min\{1, \sqrt{d/L}\}$ is at least a universal constant times $\min\{1, \sqrt{dT/n}\}$. This proves the stated lower bound for all n, d, T, not just the regime with many refreshes. The construction permits arbitrary feature-dependent Bayes signs while retaining an exactly identifiable information scale.

B Finite Voting and Shared Training Uncertainty

B.1 An exact majority-amplification identity

For odd b, define $q_b(p) = \Pr\{\mathrm{Bin}(b, p) > b/2\}$ and $g_b(p) = 2q_b(p) - 1$. Symmetry gives $q_b(1-p) = 1 - q_b(p)$. Conditioning on the first $2r+1$ Bernoulli trials yields

$$q_{2r+3}(p) - q_{2r+1}(p) = \binom{2r+1}{r} p^{r+1}(1-p)^{r+1}(2p-1). \tag{18}$$

Only initial success counts r and $r+1$ can change the majority after two further trials. Since $q_1(p) = p$, the identity implies $q_b(p) \ge p$ for $p \ge 1/2$ and $q_b(p) \le p$ for $p \le 1/2$. Thus $g_b(p)$ has the sign of $2p-1$ and $|g_b(p)| \ge |2p-1|$.

B.2 Proof of Theorem 2

Let $N_x = \sum_t \mathbf{1}\{X_t = x\}$ and, when $N_x > 0$, $\overline{Y}_x = S_x/N_x$. Conditional on the trajectory, the mean of one bootstrap majority at cell x is $a_x = g_b((1 + \overline{Y}_x)/2)$. If $\rho_x = m^{-1}\sum_{j=1}^m V_j(x)$ is the finite-vote score, then $\mathbb{E}[\rho_x \mid S] = a_x$ and $\mathrm{Var}(\rho_x \mid S) \le 1/m$.

For a nonempty cell, the following pointwise inequality separates training error from Monte Carlo error:

$$|\mu_x|\mathbf{1}\{\operatorname{sign}\rho_x \ne \operatorname{sign}\mu_x\} \le |\mu_x - \widehat{u}_x| + \frac{dN_x}{n}|\rho_x - a_x|. \tag{19}$$

If $\widehat{u}_x$ has the wrong sign, its first term already bounds the left side. Otherwise an incorrect finite-vote sign forces $|\rho_x - a_x| \ge |a_x| \ge |\overline{Y}_x|$. Multiplication by dN_x/n gives at least $|\widehat{u}_x|$, and the triangle inequality completes the argument. When $\widehat{u}_x = 0$ the first term again suffices. For an empty cell, take $\widehat{u}_x = 0$ and omit the second term; the same conclusion holds.

Average Equation (19) over cells and training samples. The first contribution is at most $\sqrt{2dT/n}$ by Equation (13). Conditional Cauchy–Schwarz bounds the second by

$$\frac{1}{d}\sum_{x=1}^{d}\mathbb{E}\left[\frac{dN_x}{n}\mathbb{E}[|\rho_x - a_x| \mid S]\right] \le \frac{1}{d}\sum_{x=1}^{d}\mathbb{E}\left[\frac{dN_x}{n\sqrt{m}}\right] = \frac{1}{\sqrt{m}}. \tag{20}$$

This calculation uses realized occupancy rather than dividing by a potentially small random cell count. Combining the two terms and the universal bound $\mathcal{E} \le 1/2$ proves the theorem. Ties at either voting stage are covered by the same inequalities.

B.3 Shared-data covariance and linear bootstrap means

For square-integrable predictions $V_j = \mathcal{A}(S, U_j)$ with conditionally independent, identically distributed seeds, total covariance and total variance give

$$\mathrm{Cov}(V_j, V_k) = \mathrm{Var}(\mathbb{E}[V_1 \mid S]), \quad j \ne k, \tag{21}$$

$$\mathrm{Var}\left(\frac{1}{m}\sum_{j=1}^{m} V_j\right) = \mathrm{Var}(\mathbb{E}[V_1 \mid S]) + \frac{1}{m}\mathbb{E}[\mathrm{Var}(V_1 \mid S)]. \tag{22}$$

The conditional cross-covariance is zero; the unconditional shared-data term need not be. For $d = 1$ and bootstrap means $L_j = b^{-1}\sum_{r=1}^{b} Y_{I_{jr}}$, independent uniform indices give

$$\mathrm{Cov}(L_j, L_k) = \mathrm{Var}(\overline{Y}_n), \qquad \mathrm{Var}\left(\frac{1}{m}\sum_j L_j\right) = \mathrm{Var}(\overline{Y}_n) + \frac{\mathbb{E}(1 - \overline{Y}_n^2)}{mb}. \tag{23}$$

For $n \geq 2T$, $|\mu| \leq 1/2$, and $\rho = 1 - 1/T$, Equation (12) gives the uniform bounds

$$\frac{3}{4}(1 - e^{-1/2})\frac{T}{n} \leq \mathrm{Var}(\overline{Y}_n) \leq \frac{2T}{n}. \tag{24}$$

For the lower bound, let $q = \lfloor T \rfloor \geq T/2$. Since $q - 1 < n/2$, $1 + 2\sum_{h=1}^{n-1}(1 - h/n)\rho^h \geq \sum_{h=0}^{q-1}\rho^h = T(1 - \rho^q) \geq T(1 - e^{-1/2})$. Multiply by $\mathrm{Var}(Y)/n \geq 3/(4n)$; $T = 1$ follows directly. Resampling does not add a second factor of T to this common uncertainty.

C Uniform-Bootstrap ERM under Absolute Regularity

C.1 Statement under Bayes containment

A blocked attainability statement also holds beyond observable cells when the hypothesis class contains the Bayes rule. Let $\mathcal{H}$ be a measurable binary class of finite VC dimension v, with growth function $\Pi_{\mathcal{H}}(s)$, and assume empirical minimizers exist. Assume $h^* \in \mathcal{H}$; label noise is allowed. For a stationary absolutely regular sequence, use the convention in which k blocks separated by at least a observations couple to independent blocks with failure at most $(k-1)\beta(a)$.

Theorem 7 (Blocked attainability under Bayes containment)**.** *Fix $\delta \in (0,1)$, $a \geq 1$, $k = \lfloor n/(2a) \rfloor \geq 1$, and $n' = 2ak$. Each voter minimizes empirical classification error on b uniform bootstrap draws from the same first n' observations. Define*

$$u = \min\left\{1, \sqrt{\frac{8\log(8\Pi_{\mathcal{H}}(n')/\delta)}{k}}\right\}, \qquad v_b = \min\left\{1, 2\sqrt{\frac{2\log(2\Pi_{\mathcal{H}}(b))}{b}}\right\}. \tag{25}$$

With probability at least $1 - \delta - 2(k-1)\beta(a)$ over training, every fixed voter count $m \geq 1$ has

$$\mathbb{E}_{\mathrm{boot}}[\mathcal{E}(H_m) \mid S] \leq \min\{1,\, 4u + 4v_b\}. \tag{26}$$

Coupling odd and even blocks controls the full empirical risk uniformly over $\mathcal{H}$. Conditional bootstrap symmetrization then controls each ERM's empirical excess. Finally, a majority can disagree with the Bayes rule only if at least half its voters do; applying this pointwise to Equation (1) proves the result (Appendix C). If $b \geq k$, the rate is $O(\sqrt{[v\log n + \log(1/\delta)]/k})$. Exponential mixing gives the $\sqrt{vT/n}$ dependence up to blocking and growth-function logarithms. The appendix retains the misspecified comparison with twice the class approximation error; it is not an agnostic attainability claim. The evaluated trees use finite-depth fitting and soft aggregation rather than this theorem's exact ERM and hard vote.

For the general comparison in this appendix, define $A_{\mathcal{H}} = \inf_{h \in \mathcal{H}} \mathcal{E}(h)$. The main theorem assumes $h^* \in \mathcal{H}$, so $A_{\mathcal{H}} = 0$. The growth function $\Pi_{\mathcal{H}}(s)$ is the maximum number of binary labelings that $\mathcal{H}$ induces on any s inputs. We assume the relevant suprema are measurable and empirical minimizers exist; a countable class suffices for the measurability issue. Approximate empirical minimization with error ξ per voter adds 2ξ to Equation (26). The bootstrap ERM uses a fixed tie rule and the observed labels, not a fitted regression function.

C.2 Uniform deviation of the original trajectory

Use the first $n' = 2ak$ observations, split into alternating odd and even blocks of length a. Write $O_j = (Z_{(2j-2)a+1}, \ldots, Z_{(2j-1)a})$ and $E_j = (Z_{(2j-1)a+1}, \ldots, Z_{2ja})$. Each subsequence of k blocks admits an independent-block coupling with mismatch probability at most $(k-1)\beta(a)$. A union bound therefore costs $2(k-1)\beta(a)$; it does not require the coupled odd and even subsequences to be independent of each other.

For either coupled block subsequence, let

$$f_h(B) = \frac{1}{a}\sum_{z \in B} \mathbf{1}\{h(x) \neq y\} \in [0,1]. \tag{27}$$

Stationarity implies $\mathbb{E}f_h(B) = R(h)$ even though observations within a block can be dependent. On k independent blocks, the usual ghost-sample symmetrization gives, for $kt^2 \geq 2$,

$$\Pr\left\{\sup_{h\in\mathcal{H}}\left|\frac{1}{k}\sum_{j=1}^{k} f_h(B_j) - R(h)\right| > t\right\} \leq 4\Pi_{\mathcal{H}}(2ka)\exp(-kt^2/8). \tag{28}$$

For completeness, introduce k independent ghost blocks. For any function attaining a deviation larger than t, the ghost mean lies within $t/2$ of its expectation with probability at least $1/2$, by variance at most $1/(4k)$ and Chebyshev's inequality. This supplies the factor two in symmetrization. Conditional on the $2k$ blocks, exchanging each original/ghost pair introduces independent Rademacher signs. Each block-loss difference is in $[-1, 1]$. Hoeffding's inequality bounds a two-sided signed deviation exceeding $t/2$ by $2\exp(-kt^2/8)$. The labeling vectors on all $2ka$ inputs number at most $\Pi_{\mathcal{H}}(2ka)$, so a union bound proves Equation (28). This is the independent-block empirical-process argument used in dependent learning (Yu, 1994).

Let $\widehat{R}_S$ be the empirical error on all n' observations. If each coupled half has uniform deviation at most u, their average does also. For $u < 1$ as defined in Equation (25), Equation (28) and a union over halves give concentration failure at most $8\Pi_{\mathcal{H}}(n')\exp(-ku^2/8) = \delta$. The condition $ku^2 \geq 2$ holds automatically. When the radius is truncated to $u = 1$, uniform deviation at most one is deterministic. Returning to the original sequence gives the event

$$\mathcal{A} = \{\sup_{h\in\mathcal{H}} |\widehat{R}_S(h) - R(h)| \leq u\}, \qquad \Pr(\mathcal{A}) \geq 1 - \delta - 2(k-1)\beta(a). \tag{29}$$

The right-hand side can be nonpositive for a poor blocking choice; such a choice provides no useful confidence statement.

C.3 Uniform bootstrap deviation and the majority comparison

Conditional on S, one voter receives b independent observations from the empirical law of S. If $\widehat{R}_b$ is its bootstrap empirical error, symmetrization and the finite-class Rademacher bound yield

$$\mathbb{E}_{\text{boot}} \sup_{h\in\mathcal{H}} |\widehat{R}_b(h) - \widehat{R}_S(h)| \leq \min\left\{1, 2\sqrt{\frac{2\log(2\Pi_{\mathcal{H}}(b))}{b}}\right\} = v_b. \tag{30}$$

To see the finite cardinality, condition on the b resampled inputs: there are at most $\Pi_{\mathcal{H}}(b)$ loss vectors in $\{0,1\}^b$. Including both signs controls the absolute supremum. Every vector has Euclidean norm at most $\sqrt{b}$, so the conditional signed maximum is at most $\sqrt{2\log(2\Pi_{\mathcal{H}}(b))/b}$. The factor two is the symmetrization factor. No independence of the original trajectory is used in this conditional calculation.

On $\mathcal{A}$, for a bootstrap ERM $\widehat{h}_b$ and any fixed comparator $h \in \mathcal{H}$, let $Z_b = \sup_h |\widehat{R}_b(h) - \widehat{R}_S(h)|$. Then

$$\begin{aligned} R(\widehat{h}_b) &\leq \widehat{R}_S(\widehat{h}_b) + u \leq \widehat{R}_b(\widehat{h}_b) + Z_b + u \\ &\leq \widehat{R}_b(h) + Z_b + u \leq \widehat{R}_S(h) + 2Z_b + u \leq R(h) + 2Z_b + 2u. \end{aligned} \tag{31}$$

Subtract R^*, take the conditional bootstrap expectation, and approximate the population infimum over h to obtain $\mathbb{E}_{\text{boot}}\mathcal{E}(\widehat{h}_b) \leq A_{\mathcal{H}} + 2u + 2v_b$. For any realized binary voters $h_1, \ldots, h_m$, the pointwise inequality

$$\mathbf{1}\{\text{sign}\textstyle\sum_{j=1}^m h_j(x) \neq h^*(x)\} \leq \frac{2}{m}\sum_{j=1}^m \mathbf{1}\{h_j(x) \neq h^*(x)\} \tag{32}$$

holds with the declared tie rule. Multiply by $|\eta(x)|$ and integrate. Combining with Equation (31) gives $\mathbb{E}_{\text{boot}}\mathcal{E}(H_m) \leq 2A_{\mathcal{H}} + 4u + 4v_b$. Under the main theorem's Bayes-containment assumption, $A_{\mathcal{H}} = 0$, which proves Equation (26). A randomized choice of one voter instead has expected excess risk at most $A_{\mathcal{H}} + 2u + 2v_b$ by linearity; that different learner is not a deterministic thresholded soft vote. The comparison does not require independent voters; the stated learner uses independent bootstrap seeds. Its bound concerns conditional expected ensemble risk for each fixed m, not a simultaneous guarantee for every realized bootstrap ensemble.

C.4 Rate interpretation and scope

For $v \le s$, Sauer's inequality gives $\Pi_{\mathcal{H}}(s) \le (es/v)^v$. If the Bayes classifier belongs to $\mathcal{H}$, then $A_{\mathcal{H}} = 0$ even with noisy labels. Taking $b \ge k$ makes the bootstrap term at most the blocked order, using monotonicity of $(\log b)/b$ beyond a constant and trivial bounds for smaller b. More explicitly, both radii are bounded, up to universal constants and truncation at one, by $\sqrt{[v\log(en') + \log(1/\delta)]/k}$. For $\beta(a) \le B\exp(-a/T)$, $B \ge 1$, taking

$$a = \left\lceil T \log \frac{2Bn}{\delta} \right\rceil, \qquad k = \left\lfloor \frac{n}{2a} \right\rfloor \ge 1 \tag{33}$$

makes the training coupling failure at most δ and gives the $\sqrt{vT/n}$ dependence up to logarithms. The theorem does not eliminate the logarithm incurred by this coupling analysis or prove a matching minimax lower bound over all absolutely regular laws. When the class is misspecified, the factor $2A_{\mathcal{H}}$ in the appendix comparison remains and does not establish a sharp agnostic approximation guarantee. The empirical tree ensembles use random feature subsampling, finite-depth fits, and soft aggregation; no claim is made that they implement the exact ERM or hard vote of this theorem.

D Graph Equivalence, Nonidentifiability, and Weighting

D.1 Proof of Theorem 3

We first establish monotonicity for the two combinatorial graph families, then treat the unweighted normalized path and the recursive learner equivalence. Spectral ordering of Robinson-type affinities is classical (Atkins et al., 1998; Fogel et al., 2016); the argument below gives the strictness and learner-matching conditions used here.

