THE RESOLUTION OF NOISY VALIDATION: MODEL SELECTION FROM REPEATED MEASUREMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

More proxy labels can estimate an observed distribution ever more precisely while leaving the best predictor unidentified. We characterize this obstacle for repeated measurements of a positive target. In a fixed finite panel allowing dependent reader errors, two admissible worlds share the complete observation law yet reverse the risks of the same nonconstant candidate pair. Matching bounds give a cubic resolution floor from second-order information, a quartic floor plus interface error from marginal third-order or matched signed information, and a sampling-limited frontier under exact transport. Candidate-disagreement directions map channel uncertainty to risk comparisons; one simultaneous matrix certifies exact selection, tolerance and an entire curve of minimum-improvement promises. Boundary-uniform moment geometry makes the information operational at finite noise. A pre-specified paid comparison tests the same-cost value of acquiring and exploiting signed information; same-information controls isolate processing gains. A stochastic-target control preserves every candidate risk difference and calibration interval while increasing evaluation variance. Together, these results identify what measurement information resolves a comparison, how strongly the available data can support an improvement, and which remaining uncertainty blocks the decision.

1 INTRODUCTION

A validation set is valuable because it supports a decision: retain the current predictor, replace it by a meaningfully better one, or establish which candidate is best. When labels themselves are measurements, increasing their number need not answer that question. More observations reduce uncertainty about a proxy distribution; they do not necessarily identify the target-risk ordering. This distinction matters whenever models are compared through repeated sensor readings, annotator responses or assays rather than directly observed targets. It determines whether the next reading should measure another object, improve calibration, or provide a different kind of information.

The elementary obstacle is retransformation: centered log noise inflates an exponentiated reading. The decisive obstacle is observational equivalence. We construct admissible latent decompositions with *the same complete joint law of inputs and all reader measurements*, but opposite winners for the same fixed, nonconstant predictors. An arbitrarily accurate estimate of that law cannot resolve the reversal. The problem is therefore not solved by a better proxy-score estimator alone.

Our advance is an information–decision chain. **The frontier:** matching bounds identify which measurement information changes an unavoidable comparison scale, even for rules using the full observed law. **Its operational meaning:** candidate-disagreement directions and one simultaneous matrix turn calibrated information into exact selection, tolerance and minimum-improvement promises. The same event supports all improvement thresholds, so the output can report how much improvement the evidence actually justifies. **Its decision value:** paid comparisons separate an information-acquisition scheme from tighter processing of the same data. A risk-difference-preserving stochastic-target control isolates evaluation variance without changing the true ranking or available calibration. The minimax frontier is the theoretical center; these controls explain how it connects to finite-budget decisions.

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.

What changes relative to existing work? Proxy-compatible loss differences preserve comparisons when an appropriate proxy is available (Patton, 2011; Hoga & Dimitriadis, 2023); partial identification bounds weak-supervision performance without true labels (Polo et al., 2024); minimax-regret decisions under partial identification are studied by Yata (2025). We identify the *measurement-dependent resolution of a fixed candidate comparison*, witnessed by identical full observation laws and matching noise orders. Independent-error deconvolution studies identification of latent distributions under different assumptions (Capitao-Miniconi et al., 2025); here dependence is allowed and comparison, rather than full distribution recovery, is the target. Binary-label budget allocation (Dorner & Hardt, 2024) and risk-controlling hyperparameter testing (Zecchin et al., 2025) address complementary decisions. Our obstruction occurs before a latent-risk comparison becomes an identifiable test. Moment bounds and concentration implement the resulting interfaces; the central new object is the information-dependent comparison frontier.

2 Measurements, comparisons and requests

Let H be the input, Y the target and Z_j reader j's log measurement. Write

$$Y = e^T \in (0,1), \qquad Z_j = T + \gamma_j + \eta_j, \qquad X_j = e^{Z_j}, \qquad \sum_{j=1}^{J} \gamma_j = 0. \tag{1}$$

The zero-sum convention fixes a panel-relative target scale; an absolute physical scale requires an anchor. Independent training freezes a finite pool $\mathcal{F}$ of size $M = |\mathcal{F}|$, with $f(H) \in [0,1]$, and an incumbent $f_0 \in \mathcal{F}$. Calibration and evaluation observe measurements, not Y. Define $R(f) = \mathbb{E}(Y - f)^2$, $\Delta_{fg} = R(f) - R(g)$ and $\mathrm{Reg}(f) = R(f) - \min_{g\in\mathcal{F}} R(g)$. Thus $\Delta_{fg} < 0$ favors f.

One assumption set. Objects are independent, while readers within an object may be dependent. The protocol declares $|\eta_j| \le b_j$, and calibration and evaluation share the biases, marginal variances and exponential means. We assume

$$\mathbb{E}[\eta_j \mid T, H] = 0, \quad \mathbb{E}[\eta_j \eta_k \mid T, H] = 0 \ (j \neq k), \quad \mathbb{E}[e^{\eta_j} \mid T, H] = G_j. \tag{2}$$

Write $v_j = \mathbb{E}\eta_j^2$ for a marginal variance; G_j is a reader-specific constant. Zero cross moments are weaker than reader independence. Training, lower-order calibration, signed calibration and evaluation use independent objects. The pool, reader set and request family are fixed before evaluation. Support and closure caps are declared inputs, not certified by sample extrema or a small observed signed mean.

For a predeclared reader, set $\lambda_j^* = e^{-\gamma_j}/G_j$ and abbreviate $X = X_j$. Then $\mathbb{E}[\lambda_j^* X \mid T, H] = Y$. Subtracting squared losses cancels the unobserved Y^2 term:

$$d_{fg}(\lambda) = f^2 - g^2 + 2(g - f)\lambda X, \qquad \Delta_{fg} = \mathbb{E} d_{fg}(\lambda_j^*). \tag{3}$$

The shared multiplier, not a separate correction per model, is the operational unknown.

A running example. For $f = .45$, $g = .55$ and $\mathbb{E}X = .515$, the contrast is $.2(.515\lambda - .5)$. Raw scoring sets $\lambda = 1$ and favors g. Information $\lambda \in [.94, .98]$ gives $\Delta_{fg} \in [-.00318, .00094]$, leaving the ranking unresolved. Refining to $[.94, .96]$ makes the entire interval negative. Finite evaluation adds a sampling radius. This teaching example uses constants; the information lower bound uses fixed nonconstant predictors.

A calibrated interval is denoted Λ, a simultaneous contrast upper bound U_{fg}, and an unresolved request $\perp$. We distinguish *exact selection*, *tolerance* ε with no increase in incumbent risk, and *strict improvement* exceeding τ. An unresolved request deploys f_0; a tolerance certificate may also return f_0, so a certified output is not necessarily an update.

3 Information determines validation resolution

Fix $P_H = \nu$ and nonconstant candidates

$$f = m - w/2, \quad g = m + w/2, \quad K = \mathbb{E}_\nu[wm] > 0, \tag{4}$$

where $0 < m_- \le m(H) \le m_+ < 1$ and $0 < w(H) \le 2\min\{m(H), 1 - m(H)\}$. Fix finite $J \ge 3$ and $\theta \in (0, 1)$. For the frontier, the common base class obeys (1)–(2) with *known* zero biases, common support radius b and marginal variances $v = \theta b^2$. Section 4 separately estimates and charges for those quantities.

Four information interfaces. $\mathcal{C}_2$ supplies the base information. $\mathcal{C}_{3,\kappa_3}$ additionally supplies $|\mathbb{E}\eta_j^3| \le \kappa_3$ for every reader. $\mathcal{C}_{\psi,\kappa_\psi}$ instead supplies $|\psi_i| \le \kappa_\psi$ for a declared triple, where, after bias removal, the observable statistic $W_i = (Z_i - Z_j)^2(Z_i - Z_k)$ obeys $\mathbb{E}W_i = \mathbb{E}\eta_i^3 + \psi_i$. The residual ψ_i collects mixed third moments; independent centered readers make it zero. $\mathcal{C}_G^{\text{full}}$ is the same full base class with exact λ_j^* supplied. These are two alternative third-order interfaces.

A population rule observes the complete $P_{H,Z}$ and the interface's side information. Let ρ^Δ be the infimum gap threshold above which a uniformly zero-error rule must decide. Let $\mathcal{V}^{\text{pop}}$ be minimax expected regret for forced two-model choice with population information, and $\mathcal{V}_n$ its version with n independent complete panels. Randomization is allowed.

Theorem 1 (The information–resolution frontier)**.** *For the fixed objects above, there exist $b_0, c, C > 0$ depending only on (ν, m, w, θ, J) such that, for $0 < b \le b_0$, $0 \le \kappa_3, \kappa_\psi \le b^3$ and every $n \ge 1$, the following orders hold. Each Θ bounds the indicated quantity above and below by positive constant multiples, uniformly over its cap and n.*

Information	ρ^Δ *and* $\mathcal{V}^{\text{pop}}$	$\mathcal{V}_n$
Second order: $\mathcal{C}_2$	$\Theta(b^3)$	$\Theta(b^3 + n^{-1/2})$
Marginal cap: $\mathcal{C}_{3,\kappa_3}$	$\Theta(\kappa_3 + b^4)$	$\Theta(\kappa_3 + b^4 + n^{-1/2})$
Signed closure: $\mathcal{C}_{\psi,\kappa_\psi}$	$\Theta(\kappa_\psi + b^4)$	$\Theta(\kappa_\psi + b^4 + n^{-1/2})$
Exact transport: $\mathcal{C}_G^{\text{full}}$	0	$\Theta(n^{-1/2})$

The regret columns require a choice; all-abstention cannot make them zero. The simultaneous certificates below may abstain. The structural construction fixes the pair and reader count independently of n; constants need not be uniform as J grows or θ approaches an endpoint. More objects remove the sampling term, whereas new valid information can change the structural term's order.

3.1 The same observations can conceal opposite winners

Choose independent centered components A, B with the same variance but different exponential means. Add a vector $\varepsilon + N$ whose negative cross moments cancel the variance of the shared component. With appropriate $c(H)$, swap which component belongs to the target:

	T	η_j	Z_j
P_A	$c(H) + A$	$B + \varepsilon_j + N_j$	$c(H) + A + B + \varepsilon_j + N_j$
P_B	$c(H) + B$	$A + \varepsilon_j + N_j$	$c(H) + A + B + \varepsilon_j + N_j$.

(5)

The entire observation is identical object by object, including all readers. Centering relative to the fixed $m(H)$ makes the risk differences $-q, +q$. A symmetric two-point/three-point swap gives $q \asymp b^4$; a cap-controlled skew mixture gives the cubic contribution. Appendix A checks support, conditional moments and equality of the supplied information.

