RECORDED SAMPLE MEMBERSHIP IN AGGREGATE-BASED PREDICTION ADJUSTMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Aggregate labels, the supervision used in learning from label proportions, can also correct a model's predictions after training. When a sample is altered by recorded replacements, knowing which instances were kept can rule out predictions that class counts alone allow. We study a setting in which a uniform sample from a group with known class totals is modified through recorded replacements. The central intervention acquires no new individual labels: it uses the retained collection record more fully. For one original subset, we derive an efficient compatibility test that determines whether the current prediction remains consistent with the full sampling record. When incompatibility is detected, the violated restrictions yield short corrections with computable lower bounds on average squared-loss improvement. We compare these corrections with full joint adjustment and with predictions already corrected using count-based aggregate bounds. A prespecified evaluation on 256 previously unused Covertype groups shows that membership-specific constraints can change predictions across several selection regimes. In the primary setting such corrections are rare, but they act on poorly predicted outputs, and short corrections recover most of the full adjustment's gain. A separate matched intervention shows that correcting an earlier prediction can discard improvement already present in the available incumbent. Together, the results give a cheap, certified way to tell when a sampling record justifies a further correction, with a guaranteed minimum improvement.

1 INTRODUCTION

An aggregate label gives class totals for a group without identifying each instance's label. These totals can be used to train a predictor through learning from label proportions (LLP) (Scott & Zhang, 2020; Busa-Fekete et al., 2025; Ma et al., 2026b), or to adjust its class-probability predictions after training (Heiser et al., 2020). We study the second use. A subset is sampled uniformly from a larger group. Some of its instances are then replaced by previously omitted members of that group, for example because the originally selected units could not be measured. Predictions are evaluated on the final subset. The collection record identifies which instances were retained from the original draw. Count summaries, such as intervals for how many instances of each class the original draw contained, do not keep this information. Reducing the record to such summaries can therefore discard restrictions on the possible labels of the retained instances.

Figure 1 isolates this distinction. Both histories have the same class totals, final scored plots and initial predictions, but different original sample memberships. In the first history, the prediction could be correct. In the second, the record rules out one predicted label, which justifies a correction.

We ask when recorded membership justifies an additional correction and how often that correction helps. In the central comparison, we obtain no new individual labels. Instead, we use the collection record more fully and compare against predictions already corrected using aggregate information. Established projection geometry supplies the loss guarantee (Heiser et al., 2020; Krijthe & Loog, 2017). What is new is to characterise the targets that remain admissible after recorded replacements and to turn incompatibility into certified corrections.

Our contributions are threefold. (i) For one original sample within a parent group, we characterise compatibility with the full collection record through a test costing $O(mK)$ operations for m sampled instances and K classes, and we identify restrictions that the tested count summaries omit

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.

(Section 3). (ii) When the current prediction violates such a restriction, the violation yields a short correction. Each computed output carries a certified lower bound on its average squared-loss improvement (Section 4). (iii) Matched experiments separate the effect of retained information from that of extra computation. A prespecified confirmation on 256 previously unused Covertype groups shows that such corrections recur on new data. They are rare in the primary setting (3 of 256 groups), and where they occur they reduce the loss of poorly predicted outputs by about 9% (Section 6). A separate comparison keeps the acquired answers fixed and changes only the prediction that the correction is centred on. Supporting spatial studies examine what happens when aggregates are applied to the wrong scored group.

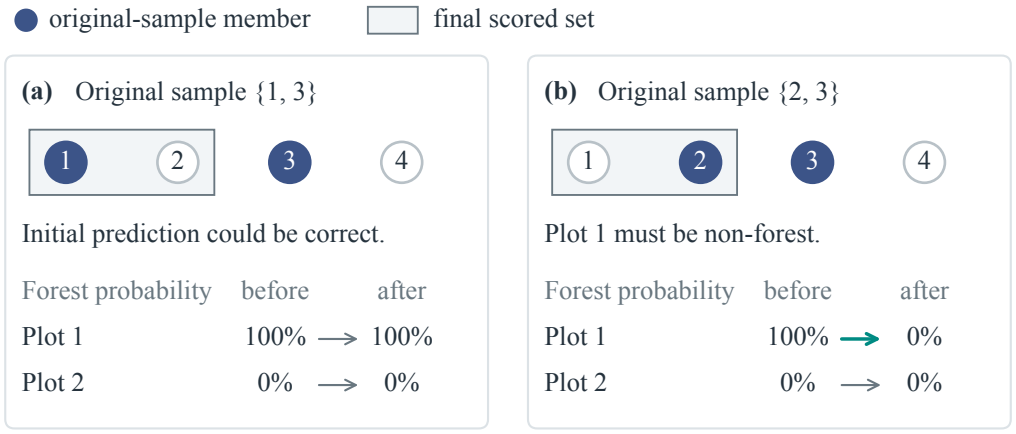


Figure 1: Same totals, different sample memberships. The parent group has four plots, one of them forest, and the original uniform sample of two plots contains that forest plot. Both histories replace one original-sample member, score plots 1 and 2, and start from the same forest probabilities. In (a), plot 1 may be the forest plot, so the prediction may be correct and no improvement is guaranteed. In (b), the forest plot lies in $\{2, 3\}$, so plot 1 must be non-forest. Setting its forest probability from 100% to 0% then gives a mean two-class squared-loss gain of 1 for every compatible target, while the label of plot 2 remains unresolved. The experiments use sampling intervals rather than this exact count.

2 Problem Setup and Relation to Prior Work

We score a fixed set of $m \geq 1$ instances. Each prediction B_i is a K-class probability vector in $\Delta_K = \{p \in \mathbb{R}^K : p \geq 0, \mathbf{1}^\top p = 1\}$, the probability simplex. Targets Y_i are one-hot class vectors in the central study and fractional compositions in the spatial studies. We measure average prediction error by

$$R(B, Y) = m^{-1} \sum_i \|B_i - Y_i\|_2^2. \tag{1}$$

The loss sums squared errors across classes and averages them across instances, without dividing by K. It is the multiclass Brier score for one-hot targets and squared composition error for fractional targets. It measures average individual prediction error, not the error of the group mean or classification accuracy.

The central comparison is $F \longrightarrow B \longrightarrow G$. Starting from the frozen predictions F, an existing correction uses replacement-aware bounds on each class count to produce B. We call this already available prediction the incumbent. A further adjustment uses recorded membership to produce G. We measure only its additional gain, $R(B, Y) - R(G, Y)$, on the same scored instances.

Let $\mathcal{C}$ contain targets allowed by the stated constraints. A direct gain certificate lower-bounds $R(B, Y) - R(G, Y)$ for every $Y \in \mathcal{C}$, not just the reference targets used to measure realised gain. Thus, it describes what the information guarantees rather than what the evaluator observes. Define the prediction domain $\mathcal{D} = \{X \in \mathbb{R}^{m \times K} : X_i \in \Delta_K \text{ for every } i\}$. For nonempty compact convex $\mathcal{C} \subseteq \mathcal{D}$, let $\Pi_{\mathcal{C}} B$ be its closest point to B in squared Frobenius distance. Classical projection gives,

for every $Y \in \mathcal{C}$,

$$R(B, Y) - R(\Pi_{\mathcal{C}} B, Y) \geq m^{-1} \|B - \Pi_{\mathcal{C}} B\|_F^2. \tag{2}$$

When $\mathcal{C}$ is the convex hull of admissible targets, this bound equals the best improvement guaranteed uniformly over that set:

$$V_{\mathcal{C}}(B) := \max_{G \in \mathcal{D}} \min_{Y \in \mathcal{C}} \{R(B, Y) - R(G, Y)\} = m^{-1} \|B - \Pi_{\mathcal{C}} B\|_F^2. \tag{3}$$

Consequently, $B \in \mathcal{C}$ rules out a strictly positive improvement for every admissible target, even when B is inaccurate. This conclusion applies only to the stated constraints, so additional justified information could change it. Equation 3 describes an ideal optimum, not a finite solver output (Appendix A.4).

Prediction adjustment. Heiser et al. (2020) establish adjustment guarantees from known class proportions and study proportion error and feature-dependent deletion. Safe projection, scoring-rule decompositions and missing-data bounds provide related foundations (Krijthe & Loog, 2017; Kull & Flach, 2015; Manski, 2005). Multicalibration also reduces squared error through group-discrepancy updates (Hébert-Johnson et al., 2018). These results explain how supplied restrictions support prediction changes.

Sampling and bag construction. MixBag uses a confidence-interval loss for augmented bags (Asanomi et al., 2023). Kubo et al. (2024) use a multivariate hypergeometric model to perturb sampled mini-bag proportions during training. Dependence-aware model selection and LLP benchmarks examine effects of bag construction (Franco et al., 2023; Brahmbhatt et al., 2024). Our predictor is instead fixed, and the original sample undergoes recorded replacements. We ask what its overlap with the final sample adds to adjustment.

Learning and adjacent constraints. Other LLP work develops learning-rate guarantees (Li et al., 2024; Busa-Fekete et al., 2025; Applebaum et al., 2026), training and pseudo-label objectives (Ma et al., 2026b; Chen et al., 2025; Ma et al., 2026a), and proportion-based adaptation (Rakotomamonjy et al., 2026). Javanmard et al. (2024) compare aggregate-response training losses. Our PVC adaptation supplies predictions, not a new training objective. Projection networks and forecast reconciliation address adjacent constrained-prediction problems (Grontas et al., 2026; Utkarsh et al., 2026; Tsiourvas et al., 2024).

3 Recorded membership and compatibility

The parent group contains N instances with known integer class totals t. Draw S_0 of size m uniformly without replacement, then replace r members with r previously omitted instances to obtain the scored set S. The overlap $I = S \cap S_0$ identifies retained original members. The hypergeometric sampling law and N, t, m give simultaneous intervals $[a, b]$ for the original sample's class counts. These are statistical bounds, not measured individual labels or exact subset totals. Adjustment uses counts, memberships and predictions. Individual reference labels are reserved for evaluation.

For candidate scored targets X, define $v(X) = \sum_{i \in S} X_i$ and $h(X) = \sum_{i \in I} X_i$. If removed instances contribute counts w, the original sample contributes $h(X) + w$, which must lie in $[a, b]$. Counts $t - v(X) - w$ must remain nonnegative for instances outside the union. Compatibility means that such a completion exists:

$$\mathcal{C}_{\text{joint}} = \{X \in \Delta_K^m : \exists w \geq 0,\ \mathbf{1}^\top w = r,\ t - v(X) - w \geq 0,\ a \leq h(X) + w \leq b\}. \tag{4}$$

Whenever the original sample's true counts lie in $[a, b]$, the true targets on the final scored set belong to $\mathcal{C}_{\text{joint}}$ after the recorded replacements. The original uniform draw gives this event at least 95% marginal coverage. The hard-label convex-hull interpretation below relies on integer endpoints, equal instance weights and one original subset. It does not extend here to arbitrary crossing groups or weighted spatial footprints.

Proposition 1 (Compatibility with recorded membership)**.** *Suppose t, a, b are integer vectors with $0 \leq a \leq b \leq t$, $\mathbf{1}^\top t = N$, $\mathbf{1}^\top a \leq m \leq \mathbf{1}^\top b$, and $0 \leq r \leq \min(m, N - m)$. For an exact row-simplex incumbent B, define*

$$\begin{gathered} \ell = \max(0, a - h(B)), \quad u = \min(b - h(B), t - v(B)), \\ B \in \mathcal{C}_{\text{joint}} \iff \ell \leq u, \quad \mathbf{1}^\top \ell \leq r \leq \mathbf{1}^\top u. \end{gathered} \tag{5}$$

The set is nonempty and is the convex hull of compatible hard scored targets under this single-subset model.

Proof idea. Removed counts must lie between ℓ and u and sum to r. Such counts exist exactly under the displayed conditions. A network-flow formulation with integer capacities gives the convex-hull result (Appendix B.1). Computing v and h costs $O(mK)$ arithmetic operations, followed by an $O(K)$ classwise check.

A probability prediction is compatible when it belongs to the convex hull of admissible hard-label assignments, without itself being a hard assignment. Proposition 1 and Equation 3 identify whether a strictly positive ideal worst-case improvement is possible, not its magnitude. Passing the 10^{-8} screen does not prove exact compatibility or impossibility.

Full joint adjustment projects B onto $\mathcal{C}_{\text{joint}}$, using a finite numerical implementation in the experiments. The count-budget control in Table 1 is a relaxation, not the strongest possible summary without identities. Our empirical attribution is limited to the tested summaries. Figure 1 shows that identities can matter in principle, not that they alone explain every observed gain.

4 Certified short corrections

Instead of solving the full projection, we can act on a violated restriction. Exact incompatibility exposes a linear inequality $\langle A, X\rangle \le \beta$ satisfied by every admissible target but violated by the incumbent. Put $g(X) = v(X) - h(X)$, the totals on newly added instances. Evaluate v and h at the current prediction to choose $J_{\text{low}} = \{k : a_k - h_k > 0\}$ and $J_{\text{up}} = \{k : b_k - h_k \le t_k - v_k\}$. With these sets fixed, the inequalities below apply to candidate targets X, using $v(X)$, $h(X)$ and $g(X)$:

$$\begin{aligned} h_k \le b_k, \quad v_k \le t_k, \quad g_k \le t_k - a_k \quad (k = 1, \dots, K), \\ \sum_{k \in J_{\text{low}}} (a_k - h_k) \le r, \\ \sum_{k \in J_{\text{up}}} h_k + \sum_{k \notin J_{\text{up}}} v_k \le \sum_{k \in J_{\text{up}}} b_k + \sum_{k \notin J_{\text{up}}} t_k - r. \end{aligned} \tag{6}$$

For example, $g_k \le t_k - a_k$ limits the newly added instances to class counts remaining outside the original sample's minimum. These at most $3K + 2$ restrictions expose any exact incompatibility. Positive parts give the lower-sum violation and coordinate minima the upper-sum restriction (Appendix B.2). They need not maximise normalised violation over all class subsets. Row-centring gives $\widetilde{A}_{ik} = A_{ik} - s_i/K$ and $\widetilde{\beta} = \beta - \sum_i s_i/K$, where $s_i = \sum_k A_{ik}$. The restriction is unchanged on simplex targets.

Proposition 2 (Certified correction from a valid restriction)**.** *First project onto a violated halfspace, then restore each probability vector. For an exact row-simplex B and valid cut with $\delta = \langle \widetilde{A}, B\rangle - \widetilde{\beta} > 0$, let*

$$H = B - \frac{\delta}{\|\widetilde{A}\|_F^2}\widetilde{A}, \qquad G_i = \Pi_{\Delta_K} H_i. \tag{7}$$

Then, for every $Y \in \mathcal{C}_{\text{joint}}$,

$$R(B, Y) - R(G, Y) \ge \frac{\delta^2}{m\|\widetilde{A}\|_F^2} + \frac{\|H - G\|_F^2}{m} > 0. \tag{8}$$

The two projections guarantee loss improvement, not satisfaction of all joint constraints. Restoring probability vectors can re-cross the selected restriction without invalidating the bound. Appendix B.2 gives an example and the proof from classical projection inequalities.

Certifying the computed output. Equation 8 assumes exact projections. We separately certify the arrays actually returned. For finite B, G, a valid original integer cut (A, β) and a fixed multiplier $\lambda \ge 0$, define

$$\underline{\Gamma}(B, G; \lambda) = \frac{\|B\|_F^2 - \|G\|_F^2 + 2[-\lambda\beta + \sum_i \min_k (G_{ik} - B_{ik} + \lambda A_{ik})]}{m}. \tag{9}$$

Table 1: Information and computation comparisons. Coordinatewise correction constructs B. Subsequent procedures start from that prediction and score the same instances. Their bounds use the same parent counts and sampling design, but retain different restrictions. Individual reference labels are used only for evaluation.

Representation	**Retained restrictions**
Coordinatewise	Original intervals expanded by r, intersected with deterministic parent bounds. Supplies the available incumbent B.
Count budget	Keeps a possible original count vector $c_0 \in [a, b]$, $\mathbf{1}^\top c_0 = m$, deterministic scored bounds and $\|v - c_0\|_1 \leq 2r$.
Joint membership	Keeps the actual overlap sum $h(X)$ and a consistent removed/outside completion in Equation 4.
Short vs full adjustment	Same joint information; changes computation and finite certificate construction.

This bounds direct gain for the computed output, even with small floating-point row-sum errors. No exact projection is required. We evaluate stored binary-rational values exactly at the chosen halfspace multiplier and at $\lambda = 0$, retain the larger bound, and round its floating-point representation downwards. Accepted stepwise bounds sum to a certificate for the update. Sharper certification and more accurate optimisation are distinct.

Algorithm 1 Recorded-membership short correction. Evaluator labels are not inputs.

Inputs: B, N, t, a, b, S_0, S and fixed step budget $L \in \{1, 3\}$.
Validate the input geometry and nonempty count conditions. On failure, return B with bound zero and a failure-status code. Otherwise set $X \leftarrow B$ and exact accumulated bound $c \leftarrow 0$.
For at most L steps:

1. Form scored/overlap sums and row-domain residuals. Stop if the maximum of joint and row-simplex residuals is at most 10^{-8}.
2. Form the candidate valid cuts and choose the largest positive $\delta^2/\|\widetilde{A}\|_F^2$ using the fixed family/class tie order. Stop if none exists.
3. Apply Equation 7 with the current X as its base. Certify the stored candidate using Equation 9 at $\lambda = \delta/\|\widetilde{A}\|_F^2$ and 0.
4. If the exact best bound is nonpositive, reject and stop. Otherwise accept $X \leftarrow G$ and add its exact bound to c.

Return X, the downward-rounded c, step statuses and domain/joint residuals. Stopping at the numerical screen or rejecting a candidate does not prove that no positive uniform improvement is possible.

Scope of the guarantee. All cuts share the event E that the original sample's true counts lie in $[a, b]$. They are not independent confidence statements. For a correction trigger T with positive probability, marginal coverage gives only $\Pr(E^c \mid T) \leq \min\{1, 0.05/\Pr(T)\}$. Rare triggers do not inherit 95% conditional coverage, and parents do not inherit simultaneous coverage. Protection applies to group-average loss, not every instance.

5 Experimental protocol

We compare adjustments of fixed Covertype predictions (Blackard, 1998) from a tabular adaptation of proportional value classification (PVC; Ma et al., 2026b) and a matched proportion-matching (PM) learner. Both use a 54–64–7 backbone, three matched seeds, training-only standardisation and aggregate-only validation. This is an adaptation of the PVC objective, not a reproduction of its published benchmark ranking. Parent groups contain 64 instances ordered by wilderness category and elevation. They are feature-defined bags rather than geographic footprints. The 512 training, 128 validation and 256 original evaluation groups exclude the separate acquisition groups, and no instances cross splits (Appendix C).

A uniform draw selects m instances. We replace its r farthest-from-water instances with the r nearest-water omitted instances, breaking ties by fixed identifiers. The initial count-summary comparison tests replacement-aware against deterministic correction, with PVC $(m, r) = (16, 4)$ as

primary. The membership follow-up adjusts the saved incumbent B on those inspected groups. The held-out confirmation then tests recurrence on previously unused groups without replacing the initial primary. The pre-test 640-epoch sensitivity and five further draws remain supporting comparisons on the original groups.

The confirmation selects 256 unused groups by feature metadata, excluding all 2,944 previously selected parents and their instances. The six selected 160-epoch models are checked against saved predictions before new-group evaluation. Two draws are coupled across $(m, r) = (16, 2), (16, 4), (32, 4), (32, 8)$. The primary is full-joint gain over B for PVC $(16, 4)$. Short corrections, PM and other regimes are secondary. Each output is one group–model-seed–draw combination, giving $256 \times 3 \times 2 = 1{,}536$ primary outputs. We average seeds and draws within parents, then parents equally. Unchanged outputs, incumbents violating previous constraints, uncovered cases and failed candidates remain included. The 2,000-resample parent bootstrap is descriptive, conditional on the construction and frozen models. Seeds and draws are dependent observations.

We report realised gain, direct certificates, output changes, coverage and individual deterioration separately. An output change uses a 10^{-8} margin. Short corrections require a strictly positive exact certificate, while full adjustment retains its 10^{-10} certificate gate, 10^{-8} feasibility checks and 10^{-7} gap targets. These procedures therefore compare finite implementations with different gates, not exact projections under identical stopping rules. Individual reference labels evaluate outcomes but never select restrictions, models or masks.

6 Results: when recorded membership changes predictions

6.1 Confirmation on previously unused groups

In the prespecified PVC $(m, r) = (16, 4)$ confirmation, full joint adjustment gives mean additional loss reduction approximately 0.000382 over B. The median parent gain is zero, and 253 of 256 parents have zero gain. The descriptive parent-bootstrap interval $[0, 0.000958]$ includes zero (Table 2 and Appendix C.3). The eight changed outputs had about twice the average loss before correction (1.01 versus 0.49), and their loss fell by 8.6% on average. Such corrections therefore recur on unused groups from the same source, but rarely, and the mean is not reliably positive.

On these same instances, mean loss falls from 0.497639 for F to 0.491823 for B and 0.491442 for G (Table 10). Most of the total decrease precedes membership-aware adjustment. The primary measures only its additional gain over B.

Table 2: Prespecified PVC confirmation on 256 unused parent groups, with $m = 16$ and $r = 4$. Gains and bounds are multiplied by 10^3. Means average three seeds and two draws within each parent, then weight parents equally. Shading identifies the full-joint primary, not significance. Short corrections are secondary. Each changing method affects eight of 1,536 outputs across three parents, and the bounds need not be equally sharp.

Method	Gain over B	Mean lower bound	Changed / 1,536
Keep incumbent	0.0000	0.0000	0
Count budget	0.0000	0.0000	0
One cut	0.2524	0.0202	8
Three cuts	0.3558	0.0241	8
Full joint (primary)	0.3818	0.0249	8

Full-joint descriptive parent-bootstrap interval: $[0, 0.958143] \times 10^{-3}$ (includes zero).

Why test recorded membership? The initial replacement-aware bounds tighten in 175 groups, yet all 768 PVC $(16, 4)$ predictions remain identical to deterministic correction. The 640-epoch sensitivity also gives zero gain. A later membership follow-up changes five outputs across three original groups, but one group dominates its gain. These concentrated results motivated held-out confirmation. Appendix C retains the original effects, positive $(32, 4)$ sensitivity and further-draw results.

6.2 INCIDENCE, MAGNITUDE AND INDIVIDUAL OUTCOMES

Eight of 1,536 primary outputs change, across three groups. Their conditional mean group-loss decrease is 0.07330, not a percentage-point accuracy increase. Thus, the small all-case mean combines rare changes with many unchanged outputs. The largest parent contributes 57.68% of positive parent-mean gain. These retrospective conditional summaries do not replace the all-group primary result.

Average improvement does not protect every individual prediction. In the full-joint primary, 25 of the $1{,}536 \times 16 = 24{,}576$ prediction–instance pairs have loss increases above 10^{-8}, or 0.102%. Among the $8 \times 16 = 128$ pairs in changed outputs, the same 25 account for 19.53%. Their mean increase is 0.04297 and their maximum is 0.19648. Seventeen exceed the existing descriptive threshold of 0.01. These pairs repeat instances across model seeds and draws, so they are not counts of distinct locations. Reporting both denominators and the increase magnitudes distinguishes overall incidence from deterioration within corrected outputs (Appendices C.3 and C.5).

The procedures also differ in their realised-gain rankings across regimes. At PVC $(16, 2)$, three correction steps slightly exceed full joint adjustment. This is consistent with Equation 3, which optimises an ideal worst-case guarantee rather than gain for the realised reference targets. The stored certificates are method-specific lower bounds and need not be equally sharp. Figure 2 separates all-case gain, correction frequency and conditional magnitude. Table 7 gives every PVC regime value.

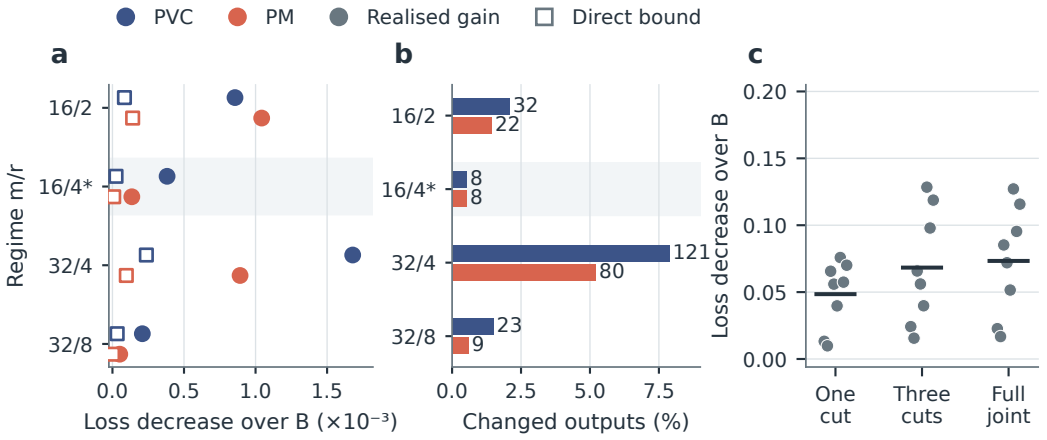


Figure 2: Frequency and magnitude of additional improvement. Rows give (m, r), and the asterisk marks the PVC primary regime. (a) Full-joint mean gains (filled) and direct bounds (open) for both learners and four regimes, including every parent. (b) Output-change frequencies and counts at the 10^{-8} coordinate margin, with $256 \times 2 \times 3 = 1{,}536$ outputs per cell. Shading identifies the PVC primary regime only. (c) The eight changed PVC primary outputs for each method, drawn from three parents. Black lines show means. Only this conditional panel excludes unchanged outputs. The dependence and all-case uncertainty are as stated in Table 2.

6.3 ADDITIONAL INFORMATION VERSUS UNFINISHED COMPUTATION

To separate new information from unfinished computation, we check the incumbent against its previous constraints. All three joint/cut procedures change 303 of 12,288 confirmation outputs. Of these full-joint cases, 297 pass the old constraints within 10^{-8} and six fail. All six count-budget changes occur in the latter group. Such gains cannot be assigned solely to new information, and numerical screening is not an exact impossibility proof. A 160-epoch diagnostic removes old violations in 21 incumbents, after which both new representations leave all 42 comparisons unchanged. The 640-epoch diagnostic retains fallbacks. Appendix C preserves the full controls and original 70/seven-case decomposition.

6.4 Short corrections and computational cost

At PVC $(16, 4)$ on the unused parent groups, one and three correction steps recover 66.12% and 93.19% of the full-joint mean gain. These ratios compare means, not guaranteed retention for each case. Both short procedures have smaller positive mean gains than full adjustment across the original, longer-training, further-draw and confirmation comparisons, and all parent medians are zero (Table 7). At 640 epochs, the short procedures change 12 outputs across eight parents, but only three parent means exceed the numerical reporting threshold of 10^{-10}. An output change therefore need not imply a substantial gain.

On 303 confirmation inputs selected by their pre-correction residuals, three additional calls per method give median runtimes of 5.293 ms for one step, 16.362 ms for three steps and 16.447 ms for full joint adjustment. Timing includes construction, screening, correction and production certification, but excludes supplemental proof extraction. All 65,985 outputs, including the all-input pass, match their saved counterparts bitwise. One step is faster in this implementation. Three steps and full adjustment have similar medians, with no consistent cost ordering across this and the historical timings. These measurements pool regimes on one hardware environment (Appendix C.6).

6.5 Supporting check: which prediction is being corrected?

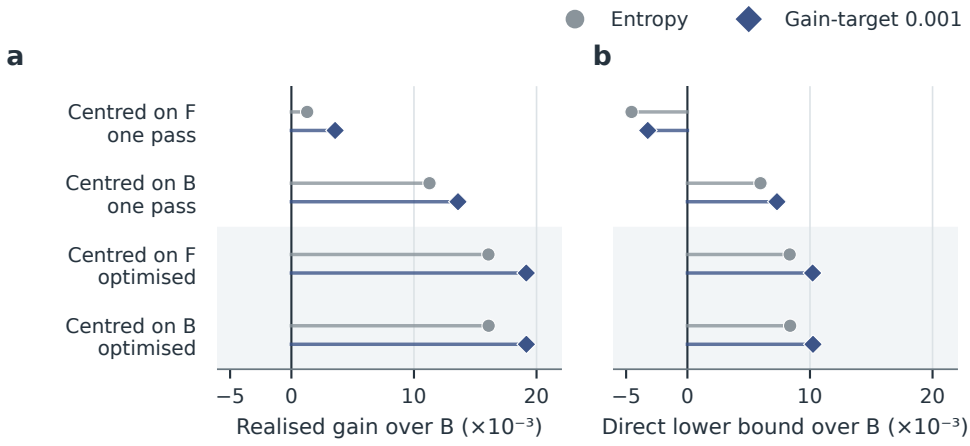


Figure 3: Additional benefit beyond the same optimised free incumbent B for all 410 Covertype groups at 24 queries. Answers are fixed within each policy. The reference is the prediction in the correction objective, not the optimiser's initial point. Gain-target 0.001 names the query policy. (a) Realised squared-loss gains and (b) sharp direct bounds average seeds within family, then weight the three learned families and groups equally. Shading groups the optimised procedures. Incumbent-centred updates protect group-average loss, not individual instances. Further optimisation adds mean benefit, and both optimised controls are retained (Appendix D.2).

The central study changes information representation around a fixed incumbent. This supporting comparison instead changes the correction's reference while holding acquired answers fixed. Each of 410 Covertype groups has 64 instances, with 16 scored and 24 labels queried among the other 48. Parent counts provide "free" information, meaning no additional queried labels, not no computation. Here B is an optimised free correction, unlike the central coordinatewise incumbent. Procedures use identical answers within each policy.

The optimised free correction reduces loss by 0.029365 relative to F, compared with 0.019020 for the one-pass free correction. When the queried-label one-pass update is centred on F, its mean additional gain over B is 0.001287 under entropy and 0.003569 under the gain-target policy. Nevertheless, it is worse than B in 73.52% and 73.36% of weighted cases, respectively. Centring the update on B raises the mean gains to 0.011266 and 0.013593. This changes the reference in the correction objective, not merely the optimiser's initial point. The answers, scored instances and original policy comparison remain fixed.

After optimisation, centring on F or B changes mean gain by only 0.000017545 under entropy and 0.000010220 under the gain-target policy, with some casewise reversals. The larger one-pass difference therefore mainly reflects preservation of the correction already available, rather than a generally superior optimiser. Figure 3 retains both optimised controls. Appendix D.2 reports the direct certificates, individual deterioration and numerical scope, and Appendix D.3 reports accuracy.

7 Discussion and limitations

These results suggest a simple way to use sampling records. After an aggregate correction, the $O(mK)$ test can be run on every group at negligible cost. Where the prediction is compatible with the record, it is left unchanged. Where the record rules the prediction out, a short correction applies. Whenever the sampling intervals contain the true counts, its certificate then guarantees a lower group-average loss for every compatible target. In the primary confirmation setting, this happened in 3 of 256 groups. How often it happens under other sampling, replacement or grouping designs remains open.

What the supporting studies add. These studies do not extend the single-subset theorem, but they identify two conditions for interpreting adjustment gains. First, an aggregate must describe the instances it is applied to. Applying a parent group's proportions to a western-quarter subset worsens loss by 13–27% in BENELUX, and a gain for the Colombia softmax model comes from a favourable change in residual errors rather than from the aggregate (Appendix E). Second, additional benefit must be measured against the prediction already available, because a correction centred on an earlier prediction can discard improvement already present in the incumbent (Figure 3). The acquisition and reconstruction studies report mixed contrasts, given in full in Appendices F and G.

