BEFORE LABELS: MINIMAX AUDITS AND EXACT INFERENCE FOR BINARY CLASSIFIER COMPARISONS

Anonymous authors
Research manuscript; not submitted

ABSTRACT

Can an audit be optimized before any evaluation labels are observed? For a frozen library of binary classifiers, we show that baseline-relative zero–one loss vectors have a label-independent outer product. Consequently, independent Bernoulli inclusion gives an exactly known joint Horvitz–Thompson covariance, although estimating the risk differences still requires sampled labels. We derive a convex allocation, Disagreement-Optimal Variance Auditing (DOVA), and an orthogonal-projection lower bound: it is minimax among nonadaptive audits that are unbiased for every binary label vector and obey an expected annotation budget. A second characterization separates known covariance from a known error law. For the independent-inclusion estimator, the entire centered law is label-independent for every label vector precisely when each informative item is sampled with probability one half or one. This yields finite-sample simultaneous regions through a standard Monte Carlo rank calibration; outside that regime, exact covariance alone does not justify Gaussian intervals. Across six public data sources, 18 trained classifier libraries, and 70 budget configurations, DOVA reduces maximum comparison variance by 27.75% on average relative to uniform sampling within the disagreement union. Half-sample regions are 39.67% narrower on average than the matched Bernstein regions. Fixed-size sampling and prediction-assisted estimators can outperform DOVA on particular label vectors, delimiting rather than contradicting its worst-case guarantee.

1 INTRODUCTION

A model update is often evaluated by asking whether it improves on an existing reference, rather than by estimating its accuracy in isolation. When several frozen candidates share the same evaluation pool, one annotation informs all of these comparisons. The acquisition problem is therefore intrinsically multivariate: an example useful for one candidate may be irrelevant to another, while a shared disagreement can inform several contrasts simultaneously. Label-efficient evaluation already exploits selective annotation and predictive surrogates (Nguyen et al., 2018; Kossen et al., 2021; 2022). We study a narrower question: which aspects of the comparison experiment can be determined exactly from the prediction matrix, before revealing evaluation labels?

This distinction matters for planning an audit. A predicted label distribution can suggest where uncertainty is concentrated, but its quality on the evaluation pool is itself uncertain. An allocation derived from such a surrogate generally optimizes a modeled rather than an observed variance. Pilot labels may improve that model, but consume budget and create an additional adaptive inference problem. In contrast, an observable design objective can be optimized without either a pilot response or an assumption about predictive calibration. The desired property is not that labels become unnecessary: the sign of an improvement remains unknown. It is that the precision of a prescribed estimation procedure can be planned separately from its unknown answer.

Binary zero–one comparisons have a simple but useful structure. When a candidate and the reference agree, their losses agree for either label. When they disagree, their loss difference has magnitude one; the unknown label changes only its sign. More strongly, for a fixed item the entire vector of comparisons changes by one common sign. Its outer product is thus already known. Independent inclusion removes cross-item covariance terms, leaving an exact covariance matrix composed of these observable outer products. This joint statement is more informative than observing that disagreement

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.

identifies useful examples: it specifies the dependence between all reference-relative estimates and the exact effect of changing each inclusion probability.

Selecting discordant examples is not a new evaluation principle (Musgrove et al., 2023). Neither inverse-probability estimation nor capped square-root allocation is new (Horvitz & Thompson, 1952; Yilmaz et al., 2021). Our first contribution connects the binary comparison identity to a precise decision problem. We minimize the largest coordinate variance under an expected annotation budget, using a convex design whose numerical optimality is checked by a primal–dual gap. An orthogonal-projection argument then shows that this is not only an optimum within a chosen family of linear estimators. It attains the worst-label minimax risk among all nonadaptive, design-unbiased audits, including nonlinear estimators and dependent subset designs. The role of independent sampling is constructive: it attains a risk that is constant over the unknown label vectors.

The qualification "design-unbiased" is essential. The lower bound requires correctness in expectation for every possible binary labeling of the frozen pool, not only for labels drawn from a particular predictive model. This is the appropriate adversarial formulation when no label model is assumed, but it is not a universal dominance statement. A reliable surrogate can reduce the variance on the actual labeling. A fixed-size design can also be more efficient on a particular finite population. Both effects appear in our experiments. They do not invalidate a worst-case result, just as the minimax result does not justify ignoring useful prior information. We retain these comparators to expose this distinction empirically.

Exact covariance raises a second question: can it be turned into exact confidence regions without conservative concentration inequalities? In general the answer is no. A Bernoulli inclusion term can be strongly skewed, and its skewness depends on the unknown label sign even when its variance does not. A one-item example has exact 80% coverage for its nominal 95% Gaussian interval despite supplying the Gaussian procedure with the true variance. This separates two notions that are easily conflated: an exactly calculated covariance and a correctly calibrated sampling distribution. A simulation from a fitted Gaussian is not an exact audit simulation merely because its covariance matrix is correct.

Our second contribution identifies a sharp exception. Sampling every informative item independently with probability one half makes each centered contribution a symmetric sign flip. Multiplication by the unknown label sign then leaves its distribution unchanged. Census items contribute no random error. Conversely, outside these two inclusion probabilities, changing a single informative label changes a third centered moment. The half-or-census characterization gives the complete label-independent error law and permits simultaneous regions calibrated by a finite Monte Carlo order statistic. The rank argument is standard; what is special here is that the random variable being calibrated is exactly observable before labels, without assuming a null value for the unknown risk differences.

We evaluate the resulting trade-off with actual fitted classifiers and explicitly identified synthetic populations. The empirical question is not whether DOVA beats uniform sampling over irrelevant agreement examples, which would be an undemanding baseline. It is whether optimizing joint disagreement structure improves over uniform sampling on the disagreement union, and how it compares with fixed-size and prediction-assisted alternatives. A separate half-sampling experiment measures the benefit of using the exact joint law instead of coordinatewise Bernstein bounds. All acquisition functions are frozen before audit labels are accessed; complete labels are used only to measure retrospective errors and comparator variances. The paper first situates these claims, then develops the design and inference results, reports the experiments and negative comparisons, and states the limitations of the finite-pool interpretation.

2 Related work

Selection-induced estimation bias is also studied in active learning (Farquhar et al., 2021).

Selective evaluation. Accuracy and F-measure estimation, active testing, and learned surrogate evaluation precede this work (Nguyen et al., 2018; Sawade et al., 2010; Kossen et al., 2021; Ha et al., 2021; Kossen et al., 2022; Marchant & Rubinstein, 2021). The closest sampling-design precedent, Yilmaz et al. (2021), already uses independent inclusions and capped optimal allocations.

Discordant-pair evaluation also precedes us (Musgrove et al., 2023), although that formulation uses known baseline performance and prevalence. Response-diversity condensation (Rubinstein et al., 2025) instead uses deterministic selection and performance prediction. We claim neither disagreement sampling nor capped square-root allocation as new. Our distinction is an observable joint covariance for binary paired contrasts, its worst-coordinate unbiased minimax characterization, and the half-sampling exact-law boundary.

Multi-model and large-language-model evaluation are established tasks (Shanmugam et al., 2025; Berrada et al., 2025). Their broader losses do not automatically satisfy our binary identity, and we do not reproduce those large-scale systems. Our finite-pool theory must be checked against the actual loss before transfer.

Prediction assistance. Survey calibration provides classical auxiliary-information estimators (Deville & S"arndal, 1992). Prediction-powered inference, PPI++, cross-prediction, and active statistical inference develop predictive correction and label acquisition (Angelopoulos et al., 2023a;b; Zrnic & Candès, 2024b;a). Further work considers Bayesian corrections, few-label regression, anytime inference, multiple predictions, and multiple tasks (Cortinovis & Caron, 2025; Eyre & Madras, 2025; Kilian et al., 2025; Brawand et al., 2026; Emmenegger et al., 2026). Our fixed-surrogate comparators are representatives, not complete reproductions of these methods. We therefore make no empirical superiority claim over their full implementations. The minimax proof permits surrogate-assisted unbiased estimators but varies all labels with the surrogate fixed, following a classical finite-population decision-theoretic perspective (Godambe & Joshi, 1965).

