SAME RISKS, DIFFERENT ERROR BARS: DISAGREEMENT-AWARE UNCERTAINTY FOR CROSS-VALIDATED MODEL SELECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

A practitioner who chooses between two fixed classifiers by cross-validation reports the selected model's cross-validated score, usually with a fold-based or binomial error bar. Fold complements overlap heavily, so the selection is re-run on almost the same data in every fold, and when the two models are close the folds flip between them. We show that this adds a reuse term to the mean squared error of the score that the usual error bars omit, and that the term is set by how the two models make mistakes *jointly*, not by their risks alone: pairs with identical risks can differ in it fifteen-fold. A pathwise identity (the cross-validated score equals the winner's empirical risk plus $1/n$ times the margins of the folds whose selection flips) yields an exact finite-sample evaluator of this error for any law of the paired 2×2 loss table and any fold count dividing n. Applied to the observed table it gives a *reuse-aware root mean squared error (RMSE)*, which apart from a floor on empty cells is the ideal paired-loss bootstrap of the select-and-score pipeline for the two fixed models, computed without resampling; its coverage is measured, not guaranteed. On a synthetic grid fixed before any run the usual intervals cover as little as 75.0% at nominal 95%, while the reuse-aware interval's worst cell is 0.895 against 0.857 for a constant widening matched to its width. On 252 cells built from open-weight language-model checkpoints its worst cell is 0.907, which an exactly calibrated interval would fall to or below with probability 0.07 (expected worst cell 0.917), against 0.877 for the binomial and Agresti–Coull intervals (probability below 10^{-5}). On our declared grid of 1212 cells of real fitted classifiers, however, a textbook Agresti–Coull interval does as well or better, mostly at $n = 200$ where the estimated table is noisiest; at $n = 1000$ the reuse-aware interval's worst cell is the higher, 0.898 against 0.878. In practice, with two fixed candidates, report the winner's empirical risk, which the identity recovers from a cross-validated score and its fold contrasts; a cross-validated score that is reported should carry the reuse-aware RMSE, not a fold-based or binomial error bar.

1 INTRODUCTION

A practitioner has two fixed candidate models, say a production model and a challenger or two released checkpoints, and a few hundred or a few thousand labelled examples with which to choose between them (Miller, 2024). Following the advice of the selection-bias literature (Varma & Simon, 2006; Cawley & Talbot, 2010), they cross-validate the *selection*: on each fold's training part they keep whichever candidate has the smaller empirical risk there, score it on the held-out part, and report the average as the winner's risk, next to a fold-based or binomial standard error.

The issue. With k folds each training complement shares a fraction $(k-2)/(k-1)$ of its observations with every other, so the selection is re-run on highly dependent data. When the two models are close, a sample within one fold of a tie makes the folds flip between candidates, selection instability in the sense of Breiman (1996), and the score picks up the loser's losses on the flipped folds. This is extra uncertainty in the reported number, and the usual error bars do not see it: a nominal 95% interval around the cross-validated score covers as little as 75.0% on our synthetic grid and 68.3% on real fitted classifiers with the fold-based standard error, and 80.4% and 74.9% with the binomial one.

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.

The key observation. How often the folds flip, and by how much, depends on how the two models make mistakes jointly, not on their individual risks. Two pairs can fail on the same examples or on different ones; only the second makes fold-level switching matter. At $n = 500$ and $k = 10$, two loss laws with the same risks $(0.20, 0.20)$ carry reuse terms 1.4×10^{-5} and 2.1×10^{-4} (Section 4). An error bar built from the risks alone cannot tell them apart.

What we propose. *A reuse-aware RMSE for the cross-validated score of the selected model* (Section 3). Its input is the two candidates' paired 0–1 losses on the evaluation sample and the fold count. For any given law of the paired 2×2 table, the one underlying McNemar's test (McNemar, 1947), an exact finite-sample evaluator computes the mean squared error of the cross-validated score about the risk of the full-sample winner, the model that will be shipped; the method applies it to the estimated table and takes the square root. The table, raw disagreement and the variance of a cross-validated estimate are all known objects, and none of them is our metric. What is new is, first, the reuse-aware mean squared error itself, whose target is the risk of the model actually shipped rather than an algorithm's average error; second, the pathwise identity that makes it computable, $\hat{r}_{\mathrm{CV}} = \widehat{R}_{\mathrm{plain}} + G/n$ with G the summed margins of the folds whose selection flips; and third, its consequence that two pairs with equal risks can carry reuse terms fifteen-fold apart, because the term depends on the joint error structure through the disagreement cells, the gap and how often the models fail together (Proposition 1).

Contributions. **(i) Method:** the reuse-aware RMSE and an evaluator, exact for any given paired law, that needs no refitting, resampling or extra data (about 20 ms at $n = 1000$). **(ii) Analysis:** a pathwise identity writing the cross-validated score as the winner's empirical risk plus the margins of the flipped folds, an exact decomposition of the resulting mean squared error with a sign result at a population tie, and the finite-sum reduction the evaluator computes (Section 4). **(iii) Evidence:** on a synthetic grid, a declared grid of real fitted classifiers and 252 cells built from open-weight language-model checkpoints the usual error bars under-cover and the reuse-aware RMSE is nearer nominal; against a constant widening matched to its average width it wins on the worst cell and the spread, so on the synthetic grid it adapts rather than merely widens (Section 5). We also report where it does not help: a textbook Agresti–Coull interval beats it on our declared real grid, although RCV has a higher worst cell than plain Agresti–Coull in three of four later frozen rounds (Table 3), and the reuse share it outputs is not a reliable flag for a wrongly chosen model. **(iv) Practice:** with two fixed candidates, report the winner's empirical risk, which lost nothing in mean squared error on every two-candidate real law we evaluated and which the identity recovers from a cross-validated score and its fold contrasts as $\hat{r}_{\mathrm{CV}} - G/n$; if the cross-validated score is reported, attach the reuse-aware RMSE. The method measures uncertainty; it does not select the better model and never changes the winner.

2 Background and Related Work

Cross-validation and its error bars. Bengio & Grandvalet (2004) show that no universal unbiased estimator of the variance of k-fold cross-validation exists, and Nadeau & Bengio (2003) and Dietterich (1998) give standard tests for comparing algorithms. Asymptotically valid intervals for a cross-validated risk (Bayle et al., 2020; Austern & Zhou, 2025) and for comparisons of candidates (Kissel & Lei, 2023) assume algorithmic stability, and ties between candidates are exactly where empirical risk minimisation loses its fast stability rate (Rubinstein & Simma, 2012); Bayle et al. (2026) show that relative stability can fail even when each candidate is stable. Bates et al. (2024) show that cross-validation tracks the average risk of models fitted on other samples rather than the model in hand, that its naive standard errors under-cover because folds are dependent, and propose the nested interval we run as a baseline. Nachum et al. (2025) decompose the same conditional error exactly for symmetric algorithms, with a term for the correlation of fold estimates across overlapping complements; that term is what we correct, and what is new is an exact finite-sum formula for it at the fold count actually used, from the paired table alone. Arlot & Celisse (2010) survey cross-validation for selection.

Selection bias and nested cross-validation. The minimum cross-validated error over a tuning grid is optimistic (Varma & Simon, 2006; Tibshirani & Tibshirani, 2009; Cawley & Talbot, 2010), which motivates nesting the selection; Wainer & Cawley (2021) argue that nesting is rarely needed in practice, and Krstajic et al. (2014) and Varoquaux (2018) document the pitfalls and the wide error bars of small-sample cross-validation. Iyengar et al. (2024) show that K-fold cross-validation need not improve on the plug-in estimate for smooth losses; our setting is the non-smooth step

their theory excludes, a selection that flips at a tie. Davison & Hall (1992) show that bootstrap and cross-validation estimates of a discrimination error rate differ in bias and variability mainly when the populations are close, the regime we study. Reusing one evaluation set for adaptive choices is also the subject of the reusable holdout (Dwork et al., 2015) and the Ladder (Blum & Hardt, 2015).

Inference after selection. Intervals for a data-selected parameter range from multiple comparisons with the best (Hsu, 1984) and false-coverage-rate adjustment (Benjamini & Yekutieli, 2005) to conditional and hybrid intervals for an estimated winner (Andrews et al., 2024) and locally simultaneous inference (Zrnic & Fithian, 2024); bootstrap corrections for the selected model's performance (Tsamardinos et al., 2018) resample the selection over a fixed set of cross-validated out-of-sample predictions. The union-bound interval in the supplement (Section S3) is the Bonferroni member of this family, in the spirit of test-set bounds (Langford, 2005). For binomial proportions, Brown et al. (2001) recommend the score interval of Wilson (1927) and the interval of Agresti & Coull (1998) over the Wald interval; both repair discreteness, not reuse, and we run them as baselines.

