CERTIFYING DATA RANKINGS UNDER DEPLOYMENT UNCERTAINTY: WHEN BENCHMARK AGREEMENT TRANSFERS AND WHEN TO ABSTAIN

Anonymous authors
Paper under double-blind review

ABSTRACT

Data products are often ranked before the downstream learner is fixed. In a prospectively frozen study on untouched human-activity data, histogram boosting and a standard multilayer perceptron (MLP) significantly prefer opposite acquisition products, drawn from different subjects, under familywise simultaneous intervals. The comparison averages over the registered acquisition and training randomness while the archive, training pools, and test cohort stay fixed, so it does not claim that the reversal will recur for new subjects or sites. The result shows why a single "best dataset" is not always the right thing to acquire. We then ask when agreement among benchmark learners can still certify a ranking under deployment uncertainty. In an exact Bayesian linear–Gaussian model with a shared target, each dataset induces an information operator and each executable learner induces a trace response. Preferring one dataset for every learner in a declared family is then exactly a dual-cone order. No finite benchmark catalog covers all responses in dimension two or more. Projection onto the benchmark cone gives the exact largest normalized reversal that every benchmark accepts, and finite datasets realize every such separating direction. On the positive side, if the convex hull of the benchmarks covers the deployment family to within ϵ, a benchmark margin γ transfers whenever $\gamma > \epsilon M$, and the auditor abstains otherwise. Learners whose response depends on the data add an exact one-sided drift penalty ρ_-. In controlled families the certificate is sound and non-vacuous: a registered sign-aware rule certifies 236 of 480 ridge pairs, which is 76.1% of the 310 pairs where a conservative oracle confirms a uniform preference, with no false certificate. The natural certificate arms return no positive certificates, and a second prospective dataset is ruled uncovered because its samples miss classes. The theorem thus certifies rankings only inside an explicitly covered response family. Learner-free scores remain useful proxies, but a ranking that holds for every deployment requires a declared family, measured coverage, and honest abstention.

1 INTRODUCTION

Which dataset should we acquire when we do not yet know which learner will be trained on it? Consider two data products from the Human Activity Recognition (HAR) archive (Reyes-Ortiz et al., 2013). Each contains 256 labeled windows. One is drawn from odd-numbered subjects and the other from even-numbered subjects, and both are evaluated on the same untouched test set of different subjects. Across 200 randomized runs of a registered acquisition and training procedure, histogram boosting significantly prefers the even-subject product (mean odd-minus-even utility -0.0359, simultaneous 99.167% interval $[-0.0500, -0.0219]$). A 32-unit MLP significantly prefers the odd-subject product (0.1576, $[0.0491, 0.2648]$). The protocol, products, learners, 200 paired seeds, multiplicity correction, and failure policy were all fixed before any labels or outcomes were opened, and a separately written verifier reproduced the result twice. The estimate averages over acquisition and training randomness on this archive, these training pools, and this test cohort, so it does not estimate how often such a reversal would occur for new subjects or sites.

Data is often acquired, screened, or priced before the final learning pipeline is fixed, which motivates learning-agnostic and model-agnostic data valuation. LAVA scores data with hierarchical class-wise

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.

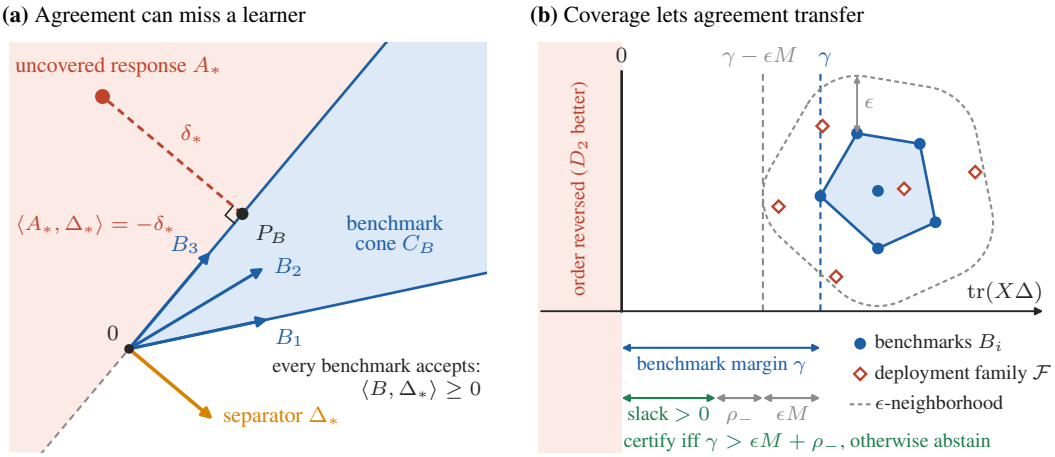

Figure 1: **Certify or abstain.** (a) Benchmarks B_i generate the cone C_B. The uncovered response A_* lies at distance δ_* from its projection P_B, and the separator Δ_* is the unit normal of the face containing P_B: every benchmark accepts it, while $\langle A_*, \Delta_* \rangle = -\delta_*$ (Theorems 6 and 9). (b) For one dataset pair, each response X sits at its utility difference $\mathrm{tr}(X\Delta)$. Benchmarks have margin at least γ, responses within ϵ of their hull keep at least $\gamma - \epsilon M$, and the rule certifies only if $\gamma > \epsilon M + \rho_-$ (Theorems 4 and 11). Tinted regions reverse the order.

optimal transport (Just et al., 2023), SAVA scales this approach to large datasets (Kessler et al., 2025), and KAIROS uses an influence objective based on the maximum mean discrepancy (MMD) (Zhu et al., 2025). These methods meet a real operational need and detect low-quality data efficiently.

The phrase *learning-agnostic* can mean two different things. A score may be computable without training any learner. Alternatively, the ranking it induces may be correct for every learner that could later be deployed. The first is a property of the algorithm. The second is a claim that has to be certified.

The difference matters when choosing between two data products. If a convenient suite of models agrees on which product is better, that agreement is evidence, but it is not a guarantee for a learner that nobody has declared. A trustworthy audit should state which learners it covers, transfer a ranking only when the margin survives the measured coverage error, and otherwise return *abstain*. Our positive result makes this selective interface precise.

We study the second meaning while keeping the evaluation distribution, the loss, and the target fixed. The only uncertainty is which information the downstream learner can use. Two data products may carry information in different directions, and two constrained learners may value those directions differently. The correct ordering over a family of learners is therefore usually partial rather than total.

Our contributions are threefold:

1. **A prospectively frozen finding of deployment sensitivity.** Two ordinary learners significantly disagree about which of two data products is better under the same target. The inference covers the registered randomization on one untouched archive (Section 8).

2. **An exact geometry of what benchmark agreement can prove.** The ordering over a declared learner family is a dual-cone order. Finite benchmark catalogs cannot cover unrestricted learners. We give the exact margin by which an uncovered learner can escape, and finite datasets that realize every separating direction (Sections 4 and 5).