Inference. Estimated-variance concentration, sampling-without-replacement bounds, and sequential confidence procedures offer complementary guarantees (Maurer & Pontil, 2009; Serfling, 1974; Bardenet & Maillard, 2015; Waudby-Smith & Ramdas, 2020; Howard et al., 2021; Waudby-Smith & Ramdas, 2024). Our elementary Bernstein interval is not claimed optimal. Half-sampling calibration uses established randomization-rank principles (Hemerik & Goeman, 2018); the new question is when the centered audit law can be calibrated without labels. It does not authorize optional stopping or repeated adaptive reuse. Extended comparisons and scope qualifications are retained in the accompanying full manuscript.

3 Method: observable design and its inferential boundary

3.1 The frozen finite-pool experiment

There are N items and $M+1$ fixed binary prediction vectors, with reference f_0 and candidates $f_1, \dots, f_M$. All features, predictions, and auxiliary scores are treated as fixed before auditing. Labels $y_i \in \{0,1\}$ are unknown fixed values; no independence or calibration model for these labels is assumed. With zero–one loss, define

$$a_{ji} = f_{ji} - f_{0i}, \qquad s_i = 1 - 2y_i, \qquad \Delta_j = \frac{1}{N}\sum_i D_{ji},$$
$$D_{ji} = \mathbf{1}\{f_{ji} \neq y_i\} - \mathbf{1}\{f_{0i} \neq y_i\} = a_{ji}s_i. \tag{1}$$

Negative Δ_j denotes improvement. Let $U = \{i : a_i \neq 0\}$, where a_i is the M-vector of prediction differences. The population denominator remains N, not $|U|$. Agreement items need no labels for these contrasts, even though they could be essential for estimating absolute accuracy.

A nonadaptive design selects a random subset $R \in \{0,1\}^N$ independently of the unknown labels, with marginal inclusion probabilities π_i and $\sum_i \pi_i \leq B$. This is an *expected*, not a hard, annotation budget. A design-unbiased estimator T is a measurable function of the selected indices, selected labels, and fixed auxiliary information such that $\mathbb{E}_R T = \Delta$ for every label vector. Any informative item must have $\pi_i > 0$. DOVA uses independent $R_i \sim \mathrm{Bernoulli}(\pi_i)$, conventionally called Poisson sampling, and the classical estimator

$$\widehat{\Delta} = \frac{1}{N}\sum_{i \in U} \frac{R_i D_i}{\pi_i}. \tag{2}$$

The independence concerns inclusion decisions, not a data-generating assumption about labels.

3.2 EXACT COVARIANCE AND A MINIMAX AUDIT

Theorem 1 (Label-free covariance and design-unbiased minimax risk)**.** *Under independent inclusion, for every binary label vector,*

$$\mathrm{Cov}_R(\widehat{\Delta}) = \Sigma(\pi) := \frac{1}{N^2}\sum_{i\in U}\left(\frac{1}{\pi_i} - 1\right) a_i a_i^\top. \tag{3}$$

Let $\mathcal{A}_B$ contain all label-independent subset designs with expected size at most B, paired with finite-second-moment estimators that are design-unbiased for every labeling. For $0 < B < |U|$,

$$\inf_{(P,T)\in\mathcal{A}_B} \sup_{s\in\{-1,1\}^N} \max_{j\le M} \mathbb{E}_P[(T_j - \Delta_j(s))^2] = \min_{\substack{0<\pi_i\le 1\ (i\in U)\\ \sum_{i\in U}\pi_i\le B}} \max_{j\le M} \Sigma_{jj}(\pi). \tag{4}$$

An independent-inclusion HT audit at a minimizing design attains this value. The lower bound allows dependent designs and nonlinear unbiased estimators.

The covariance identity follows because $D_i D_i^\top = a_i a_i^\top$ and independent sampling removes cross-item terms. For the lower bound, average over auxiliary independent uniform signs S_i, used only as a proof device. For any label-independent design, the observed variables $\phi_i = R_i S_i/\sqrt{\pi_i}$ are orthonormal in L^2. Design-unbiasedness and missing-label measurability give $\mathbb{E}[T_j\phi_i] = a_{ji}/(N\sqrt{\pi_i})$. Projection implies $\mathbb{E}T_j^2 \ge \sum_i a_{ji}^2/(N^2\pi_i)$. Subtracting $\mathbb{E}\Delta_j^2$ lower-bounds the average squared error by $\Sigma_{jj}(\pi)$; a worst-label error is at least this average. Independent HT attains the bound for every labeling. Appendix A supplies all conditioning steps and scope qualifications.

The theorem concerns the maximum *coordinatewise expected squared error*, not the expected maximum error. It allows a fixed, informative surrogate but evaluates the estimator over all labels while holding that surrogate fixed. It excludes label-adaptive acquisition and estimators biased for some labelings. A procedure requiring a hard sample cap is not covered by the attainable expected-budget statement. These distinctions explain why fixed-size or prediction-assisted methods can improve pointwise while remaining consistent with the theorem.

Convex allocation. Write $c_{ji} = a_{ji}^2/N^2$ and $v_j(\pi) = \sum_i c_{ji}(1/\pi_i - 1)$. Minimizing $\max_j v_j$ is convex. Its simplex dual uses $\lambda \in \mathbb{R}_+^M$, $\sum_j \lambda_j = 1$, with $w_i = \sum_j \lambda_j c_{ji}$. For an interior λ, the inner optimum is

$$\pi_i(\lambda) = \min\{1, \sqrt{w_i/\nu}\}, \qquad g(\lambda) = \sum_i w_i(1/\pi_i(\lambda) - 1), \tag{5}$$

where $\nu > 0$ enforces the budget. Zero-weight boundary cases are defined by limits. The derivative of g is $v(\pi(\lambda))$ wherever differentiable. For any feasible returned design and dual weights, $P = \max_j v_j$ is an upper bound and g is a lower bound on the optimum. Their relative gap $(P - g)/P$ therefore supplies a genuine stopping rule.

DOVA computes a, discards agreement items, maximizes g numerically, freezes π, queries the selected labels, and returns (2). The implementation uses SLSQP with analytic derivatives and capped water filling; it accepts a solution only when the relative gap is at most 10^{-5}. Bounds $\lambda_j \ge 10^{-10}$ avoid undefined boundary gradients without invalidating the dual lower bound. Water filling stops on the actual budget residual, and the optimizer has both iteration and time guards. One function/gradient evaluation costs $O(M|U| + |U|\log(1/\epsilon))$ for bisection accuracy ϵ; storing predictions and the coefficient matrix costs $O(MN)$. We make no worst-case iteration-complexity claim for SLSQP. For $B \ge |U|$ a census gives zero error, while $U = \varnothing$ makes every contrast exactly zero.

3.3 CONFIDENCE REGIONS: KNOWN VARIANCE IS NOT A KNOWN LAW

For independent inclusion, set $V_j = \Sigma_{jj}(\pi)$ and let $b_j = N^{-1}\max_{i:0<\pi_i<1}|a_{ji}|\max\{1, (1-\pi_i)/\pi_i\}$, with an empty maximum equal to zero. Applying Bernstein to the centered independent summands and a two-sided union bound gives simultaneous coverage at least $1-\delta$ for $[\widehat{\Delta}_j - r_j, \widehat{\Delta}_j + r_j]$, where

$$L = \log(2M/\delta), \qquad r_j = b_j L/3 + \sqrt{2V_j L + (b_j L/3)^2}. \tag{6}$$

All radii are determined before labels. However, minimizing the variance term alone need not minimize these radii, because the range term also depends on the smallest inclusions. Nor does it justify Gaussian calibration: one nonzero scalar contrast sampled with $\pi = 0.2$ has centered support $\{-1, 4\}$, probabilities $\{0.8, 0.2\}$, and variance 4. Its true-variance Gaussian 95% interval covers only the first atom and therefore has exactly 80% coverage.