Limitations. Compatibility concerns information, not accuracy. A compatible prediction can still be inaccurate, and agreement with an aggregate does not show that individual predictions are correct. The central confirmation uses one data source and one group-construction procedure, with dependent seeds and draws. The coverage argument relies on a uniform original draw, and the compatibility test treats one original sample per parent group. Several overlapping samples and partially recorded histories remain open. The exact parent counts and sampling intervals do not establish protection under noisy aggregate observations. Marginal coverage does not imply 95% coverage conditional on a correction trigger, and a group-average loss guarantee does not protect every instance. Empirical attribution is limited to the tested count summaries, because the count-budget control is not the strongest identity-free representation. The image audit reuses test evidence and remains exploratory (Appendix H).

8 Conclusion

Recorded sample membership is information that data collection often already produces, and using it requires no new labels. We give an $O(mK)$ test that shows whether this information can improve a prediction already corrected with aggregate counts. We also give short corrections that change a prediction only when the record rules it out, each with a certificate on its group-average loss that holds whenever the sampling intervals contain the true counts. On 256 previously unused Covertype groups in the primary setting, such corrections are rare: 8 of 1,536 outputs change, in 3 groups. Those outputs had about twice the average loss, and their loss fell by about 9%. One cut recovers 66% of the full adjustment's gain in about a third of its run time, and three cuts recover 93%. Recording which sampled instances were replaced costs little and makes this check possible.

AI USE STATEMENT

The research problem arose from a practical need to improve the resolution of a planetary resource map. We worked mainly with OpenAI's Codex (GPT 5.6 Sol and GPT6 Astra), which helped us find related work, develop the proofs, design the experiments, write and test the code, prepare the data, interpret the results, make the figures, and edit the text. Small synthetic cases were used only to test the code. Anthropic's Claude helped check the paper and supplement against the ICLR rules, reorganise the appendix and supplement, and edit the final text. We checked the ideas, claims, results and full text ourselves, and we take full responsibility for everything in the paper.

ETHICS STATEMENT

All experiments use publicly available datasets cited in the paper. The study involves no human subjects, personal data or new data collection. The guarantees concern group-average squared loss under the stated constraints, and individual predictions can deteriorate (Section 6.2). Aggregate-based adjustment should therefore not be treated as protection for every individual when predictions inform decisions about people or places.

REPRODUCIBILITY STATEMENT

Theory. Proposition 1 is proved in Appendix B.1. Proposition 2 and the stored-output certificate in Equation 9 are proved in Appendix B.2. The uniform-gain guarantee behind both is proved in Appendix A.4.

Experiments. Section 5 and Appendices C and C.3 give the data construction, learners, seeds, regimes and the prespecified primary. Table 27 maps each main-text proposition, figure and results table to its appendix details and to its location in the supplement.

Supplement. The supplementary ZIP contains the method code, the inputs and per-case results of the unused-parent confirmation, the plotted values for every figure, the sources of the tabulated results, and a description of the data. `summarise.py` recomputes Table 2, the matched losses in Appendix C.4 and the other confirmation tables from the per-case results, and all 217 reported values agree. `rerun.py` reruns every correction from the supplied frozen predictions in a few minutes on a laptop. On our test machine, it reproduces every primary value and 207 of the 217 reported values. The other ten belong to one secondary cell, where one of 1,536 full-joint outputs fails the solver's numerical checks and returns the unchanged prediction, as the method specifies. Earlier studies are supplied as summaries and plotted values, while their detailed inputs remain private. Neither script repeats model training.

REFERENCES

Lorne Applebaum, Travis Dick, Claudio Gentile, Haim Kaplan, and Tomer Koren. Optimal learning from label proportions with general loss functions. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2509.15145.

Takanori Asanomi, Shinnosuke Matsuo, Daiki Suehiro, and Ryoma Bise. MixBag: Bag-level data augmentation for learning from label proportions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16570–16579, 2023. URL https://arxiv.org/abs/2308.08822.

Heinz H. Bauschke and Jonathan M. Borwein. On projection algorithms for solving convex feasibility problems. *SIAM Review*, 38(3):367–426, 1996. doi: 10.1137/S0036144593251710.

Jock Blackard. Covertype. UCI Machine Learning Repository, 1998. URL https://archive.ics.uci.edu/dataset/31/covertype.

James P. Boyle and Richard L. Dykstra. A method for finding projections onto the intersection of convex sets in Hilbert spaces. In *Advances in Order Restricted Statistical Inference*, pp. 28–47. Springer, 1986. doi: 10.1007/978-1-4613-9940-7_3.

Anand Brahmbhatt, Mohith Pokala, Rishi Saket, and Aravindan Raghuveer. LLP-Bench: A large scale tabular benchmark for learning from label proportions. In *Proceedings of CIKM (Applied Research Track)*, 2024. URL https://research.google/pubs/llp-bench-a-large-scale-tabular-benchmark-for-learning-from-label-proportions/.

Robert Istvan Busa-Fekete, Travis Dick, Claudio Gentile, Haim Kaplan, Tomer Koren, and Uri Stemmer. Nearly optimal sample complexity for learning with label proportions. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 5943–5976. PMLR, 2025. URL https://proceedings.mlr.press/v267/busa-fekete25a.html.

Jueyu Chen, Wantao Wen, Yeqiang Wang, Erliang Lin, Yemin Wang, and Yuheng Jia. Robust label proportions learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0267. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/0ba7a9cf188089693eff454ac865c821-Abstract-Conference.html.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Gabriel Franco, Mark Crovella, and Giovanni Comarela. Dependence and model selection in LLP: The problem of variants. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 470–481, 2023. doi: 10.1145/3580305.3599307. URL https://www.cs.bu.edu/faculty/crovella/paper-archive/kdd23-LLP-model-selection.pdf.

Panagiotis D. Grontas, Antonio Terpin, Efe C. Balta, Raffaello D'Andrea, and John Lygeros. ΠInet: Optimizing hard-constrained neural networks with orthogonal projection layers. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/66ba03db7e50b5db5593fe759527a74f-Abstract-Conference.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Ursula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (Computationally-Identifiable) masses. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1939–1948, 2018. URL https://proceedings.mlr.press/v80/hebert-johnson18a.html.

Theodore James Thibault Heiser, Mari-Liis Allikivi, and Meelis Kull. Shift happens: Adjusting classifiers. In *Machine Learning and Knowledge Discovery in Databases: ECML PKDD 2019*, volume 11907 of *Lecture Notes in Computer Science*, pp. 55–70. Springer, 2020. doi: 10.1007/978-3-030-46147-8_4. URL https://arxiv.org/abs/2111.02529.

Adel Javanmard, Lin Chen, Vahab Mirrokni, Ashwinkumar Badanidiyuru, and Gang Fu. Learning from aggregate responses: Instance level versus bag level loss functions. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2401.11081.

Jesse H. Krijthe and Marco Loog. Projected estimators for robust semi-supervised classification. *Machine Learning*, 106:993–1008, 2017. doi: 10.1007/s10994-017-5626-8.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Shunsuke Kubo, Shinnosuke Matsuo, Daiki Suehiro, Kazuhiro Terada, Hiroaki Ito, Akihiko Yoshizawa, and Ryoma Bise. Theoretical proportion label perturbation for learning from label proportions in large bags. In *ECAI 2024*, volume 392 of *Frontiers in Artificial Intelligence and Applications*, pp. 2194–2201. IOS Press, 2024. doi: 10.3233/FAIA240740. URL https://arxiv.org/abs/2408.14130.

Meelis Kull and Peter A. Flach. Novel decompositions of proper scoring rules for classification: Score adjustment as precursor to calibration. In *Machine Learning and Knowledge Discovery in Databases*, volume 9284 of *Lecture Notes in Computer Science*, pp. 68–85. Springer, 2015. doi: 10.1007/978-3-319-23528-8_5.

Gene Li, Lin Chen, Adel Javanmard, and Vahab Mirrokni. Optimistic rates for learning from label proportions. In *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 3437–3474. PMLR, 2024. URL https://proceedings.mlr.press/v247/li24b.html.

Tianhao Ma, Ximing Li, Changchun Li, and Renchu Guan. Learning from label proportion with dual-proportion constraints. In *Findings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7583–7592, 2026a. URL https://openaccess.thecvf.com/content/CVPR2026F/html/Ma_Learning_from_Label_Proportion_with_Dual-Proportion_Constraints_CVPRF_2026_paper.html.

Tianhao Ma, Wei Wang, Ximing Li, Gang Niu, and Masashi Sugiyama. Learning from label proportions via proportional value classification. In *International Conference on Learning Representations*, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/6fc46679a7ba2ec82183cf01b80e5133-Abstract-Conference.html.

Charles F. Manski. Partial identification with missing data: Concepts and findings. *International Journal of Approximate Reasoning*, 39(2–3):151–165, 2005. doi: 10.1016/j.ijar.2004.10.006.

Alain Rakotomamonjy, Maxime Berar, and Mokhtar Z. Alaya. Unmixing mean embeddings for domain adaptation with target label proportion. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 3214–3222, 2026. URL https://proceedings.mlr.press/v300/rakotomamonjy26a.html.

Raúl Ramos-Pollán and Fabio A. González. Sentinel2 RGB chips over BENELUX with ESA World Cover for learning with label proportions, 2023a. URL https://zenodo.org/records/7935237. Version 1.0.

Raúl Ramos-Pollán and Fabio A. González. Sentinel2 RGB chips over Colombia (NE) with ESA World Cover for learning with label proportions, 2023b. URL https://zenodo.org/records/7935303. Version 1.0.

Raúl Ramos-Pollán and Fabio A. González. On-orbit model training for satellite imagery with label proportions, 2023c. URL https://arxiv.org/abs/2306.12461. arXiv:2306.12461, version 2.

Clayton Scott and Jianxin Zhang. Learning from label proportions: A mutual contamination framework. In *Advances in Neural Information Processing Systems*, 2020. URL https://papers.nips.cc/paper/2020/hash/fcde14913c766cf307c75059e0e89af5-Abstract.html.

Asterios Tsiourvas, Wei Sun, Georgia Perakis, Pin-Yu Chen, and Yada Zhu. Learning optimal projection for forecast reconciliation of hierarchical time series. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 48713–48727, 2024. URL https://proceedings.mlr.press/v235/tsiourvas24b.html.

Utkarsh Utkarsh, Danielle Maddix, Ruijun Ma, Michael W. Mahoney, and Bernie Wang. End-to-end probabilistic framework for learning with hard constraints. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/17bae482066287ce54814514cf12f248-Abstract-Conference.html.

Weiran Wang and Canyi Lu. Projection onto the capped simplex, 2015. URL https://arxiv.org/abs/1503.01002. arXiv:1503.01002.

APPENDIX CONTENTS

A Mathematical foundations and finite-return certificates **14**
A.1 Background comparisons and omission bounds . . . 14
A.2 Aggregate-loss identities and the affine diagnostic . . . 15
A.3 Finite-population identity and verification details . . . 15
A.4 Uniform-gain proof and spatial certification . . . 15
A.5 Nested residual caps and the common one-pass return . . . 16

B Recorded-membership geometry and finite correction **17**
B.1 Retaining joint membership information . . . 17
B.2 Explaining and applying the membership restrictions . . . 18

C Covertype selection studies: protocols and complete results **19**
C.1 Joint-information protocol and original results . . . 23
C.2 Attribution and short-correction comparisons . . . 24
C.3 Fixed unused-parent confirmation protocol and results . . . 25
C.4 Matched original, incumbent and adjusted losses . . . 28
C.5 Retrospective frequency, certificate and harm summaries . . . 29
C.6 Endpoint-retaining matched timing follow-up . . . 29

D Available incumbents, certificate sharpness and computation **30**
D.1 Certificate sharpness and computation are separate questions . . . 30
D.2 Updating the available prediction . . . 32
D.3 Classification accuracy as a secondary endpoint . . . 34

E Spatial data, support interventions and their interpretation **35**
E.1 Spatial implementation, controls and interpretation . . . 36
E.2 Complete primary quarter-support table . . . 37
E.3 Colombia replication: protocol and scope . . . 37
E.4 Saved-support mechanism and protected-unit audit . . . 40
E.5 Fixed eastern and central-quarter follow-up . . . 42
E.6 Composition error and magnitude of individual deterioration . . . 43

F Acquisition model, bounds and additional comparisons **43**
F.1 Restricted omitted-chip queries . . . 43
F.2 Activation, its exact hindsight cost and complementarity . . . 44
F.3 Bounds for a specified positive value . . . 45
F.4 A common one-pass repair and its paired comparison . . . 46
F.5 Acquisition protocol and further results . . . 46
F.6 Verification and scope . . . 47
F.7 Fixed-policy acquisition on a new data source . . . 48
F.8 Paired Covertype policy outcomes . . . 50
F.9 Direct benefit over the free return, harm tails and access assumptions . . . 52

G A bounded study with overlapping spatial observations **55**
G.1 Inputs, coverage and information access . . . 55
G.2 Reconstructions and the protected quantity . . . 56
G.3 Fixed estimand and full comparisons . . . 56
G.4 Numerical checks, chronology and reproduction limits . . . 57

H Exploratory image controls and count uncertainty **59**
H.1 Design and provenance . . . 59
H.2 Fitting and aggregate-only selection . . . 60
H.3 Broader models narrow the ranking claim . . . 60
H.4 Ambiguous comparisons coexist with certified repair . . . 60
H.5 Sensitivity to uncertain aggregate counts . . . 61
H.6 Numerical checks and finite returned predictions . . . 62

I Study chronology and reproducibility scope **62**
I.1 Result-to-artifact map . . . 62
I.2 Populations and decision chronology . . . 64
I.3 Measured historical optimiser sensitivity . . . 66
I.4 Counts of completed local checks . . . 66

APPENDIX ROADMAP

For the central membership result, start with the uniform-gain proof in Appendix A.4, the membership and cut derivations in Appendices B.1–B.2, and the Covertype studies in Appendix C, including the original null and the unused-parent confirmation. Appendix D.2 supports the available-incumbent comparison, and Appendix I.1 states what readers can replay. The spatial, acquisition, reconstruction and image studies answer separate supporting questions. They are not additional confirmations of the single-subset result. The guide below links each reader question to its pages.

Reader question	**Where to look**
Why do the guarantees hold?	Uniform gain, p. 15; then membership/cuts, p. 17
What are the full central protocols and results?	Selection study, p. 19; confirmation, p. 25
Do the incumbent and scored support change the conclusion?	Incumbent, p. 32; spatial support, p. 35
What do the supporting studies establish?	Acquisition, p. 43; reconstruction, p. 55; image controls, p. 59
How can the results be checked, and what is not reproduced?	Quick start and artifact map, p. 62; numerical sensitivity, p. 66

A MATHEMATICAL FOUNDATIONS AND FINITE-RETURN CERTIFICATES

Table 3: Reading key: the population, prediction and loss are different objects.

Term	Meaning in this paper
Measured support	Rows described by the supplied aggregate; often the full parent region.
Scored subset	Rows on which the prediction is evaluated; it can differ from the measured support.
Group mean	Average class/composition vector, e.g. the fraction of retained plots that are forest.
Mean individual loss	Average of per-row squared-vector errors; not the error of the group mean.
Incumbent B	The prediction already available for comparison, possibly after an earlier correction.
Direct gain certificate	A lower bound on loss reduction from that incumbent for every target in the declared set.
Attainment	Fraction of cases reaching a stated gain threshold; distinct from mean gain.

A.1 BACKGROUND COMPARISONS AND OMISSION BOUNDS

A direct comparison. For two fixed returns, the loss difference is affine in unknown targets:

$$R(G,Y) - R(F,Y) = m^{-1}(\|G\|_F^2 - \|F\|_F^2) - 2m^{-1}\langle G - F, Y\rangle. \tag{10}$$

Its extrema over compatible targets give a comparison interval. Crossing zero means the information permits opposite rankings; it does not mean the realised ranking is wrong. Appendix H evaluates this distinction with integer-count assignments.

Omitted instances change the admissible set. An exact parent mean obeys $Q = \rho q + (1-\rho)r$, where $\rho > 0$ is the retained fraction and q, r are retained/omitted compositions. Eliminating r gives the sharp set

$$T(Q,\rho) = \{q \in \Delta_K : \rho q_k \le Q_k \ \forall k\}. \tag{11}$$

The simplex also implies $q_k \ge \max\{0, (Q_k - 1 + \rho)/\rho\}$. These bounds assume nested supports and consistent weights, classes and reference time; they are not measurement-noise intervals. Let $\mathcal{C}_T$ contain row-simplex arrays with means in T.

Finite repair and a cheap control. Project $\bar{F}$ onto T, shift each row by $d = \Pi_T \bar{F} - \bar{F}$, then project $H_i = F_i + d$ onto its simplex to obtain G_i. Two projection inequalities give

$$R(F,Y) - R(G,Y) \ge \|d\|^2 + m^{-1}\sum_i \|H_i - G_i\|^2. \tag{12}$$

The bound protects group-average loss under valid information even if the final predictions do not satisfy the aggregate constraints. Iterative intersection projection (Boyle & Dykstra, 1986) can approach Equation 3; we assess each finite return separately with assignment or independently checked fractional-target bounds. Numerical feasibility, a stopping rule and a certified loss improvement are different checks.

A.2 Aggregate-loss identities and the affine diagnostic

The additive adjustment and Brier decompositions build on Kull & Flach (2015); the bounded probability adjustment is due to Heiser et al. (2020). For an equal-size partition into bags of size s, write $B(F)$ for mean squared bag-mean error and $e_i = F_i - Y_i$. The usual within/between identity is

$$R(F, Y) = B(F) + n^{-1} \sum_b \sum_{i \in b} \|e_i - \bar{e}_b\|^2. \tag{13}$$

For a uniformly random equal-sized partition independent of fixed F, Y, with $1 \le s < n$,

$$\mathbb{E}[B(F)] = aR(F, Y) + (1 - a)\|\bar{e}\|^2, \qquad a = (n - s)/(s(n - 1)). \tag{14}$$

Global counts make $\bar{e}$ observable. The corrected score $[B(F) - (1 - a)\|\bar{e}\|^2]/a$ is unbiased over partitions, not necessarily after selecting a model on the same partition or under feature ordering. These are classical identities, not new learning guarantees.

For an affine correction $G_i = F_i + \widetilde{q} - \bar{F}$ and actual retained mean q,

$$R(F, Y) - R(G, Y) = \|q - \bar{F}\|^2 - \|\widetilde{q} - q\|^2. \tag{15}$$

Row-simplex projection can only increase this gain. A negative affine gain therefore does not establish harm after projection. Equation 28 instead decomposes the actual returned full adjustment, including centred residual changes and a finite-mean remainder.

A.3 Finite-population identity and verification details

For a uniformly selected bag, $\Pr(i \in b) = s/n$ and $\Pr(i, j \in b) = s(s - 1)/(n(n - 1))$ for $i \ne j$. Expanding its squared mean residual gives Equation 14; exchangeability of bags gives the partition average. For Equation 13, the cross term between each within-bag centred residual and its bag mean is zero. These calculations are identities conditional on fixed residuals, not modelling assumptions about images.

The exact risk-difference objective is linear even for fractional label matrices because the $\|Y\|^2$ term cancels. With disjoint bags, integral class counts and unit row sums, assigning class slots yields integral optima. Labels already queried can be removed together with their counts before repeating the calculation. For overlapping or uncertain aggregates, the assignment reduction need not apply; a different constrained optimisation problem is required.

The exact repair certificates optimise over one-hot labels compatible with the observed counts. They do not depend on the realised held-out labels except through those counts. The held-out labels are used separately to measure actual risk. Equality constraints and prediction values are represented in floating point; the smallest certificate margin in the exploratory CIFAR exact-count experiment (Appendix H.4), .00390, is much larger than the checked solver and replay discrepancies. This numerical margin is evidence of stability, not a proof using interval arithmetic.

A.4 Uniform-gain proof and spatial certification

In this section, n denotes the number of scored rows (m in the main text); all losses and certificates use that same normalisation. The risk difference is affine in Y, so its minimum over $\mathcal{Y}$ equals that over its convex hull $\mathcal{C}$. Assume $\mathcal{C}$ is nonempty and contained in the row-simplex prediction domain $\mathcal{D}$. Write $P = \Pi_{\mathcal{C}}(F)$. For any $G \in \mathcal{D}$,

$$\min_{Y \in \mathcal{C}} \{R(F, Y) - R(G, Y)\} \le R(F, P) - R(G, P) \le n^{-1} \|F - P\|_F^2.$$

Projection optimality attains this bound at $G = P$, proving Equation 3. For disjoint integer counts, the transportation polytope is the required convex hull. For the spatial experiment, all fractional simplex targets with allowed means are explicitly admissible.

For Equation 15, set $d = \widetilde{q} - \bar{F}$. Expanding the loss gives the gain $2\langle d, q - \bar{F}\rangle - \|d\|^2$, which equals the displayed difference of squared norms. If $H_i = \Pi_{\Delta_K}(G_i)$, projection optimality gives $R(G,Y) - R(H,Y) \geq n^{-1}\|G - H\|_F^2$. Hence row projection cannot reduce the affine gain, but it need not preserve the supplied mean. If an independently justified bound gives $\|\widetilde{q} - q\| \leq \epsilon$, a sufficient lower gain is $\|d\|^2 - 2\epsilon\|d\|$; this expression does not supply its own calibration.

Necessity of Equation 11 follows from $0 \leq Q_k - \rho q_k \leq 1 - \rho$. For sufficiency at $0 < \rho < 1$, set $r = (Q - \rho q)/(1 - \rho)$. The bounds imply $r \in [0,1]^K$, and its entries sum to one. Thus every q in the stated set extends to an admissible omitted composition. Our chip proportions are fractional. For one-hot targets, admissible assignments instead have integer class counts; with exact integer parent and query counts, the row-simplex/cap polytope is their convex hull (Appendix F.7).

Finite-iterate certificate. Let $T = \{q \in \Delta_K : l \leq q \leq u\}$ and $h_T(p) = \max_{q \in T} p^\top q$. For any returned simplex prediction G and any vector p, define $W_i = F_i - G_i - p$. Then every admissible fractional target satisfies

$$R(F,Y) - R(G,Y) \geq \frac{\|F\|_F^2 - \|G\|_F^2}{n} - 2h_T(p) - \frac{2}{n}\sum_i \max_k W_{ik}. \tag{16}$$

Indeed, $\langle W_i, Y_i\rangle \leq \max_k W_{ik}$ and $p^\top \bar{Y} \leq h_T(p)$. Use the mean-box correction from Dykstra for p, but recompute W from the returned arrays. This residual-aware bound remains valid without assuming exact solver convergence or a numerical normal-cone equality. The support function is computed by starting at l and assigning remaining probability mass to decreasing coordinates of p up to u. It is a dual lower bound, not necessarily the sharp worst-case gain at an early iterate. Thirty small independent linear-program checks, including one- and three-iteration stops, and ten separate SLSQP projections test the implementation.

A.5 Nested residual caps and the common one-pass return

This specialised result explains the common acquisition return. Its ingredients are established bounded adjustment and projection geometry (Heiser et al., 2020); the scalar-threshold representation follows from the capped-simplex KKT conditions (see also the uniform-cap treatment of Wang & Lu (2015)). The statement specialises these ingredients to nested residual caps and the common one-pass map. It explains the observed trajectory, without a claim of theorem priority or acquisition optimality.

Statement. Let $F \in \mathbb{R}^{m \times K}$ have simplex rows and $f = \bar{F}$. For componentwise $0 \leq a \leq b$ with $\sum_k a_k \geq 1$, define

$$T_j = \{q \in \Delta_K : q \leq j\}, \quad q_j = \Pi_{T_j} f, \quad G_{j,i} = \Pi_{\Delta_K}(F_i + q_j - f), \quad j \in \{a, b\}.$$

Write $d = q_a - q_b$, $e = \bar{G}_b - q_b$ and $U_i = G_{b,i} + d$. Then, for every row-simplex target Y with $\bar{Y} \in T_a$,

$$R(G_b,Y) - R(G_a,Y) \geq \|d\|^2 - 2\langle d, e\rangle + \frac{1}{m}\|U - G_a\|_F^2 \geq \|d\|^2 + \frac{1}{m}\|U - G_a\|_F^2 \geq 0. \tag{17}$$

Moreover $q_a = \Pi_{T_a} q_b$ and $G_{a,i} = \Pi_{\Delta_K}(G_{b,i} + d)$. The sharp direct gain is strictly positive iff $q_a \neq q_b$; if $q_a = q_b$, both returned arrays and their losses are identical. Here the sharp minimum may use fractional targets or any nonempty compatible hard-label subset.

Proof. The capped-simplex mean projection has the form $q_{j,k} = \min\{j_k, f_k + \lambda_j\}$, where the least nonnegative multiplier λ_j makes the coordinates sum to one. Nesting implies $\lambda_a \geq \lambda_b$. Let $I = \{k : q_{b,k} = b_k\}$. For $k \in I$, $q_{a,k} \leq a_k \leq b_k$, so $d_k \leq 0$; for $k \notin I$, $q_{b,k} = f_k + \lambda_b$. These two cases give

$$q_{a,k} = \min\{a_k, q_{b,k} + \lambda_a - \lambda_b\}, \qquad \langle d, q - q_a\rangle \geq 0 \quad (q \in T_a).$$

Thus mean projections compose. Set $H_i = F_i + q_b - f$, whose coordinates sum to one. Its row projection is

$$G_{b,i} = H_i - \theta_i \mathbf{1} + v_i, \qquad v_i = (\theta_i \mathbf{1} - H_i)_+.$$

The row threshold obeys $0 \le \theta_i \le \lambda_b$: at zero the positive parts sum to at least one, while at λ_b each positive part is at most F_{ik}. For $k \notin I$, $H_{ik} = F_{ik} + \lambda_b \ge \theta_i$, so $v_{ik} = 0$. Therefore positive truncation residuals occur only where $d_k \le 0$. Since $\sum_k d_k = 0$,

$$e = -\bar{\theta}\mathbf{1} + \bar{v}, \qquad \langle d, e\rangle = \langle d, \bar{v}\rangle \le 0.$$

Next, $U_i = G_{b,i} + d$ also sums to one, so its simplex threshold γ_i is nonnegative. If $v_{ik} > 0$, then $G_{b,ik} = 0$ and $d_k \le 0$, hence both $(U_{ik} - \gamma_i)_+$ and $(U_{ik} - v_{ik} - \gamma_i)_+$ vanish. On all other coordinates they agree. Using $F_i + q_a - f = U_i + \theta_i\mathbf{1} - v_i$, threshold $\theta_i + \gamma_i$ gives the same row return. This proves row composition. Finally, for $q = \bar{Y} \in T_a$,

$$\begin{aligned} R(G_b, Y) - R(U, Y) &= 2\langle d, q - \bar{G}_b\rangle - \|d\|^2 \\ &= \|d\|^2 + 2\langle d, q - q_a\rangle - 2\langle d, e\rangle \ge \|d\|^2 - 2\langle d, e\rangle. \end{aligned}$$

Row projection adds at least $m^{-1}\|U - G_a\|_F^2$, proving the bound. If $d \ne 0$ the bound is positive, so the returned arrays cannot agree. If $d = 0$, their definitions give identical arrays. This proves the stated equivalences. □

Acquisition scope and boundaries. For purchased sets $A_0 \subseteq A_1$, choose residual caps

$$b = \frac{NQ - \sum_{j\in A_0} Y_j}{m}, \qquad a = \frac{NQ - \sum_{j\in A_1} Y_j}{m}.$$

Nonnegative answers give $a \le b$; compatibility ensures total residual mass at least m. Both fractional-composition and hard-label queries qualify. Zero caps, redundant caps above one and active-set ties are permitted. The lower bounds in Equation 11 follow from upper caps and unit sum, so they do not change the result. Unrelated lower bounds, nonnested policies, nonuniform measurement weights, noisy observations and comparisons with a separately optimised free return are outside the statement. The result concerns the exact map; saved analyses retain numerical tolerances. In particular, a change of projected mean cannot be erased by the second projection under these assumptions.

Verification. Two independently implemented finite check families cover rational hard-label assignments and fractional LP/QP cases, with maximum bound excesses 2.09×10^{-16} and 2.78×10^{-17}. Some real-cap examples admit only the fractional interpretation. Exact counterexamples demonstrate failure outside nested upper caps. These checks support implementation consistency; the algebraic proof supplies the guarantee. Appendix I.4 gives the detailed enumeration counts.

B Recorded-membership geometry and finite correction

B.1 Retaining joint membership information

We use the single-subset notation and information set in Equation 4. The original interval event implies that the true scored targets satisfy this set after the recorded replacements. The event is marginal over the original uniform draw, not simultaneous across draws or parents.

Exactness within the stated hard-label model. Every compatible full-parent assignment induces the constraints above by summing removed rows. Conversely, feasible w can be divided equally among the r removed rows, and $t - v - w$ among the $N - m - r$ rows outside the union. These are simplex rows; an empty group has zero aggregate. Thus the lift exactly eliminates unscored rows from the full fractional system. With integer totals and endpoints, that system is a network-flow polytope: each row sends one unit to a class via its original-subset or complement node, original-subset class arcs have capacities $[a_k, b_k]$, and class sinks receive t_k. Integral capacities imply an integral flow polytope. Projecting its vertices onto S shows that (4) is the convex hull of compatible hard scored targets. This argument concerns one subset inside one parent, not arbitrary overlapping or weighted measurements.

Count-budget control and incremental comparison. An intermediate relaxation keeps $c_0 \in [a, b]$, $\mathbf{1}^\top c_0 = m$, deterministic bounds on v, and $\|v - c_0\|_1 \le 2r$. It retains a shared replacement budget but omits specific overlap membership. It is not an exact parent-compatibility characterisation: $N = 8, m = 3, r = 1, t = (3, 5), c_0 = v = (3, 0)$ passes these conditions, although the four distinct union rows cannot all belong to a class with only three parent members.

Both new repairs project about the same available replacement-aware return B, directly certifying $R(B, Y) - R(G, Y)$. Exact old-set membership of B precludes positive uniform improvement under the old information by Equation 3; old solver optimality about F is not required. Numerical membership alone is reported as an approximation, not an exact impossibility result.

B.2 Explaining and applying the membership restrictions

The fixed-B membership test in Equation 5 was already used by the joint solver. The follow-up makes its violated restrictions explicit and compares short corrections. All statements use the same original-sample confidence event as Equation 4.

Complete separation without enumerating class subsets. Write $g = v - h$. Eliminating the removed counts gives the following equivalent inequalities on row-simplex arrays:

$$h_k \le b_k, \qquad v_k \le t_k, \qquad g_k \le t_k - a_k, \tag{18}$$

$$\sum_{k \in J} (a_k - h_k) \le r \quad \text{for all } J, \tag{19}$$

$$\sum_{k \in J} h_k + \sum_{k \notin J} v_k \le \sum_{k \in J} b_k + \sum_{k \notin J} t_k - r \quad \text{for all } J. \tag{20}$$

Indeed, $\ell_k \le u_k$ expands into Equation 18 and the assumed $a_k \le b_k$. The sum of positive parts is the maximum of $\sum_J (a_k - h_k)$, and the sum of coordinate minima is the minimum of $\sum_J (b_k - h_k) + \sum_{k \notin J} (t_k - v_k)$. Hence the two sum conditions give Equations 19–20. At a particular prediction, take $J_{\text{low}} = \{k : a_k - h_k > 0\}$ and $J_{\text{up}} = \{k : b_k - h_k \le t_k - v_k\}$. Together with the $3K$ coordinate restrictions, these give at most $3K + 2$ candidates and expose a violation whenever exact membership fails. The candidates need not maximise the normalised gain score over every possible class subset. For integer inputs $0 \le a \le b \le t$, $\sum t = N$ and $0 \le r \le \min(m, N - m)$, the information set is nonempty precisely when $\sum a \le m \le \sum b$; no feasibility solve is required for this check.