Corollary 2 (A consequence for abstention)**.** *On two admissible opposite-winner worlds with the same complete observed law and supplied information, any rule whose unconditional probability of a wrong explicit choice is at most α in each world satisfies $P(\text{commit}) \le 2\alpha$.*

Under the common input law, each of the two choice probabilities is bounded by the opposite world's error constraint. Their sum proves the claim. Thus some abstention is demanded by the information, not simply by conservative inference.

For the remaining sampling term, use $Y = m(H) + r_0 S$ with $P_\pm(S = 1) = (1 \pm 1/(4\sqrt{n}))/2$ and independent symmetric errors $\pm b\sqrt{\theta}$. The two worlds have the same *exact* side information $G_j = \cosh(b\sqrt{\theta})$ and observed KL divergence at most $1/4$. Testing therefore costs $\Omega(n^{-1/2})$, even with exact transport. The empirical contrast gives the matching upper bound. This endpoint belongs to the same full class, not to an unspecified narrow subclass.

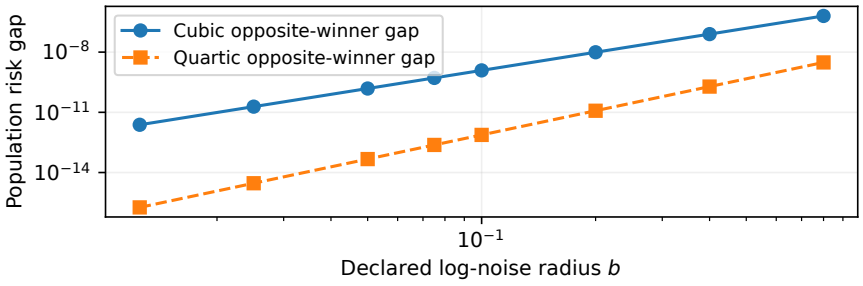

Figure 1: **The obstruction is a risk reversal, not an estimator bias.** Population risk gaps in complete-law paired constructions with $J = 3$. These are actual opposite-winner gaps, not multiplier-interval widths. Different orders have different constants; the numerical construction corroborates the analytical lower bound.

4 From Information to Certified Model Decisions

4.1 Calibrate the Directions on Which Models Disagree

The measurement model supplies one shared multiplier. More generally, let $r_\lambda(H) = \mathbb{E}[Y - \lambda X \mid H]$ and $\mathcal{D} = \mathrm{span}\{f - f_0 : f \in \mathcal{F}\}$. Conditional expectation gives

$$\Delta_{fg} - \mathbb{E}d_{fg}(\lambda) = 2\langle g - f, \Pi_{\mathcal{D}} r_\lambda \rangle. \tag{6}$$

All comparisons are exact if and only if $\Pi_{\mathcal{D}} r_\lambda = 0$. This characterizes the required calibration; the implementation estimates the shared multiplier. Write $a_{fg} = \mathbb{E}(f^2 - g^2)$ and $\beta_{fg} = 2\mathbb{E}[(g - f)X]$. The contrast is $a_{fg} + \beta_{fg}\lambda$, so its width over an interval is exactly $|\beta_{fg}|\,|\Lambda|$. Channel uncertainty matters in proportion to the direction in which the candidates disagree. Appendix B proves the identity and gives crossing predictors with $\beta_{fg} = 0$ and a nonzero risk gap.

4.2 Observable Information and Finite-Noise Precision

Reader differences eliminate T object by object. For matched readers, $A_j = Z_j - J^{-1}\sum_k Z_k$ has mean γ_j and $D_{jk} = Z_j - Z_k$ has variance $V_{jk} = v_j + v_k$. Independent object pairs give $Q_{jk,r} = (D_{jk,2r-1} - D_{jk,2r})^2/2$ with mean V_{jk}. Bounded-mean intervals, followed by $v_j = (V_{jk} + V_{j\ell} - V_{k\ell})/2$, estimate the bias and variance. The sample size for Q is $\lfloor N_0/2 \rfloor$, not N_0.

Given zero mean, support $[-b, b]$ and exact variance v, the attainable second-order bounds are

$$L_b(v) = \frac{ve^{-b} + b^2 e^{v/b}}{b^2 + v}, \qquad U_b(v) = \frac{ve^{b} + b^2 e^{-v/b}}{b^2 + v}. \tag{7}$$

They increase with v, giving the finite-calibration interval

$$\Lambda_2 = \left[e^{-\gamma_U}/U_b(v_U),\ e^{-\gamma_L}/L_b(v_L)\right]. \tag{8}$$

The signed statistic introduced in Section 3 satisfies, after removing biases,

$$\mathbb{E}W_i = \mu_{3,i} + \psi_i, \qquad \mu_{3,i} = \mathbb{E}\eta_i^3. \tag{9}$$

A declared $|\psi_i| \le \kappa_\psi$, independent signed sampling and bias-error propagation give $\mu_{3,i} \in [s_L, s_U]$. Taylor's formula supplies a lower exponential endpoint $1 + v_L/2 + s_L/6 + v_L^2/24 - e^b b^5/120$ and an upper endpoint using $(v_U, s_U, b^2 v_U, +e^b b^5/120)$; intersect these with (7). A marginal cap instead supplies $[s_L, s_U] = [-\kappa_3, \kappa_3]$. Appendix B derives the signed bias correction. Empty intervals remain unresolved.

The distance to a moment boundary controls precision. For exact $0 < v < b^2$ and $s = \mathbb{E}\eta^3$, define $\xi = bs/[v(b^2 - v)]$. Feasibility gives $|\xi| \le 1$. The sharp scalar bounds $L_*(v, s), U_*(v, s)$ obey the finite-noise inequality

$$\frac{e^{-b}}{24} D_4(v, s) \le U_* - L_* \le \frac{e^{b}}{24} D_4(v, s), \qquad D_4(v, s) = v(b^2 - v)(1 - \xi^2). \tag{10}$$

Algorithm 1 A frozen pool, calibrated information and a requested decision

Require: Frozen $\mathcal{F}$, incumbent f_0, independent calibration/evaluation, declared support and interface, α, ε, τ.

1: Estimate bias/variance intervals and, when available, the signed interval with its closure cap.
2: Form Λ by (8), the Taylor intersection or dual refinement. If empty, return $\perp$.
3: Compute the two endpoint contrasts, variances and domain ranges; use (11) to form U.
4: $S_{\text{exact}} \leftarrow \{f : U_{fg} < 0 \text{ for every } g \neq f\}$.
5: $S_{\text{tol}} \leftarrow \{f : \max_g U_{fg} \leq \varepsilon,\ U_{ff_0} \leq 0\}$.
6: $S_{\text{imp}} \leftarrow \{f \neq f_0 : U_{ff_0} < -\tau\}$.
7: For exact selection choose the first qualifying index; for tolerance minimize $\max_g U_{fg}$; for improvement minimize U_{ff_0}, with fixed index tie-breaking. An empty requested set returns $\perp$.
8: Deploy f_0 on $\perp$. Record certified outputs and actual switches separately.

Thus at $v = \theta b^2$ the moment-summary width scales with $b^4\theta(1-\theta)(1-\xi^2)$, including when ξ approaches a boundary; it is exactly zero at $|\xi| = 1$. Appendix B.3 gives the attaining laws and a Hermite-remainder proof. These are scalar moment sets, whereas Theorem 1 concerns the full joint observation law.

With uncertain moments, *dual refinement* uses rational cubic upper and lower polynomials for e^x and propagates the entire calibrated moment rectangle. Intersecting their bounds with the preceding envelopes tightens Λ without new readings or statistical failure events. Its arithmetic certificate checks this deterministic processing; the statistical decision certificate below controls the requested model comparison.

4.3 One simultaneous matrix, three requests

Condition on independent training and calibration. On $n \geq 2$ evaluation objects compute $\bar{d}_{fg}(\lambda)$ and the unbiased variance $\widehat{\sigma}^2_{fg}(\lambda)$. Let $\mathcal{R}_{fg}(\lambda)$ be a valid range width over the declared input/reading domain. Since $d_{gf} = -d_{fg}$, one two-sided event covers both directions. The two endpoints of Λ therefore require $2\binom{M}{2} = M(M-1)$ events. Put $\delta = \alpha_{\text{eval}}/[M(M-1)]$ and use Maurer & Pontil (2009):

$$r_{fg}(\lambda) = \sqrt{\frac{2\widehat{\sigma}^2_{fg}(\lambda)\log(4/\delta)}{n}} + \frac{7\mathcal{R}_{fg}(\lambda)\log(4/\delta)}{3(n-1)}, \qquad U_{fg} = \max_{\lambda \in \partial\Lambda}\{\bar{d}_{fg}(\lambda) + r_{fg}(\lambda)\}. \quad (11)$$

The range comes from the declared domain, not sample extrema; $\partial\Lambda$ denotes the two endpoints and $U_{ff} = 0$.

Theorem 3 (Simultaneous decision certificates)**.** *If calibration covers λ^* with probability at least $1 - \alpha_{\text{cal}}$, then with probability at least $1 - \alpha_{\text{cal}} - \alpha_{\text{eval}}$ every pair satisfies $\Delta_{fg} \leq U_{fg}$. On this event Algorithm 1's exact output is uniquely best, its tolerance output has regret at most ε and does not increase incumbent risk, and its improvement output reduces risk by more than τ.*

The proof covers endpoint means and uses the affinity of their population contrasts. For an improvement update A with gain Γ, it gives the *unconditional* guarantee $P(A, \Gamma \leq \tau) \leq \alpha$. A positive gain below τ is a shortfall, even without harm. The conditional proportion among updates is a different quantity.

One matrix supports an entire improvement curve. Define the strongest available lower bound and its fixed-tie-breaking challenger by

$$\widehat{\Gamma}_{\max} = \max_{f \neq f_0}(-U_{ff_0}), \qquad f^\dagger \in \arg\max_{f \neq f_0}(-U_{ff_0}). \quad (12)$$

On the same event, $\Gamma_{f^\dagger} \geq \widehat{\Gamma}_{\max}$, so *every* $\tau \geq 0$ below this bound is a valid improvement promise. Thus $P\{\exists \tau \geq 0 : \widehat{\Gamma}_{\max} > \tau,\ \Gamma_{f^\dagger} \leq \tau\} \leq \alpha$, without a new union bound over thresholds. With no alternative candidate or invalid calibration, set the bound to $-\infty$ and retain f_0. This reuses the fixed matrix, not unbudgeted models, readers or repeated looks at new data. Choosing an empirical performance peak still does not redefine a pre-specified experimental endpoint.

What a tighter interval buys. On identical evaluation data, with the same domain ranges and failure allocation, $\Lambda_S \subseteq \Lambda_2$ implies $U^S_{fg} \leq U^2_{fg}$. The empirical mean is affine and the standard deviation and range are convex in λ; the interval maximum is attained at an endpoint. Valid refinement expands each qualifying set. Buying information, in contrast, changes the data allocation as well as the interval.