Theorem 2 (Half-or-census characterization)**.** *For the independent-inclusion HT estimator, the joint law of $\widehat{\Delta} - \Delta$ is independent of the binary label vector if and only if $\pi_i \in \{1/2, 1\}$ for every $i \in U$. In that case it equals the law of $N^{-1} \sum_{i:\pi_i=1/2} \xi_i a_i$, where the ξ_i are independent uniform signs.*

Symmetry proves sufficiency. For necessity, choose an informative coordinate of item i. Its third centered moment is proportional to $s_i(1 - \pi_i)(1 - 2\pi_i)/\pi_i^2$; flipping only s_i changes the total third moment unless π_i is one half or one. Thus knowing the exact covariance is strictly weaker than knowing the centered law.

For the implemented half-union design, let $\sigma_j^2 = N^{-2} \sum_i a_{ji}^2$ and omit coordinates with $\sigma_j = 0$. Simulate K independent copies of $Z = \max_j |N^{-1} \sum_i \xi_i a_{ji}|/\sigma_j$ before collecting labels. Use order statistic $Z_{(k)}$ with $k = \lceil (1 - \delta)(K + 1) \rceil$ and radii $r_j = \sigma_j Z_{(k)}$, taking infinity if $k > K$. Exchangeability of the actual and simulated statistics gives simultaneous coverage at least $1 - \delta$, including calibration randomness; ties are conservative. Conditional on one stored calibration sample, coverage need not equal 95%. Half sampling is a separate design, not a cost-free confidence upgrade for arbitrary optimized inclusions.

Boundary beyond binary losses. For general fixed contrast vectors $D_i(y)$, independent-design covariance is label-free for every design precisely when $D_i(y)D_i(y)^\top$ is label-free for each item. Equivalently, each item's possible contrast vectors must lie in $\{v_i, -v_i\}$, or be identically zero. The condition is necessary by taking all other items at census. Multiclass zero–one loss already fails: predictions 0 and 1 yield squared differences $1, 1, 0$ for labels $0, 1, 2$. This characterization prevents silently extending the method to losses that do not possess its algebraic structure.

What is fixed before annotation. The reference, all candidate predictions, the population weights, budget, and confidence level are fixed inputs to the experiment. The design may exploit arbitrary feature-derived scores without using evaluation responses. After probabilities are frozen, a Bernoulli draw decides which labels are acquired, and a queried label is reused across every comparison. The realized count has mean $\sum_i \pi_i$ and variance $\sum_i \pi_i(1 - \pi_i)$; neither identity turns an expected budget into a hard cap. Stopping acquisition when the count reaches B would change the design and invalidate the independent-inclusion covariance and error-law arguments. Likewise, conditioning the resulting interval on the realized count requires a separate analysis. This is why the fixed-size comparator is implemented as its own estimator and confidence procedure rather than by truncating DOVA's sample.

4 Experimental design

Public data and splits. Six public packaged data sources provide computational test beds, not large-scale annotation benchmarks (Scikit-learn developers, 2026; Statsmodels developers, 2026a;b; Horst et al., 2020). We retain the breast-cancer binary response and derive digits parity, wine cultivar zero versus others, positive versus zero Fair affairs, any versus no RAND physician visit, and Adelie versus other penguins. RAND is a label-independent seeded subset of 10,000 of 20,190 records; penguins retain rows complete on four numerical measurements. Response columns are excluded from features. No clinical, social, or ecological interpretation of fitted predictions is intended. Seeds $0, 1, 2$ produce stratified 45% training, 15% validation, and 40% audit splits, up to rounding. Offline benchmark stratification uses source labels; it is not presented as an unlabeled production step. Model selection uses only training and validation data. The acquisition function accepts predictions, probabilities, a frozen proxy, and a budget, but not audit labels. Complete audit labels are accessed after proposals are constructed to simulate annotation and measure error. Arrays and split indices are saved.

Models and tuning. Each pool contains logistic regression, random forest, extra trees, histogram gradient boosting, nearest neighbors, Gaussian naive Bayes, a decision tree, and linear discriminant

Table 1: Actual processed datasets. N_0 precedes splitting; d excludes the response. These are six sources, not six large-scale audit benchmarks.

Task	N_0	d	Positive labels
Breast cancer	569	30	357
Digits parity	1,797	64	906
Wine binary	178	13	59
Fair binary	6,366	8	2,053
RAND binary	10,000	9	6,873
Penguins binary	342	4	151

analysis. Three configurations per family are selected by validation error, breaking ties with validation log loss. Logistic regression is the reference, giving seven comparisons. The proxy uses the best validation-log-loss model among all 24 fits and shrinkage $\eta_i = 0.9p_i + 0.05$. Across 18 pools, 432 models are actually fitted. Applicable standardization is learned from training data. Appendix B specifies grids, stopping rules, and tree sizes.

Eight valid methods and an invalid ablation. Uniform-all independently samples the entire pool; uniform-union samples only U. Mean-disagreement minimizes average rather than maximum variance. Entropy uses the ensemble mean probability, floor 0.05, and exponent selected from $\{0.5, 1, 2\}$ by the exact pre-label maximum variance. DOVA optimizes that same maximum directly.

Uniform-CV and proxy-CV use $\mu_{ji} = a_{ji}(1 - 2\eta_i)$ and estimator $N^{-1}\sum_i \mu_i + N^{-1}\sum_i R_i(D_i - \mu_i)/\pi_i$. The former samples uniformly on U; the latter applies the same minimax solver to surrogate coefficients $4\eta_i(1 - \eta_i)c_{ji}$. Both remain unbiased for a wrong proxy. Their Bernstein intervals use the valid residual envelope $2|a_{ji}| \max(\eta_i, 1 - \eta_i)$, not the retrospective residual variance. These fixed-surrogate representatives are not full PPI++ or active-surrogate implementations. Union-SRS samples exactly B distinct union items and uses the expansion estimator with a conservative Serfling interval. An ablation removes DOVA's inverse-inclusion weights; it is explicitly invalid and excluded from valid-method rankings and coverage guarantees.

Budgets and evaluation. Fractions $\{0.1, 0.25, 0.5, 0.8\}$ of $|U|$ are rounded to positive integer budgets and deduplicated, producing 70 real-data configurations. Each method receives 1,000 annotation repetitions; eight valid methods therefore yield 560,000 replicate estimates. Poisson methods share random numbers within a configuration; SRS uses uniform subsets. These repetitions are finite-pool Monte Carlo draws, not additional independent datasets or human labels. The primary metric is exact maximum coordinate variance, evaluated retrospectively for all methods. Empirical MSE, bias, coverage, and label-count distributions are also saved. Ratios are computed within configurations before averaging.

Fifteen synthetic configurations vary five geometries at three budgets; ten more vary proxy reversal at two budgets. Each uses 2,000 annotation draws per method. Half-sampling uses 9,999 calibration draws and 1,000 audit draws for each real pool. Exhaustive tests verify covariance, the half law, and small nonlinear-estimator lower-bound programs. NumPy, SciPy, and scikit-learn are used with one BLAS thread (Harris et al., 2020; Virtanen et al., 2020; Pedregosa et al., 2011). The recorded expanded main run took 26.18 CPU-environment seconds, excluding the separate proxy stress run, research, and writing; it is not a runtime promise for another machine.

5 Results

Disagreement structure improves the intended objective. Across the 70 real-data configurations, DOVA's maximum variance is smaller than uniform-union's in all 70. The mean within-configuration ratio is 0.7225, corresponding to a 27.75% reduction; the median ratio is 0.7442. Ratios range from 0.4652 to 0.8823. This gain is not solely the removal of agreement items, since the denominator already excludes them. Table 2 reports source-level averages, while Figure 1 shows budget dependence. The maximum returned relative primal–dual gap for DOVA and proxy-CV is 8.503×10^{-6}, and the

Table 2: Mean within-configuration ratios of maximum variance to uniform-union (= 1). Lower is better. All six source-level rows are shown; optimized proxy correction and fixed-size sampling sometimes beat DOVA.