3 Proposed Method

Setting and target. Two fixed candidate classifiers h_- and h_+, chosen without looking at the evaluation sample, and an iid evaluation sample of size n with 0–1 loss. The practitioner runs k-fold cross-validation of the selection: on the complement of each fold they keep the candidate with the smaller empirical risk (ties go to h_-) and score it on the held-out fold; $\hat{r}_{\mathrm{CV}}$ is the average of these held-out losses. They ship the full-sample winner $\hat{h}_n$, the candidate with the smaller empirical risk on the whole sample, and report $\hat{r}_{\mathrm{CV}}$ as its risk. The target of the error bar is therefore $R_P(\hat{h}_n)$, the population risk of the model that is shipped, and the quantity to estimate is $\mathrm{MSE}_{\mathrm{CV}} = \mathbb{E}\big(\hat{r}_{\mathrm{CV}} - R_P(\hat{h}_n)\big)^2$. We assume k equal folds ($k \mid n$) drawn independently of the data.

Reuse-aware RMSE (RCV, reuse-corrected cross-validation).
Input: the paired 0–1 losses $(\ell_{-,i}, \ell_{+,i})_{i=1}^n$ of the two fixed candidates on the evaluation sample, and the fold count k.
Step 1. Form the paired 2×2 table $\hat{p} = (\hat{p}_{00}, \hat{p}_{01}, \hat{p}_{10}, \hat{p}_{11})$, where $\hat{p}_{uv}$ is the fraction of points on which h_- has loss u and h_+ has loss v. Floor the two disagreement cells at $1/(2n)$ and renormalise (fixed in advance; it binds only when a cell is empty).
Step 2. Run the exact finite-sample evaluator (Proposition 3) at $(n, k, \hat{p})$: it treats $\hat{p}$ as the law of an iid loss pair and returns the exact mean squared error of the cross-validated score about the selected model's risk, together with its reuse component.
Step 3. Read off $\widehat{\mathrm{MSE}}_{\mathrm{CV}} = \widehat{\mathbb{E}Z^2} + \hat{\Delta}_{\mathrm{reuse}}$, the plain estimate's error plus the reuse term (Section 4).
Step 4. Take the square root: $\mathrm{RMSE}_{\mathrm{RCV}} = \big(\widehat{\mathrm{MSE}}_{\mathrm{CV}}\big)^{1/2}$.
Output: the error bar $\mathrm{RMSE}_{\mathrm{RCV}}$, the nominal normal-approximation interval $\hat{r}_{\mathrm{CV}} \pm z_{1-\alpha/2}\,\mathrm{RMSE}_{\mathrm{RCV}}$ ($\alpha = 0.05$, $z_{0.975} = 1.96$; coverage not guaranteed), and the reuse share $\hat{s} = \hat{\Delta}_{\mathrm{reuse}}/\widehat{\mathrm{MSE}}_{\mathrm{CV}}$.

What it does and does not do. The reuse-aware RMSE is an uncertainty measurement for a number the practitioner already reports. It does not choose between the models, does not change the winner, and does not say whether the winner is the better model. It is a root mean squared error, not a standard error, because the squared selection bias of the score is part of $\mathrm{MSE}_{\mathrm{CV}}$. Apart from the floor, and for the two fixed candidates, it is the ideal paired-loss bootstrap estimate of the score's RMSE, computed exactly rather than by resampling: resampling the n loss pairs and re-running the fold selection draws them iid from $\hat{p}$, and in that resampled world each candidate's risk is its empirical risk, which are the law and the target step 2 evaluates (Section S2 checks this by simulation); like any bootstrap, it inherits the sampling error of $\hat{p}$. Exactness refers to the evaluator: given a law, its finite sums equal $\mathrm{MSE}_{\mathrm{CV}}$ in exact arithmetic, with no simulation; the released floating-point implementations are deterministic and agree with exact rational arithmetic to better than 10^{-12} in relative error on every law checked (Section S2). The error bar plugs in the estimated, floored table, so it inherits that table's sampling error; no finite-sample coverage guarantee is claimed, and

its calibration is measured (Section 5). A practitioner who needs a guarantee for the selected model's risk can use the union-bound interval of Section S3, which attaches to the winner's empirical risk.

Computation. Folds are exchangeable and each is independent of its complement, so every moment of the score's error reduces to finite sums over the laws of the contrast on $m = n/k$, $n - m$, $n - 2m$ and n points (Proposition 3). Nothing is simulated or resampled. The released tool takes 12 to 27 ms at $n = 1000$, $k = 10$ on two cores. The step 1 floor binds when a disagreement cell is empty, in 15.2% to 17.2% of replicates at the cells where the interval fails worst. At the eight worst synthetic and real cells, 200,000 replicates give coverage 0.937 to 0.962 when the evaluator is handed the population table and 0.833 to 0.926 with the plug-in table, so the under-coverage there comes from estimating the table (Section 6).

A worked example. From the digits pool of Section 5, 500 evaluation points compare a frozen logistic regression (h_-, pool risk 0.0436) with a frozen RBF support vector machine (h_+, pool risk 0.0371) at $k = 10$. The observed table is $(\hat{n}_{00}, \hat{n}_{01}, \hat{n}_{10}, \hat{n}_{11}) = (469, 10, 11, 10)$: the models disagree on 21 points, 11 of them in favour of h_+, which wins the full-sample comparison by one error, so the winner's empirical risk is 0.040. Three fold complements select h_- instead, and on those folds h_- makes 2, 2 and 1 more errors than h_+, so the cross-validated score is $0.040 + 5/500 = 0.050$. Its fold-based standard error is 0.0068, its binomial one 0.0097, and $\mathrm{RMSE}_{\mathrm{RCV}} = 0.0104$ with a reuse share of 30.0%. The three intervals are 0.050 ± 0.013, 0.050 ± 0.019 and 0.050 ± 0.020.

4 Theoretical Analysis

Notation. For one observation let $(\ell_-, \ell_+) \in \{0,1\}^2$ be the two losses, with $p_{uv} = P(\ell_- = u, \ell_+ = v)$, and put $X = \ell_- - \ell_+ \in \{-1, 0, 1\}$ and $B = (\ell_- + \ell_+)/2$. The risks are $R_P(h_-) = p_{10} + p_{11}$ and $R_P(h_+) = p_{01} + p_{11}$; their difference is $\mu = \mathbb{E}X = p_{10} - p_{01}$ and their mean is $\beta = \mathbb{E}B$. The loss pairs are iid over the sample. Let $T = \sum_i X_i$, let $A_j = \sum_{i \in F_j} X_i$ be the contrast on fold F_j, and let $\tau(t) = +1$ for $t > 0$ and -1 otherwise, so the full sample selects h_+ iff $\tau(T) = +1$ and fold j's complement selects h_+ iff $\tau(T - A_j) = +1$. Write $\widehat{R}_{\mathrm{plain}}$ for the winner's empirical risk, $Z = \widehat{R}_{\mathrm{plain}} - R_P(\hat{h}_n)$ for its error,

$$G = \sum_{j:\, \tau(T - A_j) \neq \tau(T)} |A_j|$$

for the summed margin of the folds whose selection flips, $N_D = \#\{i : X_i \neq 0\}$ for the number of disagreements, and, when $p_{00} + p_{11} > 0$, $\beta_0 = p_{11}/(p_{00} + p_{11})$ for the probability that both candidates are wrong on a point where they agree.

Proposition 1 (Pathwise identity and exact decomposition)**.** *On every sample, $\hat{r}_{\mathrm{CV}} = \widehat{R}_{\mathrm{plain}} + G/n$, and $G \geq |T|$ whenever $G > 0$. For every law define $\Delta_{\mathrm{reuse}} := \mathrm{MSE}_{\mathrm{CV}} - \mathbb{E}Z^2$. If $p_{00} + p_{11} > 0$, so that β_0 is defined,*

$$n^2 \Delta_{\mathrm{reuse}} = (1 - 2\beta_0)\,\mathrm{Cov}(G, N_D) + \mathbb{E}\big[G(G - |T|)\big] + n\mu\,\mathbb{E}[\tau(T)G]$$

and $G(G - |T|) \geq 0$ on every sample.

In words: cross-validating a fixed two-way selection can only raise the reported risk, and by exactly the flipped folds' margins; a fold flips only if its own contrast is at least $|T|$ in magnitude and has the sign of $\tau(T)$. The reuse term has three parts: the covariance of the shift with the number of disagreements, a non-negative second-moment term, and a drift term that vanishes at a tie. Everything depends on the joint table through p_{01}, p_{10} and β_0, not only through the two risks.