3. **A selective certificate tested under frozen protocols.** Convex-hull coverage, the benchmark margin, and signed response drift combine into a rule that either certifies or abstains (Sections 3 and 6). Controlled studies show regimes where the rule is sound and informative. On natural data it certifies nothing, and we report these null and uncovered outcomes rather than redefine them as success.

2 Related Work and Positioning

Data valuation. Data Shapley defines the value of data relative to a specified learning algorithm and utility (Ghorbani & Zou, 2019). LAVA instead optimizes a model-agnostic, class-wise Wasserstein proxy (Just et al., 2023). SAVA computes this proxy in batches (Kessler et al., 2025), and KAIROS uses a closed-form MMD influence (Zhu et al., 2025). These objectives work well for distribution matching, corruption detection, and scalable ranking. Our question is different: when does a single scalar ordering certify the utility order over an unknown family of deployments?

Validation-free and task-agnostic valuation also includes moment-based and distribution-level approaches (Mohammadi Amiri et al., 2023; Xu et al., 2024). These works change the object being valued or rely on a distributional proxy. In our setting the Bayesian target and the evaluation loss stay fixed, and the only uncertainty is the response of the downstream learner. Data Distribution Valuation, for example, gives finite-sample comparison policies for an MMD-based value of vendor distributions under Huber heterogeneity (Xu et al., 2024). It certifies the chosen distributional proxy, not the utility order over a family of retrained learners.

OpenDataVal finds that no valuation method is best across all downstream tasks (Jiang et al., 2023), and Tamine et al. (2026) describe geometrically how semivalues depend on the choice of utility. We instead hold the target utility fixed and vary the learner. This lets us derive exact escape margins and coverage conditions rather than only observe that rankings vary.

Predictive $\mathcal{V}$-information formalizes information relative to a computationally constrained family of predictors (Xu et al., 2020), and its use for comparing dataset difficulty makes model-relative usable information explicit (Ethayarajh et al., 2022). We do not claim that observer- or model-relative information is new. Our object is the order between two acquisition products, together with an exact escape margin for finite benchmarks and a condition under which the order transfers to a covered family.

LARP protects a pre-specified set of learners by optimizing worst-case loss under corruption (Minchev et al., 2026). Its motivation is close to ours, but it designs a prefiltering rule for one dataset, whereas we compare two candidate products and ask whether benchmark agreement certifies their order. Utility underspecification can make semivalue point valuations arbitrary (Diehl & Wilson, 2025), and a Data Shapley vector can be insufficient for subset selection (Wang et al., 2024). Those obstructions vary the set utility or its compression, not the learner. DaSH selects whole datasets under resource constraints (Zhou et al., 2026), and DsDm makes selection model-aware for specified target tasks and a given learning algorithm (Engstrom et al., 2024). Bayesian data-source selection has also been posed as cost-constrained submodular covering for a specified learning model (Ye et al., 2021). Our uncertainty is instead which response in a declared family will use the acquired data.

Robust valuation and comparison of experiments. Distributionally robust data valuation protects against uncertainty in the validation distribution (Lin et al., 2024), whereas we hold the target distribution fixed. KAIROS bounds a distributional risk difference for a fixed continuous loss using marginal and conditional discrepancies (Zhu et al., 2025). That bound does not order two downstream algorithms after each is retrained on a different candidate dataset. Our certificate addresses this ordering question, but only for response families that are explicitly represented and covered. Bayesian information matrices and the comparison of statistical experiments are classical (Blackwell, 1953; Pukelsheim, 2006), as are convex duality and cone projection (Boyd & Vandenberghe, 2004). **Novelty boundary.** We do not claim Blackwell comparison, PSD self-duality, cone projection, or the Cauchy–Schwarz inequality as new. Our contribution is their synthesis for data ranking. We map an executable learner family to a response cone. We convert a separator of a finite catalog into attainable finite datasets with an exact normalized escape margin. We then combine response-family coverage with signed, data-dependent drift in a selective ranking rule. We make no claim that arbitrary classifiers or observed-label utilities satisfy this trace model.

Projective covering and PSD approximation. Packings of finitely many points in real and complex projective spaces have long been studied (Dhillon et al., 2008), and random covering radii on smooth manifolds have known asymptotics (Reznikov & Saff, 2016). Polyhedral approximation of the trace-one PSD body is also established. Fawzi (2021) proves lower bounds on extension complexity and discusses rank-one constructions from sphere nets. We do not claim a new general theorem

in projective or convex geometry. We use only the qualitative obstruction for finite catalogs and the exact cone distance for each instance, and we make no claim about how many benchmarks are needed.

3 Same-Target Executable Learning Model

We work in a Bayesian linear–Gaussian model in which every learner is scored against the same target, so learners differ only in which information they can use. Let $\theta \sim \mathcal{N}(0, I_d)$ be the unknown parameter. A candidate dataset, or design, D consists of a finite matrix $X_D \in \mathbb{R}^{n_D \times d}$ and observations

$$y_D = X_D\theta + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, I_{n_D}).$$

We write $H_D = X_D^\top X_D$ for its information matrix, $\Sigma_D = (I_d + H_D)^{-1}$ for the posterior covariance, $m_D = \mathbb{E}[\theta \mid y_D, X_D]$ for the posterior mean, and

$$K_D = I_d - (I_d + X_D^\top X_D)^{-1}.$$

The matrix K_D measures how much the design reduces uncertainty about θ, and we call it the information operator. A deployment learner indexed by an orthogonal projection $P = P^\top = P^2$ can use only the directions in $\mathrm{range}(P)$ and returns Pm_D. Every learner is judged by the same squared error on the full parameter, and its utility is the resulting reduction in risk:

$$U_P(D) = \mathbb{E}\|\theta\|_2^2 - \mathbb{E}\|\theta - Pm_D\|_2^2.$$

Proposition 1 (Exact trace utility)**.** *For every finite design and projection learner,*

$$U_P(D) = \mathrm{tr}(PK_D), \qquad 0 \preceq K_D \prec I_d.$$

Nonnegative weighted combinations of rank-one projection utilities give a positive semidefinite (PSD) response A with utility $U_A(D) = \mathrm{tr}(AK_D)$. This represents an aggregate utility. It does not claim that every PSD matrix is itself an orthogonal projection estimator.

The trace form is not limited to projection learners. The next identity extends it to any linear or spectral learner, and it also shows where the fixed-response geometry used later stops applying.

Proposition 2 (Executable response of a linear estimator)**.** *Under the model above, let a linear estimator return $\widehat{\theta} = GX_D^\top y_D$ for any $G \in \mathbb{R}^{d\times d}$. Define*

$$\Phi_D = G(I_d + H_D), \qquad A_{\Phi_D} = I_d - (I_d - \Phi_D)^\top (I_d - \Phi_D).$$

For prior-risk reduction $U_G(D) = \mathbb{E}\|\theta\|_2^2 - \mathbb{E}\|\theta - \widehat{\theta}\|_2^2$,

$$U_G(D) = \mathrm{tr}(A_{\Phi_D} K_D).$$

No symmetry or commutation assumption on G is required. Moreover, $A_{\Phi_D} \succeq 0$ iff $\|I_d - \Phi_D\|_{\mathrm{op}} \le 1$.