Task	DOVA	Mean	Entropy	Fixed size	Unif.-CV	Proxy-CV
Breast cancer	0.696	0.993	0.929	0.887	0.789	0.561
Digits parity	0.691	1.223	1.063	0.518	0.117	0.223
Fair binary	0.737	1.073	0.977	0.997	0.973	0.717
Penguins binary	0.782	1.113	1.067	0.694	1.498	1.792
RAND binary	0.727	1.194	1.004	0.958	0.867	0.698
Wine binary	0.706	1.059	1.014	0.587	0.817	0.490

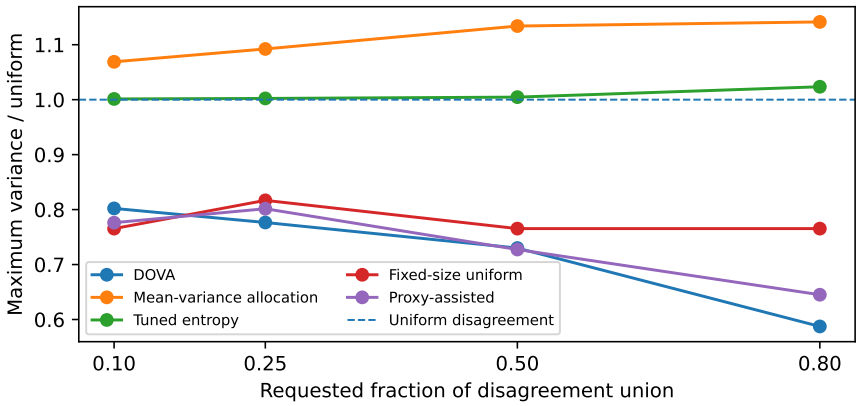


Figure 1: Maximum variance relative to uniform-union, averaged over the real-data configurations at each requested budget fraction. DOVA improves its label-free worst-coordinate objective; a proxy-assisted estimator or fixed-size design may be better for the actual labels. The latter comparisons do not contradict a worst-label minimax statement. The small-pool integer budgets are recorded in the raw results.

median DOVA solve time is 0.0169 seconds. The numerical objective is therefore optimized rather than inferred from a qualitative acquisition heuristic.

Stronger comparators prevent a dominance claim. DOVA improves on fixed-size uniform sampling in 46 of 70 configurations, but the mean ratio to that comparator is 1.0499: DOVA is approximately 4.99% worse under this equal-configuration average. A win count and an average relative loss answer different questions, so we report both. Proxy-CV has lower variance than DOVA in 54 of 70 configurations; uniform-CV does so in 22. The strong proxy effect on digits is especially visible in Table 2. Conversely, proxy-CV's mean ratio to uniform-union exceeds one for the small penguins task. A label-agnostic worst-case guarantee is thus useful precisely because pointwise proxy quality cannot be guaranteed from the prediction matrix alone; it is not a reason to discard a genuinely accurate proxy.

Ablations identify what each component contributes. Replacing the maximum-variance objective with a mean-variance allocation is not harmless. The source-average mean-disagreement ratio reaches 1.223 on digits and 1.194 on RAND, whereas DOVA's corresponding ratios are below one. The objective determines which comparison is protected, not just how many labels are allocated to disagreements. Removing the disagreement restriction wastes budget and can greatly inflate variance when candidate predictions are similar. Removing the proxy correction or changing its allocation separates estimator quality from design quality. Finally, the unweighted ablation is reported as a biased procedure, not as a competing valid audit: a small empirical squared error from selective unweighted averaging would not establish correctness for all labelings. Its per-configuration biases and errors remain in the artifact rather than being selectively omitted.

Exact half-sampling regions are narrower than the matched bound. For the identical half-union estimator and annotation design, the exact-law regions reduce mean coordinate radius by 39.67% on average across the 18 pools relative to Bernstein. Figure 2 includes every pool. Observed simultaneous coverage ranges from 0.943 to 1.000, with mean 0.9681; these are finite Monte Carlo frequencies, not asserted exact conditional probabilities. The theorem includes the independent calibration randomness, whereas each pool's calibration draw is held fixed across its 1,000 audit repetitions. This distinction and sampling variation explain why an individual frequency below 0.95 is not by itself a counterexample. DOVA's Bernstein coverage is at least 0.995 in every real configuration, illustrating conservatism rather than a need for approximate Gaussian replacement.

Synthetic and numerical verification. The explicit proxy-reversal experiment holds the labels and disagreement geometry fixed while changing only the proxy. DOVA's exact variance is invariant; the proxy estimators' variance changes with misspecification (Appendix C). This is a designed sensitivity analysis, not observed real-data distribution shift. Exhaustive covariance calculations agree to 5.56×10^{-17} in the stored five-item test. An independent primal solve checks the dual implementation on 12 small coefficient matrices. Finite least-norm programs over all observed-data functions also match the unbiased lower-bound calculation for independent, fixed-size, and coupled subset designs. These checks support implementation consistency; they do not replace the symbolic proofs or constitute independent peer review.

Empirical maximum coordinate MSE has a mean ratio of 0.7744 to uniform-union, distinct from the exact variance ratio of 0.7225. Finite repetitions and taking a maximum over noisy coordinate estimates affect this diagnostic. We report both quantities without replacing the exact objective by its favorable Monte Carlo realization. No dataset-level significance test is inferred from the many annotation repetitions.

6 Discussion

The main distinction is between uncertainty about the answer and uncertainty about an estimator's precision. Binary paired contrasts leave the first uncertainty intact while removing the second for independent HT sampling. This supports auditable acquisition decisions: every inclusion probability has a mathematically specified effect on each comparison variance before any audit response is observed. It also gives an unusually clean benchmark for acquisition heuristics. Because the objective is directly computable, the baseline cannot be excused or disadvantaged by a poorly estimated variance model; the optimizer's claimed improvement can be checked independently of annotation simulations.

The worst-label perspective is deliberately conservative. An accurate proxy encodes information that the adversarial class refuses to assume, and the residual estimators profit from that information in many configurations. This does not make the theorem vacuous: it identifies the price of demanding unbiasedness for every labeling without trusting a predictive model. At the same time, the fixed-size results show that even an elementary alternative can have better pointwise variance. A practitioner choosing an audit should therefore specify whether the primary requirement is worst-case protection, a hard annotation cap, or efficiency under a credible proxy, rather than treating these requirements as interchangeable.

The half-sampling result exposes a separate design objective. An optimized unequal-inclusion audit may achieve a smaller variance but a more skewed centered law. A symmetric half-sampling audit sacrifices allocation flexibility in exchange for exact distributional calibration from predictions alone. It is not generally meaningful to compare the two only through a single interval-width number without matching annotation budgets and sampling laws. Our width experiment compares two confidence procedures for the same half-sampling design; it does not assert that half sampling is globally optimal for width or cheaper than DOVA. Mixed half-and-census designs are covered by the theorem but are not optimized in the experiments.

The framework is also useful for identifying invalid extensions. Reusing the covariance formula after adding a newly trained model, changing the loss to a continuous score, or selecting labels adaptively would change assumptions that the proof uses. A simultaneous box allows selection among the frozen comparisons, but not the retroactive invention of new ones. Likewise, the exact error law concerns the fixed evaluation pool. It is not a statement about training variability or performance under a

shifted deployment distribution. These boundaries suggest concrete research directions: biased minimax estimators, label-adaptive designs with valid sequential inference, and expected-budget versus hard-budget optimality gaps. None is claimed to be solved by the present implementation.

7 LIMITATIONS

The empirical study uses six accessible data sources, including very small wine and penguins tasks, and derived binary responses rather than a large, externally curated audit benchmark. Three training splits per source do not support a broad statistical claim about all learning problems. Annotation repetitions estimate conditional finite-pool behavior and must not be counted as additional independent datasets. No human annotation costs, annotation noise, or operational deployment outcomes were measured. The modest classifier libraries do not demonstrate scaling to foundation models.

The principal optimality result requires design-unbiasedness under every binary labeling and a label-independent subset design with an expected budget. It says nothing about minimax biased estimators or label-adaptive acquisition. Its independent-inclusion implementation can exceed the expected label count on a particular run. Label-free covariance fails for generic multiclass or continuous losses, and finite-pool confidence does not establish population or distribution-shift coverage. The Bernstein bound is conservative and is not compared with every modern interval construction. The exact-law calibration applies only to half-or-census inclusions and includes calibration randomness.