For a connected graph on s ordered vertices, let $D \in \mathbb{R}^{(s-1)\times s}$ be the first-difference operator $(Dx)_i = x_{i+1} - x_i$. Since $L\mathbf{1} = 0$, there is a unique $(s-1)$-dimensional operator M satisfying $DL = MD$. Directly writing $x_\ell = x_1 + \sum_{j<\ell}(Dx)_j$ gives

$$M_{ij} = \begin{cases} \displaystyle\sum_{\ell \le j} (W_{i+1,\ell} - W_{i,\ell}), & j < i, \\ \displaystyle\sum_{\ell \le i} W_{i+1,\ell} + \sum_{\ell > i} W_{i,\ell}, & j = i, \\ \displaystyle\sum_{\ell > j} (W_{i,\ell} - W_{i+1,\ell}), & j > i. \end{cases} \tag{34}$$

For a positive weighted nearest-neighbor path, M is tridiagonal. Its lower and upper off-diagonal entries are strictly negative. For $W_{ij} = \psi(|i-j|)$ with strictly decreasing positive ψ, every off-diagonal entry is strictly negative: when $j < i$, vertex $i+1$ is farther than i from every $\ell \le j$; when $j > i$, vertex i is farther than $i+1$ from every $\ell > j$. Thus both cases give an irreducible matrix with nonpositive off-diagonal entries. The case $s = 2$ is scalar and immediate.

The restriction of D to $\mathbf{1}^\perp$ is an isomorphism onto $\mathbb{R}^{s-1}$. Hence M is similar to the restriction of the symmetric Laplacian to $\mathbf{1}^\perp$, and its eigenvalues are precisely $\lambda_2(L), \dots, \lambda_s(L)$. Choose c larger than $\lambda_s(L)$ and every diagonal entry of M. The matrix $cI - M$ is nonnegative and irreducible. Perron–Frobenius gives a simple largest eigenvalue with a strictly positive eigenvector. Since all eigenvalues of M are real and positive, this largest eigenvalue is $c - \lambda_2(L)$. Therefore the Fiedler eigenvalue is simple and Dv is strictly positive or strictly negative. Orthogonality $v^\top \mathbf{1} = 0$ makes an orientation with $v_1 > 0$ strictly decreasing. Its largest $\lceil s/2 \rceil$ coordinates are the first such indices. Every child is an induced interval of the same graph family, so the argument repeats.

For the unweighted combinatorial path one can also verify the eigenvalues $2\{1 - \cos(k\pi/s)\}$ and vectors $\cos(k\pi(i-1/2)/s)$ directly. For the symmetric normalized path, similarity to the random-walk operator instead gives

$$\lambda_k = 1 - \cos\left(\frac{k\pi}{s-1}\right), \qquad v_i^{(k)} = \sqrt{\deg(i)}\cos\left(\frac{k\pi(i-1)}{s-1}\right), \quad k = 0, \dots, s-1. \tag{35}$$

At $k = 1$, the left $\lfloor s/2 \rfloor$ coordinates are positive and the right ones negative; for odd s, the central coordinate is zero. The largest $\lceil s/2 \rceil$ entries therefore give the same left interval. This uses their sign

separation, not global monotonicity of the normalized vector, which can fail near the endpoints. This proves the normalized-path part of Theorem 3; Corollary 4 covers normalized bands by a different argument below.

Induction through the requested bisection levels produces exactly recursive midpoint intervals. Canonically order groups by their smallest index. With identical head assignments, each corresponding learner receives the same ordered observations, bootstrap indices, and random seed. Its fitted prediction is then identical on every input. The same subsequent aggregation preserves equality, including thresholding individual heads and taking a hard majority. This is a pathwise identity under any data-generating law. It does not assert that blocking itself equals uniform bagging or that an unmatched implementation has the same budgets.

D.2 Proof of Corollary 4

Let $s \geq 2w+1$, with vertices $1, \ldots, s$, and write $D_\nabla x_i = x_{i+1} - x_i$. In the combinatorial difference operator of Equation (34), a band affinity gives

$$M_{ij} = \begin{cases} -\mathbf{1}\{1 \leq i-w \leq j < i\}, & j < i, \\ -\mathbf{1}\{i < j \leq i+w,\ i+w+1 \leq s\}, & j > i. \end{cases} \tag{36}$$

Thus row i has a backward link exactly when $i \geq w+1$, and a forward link exactly when $i \leq s-w-1$. Interpret a strictly negative entry as a directed link from its row to its column. Vertices $1, \ldots, w$ each reach $w+1$ by a forward link, while $w+1$ reaches all of them by backward links. Every vertex above $w+1$ can repeatedly move backward to this core. Conversely, forward-neighbor links reach all vertices through $s-w$; row $s-w-1 \geq w$ also reaches the remaining vertices through $s-1$. Hence the digraph is strongly connected. The similarity and Perron–Frobenius argument in the preceding proof gives a simple λ_2 and strictly one-signed differences. The resulting combinatorial vector is strictly monotone.

For normalization, distinguish the degree matrix D_{deg} from the difference operator. Set

$$d_i = \min(w, i-1) + \min(w, s-i), \qquad P = D_{\mathrm{deg}}^{-1} W, \qquad L_{\mathrm{rw}} = I - P. \tag{37}$$

Since $L_{\mathrm{rw}}\mathbf{1} = 0$, there is a unique matrix $\widetilde{M}$ satisfying $D_\nabla L_{\mathrm{rw}} = \widetilde{M} D_\nabla$. For $j < i$, its entry is the difference of left-tail transition probabilities,

$$\widetilde{M}_{ij} = F_{i+1}(j) - F_i(j), \qquad F_i(j) = \frac{(j - \ell_i + 1)_+}{d_i}, \quad \ell_i = \max(1, i-w). \tag{38}$$

If $\ell_{i+1} = \ell_i$, then $i \leq w$ and $d_{i+1} \geq d_i$, so the difference is nonpositive. Otherwise the left endpoint increases by one. If $q = j - \ell_i + 1 > 0$, the next numerator is $q-1$ and $d_{i+1} \geq d_i - 1$. Since $j < i$ and vertex i has a right neighbor, $q < d_i$. Therefore

$$\frac{q-1}{d_{i+1}} \leq \frac{q-1}{d_i - 1} < \frac{q}{d_i}. \tag{39}$$

For $q \leq 0$, both tails are zero. Every combinatorial backward link is consequently a strict normalized link. Reflection of the interval gives the upper off-diagonal entries and preserves every combinatorial forward link. Thus $\widetilde{M}$ has nonpositive off-diagonal entries and is irreducible.

The subspace $\{h : d^\top h = 0\}$ is invariant under L_{rw}, and D_∇ maps it isomorphically onto $\mathbb{R}^{s-1}$. The operator on this subspace is similar to the nonzero restriction of $L_{\mathrm{sym}} = I - D_{\mathrm{deg}}^{-1/2} W D_{\mathrm{deg}}^{-1/2}$, so its eigenvalues are real and positive. Applying Perron–Frobenius to $cI - \widetilde{M}$ proves that the generalized Fiedler vector h is simple and strictly monotone. Reflection commutes with L_{rw}; simplicity implies that a reflected eigenvector is either h or $-h$. Strict monotonicity excludes the former. Orienting $h_1 > 0$ gives

$$h_i = -h_{s+1-i}, \qquad h_i > 0 \ (i < (s+1)/2), \qquad h_i < 0 \ (i > (s+1)/2). \tag{40}$$

The symmetric normalized Fiedler vector is $v = D_{\mathrm{deg}}^{1/2} h$. Positive degrees preserve these signs, including the central zero for odd s. Its largest $\lceil s/2 \rceil$ coordinates therefore form the left midpoint interval even though their ordering within either half need not be monotone. This proves both balanced-split conclusions.

Every midpoint child is itself an induced band interval with the same width parameter. Recursion therefore preserves equality whenever all parents being split have size at least $2w + 1$. For $\ell \geq 1$ levels, the sufficient condition is

$$\left\lfloor \frac{n}{2^{\ell-1}} \right\rfloor \geq 2w + 1. \tag{41}$$

Equality of the trained predictors follows from the same allocation and seed matching as Theorem 3. A two-vertex interval is separately a single-edge graph and has its trivial midpoint split.

The threshold also explains the loss of strict ordering below it. For $3 \leq s \leq w + 1$, the graph is complete and its Fiedler eigenspace has dimension $s - 1$. For $w + 2 \leq s \leq 2w$, the consecutive vertices from $s - w$ through $w + 1$ are all adjacent to every other vertex, so there are at least two universal central vertices. The graph is not complete. Its combinatorial $\lambda_2 < s$: all nonzero Laplacian eigenvalues are at most s, and equality for the smallest would force the complete graph. A Fiedler vector orthogonal to $\mathbf{1}$ consequently has zero at every universal vertex, since there $(Lv)_i = sv_i$. For the normalized operator, $\lambda_2 < s/(s-1)$: the nonzero eigenvalues average $s/(s-1)$, and equality of all of them would imply a complete graph. Subtracting the generalized eigen-equations at two universal vertices gives

$$\left(\frac{s}{s-1} - \lambda_2 \right) (h_i - h_j) = 0. \tag{42}$$

Their generalized and symmetric normalized coordinates are therefore equal. This proves the strict-ordering boundary; it does not assert that every undersized interval yields a noncontiguous median partition.

D.3 A NONCONTIGUOUS SPLIT WITHOUT AN EIGENVALUE TIE

On four ordered vertices, set $W_{13} = W_{31} = W_{24} = W_{42} = 1$, all other off-diagonal affinities to $\varepsilon \in (0, 1)$, and diagonal entries to zero. Every degree is $1 + 2\varepsilon$. The combinatorial eigenvalues are

$$0, \quad 4\varepsilon, \quad 2 + 2\varepsilon, \quad 2 + 2\varepsilon, \tag{43}$$

and the simple Fiedler eigenvector is $(1, -1, 1, -1)/2$. Its positive and negative sets are $\{1, 3\}$ and $\{2, 4\}$. Because the graph is regular, the normalized Laplacian has the same eigenvectors. Both median partitions are noncontiguous and have a strict gap between the two selected coordinate levels. Consequently, an unrestricted affinity can escape midpoint blocking for structural reasons, not just numerical ambiguity. The audit uses $\varepsilon = .1$ and recovers this eigenspace numerically.

D.4 WHY AN INDEX GAP DOES NOT ESTIMATE TRANSITION MIXING

A graph constructed solely from a fixed index rule, including $|i - j| \leq w$, has the same distribution under every process of the same length. Choose two transition laws with distinct gaps g_1, g_2. For any graph-only estimator and $\varepsilon = |g_1 - g_2|/3$, the two success intervals are disjoint, while the estimator's output law is identical under both processes. Its failure probability is therefore at least $1/2$ under one law. For a simple path,

$$\lambda_2(L_{\text{norm}}) = 1 - \cos\left(\frac{\pi}{n-1} \right) = \Theta(n^{-2}), \qquad n \geq 3. \tag{44}$$

This eigenvalue changes with graph size, not with the refresh parameter. Observed transition information or further identifiable modeling assumptions are needed to estimate the latter.

D.5 BALANCED LINEAR CONSERVATION AND UNIFORM OPTIMALITY

For a balanced partition $I_1, \ldots, I_m$ with group size $p = n/m$, each observation has coefficient $1/(mp) = 1/n$ in the uniform average of group means. More generally, $m^{-1} \sum_j a_j = n^{-1}\mathbf{1}$ directly implies $m^{-1} \sum_j a_j^\top U = n^{-1}\mathbf{1}^\top U$. A reduction in within-group covariance is offset by changes in cross-group covariance whenever this identity holds.

Proposition 8 (Uniform optimum and a feasible descent direction)**.** *Let $K \succeq 0$, $q = \mathbf{1}/m$, and $\kappa > 1$. Uniform weights minimize $w^\top K w$ on $\mathcal{W} = \{w \geq 0 : \mathbf{1}^\top w = 1, \|w\|_2^2 \leq \kappa/m\}$ if and only*

if $K\mathbf{1} = c\mathbf{1}$ for some scalar c. Define $\Pi = I - \mathbf{1}\mathbf{1}^\top/m$ and $g = \Pi Kq$. If $g \neq 0$, any positive t satisfying

$$t \leq \min\left\{\frac{1}{m\|g\|_\infty}, \frac{1}{\lambda_{\max}(K)}, \sqrt{\frac{\kappa - 1}{m\|g\|_2^2}}\right\} \tag{45}$$

gives a feasible $w = q - tg$ with $q^\top Kq - w^\top Kw \geq t\|g\|_2^2 > 0$.

Proof. Uniform q lies in the relative interior of the simplex and strictly satisfies the norm cap. First-order optimality requires Kq to be orthogonal to the zero-sum subspace, equivalently $K\mathbf{1} = c\mathbf{1}$. Conversely, when this holds, $w^\top Kw - q^\top Kq = (w-q)^\top K(w-q) \geq 0$ for every simplex w.

For the descent construction, $\mathbf{1}^\top g = q^\top g = 0$. The first bound on t ensures nonnegativity, and $\|q - tg\|_2^2 = 1/m + t^2\|g\|_2^2 \leq \kappa/m$. Since $g^\top Kq = \|g\|_2^2$,

$$q^\top Kq - (q-tg)^\top K(q-tg) = 2t\|g\|_2^2 - t^2 g^\top Kg \geq (2t - t^2\lambda_{\max}(K))\|g\|_2^2 \geq t\|g\|_2^2. \tag{46}$$

□

For $\kappa = 1$, only the uniform vector is feasible. If $\|K - C\|_{\mathrm{op}} \leq \varepsilon$ for an actual error matrix C, then

$$q^\top Cq - w^\top Cw \geq q^\top Kq - w^\top Kw - \varepsilon(\|q\|_2^2 + \|w\|_2^2). \tag{47}$$

A numerical proxy gain must exceed this uncertainty to certify an actual gain.

Proposition 9 (Combinatorial graph filters force uniformity)**.** *For a combinatorial Laplacian L and a PSD spectral function $K = f(L)$, $K\mathbf{1} = f(0)\mathbf{1}$, so uniform weights minimize $w^\top Kw$ over the simplex. If A averages a balanced partition with group size p, then $Af(L)A^\top\mathbf{1} = f(0)\mathbf{1}/p$ and the same conclusion holds for group weights.*

Proof. The identities follow from $L\mathbf{1} = 0$, $f(L)\mathbf{1} = f(0)\mathbf{1}$, $A\mathbf{1} = \mathbf{1}$, and $A^\top\mathbf{1} = \mathbf{1}/p$. Apply Proposition 8, or its simplex version with an inactive cap. □

This obstruction applies to bare heat-kernel and resolvent proxies. It does not automatically apply to normalized Laplacians on irregular graphs or to a residual matrix projected through a graph basis. The latter can retain nonconstant residual scales.

D.6 A BOUNDED HETEROGENEOUS-SCORE CONSTRUCTION

For $m = 2p$, let

$$C = c\mathbf{1}\mathbf{1}^\top + \mathrm{diag}\left(\sigma_A^2 I_p + \tau_A \mathbf{1}_p\mathbf{1}_p^\top, \sigma_B^2 I_p + \tau_B \mathbf{1}_p\mathbf{1}_p^\top\right), \tag{48}$$

with nonnegative parameters. Convexity permits averaging weights within either exchangeable group without increasing the objective. Set $A = \tau_A + \sigma_A^2/p$ and $B = \tau_B + \sigma_B^2/p$. The resulting variance is $c + As^2 + B(1-s)^2$. If $A + B > 0$, elementary minimization gives

$$s^* = \frac{B}{A+B}, \qquad V^* = c + \frac{AB}{A+B}, \qquad V(1/2) - V^* = \frac{(A-B)^2}{4(A+B)}. \tag{49}$$

The per-score optimum has squared norm $[s^{*2} + (1-s^*)^2]/p \leq 2/m$, so $\kappa = 2$ allows it. The gain is strict exactly when $A \neq B$. Taking $A = a/n$, $B = b/n$, $c = d_0/n$ gives an order-n^{-1} absolute gain and a nonvanishing relative gain when $a \neq b$. The common component cannot be removed by simplex averaging.