Locate the remaining decision cost. For the *actual* interval $\widehat{\Lambda}_p = [\widehat{c}_p - \widehat{r}_p, \widehat{c}_p + \widehat{r}_p]$, let $\widehat{h}^p_{fg} = a_{fg} + \beta_{fg}\widehat{c}_p + |\beta_{fg}|\widehat{r}_p$. For comparison to f_0, write $\Gamma_f = -\Delta_{ff_0}$, $B^p_f = \widehat{h}^p_{ff_0} - \Delta_{ff_0}$ and $V^p_f = U^p_{ff_0} - \widehat{h}^p_{ff_0}$. Then exactly

$$U^p_{ff_0} = -\Gamma_f + B^p_f + V^p_f, \qquad U^p_{ff_0} < -\tau \iff \Gamma_f - \tau > B^p_f + V^p_f. \tag{13}$$

Here B^p_f combines summary slack, closure and finite calibration; V^p_f is evaluation loss. Writing $\beta = \beta_{ff_0}$, the population upper contrast uses $\lambda_{\mathrm{act}} = \widehat{c}_p + \mathrm{sgn}(\beta)\widehat{r}_p$, giving $B^p_f = \beta(\lambda_{\mathrm{act}} - \lambda^*)$ (zero for $\beta = 0$). The relevant endpoint, not width alone, controls this loss. On joint coverage, $0 \leq B^p_f \leq |\beta|\,|\widehat{\Lambda}_p|$ and $0 \leq V^p_f \leq 2r^p_{ff_0,\max}$. These are post-selection diagnostics of the actual certificate; Appendix B gives their proofs and budget implications.

5 Evidence: which information supports which decision?

Complete-law pairs test Theorem 1; scalar moment laws test (10); paid comparisons test the simultaneous promises and actual losses in (13). We distinguish the pre-specified same-cost comparison from same-information processing, threshold reanalysis and a controlled increase in target randomness. Appendices C–D specify every cohort and retain unfavorable outcomes.

A common request and reading ledger. Seven candidates learn from 512 proxy-labeled objects on 25 input states; the corrected-cubic incumbent is fixed. Independent lower calibration, signed calibration and evaluation cost

$$B_{\mathrm{total}} = 3N_{\mathrm{train}} + 3N_0 + 3N_3 + n. \tag{14}$$

The signed scheme uses 35/35/30 percent of its validation readings for these last three stages. Native second-order controls redirect the entire unused signed budget to evaluation or lower calibration and reallocate its failure probability. All use the same declared support and bias bound, with $\alpha = .05$, $\tau = 10^{-4}$ and $\varepsilon = .002$. Exact transport is a supplied reference, whose acquisition cost is not measured. The ledger counts simulated readings; physical verification of support and closure caps is separate.

For an update A with true gain Γ, a useful update reaches τ; a shortfall does not, and harm is the subset $\Gamma < 0$. Rates use all pools. Deployed regret includes incumbent fallback. Targets enter only post-selection scoring; exact and tolerance requests have their own outputs and failures.

5.1 Full-law reversals and moment-summary precision

The complete-law cap grid includes unsaturated and saturated mixtures. Its 34 couplings preserve all observed readers; reconstruction discrepancy is at most 4.44×10^{-16}. On 1,200 independent panel-count draws, all 3,600 paired rule outputs agree between worlds: forced-choice errors sum to one, while the evaluated certificates remain unresolved. Separately, 20,000 local two-world trials give $\sqrt{n}$-scaled regret between .004335 and .004575, with target gaps shrinking as $n^{-1/2}$.

For the scalar and learned-pool experiments, independent reader errors have law

$$P_u = u\left(\tfrac{2}{3}\delta_{-b/2} + \tfrac{1}{3}\delta_b\right) + (1-u)\left(\tfrac{1}{2}\delta_{-b/\sqrt{2}} + \tfrac{1}{2}\delta_{b/\sqrt{2}}\right), \quad 0 \leq u \leq 1. \tag{15}$$

Variance is $b^2/2$, closure is zero, and $s = ub^3/4$, so $D_4 = b^4(1-u^2)/4$. A 231-cell, 100-digit check spans three variances and both moment boundaries, with zero boundary width. Deterministic-target learning cohorts evaluate the moment-summary procedure: their full laws identify $Y(H)$ through zero conditional cross covariance. Appendix B.6 proves this distinction and the stochastic control below; full-law impossibility is tested by the paired worlds.

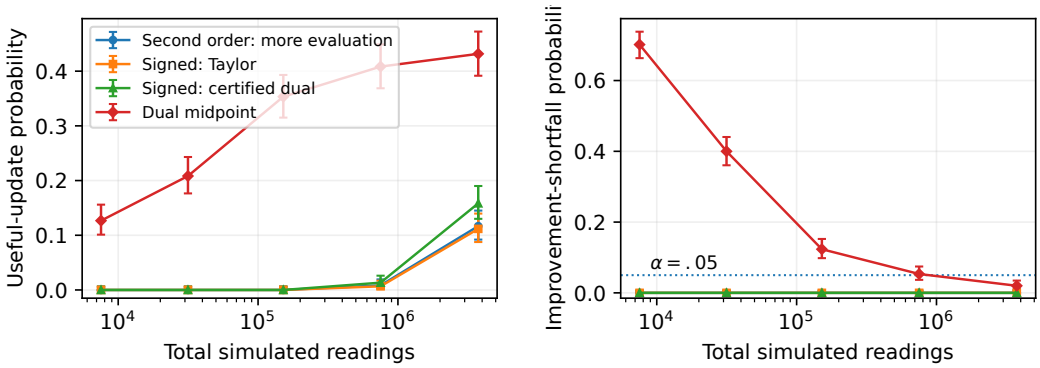

Figure 2: **The same request has two scored outcomes: useful updates and shortfalls.** The confirmation uses 600 paired pools per budget; bars are pointwise 95% binomial intervals. Budgets share pool seeds. The dotted line is $\alpha = .05$. Point rules exchange a broader update set for shortfalls; the curves evaluate a moment-summary procedure.

5.2 Same-cost acquisition and same-information processing

The independent interior confirmation fixed 600 training/channel seeds, five budgets and one primary endpoint before outcomes: the largest-budget useful-update difference between certified dual and second order with more evaluation. Configuration and code hashes were locked locally. All certificate methods use the same three-request family and unordered-pair allocation in (11).

At $B_{\text{total}} = 3{,}751{,}536$, the signed scheme certifies 95 useful updates versus 70 for more evaluation: 4.17 percentage points, paired 95% bootstrap interval $[1.67, 6.67]$, and pre-specified two-sided discordance $p = .00155$ (42 dual-only, 17 second-order-only successes). Table 1 reports the same-request utility tradeoff.

Rule	Updates	Useful	Shortfall	Regret $\times 10^4$
Retain incumbent	0	0	0	0.979
Second order, more evaluation	70	70	0	0.812
Second order, more calibration	41	41	0	0.868
Signed (Taylor)	67	67	0	0.745
Signed (certified dual)	95	95	0	0.634
Dual midpoint (point rule)	271	259	12	0.217
Low-order point	264	246	18	0.234
Exact transport reference	180	180	0	0.392

Table 1: Independent interior confirmation, 600 pools, largest common reading budget. The request is improvement exceeding τ. Shortfall includes harm; regret includes fallback. Exact transport is supplied reference information. Only dual versus more evaluation is the primary comparison.

The two gains have different comparators. On the same signed data, dual yields 28 more useful updates than Taylor (95 versus 67); Taylor's net difference from native second order is -3. Thus the overall net gain decomposes exactly as $25 = 28 - 3$. The first comparison tests the whole acquisition-and-processing scheme; the 28-update contrast isolates processing with identical data. These secondary differences are not causal mediation percentages, and their interval endpoints are not additive. Same-data event-accounting controls, now in Appendix D, preserve the guarantee for every method.

Across the five budgets, dual-midpoint shortfalls are $421, 240, 74, 32, 12$ per 600 pools. At high budgets point scoring has lower mean regret; at small budgets certificates have zero shortfalls but also no updates. Zero shortfalls in one 600-pool cell has a one-sided 95% upper limit of $.498\%$, not a grid-wide guarantee. Both outcomes are needed to interpret reliability.

A certificate also reports how much improvement is supported. Equation (12) turns each existing matrix into an entire request curve. At the maximum budget, dual certifies 394 positive improvements, 95 exceeding the original 10^{-4} request, and 12 exceeding 2×10^{-4}. Figure 3 is a secondary reanalysis; the original primary threshold is unchanged. The simultaneous theorem supports all thresholds on one matrix, while empirical rate comparisons retain their separate statistical roles.



Figure 3: **Available evidence supports a range of promise strengths.** Update rates from the same 600 confirmation matrices at 3,751,536 readings. The vertical line marks the original primary threshold. Lines connect nine exploratory thresholds; point-rule updates are not simultaneous certificates. Full shortfall counts are in Appendix D.

Acquisition and additional evaluation cross different thresholds. Exploratory interior, boundary and valid-nonzero-closure cohorts give dual/more-evaluation useful counts $83/52$, $147/60$ and $39/65$ at the same cost (600 pools each, kept separate from confirmation). Table 2 explains the reversal: signed information improves the population interval gate but reduces the evaluation sample. In the confirmation, 53 pools pass both methods, 42 pass only dual and 17 only second order; 159 of the 488 passing neither still have a true improvement opportunity. Active-endpoint distributions for each method's observable challenger are reported in the supplement.

Cohort	Rule	Opportunity	Interval gate	Finite gate
Confirmation	Second order, more evaluation	271	140	70
Confirmation	Signed (Taylor)	271	170	67
Confirmation	Signed (certified dual)	271	202	95
Nonzero closure	Second order, more evaluation	321	143	65
Nonzero closure	Signed (certified dual)	321	158	39

Table 2: **Opportunity, current-interval gate, and finite certificate**, at the largest budget. Each cohort has 600 pools. The first two counts use the generating law only after selection; they never filter pools or enter the algorithm. Nonzero-closure rows are exploratory evidence.

5.3 Increase target randomness without changing any risk comparison

Freeze the same 600 learned pools and calibration records, and replace the evaluation target by $Y_a = \mu(H) + aS$, with independent fair $S \in \{-1, 1\}$ and $a \in \{0, .025, .075\}$. The target remains in $(0, 1)$ and the error channel is unchanged. For every frozen candidate and fixed calibrated endpoint,

$$R_a(f) = R_0(f) + a^2, \qquad \Delta_{fg}^{(a)} = \Delta_{fg}^{(0)}, \tag{16}$$

$$\mathrm{Var}(d_{fg}^{(a)}(\lambda)) = \mathrm{Var}(d_{fg}^{(0)}(\lambda)) + 4a^2\lambda^2 e^{2\gamma}\mathbb{E}[(g-f)^2]\mathbb{E}e^{2\eta}. \tag{17}$$

Reader differences remove T_a exactly, so calibration intervals and population loss B remain fixed. The added contrast has conditional mean zero, proving the risk identity and orthogonal variance increment; Appendix B.6 supplies the full argument.