Proof idea. (1) Each point's loss under the candidate with sign s is $B_i - sX_i/2$, so $\hat{r}_{\mathrm{CV}}$ and $\widehat{R}_{\mathrm{plain}}$ differ only on folds whose complement picks the other candidate. (2) On such a fold removing A_j changes the sign of T, which forces A_j to carry the sign of $\tau(T)$ and $|A_j| \geq |T|$; each contributes $|A_j|/n$. (3) Expanding $(Z + G/n)^2$ pairs G with the three parts of Z (Section S1). (4) Given the contrasts, the agreement points are independent Bernoulli(β_0) draws, which turns $\mathbb{E}[(\bar{B} - \beta)G]$ into the covariance term. The delicate step is (2) at $T = 0$, where ties go to h_-.

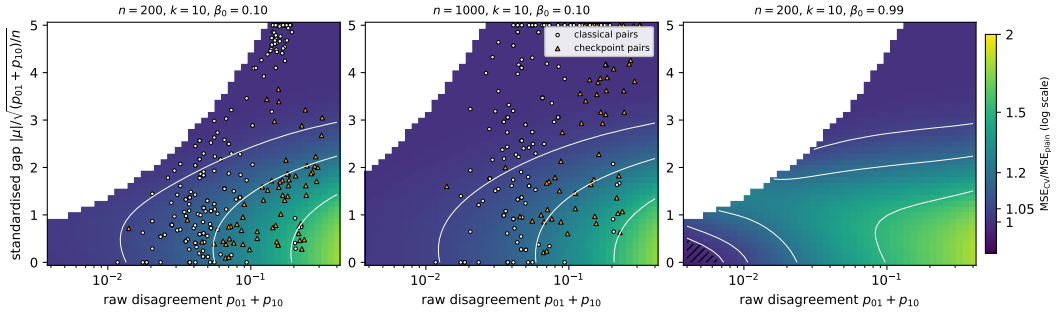

Figure 1: The exact ratio $\mathrm{MSE}_{\mathrm{CV}}/\mathbb{E}Z^2$ at $k = 10$ over the paired-loss table (log colour scale, white contours at 1, 1.05, 1.2, 1.5), with the 202 classical (scikit-learn) pairs and 63 checkpoint laws overlaid (points above a gap of 5 are drawn at the top edge). Left and centre: $\beta_0 = 0.10$, at $n = 200$ and $n = 1000$. Right: $\beta_0 = 0.99$ at $n = 200$; hatched cells have a negative reuse term. The colour bar's $\mathrm{MSE}_{\mathrm{plain}}$ is $\mathbb{E}Z^2$. Overlaid laws are placed by their disagreement and gap only; the colour is the panel's β_0, not each law's own.

Proposition 2 (Sign at a population tie)**.** *If $p_{10} = p_{01}$ and $p_{00} + p_{11} > 0$, then $\mathbb{E}G = \mathbb{E}|T|/2$, and*

$$n^2\Delta_{\mathrm{reuse}} = \left(\tfrac{1}{2} - \beta_0\right)\mathrm{Cov}(|T|, N_D) + \mathbb{E}\big[G(G - |T|)\big], \qquad \mathrm{Cov}(|T|, N_D) \geq 0.$$

So $\Delta_{\mathrm{reuse}} \geq 0$ whenever $p_{11} \leq p_{00}$.

At a tie the mean shift $\mathbb{E}G/n$ equals the plain estimate's selection bias $\mathbb{E}|T|/(2n)$, so the cross-validated score is unbiased there, and the reuse term is the variance it adds minus the squared selection bias it removes. The condition $p_{11} \leq p_{00}$ says the two models are wrong together no more often than right together; outside it, or off the tie, the sign is not guaranteed, and at a tie with $p_{11} > p_{00}$ it can be negative (Section 5). *Proof idea.* Condition on which points are disagreements. At a tie their signs are fair coins, so a fold's contrast is independent of its complement's and has mean zero; averaging the identity then gives $\mathbb{E}G = \mathbb{E}|T|/2$ and $\mathrm{Cov}(G, N_D) = \frac{1}{2}\mathrm{Cov}(|T|, N_D)$. Given $N_D = d$, $|T|$ is the distance of a d-step fair walk from the origin, whose mean does not decrease with d, so the covariance is non-negative. Full proofs are in Section S1.

Proposition 3 (Exact finite-sum evaluator)**.** *For every four-cell law (including $p_{01} = p_{10} = 0$, where no fold can flip), every $k \geq 2$ and every $n = mk$, both $\mathrm{MSE}_{\mathrm{CV}}$ and $\mathbb{E}Z^2$ are explicit finite sums over the laws of $\sum_{i \leq N} X_i$ and $\mathbb{E}[\sum_{i \leq N}(\ell_{-,i} + \ell_{+,i})\mathbf{1}\{\sum_{i \leq N} X_i = t\}]$ for $N \in \{m,\, n-m,\, n-2m,\, n\}$, computable in $O(n)$ arithmetic operations.*

The proof and the formulas are in Section S2. The argument uses only two facts: each fold is independent of its complement, and two folds are conditionally independent given the rest of the sample. The evaluator is exact for the law it is given; in the method that law is the floored plug-in table $\hat{p}$, so the method's accuracy is an empirical question, answered in Section 5.

Same risks, different reuse. Because $\mathrm{MSE}_{\mathrm{CV}}$ is a function of the whole table, the risks do not determine it. At $n = 500$ and $k = 10$ the tables $(0.79, 0.01, 0.01, 0.19)$ and $(0.65, 0.15, 0.15, 0.05)$ both give risks $(0.20, 0.20)$. In the first the models make the same mistakes, disagree on 2% of points, and the folds rarely flip; in the second they disagree on 30%. Their exact reuse terms are 1.4×10^{-5} and 2.1×10^{-4}, and the ratios $\mathrm{MSE}_{\mathrm{CV}}/\mathbb{E}Z^2$ are 1.045 and 1.683. An error bar that is a function of the two risks alone gives both tables the same width. Figure 1 maps the ratio over the whole table: at $\beta_0 = 0.10$ it is at least one on every grid cell and largest at a tie with large disagreement; at $\beta_0 = 0.99$ it dips below one only at grid cells where the models disagree on at most 1.37 points in expectation (Section S5).

5 Experiments

Protocol and baselines. The synthetic grid, the real grid and the width-matched control follow a protocol fixed and hashed before any run; every later study is dated in a deviation register in the artifact (Section S6). Nominal level is 95%. Each interval is summarised over configurations by its

Table 1: Synthetic grid, 340 cells: coverage of a nominal 95% interval for the risk of the full-sample winner.

error bar	worst	spread	mean
fold-based	0.750	0.192	0.866
binomial	0.804	0.156	0.909
bootstrap	0.804	0.156	0.910
binomial × constant matched to RCV's width	0.857	0.128	0.940
RCV (reuse-aware RMSE)	**0.895**	**0.075**	**0.938**

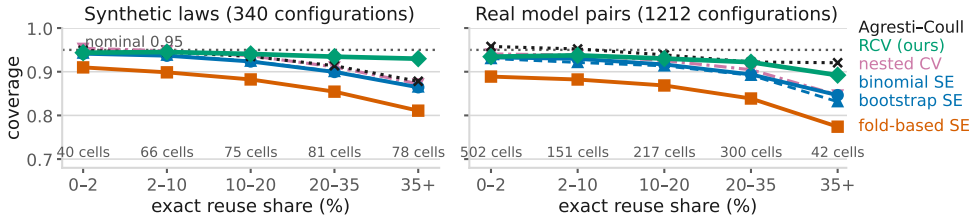

Figure 2: Coverage of a nominal 95% interval for the risk of the full-sample winner, banded by exact reuse share; each point is the unweighted mean over the configurations in that band, whose count is printed above the axis. Left: 340 synthetic cells. Right: all 1212 cells of the declared real grid. The fold-based, binomial and bootstrap standard errors degrade as the reuse share grows; nested cross-validation (Bates et al., 2024) is wider and still follows the binomial one down. RCV stays nearer nominal on the synthetic laws but falls to 0.892 in the top real-pool band, and Agresti–Coull, which repairs discreteness and not reuse, beats it on the declared real grid in every band but one, a tie. SE: standard error; CV: cross-validation.

worst cell, its spread (best cell minus worst) and its mean, and each grid is read beside its noise floor, the worst cell and spread that an exactly calibrated interval would show from finite replicates alone. Three baselines are current practice: the fold-based standard error $\mathrm{sd}(\text{fold means})/\sqrt{k}$, the binomial (Wald) standard error $\sqrt{\hat{r}(1-\hat{r})/n}$, and a bootstrap of the held-out loss vector, which does not re-run the selection (RCV is, apart from its floor, the bootstrap that does). Two repair the discreteness of the binomial interval rather than the reuse: Wilson and Agresti–Coull. The control is the binomial standard error times one global constant matched to RCV's average width: without it one cannot tell whether a method adapts or is merely wider, and the protocol named it the decisive comparison. The last baseline is the nested cross-validation interval of Bates et al. (2024) with its bias correction and 200 random splits. Every interval is scored against the same target, the risk of the full-sample winner.