A bounded realization includes independent Rademacher coordinates $U_0, U_A, U_B, U_1, \ldots, U_m$ in the observable input. For j in group G, set

$$f_j(X) = \eta(X) + \sqrt{c}\,U_0 + \sqrt{\tau_G}\,U_G + \sigma_G U_j, \tag{50}$$

where the regression function depends on an additional independent input coordinate, and sample $Y \in \{-1, 1\}$ with conditional mean $\eta(X)$. If $|\eta(X)| + \sqrt{c} + \sqrt{\tau_G} + \sigma_G \leq 1$, the score is bounded and its error matrix relative to η is exactly Equation (48). The group structure here is part of the construction, not inferred from residual labels.

E Residual Calibration and Surrogate Guarantees

Proposition 10 (Residual-risk oracle). *For every fixed candidate family r, let $f_r : \mathcal{X} \to [-1,1]^{m_r}$ and let $\widehat{Q}_r$ estimate the matrix in Equation* (5)*. Suppose $\max_{r,i,j} |(\widehat{Q}_r - Q_r)_{ij}| \le u$ and $K_r \succeq 0$ has $\|K_r - \widehat{Q}_r\|_{\mathrm{op}} \le d_r$. Over $\mathcal{W}_r = \{w \ge 0 : \mathbf{1}^\top w = 1, \|w\|_2^2 \le \kappa_r/m_r\}$, $\kappa_r \ge 1$, let $(\widehat{r}, \widehat{w})$ minimize $J_r(w) = w^\top K_r w + d_r\|w\|_2^2$ to objective tolerance ζ. With $L_r(w) = \mathbb{E}_X[(w^\top f_r(X) - \eta(X))^2]$,*

$$\mathcal{E}(\mathrm{sign}(\widehat{w}^\top f_{\widehat{r}})) \le \left[\min_{r, w \in \mathcal{W}_r} \{L_r(w) + 2u + 2d_r\|w\|_2^2\} + \zeta\right]^{1/2}. \tag{51}$$

E.1 Proof of Proposition 10

Condition on training so every score function and graph basis is fixed. Since $\mathbb{E}[Y - \eta(X) \mid X] = 0$ and $\mathbf{1}^\top w = 1$, expansion gives

$$w^\top Q_r w = L_r(w) + \nu, \qquad \nu = \mathbb{E}[1 - \eta(X)^2]. \tag{52}$$

Individual scores need not be unbiased or consistent. They must share the target distribution. Nonnegative simplex weights and the simultaneous entrywise event imply

$$|w^\top(\widehat{Q}_r - Q_r)w| \le \sum_{i,j} w_i w_j u = u, \qquad |w^\top(K_r - \widehat{Q}_r)w| \le d_r\|w\|_2^2. \tag{53}$$

Thus, for $J_r(w) = w^\top K_r w + d_r\|w\|_2^2$,

$$w^\top Q_r w \le J_r(w) + u \le w^\top Q_r w + 2u + 2d_r\|w\|_2^2. \tag{54}$$

The left inequality and objective optimality to tolerance ζ yield, for any comparator (r, w),

$$\begin{aligned} \widehat{w}^\top Q_{\widehat{r}}\widehat{w} &\le J_{\widehat{r}}(\widehat{w}) + u \\ &\le J_r(w) + u + \zeta \le w^\top Q_r w + 2u + 2d_r\|w\|_2^2 + \zeta. \end{aligned} \tag{55}$$

Subtract the common ν and minimize over comparators. For any score g, a sign disagreement with η implies $|\eta| \le |g - \eta|$, so $\mathcal{E}(\mathrm{sign}\, g) \le \mathbb{E}|g - \eta| \le \sqrt{\mathbb{E}(g-\eta)^2}$. This proves Equation (51). The bound is uniform over calibration-dependent simplex weights. An unconditional expectation bound adds at most the calibration-event failure probability, since excess classification risk is at most one.

The same event gives a direct within-family squared-error certificate. For $q = \mathbf{1}/m$, if

$$q^\top K q - w^\top K w > 2u + d(\|q\|_2^2 + \|w\|_2^2), \tag{56}$$

then $L(w) < L(q)$. Taking $K = \widehat{Q}$ removes the approximation term. Neither inequality orders classification risks in general: at constant $\eta = .2$, the score $-.01$ has squared error $.0441$ while $.5$ has squared error $.09$, yet their classification risks are $.6$ and $.4$, respectively. This is why final validation uses the thresholded classifiers rather than their residual objectives.

E.2 Independent and dependent calibration

For each binary label, every coordinate of $f_r(X) - Y\mathbf{1}$ has the same weak sign and magnitude at most two. Every residual-product entry lies in $[0, 4]$. Hence, on k_c independent observations,

$$\Pr\{|(\widehat{Q}_r - Q_r)_{ij}| > u\} \le 2\exp(-k_c u^2/8). \tag{57}$$

A union bound over $\sum_r m_r^2$ entries proves the main-text radius.

Proposition 11 (Calibration from a thinned trajectory). *Suppose frozen score functions are independent of a stationary absolutely regular calibration trajectory. Retain k_c observations with successive gaps at least a. For*

$$u = \sqrt{\frac{8\log(2\sum_r m_r^2/\delta_c)}{k_c}}, \tag{58}$$

the event $\max_{r,i,j} |(\widehat{Q}_r - Q_r)_{ij}| \le u$ has probability at least $1 - \delta_c - (k_c - 1)\beta(a)$. It also implies $\|\widehat{Q}_r - Q_r\|_{\mathrm{op}} \le m_r u$ for each r.

Proof. Successive Berbee couplings replace all retained observations by independent stationary observations with probability of any mismatch at most $(k_c - 1)\beta(a)$. Apply the preceding Hoeffding union bound to that independent sample. The actual sample satisfies the event whenever both the independent concentration event and coupling success hold; a union bound on failures does not condition the independent law on coupling success. Finally $\|E\|_{\mathrm{op}} \le \|E\|_{\mathrm{F}} \le m_r \max_{i,j} |E_{ij}|$. □

When construction and calibration occupy the same trajectory with a separating gap a_0, first decouple the entire future calibration block from the construction sigma-field, adding $\beta(a_0)$. Further data reuse requires accounting for every separation. The generic remainder $u = O(k_c^{-1/2})$ enters inside the square root in Equation (51); no refresh-family rate is asserted for arbitrary fitted scores.

F Paired Protection, Power, and Lower Bounds

F.1 The separate-interval comparison

Condition on construction and calibration. Couple the retained validation observations once to independent observations with target law π. A mismatch occurs with probability at most $\gamma = (k-1)\beta(a)$. On the independent sample, Hoeffding gives $\Pr(|\widehat{R}_r - R(H_r)| > e) \le 2e^{-2ke^2}$. Taking $e = \sqrt{\log(2M/\delta)/(2k)}$ and a union over the frozen family gives simultaneous risk intervals with failure probability δ. Adding the common coupling failure yields $\delta + \gamma$.

On this event, empirical risk minimization satisfies

$$R(H_{\widehat{r}}) \le \widehat{R}_{\widehat{r}} + e \le \widehat{R}_{r^*} + e \le R(H_{r^*}) + 2e, \tag{59}$$

where r^* minimizes population risk. The protected rule is different: whenever $\widehat{R}_0 - \widehat{R}_r > 2e$, $R(H_r) < R(H_0)$, and otherwise it returns H_0. It can therefore decline a useful but unresolved candidate. If the observed gain is fixed at $g > 0$, its threshold can only pass when

$$k > \frac{2\log(2M/\delta)}{g^2}. \tag{60}$$

This inequality is a threshold calculation, not a power guarantee: observed gains vary with validation data.

For total failure budget α, use $\delta + \gamma \le \alpha$. If $\beta(a) \le Be^{-a/T}$ and $B \ge 1$, the choice $\delta = \alpha/2$ and $a \ge T\log(2BN/\alpha)$ suffices for $\gamma \le \alpha/2$, since $k - 1 \le N$. The raw trajectory length N and retained count k must both be reported. A previous construction segment requires an additional construction-to-validation coupling term. Stronger Markov concentration can reduce the cost of thinning under additional assumptions (Paulin, 2012); thinning is not asserted to be optimal.

F.2 Exact paired protection

For a frozen candidate define

$$U_i = \mathbf{1}\{H_0(X_i) \ne Y_i\} - \mathbf{1}\{H_r(X_i) \ne Y_i\} \in \{-1, 0, 1\}. \tag{61}$$

In binary classification, disagreeing predictions have exactly one error, so $\mathbb{E}U_i = \Delta_r$ and $\mathbb{E}U_i^2 = \omega_r$. For independent target observations,

$$B_r \mid D_r = d \sim \mathrm{Bin}\left(d, \frac{1}{2} + \frac{\Delta_r}{2\omega_r}\right), \quad \omega_r > 0. \tag{62}$$

Under $\Delta_r \le 0$, the conditional success probability is at most $1/2$. For every t, the event that the upper tail at $1/2$ is at most t is an upper set in B_r. Its probability is largest at $1/2$ and is at most t by construction of the discrete tail. Averaging over D_r gives $\Pr(p_r \le t) \le t$. If $\omega_r = 0$, then $D_r = 0$ almost surely and $p_r = 1$.

A union bound over at most J non-improving candidates yields

$$\Pr\{\exists r : \Delta_r \le 0,\ p_r \le \delta/J\} \le \delta. \tag{63}$$

The candidate tests may be arbitrarily dependent. Any validation-dependent choice among qualifying candidates retains the guarantee. Holm's step-down correction is also valid under arbitrary dependence, but all numerical comparisons here use the fixed Bonferroni threshold. For dependent validation, the same observation-level coupling preserves every B_r, C_r whenever it succeeds; add γ once, not once per candidate.

For $D_r > 0$, Hoeffding under the conditional null gives

$$p_r \le \exp\left(-\frac{(B_r - C_r)^2}{2D_r}\right) \quad \text{when } B_r > C_r. \tag{64}$$

Consequently the sufficient condition

$$B_r - C_r > \sqrt{2D_r \log(J/\delta)} \tag{65}$$

approves a candidate. In terms of empirical gain, this threshold is $\sqrt{2\widehat{\omega}_r \log(J/\delta)/k}$, where $\widehat{\omega}_r = D_r/k$. The exact tail can approve additional cases.

F.3 Power

For one beneficial candidate, abbreviate $\Delta = \Delta_r$, $\omega = \omega_r$, $S = B - C$, and $D = B + C$. Then $\mathbb{E}S = k\Delta$, $\mathrm{Var}(U_i) \le \omega$, $|U_i - \Delta| \le 2$, and $D \sim \mathrm{Bin}(k, \omega)$. Bernstein's inequality and a multiplicative Chernoff bound give

$$\Pr\{S \le k\Delta/2\} \le \exp\left(-\frac{k\Delta^2}{8(\omega + \Delta/3)}\right) \le \exp\left(-\frac{3k\Delta^2}{32\omega}\right), \tag{66}$$

$$\Pr\{D > 2k\omega\} \le \exp(-k\omega/3). \tag{67}$$

Let $L_0 = \max\{\log(J/\delta), \log(2/\epsilon)\}$. Under $k \ge 32\omega L_0/\Delta^2$, the failure probabilities sum to at most $2e^{-L_0} \le \epsilon$, using $\Delta \le \omega$. Outside these events,

$$S > k\Delta/2 > 2\sqrt{k\omega \log(J/\delta)} \ge \sqrt{2D \log(J/\delta)}. \tag{68}$$

The strict middle inequality follows from $k\Delta^2 \ge 32\omega L_0 > 16\omega \log(J/\delta)$. Equation (65) implies qualification. Adding coupling failure proves the dependent statement. At a fixed absolute gain, reducing feasible disagreement improves this cost; at a fixed conditional advantage Δ/ω, rarer disagreements instead require more observations.

F.4 Necessary cost of a protected comparison

Let $X \in \{0, 1\}$ with $\Pr(X = 1) = \omega$. Fix $H_0(x) = -1$, $H_1(0) = -1$, and $H_1(1) = +1$. Under both laws set $Y = -1$ at $X = 0$. At $X = 1$, set

$$\Pr_{+}(Y = 1 \mid X = 1) = \frac{1}{2} + \frac{\Delta}{2\omega}, \qquad \Pr_{-}(Y = 1 \mid X = 1) = \frac{1}{2} - \frac{\Delta}{2\omega}. \tag{69}$$

The gains are $+\Delta$ and $-\Delta$ while disagreement remains ω. For $0 < \Delta \le \omega/2$,

$$\mathrm{KL}(P_+ \| P_-) = \mathrm{KL}(P_- \| P_+) = \Delta \log \frac{\omega + \Delta}{\omega - \Delta} \le \frac{4\Delta^2}{\omega}. \tag{70}$$

For $s = \Delta/\omega \le 1/2$, the bound follows from $\log((1 + s)/(1 - s)) = 2\int_0^s (1 - t^2)^{-1} dt \le 2s/(1 - s^2) \le 4s$.

Let A be the event of candidate deployment. If $\Pr_+(A) \ge 1 - \epsilon_0$ and $\Pr_-(A) \le \delta_0$, with $\delta_0 + \epsilon_0 < 1$, independent observations and data processing imply

$$k\,\mathrm{KL}(P_+ \| P_-) \ge \mathrm{kl}(\Pr_{+}(A), \Pr_{-}(A)) \ge \mathrm{kl}(1 - \epsilon_0, \delta_0). \tag{71}$$

Binary relative entropy increases as its ordered arguments move apart. Apply the same argument to A^c under reversed laws to obtain $\mathrm{kl}(1 - \delta_0, \epsilon_0)$. Combining both inequalities with Equation (70) proves Equation (8). Randomized deployment rules are covered by including their independent seed in the data-processing step.

Algorithm 1 Protected comparison with an explicit risk target

Require: Construction and calibration data, candidate specifications, audit design, failure budget α

1: Fit matched learners, calibrate weights, and freeze every classifier.
2: Deduplicate identical specifications; retain uniform baseline H_0 and let J count alternatives.
3: **if** $J = 0$ **then**
4: **return** H_0
5: **end if**
6: Declare either population risk under a justified common target law, or the average conditional audit target.
7: Evaluate all candidates on the same observations, with a fixed reveal order.
8: **if** the population-risk target is declared **then**
9: Use the justified independence or coupling design, with $\delta + \gamma \le \alpha$.
10: Qualify candidates satisfying Equation (6) at level δ/J.
11: **else**
12: Use the fixed scale grid and qualify positive bounds in Equation (10).
13: **end if**
14: At the declared endpoint, select the qualified candidate with largest empirical gain, or H_0 if none qualifies.
15: Evaluate any later-period transfer separately, without changing the audit decision.

The construction also provides a raw-trajectory lower bound for refresh validation. Let the validation states follow the refresh kernel with stationary law P_+ or P_- and relaxation time T_v. Revealing refresh indicators shows

$$\mathrm{KL}(\mathbb{P}_+^N \| \mathbb{P}_-^N) \le \left(1 + \frac{N-1}{T_v}\right) \frac{4\Delta^2}{\omega}. \tag{72}$$

Any rule with the same protection and power must therefore replace k in the necessary inequality by $1 + (N-1)/T_v$. This is an information lower bound for that explicit Markov family. It neither asserts optimality of thinning nor removes the logarithmic and coupling costs from a particular sufficient construction.