We fixed three budgets and a shared count coupling across amplitudes before generating fresh evaluation samples. Calibration is reused but still charged. At the largest budget, dual's interval gate stays at 202 pools, whereas finite useful updates are $91, 90, 82$ across amplitudes; second order gives $71, 69, 66$ and Taylor $65, 62, 58$. All three certificate methods show zero shortfalls. The dual decrease at .075 is 1.50 percentage points (secondary paired 95% interval $[.67, 2.50]$). These are new evaluation draws, not a replacement for the 95/70 primary result. The control changes evaluation variance while keeping the true opportunity set fixed; finite-sample decisions need not be monotone pool by pool.

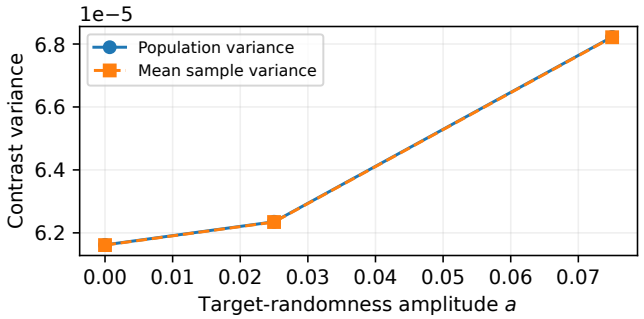

Figure 4: **Target randomness changes precision, not the true comparison.** At the largest budget, the exact variance and mean sample variance agree for the corrected-cubic incumbent versus central corrected-quintic fit at each pool's dual upper endpoint. Means use the same 600 pools; the target amplitude changes no risk difference or calibration interval.

5.4 Real features and natural measurement diagnosis

On the 442-row diabetes data (Efron et al., 2004), ridge and kernel-ridge models learn simulated proxies. Three fixed 265/88/89 row partitions use 200 repetitions per channel; held-out $Y = y/(y + 200)$ supplies scoring only. At 6,000,795 readings, dual/more-evaluation useful counts are $80/57, 111/53, 118/95$ for skew and $39/39, 45/34, 64/57$ for nonzero closure. Partitions overlap and remain separate. Larger budgets draw fresh noise on their fixed empirical laws, not new patients.

Chákṣu Remidio tables (Kumar et al., 2023) supply 1,073 positive ratios and 100 panel/support scenarios without latent-target observations. Current intervals support no requested action. At fixed multiplier centers, removing either width or evaluation radius alone unlocks no improvement; removing both admits 80 scenarios, leaving 20 unfavorable centers. This diagnoses a joint bottleneck conditional on current centers. Appendix D retains all feature controls and reader corners. The arithmetic stage takes median 11.22 ms per conversion in the recorded run; tighter processing also has a computation cost.

6 Conclusion

Measurement information determines the risk-comparison resolution that proxy-sample size alone cannot overcome. Complete-law reversals establish this frontier; candidate directions and finite-noise geometry make calibrated information operational. One simultaneous matrix then reports which model decisions, and how strong an improvement, the evidence supports. Same-cost, same-information and risk-difference-preserving controls separate acquisition, processing and evaluation uncertainty. The resulting design question is precise: which unresolved comparison margin can the next unit of measurement or calibration change?

AI USE STATEMENT

Generative AI tools were used to aid manuscript restructuring, language editing, and documentation. We have reviewed all AI-assisted work. The mathematical results, proofs, computational checks, and reported numbers were produced and verified by the authors, who take full responsibility for the final content of the paper.

REPRODUCIBILITY STATEMENT

The supplement contains complete proofs, configurations, frozen predictions, finite-atom counts, contrast matrices and scripts that generate the tables and figures. A separate operational inference module accepts observations, predictions and declared protocol information, not latent risks. Exact rational records verify the moment-processing stage. The stochastic-target control retains the existing frozen pools and calibration records, charges their costs and generates independent evaluation counts; its analysis is separate from the original confirmation primary. Experiment statistics identify their independent repetition units and retain unfavorable controls. The anonymous code package excludes internal review and delivery documents.

ETHICS STATEMENT

Real-feature experiments use a packaged public regression dataset only for conditional semi-synthetic validation. Natural-reader analysis uses the released numerical annotation tables and preserves the documented exclusions and provenance. No images are newly labeled and no individuals are recruited. Neither the empirical held-out law nor reader-panel sensitivity establishes clinical benefit or patient-level independence. The mathematical target scale and any external anchoring requirements are stated explicitly.

REFERENCES

Jérémie Capitao-Miniconi, Elisabeth Gassiat, and Luc Lehéricy. Deconvolution of repeated measurements corrupted by unknown noise. *Electronic Journal of Statistics*, 19(2):4141–4173, 2025. doi: 10.1214/25-EJS2415.

Florian E. Dorner and Moritz Hardt. Don't label twice: Quantity beats quality when comparing binary classifiers on a budget. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 11544–11572, 2024. URL https://proceedings.mlr.press/v235/dorner24a.html.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004. doi: 10.1214/009053604000000067.

Yannick Hoga and Timo Dimitriadis. On testing equal conditional predictive ability under measurement error. *Journal of Business & Economic Statistics*, 41(2):364–376, 2023. doi: 10.1080/07350015.2021.2021923. URL https://arxiv.org/abs/2106.11104.

J. R. Harish Kumar, Chandra Sekhar Seelamantula, J. H. Gagan, Yogish S. Kamath, Neetha I. R. Kuzhuppilly, U. Vivekanand, Preeti Gupta, and Shilpa Patil. Chákṣu: A glaucoma specific fundus image database. *Scientific Data*, 10(1):70, 2023. doi: 10.1038/s41597-023-01943-4. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC9898274/.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Conference on Learning Theory*, 2009. URL https://www.cs.mcgill.ca/~colt2009/papers/012.pdf.

Andrew J. Patton. Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, 160(1):246–256, 2011. doi: 10.1016/j.jeconom.2010.03.034. URL https://public.econ.duke.edu/~ap172/Patton_vol_proxies_JoE_2011.pdf.

Felipe Maia Polo, Subha Maity, Mikhail Yurochkin, Moulinath Banerjee, and Yuekai Sun. Weak supervision performance evaluation via partial identification. In *Advances in Neural Information Processing Systems*, volume 37, pp. 135704–135740, 2024. doi: 10.52202/079017-4309. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f4c6bec746b0aeca8c2cd15096f1ad1f-Abstract-Conference.html.

E. Seneta and N. C. Weber. Attainable bounds for expectations. *Journal of the Australian Mathematical Society (Series A)*, 33(3):411–420, 1982. doi: 10.1017/S144678870001884X.

Kohei Yata. Optimal decision rules under partial identification, 2025. URL https://arxiv.org/abs/2111.04926v4. arXiv:2111.04926, version 4.

Matteo Zecchin, Sangwoo Park, and Osvaldo Simeone. Adaptive learn-then-test: Statistically valid and efficient hyperparameter selection. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 74018–74036. PMLR, 2025. URL https://proceedings.mlr.press/v267/zecchin25a.html.

A COMPLETE PROOF OF THE INFORMATION–RESOLUTION FRONTIER

This appendix uses the fixed (ν, m, w, θ, J) of Theorem 1. Constants below can depend on these objects, but not on n or the supplied cap. Set $K = \mathbb{E}[wm]$ and $\bar{w} = \mathbb{E}w$. For a population rule D with values $f, g, \perp$, let ρ^{Δ} be the infimum of $r \geq 0$ for which some uniformly zero-error D commits whenever $|\Delta_{fg}| > r$. The infimum includes randomized rules. Forced-choice values are $\inf_D \sup_P \mathbb{E}_P \operatorname{Reg}(D)$ with either the complete population law or n independent observed panels and the declared side information as input.

A.1 A SHARP MOMENT INTERVAL AND THE UPPER BOUNDS

Lemma 4 (Exponential moment envelope)**.** *If $|\eta| \leq b$, $\mathbb{E}\eta = 0$ and $\mathbb{E}\eta^2 = v \in [0, b^2]$, then $L_b(v) \leq \mathbb{E}e^{\eta} \leq U_b(v)$ with the functions in* (7)*; both bounds are attained. If also $|\mathbb{E}\eta^3| \leq \kappa$, intersect this interval with $[L_3, U_3]$, where*

$$\begin{aligned} L_3 &= 1 + v/2 - \kappa/6 + v^2/24 - e^b b^5/120, \\ U_3 &= 1 + v/2 + \kappa/6 + b^2 v/24 + e^b b^5/120. \end{aligned} \tag{18}$$

For $v = \theta b^2$, the intersection $[L, U]$ satisfies $L \geq 1$ and

$$U - L \leq \kappa/3 + \theta(1-\theta)b^4/24 + e^b b^5/60. \tag{19}$$

Proof. For the lower bound, take the quadratic Hermite interpolant q_L matching e^x at $-b$ and matching both value and derivative at $t = v/b$. Its remainder is

$$e^x - q_L(x) = e^{\xi_x}(x+b)(x-t)^2/6 \geq 0 \quad (-b \leq x \leq b).$$

The expectation of a quadratic depends only on the first two moments. The law on $\{-b, v/b\}$ with probabilities $v/(b^2+v)$ and $b^2/(b^2+v)$ has those moments and attains equality at both support points. Its exponential mean is $L_b(v)$. For the upper bound, interpolate twice at $-v/b$ and once at b. The remainder has factor $(x+v/b)^2(x-b) \leq 0$; the attaining law has support $\{-v/b, b\}$. Degenerate endpoints follow directly or by continuity. This is the classical moment-envelope argument underlying Seneta & Weber (1982).