Synthetic grid: under-coverage and adaptivity. 340 configurations span risk gaps 0 to 0.04, raw disagreement 0.02 to 0.30, n from 200 to 2000 and $k \in \{5, 10, 20\}$, with 2000 replicates in 204 cells and 1000 or 400 in the rest. The usual error bars under-cover, the fold-based one worst (Table 1). The width-matched control equates mean width to five decimals (0.01590 against 0.01590) and nearly the mean coverage (0.940 against 0.938), and still separates: worst cell 0.857 against 0.895 and spread 0.128 against 0.075, so RCV's gain is where it puts its width, not how much it uses. On all 684 admissible configurations of the declared design the comparison repeats (0.852 and 0.131 against 0.895 and 0.075, at the swept constant 1.09 nearest the matched 1.094). Replicate noise alone would give a spread of 0.047 and a worst cell of 0.925. Banded by exact reuse share (Figure 2, left), the correction is inert below a 2% share, moving coverage from 0.941 to 0.944, and large above 35%, moving it from 0.864 to 0.930. Wilson and Agresti–Coull, which repair discreteness, both reach worst cell 0.8245 and spread 0.1425 here at their own centres. The nested cross-validation interval is 7% wider than RCV yet has worst cell 0.810.

Real fitted classifiers. Six classifiers (logistic regression, RBF support vector machine, random forest, k-nearest neighbours, naive Bayes, decision tree) are trained on part of each of three scikit-learn datasets (digits, wine, breast cancer) under five split seeds and frozen; their paired losses on the held-out pool define the population, so both true risks are exact. All $\binom{6}{2} = 15$ pairs are used:

Table 2: Declared real grid, 1212 cells: coverage of a nominal 95% interval. Wilson and Agresti–Coull are at their own centres; centred at the cross-validated score, as in Table 3, Agresti–Coull gives 0.876/0.124/0.938 and Wilson 0.784/0.216/0.929.

error bar	worst	spread	mean
fold-based	0.683	0.257	0.868
binomial	0.749	0.216	0.917
bootstrap	0.752	0.204	0.913
Wilson (2.5% narrower than RCV)	0.832	0.145	0.936
Agresti–Coull (1.5% wider than RCV)	0.868	0.127	0.944
RCV (reuse-aware RMSE)	**0.827**	**0.142**	**0.929**

Table 3: Four later rounds, each frozen before running, on unused public pools: worst cell / spread over 180 cells. $f = 1/\sqrt{1-\hat{s}}$ is the reuse factor, with $\hat{s}$ clipped to $[0, 0.99]$ so that $f \geq 1$ (the clip never bound); round 4 used 4,000 replicates per cell, the others 1,000; $*$ marks each round's declared subject; the constant row is Agresti–Coull scaled to the mean width of AC$\times f$. The constants that decided rounds 1 and 2 were matched to the Wilson form's width (0.890/0.081; 0.917/0.056, equal to the row's by coincidence); read after the fact, AC$\times f$ fails round 2 too.

round (pools)	1 (Adult, Cover)	2 (MAGIC, Bank)	3 (Shop., Diab.)	4 (Occ., Dota 2)
Agresti–Coull (AC)	0.890/0.074	0.909/0.064	0.902/0.067	0.883/0.117
RCV (reuse-aware RMSE)	0.918/0.047	0.913/0.055	0.919/0.052	0.827/0.134
Wilson $\times f$	0.924/0.041*	0.913/0.055*	0.921/0.049	0.912/0.088
AC $\times f$	0.928/0.038	0.913/0.063	0.928/0.042*	0.931/0.069*
AC $\times$ constant matched to it	0.904/0.070	0.917/0.056	0.920/0.058	0.920/0.081
subject passes the five-part rule	yes	no	yes	yes

225 cases, of which 202 have both disagreement cells occupied. The declared grid crosses them with $n \in \{200, 500, 1000\}$ and $k \in \{5, 10\}$: 1212 cells, 1000 replicates each, with a noise floor of 0.926 for the worst cell and 0.045 for the spread. The median reuse share is 7.1%, falling from 15.7% at $n = 200$ to 1.3% at $n = 1000$, with maximum 49.2%: the correction is usually small and occasionally decisive, and the score alone cannot say which. RCV lifts the worst cell from 0.683 (fold-based) and 0.749 (binomial) to 0.827 at 6.5% more width than the binomial interval on average (Table 2).

Where a simpler interval wins. On the declared real grid RCV beats the width-matched control for the wrong reason: at the matched constant the inflated binomial interval has worst cell 0.757 against RCV's 0.827, because in 123 of the 1212 cells the score is exactly zero in over 1% of replicates and the binomial radius collapses. Against intervals that do not collapse RCV loses: Agresti–Coull beats it on every summary while only 1.5% wider, and Wilson on the worst cell and the mean while 2.5% narrower (both at their own centres). The loss is concentrated at $n = 200$, where the plug-in table is noisiest: by sample size RCV's worst cell is 0.827, 0.874 and 0.898 at $n = 200$, 500 and 1000, against Agresti–Coull's 0.868, 0.873 and 0.878. On four later rounds frozen before running, on eight unused public datasets with pools of 8,000 to 100,000 rows, RCV covered better than plain Agresti–Coull in three and worse in one (Table 3; details in Section S4). We read this as the boundary of the mechanism rather than a verdict between intervals: the correction tracks the reuse term where that term is large (Figure 2, left), but it inherits the sampling error of the estimated table, which is largest at the smallest n, where the real-pool loss is concentrated. What this paper claims is the reuse mechanism and its exact accounting, not a universally best interval.

Open-weight language-model checkpoints. Choosing between two released checkpoints on a benchmark is the setting of Section 1 in its modern form. Two rosters, each fixed before it ran and recorded in the deviation register, give 21 pairs of Qwen2.5, SmolLM2 and TinyLlama checkpoints (base against instruction-tuned, and across sizes and families), scored zero-shot on ARC-Challenge, HellaSwag and OpenBookQA; each pair's frozen losses on a benchmark define an exact law, 63 in all, crossed with $n \in \{500, 1000\}$ and $k \in \{10, 20\}$ into 252 cells of 400 replicates (Section S4). The reuse-aware interval's worst cell is 0.907 and its spread 0.073, against a noise floor of 0.917 and

Table 4: The reuse factor carried to Agresti–Coull (AC), worst cell / spread / mean at the common cross-validated centre. $f = 1/\sqrt{1-\hat{s}}$, unclipped here; the constant row widens AC by one factor matched to the mean width of AC$\times f$ (7.9% on the real grid, 12.8% on the synthetic one). AC$\times f$ was specified after frozen rounds 1 and 2, so on these two grids it is post hoc; rounds 3 and 4 (Table 3) tested it prospectively.

interval	declared real grid (1212 cells)	synthetic grid (340 cells)
Agresti–Coull	$0.876/0.124/0.938$	$0.808/0.151/0.916$
RCV (reuse-aware RMSE)	$0.827/0.142/0.929$	$0.895/0.075/0.938$
AC $\times$ constant matched to AC$\times f$	$0.891/0.109/0.953$	$0.857/0.128/0.948$
AC $\times f$ (reuse factor)	$\mathbf{0.907/0.093/0.953}$	$\mathbf{0.920/0.052/0.948}$

0.061: an exactly calibrated interval would have a worst cell this low with probability 0.07, and one as low as 0.877 with probability below 10^{-5}. The fold-based, binomial and bootstrap intervals reach 0.843, 0.877 and 0.875. Added after the fact on the same replicates, Agresti–Coull at the common centre gives 0.877 and 0.095 and covers less than the reuse-aware interval in 206 cells and more in none; a Wald or Agresti–Coull half width widened by a constant matched to the reuse-aware width reaches 0.892 and 0.085, with 19 and 18 cells below 0.925 against 6, though it ties the reuse-aware interval cell by cell (median paired difference 0). Every checkpoint cell has $n \geq 500$, where the plug-in table is less noisy than at the $n = 200$ that carries the classical loss.