For a spectral filter $G = g(H_D)$, each eigenvalue h of H_D contributes $\frac{h}{1+h}[1 - (1 - (1+h)g(h))^2]$ to the utility. The projection learner above has $G = P(I + H_D)^{-1}$, so $\Phi_D = P$ and Proposition 1 is recovered. A ridge learner with a fixed penalty is different. Its response A_{Φ_D} generally depends on H_D and therefore changes with the dataset. The next result handles this case directly instead of replacing it with a single fixed response.

Proposition 3 (Exact response-drift decomposition)**.** *Let a learner family have symmetric response $A_\alpha(H)$ and utility $U_\alpha(D) = \langle A_\alpha(H_D), K_D\rangle$. For two designs, write $\Delta_K = K_1 - K_0$. Then*

$$U_\alpha(D_1) - U_\alpha(D_0) = \langle A_\alpha(H_0), \Delta_K\rangle + \langle A_\alpha(H_1) - A_\alpha(H_0), K_1\rangle\,.$$

For ridge $G_\lambda(H) = (H + \lambda I)^{-1}$,

$$A_\lambda(H) = I - (\lambda - 1)^2 (H + \lambda I)^{-2}.$$

Thus $\lambda = 1$ is the Bayes learner under unit prior and noise variance, and its response is constant, whereas any other fixed λ gives a response that depends on the dataset. More generally, for prior covariance $\tau^2 I$ and noise variance $\sigma^2 I$, the Bayes ridge penalty in the original coordinates is $\lambda_{\mathrm{B}} = \sigma^2/\tau^2$. Rescaling by λ_{B} recovers the unit-calibration formulas below. The values 1 and $1/2$ therefore depend on this calibration and are not universal recommendations for regularization.

We can now state the certificate. It compares two designs using m benchmark responses, measures how far the deployment family lies from them, and charges a penalty for adverse drift.

Theorem 4 (Sign-aware dataset-pair coverage with response drift)**.** *Let $B_1, \dots, B_m$ be benchmark responses. Suppose every $A_\alpha(H_0)$ is within Frobenius distance ϵ of their convex hull, $\langle B_j, \Delta_K \rangle \geq \gamma$ for all j, $\|\Delta_K\|_F \leq M$, and for some $\rho_- \geq 0$ the adverse response drift obeys*

$$\langle A_\alpha(H_1) - A_\alpha(H_0), K_1 \rangle \geq -\rho_- \quad \text{for every } \alpha.$$

Then $U_\alpha(D_1) - U_\alpha(D_0) \geq \gamma - \epsilon M - \rho_-$ for every α. Hence, if $\gamma > \epsilon M + \rho_-$, every learner in the family prefers D_1 and the procedure certifies this preference. Otherwise it abstains.

Statistical meaning. Theorem 4 is deterministic given its inputs. Appendix C gives a version that holds with probability $1 - \delta$ when all four inputs are replaced by simultaneous confidence bounds, and it states what such bounds require. Our natural studies with observed labels do not meet these requirements, so we treat them as diagnostics only.

The smallest valid drift allowance is $\rho_- = \sup_\alpha [-\langle A_\alpha(H_1) - A_\alpha(H_0), K_1 \rangle]_+$. When only a norm bound is available, the Cauchy–Schwarz inequality gives the conservative choice $\rho_- = \sup_\alpha \|A_\alpha(H_1) - A_\alpha(H_0)\|_F \|K_1\|_F$. For ridge, let $\mu_i = \lambda_{\min}(H_i)$. The resolvent identity then gives

$$\|A_\lambda(H_1) - A_\lambda(H_0)\|_F \leq (\lambda - 1)^2 \left(\frac{1}{\lambda + \mu_1} + \frac{1}{\lambda + \mu_0} \right) \frac{\|H_1 - H_0\|_F}{(\lambda + \mu_1)(\lambda + \mu_0)}.$$

This bound accounts for conditioning. It is never worse than $2(\lambda - 1)^2 \lambda^{-3} \|H_1 - H_0\|_F$ and can be much tighter for informative designs. Moreover, $A_\lambda(H) \succeq 0$ for all $H \succeq 0$ if and only if $\lambda \geq 1/2$, which is why our core family starts at $\lambda = 1/2$ after unit rescaling. For $\lambda < 1/2$, weakly informed directions can produce indefinite responses. A signed-response order could be studied there, but PSD self-duality and the present cone certificate no longer apply uniformly, and we make no claim for that regime. Note that the drift term is multiplied by $\|K_1\|_F$, not by $\|\Delta_K\|_F$. The tempting replacement by $\|\Delta_K\|_F$ is false in general. The cone order in the next section is the exact special case of a fixed response.

4 Attainability and Learner-Family Order

We now characterize which information operators finite datasets can produce, and what it means for every learner in a family to prefer one dataset.

Theorem 5 (Information-operator characterization)**.** *A symmetric matrix K equals K_D for a finite real-valued Gaussian design if and only if $0 \preceq K \prec I_d$.*

Theorem 6 (Direction-complete dataset differences)**.** *For every nonzero symmetric Δ, finite designs D_1, D_2 and $c > 0$ exist such that*

$$K_{D_1} - K_{D_2} = c\Delta.$$

Any $0 < c < 1/\|\Delta\|_{\mathrm{op}}$ is valid.

Let $\mathcal{F}$ be a declared family of PSD response operators and define

$$C_{\mathcal{F}} = \mathrm{cl}\,\mathrm{cone}\{A_L : L \in \mathcal{F}\}.$$

Write $D_1 \succeq_{\mathcal{F}} D_2$ when every learner in $\mathcal{F}$ weakly prefers D_1.

Theorem 7 (Exact family-ordering cone)**.**

$$D_1 \succeq_{\mathcal{F}} D_2 \quad \Longleftrightarrow \quad K_{D_1} - K_{D_2} \in C_{\mathcal{F}}^*.$$

This is a preorder. Define $D \sim_{\mathcal{F}} D'$ iff every family response assigns equal utility to D and D'. The induced relation on dataset equivalence classes is a partial order.

If $C_{\mathcal{F}} = \mathbb{S}_+^d$, self-duality reduces the family order to the Loewner order. A valuation that is correct for every learner in the family must therefore represent a cone order that depends on the family. In general, this is not a scalar total order.

5 Finite Benchmarks and Exact Escape

In practice, a declared family is often represented by a finite set of benchmark learners. This section shows what such a set cannot guarantee and by exactly how much it can fail.