Proof of Proposition 2. Let H and G be as in Equation 7. Projection onto the valid halfspace gives $\|B - Y\|_F^2 - \|H - Y\|_F^2 \ge \|B - H\|_F^2 = \delta^2 / \|\widetilde{A}\|_F^2$. Projection of H onto the row simplexes gives $\|H - Y\|_F^2 - \|G - Y\|_F^2 \ge \|H - G\|_F^2$. Adding and dividing by m proves Equation 8. This uses classical projection geometry, related to group-discrepancy updates in multicalibration (Hébert-Johnson et al., 2018). Row-centring minimises the normal's norm within its row-equality equivalence class; it does not imply realised-loss dominance. The final simplex projection can recross the cut, so full joint feasibility is not required.

For a concrete recrossing example, use three scored binary rows with total forest count one, and forest probabilities $(0, 1, 1)$. Projection onto the centred sum-at-most-one halfspace gives $(-1/3, 2/3, 2/3)$; row-simplex projection gives $(0, 2/3, 2/3)$. Its forest sum is $4/3$, yet the minimum mean two-class squared-loss improvement over compatible hard targets is $8/27$. Constraint satisfaction and improvement certification answer different questions.

Proof of the stored-output certificate. For $c = G - B$ and $\lambda \ge 0$, set $D(\lambda) = -\lambda\beta + \sum_i \min_k (c_{ik} + \lambda A_{ik})$. Equation 9 is the corresponding lower bound.

The proof uses $c : Y = (c + \lambda A) : Y - \lambda A : Y \ge D(\lambda)$ and expands the quadratic contrast. It requires only the target's simplex rows and the valid cut, not exact simplex membership of the stored predictions. We evaluate the original integer cut and binary-rational values of B, G exactly, using the fixed halfspace multiplier $\lambda = \delta / \|\widetilde{A}\|_F^2$ and the zero-multiplier alternative. Each scalar evaluation needs $O(mK)$ arithmetic operations, without an LP; this is not a bit-complexity or wall-time claim. The reported floating bound is rounded downward.

Each step begins with the same 10^{-8} numerical membership screen as the full method. A screened no-op is not a proof of exactly zero optimal opportunity. Otherwise, choose the largest positive $\delta^2 / \|\widetilde{A}\|_F^2$ among the candidates, with fixed family/class tie order, and certify the actual output. A nonpositive step certificate stops at the preceding accepted prediction. One and three steps are fixed alternatives. Three accepted direct bounds telescope between stored successive endpoints; the exact

sum is rounded only at the end. Returned domain and joint residuals are reported separately. Constructing dense candidate matrices adds cost beyond the scalar verifier and is included in measured runtimes.

Selection does not preserve a marginal coverage level. Let E be the original interval event and T a trigger with positive probability. From $\Pr(E^c) \leq \alpha$ one obtains only

$$\Pr(E^c \mid T) \leq \min\{1, \alpha / \Pr(T)\}. \tag{21}$$

A rare trigger therefore does not inherit 95% conditional coverage. Derived cuts and successive steps share E, whereas independent draws or parents do not automatically share one event. We report all-case and triggered/changed-case truth containment, retaining uncovered no-ops and updates. Empirical coverage of observed triggers is not a conditional theorem or individual-row guarantee.

C COVERTYPE SELECTION STUDIES: PROTOCOLS AND COMPLETE RESULTS

This appendix preserves the three distinct stages: the original coordinatewise count study and its null result, the joint-membership follow-up on inspected parents, and the fixed confirmation on unused parents (Appendix C.3). The later matched-loss, harm and timing summaries describe those saved comparisons; they do not replace the earlier primaries.

Question and relation to prior work. This bounded diagnostic asks whether known departures from uniform subset selection leave useful correction for a contemporary proportion-supervised predictor. PVC (Ma et al., 2026b) supplies the recent count-likelihood objective; a matched proportion-matching (PM) objective controls the backbone and optimisation. Sampled sub-bag uncertainty and hypergeometric modelling are established (Asanomi et al., 2023; Kubo et al., 2024). The construction below combines finite-population intervals, elementary count sensitivity and established adjustment. We claim neither a new general theorem nor reproduction of PVC's published benchmark ranking.

Inputs and information boundary. We reuse the Covertype source and size-64 feature-defined groups, excluding every row in the 2,048 previously selected groups. A fixed seed selects 896 remaining groups: 512 training, 128 validation and 256 evaluation, with disjoint rows across splits. Selection uses group metadata, not targets. All models use 54 inputs, one 64-unit ReLU hidden layer and seven outputs, train-only standardisation, Adam learning rate .001, weight decay .0001, batches of 64 groups and seeds 0–2. Corresponding seeds use identical initialisation and batch orders. Training sees features and parent counts; validation uses only its aggregate objective. Separate evaluator labels are opened after all predictors are frozen and hashed. Reusing the dataset and bag construction does not constitute an independent new dataset or project-wide untouched confirmation.

Objective adaptation and training duration. For logits z_{ik}, PVC uses $p_{ik} = \mathrm{sigmoid}(z_{ik})$ and minimises the sum over classes of the negative log probability of the observed parent count under independent Bernoulli probabilities $p_{1k}, \ldots, p_{Nk}$. Stable log-space dynamic programming computes this Poisson-binomial likelihood. PM minimises cross-entropy between the parent proportions and the mean softmax prediction. Both release softmax probability vectors, preserving the inspected PVC implementation's sigmoid-training/softmax-evaluation convention. Our tabular backbone, matched default initialisation and exact likelihood implementation differ from the authors' image pipeline; their classifier-bias initialisation is not reproduced. Small-bag enumeration, autograd and finite differences check the implemented objective.

The primary horizon is 160 epochs, with checkpoint selection every 20 epochs by each learner's native aggregate-validation loss. All six selected the final checkpoint. Before inspecting any test outcomes, we therefore fixed one 640-epoch sensitivity with identical settings and checkpoint rule; it exactly reproduced all original selected parameters at epoch 160. The primary remains the 160-epoch analysis. No later horizon, calibration or test-driven model selection was added.

Scored subsets and controls. Within each evaluation parent, one fixed-seed uniform draw retains $m = 16$ or 32 rows. For $r \in \{0, 1, 2, 4, 8\}$, exchange exactly r initially retained rows farthest from water for r initially omitted rows nearest water, breaking ties by row ID. Also score the fully

nearest-water subset, recording its realised replacement count relative to the original draw. This is a synthetic feature-selection intervention, not a geographical survey or an observed missingness frequency. Masks are dependent within a parent.

For each frozen prediction and selected subset, compare unchanged predictions, the parent mean imposed as an exact subset mean, deterministic bounds, sampling intervals, replacement-aware intervals and the evaluator-only exact-subset oracle. All gains refer to the unchanged prediction on those same scored rows. The parent-mean control is intentionally misspecified when subset composition differs.

Confidence-qualified information. Fix parent counts T_k, population size N, retained size m and K classes. The deterministic bounds are

$$D_k^- = \max(0, T_k - (N - m)), \qquad D_k^+ = \min(T_k, m).$$

Let H_k be the CDF of $\mathrm{Hypergeom}(N, T_k, m)$ and define

$$a_k = H_k^{-1}(\alpha/(2K)), \qquad b_k = H_k^{-1}(1 - \alpha/(2K)), \quad \alpha = .05,$$
$$L_k(r) = \max(D_k^-, a_k - r), \quad U_k(r) = \min(D_k^+, b_k + r).$$

Discrete quantiles and a union bound imply simultaneous class-count coverage at least $1 - \alpha$ for the original uniform draw. Each replacement changes each class count by at most one. Hence the same event implies $L_k(r) \le C_k \le U_k(r)$ for every class of the coupled selected subset. Intersecting with the always-valid deterministic bounds preserves that event.

This probability is over the original uniform draw, not necessarily conditional on its realised replacement count. The same event protects all coupled selections from that draw, but is not a joint guarantee over all 256 evaluation parents. Known coupling is essential; an arbitrary undocumented selected subset does not inherit the claim. Projection safety is conditional on truth belonging to the supplied set. No group-level statement protects every individual row.

Finite numerical returns. Let D be the product of row simplexes, $B = [L, U]$ the box of class counts, and $f(X) = \frac{1}{2}\|X - F\|_F^2$. For any multiplier λ, set $G = \Pi_D(F - \lambda)$ and

$$d(\lambda) = f(G) + \lambda^\top \sum_i G_i - \sigma_B(\lambda), \qquad \sigma_B(\lambda) = \sum_k [\max(\lambda_k, 0)U_k + \min(\lambda_k, 0)L_k].$$

Strong convexity of $f(X) + \lambda^\top \sum_i X_i$ on D gives, for every compatible Y,

$$f(Y) + \lambda^\top \sum_i Y_i \ge f(G) + \lambda^\top \sum_i G_i + \tfrac{1}{2}\|Y - G\|_F^2, \qquad R(F, Y) - R(G, Y) \ge 2d(\lambda)/m.$$

Thus the bound concerns the actual returned G, not a presumed exact optimiser. Batched Dykstra uses at most 2,000 iterations. The implementation additionally requires count residual and absolute objective–dual gap at most 10^{-8}; empty sets, nonconvergence or a negative bound return F and zero certificate. Reported nonzero bounds subtract 10^{-10}. These floating-point checks are not interval-arithmetic verification. We compare an independently formulated direct worst-case LP and an SLSQP projection on 24 metadata-fixed cases per horizon: the first three sorted parent IDs, both m, $r = 0/4$, sampling/replacement sets and PVC seed 0. The first evaluation stopped on an SLSQP line-search flag despite objective agreement within 1.1×10^{-12}. One documented checker repair retained this failure and the unchanged production corrections. The final gate uses an independent LP dual bound, including its box-residual term, to bound the returned projection's first-order gap by 10^{-7}; LP primal feasibility and primal–dual gap must meet 10^{-8}. SLSQP status remains reported and its objective must agree within 10^{-7}. A status flag alone is not the optimality test. Test losses assess outcomes and never construct non-oracle corrections.

Fixed analysis. The D1 primary is replacement-aware minus deterministic realised gain for PVC at $m = 16, r = 4$, averaging its three seeds within each of 256 parents and then parents equally. A descriptive paired parent bootstrap uses 2,000 resamples. This is a separate estimand from both earlier primaries. The 640-epoch counterpart is a sensitivity. Coverage counts each parent once, not once per fitted seed; harmed-case and certificate-attainment indicators are computed before equal-seed averaging. Attainment at .001 is a fixed analysis threshold, not an operational requirement. Every declared support, replacement level, learner and seed is retained.

Results and the null comparison. The fixed primary is exactly zero, with zero median and a degenerate empirical bootstrap interval $[0, 0]$: all 256 seed-averaged paired differences are zero. All 768 PVC returned arrays are exactly identical between the two corrections. This is not cancellation or a primary-setting convergence failure. A post-result saved-array diagnostic finds narrower replacement bounds in 175 parents, but the deterministic returns already satisfy them within 2.4×10^{-10}. The additional constraints are inactive at these numerical returns. A degenerate resampling interval does not imply zero uncertainty on future groups. The 640-epoch sensitivity gives additional mean gain 0.00000000, with interval $[0.00000000, 0.00000000]$. Its 768 of 768 PVC pairs are exactly identical. Neither horizon replaces the earlier acquisition or bridge primary. Seven PVC and 22 PM paired cases at the longer horizon return the original prediction under the certificate-margin fallback. Their zero certificates remain valid although those unchanged predictions can violate shared zero-count constraints; the newly tightened bounds remain slack. All other primary-setting returns satisfy the replacement bounds within 2.4×10^{-10}.

For original-horizon PVC at $m = 16$, the extra replacement-aware gain over deterministic correction is .006127 at $r = 0$, .001702 at $r = 1$, approximately .000009 at $r = 2$, and zero at $r = 4$. At $m = 32, r = 4$, extra gain remains .002553 for PVC but only .000096 for PM. These were declared secondary settings, not replacements for the primary. With fully nearest-water selection, unexpanded sampling intervals cover only 78.125%/71.09375% of parents for $m = 16/32$, although mean gains remain positive. PVC harmed-case fractions there are 2.99%/4.30%; PM fractions are 5.08%/6.38%. Mean benefit therefore does not verify the information assumption. Replacement-aware coverage is 100% on the altered supports in this sample; at the original random $m = 32$ support it is 98.83%. These are descriptive frequencies under one fixed draw per parent.

The two horizons contain 92,160 scored output records each, with 113 and 204 nonconverged no-ops, respectively. None occurs in the primary correction pair. Every prescribed mathematical check passes in the two complete runs (24 per horizon); SLSQP nevertheless reports 8 and 8 unsuccessful status flags, retained with the agreeing objectives and independent LP bounds. The failed first attempt made 13 additional check calls before stopping. These are repeated numerical comparisons on fixed cases, not 61 independent data replications. The other no-op conditions and all per-case diagnostics remain in the saved tables. Finite output gains are not asserted to be exact information optima.

Baseline quality and training sensitivity. Table 4 reports unchanged-prediction loss and accuracy on each scored support; every correction's loss is that baseline minus its tabulated gain. Accuracy is descriptive, and our count-objective/softmax-output adaptation does not establish a general ranking between learners. Figure 4 reports all aggregate-validation curves. At the longer horizon, PVC selects epochs 480, 480 and 340; PM selects 620, 640 and 620. One PM seed still selects the longer cap. No further training search is made.

Table 4: Original coordinatewise study: unchanged predictions on every declared support. Entries are squared probability-vector loss (lower is better) and accuracy in percent (higher is better), equally averaging the three seeds and 256 groups. These descriptive adaptation results are not a published-method ranking.

		PVC 160		PM 160		PVC 640		PM 640	
m	Selection	Loss	Acc.	Loss	Acc.	Loss	Acc.	Loss	Acc.
16	$r = 0$	0.45473	70.18	0.44821	68.46	0.46199	69.99	0.46173	68.89
16	$r = 1$	0.46791	69.30	0.45646	67.79	0.47330	69.31	0.46928	68.05
16	$r = 2$	0.47288	68.96	0.46076	67.49	0.47797	68.96	0.47416	67.59
16	$r = 4$	0.48694	67.90	0.47273	66.63	0.48945	68.26	0.48658	66.57
16	$r = 8$	0.51543	66.01	0.50169	64.65	0.51486	66.62	0.51450	64.71
16	Full	0.52341	65.31	0.50567	64.27	0.51950	66.45	0.51603	64.42
32	$r = 0$	0.45246	69.95	0.44065	68.98	0.45776	70.32	0.44951	69.90
32	$r = 1$	0.45845	69.57	0.44535	68.60	0.46302	70.00	0.45482	69.44
32	$r = 2$	0.46229	69.37	0.44866	68.34	0.46640	69.82	0.45805	69.12
32	$r = 4$	0.46656	69.14	0.45269	68.14	0.46977	69.73	0.46323	68.67
32	$r = 8$	0.47276	68.63	0.45908	67.56	0.47422	69.29	0.46904	68.08
32	Full	0.48746	67.70	0.47275	66.65	0.48631	68.39	0.48143	67.01

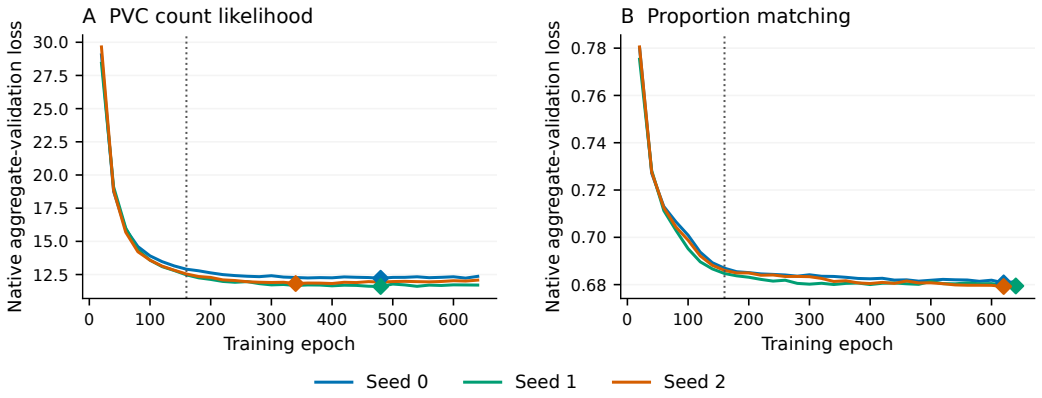


Figure 4: Original coordinatewise study: native aggregate-validation losses for the fixed longer-horizon sensitivity. Diamonds mark selected checkpoints; the dotted line marks the original 160-epoch horizon. The two objectives use different vertical scales and are not directly comparable. Selection uses no individual validation or evaluation targets. The first 160 epochs exactly reproduce the original selected parameters for all six fits.

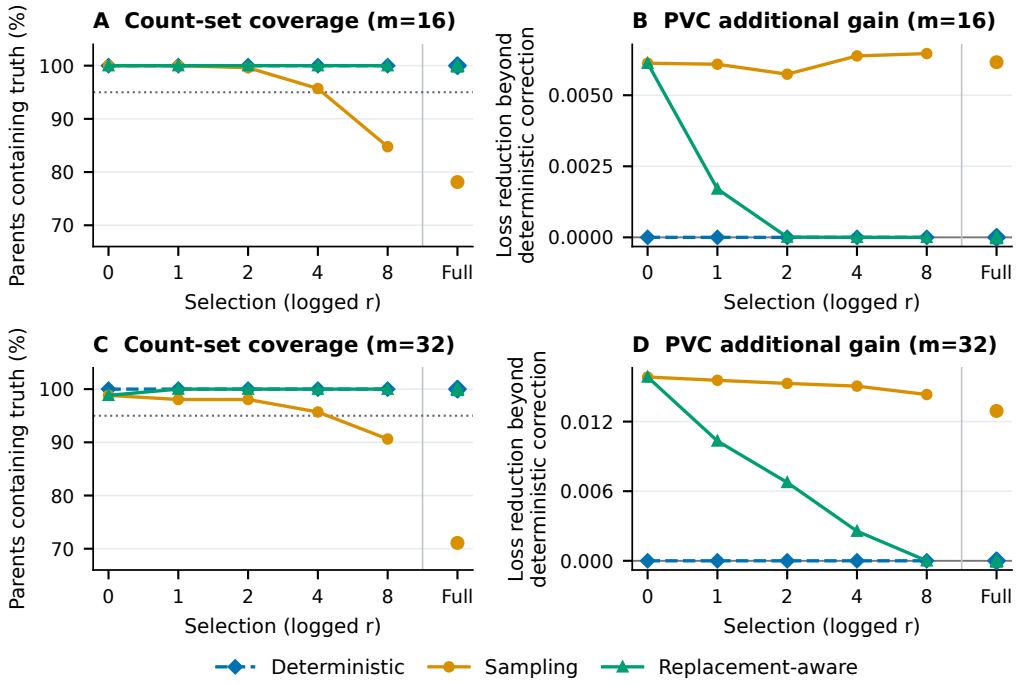


Figure 5: Original coordinatewise study: count-set validity and additional utility at the original 160-epoch horizon, for $m = 16$ (A, B) and $m = 32$ (C, D). (A, C) Fraction of 256 parents whose full class-count vector lies in the set (dotted: 95%). (B, D) PVC equal-seed/equal-parent gain beyond deterministic correction. The primary is PVC at $m = 16, r = 4$. Full nearest-water selection has variable logged replacements and is separate from the fixed-r curve. PM, complete parent-mean/oracle controls and the 640-epoch sensitivity remain in Appendix C. These are dependent descriptive comparisons.

C.1 JOINT-INFORMATION PROTOCOL AND ORIGINAL RESULTS

The joint set and its proof appear in Appendix B.1; finite short-correction mathematics is in Appendix B.2.

Locked follow-up and verification. We reuse all fitted D1 predictions. The primary is PVC, 160 epochs, original masks, $m = 16, r = 4$, averaged over three seeds within each of 256 parents. Secondary original-mask cells are $(m, r) = (16, 0), (16, 2), (32, 0), (32, 4), (32, 8)$ for PVC and PM. A 640-epoch sensitivity uses only $m = 16, r = 4$. Five additional independent initial draws reuse the 160-epoch PVC models at that setting; seeds and draws are averaged within parent. Five draws were fixed from synthetic runtime before new evaluation outcomes, reducing the proposed twenty to remain within a four-hour cap. Repeated draws are randomisation replications on the same parents, not independent datasets or confidence sets to intersect. This is a follow-up on inspected Covertype data.

SLSQP uses analytic gradients, at most 500 iterations and $\texttt{ftol} = 10^{-14}$. Recomputed candidate and LP primal residuals must be at most 10^{-8}; normalised first-order and LP primal–dual gap targets are 10^{-7}. An independent LP supplies a direct loss bound, including bounded-variable correction for dual stationarity residuals. Exact binary-rational arithmetic evaluates this scalar on the stored predictions, constraints and duals, rounding its float representation down. This certifies a bound for the supplied set, not an exact QP optimum. Updates below a 10^{-10} certificate threshold and failed numerical checks return B unchanged with a labelled zero certificate. Numerical-compatible no-ops have an uncomputed optimisation gap. Independent tiny-system enumeration, membership, projection and perturbed-dual checks precede real outcomes. Saved-output verification distinguishes proof arithmetic from solver replay.

Results and their concentration. The original joint follow-up primary is $+.000814757$, with direct mean certificate $.000099486$, parent median zero and descriptive 95% parent-bootstrap interval $[0, .002422579]$. Only 5/768 PVC returns change, across three parents; the largest parent contributes 97.91% of positive mean gain. The count-budget relaxation returns all 768 incumbents unchanged. No primary incumbent violates the old set numerically, and no new primary solve fails. At 640 epochs the gain is $.000654533$ (interval $[0, .001914027]$), again with five changed returns. Seven old-set-invalid PVC incumbents remain unchanged under the full/count procedures; these are not evidence of exact old-set membership.

The five further draws yield gain $.000062353$, mean certificate $.000013066$ and interval $[.000009152, .000144741]$, with 14/3,840 changed returns and zero median. Joint truth membership is 1,279/1,280 parent–draw cases; all three model outputs for the uncovered case remain unchanged. Neither the positive mean nor its descriptive interval establishes a population-wide or practically large benefit. Classification accuracy is unchanged for the original primary and changes by $-.004883$ percentage points over further draws. Individual loss increases above .01 occur in .024414%/.034180% of scored row–model cases (original/further draws), despite no observed harmed group mean. Squared-loss safety is not an accuracy or individual-instance guarantee.

Baseline computation and verification limits. Across the original 160-epoch cells, 21 old-incumbent cases violate old-set constraints at 10^{-8} (42 representation comparisons). A separately checked coordinatewise correction about the original predictor removes that numerical discrepancy; both new representations then return it unchanged in all 42 sensitivities. Thus gains in these cases cannot be uniquely credited to extra information. At 640 epochs, 29 old-incumbent cases generate 58 representation comparisons; polishing also returns a certificate-margin no-op, so that sensitivity does not establish exact old-set feasibility. All these actual-incumbent outcomes remain in Table 7.

Saved-output checks cover all actual-incumbent records, baseline-computation sensitivities, masks and hashes. Candidate bounds, including rejected candidates, receive exact scalar checks; sensitivities lacking dual payloads receive only loss/realised-bound replay. Optimisers and training are not rerun. Appendix I.4 gives the check counts and corruption tests, and Table 28 lists what is supplied for each component.

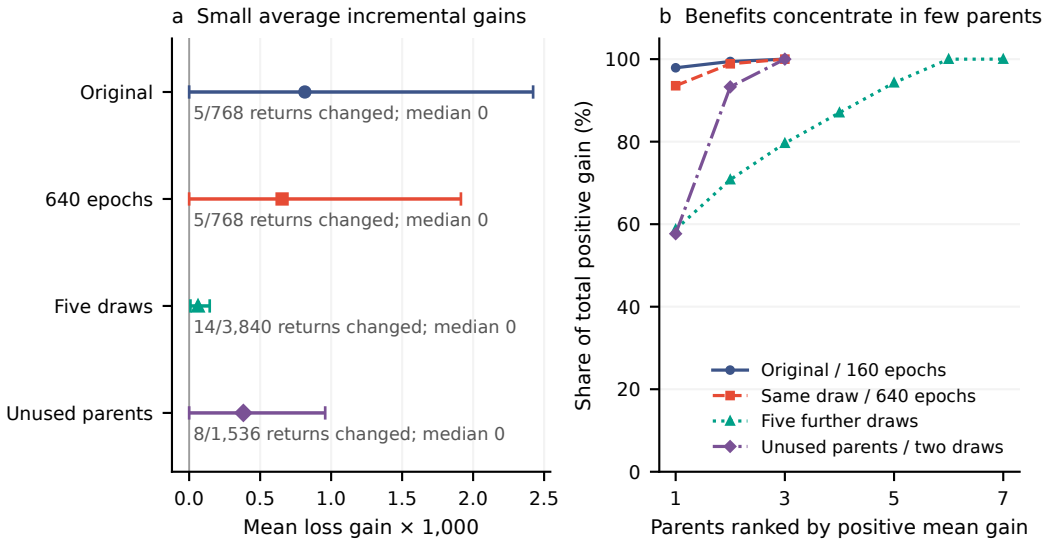


Figure 6: Membership information can help while gains remain sparse. PVC at $m = 16, r = 4$; each repair is centred on the same replacement-aware incumbent. (a) Mean incremental loss gain and descriptive parent-bootstrap intervals; all cases enter the means. (b) Concentration of positive parent-mean gains; zero-gain parents are omitted from the ranking only. Seeds and draws are averaged within parent. The first three comparisons reuse 256 parents; the last uses 256 unused parents and two draws from the same source. The count-budget control is unchanged in all four comparisons.

C.2 Attribution and short-correction comparisons

Cached attribution audit. Reconstructing the membership tests for all 14,592 cached joint model cases agrees with stored residuals within 1.1×10^{-14}. At original 160-epoch PVC $m = 32, r = 4$, the 70 numerically old-compatible/joint-incompatible changed predictions span 26 parents; seven old-incompatible changes span another three. Their total gains over seed outputs are 1.787393274 and .301989920, respectively. The first stratum separates newly active membership restrictions from detected old-set violations at the 10^{-8} screen; it is not an exact old-set membership proof. The original joint follow-up primary has five such triggers across three parents. Across five further draws, 14 such triggers occupy eight parent–draw events across seven parents, all with initial-event and joint-target coverage. Overall, those further draws have initial-event coverage 1,274/1,280 and joint-target coverage 1,279/1,280: compatible completion can exist even when the true original-sample count misses its interval. All denominators distinguish parents, draws and repeated model outputs.

Cached short-correction comparisons. At PVC $m = 16, r = 4$, one/three-cut mean gains over the actual incumbent are .000303/.000638 on the original 160-epoch masks, .000296/.000534 at 640 epochs, and .000031934/.000045300 over the five further draws. Seeds and draws are averaged within each of 256 parents; all medians are zero. On the original masks, both cuts and full joint change 5/768 outputs across three parents; further draws give 14/3,840 changes across seven parents. Two of those five draws have zero gain for every method and remain in the pooled result. At 640 epochs, cuts change 12/768 outputs across eight parents, including the seven old-invalid outputs left unchanged by full joint. Only three parent means exceed 10^{-10} under either cut: the additional changes do not imply material additional benefit. Three cuts have lower mean gain than full joint in all three comparisons, but exceed it on one parent in the original comparison and one after five-draw pooling. This is consistent with uniform-guarantee optimality, which does not order realised gains case by case.

Cached computation comparison. The residual-selected repeated timings use 214 original-horizon cases across both learners and all declared regimes, 42 longer-horizon cases across both

Table 5: Original coordinatewise study: complete correction gains at $m = 16$, for both learners and training horizons. Positive means lower Brier loss than unchanged predictions. Coverage is observed truth membership (percent), counted once per parent and identical across learners/horizons. Parent-mean and unexpanded sampling controls can exclude truth; their gains are not unconditional safety claims. Corrected losses follow from Table 4.

Selection	Correction	Coverage	PVC 160	PM 160	PVC 640	PM 640
$r = 0$	Parent mean	6.64	0.056429	0.049102	0.056648	0.050180
	Deterministic	100.00	0.006415	0.009741	0.005716	0.008865
	Sampling	100.00	0.012542	0.013664	0.010573	0.011265
	Replacement	100.00	0.012542	0.013664	0.010573	0.011265
	Subset oracle	100.00	0.079750	0.073166	0.081067	0.076294
$r = 1$	Parent mean	4.30	0.057275	0.046463	0.056267	0.049340
	Deterministic	100.00	0.005785	0.009082	0.005030	0.008648
	Sampling	100.00	0.011876	0.012398	0.009843	0.011047
	Replacement	100.00	0.007487	0.010305	0.005922	0.009101
	Subset oracle	100.00	0.081822	0.071766	0.082089	0.075581
$r = 2$	Parent mean	6.25	0.055933	0.044893	0.054646	0.049398
	Deterministic	100.00	0.006072	0.009791	0.005124	0.009994
	Sampling	99.61	0.011813	0.013106	0.010098	0.012618
	Replacement	100.00	0.006081	0.009791	0.005124	0.009994
	Subset oracle	100.00	0.086012	0.075707	0.084815	0.080147
$r = 4$	Parent mean	4.69	0.057779	0.042989	0.053728	0.046515
	Deterministic	100.00	0.007466	0.012103	0.006011	0.012176
	Sampling	95.70	0.013849	0.016572	0.011459	0.016239
	Replacement	100.00	0.007466	0.012103	0.006011	0.012176
	Subset oracle	100.00	0.102667	0.088295	0.098983	0.092395
$r = 8$	Parent mean	3.12	0.049725	0.033330	0.044015	0.037799
	Deterministic	100.00	0.009565	0.015990	0.007515	0.014434
	Sampling	84.77	0.016031	0.021435	0.011817	0.019207
	Replacement	100.00	0.009565	0.015990	0.007515	0.014434
	Subset oracle	100.00	0.131094	0.113546	0.126749	0.119976
Full	Parent mean	3.12	0.037934	0.018153	0.030295	0.023938
	Deterministic	100.00	0.010786	0.017635	0.009387	0.016537
	Sampling	78.12	0.016949	0.022223	0.013534	0.019903
	Replacement	100.00	0.010786	0.017635	0.009387	0.016537
	Subset oracle	100.00	0.139715	0.118408	0.133454	0.126980

learners, and 14 further-draw PVC cases, each repeated three times. Selection uses the pre-correction joint residual, not old-set validity or observed gain. Median complete-call times are 3.25–5.03 ms for one cut, 9.47–15.93 ms for three cuts and 9.31–15.09 ms for full joint across these batches. One cut is cheaper; three cuts have lower mean but slightly higher median time than full joint in every batch. All-case timings include the many screened no-ops and are reported separately. Effects use saved full/count outputs; fresh comparator calls on identical inputs supply timing and status, but their endpoint arrays were not retained for an equality check. These timings do not establish an exactly verified frontier for the same timed output arrays.

C.3 Fixed unused-parent confirmation protocol and results

The confirmation selects 256 previously unused 64-row Covertype parents from the existing wilderness/elevation grouping. All 2,048 historical and 896 later training, validation and evaluation groups are excluded by both parent ID and individual row ID. Selection uses a fixed random ordering of the eligible IDs, without labels or predictions. It is a new-parent test on the same source and constructed grouping, not an independent dataset. All six original 160-epoch PVC/PM checkpoints and train-only transforms remain frozen; inference must reproduce the old stored predictions within 10^{-12} before new outcomes are evaluated.