Carrying the correction to a textbook interval. The two failures are separable: Agresti–Coull repairs discreteness and ignores reuse, and RCV does the reverse. Because the evaluator returns the reuse share, the same factor can be applied to another interval: scale its half width by $f = 1/\sqrt{1-\hat{s}} = \mathrm{RMSE}_{\mathrm{RCV}}/(\widehat{\mathbb{E}Z^2})^{1/2}$, the ratio of the reuse-aware to the plain root mean squared error, which widens the interval exactly when $\hat{s} > 0$. Applied to Agresti–Coull at the cross-validated centre, on both grids of Table 4 it has a better worst cell, spread and mean than Agresti–Coull and RCV, at 9.5% and 2.9% more width than RCV, and a better worst cell and spread than a width-matched constant; those grids were read before the form was chosen, so they motivate it and do not test it. The form was chosen after the first two frozen rounds, whose declared subject, the Wilson form, passed one and failed one; it was then declared the subject of rounds 3 and 4 before they ran and passed all five pre-specified conditions in both (Table 3), by 0.008 and 0.012 (unrounded) in worst-cell coverage over a width-matched constant, with single-cell Monte Carlo standard errors of about 0.008 and 0.004, and read after the fact it fails round 2. We therefore offer it as an exploratory extension, a post-hoc candidate that passed two sealed tests on four unused datasets, not as a validated method or a coverage guarantee.

What the reuse share does not do. The correction never changes which candidate is chosen, so the only way it could warn is through its size, and it does not. Over thirty-two synthetic configurations the reuse share reaches mean AUROC 0.627 for the event that the full-sample selection picked the higher-risk candidate, against 0.730 for the paired z-statistic a practitioner already has and 0.475 for raw disagreement. On the 26 cells of a separate abstention study, abstaining when the share exceeds 0.10, 0.20 or 0.30 discards 53%, 34% and 21% of comparisons and leaves binomial coverage 0.852, 0.892 and 0.914 on the kept samples: the flag drops samples that report large reuse and keeps those whose plug-in table understates it, so coverage falls as the threshold tightens. Widen the interval; do not screen on the share. The reuse term itself can be negative outside Proposition 2's condition: in 1.6% of 20,000 swept tie laws with $\beta_0 > \frac{1}{2}$, all with at most 2.26 expected disagreements.

Which number to report. Proposition 1 also compares the score with the winner's empirical risk. On the 202 real cases at $k = 10$ and $n = 1000$ the ratio $\mathrm{MSE}_{\mathrm{CV}}/\mathbb{E}Z^2$ is never below 1, with median 1.010 and maximum 1.848 (1.180 and 1.967 at $n = 200$); on 120 further fixed pairs from eight larger public tables, added post hoc, it is never below 1 in 1080 cases; and on 21 pairs of open-weight language-model checkpoints on three benchmarks, 504 exact evaluations, none is below one. On every two-candidate real law we evaluated, a practitioner who has not yet cross-validated the selection loses nothing in mean squared error by reporting the winner's empirical risk; one who has can recover it exactly as $\hat{r}_{\mathrm{CV}} - G/n$, and the reuse-aware RMSE is the error bar for the score they did report.

Refitted learners. The identity of Proposition 1 extends pathwise to learners refitted in every fold when each fold selects on the other folds' held-out losses (Section S3); the exact evaluator does not. In three studies frozen before running, against the error of the shipped model, nesting was less accurate than the flat score in 33 of 54 configurations with five or six candidates and more accurate in 21 of 36 with random searches of 200 in the second study and 11 of 18 in the third (Table 5): the advice not to nest holds for a handful of candidates, not for tuning searches. The reuse-aware RMSE is not claimed for this setting.

6 Limitations and Conclusion

The analysis is for two fixed candidates under 0–1 loss with equal random folds; a probe with three or six fixed candidates (archived technical report) found the cross-validated score more accurate than the winner's empirical risk in only 2 of 225 cells, and refitting on the same sample is outside the evaluator. The interval is calibrated empirically, not certified. Its main remaining weakness is that the 2×2 table is itself estimated: the worst real cells are small-n cells where the plug-in table is noisiest, and propagating that uncertainty is open. Near a tie the reuse term moves with the gap on the $1/\sqrt{n}$ scale at which the plug-in gap is itself uncertain, so there its relative error does not vanish as n grows, as for the bootstrap of other non-differentiable functionals (Andrews, 2000; Fang & Santos, 2019). The reuse share is not a detector of a wrongly chosen model.

Conclusion. Cross-validating a selection between two fixed models re-runs it on overlapping data, and the resulting reuse term depends on how the models err jointly, not only on their risks. The reuse-aware RMSE computes exactly the mean squared error the cross-validated score would have if the observed, floored paired table were the true law, at the given fold count. It reduces the under-coverage of the fold-based, binomial and bootstrap error bars and adapts to the amount of reuse rather than widening uniformly; it is beaten on the declared real grid by the Agresti–Coull interval, though it covers better than plain Agresti–Coull in three of four later frozen rounds and on the checkpoint pools; scaling Agresti–Coull by the reuse factor it returns, a form chosen after two frozen rounds, passed the two run after it was fixed, by small margins, and remains an exploratory extension without a coverage guarantee. It measures uncertainty without saying which model is better. In practice, with a few fixed candidates we would report the winner's empirical risk, with the union-bound interval of Section S3 when a guarantee is needed; a reported cross-validated score should carry its reuse-aware RMSE rather than a fold-based or binomial error bar, and the reuse share should not be used to screen; with refitted learners over large searches, nesting pays.

Reproducibility Statement

`REPRODUCIBILITY_ARTIFACT.zip` carries the LaTeX source, the standalone evaluator with its tests (`python3 tool/test_rcv.py`), every study script and per-cell result file, the frozen loss pools, the experiment protocol that was fixed and hashed before any run (the hash is self-recorded in the artifact, not externally timestamped) with a dated log of every deviation from it, and a script, `python3 code/check_final_numbers.py`, that finds every numeric literal in this paper and its supplement and fails unless each value is recomputed from the result files, listed as a design constant, tied by hand to the check that recomputes it, or exempted as a small integer, level or year; it matches values, not positions (Section S6). Secondary results that the claims here do not need, including a large-fold asymptotic law for the reuse term and its extensions, are kept in an archived technical report in the artifact and are not part of this submission.

Ethics Statement

This work uses only public benchmark datasets (scikit-learn's loaders; the UCI archives Adult, Covertype, MAGIC Gamma Telescope, Bank Marketing, Online Shoppers Purchasing Intention, Diabetes 130-US Hospitals, Occupancy Detection and Dota 2 Games Results; and the language-model benchmarks ARC-Challenge, HellaSwag and OpenBookQA scored with open-weight checkpoints) and synthetic data; it involves no new data collection, no identifiable personal data and no deployment. Its subject is the honesty of reported uncertainty for a model-selection procedure, and we report where the method does not help alongside where it does.

AI USE STATEMENT

Generative AI systems (large language models) were used extensively in this work: in formulating and refining the mathematical claims and their proofs, designing the method and the experimental protocol, writing the code, running and analysing the experiments, literature search, figure and table generation, and drafting and editing the text. The safeguards are these: the paper keeps only the results needed for its claims, and each retained proof has been checked step by step by the authors; every finite identity was additionally checked by exhaustive enumeration and against an independent reference implementation; a released script ties the numbers in the text to the result files (Section S6); the protocol was frozen before the runs and every deviation is logged; and negative results are reported rather than removed. No review independent of the author list has been obtained at submission. The authors take responsibility for the final content, including all text, claims and artifacts produced with the aid of generative AI.

REFERENCES

Alan Agresti and Brent A. Coull. Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, 52(2):119–126, 1998. doi: 10.1080/00031305.1998.10480550.

Donald W. K. Andrews. Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica*, 68(2):399–405, 2000. doi: 10.1111/1468-0262.00114.

Isaiah Andrews, Toru Kitagawa, and Adam McCloskey. Inference on winners. *The Quarterly Journal of Economics*, 139(1):305–358, 2024. doi: 10.1093/qje/qjad043.

Sylvain Arlot and Alain Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79, 2010.

Morgane Austern and Wenda Zhou. Asymptotics of cross-validation. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 61(4):2804–2865, 2025. doi: 10.1214/24-AIHP1488.

Stephen Bates, Trevor Hastie, and Robert Tibshirani. Cross-validation: What does it estimate and how well does it do it? *Journal of the American Statistical Association*, 119(546):1434–1445, 2024. doi: 10.1080/01621459.2023.2197686.

Alexandre Bayle, Lucas Janson, and Lester Mackey. The relative instability of model comparison with cross-validation. In *International Conference on Machine Learning (ICML)*, 2026. arXiv:2508.04409.

Pierre Bayle, Alexandre Bayle, Lucas Janson, and Lester W. Mackey. Cross-validation confidence intervals for test error. In *Advances in Neural Information Processing Systems*, volume 33, pp. 16339–16350, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/bce9abf229ffd7e570818476ee5d7dde-Abstract.html.