Taylor's formula gives $\mathbb{E}e^{\eta} = 1 + v/2 + \mathbb{E}\eta^3/6 + \mathbb{E}\eta^4/24 + R_5$, $|R_5| \leq e^b b^5/120$. Jensen and support give $v^2 \leq \mathbb{E}\eta^4 \leq b^2 v$, proving (18)–(19). Jensen also gives $L_b(v) \geq 1$. Expanding the attained bounds at $v = \theta b^2$ yields $U_b - L_b = \theta(1-\theta)b^3/3 + O(b^5)$. □

For a reader with zero bias, $X = Ye^{\eta}$, $G = \mathbb{E}e^{\eta}$, and stability gives $\mathbb{E}[wX] = G\mathbb{E}[wY]$. Put $M_X = \mathbb{E}[wX]$. The population contrast obeys

$$\mathbb{E}[wY] \in [M_X/U, M_X/L], \qquad \Delta_{fg} = 2(\mathbb{E}[wY] - K). \tag{20}$$

Choose f if $M_X/L < K$, g if $M_X/U > K$, and otherwise abstain. The rule has zero errors. At abstention $KL \leq M_X \leq KU$, which implies

$$|\Delta_{fg}| \leq 2K(U/L - 1) = O(\kappa + b^4).$$

Forcing either choice at abstention yields the same-order regret upper bound. For finite samples, let $\lambda_m = (1/L + 1/U)/2$ and choose by the sign of $P_n d_{fg}(\lambda_m)$. Wrong selection implies that the absolute true gap is no larger than the discrepancy from this empirical contrast, hence

$$\operatorname{Reg}(\widehat{f}) \leq |(P_n - P)d_{fg}(\lambda_m)| + 2\|f - g\|_{\infty}\mathbb{E}X\,|\lambda_m - 1/G|. \tag{21}$$

The second term is at most $\|f - g\|_{\infty}(U/L - 1)$ because $\mathbb{E}X \leq U$. For $b \leq b_0$, the contrast is uniformly bounded and the first term has expectation at most $\sqrt{\operatorname{Var}(d)/n} = O(n^{-1/2})$. This proves the marginal-cap upper bounds. Taking $\kappa = b^3$ gives $\mathcal{C}_2$; the expected minimax bound does not require a confidence logarithm.

A.2 One complete-law construction supplies both structural lower bounds

Define

$$c_{\theta,J} = \min\left\{\frac{\sqrt{\theta}}{2\sqrt{J}}, \frac{1-\sqrt{\theta}}{2(J+1)}\right\}, \quad a = c_{\theta,J} b, \quad s = \sqrt{\theta b^2 - J a^2}. \tag{22}$$

Take independent random objects A, B, ε, N, also independent of H. Initially $A = \pm a$ equally, and $B \in \{0, \sqrt{2}a, -\sqrt{2}a\}$ with probabilities $(1/2, 1/4, 1/4)$. The vector ε is a random permutation of $((J-1)a, -a, \ldots, -a)$ multiplied by an independent fair sign. Coordinates $N_j = \pm s$ are independent fair signs. Let

$$r_A = \cosh a, \quad r_B = (1 + \cosh(\sqrt{2}a))/2, \quad r = (r_A + r_B)/2, \quad c(H) = \log(m(H)/r).$$

Set $(T_A, \eta_j^A) = (c(H) + A, B + \varepsilon_j + N_j)$ and $(T_B, \eta_j^B) = (c(H) + B, A + \varepsilon_j + N_j)$. Both give exactly

$$(H, Z_1, \ldots, Z_J) = \big(H, c(H) + A + B + \varepsilon_1 + N_1, \ldots, c(H) + A + B + \varepsilon_J + N_J\big). \tag{23}$$

Thus the complete observed laws, not just a selected set of moments, coincide.

Admissibility is preserved by the swap. The variables A, B have mean zero and variance a^2. Also $\mathbb{E}\varepsilon_j^2 = (J-1)a^2$ and, by exchangeability and $\sum_j \varepsilon_j = 0$, $\mathbb{E}\varepsilon_j\varepsilon_k = -a^2$ for $j \neq k$. The shared component's variance cancels this negative cross moment. Therefore $\mathbb{E}\eta_j\eta_k = 0$ and $\mathbb{E}\eta_j^2 = Ja^2 + s^2 = \theta b^2$. Each world's error vector is independent of its own (T, H), so the centering, cross-moment and exponential-stability assumptions hold conditionally too. Joint central symmetry makes all cubic error moments zero. Finally,

$$|\eta_j| \leq (J+1)a + s \leq (1 + \sqrt{\theta})b/2 < b. \tag{24}$$

For sufficiently small b, fixed $m_+ < 1$ gives $T_A, T_B < 0$, so both targets lie in $(0, 1)$. Both worlds give the same declared second-order and zero-third-moment side information.

Their target means given H are mr_A/r and mr_B/r. Inserting these in (20) gives opposite risk differences $-q_4, +q_4$, with

$$q_4 = K(r_B - r_A)/r = K\{a^4/24 - a^6/60 + O(a^8)\} = \Theta(b^4). \tag{25}$$

A uniformly zero-error rule must abstain on this shared population input, so $\rho^\Delta \geq q_4$. For a forced randomized choice the two error probabilities add to one; the larger expected regret is at least $q_4/2$, for the population experiment and for every n.

A continuous cubic obstruction. Replace B by B_t: with probability $1 - t$, a fresh fair $\pm a$, and with probability t, a variable on $\{-a/2, 2a\}$ with probabilities $(4/5, 1/5)$. It has mean zero, variance a^2 and third moment $3ta^3/2$. Choose

$$t = \min\{1, 2\kappa/(3a^3)\}, \qquad r_t = \mathbb{E}e^{B_t}, \qquad r = (r_A + r_t)/2. \tag{26}$$

Redefine $c(H) = \log(m(H)/r)$ and make the same swap. The support bound (24) still holds because $|B_t| \leq 2a$; all second-order and conditional checks above are unchanged. The error third moments are $3ta^3/2$ in world A and zero in world B, meeting the same cap. Moreover

$$r_t - r_A = t\{(4e^{-a/2} + e^{2a})/5 - \cosh a\} = t\{a^3/4 + O(a^4)\}.$$

Consequently the opposite gap satisfies $q_3 = K(r_t - r_A)/r \geq Kta^3/8$ for sufficiently small b. Since $ta^3 = \min\{c_{\theta,J}^3 b^3, 2\kappa/3\}$ and $\kappa \leq b^3$, this is at least $c\kappa$, uniformly over the cap. The same-law argument yields all $\Omega(\kappa)$ structural lower bounds for the marginal interface. Together with (25), this proves $\Omega(\kappa + b^4)$, by $\max(x, y) \geq (x + y)/2$.

A.3 Signed closure: identification, upper bound and same-law lower bound

With known biases removed, expand $W_i = (\eta_i - \eta_j)^2(\eta_i - \eta_k)$ to obtain

$$\begin{aligned} \mathbb{E}W_i &= \mu_{3,i} + \psi_i, \quad \mu_{3,i} = \mathbb{E}\eta_i^3, \\ \psi_i &= \mathbb{E}[-\eta_i^2\eta_k - 2\eta_i^2\eta_j + 2\eta_i\eta_j\eta_k + \eta_i\eta_j^2 - \eta_j^2\eta_k]. \end{aligned} \tag{27}$$

Because $|W_i| \le 8b^3$, an independent N-object mean has radius $8b^3\sqrt{2\log(2/\delta)/N}$ at failure probability δ. Adding κ_ψ gives h_3. Substitution of $(\widehat{\nu} \pm h_3)/6$ for the third-moment term in (18) gives width at most $h_3/3 + \theta(1-\theta)b^4/24 + e^b b^5/60$. At population level the sampling radius is zero, proving the $O(\kappa_\psi + b^4)$ upper bound by (20).

For expected finite-sample regret, split n into $N = \lfloor n/2 \rfloor$ calibration and $n - N$ evaluation objects. Project the calibration mean of W_i onto $[-b^3, b^3]$ to give $\widetilde{\nu}$. Since $\mu_{3,i}$ is in this interval, projection is nonexpansive relative to it, and

$$\mathbb{E}|\widetilde{\nu} - \mu_{3,i}| \le 8b^3/\sqrt{N} + \kappa_\psi.$$

Take $\widehat{G} = 1 + v/2 + \widetilde{\nu}/6 + (v^2 + b^2 v)/48$. For small b, $\widehat{G} \ge 1/2$, and

$$\mathbb{E}|\widehat{G} - G| \le (8b^3/\sqrt{N} + \kappa_\psi)/6 + (b^2 v - v^2)/48 + e^b b^5/120. \tag{28}$$

As $G \ge 1$, taking reciprocals multiplies this bound by at most two. Conditional on calibration, the argument of (21) on the independent evaluation split gives $O(\kappa_\psi + b^4 + n^{-1/2})$. For $n < 4$, bounded regret is covered by enlarging the constant.

For the lower bound, the quartic swap has all mixed cubic moments zero and hence $\psi_i = 0$. In the cubic swap the common A or B_t disappears from reader differences. The remaining difference vector of $\varepsilon + N$ is centrally symmetric, so $\mathbb{E}W_i = 0$ in both worlds. Equation (27) gives $\psi_i^A = -3ta^3/2$, $\psi_i^B = 0$. Choosing t by (26) with κ_ψ therefore gives an $\Omega(\kappa_\psi)$ obstruction with the same supplied closure cap. This also explains why a small observed signed mean cannot itself certify closure.

Closure does not require zero marginal skew. The quartic obstruction persists with a known nonzero common third moment. In its construction replace each N_j by a centered two-point law on $\{-s/r_\theta, r_\theta s\}$ with probabilities $r_\theta^2/(1+r_\theta^2)$ and $1/(1+r_\theta^2)$, where $r_\theta = (1+3\sqrt{\theta})/(4\sqrt{\theta}) > 1$. Its variance is s^2 and its third moment is $s^3(r_\theta - r_\theta^{-1}) = \Theta(b^3)$, the same in both worlds. Support is bounded by $(J+1)a + r_\theta s \le (3+\sqrt{\theta})b/4 < b$. The symmetric common component has zero mixed cubic moments; independence and centering remove mixed terms containing an N_j. Thus $\psi_i = 0$ in both worlds although the marginal third moment is nonzero. Their latent target means and q_4 are unchanged. Conversely, errors $a(U, V, UV)$ from independent fair signs have symmetric marginals and zero cross second moments but $\psi_i = 2a^3$, proving that marginal symmetry alone is not closure.

A.4 A SHARED SAMPLING OBSTRUCTION, INCLUDING EXACT TRANSPORT

Choose $0 < r_0 < \min(m_-, 1 - m_+)$, independently of b, n. Let S be independent of H, set $Y = m(H) + r_0 S$, and in two worlds put

$$\Pr_{\pm}(S = 1) = (1 \pm \epsilon)/2, \quad \epsilon = (4\sqrt{n})^{-1}, \quad \eta_j = \pm b\sqrt{\theta} \text{ independently and symmetrically.} \tag{29}$$

The worlds belong to every nonnegative cap class, have opposite contrasts $\Delta_\pm = \pm 2r_0\bar{w}\epsilon$, and use the same measurement kernel. Data processing and the Bernoulli divergence give

$$\mathrm{KL}(P_+^{n,\mathrm{obs}} \| P_-^{n,\mathrm{obs}}) \le n\epsilon \log \frac{1+\epsilon}{1-\epsilon} \le 4n\epsilon^2 = 1/4. \tag{30}$$

Pinsker gives $\mathrm{TV} \le 1/(2\sqrt{2}) < 1/2$, so the sum of the two testing errors is greater than $1/2$. The worst expected regret is at least $2r_0\bar{w}\epsilon/4 = r_0\bar{w}/(8\sqrt{n})$.