Our prediction-assisted comparators implement a fixed-surrogate correction, not all the mechanisms of active testing, ASE, PPI++, or recent multiple-prediction methods. Consequently the experiments cannot establish a state-of-the-art empirical ranking against those methods. The literature survey is curated rather than exhaustive, and the theoretical novelty assessment has not received independent expert verification. The artifact includes executed code and internal consistency checks, but no independent external review or conference acceptance is represented as completed.

8 CONCLUSION

For frozen binary classifier comparisons, the prediction matrix determines more than a useful region to label. It determines every per-item contrast outer product, and hence the exact joint covariance of an independent-inclusion HT audit. We use that structure to construct DOVA, an implementable convex design with a checked primal–dual gap, and to establish its worst-label optimality among nonadaptive design-unbiased expected-budget audits. The result specifies an exact decision criterion without asserting pointwise superiority over methods that exploit reliable predictive information.

A second characterization distinguishes known covariance from a label-independent centered distribution. Half-or-census inclusion is precisely the regime where the entire HT error law is observable from the frozen predictions. Standard finite Monte Carlo rank calibration then yields simultaneous regions without a Gaussian approximation. The explicit skewed counterexample explains why the covariance identity alone is insufficient outside that regime.

Executed experiments support variance improvements over a strong disagreement-uniform baseline and narrower half-sampling regions than a matched Bernstein bound, while retaining cases where fixed-size and proxy-assisted methods are better. The accompanying data, allocation records, replicate errors, tests, proofs, and rendering sources make those claims inspectable. Extending the analysis to hard sample caps, biased estimation, or label-adaptive designs requires new results rather than a change of terminology; those are the natural next questions left open by this study.

AI USE DISCLOSURE

This research artifact was developed with a generative AI assistant. AI assistance covered topic selection and literature retrieval; theoretical formulation, claims and proofs; synthetic experiment specification; software implementation and debugging; data transformation; execution and analysis of experiments; figure and table code; and drafting and revision. Numerical results were produced by executing the supplied Python programs, not by language-model simulation. Reviews in the artifact are internal self-critiques, not independent human or external-model reviews. No external model API experiment or conference submission was performed. Independent human verification and authorization by the eventual authors have not been represented as completed. The eventual submitting authors must verify the work, accurately report this assistance in the submission form, and take responsibility for the final submission.

REFERENCES

Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-Powered Inference. *Science*, 382(6671):669–674, 2023a. doi: 10.1126/science.adi6000. URL https://arxiv.org/abs/2301.09633.

Anastasios N. Angelopoulos, John C. Duchi, and Tijana Zrnic. PPI++: Efficient Prediction-Powered Inference, 2023b. URL https://arxiv.org/abs/2311.01453. arXiv:2311.01453; preprint.

R'emi Bardenet and Odalric-Ambrym Maillard. Concentration inequalities for sampling without replacement. *Bernoulli*, 21(3):1361–1385, 2015. doi: 10.3150/14-BEJ605. URL https://arxiv.org/abs/1309.4029.

Gabrielle Berrada, Jannik Kossen, Muhammed Razzak, Freddie Bickford Smith, Yarin Gal, and Tom Rainforth. Scaling Up Active Testing to Large Language Models, 2025. URL https://arxiv.org/abs/2508.09093. arXiv:2508.09093; preprint.

Nicholas Brawand, Nima Leclerc, Anhthy Ngo, Matthew Peterson, Sriram Vishwanath, Laith Alhussein, and Ben Wellner. Active Multiple-Prediction-Powered Inference, 2026. URL https://arxiv.org/abs/2605.08429. arXiv:2605.08429; preprint.

Stefano Cortinovis and Francois Caron. FAB-PPI: Frequentist, Assisted by Bayes, Prediction-Powered Inference. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 11328–11356, 2025. URL https://proceedings.mlr.press/v267/cortinovis25a.html.

Jean-Claude Deville and Carl-Erik S"arndal. Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association*, 87(418):376–382, 1992. doi: 10.1080/01621459.1992.10475217. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1992.10475217.

Nicolas Emmenegger, Ellery Stahler, and Chara Podimata. Prediction-Powered Inference Across Many Tasks for AI Evaluation & Social Science Research, 2026. URL https://arxiv.org/abs/2605.29249. arXiv:2605.29249; preprint.

Benjamin Eyre and David Madras. Regression for the Mean: Auto-Evaluation and Inference with Few Labels through Post-hoc Regression. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 15596–15612, 2025. URL https://proceedings.mlr.press/v267/eyre25a.html.

Sebastian Farquhar, Yarin Gal, and Tom Rainforth. On Statistical Bias In Active Learning: How and When To Fix It. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2101.11665.

V. P. Godambe and V. M. Joshi. Admissibility and Bayes Estimation in Sampling Finite Populations. I. *The Annals of Mathematical Statistics*, 36(6):1707–1722, 1965. doi: 10.1214/aoms/1177699799. URL https://www.jstor.org/stable/2239112.

Huong Ha, Sunil Gupta, Santu Rana, and Svetha Venkatesh. ALT-MAS: A Data-Efficient Framework for Active Testing of Machine Learning Algorithms, 2021. URL https://arxiv.org/abs/2104.04999. arXiv:2104.04999; preprint.

Charles R. Harris et al. Array programming with NumPy. *Nature*, 585:357–362, 2020. doi: 10.1038/s41586-020-2649-2. URL https://doi.org/10.1038/s41586-020-2649-2.

Jesse Hemerik and Jelle Goeman. Exact testing with random permutations. *TEST*, 27:811–825, 2018. doi: 10.1007/s11749-017-0571-1. URL https://link.springer.com/article/10.1007/s11749-017-0571-1.

Allison Marie Horst, Alison Presmanes Hill, and Kristen B. Gorman. *palmerpenguins: Palmer Archipelago (Antarctica) penguin data*, 2020. URL https://allisonhorst.github.io/palmerpenguins/.

D. G. Horvitz and D. J. Thompson. A Generalization of Sampling Without Replacement from a Finite Universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952. doi: 10.1080/01621459.1952.10483446. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1952.10483446.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021. doi: 10.1214/20-AOS1991. URL https://arxiv.org/abs/1810.08240.

Valentin Kilian, Stefano Cortinovis, and François Caron. Anytime-valid, Bayes-assisted, Prediction-Powered Inference, 2025. URL https://arxiv.org/abs/2505.18000. arXiv:2505.18000; preprint.

Jannik Kossen, Sebastian Farquhar, Yarin Gal, and Tom Rainforth. Active Testing: Sample-Efficient Model Evaluation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 5753–5763, 2021. URL https://proceedings.mlr.press/v139/kossen21a.html.

Jannik Kossen, Sebastian Farquhar, Yarin Gal, and Tom Rainforth. Active Surrogate Estimators: An Active Learning Approach to Label-Efficient Model Evaluation, 2022. URL https://arxiv.org/abs/2202.06881. arXiv:2202.06881; preprint.

Neil G. Marchant and Benjamin I. P. Rubinstein. Needle in a Haystack: Label-Efficient Evaluation under Extreme Class Imbalance. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2021. URL https://arxiv.org/abs/2006.06963v2.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein Bounds and Sample Variance Penalization, 2009. URL https://arxiv.org/abs/0907.3740. arXiv:0907.3740; preprint.

Donald Musgrove, Andrew Radtke, and Tarek Haddad. Discordant pair analysis for sample efficient model evaluation. *Scientific Reports*, 13:20987, 2023. doi: 10.1038/s41598-023-48017-4. URL https://www.nature.com/articles/s41598-023-48017-4.

Phuc Nguyen, Deva Ramanan, and Charless Fowlkes. Active Testing: An Efficient and Robust Framework for Estimating Accuracy. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 3759–3768, 2018. URL https://proceedings.mlr.press/v80/nguyen18d.html.

Fabian Pedregosa et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. URL https://jmlr.org/papers/v12/pedregosa11a.html.