Yoshua Bengio and Yves Grandvalet. No unbiased estimator of the variance of K-fold cross-validation. *Journal of Machine Learning Research*, 5:1089–1105, September 2004. URL https://www.jmlr.org/papers/v5/grandvalet04a.html.

Yoav Benjamini and Daniel Yekutieli. False discovery rate-adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association*, 100(469):71–81, 2005. doi: 10.1198/016214504000001907.

Avrim Blum and Moritz Hardt. The Ladder: A reliable leaderboard for machine learning competitions. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 1006–1014, 2015.

Leo Breiman. Heuristics of instability and stabilization in model selection. *The Annals of Statistics*, 24(6):2350–2383, 1996. doi: 10.1214/aos/1032181158.

Lawrence D. Brown, T. Tony Cai, and Anirban DasGupta. Interval estimation for a binomial proportion. *Statistical Science*, 16(2):101–133, 2001. doi: 10.1214/ss/1009213286.

Gavin C. Cawley and Nicola L. C. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11:2079–2107, 2010. URL https://www.jmlr.org/papers/v11/cawley10a.html.

A. C. Davison and Peter Hall. On the bias and variability of bootstrap and cross-validation estimates of error rate in discrimination problems. *Biometrika*, 79(2):279–284, 1992. doi: 10.1093/biomet/79.2.279.

Thomas G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1923, 1998. doi: 10.1162/089976698300017197. URL https://doi.org/10.1162/089976698300017197.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638, 2015.

Zheng Fang and Andres Santos. Inference on directionally differentiable functions. *The Review of Economic Studies*, 86(1):377–412, 2019. doi: 10.1093/restud/rdy049.

Jason C. Hsu. Constrained simultaneous confidence intervals for multiple comparisons with the best. *The Annals of Statistics*, 12(3):1136–1144, 1984. doi: 10.1214/aos/1176346732.

Garud Iyengar, Henry Lam, and Tianyu Wang. Is cross-validation the gold standard to estimate out-of-sample model performance? In *Advances in Neural Information Processing Systems*, volume 37, pp. 94736–94775, 2024. doi: 10.52202/079017-3003. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ac4106bcfff33140de7799d03daeb8a4-Abstract-Conference.html.

Nicholas Kissel and Jing Lei. Black-box model confidence sets using cross-validation with high-dimensional Gaussian comparison, 2023. URL https://arxiv.org/abs/2211.04958v2. Version 2, revised 13 November 2023.

Damjan Krstajic, Ljubomir J. Buturovic, David E. Leahy, and Simon Thomas. Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of Cheminformatics*, 6:10, 2014.

John Langford. Tutorial on practical prediction theory for classification. *Journal of Machine Learning Research*, 6:273–306, 2005.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947. doi: 10.1007/BF02295996. URL https://doi.org/10.1007/BF02295996.

Evan Miller. Adding error bars to evals: A statistical approach to language model evaluations, 2024. URL https://arxiv.org/abs/2411.00640.

Ido Nachum, Rüdiger Urbanke, and Thomas Weinberger. The structure of cross-validation error: Stability, covariance, and minimax limits, 2025. URL https://arxiv.org/abs/2511.03554. Version 2, revised 8 January 2026.

Claude Nadeau and Yoshua Bengio. Inference for the generalization error. *Machine Learning*, 52(3):239–281, 2003. doi: 10.1023/A:1024068626366. URL https://doi.org/10.1023/A:1024068626366.

Benjamin I. P. Rubinstein and Aleksandr Simma. On the stability of empirical risk minimization in the presence of multiple risk minimizers. *IEEE Transactions on Information Theory*, 58(7):4160–4163, 2012. doi: 10.1109/TIT.2012.2191681.

Ryan J. Tibshirani and Robert Tibshirani. A bias correction for the minimum error rate in cross-validation. *The Annals of Applied Statistics*, 3(2):822–829, 2009. doi: 10.1214/08-AOAS224.

Ioannis Tsamardinos, Elissavet Greasidou, and Giorgos Borboudakis. Bootstrapping the out-of-sample predictions for efficient and accurate cross-validation. *Machine Learning*, 107(12):1895–1922, 2018. doi: 10.1007/s10994-018-5714-4.

Sudhir Varma and Richard Simon. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7:91, 2006. doi: 10.1186/1471-2105-7-91.

Gaël Varoquaux. Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage*, 180:68–77, 2018.

Jacques Wainer and Gavin Cawley. Nested cross-validation when selecting classifiers is overzealous for most practical applications. *Expert Systems with Applications*, 182:115222, 2021. doi: 10.1016/j.eswa.2021.115222.

Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927. doi: 10.1080/01621459.1927.10502953.

Tijana Zrnic and William Fithian. Locally simultaneous inference. *The Annals of Statistics*, 52(3): 1227–1253, 2024. doi: 10.1214/24-AOS2391.

S1 PROOFS OF PROPOSITIONS 1 AND 2

Write $\ell_{\pm,i}$ for the two losses of observation i. Since $\ell_+ = B - X/2$ and $\ell_- = B + X/2$, a candidate chosen by the sign $s \in \{+1, -1\}$ ($s = +1$ for h_+) has losses $B_i - sX_i/2$.

Pathwise identity. The full sample selects $s = \tau(T)$, so $\widehat{R}_{\text{plain}} = \bar{B} - \tau(T)T/(2n) = \bar{B} - |T|/(2n)$, using $\tau(t)t = |t|$ for every integer t (including $t = 0$, where $\tau = -1$). Fold j is scored by the candidate its complement selects, $\tau(T - A_j)$, so $\hat{r}_{\text{CV}} = \bar{B} - (2n)^{-1} \sum_j \tau(T - A_j)A_j$. Since $\sum_j A_j = T$,

$$\hat{r}_{\text{CV}} - \widehat{R}_{\text{plain}} = -\frac{1}{2n} \sum_j A_j \big\{\tau(T - A_j) - \tau(T)\big\}.$$

Only folds with $\tau(T - A_j) \neq \tau(T)$ contribute. If $T > 0$, such a fold has $T - A_j \leq 0$, so $A_j \geq T > 0$ and its term is $-2A_j = -2|A_j|$. If $T \leq 0$, it has $T - A_j > 0$, so $A_j < T \leq 0$ and its term is $2A_j = -2|A_j|$. In both cases the flipped fold has $|A_j| \geq |T|$, and the sum equals $-2G$, which proves $\hat{r}_{\text{CV}} = \widehat{R}_{\text{plain}} + G/n$ and $G \geq |T|$ when $G > 0$.

Decomposition (for $p_{00} + p_{11} > 0$). The shipped model's risk is $R_P(\hat{h}_n) = \beta - \tau(T)\mu/2$, so $Z = (\bar{B} - \beta) - |T|/(2n) + \tau(T)\mu/2$. Expanding $\mathbb{E}(Z + G/n)^2 - \mathbb{E}Z^2$,

$$n^2 \Delta_{\text{reuse}} = 2n\, \mathbb{E}[(\bar{B} - \beta)G] + \mathbb{E}\big[G^2 - |T|G\big] + n\mu\, \mathbb{E}[\tau(T)G].$$

G is a function of the contrasts (X_i) alone. Given them, $B_i = \frac{1}{2}$ on every disagreement and $B_i \in \{0, 1\}$ is Bernoulli(β_0), independently, on every agreement. Hence $\mathbb{E}[\bar{B} \mid X] = \beta_0 + (\frac{1}{2} - \beta_0)N_D/n$, $\beta = \beta_0 + (\frac{1}{2} - \beta_0)\mathbb{E}N_D/n$, and

$$\mathbb{E}[(\bar{B} - \beta)G] = \mathbb{E}\big[G\,(\mathbb{E}[\bar{B} \mid X] - \beta)\big] = \frac{\frac{1}{2} - \beta_0}{n} \operatorname{Cov}(G, N_D).$$

Substituting gives the display of Proposition 1. Finally $G(G - |T|) \geq 0$ because $G = 0$ or $G \geq |T|$.