Crucially, the supplied exact side information in these worlds is identical:

$$G_j = \cosh(b\sqrt{\theta}), \qquad \lambda_j^* = 1/\cosh(b\sqrt{\theta}). \tag{31}$$

It depends on neither the world sign nor n. Adding it to the observations does not change (30), so the $\Omega(n^{-1/2})$ lower bound holds in $\mathcal{C}_G^{\mathrm{full}}$ itself. At population level, (3) determines the risk difference exactly, giving zero structural resolution and regret. Its bounded empirical mean gives the matching $O(n^{-1/2})$ upper bound. This establishes the complete exact-transport row without imposing an unspecified additional local condition.

Finally, combine the structural and sampling lower bounds by $\max(x, y, z) \ge (x + y + z)/3$. Different least-favorable subfamilies may supply different terms; no single world must attain all three. This completes Theorem 1. For Corollary 2, under the common data and algorithm-randomness law Q, the probabilities p_f, p_g of choosing f, g are each bounded by α using the opposite world's error constraint. Their disjoint union has probability at most 2α.

B CALIBRATION, MOMENT GEOMETRY AND DECISION CERTIFICATES

B.1 COMPARISON-SPECIFIC CALIBRATION

Stability gives $\mathbb{E}[\lambda_j^* X_j \mid T, H] = Y$; square expansion proves (3). For arbitrary λ, the tower property gives $\Delta_{fg} - \mathbb{E}d_{fg}(\lambda) = 2\langle g - f, r_\lambda\rangle = 2\langle g - f, \Pi_{\mathcal{D}} r_\lambda\rangle$. The finite-dimensional space $\mathcal{D}$ is closed. Orthogonality suffices; taking $g = f_0$ proves necessity for each generator, hence their span. Cauchy–Schwarz bounds the discrepancy by $2\|\Pi_{\mathcal{D}} r_\lambda\|_2 \|g - f\|_2$. An estimated population norm needs its own coverage. Affinity gives interval width $|\beta_{fg}|\,|\Lambda|$.

For an explicit crossing-predictor example on two equiprobable states, set $Y = (.3, .6)$, $d = g - f = (2e, -e)$, $f = .5 - d/2$, $g = .5 + d/2$, $0 < e \le .5$. Then $\mathbb{E}[dY] = 0$, so $\beta_{fg} = 2e^\gamma G\mathbb{E}[dY] = 0$, but $\Delta_{fg} = -\mathbb{E}d = -e/2 < 0$. This is not the ordered hard pair in Theorem 1.

B.2 LOWER CALIBRATION AND THE SIGNED PERTURBATION

Subtracting before expectation gives $A_j = \gamma_j + \eta_j - J^{-1}\sum_k \eta_k$, so $\mathbb{E}A_j = \gamma_j$ without moments of T, and $|A_j - \gamma_j| \le B_j = 2b(1 - 1/J)$. Hoeffding radii are $B_j\sqrt{2\log(2J/\alpha_\gamma)/N_0}$; the implementation uses empirical Bernstein with range width $2B_j$. Conditional centering and zero cross moments give $\mathrm{Var}(D_{jk}) = v_j + v_k = V_{jk}$. Independent object pairs yield $\mathbb{E}Q_{jk,r} = V_{jk}$ and $0 \le Q_{jk,r} \le 8b^2$. Intervals for V_{jk} clipped to $[0, 2b^2]$ imply

$$v_j \in \left[\frac{V_{jk,L} + V_{j\ell,L} - V_{k\ell,U}}{2}, \frac{V_{jk,U} + V_{j\ell,U} - V_{k\ell,L}}{2}\right] \cap [0, b^2]. \tag{32}$$

Intersect valid triples for additional readers, using $\lfloor N_0/2 \rfloor$ independent pairs; empty intersections are unresolved. A union bound permits shared bias/variance data. With $r = v/b^2$, derivatives of L_b, U_b have positive factors times $e^{br}\{e^{-b(1+r)} + b(1+r) - 1\}$ and $e^{-br}\{e^{b(1+r)} - 1 - b(1+r)\}$, proving the monotonicity used in (8).

Condition on lower calibration. With $A = \eta_i - \eta_j$, $B = \eta_i - \eta_k$ and residual bias differences d, e, expansion and zero cross moments give

$$\mathbb{E}[(A + d)^2(B + e)] = \mu_{3,i} + \psi_i + e(v_i + v_j) + 2dv_i + d^2 e. \tag{33}$$

If each bias error is at most r_γ, then $|d|, |e| \le 2r_\gamma$, the extra bias is at most $8b^2 r_\gamma + 8r_\gamma^3$, and $|(A + d)^2(B+e)| \le 8(b+r_\gamma)^3$. Thus an independent signed mean gives $\mu_{3,i} \in [\widehat{\nu} - h_3, \widehat{\nu} + h_3] \cap [-b^3, b^3]$, where

$$h_3 = r_W + \kappa_\psi + 8b^2 r_\gamma + 8r_\gamma^3, \qquad r_W = 8(b + r_\gamma)^3\sqrt{2\log(2/\alpha_W)/N_3}. \tag{34}$$

The implemented Bernstein alternative uses range width $2(2b + d_{\max})^2(2b + e_{\max})$ and bias bound $e_{\max}(v_{i,U} + v_{j,U}) + 2d_{\max}v_{i,U} + d_{\max}^2 e_{\max}$, for $d_{\max} = r_{\gamma,i} + r_{\gamma,j}$ and $e_{\max} = r_{\gamma,i} + r_{\gamma,k}$. Conditioning then integrating justifies its failure allocation. Taylor propagation and intersection with the second-order envelope preserve coverage.

B.3 SHARP SCALAR THIRD-MOMENT GEOMETRY

Proposition 5 (Attainable exponential bounds at three moments)**.** *For $|\eta| \le b$, $\mathbb{E}\eta = 0$, $\mathbb{E}\eta^2 = v \in (0, b^2)$, $s = \mathbb{E}\eta^3$ satisfies $|s| \le v(b^2 - v)/b$. Set*

$$\begin{aligned} z &= s/v, & x_\pm &= (z \pm \sqrt{z^2 + 4v})/2, & L_* &= \frac{x_+ e^{x_-} - x_- e^{x_+}}{x_+ - x_-}, \\ t &= -s/(b^2 - v), & p_0 &= \frac{b^2 - v}{b^2 - t^2}, & p_\pm &= \frac{v \mp bt}{2b(b \mp t)}. \end{aligned} \tag{35}$$

Then $L_* \le \mathbb{E}e^\eta \le U_* := p_- e^{-b} + p_0 e^t + p_+ e^b$, and both bounds are attainable.

Proof. Nonnegative polynomials give $\mathbb{E}[(b \mp \eta)(\eta \pm v/b)^2] = bv - v^2/b \mp s \ge 0$, proving feasibility. At the positive boundary, $\eta \in \{-v/b, b\}$ almost surely, with uniquely determined masses $b^2/(b^2 + v), v/(b^2 + v)$; reflection handles the negative boundary.

Feasibility gives $-b \le x_- < 0 < x_+ \le b$. Put masses $x_+/(x_+ - x_-), -x_-/(x_+ - x_-)$ at these nodes. They sum to one, have zero mean and, from $x^2 = zx + v$, moments v, s. The cubic Hermite interpolant q_L matches e^x and its derivative at both nodes, with remainder

$$e^x - q_L(x) = \tfrac{e^{\zeta_x}}{24}(x - x_-)^2(x - x_+)^2 \ge 0.$$

A cubic's expectation depends only on $(1, 0, v, s)$; it equals L_* for this law, proving the lower bound and attainment.

For the upper law, $|t| \le v/b < b$ makes p_-, p_0, p_+ nonnegative. Their sum and first two moments are $1, 0, v$. The node polynomial $x^3 - tx^2 - b^2x + b^2t = 0$ gives third moment $t(v - b^2) = s$. Hermite interpolation at $-b, b$ and doubly at t has remainder $e^{\zeta_x}(x + b)(x - t)^2(x - b)/24 \le 0$. Expectation proves $\mathbb{E}e^\eta \le U_*$, attained by this law. □

At $v = 0$, $G = 1$; at $v = b^2$, zero mean forces the fair endpoint law and $G = \cosh b$. To prove (10), integrate the lower remainder under the upper law. Multiplying its node polynomial by x gives $\mathbb{E}_{P_U}\eta^4 = b^2v - s^2/(b^2 - v)$, hence $\mathbb{E}_{P_U}\{\eta^2 - (s/v)\eta - v\}^2 = D_4$. Also $\mathbb{E}_{P_U} q_L = L_*$. Bounding e^{ζ_x} by e^{-b}, e^b gives the asserted two-sided bound, including zero boundary width. Since $D_4 = b^4\theta(1 - \theta)(1 - \xi^2)$, approaching $1 - |\xi| \asymp b^a$ gives scalar width $\Theta(b^{4+a})$, without changing the full-class frontier. Exact bias yields $|\Lambda_*| = e^{-\gamma}(U_* - L_*)/(L_*U_*)$ and risk-difference width $|\beta_{fg}|\,|\Lambda_*|$.

B.4 A COMPUTABLE CERTIFICATE WITH UNCERTAIN CALIBRATION

On calibration coverage, the true moments lie in $\mathcal{K} = \{(v, s) : v_L \le v \le v_U,\ s_L \le s \le s_U,\ 0 \le v \le b^2,\ |s| \le v(b^2 - v)/b\}$. Optimizing L_*, U_* over $\mathcal{K}$ would give sharp scalar envelopes. The implementation instead uses a conservative polynomial dual. Any distinct rational lower nodes $\ell, r \in [-b, b]$ and upper node $t \in (-b, b)$ give Hermite bounds by the preceding remainder signs. If rational approximations have uniform errors δ_L, δ_U, then $\widetilde{q}_L - \delta_L \le e^x \le \widetilde{q}_U + \delta_U$. Zero mean removes coefficient c_1, leaving rectangle extrema

$$\begin{aligned} G_L^d &= c_0^L + \min(c_2^L v_L, c_2^L v_U) + \min(c_3^L s_L, c_3^L s_U) - \delta_L, \\ G_U^d &= c_0^U + \max(c_2^U v_L, c_2^U v_U) + \max(c_3^U s_L, c_3^U s_U) + \delta_U. \end{aligned} \tag{36}$$

Infeasible corners only enlarge the envelope. Intersect with second-order/Taylor bounds and $G \ge 1$, then set $\Lambda = [e^{-\gamma_U}/G_U, e^{-\gamma_L}/G_L]$; an empty intersection gives no output. No statistical failure event is added.