Algorithm 1 keeps the two risk targets distinct. Without a justified common target law and dependence bound, an exact-binomial result is only an independent-observation reference; the temporal branch can instead certify the declared conditional audit target.

G Temporal Certification of Conditional Improvement

G.1 Proof of Theorem 6

Condition on the construction information $\mathcal{F}_0$, so the classifiers are fixed. For $u \in \{-1, 0, 1\}$ and $0 < \lambda < 1$, put $\psi(\lambda) = -\log(1-\lambda) - \lambda$. Then

$$\exp\{\lambda u - \psi(\lambda) u^2\} \le 1 + \lambda u. \tag{73}$$

The cases $u = 0$ and $u = -1$ are equalities. For $u = 1$, the inequality is equivalent to $2\lambda \le \log((1+\lambda)/(1-\lambda))$, which follows by integrating $2/(1-z^2) \ge 2$ from zero to λ. This is a binary-loss specialization of the self-normalized exponential construction used in confidence sequences (Howard et al., 2018; 2021).

Let $a_{r,t} = \mathbb{E}[U_{r,t} \mid \mathcal{F}_{t-1}]$. Taking conditional expectations in Equation (73) gives

$$\mathbb{E}[\exp\{\lambda U_{r,t} - \psi(\lambda) U_{r,t}^2\} \mid \mathcal{F}_{t-1}] \le 1 + \lambda a_{r,t} \le \exp(\lambda a_{r,t}). \tag{74}$$

Consequently,

$$M_{r,t}^{\lambda} = \exp\left\{ \lambda \left(S_{r,t} - \sum_{i=1}^{t} a_{r,i} \right) - \psi(\lambda) D_{r,t} \right\}, \qquad M_{r,0}^{\lambda} = 1, \tag{75}$$

is a nonnegative supermartingale under arbitrary adapted dependence. Ville's inequality yields $\Pr(\sup_t M_{r,t}^{\lambda} \ge JG/\alpha \mid \mathcal{F}_0) \le \alpha/(JG)$. A union bound over J candidates and G fixed scales,

followed by averaging over $\mathcal{F}_0$, gives probability at least $1-\alpha$ for all inequalities

$$\sum_{i=1}^{t} a_{r,i} \geq S_{r,t} - \frac{\psi(\lambda)}{\lambda} D_{r,t} - \frac{\log(JG/\alpha)}{\lambda} \quad \text{for every } r, t, \lambda. \tag{76}$$

Maximizing the right side over scales and dividing by t proves the theorem. No independence between candidate comparisons, no exchangeability of panel entities, and no stationary conditional mean is used. Adaptive choice among candidates with positive lower bounds is covered by the same simultaneous event. A zero-disagreement candidate cannot qualify because its lower bound is strictly negative.

The meaning of the target is part of the result. Each $a_{r,t}$ is a conditional expected reduction in binary loss before the next audit example is revealed, with the filtration containing earlier revealed examples. Their average describes this observed sequence of prediction opportunities. Under independent observations from a common law, it is the fixed population difference $R(H_0) - R(H_r)$. Under arbitrary dependence, it is generally random and history-specific. The theorem does not transform retrospective audit evidence into a guarantee for a shifted future suffix. Including additional predictable information in the filtration changes the conditional target; the same proof applies when that information is declared before the corresponding outcome.

G.2 Power in the independent comparison model

The temporal certificate is more conservative than the exact conditional-binomial rule in the independent model, because it protects all times and permits changing conditional means. It nevertheless retains disagreement-sensitive power over the gain range covered by its fixed scale grid. Suppose $\mathbb{E}U_t = \Delta > 0$, $\mathbb{E}U_t^2 = \omega$, and the observations are independent. Let $L_0 = \max\{\log(JG/\alpha), \log(2/\epsilon)\}$. If the grid contains a scale satisfying

$$\frac{\Delta}{8\omega} \leq \lambda \leq \frac{\Delta}{4\omega}, \tag{77}$$

then a sufficient label budget for qualification with probability at least $1-\epsilon$ is

$$k \geq \frac{64\omega}{\Delta^2} L_0. \tag{78}$$

Indeed, the Bernstein and Chernoff bounds in Appendix F make $S_k > k\Delta/2$ and $D_k \leq 2k\omega$ hold together with probability at least $1-\epsilon$ under this budget. Since $\lambda \leq 1/4$,

$$\frac{\psi(\lambda)}{\lambda} = \sum_{j=2}^{\infty} \frac{\lambda^{j-1}}{j} \leq \frac{\lambda}{2(1-\lambda)} \leq \frac{2}{3}\lambda. \tag{79}$$

On those events, the lower-bound numerator at this scale exceeds

$$\frac{k\Delta}{2} - \frac{4}{3}\lambda k\omega - \frac{L_0}{\lambda} \geq \frac{k\Delta}{6} - \frac{8\omega L_0}{\Delta} > 0. \tag{80}$$

This proves Equation (78). A fixed finite grid need not satisfy Equation (77) for arbitrarily small Δ/ω. The exact paired theorem, not this finite-grid statement, supplies the unrestricted fixed-confidence comparison lower and upper bounds.

G.3 Geometric scales and their coverage range

Corollary 12 (Coverage of a finite dyadic grid)**.** *Let $K \geq 2$ and $\Lambda_K = \{2^{-j} : j = 1, \ldots, K\}$. If $0 < \Delta \leq \omega$ and $\Delta/\omega \geq 2^{2-K}$, this grid contains a scale in Equation* (77)*. Consequently, Equation* (78) *holds with $G = K$. For $K = 9$, the sufficient coverage range is $\Delta/\omega \in [1/128, 1]$.*

Proof. Put $r = \Delta/\omega$ and $j = \lceil \log_2(4/r) \rceil$. Since $r \leq 1$, $j \geq 2$; since $r \geq 2^{2-K}$, $j \leq K$. The defining ceiling gives $r/8 < 2^{-j} \leq r/4$, with the endpoints of the required closed interval also allowed. Apply the power argument above. The additional logarithmic term for replacing $K = 5$ by $K = 9$ is exactly $\log(9/5)$, not an additional label count. □

The lower cutoff is a sufficient coverage condition, not a necessary condition for positive power. A finite grid still excludes arbitrarily small ratios. A countable mixture or stitched boundary can remove a fixed cutoff with scale-dependent error allocation (Howard et al., 2021); the experiments here retain the explicit nine-scale finite rule and its exact correction.

G.4 An audit-only guarantee cannot identify future transfer

Proposition 13 (Indistinguishable audit laws and opposite future gains)**.** *Fix two classifiers H_0, H_1 that disagree at some input. Let a rule approve H_1 using only an audit history, with probability p under an audit law P_A. Without a restriction linking audit and future laws, two extensions of the same P_A can make H_1 strictly better or strictly worse than H_0 on every future prediction. In the harmful extension, unconditional harmful-deployment probability is p, and its conditional probability given approval is one when $p > 0$.*

Proof. Choose an input x where the predictions differ. The first extension retains the complete audit law and sets every future input to x with label $H_1(x)$. The second has the identical audit law and future inputs but labels $H_0(x)$. The risk differences are $+1$ and -1, respectively. The audit-measurable approval event has the same probability p under both extensions. In the second, that event is exactly harmful deployment. An independent switch between the two extensions with probability q therefore yields harmful-deployment probability qp. No audit procedure can distinguish these extensions before future information is supplied. □

The proposition does not challenge the conditional-risk guarantee in Theorem 6; it identifies a different target. For example, an independently justified bound $\Delta_{r,B} \geq \overline{\Delta}_{r,k} - b_r$ would transfer the guarantee whenever $\underline{\Delta}_{r,k} > b_r$. Setting $b_r = 0$ without such a link would assert the transfer assumption rather than establish it. The regime-shift experiment quantifies this boundary with bounded, nondeterministic outcomes rather than the extremal construction in the proof.

H Numerical Protocols and Extended Results

H.1 A joint known-risk family for both safeguards

Let $Y \in \{-1, 1\}$ satisfy $\Pr(Y = 1) = .30$. Conditional on Y, draw $J = 12$ independent binary features with

$$\Pr(X_j = 1 \mid Y = 1) = \frac{\omega + \Delta_j}{.60}, \qquad \Pr(X_j = 1 \mid Y = -1) = \frac{\omega - \Delta_j}{1.40}. \tag{81}$$

The evaluated parameters keep both probabilities in $[0, 1]$. Freeze $H_0 = -1$ and $H_j = 2X_j - 1$. Then

$$R(H_0) = .30, \quad R(H_j) = .30 - \Delta_j, \quad \Pr(H_j \neq H_0) = \omega. \tag{82}$$

Distinct candidate coordinates remain distinct classifiers even when their population risks coincide. In the mixed family, $(\Delta_1, \ldots, \Delta_{12}) = (.01, 0, 0, 0, 0, 0, -.01, -.01, -.01, -.01, -.01, -.01)$. The all-null and all-harmful controls set every gain to 0 or $-.01$. The positive control sets the first gain to .08 and leaves the other gains as in the mixed family.

For each complete independent validation sample, generate $N_+ \sim \mathrm{Bin}(k, .30)$. Conditional on N_+, the exact joint sufficient counts are

$$B_j \sim \mathrm{Bin}\left(N_+, \frac{\omega + \Delta_j}{.60}\right), \qquad C_j \sim \mathrm{Bin}\left(k - N_+, \frac{\omega - \Delta_j}{1.40}\right), \tag{83}$$

independently across coordinates within each label group. This is distributionally identical to drawing all features on the same k labeled points; it retains the common baseline and the dependence among candidate comparisons. Both safeguards consume these identical counts. Table 3 gives deployment probabilities and risks; Table 5 preserves their exact counts. Every rule selects the approved candidate with largest empirical gain and then the smallest candidate index on ties. Population deployed risk is exactly $.30 - \Delta_{\mathrm{selected}}$, with gain zero at fallback.

Settings are enumerated by increasing $\omega \in \{.05, .10, .20, .40\}$, with $k = 2{,}000$ then $8{,}000$ within each setting, followed by the null, harmful, and positive controls. Setting index $i = 0, \ldots, 10$ uses NumPy's `default_rng(31000+i)` and 20,000 repetitions. Binomial survival functions avoid explicit summation of large binomial coefficients. No candidate is chosen or changed using these repetitions.

Table 3: Known-risk protected deployment on the same frozen family and validation points. Mixed settings have $\Delta = .01$, $M = 13$, $R(H_0) = .30$, and 20,000 repetitions. P_+ is useful deployment, P_- harmful deployment, and P_0 fallback. The remaining probability is deployment of an equal-risk candidate. Separate intervals have $P_+ = P_- = 0$, $P_0 = 1$, and deployed risk .30 in every row. Parentheses give the Monte Carlo standard error of deployed risk.

ω	k	Paired P_+	Paired P_-	Paired P_0	Paired risk
.05	2,000	.2308	0.00000	.7580	.297692 (.000030)
.05	8,000	.9078	0.00000	.0901	.290922 (.000020)
.10	2,000	.0970	0.00000	.8877	.299030 (.000021)
.10	8,000	.5575	0.00000	.4353	.294425 (.000035)
.20	2,000	.0453	.00060	.9385	.299553 (.000015)
.20	8,000	.2559	0.00000	.7319	.297441 (.000031)
.40	2,000	.0230	.00130	.9629	.299783 (.000011)
.40	8,000	.1046	.00005	.8854	.298955 (.000022)

Table 4: Control families with $k = 8{,}000$ and 20,000 independent repetitions. Counts are deployments, not candidate-wise rejections. Non-improving deployments include equal-risk and harmful choices. Brackets are marginal 95% Clopper–Pearson intervals for the non-improving deployment probability.

Family	Rule	Useful	Harmful	Fallback	Non-improving probability
All null	Separate	0	0	20,000	0.0000 [0.0000,0.0002]
All null	Paired	0	0	19,152	0.0424 [0.0396,0.0453]
All harmful	Separate	0	0	20,000	0.0000 [0.0000,0.0002]
All harmful	Paired	0	0	20,000	0.0000 [0.0000,0.0002]
Positive control	Separate	20,000	0	0	0.0000 [0.0000,0.0002]
Positive control	Paired	20,000	0	0	0.0000 [0.0000,0.0002]

The largest observed harmful-deployment rate among the mixed settings is $26/20{,}000 = .00130$. A zero harmful count has a two-sided 95% Clopper–Pearson upper endpoint of .0001844, rather than proving zero population harm. The intervals in Table 4 describe Monte Carlo uncertainty within each setting and are not simultaneous across the entire experiment. For a count s among R repetitions, the endpoints are beta quantiles at .025 and .975, with the usual endpoints zero at $s = 0$ and one at $s = R$.

Table 5: Run counts underlying Table 3. Every row has 20,000 repetitions and uses the paired safeguard. Equal-risk deployment is separated from fallback and harm.

ω	k	Useful	Harmful	Equal risk	Fallback
.05	2,000	4,615	0	226	15,159
.05	8,000	18,156	0	42	1,802
.10	2,000	1,940	0	307	17,753
.10	8,000	11,149	0	146	8,705
.20	2,000	906	12	312	18,770
.20	8,000	5,117	0	245	14,638
.40	2,000	460	26	256	19,258
.40	8,000	2,091	1	199	17,709

H.2 Exact qualification power and validation cost

For a fixed candidate, $D \sim \mathrm{Bin}(k, \omega)$ and $B \mid D = d \sim \mathrm{Bin}(d, q)$, where $q = (1 + \Delta/\omega)/2$. Let

$$c_{\mathrm{pair}}(d) = \min\{b : \Pr(\mathrm{Bin}(d, 1/2) \geq b) \leq \alpha/J\}, \tag{84}$$

$$c_{\mathrm{sep}}(d, k) = \left\lfloor \frac{d + k\sqrt{2\log(2M/\alpha)/k}}{2} \right\rfloor + 1. \tag{85}$$

Table 6: Exact independent-label budgets attaining 90% marginal qualification power for gain $\Delta = .01$, $M = 13$, and $\alpha = .05$. Powers integrate over the random discordant count. The normalized budget measures the ω/Δ^2 scaling.

Disagreement ω	.05	.10	.20	.40
Paired labels k	7,803	15,528	30,911	61,649
$k\Delta^2/\omega$	15.61	15.53	15.46	15.41

An unattainable critical count is interpreted as giving zero rejection probability. The marginal powers are

$$\mathcal{P}_s(k;\Delta,\omega) = \sum_{d=0}^{k} \Pr\{\mathrm{Bin}(k,\omega) = d\} \Pr\{\mathrm{Bin}(d,q) \geq c_s(d,k)\}. \tag{86}$$

The paired critical count does not otherwise depend on k. Numerical summation retains discordant counts between binomial quantiles 10^{-13} and $1 - 10^{-13}$, bounding omitted mass by 2×10^{-13}. Starting at $k = 64$, doubling finds an upper budget with power at least .90; integer bisection finds a crossing between the bracketing budgets. Because discrete power can have small oscillations, these are verified attaining budgets, not claims of globally minimal integer sample size. No normal approximation is used.

Table 7: Complete exact-power sweep. The normalized budget is $k_{\mathrm{pair}}\Delta^2/\omega$ and the ratio is $k_{\mathrm{sep}}/k_{\mathrm{pair}}$. Each displayed budget attains at least 90% marginal qualification power for its rule under Equation (86).