Sign at a tie. Let $p_{10} = p_{01}$, so $\mu = 0$ and the last term vanishes. Condition on the set $\mathcal{D}$ of disagreeing positions. Given $\mathcal{D}$ the nonzero X_i are independent fair signs, so for each fold A_j is independent of $T - A_j$ and $\mathbb{E}[A_j \mid T - A_j, \mathcal{D}] = 0$. Therefore $\mathbb{E}[\sum_j A_j \tau(T - A_j) \mid \mathcal{D}] = 0$, while $\mathbb{E}[\tau(T)T \mid \mathcal{D}] = \mathbb{E}[|T| \mid \mathcal{D}]$. Taking the conditional expectation of the identity $-2G = \sum_j A_j \tau(T - A_j) - \tau(T)T$ gives $\mathbb{E}[G \mid \mathcal{D}] = \frac{1}{2}\mathbb{E}[|T| \mid \mathcal{D}]$. Since N_D is a function of $\mathcal{D}$, $\mathbb{E}G = \frac{1}{2}\mathbb{E}|T|$ and $\mathbb{E}(GN_D) = \frac{1}{2}\mathbb{E}(|T|N_D)$, so $\operatorname{Cov}(G, N_D) = \frac{1}{2}\operatorname{Cov}(|T|, N_D)$, and $(1 - 2\beta_0)\operatorname{Cov}(G, N_D) = (\frac{1}{2} - \beta_0)\operatorname{Cov}(|T|, N_D)$. Given $N_D = d$, T is a d-step simple symmetric random walk S_d, and $d \mapsto \mathbb{E}|S_d|$ is nondecreasing ($\mathbb{E}[|S_d + \xi| \mid S_d] \geq |S_d|$ by Jensen's inequality for an independent fair sign ξ), and $\operatorname{Cov}(|T|, N_D) = \operatorname{Cov}(\phi(N_D), N_D)$ with $\phi(d) = \mathbb{E}|S_d|$. A nondecreasing function of N_D has non-negative covariance with N_D, so $\operatorname{Cov}(|T|, N_D) \geq 0$. With $\beta_0 \leq \frac{1}{2}$, which is $p_{11} \leq p_{00}$, both terms are non-negative. □

Checks. Both identities were verified by exhaustive enumeration of all 4^n samples at small n (agreement to 10^{-14} in mean squared error), including off-tie laws and a law with a negative reuse term. The tie condition is not cosmetic: a sweep found 321 negative tie laws with $\beta_0 > \frac{1}{2}$ among 20,000 and none among 55,000 with $\beta_0 \leq \frac{1}{2}$.

S2 THE EXACT EVALUATOR (PROPOSITION 3)

Let $D = \ell_- + \ell_+ = 2B$, $\bar{d} = \mathbb{E}D$, $d_2 = \mathbb{E}D^2$, $x_2 = \mathbb{E}X^2$, and keep $\mu = \mathbb{E}X$ and $\tau(0) = -1$. For N observations put $T_N = \sum_{i \leq N} X_i$, $D_N = \sum_{i \leq N} D_i$,

$$\pi_N(t) = P(T_N = t), \quad \lambda_N(t) = \mathbb{E}[D_N \mathbf{1}\{T_N = t\}], \quad b_N = \sum_t \tau(t)\pi_N(t),$$

$$u_N = \sum_t \tau(t)\lambda_N(t), \qquad f_m(s) = \sum_r r\, \pi_m(r)\tau(r + s),$$

with $\pi_0(t) = \mathbf{1}\{t = 0\}$ and $\lambda_0 = 0$. With $v = n - m$ and $w = n - 2m$, define

$$g_1 = n\mu b_v, \qquad g_2 = k\{mx_2 + m(m-1)\mu^2\} + k(k-1)\sum_t \pi_w(t) f_m(t)^2,$$

$$e_{DG} = k\big[\{m\mu + m(m-1)\mu\bar{d}\}b_v + m\mu u_v\big], \qquad e_{D^2} = nd_2 + n(n-1)\bar{d}^2,$$

$$q_1 = n\bar{d} - g_1, \qquad q_2 = e_{D^2} - 2e_{DG} + g_2, \qquad q_\tau = u_n - k\sum_t \pi_v(t)\tau(t) f_m(t),$$

and $\mathcal{M}(y_1, y_2, y_\tau) = \frac{y_2}{4n^2} - \frac{\bar{d}y_1 - \mu y_\tau}{2n} + \frac{\bar{d}^2 + \mu^2 - 2\bar{d}\mu b_n}{4}$. Then

$$\mathrm{MSE}_{\mathrm{CV}} = \mathcal{M}(q_1, q_2, q_\tau),$$

$$\mathbb{E}Z^2 = \mathcal{M}\Big(n\bar{d} - \sum_t |t|\pi_n(t),\ e_{D^2} + nx_2 + n(n-1)\mu^2 - 2\sum_t |t|\lambda_n(t),\ u_n - n\mu\Big).$$

Proof. Put $\Gamma = \sum_j A_j \tau(T - A_j)$ and $Q = D_n - \Gamma$, so $\hat{r}_{\mathrm{CV}} = Q/(2n)$ by the first display of Section S1. The target is $R_P(\hat{h}_n) = (\bar{d} - \mu\tau(T))/2$; expanding $\mathbb{E}\big(Q/(2n) - (\bar{d} - \mu\tau(T))/2\big)^2$ with $\tau^2 = 1$ gives $\mathcal{M}(\mathbb{E}Q, \mathbb{E}Q^2, \mathbb{E}[Q\tau(T)])$. A fold is independent of its complement, so $\mathbb{E}\Gamma = \sum_j \mathbb{E}A_j\, \mathbb{E}\tau(T - A_j) = g_1$ and $\mathbb{E}[D_n\Gamma] = k\{\mathbb{E}[D_m A_1]b_v + \mathbb{E}[A_1]u_v\} = e_{DG}$, using $DX = X$ (so $\mathbb{E}[D_m A_1] = m\mu + m(m-1)\mu\bar{d}$) and $\mathbb{E}A_1 = m\mu$. For two different folds, condition on the rest W of the sample: A_1 and A_2 are then independent, and $\mathbb{E}[A_1\tau(A_2 + W)A_2\tau(A_1 + W) \mid W] = \mathbb{E}[A_1\tau(A_1 + W) \mid W]\,\mathbb{E}[A_2\tau(A_2 + W) \mid W] = f_m(W)^2$; with the diagonal $\mathbb{E}A_1^2 = mx_2 + m(m-1)\mu^2$ this gives $\mathbb{E}\Gamma^2 = g_2$, and $\mathbb{E}Q^2 = q_2$. Conditioning on one complement gives $\mathbb{E}[\Gamma\tau(T)] = k\sum_t \pi_v(t)\tau(t)f_m(t)$, so $\mathbb{E}[Q\tau(T)] = q_\tau$. For the plain estimate replace Q by $Q_R = D_n - |T|$; its moments use $\mathbb{E}[D_n|T|] = \sum_t |t|\lambda_n(t)$, $\mathbb{E}T^2 = nx_2 + n(n-1)\mu^2$ and $Q_R\tau(T) = D_n\tau(T) - T$.

For the operation count let $g(z) = p_{01} + (p_{00} + p_{11})z + p_{10}z^2$ and $h(z) = p_{01} + 2p_{11}z + p_{10}z^2$; then $\pi_N(t) = [z^{N+t}]g(z)^N$ and $\lambda_N(t) = N[z^{N+t}]h(z)g(z)^{N-1}$. If $a_i = [z^i](c_0 + c_1 z + c_2 z^2)^N$, differentiating gives $(i+1)c_0\, a_{i+1} = (N-i)c_1\, a_i + (2N - i + 1)c_2\, a_{i-1}$, which yields all $2N + 1$ coefficients in $O(N)$ operations when $c_0 > 0$ (reverse the polynomial when $c_0 = 0 < c_2$), in exact arithmetic. Applied to g^N it gives π_N; applied to g^{N-1} and followed by one multiplication by the degree-two polynomial h, a three-term convolution of $2N - 1$ coefficients, it gives λ_N, again in $O(N)$. Both released implementations instead obtain the contrast distributions by repeated convolution. Prefix sums give every $f_m(t) = m\mu - 2\sum_{r \le -t} r\pi_m(r)$ in constant time, and every remaining sum has $O(n)$ terms. The remaining case $c_0 = c_2 = 0$ is the law with no disagreement, $p_{01} = p_{10} = 0$: then $g(z) = z$, so $\pi_N(t) = \mathbf{1}\{t = 0\}$ and $\lambda_N(t) = N\bar{d}\,\mathbf{1}\{t = 0\}$ directly, $X \equiv 0$, no fold can flip, $G \equiv 0$ and $\mathrm{MSE}_{\mathrm{CV}} = \mathbb{E}Z^2$. The released evaluator assumes $p_{01}, p_{10} > 0$, which the floor in step 1 of the method guarantees. □

Verification. An independent reference implementation, which selects a candidate on each fold complement and scores it directly without the reduction, agrees with it in 728 rational designs over 91 distinct laws; a literal enumeration covers 25 designs and 64,000 observation sequences. Two floating-point implementations are released. The one behind the real-pair ratios of Section 5 uses a pairwise fold-sum decomposition that prunes contrast atoms below 10^{-17}; over 808 real-pool cells (law and fold count) its largest disagreement with exact rational arithmetic is 5.80×10^{-17} in MSE (6.62×10^{-13} relative). The one inside the released tool keeps every atom and agrees with exact rational arithmetic to 1.3×10^{-14} in relative MSE on 18 anchor laws. Both are deterministic and accurate to these tolerances rather than literally exact. Without the floor, the evaluator at the observed table is the ideal paired-loss bootstrap (resample the loss pairs, re-run the fold selection) of the select-and-score pipeline for the two fixed candidates (Section 3); no bootstrap statement is made for refitted learners; a Monte Carlo bootstrap with 2,000,000 resamples agrees with it on four observed tables to within 1.9 Monte Carlo standard errors. These checks corroborate the finite identities; they do not replace the proof.