Alexander Rubinstein, Benjamin Raible, Martin Gubri, and Seong Joon Oh. DISCO: Diversifying Sample Condensation for Efficient Model Evaluation, 2025. URL https://arxiv.org/abs/2510.07959. arXiv:2510.07959; preprint.

Christoph Sawade, Niels Landwehr, and Tobias Scheffer. Active Estimation of F-Measures. In *Advances in Neural Information Processing Systems*, 2010. URL https://neurips.cc/virtual/2010/poster/2128.

Scikit-learn developers. Toy datasets, 2026. URL https://scikit-learn.org/stable/datasets/toy_dataset.html.

R. J. Serfling. Probability Inequalities for the Sum in Sampling without Replacement. *The Annals of Statistics*, 2(1):39–48, 1974. doi: 10.1214/aos/1176342611. URL https://doi.org/10.1214/aos/1176342611.

Divya Shanmugam, Shuvom Sadhuka, Manish Raghavan, John Guttag, Bonnie Berger, and Emma Pierson. Evaluating multiple models using labeled and unlabeled data. In *Advances in Neural Information Processing Systems*, 2025. doi: 10.52202/085713-1029. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/2c0e78cc177dfeca260ef990a8c99209-Abstract-Conference.html.

Statsmodels developers. Affairs dataset, 2026a. URL https://www.statsmodels.org/stable/datasets/generated/fair.html.

Statsmodels developers. RAND Health Insurance Experiment Data, 2026b. URL https://www.statsmodels.org/stable/datasets/generated/randhie.html.

Pauli Virtanen et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2. URL https://doi.org/10.1038/s41592-019-0686-2.

Ian Waudby-Smith and Aaditya Ramdas. Confidence sequences for sampling without replacement. In *Advances in Neural Information Processing Systems*, 2020. URL https://arxiv.org/abs/2006.04347.

Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B*, 86(1):1–27, 2024. doi: 10.1093/jrsssb/qkad009. URL https://doi.org/10.1093/jrsssb/qkad009.

Emine Yilmaz, Peter Hayes, Raza Habib, Jordan Burgess, and David Barber. Sample Efficient Model Evaluation, 2021. URL https://arxiv.org/abs/2109.12043. arXiv:2109.12043; preprint.

Tijana Zrnic and Emmanuel J. Candès. Active Statistical Inference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 62993–63010, 2024a. URL https://proceedings.mlr.press/v235/zrnic24a.html.

Tijana Zrnic and Emmanuel J. Candès. Cross-Prediction-Powered Inference. *Proceedings of the National Academy of Sciences*, 121(15):e2322083121, 2024b. doi: 10.1073/pnas.2322083121. URL https://arxiv.org/abs/2309.16598.

A Complete proofs and scope of the minimax result

A.1 The binary identity and covariance

For $f, y \in \{0,1\}$, zero–one loss is $f + y - 2fy$. Subtracting the reference's loss yields $(f_j - f_0)(1-2y)$, proving (1). For an arbitrary label-independent subset design with positive informative marginals, linearity gives $\mathbb{E}_R\widehat{\Delta} = N^{-1}\sum_i D_i = \Delta$. Under independent inclusion, the centered estimator is $N^{-1}\sum_i (R_i/\pi_i - 1)D_i$. Its summands are independent, have mean zero, and have covariance $(1-\pi_i)D_iD_i^\top/\pi_i$. Since $(1-2y_i)^2 = 1$, the outer product is $a_ia_i^\top$. Summing and dividing by N^2 proves (3). No limit argument or model for the labels is involved. In particular, a poorly calibrated classifier does not invalidate the covariance calculation.

The equality is specific to the independent-inclusion construction. For a dependent design with pairwise inclusion probability π_{ik}, the covariance additionally contains

$$\frac{1}{N^2}\sum_{i\neq k}\frac{\pi_{ik}-\pi_i\pi_k}{\pi_i\pi_k}a_ia_k^\top s_is_k. \tag{7}$$

These terms generally depend on labels. The minimax lower bound below nevertheless applies to dependent designs because its auxiliary average cancels cross-sign terms. This is the precise difference between the upper-bound construction and the lower-bound comparison class.

A.2 A lower bound for every design-unbiased estimator

Fix a label-independent law P for R. Any ancillary estimator randomness can be included in the observed experiment, independent of the label vector. An estimator may use all selected indices and labels jointly, with arbitrary nonlinear dependence. If $a_i \neq 0$ but $\pi_i = 0$, flip only label i: the observed-data distribution is unchanged while at least one target coordinate changes. Unbiasedness for both labelings is impossible. It is therefore sufficient to consider positive informative marginals and omit zero columns.

Introduce an auxiliary random vector S of independent uniform signs, independent of R and ancillary randomness. This uniform prior is a lower-bound device, not an assumption that real labels are random, balanced, or independent. For each $i \in U$, set $\phi_i = R_iS_i/\sqrt{\pi_i}$. Regardless of dependence within R,

$$\mathbb{E}\phi_i^2 = 1, \qquad \mathbb{E}[\phi_i\phi_k] = \frac{\mathbb{E}[R_iR_k]\mathbb{E}[S_iS_k]}{\sqrt{\pi_i\pi_k}} = 0 \quad (i \neq k). \tag{8}$$

For any coordinate T_j of a design-unbiased estimator,

$$\mathbb{E}[T_jS_i] = \mathbb{E}_S[\Delta_j(S)S_i] = a_{ji}/N. \tag{9}$$

On the event $R_i = 0$, T_j is a function of observations not containing S_i. Independence of the prior signs and label independence of the design imply $\mathbb{E}[T_jS_i\mathbf{1}\{R_i = 0\}] = 0$. Consequently,

$$\mathbb{E}[T_j\phi_i] = \frac{a_{ji}}{N\sqrt{\pi_i}}. \tag{10}$$

Bessel's inequality, or simply expansion of the nonnegative squared residual after orthogonal projection, yields

$$\mathbb{E}[T_j^2] \;\geq\; \sum_{i\in U}(\mathbb{E}[T_j\phi_i])^2 = \frac{1}{N^2}\sum_{i\in U}\frac{a_{ji}^2}{\pi_i}. \tag{11}$$

Design-unbiasedness also gives $\mathbb{E}[T_j\Delta_j(S)] = \mathbb{E}[\Delta_j(S)^2]$. Since $\mathbb{E}\Delta_j(S)^2 = N^{-2}\sum_i a_{ji}^2$, the prior-average mean squared error obeys

$$\mathbb{E}_S\mathbb{E}_P[(T_j - \Delta_j(S))^2] = \mathbb{E}[T_j^2] - \mathbb{E}[\Delta_j(S)^2] \geq \frac{1}{N^2}\sum_i a_{ji}^2(1/\pi_i - 1). \tag{12}$$

Taking a maximum after this lower bound is legitimate because

$$\sup_s \max_j \mathbb{E}_P[(T_j - \Delta_j(s))^2] \geq \max_j \mathbb{E}_S\mathbb{E}_P[(T_j - \Delta_j(S))^2]. \tag{13}$$

Every design in $\mathcal{A}_B$ has marginals summing to at most B, so minimizing over designs and estimators is bounded below by the right side of (4). Independent HT at an optimal marginal vector attains that bound for every label vector, by the covariance identity and unbiasedness. This proves equality. For $0 < B < |U|$, a feasible strictly positive design exists; any bounded objective sublevel keeps each informative inclusion away from zero. Compactness on such a sublevel gives existence of a minimizer. The census and empty-union cases are immediate.

The comparison class includes random mixtures of nonadaptive designs after marginalizing their design randomness. It excludes label-adaptive sampling: then R need not be independent of the auxiliary signs, and the orthogonality argument need not hold. It also excludes biased estimators. For example, with one target sign and no observed label, returning zero is a valid Bayesian or shrinkage decision but cannot be design-unbiased for both sign values. Finally, a hard-cap class is a stricter class than an expected-budget class; the independent construction is not generally feasible under a hard cap. The theorem's value is an attainable optimum for the latter, not a claim of attainable hard-cap optimality.