Δ	ω	Paired k	Separate k	Ratio	Normalized
.005	.05	31,057	584,713	18.83	15.53
.005	.10	61,822	621,594	10.05	15.46
.005	.20	123,298	675,504	5.48	15.41
.005	.40	246,228	755,827	3.07	15.39
.010	.05	7,803	146,194	18.74	15.61
.010	.10	15,528	155,418	10.01	15.53
.010	.20	30,911	168,898	5.46	15.46
.010	.40	61,649	188,981	3.07	15.41
.020	.05	1,944	36,552	18.80	15.55
.020	.10	3,901	38,861	9.96	15.60
.020	.20	7,764	42,207	5.44	15.53
.020	.40	15,455	47,228	3.06	15.46

With M frozen classifiers and a directly evaluated sample of size k, certification requires k labels and Mk binary loss evaluations; both rules share these predictions. Exact-tail evaluation adds one binomial survival-function call per nonbaseline candidate. These are operation counts, not measured latency reductions. Training, graph construction, and weight calibration are excluded from this known-risk comparison because its classifiers are fixed analytically. All candidates incur the same zero fitting cost. For a dependent trajectory, raw acquisition cost is N, not k, and the coupling allowance must be added before using a nominal confidence level.

H.3 Covariate-refresh experiment

The training law is exactly Equation (2). For each (n, T), set $d = 8$ and $\mu_x = .5\sqrt{dT/n}(-1)^{x-1}$. Start from stationarity. Generate independent geometric refresh-block lengths with success probability $1/T$, truncate the last block at observation n, and draw a new (X, Y) from π_θ for each block. This representation is exact and requires no burn-in. At $T = 1$, a multinomial count draw is equivalent for the evaluated count-based learners.

For each cell, compute the conditional bootstrap-majority probability q_{31} exactly and draw the number of positive voters as $\mathrm{Bin}(m, q_{31})$, with $m = \lceil n/(dT) \rceil$. This exactly reproduces the finite hard vote without separately materializing every bootstrap label. Empty cells predict $+1$. Compute excess risk by averaging $|\mu_x|$ over incorrectly classified cells, and compute Bayes risk

as $R^* = (1 - d^{-1}\sum_x |\mu_x|)/2$. Thus no estimated Bayes risk or finite test-label noise enters this experiment. Each setting uses 2,000 independent trajectories from `default_rng(41000+n+T)`; a continuous generator stream supplies the repetitions.

Table 8: Exact population excess risk on the covariate-refresh family, averaged over 2,000 independent training trajectories. Parentheses are Monte Carlo standard errors. The final column divides finite-vote risk by $\sqrt{dT/n}$. All settings use $d = 8$ and $b = 31$; margins follow the local-alternative rule in the text.

T	n	m	Empirical cell sign	Uniform finite vote	Normalized
1	2,048	256	.00974 (.00011)	.00978 (.00011)	.1564
1	8,192	1,024	.00481 (.00006)	.00490 (.00006)	.1567
1	32,768	4,096	.00245 (.00003)	.00248 (.00003)	.1589
8	2,048	32	.03118 (.00033)	.03131 (.00033)	.1771
8	8,192	128	.01554 (.00017)	.01558 (.00017)	.1763
8	32,768	512	.00791 (.00009)	.00791 (.00009)	.1791
32	2,048	8	.06251 (.00069)	.06303 (.00068)	.1783
32	8,192	32	.03135 (.00033)	.03153 (.00033)	.1784
32	32,768	128	.01573 (.00016)	.01578 (.00016)	.1786

H.4 Recursive temporal-graph audit

For each $n \in \{63, 64, 127, 128, 255, 256, 511, 512\}$, construct either $W_{ij} = \mathbf{1}\{0 < |i - j| \le w\}$, with $w \in \{2, 4, 8\}$, or $W_{ij} = \exp(-|i - j|/\tau)$ for $i \ne j$, with $\tau \in \{1, 4, 16\}$. Diagonal entries are zero. Apply four recursive bisection levels using both the combinatorial and symmetric normalized Laplacians, recomputing the operator on the actual induced child vertices. Each root contributes fifteen splits, giving 120 splits for each family and normalization and 1,440 overall. The eigensolver uses `scipy.linalg.eigh`, driver `evr`, and the first four eigenpairs. Orient the second vector to have a nonnegative first coordinate and use a stable decreasing-coordinate sort; no midpoint substitution is used.

A split is contiguous only when both returned children are intervals in the original index order. Its midpoint distance is half the symmetric difference between the selected child and the first $\lceil s/2 \rceil$ current vertices; for even s, take the smaller distance under exchanging child names. This is a misassigned-vertex distance, not an asserted cut location for a noninterval. A tied median has a gap between the two relevant ordered coordinates at most 10^{-10}. A repeated Fiedler eigenvalue is detected when $\lambda_3 - \lambda_2 \le 10^{-9} \max\{1, \|L\|_{\mathrm{op}}\}$. The record retains both diagnostics, the actual child IDs, and Euclidean eigenpair residuals.

All 96 root splits are midpoint intervals, and all 720 exponential-kernel splits are contiguous. The band exceptions occur deeper in recursion on subgraphs of size 7, 8, or 15. Width eight produces 16 repeated-Fiedler cases per normalization; width four has none, but tied median coordinates can still give a noninterval under floating-point sorting. All 51 noncontiguous splits lie in the union of these two ambiguity conditions. Every exceptional interval is below the $s \ge 2w + 1$ threshold in Corollary 4. Thus the band theorem explains where strict ordering is guaranteed and where the audited implementation can encounter ambiguity; it does not remove the recorded undersized cases. Table 9 includes these outcomes rather than filtering them out. The counterexample of Appendix D.3 has a simple Fiedler eigenspace and shows a separate structural departure.

A second audit varies $w = 1, \ldots, 16$ and all consecutive sizes from $\max\{3, w + 1\}$ through $\max\{3, w + 1\} + 64$. The theorem regime contains 905 checks per Laplacian. All 1,810 balanced splits are midpoint splits. Every combinatorial vector is strictly monotone and every normalized generalized vector $D_{\mathrm{deg}}^{-1/2} v$ is strictly monotone. For the combinatorial and normalized operators, respectively, the smallest adjacent oriented differences are .0003797 and .0001394, the smallest second-to-third eigenvalue gaps are .006590 and .003395, and the maximum Euclidean eigenpair residuals are 2.18×10^{-14} and 1.57×10^{-15}. Reflection antisymmetry is also checked. These are finite-precision checks of the specified interval family, not a substitute for the proof. The executable `band_extension.py` records the undersized cases separately rather than treating them as theorem failures.

Table 9: Recursive temporal-graph audit. Each entry includes all 120 splits, including ties. C and N denote combinatorial and symmetric normalized Laplacians. Distance is the maximum number of vertices misassigned relative to the midpoint split, up to exchanging equal-size children.

Affinity	Contiguous C	Contiguous N	Max. distance C/N
Band $w = 2$	$120/120$	$120/120$	$0/0$
Band $w = 4$	$119/120$	$104/120$	$1/1$
Band $w = 8$	$102/120$	$104/120$	$2/2$
Exponential $\tau = 1$	$120/120$	$120/120$	$0/0$
Exponential $\tau = 4$	$120/120$	$120/120$	$0/0$
Exponential $\tau = 16$	$120/120$	$120/120$	$0/0$

H.5 Learning the heterogeneous-score graph

Use $m = 32$ scores and independent Rademacher coordinates $S, U_0, U_A, U_B, U_1, \ldots, U_{32}$. Set $\eta(X) = .15S$, sample $Y \in \{-1, 1\}$ with this conditional mean, and define

$$f_j(X) = \eta(X) + .03U_0 + .10U_j + \begin{cases} .60U_A, & j \le 16, \\ .04U_B, & j > 16. \end{cases} \tag{87}$$

Every score has absolute value at most $.88$, and the Bayes risk is $.425$. The two residual groups specify the data-generating law; their identities are not passed to the learned graph. For each repetition, an independent unlabeled pilot of 256 score vectors supplies squared distances

$$D_{ij}^{\mathrm{head}} = \frac{1}{256}\sum_{t=1}^{256}(f_i(X_t) - f_j(X_t))^2, \qquad W_{ij} = \exp(-D_{ij}^{\mathrm{head}}/\tau), \quad W_{ii} = 0, \tag{88}$$

where τ is the median off-diagonal distance, floored at 10^{-12}. The projector is onto the two lowest eigenvectors of the combinatorial head Laplacian, including the constant vector. Its rank is fixed before calibration.

Compare uniform weights, full residual stacking, learned-graph projection, full residual stacking with the same learned-graph penalty dI, and an independent random rank-two projection. The random projection contains the normalized constant vector and a Gaussian direction projected onto its orthogonal complement; its own discrepancy d is used in the same robust objective. Every fitted method has the same calibration labels and the cap $\|w\|_2^2 \le 2/32$. The graph and its matched-penalty control receive extra unlabeled pilot scores; this acquisition cost is not counted as a reduction in total observations or runtime.

At each $k_c \in \{32, 128, 512\}$, use 100 independent pilot/calibration repetitions. Calibration seeds are $92000 + 100k_c + s$, pilot seeds $120000 + 100k_c + s$, and random-projection seeds $150000 + 100k_c + s$, with $s = 0, \ldots, 99$. A common test-feature pool of $50,000$ observations has seed 91000. Excess risk is the pool average of $.15\mathbf{1}\{\mathrm{sign}(w^\top f(X)) \ne \mathrm{sign}\, S\}$, integrating over the known conditional label law; squared error averages $(w^\top f(X) - \eta(X))^2$. Thus there is no sampled test-label noise. The test-feature pool is fixed across repetitions, so uncertainty is conditional on that pool. Uniform weighting has no pilot/calibration variation.

The mean paired learned-minus-full excess differences at $k_c = 32, 128, 512$ are $-.001313$, $-.000991$, and $-.000298$, respectively. Their marginal 95% Student intervals are $[-.001671, -.000955]$, $[-.001208, -.000773]$, and $[-.000375, -.000221]$. These use the standard deviation of the one hundred paired differences, divided by ten, and a t_{99} critical value. They are not simultaneous across calibration sizes or comparators. The corresponding learned-minus-matched-penalty intervals are $[-.001947, -.001180]$, $[-.001445, -.000953]$, and $[-.000422, -.000256]$. The gain therefore survives this specific penalty control, but does not establish a graph advantage outside the constructed score family.

H.6 Scalar-series targets and chronological allocation

The scalar-series comparisons use packaged, versioned observations rather than repeated draws from an audit pool. Climate and electricity enter the main comparison; the smaller El Niño and sunspot

Table 10: Heterogeneous-score ablations. Excess risk is mean ± standard deviation over 100 paired pilot/calibration repetitions, conditional on the same 50,000 test features. Mean squared error integrates the known regression function. s_A is total mass on the noisier group, used only for analysis. Neither learned graph nor random projection receives group labels.

k_c	Weighting	Excess risk	Squared error	s_A
32	Uniform	$.07509 \pm 0.00000$	.09173	.5000
32	Full residual	$.02242 \pm .02921$	.02595	.1783
32	Learned graph	$.02111 \pm .02981$	.02511	.1597
32	Full with matched penalty	$.02267 \pm .02940$	.02613	.1794
32	Random rank two	$.07509 \pm .00001$	.08915	.4926
128	Uniform	$.07509 \pm 0.00000$	.09173	.5000
128	Full residual	$.00830 \pm .01349$	.00870	.0986
128	Learned graph	$.00731 \pm .01331$	.00794	.0814
128	Full with matched penalty	$.00851 \pm .01378$	.00880	.0996
128	Random rank two	$.07509 \pm .00001$	.08750	.4877
512	Uniform	$.07509 \pm 0.00000$	.09173	.5000
512	Full residual	$.00140 \pm .00269$	.00443	.0491
512	Learned graph	$.00110 \pm .00254$	.00415	.0362
512	Full with matched penalty	$.00144 \pm .00279$	.00446	.0495
512	Random rank two	$.07509 \pm 0.00000$	.08878	.4916

tasks are retained below as limited-budget secondary checks. The weekly Mauna Loa series contains $2,284$ dates and $2,225$ observed concentrations (Keeling & Whorf, 2004). The electricity task uses the `mstl_elec_vic.csv` fixture distributed with statsmodels $0.14.6$, containing 3,600 hourly values from January 1–May 29, 2012. It is the Victoria demand series used in the package's MSTL example, aggregated to hourly sums of the half-hour demand measurements (statsmodels Developers, 2026a). We use this exact 150-day fixture, not the longer source collection or the differently truncated documentation example.

The El Niño task uses 732 monthly Niño 1+2 sea-surface temperatures from 1950–2010, distributed from NOAA's ERSST.V3B source (statsmodels Developers, 2026b). Months are read in calendar order. The sunspot task uses the packaged historical series of 309 annual observations from 1700–2008, rather than a later revision of the international sunspot number (statsmodels Developers, 2026c). Exact value-array hashes below identify all four data versions.

For any scalar series c_t, include an example only when $c_{t-5}, \dots, c_{t+1}$ are finite. Its target is $y_t = \mathbf{1}\{c_{t+1} > c_t\}$; ties are zero, and residual fitting converts this encoding to $\{-1, 1\}$. Its initial six features are

$$(c_t, c_t - c_{t-1}, c_{t-1} - c_{t-2}, \dots, c_{t-4} - c_{t-5}). \tag{89}$$

For weekly climate, append annual sine and cosine at phase $2\pi\,\mathrm{dayofyear}(t)/365.25$. For monthly El Niño append the phase $2\pi(\mathrm{month}(t) - 1)/12$. For hourly electricity append daily phase $2\pi\,\mathrm{hour}(t)/24$ and weekly phase $2\pi(\mathrm{weekday}(t) + \mathrm{hour}(t)/24)/7$, each as sine and cosine. Annual sunspots use no calendar phase. There is no interpolation, scaling fit, or look-ahead transformation.

For q eligible examples, boundaries are $0, \lfloor .50q \rfloor, \lfloor .65q \rfloor, \lfloor .80q \rfloor, q$. Remove the first six eligible examples from each non-training split. This prevents shared lag-target windows across boundaries but does not prove independence of the remaining observations. Table 11 gives exact retained counts and learner budgets; Table 12 gives the periods. Dates refer to feature times, not next-step target times.

H.7 Small-budget secondary series

Table 13 retains the complete scalar-series comparison. El Niño supplies only 103 audit and 140 later examples; sunspots supplies 40 and 55. Several mean differences therefore represent only a few changed labels, and neither task approves a replacement under the exact-binomial reference. They do not support the main claim about useful real-budget certification. The larger production and housing panels in Appendix I test that question with more than two thousand audit rows, while retaining dependence explicitly in the target.

Table 11: Chronological allocation and matched budgets. Split sizes exclude six-example embargoes. b is presentations per tree, L is expected stationary or fixed circular block length, and leaf is the minimum leaf size. Every fitted ensemble has 64 trees.

Task	Train	Calibration	Audit	Test	b	L	Leaf
Climate	1050	309	309	415	262	11	8
Electricity	1797	533	533	713	449	13	8
El Niño	363	102	103	140	90	8	5
Sunspots	151	39	40	55	37	6	2

Table 12: Retained chronological periods. Endpoints identify feature dates; the target is the next observation. Positive fraction denotes an increase, not a threshold chosen from later data.