Two initial uniform draws per parent are coupled across $(m, r) = (16, 2), (16, 4), (32, 4), (32, 8)$. The fixed near-water replacement rule and confidence intervals are unchanged. The primary is the full-joint gain over the actual replacement-aware incumbent for PVC at $m = 16, r = 4$, averaging

Table 6: Original coordinatewise study: complete correction gains at $m = 32$, for both learners and training horizons. Positive means lower Brier loss than unchanged predictions. Coverage is observed truth membership (percent), counted once per parent and identical across learners/horizons. Parent-mean and unexpanded sampling controls can exclude truth; their gains are not unconditional safety claims. Corrected losses follow from Table 4.

Selection	Correction	Coverage	PVC 160	PM 160	PVC 640	PM 640
$r = 0$	Parent mean	6.25	0.055190	0.044963	0.054222	0.046820
	Deterministic	100.00	0.009617	0.015207	0.007019	0.013079
	Sampling	98.83	0.025457	0.021296	0.022221	0.019182
	Replacement	98.83	0.025457	0.021296	0.022221	0.019182
	Subset oracle	100.00	0.064358	0.054334	0.063876	0.057481
$r = 1$	Parent mean	7.03	0.055706	0.044559	0.054219	0.046669
	Deterministic	100.00	0.009575	0.015592	0.006980	0.013533
	Sampling	98.05	0.025133	0.021111	0.021871	0.019333
	Replacement	100.00	0.019903	0.018570	0.016695	0.016722
	Subset oracle	100.00	0.065195	0.053729	0.064148	0.057056
$r = 2$	Parent mean	7.42	0.055172	0.043485	0.053182	0.045101
	Deterministic	100.00	0.009189	0.015418	0.006490	0.012952
	Sampling	98.05	0.024478	0.020720	0.021159	0.018665
	Replacement	100.00	0.015943	0.017051	0.012510	0.014463
	Subset oracle	100.00	0.065808	0.053712	0.064417	0.056056
$r = 4$	Parent mean	7.42	0.052923	0.040928	0.050270	0.043363
	Deterministic	100.00	0.009929	0.016537	0.007258	0.014992
	Sampling	95.70	0.024985	0.021464	0.020872	0.019948
	Replacement	100.00	0.012482	0.016633	0.008814	0.015046
	Subset oracle	100.00	0.068455	0.056079	0.066353	0.058937
$r = 8$	Parent mean	4.69	0.049984	0.036613	0.046142	0.040068
	Deterministic	100.00	0.012070	0.018273	0.009889	0.016931
	Sampling	90.62	0.026403	0.022676	0.022297	0.021845
	Replacement	100.00	0.012070	0.018273	0.009889	0.016931
	Subset oracle	100.00	0.073084	0.059267	0.069605	0.063460
Full	Parent mean	5.08	0.045931	0.031083	0.040126	0.033161
	Deterministic	100.00	0.013743	0.020812	0.011473	0.017997
	Sampling	71.09	0.026658	0.024864	0.021602	0.022620
	Replacement	100.00	0.013743	0.020812	0.011473	0.017997
	Subset oracle	100.00	0.083788	0.068715	0.079220	0.072055

seeds and both draws within each parent, then parents equally. One/three cuts, PM and the other regimes are prespecified secondary comparisons. No historical primary is replaced. Bootstrap intervals resample parents 2,000 times and remain descriptive, conditional on this construction and the frozen models. No-op, empty, old-incompatible, failed-certificate and uncovered cases remain in the denominators.

The fixed sample size targets modest recurrence precision. Under a parent-level binomial planning model, zero events among 256 would imply a one-sided 95% upper rate of 1.16%; 13 events would yield a Wilson interval of approximately [2.99%,8.49%]. This does not establish power for sparse mean gains: the old primary's parent standard deviation gives a nominal mean half-width .00156 at 256, larger than its mean .000815. Two draws and six predictors are dependent observations within each parent. The study neither grows its sample after a disappointing result nor treats rare triggered-case coverage as a precise conditional guarantee.

The historical runtime comparison uses identical stored incumbents and information, including construction, numerical screening, correction and certificate arithmetic within each method's timer. A fixed cyclic order covers all five alternatives; three additional timing repetitions use cases selected only by pre-correction joint residual, not realised gain. Common interval computation and evaluation/I/O are outside individual method timers and inside total-stage runtime. Stored full/count outputs provide the effect comparison; fresh calls provide the timing reference. CPU threads are fixed at two with no GPU and 12 GiB maximum memory. Stage hashes, immutable outputs and

the numerical/environment scope are retained in the private historical archive. The later endpoint-retaining follow-up is distinguished in Appendix C.6.

Confirmation results. The fixed PVC $m = 16, r = 4$ full-joint mean is .000381794321, with descriptive parent-bootstrap interval $[0, .000958143197]$, median zero and 253/256 zero-gain parents. The eight changed model–draw returns occupy three parents: one changes all three seeds in the first draw, another all three seeds in the second, and the third changes one seed in both. The largest parent supplies 57.68% of positive pooled-parent gain. Count budget changes none. One/three cuts have mean gains .000252440994 and .000355807497; each changes the same eight primary returns. This is evidence of sparse recurrence on unused parents from the same construction, not an independent dataset or a precisely established positive population mean.

Table 7 pools every declared draw within parent and reports all method/setting combinations. Table 9 retains both fresh draws separately. Across the fresh settings, the joint procedure and each cut method change 303/12,288 model cases. Under the 10^{-8} numerical screen, 297 changed incumbents are old-compatible and six are old-incompatible; the six count-budget changes are all in the latter stratum. Thus those six gains are not evidence of membership-specific information. All saved coordinate, count and joint sets contain their scored reference targets. The initial interval event holds in 254/256 parents for the second draw at $m = 32$ and in every other draw/sample-size cell; paired replacement counts reuse that initial draw. Coverage rows and model seeds are not independent replicates.

Positive group gain does not protect every row. The fresh primary has 25 prediction–row pairs with squared-loss rise above 10^{-8} among $1,536 \times 16 = 24,576$ scored prediction–row pairs (0.102%). These are model/draw-specific pairs, not distinct physical locations. Table 9 reports all-case row-harm rates for each fresh cell; all six count-budget updates have zero observed row harm at this threshold.

Table 7: All declared short-correction comparisons. Gains are mean squared-vector loss reductions over the same saved incumbent, multiplied by 10^3. Three seeds and every declared draw are averaged within parent, then 256 parents equally. Original uses six regimes and both learners; 640 epochs is the fixed primary-setting sensitivity; five further draws use PVC only; unused parents use four regimes, two draws and both learners. No-op and old-incompatible cases remain included. Full joint has no casewise realised-loss dominance claim.

Study	Learner	m/r	Keep B	Count	One cut	Three cuts	Full joint
Original	PM	16/0	0.0000	0.0000	0.0000	0.0000	0.0000
Original	PVC	16/0	0.0000	0.0000	0.0000	0.0000	0.0000
Original	PM	16/2	0.0000	0.0000	1.0230	1.8644	2.2008
Original	PVC	16/2	0.0000	0.0000	0.6248	1.2460	1.5534
Original	PM	16/4	0.0000	0.0000	0.3993	0.8068	0.9961
Original	PVC	16/4	0.0000	0.0000	0.3030	0.6384	0.8148
Original	PM	32/0	0.0000	0.1652	0.0082	0.0225	0.1652
Original	PVC	32/0	0.0000	0.2239	0.0141	0.0331	0.2239
Original	PM	32/4	0.0000	0.2232	0.7891	1.2480	1.4264
Original	PVC	32/4	0.0000	0.3932	1.5301	2.3873	2.7206
Original	PM	32/8	0.0000	0.0542	0.0574	0.1032	0.1488
Original	PVC	32/8	0.0000	0.1137	0.4338	0.6279	0.7259
640 epochs	PM	16/4	0.0000	0.0000	0.3584	0.5943	0.7824
640 epochs	PVC	16/4	0.0000	0.0000	0.2961	0.5339	0.6545
Five draws	PVC	16/4	0.0000	0.0000	0.0319	0.0453	0.0624
Unused	PM	16/2	0.0000	0.0000	0.6167	1.0117	1.0429
Unused	PVC	16/2	0.0000	0.0000	0.5099	0.8701	0.8564
Unused	PM	16/4	0.0000	0.0000	0.0795	0.1403	0.1354
Unused	PVC	16/4	0.0000	0.0000	0.2524	0.3558	0.3818
Unused	PM	32/4	0.0000	0.0283	0.5387	0.8130	0.8918
Unused	PVC	32/4	0.0000	0.1096	0.9855	1.5100	1.6788
Unused	PM	32/8	0.0000	0.0000	0.0353	0.0516	0.0511
Unused	PVC	32/8	0.0000	0.0267	0.1241	0.1710	0.2088

Table 8: Historical complete-call times in milliseconds, shown as mean/median; counts are calls per method. All inputs receive one call. Selected inputs receive three further calls and are selected by pre-correction joint residual, without truth or realised gain. Rows aggregate the complete study settings in Table 7; they are not primary-setting-only timings. Common intervals and evaluator/I/O work are excluded from method timers. One cut costs less than full joint in these comparisons. Three cuts have lower means but slightly higher selected-case medians, so there is no consistent speed advantage. These timed comparator arrays were not saved; their effect and timing endpoint equality cannot be independently checked. Table 12 reports the separate endpoint-retaining follow-up.

Study	Inputs	Calls	Keep B	Count	One cut	Three cuts	Full joint
Original	All	9216	0.001/0.001	0.445/0.265	0.302/0.204	0.499/0.190	0.621/0.269
Original	Selected	642	0.003/0.003	4.989/0.333	4.356/5.033	12.962/15.935	15.632/15.090
640 epochs	All	1536	0.001/0.001	0.555/0.242	0.315/0.218	0.507/0.205	0.526/0.246
640 epochs	Selected	126	0.002/0.001	10.231/13.423	3.540/3.648	9.886/10.036	10.079/9.575
Five draws	All	3840	0.001/0.001	0.234/0.232	0.213/0.196	0.230/0.185	0.260/0.216
Five draws	Selected	42	0.002/0.002	0.293/0.293	3.237/3.250	9.377/9.470	10.388/9.309
Unused	All	12288	0.001/0.001	0.288/0.270	0.334/0.214	0.592/0.200	0.671/0.278
Unused	Selected	909	0.003/0.003	0.760/0.346	4.993/5.315	15.112/16.299	16.592/15.794

Table 9: All unused-parent cells, preserving draws separately. Gains are multiplied by 10^3. Each cell has 768 model returns (three seeds for 256 parents); count/joint changed-return counts share this denominator. The parent column counts parents with a changed joint return, not independent model records. Harm is the percentage of individual prediction–row pairs whose joint-return squared loss rises by more than 10^{-8}, with unchanged returns included. Every parent median is zero. Truth containment of all scored sets is 100%; the original interval event is 254/256 for draw1 at $m = 32$ and 256/256 elsewhere.

Draw	Learner	m/r	Count gain	Joint gain	Changed C/J	Parents J	Harm (%)
0	PM	16/2	0.0000	0.8908	0/13	7	0.236
0	PVC	16/2	0.0000	0.7271	0/15	6	0.521
0	PM	16/4	0.0000	0.0074	0/2	2	0.024
0	PVC	16/4	0.0000	0.3013	0/4	2	0.114
0	PM	32/4	0.0000	0.9251	0/37	16	1.388
0	PVC	32/4	0.0534	1.6940	1/54	25	2.389
0	PM	32/8	0.0000	0.0822	0/7	3	0.203
0	PVC	32/8	0.0534	0.2169	1/12	7	0.342
1	PM	16/2	0.0000	1.1949	0/9	3	0.122
1	PVC	16/2	0.0000	0.9856	0/17	7	0.659
1	PM	16/4	0.0000	0.2633	0/6	2	0.073
1	PVC	16/4	0.0000	0.4623	0/4	2	0.090
1	PM	32/4	0.0567	0.8584	1/43	19	1.457
1	PVC	32/4	0.1658	1.6637	3/67	27	3.137
1	PM	32/8	0.0000	0.0200	0/2	2	0.073
1	PVC	32/8	0.0000	0.2007	0/11	6	0.427

C.4 Matched original, incumbent and adjusted losses

This retrospective summary uses the same saved PVC $(m, r) = (16, 4)$ confirmation outputs. The original prediction F, replacement-aware incumbent B and full-joint output G use identical scored instances and reference labels. Average the three seeds and two draws within each parent, then all 256 parents equally. The decomposition $R(F) - R(G) = [R(F) - R(B)] + [R(B) - R(G)]$ holds exactly before floating-point evaluation. The computed B-to-G mean reproduces the saved primary exactly; the floating decomposition residual is 8.67×10^{-19}. This does not introduce a new primary or uncertainty estimate.

Table 10: Matched confirmation risks and gains in raw squared-loss units. Every mean includes all 256 parents and 1,536 model–draw outputs. The existing correction supplies most of the measured total reduction. Only the last row is the fixed confirmation primary; its uncertainty and sparse incidence remain as in Table 2.

Quantity	Mean
$R(F)$	0.497639313952
$R(B)$	0.491823419948
$R(G)$	0.491441625626
$R(F) - R(B)$	0.005815894004
$R(B) - R(G)$	0.000381794321

C.5 Retrospective frequency, certificate and harm summaries

These summaries read the same archived returns; they do not rerun corrections or change any primary, setting, weight or threshold. Seeds are averaged within draw, draws within parent and parents equally. Conditional output gains instead average the changed model–draw outputs, and cannot be interpreted as future benefit in a prospectively identifiable subgroup. All unchanged cases remain in the unconditional means. The fresh primary has eight changed model–draw outputs across three parent groups. Its mean conditional gain is .07330451; the mean among changed parents, retaining each parent's complete seed/draw averaging, is .03257978. The different values reflect different denominators. At row loss rise $> 10^{-8}$, 25 harmed observations have mean increase .04297416 and maximum .19647641. The same 25 exceed .001; 17 exceed .01. At strict > 0, 31 observations are counted, including six below the original 10^{-8} reporting margin. No new harm threshold is introduced.

Table 11: Full-joint unused-parent results for every learner/regime. Gain and certificate are multiplied by 10^3; conditional gain is in raw loss units. Changed counts use the original 10^{-8} coordinate margin, out of 1,536 outputs and 256 parents per cell. Harm percentages at loss rise $> 10^{-8}$ use all rows / rows within changed outputs. These denominators count repeated prediction–row observations. Only PVC 16/4 is the fixed confirmation primary; every parent median is zero.

Learner	m/r	Gain	Bound	Outputs	Parents	Cond. gain	Harm (%)
PVC	16/2	0.856379	0.083721	32	8	0.04111	0.590/28.32
PM	16/2	1.042878	0.141070	22	7	0.07281	0.179/12.50
PVC	16/4	0.381794	0.024884	8	3	0.07330	0.102/19.53
PM	16/4	0.135360	0.008655	8	4	0.02599	0.049/9.38
PVC	32/4	1.678828	0.236876	121	32	0.02131	2.763/35.07
PM	32/4	0.891788	0.096824	80	24	0.01712	1.422/27.30
PVC	32/8	0.208781	0.032893	23	11	0.01394	0.385/25.68
PM	32/8	0.051125	0.000798	9	4	0.00873	0.138/23.61

Positive certificates and those meeting the full solver's 10^{-10} margin are counted separately. Both equal changed-output counts in fresh cells; at 640 epochs, one/three cuts have 12 versus five PVC certificates and 30 versus eight PM certificates. Small positive cut bounds remain included. All-stage CSVs retain both definitions and numerical statuses.

C.6 Endpoint-retaining matched timing follow-up

This bounded follow-up reruns only the frozen correction procedures on the unchanged unused-parent inputs. Every one of the 12,288 model–parent inputs receives one call from each of five methods. The 303 inputs with pre-correction joint residual strictly above the original 10^{-8} screen receive three further calls per method, giving 65,985 calls overall. Selection is sealed before accessing evaluator labels or saved outcomes. The fixed cyclic method order is retained; timing repeats do not enter the scientific gain estimates. The full-joint PVC $(16, 4)$ confirmation primary, all-parent weighting and uncertainty are unchanged.

Table 12: Matched timing follow-up, milliseconds. Each method has 12,288 all-input calls and 909 further calls on 303 residual-selected inputs. Columns pool both learners and all four fresh regimes. Construction, screening, correction and production certification are included; shared intervals, evaluation, I/O and supplemental proof extraction are excluded. The all-input medians are dominated by unchanged returns. These are measurements on one CPU environment, not distribution-free speed guarantees.

	All inputs		Residual-selected	
Method	Mean	Median	Mean	Median
Keep B	0.001	0.001	0.002	0.002
Count	0.280	0.263	0.769	0.339
One cut	0.318	0.201	4.979	5.293
Three cuts	0.560	0.195	15.123	16.362
Full joint	0.673	0.272	17.224	16.447

The timer starts before each fresh system/adjuster construction and ends at the frozen method's return, including its production certificate arithmetic. The wrapper then retains the actual endpoint, status, elapsed nanoseconds and production bound. For changed full/count returns, a supplemental LP certificate is extracted outside the timer and tied to that endpoint; this extra verification work is not part of the reported method cost. The run uses two CPU threads, no GPU, Python 3.11.16, NumPy 2.4.6 and SciPy 1.17.1. Historical environment identity is not established, so this is a separately reported measurement, not a replacement for Table 8.

All 65,985 timed endpoints equal their historical counterparts bitwise, and repeated endpoints equal the first pass. Saved-output verification checks all calls, 4,848 accepted cut-step bounds and 1,236 full/count certificate payloads with exact scalar arithmetic. A separate reporting check links statuses, production bounds, nanosecond timers and parent/cell summaries to these outputs. Two all-input timing means differ from local floating-point reduction by $1.0842\,10^{-19}$ seconds; their errors against exact rational input means lie within the analytical floating-point summation bound. Strict bitwise summary reproduction therefore fails for those two means, while endpoints and the reported table precision agree. No experimental tolerance is changed. The private historical archive contains the retained endpoints and proof readers; the supplied supplement contains timing summaries only.

D Available incumbents, certificate sharpness and computation

Appendix D.1 separates a sharper certificate from a better computed output. Appendix D.2 gives the available-incumbent comparison in Figure 3, and Appendix D.3 reports classification accuracy as a secondary endpoint.

D.1 Certificate sharpness and computation are separate questions

For a fixed one-pass output G_A, distinguish its analytic certificate L_A, its sharp certificate L_A^* and the best uniform value V_A:

$$L_A \le L_A^* := \min_{Y \in \mathcal{C}_A} \{R(F,Y) - R(G_A,Y)\} \le V_A = m^{-1}\|F - \Pi_{\mathcal{C}_A} F\|_F^2.$$

Tightening the first inequality changes the certificate without changing the output or actual loss. Improving the second may require additional computation and changes the output. Neither $L_A^* - L_0^*$ nor $V_A - V_0$ is a direct lower bound on the loss difference between paid and free returns.

Exact slack of the analytic one-pass certificate. Write $q = \Pi_T f$, $d = q - f$, $H_i = F_i + d$ and $G_i = H_i - \theta_i \mathbf{1} + v_i$, with $v_i = (\theta_i \mathbf{1} - H_i)_+$. Expanding the loss gives

$$R(F,Y) - R(G,Y) = L + 2\langle d, \bar{Y} - q\rangle + \frac{2}{m}\sum_i \langle v_i, Y_i\rangle, \quad L = \|d\|^2 + m^{-1}\|H - G\|_F^2. \tag{22}$$

Both slack terms are nonnegative. The certificate is sharp precisely when some compatible target makes both zero. This can fail even for exact mean information: take $F_1 = e_1, F_2 = e_2$

and required five-class mean $q = (.10, .10, .01, .39, .40)$. Both row thresholds are .13, yielding $G_1 = (.47, 0, 0, .26, .27)$ and $G_2 = (0, .47, 0, .26, .27)$. The analytic value is .843, but the forced third-class mass makes the sharp value .8454. Targets $(.20, 0, .01, .39, .40)$ and $(0, .20, .01, .39, .40)$ attain it. Repeating each prediction 50 times gives an exact hard-label version with counts $(10, 10, 1, 39, 40)$. The gap is certificate conservatism, not a new improvement in predictions or a convergence failure.

An iteratively optimised free return can win. For binary first-class probabilities, take $F = (.9, .1)$ on two retained rows, a third omitted row with target zero, and parent first-class mean .2. Free retained means satisfy $q_1 \leq .3$. The one-pass free return has first-class probabilities $(.7, 0)$, while the exact free projection is $P_0 = (.6, 0)$. Querying the omitted zero makes $q_1 = .3$ exact, but leaves the one-pass paid output $(.7, 0)$ unchanged. The compatible fractional target $Y = P_0$ gives $R(P_0, Y) - R(G_A, Y) = -.01$. Repeating each row five times and choosing three positive retained labels in the first block gives a hard-label example with parent counts $(3, 12)$ and risks .34 original, .25 one-pass and .24 optimised free. Thus the particular comparator matters.

A fixed full-data information/computation comparison. The retrospective diagnostic keeps all 410 near-water Covertype groups and eight configurations, both saved 24-query policies and the same retained rows. Let G_0, P_0, G_A, P_A denote one-pass free, numerically optimised free, one-pass paid and numerically optimised paid returns. Here P_0 is the incumbent B in the main text. Free capacities are parent counts; paid capacities subtract the saved answers. All four returns are fitted to the original F, not sequentially projected from one another. Appendix D.2 tests the complementary intervention that starts from the optimised free return. The original primary G_A^{target} versus G_A^{entropy} remains $+.002281224216$. New comparisons are retrospective diagnostics.

Among cases where G_A loses to P_0, 91.485%/92.242% of the entropy/target weighted mass has G_A unchanged from G_0 at the declared output tolerance. Most disadvantage therefore reflects an unchanged cheap return versus better free computation, not a harmful change caused by answers. The worst observed G_A-versus-P_0 gain is $-.166397$. Softmax's mean is negative for both policies ($-.002267$ entropy, $-.000111$ target). The constant control differs: positive means .043065/.039989 and no observed adverse cases at 10^{-8} tolerance. More computation is not itself casewise dominance: P_0 loses to G_0 in .1084% of weighted learned cases, while P_A loses to G_A in .2168%/.1897% for entropy/target. The three negative P_A-versus-P_0 direct bounds are MLP seed 0 at bag index 285 under both policies (actual +.012897, bound $-.007089$), and bag index 389 under entropy (actual +.012242, bound $-.001150$). Their compatible worst-case hard-label assignments rule out universal dominance of these finite optimised returns. The nested-one-pass theorem does not apply to this different map.

Certifier sensitivity. For every fixed free and paid one-pass output, analytic and sharp baseline certificates agree within 6.67×10^{-16}, including constant. Thus no attainment classification changes by sharpening the certifier here. At .001, baseline attainment rises from 57.6152% for G_0 to 63.9566% for P_0, and from 66.2873%/66.5312% for G_A to 71.2195%/71.7615% for P_A. Direct G_A-versus-P_0 attainment is 16.9106%/17.9404%, whereas P_A-versus-P_0 attainment is 23.4959%/22.8997%. Comparator choice can therefore change even the direction of a coverage difference. This neither replaces the primary nor contradicts the constructed strict certificate gap above.

Numerical solver and independent checks. For capacities c, the projection minimises $\frac{1}{2}\|P - F\|_F^2$ subject to simplex rows and $\sum_i P_{ik} \leq c_k$. Nonnegative class multipliers λ give $P_i(\lambda) = \Pi_{\Delta_K}(F_i - \lambda)$ and dual value

$$D(\lambda) = \tfrac{1}{2}\|P(\lambda) - F\|_F^2 + \lambda^\top \left(\sum_i P_i(\lambda) - c \right).$$

L-BFGS-B minimises $-D$, followed when needed by dual-coordinate polishing. A row-preserving transfer of small capacity surpluses produces a feasible primal return. Every state's reported normalised gap is $2(P_{\text{obj}} - D)/m$; the maximum is 9.40×10^{-10} across all 9,840 unique states, below the declared 10^{-9} target. Maximum row-sum error is 3.34×10^{-16}, with no positive final capacity violation. This establishes numerical proximity, not equality to an exact projection. Every sharp guarantee is separately evaluated on the actual finite returned array.

Table 13: Same inputs, different information and computation. Mean gains and direct bounds use equal learned-family/bag weights; indicators precede seed averaging. G denotes one pass and P the finite optimised return.

Quantity	Entropy	Target .001
Free G_0 gain over original	.019020285	.019020285
Free P_0 gain over original	.029364780	.029364780
Paid G_A gain over original	.030652115	.032933340
Paid P_A gain over original	.045439645	.048526566
G_A realised gain over P_0	.001287336	.003568560
G_A direct lower bound over P_0	$-.004555360$	$-.003234569$
G_A actually worse than P_0 (%)	73.5230	73.3604
G_A admits worse target than P_0 (%)	75.9621	74.6883
P_A realised gain over P_0	.016074866	.019161786
P_A direct lower bound over P_0	.008343945	.010215343
P_A actually worse than P_0 (%)	0	0
P_A admits worse target than P_0 (%)	.0542	.0271

Independent primal SLSQP and objective-scaled LP checks give maximum projection discrepancy 5.71×10^{-9} and assignment-endpoint discrepancy 1.02×10^{-11}. Original gain/certificate replay error is at most 3.61×10^{-16}. An initial LP-tolerance failure and 28 above-target gaps remain recorded. Gap-triggered polishing corrected those states in at most three sweeps, changing scientific values by at most 4.11×10^{-9} without changing sign or attainment classifications. No state was removed. This sensitivity is neither a wall-clock benchmark nor a measurement-cost analysis. Appendix I.4 gives the check counts and selection details.

D.2 Updating the available prediction

Let B be any available matrix of row-simplex predictions, and let T_A be a nonempty closed convex subset of the class simplex containing the true retained mean after the acquired answers. Set $d = \Pi_{T_A}\bar{B} - \bar{B}$, $H_i = B_i + d$ and $G^{\mathrm{inc}}_{A,i} = \Pi_{\Delta_K} H_i$. We write $S = G^{\mathrm{inc}}_A$ within this appendix to match the retained comparison tables and figures. The two projection inequalities underlying Equation 12 give

$$R(B,Y) - R(S,Y) \geq \|d\|^2 + m^{-1}\|H - S\|_F^2 \geq 0. \tag{23}$$

The first inequality uses the projection of $\bar{B}$ onto a set containing $\bar{Y}$; the second uses row projections onto sets containing each Y_i. No optimality of B is needed, and the new information set need not be nested in the old one. The scored rows, their weights and the target must remain fixed. The statement holds for each realised set of valid answers, including adaptive queries. It is a direct corollary of established projection geometry, not a new adjustment algorithm.

If $\bar{B} \in T_A$, then $d = 0$ and $S = B$; otherwise the guaranteed decrease is strictly positive. Full intersection projection $Q = \Pi_{\mathcal{C}_A} B$ instead maximises the uniform guarantee relative to B among row-simplex returns, by Equation 3. This changes the objective from distance to the earlier predictor F to distance to B. Merely initialising a distance-to-F solver at B would be a warm start of the original problem, not this incumbent-centred correction.

The difference of optimal baseline values is an upper bound. Let $\mathcal{C}_A \subseteq \mathcal{C}_0$ be nonempty compact convex target sets, and in this paragraph use the exact projections $B = \Pi_{\mathcal{C}_0} F$, $P_A = \Pi_{\mathcal{C}_A} F$ and $Q = \Pi_{\mathcal{C}_A} B$. Write $\|X\|_m^2 = m^{-1}\|X\|_F^2$ and $V_{\mathcal{C}}(X) = \|X - \Pi_{\mathcal{C}} X\|_m^2$. The established projection inequalities imply

$$V_{\mathcal{C}_A}(F) - V_{\mathcal{C}_0}(F) \geq V_{\mathcal{C}_A}(B) + \|P_A - Q\|_m^2 \geq V_{\mathcal{C}_A}(B). \tag{24}$$

This is a consequence of classical convex geometry (Bauschke & Borwein, 1996), not a new adjustment principle. When the target sets are the relevant convex hulls, the V terms are the minimax values in Equation 3. Their difference relative to F can exceed even the best attainable direct guarantee relative to B; it is not a lower bound on the paid return's improvement over that incumbent.

To prove the result, expand the two squared distances:

$$V_{\mathcal{C}_A}(F) - V_{\mathcal{C}_0}(F) - V_{\mathcal{C}_A}(B) = \|P_A - Q\|_m^2 + \frac{2}{m}\langle F - B, B - P_A\rangle + \frac{2}{m}\langle Q - B, P_A - Q\rangle. \tag{25}$$

Both inner products are nonnegative: P_A belongs to both target sets, so the variational inequalities for B and Q apply. The excess thus combines disagreement between the two paid optima with two projection-normal slack terms. Equality in the first inequality of Equation 24 requires both inner products to vanish. The entire excess vanishes only when $P_A = Q$ and the old normal term is zero. Identical paid outputs alone do not make the values equal. This comparison of guarantees implies no ordering of realised losses.

The geometric inequality remains true for any feasible $P_A \in \mathcal{C}_A$ when its squared distance replaces $V_{\mathcal{C}_A}(F)$; its optimality is needed only for the minimax-value interpretation. In contrast, exactness of B and Q is essential to the nonnegative-normal argument.

Finite-output consistency allowance. For a feasible approximate projection P of X onto a convex set $\mathcal{C}$, let a valid dual lower bound give normalised primal–dual gap $\delta = 2(P_{\mathrm{obj}} - D)/m$. Strong convexity yields $\|P - \Pi_{\mathcal{C}} X\|_m \le \sqrt{\delta}$. For $Z \in \mathcal{C}$,

$$\frac{1}{m}\langle P - X, Z - P\rangle \ge -\frac{\delta}{2} - \sqrt{\delta}\,\|Z - P\|_m. \tag{26}$$

Indeed, with $U = \Pi_{\mathcal{C}} X$ and $e = P - U$, optimality gives $\langle U - X, Z - U\rangle \ge 0$, while the objective gap bounds $m^{-1}\langle U - X, e\rangle + \|e\|_m^2/2$ by $\delta/2$. Expanding the left side and using Cauchy–Schwarz proves the bound.

For the actual finite arrays, define

$$E = \|F - P_A\|_m^2 - \|F - B\|_m^2 - \|B - Q\|_m^2 - \|P_A - Q\|_m^2.$$

The expansion in Equation 25 remains an algebraic identity. If the three arrays are exactly feasible and the free and incumbent-centred gaps δ_0, δ_I are valid, then

$$E \ge -\delta_0 - \delta_I - 2\sqrt{\delta_0}\|B - P_A\|_m - 2\sqrt{\delta_I}\|Q - P_A\|_m. \tag{27}$$

No original-start paid gap is needed for this allowance, only feasibility of P_A. Floating-point feasibility and gap receipts provide a numerical robustness diagnostic; they do not rigorously enclose these premises as interval arithmetic would. Direct guarantees below evaluate the actual returned arrays separately.

Fixed-endpoint interpretation. The retrospective incumbent diagnostic retains all 410 Covertype bags, eight configurations, both saved policies and the same 24 acquired labels; there are 6,560 returned comparisons. It changes neither training nor query orders nor the original primary contrast. The squared-distance quantities in Table 14 are finite-output proxies for the exact values above. Means average neural seeds within family, then the three learned families and bags equally; constants remain separate.

Most of the excess comes from the old normal term ($.002252515/.001952622$), rather than disagreement between the paid returns; the new normal term is approximately 6.3×10^{-9} under both policies. All 6,560 algebraic identities agree within 2.79×10^{-16}. The minimum raw E is -8.77×10^{-10}, and no residual violates the conditional allowance in Equation 27 beyond 2×10^{-12}. All incumbent-centred projection gaps are below 10^{-9}; their sharp direct bounds differ from the squared-distance proxies by at most 3.53×10^{-9}. These numerical checks support the comparator interpretation without replacing finite-return certification or claiming that exact projection has been computed.