A.3 Convexity, duality, and the solver certificate

The epigraph primal is $\min_{\pi,t} t$ subject to $v_j(\pi) \le t$, positive informative inclusions bounded by one, and $\sum_i \pi_i \le B$. Each v_j is convex because its Hessian is diagonal with nonnegative entries $2c_{ji}/\pi_i^3$. For $0 < B < |U|$, taking all inclusions equal to $B/(2|U|)$ and a sufficiently large t supplies strict feasibility. Slater's condition therefore gives strong duality. Multipliers on the M epigraph constraints must sum to one after minimizing over t, producing the simplex variable λ.

For strictly positive w_i, the inner problem is strictly convex. A multiplier ν on the budget gives per-item derivative $-w_i/\pi_i^2 + \nu$, so the solution is the capped square-root rule (5). The budget is used fully before census. For any feasible λ',

$$g(\lambda') \le \sum_j \lambda'_j v_j(\pi(\lambda)) = g(\lambda) + (\lambda' - \lambda)^\top v(\pi(\lambda)), \tag{14}$$

so the variance vector is a supergradient of the concave dual. With unique interior solutions it is the ordinary gradient. A returned feasible π supplies the primal upper bound P, and a dual point supplies a lower bound. No optimizer success flag alone certifies the answer. All certificates in the implementation are numerical floating-point checks, not formally verified interval-arithmetic certificates.

A.4 Bernstein regions

For $0 < \pi_i < 1$, the centered scalar summand is $X_{ji} = N^{-1} a_{ji} s_i (R_i/\pi_i - 1)$. Its two possible absolute magnitudes are $|a_{ji}|/N$ and $|a_{ji}|(1-\pi_i)/(N\pi_i)$. Thus $|X_{ji}| \le b_j$. Census terms are exactly zero and do not enter the maximum. The variance sum is V_j. Bernstein's inequality gives

$$\Pr\{|\widehat{\Delta}_j - \Delta_j| \ge r\} \le 2\exp\{-r^2/[2(V_j + b_j r/3)]\}. \tag{15}$$

Solving $r^2 \ge 2L(V_j + b_j r/3)$ yields (6). Summing the two-sided error probability over the fixed M coordinates proves simultaneous coverage. When $V_j = b_j = 0$, the estimate is deterministic and its radius is zero. This argument is for one frozen audit, not optional stopping over successively selected audit sizes.

A.5 Proof of the half-or-census characterization

At $\pi_i = 1/2$, $R_i/\pi_i - 1 = 2R_i - 1$ is a uniform sign. Multiplying by a fixed unknown sign s_i preserves its distribution and independence across items. Items at census add zero centered error. This proves label independence of the joint law, including arbitrary dependence across its M coordinates induced by the known vectors a_i.

For necessity, suppose an informative item i has $\pi_i \notin \{1/2, 1\}$. Choose j with $a_{ji} \ne 0$. Since the summands are independent and mean zero, their third centered moments add, and

$$\mathbb{E}\left[\left(\frac{a_{ji}s_i}{N}\right)^3 \left(\frac{R_i}{\pi_i} - 1\right)^3\right] = \frac{a_{ji}^3 s_i}{N^3} \frac{(1-\pi_i)(1-2\pi_i)}{\pi_i^2}. \tag{16}$$

The coefficient of s_i is nonzero. Flipping only s_i therefore changes the third moment of coordinate j while leaving all other terms unchanged. A distribution with a changed third moment cannot be identical. All relevant moments are finite because the pool is finite and informative inclusions are positive. This proves necessity for invariance across all binary label vectors, not merely invariance for a particular pair of labelings.

A.6 Finite Monte Carlo calibration

Let Z_0 denote the actual maximum standardized centered error under the half-union design, and $Z_1, \ldots, Z_K$ its independent simulated copies generated from the known sign law. They are exchangeable. If $k \le K$, the event $Z_0 > Z_{(k)}$ implies that the rank of Z_0 among all $K+1$ values exceeds k, under any independent tie breaking. Therefore

$$\Pr\{Z_0 > Z_{(k)}\} \le \frac{K+1-k}{K+1} \le \delta. \tag{17}$$

With ties and no randomized rejection, the inequality is conservative. If $k > K$, an infinite radius is valid. Zero-variance coordinates have exactly zero error and are omitted from standardization. In the experiments $K = 9999$, $\delta = 0.05$, and $k = 9500$, where order statistics are one-indexed. A floating-point implementation selects array position 9499. An interpolated percentile need not implement this rule.

A.7 General loss characterization

Under independent inclusions the covariance is $N^{-2}\sum_i (1/\pi_i - 1) D_i(y_i) D_i(y_i)^\top$. Label-independent outer products are sufficient. For necessity over all designs, take all items except one at census and choose a non-census inclusion for the remaining item. Its covariance then reveals its outer product up to a positive factor, so that outer product must be constant over labels. If one nonzero vector v has $ww^\top = vv^\top$, equality of the ranges gives $w = \alpha v$ and equality of norms gives $\alpha \in \{-1, 1\}$. A zero outer product requires the vector itself to be zero. This proves the stated sign-orbit characterization.

B Implementation and dataset details

Training configurations. The logistic grid is $C \in \{0.01, 0.1, 1\}$ with at most 3,000 iterations and tolerance 10^{-8}. Random forests and extra trees each contain 80 trees, use minimum leaf size two, and vary maximum depth over $\{3, 8, \text{unlimited}\}$. Histogram gradient boosting varies maximum leaf nodes over $\{7, 15, 31\}$, uses at most 180 iterations, and enables early stopping with patience 12 and tolerance 10^{-7}. Nearest neighbors use $k \in \{3, 11, 31\}$. Gaussian naive Bayes varies variance smoothing over $\{10^{-11}, 10^{-9}, 10^{-7}\}$. The decision tree uses minimum leaf size three and depths $\{3, 6, \text{unlimited}\}$. Linear discriminant analysis uses the least-squares solver and shrinkage $\{0, 0.1, 0.5\}$. Logistic, nearest-neighbor, naive-Bayes and discriminant pipelines standardize from training data. Tree models do not. Finite tree construction is not an unconverged iterative optimizer; iterative training and numerical design optimization have their recorded stopping criteria.

Data provenance. The processed datasets contain 569 breast-cancer rows, 1,797 digits, 178 wine rows, 6,366 Fair records, 10,000 selected RAND records, and 342 complete-measurement penguins. The original packaged data documentation is cited in the main text. No records are matched to named individuals or interpreted as recommendations. RAND selection uses seed 20260911 before consulting its response. Binary transformations and SHA-256 checksums are in `data/catalog.json`; training, validation, and audit indices are in the pool NPZ files. Source dataset permissions remain with their original providers; the accompanying code license does not relicense those data.

Solver configuration. Water filling rescales positive weights before bisection and stops when the absolute budget residual is at most $10^{-13}\max(1, B)$. Its safety cap is 200 iterations. SLSQP uses analytic objective derivatives, function tolerance 10^{-13}, simplex bounds $10^{-10} \le \lambda_j \le 1$, and a total cap of 1,500 iterations over at most ten Hessian restarts. Convergence is accepted only at relative primal–dual gap at most 10^{-5}. A time guard interrupts optimization at 80% of its 30-second budget.

The full experiment runner also has an 80% resource guard and persists completed configurations. No failed or unconverged proposal is included in the reported results. The supplied audit script independently checks saved probabilities, objective values, and replicate-derived metrics.

Fixed-size comparison. Writing $u = |U|$ and $k = B$, uniform sampling of exactly k union items yields $\widehat{\Delta}_j = (u/N)\overline{D}_{j,S}$. Its retrospective variance is

$$(u/N)^2(1 - k/u)\,\frac{S^2_{j,U}}{k}, \qquad S^2_{j,U} = \frac{1}{u-1}\sum_{i\in U}(D_{ji} - \overline{D}_{j,U})^2. \tag{18}$$

This variance uses the complete labels only for experimental measurement. For inference, the implementation instead applies the range-2 Serfling bound, with common radius $(u/N)\sqrt{2(1-(k-1)/u)\log(2M/\delta)/k}$ for $k < u$, and zero for a census. It does not pretend the retrospective variance was available to the live audit.