Task	Split	First feature time	Last feature time	Positive fraction
Climate	Train	1958-08-09	1981-07-04	.5133
Climate	Calibration	1981-08-22	1987-11-14	.5437
Climate	Audit	1988-01-02	1993-11-27	.5405
Climate	Test	1994-01-15	2001-12-22	.5036
Electricity	Train	2012-01-01 05:00	2012-03-16 01:00	.5431
Electricity	Calibration	2012-03-16 08:00	2012-04-07 12:00	.4953
Electricity	Audit	2012-04-07 19:00	2012-04-29 23:00	.5009
Electricity	Test	2012-04-30 06:00	2012-05-29 22:00	.3899
El Niño	Train	1950-06-01	1980-08-01	.4821
El Niño	Calibration	1981-03-01	1989-08-01	.4608
El Niño	Audit	1990-03-01	1998-09-01	.4854
El Niño	Test	1999-04-01	2010-11-01	.4714
Sunspots	Train	1705	1855	.4172
Sunspots	Calibration	1862	1900	.3846
Sunspots	Audit	1907	1946	.4000
Sunspots	Test	1953	2007	.3455

Table 13: Scalar-series later errors, mean ± standard deviation over ten learner seeds on fixed periods. The smaller El Niño and sunspot tasks have only 140 and 55 test examples. All methods use 64 trees and matched row-presentation budgets within each task. Stacking reuses the uniform heads. These standard deviations do not quantify uncertainty across independent series.

Method	Climate	Electricity	El Niño	Sunspots
Uniform bagging	$.2800 \pm .0055$	$.3386 \pm .0286$	$.1071 \pm .0095$	$.1564 \pm .0176$
Contiguous / path, $P = 4$	$.2930 \pm .0095$	$.3473 \pm .0282$	$.1007 \pm .0086$	$.1273 \pm .0284$
Random balanced, $P = 4$	$.2812 \pm .0062$	$.3391 \pm .0272$	$.1007 \pm .0137$	$.1727 \pm .0312$
Contiguous / path, $P = 8$	$.2957 \pm .0132$	$.3463 \pm .0297$	$.0979 \pm .0048$	$.1673 \pm .0372$
Random balanced, $P = 8$	$.2858 \pm .0092$	$.3238 \pm .0168$	$.1036 \pm .0084$	$.1982 \pm .0200$
Stationary bootstrap	$.2827 \pm .0062$	$.3480 \pm .0220$	$.1007 \pm .0092$	$.1382 \pm .0334$
Circular block bootstrap	$.2853 \pm .0130$	$.3407 \pm .0299$	$.0986 \pm .0074$	$.1564 \pm .0287$
Full residual stacking	$.2742 \pm .0062$	$.3133 \pm .0180$	$.1043 \pm .0096$	$.1382 \pm .0260$
Graph residual stacking	$.2793 \pm .0061$	$.3311 \pm .0205$	$.1036 \pm .0061$	$.1509 \pm .0243$

H.8 Matched learners, resampling, and graph calibration

Every routing or resampling family uses $m = 64$ trees, maximum depth four, square-root feature subsampling, and $b = \lfloor n_{\mathrm{tr}}/4 \rfloor$ row presentations per tree. Minimum leaf size is $\max\{2, \min\{8, \lfloor b/16 \rfloor\}\}$. These rules are fixed across datasets; they are not selected on audit or later labels. For learner seed $s = 0, \ldots, 9$ and head $j = 0, \ldots, 63$, bootstrap and fitting seeds are $61000 + 1000s + j$ and $71000 + 1000s + j$.

Uniform bagging samples rows independently from the entire training split. Path-four and path-eight use recursive midpoint intervals; their contiguous copies are removed by Theorem 3. Random-balanced partitions use a permutation generated by seed $51000 + s$, first for four groups and then for eight, with the same group-size multiset as the corresponding midpoint partition. Groups are canonically ordered by their smallest index, and head j uses group $j \bmod P$. Scores are $2\widehat{p}_j - 1$ and unweighted aggregation thresholds their mean at zero. The number of row presentations is matched, while the number of distinct training points seen by a tree can differ as part of the resampling intervention.

For both block-bootstrap methods, set $L = \lceil n_{\mathrm{tr}}^{1/3} \rceil$ from training size alone. Stationary bootstrap starts at a uniformly chosen training index, continues to the next index modulo n_{tr} with probability $1 - 1/L$, and otherwise restarts uniformly. Circular bootstrap concatenates length-L circular blocks with independent uniform starts. Both stop after exactly b row presentations. These schemes are the stationary and circular resampling constructions (Politis & Romano, 1991; 1994), applied to fitting the same base learner. L is a deterministic baseline setting, not an estimate of a transition mixing time.

Full and graph stacking reuse the uniform ensemble's 64 scores. Their empirical residual matrix uses only calibration data. For the graph, compute Equation (88) on the uniform heads' training predictions, replacing the pilot size by n_{tr}, and take the lowest eight Laplacian eigenvectors. Both weight fits impose $\|w\|_2^2 \le 2/64$ and the simplex constraint. Thus there are nine distinct specifications: uniform, path-four, random-four, path-eight, random-eight, stationary bootstrap, circular bootstrap, full stacking, and graph stacking. The residual weights and all candidate choices are frozen before audit outcomes are evaluated.

Seven families require 64 new trees each; the two stacking candidates reuse existing trees. Total fitting therefore uses $448b$ row presentations per seed and task, before score evaluation and calibration. Certification of the frozen nine candidates uses k acquired labels and $9k$ binary loss evaluations. All construction, score evaluation, eigensolving, and weight fitting are additional costs. These are operation counts; no measured latency reduction or total-cost superiority is asserted.

H.9 Actual-budget paired counts and dependence sensitivity

The full-budget reference comparison uses each retained audit example once with $\delta = .05$ and $J = 8$. For every candidate, the same examples determine corrected errors B_r, introduced errors C_r, disagreement D_r, and the exact binomial tail. The reference selector takes the largest empirical

Table 14: Climate audit with each retained label used once. All eight alternatives and ten seeds enter the minimum raw p-value. Testing uses $\delta = .025$, with an additional .025 reserved for coupling. The $\beta(a)$ ceiling assumes $\beta(g) \le .0125$; neither envelope is asserted for climate. The power column uses disagreement .10; a dash means no feasible gain attains 90% power.

Lag a	Labels k	Required $\beta(a) \le$	Min. p	90% power gain	Approvals
1	309	4.06×10^{-5}	.1445	.0704	0/10
2	155	8.12×10^{-5}	.1094	.0941	0/10
4	78	1.62×10^{-4}	.0625	—	0/10
8	39	3.29×10^{-4}	.0625	—	0/10
16	20	6.58×10^{-4}	.2500	—	0/10

gain among qualifiers, then the lowest candidate index on ties. Table 15 reports reference approvals, fallback, and later risk. Their denominator is ten learner seeds on a fixed sequence, not ten independent validation trajectories. There is no known population risk on these real tasks, so a harmful-deployment probability is not estimated by relabeling later test differences as truth.

Table 15: Actual-budget independent-observation reference decisions at $\delta = .05$, $J = 8$. Approvals count seeds with any qualified replacement; the remaining seeds fall back. Errors are mean $\pm$ standard deviation across ten learner seeds on fixed periods. Unknown dependence prevents interpreting these approvals as population-safe certifications.

Task	Decision rule	Approvals	Audit error	Later error
Climate	Uniform	—	$.2615 \pm .0059$	$.2800 \pm .0055$
Climate	Paired reference	0/10	$.2615 \pm .0059$	$.2800 \pm .0055$
Climate	Separate reference	0/10	$.2615 \pm .0059$	$.2800 \pm .0055$
Electricity	Uniform	—	$.3338 \pm .0095$	$.3386 \pm .0286$
Electricity	Paired reference	5/10	$.3180 \pm .0143$	$.3264 \pm .0270$
Electricity	Separate reference	0/10	$.3338 \pm .0095$	$.3386 \pm .0286$
El Niño	Uniform	—	$.0777 \pm .0129$	$.1071 \pm .0095$
El Niño	Paired reference	0/10	$.0777 \pm .0129$	$.1071 \pm .0095$
El Niño	Separate reference	0/10	$.0777 \pm .0129$	$.1071 \pm .0095$
Sunspots	Uniform	—	$.1050 \pm .0258$	$.1564 \pm .0176$
Sunspots	Paired reference	0/10	$.1050 \pm .0258$	$.1564 \pm .0176$
Sunspots	Separate reference	0/10	$.1050 \pm .0258$	$.1564 \pm .0176$

For sensitivity to a coupling allowance, set $\alpha = .05$, $\delta = .025$, reserve .0125 for separating construction from audit, and reserve .0125 for internal audit dependence. At lag $a \in \{1, 2, 4, 8, 16\}$ retain eligible audit indices $0, a, 2a, \ldots$. With k retained points, a sufficient condition is

$$\beta(g) \le .0125, \qquad \beta(a) \le \frac{.0125}{k-1}, \qquad \delta + \beta(g) + (k-1)\beta(a) \le .05. \tag{90}$$

Here g is the separation from the final construction/calibration observation to the audit block, and β refers to the joint example process. The time index is eligible-example order; dropping incomplete windows does not itself establish absolute regularity in that index. Table 16 gives the complete thinning sensitivity. These envelopes are explicit requirements, not fitted or verified bounds for the datasets. We therefore report the resulting test decisions as dependence sensitivity only. The six-example embargo establishes disjoint windows, not the numerical inequalities in Equation (90).

The power entries in Table 14 use the exact independent paired law, with $D \sim \mathrm{Bin}(k, \omega)$ and $B \mid D \sim \mathrm{Bin}(D, 1/2 + \Delta/(2\omega))$, at $J = 8$ and $\delta = .025$. Bisection in $\Delta \in [0, \omega]$ finds the value attaining 90% qualification probability. An empty power entry means that even $\Delta = \omega$ cannot achieve that power at the given k; it does not represent a missing calculation. At disagreement .20, the corresponding climate gains for $k = 309, 155, 78$ are .1018, .1402, and .1868; the smaller budgets remain infeasible. The power calculation assumes the independent target model. Coupling success can transfer it with the stated additional failure probability, but it cannot validate an unknown mixing envelope or a shifted future distribution.

Table 16: Thinning sensitivity for all four series, with $\delta = .025$, $J = 8$, and the coupling requirements in Equation (90). Minimum p is the smallest raw paired tail across the eight comparisons and ten learner seeds. Approvals remain reference decisions until the coupling assumptions are established.

Task	Lag	Retained k	Minimum p	Any approval
Climate	1	309	.14453	0/10
Climate	2	155	.10938	0/10
Climate	4	78	.06250	0/10
Climate	8	39	.06250	0/10
Climate	16	20	.25000	0/10
Electricity	1	533	.00044	4/10
Electricity	2	267	.00014	2/10
Electricity	4	134	.01953	0/10
Electricity	8	67	.01953	0/10
Electricity	16	34	.18750	0/10
El Niño	1	103	.06250	0/10
El Niño	2	52	.12500	0/10
El Niño	4	26	.50000	0/10
El Niño	8	13	1.00000	0/10
El Niño	16	7	1.00000	0/10
Sunspots	1	40	.12500	0/10
Sunspots	2	20	.12500	0/10
Sunspots	4	10	.25000	0/10
Sunspots	8	5	.50000	0/10
Sunspots	16	3	.50000	0/10

Table 17: Finite-pool climate oracle with seven candidates and 20,000 replacement draws from the 309-point audit pool. Errors are mean $\pm$ standard deviation over ten learner seeds. Approvals use $J = 6$ and the fixed pool as target; they do not certify the later test period. Deployed rows include fallback.

Method	Audit error	Later test error	Approvals
Uniform bagging	$.2615 \pm .0059$	$.2800 \pm .0055$	—
Path / contiguous, $P = 4$	$.2751 \pm .0137$	$.2930 \pm .0095$	0/10
Random balanced, $P = 4$	$.2683 \pm .0078$	$.2812 \pm .0062$	1/10
Path / contiguous, $P = 8$	$.2799 \pm .0121$	$.2957 \pm .0132$	0/10
Random balanced, $P = 8$	$.2676 \pm .0057$	$.2858 \pm .0092$	1/10
Full residual stacking	$.2641 \pm .0085$	$.2742 \pm .0062$	3/10
Graph residual stacking	$.2638 \pm .0075$	$.2793 \pm .0061$	1/10
Paired deployment	$.2563 \pm .0030$	$.2814 \pm .0053$	6/10
Separate-interval deployment	$.2615 \pm .0059$	$.2800 \pm .0055$	0/10

H.10 Finite-pool oracle for the earlier climate comparison

For comparison with the actual-budget audit, the finite-pool oracle retains the earlier seven-candidate family without block-bootstrap candidates and uses $J = 6$. For each learner seed, draw $20,000$ indices with replacement from the fixed 309-point climate audit pool, using seed $81000 + s$, and compare both safeguards on the same draws. Conditional on the pool A, these draws are independent from $\pi_A = |A|^{-1} \sum_{i \in A} \delta_{(x_i, y_i)}$. The protected target is therefore the pool risk R_A, not the risk on an independent climate trajectory. All 309 pool labels occur in each run.

The paired oracle approves in $6/10$ seeds, reducing mean pool error from .26149 to .25631, while later error changes from .28000 to .28145 (Table 17). As every pool label is already available, R_A can also be computed without a test. Increasing the number of replacement draws can increase certainty about this empirical ordering without adding a single independent climate observation. This oracle is consequently excluded from claims about real-budget certification or transfer to later data. The actual-budget study adds two block-bootstrap candidates and uses nine-way multiplicity; the family sizes are not silently interchanged.

I LARGER CHRONOLOGICAL PANELS AND TEMPORAL DEPLOYMENT

I.1 DATA IDENTITY, TARGETS, AND OBSERVATION ORDER

The production panel uses the `meat.csv` dataset distributed with plotnine 0.15.3. Its 960 monthly rows cover January 1944 through December 2023. Values are U.S. production weights in millions of pounds, sourced by the package from the USDA Economic Research Service (Plotnine Contributors, 2026). The housing panel uses the package's `txhousing.csv`: 8,602 original records from 187 calendar months. Its `sales` field is the number of housing sales in a Texas multiple-listing-service area, not median price or dollar volume (ggplot2 Contributors, 2026). These are distinct dependent prediction tasks, not surrogate copies of the climate series.

We form complete monthly calendars without interpolating missing values. An entity is retained when its training period contains at least 24 complete lag-target windows. The production panel retains 6 categories: beef, broilers, lamb and mutton, pork, turkey, and veal. Other chicken is excluded by the training-availability rule. Housing retains 38 MLS areas; their names are stored in the numerical manifest. Entity eligibility uses no audit performance. A row is eligible only if its entity's values at $t-5,\ldots,t+1$ are all finite. The target is $y_{e,t}=\mathbf{1}\{c_{e,t+1}>c_{e,t}\}$, with equal values assigned zero.

Features contain the entity's current level, five backward differences, annual sine and cosine, and an entity one-hot vector. Production therefore has 14 features and housing 46. Features use no future value; the next month's value is used only to define the target and determine whether that target exists. We fit no scaling or imputation transform. Within every split, observations are ordered first by calendar month and then by alphabetical entity name. This order defines the filtration in Equation (9). Neither multiple entities in one month nor consecutive months are assumed independent. Predictions from every frozen candidate are evaluated on this same order.

Let T_{cal} be the number of calendar months. Split boundaries are $0,\lfloor.30T_{\mathrm{cal}}\rfloor,\lfloor.40T_{\mathrm{cal}}\rfloor,\lfloor.80T_{\mathrm{cal}}\rfloor,T_{\mathrm{cal}}$. The first six calendar months of each non-training segment are embargoed. Within each segment, only complete eligible rows are used. Thus no lag-target window is shared across the construction, calibration, audit, and later-test boundaries. The embargo does not imply independence, which is why the operational temporal rule uses Theorem 6. The long audit segments were fixed as part of the panel design, before their candidate outcomes were computed. Table 18 reports exact counts and periods.

Table 18: Panel splits after six-month embargoes and complete-window filtering. Rows count entity-month prediction examples; months count calendar cross-sections with at least one example. Dates identify feature months, not target months.