Theorem 8 (Finite-benchmark incompleteness)**.** *For $d \geq 2$, no cone generated by finitely many fixed PSD responses equals $\mathbb{S}^d_+$. Consequently, finite Gaussian designs exist that every benchmark weakly orders as $D_1 \succeq D_2$, while an unbenchmarked rank-one projection learner strictly prefers D_2.*

Let C_B be the cone generated by the finite benchmark responses, and let $A_* \notin C_B$ be a normalized response that the benchmarks do not cover. We measure distances between symmetric matrices in the Frobenius norm. Figure 1(a) illustrates the construction.

Theorem 9 (Distance-exact reversal)**.** *Let $\delta_* = \operatorname{dist}_F(A_*, C_B) > 0$. There is a unit-Frobenius Δ_* such that*

$$\operatorname{tr}(B\Delta_*) \geq 0 \quad \forall B \in C_B, \qquad \operatorname{tr}(A_*\Delta_*) = -\delta_*.$$

No accepted unit direction produces a larger reversal against A_*.

Corollary 10 (Strict attainable escape)**.** *Assume benchmark responses and A_* have trace one. Then $\widetilde{\Delta} = \Delta_* + (\delta_*/2)I_d$ has benchmark margins at least $\delta_*/2$ and uncovered margin $-\delta_*/2$. After any admissible attainability scaling, finite Gaussian designs realize these strict opposing margins.*

The normalization matters. Without it, a common additive margin could be made arbitrarily large or small by rescaling.

6 Positive Certification and Regret

The obstruction above concerns unrestricted learners. When the deployment family is known to lie close to the benchmarks, agreement does transfer. Let $\mathcal{F}$ be a compact normalized deployment family and let $B_1, \ldots, B_m$ be normalized benchmark responses.

Theorem 11 (Deployment-geometry coverage certificate)**.** *Suppose each $A \in \mathcal{F}$ is within Frobenius distance ϵ of $\operatorname{conv}(B_1, \ldots, B_m)$. For $\|\Delta\|_F \leq M$, if $\operatorname{tr}(B_i\Delta) \geq \gamma$ for all i, then*

$$\operatorname{tr}(A\Delta) \geq \gamma - \epsilon M \qquad \forall A \in \mathcal{F}.$$

A ranking is therefore certified only when $\gamma > \epsilon M$ (Figure 1b). Otherwise the right output is to abstain for the declared family, to expand the benchmark catalog, or to make an explicitly robust decision.

When the certificate abstains, a decision may still be needed. For fixed $\Delta = K_{D_1} - K_{D_2}$, define the largest advantage of each dataset over the family,

$$\alpha = \sup_{A \in \mathcal{F}} [\operatorname{tr}(A\Delta)]_+, \qquad \beta = \sup_{A \in \mathcal{F}} [-\operatorname{tr}(A\Delta)]_+.$$

Theorem 12 (Exact two-dataset minimax regret)**.** *The optimal deterministic deployment-blind worst-case regret is $\min(\alpha, \beta)$. If D_1 is selected with probability p,*

$$R(p) = \max\{(1-p)\alpha, p\beta\}.$$

For $\alpha + \beta > 0$, the optimum is $p_* = \alpha/(\alpha+\beta)$ with value $\alpha\beta/(\alpha+\beta)$. If either one-sided width is zero, a universally optimal dataset exists.

7 Experimental Design

Every confirmatory study followed the same sequence: a written protocol, an implementation freeze, pilot runs on separate seeds, and a confirmation run on disjoint seeds. Validity gates checked source hashes, deterministic replay, sample integrity, solver behavior, finite outputs, and unique configurations. A scientific null result was never treated as a pipeline failure. Table 1 lists the unit of evidence and the scope of inference for each study, and the supplementary material contains the full protocols, outputs, rulings, environments, and checksums. The configuration and cell counts

Table 1: **Experiment registry and evidence scope.** "Covered" means that the reported certificate is licensed by the exact trace-family model. Diagnostic classifier results are never promoted to theorem evidence.

Study	Evidence unit and theorem scope	Retained outcome
Exact and controlled	Finite Gaussian designs, declared trace responses (covered)	Exact escape and attainable reversals
Declared family	900 fresh pairs, 512 declared responses (covered)	Sound, non-vacuous certification
Continuous ridge	480 frozen pairs, $\lambda \in [0.5, 2]$ (covered)	236 sign-aware certificates, 0 false
Credit and Bank	100 acquisition/training seeds per size, observed-label learners (diagnostic)	Credit crossover, no Bank reversal
Earlier natural sources	Three frozen pairs, Gaussian certificate arm covered, observed labels diagnostic	0 certificates, reversal gate failed
Sealed HAR and Covertype	200 randomized runs on a fixed archive and test cohort, only the Gaussian ridge arm covered	HAR sensitive, 0 certificates, Covertype uncovered

reported below count algorithmic stress tests, not independent deployment populations. In the natural studies, the resampling unit is one registered acquisition and training seed on a fixed archive. We never present repeated optimizer runs or multiple cells as new subjects, sites, or domains.

All natural data products were fixed before any outcome was inspected. The credit products expose either repayment-history or liquidity measurements (Yeh & Lien, 2009). The Bank products expose documented client fields or campaign fields, with the post-contact call duration excluded (Moro et al., 2011). The sealed study compares HAR training pools from odd-numbered and even-numbered subjects, and Covertype products from the Rawah and Comanche Peak wilderness areas (Reyes-Ortiz et al., 2013; Blackard, 1998). HAR keeps its original test partition, whose subjects do not appear in training. Its 200 seeds vary which rows are acquired and how the models are trained, not the archive or the test subjects. The six sealed contrasts (three learners on two datasets) use simultaneous 99.167% paired seed-cluster bootstrap intervals, which is a Bonferroni correction at level $0.05/6$. Appendix C details the resampling and the Monte Carlo error. Proportions carry Wilson score intervals (Wilson, 1927), and the sealed study uses scikit-learn learners (Pedregosa et al., 2011). A separate Gaussian ridge analysis is never used to certify the observed-label classifiers.

8 Results

Exact and sampled obstruction. All 15 algebraic configurations pass, with reconstruction error below 1.7×10^{-15}. In dimension 16, a catalog of 64 random responses still leaves an audited cone distance of 0.958. Specialist crossover occurs in all 900 controlled configurations and in 1,752/1,800 (97.3%) IID configurations. Appendix C describes both designs.

Certificate calibration. For the broad family the rule is valid but always abstains, because the coverage radius is still 0.541 with 40 benchmarks. Over 900 new declared-family pairs and 512 responses, every lower bound holds and no false certification occurs. As the novel mass outside the catalog grows from 0 to 20%, strong pairs are certified in 300/300 cases, while certification of moderate pairs falls from 299/300 to 195/300. Pairs with equal acquisitions are never certified.