Rational sums $\sum_{k=0}^{40} x^k/k!$ enclose e^x with remainder $3^{\max(1,\lceil|x|\rceil)}|x|^{41}/41!$ for $|x| \le 4$, using $e < 3$. Exact inversion of the Hermite system propagates data errors as $|A^{-1}|\epsilon$; their b^k-weighted sum bounds polynomial error. Denominator-2^{80} coefficient rounding includes its error; denominator-2^{24} center-based nodes are checked on the support. Node optimality is unnecessary. Bias exponentials, reciprocals and float conversion are outward bounded. Records contain all nodes, coefficients, errors and endpoints; unsupported inputs are rejected and degeneracy falls back. This verifies deterministic processing conditional on supplied intervals, not the statistical assumptions.

B.5 SIMULTANEOUS COVERAGE, REFINEMENT AND MARGIN

Conditional on independent calibration, endpoints are fixed. The two-sided Bernstein bound (Maurer & Pontil, 2009) and a union bound over $M(M - 1)$ unordered-pair endpoint events cover all ordered means, since $d_{gf} = -d_{fg}$. Population affinity covers every interior multiplier. Integrating over calibration proves Theorem 3. Contrasts may depend on one another. Searching R declared states and J' readers requires $RJ'M(M - 1)$ events.

The gates follow from $\Delta \le U$ and $\mathrm{Reg}(f) = \max_g \Delta_{fg}$. On the coverage event E, an improvement update A has $\Gamma > \tau$, proving $P(A, \Gamma \le \tau) \le \alpha$. Since risks are in $[0, 1]$, good updates gain at least τ and bad ones at least -1, giving

$$\mathbb{E}[\Gamma \mathbf{1}_A] \ge \tau P(A) - (1 + \tau)\alpha. \tag{37}$$

Similarly a tolerance output event C gives $\mathbb{E}\,\mathrm{Reg}(\widehat{f}_{\mathrm{deploy}}) \le \varepsilon P(C) + \mathbb{E}[\mathrm{Reg}(f_0)\mathbf{1}_{C^c}] + \alpha$. For the entire curve, the same event gives $\Gamma_{f^\dagger} \ge \widehat{\Gamma}_{\max}$, making $\widehat{\Gamma}_{\max} > \tau \ge \Gamma_{f^\dagger}$ impossible for any $\tau \ge 0$.

Thus the union of threshold failures is contained in E^c. Fixed tie-breaking ensures measurability; invalid calibration or no alternative yields $-\infty$ and no output. Performance bands are different: K predeclared rates over N independent pools admit a Hoeffding union radius $\sqrt{\log(2K/\delta)/(2N)}$.

For refinement, the centered sample vector is affine in λ, its norm (standard deviation) is convex, and a fixed-domain range is a supremum minus an infimum of affine functions. Therefore $\bar{d} + r$ is convex and its interval supremum occurs at an endpoint. Identical data, ranges and failure allocation give $U^S \le U^2$ for $\Lambda_S \subseteq \Lambda_2$, expanding qualifying sets without implying lower realized risk of a different selected candidate.

An affine population contrast with coefficient $\beta \neq 0$ uses endpoint $c + \mathrm{sgn}(\beta) r$. Subtracting its value at λ^* gives $B = |\beta|\{r + \mathrm{sgn}(\beta)(c - \lambda^*)\}$; coverage implies $0 \le B \le 2|\beta| r$, and $\beta = 0$ gives $B = 0$. For the same candidate, $B_T - B_D = \beta(c_T - c_D) + |\beta|(r_T - r_D) \ge 0$ on nested intervals. Adding and subtracting the actual population supremum proves (13); endpoint coverage gives $\mathbb{E}d \le \bar{d} + r \le \mathbb{E}d + 2r$, hence $0 \le V \le 2r_{\max}$. On failed coverage, the algebra still holds and losses are not clipped.

For a population outer interval, write $h^p_{fg} = \sup_{\Lambda_p^{\mathrm{pop}}} \mathbb{E} d_{fg}$ and $e_p = d_H(\widehat{\Lambda}_p, \Lambda_p^{\mathrm{pop}})$. The contrast is $|\beta_{fg}|$-Lipschitz, so the suprema differ by at most $|\beta_{fg}| e_p$. Endpoint coverage yields

$$h^p_{fg} - |\beta_{fg}| e_p \le U^p_{fg} \le h^p_{fg} + |\beta_{fg}| e_p + 2 r^p_{fg,\max}. \tag{38}$$

An upper bound below $-\tau$ suffices; a lower bound at least $-\tau$ rules out that candidate. Bounded calibration gives $r_\gamma = O(b\sqrt{t/N_0})$, $r_v = O(b^2\sqrt{t/N_0})$, $r_W = O(b^3\sqrt{t/N_3})$. Matching $O(b^4)$ requires, sufficiently, $N_3 \gtrsim tb^{-2}$, $N_0 \gtrsim tb^{-4}$ for unknown variance or tb^{-6} for unknown bias, and $\kappa_\psi = O(b^4)$; nonvanishing contrast variance also requires $n \gtrsim tb^{-8}$. Here t is a confidence logarithm. For positive fixed coefficients, Lagrange differentiation $a_\ell/(2N_\ell^{3/2}) = \zeta c_\ell$ gives the surrogate allocation $N_\ell = B(a_\ell/c_\ell)^{2/3}/\sum_k a_k^{2/3} c_k^{1/3}$. If $m_p = -\tau - h^p_{ff_0} > 0$ and statistical error is at most $A_p/\sqrt{B}$, then $B > (A_p/m_p)^2$ suffices; $m_p \le 0$ cannot pass the population envelope. These are protocol-specific sufficient scales, not optimal acquisition claims, and coefficients cannot use evaluation risks.

B.6 Risk-difference-preserving stochastic-target control

Let $Y_a = \mu(H) + aS$ under the amplitude restriction in the main text. Condition on the frozen pool and independent calibration. Error independence from (H, S) preserves all conditions in (2) for $T_a = \log Y_a$. Expanding the squared loss and using $\mathbb{E}[S \mid H] = 0$ proves (16); all gains, regrets and opportunities are invariant. Differences $Z_{j,a} - Z_{k,a} = \gamma_j - \gamma_k + \eta_j - \eta_k$ preserve every calibrated statistic, interval and true multiplier under common error realizations.

The added contrast $2(g - f)\lambda e^{\gamma+\eta} aS$ has conditional mean zero given (H, η), hence is orthogonal to $d^{(0)}$. Its second moment proves (17), also conditional on random calibrated endpoints. Population means and B remain fixed; empirical variance and V need not increase on every realization.

When log conditional second moments exist, total covariance and (2) give $\mathrm{Cov}(Z_j, Z_k \mid H) = \mathrm{Var}(T \mid H)$. Zero cross covariance forces deterministic T in every compatible decomposition, with $Y(H) = \exp\{J^{-1}\sum_j \mathbb{E}[Z_j \mid H]\}$ under zero-sum bias. For $a > 0$, instead $\exp\{\mathbb{E}[T_a \mid H]\} = \sqrt{\mu(H)^2 - a^2} < \mu(H)$ and $\mathrm{Var}(T_a \mid H) = \frac{1}{4}[\log\{(\mu(H) + a)/(\mu(H) - a)\}]^2 > 0$. This removes the zero-variance convenience but does not prove nonidentification; the swap does. No log-moment assumption is added to the general frontier.

C Experimental protocols and reproducible units

Randomization, inference and analysis. SHA-256-keyed PCG64DXSM streams separate training, lower calibration, signed calibration and evaluation. Multinomial counts encode independent finite-atom draws. The confirmation uses seeds 9000–9599 and one locally locked primary comparison; its code and configuration hashes precede outcomes. The discordance test and 10,000 paired bootstrap resamples assess the primary; threshold and mechanism analyses are secondary. Exploratory cohorts retain their original Taylor-primary/dual-secondary roles. Cross-budget and cross-amplitude resampling keeps each pool together.

Bias/variance/signed/evaluation failure allocations are $.01/.01/.01/.02$; native second order uses $.015/.015/.02$. Confirmation and the stochastic control use $M(M-1)$ endpoint events; ordered controls and exploratory cohorts use $2M(M-1)$. Gates require $U < -\tau - 10^{-12}$; scores within 10^{-12} of τ count as reaching it. The mathematical guarantee remains strict. Point exact scoring uses the minimum-risk set; tolerance outputs and switches are distinct. Rate bars are pointwise 95% Clopper–Pearson intervals and zero-failure limits are one-sided.

Complete-law and scalar mechanisms. The grid fixes $\theta = .5$, equiprobable $H = \pm 1$, $m = .5 + .15H$, $w = .2$, $J = 3, 5$, and $b = .0125, .025, .05, .075, .1, .2, .4, .8$. Caps are $0, .1\kappa_{\mathrm{sat}}, .5\kappa_{\mathrm{sat}}, \kappa_{\mathrm{sat}}, 2\kappa_{\mathrm{sat}}, b\kappa_{\mathrm{sat}}, \sqrt{b}\kappa_{\mathrm{sat}}, b^3$ with $\kappa_{\mathrm{sat}} = 3a^3/2$; couplings have 1,024 objects. The sampling witness uses $r_0 = .1, b = .4$, five $n = 100 \times 4^k$ ($k = 0, \dots, 4$), and 2,000 repetitions per sign and n. Its one-reader evaluation checks the rate; the proof covers panels. Finite-interface diagnostics use both swaps, $J = 3, b = .4, .8$, $n = 10^3, 10^4, 10^5$, 100 count draws per cell, cubic $t = .5$ and shared cap $3a^3/4$, without world-specific exact transport. The 100-digit scalar grid uses $\theta = .2, .5, .8$, the seven radii excluding .075, and $\xi = 0, \pm.5, \pm.9, \pm.99, \pm.999999, \pm 1$.

Channels, candidates and cost. Learning uses $J = 3, b = .8, \gamma = (.06, -.015, -.045)$, declared $|\gamma_j| \le .1$, variance $b^2/2$, and errors independent of (T, H). Equation (15) supplies interior $u = .5$ and boundary $u = 1$ channels. The dependent closed channel multiplies independent fair signs by common $R = b/2, b$ with probabilities $2/3, 1/3$. The nonclosed channel mixes independent signs (80%) and (U, V, UV) (20%), scaled by $a = b/\sqrt{2}$: cross second moments vanish, $\mu_3 = 0$, and $\psi = .4a^3$.