D.2.1 Incumbent intervention: protocol, outcomes and checks

The intervention applies one pass to the saved finite B, and separately projects B onto the paid feasible set using the solver in Appendix D.1. Inputs, training, masks, all eight configurations and

Table 14: Reference-dependent projection distances at the fixed endpoint. These are finite-output distances, not assertions of exact minimax values.

Quantity	Entropy	Target .001
$\|F - P_A\|_m^2 - \|F - B\|_m^2$	.010636317	.012202206
$\|B - Q\|_m^2$	.008382574	.010247860
Difference	.002253743	.001954346
$\|P_A - Q\|_m^2$	.000001222	.000001718

the saved entropy and target .001 query permutations remain fixed. Each of the 410 bags retains 16 of 64 rows; the same 24 of 48 omitted labels set each paid capacity. Means below use the three learned families and bags equally after averaging seeds within family. Indicators precede averaging. The 2,870 learned configuration cases per policy are not independent replications. The 410 constant cases remain a separate control.

The old original-start G_A remains the primary policy endpoint; replacing it by S is a retrospective intervention. The new target-minus-entropy contrast for S is $+.002326819$, with 104 positive, 97 negative and 209 tied family-averaged bags; for Q it is $+.003079596$, with 107/94/209. Both medians are zero. The original $+.002281224$ and its descriptive bootstrap interval are not replaced or reassigned to these new contrasts. For S, paired .001 direct attainment has 16.721% both, 6.287% entropy only, 5.827% target only and 71.165% neither. Switching entropy to target benefits 8.611% of rows and harms 9.251%, with conditional magnitudes .075238 and .044880; these magnitudes are weighted gain/loss mass divided by the corresponding weighted row fraction. The constant control favours entropy (.043065 versus .039989 mean gain over B); its S and Q coincide up to numerical precision. All family and wilderness contrasts are retained in the private historical archive.

Further incumbent-centred optimisation adds .004826596/.005579373 mean gain over S. No actual or compatible group harm against S exceeds the stated tolerance here, but arbitrary projections need not have that ordering. Similarly Q exceeds original-start P_A by only .000017545/.000010220 on average, and loses in .217%/.244% of weighted cases (nine raw cases across policies). S also loses to original-start G_A in 18 raw cases. These comparisons imply no general dominance over other paid returns; Equation 23 specifically protects B.

The private historical archive retains all 6,560 endpoint comparisons and 104,960 row records. Label/count identities, permutations, row identities, capacities, original risks and input hashes pass. All 6,560 new optimised states meet the 10^{-9} normalised primal–dual gap target (maximum 8.81×10^{-10}), with maximum row-sum error 6.66×10^{-16}. The same five metadata-selected bags (0,37,101,205,409) supply 240 separately formulated assignment LP checks and 80 independent primal QPs; maximum assignment error is 1.02×10^{-11}, and maximum primal objective and coordinate differences are 1.89×10^{-9} and 1.53×10^{-9}. All direct sharp guarantees are evaluated on the returned arrays. The one-pass analytic and sharp guarantees agree within 4.45×10^{-16}. The calculation uses one CPU thread and makes no new timing claim.

D.3 Classification accuracy as a secondary endpoint

Argmax accuracy is evaluated from the same saved probabilities and queried labels, using the original seed/family/bag hierarchy. Before correction, the learned-family average is 65.169%; the free one-pass and optimised incumbents reach 66.103% and 66.702%. Original-centred paid entropy/target updates reach 66.7666%/66.7446%, despite target's fixed $+.002281224$ squared-loss advantage. The accuracy difference is only $-.0220$ percentage points: it illustrates endpoint disagreement, not practically important entropy superiority. Incumbent-centred one-pass updates reach 67.3933%/67.3476%; their optimised counterparts reach 67.6846%/67.6795%. No classification-accuracy guarantee follows from the squared-loss projection result. These are retrospective summaries of the same 410 bags, not additional independent experiments.

Table 15: Incumbent-centred outcomes at 24 queries, relative to the same finite optimised free B. S uses one pass and Q finite intersection projection. Gains use mean squared probability-vector loss. Harm/benefit tolerance is 10^{-8}; unchanged coordinates use 10^{-12} and need not complement these risk indicators.

	One pass S		Optimised Q	
Quantity	Entropy	Target .001	Entropy	Target .001
Actual gain	.011265815	.013592634	.016092411	.019172007
Sharp direct guarantee	.005959227	.007306816	.008382574	.010247860
Group benefit (%)	27.073	26.775	27.127	26.802
Group harm (%)	0	0	0	0
Unchanged predictions (%)	72.846	73.117	72.846	73.117
Direct .001 attainment (%)	23.008	22.547	23.577	22.954
Individual row benefit (%)	21.160	20.667	21.162	20.667
Individual row harm (%)	2.075	2.346	2.073	2.348

E SPATIAL DATA, SUPPORT INTERVENTIONS AND THEIR INTERPRETATION

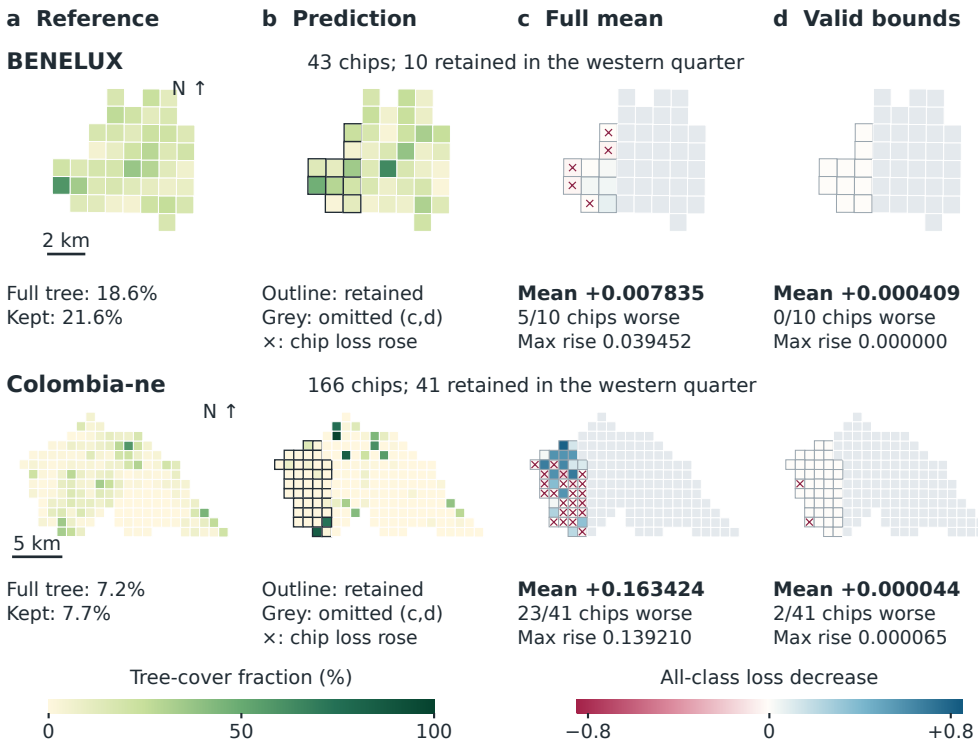


Figure 7: Land-cover composition and loss changes on actual chip footprints. Each group was selected by proximity to median eligible chip count before viewing outcomes; both use softmax seed 0 and the western quarter. Reference tree cover is evaluator-only; losses use all five classes. Full-mean adjustment improves both selected group means while worsening chips; BENELUX's pooled western result is negative (Figure 8). Valid bounds give small group gains. "Max rise" reports the largest individual loss increase, distinguishing its size from its frequency. Composition and signed-loss scales are common; geographic scales differ. Crosses mark loss increases above 10^{-8}; gaps contain no released chips.

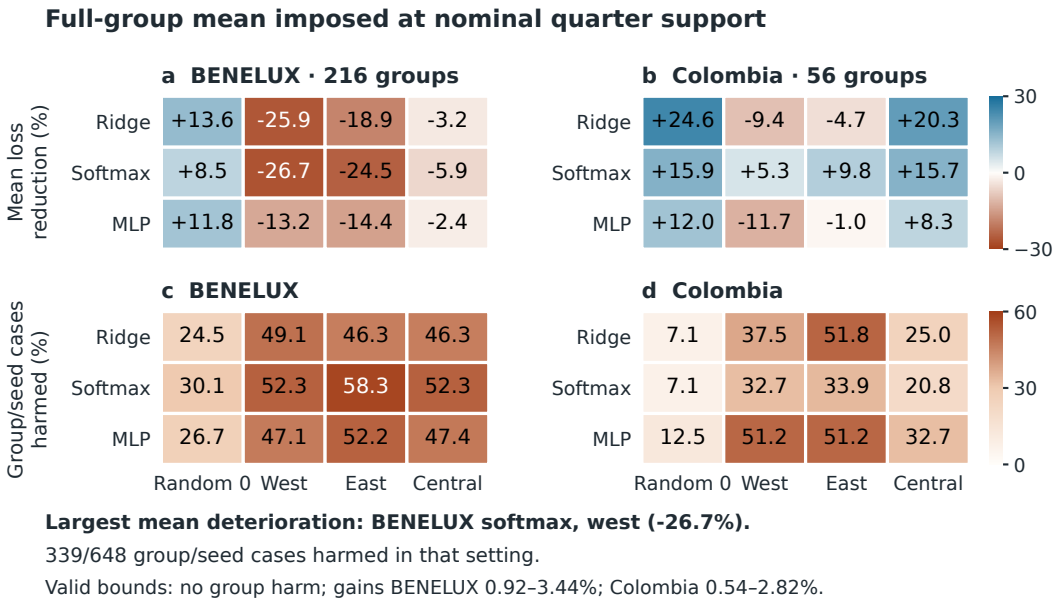


Figure 8: Full-parent correction on nominal quarter supports. Top: relative reduction in equally weighted group loss; negative is worse. Bottom: harmed group/seed cases (10^{-8} tolerance). **Random 0 uses one fixed mask**; Table 16 pools five. Neural rows pool three seeds; ridge uses one predictor. The footer gives valid-bound controls. Policies and model seeds reuse the 216 BENELUX and 56 Colombia groups.

E.1 Spatial implementation, controls and interpretation

The BENELUX release (Ramos-Pollán & González, 2023a) has a published MD5 that matches the 2,009,485,925-byte download; its SHA-256 and compact derived-input hashes are recorded. A restricted NumPy-only unpickler decoded all 72,213 records. Each contains 10,000 valid reference pixels, and every stored fine proportion matches counted labels. The raw label map is grouped as in Ramos-Pollán & González (2023c): infrequent classes 7–11, tree 1, vegetation 2/3/6, cropland 4 and built-up 5. There are 644/205/233 train/validation/test groups, with no group shared between splits. Two training groups contain one chip and therefore reveal its composition through the aggregate. The actual archive split counts differ from the paper-reported counts; this is a declared adaptation.

We independently extract ImageNet ResNet-18 features using the official torchvision IMAGENET1K_V1 weights, with full weight-file SHA-256 recorded. The whole RGB chip is scaled by 255, resized from 100 to 224 pixels using antialiased bilinear interpolation, and normalised by ImageNet means and standard deviations. Unlike the default ImageNet transform, no centre crop discards chip support. Inference uses float32, evaluation mode and no AMP/TF32. Replaying 32 chips at a different batch size differs by at most 1.38×10^{-5}. Train-only feature standardisation, covariance eigendecomposition and PCA-coordinate standardisation produce 64 inputs.

Ridge fits an affine surrogate to equally weighted bag means, then applies a nonlinear row-simplex decoder; neural training instead aggregates decoded probabilities. This baseline distinction precludes a causal claim about model capacity alone. Ridge uses an unpenalised intercept, penalties .01/.1/1/10/100/1000 and row-simplex decoding. Neural heads are softmax linear or one-hidden-layer ReLU MLP64, trained by Adam with learning rate .01, seeds 0/1/2, weight decay 0/.001 and 800 full-batch updates. Selection considers checkpoints 50/100/200/400/800 and uses only equally weighted validation bag-mean squared error. Deterministic sorted cumulative sums compute neural bag means in float64. All 66 candidate scores are retained. Independent NumPy replay of the six selected neural heads agrees within 1.98×10^{-7}; a separate augmented ridge least-squares solve agrees within 6.77×10^{-16}.

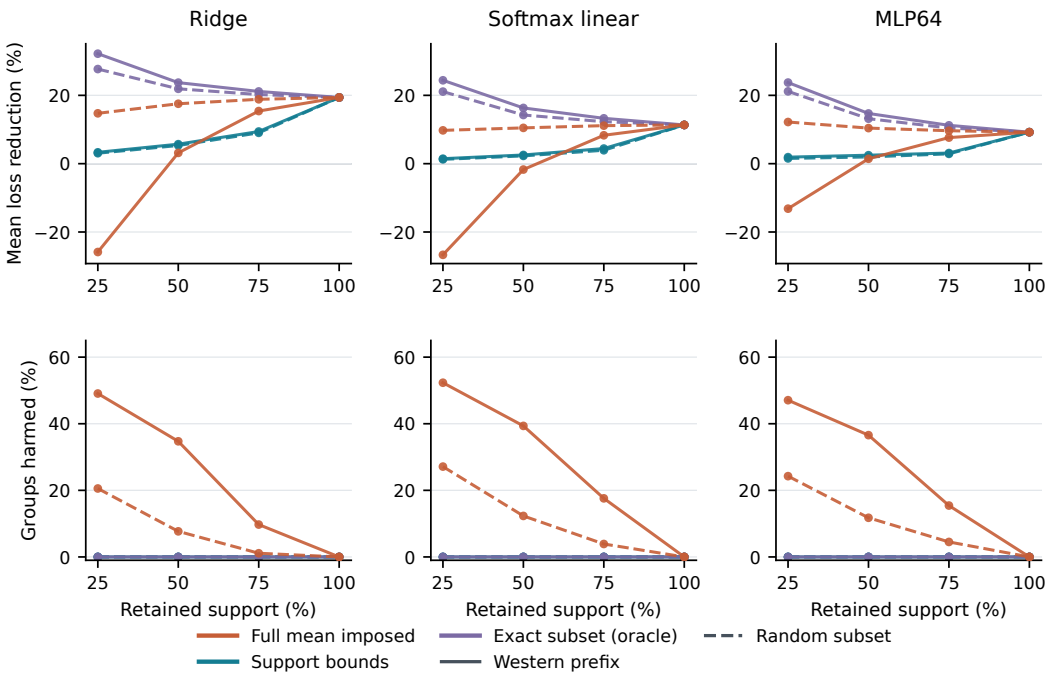


Figure 9: BENELUX: support mismatch has different consequences under random and geographic omission. Top: relative reduction in equally weighted mean group squared loss; negative values indicate actual harm. Bottom: fraction of group/seed/mask cases with loss increase greater than 10^{-8}. Solid and dashed lines represent western and random subsets; markers are the four nominal coverage settings, using rounded-down counts within each group. Full-coverage points are shared. Neural seeds and random masks are averaged descriptively. The constant control is retained below; Figure 7 shows metadata-selected chip footprints.

The selected MLP heads have lower validation aggregate error than ridge, but higher held-out chip-composition loss. This is an observed generalisation contrast; differences in validation/test geography mean it does not isolate the aggregate objective as its cause. Full-support equal-group baseline risks are .2174, .0746, .0814 and .1033 for constant, ridge, softmax and MLP respectively. The constant is kept as a weak-model control rather than used to inflate learned-model gains.

Dykstra alternates a mean-box projection with row-simplex projection, using tolerance 10^{-9} and 1,500 iterations. Each bag stops separately; capped iterates remain in every summary. Certificate positivity and observed harm use a 10^{-8} margin. Equal-group summaries weight each eligible support equally; neural seeds and random masks repeat those same supports. There is no calibrated confidence interval or claim of spatial independence. The evaluator saves every subset ID, group result, loss, certificate, convergence diagnostic and descriptive summary.

E.2 Complete primary quarter-support table

Table 16 retains the primary random/western comparison, including the constant control and five-mask random ensemble. Figure 8 instead shows one fixed random mask alongside the geographic follow-ups.

E.3 Colombia replication: protocol and scope

The Colombia release (Ramos-Pollán & González, 2023b) contains 69,193 chips. Published MD5 and byte count match the 1,886,667,211-byte archive; the licence is CC-BY-4.0. A restricted NumPy-only audit identifies exactly one entirely invalid reference chip, in a validation group. Ex-

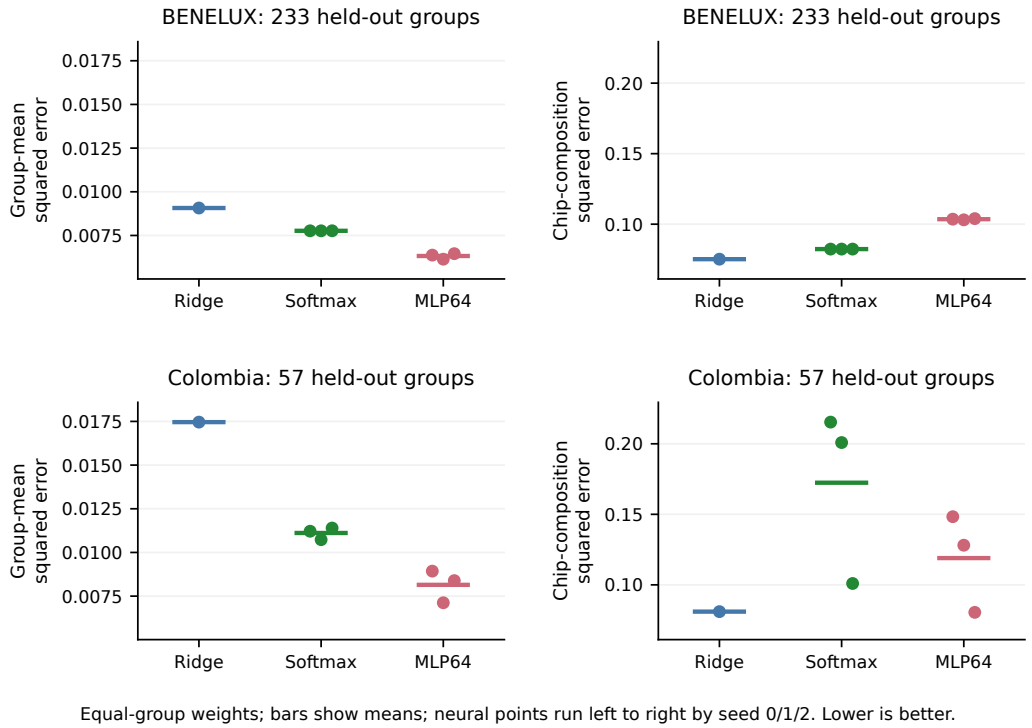


Figure 10: Same-support ranking in both regions. Each row uses its identical held-out groups for aggregate and chip loss; points retain all three neural seeds and bars show their means. Columns use different losses and scales, matched across regions. BENELUX reverses all 15 learned-family pairs; Colombia reverses seven and retains a strong MLP seed. The secondary BENELUX analysis motivated the forward regional ranking check. Neither comparison confounds validation with test geography.

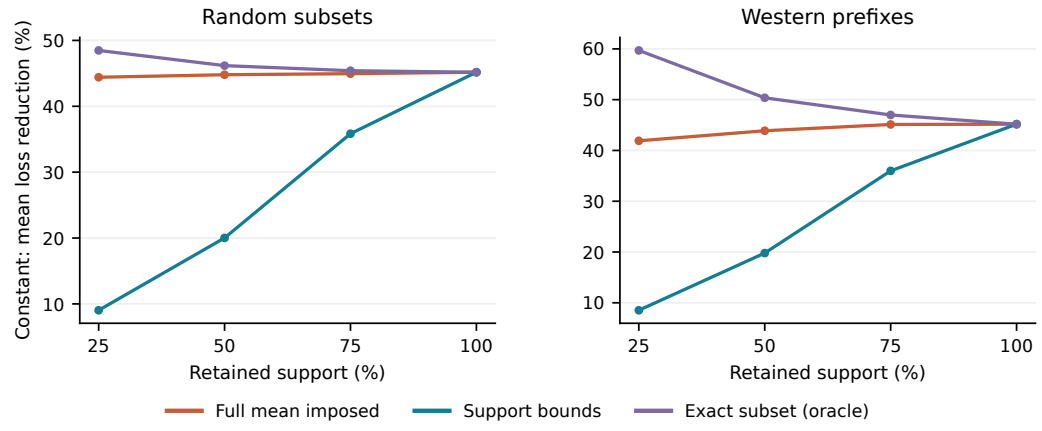


Figure 11: The BENELUX constant control improves on average under both omission policies, including the mismatched full mean. Its baseline error is larger than those of learned heads. This preserves counterevidence to a universal-harm claim. Positive average gain does not imply that every group benefits.

Table 16: Nominal quarter-support results. Positive gains reduce loss. Baseline risk uses western subsets; gains are ratios of equally weighted mean group losses. Neural rows pool three seeds; random results also pool five masks. Harm is the fraction of group/seed cases worsened by imposing the full mean on western subsets. Repeated configurations are dependent.

Family	Baseline risk	Full mean: gain (%) Random	West	Bounds: gain West (%)	West harm (%)
BENELUX: 216 eligible groups					
Constant	0.2282	44.41	41.90	8.53	11.57
Ridge	0.0747	14.74	−25.86	3.32	49.07
Softmax	0.0823	9.75	−26.68	1.46	52.31
MLP64	0.1099	12.19	−13.17	1.91	47.07
Colombia: 56 eligible groups					
Constant	0.2526	59.39	54.16	11.85	16.07
Ridge	0.0801	26.39	−9.45	1.58	37.50
Softmax	0.1656	19.32	5.30	2.51	32.74
MLP64	0.1128	14.49	−11.75	0.54	51.19

cluding it reduces that group's support from 355 to 354 chips and validation from 14,075 to 14,074 chips. Training remains 40,107 chips in 125 groups; validation has 62 groups; test remains 15,011 chips in 57 groups. Every retained chip has 10,000 valid reference pixels. One training group and two validation groups are singletons. No group crosses splits; neighbouring groups need not be spatially independent.

The five-class mapping, discrete released-chip support and uniform reference means match BENELUX. Polygon-level WorldCover targets are absent in the commune metadata; per-chip mixture targets remain recorded and are not imposed as exact group means. The forward protocol was written before Colombia feature extraction and fitting, after BENELUX outcomes. It is a recorded local protocol, not external preregistration.

All feature transforms, grids, optimiser steps and checkpoints remain unchanged. The first-32 feature replay differs by at most 1.10×10^{-5}; independent selected-head replay differs by at most 1.43×10^{-7}, and the augmented ridge solve by 1.51×10^{-15}. Ridge again selects penalty 10. All neural selections use update 800, with each weight decay chosen by the same aggregate-only validation rule. The unchanged ridge penalties apply to a summed group-residual objective, so the smaller training-group count changes their effective regularisation. Region differences cannot be attributed solely to geography or capacity.

All 56 eligible groups contain at least 16 chips. The actual quarter-support fraction ranges from $2/9$ to $1/4$ and retains 3,731 chips. The 19 settings, three information controls, 10^{-9} stopping tolerance, 1,500-iteration cap and 10^{-8} reporting margin are identical to BENELUX. The independent real-QP checks use the three smallest eligible groups, with 27/30/34 chips; the BENELUX verifier's at-most-24-chip diagnostic filter would be empty here. This adaptation selects verification cases by metadata, not by model outcomes.

Across 25,536 Colombia records, 5,414 reach the iteration cap, with maximum stopping residual 1.95×10^{-4}. These are returned numerical adjustments, not all converged projections. Valid-information lower certificates never exceed actual gain by more than 6.67×10^{-16}. Independent verification reconstructs every mask, baseline and denominator; 81 real-data SLSQP checks reproduce gains within 1.41×10^{-9}. The numerical limits remain part of the reported fixed-budget experiment. A separate numerical stress check selects the largest-residual case per control globally and within western-quarter support, producing six cases from one group. Increasing only their cap to 20,000 satisfies the stopping rule in 6,301–17,610 iterations. Gain changes by at most 1.11×10^{-4} and no harm/certificate signs change, but an individual prediction entry changes by .202. Thus risk stability in these six cases does not imply pointwise convergence or a bound for all capped cases. Primary outputs remain unchanged.

The pooled softmax exception also contains seed variation: western-quarter gains are 10.62%, 9.32% and -14.51% for seeds 0/1/2. All three MLP seeds worsen, by 4.68%, 10.37% and 26.84%. These selected-configuration differences do not isolate weight decay or architecture as their cause.

Secondary weighting sensitivity. After BENELUX and Colombia ranking outcomes, while the Colombia intervention was running, we declared a complete retained-chip-weighted re-analysis of both regions. It preserves every family, coverage and control and leaves the primary equal-group estimand unchanged. For the original random/western policies, all quarter-support gain signs agree. One non-quarter sign changes: BENELUX MLP under the wrong full mean at western half-support improves by 1.43% under equal-group weighting but worsens by 2.34% under equal-chip weighting. These are different finite-dataset estimands, not competing significance tests.

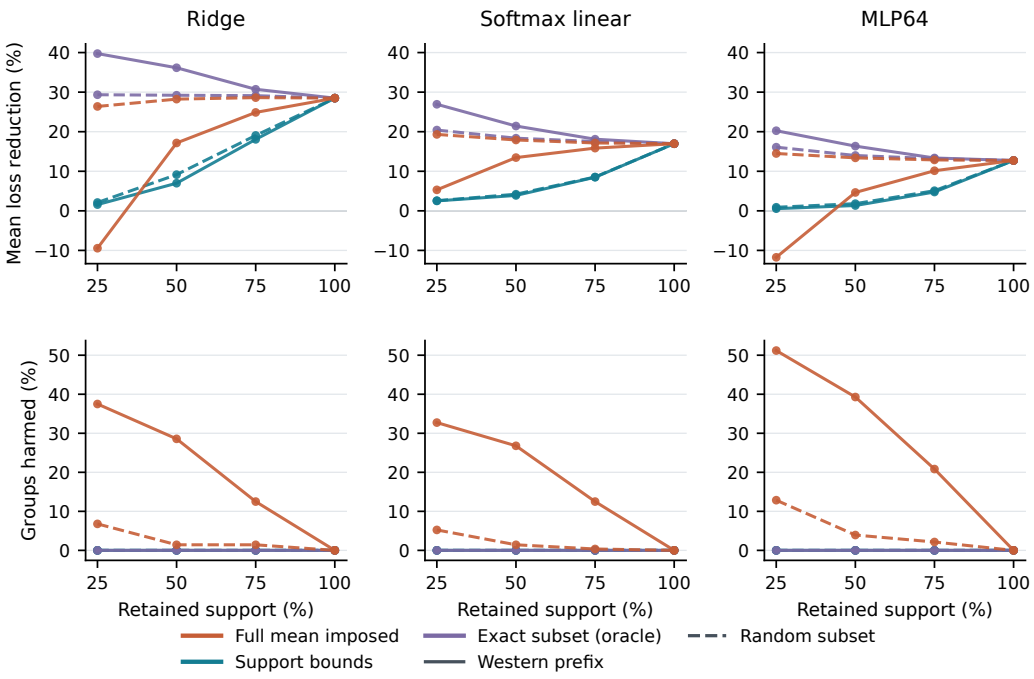


Figure 12: Forward Colombia support replication with the same frozen settings. Means pool the three selected neural seeds and, for random omission, five masks. Softmax retains positive mean gains under the western policy, although individual group/seed cases are harmed. The other two families reproduce the quarter-support mean reversal. Valid-bound and oracle controls have no observed harm in mean chip loss within groups. Axes have common scales within each row; the smaller valid gains remain visible.

E.4 Saved-support mechanism and protected-unit audit

This secondary analysis uses the existing reference labels after the primary outcomes were known. It retains all eight predictors (constant, ridge, three softmax and three MLP seeds), all 216/56 eligible groups, and archived arrays for full support, random-mask-0 quarters and western quarters. The single random mask is not the five-mask average in Table 16. All 19,584 group/model/control gains and certificates replay within 7.22×10^{-16}; all 124,032 historical solver records are summarised separately. No training or target selection changed.

The support comparison cited in the Discussion has opposite regional outcomes: full-parent proportions applied to western quarters worsen BENELUX learned-family loss by 13.17–26.68%, while Colombia softmax improves 5.30%. For the latter, mean-shift contribution -0.017159 is offset by centred residual improvement $+0.025942$, yielding $+0.008783$. This reference-based decomposition neither validates the wrong mean nor forecasts the exception from aggregates alone.

Exact decomposition and its limit. Write $F_c = F - \bar{F}$, $G_c = G - \bar{G}$, $Y_c = Y - q$, and $C = n^{-1}(\|F_c - Y_c\|_F^2 - \|G_c - Y_c\|_F^2)$. The within/between identity gives

$$R(F,Y) - R(G,Y) = B - D + C + \eta, \qquad B = \|\bar{F} - q\|^2, \quad D = \|Q - q\|^2, \tag{28}$$

where $\eta = \|Q-q\|^2 - \|\bar{G}-q\|^2$. For mean residual $r = \bar{G} - Q$, its remainder is $\eta = -2\langle Q - q, r\rangle - \|r\|^2$. The largest identity discrepancy is 5.56×10^{-16}. The centred term is a residual-error change, not simply prediction variance. Across 4,352 quarter-support full-mean records, 297 have negative affine but positive actual gain at the 10^{-8} margin; none show the reverse sign change. Nevertheless, C is negative at the reporting margin in 199 records and reaches $-.095028$.

The observed one-sided sign changes are not a theorem. For two binary first-class probabilities $F = (.9, .3)$, fractional targets $Y = (1, .41)$ and imposed mean .8, the exact intersection projection is $G = (1, .6)$. Under the two-class Brier convention the affine gain is $+.004$, centred gain is $-.018$, and actual gain is $-.014$. The same example extends to one-hot targets by repeating each prediction 100 times with 100/41 positives in the respective blocks. A single row-simplex projection after the affine shift retains the affine lower bound; reimposing the wrong mean through intersection projection does not.

For Colombia western softmax, seeds 0/1/2 have affine gains $-.016430, -.015535, -.019510$, centred gains $+.038380, +.033725, +.005721$, and actual gains $+.021950, +.018190, -.013789$. The 151 converged records retain mean affine gain $-.013669$ and actual gain $+.011647$; the 17 capped records average $-.016655$ actual gain. These selected strata do not estimate a causal effect of convergence.

Group influence and fine-unit effects. Leave-one-group-out removes all seeds of one physical group together. Equal-group western gain ranges are $[-26.81, -22.67]\%$, $[-27.50, -23.75]\%$ and $[-13.77, -10.98]\%$ for BENELUX ridge/softmax/MLP; Colombia ranges are $[-13.04, -2.77]\%$, $[3.84, 8.19]\%$ and $[-14.07, -7.60]\%$. Every primary sign survives. However, Colombia softmax's chip-weighted gain ranges from $-.158\%$ to 7.05% across these deletions, versus 4.06% overall. Between 23–28 BENELUX or 7–10 Colombia groups account for half the absolute learned-family contribution; tail magnitude still matters. No spatial independence or calibrated confidence interval is asserted.

Valid western bounds harm 1.26/2.72/4.21% of BENELUX ridge/softmax/MLP chip records and 1.37/4.97/3.61% in Colombia. Corresponding subset-oracle rates are 30.29/37.11/41.63% and 33.99/42.59/45.63%. Class-average errors can worsen too: BENELUX softmax built-up gain is $-.000135$ while total gain is $+.001202$; Colombia softmax vegetation gain is $-.000554$ while total gain is $+.004153$. These small class effects do not establish downstream harm.