Credit-product diagnostic. All 300 credit-default configurations pass their validity gates. Here D_H exposes repayment history and D_L exposes liquidity. Crossover between a history specialist and a liquidity specialist rises from 36/100 at 64 records to 69/100 at 128 and 95/100 at 256 (Wilson 95% interval 88.8–97.8%). At 256 records, linear and boosted generalists select D_H in 100/100 cases, while the liquidity specialist selects D_L in 95/100. Literal LAVA selects D_H in 78/100 cases. It agrees with the history specialist in 78% of runs but with the liquidity specialist in only 27%. Conditional MMD agrees with the liquidity specialist in only 8% of cases. Mean AUCs confirm that

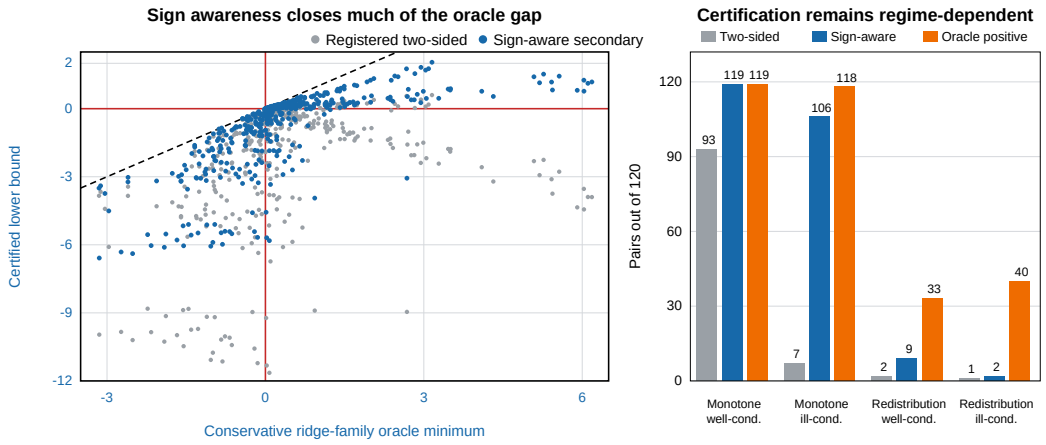

Figure 2: **What limits selective certification?** Continuous ridge core family with nine actively selected benchmarks. The registered two-sided primary rule certifies 103/480 pairs, and the preregistered secondary sign-aware rule certifies 236/480, with zero false certifications for either rule. Among the 310 pairs where the conservative family oracle is positive, the sign-aware rule recovers 76.1%, compared with 33.2% for the two-sided rule. Orange bars count these oracle-positive pairs. The remaining abstentions in the redistribution regimes mostly reflect unfavorable benchmark margins rather than coverage error.

the two products carry complementary signal. The history specialist scores 0.689 on D_H and 0.500 on D_L, and the liquidity specialist scores 0.500 on D_H and 0.640 on D_L.

Natural banking audit. The banking audit passes every pipeline gate but fails its primary scientific gate. At 256 records, logistic regression prefers the campaign attributes in 100/100 runs and boosting does so in 98/100, and only 2/100 paired runs reverse (Wilson 95% interval 0.6–7.0%). Both mean client-minus-campaign log-utility contrasts are negative, -0.056 for logistic regression (bootstrap 95% CI $[-0.060, -0.052]$) and -0.065 for boosting ($[-0.071, -0.058]$). Literal LAVA and conditional MMD nevertheless prefer the client product in 100/100 runs, so they agree with logistic regression in 0/100 runs and with boosting in 2/100. On this frozen product pair, their distributional ordering is confident but opposite to both standard deployments.

Continuous ridge-family calibration. We test Theorem 4 on a frozen study of 480 structured pairs over the continuous core family $\lambda \in [0.5, 2]$. No false certification occurs. With nine benchmarks, active selection of the benchmark values improves coverage on 480/480 pairs and reduces the median coverage radius by 15.3% relative to an equally spaced logarithmic grid (paired-bootstrap 95% interval for the median log ratio $[-0.171, -0.163]$). How useful the certificate is depends strongly on the regime. The registered two-sided rule certifies 93/120 well-conditioned monotone acquisitions, but only 7/120 ill-conditioned monotone pairs and 2/120 well-conditioned and 1/120 ill-conditioned redistribution pairs. The preregistered secondary sign-aware rule does not charge for drift that favors the preferred dataset. It raises these counts to 119/120, 106/120, 9/120, and 2/120, again with zero false certifications (Figure 2). The conservative family oracle is positive on 310/480 pairs. The sign-aware rule recovers 236/310 (76.1%) of these opportunities, compared with 103/310 (33.2%) for the registered two-sided rule.

Earlier natural source products. On frozen source-product pairs from the Bank (Moro et al., 2011), Heart (Janosi et al., 1989), and Gas (Vergara, 2012; Vergara et al., 2012) datasets, the rule produces no false certificates but also no positive ones. The registered gate, which required an observed-label reversal on two of the three datasets, fails (Appendix C).

Prospective natural deployment sensitivity. The final sealed study passes every pipeline and multiplicity gate. On HAR, all 200 paired randomized executions succeed for all three learners. Conditional on the fixed archive, training pools, and test cohort, histogram boosting (HGB) has

a negative odd-minus-even log-utility contrast (mean -0.0359, simultaneous 99.167% interval $[-0.0500, -0.0219]$). The MLP has a positive contrast (0.1576, $[0.0491, 0.2648]$), and logistic regression is unresolved (-0.00294, $[-0.0120, 0.00607]$). The registered output is therefore *sensitive*, because two ordinary learners prefer opposite products under this finite-archive randomization estimand. A post-freeze direct MLP-minus-HGB interaction check also excludes zero (mean 0.1935, simultaneous max-t interval $[0.0944, 0.2927]$), but it was not preregistered and does not replace the primary decision (Appendix C). On Covertype, every size-256 sample from Rawah and Comanche Peak misses at least one of the seven classes, so the frozen failure policy returns *uncovered* without repair. The Gaussian ridge-family arm produces zero false certificates but also zero positive certificates on both datasets. In this arm, 108/200 HAR pairs are oracle-sensitive and the rest are uncovered. Two independent verification runs reproduce all intervals, state decisions, and sentinel certificate quantities.

9 Discussion

From learner-free ranking to selective certification. A score can be learner-free to compute without being correct for every learner. This does not invalidate LAVA, SAVA, KAIROS, or other ranking proxies. Our certificate adds a layer that declares a response family, measures coverage and margin, and requires $\gamma > \epsilon M + \rho_-$ when responses drift with the data. It is informative in the controlled and ridge families but certified no sealed natural pair, leaving its natural-data usefulness open.

Limits of the practical case. When the learner family is finite and cheap to train, evaluating every learner directly is preferable. The ridge study is a calibration test, not an economic argument against sweeping the penalty. Reusable certificates may help for costly combinatorial families, but we do not establish a cost advantage for modern machine learning on natural data.

Natural interpretation and negative gates. Credit crossover rises with sample size because specialists that use different measurements exploit them more reliably as data grows. In Bank, the ordinary learners agree on the campaign fields, while two learner-free objectives unanimously choose the client fields. HAR instead returns *sensitive*. The banking reversal gate, the source-usefulness gate, and the two-of-three reversal gate all fail, the sealed certificate yield is zero, and Covertype is uncovered. These null results show that positive certification remains difficult. They support neither a universal ranking nor a claim that the certificate is non-vacuous on natural data.