On 25 uniform input states, $\mu = .42 + .15\sin(\pi H) + .06H^2$. Legendre fits of degrees 1,3,5 use penalty .01 and the mean of three proxy readings on 512 training objects. With $s_0 = e^{-b^2/4}$, candidates are the proxy mean, s_0 times linear/cubic fits, $(s_0 - .015), s_0, (s_0 + .015)$ times the quintic fit, and raw quintic, clipped to $[0, 1]$; corrected cubic is incumbent. For validation budget B, $N_0 = 2\lfloor .35B/6 \rfloor$, $N_3 = \lfloor .35B/3 \rfloor$, $n = B - 3N_0 - 3N_3$. Native controls allocate 35/0/65 or 70/0/30. Budgets are $6000, 30000, 150000, 750000, 3750000$ plus 1,536 training readings. Each exploratory synthetic cohort has 600 pools; its fixed seed ranges are recorded separately in the supplied protocol files.

Frozen-pool stochastic control. The existing 600 confirmation pools and per-protocol calibration intervals are held fixed. A separate local lock fixes $a = 0, .025, .075$, budgets $150000, 750000, 3750000$, methods, generator and dependencies before fresh evaluation counts. Each joint (H, η, S) atom has original probability times $1/2$. The same count array is used across amplitudes within a protocol, while protocols retain their distinct evaluation counts and streams. All used calibration remains charged; the exact-transport reference is labeled separately. Targets score outputs and variance identities only. A metadata-key error was corrected before any samples were generated; the unchanged design and both code locks are archived. No seeds or amplitudes were selected from outcomes.

Real features and natural readers. Diabetes targets are $Y = y/(y + 200)$. Each fixed 265/88/89 partition standardizes inputs on training rows, learns ridge (penalty 5) and radial-basis kernel ridge (penalty 1, kernel parameter .1) from proxies, and restores the centered response mean. Candidates are proxy mean, raw ridge, s_0 ridge, $(s_0 - .02), s_0, (s_0 + .02)$ kernel fits, and raw kernel; corrected kernel is incumbent. Each partition/channel has 200 repetitions and budgets $6000, 60000, 600000, 6000000$ plus 795 training readings, drawing fresh noise on the held-out empirical law.

Natural-reader ratios exclude the nonpositive object `IMG_3480`, leaving 1,073 objects: 566 fit, 243 calibrate, 264 evaluate. Candidates are the $.1, .25, .4, .5, .6, .75, .9$ fitting quantiles, with median incumbent. Four radii $.35, .5, 1, 2$, five full-panel readers and five four-reader removal panels form 100 scenarios, each with $\alpha_\gamma = \alpha_v = .0125$, $\alpha_{\mathrm{eval}} = .025$, $\varepsilon = .01, \tau = 0$. Coverage is per panel/radius. At fixed multiplier center c, scales $a, t \in \{0, .05, \dots, 1\}$ use $U_{fg}(a, t) = \max_{\lambda = c \pm ar_0}\{\bar{d}_{fg}(\lambda) + tr_{fg}(\lambda)\}$; endpoint variances and domain ranges are recomputed. Scaling t is a sensitivity diagnostic, not a new sample size. Reader removal changes the panel-relative scale; no absolute anchor or patient-level independence is inferred.

D SUPPLEMENTARY DECISION EVIDENCE AND COMPUTATION

Budget and promise strength. Table 3 reports useful updates/shortfalls, preserving point-rule utility, small-budget abstention and the nonzero-closure reversal. The original $\tau = 10^{-4}$ primary remains fixed. Table 4 varies only the request on the existing maximum-budget matrices. Its pointwise rates do not justify choosing a new experimental primary. The continuous-threshold guarantee concerns the validity of one matrix, not empirical superiority at every threshold.

Cohort	Validation cost	Second order	Taylor	Dual	Point	Regret: D/P $\times 10^4$
Confirmation	6,000	0/0	0/0	0/0	76/421	0.979/3.287
Confirmation	30,000	0/0	0/0	0/0	125/240	0.979/0.625
Confirmation	150,000	0/0	0/0	0/0	212/74	0.979/0.259
Confirmation	750,000	5/0	4/0	8/0	245/32	0.934/0.217
Confirmation	3,750,000	70/0	67/0	95/0	259/12	0.634/0.217
Interior	6,000	0/0	0/0	0/0	70/427	1.006/3.861
Interior	30,000	0/0	0/0	0/0	137/243	1.006/0.632
Interior	150,000	0/0	0/0	0/0	222/70	1.006/0.274
Interior	750,000	5/0	2/0	3/0	268/25	0.991/0.209
Interior	3,750,000	52/0	50/0	83/0	277/11	0.697/0.210
Nonzero closure	6,000	0/0	0/0	0/0	173/345	1.072/2.824
Nonzero closure	30,000	0/0	0/0	0/0	222/196	1.072/0.331
Nonzero closure	150,000	0/0	0/0	0/0	239/72	1.072/0.308
Nonzero closure	750,000	8/0	1/0	1/0	235/55	1.069/0.352
Nonzero closure	3,750,000	65/0	33/0	39/0	216/72	0.963/0.392

Table 3: Useful updates/shortfalls per 600 pools for the improvement request. Second order buys more evaluation; Point is dual midpoint. Budgets exclude 1,536 common training readings. Regret includes fallback. Cohorts remain separate.

$\tau \times 10^4$	Second order	Taylor	Dual	Point
0	381/0	377/0	394/0	478/4
0.1	357/0	354/0	371/0	469/4
0.5	258/0	256/0	280/0	406/10
1	70/0	67/0	95/0	259/12
2	0/0	6/0	12/0	66/1
5	0/0	0/0	0/0	0/0

Table 4: Same 600 confirmation pools at 3,751,536 readings: useful updates/shortfalls as the request varies. Thresholds are in 10^{-4} units; the original primary is the row $\tau \times 10^4 = 1$.

Stochastic-target response. The new control preserves each frozen pair's risk difference and each method's population interval gate. Table 5 includes every amplitude and budget. All certified methods show zero shortfalls, exact-choice errors and tolerance errors; the point utility remains higher at large budgets. At maximum cost, the dual midpoint's useful/shortfall counts are $264/16, 265/15, 263/16$ across amplitudes. The $a = 0$ row uses fresh evaluation counts and is distinct from the original confirmation. At $a = .075$ the dual useful-update change from this control is -1.50 percentage points, paired 95% interval $[-2.50, -.67]$; this is a secondary effect on existing pools. Across 1,587,600 candidate-endpoint checks, the exact variance identity's largest floating residual is 1.39×10^{-17}. The proof, rather than this grid, establishes the identity.

Active endpoints and matched comparators. For each method, the contender is the observable $\arg\min_{f \neq f_0} U_{ff_0}$ even when it abstains. Grouping by primary dual/second-order success gives 53 both, 42 dual-only, 17 second-order-only and 488 neither. In the 17 second-order-only pools, median $(\Gamma - \tau, B, V)$ in 10^{-4} units is $(.218, .031, .143)$ for second order and $(.218, .011, .264)$ for dual: a tighter interval can lose to a larger evaluation penalty. In the 42 dual-only pools, dual and Taylor have median B of .679 and 1.079, respectively. Full distributions, signs of β and candidate identities are supplied; the two methods can propose different candidates, so group medians are not an additive

Validation cost	a	Second order	Taylor	Dual	Point	Dual exact	Dual tol.
150,000	0	0/0	0/0	0/0	215/60	0	598
150,000	0.025	0/0	0/0	0/0	214/62	0	598
150,000	0.075	0/0	0/0	0/0	210/68	0	598
750,000	0	9/0	4/0	10/0	249/25	15	600
750,000	0.025	8/0	3/0	10/0	251/23	15	600
750,000	0.075	8/0	1/0	6/0	250/21	14	600
3,750,000	0	71/0	65/0	91/0	264/16	133	600
3,750,000	0.025	69/0	62/0	90/0	265/15	132	600
3,750,000	0.075	66/0	58/0	82/0	263/16	131	600

Table 5: Risk-difference-preserving control, 600 existing frozen pools per cell. Entries are useful/shortfall counts. Dual exact/tolerance count certified outputs, not necessarily switches. All budgets additionally charge 1,536 training readings and their reused calibration; amplitude changes only fresh evaluation observations.

causal decomposition. Of the neither pools, 159 still have a true improvement opportunity; these groups never filter evaluation or choose the protocol.

Same-guarantee efficiency. Ordered/unordered event allocations give maximum-budget useful counts $69/70, 41/41, 62/67, 92/95$ for more-evaluation second order, more-calibration second order, Taylor and dual, respectively; exact counts are $27/28, 28/28, 74/75, 132/134$. Same-data matrix containment holds for 12,000 controls. Pure processing adds 28 dual-over-Taylor successes without a Taylor-only success; the full same-cost comparison retains 17 second-order-only successes.

Additional channels and requests. At the boundary, Taylor/dual/midpoint useful counts are $100/147/308$; midpoint has five shortfalls and regret $.192 \times 10^{-4}$ versus dual's $.592 \times 10^{-4}$. In the three-channel exploratory cohort, Taylor/native second-order useful counts are $79/49, 57/56, 31/58$, while exact certificates are $129/31, 55/8, 17/11$ for boundary, dependent closure and nonzero closure. Thus request choice also matters; Taylor's 600 tolerance outputs at boundary produce only 370 switches. All method-level records remain available. In the separate nonzero-closure cohort, second order/Taylor/doubled-cap Taylor/dual yield $65/33/29/39$ useful updates. An invalid zero cap yields 51 without observed shortfall, which does not validate closure. A known-nuisance stress pair $f_0 = Y + .0005, f_1 = Y - .0035$ increases risk by 1.2×10^{-5}: at $n = 10^8, N_3 = n/10, \tau = 0$, invalid closure gives 600 harmful switches per 600 seeds and valid closure retains f_0.

Partition	Channel	More eval.	More cal.	Taylor	Dual	Dual shortfall
1	Skew	57	58	70	80	0
1	Nonclosed	39	40	39	39	0
2	Skew	53	52	91	111	0
2	Nonclosed	34	30	40	45	0
3	Skew	95	95	110	118	0
3	Nonclosed	57	58	60	64	0

Table 6: Real-feature useful updates per 200 repetitions at 6,000,795 readings, with zero dual shortfalls. Native controls purchase more evaluation or calibration. Overlapping partitions remain separate.

For natural readers, the four fixed-center (exact, tolerance, improvement) counts are $(0, 0, 0)$ with no uncertainty removed, $(0, 0, 0)$ with multiplier width removed, $(0, 29, 0)$ with evaluation radius removed, and $(100, 100, 80)$ with both removed. These are sensitivity outputs, not newly certified confidence levels. Fixing bias/variance rather than multiplier centers yields $91/100/41$ exact/tolerance/improvement outputs after removing penalties, versus $100/100/80$ at the fixed multiplier point. For the 3,000 recorded arithmetic conversions, median time is 11.22 ms (95th percentile 13.24 ms), record size 1,335 bytes and fallback count zero; matrix formation takes .115 ms.