Convergence and information. Across the 19 original settings, BENELUX cap counts are 76/130/68 for full mean/bounds/oracle, each out of 32,832. Colombia counts are 1,715/1,924/1,775, each out of 8,512. Every BENELUX western-quarter return converged; Colombia western-quarter counts are 38/38/36 out of 448 per control. On Colombia western-quarter pairs where both bounds and oracle converged, mean absolute oracle-minus-bounds gains remain .033126 (ridge, 50 records), .042933 (softmax, 146) and .024116 (MLP, 153). This does not supply a global optimality-gap bound.

For the declared fractional-target class $\mathcal{C}_T = \{Y : Y_i \in \Delta_K, \bar{Y} \in T\}$, $F \in \mathcal{C}_T$ implies that the admissible target $Y = F$ makes every nonzero modification harmful. Thus optimal uniform gain is zero independently of iterations. With 10^{-12} slack required in every mean-box coordinate, western strict-membership counts are 77/216, 254/648 and 368/648 for BENELUX ridge/softmax/MLP, and 7/56, 13/168 and 37/168 for Colombia. Pooled fractions are 46.23% and 14.54%; mean realised oracle gains on those strata are .023535 and .039120. Random-mask-0 fractions are 47.62% and 10.97%, with gains .019988 and .032815. Every stratum oracle return has a positive applicable certificate. Across these learned strata, minimum mean-coordinate slack is 3.57×10^{-6}; normalising row sums changes entries by at most 4.45×10^{-16} and retains membership. This is a numerical membership check, not interval arithmetic. The conclusion concerns the stated fractional class; justified semantic or physical restrictions could exclude some admissible targets.

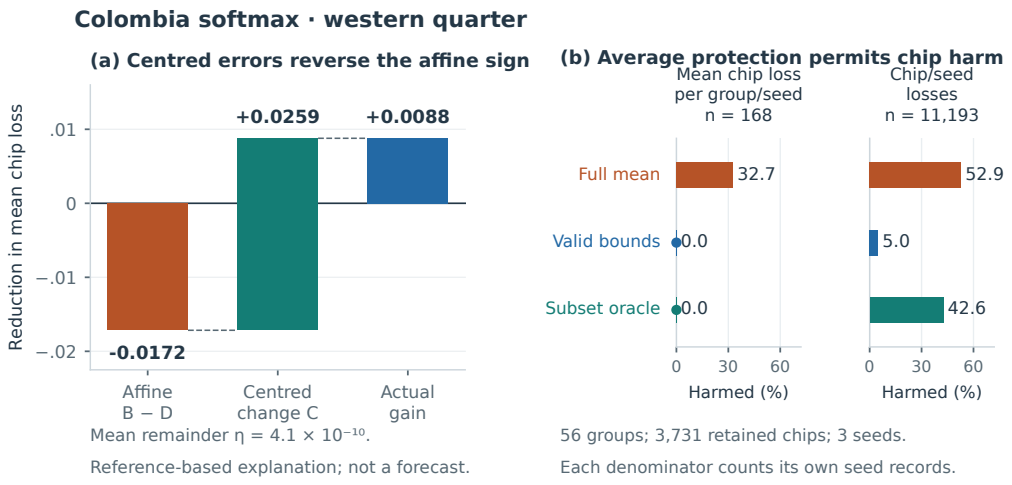


Figure 13: Colombia softmax on western-quarter support. (a) The positive centred residual change reverses the affine diagnostic; the mean-constraint remainder is approximately 4.1×10^{-10}. (b) Harm frequencies have separate denominators: 168 group/seed mean-chip-loss cases and 11,193 chip/seed losses (56 groups, 3,731 chips, three seeds). Valid bounds and exact-subset information protect group mean loss while permitting chip harm. Capped numerical iterates are included.

E.5 Fixed eastern and central-quarter follow-up

Both new rules were fixed after the western/random outcomes but before their own evaluation; this is local predeclaration, not external preregistration. Eastern selection orders chip centres by descending longitude, then ascending latitude and chip ID. Central selection orders distance from the unweighted group chip-centre centroid, using $d_x = \cos(\pi\bar{\phi}/180)(\lambda - \bar{\lambda})$ and $d_y = \phi - \bar{\phi}$, then ascending longitude, latitude and chip ID for ties. Coordinates are in degrees; this is a local distance ordering, not a new physical-area support. Available centres selected this way need not form a contiguous patch. Neither rule uses reference labels or prediction errors.

Each retains $\max(1, \lfloor n_g/4 \rfloor)$ of the same eligible group's chips. BENELUX retains 3,490 of 14,269 eligible chips, Colombia 3,731 of 15,009. Actual fractions range from 4/19 to 1/4 and 2/9 to 1/4; all bounds use actual fractions. All eight fixed predictors and three information controls are retained. The reference pipeline, solver and stopping choices are unchanged.

Table 17: Additional quarter-support gains (%). Full-mean columns compare weighting choices; bounds/oracle columns use the primary equal-group estimand. Positive values reduce loss.

Region / subset	Family	Full: group	Full: chip	Bounds	Oracle
BENELUX east	Ridge	−18.93	−14.11	3.44	31.02
	Softmax	−24.47	−20.63	1.30	21.31
	MLP	−14.45	−12.97	1.05	20.96
BENELUX central	Ridge	−3.16	−10.43	3.38	30.90
	Softmax	−5.94	−14.45	1.63	22.89
	MLP	−2.38	−10.46	.92	20.70
Colombia east	Ridge	−4.75	−24.94	2.21	37.54
	Softmax	9.79	−1.54	2.46	24.98
	MLP	−1.02	−17.79	.59	23.69
Colombia central	Ridge	20.28	11.73	2.82	38.48
	Softmax	15.69	8.46	2.54	23.84
	MLP	8.28	.72	.80	19.10

All 13,056 group/model/mask/control returns were independently replayed from saved arrays and dual vectors; every bounded-simplex support value was checked by an independent five-variable linear programme. Maximum gain and certificate replay errors were 10^{-16} and 7.22×10^{-16}. Independent SLSQP checks on the three smallest eligible groups per region cover 288 real-data cases and agree in gain within 9.18×10^{-10}. They do not establish convergence of every larger solve. BENELUX has no capped solves; Colombia retains 239/2,688, maximum stopping residual 7.86×10^{-5}. No applicable certificate exceeds realised gain by more than 2.78×10^{-16}. The final plotting repair left all scientific arrays, selection and fitting unchanged.

E.6 Composition error and magnitude of individual deterioration

A retrospective saved-output analysis expresses error in composition percentage points. For each chip, component MAE is $(100/5)\sum_k |F_{ik} - Y_{ik}|$; it is not pixel classification error. We average chips within western-quarter groups, seeds within family, then the three learned families and groups equally. Table 18 retains the unchanged, mismatched-parent, valid-bound and exact-subset controls. References and predictions are unchanged. These secondary summaries do not inherit the squared-loss guarantee. For valid bounds, the mean numerical-sign harm rates are 2.756%/3.049%,

Table 18: Saved western-quarter predictions: component MAE (percentage points) and percentage of chip/seed outcomes with squared-vector-loss increase above .01 relative to the original prediction. The threshold describes magnitude, not a calibrated application cost. Groups: 216 BENELUX and 56 Colombia; settings are dependent.

	MAE (pp)		Loss rise $>$ .01 (%)	
Correction	BENELUX	Colombia	BENELUX	Colombia
Unchanged	8.483	7.580	0.000	0.000
Full parent mean	9.387	7.635	44.079	35.803
Valid support bounds	8.274	7.304	0.046	0.030
Exact subset mean	7.175	6.076	19.871	20.746

falling to .046%/.030% above .01. The exact-subset control improves average component MAE further, but substantial chip deterioration remains. Complete family/component results at thresholds $10^{-8}, 10^{-4}, 10^{-3}, .01$ are retained in the private historical archive. This magnitude analysis complements the original family ranges rather than replacing their denominators.

F Acquisition model, bounds and additional comparisons

This appendix supplies the assumptions and derivations behind Equations 31 and 30, the matched one-pass comparison, and the complete spatial acquisition curves below. The arguments specialise classical projection and threshold geometry; we make no theorem-priority or optimal-policy claim.

For the original fixed Covertype policy contrast, read Appendix F.7 and the paired outcomes in Appendix F.8. The access model and common correction are in Appendices F.1 and F.4; the intervening activation and value bounds give further derivations. Appendix F.9 compares paid and free returns directly and reports harm tails and access assumptions. These policy comparisons use a one-pass free return. The separate optimised-incumbent intervention is in Appendices D.1 and D.2; it is not the same free comparator.

F.1 Restricted omitted-chip queries

There are N equally weighted chips, of which a fixed retained set $\mathcal{R}$ contains m and its complement $\mathcal{O}$ contains $M = N - m$. Let $\rho = m/N \in (0, 1)$, $f = m^{-1}\sum_{i\in\mathcal{R}} F_i$, and let Q be the exact full-group composition. Predictions and target composition vectors belong to Δ_K. The loss throughout acquisition is $R(F, Y) = m^{-1}\sum_{i\in\mathcal{R}} \|F_i - Y_i\|^2$.

One permitted query reveals the exact composition vector of one omitted chip, at one vector-query unit. All policies observe the same predictions, memberships, coordinates, Q and purchased answers. Unqueried reference targets remain with the evaluator. In particular, purchased targets never

replace predictions in the retained loss. Omitted-chip predictions are an explicit input assumption. Exact reference-vector access and equal query costs are emulated restrictions, not measured annotation costs or an operational missing-image mechanism. An exact retained-mean query, or an omitted-total query combined with Q, would identify the retained mean in one aggregate query. That different action is not assigned the price of one chip-vector query.

For purchased set $A \subseteq \mathcal{O}$, define $s_A = N^{-1} \sum_{j \in A} Y_j$, $a = |A|/N$, and $u = 1 - \rho - a$. Eliminating the remaining omitted vectors gives the exact mean set

$$T_A = \{q \in \Delta_K : Q - s_A - \rho q \geq 0\}, \qquad \mathcal{C}_A = \{Z : Z_i \in \Delta_K,\ \bar{Z} \in T_A\}. \tag{29}$$

Indeed, the nonnegative residual sums to u. If $u > 0$, assigning its normalised value to every unqueried omitted chip realises any $q \in T_A$; if $u = 0$, the residual must vanish. Lower coordinate bounds are implied by the upper bounds and the unit-sum constraint. Equivalently, the existing support formula in Equation 11 applies with

$$Q' = \frac{NQ - \sum_{j \in A} Y_j}{N - |A|}, \qquad \rho' = \frac{m}{N - |A|}.$$

These statements use the declared fractional-simplex target class and common weights, classes and reference time. No independent measurement error model is supplied.

F.2 Activation, its exact hindsight cost and complementarity

Set $t = Q - \rho f$. Substituting f into Equation 29 proves the threshold statement in Equation 31: $F \in \mathcal{C}_A$ iff $s_{Ak} \leq t_k$ for every class. Equation 3 then yields the value in Equation 30,

$$V(A) = \frac{1}{m} \|F - \Pi_{\mathcal{C}_A} F\|_F^2, \tag{30}$$

$$V(A) > 0 \iff \max_k (s_{Ak} - t_k) > 0. \tag{31}$$

This is an optimal uniform-improvement statement, independent of an iterative solver. Nested purchases make $V(A)$ nondecreasing. That fact alone does not order the realised losses of arbitrary returned corrections. For the particular common one-pass map used here, the stronger nested-cap result in Appendix A.5 does establish nondecreasing realised group gain along an exact nested trajectory.

For initially feasible F, sort the true omitted values separately in each class as $Y_{(1)k} \geq \cdots \geq Y_{(M)k}$. The exact minimum equal-cost query count over all subsets is

$$r_* = \min_k \min \left\{ r : \sum_{j=1}^{r} Y_{(j)k} > N t_k \right\}. \tag{32}$$

An empty minimum is infinity; initially infeasible predictions have cost zero. Every activating subset must cross one class threshold and cannot contain more mass in that class than its top-r prefix. Conversely, the winning prefix attains the crossing. This proves Equation 32. The computation uses hidden labels and is evaluator-only. It is neither an implementable policy nor the optimal cost for a specified positive gain. Unequal query costs do not admit this simple sorting argument.

At census, $s_{\mathcal{O}} - t = \rho(f - q_{\text{true}})$, where q_{true} is the retained reference mean. Hence $V(\mathcal{O}) > 0$ iff $f \neq q_{\text{true}}$. Large individual errors can coexist with zero census value: two retained predictions $(1/2, 1/2)$ and targets $(1, 0)$, $(0, 1)$ have matching means and loss $1/2$ each. Moreover, when $t \geq 0$, the initial information admits $Y_{\mathcal{R}} = F$ and all omitted targets equal to $t/(1-\rho)$. Thus no policy can guarantee activation uniformly over all initially compatible fractional-target worlds, even at census.

A zero plateau. Take $N = 4$, $m = 2$, both retained predictions $(1/2, 1/2)$, both retained targets $(0, 1)$ and both omitted targets $(1, 0)$. Then $Q = (1/2, 1/2)$, $t = (1/4, 1/4)$ and

$$V(\varnothing) = V(\{1\}) = V(\{2\}) = 0, \qquad V(\{1, 2\}) = 1/2.$$

Each single query only reaches a threshold; their union crosses it. This violates the submodular inequality and exhibits complementarity. More generally, if every residual allocation $t_k - s_{Ak} \geq 1/N$, every possible single-chip answer leaves F feasible. A one-step value score then has no signal, even though a batch may help. This does not establish a new nonmyopic acquisition principle or a general greedy approximation guarantee.

F.3 BOUNDS FOR A SPECIFIED POSITIVE VALUE

Write $h_k = (s_{Ak} - t_k)_+$, $H_A = \sum_k h_k$ and $J = \{k : h_k > 0\}$. If $H_A = 0$, then $V(A) = 0$. Otherwise $|J| < K$ and

$$\frac{\sum_k h_k^2 + H_A^2/(K - |J|)}{\rho^2} \le \operatorname{dist}(f, T_A)^2 \le V(A),$$
$$V(A) \le \frac{2}{\rho^2} \sum_{k \in J} \frac{h_k^2}{f_k} \le \frac{2H_A}{\rho}. \tag{33}$$

For the lower bound, every feasible mean reduces each violated coordinate by at least h_k/ρ. Its other $K - |J|$ coordinates must receive total mass at least H_A/ρ; Cauchy–Schwarz bounds their squared change. Jensen's inequality then bounds the mean squared row movement below by the squared mean movement.

For the upper bound, put $v_k = h_k/\rho$. Remove fraction v_k/f_k of each violated class from every prediction row and redistribute the removed mass among the nonviolated classes to attain a feasible mean. Such redistribution exists because the mean caps admit unit total mass. A violated class has $0 < v_k \le f_k$, so division by zero cannot occur. If row i loses $\alpha_i = \sum_{k \in J}(v_k/f_k)F_{ik}$, the squared source change is at most $\sum_{k \in J}(v_k/f_k)^2 F_{ik}$. The squared destination change is at most α_i^2, bounded by the same quantity using Cauchy–Schwarz and $\sum_{k \in J} F_{ik} \le 1$. Averaging gives a feasible squared movement at most $2\sum_{k \in J} v_k^2/f_k$. Projection minimises that movement; $v_k \le f_k$ gives the final inequality. If all retained prediction rows equal f, the mean-distance lower bound is attained exactly. In particular,

$$\frac{4H_A^2}{K\rho^2} \le V(A) \le \frac{2H_A}{\rho}, \tag{34}$$

where the lower bound also uses $1/|J| + 1/(K - |J|) \ge 4/K$ for $0 < |J| < K$.

For initially feasible F, $s_{Ak} \le a$ implies that activation requires $a > \min_k t_k$. A value $V(A) \ge \epsilon > 0$ further requires

$$a \ge \min_k t_k + \frac{\rho\epsilon}{2},$$
$$a \ge \min_{k: f_k > 0} \left[t_k + \rho\sqrt{\epsilon f_k/2} \right]. \tag{35}$$

The first term uses $H_A \le (a - \min_k t_k)_+$. For the second, maximise the convex function $\sum_{k: f_k > 0}(s_k - t_k)_+^2/f_k$ over $s \ge 0$, $\sum_k s_k = a$; its maximum occurs at a vertex ae_k. These are necessary budgets, not sufficient acquisition guarantees. For $K \ge 2$, conversely, a single excess $h_k \ge \rho\sqrt{\epsilon(K-1)/K}$ makes the mean-distance lower bound at least ϵ.

An upper bound over every subset at a fixed budget. For a class subset $B \subseteq [K]$, write $Y_j(B) = \sum_{k \in B} Y_{jk}$ and $t(B) = \sum_{k \in B} t_k$. With a separate descending sort for each B,

$$\max_{A \subseteq \mathcal{O}:\, |A| = r} H_A = \max_{B \subseteq [K]} \left\{ \frac{1}{N} \sum_{j=1}^{r} Y_{(j)}(B) - t(B) \right\} =: H_{\max}(r). \tag{36}$$

Indeed, a sum of positive parts equals the maximum sum over coordinate subsets. Interchanging the two finite maxima leaves, for each B, the sum of its largest r scalar contributions. The empty class subset supplies zero. Our five-class setting has only $2^5 = 32$ subsets. Every r-query strategy therefore has $V(A) \le 2H_{\max}(r)/\rho$. A second upper bound substitutes each class's separate top-r excess into the quadratic upper bound in Equation 33. Those coordinate maxima need not occur together, so this can be loose. We take the smaller upper bound. If it is below ϵ, no subset of that size permits optimal value ϵ. This evaluator-only calculation is an impossibility bound, not an exact useful-gain optimum.

F.4 A COMMON ONE-PASS REPAIR AND ITS PAIRED COMPARISON

At every acquisition budget, let $q_0 = \Pi_{T_A}(f)$ and return

$$\widetilde{F}_i = F_i + q_0 - f, \qquad G_i = \Pi_{\Delta_K}(\widetilde{F}_i),$$
$$L_{\text{one}} = \|q_0 - f\|^2 + \frac{1}{m}\sum_i \|\widetilde{F}_i - G_i\|^2. \tag{37}$$

Mean-set projection gives $\langle q_0 - f, q - q_0\rangle \geq 0$ for every $q \in T_A$. Expanding the affine-shift loss difference therefore gives at least $\|q_0 - f\|^2$ improvement. Each row-simplex projection adds at least $\|\widetilde{F}_i - G_i\|^2$. Thus $R(F,Y) - R(G,Y) \geq L_{\text{one}}$ for every $Y \in \mathcal{C}_A$. These are standard projection inequalities underlying established adjustment (Heiser et al., 2020). The final mean need not lie in T_A. The certificate is valid for the returned prediction but is not generally equal to $V(A)$; the single-coordinate sufficient condition above also ensures $L_{\text{one}} \geq \epsilon$. Appendix A.4 separately treats the certificates used for archived iterative returns.

The paired comparison uses all 216 BENELUX and 56 Colombia eligible groups, eight frozen configurations, and four quarter-support policies. It contains 8,704 group/model/policy pairs, including 7,616 learned-model pairs. The unchanged, one-pass, archived support-valid iterative, mismatched-full-mean and exact-subset-oracle arms give 43,520 records. No production iterative solves were rerun.

Across the 24 learned-family/region/policy summaries in Table 19, one pass retains 25.90–77.47% of iterative realised gain under equal-group weighting and 26.10–77.48% under retained-chip weighting. At the 10^{-8} comparison margin, iterative gain is larger in 4,355 pairs, 4,349 are tied and none favours one pass. The constant control accounts for 1,088 ties. This is an observed utility difference, not a dominance theorem or a timed compute–utility comparison.

Both methods have zero observed harmful group-average returns, yet one pass harms 0.77–5.45% of learned chip/seed records across the family/region/policy summaries, versus 0.95–5.57% for the iterative method. Their largest individual loss increases are 0.150663 and 0.158378. One-pass means exceed a valid-set coordinate bound by more than 10^{-9} in 4,368 of 7,616 learned records, with maximum residual 0.095253; this does not invalidate Equation 37. All 150 capped support-valid iterative returns occur in Colombia and remain included. The certificate-positivity classifications agree in 8,703 of 8,704 pairs; the sole disagreement straddles the 10^{-8} reporting margin. All 3,288 strictly interior original cases are numerically unchanged under both methods.

F.5 ACQUISITION PROTOCOL AND FURTHER RESULTS

The main acquisition study uses all eight configurations in both regions on random-mask-0 and western quarter supports: 4,352 group/model/support cases. Policies are random permutations with seeds 0, 1 and 2, spatial farthest-first, descending predictive entropy, a myopic positive-excess proxy and predicted class-threshold planning. Their 30,464 trajectories give 213,248 records at omitted-chip fractions $0, .05, .10, .25, .50, .75, 1$, rounded up to integer counts. Repeated counts retain their actual query numbers. The initial correctness pilot uses the first 12 archived groups per region, all models and western support; its 1,344 trajectories are a subset of the final design.

At each step, threshold planning sorts remaining predicted class contributions, chooses the class with the shortest predicted crossing prefix and buys its highest-contribution chip. Ties resolve by class index and a frozen seed-0 rank. When no predicted route crosses, it maximises normalised progress; after crossing, it uses the myopic score. That score is the largest positive predicted class excess after one query, with frozen random ties. It is a proxy, not an expected gain or a projection using an assumed feasible answer. Only purchased answers update the allocations. Streams are evaluated separately per predictor; costs do not represent a shared acquisition campaign.

Activation uses a 10^{-12} class-mass margin; positive returned certificates and observed harm use 10^{-8}. Neither is a validated practical-utility cutoff. Figure 14 retains every original strategy. Among initially feasible learned western cases, mean activation fractions of omitted chips are 54.14% for threshold planning versus 69.56% for random in BENELUX, and 44.31% versus 56.05% in Colombia. The corresponding clairvoyant fractions are 29.47% and 14.82%. These are means over

Table 19: One-pass / archived iterative relative loss reduction (%). Each percentage divides equally weighted mean group gain by equally weighted mean baseline loss; neural families pool three seeds. Random 0 is one archived mask, not the historical five-mask ensemble. Both methods use the same valid support information.

Region	Family	Random 0	West	East	Central
BENELUX	Ridge	1.57 / 2.88	1.88 / 3.32	2.06 / 3.44	2.18 / 3.38
	Softmax	1.02 / 1.40	1.02 / 1.46	0.99 / 1.30	1.26 / 1.63
	MLP64	0.27 / 1.04	0.55 / 1.91	0.32 / 1.05	0.35 / 0.92
Colombia	Ridge	1.22 / 2.21	0.90 / 1.58	1.26 / 2.21	1.66 / 2.82
	Softmax	1.09 / 2.25	1.08 / 2.51	1.06 / 2.46	1.07 / 2.54
	MLP64	0.25 / 0.58	0.22 / 0.54	0.26 / 0.59	0.33 / 0.80

group/model cases, with random acquisition repetitions averaged within cases; they are not universal annotation savings.

After observing that earlier activation need not give the largest realised gain, a development-informed follow-up fixed three illustrative targets $\epsilon \in \{10^{-4}, 10^{-3}, 10^{-2}\}$ without selecting a winner. It adds $\rho\sqrt{\epsilon(K-1)/K}$ to every class threshold and otherwise uses the same routing and post-crossing fallback. All 13,056 follow-up trajectories and 91,392 budget records are retained. The target is a sufficient single-class overshoot, not a claim that this route minimises the cost of achieving ϵ. Figure 15 separates the returned certificate from realised gain. Census supplies the same exact retained-mean information to every strategy. These follow-ups reuse inspected cohorts and are not independent confirmation or a tuned-model comparison.

F.6 Verification and scope

The paired one-pass comparison checked all 137 input-manifest files, 384 independent primal LPs and 192 five-variable mean-projection QPs. Its maximum applicable certificate-minus-realised-gain discrepancy was 3.34×10^{-16}. Independent mean-QP coordinate discrepancies reached 5.16×10^{-7}, with objective discrepancies at most 4.77×10^{-13}; these are finite optimiser tolerances.

Separate acquisition verification used 216 certificate LPs and 348 projection QPs, including exhaustive subsets for activation and useful-gain bounds. It reconstructed 432 complete selected query orders and 3,024 budget rows; maximum gain/certificate discrepancy was 1.90×10^{-16}. All 314,048 saved pilot, full and follow-up budget records satisfied certificate $\leq$ realised gain within 4.00×10^{-15}. These counts include overlapping pilot/full configurations and repeated groups. The final follow-up package corrects a counter's label to "before gain target" while preserving the original query orders and CSV data bodies exactly; it does not rerun the experiment.

The checks do not establish exact arithmetic, iterative optimality, independent reference accuracy, operational vector-query access or transfer to an untouched population. Every aggregate and query answer shares the existing WorldCover reference pipeline. The protected risk remains mean chip-composition loss within the fixed retained group; per-chip, per-class and pixel-level protection do not follow.

Half-budget outcomes and limits. Among initially feasible western cases, entropy's mean activation fractions are 64.0% in BENELUX and 45.2% in Colombia, compared with threshold planning's 54.1% and 44.3%. The BENELUX denominator includes five boundary-feasible cases excluded by the earlier strict-box test. Every reported learned western case eventually activates, but theory does not promise this universally.

At half budget, threshold planning exceeds entropy on BENELUX (.001825 versus .001579 realised reduction) but falls below it on Colombia (.001568 versus .001814). The three development-informed Colombia gain targets give .001782, .002394 and .002336. The middle target's advantage in this inspected comparison motivated its subsequent fixed Covertype test; this was not a prospectively validated spatial winner. At nominal quarter budget, all-subset upper bounds exclude optimal uniform gain .001 in at least 52.0% of initially feasible western BENELUX cases and 33.3% in

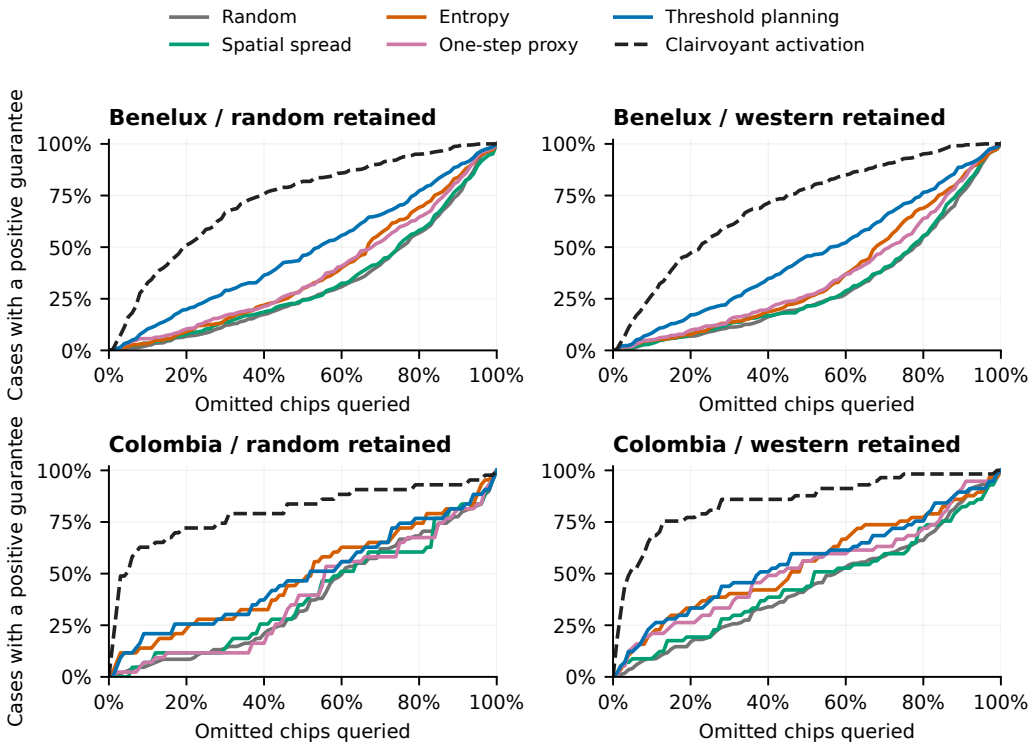


Figure 14: Activation under every original strategy on initially feasible learned group/model cases. The two BENELUX panels contain 722 and 704 cases for random and western retained supports; the Colombia panels contain 43 and 57. The dashed clairvoyant curve uses hidden targets and Equation 32. Neural seeds and repeated acquisition orders reuse physical groups. Activation concerns positive optimal uniform value at the numerical margin, not a specified useful gain.

Colombia. These are sufficient impossibility counts, not exact useful-gain optima or calibrated utility requirements.

F.7 Fixed-policy acquisition on a new data source

The Covertype follow-up fixes its data construction, fitting grid, policies and primary contrast before downloading the dataset or inspecting outcomes. This is a timestamped local protocol, not external preregistration. The routing target .001 was selected using the earlier spatial development results. All three existing targets and all comparators remain reported, regardless of their fresh results.

Source, bags and splits. Covertype (Blackard, 1998) provides 581,012 observations with 54 cartographic features and seven cover classes. Its labels are from US Forest Service records, a different source from WorldCover. Released features identify four wilderness categories but do not include coordinates. Within each category, rows are sorted by elevation then original row index and partitioned into complete 64-row bags. Bag construction discards 212 incomplete-tail rows. A seed-20260908 permutation then selects 2,048 of the 9,075 complete bags; the first 1,228 train, next 410 validate and last 410 test. No row crosses a split. These are constructed feature bags, not physical footprints or independent ecological samples; all four categories occur in every split.

Each test bag retains 16 rows. The primary `near_water` rule takes the lowest horizontal-distance-to-hydrology values, breaking ties by original row ID. A seed-0 random permutation stream defines the second support. The scoring support remains fixed throughout acquisition. Features and omitted-row predictions remain available; an omitted query returns one exact class label, represented as a seven-dimensional one-hot vector. This has a different unmeasured cost from a five-class chip-composition query.

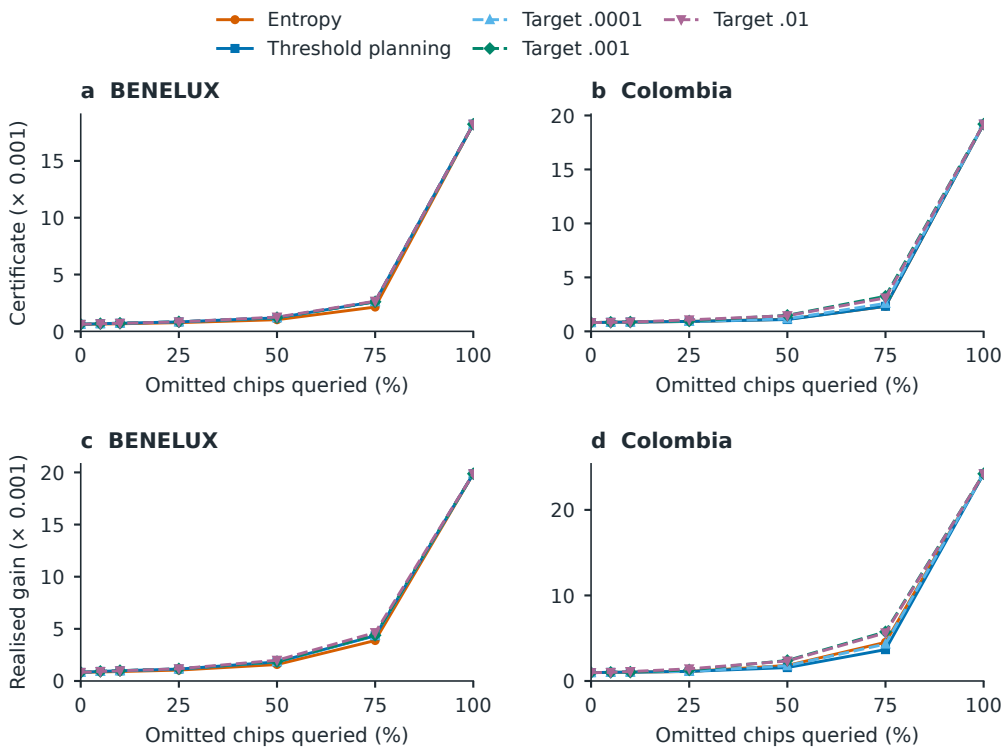


Figure 15: All three fixed gain targets on western retained supports, compared with entropy and activation-only threshold planning. Panels show the mean one-pass certificate and realised mean chip-loss reduction over all learned group/model cases (1,512 BENELUX; 392 Colombia), with equal case weights. These denominators include initially active cases and differ from Figure 14. Curves connect the seven evaluated budgets descriptively. The target grid was fixed after the original acquisition results; no target is designated a selected method.