10 Limitations

The theory is finite-dimensional, Bayesian linear–Gaussian, and pre-posterior, and it places no restriction on row norms. It does not value label surprise or corruption. Arbitrary deep training need not admit a trace response, and the drift result does not carry over to observed-label or deep learners. The controlled studies are synthetic and low-dimensional, and the real-outcome products use constructed groups of fields. The HAR intervals capture the randomness of acquisition and training on one archive and one test cohort, not newly sampled subjects or sites. Covertype lacks the registered class support, and the natural certificate yields are zero.

11 Conclusion

The value of data is generally not a property of the dataset alone. Under a fixed target, it is an order determined jointly by the information in the dataset and by what the learner can use. Finite benchmark agreement cannot certify an unrestricted deployment, but the limitation is quantitative rather than absolute. The distance to the benchmark cone gives an exact, attainable escape margin. Convex-hull coverage of the deployment family gives a positive certificate for safe use, and a sign-aware drift penalty extends it to linear learners whose response depends on the data. Minimax regret measures the cost of choosing without coverage. Learner-free scores remain useful, and exact rankings become certifiable once the deployment family is declared, covered, and reported. The HAR reversal shows that this scope matters beyond constructed specialists, and the null yields on natural data show that soundness alone does not yet give practical coverage.

AI USE STATEMENT

In this work, we used generative AI tools for the following tasks with required disclosure: feedback on theoretical models and claim scope, formulating mathematical claims, proof development and cross-checking, experimental design, implementing methods and artifact code, and interpreting results. Additionally, we used generative AI tools for the following tasks with recommended disclosure: literature discovery, figure and table design, and manuscript editing. We have reviewed all AI-assisted work. AI-suggested mathematical claims were checked by explicit derivations and numerical falsification tests. All confirmatory claims reported here are tied to frozen protocols and retained outputs, and AI-suggested empirical analyses were not retroactively promoted to preregistered outcomes. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

All data used in this work are synthetic or come from publicly available benchmark datasets. No new data were collected, and no analysis attempts to identify individuals. Excluding demographic fields does not establish fairness, and no banking model is proposed for deployment. Uncovered responses are not automatically security vulnerabilities, and we make no security claim.

REPRODUCIBILITY STATEMENT

The assumptions are stated in Sections 3–6, and complete proofs of all stated results are given in Appendix A. Appendix C gives the resampling, multiplicity, and Monte Carlo details. The anonymous supplementary material contains the frozen protocols, source and dataset hashes, environment records, run manifests, outputs, independent verification code, and claim-to-artifact checks used for every confirmatory number. Its README gives a fast verification path (file manifest, theorem arithmetic, and manuscript-claim audit) that needs no external downloads. Third-party UCI datasets are not redistributed, but their source URLs and SHA-256 hashes are recorded in the supplementary source audits. Failed scientific gates and zero-yield arms are retained. Reproducing the manuscript does not require reopening or redefining any frozen decision rule.

REFERENCES

J. A. Blackard. Covertype. UCI Machine Learning Repository, 1998. URL https://doi.org/10.24432/C50K5N.

D. Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24(2):265–272, 1953.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

I. S. Dhillon, R. W. Heath, Jr., T. Strohmer, and J. A. Tropp. Constructing packings in Grassmannian manifolds via alternating projection. *Experimental Mathematics*, 17(1):9–35, 2008.

H. Diehl and A. C. Wilson. Semivalue-based data valuation is arbitrary and gameable. *arXiv preprint arXiv:2506.12619*, 2025.

L. Engstrom, A. Feldmann, and A. Madry. DsDm: Model-aware dataset selection with datamodels. In *International Conference on Machine Learning (ICML)*, 2024.

K. Ethayarajh, Y. Choi, and S. Swayamdipta. Understanding dataset difficulty with $\mathcal{V}$-usable information. In *International Conference on Machine Learning (ICML)*, 2022.

H. Fawzi. On polyhedral approximations of the positive semidefinite cone. *Mathematics of Operations Research*, 46(4):1479–1489, 2021.

A. Ghorbani and J. Zou. Data Shapley: Equitable valuation of data for machine learning. In *International Conference on Machine Learning (ICML)*, 2019.

A. Janosi, W. Steinbrunn, M. Pfisterer, and R. Detrano. Heart disease. UCI Machine Learning Repository, 1989. URL https://doi.org/10.24432/C52P4X.

K. Jiang, W. Liang, J. Zou, and Y. Kwon. OpenDataVal: A unified benchmark for data valuation. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023.

H. A. Just, F. Kang, J. T. Wang, Y. Zeng, M. Ko, M. Jin, and R. Jia. LAVA: Data valuation without pre-specified learning algorithms. In *International Conference on Learning Representations (ICLR)*, 2023.

S. Kessler, T. Le, and V. Nguyen. SAVA: Scalable learning-agnostic data valuation. In *International Conference on Learning Representations (ICLR)*, 2025.

X. Lin, X. Xu, Z. Wu, S.-K. Ng, and B. K. H. Low. Distributionally robust data valuation. In *International Conference on Machine Learning (ICML)*, 2024.

K. Minchev, D. I. Dimitrov, and N. Konstantinov. LARP: Learner-agnostic robust data prefiltering. *Transactions on Machine Learning Research*, 2026.

M. Mohammadi Amiri, F. Berdoz, and R. Raskar. Fundamentals of task-agnostic data valuation. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2023.

S. Moro, R. Laureano, and P. Cortez. Using data mining for bank direct marketing: An application of the CRISP-DM methodology. In *European Simulation and Modelling Conference (ESM)*, 2011.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

F. Pukelsheim. *Optimal Design of Experiments*. SIAM, 2006.

J. L. Reyes-Ortiz, D. Anguita, A. Ghio, L. Oneto, and X. Parra. Human activity recognition using smartphones. UCI Machine Learning Repository, 2013. URL `https://doi.org/10.24432/C54S4K`.

A. Reznikov and E. B. Saff. The covering radius of randomly distributed points on a manifold. *International Mathematics Research Notices*, 2016(19):6065–6094, 2016.

M. Tamine, B. Heymann, M. Vono, and P. Loiseau. On the impact of the utility in semivalue-based data valuation. In *International Conference on Learning Representations (ICLR)*, 2026.

A. Vergara. Gas sensor array drift dataset. UCI Machine Learning Repository, 2012. URL `https://doi.org/10.24432/C5RP6W`.

A. Vergara, S. Vembu, T. Ayhan, M. A. Ryan, M. L. Homer, and R. Huerta. Chemical gas sensor drift compensation using classifier ensembles. *Sensors and Actuators B: Chemical*, 166–167:320–329, 2012.

J. T. Wang, T. Yang, J. Zou, Y. Kwon, and R. Jia. Rethinking data Shapley for data selection tasks: Misleads and merits. In *International Conference on Machine Learning (ICML)*, 2024.