Prediction-assisted comparison. For fixed μ_i, the residual estimator is unbiased because $\mathbb{E}_R[R_i(D_i - \mu_i)/\pi_i] = D_i - \mu_i$. Its true conditional variance uses $(D_{ji} - \mu_{ji})^2$, which is generally label-dependent. Under a Bernoulli proxy model with probability η_i, the proxy residual second moment is $4a^2_{ji}\eta_i(1-\eta_i)$; this motivates proxy-CV's design without making the proxy model an assumption of unbiasedness. Its confidence bound uses the larger valid envelope stated in the main text. The fixed coefficient of one in this residual correction is not PPI++'s data-adaptive coefficient.

C Additional empirical evidence

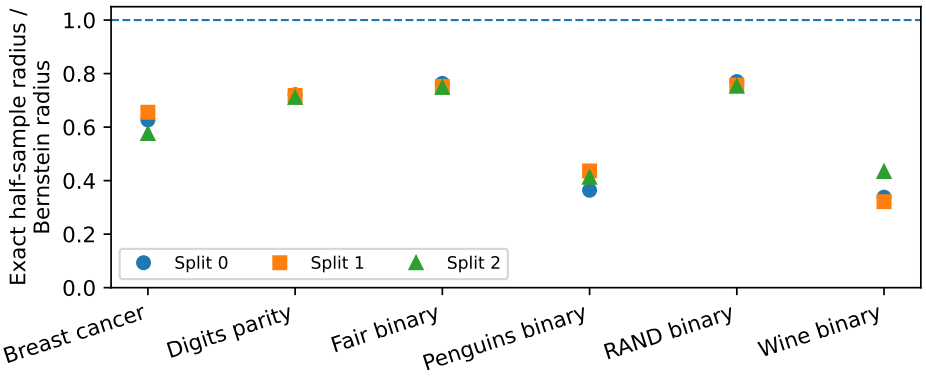


Figure 2: Every trained real-data pool in the half-sampling comparison. Each point is the ratio of mean coordinate radii for exact-law calibration and the matched Bernstein bound. Both procedures use the same independent half-union estimator and the same expected number of labels.

Synthetic populations. The first experiment fixes $N = 400$ and six comparison rows. Nested supports have sizes $20, 40, 80, 160, 240, 320$; disjoint supports are six consecutive blocks of 50; overlapping supports use cyclic shifts of 30 and widths $50 + 30j$; the bridge geometry combines a common block of 40 with six disjoint private blocks of 20; identical supports repeat a block of 180. A zero reference and the support indicators give genuine binary prediction matrices. Labels are sampled once with probabilities $0.15 + 0.7i/N$ and then frozen across annotation repetitions. Fractions $0.1, 0.25, 0.5$ create 15 configurations. These controlled geometries test the mathematical design, not trained model performance.

The separate proxy-stress population has 400 fixed labels generated with seed 1118 and probabilities $p_i = 0.02 + 0.96i/N$. Its six cyclic support widths are $90 + 20j$ with shifts of $35j$. For reversal level $t \in \{0, 0.25, 0.5, 0.75, 1\}$ the proxy is $0.05 + 0.9[(1-t)p_i + t(1-p_i)]$. Only this proxy changes. Two budget fractions, 0.1 and 0.5, create ten additional configurations. The squared proxy residual is measured on the same frozen labels for all reversal levels; random numbers are not used to manufacture a desired performance curve.

Table 3: All-method aggregate diagnostic results on real-data configurations. Variance and MSE columns average within-configuration ratios to uniform-union. Coverage is the minimum observed frequency; no coverage is attached to the invalid ablation.

Method	Variance ratio	MSE ratio	Min. coverage	Mean max. abs. bias
uniform_all	17.681	18.214	0.994	0.00283
uniform_union	1.000	1.000	0.995	0.00095
mean_disagreement	1.110	1.112	0.996	0.00109
entropy	1.008	1.007	0.995	0.00101
dova	0.723	0.774	0.995	0.00112
union_srs	0.777	0.786	0.999	0.00084
uniform_cv	0.835	0.831	0.999	0.00074
proxy_cv	0.736	0.748	1.000	0.00085
unweighted_invalid	–	1.482	–	0.01419

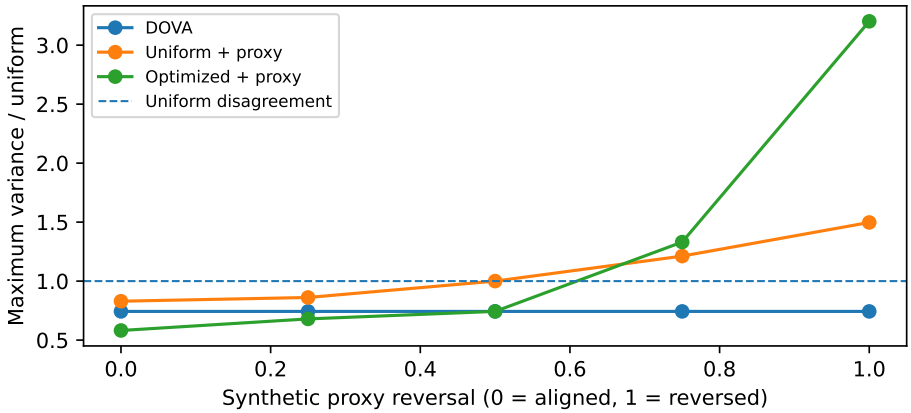


Figure 3: Explicitly synthetic proxy misspecification at requested budget fraction 0.5. The unchanged prediction geometry keeps DOVA's exact variance fixed. Corrective estimators can exploit an aligned proxy or suffer under its reversal. This is not a real-data shift experiment.

Numerical verification. The test suite enumerates 32 label vectors and 32 inclusion masks for a five-item covariance identity, and 16 label vectors with 16 masks for a four-item half-law check. Twelve small numerical design instances compare the dual optimizer with a separately implemented primal formulation. A separate all-functions calculation enumerates observed-data states and solves the weighted least-norm unbiasedness equations. For four items it includes 81 possible observation states under unequal independent inclusion, 24 under uniform size-two sampling, and 17 under an all-or-none design. In each case the computed minimum-prior-second-moment estimator matches HT to below 7×10^{-15}. The dependent designs have the same prior-average lower bound but larger worst-label risk in these examples. This verifies a finite instance of the proof mechanism rather than an exhaustive search over all finite populations.

Statistical reporting. Per-pool Clopper–Pearson intervals for measured coverage are included in the CSV files. These quantify Monte Carlo variation with one fixed calibration draw. They are not an assertion of exact conditional coverage for that calibration draw. A descriptive bootstrap over task–seed pools is also saved but is not used for an inferential claim, since splits of a common source are dependent. The manuscript does not report a t-test, ANOVA, dataset-level significance, or a claim that 560,000 masks are independent benchmark replications. Exact variances are deterministic functions of the saved finite population and allocation, so their comparisons require no simulation-based significance test.

D EVIDENCE MAP AND REPRODUCIBILITY PROCEDURE

The artifact's `README.md` contains commands for installation, tests, experiment execution, analysis, and rebuilding the PDF. `results/v2/summary.csv` and `results/v2/results.json` contain all configurations; `allocations/` records actual probabilities and proxy offsets; `replicates/` stores errors and label counts. `paper_facts.json` supplies numerical macros and tables through `make_tables.py`. The bibliography ledger preserves citation keys, source URLs, available DOIs, and whether verification reached full methods, an abstract, or metadata. No placeholder citation, fabricated external review, or unexecuted experiment is used to fill a missing result.

Version one contains an actual pilot, initial manuscript, internal review, and a refinement decision. Version two expands the data, comparisons, misspecification checks, theorem scope, and manuscript. The review files distinguish internally addressed concerns from remaining novelty and independent-verification risks. This evidence map is a navigation aid for the supplied artifact, not a new methodological contribution. The current PDF is a research manuscript prepared with the ICLR 2027 style; it has not been submitted or independently endorsed.