Panel	Split	Rows	Months	First feature	Last feature
Production	Train	1,314	283	1944-06-01	1967-12-01
Production	Calibration	540	90	1968-07-01	1975-12-01
Production	Audit	2,268	378	1976-07-01	2007-12-01
Production	Later test	1,110	185	2008-07-01	2023-11-01
Housing	Train	1,813	51	2000-06-01	2004-08-01
Housing	Calibration	456	12	2005-03-01	2006-02-01
Housing	Audit	2,615	69	2006-09-01	2012-05-01
Housing	Later test	1,167	31	2012-12-01	2015-06-01

I.2 MATCHED FITTING AND CALENDAR-BLOCK COMPETITORS

Every routing or resampling family uses 64 depth-four trees with square-root feature subsampling. The per-tree row-presentation count is $b=\lfloor n_{\mathrm{tr}}/4\rfloor$, giving 328 on production and 453 on housing. The minimum leaf size is eight. For learner seed $s\in\{0,\ldots,9\}$ and head $j\in\{0,\ldots,63\}$, bootstrap and fitting seeds are $61000+1000s+j$ and $71000+1000s+j$. The seed schedules match the scalar-series comparisons. Random-balanced partitions use seed $51000+s$, with four-group then eight-group partitions generated in that order and sizes matched to recursive midpoint groups.

For panel block bootstraps, the sampling unit is a calendar-month cross-section. Let T_{tr} be the number of training months with eligible rows and set $L=\lceil T_{\mathrm{tr}}^{1/3}\rceil$, giving 7 months for production and 4 for

housing. Stationary bootstrap chooses a uniform starting month, continues to the next month modulo T_{tr} with probability $1-1/L$, and otherwise restarts uniformly. Circular bootstrap concatenates blocks of L consecutive circular months with independent uniform starts. Each visited month contributes all its observed eligible entities in alphabetical order; the final cross-section is truncated when the head reaches exactly b row presentations. This preserves the available within-month cross-section except at that budget boundary. It is not an independent bootstrap of entity-month rows. The block-length choice depends only on training size and is not a fitted mixing-time estimate.

The full and graph weighting procedures reuse the same uniform trees. Calibration constructs $\widehat{Q}$ from labeled calibration rows. The graph uses mean squared training-score distances, exponential affinity with median off-diagonal scale, and the lowest eight combinatorial-Laplacian eigenvectors. Its robust objective includes the actual discrepancy $d = \|P\widehat{Q}P - \widehat{Q}\|_{\mathrm{op}}$; full stacking has $P = I$ and $d = 0$. Both satisfy the simplex and $2/64$ squared-norm cap. There are nine candidate specifications after path/contiguous deduplication. All scores, weights, ranks, partitions, and learner seeds are fixed before audit outcomes determine a deployment decision.

Seven fitting families require 448 trees per seed, while both stacking candidates reuse existing trees. Training costs are therefore $448b$ row presentations, excluding prediction, eigensolving, and calibration. An audit of size k acquires k labels and evaluates nine losses per label. The temporal certificate adds nine scalar-scale calculations per candidate; it does not acquire additional labels. These are counted operations, not measured end-to-end latency savings. The full benchmark grid costs more than deploying the uniform-plus-stacking pair alone.

I.3 Candidate-level outcomes and the protected decision

At the fixed audit endpoint, Equation (10) uses $\alpha = .05$, $J = 8$, and $\Lambda_9 = \{2^{-j} : j = 1, \ldots, 9\}$. Every label is used once. The selector chooses the largest empirical gain among candidates with a positive lower bound and uses the original candidate ordering for ties. For comparison, we also apply exact paired-binomial and separate-interval rules at the independent-observation reference level .05. The latter two rules are not interpreted as temporal population certificates. Table 19 reports all candidate-level audit errors and qualification counts. Table 20 gives the full-stacking paired counts and temporal lower bound for every production seed.

Table 19: Panel audit errors and candidate qualification. Errors are mean $\pm$ standard deviation over ten learner seeds; qualification counts positive temporal lower bounds at the full audit budget. Multiple candidates can qualify in one seed. The target is average conditional audit improvement, not later-period population risk.

	Production		Housing	
Method	Audit error	Qualifies	Audit error	Qualifies
Uniform bagging	$.3127 \pm .0100$	—	$.2650 \pm .0055$	—
Contiguous / path, $P = 4$	$.3215 \pm .0097$	$0/10$	$.2667 \pm .0078$	$0/10$
Random balanced, $P = 4$	$.3109 \pm .0139$	$1/10$	$.2683 \pm .0048$	$0/10$
Contiguous / path, $P = 8$	$.3238 \pm .0188$	$0/10$	$.2814 \pm .0110$	$0/10$
Random balanced, $P = 8$	$.3153 \pm .0138$	$0/10$	$.2740 \pm .0043$	$0/10$
Stationary bootstrap	$.3138 \pm .0122$	$1/10$	$.2775 \pm .0217$	$0/10$
Circular block bootstrap	$.3130 \pm .0057$	$0/10$	$.2685 \pm .0127$	$0/10$
Full residual stacking	$.2870 \pm .0097$	$6/10$	$.2802 \pm .0116$	$0/10$
Graph residual stacking	$.3094 \pm .0099$	$0/10$	$.2883 \pm .0109$	$0/10$

On production, full stacking has smaller later error than its matched uniform ensemble in all 10 seeds. Its mean later difference is $-.017658$; a marginal 95% Student interval over the paired learner-seed differences is $[-.022482, -.012833]$. Housing reverses this comparison in all 10 seeds, with difference $+.019623$ and interval $[.012887, .026359]$. These intervals use the standard deviation of ten paired differences divided by $\sqrt{10}$ and the t_9 critical value. They describe learner randomization conditional on fixed data periods; they are neither simultaneous across tasks nor estimates of variation across independent temporal environments.

Production's final temporal selector approves a replacement in 7 seeds: full stacking in 6 and stationary bootstrap in 1. The remaining 3 predictions are exactly their uniform baseline. Every

Table 20: Production full-stacking audit counts at $k = 2,268$, its nine-scale temporal lower bound, and the actual final selection. The last column is selected-minus-uniform error on the untouched later period; zero denotes fallback. Selection uses only audit outcomes. At seed nine, full stacking does not qualify, but stationary bootstrap does.

Seed	B	C	Lower bound	Selected	Later difference
0	97	45	$+0.00066$	Full stacking	-0.0171
1	63	40	-0.00953	Uniform	$+0.0000$
2	135	70	$+0.00221$	Full stacking	-0.0108
3	159	66	$+0.01323$	Full stacking	-0.0207
4	175	78	$+0.01313$	Full stacking	-0.0243
5	87	43	-0.00207	Uniform	$+0.0000$
6	108	49	$+0.00275$	Full stacking	-0.0234
7	67	38	-0.00702	Uniform	$+0.0000$
8	131	66	$+0.00274$	Full stacking	-0.0234
9	112	58	-0.00031	Stationary	-0.0288

actual replacement improves on the fixed later suffix. This count is an observed transfer diagnostic, not an estimate of population harmful-deployment probability. A true future risk is not known on these datasets, and the conditional audit target does not assert stationarity or prevent later distribution shift. Housing, climate, and electricity have no qualifying temporal replacement. Recomputing the climate and electricity endpoint lower bounds from their stored paired counts gives maximum lower bounds of $-.04413$ and $-.01411$, respectively, over all candidates and seeds.

The guarantee is at level $.05$ for each fitted candidate family, simultaneously over its eight alternatives and all audit prefixes. The ten reported learner seeds do not share one $.05$ familywise budget, and no simultaneous cross-dataset guarantee is asserted. Only the fixed endpoint decision is used in the transfer comparison. First-crossing times are logged for diagnostics and do not determine a different test period or early deployment.

I.4 Numerical checks and data hashes

Across the 40 panel weight fits, the maximum capped-simplex violation is 2.22×10^{-16} and the maximum final conditional-gradient gap is 9.16×10^{-8}. Panel graph bases have maximum residual $\|LU - U\Lambda\|_{\mathrm{op}}$ of 3.12×10^{-14}. These float64 diagnostics measure numerical feasibility and objective accuracy, not statistical risk. The programs save per-seed predictions, split indices, entity and month IDs, all paired counts, temporal lower bounds, selected candidates, partition hashes, and objective diagnostics. The raw CSV byte hashes are

Production	`b587c1e758ae43d87fe86fac0efc73da40`
	`aaeaf4046eba370f70fced554564a5`
Housing	`45d1e81f95bd6ee77f0f3b1e7c873cc8`
	`d3856b1325e880c328c88febc6a82286`

J Power and Coverage of Temporal Certification

J.1 The same independent known-risk comparisons

We apply Equation (10) to the exact joint counts of Appendix H.1, using the same setting seeds, labels, candidates, and $20,000$ repetitions. The scale grid is $\Lambda_9 = \{2^{-j} : j = 1, \ldots, 9\}$ and $\alpha = .05$ across $J = 12$ alternatives. Table 21 reports the decisions and deployed risk. The independent risk of a selected candidate is exactly $.30 - \Delta_r$, so this experiment can distinguish useful, harmful, equal-risk, and fallback decisions without estimating future risk. The rule preserves the broader temporal target and simultaneous-time protection. Its scale-grid correction differs from the five-scale ablation in Table 22; all candidate functions and validation outcomes remain matched.

Table 21: Nine-scale temporal-certificate decisions on the independent joint known-risk model, with 20,000 repetitions per row. P_+, P_-, $P_=$, and P_0 denote useful, harmful, equal-risk, and fallback deployment. The final rows are the all-null, all-harmful, and positive-control families. Values are frequencies over complete validation samples; their numerators are retained in the numerical manifest.

Δ	ω	k	P_+	P_-	$P_=$	P_0	Risk
0.010	0.05	2,000	0.00400	0.00000	0.00000	0.99600	0.299960
0.010	0.05	8,000	0.32865	0.00000	0.00005	0.67130	0.296713
0.010	0.10	2,000	0.00165	0.00000	0.00000	0.99835	0.299983
0.010	0.10	8,000	0.09940	0.00000	0.00005	0.90055	0.299006
0.010	0.20	2,000	0.00040	0.00000	0.00000	0.99960	0.299996
0.010	0.20	8,000	0.01235	0.00000	0.00015	0.98750	0.299876
0.010	0.40	2,000	0.00035	0.00000	0.00005	0.99960	0.299996
0.010	0.40	8,000	0.00455	0.00000	0.00010	0.99535	0.299954
0.000	0.10	8,000	0.00000	0.00000	0.00035	0.99965	0.300000
-0.010	0.10	8,000	0.00000	0.00000	0.00000	1.00000	0.300000
0.080	0.20	8,000	1.00000	0.00000	0.00000	0.00000	0.220000

J.2 Exact power of five versus nine scales

For $\Lambda_K = \{2^{-j} : j = 1, \dots, K\}$, define the cumulative evidence boundary

$$b_K(d) = \min_{\lambda \in \Lambda_K} \left\{ \frac{\psi(\lambda)}{\lambda} d + \frac{\log(JK/\alpha)}{\lambda} \right\}. \tag{91}$$

The endpoint qualifies exactly when $2B - D > b_K(D)$. For independent observations, its marginal qualification power is

$$\mathcal{P}_K(k; \Delta, \omega) = \sum_{d=0}^{k} \Pr\{\mathrm{Bin}(k, \omega) = d\} \Pr\left\{ \mathrm{Bin}\left(d, \frac{1}{2} + \frac{\Delta}{2\omega} \right) \geq \left\lfloor \frac{d + b_K(d)}{2} \right\rfloor + 1 \right\}. \tag{92}$$

This integrates the random discordant count and uses the strict positivity decision without a normal approximation. Summation between binomial quantiles 10^{-13} and $1 - 10^{-13}$ omits mass at most 2×10^{-13}. Both grids use $\alpha = .05$, $J = 12$, the same classifier and target law, and their own correct scale correction. Table 22 isolates the scale change from any fitting or label change.

Table 22: Exact fixed-candidate marginal qualification power at $\Delta = .01$, $\omega = .40$, $J = 12$, and $\alpha = .05$. The five-scale grid ends at $1/32$; the nine-scale grid ends at $1/512$. Both protect all audit times. These are exact-power calculations, not frequencies of winning a multicandidate selection. The displayed value 1.000000 is rounded.

Labels k	Five scales	Nine scales
8,000	0.007387	0.004602
32,000	0.162862	0.125276
128,000	0.851642	0.955358
512,000	0.999836	1.000000

At the mean discordant count $d = k\omega$ and $k = 8,000$, both grids minimize their boundary at $\lambda = 1/16$. Thus adding smaller scales does not reduce that boundary, while its error correction increases. At $k = 128,000$, the minimizing scales are $1/32$ and $1/64$, respectively. This accounts for the change of ordering in Table 22. The conservative power-proof scale $1/256$ and the boundary-minimizing scale answer different questions and need not coincide.

Doubling and integer bisection find verified attaining 90% budgets. At $\Delta = .01$ and disagreements $.05, .10, .20, .40$, the five-scale budgets are $13,629$, $26,620$, $52,620$, and $149,304$; the nine-scale budgets are $14,502$, $28,322$, $55,991$, and $111,337$. These are attaining integer crossings rather than a claim of global integer minimality under discrete-power oscillations. At $\Delta = .005$, $\omega = .40$, the nine-scale budget is $444,072$. Every five-scale value has nonpositive mean log-evidence growth there, so its long-budget behavior cannot support a consistent positive-power claim. The complete sweep is stored in the manifest. Expanding a grid trades a smaller minimum scale for the $\log K$ correction; it does not dominate a narrower grid at every operating point.

J.3 PREDICTABLE REGIMES WITH TEMPORALLY VARYING DISAGREEMENT

Each setting consists of 2,000 independent trajectories of length 4,000. A binary regime starts uniformly. After each outcome it is independently refreshed to a fair bit with probability $1/T$, for $T \in \{1, 50, 200\}$; otherwise it persists. The current regime is revealed before generating the next example. Its disagreement probability is $\omega_t = .04$ in one regime and .36 in the other. Thus the conditional variance can remain high or low for long periods without a common conditional risk difference.

Set $Y_t \in \{-1, 1\}$ with $\Pr(Y_t = 1) = .30$, and, conditional on the shared label and predictable regime, generate independent candidate coordinates using

$$\Pr(X_{j,t} = 1 \mid Y_t = 1, \mathcal{F}_{t-1}) = \frac{\omega_t + \Delta_{j,t}}{.60}, \qquad \Pr(X_{j,t} = 1 \mid Y_t = -1, \mathcal{F}_{t-1}) = \frac{\omega_t - \Delta_{j,t}}{1.40}. \tag{93}$$

Freeze $H_0 = -1$ and $H_j = 2X_j - 1$. The mixed family has $\Delta_{j,t} = \theta_j \omega_t$, with $\theta = (.2, 0, 0, 0, 0, 0, -.1, -.1, -.1, -.1, -.1, -.1)$. The null family has every $\theta_j = 0$. All displayed probabilities lie in $[0, 1]$. The true accumulated conditional gain is computed exactly as $\sum_{i \le t} \theta_j \omega_i$ along each regime realization. Both candidates and their conditional advantages are fixed by this generator, not selected from favorable runs.

We check the simultaneous lower bound at every prefix and for every candidate, then apply the endpoint deployment rule separately. A coverage failure means that some lower bound exceeds its true accumulated conditional target; a null approval means that some non-improving candidate has a positive bound. Table 23 distinguishes these events from endpoint deployment. The regime refresh time affects the simulated process but is not supplied to the certificate. Seeds and exact generation order are in `temporal_stress.py`.