E. B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

X. Xu, S. Wang, C.-S. Foo, B. K. H. Low, and G. Fanti. Data distribution valuation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Y. Xu, S. Zhao, J. Song, R. Stewart, and S. Ermon. A theory of usable information under computational constraints. In *International Conference on Learning Representations (ICLR)*, 2020.

L. Ye, A. Mitra, and S. Sundaram. Near-optimal data source selection for Bayesian learning. In *Learning for Dynamics and Control Conference (L4DC)*, 2021.

I.-C. Yeh and C.-H. Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2):2473–2480, 2009.

X. Zhou, Y. Zeng, R. Jin, and I. Lourentzou. Hierarchical dataset selection for high-quality data sharing. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2026.

J. Zhu, P. P. Prashant, A. Cloninger, and B. Salimi. KAIROS: Scalable model-agnostic data valuation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

A PROOFS

Proof of Proposition 1. Write $\theta = m_D + e_D$, where posterior error e_D is orthogonal in L_2 to every function of the observations, including m_D. Then

$$\theta - Pm_D = e_D + (I - P)m_D,$$

and the expected cross term vanishes. Total covariance gives $\mathrm{Var}(m_D) = I - \Sigma_D = K_D$, hence

$$\mathbb{E}\|\theta - Pm_D\|^2 = \mathrm{tr}(\Sigma_D) + \mathrm{tr}((I - P)K_D(I - P)).$$

Because P is an orthogonal projection, trace cyclicity gives $\mathrm{tr}((I-P)K_D(I-P)) = \mathrm{tr}((I-P)K_D)$. Subtracting from baseline risk d yields $U_P(D) = \mathrm{tr}(PK_D)$. Eigenvalues of K_D are $h/(1+h)$ for eigenvalues $h \geq 0$ of H_D, so they lie in $[0, 1)$. □

Proof of Proposition 2. Write $H = H_D$. Since $\widehat{\theta} = GH\theta + GX_D^\top \varepsilon$ and the two Gaussian terms are independent,

$$\begin{aligned} U_G(D) &= d - \mathrm{tr}((I - GH)(I - GH)^\top) - \mathrm{tr}(GHG^\top) \\ &= 2\,\mathrm{tr}(GH) - \mathrm{tr}(GH(H + I)G^\top). \end{aligned}$$

Let $K = H(I + H)^{-1}$ and $\Phi = G(I + H)$. Expanding $A_\Phi = \Phi + \Phi^\top - \Phi^\top\Phi$, using that H, $I + H$, and K commute, and applying trace cyclicity gives

$$\mathrm{tr}(A_\Phi K) = 2\,\mathrm{tr}(GH) - \mathrm{tr}(GH(H + I)G^\top).$$

Finally, $A_\Phi = I - (I - \Phi)^\top(I - \Phi) \succeq 0$ iff every singular value of $I - \Phi$ is at most one, equivalently $\|I - \Phi\|_{\mathrm{op}} \leq 1$. □

Proof of Proposition 3. Add and subtract $\langle A_\alpha(H_0), K_1\rangle$ in $\langle A_\alpha(H_1), K_1\rangle - \langle A_\alpha(H_0), K_0\rangle$. For ridge, $\Phi = (H + \lambda I)^{-1}(H + I)$, so $I - \Phi = (\lambda - 1)(H + \lambda I)^{-1}$. Substituting into $A_\Phi = I - (I - \Phi)^\top(I - \Phi)$ gives the stated formula. □

Proof of Theorem 4. For any α, choose $\bar{B}$ in the benchmark convex hull with $\|A_\alpha(H_0) - \bar{B}\|_F \leq \epsilon$. Convexity gives $\langle \bar{B}, \Delta_K\rangle \geq \gamma$, and Cauchy–Schwarz gives $\langle A_\alpha(H_0) - \bar{B}, \Delta_K\rangle \geq -\epsilon M$ and $\langle A_\alpha(H_1) - A_\alpha(H_0), K_1\rangle \geq -\rho_-$. Combining these inequalities with Proposition 3 proves the claim.

For the ridge bound, let $R_i = (H_i + \lambda I)^{-1}$. Since $A_\lambda(H) = I - (\lambda - 1)^2 R^2$,

$$\|A_\lambda(H_1) - A_\lambda(H_0)\|_F \leq (\lambda - 1)^2(\|R_1\|_{\mathrm{op}} + \|R_0\|_{\mathrm{op}})\|R_1 - R_0\|_F.$$

The resolvent identity and $H_i \succeq 0$ yield $\|R_i\|_{\mathrm{op}} = (\lambda + \mu_i)^{-1}$ and

$$\|R_1 - R_0\|_F \leq \frac{\|H_1 - H_0\|_F}{(\lambda + \mu_1)(\lambda + \mu_0)},$$

proving the displayed conditioning-aware bound. Finally, $I - \Phi = (\lambda - 1)(H + \lambda I)^{-1}$ has maximum operator norm over $H \succeq 0$ equal to $|\lambda - 1|/\lambda$. Proposition 2 therefore gives uniform response positivity iff $|\lambda - 1|/\lambda \leq 1$, equivalently $\lambda \geq 1/2$. □

Proof of Theorem 5. Necessity follows above. Conversely, for $0 \preceq K \prec I$, define $H = K(I-K)^{-1}$. Spectral calculus shows $H \succeq 0$. Choose $X = H^{1/2}$, so $X^\top X = H$. Then

$$I - (I + X^\top X)^{-1} = I - (I + K(I - K)^{-1})^{-1} = K.$$

□

Proof of Theorem 6. Let $\Delta = \Delta_+ - \Delta_-$ be its spectral positive/negative decomposition. For $0 < c < 1/\|\Delta\|_{\mathrm{op}}$, set $K_1 = c\Delta_+$ and $K_2 = c\Delta_-$. Both are PSD strict contractions, so both are attainable by Theorem 5. Their difference is $c\Delta$. □

Proof of Theorem 7. Universal weak preference is the intersection of halfspaces $\{\Delta : \mathrm{tr}(A_L\Delta) \geq 0\}$. Taking nonnegative response combinations and closure does not change this intersection, yielding $C^*_{\mathcal{F}}$. Reflexivity and transitivity follow from cone membership. If both a dataset difference and its negative lie in the dual, every family response pairs to zero. This is exactly equality under $\sim_{\mathcal{F}}$, which restores antisymmetry on equivalence classes. □

Proof of Theorem 8. A finitely generated cone is polyhedral. For $d \geq 2$, $\mathbb{S}^d_+$ is not polyhedral because it has infinitely many rank-one extreme rays. Thus the finite benchmark cone C_B is a strict closed subcone of $\mathbb{S}^d_+$. The finite-dimensional bipolar theorem gives the strict reverse inclusion $\mathbb{S}^d_+ \subsetneq C^*_B$. Choose $\Delta \in C^*_B \setminus \mathbb{S}^d_+$. Since Δ is not PSD, a unit vector u exists with $u^\top \Delta u < 0$, and the response $uu^\top$ reverses the benchmark order. Theorem 6 realizes a positive multiple of Δ using finite designs. □