Fitting and information boundary. All 54 features are standardised using training rows only; zero-variance scales become one. We reuse constant, simplex-decoded ridge, softmax linear and one-hidden-layer MLP64, with neural seeds 0/1/2. Ridge penalties are $.01, .1, 1, 10, 100, 1000$. Neural Adam uses learning rate .01, weight decay 0/.001 and 800 full-batch steps; candidate checkpoints are 50/100/200/400/800. Selection minimises validation bag-mean loss after individual predictions. There are eight selected configurations and 66 validation candidates. No test target score selects a parameter, checkpoint or policy. Test labels are present in the source process while constructing bag counts; the boundary is enforced non-use in fitting/selection, not physical isolation. Parameters and predictions are saved and hashed before acquisition. Independent reconstruction verifies raw-row grouping, splits, training-only scaling and prediction replay.

Policies and exact hard-label geometry. We reuse random queries (seeds 0/1/2), entropy, a one-step proxy, threshold planning and gain targets $10^{-4}, 10^{-3}, 10^{-2}$. Feature spread replaces the unavailable spatial distance with Euclidean distance over the first ten standardised continuous features; it starts at the smallest original row ID and uses that tie break subsequently. Aspect remains raw azimuth, so this comparator has a seam at $0/360$ degrees. Adaptive policies receive only predictions, parent counts and previously queried answers. Full orders are stored. There are 6,560 cases, 65,600 orders and 459,200 budget records, using 0, 3, 5, 12, 24, 36 and 48 of the 48 omitted labels.

For exact one-hot counts, write $c = NQ - \sum_{j \in A} Y_j \in \mathbb{Z}_+^7$. The retained target polytope is

$$\mathcal{C}_A = \{Y \geq 0 : Y\mathbf{1} = \mathbf{1}, \quad Y^\top \mathbf{1} \leq c\}.$$

This is a capacitated bipartite flow polytope with integral capacities. Its vertices are integral, hence one-hot, and every vertex extends to the remaining omitted rows using the residual counts. Thus

it is the exact convex hull of admissible retained hard-label assignments. The minimax identity applies to this hull; it is not an unproved replacement of integer labels by arbitrary fractions. This uses standard network-flow integrality, not a new theorem. For $K = 7$, the gain-routing overshoot is $\rho\sqrt{6\epsilon/7}$; the any-subset upper bound enumerates 128 class subsets. The quadratic upper bound retains every positive $\bar{F}_k$, including tiny values.

Primary result and its size. The primary contrast is actual one-pass gain for target .001 minus entropy at 24 queries under `near_water`. Random repetitions are first averaged within each predictor/bag case; neural seeds are averaged within family, then ridge/softmax/MLP receive equal weight, then 410 bags equally. This differs from the earlier spatial group/model-equal pooling. The fixed contrast is $+.0022812$, compared with uncorrected mean loss .5220521. Its descriptive paired-bag 95% bootstrap interval is $[.0005270, .0043049]$ (2,000 resamples, seed 20260909). Each resample carries all models and policies of a bag together; the interval does not imply independent geographic coverage.

The effect is 0.437% of original loss, 7.44% of entropy's total gain or 19.61% of its gain beyond the free correction. These denominators answer different questions. Of 410 paired bag differences, 103 are positive, 98 negative and 209 numerical ties at 10^{-10}; the median is zero. All seven learned predictor means favour target .001, but the constant control favours entropy by .0030762. The positive mean does not mean that most groups benefit from switching policies.

Activation is almost free here. Every test bag omits at least two classes: class-richness counts are 6, 224, 129, 33 and 18 bags with one through five represented classes. A positive neural prediction for an absent class makes the original prediction inadmissible without queries. Only five near-water ridge cases and one random ridge case are initially feasible; none of the 2,460 neural cases per support is feasible. Therefore this dataset supplies very little evidence about first-activation cost. Six pure bags reveal every retained label from the parent counts and all six have zero policy contrast; 203 other paired ties occur in nonpure bags.

Formal activation can also have small magnitude. For near-water MLP64, all 1,230 cases activate at zero queries, but only 963 have one-pass certificates above 10^{-8} and 261 reach .001; the median certificate is 2.77×10^{-6}. The corresponding random-support counts are 1,230, 983 and 213. These post-outcome diagnostics describe the distinction between activation and useful gain; .001 is the routing sensitivity target, not a calibrated practical threshold. We retain all cases in the primary comparison. At omitted census every trajectory activates, and policies share the same information and corrected output.

All 459,200 saved budget records have zero observed harmed group means; maximum certificate excess over actual gain is 4.44×10^{-16}. Independent checks reconstruct every first-activation oracle, replay 640 complete orders and 4,480 budget states, and solve 128 real-case target LPs plus 128 separate mean QPs. Synthetic seven-class checks add 60 full projection QPs and 60 integral target LPs. Row-level harm remains possible. These numerical checks distinguish verifying outputs from establishing broad transfer. The new dataset supports a bounded positive average useful-gain result; it does not validate the original two regions' activation-cost advantage, independently calibrated instrument data, query noise, ecological extrapolation or universal superiority.

F.8 Paired Covertype policy outcomes

These retrospective diagnostics retain all 410 near-water bags, all eight configurations and both saved 24-query policies. Indicators are formed for each configuration before averaging neural seeds, then the three learned families and bags equally. The constant comparator is separate. No model, mask, query order, target threshold or primary weighting was changed. The full-precision primary contrast remains $+.002281224215686082$. Attainment uses $\epsilon - 10^{-12}$, row signs use $\pm 10^{-8}$, bag signs use $\pm 10^{-10}$ and output equality uses maximum coordinate discrepancy 10^{-12}.

Here direct paid/free attainment compares against the one-pass free return, not the optimised incumbent in Appendix D.2; Appendix F.9.1 defines the direct bound. Baseline attainment differs by +.243902 percentage points, with 5.880759% discordant cases. Direct attainment differs by $-.460705$ points, with 12.113821% discordant cases. The latter's both/target-only/entropy-

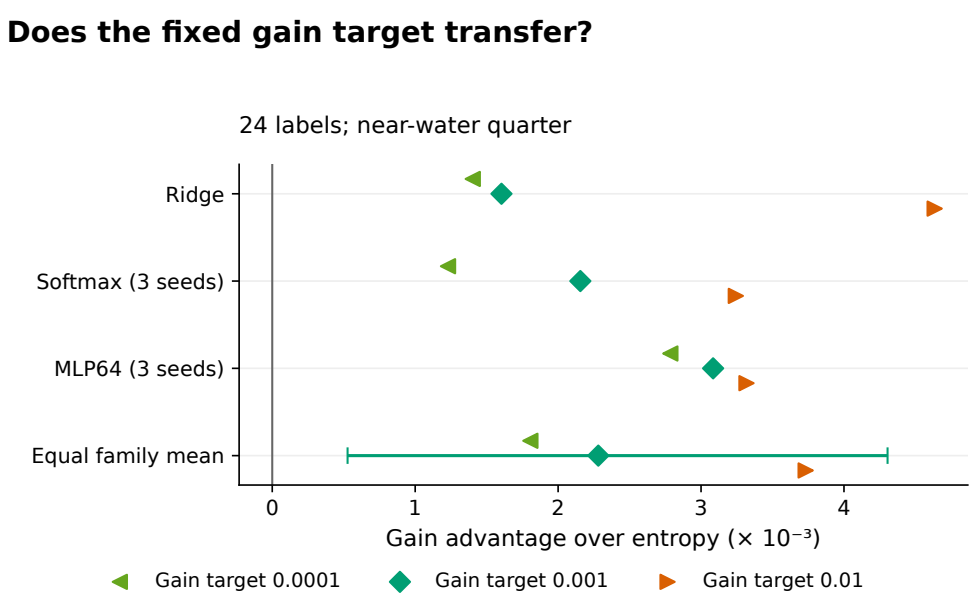


Figure 16: Fresh primary-setting contrasts by learned family, retaining all three targets. The whisker is the descriptive paired-bag 95% bootstrap interval for the fixed .001 target under equal family/bag weighting. Family markers are point summaries, not independent replications. Positive mean contrasts coexist with zero median and a negative constant control.

Table 20: Heterogeneity of the fixed target-minus-entropy contrast. Wilderness IDs use the released indicator order (shown one-based). Leave-one-category-out contrasts are descriptive recomputations, not refitted geographical holdouts.

Category	Test bags	Mean contrast	Leave this category out
Wilderness 1	176	−.000139	+.004102
Wilderness 2	23	+.005582	+.002085
Wilderness 3	192	+.004248	+.000549
Wilderness 4	19	+.000833	+.002352

only/neither conditional mean contrasts are +.007724, +.047046, −.028109 and +.000014. Mean advantage reflects magnitudes and beneficiaries, not simply coverage.

Ties are output ties in the learned models. All 209 tied primary bags have identical outputs in every learned configuration (maximum discrepancy 2.23×10^{-16}). In 189, all learned paid outputs also equal free; in 20, at least one learned output changes from free while both policies agree. Only 181 agree across all eight configurations: the constant comparator differs in 28 of these tied bags. No learned primary tie requires cancellation across nonzero seed/family risks.

The expected positive and negative parts of the target-minus-entropy row gain are .00649019749 and .00420897328. Their difference gives the primary mean, up to the explicitly retained negligible tolerance-band contribution. Even the 103 positive-primary bags contain 9.499730% harmed rows. Both paid policies harm 1.881775% of rows relative to the original predictor; target-only harms .716463% and entropy-only .409892%. Another 8.456978% of rows worsen when switching to target although neither paid output is worse than the original predictor: policy-switch harm often loses part of an existing improvement. These denominators describe model-row outcomes, not independent human subjects or geographic samples.

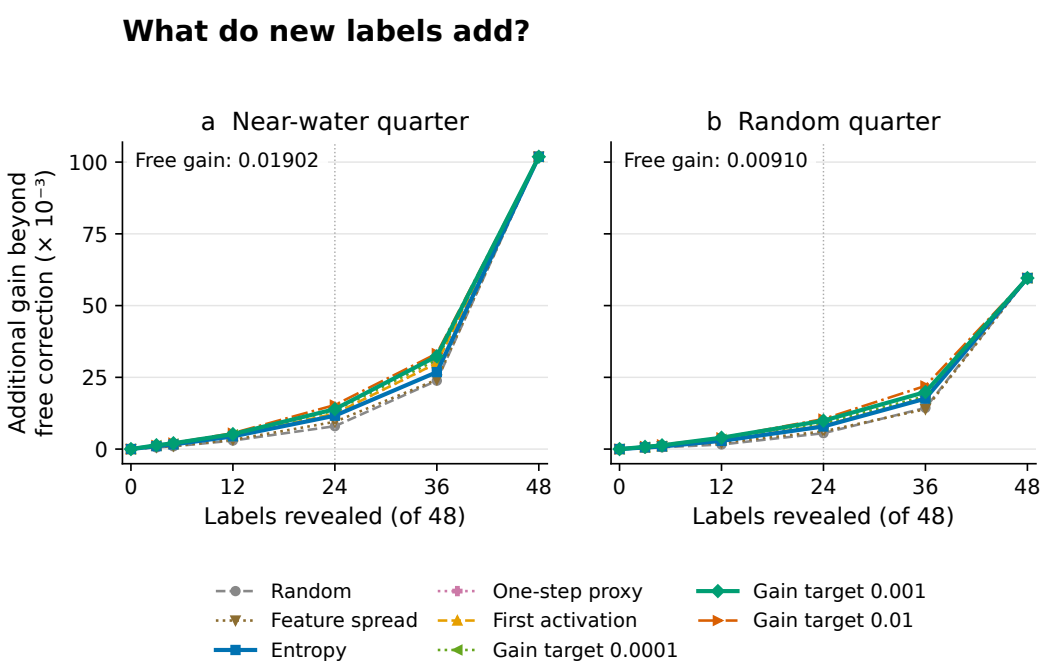


Figure 17: Additional realised gain beyond the correction available from parent counts alone. The fixed near-water test and random-support sensitivity show all policies and targets. At census, all orders yield identical information. The curves use all 410 bags with equal learned-family weights; these are class-label queries in constructed bags, not physical footprint measurements.

Table 21: Covertype: all acquisition rules at 24 of 48 queries. Entries are actual reductions in squared probability-vector loss (higher is better), with equal learned-family and bag weights. Free correction already gains .019020 on near-water and .009100 on random support. The primary target was .001; .01's larger observed gain is a retained sensitivity, not a replacement primary result.

Rule	Near-water quarter	Random quarter
Random (three orders)	.026997	.014637
Feature spread	.028525	.015268
Entropy	.030652	.016970
One-step proxy	.031353	.017317
Activation threshold	.031377	.017317
Gain target .0001	.032458	.018146
Gain target .001 (fixed primary)	.032933	.018989
Gain target .01	.034384	.019578

F.9 Direct benefit over the free return, harm tails and access assumptions

These retrospective analyses use the frozen predictions, acquisition orders, reference targets and records. The original target-.001 versus entropy primary, its weights and descriptive interval remain unchanged. They explain inspected outcomes and are not new independent tests.

F.9.1 A direct comparison with the free return

For free return G_0 and paid return G_A, the direct target is

$$W(A) = \min_{Y \in \mathcal{C}_A} \{R(G_0, Y) - R(G_A, Y)\}. \tag{38}$$

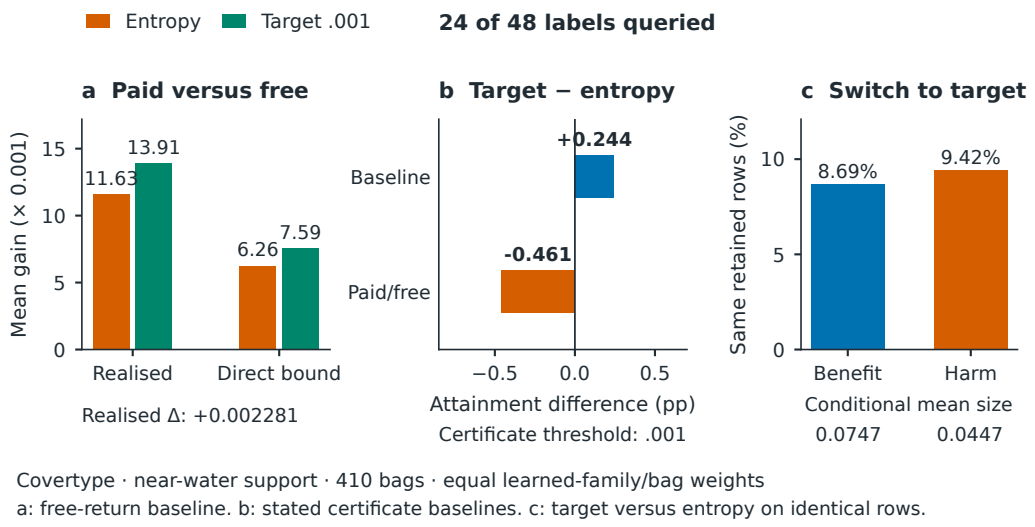


Figure 18: Identical Covertype cases and query budgets, different evaluation questions. (a) Realised and directly certified gain beyond free correction. (b) Target-minus-entropy certificate attainment at .001 reverses sign when measured incrementally. (c) Switching entropy to target benefits fewer rows than it harms, but with larger conditional magnitude. Indicators precede seed-within-family, three-family and 410-bag averaging. Losses use squared probability vectors. Paired cells appear below, and Appendix F.9 defines the direct comparison.

Table 22: Paired .001 attainment. Fractions classify configurations before weighting; contributions are unconditional parts of the primary mean.

Certificate	Joint cell	Cases (%)	Mean contribution
Baseline analytic	Both	63.468835	+.001146303
	Target only	3.062331	+.001808673
	Entropy only	2.818428	−.000681757
	Neither	30.650407	+.000008005
Direct paid/free	Both	16.802168	+.001297724
	Target only	5.826558	+.002741151
	Entropy only	6.287263	−.001767289
	Neither	71.084011	+.000009638

For $m = 16$ retained Covertype rows and r queried omitted labels, let c_k be parent class counts minus queried class counts. Then $\sum_k c_k = 64 - r \geq 16$. Feasible hard retained labels have one class per row and class totals at most c_k. Every such assignment can be completed with the remaining $48 - r$ omitted labels. Its convex hull is the corresponding capacitated transportation polytope. Therefore Equation 38 is computed by assigning retained rows to class slots with linear costs $2(G_A - G_0)_{ik}/m$, plus constant $(\|G_0\|_F^2 - \|G_A\|_F^2)/m$. This uses established assignment integrality and the affine loss identity. A lower bound for each separate comparison with F cannot be subtracted to obtain this minimum.

Independent capacitated LP checks agree with assignment endpoints within 8.33×10^{-17}. Original gain/certificate replay agrees within 3.99×10^{-16}, and direct lower endpoints exceed realised increments by at most 2.23×10^{-16}. This is numerical verification, not interval-arithmetic proof on every input. Appendix I.4 gives the pilot selection and endpoint counts.

Thus the target's larger mean increment need not give more cases exceeding a small incremental threshold. At census, both policies yield mean realised/direct increments .101861/.093495 over free correction. No direct endpoint is below -10^{-8} in the saved states; observed nonnegative endpoints do not prove a universal dominance relation for arbitrary future returns. Subtracting separate certificates exceeds the direct endpoint in 6.016% of weighted entropy cases and 5.772% of target cases at

Table 23: Same-row comparisons. Conditional sizes average among affected rows; maxima span all included model-row outcomes. Positive benefit means the first named return has lower loss.

Comparison	Benefit (%)	Harm (%)	Benefit size	Harm size	Max harm
Target vs entropy	8.687331	9.424119	.074709	.044662	1.035271
Target vs free	21.001016	2.356030	.078407	.108373	1.035271
Entropy vs free	21.510840	2.086721	.064976	.112380	.711499

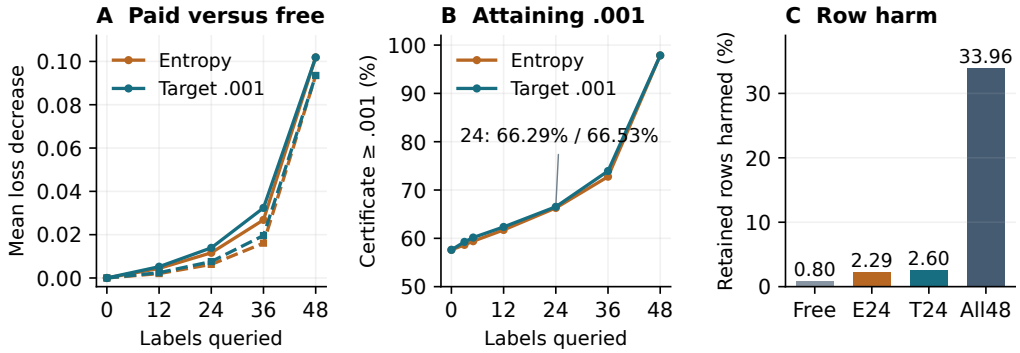


Figure 19: Full-budget and original-baseline view. (a) Realised and directly certified paid/free gains at five saved integer budgets. (b) Baseline analytic certificate attainment at .001. (c) Original-predictor row harm, which is a different comparison from the policy-switch comparison in Appendix F.8. All panels use equal learned-family/bag weights; census is an information endpoint, not a recommended operating budget.

half budget, without producing a positive-versus-negative contradiction here. Raw case counts and weighted rates are kept separate.

F.9.2 Row-harm tails

Row harm means baseline loss minus corrected loss below -10^{-8}. Conditional severity divides the weighted positive loss increase by the weighted harm rate: draw a case under its declared weights, then a retained row uniformly. We reconstructed all 459,200 Covertype states (7,347,200 row-state losses); summary discrepancies are at most 2.23×10^{-16} and harm counts agree exactly. Near-water target .001 at 24 queries has maximum single-row increase 1.03527. Random-support harms are .208% free, 1.413% entropy at 24, 1.955% target at 24 and 36.430% at census. These use the original predictor as baseline, not a direct row comparison between policies. No scalar cost trades row harm against mean gain.

F.9.3 Access assumptions and the contribution boundary

A hypothetical use case is an existing full-region reference total with imagery-derived predictions everywhere, but costly new local validation outside a fixed scoring region. If complete retained-area annotation is unavailable as one cheap action, purchasing individual omitted labels narrows the admissible retained total. This explains the restricted access model; it is not evidence that actual annotators or cloud masks have the simulated costs. Parent counts, retained membership, predictions and acquired answers suffice for adjustment, membership tests and certificates. Realised loss, the Colombia residual decomposition, clairvoyant activation and useful-value bounds require evaluator-only reference targets. A measured retained total would be a different action with different information and cost.

Table 24: At 24 near-water queries, additional improvement over the free return. All percentages retain original learned-family/bag weights; positivity uses 10^{-8}. Threshold attainment uses $\epsilon - 10^{-12}$.

Quantity	Entropy	Target .001
Mean realised increment	.011632	.013913
Mean direct lower increment	.006264	.007590
Positive direct lower increment (%)	27.073	26.694
Numerically identical output (%)	72.846	73.117
Realised increment at least .001 (%)	24.959	24.417
Direct lower increment at least .001 (%)	23.089	22.629
Direct lower increment at least .01 (%)	13.442	13.631

The empirical distinction from prior deletion/proportion-error adjustment work is the matched account of spatially selected scored support, mixed regional/model outcomes, local harms under group protection, and the separation of activation, target attainment and incremental paid-versus-free value under one declared acquisition contract. The projection and threshold algebra remain established or elementary. For model comparison, a sufficient condition already follows from Equation 10: if $G - F$ is constant within each exact-count bag, its inner product with unknown labels depends only on the supplied counts, so the interval collapses to a known contrast (which can still tie). Covertype tests useful repair acquisition, not improved model-selection identification. No noisy-observation guarantee, optimal acquisition policy or application-calibrated utility threshold is established.

G A BOUNDED STUDY WITH OVERLAPPING SPATIAL OBSERVATIONS

This follow-up changes the observation geometry while reusing the earlier spatial predictors. Its primary asks whether a fixed learned proposal outperforms a declared nonlearned reconstruction after eight finer observations. Gain from extra information and incremental-versus-joint updating are separate secondary questions. The design, selection, weights and evaluator were fixed before this study's reconstruction accuracy was inspected; the underlying test groups had already been used in earlier experiments. No models were retrained or selected using these outcomes.

G.1 INPUTS, COVERAGE AND INFORMATION ACCESS

The saved five-class WorldCover compositions, group membership, coordinates and eight prediction configurations are those of Appendix E.1: constant, ridge, and three seeds each of softmax and MLP64. Each chip has 10,000 valid reference pixels. The target is agreement with this reference product, not independently measured surface truth. Uniform means of released chips are neither physical commune totals nor instrument readings.

Select groups with 16–256 chips for which both fixed layouts pass the metadata checks below. If more than 64 remain in a region, sort by chip count and stable ID and take 64 equally spaced indices $\lfloor k(m-1)/63 \rfloor$. Of 233 BENELUX test groups, 216 have at least 16 chips, 213 meet the size cap and 205 pass both layouts; 64 are selected, containing 4,006 chips. Of 57 Colombia groups, 56 have at least 16 chips and 37 meet the size cap; all 37 pass both layouts, containing 4,528 chips. The 19 large Colombia groups are excluded before outcomes, narrowing coverage relative to the earlier omission audit. Both layouts reuse each selected group, giving 202 numerical cases and 1,616 configuration rows. They are not independent geographic replications.

For a full group S with n chips, the initial observation is its uniform mean $a_0 Y = y_0$. Each finer observation averages $w = \max(4, \lceil n/4 \rceil)$ chips. We release four and then eight pooled composition vectors, retaining all earlier observations and scoring the same full group at every budget. Individual reference vectors remain evaluator-only except for these declared aggregates. Scored chips are therefore aggregate-covered, not wholly unobserved. A composition vector has at most four independent scalar degrees of freedom; its unit cost is an abstract information budget, not a measured annotation or instrument expense.

Fixed layouts. For strips, order chips by longitude, latitude and stable ID. Eight windows begin at $\lfloor j(n-w)/7 \rfloor$, $j = 0, \ldots, 7$, and arrive in order $0, 2, 4, 6, 1, 3, 5, 7$. For neighbourhoods, the first centre is nearest the coordinate median and subsequent centres follow farthest-point selection; each observes its w nearest chips. Distances use latitude and longitude scaled by the cosine of mean latitude; ID breaks ties. This is a local coordinate approximation. A strip of released chips need not be physically contiguous.

Require eight distinct fine rows, a connected overlap graph at budget eight, and strictly increasing observation rank at budgets zero, four and eight. Connectivity at budget four is not required. Using integer incidence rows and exact rational arithmetic, reject a geometry if any single-chip coordinate vector lies in the row span of the full observation matrix. This metadata condition does not exclude pure-class answers that can identify some targets through nonnegativity. Both layouts must pass; no replacement layout is sought based on accuracy. The BENELUX metadata exclusions comprise three oversized groups and eight further geometry failures, with one of those failures satisfying more than one reason.

G.2 Reconstructions and the protected quantity

Let $M_t = [a_0; A_t]$ contain the coarse and first t finer observation rows, and define

$$\mathcal{C}_t = \{X \in \Delta_5^n : M_t X = M_t Y\}, \qquad t \in \{0, 4, 8\}. \tag{39}$$

The learner receives M_t and its answers, not the individual rows of Y. Projection and evaluation give each chip equal weight. For a frozen proposal F, define

$$B = \Pi_{\mathcal{C}_0} F, \quad J_t = \Pi_{\mathcal{C}_t} F, \quad I_0 = B, \quad I_4 = \Pi_{\mathcal{C}_4} B, \quad I_8 = \Pi_{\mathcal{C}_8} I_4. \tag{40}$$

J_t incorporates all observations around the original proposal; I_t updates the return actually available at the preceding stage. Solving distance-to-F jointly with hard observation constraints is exactly the same estimator as posthoc projection of F; we do not count them as separate methods.

The nonlearned comparator N_t minimises, over $\mathcal{C}_t$,

$$\frac{1}{|E|} \sum_{(i,j) \in E} \|X_i - X_j\|^2 + \frac{.01}{n} \sum_i \|X_i - u\|^2, \qquad u = (.2, .2, .2, .2, .2). \tag{41}$$

E is the undirected union of four-nearest-neighbour coordinate edges, with ID tie-breaking. The coefficient is fixed, without test-dependent tuning. The positive quadratic term makes the objective strictly convex. At budget zero, N_0 is the constant whole-group mean, as is B for the constant proposal. Unchanged F, B and N_0 are retained as controls. This compares useful fixed priors, not equal training expenditure or architecture capacity: the learned proposals additionally embody earlier RGB training.

Classical projection geometry protects the full-group loss of each exact incremental update relative to its preceding incumbent when $Y \in \mathcal{C}_t$. It does not establish a universal realised-loss ordering between I_t, J_t and N_t. Two exactly feasible maps X, Z satisfy $M_t(X - Z) = 0$; their differences lie in the observation nullspace. Projections onto nested affine sets compose when the relevant affine projections satisfy the simplex inequalities. Active nonnegativity constraints permit joint/incremental differences, but do not force them or determine their realised sign. These are explanatory consequences of existing geometry (Bauschke & Borwein, 1996), not new theorem claims or a causal diagnosis of the empirical deficits.

G.3 Fixed estimand and full comparisons

Loss is $L(X, Y) = n^{-1} \sum_i \sum_{k=1}^5 (X_{ik} - Y_{ik})^2$, without an additional division by five. Average neural seeds within family, then ridge/softmax/MLP64 and both layouts equally within group; average groups within region and the two regions equally. Constants remain separate. This averages evaluation results, not predictions into an ensemble.

The fixed primary is $\Delta_8 = \mathrm{mean}[L(N_8, Y) - L(J_8, Y)]$, positive when joint learned reconstruction is better. The prespecified useful-positive rule requires a pooled lower numerical endpoint at least .001 and positive lower endpoints in both regions. It fails: $\Delta_8 = -.014390297557002907$, with numerical interval $[-.01440002, -.01438058]$. Table 25 preserves the contrary ridge result and

all zero/additional-information comparisons. The pooled coarse advantage changes sign after finer observations; neither individual region reverses sign. Both layouts give the same regional signs: graph-minus-joint means are +.01319337/+.00198980 in BENELUX and -.03449286/-.03825150 in Colombia for strips/neighbourhoods, giving pooled -.01064975/-.01813085. A model chosen after inspecting these results would answer a different question.

All 101 group averages improve under both learned paths. No learned configuration has joint or incremental loss deterioration greater than 10^{-8} relative to B in the reference evaluation. This observation is not a universal joint-versus-incumbent guarantee. Incremental steps have positive direct full-group lower certificates throughout; joint J_8 versus B has positive direct certificates in 87.83% of weighted learned configurations. Direct certificates minimise the affine loss contrast over the declared feasible truth set; they are not probabilities.

Incremental-versus-coarse outcomes are 62.87% benefit, 35.08% harm and 2.05% tie under the hierarchical chip/configuration weighting. These are not proportions of distinct geographic locations. Joint beats incremental by more than 10^{-8} in 79/896 BENELUX and 16/518 Colombia learned configurations despite the opposite mean ordering. Joint beats graph in 36/64 BENELUX and 3/37 Colombia family/layout-averaged groups. The separate reference-based Colombia residual decomposition in Appendix E.4, Equation 28 concerns parent totals applied to subsets; it does not explain these reconstruction differences.

Spatial-contrast error averages $\|(X_i - X_j) - (Y_i - Y_j)\|^2$ over the fixed graph edges, then follows the same configuration hierarchy. Pooled errors are .07975148 for graph, .13393455 for joint and .12967654 for incremental reconstruction. Even in BENELUX, graph has lower contrast error (.10591506 versus .11591334 for joint) despite higher composition loss. This evaluates a declared contrast, not independently validated boundary recovery.

G.4 Numerical checks, chronology and reproduction limits

The main computations use a feasible active-set quadratic solver, observation-only rational feasibility repair and primal–dual bounds. Independent full-variable formulations check the six smallest selected groups per region under both layouts at eight additional finer observations (24 cases): ridge joint projection, its direct LP bound, and graph reconstruction. The final independent quadratic checks use Clarabel 0.11.1; they do not reuse the main solver's variable elimination or Hessian builder. Earlier numerical failures and two bounded repair stages are retained in the private historical archive. The final checker reused all main returns and bound proofs. Across the original and repeated independent checks, 72 cumulative case checks were performed; these are numerical diagnostics, not additional scientific samples.

All 202 cases pass before the unchanged evaluator runs: 8,080 projection returns, 404 shared graph returns, 8,080 direct bounds and the final 24 independent cases. Rational feasible witnesses and dual objectives give conservative optimisation gaps and solution-distance allowances; float64 return conversion is bounded separately. Projection acceptance requires normalised gap at most 10^{-9}, graph gap at most 10^{-11}, per-return loss allowance below 10^{-4} and residual at most 10^{-8}. The allowance is propagated through the declared averaging hierarchy; its primary interval is not statistical uncertainty. Exact rational aggregation independently reproduces the primary. All direct bounds are checked against the exact saved-reference losses after numerical validity is established. This distinct arithmetic does not retroactively strengthen the earlier experiments' checks.