Table 23: Nine-scale predictable-regime stress test: 2,000 length-4,000 trajectories per setting. Coverage failure checks all candidates and prefixes against the known conditional target. False approval also checks every prefix. Useful, harmful, and fallback columns concern the endpoint decision; the two other all-null decisions at $T = 1$ select an equal-risk candidate.

T	Family	Coverage failures	False approvals	Useful	Harmful	Fallback
1	null	3	3	0	0	1998
1	mixed	1	0	1872	0	128
50	null	5	5	0	0	2000
50	mixed	3	1	1839	0	161
200	null	1	1	0	0	2000
200	mixed	1	0	1755	0	245

Across the three null settings, any-time coverage failures are 3, 5, and 1 out of $2,000$, respectively. The largest frequency, .0025, is below the nominal $.05$ but does not prove coverage by itself. In the mixed family, useful endpoint deployments are 1,872, 1,839, and 1,755, with zero harmful or equal-risk endpoint deployments. A zero harmful count in 2,000 independent trajectories has marginal two-sided 95% Clopper–Pearson upper endpoint .001843, not a zero population probability. The study isolates certification under a fully known adapted law; it does not validate a Markov model for any real panel.

K WITHIN-AUDIT DRIFT AND RECENT-WINDOW CERTIFICATION

K.1 A FIXED RESTART CHANGES THE CERTIFIED TARGET

The full-history certificate averages conditional advantages from the beginning of the audit. An early improvement can therefore outweigh a later reversal even when every forthcoming prediction favors the baseline. A restart removes those early observations from the target, rather than reinterpreting a full-history guarantee as a statement about the current regime.

Fix an audit endpoint n and a lookback $h \le n$ before inspecting audit outcomes, and put $s = n - h$. Write $a_{r,t} = \mathbb{E}[U_{r,t} \mid \mathcal{F}_{t-1}]$, $S_{r;s,t} = \sum_{i=s+1}^{t} U_{r,i}$, and $D_{r;s,t} = \sum_{i=s+1}^{t} U_{r,i}^2$. With the unchanged

scale grid Λ_9, define

$$\underline{\Delta}_{r;s,t} = \frac{1}{t-s} \max_{\lambda \in \Lambda_9} \left\{ S_{r;s,t} - \frac{\psi(\lambda)}{\lambda} D_{r;s,t} - \frac{\log(9J/\alpha_s)}{\lambda} \right\}. \tag{94}$$

At the fixed endpoint this bound concerns the average conditional gain over the last h audit observations. The resulting rule retains the same frozen candidates, approves only positive lower bounds, and selects the largest empirical gain within that window. Neither the window nor its start is chosen to maximize an observed bound.

Proposition 14 (Predeclared restarts)**.** *For any deterministic start s and $\alpha_s \in (0,1)$, with conditional probability at least $1-\alpha_s$ given $\mathcal{F}_s$, simultaneously over $r \in [J]$ and $t > s$,*

$$\frac{1}{t-s} \sum_{i=s+1}^{t} a_{r,i} \geq \underline{\Delta}_{r;s,t}. \tag{95}$$

For a predeclared finite set of starts $\mathcal{S}$, the same inequalities hold simultaneously over all starts with failure probability at most $\sum_{s \in \mathcal{S}} \alpha_s$. No independence between the windows is required.

Proof. Condition on $\mathcal{F}_s$. For each candidate and scale, start

$$M^{\lambda}_{r;s,t} = \exp\left\{ \lambda \sum_{i=s+1}^{t} (U_{r,i} - a_{r,i}) - \psi(\lambda) \sum_{i=s+1}^{t} U_{r,i}^2 \right\}, \qquad M^{\lambda}_{r;s,s} = 1. \tag{96}$$

The conditional exponential inequality in Equation (74) makes this a nonnegative supermartingale. Ville's inequality at threshold $9J/\alpha_s$ and a union bound over the unchanged nine scales and J candidates give Equation (95). Averaging over $\mathcal{F}_s$ and taking a union over starts gives the final assertion. The construction is the fixed-restart specialization of the same confidence-sequence argument (Howard et al., 2018; 2021). □

One full-history process already covers every audit prefix at its original level α. In contrast, restarting at several origins introduces several processes and needs a restart allocation, such as $\alpha_s = \alpha/|\mathcal{S}|$. Repeatedly selecting an uncorrected lookback from observed outcomes is not covered. Our comparisons evaluate separately specified rules; they do not select between full-history and trailing-window outputs after seeing their future errors.

Recency alone does not guarantee future improvement. For example, suppose a specified $\rho \geq 0$ bounds $|a_{r,t+1} - a_{r,t}|$ pathwise over the last h audit observations and the next b observations. Averaging the inequality $a_{r,n+j} \geq a_{r,n-i} - \rho(i+j)$ over $i = 0, \ldots, h-1$ and $j = 1, \ldots, b$ gives

$$\frac{1}{b} \sum_{j=1}^{b} a_{r,n+j} \geq \frac{1}{h} \sum_{i=0}^{h-1} a_{r,n-i} - \rho \frac{h+b}{2} \geq \underline{\Delta}_{r;n-h,n} - \rho \frac{h+b}{2} \tag{97}$$

on the restart coverage event. The first inequality is an assumption-dependent transfer calculation, not a bound established for our real series. It identifies the tradeoff: decreasing h reduces potential drift bias but increases the statistical cost of using fewer labels.

K.2 One continuing process with drift inside the audit

There are $2,000$ independent trajectories per setting, an audit of length $n \in \{4,000, 16,000\}$, and a future segment of length $h = n/4$. A binary regime starts uniformly and, after each observation, refreshes to a fair bit with probability $1/50$. It otherwise persists. The same regime path continues through the audit boundary: there is no independent phase-B initialization or post-audit reversal coin. Before each outcome, the regime reveals disagreement $\omega_t \in \{.04, .36\}$.

Use the shared-label model in Equation (93), with $\Pr(Y_t = 1) = .30$, $J = 12$, and frozen classifiers $H_0 = -1$, $H_j = 2X_j - 1$. The drifting first candidate has

$$\theta_{1,t} = .3 - .4 \frac{t-1}{n-1}, \qquad \Delta_{1,t} = \omega_t \theta_{1,t}, \qquad t = 1, \ldots, n+h. \tag{98}$$

Thus its conditional advantage declines during the audit, is $-.1\omega_n$ at the endpoint, and continues declining throughout the future segment. Candidates two through six have zero advantage, while candidates seven through twelve have $\theta_{j,t} = -.1$. The stable positive control instead fixes $\theta_{1,t} = .2$ throughout both segments. All conditional feature probabilities are in $[0, 1]$.

Both rules see identical candidates and outcomes. Full history uses all n audit examples; the recent-window rule restarts at the fixed point $n - h$ and uses the last h. Each separately specified rule uses $\alpha = .05$, the same candidate correction, and exactly the same nine scales. The decision is made once, at time n, and the selected classifier remains fixed throughout the future segment. The shorter window is not credited with a lower acquisition cost: the full trajectory is generated and only its certification window differs.

For $W = \{1, \ldots, n\}$ or $W = \{n - h + 1, \ldots, n\}$, the audit target is $\Delta_{r,W} = |W|^{-1} \sum_{t \in W} \theta_{r,t}\omega_t$. The exact future target is $\Delta_{r,B} = h^{-1} \sum_{t=n+1}^{n+h} \theta_{r,t}\omega_t$. A false audit approval has $\Delta_{\widehat{r},W} \leq 0$; future harm has $\Delta_{\widehat{r},B} < 0$. Exact future risk is $.30 - \Delta_{\widehat{r},B}$ on approval and .30 on fallback. We also record empirical future loss, but do not use it to classify a simulated population gain as beneficial or harmful.

Table 24: Within-audit drift and a fixed recent-window restart. Each setting has 2,000 independent trajectories. App., FA, Harm, and FB are endpoint approval, false audit approval, future-harmful deployment, and fallback counts. Every rule uses the same nine scales at $\alpha = .05$ over twelve alternatives. Risk is exact future conditional risk averaged over trajectories, including fallback; parentheses give Monte Carlo standard errors. Full and recent rules are compared on identical outcomes.

Audit n	Scenario / rule	App.	FA	Harm	FB	Future risk
4,000	Stable / full	1,852	0	0	148	.26303 (.00032)
	Stable / recent	140	0	0	1,860	.29718 (.00024)
	Drift / full	228	0	228	1,772	.30342 (.00022)
	Drift / recent	0	0	0	2,000	.30000 (0)
16,000	Stable / full	2,000	0	0	0	.25967 (.00011)
	Stable / recent	1,842	0	0	158	.26289 (.00026)
	Drift / full	1,870	0	1,870	130	.32810 (.00019)
	Drift / recent	1	1	0	1,999	.30000 (0)

The full-history target remains positive even when the endpoint advantage has reversed (Table 24). At $n = 16,000$, its first-candidate mean audit target is .01999, compared with $-.01000$ for the recent window and $-.03006$ for the future. Every one of the $1,870$ full-history approvals selects candidate one and becomes harmful from the first future observation. None is a false approval for its own audit target. The corresponding harmful-deployment frequency is .9350, with marginal 95% Clopper–Pearson interval $[.92329, .94541]$. The recent rule has no future-harmful deployment; its one approval selects an equal-risk candidate and is a false audit approval. Its zero harmful count has upper confidence endpoint .001843, not a demonstrated zero population probability.

Restarting pays for that recency with power. Under the stable control at $n = 16,000$, useful deployment falls from 1.0000 to .9210, and the mean future risk increases from .25967 to .26289. At $n = 4,000$ the recent window contains only $1,000$ observations and useful deployment is .0700. This is a bias–power tradeoff, not uniform dominance of a shorter audit. A restart protects its recent conditional target but cannot prevent an arbitrary change immediately after the endpoint.

The first-candidate full-history target is positive in all $2,000$ drift trajectories at each audit length; its recent-window target is negative in all of them. Endpoint coverage failures, distinct from false approvals, are 2 and 0 for full and recent rules in the shorter stable control, 2 and 2 in the longer stable control, 0 and 0 in the shorter drift setting, and 0 and 1 in the longer drift setting. These frequencies assess the implemented endpoints, not all possible restart times. The separate stress test in Table 23 assesses all-prefix coverage.

K.3 Rolling-origin transfer on the production panel

The original 378-month audit supports a second transfer diagnostic without changing its trained candidates. Keep each seed's nine classifiers frozen at the original calibration boundary. Set audit-

prefix endpoints after 120, 156, 192, 228, 264, 300, and 336 audit months. At each endpoint, evaluate the next 36 months after a six-month embargo. All seven future blocks lie inside the original audit calendar and are mutually disjoint. The origins form a fixed retrospective evaluation grid; the candidates still use only the original training and calibration segments. Labels from an earlier evaluation block can become ordinary past observations at a later origin. The procedure is rolling evaluation of fixed candidates, not rolling refitting.

Compare the expanding full history with a fixed 120-month trailing window. The full-history process already controls all seven origins simultaneously at $\alpha = .05$ for each learner seed. For the seven distinct recent-window starts, allocate $\alpha_s = .05/7$ to each restart. For transparency, Table 25 also reports a per-origin recent-window analysis at .05 and an unnecessarily conservative $.05/7$ full-history control. The per-origin recent result is not given a simultaneous seven-origin guarantee.

Each forward block has 216 entity-month examples. A forward reversal means that the selected classifier makes more errors than its matched uniform ensemble on this block. This is an observed paired-error comparison, not known population harm. The 70 seed-origin decisions reuse ten fitted families and seven calendar blocks, so we do not attach an independent-binomial interval to their pooled reversal proportion. The full-history and recent-window procedures are separately specified comparisons, not alternatives selected from future performance.

Table 25: Rolling-origin production evaluation with seven fixed calendar origins and ten learner seeds. Each rule has 70 decisions and uses the unchanged nine scales. Better and Worse count approved replacements with smaller or larger forward-block errors than the matched baseline; there are no approved ties. FB counts fallback. Error pools the same 216 observations per forward block and includes fallback. Uniform error is .32480. The first two rows each control all seven origins at .05 per learner seed; the final two are sensitivity analyses.

Rule	Origin-level budget	App.	Better	Worse	FB	Error
Full history	.05, one process	23	21	2	47	.31614
Recent 120 months	$.05/7$, restarts	0	0	0	70	.32480
Full history	$.05/7$, one process	12	11	1	58	.32077
Recent 120 months	.05 per origin	1	1	0	69	.32467

The expanding certificate approves 23 replacements, of which 21 improve and 2 worsen forward error (Table 25). Both reversals occur at the 300-month origin for full stacking: seed 6 changes 75 baseline errors to 76, and seed 8 changes 81 to 84. Thus the reversals involve small error counts on fixed blocks, not a measured population failure probability. The full procedure nevertheless lowers pooled forward error from .32480 to .31614. At the same simultaneous confidence level, the recent rule retains uniform throughout. Its 720-label window and restart correction leave insufficient evidence to approve a change. This real-data result complements the drift experiment: recent information can prevent an outdated decision, but discarding earlier observations can also remove the power needed to use a beneficial replacement.

K.4 Execution and retained numerical records

The drift generator uses `default_rng(410000+n)` for drift and `default_rng(411000+n)` for the stable control. Within each step it draws the shared label, candidate discordances, and the regime refresh in that order. It accumulates exact conditional targets and stores full-history and recent-window counts, bounds, selections, and future targets for every repetition. Marginal binomial intervals use Clopper–Pearson endpoints; the risk standard errors use the standard deviation of exact per-trajectory future risks divided by $\sqrt{2000}$. No Monte Carlo interval is interpreted as simultaneous across the four settings or the two procedures.

The rolling-origin program reads the recomputed production predictions, month identifiers, and split indices from the same frozen model specifications used in the main table. It saves every window's paired counts, lower bounds, selected classifier, forward error counts, and confidence allocation. It performs no additional fitting and does not treat overlapping audit windows as independent observations. Both programs and their complete summary records are embedded in the source.

L Executable Record and Numerical Diagnostics

L.1 Numerical optimization and reproducible execution

The programs use Python 3.13.5, NumPy 2.3.5, SciPy 1.17.0, scikit-learn 1.8.0, statsmodels 0.14.6, and pandas 2.2.3, with CPU float64 arithmetic. BLAS and OpenMP thread counts are one. The LaTeX source writes executable programs for the known-risk/refresh and finite-pool comparisons, capped-simplex solver, temporal/learned-graph audits, scalar series, band extension, temporal certificate, predictable-regime stress test, panels, grid-power ablation, scalar recertification, within-audit drift, and rolling-origin evaluation. The panel data use plotnine 0.15.3. The separate numerical manifest records input identities, summaries, decisions, and diagnostics.

For a PSD quadratic objective $w^\top Aw$, the revised solver starts with SLSQP using analytic gradients, maximum $2,000$ iterations, and tolerance 10^{-11}. It optimizes the equivalent tangent-space objective around $q = \mathbf{1}/m$ to reduce the common-label term. A feasibility repair clips negative coordinates, renormalizes, and contracts toward q if necessary. It then uses conditional-gradient polishing. The linear minimizer over the capped simplex is obtained by water filling and scalar bisection; the gap $(2Aw)^\top(w - v)$ is a convex objective-gap diagnostic. This separates a numerical optimization check from the sampling radius in Proposition 10.

Across all revised weight fits, the maximum constraint violation is 4.44×10^{-16} and the maximum final conditional-gradient gap is 1.00×10^{-7}. These are float64 numerical diagnostics, not interval-arithmetic certificates. SLSQP reports unsuccessful termination in 9 heterogeneous-score fits; feasibility repair and conditional-gradient polishing nevertheless bring their recorded gaps below the same tolerance. The maximum polishing count is 9 steps. No empirical guarantee is inferred solely from SLSQP's success flag.