Proof of Theorem 9. Let P_B be the Euclidean projection of A_* onto C_B and set $R = A_* - P_B$. Moreau decomposition gives $\langle R, B\rangle \leq 0$ for all $B \in C_B$, $\langle R, P_B\rangle = 0$, and $\|R\|_F = \delta_*$. Set $\Delta_* = -R/\|R\|_F$. The stated margins follow. For any accepted unit Δ and any $P \in C_B$,

$$\begin{aligned} -\langle A_*, \Delta\rangle &= \langle P - A_*, \Delta\rangle - \langle P, \Delta\rangle \\ &\leq \langle P - A_*, \Delta\rangle \leq \|P - A_*\|_F. \end{aligned}$$

Infimizing over P proves tightness. □

Proof of Corollary 10. Let $\widetilde{\Delta} = \Delta_* + (\delta_*/2)I$. Trace normalization gives $\mathrm{tr}(B_i\widetilde{\Delta}) \geq \delta_*/2$ and $\mathrm{tr}(A_*\widetilde{\Delta}) = -\delta_*/2$. Apply Theorem 6 with any admissible positive scaling. □

Proof of Theorem 11. For each A, choose B in the benchmark convex hull with $\|A - B\|_F \leq \epsilon$. Convexity gives $\mathrm{tr}(B\Delta) \geq \gamma$, while Frobenius Cauchy–Schwarz gives

$$\mathrm{tr}(A\Delta) = \mathrm{tr}(B\Delta) + \mathrm{tr}((A - B)\Delta) \geq \gamma - \epsilon M.$$

□

Proof of Theorem 12. For responses with $\mathrm{tr}(A\Delta) \geq 0$, regret occurs only when D_2 is selected and has supremum $(1 - p)\alpha$. For negative pairings, regret occurs only when D_1 is selected and has supremum $p\beta$. Their maximum is $R(p)$. Deterministic endpoints yield $\min(\alpha, \beta)$. When both widths are positive, equalizing the two affine terms gives the stated p_* and value. In the one-sided cases, the universally weakly optimal dataset is chosen. □

B Two-Axis Exact Witness

Let $X_1 = \sqrt{a}\, e_1^\top$, $X_2 = \sqrt{a}\, e_2^\top$, $a > 0$, and $\lambda = a/(1 + a)$. Then $K_1 = \lambda e_1 e_1^\top$ and $K_2 = \lambda e_2 e_2^\top$. Projection learners $P_1 = e_1 e_1^\top$ and $P_2 = e_2 e_2^\top$ have utilities

	D_1	D_2
P_1	λ	0
P_2	0	λ

under the same full squared-error target. Every deterministic blind choice has worst-case regret λ, while the optimal 50/50 randomization has regret $\lambda/2$.

C Additional Experimental Detail

Simultaneous uncertainty wrapper. If random one-sided bounds satisfy, jointly with probability at least $1 - \delta$,

$$\gamma \geq \underline{\gamma}, \qquad \epsilon \leq \overline{\epsilon}, \qquad M \leq \overline{M}, \qquad \rho_- \leq \overline{\rho},$$

then Theorem 4 implies that $\underline{\gamma} > \overline{\epsilon}\,\overline{M} + \overline{\rho}$ certifies the same uniform preference with probability at least $1 - \delta$. This wrapper is not a generic estimator for black-box responses. Each application must justify all four simultaneous bounds without leaking information from the confirmation outcomes.

Controlled exact symmetry. Binary validation and test samples contain signal in both coordinates. D_1 exposes only the signal in coordinate 1, and D_2 only the signal in coordinate 2. At equal signal, closure under swapping the coordinates forces all audited coordinate-invariant scores to tie, while the two restricted learners strictly reverse.

IID robustness. Learners use both features but receive opposite diagonal regularization, $(0.01, 1)$ and $(1, 0.01)$. An isotropic control uses $(0.01, 0.01)$. Samples are IID Gaussian. The 1,752/1,800 crossover count shows that the result is not an artifact of exact antithetic symmetry.

Earlier natural source-product audit. The Bank contact-mode, Heart collection-site, and Gas time-batch pairs were frozen with a shared representation before any label was inspected. Under controlled Gaussian labels, the drift-aware rule has zero false certifications and zero simultaneous-band violations over 1,500 certificate cells and 4,500 Monte Carlo comparisons, but zero positive certifications on every dataset. With observed labels, no pair establishes a significant reversal between ordinary learners. One learner is decisive on Bank, two agree on Heart, and all three agree on Gas. Both the usefulness gate and the two-of-three reversal gate therefore fail.

Conservative coverage witness. Each deployment response is a mixture of 40 primitive catalog responses plus a controlled novel rank-one component. For each benchmark prefix, the represented weights are retained, and all omitted and novel mass is redistributed over the included primitives. This gives an explicit feasible convex-hull witness, so its maximum residual is a conservative upper bound on the coverage radius.

Sealed natural diagnostic. The HAR and Covertype archives were absent when the protocol was frozen. Their source URLs, licenses, and SHA-256 hashes were recorded before parsing. The registered 200 paired seeds exceed the normal-approximation requirement of 171 seeds for two opposite standardized effects of magnitude 0.30 at per-contrast level $\alpha = 0.05/6$ and 90% marginal power. The label-blind pilot was reproduced byte for byte before the targets were loaded. The separately coded result verifier recomputed every sample's class-support decision and bootstrap interval, checked all 400 Gaussian state decisions, and independently reconstructed 56 sentinel certificate quantities to a maximum discrepancy of 7.11×10^{-15}. The Covertype class-support failure and both zero certificate yields are retained without changing the sample size, balancing classes, or adding datasets. The frozen intervals use 20,000 paired seed-cluster percentile-bootstrap resamples. The Bonferroni correction gives the simultaneous level $1 - 0.05/6 = 99.167\%$. Each two-sided tail then has probability $0.05/12$, whose nominal Monte Carlo standard error is 4.55×10^{-4} at 20,000 resamples.

Post-freeze HAR interaction robustness check. The exploratory analysis arranges the 200 complete HAR acquisition and training seeds into paired logistic, HGB, and MLP contrast vectors and defines $q = \Delta_{\mathrm{MLP}} - \Delta_{\mathrm{HGB}}$. It resamples whole seeds, recomputes centered studentized statistics, and takes the 95th percentile of the maximum absolute statistic over the three learner means and q. With 200,000 resamples and the fixed random seed 2026091801, the max-t critical value is 2.4134. The interaction mean is 0.1935 (simultaneous 95% interval $[0.0944, 0.2927]$). The nominal upper-tail Monte Carlo standard error is 4.87×10^{-4}. This analysis was specified after the registered E5 outcome was known. It is therefore robustness evidence only, not a confirmatory endpoint.