The supplement provides the plotted values for Figure 20. The private historical bridge archive retains separated inputs, returned arrays, saved certificates, rational proof payloads and the fixed evaluator with portable paths; these detailed fixtures are not included in the supplement. A completed fresh extraction outside the original files reproduced every per-configuration rational loss and the exact pooled primary for all 202 cases, and checks all 8,080 saved direct bounds against realised contrasts. The calculation body is unchanged; its path/metadata adaptation is documented. That initial replay audits recorded numerical gates and payload hashes without re-executing all witness and dual arithmetic. A separate proof verifier now checks all 16,564 feasible rational witnesses, 8,484 reconstruction certificates and 8,080 direct certificates. It exactly reconstructs the 8,080 projection gaps and proves the 404 historical graph bounds by a rational inverse-residual bound. Graph inverse witnesses retained in the private historical archive permit an arithmetic-only replay without optimisation or linear solves; all 202 cases pass, and five deliberate mathematical corruptions are rejected.

Table 25: Fixed overlapping-observation comparisons. Losses sum five class errors before averaging chips. Contrasts are differences of losses, not accuracy percentages. Positive graph-minus-joint contrasts favour the learned proposal. Constants do not enter the primary.

Quantity	BENELUX	Colombia	Equal-region mean
Original proposal F	0.09087370	0.12728538	0.10907954
Coarse incumbent B	0.07857942	0.10141236	0.08999589
Graph N_0	0.12102736	0.07890476	0.09996606
Joint J_4	0.06583426	0.08657602	0.07620514
Incremental I_4	0.06533268	0.08460326	0.07496797
Graph N_4	0.08329235	0.05084882	0.06707059
Joint J_8	0.05638912	0.07829575	0.06734244
Incremental I_8	0.05553103	0.07515024	0.06534063
Graph N_8	0.06398071	0.04192357	0.05295214
Coarse advantage $L(N_0) - L(B)$	0.04244794	-0.02250760	0.00997017
Primary $L(N_8) - L(J_8)$	0.00759159	-0.03637218	-0.01439030
Additional gain $L(B) - L(J_8)$	0.02219029	0.02311661	0.02265345
Additional gain $L(B) - L(I_8)$	0.02304839	0.02626212	0.02465525
Incremental advantage $L(J_8) - L(I_8)$	0.00085809	0.00314551	0.00200180
ridge $L(N_8) - L(J_8)$	+0.02011012	+0.00184464	+0.01097738
softmax $L(N_8) - L(J_8)$	+0.00872319	-0.07118178	-0.03122929
mlp64 $L(N_8) - L(J_8)$	-0.00605856	-0.03977940	-0.02291898
constant $L(N_8) - L(J_8)$	-0.00668575	-0.00594924	-0.00631749

Independent saved arrays also receive feasibility, objective and agreement checks. Solver-internal dual diagnostics cannot be replayed because their dual vectors were not exported; no independent solver campaign is repeated. Checking a bound against one realised target remains weaker than the separate compatible-target proof. No field validation, noisy-observation safety or transfer to other data is established.

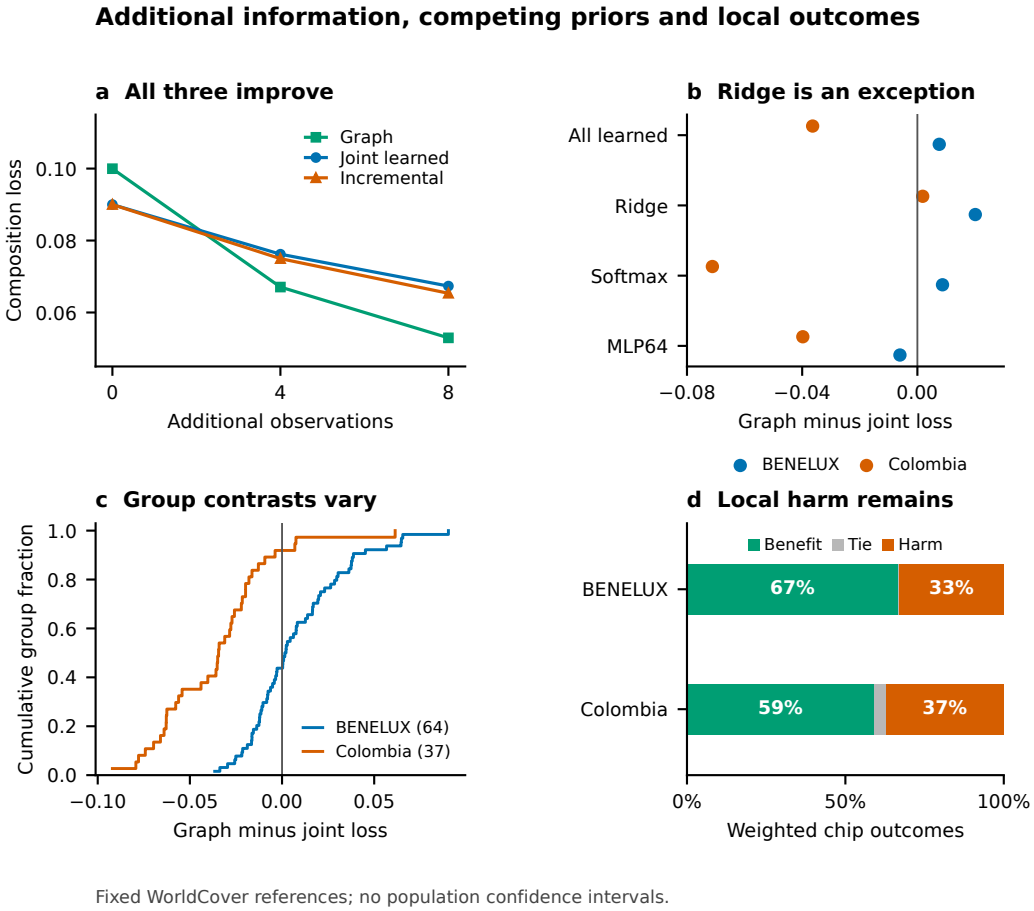


Figure 20: The same finer observations improve reconstruction without establishing general learned-prior superiority. (a) Loss at cumulative budgets zero, four and eight, retaining the coarse observation. (b) Region/family comparisons at eight; positive favours joint learned reconstruction. (c) Group contrasts after averaging families and layouts, retaining all groups. (d) Weighted chip benefit, tie and harm for incremental versus coarse reconstruction; sign margin 10^{-8}. Weights follow the declared hierarchy, not unique-chip counts. No statistical confidence intervals are shown.

H Exploratory image controls and count uncertainty

H.1 Design and provenance

We reuse cached frozen ResNet-18 features (He et al., 2016) for CIFAR-10 (Krizhevsky, 2009) and a PCA64 transform fitted on the 50,000 training features. The cache is described by its original extraction code as ImageNet-pretrained (Deng et al., 2009); we verify the cache hash but do not independently re-extract it. A fixed permutation divides the training set into 19,968 fitting instances and 9,984 aggregate-validation instances. Each pool is partitioned independently into random bags of size 16 or 64 for three seeds. All four families share these pools and partitions.

The four matched families use fixed aggregate-only selectors; their grids, solver checks and independent replay are specified below. Three disjoint target cohorts each contain 3,072 instances assembled from the previous audit's three phases. They are paired across training bag sizes and reused across families. The remaining 784 test instances are excluded. These are not historically untouched test data: earlier neural experiments already used test labels for selection/reporting. Consequently, this is an exploratory extension on one dataset, not external confirmation.

For each fixed candidate set, we evaluate random, PC1-sorted and PC2-sorted target bags of sizes 16, 64 and 128. Random aggregation has 32 partitions per cell. We retain all within-family selector pairs, including duplicate selections, and all six cross-family pairs using the KL selector. A practical-gap stratum, declared before fitting, requires an absolute Brier difference of at least .01. Partitions and reused cohorts are not counted as independent datasets.

H.2 Fitting and aggregate-only selection

The families are clipped and normalised ridge, softmax linear, a two-hidden-layer ReLU MLP of width 128, and constant logits on 64 training-feature clusters. Ridge uses $\lambda_j = 10^{-8+0.2j}$ for $j = 0, \ldots, 80$, with squared bag-residual loss averaged over fitting bags and all 65 coefficients penalised, including the intercept. Its decoder clips to $[0, 1]$, normalises each row and uses the fitting-pool class prior for zero-sum rows. Neural models minimise mean bag KL using full-batch Adam with learning rates .01, .001 and .1 for softmax, MLP128 and cluster64, respectively. Weight decay is 0 or .01; selection considers updates $0, 20, \ldots, 400$ independently by validation bag KL, bag Brier and corrected Brier. Test labels never enter these new selectors. The MLP is a bounded CPU variant of the original architecture, not an exact reproduction of the historical neural study.

The split permutation uses RNG seed 314159. Fitting and validation bag seeds are $7000 + j$ and $8000 + j$, $j = 0, 1, 2$. K-means++ uses 64 clusters, 30 iterations and seed 271828 on fitting features only. PCA coordinates share one scalar divisor, the largest training coordinate standard deviation plus 10^{-12}; they are not individually whitened. Unlike CIFAR ridge, spatial ridge leaves the intercept unpenalised.

H.3 Broader models narrow the ranking claim

The original ridge audit reproduced six historical paths, then compared consistently decoded fixed candidates. In its final sequential phase, PC1-sorted bags reversed the ranking in all 12 primary configurations; 11 remained reversed after mean-bias correction. Corrected random partitions reversed 10 of 384 trials. The underlying individual-risk advantages were only .36–2.42% relative. These results concern two nearby ridge selectors, not all useful predictors.

The matched broader experiment materially changes the interpretation (Table 26). PC1-sorted bags preserve all 108 cross-family rankings, including all 105 practical-gap comparisons. The correction changes none of those decisions. At bag size 128, PC2 reverses one of 36 cross-family comparisons. Random raw rankings reverse in 1, 20 and 51 of 1,152 trials for sizes 16, 64 and 128; corrected counts are 1, 25 and 56. Thus feature ordering is not uniformly more dangerous than random grouping.

Twelve within-family selector comparisons remain non-identical; all have Brier gaps below .01. PC1 raw rankings are wrong in 4/12, 4/12 and 6/12 at sizes 16, 64 and 128. This retains a close-choice sensitivity finding while rejecting an unrestricted cross-family failure claim. Some selector comparisons share exactly the same endpoints, so these fractions describe the declared comparison list rather than independent effects.

H.4 Ambiguous comparisons coexist with certified repair

For exact identification we use one random partition and both feature-ordered partitions at each size. All 324 cross-family comparisons have intervals spanning zero. This is compatible with the accurate observed rankings above: the intervals allow arbitrary within-bag class assignments, whereas the realised images contain semantic structure.

Using the same target counts, we repair each of the four KL-selected predictions across the six paired cells and nine target geometries. All 216 returned predictions have strictly negative upper endpoints for their Brier difference relative to the original prediction. Observed relative reductions range from 1.08% to 22.28%, with median 4.71%; absolute decreases range from .00420 to .07414. The smallest worst-case certified decrease is .00390. Mean relative improvements at bag size 16 are 12.19%, 17.30%, 17.97% and 14.10% for ridge, softmax linear, MLP128 and cluster64 respectively.

The full iterative adjustment improves over the one-pass affine/simplex control in every tested case, by .00915 Brier on average. This is an observed comparison, not a general dominance theorem

Table 26: Individual performance of KL-selected models. Each row averages six paired training-size/cohort cells. Ranges describe these cells; they are not confidence intervals.

Family	Mean Brier	Brier range	Mean accuracy
Ridge	.4595	.4300–.4877	79.73%
Softmax linear	.2496	.2177–.2823	82.92%
MLP128	.2317	.2076–.2594	83.97%
Cluster64	.4265	.4081–.4555	70.41%

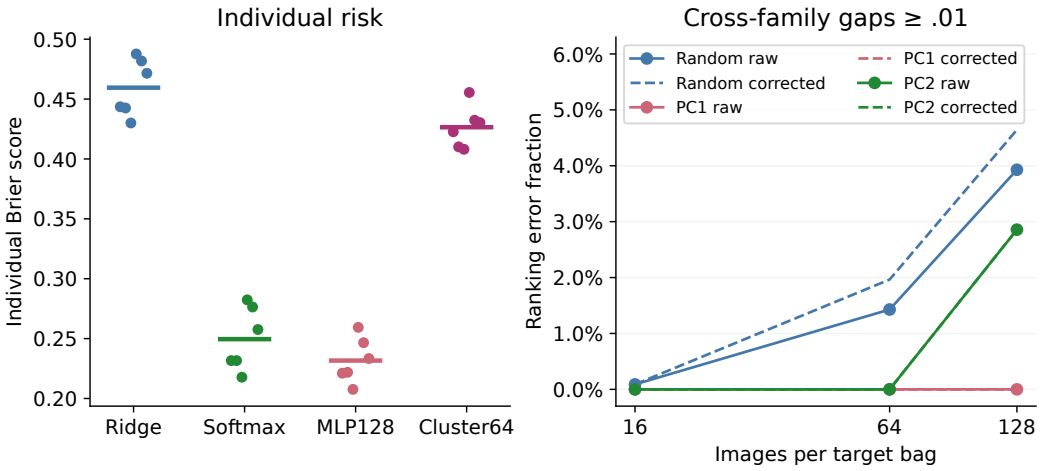


Figure 21: Matched broader-model evidence. Left: individual Brier for KL-selected predictors, with a mean line and six paired cells per family. Right: cross-family ranking errors restricted to predeclared practical gaps; dashed curves apply the random-partition correction. Repeated random partitions do not constitute independent datasets.

between the two procedures. Bag-only constant predictions are retained as a further control in the private historical archive.

H.5 Sensitivity to uncertain aggregate counts

We reuse all 24 PC1/bag-64 configurations: four model families, two training bag sizes and three existing cohorts. Let N_{bk} be the nominal count of class k in bag b. For allowance $r \in \{0, 1, 2, 4, 8\}$, replace each count equality by the integer interval $[\max(0, N_{bk} - r), \min(64, N_{bk} + r)]$, retaining row simplices and bag totals. An allowance of one bounds each class count separately; it is not a budget of one mislabelled instance per bag. Nominal counts derive from the evaluation labels, which remain fixed and feasible at every allowance.

For each enlarged set we independently certify both the saved exact-count adjustment and a new classical interval projection. At $r = 1$, corresponding to ± 1.5625 percentage points per class, only 6/24 saved adjustments retain a strictly positive worst-case Brier decrease. The interval projection retains one in 24/24, with median realised relative improvement 2.648%, compared with 4.955% for the saved adjustment. At $r = 2$ and $r = 4$, the interval projection still has positive certificates in all 24 configurations, while median realised gains fall to 1.240% and .351%. At $r = 8$, only 12/24 interval projections retain strictly positive certificates. Positivity requires a margin above 10^{-8}.

These are related configurations, not independent datasets. The saved exact adjustment continues to improve realised Brier throughout: its predictions and the actual labels never change. A negative worst-case decrease supplies an adverse alternative assignment permitted by the enlarged set. It does not demonstrate actual harm from faulty counts. The allowances are sensitivity parameters, with no calibrated confidence level or claim of physical plausibility for every assignment.

a Positive worst-case certificates (out of 24)

Allowance (percentage points)	0	1.56	3.13	6.25	12.5
Keep exact-count adjustment	**24**	**6**	**0**	**0**	**0**
Use interval projection	**24**	**24**	**24**	**24**	**12**

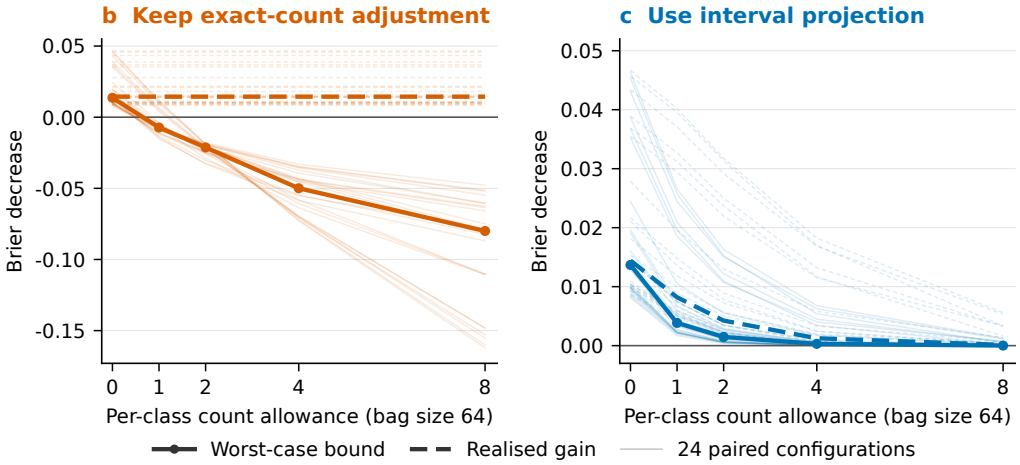


Figure 22: Certificate sensitivity across the 24 fixed PC1/bag-64 configurations. The table reports positive certificates out of 24. Faint curves retain all paired configurations and bold curves give medians; solid curves show worst-case improvement and dashed curves realised improvement on unchanged labels. The two lower panels have different vertical scales. Enlarging the information set can invalidate a saved adjustment's guarantee without changing its actual performance. Interval projection preserves smaller guarantees. Connecting lines interpolate the five evaluated allowances descriptively.

H.6 Numerical checks and finite returned predictions

Only 168/216 iterative solves satisfy the strict 10^{-9} stopping tolerance within 1,500 iterations. The maximum residual stopping error is 5.85×10^{-8}. We preserve the other 48 as incomplete solves and certify their returned predictions through assignment bounds; we do not relabel them exact projections. Eight independent small quadratic-program checks match projected values within 1.79×10^{-9}, and exhaustive feasible-label enumeration verifies assignment endpoints within 9.37×10^{-17}. Independent NumPy replay of all 72 selected prediction arrays differs from the saved PyTorch/NumPy outputs by at most 1.50×10^{-6}.

Figure 22 shows all paired configurations. All 240 endpoint witnesses independently reproduce their direct squared-loss differences within 4.30×10^{-16}. Of 120 interval solves, 119 meet the strict stopping criterion; one reaches the 2,000-iteration cap and retains an independently positive certificate. Zero-allowance results reproduce the prior exact-count analysis. The worst-case optimisation is over compatible labels without image semantics, so the same conservatism applies as in exact comparison.

I Study chronology and reproducibility scope

I.1 Result-to-artifact map

The supplement has two scripts:

```
python summarise.py
python rerun.py --out my_rerun
python summarise.py --results my_rerun
```

`summarise.py` recomputes the reported confirmation values (Table 2, the numbers in Section 6, the unused-parent rows of Table 7, and Tables 9, 10 and 11) from the saved per-case results, and compares each with the value printed here. All 217 values agree. `rerun.py` starts from the supplied frozen predictions, parent class counts and group memberships. It redraws the original samples, applies the recorded replacements, builds the incumbent and computes all five corrections for the 61,440 cases, then compares them with the saved results. On our test machine it reproduces the frozen and incumbent losses exactly, and every case except two has the same status and a gain within 10^{-12}. Both exceptions are full joint projections for one group in the secondary regime $m = 32$, $r = 4$. In one, the gain differs by 8×10^{-10} with the same status. In the other (PVC), the projection fails the solver's numerical checks and returns the unchanged prediction, which the method specifies as its fallback, whereas the original run updated it. As a result, 207 of the 217 values agree after the rerun, including every primary value. The other ten all belong to that secondary cell. Both scripts were run from a fresh extraction of the submitted ZIP with Python 3.10.9, NumPy 1.26.4 and SciPy 1.13.1 on macOS (Apple silicon). `README.md` describes every file, `docs/DATA.md` describes the data and its licence, and `paper_data/` holds the plotted values for every figure and the sources of the tabulated results. Appendix I.4 gives the counts of the completed local checks cited in this appendix.

Table 27 maps each main-text result to its appendix details and supplement location. Paths are relative to the ZIP root.

Table 27: Where to find each main-text result. "Rerun" means that `rerun.py` recomputes the result from the supplied inputs and `summarise.py` recomputes the reported values. "Data" means that the supplement holds the plotted or tabulated values, but not the detailed replay fixtures.

Paper item	Details	Supplement location	Access
Proposition 1	Proof, App. B.1	`method/joint_adjustment.py`	Code
Proposition 2; Eq. 9	Proofs, App. B.2	`method/short_corrections.py`	Code
Uniform-gain guarantee	Proof, App. A.4	Not applicable	Proof only
Figure 1	Constructed example	`paper_data/figure_01/`	Data
Table 2; confirmation in Section 6	App. C.3	`results/corrections.csv.gz`	Rerun
Figure 2	App. C.3	`paper_data/figure_02/`; `results/`	Rerun and data
Short corrections and matched losses in Section 6	Apps. C.2, C.4, C.6	`results/corrections.csv.gz`; `summaries/matched_timing/`	Rerun; timing summary only
Figure 3	App. D.2	`paper_data/figure_03/`	Data
Appendix figures 4–22	Apps. C–H	`paper_data/figure_NN/`, numbered as in the paper	Data
Appendix tables built from saved records	Apps. C, C.6	`paper_data/tables/`	Data
Numerical sensitivity	App. I.3	`summaries/numerical_sensitivity/`	Data

Table 28 separates what the supplement lets a reader rerun from completed local checks of earlier studies. The larger historical archive is retained privately and is not part of this supplement. Its detailed fixtures are therefore unavailable for reader replay, and the supplied summaries do not replace them. Neither script repeats model training.

Table 28: Evidence access and completed check scope. Historical checks listed here do not imply that their detailed fixtures are supplied. Record counts are saved comparisons, not independent datasets; exact scalar proofs and saved-output arithmetic are distinct.

Main result	Supplied access	Completed check and scope
Coordinatewise null	Summaries; detailed fixtures private	184,320 saved losses; both zero primaries; no training
Original joint follow-up	Summaries; detailed fixtures private	29,184 records, 375 exact candidate bounds
Unused-parent confirmation	Inputs, per-case results and code / `rerun.py`, `summarise.py`	61,440 cases rerun; 24,576 full/count records and 309 candidate bounds checked locally
Short-cut comparisons	Unused-parent cases supplied; earlier stages private	All four stages checked locally: 134,400 records; 2,192 exact accepted-step bounds
Incumbent comparison	Summaries; detailed fixtures private	6,560 finite returns, 104,960 row losses; solver diagnostics are hash-checked
Spatial/acquisition	Summaries; detailed fixtures private	Saved-input substage checks; separate historical optimiser sensitivity
Bridge negative primary	Summaries; detailed fixtures private	202 cases; saved witnesses and certificate arithmetic checked locally; no new optimisation
Matched timing follow-up	Timing summaries; endpoint fixtures private	65,985 endpoints; exact scalar proofs and reporting linkage; summaries supplied
Figure and table values	`paper_data/`	Values from which the figures and tables were produced; no rendering scripts

I.2 Populations and decision chronology

Table 29: Matched evidence, with distinct populations and denominators. Repeated seeds, policies, layouts and horizons are dependent settings. Extra records are not extra independent datasets.

Question	Data and scored unit	Choice changed / held fixed
Support validity	BENELUX 216 / Colombia 56 groups; fractional chip targets	Retained geography / parent totals, fitted predictions and correction
Selection utility	Covertype; 256 original + 256 unused parents, 16/32 scored rows	Logged selection and count bounds / coupled masks, PVC and PM predictions
Acquisition utility	Covertype; 410 bags, 16 scored rows	Query rule / 24-label budget, free correction and repair
Incumbent baseline	Same 410 acquisition bags	Starting prediction and optimisation / queried labels, rows and weights
Reconstruction prior	BENELUX 64 / Colombia 37 groups; full chips	Learned proposal or graph / same finer observations and scored population

Table 30: When decisions were fixed and what each stage can establish.

Stage	Fixed or inspected information	Evidential role
CIFAR/spatial development	Historical test reuse; BENELUX precedes Colombia; later omission and gain-target diagnostics	Exploratory audit and same-pipeline extension
Covertype fixed follow-up	Target .001, near-water, 24 queries and primary weights fixed before download; construction/label-access caveats retained	New-source test of that fixed contrast, not untouched policy development overall
Retrospective acquisition diagnostics	Reused all saved policies and outcomes	Direct paid/free comparison and harm tails; explanation only
Paired/computation diagnostics	Fixed same 410 bags and outputs; new paired classifications, sharp certificates and stronger-free solves	Explanation and comparator sensitivity; no new primary, training or routing
Incumbent-start intervention	Same predictions, queried labels, scored rows and weights	Retrospective reference-change test; no rerouting or new primary
Overlapping-observation follow-up	Same frozen spatial predictors; metadata-fixed groups/layouts; accuracy evaluated after numerical gates	Controlled reconstruction comparison on reused reference products; negative primary retained
Contemporary-learner support diagnostic	New disjoint Covertype rows, fixed support/primary before fitting; duration sensitivity before test outcomes; one checker repair	Objective adaptation and confidence-qualified support audit; previous primaries retained
Joint-membership follow-up	Same inspected count-study parents and predictions; comparison and five extra draws fixed before correction outcomes	Sparse gains beyond coordinatewise/count summaries; no new population
Selection-record audit and confirmation	Cached mechanism audit; fixed one/three-cut methods; unused-parent protocol before new outcomes	Separate information from computation; same-source confirmation, with old primaries retained

What the checks establish. The supplied scripts rerun and summarise the unused-parent confirmation. The following paragraphs describe completed local checks, including historical components whose detailed fixtures are not supplied. These checks reproduce saved-output arithmetic, not raw-data preparation, model training or hardware timing. Appendix I.4 gives their counts.

The independent 400-state audit reconstructs corrections on five metadata-selected bags and solves an expanded LP over all unqueried rows. Seventeen of its 80 states at 24 queries have positive direct bounds, unlike the earlier four-bag pilot's unchanged outputs. The incumbent extension checks assignment bounds and primal optimisation states; its saved finite returns define the reported comparisons. Regeneration can change tolerance-based equality and row-harm indicators despite nearly identical mean gains, as quantified below.

The bridge reader checks fixed evaluation arithmetic, numerical-gate summaries, saved feasible witnesses, direct certificates and error allowances. It does not rerun QP/LP solvers or recover unexported independent-solver internal diagnostics (Appendix G). Count-learning checks cover both horizons and numerical independent-check records; they do not retrain learners or certify every floating-point solve exactly.

Membership and short-correction readers distinguish numerical membership/loss replay from exact scalar certification. Rejected candidates remain distinct from unchanged outputs. Cases lacking dual payloads receive narrower arithmetic checks. Tiny synthetic fixtures are engineering checks, not reproduction of a scientific primary. The fresh reader retains all 256 parents, two draws and matched seeds; these repeated observations are averaged within parent.

Historical timed comparator outputs were not retained, so their endpoint identity is unavailable. The separate matched-timing follow-up retains and verifies endpoints (Appendix C.6); its replay still does not repeat hardware timings.

I.3 Measured historical optimiser sensitivity

The historical equality flag compares paid optimised returns starting from B and F at maximum coordinate difference $\leq 10^{-12}$: neither bitwise equality nor a no-change test relative to B. Table 31 compares two retained regenerations without new solves. Means weight learned families and bags equally after within-family seed averaging; raw counts use 2,870 learned model/bag records per policy. Constants remain separate.

Table 31: Existing incumbent regeneration sensitivity. Equality rates use original weights; flip counts are unweighted. Endpoint drift compares each incumbent-centred return with its saved original, not the two methods used by the equality test. Mean-gain drift is versus B.

Attempt	Policy	Equality fraction	Flips	Max. drift	Mean-gain drift
Original	Entropy	.833062	—	—	—
Initial regeneration	Entropy	.831436	64	$3.958\,10^{-9}$	$-2.825\,10^{-13}$
Exact-input repeat	Entropy	.832249	11	$2.897\,10^{-9}$	$-1.015\,10^{-12}$
Original	Gain target	.821951	—	—	—
Initial regeneration	Gain target	.822222	67	$3.958\,10^{-9}$	$+4.848\,10^{-13}$
Exact-input repeat	Gain target	.823035	12	$3.526\,10^{-9}$	$-1.204\,10^{-14}$

All 21 original attainment indicators remain unchanged for both attempts, policies, learned models and constants (137,760 flags per attempt). Incumbent-centred coordinate drift stays below 10^{-8}; this does not revise the equality tolerance. Row-benefit/harm fractions at the existing 10^{-8} sign threshold can change: the initial entropy row-harm fraction shifts by -0.0000677507 under learned-family weights; its exact-input repeat is unchanged. Small gain drift does not imply identical outputs or thresholded indicators.

The initial attempt's original-start input archive is unavailable at its recorded path: its equality flags use retained records, without independent reconstruction of both endpoints. Original and exact-input-repeat flags were independently rederived; both attempts retain incumbent-centred arrays for direct drift measurement. This historical comparator sensitivity does not assess the 253 zero-gain fresh parents or 209 acquisition-policy ties, whose predicates and components differ.

I.4 Counts of completed local checks

This subsection records the counts of completed local checks cited elsewhere in the appendix. Most concern studies whose detailed inputs are not included in the supplement. None is an additional experiment.

Finite-return certificates (Appendix A). An exact rational check covers 240 generated cases, 1,147 compatible hard-label assignments and 170 changed projected means. Eighty additional fractional cases use separate LP/QP checks, with maximum numerical bound excess 2.09×10^{-16}. A second implementation enumerates active faces instead of reusing the threshold code. It checks 120 rational cases, 120 fractional LPs and 1,028 compatible hard-label assignments, with maximum bound excess 2.78×10^{-17}. Some arbitrary real-cap examples have no hard assignment, so only their fractional statement is tested.

Joint-information follow-up (Appendix C). All 29,184 actual-incumbent records and 100 baseline-computation sensitivities pass saved-output loss replay, together with 9,216 coverage records. The reader recomputes 375 exact candidate certificates, including 138 rejected candidates whose actual return is unchanged, and rejects three deliberately corrupted payloads. Sensitivity cases without saved dual payloads receive loss and realised-bound replay only. The readers take 6.10, 1.00 and 2.44 seconds for the original 160-epoch, 640-epoch and further-draw stages, which contain 18,432, 3,072 and 7,680 records with 290, 71 and 14 exact bounds.

Certificate sharpness and computation (Appendix D.1). Independent primal SLSQP checks all 120 information states on five fixed metadata-selected groups, from a separate feasible starting

point. Independently formulated, objective-scaled LPs check 560 assignment endpoints. An initial LP-tolerance failure and a full run with 28 gaps above target are retained. A gap-triggered polish corrected those states in at most three sweeps, and no state was removed.

Direct comparison with the free return (Appendix F.9.1). All eight frozen configurations, 410 test bags, the entropy and target-.001 policies and query counts 0, 12, 24, 36 and 48 yield 32,800 direct endpoints. The first four bags in source order were fixed for a pilot before any new summaries were computed. On those bags, all eight configurations, both policies and budgets 0, 24 and 48 give 192 independent capacitated LP checks.

Acquisition and incumbent audits (Appendices F and D.2). An independent 400-state audit uses metadata-selected bag indices 0, 37, 101, 205 and 409, every configuration, and both policies at 0, 12, 24, 36 and 48 queries. At 24 queries, 17 of its 80 states have positive direct bounds. The incumbent extension checks 240 assignment bounds and 80 primal optimisation states on the same five bags.

Selection-record comparisons (Appendix C). The four five-method comparisons contain 46,080 original-horizon, 7,680 longer-horizon, 19,200 further-draw and 61,440 unused-parent records. Of their nontrivial scalar bounds, 756, 168, 56 and 1,212 respectively receive exact rational checks.