S3 A GUARANTEE FOR THE SELECTED MODEL, AND REFITTING

Proposition 4 (Union-bound interval for the selected model)**.** *Let $h_1, \ldots, h_d$ be fixed before the sample is drawn, let C_j be the two-sided Clopper–Pearson interval at level $1 - \alpha/d$ from candidate*

j's error count on the whole sample, and let $\hat{h}_n$ be any candidate chosen from the sample. Then $\mathbb{P}\big(R_P(\hat{h}_n) \in C_{\hat{h}_n}\big) \geq 1 - \alpha$.

Proof. Each error count is Binomial$(n, R_P(h_j))$ because h_j is fixed, so $\mathbb{P}(R_P(h_j) \notin C_j) \leq \alpha/d$; by the union bound all d intervals contain their risks with probability at least $1 - \alpha$, and on that event so does the interval of whichever candidate is chosen. □

Over the 606 cells of the real grid without folds, computed exactly (not simulated) by a forward recursion over the joint error counts, this interval's worst coverage is 0.975 at 24% more half width than the binomial interval. On the same cells the Wald interval at the winner's empirical risk falls to 0.717 and the Clopper–Pearson interval at level $1 - \alpha$, which is not certified here, reads 0.951. The interval $C_{\hat{h}_n}$ is computed from the winner's full-sample error count and is reported unchanged alongside $\hat{r}_{\mathrm{CV}}$; translating its endpoints by the observed G/n would not preserve its coverage of $R_P(\hat{h}_n)$.

Refitting. Let $\mathcal{A}_1, \ldots, \mathcal{A}_d$ be learning algorithms, $\ell \in [0, 1]$ a loss, and $F_1, \ldots, F_k$ a partition chosen without the labels. Let $e_j(a)$ be the summed held-out loss over F_j of $\mathcal{A}_a$ trained on the complement, $L(a) = \sum_j e_j(a)$, the flat winner $\hat{a} = \arg\min_a L(a)$ and the fold-j winner $\hat{a}_j = \arg\min_a [L(a) - e_j(a)]$, ties broken by index. The flat score is $L(\hat{a})/n$ and the leave-fold-out nested score is $n^{-1} \sum_j e_j(\hat{a}_j)$.

Proposition 5 (Flip identity under refitting)**.** *On every sample, $n^{-1} \sum_j e_j(\hat{a}_j) - L(\hat{a})/n = G/n$ with $G = \sum_j [e_j(\hat{a}_j) - e_j(\hat{a})]$, and each summand is zero when $\hat{a}_j = \hat{a}$ and at least $L(\hat{a}_j) - L(\hat{a}) \geq 0$, and positive, when $\hat{a}_j \neq \hat{a}$. For two fixed candidates this G is the G of Proposition 1.*

Proof. If $\hat{a}_j \neq \hat{a}$, then $L(\hat{a}) \leq L(\hat{a}_j)$ and $L(\hat{a}_j) - e_j(\hat{a}_j) \leq L(\hat{a}) - e_j(\hat{a})$, not both with equality, since equality in both would make the index tie-break choose the same candidate in both. Rearranging the second gives $e_j(\hat{a}_j) - e_j(\hat{a}) \geq L(\hat{a}_j) - L(\hat{a}) \geq 0$, strictly if either inequality is strict. Summing over j and using $\sum_j e_j(\hat{a}) = L(\hat{a})$ gives the identity. For fixed candidates with h_- first, $e_j(-) - e_j(+) = A_j$ and $L(-) - L(+) = T$, and the flipped folds contribute $|A_j|$ as in Section S1. □

The identity does not cover the inner-loop nested cross-validation that practitioners run, and the exact evaluator does not apply once learners are refitted; those settings are measured, not proved (Section S4).

S4 Additional experimental details

Grids. The synthetic protocol declares 900 combinations of risk gap, disagreement, risk level, n and k; its feasibility rule (all four cells strictly positive) excludes 18 laws, which account for 216 of them, and all 684 remaining were run; Table 1 is the 340-cell subset run first, cut for runtime by a logged deviation made before any cell was inspected. The real grid (Table 2) uses a stratified 40% training split of each dataset, leaving pools of 1079 (digits), 107 (wine) and 342 (breast cancer) points, sampled with replacement.

Checkpoint pools. Zero-shot multiple choice, each option scored by its length-normalised log-likelihood as a continuation of the question, argmax over options, ties to the first: the ARC-Challenge test set (1172 items), the first 2000 HellaSwag validation items and the OpenBookQA test set (500 items), one frozen 0–1 loss vector per checkpoint and benchmark. The first roster is four Qwen2.5 checkpoints (two sizes, base and instruction-tuned); four Llama-3.2 checkpoints declared with it failed to load and were not replaced. The second is six SmolLM2 and TinyLlama checkpoints (base and instruction- or chat-tuned). Pairs are formed within a roster, 6 and 15 per benchmark, and samples are drawn with replacement from each pool's joint law.

Four later frozen rounds. Each round froze its protocol, code and pass rule before drawing a replicate and used two public datasets not used before, with a six-model roster and pools of 8,000 to 100,000 rows; 180 cells each. Table 3 in Section 5 gives worst cell / spread at the common cross-validated centre; the declared subject passes only if, on the pooled grid, it beats the

width-matched constant on both the worst cell and the spread, its mean coverage lies in $[0.94, 0.96]$, its mean absolute coverage deviation is no larger than Agresti–Coull's, and on neither dataset is its worst cell more than 0.01 below Agresti–Coull's.

Refitted learners. Three studies, each frozen before its first replicate, compared inner-loop nested cross-validation with the flat cross-validated score on six UCI tables, $n \in \{200, 500, 1000\}$, five outer and five inner folds, against the error of the model refitted on the whole sample, with labels from paired bootstrap 95% intervals of the MSE ratio (Table 5). The third study also tested, as a declared endpoint, a rule found after the first two: the average of the flat and nested scores stayed within 10% of the better of the two in 48 of 54 configurations (threshold 44; post hoc on the first two studies, 147 of 162).

Table 5: Nested against flat cross-validation with refitted learners: configurations where nesting was less accurate / indistinguishable / more accurate.

study	candidates	worse / n.s. / better
1	five families, or a six-value grid of one learner	$33/17/4$
2	random searches of 10 / 50 / 200	$7/23/6$, $2/19/15$, $0/15/21$
3	new family, searches of 10 / 50 / 200	$6/9/3$, $3/9/6$, $0/7/11$

S5 The reuse ratio over the table

Figure 1 (Section 4) evaluates $\mathrm{MSE}_{\mathrm{CV}}\,/\mathbb{E}Z^2$ exactly at $k = 10$ over the paired-loss table, with the real and checkpoint laws overlaid. At $\beta_0 = 0.10$, where Proposition 2's condition $p_{11} \le p_{00}$ holds (its sign guarantee is exact only at a tie), the ratio is at least one on every one of the 1096 grid cells at $n = 200$ and 1544 at $n = 1000$, largest (1.77 and 1.74) at a tie with large disagreement. At $\beta_0 = 0.99$, where the condition fails, 19 cells have a negative reuse term, all with at most 1.37 expected disagreements, and the lowest ratio is 0.960 (both on the plotted $n = 200$ grid).

S6 Reproducibility and the deviation register

The protocol for the synthetic grid, the real grid and the width-matched control was fixed and hashed before any run; each later study (the four frozen rounds and the three refitting studies) froze its own protocol and code the same way, with the hash written to a freeze file before the first replicate; these hashes are self-recorded, not externally timestamped. A dated register in the artifact logs every deviation from a frozen protocol, what it changed and why, including the failed second round and the defects found in our own code. `code/check_final_numbers.py` reruns the claim checks of the archived full-length version against the result files, recomputes each number added or reworded for this version, ties each number in a reworded sentence to the check that recomputes it, exempts listed design constants and small integers, levels and years, and fails on any numeric literal in none of these classes and on any check whose recomputed value differs from the printed one. It matches values rather than positions, so a wrong number equal to a correct one printed elsewhere would pass; it guards against drift between the text and the results and does not show that the results are right. The archived technical report in the artifact keeps the secondary results removed from this submission.